NOVA IMS Information Management School

DOCTORATE PROGRAM

Doctorate in Information Management

Specialization in Decision Support Systems

Modelling Decisions in Banking Supervision
A Machine Learning Approach

Pedro Arteaga Guerra

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor in Information Management, under the supervision of Prof. Mauro Castelli (Supervisor) and Prof. Nadine Côrte-Real (Co-supervisor)

March, 2022

Information Management School
NOVA University Lisbon

Acknowledgements

To my supervisor Prof. Mauro Castelli for his invaluable support, guidance and experience. To a great sparring partner.

Also to my co-supervisor and colleague Prof. Nadine Côrte-Real for her timely insights and critical business views.

To my wonderful parents, who grew into my closest friends, for teaching me when to walk and when to fly.

To my beloved and magnificent wife, with whom I share the greatest joys, for being key to my successes and my haven in desperate times. She stood by me through all my travails, my absences, my fits of pique and impatience. She gave me support and help, discussed ideas and prevented several wrong turns.

To my children, the courageous, adorable and rebellious Miguel and Tomás, who teach me every day to be patient and understanding. From the very beginning, with arms wide open, I hope to live up to them and watch them do better than me.

To my friends and family, *la crème de la crème*, the very very few that prevail laughing with me.

A special thank you to my most understanding bosses, João Pedro Gomes and Luís Costa Ferreira, who gave me the luxury of time and provided the means to see this project through.

Publications

Machine Learning Applied to Banking Supervision: a Literature Review
Pedro Guerra and Mauro Castelli
Risks, 2021, 9, no. 7: 136
https://doi.org/10.3390/risks9070136

Machine learning for liquidity risk modelling: A supervisory perspective
Pedro Guerra, Mauro Castelli, Nadine Côrte-Real
Economic Analysis and Policy, 2022, Volume 74, Pages 175-187
https://doi.org/10.1016/j.eap.2022.02.001

Approaching European Supervisory Risk Assessment with SupTech: A Proposal of an Early Warning System
Pedro Guerra, Mauro Castelli, Nadine Côrte-Real
Risks, 2022, 10, no. 4: 71
https://doi.org/10.3390/risks10040071

Begin at the beginning, the King said gravely, "and go on till you come to the end: then stop."

—Lewis Carroll, *Alice in Wonderland*

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

—Sir Arthur Conan Doyle, *Sherlock Holmes*

Contents

Chapter 1

Introduction

Machine learning (ML) has revolutionised data analysis over the past decade. Like innumerous other industries heavily reliant on accurate information, banking supervision stands to benefit greatly from this technological advance. However, there is an increasing tendency of retrieving data in every context, hoping that it might support or reveal an unseen pattern. Regulatory data is no exception. Even though this data obeys a strict taxonomy, the total amount of data makes it impossible to go through it using a case by case approach.

National Central Banks (NCBs) have the complex mission to ensure the stability of prices and of the financial landscape as a whole. To accomplish that mission NCBs rely, among other tools, on the Supervisory Review and Evaluation Process (Bank). This is a yearly assessment of how each bank is performing, that summarises supervisory findings for that year according to:

- How sustainable its business model is;
- The governance policies and their implementation;
- How the capital buffers can absorb the effects of adverse economical scenarios;
- The liquidity available to meet short-term needs.

To a great extent, the SREP process depends on the data reported by the banks, which is generating at increasingly higher rates. Supervisors go through carefully selected indicators in order to obtain an overview of each bank's financial outlook. Reporting requirements are defined and reviewed at European level. This step ensures harmonisation of the reported concepts as well as a broader perspective on a bank's finances through the analysis and interpretation of ever-increasing amounts of information. Supervisors are thus provided with a wide scope of financial figures, with the aim of giving an accurate perspective of how banks operate.

The magnitude of data retrieved within the regulatory framework is overwhelming for traditional approaches. Despite the fact that NCBs are dealing with structured financial data, going manually through hundreds of balance sheets, each of which containing thousands of financial figures is impractical, even for the largest teams. This naive approach is undoubtedly error prone, making a timely assessment of distress events unfeasible. Furthermore, traditional approaches also fall short on testing alternative economic conditions by impacting key indicators.

Conventional business intelligence can already uplevel financial data analysis by providing organised views of the reported data, through standardised reports and

interactive dashboards (Broeders and Prenio, 2018; di Castri et al., 2019). The use of existing, or innovative, technology to support the supervisory processes is denominated Sup-Tech, and it is a trending topic in the banking context. The analytical tools currently at hand deliver an aggregate view of the available data and also allow the combination of reported data-points to provide new key indicators. However, this *ad-hoc* method only looks at past events, and it is limited to the regulatory framework. It seldom considers any specificity hidden in the data that might lead to missing a crisis event, it overlooks the decision processes on top of that information, and disregards the value of what-if analysis that can only be carried out as exploratory exercises.

The financial sector's characteristical risk-aversion allied with the fast-paced revisions to the regulatory context, have prevented the early adoption of alternative solutions to this problem. The lack of qualified human resources has also been one of the main hurdles to changing how supervision is brought about (Doerr et al., 2021). Several authors showcase sup-tech initiatives throughout central banks, and the potential shown by machine learning to redefine supervisory processes (Chakraborty and Joseph, 2017; Broeders and Prenio, 2018; Beerman et al., 2021; Hertig, 2021). These works stress the gap that this thesis proposes to address and that are evidenced by its key components:

- Machine learning to support supervisory risk assessment;
- A European level standardised risk measure, Risk Assessment System (RAS), the quantitative pillar of the SREP methodology - assessing a bank's risk according to liquidity, credit, market, operational, and profitability;
- A multi-class formulation of the risk assessment ML problem;
- Real-world supervisory data from the Portuguese banking sector (March 2014 until August 2021).

In this thesis we use machine learning techniques to model the risk assessment process and present a comparison of their performance.

The novel contributions of our work are threefold:

- Banks and consultancy companies are the first to benefit from a comprehensive perspective on which risk assessment approaches are available. The results provide a guideline on how to leverage on ML to anticipate risk, through a compendium of ML techniques and risk measures. Decision support systems based on the models presented in this work allow a bank to proactively monitor and adjust its own risk exposure, not only from a business model perspective but also from the regulatory compliance standpoint.
- Central banks can find a collection of machine learning techniques used for financial data, and leverage on that information to develop sup-tech initiatives. Early Warning Systems are among the many innovative projects being developed by the ECB, Bank of International Settlements (BIS) and worldwide NCBs, using quantitative data.
- Academia can use this this work to expand the usage of ML in the regulatory context. This will hopefully serve as a pillar to further developments on the applications of machine learning in the supervisory context.

There are some limitations to this work that can trigger future developments in the area. Firstly, we envision the expansion of the sample to European level. The ECB retrieves data from all National Central Banks, holding data for all banks in the Euro-area (roughly 3150 institutions). Expanding the sample would increase model robustness. Another limitation is the lack of contextual data to evaluate the ML models. By providing risk specific data we could top off the accuracy of the resulting models. Finally, we suggest combining the quantitative approach in this work with the qualitative data available in reports and internal notes from supervisors, through the use of Natural Language Processing (NLP).

Chapter 2

Machine Learning Applied to Banking Supervision: a Literature Review

2.1 Introduction

Decision support systems had their genesis in the 1960s (Burstein et al., 2008). Perhaps because of the exposure risk and magnitude of revenues generated, the financial sector has been a particularly avid driver for developing these technologies.

Predicting how financial institutions will perform and whether they will create value is key for every contender in this field - financial institutions, central banks, consultancy companies, and academia. Consequently, the use of new technology and methods to support risk assessment tasks (fin-tech) is a rising trend in this sector (Milian et al., 2019). In recent years, machine learning (ML) methods and, to some extent, deep learning (DL), have been used for the assessment of credit risk, and more broadly, predicting bank failures. Currently, traditional statistical methods are still commonly used for this purpose. Nevertheless, machine learning techniques are overcoming traditional approaches by allowing practitioners to module past decisions, exploit them for other scenarios, and predict future chaotic phenomena.

This review intends to provide a comprehensive picture of how machine learning techniques have been used so far in risk assessment from a central bank's perspective. Thus, the scope of this work encompasses credit institutions and investment firms since those are the ones the European Banking Authority (EBA) regulation focuses on (Authority, 2013). Henceforth, the term *institutions* will be used to refer to both.

The above-mentioned regulation establishes the standardisation of reporting requirements under the Single Supervisory Mechanism (SSM) (Commission, 2015). As a consequence, this study focuses on the European banking sector. Although we are aware of the importance of insurance, pension funds, securities, and markets in the financial sector, these are subject to different regulations and would benefit from a dedicated study. This work intends to contribute to several stakeholders in the supervisory landscape:

1. Institutions can have a comprehensive perspective on which risk assessment approaches are available and how they can evaluate their own exposures.

2. Central banks can acquire an integrated view of several validated methodolo-

gies for risk assessment. These can be the pillars of their next decision support systems by laying down the technologies supporting risk assessment processes. Furthermore, this work can also incite surveys and case studies on the use and adoption of ML at central banks.

3. Consultancy companies will benefit from a compendium of ML techniques and risk measures, to better support their clients.

4. Academia receives an important contribution that gathers an extensive number of papers on risk assessment and collates the identified methodologies from a supervisory perspective. This will hopefully serve as a stepping stone for future developments in this area, and provide a baseline for testing new methodologies.

This paper is organised as follows: it starts by justifying the methodology and describing how the references were selected. The results section gathers similarities among published scientific knowledge and presents the most relevant works that influence this field. The last section provides a space for discussing lessons learned and future work.

2.2 Methodology

This research was conducted through a series of exploratory steps on the topics of machine learning, banking, risk assessment, and banking supervision. The initial objective was to evaluate how machine learning techniques were being used at central banks. Additionally, we intended to analyse how these methods were informing the analytical capabilities of supervisors. We then refined a search query broad enough to return a set of articles we could work on. The following subsections describe a step-by-step guide for the reference search and selection.

2.2.1 Engines

This literature review relies on three search engines: *Springer Link*, *ScienceDirect*, and *Google Scholar*, queried until June 2021. The first and second search engines are extensively renowned for their trustworthiness and for selecting top journals for their results. The last one provides an extensive overview of all articles published in English (Gusenbauer, 2019).

2.2.2 Query

Through extensive addition and diversification of search terms, we refined the search query to the following: ***"machine learning" and ("bank" or "banking" or "supervision")***.

The underlying reasoning is that machine learning techniques are the focal point of this review article. The added value comes from analysing their potential applications to the banking sector, specifically banking supervision. No limitation concerning the year of publication was applied. Overlapping results are addressed in our secondary analysis. Furthermore, no filter regarding type or place of publication was applied, since the included papers' journals of publication were evaluated

and classified after screening. Additionally, to keep up with new publications, we defined an alert in *Google Scholar* with this query. Finally, we pay close attention to Mendeley's alerts for articles related to the set gathered in this review.

2.2.3 Steps

The following subsections detail every step of the selection process summarise in the following PRISMA diagram 2.1.

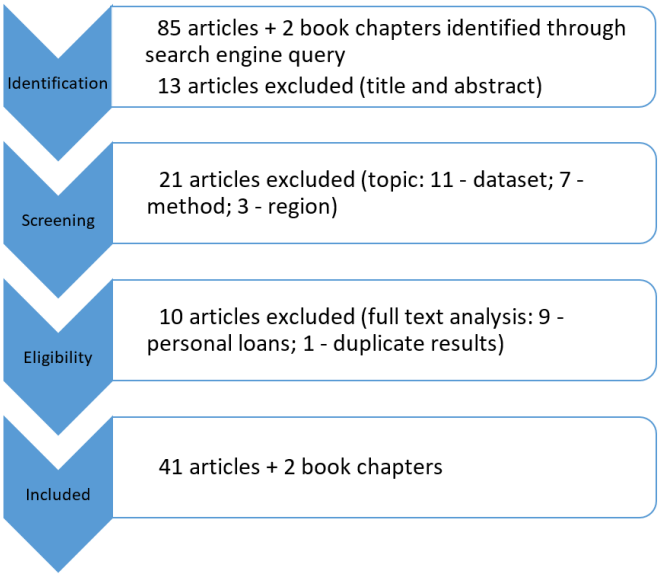


Figure 2.1: PRISMA diagram detailing the selection process of the identified articles.

Table A.3 lists the selected papers, providing a single-sentence summary of their content.

Identification

The research query identified 85 articles and two books, from the three search engines. All the papers were published in English, in several different journals, and spanned from 2000 to 2021. This first step involved title and abstract analysis, and excluded 14 articles for lack of relevance.

Screening

In this phase, the main topics of each article were analysed, resulting in the exclusion of 21 papers, based on the following criteria:

- Dataset: when the analysed paper used data other than the banking sector, it was discarded. We are aware that applications of ML to the stock market are a trendy topic in the literature, and that the insurance and pension funds sector is of great importance in the Eurozone. Nevertheless, the regulation is substantially different, and they would merit from a different study and approach;

- Methodology: risk assessment exercises are historically based on quantitative data, combined with expert judgment. Furthermore, it is the quantitative data that holds the largest amount of information regarding risk exposure practices. Therefore, we focus our analysis on quantitative methods, for which a risk assessment classification has already been assigned (leveraging on previous knowledge through supervised learning). We thus excluded works concerning unsupervised learning methods, or sentiment analysis (qualitative);

- Region: this criterion is closely related to the first, since regulation changes according to geography. We chose to focus mainly on works based upon institutions operating in the Eurozone. Nonetheless, relevant works by other central banks were considered eligible.

Eligibility

The next step required a thorough analysis of each paper, to verify its sources and classify the journal it was published in (quartile of impact). Papers were analysed from 2021 backward to identify any overlapping results or new or improved methodologies, resulting in the exclusion of ten more articles: nine being personal loans related and one duplicate result.

The scope of this review is the application of ML techniques to risk assessment from a supervisory perspective, which includes at best how institutions are addressing their risk assessment exercises. The data and predictors used to evaluate an individual credit application (personal loan) differ substantially from the data used by banks from a corporate perspective, and even more from the data collected in the regulatory context. As such, works regarding credit risk for individual applicants were also excluded.

Considered papers

The final article base consists of 41 papers and two books, published from 2000 until 2021, selected through the steps mentioned. In the next section, we will describe the similarities among the papers, as well as the methods applied and respective banking areas.

2.3 Results

2.3.1 Distribution

Based on the reviewed works from the previous section, the following paragraphs describe how machine learning techniques have been used in the banking sector. Our research intends to provide a future reference on how these technologies address and support the risk assessment process, in particular from a central bank's perspective. These results solely reflect the analysis of the papers selected for this review. They represent neither the total of publications throughout these years nor the distribution of topics for all publications.

Table A.1 summarises the selected articles, referenced by author, year of publication, affiliation and number of citations. Additionally, table A.2 lists the journals from the selected articles.

The most common topic on these papers is credit risk related (nearly 34% of references), as shown in Figure 2.2.

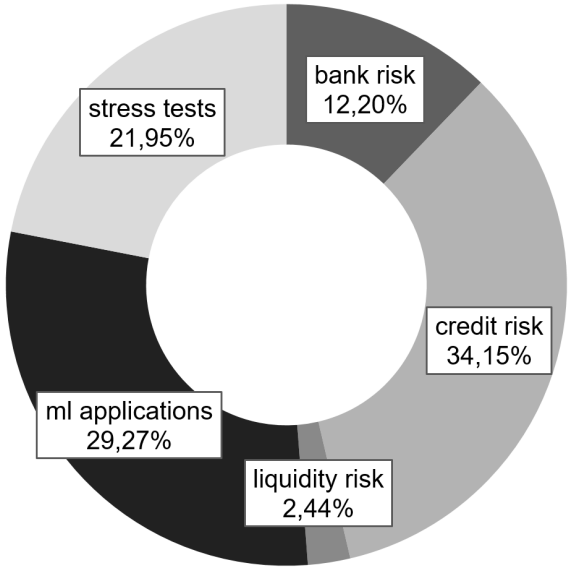


Figure 2.2: Distribution of articles according to main topic.

The second major category relates to "ML application" (surveys, fin-tech and sup-tech, as per the division suggested by Broeders and Prenio (2018), the use of innovative technologies by supervisory agencies to support their processes) along with "stress tests". The remainder of the results focuses either on "bank risk" more broadly, or on specific topics for supervision such as liquidity risk and other banking risk perspectives. Another relevant aspect is the publication date of these articles, ranging from 2000 to 2021 and distributed as shown in Figure 2.3.

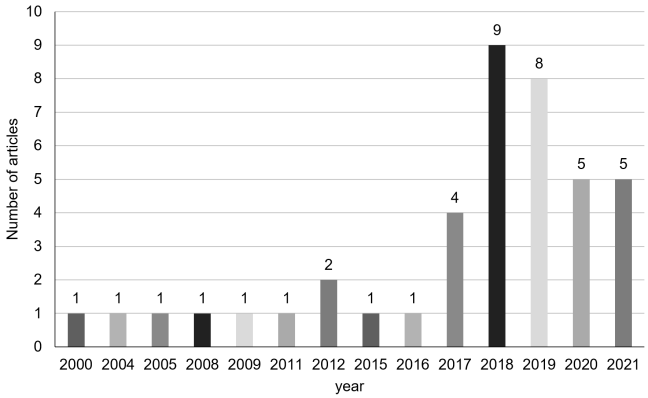


Figure 2.3: References according to year of publication.

Importantly, although ML applied to the financial sector has been present since 2000, by 2015 the intersection of these knowledge areas gained a huge interest. This translated to increasing numbers of publications in this field, with the majority of relevant articles in this study being published from 2017 onward. Table A.4 lists the machine learning methods applied by each author as well as the datasets that supported each research.

2.3.2 Evolution

The selected papers were organised by date of publication. Publication intervals were defined based on relevant events in the banking sector, technological evolution, and the number of papers per interval. The first slot ranging from 2000 to 2011 encompasses the effects of the financial crisis of 1999 and 2008. The second range (from 2012 to 2016) still reflects several studies based on the 2008 crisis, but with a more mature insight. In this period there is also a trending increase of ANN models. The third slot encompasses the years of 2017-2018, which show a significant increase in publications intersecting ML and the banking sector.

The final interval (2019 to the current date) depicts important ML applications to the financial market in general. Studies in this period reveal an increased ponderation of the uses and impacts of machine learning in banking supervision, with several publications from banking authorities.

2000-2011

Six papers were identified from this period. They mostly focus on stress tests although three of them engage on the topic of credit risk and default risk.

Early in this period, Galindo and Tamayo (2000) identified the risk assessment task as crucial for an efficient use of resources. They used an error curve methodology to compare model precision and concluded that tree-based models outperform ANNs, KNN and probit. This sets forward the finding that tree-based models are more appropriate to structured data, as opposed to ANNs.

Hillegeist et al. (2004) proposed a new method for assessing bankruptcy probability. Based on the Black–Scholes–Merton option-pricing model, this method was compared to the well-known Z-score (Altman, 1968) and O-score (Ohlson, 1980), obtaining superior results. These authors stressed the need for a standardised risk assessment measure mainly for comparability purposes.

Min and Lee (2005) presented a paper that compares statistical and artificial intelligence methods, with the latter outperforming the former in the classification of bankruptcy. Although this study focuses on credit risk assessment for heavy industry firms in Korea, we included it in our sample for a compelling reason. It is a clear example of machine learning methods outperforming conventional statistics and it uses a set of predictors (financial ratios) easily mapped to regulatory financial reporting since they are based on balance sheet entries. Angelini et al. (2008) based their work on the Basel II capital requirements and the need for a system to assess credit risk. The main objective of this work is to evaluate the possibility of using neural networks to estimate the probability of default of a borrower (Italian small companies). In spite of some ANNs being used, the comparison of classic machine learning models to conventional statistical methods was the more recurrent approach. Furthermore, the risk definition used to evaluate the data sets was based on the probability of default. This is explained by the fact that the datasets are mostly from loan applications, either from small and medium enterprises or personal loans (housing included). These findings contradict Galindo and Tamayo (2000) as well as more recent developments in this area. ANNs have been proved to excel in time-series, image, and voice recognition, as opposed to their performance using structured data.

Additionally, some articles used financial ratios and CAMELS rating model (an

international rating system used by regulatory banking authorities to rate financial institutions) to assess an institution's performance (stress testing and bankruptcy prediction). Assessing the health of a bank is crucial to prevent its failure and contain the systemic risk its failure or losses represent. The work of Boyacioglu et al. (2009) identifies this assessment as an original classification problem. The authors use the CAMELS method to select the most relevant predictors. Using this method, neural networks were shown to outperform multivariate statistical methods for a Turkish banking sector use case.

Chaudhuri and De (2011) considers Basel II definition of risk to select features for the models. In this case, ANNs are not as frequently used as other conventional ML techniques, such as support vector machines and k-nearest neighbours. As a consequence, the authors focus on the optimisation of those models to the problem at hand (i.e. nature of the dataset).

2012-2016

In this period, articles mostly reflect the first insights gained from the 2008 financial crisis.

Having identified the lack of a comprehensive method to incorporate circumstantial aspects into the banking default risk predictive models, Ribeiro et al. (2012) reported that SVM+ outperformed other methods that did not include non-financial information. Hammer et al. (2012) showed that Logical Analysis of Data (LAD) is an accurate method by reverse-engineering Fitch risk ratings. The authors stated that LAD can be used as an internal rating system that is Basel compliant.

Iturriaga and Sanz (2015) took a different approach to this matter. First, they used self-organising maps (SOM) to profile distressed banks. This unsupervised learning method is competitive so it thrives to reach the right pattern, the representation of bankruptcy for a bank. Afterward, the authors applied multi-layer perceptrons to assess a bank's risk in several time frames, obtaining very promising results predicting bankruptcy for commercial banks. This two-step approach is the first in this selection of papers to recognise the benefits of a pre-processing phase to map the bankruptcy layout of a bank. Although previous research has shown better results using conventional ML, the success shown by this perceptron model suggests it is adequate to model the time evolution of quantitative data.

A new approach to credit scoring using an ensemble model was proposed by Ala'raj and Abbod (2016). These authors combine several data filtering and feature selection methods before evaluating model performance, and compare the most traditional classifiers with their method. The results are validated on several public datasets and their accuracy assessed under several measures: average accuracy, area under the curve (AUC), H-measure, and Brier Score. This is the first paper in our sample showing that ensembles outperform single models for classification problems.

2017-2018

These two years showed a more than 60% increase in publications in the intersection of ML and banking sector. As highlighted by Strydom and Buckley (2019), the technological evolution allowed for the development of deep learning (DL) models, as well as new ensemble methods like extreme gradient boosting (XGBoost). Although

the DL's first reappearance happened in 2012 (Krizhevsky et al.), its application to financial risk only came to light in 2016-2017.

Traditional ML and classical statistical approaches are still the cornerstones of most of these articles. However, an increasing trend is noticeable in the use of ANN-based models mainly due to bigger datasets and enhanced computing power.

Abellán and Castellano (2017) build on their previous work showing how ensembles achieve better results in credit risk assessment than single models, validating the findings of Ala'raj and Abbod (2016). The authors stress the importance of individual model performance as a criterion for ensemble selection. Although the authors emphasize their own tree-based model (Credal Decision Tree, CDT), the main finding of their work is the corroboration of the hypothesis that ensembles outperform single classifiers.

Prompted by the 2008 Global Financial Crisis and the need to foresee signals of financial instability, Italian authors Pompella and Dicanio (2017) developed an Early Warning System (EWS) to help uncover distress signs for banks. This credit risk model allows users to discriminate stable from likely-to-fail banks and might be useful in adjusting rating assignments by Rating Agencies. The authors suggest its implementation in regulators to support the supervisory process.

Xia et al. (2017) present an extreme gradient boosting model (XGBoost by Chen and Guestrin (2016)) that consistently outperforms baseline models. The authors stress the importance of model-based feature selection as well as the use of Bayesian hyper-parameter optimisation to achieve better predictive results. Although personal credit risk is not the main topic of interest in this review, this study shows the advantages of boosting techniques and the importance of an interpretable model for decision making. This type of models have won several Kaggle competitions and are consistently showing excellent results with structured data.

Chakraborty and Joseph (2017) from the Bank of England introduce a central bank perspective on machine learning and its applications. The authors provide an overview of machine learning models and model validation to support the presentation of three case studies. As a final note, this work acknowledges the amount of available data as an important vector in decision support systems based on machine learning at central banks and other offices. As previously stated, agency papers as this one are paramount in understanding the use of machine learning in these contexts, providing use cases and areas of interest for future work.

Alessi and Detken (2018) contribute with another EWS to detect excessive credit growth. This phenomenon is usually at the root of systemic risk to financial stability and its early detection can help avoid cases of bankruptcy. The authors use Random Forest classifier model with credit and real estate predictors. Their work pioneers in the domain of risk assessment from the perspective of central banks, thus setting peer practitioners in their future path. Moreover, the work reinforces that ensembles consistently outperform single models. Other authors successfully use extreme gradient boosting to develop a credit risk model for financial institutions (Chang et al., 2018). Those tools promise significant support (i.e. low error rate) for risk assessment in loans.

The Central Bank of Greece also provides a thorough analysis based on post-2008 crisis loan data from Greek banks, by Petropoulos et al. (2018). This study sets a milestone for the use of advanced ML techniques from a supervisory perspective. Furthermore, it leverages the resulting model to create an EWS that will support

subsequent decisions in loan approval. Similar to what Iturriaga and Sanz (2015) have shown, modeling a timeline evolution is where neural networks (in this case deep neural networks, DNN's) excel. Another important result is that DNNs can perform just as well as XGBoost, showcasing how precisely deep learning models adapt to structured data.

Tavana et al. (2018) present a study that directly addresses liquidity risk, which is the most rapidly devastating risk a bank is exposed to. In this paper, the authors present an artificial neural network model combined with a Bayesian network (BN) to assess liquidity risk using solvency as a proxy. This combined approach models the liquidity risk indicator through the ANN and the probability of occurrence through the BN. The results show this approach distinguishes the most critical factors for liquidity in this dataset.

Broeders and Prenio (2018) conduct a study that compiles the experience of early users of innovative technology in financial supervision (sup-tech). The authors structure a definition of sup-tech and show how it is used for data collection and analytics. These two applications have different initiators in supervisory agencies. Data collection tends to be initiated by management decisions and projects whereas analytics usually start out as research questions or analysis queries from supervision units. A conductive thread of all use cases is the sharing of the experience of some early adopters and the impact those technologies are having on the organisation. Similar studies, such as the one conducted by Chakraborty and Joseph (2017) are essential for compiling, sharing, contrasting the several approaches throughout central banks and other agencies.

The Federal Reserve provides a broader perspective, analysing how the use of machine learning and big data will impact compliance aspects (Jagtiani et al., 2018). The authors also stress the need to identify the risks that these technologies carry when applied to the financial market.

Gogas et al. (2018) propose a methodology that separates solvent and failed banks, using machine learning models. The authors present an alternative tool for stress-testing that outperforms the O-score. Their approach is based on a support vector machine model that helps to define a boundary between solvent and insolvent banks, converting this issue into a classification problem. Kupiec (2018) presents a related study that stresses the need for new methodologies to validate conventional bank stress tests.

As a final reference for this period, Le and Viviani (2018) also tackle the problem of bank failure prediction using machine learning and classical financial ratios. One important aspect of this work is that the authors use ratios from 5 different risk perspectives: Loan quality, Capital quality, Operations efficiency, Profitability, and Liquidity. This work validates yet again that machine learning methods outperform traditional statistics. However, these authors do not explore the possibility of using ensembles, which have already been proven to be top performers in classification problems.

2019-2021

Credit and banking risks are essential for a balanced economy; trying to prevent systemic repercussions stemming from them is considered of the utmost importance. Similarly to earlier periods, these risks maintain a privileged spot in research. Still, it was on ML application we saw the most significant increase in publications. This

suggests the demand for coordination and a global perspective on the developments conquered so far in this area.

Leo et al. (2019) produce a thorough review on how machine learning has been used at banks for risk assessment. This paper offsets the industrial and academic claim for ML application *versus* real-life practices, highlighting a series of perspectives where risk management has been poorly applied. Climent et al. (2019) develop an insightful study that aims to identify a set of financial predictors that best model a bank's financial distress. To this end, the authors apply an XGBoost based model to a set of indicators that might predict a bank failure in the Eurozone. The set of selected indicators (Total assets, Loan loss provisions/net interest revenue, Equity/net loans and Interbank ratio) are shown to best help regulators monitor financial distress for those banks. From a technical perspective, this work reinforces the choice of XGBoost for classification problems using structured data. A recent study by Wang et al. (2021) deconstructs the use of logit as the base classifier for EWS developed to predict banking crisis. In fact, the authors use random forest classifier to simulate expert decision, obtaining a generalisation capability above 80% area under the curve (AUC).

Kou et al. (2019) compare several ongoing researches concerning the applications of machine learning methods to the detection of systemic risk events, that is, financial distress phenomena that affect several markets or geographic regions. They also propose the use of big-data analysis to assess systemic risk.

Soui et al. (2019) address the issue of comprehensibility of machine learning models for credit risk assessment. Interestingly, in this study, interpretability was mentioned as one of the barriers for adopting ML models in day-to-day decision making. In an attempt to circumvent this problem, the authors proceeded to develop an evolutionary algorithm to approach credit risk assessment as an optimisation problem: minimising complexity while maximising accuracy.

A recent review by Dastile et al. (2020) comparing statistical and ML learning models for credit scoring showed that ensembles outperform single classifiers, confirming the results of previously mentioned works. The authors identify model explainability and the ability to deal with imbalanced datasets, as the main issues to deal with when modelling credit risk. Deep learning models also show promising results, although they have not been extensively explored for credit risk assessment. The authors identify the lack of interpretability as the main barrier for adopting deep learning for credit risk assessment.

Banco de España (Alonso and Carbo, 2021) published a comparison of several well-known machine learning algorithms for credit default prediction, showing significant improvements over logit. The authors estimate that implementing XGBoost-mediated assessment could lead to savings of up to 17% of capital requirements under current ECB regulation. Antunes (2021) from the Central Bank of Brazil presents a solid argument to maintain supervisory on-site inspections. The author compares two machine learning models, one trained with portfolio ratings assessed by the banks themselves, and the other based on past ratings obtained through on-site inspections. The results show that the overall performance is consistently higher when using data retrieved through inspections.

This is the period with the most ML applications papers identified (with a total of 9 out of 13). They span from insights on how AI will continue to revolutionise industries and change social behaviour (Dwivedi et al., 2021), to more practical

approaches on how to incorporate ML in financial services (Lee and Shin, 2020). Milian et al. (2019) also provide a list comparing fin-tech definitions, how it is supported by digital transformation, and the financial risks associated with the use of ML.

A comprehensive study from 2019 by di Castri et al. (2019) focuses on the definition of sup-tech and highlights the need for a more precise notion of what to include as "innovative technology" at the service of a financial authority. It presents several use cases and classifies the technologies onto maturity levels (named in the paper as "generations"), concluding that the identified initiatives (applications of innovative technologies to support the activities carried out by financial regulators and authorities) are mostly experimental. The authors suggest an international coordination effort and alignment to create synergies that leverage sup-tech development.

The Bank of Italy presented a use case for a classification problem (deducing the institutional sector code of a company based on its characteristics) (Massaro et al., 2020). Although this work is not related to risk assessment, it provides an excellent example of a production-ready application of ML to supervisory tasks.

Alonso and Carbo (2020) from Banco de España stress the need for a joint strategy to assess ML models to increase transparency and promote adherence to this technology. The authors conclude ML models increase the predictive capability of a credit default classifier by 20%. The study also identifies factors in credit risk management that might increase supervisory costs.

Driven by the recent progress in financial technology, Huang et al. (2021) acknowledge the complex and hierarchical nature of financial data and the technological barriers found when using statistics and classic ML. The authors then proceed to apply advanced deep learning methods and make use of several graphic processors to improve computation.

As a final remark regarding ML applications, Doerr et al. (2021), from the Bank of International Settlements, presented a policy briefing on the European Money and Finance Forum, evaluating to what extent central banks are making use of ML and big data. The authors conclude that although central banks are acquainted with big data, there exists a persistent need for specialised knowledge on how to use ML throughout these organisations.

Stress tests are also referenced in these years. In a 2019 study, Kolari et al. (2019) hypothesise that stress tests themselves are more of an assessment of a bank's ability to deal with the risks it is exposed to. This statement challenges the common conception of stress tests as a marker of a bank's resilience to adverse alternative macroeconomic scenarios. For this purpose, the authors develop an early warning system to assess how European banks will perform on stress tests. These authors suggest surviving stress tests depends largely on the underlying risk dimensions of individual banks. Moreover, this paper reaffirms boosting techniques as winning solutions, not only for this sort of classification problems but also when applied to structured data. As a future work, the authors recommend a similar approach using regulatory data.

In the same line of investigation, an EWS was developed by Filippopoulou et al. (2020) to predict bank systemic risks in the Eurozone. This study starts by analysing the importance of the indicators that are usually applied and presents a model that detects a systemic crisis one to four years beforehand. In spite of using a classic multivariate binary logistic regression model, the methodology adopted for this EWS

shows promising results and can be a reference for future developments in this area.

2.3.3 Datasets

Most central banks and supervisory agencies do not make their datasets available for confidentiality reasons. This is true for several types of data, such as credit responsibilities and supervisory data (Authority, 2013).

As depicted in Figure 2.4, regardless of the research topic, most datasets used in these papers are public. The main reason for this is that most researchers cannot gain access to validated supervisory data. Another relevant aspect is that central banks and supervisory agencies have just begun to engage in programs where ML development strategies were in place. These developments are starting to appear, as can be seen by the growing number of titles under the "ML applications" topic. Table A.4 lists the datasets used in each paper.

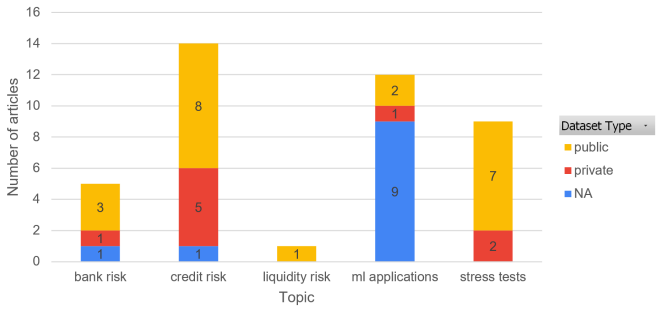


Figure 2.4: Dataset types by research topic (*NA* - not applicable)

Some rating agencies, central banks and other institutions provide datasets to support research projects. A good example is *Banco de Portugal* BPLIM (de Portugal, 2021), a micro-data research laboratory that provides up-to-date anonymised datasets available for national and international researchers. Another example is Moody's DataHub (Moody's, 2021), that provides a cloud-based platform containing eligible data alongside affiliated third-party participants.

2.3.4 Related Work

In this research, we have found few papers strictly addressing the use of machine learning techniques for supervisory risk assessment. As a consequence, we have broadened our research question to include banking risk assessment and machine learning in the financial sector. This reasoning is thoroughly presented in section 2.2.2.

Nonetheless, we found some works that support the purpose of this review. di Castri et al. (2019) is a survey that summarises the activities that can be considered as an application of innovative technology to supervisory purposes. The authors also present a series of use cases, mostly experimental and originated by supervisory agencies. Kou et al. (2019) list the most common methodologies - ML, big data analysis and sentiment analysis - to address systemic risk in the banking sector. Last, and closest to this research, Leo et al. (2019) contribute with a literature review that brings to light how machine learning is currently being used in the banking sector. The authors stress that contrarily to what might be expected

due to the magnitude of financial consequences involved, the real-life use of these sophisticated technologies is in fact under-used and poorly developed.

The authors' specific knowledge of banking context, namely projects within *Banco de Portugal* and European Central Bank, allowed them to propose a reliable proxy for the scarcity of published works on this topic. To establish the ideal perspective, we evaluated how risk assessment is carried out in the banking industry, and central banks in the SSM. On the other hand, we investigated how ML is being used for risk assessment in banks. Additionally, we referenced various surveys from central banks to depict and support our statements regarding the use of innovative technologies for supervisory purposes.

In this sense, although this review is sustained by a proxy and there is a paucity of related works from a central bank perspective, the authors propose this review as a starting point for researchers and industry stakeholders. We aggregate relevant contributions to support and ignite the use of ML in risk assessment exercises, from a central bank or supervisory agency perspective.

2.3.5 Global Analysis

The set of papers identified in this review includes diverse approaches to risk assessment. We have selected some works that use a specific bankruptcy indicator (such as the Altman score or the O-score). However, most of the authors set forth from a set of financial ratios and, knowing the final result, try to model that knowledge through supervised learning. Most of these approaches convert the problem at hand to a classification task, for example, "failure" or "no failure" of a bank.

Another interesting aspect is how the datasets are designed. Most of these works use public datasets to validate a certain approach, even though some of these datasets are specifically collected to depict financial crises. The set of features available in these datasets often reflect a certain industry perspective of risk assessment. For instance, many datasets focus on credit and profitability ratios, since both are two crucial vectors for the industry: how a bank performs and how it is exposed to its main business model.

As a final remark, although most of the selected works come from the academia, we would like to mention the five papers published from 2017 until now by central banks. Alessi and Detken (2018), from the European Central Bank (ECB) and European Commission, have a significant number of citations (135 by the end of 2020) and present an important EWS that can support everyday processes. Also, Chakraborty and Joseph (2017) from the Bank of England give a great contribution with a broad view of what is being done with ML in this context. By presenting some use cases, they also turn the spotlight on the successes of these approaches. The Bank of Greece presents an insightful use case by Petropoulos et al. (2018) for credit risk analysis.

From a more strategic point of view, Jagtiani et al. (2018) from Federal Reserve Banks depict the impacts, roles and possible risks of using ML at central banks.

Although not related to risk assessment, a recent study Massaro et al. (2020) from Bank of Italy presents a production-ready solution of the application of ML techniques to everyday central bank tasks. This is one of the most recently works, showing how ML can make a difference in day to day tasks.

2.4 Conclusion

This review provides a comprehensive picture of how machine learning techniques have been used so far in risk assessment from a central bank's perspective. It is organised by timeline and topic. All of the presented topics relate to some extent to the supervisory activity and to dimensions of analysis that are part of the day-to-day processes. As a consequence of the SSM legislation and the EBA reporting requirements, this work focused on the European banking sector.

The majority of the selected papers reflect upon the credit scoring problem. This stems largely from the fact that granting loans is the core business of most of the commercial banking sector. Stress testing in the form of bankruptcy prediction is also in the spotlight since it is strongly connected with regulators' compliance. There are several other risks a bank is exposed to that require their own studies, such as liquidity or operational risk. However, focusing on those risks is more of a compliance issue, rather than a business model perspective.

Some studies benefited from more structure and clarity, which is useful for comparability purposes. The more structured studies answer the questions of which problem they are addressing (a measure of risk and its perspective, a stock index, portfolio pricing, etc.), ML techniques that were applied, and variables considered. They also offer insight into the datasets they were based upon, and clarify the methods used to assess the models' precision and prediction capability. The lack of this organised approach evidenced in some articles made it more difficult to review and condense the information published across the broad spectrum of expertise found. As a consequence, interpreting data originating in different geographies and diverse banks' business models proves to be a challenging task. International consensus must be established regarding terminology, analysis methods and result reporting, as pertaining to this field. The authors advocate for a universal risk assessment methodology, classifying bank risk according to preset parameters and based on the same data, regardless of their location or business model. To this end and taking advantage of the central bank's perspective, the authors suggest the use of the Supervisory Review and Evaluation Process (SREP), namely, one of its pillars, the Risk Assessment System (RAS). This methodology is used by the ECB and applied, to some extent, to every institution in the SSM. Through the application of such a broad methodology, results of analysis and ML application are more comparable to an already established practice.

Another relevant aspect is the paucity of data published from a supervisory perspective. The reviewed papers mainly focus on credit risk and stress tests using public data. Despite being useful in assessing the financial health of a credit institution, they seldom use data collected through supervisory directives. Scenario testing, sometimes used as a synonym for stress testing, is another decision support system that greatly increases the analytical capabilities of supervisors. The authors emphasise the importance of landmark publications such as the EWS proposed by Filippopoulou et al. (2020), using data gathered in the aftermath of the 2008 economic collapse (European Central Bank Macroprudential Database). These systems are especially relevant since they function as a daily tool for analysts, and strongly benefit from supervisory data. The EWS developed by Alessi and Detken (2018) has also had an enormous impact in the literature by presenting a solution for anticipating banking crisis, using random forests.

As a final remark, we point out that many of these studies rely on public datasets. This often implies they are not as recent as desired since the data might not include the more recent events. For instance, a dataset from 2005 to 2011 captures the market behaviour before the crisis, the crisis itself, and a fraction of the decline of the market. It would be useful to model the behaviour of the institutions with the new regulation as well as the economic recovery seen later until 2019.

2.4.1 Limitations and future work

This study proposed to select and review the literature regarding the applications of machine learning to banking supervision. However, since this is a rather specific topic and the regulation has suffered a thorough revision after the 2008 financial crisis, our review falls short on papers that address solely this issue. There is some literature published by central banks and other agencies, but these works are mostly surveys, assessments of adoption, or definition of new concepts. As a consequence, the research query was broadened to include works from other perspectives:

- Assessment of credit defaults (the topic most explored in the reviewed literature);
- New stress test methodologies;
- Systemic risk detection;
- Other surveys regarding fin-tech and sup-tech.

All these topics are pillars of financial analysis and as such, they relate in a direct and crucial manner to proper supervision. Nevertheless, they are all collateral aspects and do not correspond to the core of the supervisory process itself.

Another aspect worth mentioning is the fact that our work is not a detailed review of the literature cited within it. Due to the heterogeneous structure of the included literature, we opted for a broader approach when comparing them. Each topic would merit an individual in-depth analysis and review, which was not warranted in the scope of this article. The authors believe this review will provide a stepping stone for supervisors, analysts, consultants, or academics that desire to further explore machine learning as a tool for banking risk assessment.

Chapter 3

Machine Learning for Liquidity Risk Modelling: a Supervisory Perspective

3.1 Introduction

Ever since the 1990s, the financial sector has stimulated the development of decision support systems (Zopounidis et al., 1997). Classic statistical methods, like linear or logistic regressions, have been a pillar of those systems and financial analytical models in general. More recently, machine learning (ML) has been gaining thrust as the preferable tool, mainly due to the vast amount of data collected and the increasing computational power available. ML has been proven to unveil previously undetected complex data patterns, which are almost impossible to model. Furthermore, a recent study from the Bank of England (Hertig, 2021) emphasises machine learning as a growing technology for supervisory processes, mainly for detecting illegal market practices. These findings support expanding the use of ML for other supervisory tasks, namely, risk assessment processes. The findings in (Guerra and Castelli, 2021) also settle the intersection of this two knowledge areas as the way forward.

3.1.1 Risk assessment measures

A universally accepted risk assessment methodology has always been a hot debate topic for researchers in this area. The methodologies used are increasing in sophistication as sup-tech is incorporated into day-to-day tasks. Additionally, this happens due to the necessity to accurately measure the risks banks are incurring in, from several perspectives. On the other hand, the need to comply with new regulatory requirements also triggers new approaches to risk analysis.

Several have proposed methods for bankruptcy probability assessment (Hillegeist et al., 2004; Ribeiro et al., 2012; Climent et al., 2019; Leo et al., 2019; Wang et al., 2021). The method presented by Hillegeist et al. (2004), Black–Scholes–Merton option-pricing model, outperforms two other well-known and reliable measurements: Z-score (Altman, 1968) and O-score (Ohlson, 1980). The authors stress the need for a standardised methodology to support the comparability between institutions.

Most of the current literature models risk classification according to a binary

classification, for instance, "failure" or "no failure" of a bank. This target variable is derived from a set of financial ratios, most often from public or proxy datasets.

For this study, we consider the classification method presented in a well-established and widely approved methodology for risk measurement - the Supervisory Review and Evaluation Process (SREP) (Bank) - defined by the ECB in cooperation with the National Competent Authorities (NCAs). This is the process through which supervisors periodically assess and measure the risk for each bank from five perspectives: liquidity, credit, market, operational, and profitability. The authors support our risk classification on the automatic Risk Assessment System (RAS), which is then reclassified according to expert judgement.

This methodology uses real supervisory data collected through the European Banking Authority (EBA) directive for Implementing Technical Standards (Authority, 2013), within the scope of the Single Supervisory Mechanism (SSM) (Commission, 2015). Data is used to classify each institution in terms of its risk level, according to the automatic risk assessment system from the SREP process. These observations range from 2014 until March 2021. The data used in this research is extensively validated, thus ensuring a positive correlation with liquidity risk assessment capabilities (Ng, 2011).

3.1.2 Machine learning for risk assessment

Risk assessment is a predominantly quantitative exercise, often adjusted through expert judgement. The use of machine learning methods from a central bank perspective is a recent topic of interest, not only from NCAs and other agencies' perspective, but also from the academic point of view.

Since the early 2000s, risk assessment has been identified as a top priority for the efficient use of financial resources (Galindo and Tamayo, 2000). Early in that decade, the same authors established that tree-based models are more adequate in prediction tasks when compared to artificial neural networks (ANN), using structured data. This result is reinforced by other publications, throughout the years. Kolari et al. (2019) specifically address stress testing, suggesting it is an assessment of a bank's ability to deal with the risk it is exposed to, rather than the bank's actual resilience.

Recent technological evolution has been supporting the development of more sophisticated models (Strydom and Buckley, 2019), like deep learning (DL) models, as well as new ensemble methods like extreme gradient boosting (XGBoost) (Abellán and Castellano, 2017), due to their capability to capture the complexity of this type of phenomenon. DL first reappeared in 2012 with ImageNet (Krizhevsky et al.). However, DL was applied to financial risk assessment only in 2016. Dastile et al. (2020) confirm DL as a promising tool in risk assessment, in particular for credit risk. They hypothesise extrapolating this approach to other risk perspectives, although the lack of interpretability of DL is seen by these authors as the main barrier for adopting this approach.

At the same time, several studies showcase the level of precision with which deep learning models adapt to structured data. Petropoulos et al. (2018) expand on the use of advanced ML techniques from a supervisory perspective. These authors developed an Early Warning System (EWS) for credit risk prediction, using data from Greek banks' corporate loans (Bank of Greece; 2005-2015).Although XGBoost emerged as the best model, DNNs also presented promising results. Similarly to

what Iturriaga and Sanz (2015) have demonstrated, modelling a timeline evolution is where neural networks excel (in this case, deep neural networks - DNN's).

As in bankruptcy prediction, using machine learning to model a risk assessment usually sums up to a classification task where the developed model assigns a binary result to a certain observation of context: "fail" or "no fail". This means that for a set of independent variables/indicators, that represent a bank's context in a certain period, the model will first learn, then predict, whether that bank will go bankrupt or not, with a particular degree of certainty.

On the business side, it is crucial to understand how banks, national competent authorities and other agencies are adapting to this evolution. In particular, we are interested in how central banks use innovative technologies to leverage their analytical capabilities, namely for risk assessment.

According to what Stock and Watson (2001) formulate that macroeconometricians at policy institutions do, NCAs are responsible for:

1. Summarising and analysing data;
2. Forecasting the key macroeconomic variables;
3. Conducting risk analysis and balance of uncertainties;
4. Performing structural/causal analysis, as well as scenario analysis;
5. Making decisions, communicating them and justifying these decisions vis-a-vis the public.

A study conducted by Broeders and Prenio (2018) showcases the experience of early users of innovative technology in supervision (sup-tech). This work presents a new definition of sup-tech and shows how it is used for data collection and analytics. Chakraborty and Joseph (2017) published a similar study where the authors compile, present and compare the approaches adopted by NCAs and other agencies. As noted before, the amount of available data emerges as an important vector for the development of decision support systems based on ML.

Massaro et al. (2020) present a production-ready solution using ML to support a NCA's everyday tasks. Although this work is not a risk assessment tool, it proves how these NCAs can leverage on sup-tech.

We found only one paper addressing risk assessment using ML, from a supervisory perspective (Filippopoulou et al., 2020). The EWS developed by these authors is of great relevance for central banks. It addresses risk assessment, but most importantly, it uses real data gathered in the aftermath of the 2008 economic collapse (European Central Bank Macroprudential Database). Pompella and Dicanio (2017) also propose an EWS to alert for banks' distress signs. The authors propose a credit risk model to help adjusting rating assignments by the responsible agencies. Along with Filippopoulou et al. (2020), these findings suggest EWS as reliable instruments supporting supervisory processes.

The paucity of studies such as the one just mentioned, is a gap we propose to address. To the best of our knowledge, there are no papers addressing liquidity risk assessment from a supervisory perspective. Additionally, this work uses real-world data collected at a central bank in the context of supervisory directives. The fact that this type of datasets are privileged and therefore confidential further justifies the nonexistence of similar studies.

The few studies addressing risk assessment with ML techniques use public or proxy datasets. These early works set the tone for the particular use case of central banks. In the supervisory context, data is confidential and the processes are supported by European-wide legislation, thus making these papers more likely to stem from joint works with NCAs. Additionally, we do not use a sample dataset but rather the entire population: the Portuguese banking sector. Also supporting the novelty of this work is the risk assessment methodology used: the quantitative pillar of SREP, the Risk Assessment System (RAS). We model the risk assessment task through a classification problem. As opposed to the papers cited above, we propose expanding the usual binary classification into multiple classes, according to banks' risk level and as established in the RAS methodology:

1. low risk;
2. medium-low risk;
3. medium risk;
4. high risk.

This approach ensures that we can look through the same lenses at all banks in the Euro-area, making these assessments comparable, replicable and transparent.

In this work, we decide to consider solely liquidity risk due to its high importance to a bank's financial health (Vento and Ganga, 2009). A liquidity crisis can lead a bank to bankruptcy in less than a week (Shah et al., 2018). Therefore, it is of the utmost importance to deliver innovative tools that increase the current analytical capabilities of central banks. We aim to provide a solid base for a scenario analysis tool.

3.2 Methodology

The fundamental purpose of machine learning (ML) is extracting predictions from underlying data (or Big Data). Generally, Machine Learning algorithms are applied to data to get insights from it. In this case we are using Cross Sectional Data, that can be captured at any point in time. Using information from previously observed circumstances (cross sectional data), ML algorithms can predict values pertaining to events that have yet to occur.

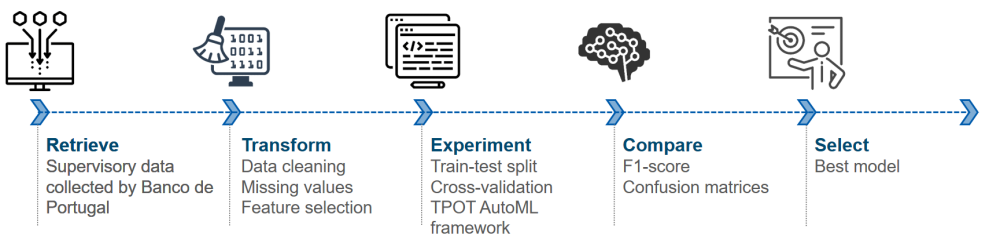


Figure 3.1: Methodology process overview.

Figure 3.1 displays the steps performed in the experimental phase of this study. We started by retrieving the data from the *Banco de Portugal* production database

for supervisory data. This dataset includes all the available features, as well as the pre-computed target – the RAS score for liquidity risk. Data transformation comprises data cleaning, implementing a strategy to deal with missing values, and the feature selection process. In the experiment phase, we compare three different approaches to evaluate the ML algorithms for this task: the classic train-test split, the more accurate cross-validation, and the TPOT AutoML framework (Olson et al., 2016). We then use the f1-score and the confusion matrices to compare the results and finally, select the best model. In future use, this model can be deployed as an Early Warning System making predictions for the liquidity risk level.

In this section we will describe the methods used in this research, from data gathering to model performance evaluation.

3.2.1 The Data

This study relies on supervisory data collected by *Banco de Portugal* (Portuguese Central Bank - BdP) within the Capital Requirements Regulation (CRR) and Capital Requirements Directive IV (CRD IV) Parliament (2013). The data ranges from March 2014 until March 2021. Depending on its nature, some data is gathered monthly while in other cases it is gathered quarterly (Authority, 2013). Due to confidentiality issues, the dataset used in this study cannot be made available for public consult.

Data is extracted via SQL query from BdP's production database into a *comma-separated-values* (csv) file to be imported using the Python programming language. An extraction routine was implemented to assure consistency and automation in data gathering. No filter is applied regarding reference date, institutions or level of consolidation. The extraction is structured in two steps:

1. First, the features are selected from the reported data. These belong to the 4 main reporting frameworks for banking supervision: Financial Reporting, Common Reporting, Asset Encumbrance and Funding Plans. This set encompasses all possible predictors.

2. The target variables are selected. These are computed through a corporate calculation process but all intermediate variables are discarded, in order to avoid any possible mathematical relation between features and target.

The data resides in a relational database where each row represents a reported value. This means that in the data source, several rows represent a single observation. During extraction, data is anonymised using MD5 algorithm within a hash function. This step assures the same identifier for every row in the same observation. The base dataset has the following topology:

1. **ID** - a hash code representing each observation's identifier;

2. **variable** - a code with business meaning that represents each reported value;

3. **val** - the actual numeric value of the variable.

3.2.2 Transformations

A python routine imports the CSV file, preparing the data for machine learning algorithms. The first step is pivoting the data set so that each of the resulting lines corresponds to an observation. Subsequently, we go through the data cleaning process that starts by discarding the target columns that fall out of the liquidity context. By this stage, each row corresponds to a single observation, and the last column represents our target variable (the RAS liquidity risk score). The other columns portray all the features available in our dataset.

The next steps delineate under which circumstances a row or column is discarded from our dataset:

1. Rows for which the target variable is null.

2. Rows that have a target variable 0. This value represents a non-applicable observation.

3. Rows where all features/columns are null.

4. Null columns: every column/feature has at least one reported value. After completing the previous steps, we must confirm that every feature still has values.

Finally, we deal with missing values for each feature. As pointed out by Madley-Dowd et al. (2019), multiple imputations can attain unbiased results up until 90% of missing data. Since in our dataset we have at most 20% of missing values, we do not discard observations based on this criteria. Instead, we use the median to fill out the missing values, which is the most adequate strategy for numeric datasets where the features present different distributions (Acuna and Rodriguez, 2004). If within the same feature/column we have similar mean and median it is indifferent which strategy to use. The use of the median gives a more appropriate idea of data distribution. After undergoing this process, the final sample included 5299 observations.

3.2.3 Feature Selection

The selection of the most relevant predictors is an important step, not only for reducing computational time, but also to compare and contrast with the business perspective, the ECB Risk Assessment Methodology. After cleaning the data and dropping some non-representative features we are still dealing with the total universe of available data.

For the feature selection process we used Random Forest Classifier with an 85% threshold for the feature importance. This method was chosen due to its ability to rank the purity of each node (gini impurity): greatest impurity decrease occurs at the top of the tree (near root level) whereas smaller impurity decrease are is observed at the end (near leaf nodes). When this algorithm prunes below a particular node, it creates a subset of the most important features.

Through this strategy, we are able to technically assess the relevance of each feature regarding the variable we want to predict and select the ones that explain 85% (the importance threshold defined in the algorithm) of our target variable. The

final dataset has a total of 3409 features selected from a universe of 82559 predictors, and 5299 observations.

Afterwards, we compare the similarity of the obtained features with the ones the methodology highlights. This, per se, is a useful analysis since it gives hints to the analysts on which indicators to monitor more closely.

For the purpose of reducing computational time we have also considered, at first, the Principal Component Analysis (PCA). Although this method is associated with dimensionality reduction, its use compromises model explainability. At the same time, PCA loses track of the features that better represent our target variable, by projecting the feature space into a lower dimensional space.

At the end of this process we compute the correlation matrix of the dataset to assure there is not a high correlation between features and target. This would suggest that a certain feature represents the same phenomena as the target. The correlation indices range between a positive 26% and a negative 32%.

3.2.4 Experiments

The experiments carried out to assess and compare the performance of each model were organised in three separate phases, each of which is explained in the following subsections. First, we adopted the most straight-forward approach of splitting the data into two sets, the train and test sets. Afterwards, we use cross validation to measure the average performance of each model, considering every observation for either training or testing. Finally, we use an auto-ml library, TPOT (Olson et al., 2016), to have another evaluation perspective.

For each of the three approaches we calculate a measure of performance/scoring for both train and test sets. Furthermore, we compute the confusion matrix for a precise picture of each model's prediction.

We have selected a list of some of the most common machine learning algorithms used for classification problems. For the purpose of these experiments we have selected scikit-learn implementation of the following models:

1. Logistic Regression (LG) by Cox (1958), or Multinomial Logistic Regression, is an extension of the Bivariate Logistic Regression proposed by McCullagh and Nelder in 1989 (Glonek and McCullagh, 1995) for problems with more than two discrete outcomes. The original approach was designed for binary problems, and the target variable was modelled through a binomial probability distribution function. In its multiclass form, the probability is distributed by the number of classes of the problem at hand. In this paper, we used the scikit-learn implementation of the Logistic Regression for multi-classes (Pedregosa et al., 2011).

2. Support Vector Machine Classifier (SVC) - or Multi-class Support Vector Machine - is a generalisation proposed by Weston and Watkins (1998) of the binary classification Support Vector. Instead of computing the probability of an observation corresponding to a certain class (like the Logistic Regression), this method represents all datapoints in an n-dimensional space, and aims at creating a boundary, called a hyperplane, that separates the datapoints into classes. The algorithm tries to maximise the distance between the boundary and the nearest datapoints. Real-world data is seldom linearly separable, so

it becomes computationally expensive to project all data into a higher dimensional space for calculating the distances to the optimal boundary. To overcome this computational hurdle, SVM uses the kernel trick, a method that uses a kernel function that takes two vectors/datapoints in the original space and computes their dot product in the feature space. Since the vectors are normalised the result is related to the Euclidean distance of both vectors - the distance we wanted to compute. In other words, this method shortcuts the computation of the distances from the datapoints to the possible hyperplanes, by performing them in the original n-dimensional space, thus reducing wall time (Adankon and Cheriet, 2009). We have used scikit-learn implementation of SVM based on libsvm library.

3. Naive Bayes Classifier (NBC) is a supervised learning method based on Bayes theorem, based upon the statistical independence of features. This simplified approach to learning shows it is up to par with more sophisticated classifiers, namely when dealing with high dimensionality and complex classification problems (Rish, 2001). Naive Bayes algorithms are thus very efficient to train and require little data to converge. This derives from the fact that they only require to compute the probability of each class, the conditional probabilities of each input value given a certain class, and the mean and standard deviation values of each attribute for each class. In this paper, we use the Gaussian Naive Bayes implementation from scikit-learn which presupposes a Gaussian distribution of the features.

4. Random Forest Classifier (RFC) is a learning method that combines tree predictors working together to minimise the error (Breiman, 2001). As thoroughly explained by Fawagreh et al. (2014), each decision tree in the forest is a base classifier using a sample of the instances in-bag, hence the bagging technique. The trees are combined through a voting system - one vote per tree - where the forest chooses the class with most votes. Another aspect that improved the randomness of the trees was the use of the Gini index - features with the highest index are used to split the inner node of the tree. This algorithm presents great results when dealing with data noise and avoiding overfit, and handles large datasets with high dimensionality. Here again, we are using its scikit-learn implementation.

5. Extreme Gradient Boosting (XGBC) Classifier proposed by Chen and Guestrin (2016) is a machine learning algorithm used for tree boosting that uses data compression and sharding (a data partition technique) to scale to large amounts of data. Due to its capability to avoid overfitting and its efficient use of large amounts of data, it has become one of the most popular ML methods in the last few years (Sahin, 2020). Having Friedman (2001) gradient boosting technique as its pillar, XGBoost uses a differentiable loss function and optimises it with gradient descent algorithm, in order to build an ensemble of classification trees. For this algorithm, we have used the authors' implementation package (Chen and Guestrin, 2016).

The TPOT auto-ml library automatically selects the best model and we use that result to compare with the others.

In order to have all features in a similar scale we have applied a scaling method when preprocessing the data. MinMaxScaler was the best choice since it preserves the shape of the original distribution. It does not significantly change the information embedded in the original data. Note that MinMaxScaler does not reduce the importance of outliers. The default range for the feature returned by MinMaxScaler is 0 to 1.

Here we present a list of the main characteristics of the experiments' environment:

1. Lenovo ThinkPad P50 with an Intel Xeon processor (2.8GHz), 32 GB of RAM, 1 TB SSD;
2. Windows 10 64-bits;
3. Python 3.9.1 64-bits;
4. Pandas 1.2.0;
5. scikit-learn 0.24.0;
6. TPOT 0.11.7.

Performance measures

We used two different tools for comparison purposes: the confusion matrix and the f1-score. The confusion matrix is the most detailed view of how a particular machine learning model is performing in a classification problem (Tharwat, 2018). Through this tool, we are able to assess each of our model's predictions and compare them with the correct value.

	Population = P + N	Predicted class	
		Positive (PP)	Negative (PN)
Actual class	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Figure 3.2: Example of a confusion matrix for a binary classification problem.

Figure 3.2 shows a generic matrix for a binary classification problem where we can observe each possible classification:

- True positive (TP) corresponds to the model correct hits.
- False negative (FN) represents every missed case, where the model underestimated.
- False positive (FP) represents false alarms, where the model overestimated.
- True negative (TN) corresponds to the correct rejections made by the model.

For our specific problem where we have four classes representing the risk levels for liquidity, we will have a *4X4* matrix for each model, which is simply a generalisation of the one just presented.

There are several metrics that one can extract from these statistics. However we will focus on the f1-score and two others derived from it, precision - or positive predictive values, that is the number of positive results that are true positives - and recall - also known as sensitivity or true positive rate, which measures the number of positive hits among all the positives:

- f1-score represents the harmonic mean of precision and recall. It is most suited for uneven class distributions, as is the case of our dataset . It is calculated as

$$f1 = 2 * \frac{precision * recall}{precision + recall} \tag{3.1}$$

where

$$precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$recall = \frac{TP}{TP + FN} \tag{3.3}$$

Train-test split

Our first approach to evaluating the performance of each model is through a train-test split of the available sample data. As a general principle, we used 80% of the data for training and 20% for testing.

The assessment is organized as follows:

1. Use the MinMaxScaler, as specified above;
2. Iterate through all machine learning models;
3. Fit the model to the data;
4. Assess train and test scores;
5. Compute the confusion matrix;
6. Store the results.

Cross-validation

When we are dealing with small to medium datasets a simple train-test split will most likely misrepresent our real-world problem by missing some classes. This is the main indication for using cross-validation, where every single observation is eligible for the train and test sets. The technique consists of splitting the dataset into a specific number of folds, or partitions, and iterating through the partitions.

In Figure 3.3 we picture how a 5-fold cross-validation example would process. First, the dataset is split into 5 folds. Then, in each of the five iterations, one of the folds assumes the role of *test fold* and the other four as *training fold*. In each iteration, the machine learning algorithms are trained on the *training fold*, and their performance is assessed on the *test fold*. By the end of the five iterations, the average

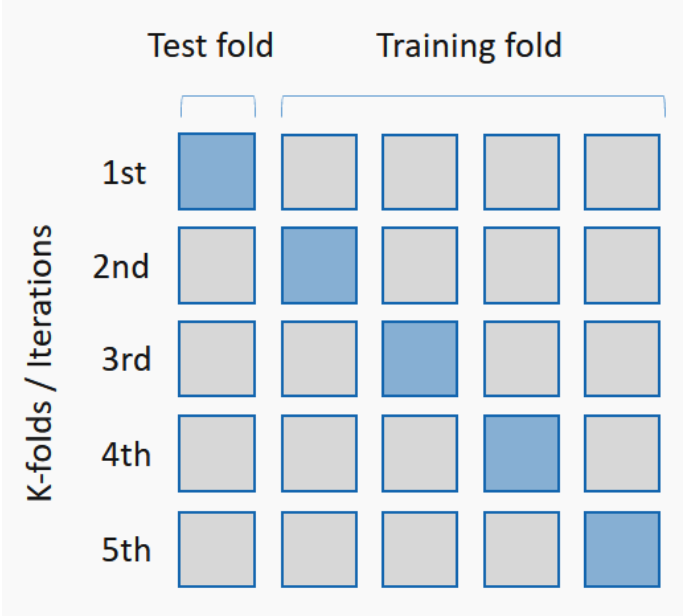


Figure 3.3: Example of a 5-fold cross-validation process.

of the performance obtained on each iteration is the value considered for comparison. Cross-validation is the preferred method for assessing model performance because it gives models the opportunity to train on multiple train-test splits. This will better indicate how well a model will perform on unseen data. Conversely, a simple train-test split is dependent on just a single data split which can overestimate the overall performance.

In this experiment we used StratifiedKFold (a form of cross-validation) to preserve the percentage of samples among classes. The purpose of this specific form is for the test to be as close as possible to the whole dataset. The stratification ensures class frequencies of the partitions are equal to the complete dataset. This is particularly advantageous in an imbalanced dataset scenario, where this method ensures every class is represented.

The use of cross-validation can also raise some issues. Since we are assessing performance of a model on several splits, situations may arise where data leaks from one iteration to another. In other words, data leakage can happen when we are learning from both the testing and training set. If we do any pre-processing outside the cross-validation algorithm, we will bias our results and most likely overfit our model. To avoid this common problem we feed our cross-validation cycle the entire dataset and perform every transformation within each iteration. Although the authors concede that this repetition takes its toll on performance, the extra step assures no data is leaking from each of the splits or iterations.

F1-score is used as a performance measure since it keeps a balance between precision and recall. Furthermore, since we observe uneven class distribution in the dataset, F1-score is more appropriate than AUC (F1 gives a score for a specific threshold, whereas AUC averages over all possible thresholds).

Confusion matrix was selected as the best tool for describing performance on a classification model. This is an NxN matrix where N is the number of classes in our classification problem (as mentioned earlier, classes 1, 2, 3, and 4 representing the risk tiers for any given financial institution).

Similarly to the train-test split approach, we cycled through each model as follows:

1. Define a pipeline to streamline scaling, using the MinMaxScaler, and training;
2. Define a 10-fold cross-validation process using StratifiedKFold;
3. Use cross-val-predict to obtain predictions for each element (we want to compare and contrast predictions obtained from different models);
4. Compute the F1-score as well as the confusion matrix;
5. Store the results.

TPOT - An AutoML approach

As noted by Zöller and Huber (2019), AutoML has been around since the early 90's with the automated selection of algorithms via grid search - an exhaustive search method that automates the process of trying all possible combinations of a given set of hyper-parameters. However, only in 2018 did the first commercial full-pipeline solutions come to light. In order to keep up with these developments and to provide another perspective on how to approach this research, we used an autoML tool, TPOT, that optimizes machine learning pipelines through genetic programming.

The use of this framework is straightforward and the parameters used were:

1. **generations** is an evolutionary computation concept that determines the number of iterations in the optimisation process. It gives the tool more time to find an optimal solution. This parameter was set to 5;
2. **population_size** is also an evolutionary computation concept that represents the number of possible solutions (as a subset of the total population, or number of solutions) to be considered in each generation. This parameter was set to 100;
3. **cv** specifies the number of cross-validation folds in a StratifiedKFold. This parameter was set to 10;
4. **verbosity** is a definition of how much information TPOT will provide during run-time. This parameter was set to 5;
5. **random_state** is a random number generator to assure TPOT will provide the same results given the same inputs. This parameter was set to 42.

This framework takes the full dataset and saves a portion (in our case, predefined as 20%) of randomly selected observations for validating the best model. Both the model's score and its confusion matrix are assessed based on the unseen data.

3.3 Results and Discussion

In this section we present how the selected models compare when solving the liquidity risk problem, organised by approach: train-test split, cross-validation and autoML tool (TPOT).

Supervisors were closely engaged in this process, namely validating the underlying methodology and contributing to the feature selection process. One of the innovative aspects of this work is the risk assessment methodology used to classify each bank - the quantitative pillar of ECB's Risk Assessment Methodology - that assigns a score of one to four to each bank according to its liquidity risk level: low, medium-low, medium, or high risk.

RAS is used for all banks in the euro-area, and it is developed and maintained through a joint task-force from the ECB and the National Central Banks (NCBs). To the best of our knowledge, there are no studies using this methodology, nor considering this multi-class risk score.

The processing and evaluation times for each of the approaches were:

1. Train-test split: 1 minute and 18 seconds;
2. Cross-validation: 8 minutes and 50 seconds;
3. TPOT framework: 21 hours, 34 minutes and 51 seconds.

The execution times were measured using *%%time* python statement within jupyter notebook. This statement returns the wall time for the cell under evaluation. In this case, each approach is in a cell of its own.

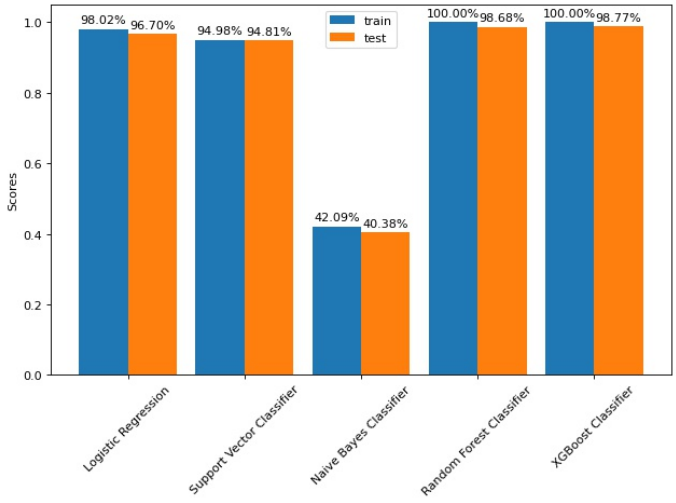
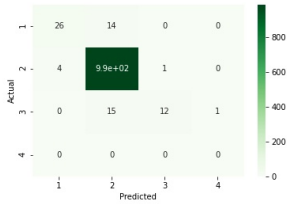


Figure 3.4: Precision scores of each model, using train-test split approach.

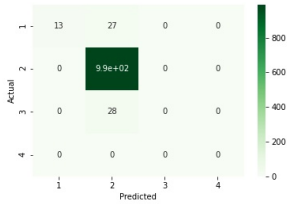
Figure 3.4 pictures the train-test split approach. This graph compares each of the selected models, contrasting their train and test scores as well as the models amongst themselves.

The overall picture stresses that Naive Bayes classifier is inadequate for this problem, at least without further tuning. In the authors' opinion, the hypothetical gains do not compensate for the processing time required for the tuning process.

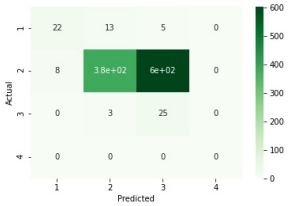
Figure 3.5 gives a more exact notion of each model's precision through its confusion matrix. The Logistic Regression (3.5a) presents a solid score. Nonetheless,



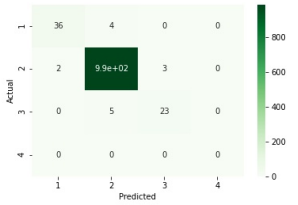
(a) Logistic Regression.



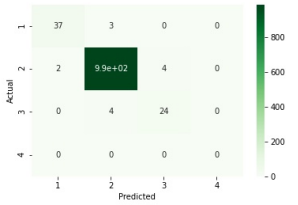
(b) Support Vector Machine classifier.



(c) Naive Bayes classifier.



(d) Random Forest classifier.



(e) Extreme Gradient Boosting classifier.

Figure 3.5: Confusion matrices generated when evaluating the above mentioned models, using train-test split approach.

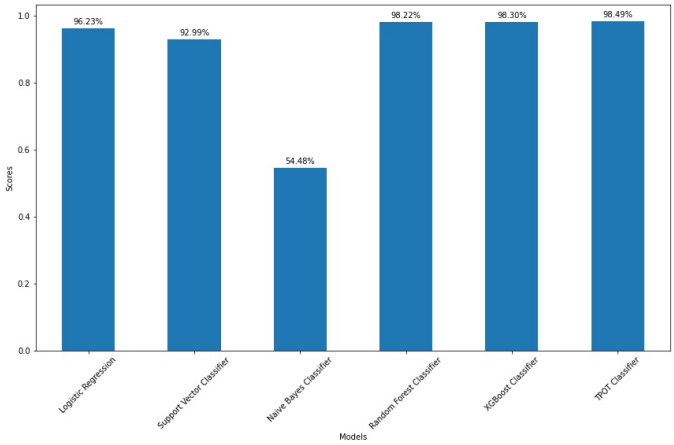
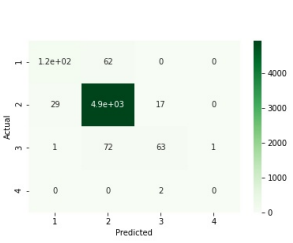


Figure 3.6: Precision f1-scores of each model, using cross-validation approach.

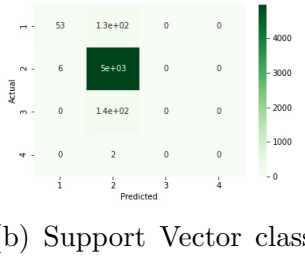
its confusion matrix clearly demonstrates a difficulty detecting a risk score of 3 or 4. The Support Vector classifier (3.5b) fares even worse, not detecting any scores of classes 3 or 4, and predicting very few class 1 scores.

Random Forest (3.5d) and XGBoost (3.5e) both show consistently good performance. Unsurprisingly, however, they both make similar mistakes, confusing predominantly the same risk classes. These results imply the need for a more robust approach.

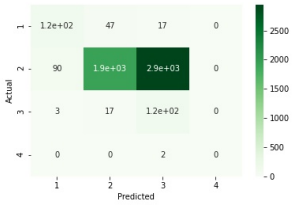
Figure 3.6 shows the comparison of the same models using the f1-scores to evaluate the cross-validation process. In relative terms, the picture shows the same distribution as the train-test split approach, although with cross-validation we have used the whole dataset. The reasoning behind this is that cross-validation already considers several train and tests splits (k splits for k-fold cross-validation), ensuring that the final score is not biased to a particular random split, and we are not



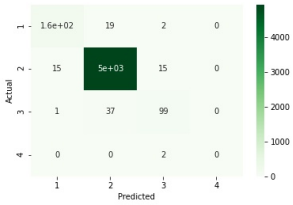
(a) Logistic Regression.



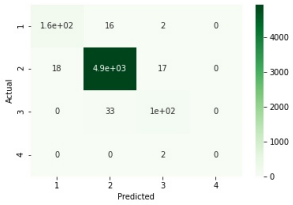
(b) Support Vector classifier.



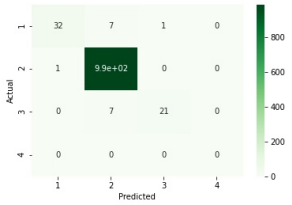
(c) Naive Bayes classifier.



(d) Random Forest classifier.



(e) Extreme Gradient Boosting classifier.



(f) TPOT classifier autoML framework.

Figure 3.7: Confusion matrices generated when evaluating the above mentioned models, using cross-validation approach.

excluding observations that can compromise the final score.

On the same premise as before, we use the confusion matrix to better assess each model's precision. Figure 3.7 provides a comprehensive view of the classifications accomplished by each model. Although the Logistic Regression (3.7a), used as a benchmark, shows a 96% f1-score with cross-validation, when we dive into each individual classification we realise that this is not the case. Class 2 shows 99% of correct classifications. However, classes 1 and 3 show 39.9% and 48.4% of misclassifications, respectively.

Support Vector Classifier (3.7b) and Naive Bayes (3.7c) models both show several misclassifications. The former shows good results detecting class 2, but not the other classes. The latter, although it got the worse f1-score, shows an average performance regarding classes 1 and 3. Class 2 presents 64% of misclassifications.

Random Forest classifier (3.7d) and XGBoost (3.7e) present very similar results. The misclassifications occur in the same classes and differ only by a couple of observations. This can be explained by the fact that both models are based on decision trees.

TPOT classifier (3.7f) takes a step further by significantly increasing precision in classes 1, 2, and 3. Although this precision gain comes at the cost of 21 hours of training, this approach proves itself worthwhile due the financial impacts that liquidity risk assessment might have.

Contrarily to cross-validation, where we used the full dataset to train and validate the best model (see section 3.2.4), with TPOT, we set aside part of the dataset to validate the scores of the model as well as its confusion matrix. This explains the smaller number of observations in its confusion matrix when compared to others in figure 3.7.

One aspect that is common to all models, regardless of the approach, is the fact that none correctly classifies class 4. This can be easily explained by figure 3.8 that shows how imbalanced our dataset is - by providing only two class 4 observations, we

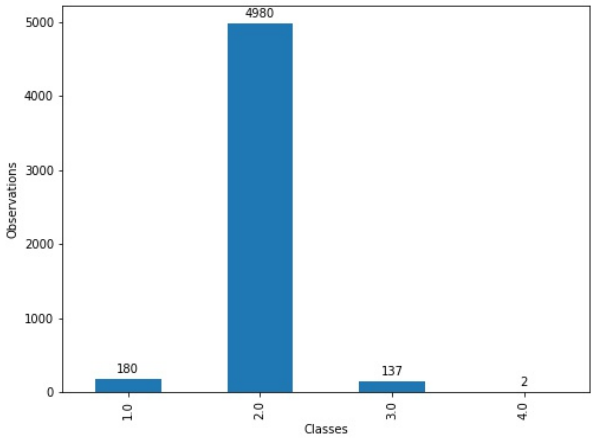


Figure 3.8: Number of observations per target class in the dataset.

make it almost impossible for the models to establish a pattern. We considered using oversampling/undersampling techniques, however, we decided not to due to the nature of the problem and the fact that this represents the frequency of occurrence of each class in the real world.

As an overall consideration, we can add that the machine learning techniques clearly outperform the traditional Logistic Regression approach. Although we can be misled by a high score, an in-depth analysis of the confusion matrices clearly expose the classification advantages of ML techniques when compared to classic statistical methods.

These findings can bring a great advantage for regulators when considering risk assessment tools. The best performing model can be set up as a decision support system, either as stand-alone stress-testing tool, or as part of an EWS. Furthermore, this research will open a new path to address and support the overall risk assessment exercise, as part of the SREP process. If all other risk perspectives in this methodology present similar results, this work will foster a new approach to the SREP decisions, by combining all its components and establishing a baseline for each institutions' capital requirements.

3.4 Conclusion

A proper risk assessment methodology requires a thorough evaluation of a credit institution's practices as well as an integrated analysis of the uncertainty factors involved. In this study, we focused on liquidity risk assessment through the lenses of the SREP methodology established by the ECB. This step is the linchpin of this work since it establishes a widely accepted methodology for risk assessment.

3.4.1 Practical and theoretical implications

This paper intends to contribute to the growing body of knowledge regarding the use of ML techniques on sup-tech solutions. It specifically proposes a comprehensive and innovative approach, that considers all the knowledge expressed in historical data to

support and envisage a critical European supervisory business process (ECB RAS).

The novelty of this work comprises two main vectors:

1. The use of a European-wide risk assessment methodology, the quantitative pillar of ECB's Risk Assessment System, guaranteeing a transparent approach to risk, as well as comparable results across the Euro-area. We classify the observed banks according to their risk level.

2. A dataset that represents the most up-to-date reality in the Portuguese supervisory context: we use real-world data retrieved within the scope of the SSM context from 2014 to March 2021.

Based on these two pillars, we compared several well-established machine learning algorithms to a traditional statistical method to evaluate which would best model this decision process.

Results showed that not only we can model this classification problem, but also the ML techniques used in this work clearly outperform the classic approach. Moreover, XGBoost stands out as the best quick approach to solve this classification problem. Conversely, the autoML framework TPOT takes 21 hours to evaluate but with very few misclassifications. Given that liquidity risk is an extremely sensitive aspect of a credit institution's health, we believe the precision gains compensate for the added processing time.

This paper intends to be the outset of new approach to risk assessment from several perspectives:

- For NCAs, an EWS based on this model can significantly increase the robustness of the RAS decision process.

- The findings of this paper can leverage the application of this methodology to other perspectives, like market, operational, rendibility and credit risks.

- This work also benefits banks and consultancy companies when implementing similar decision support systems. Although the underlying methodology is confidential, banks have the required data and the scores assigned by the NCA. This should enable them to better analyse their risk profile in terms of regulatory compliance.

3.4.2 Limitations and future work

Throughout this study we have identified several aspects that could be revisited in order to improve the results. The first issue is related to the dataset. The fact that the dataset is imbalanced hinders the detection of some classes. As shown in the previous section, class 4 contains just two observations which makes it difficult to model that particular decision process.

Another limitation is the fact that this dataset only reflects the Portuguese context. Further investigation would benefit from the use of all central banks' data, thus reflecting a broader picture of the supervisory landscape. Furthermore, the increase in the dataset size would strengthen the validation of the ML models.

The inclusion of quantitative data from non-supervisory frameworks is also extremely beneficial for improving the model's robustness. Data from financial markets, payment systems and macroeconomic indicators provide a context that supports the overall risk assessment.

We also find relevant to include expert judgement to reinforce to final risk assessment. To this end, qualitative data sources like internal notes and risk assessment reports, as well as risk scores reassigned by the supervisors should contribute to the model's learning phase.

Finally, we believe the other risk perspectives comprised in the SREP methodology should also be addressed using the same methodology. Ultimately, combining all risk perspectives could be a stepping stone for regulators as a support of the SREP exercise.

Chapter 4

Approaching European Supervisory Risk Assessment with SupTech: A Proposal of an Early Warning System

4.1 Introduction

In recent years, the use of decision support systems has skyrocketed, with machine learning (ML) spearheading the change. The financial industry has always been one of the main drivers for that development (Zopounidis et al., 1997). As the amount of data collected soars and computing power rises to meet the challenge, the use of classical statistics such as linear and logistic regressions is gradually declining. Although they were once the mainstay of decision support systems, nowadays they tend to be recalled sporadically, and mainly for their better comprehensibility in comparison to most ML models (Yang and Wu, 2021). The current research problem is how to leverage on ML to support risk assessment processes at central banks, using quantitative supervisory data.

Recent uses of machine learning have unveiled data patterns that were as of yet undiscovered (Huang et al., 2021). These applications have also expanded to the fields of regulation and supervision, as described by Hertig (2021). For supervisory purposes, there has been a huge increase of interest in developing sup-tech tools. As Beerman et al. (2021) reported, the number of ongoing ML projects in this field skyrocketed from 12 in 2019 to 71 as of December 2021. The pandemic forced an off-site approach to what was previously required to be done in person. In the past two years we have seen an increasingly higher number of production-ready systems applying ML to support central banks' tasks (Massaro et al., 2020). From the specific standpoint of supervision, the work from Filippopoulou et al. (2020) is a watershed in EWS development at central banks, using EBC Macro-prudential Database to address credit risk. This work, along with the EWS proposed by Pompella and Dicanio (2017), supports the importance of these systems to support rating assignments and alert for distress signals.

The amount of data retrieved in the supervisory framework is overwhelmingly high (Authority, 2013). Additionally, supervisors often ask for complementary information, either quantitative or qualitative. Even thought National Central Banks

(NCBs) are equipped with business intelligence systems that allow them to organise most quantitative information, data analysis is mostly done on a *ad-hoc* manner that is impractical for a prompt spotting of risky events (Broeders and Prenio, 2018). Besides, this method only looks at past events, making it impossible to systematically test alternative economic scenarios. Furthermore, we must mention that risk methodologies might vary, making it difficult to compare not only the assessments, but also the evolution of the classifications.

Traditional approaches already set out an organised perspective of the reported data, through dashboards and reports that provide aggregated and specific views of key indicators (di Castri et al., 2019). However, these approaches only consider past events and they are constrained by the regulatory framework (not to mention, they lack what-if analysis and decision processes built on that data). The use of innovative technologies to support supervisory processes is defined by Broeders and Prenio (2018); Doerr et al. (2021) as sup-tech, and these authors summarise the barriers of adoption in three main items:

1. Frequent regulatory updates;
2. Conservative industry;
3. Lack of qualified human resources.

From a data perspective, Early Warning Systems for predicting banking crisis have also been in the spotlight. Casabianca et al. (2019); Consoli et al. (2021) are some of the many examples of landmark findings in that area, along with the previously mentioned Filippopoulou et al. (2020). However, none of these authors explore the information available in the European supervisory framework.

In a previous work, we have addressed the issues of using a single risk methodology, selecting literature-supported ML models to evaluate the risk level of banks, and using up-to-date real-world supervisory data from the Portuguese banking sector (Guerra et al., 2022). The previous work addressed the concept of liquidity risk since it is crucial for a bank's ability to operate (Vento and Ganga, 2009) and it can render a bank nonoperational in a matter of days (Shah et al., 2018). In our paper we expand previous findings to the other risk perspectives comprised in the Supervisory Review and Evaluation Process (SREP).

In our current study, we have extended the sample from March 2014 until August 2021. This data is extensively validated by *Banco de Portugal* and European Central Bank (ECB) quality assurance processes. The quality of gathered information allows for accurate assessment, thus ensuring a positive correlation between risk prediction and the observed phenomena (Ng, 2011).

Another key component of our approach is the way we set up the classification problem. Contrary to what is commonly found in the literature, we reiterate the importance of considering a multi-tier classification approach to this problem. Our data being provided by real world context, we feel highly confident in expanding from the *fail/no-fail* classes and adopting the four classes comprised in the RAS methodology, a European-wide risk assessment methodology:

1. Low-risk;
2. Medium-low risk;

3. Medium risk;

4. High risk.

Our work also showcases literature-backed ML models for structured financial data that support the efficiency of supervisory processes.

Based on the findings of our study, we provide a comprehensive guidance for the development of valuable supervisory use-cases enhanced by innovative techniques.

The purpose of this work is to leverage on the above-mentioned aspects, and expand the academic body of knowledge of quantitative risk assessment for prudential supervision. From a supervisor's standpoint, we aim to bring better insights into the data and attain higher efficiency - automating resource intensive tasks and freeing up analysts for more integrative analysis (Beerman et al., 2021). As pointed out, there is room for improvement in this field, since less than 25% of sup-tech systems are exclusively intended for quantitative purposes. Following in that lead, this work develops a methodology to address each of the risk perspectives in the RAS methodology: credit, market, operational and profitability.

4.1.1 Related work

The use of machine learning for risk assessment has been a highly debated topic, both from an academic and industry standpoint. Since the 2000s (Galindo and Tamayo, 2000), risk assessment has been recurrently identified as a top-priority investment for developing the data literacy of financial institutions. As recently shown by Antunes (2021), risk assessment by central banks is paramount for accurate supervision and less biased then the self-assessments carried out by the banks themselves.

Additionally, Galindo and Tamayo (2000) established that tree-based models perform consistently better than artificial neural networks (ANNs) considering structured financial data. This finding is one of the pillars of our approach and it has been confirmed by several other authors (Xia et al., 2017; Chen and Guestrin, 2016; Climent et al., 2019).

In their literature review, Leo et al. (2019) highlighted the popularity of machine learning applications for risk management in banking industry, while also noting the experimental nature of most approaches. The authors also remark the discrepancy between the high level of academic research concerning this area *versus* the *de facto* industry applications.

This debate has focused around two main issues:

- Finding the right risk assessment measure.

- Finding the adequate machine learning algorithm to build a risk assessment model.

Guerra and Castelli (2021) studied both of these aspects appraising several methodologies for assessing distress signals. This review spans from 2004, when Hillegeist et al. (2004) turned the page on two landmark methods (the Z-score (Altman, 1968) and the O-score (Ohlson, 1980)) by proposing the use of the Black–Scholes–Merton option-pricing model, up until 2019, when Kou et al. (2019) listed the most common methodologies for assessing systemic risk on the financial system.

On the same topic, Climent et al. (2019) used XGBoost to identify the best predictors of bank failure and develop a classification model to label failed and non-failed banks in the Eurozone. The data used in their study comprised 25 annual financial ratios for commercial banks.

The majority of current literature converts the risk assessment problem into a binary classification task, where each bank is labelled as "failed or likely to fail" or "no fail" (Climent et al., 2019; Kolari et al., 2019; Leo et al., 2019; Filippopoulou et al., 2020; Wang et al., 2021). These studies usually rely on public datasets, where the target variable is derived from a set of financial ratios.

At central banks, as clearly pointed out by Stock and Watson (2001), economists are responsible for conducting risk analysis and performing scenario testing.

Since the appearance of the Single Supervisory Mechanism (Commission, 2015) we are bearing witness to a standardisation of reporting requirements and methodologies. The heterogeneous landscape of financial performance measures identified in the literature has been increasingly replaced by the use of the Supervisory Review and Evaluation Process (SREP) (Bank), leading us to leverage on this risk assessment methodology. SREP is an ongoing work by the European Central Bank (ECB) and the National Central Banks (NCBs) that provides an integrated view on each bank according to five risk perspectives: liquidity, credit, market, operational and profitability. The Risk Assessment System (RAS) is the quantitative pillar of the methodology, and it is the focal point of this work.

Selecting the adequate machine learning methods applied to central banking, we found that it recently became a hot topic from both an academic and NCBs standpoint (Lee and Shin, 2020; Huang et al., 2021; Wang et al., 2021; Alonso and Carbo, 2020; Antunes, 2021). Beerman et al. (2021) report that the pandemic prompted NCBs to rely on sup-tech solutions in their everyday processes. Several of the surveyed authorities already have operational systems. For instance, Central Bank of Brazil has a tool that examines the whole credit portfolio of a bank to detect exposures with unrecognised expected losses; Bank of Spain is applying inference maps to model the relationships between borrower and evaluate the risk impact; and the Monetary Authority of Singapore is developing a tool to automate data analysis so that supervisors can rely on complete datasets, instead of sampling. For this reason, we expanded our research to applications of ML to risk assessment. By broadening this research, we can evaluate how ML has been used for financial structured data and then focus on the central bank case.

Stress-testing is one of the many forms of risk assessment that is particularly used at central banks. Kolari et al. (2019) challenged the concept of a bank's resilience by suggesting that it mostly represents a bank's ability to deal with a specific risk supported by its own capacity to absorb it. In such a setting, applying a risk-focused methodology like SREP allows supervisors to better assess the root causes of what might otherwise be perceived as a general business model issue.

Chakraborty and Joseph (2017) presented a series of ML applications for financial problems and they analysed the most frequently used algorithms, like tree-based ensembles, artificial neural networks and clustering techniques. The authors also showcase three use-cases at central banks, that establish ML as a better solution than traditional statistics. The most relevant for our work is one that develops a series of alerts (EWS) based on the balance sheet structure of a bank, in a supervisory context. This shows not only how relevant supervisory data is for a proactive risk

assessment, but also how it can be used to sense the risk proclivity of supervised institutions.

Recent technological developments have allowed newer and more complex models to emerge (Strydom and Buckley, 2019), such as deep learning (DL) and extreme gradient boosting (XGBoost) (Abellán and Castellano, 2017). Evidence shows those analysis methods have a unique capacity of capturing the intricacies of financial phenomena (Ribeiro et al., 2012; Huang et al., 2021).

Iturriaga and Sanz (2015) showed that modelling time series is where DL excels. Also Petropoulos et al. (2018) leverage on DL's precision and develop an Early Warning System (EWS) for predicting failure of Greek banks (data in 2005-2015). This is a landmark report on the use of advanced ML in a daily supervisory context. Wang et al. (2021) proposed an add-on to the conventional logit-based EWS, which involves simulating expert voting through a Random Forest based system, and that showed valuable results in predicting systemic crises.

Broeders and Prenio (2018) organise supervisory innovation concepts and present a series of use-cases where early adopters are implementing innovative approaches (sup-tech), converting retrieved data into predictive indicators. These works are of great importance to systematise how to implement this technology. The increasing amount of available data is one of the main drivers for the development of ML-based systems, as Chakraborty and Joseph (2017) also have claimed. Banking supervision acknowledges the benefits of innovative technologies and the importance of keeping up with the variety of sup-tech initiatives being developed. These initiatives have the potential to dramatically change the supervisory process; anticipating the consequences of current behaviours instead of belatedly reacting to past events. The same authors explore several use-cases from the Central Bank of the Republic of Austria (OeNB), Monetary Authority of Singapore (MAS), Securities and Exchange Commission (SEC), among others. Business process effectiveness, cost reduction and increased analytical capability, are noted as the main drivers for the sup-tech endeavour. These supervisory agencies report several challenges in exploring and implementing these technologies, such as:

- The technical know-how and appropriate infrastructure to support these analytical solutions;
- The legal framework to support the use of the relevant information;
- The internal support from management to invest in these initiatives and from the end-users, to provide the expert knowledge and to use and promote the new analytical tools.

Board (2020) also shows how the balance of supply and demand ignited the development and use of sup-tech tools. From the demand side, these authors mention, among other aspects, enhanced supervisory and regulatory requirements and improved risk management capabilities, where the automation of data retrieval and summarisation can drastically improve supervisory processes. From the supply side, the increasing availability of data and new analytical methods are among the top supporters of the above-mentioned regulatory necessities. Listed benefits of implementing these ground-breaking tools include:

- Enhanced analytical capabilities;

- Enriched visuals, stemming from state-of-the-art data collection to sophisticated dashboards;
- Reduced costs, as a consequence of automation.

Nevertheless, adopting new analytical processes inevitably brings on fresh challenges. Recognising this aspect, Jagtiani et al. (2018) expand on the impacts of these new analytical solutions and possible risks of adopting them, such as:

- Third-party vendor risk, where banks give access to outside specialists - data scientists and business users involved in setting up the tool - that can lead to data breaches. Additionally, if the vendor is a dominant player in the market, that circumstance can create a single point of failure in the financial system.
- Cyber-security risk, which is related to the previous topic, as vendors might not comply with supervisors' security requirements. Additionally, by allowing for external sources of data, banks and central banks become exposed to that channel and the information therein contained.
- Model risk, where systems based on complex machine learning models or even black-boxes make decisions that might not make sense from a business perspective, hence providing wrong predictions.

Another factor with major impact in ML use is the comprehensibility of the models. Although ML models are seldom capable of explaining prediction, they consistently outperform the classic approaches. Dastile et al. (2020) published a systematic literature review contrasting these techniques for a credit scoring problem, and they stress the lack of interpretability of DL as the main barrier for adoption in supporting financial decisions.

It is worth bearing in mind that pointing out the direction of future research is as important as signalling risks associated with implementing ML models. Kou et al. (2019) present a thorough report on state-of-the-art applications and ML techniques to assess systemic risk. Based on the existing technology, they suggest several future work areas, like big-data analysis, data-driven research and policy analysis with data science.

4.2 Methodology

Developments in the area of data science and machine learning usually fall into one of two categories: developing a new computational method to better solve an existing problem; or alternatively, using the existing methods to address a new problem. In this work, we aim to address a problem that was yet to be solved using machine learning: supervisory risk modelling.

Figure 4.1 illustrates how we attained our objective in a step-by-step diagram, as a development of what was presented in Guerra et al. (2022). The first step comprises a data retrieval process from *Banco de Portugal* supervisory data system, including a wide set of features and the target variables we want to model. Next, there is a transformation process that is responsible for cleaning the data, dealing with missing values and selecting the most significant features. In the following step, we compare the ML models for this task using train-test split, cross-validation

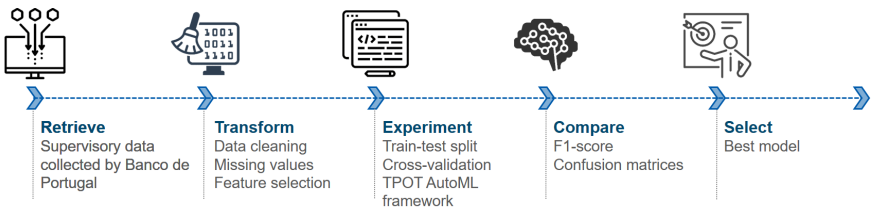


Figure 4.1: Methodology process overview.

and the TPOT AutoML framework (Olson et al., 2016). The f1-score and confusion matrices are used to compare the results and select the best model that can then support an Early Warning System for the RAS risk perspectives.

In this section we present the steps carried out in this research, beginning with explaining how the data was retrieved, what transformations were required, which features were selected and its criteria, and finally, how the models' performance was evaluated.

4.2.1 The Data

One of the main pillars of this paper is the supervisory data collected by *Banco de Portugal* (BdP) within the Capital Requirements Regulation (CRR) and Capital Requirements Directive IV (CRD IV) (Parliament, 2013). Our data ranges from March 2014 until August 2021 and most of the data used for the purposes of this paper is quarterly (Authority, 2013). Due to confidentiality issues, the dataset used in this study cannot be made available for public consult.

Data is extracted via an SQL query from BdP's production database (with no filters regarding reference date, banks or their consolidation level) into a *comma-separated-values* (csv) file. The result set is imported using a Python script within Jupyter notebooks. An extraction routine was implemented to assure consistency and automation in data gathering.

To account for all possible predictors, we have selected our feature space from the four main reporting frameworks for banking supervision: Financial Reporting, Common Reporting, Asset Encumbrance and Funding Plans.

The data resides in a relational database where each row represents a reported value. This means that in the data source, several rows represent a single observation. During extraction, data is anonymised using MD5 algorithm within a SQL's hashing function. This step assures the same identifier for every row in the same observation. The extracted dataset follows this column schema:

1. **ID** - a hash code representing each observation's identifier;
2. **variable** - a code with business meaning that represents each reported value;
3. **val** - the actual numeric value of the variable.

4.2.2 Transformations

Preparing the data for machine learning algorithms is the single most critical stage in such studies and projects. The first step in our study is to pivot the data with

the aim of having each row corresponding to one observation. This transformation uncovers the sparsity of our feature space, requiring null columns to be dropped.

Another important step is to focus the dataset on the risk perspective to be evaluated. In our study we are addressing credit, market, operational and profitability risks. When investigating one risk perspective - one specific target variable - we drop all the others. This might lead to invalid observations, that is, observations that only made sense for a certain risk. As a consequence, we discard the rows for which the selected target value is null.

Dealing with missing values is the final step of the transformations phase. Our dataset is exclusively numeric and each column/feature has its own distribution. Therefore, we opted for inputting the missing values of each feature with the median, since it provides a more accurate perspective on the data's distribution when dealing with up to twenty percent of missing values (Acuna and Rodriguez, 2004).

By the end of these steps our dataset consists of 9262 rows and 82576 columns.

4.2.3 Feature Selection

As we saw in the previous subsection, this dataset is extremely sparse - here the inaccuracy of the term "extremely" endeavours to capture the fact that this is a wide dataset (more features than observations). Although we have considered using Principal Component Analysis (PCA), this method compromises model comprehensibility. Since it projects the original features into a lower dimensionality feature space, there is always information loss from discarding the components with less variance/information. The selection criteria is based on the covariance matrix, and does not account for the target variable to be studied. As this dataset comprises five different target variables - one per risk - PCA might exclude features regardless of their contribution to a specific target.

To address the above-mentioned issues we have used the Random Forest feature selection algorithm, with an 85% threshold for feature importance. Tree-based models are best to perform this task since they not only take into account the target variable to be explained, but also *a priori* they rank features according to how well they improve the purity of nodes (gini impurity). The closer a node is to the root, the greater impurity decrease occurs (i.e. the "cleaner" data becomes). Contrarily, leaf nodes have smaller impurity decrease. Hence, pruning below a certain node results in a subset of the most important features.

This method allowed us to technically assess the list of features that explain at least 85% of our target variable. From the original total of 82576 features we selected 2608 features - for credit risk. This number varied for different target variables.

As a final check, we have computed the correlation matrix for each target variable to assure features and target were not highly correlated - Pearson's correlation coefficient less that 0.3.

4.2.4 Experiments

In the following subsections we lay out the three approaches followed to assess the best machine learning model:

- Train-test split: simply splitting the dataset in train and test sets.

- Cross-validation: using different partitions of the data to test and train the model on every observation, iteratively.
- TPOT Auto ML: an auto ML framework by Olson et al. (2016), for comparison purposes.

These approaches provide a performance measure that summarises the generalisation capability of every model and allows for a reliable and fast comparison among models. F1-score was used as a performance measure since it keeps a balance between precision and recall. Furthermore, since we observe uneven class distribution in the dataset, F1-score is more appropriate than the Area Under the Curve (AUC) (f1-score gives a score for a specific thresholds, whereas AUC averages over all possible thresholds). For a full detail of each evaluation, the confusion matrices are also provided.

For the purposes of this study we selected and evaluated the performance of each of the following models:

- Logistic Regression (LG); used only for benchmarking;
- k-Nearest Neighbours Classifier (kNN);
- Random Forest Classifier (RFC);
- Extreme Gradient Boosting Classifier (XGBC).

The TPOT framework is an AutoML framework that makes use of genetic programming to optimise the process of finding the best model to the problem at hand. This is a rising trend in the usage of machine learning and we have included it in order to to evaluate its adequacy to this problem.

All three approaches comprise an optimisation phase, where we experiment with a range of values for the hyper-parameters of each of the considered models. For both the train-test split and cross-validation we carried out a 5-fold cross validated grid search for the specific parameters of each model. The TPOT framework has an optimisation step within its pipeline that is fully documented.

Just before feeding the data to the ML algorithms, we used the MinMaxScaler to fit the features in the same scale. This approach preserves outliers and the original distribution of each feature, hence conserving the information embedded in the data.

All the experiments were executed at *Banco de Portugal* using its computing infrastructure. The specifications of the node assigned to these experiments were the following:

- 4 Intel(R) Xeon(R) CPUs E7-8891 v4 @ 2.80GHz, 32 GB of RAM, 1 TB SSD;
- Ubuntu 20.04.3 LTS;
- Python 3.8.10;
- Pandas 1.2.0;
- scikit-learn 0.24.0;
- TPOT 0.11.7.

Train-test split

Train-test split is the standard approach to model evaluation and the one we have used to begin with. The initial three-fold split was 60-20-20 for train, validation and test sets, respectively, and we organised the experiment in the following steps:

1. Prepare the data as specified in the previous subsection;
2. Iterate through the machine learning models considered previously;
3. Fit each model to the training data;
4. Use the validation set to run an hyper-parameter optimisation process;
5. Compute the relevant scoring measures for train and test phases, along with the confusion matrices;
6. Persistently store the results.

Cross-validation

Train-test split provides a good approximation of a model's potential performance on a specific dataset. However, for small to medium datasets splitting the data might prove inaccurate, since the training set will probably misrepresent our universe of events, and overestimate the overall performance of the model.

In order to avoid this pitfall, we have used StatifiedKFold, a specific implementation of cross-validation within *scikit-learn* that preserves the proportion of samples among classes.

Cross validation splits the dataset in a specified number of folds and provides models of each of the folds as train and tests sets. This strategy balances the scores of the several splits, providing a more accurate view of how the model will perform on unseen data.

Despite its numerous advantages, the use of cross validation will concurrently entail difficulties, the most common being data leakage. This happens when the model trains from both training and test sets. The authors avoid this problem by providing the cross validation function the complete dataset and performing the necessary data transformations within each iteration. Arguably this approach comes at a cost, but the benefit of assuring that no data leakage will happen largely compensates for the performance deterioration.

After training the models on the data, choosing the proper performance measures is key to correctly evaluating and comparing each resulting model. For this experiment we have chosen the f1-score as an overall performance measure and the confusion matrix for a detailed view on each model's classification decisions:

- f1-score represents the harmonic mean of precision and recall. It is most suited for uneven class distributions, as it is the case of our dataset. It is calculated as

$$f1 = 2 * \frac{precision * recall}{precision + recall} \tag{4.1}$$

where

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

- The confusion matrix is an NxN matrix with each of the rows representing each class prediction, and the columns each actual value provided. In our case, classes 1, 2, 3, and 4 represent the risk tier of a given bank.

To achieve these measurements, we have evaluated each model through the following steps:

1. Defining a pipeline for scaling and training - MinMaxScaler;
2. Establishing a 5-fold cross validation - StratifiedKFold;
3. Using *cross-val-predict* to assess each model's generalisation capability;
4. Performing a nested cross-validated loop to tune the hyper-parameters of each model;
5. Computing the performance measures - f1-score and confusion matrices;
6. Storing the results.

Optimisation process

In order to improve the robustness of the two previous approaches, we carried out an optimisation step, as mentioned in steps listed above. This hyper-parameter tuning phase is in everything similar for train-test and cross-validation, except for the data used to optimise those parameters. For the latter, we used 20% of the data corresponding to the validation set described in the respective subsection. For the former, we used a double, or nested, cross-validation. This is the preferred way for avoiding the bias created by selecting and tuning a model in the same data (Cawley and Talbot, 2010).

In both approaches we used GridSearchCV with five-fold cross-validation to tune the hyper-parameters of each model. The parameters tuned for the purposes of this work were the following:

- Logistic Regression
  - **C** - 100, 10, 1.0, 0.1, 0.01.
  - **penalty** - none, l1, l2, elasticnet.
  - **solver** - newton-cg, lbfgs, saga.
- k-Nearest Neighbours Classifier
  - **n_neighbors** - {1,2,3,. . . ,21}.
  - **metric** - euclidean, manhattan, minkowski.
  - **weights** - uniform, distance.
- Random Forest Classifier
  - **n_estimators** - 10, 100, 500.

  - **max_features** - sqrt, log2.
  - **criterion** - gini, entropy.

- Extreme Gradient Boosting Classifier
  - **max_depth** - 3, 7, 9.
  - **n_estimators** - 10, 100, 500.
  - **learning_rate** - 0.001, 0.01, 0.1.

TPOT - An AutoML approach

The first automation initiatives in model selection through grid search and similar methods were described in the 90s. The term *AutoML* has been used ever since, and a commercial version made its debut in 2018 (Zöller and Huber, 2019).

In this work we opted for TPOT AutoML framework to accompany with this rising trend and compare it using our real-word problem. This framework employs genetic programming techniques to hone the model selection pipeline, by providing state-of-the-art optimisation methods to a broader audience:

1. **generations** - selected value was 5 - represents the number of iterations given to the optimisation pipeline. By increasing the number of iterations, we increase the chances of finding a better (or even optimal) solution, always at the cost of time and computational resources.

2. **population_size** - selected value was 50 - is the number of solutions in each generation, as a subset of the total population of solutions.

3. **cv** - selected value was 5 - is the number of folds considered in the cross validation function - StratifiedKFold.

4. **verbosity** - selected value was 3 - determines the amount of information TPOT shows to the user during run-time.

5. **scoring** - selected value was *f1* - determines the scoring method for the models.

6. **n_jobs** - selected value was 16 - determines the number of processes to be used in parallel.

7. **random_state** - selected value was 42 - is the random generator seed used to assure the same results across executions, given the same inputs.

By following these structured steps - feature preprocessing, feature selection, model training, optimisation and scoring - we assure the comparability of the three approaches.

4.3 Results and Discussion

In this section we present the results and their discussion, structured by risk perspective. We analyse how our observations span through the risk classes. Each risk perspective is described in terms of performance of the models assessed using the forementioned approaches: train-test split, cross-validation and TPOT.

Figure 4.2 shows the distribution of observations in our dataset by risk class. Most risks show a balanced distribution of classes, except for credit risk, which has significantly more observations on class 2. Using oversampling and undersampling techniques to deal with this issue was pondered. However, this distribution reflects the frequency of each class in the real world, as a consequence we decided not to balance the dataset.

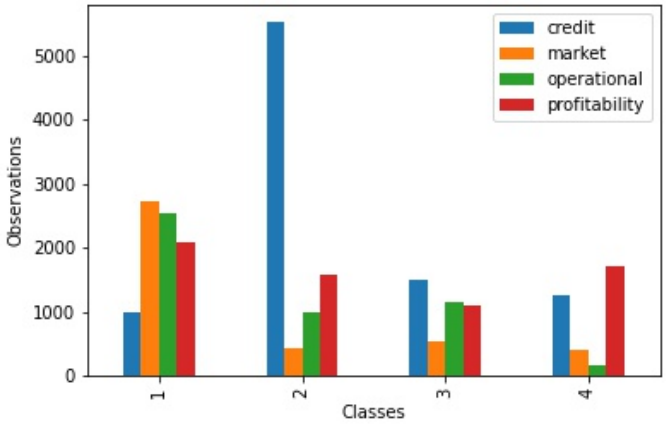


Figure 4.2: Number of observations in the dataset per target class, per target variable.

4.3.1 Credit Risk

The evaluation of credit risk was performed in a sample with 9262 observations, and 2608 features after the feature selection process. The processing wall time used to evaluate the performance of the listed models for credit risk were:

1. Train-test split: 28 minutes and 48 seconds;
2. Cross-validation: 1 hour, 57 minutes and 44 seconds;
3. TPOT framework: 2 days, 15 hours, 20 minutes and 34 seconds.

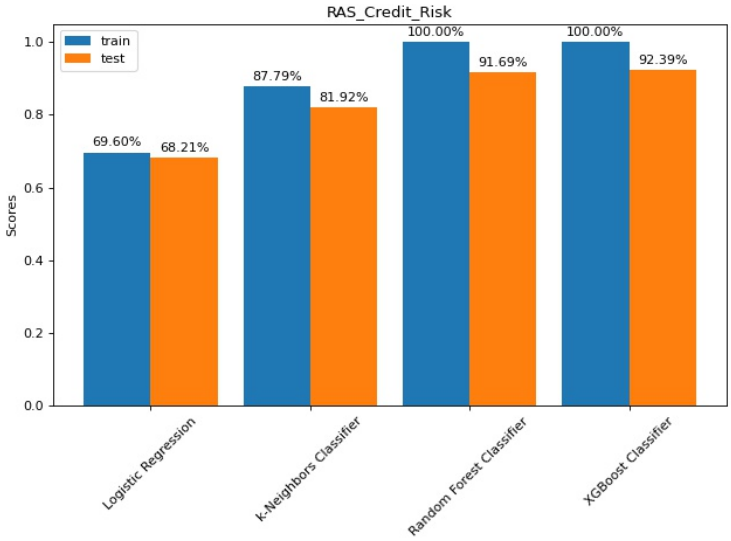


Figure 4.3: F1-scores of each model, using train-test split approach.

Figure 4.3 shows the results of each model comparing its training and test scores.

The Logistic Regression presents average results, slightly below 70% in both train and test sets. This could suggest that we are dealing with more complex decision boundaries. k-Nearest Neighbours shows better results, suggesting that the classes in our dataset are not linearly separated and they might not be independent. This is often the case with financial reporting data: variables are not completely independent from each other and the heterogeneity of the data creates more complex boundaries between classes.

When applying tree-based models with ensembles, like Random Forest and Extreme Gradient Boosting (XGBoost), we see a 10 to 20% increase in test performance. The work by Chang et al. (2018) shows how these techniques capture the heterogeneous structure of financial data, making these models the most adequate choice.

Figure A.1 in the Annexes, shows the detailed classifications of each model through their confusion matrices, supporting the above mentioned findings.

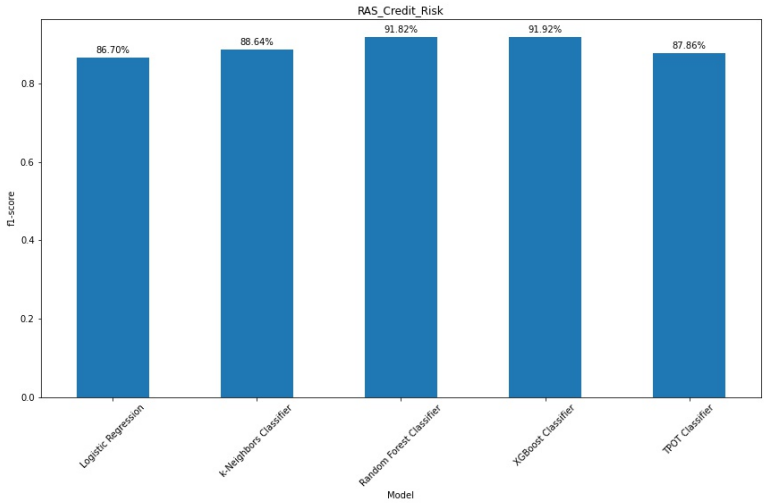


Figure 4.4: F1-scores of each model, using cross-validation approach.

For a more precise view on how the intricacies of the data affect the performance of these models, we applied cross-validation with with a hyper-parameter optimisation process to assess the f1-score of each model (Figure 4.4). As already mentioned, this measure represents the harmonic mean of precision and recall, ideal for multi-label classifiers and imbalanced datasets. The figure also includes the results of TPOT, the autoML framework considered in this study. By using the whole dataset for the several train-test splits in cross-validation, we ensure that the final score is not biased to any specific split, missing some particular event that compromises the models' performance on unseen data.

In terms of relative performance, the models performed similarly when compared to each other. XGBoost presents the best performance, even when compared to TPOT - this last framework being resource-intensive and needing almost three days to optimise its pipeline. Furthermore, in terms of time and performance gains, it does not outperform XGBoost.

The confusion matrices in Figure A.2 offer a detailed perspective on each model's classification decision. As with the train-test split, we see k-Nearest Neighbours being penalised by class 2 imbalance for credit risk and XGBoost missing the least classifications.

4.3.2 Market Risk

This subsection presents the results of evaluating the market risk perspective. This sample is composed by 4080 observations and 3539 features, selected through Random Forest feature selection process. The wall time of each of the approaches was:

1. Train-test split: 4 minutes and 14 seconds;
2. Cross-validation: 16 hours, 40 minutes and 8 seconds;
3. TPOT framework: 18 hours, 44 minutes and 16 seconds.

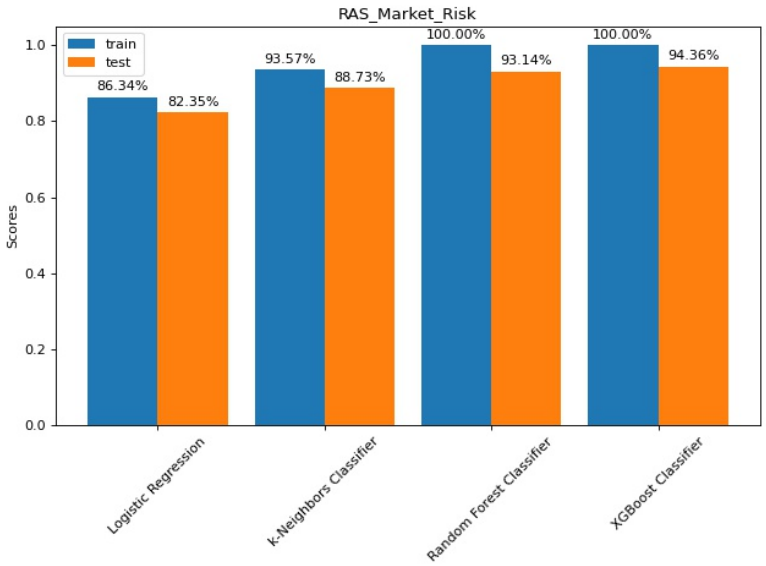


Figure 4.5: F1-scores of each model, using train-test split approach.

The results of the train-test split evaluation are shown in figure 4.5. Here the results show a distribution similar to what we observed with credit risk however, the scores are slightly better.

The Logistic Regression results suggest that we face linear (or close to linear) boundaries between classes. This reading is also supported by the fact that its score is closer to k-Nearest Neighbours'.

Still, the use of ensemble tree-based models show a significant increase in performance. The spike is not as prominent as with credit risk, and Random Forest has again a similar, but lower, score than XGBoost - on the order of the decimal percentage points. Figure A.3 shows the confusion matrices for the train-test split evaluation.

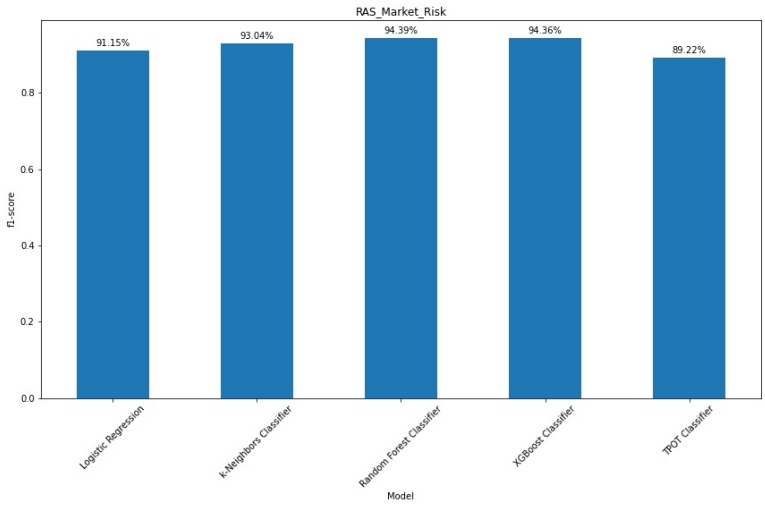


Figure 4.6: F1-scores of each model, using cross-validation approach.

However, a random train-test split might give an undervalued or overvalued perspective of a model's performance. To validate these findings we applied cross-validation with f1-score to the whole dataset. The results of this process are shown in figure 4.6 along with the evaluation of the TPOT framework.

The models show similar scores when compared to each other, with Logistic Regression faring close to the k-Nearest Neighbours. Contrarily to what we observed in the train-test split, a more discerning look at the results shows that Random Forest classifier slightly outperforms XGBoost. TPOT comes in third place in terms of performance, and it becomes even less appealing if we consider its wall time. Figure A.4 presents the confusion matrices for this classification process.

4.3.3 Operational Risk

The sample provided to evaluate operational risk has 4819 observations and 3447 features. The wall time needed to evaluate the models on this sample was:

1. Train-test split: 5 minute and 19 seconds;
2. Cross-validation: 18 hours, 13 minutes and 52 seconds;
3. TPOT framework: 2 days, 15 hours, 31 minutes and 41 seconds.

The train-test split results shown in figure 4.7 paint a different picture than the other perspectives. Although we can observe a similar distribution of results, the Logistic Regression presents below-average results on unseen data. Furthermore, the k-Nearest Neighbours classifier exhibit a slight improvement to the previous model.

Random Forest and XGBoost classifiers again come into the spotlight, with the latter showing a modest advantage of less than two percentage points. Figure A.5 shows the confusion matrices for the train-test split, for a detailed view of each classification.

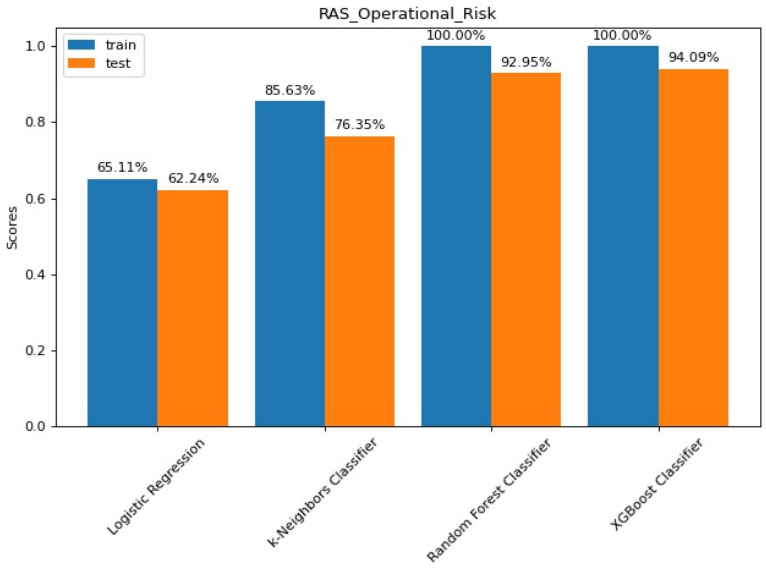


Figure 4.7: F1-scores of each model, using train-test split approach.

Applying cross-validation to this sample reveals several performance adjustments. Our non-tree-based models - the Logistic Regression, and k-Nearest Neighbours classifier - expressed an increase in their score, due to the optimisation process.

For Random Forest and XGBoost we see minor adjustments in the f1-score, however, their performance difference is consistent with the train-test split approach. This finding confirms the ability to grasp the heterogeneity of regulatory financial data. The TPOT framework is again in third place, revealing to be a poor choice due to the more than two and a half days of processing. Figure A.6 shows the confusion matrices of the cross-validation process, for a detailed view of the classifications of each model.

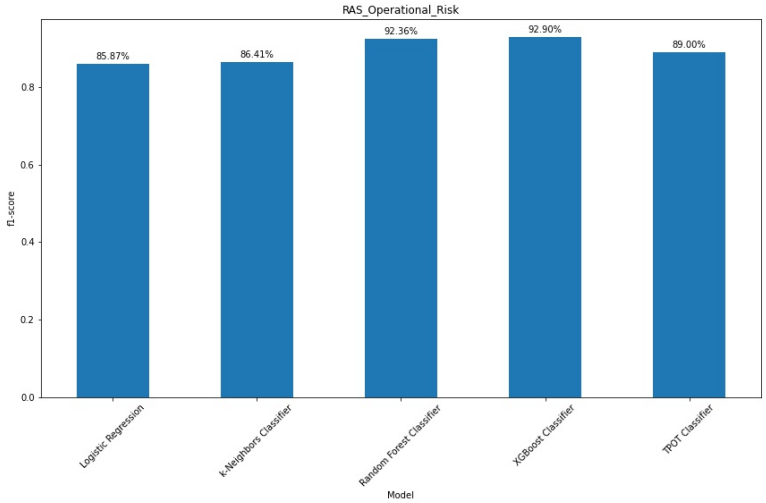


Figure 4.8: F1-scores of each model, using cross-validation approach.

4.3.4 Profitability Risk

As for our final risk perspective - profitability - we used a sample of 6448 observations and 3177 features. The processing and evaluation times for each of the approaches were:

1. Train-test split: 9 minute and 14 seconds;
2. Cross-validation: 1 day, 2hours, 25 minutes and 58 seconds;
3. TPOT framework: 1 day, 11 hours, 56 minutes and 42 seconds.

This is the risk perspective with the worse overall results. Figure 4.9 shows the train and tests scores for each model. Logistic Regression, k-Nearest Neighbours present a paltry performance. Even Random Forest and XGBoost show some decrease in performance, although still presenting good results. Figure A.7 pictures the detailed classifications of these models through the confusion matrices.

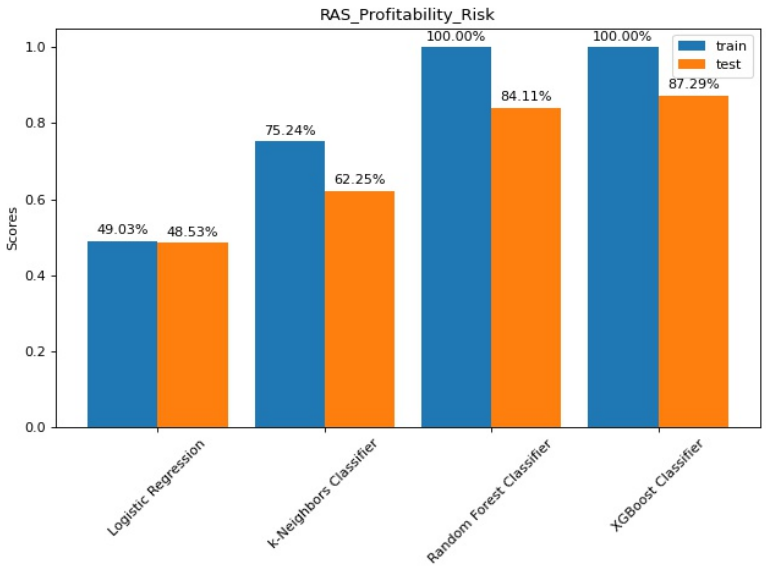


Figure 4.9: F1-scores of each model, using train-test split approach.

The cross-validation process corrects for any misclassification resulting from a unfavourable train-test split. In figure 4.10 we show the f1-scores for each model, including the TPOT framework.

Just as with train-test split, the Logistic Regression, and k-Nearest Neighbours present a low score, when compared to the other algorithms and their performance in other risk perspectives. Although this seems not related to class imbalance (see figure 4.2), the complexity of the decision boundaries and the dependence of some of the features might be the root cause for these foundering results.

Even so, the Random Forest and XGBoost show good results, with the latter again outperforming the former. The TPOT framework, comes in third with average results and one and a half day of processing, again making it an unsatisfactory alternative for this task. See figure A.8 for the confusion matrices of the cross-validation process.

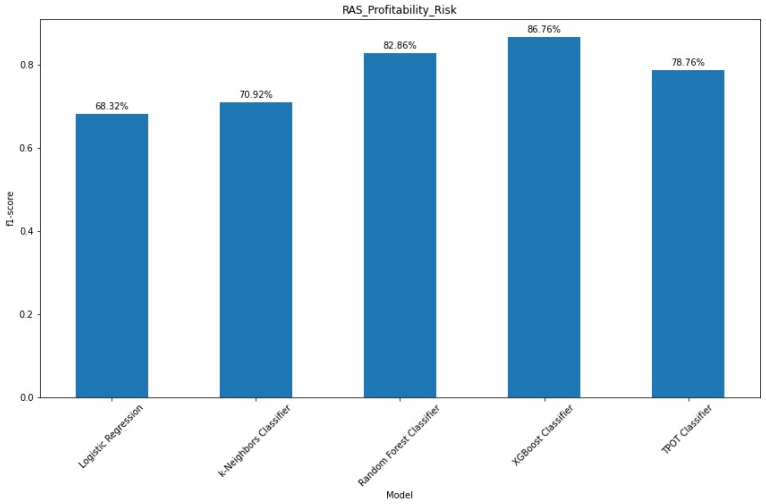


Figure 4.10: F1-scores of each model, using cross-validation approach.

4.3.5 Final remarks

Following our previous work (Guerra and Castelli, 2021), we clearly defined the required elements for modelling the supervisory risk assessment process comprised in RAS. First, we suggested the use of SREP's quantitative pillar - Risk Assessment System - as a standard methodology to compare the banks at European level. This methodology is already established across the Euro-area, hence making it the ideal choice for the task. Moreover, most works in this area adopt a binary classification of the risk level of the banks, limiting the classification to "failure" or "no failure". As mentioned before, this approach lacks the flexibility required for central banks to detect the effect of distress events gradually and earlier in time. This is accomplished through the progressive multi-class scale provided in the RAS. Additionally, we identified a research gap specifically addressing the supervisory use-case. Using real-world supervisory data, designed and retrieved for regulatory purposes, it has been proven to provide the most accurate outlook (Broeders and Prenio, 2018; di Castri et al., 2019; Massaro et al., 2020; Filippopoulou et al., 2020).

We tested the above mentioned elements and successfully modelled the liquidity risk of a bank (Guerra et al., 2022). Based on those findings, we set out to generalise the methodology and model the remaining risk perspectives comprised in the RAS: credit, market, operational and profitability.

From a technical standpoint, we confirmed that an optimised XGBoost outperformed the other considered models. This is accordance with previous literature results suggesting XGBoost performs best with structured financial data. In addition to that, we have tested it against the auto ML framework TPOT, a rising trend in the field. The results showed that due to the characteristics of the dataset - large number of features and sparse dataset - computing time was extremely taxing, even with low parameters for the GP algorithm. It might be interesting to reduce the number of features to fewer than ten, and see how TPOT performs.

From a business perspective, the novelty within the presented results is the fact that we are modelling a multi-class decision process with real-world supervisory data. Whereas other works have not explored supervisory data, we rely on the European regulatory framework and the data collected within it. This data is the pillar of supervisory processes and brings the structure and context to our models. By relying on these models, we can develop early warning systems capable of anticipating distress events, considering the risk measures above, and also give supervisors a tool to test alternative economic scenarios to prevent pitfalls.

4.4 Conclusion

Streamlining an effective supervisory methodology requires an integrated view of the risks a credit institution is subject to. In our previous work we have successfully modelled liquidity risk according to SREP methodology. Once that pillar was set, we were able to apply the same modelling techniques to the other risk perspectives comprised in the Supervisory Review and Evaluation Process (SREP) and its Risk Assessment System (RAS): credit, market, operational and profitability.

Based on the quantifiable mainstay of ECB's Risk Assessment Methodology, we classified credit institutions from the Portuguese banking sector according to their risk level on each of the perspectives encompassed in the methodology. We used

real-life supervisory data and modelled this decision process by comparing several machine learning techniques, benchmarked against a widely used statistical method.

We have reached significant results clearly establishing that this decision process can be modelled and that the ML techniques used outperform the classic statistical approaches.

Regulatory supervisory data is highly correlated and heterogeneous, making the decision boundaries of this exercise a challenging task. Additionally, real-world events are seldom represented by balanced data. All risk levels are observed but with occurrences that are subject to events in a specific point in time. The complexities of such reality were best represented by ensemble tree-based models, like Random Forest and XGBoost classifiers. These models can capture the heterogeneous nature of financial data and establish clear decision boundaries with little error - f1-score between 87% and 94%. These results were obtained after applying an optimisation process within the cross-validation cycle.

Given the computational resources available and the cutting edge genetic programming optimisation pipeline available through TPOT, we expected it to outperform XGBoost. However, TPOT consistently came in third regarding f1-score, being outperformed by Random Forest and XGBoost. Its long processing times can be explained by the dedicated optimisation process, and the fact that our dataset is sparse (82576 features). The feature selection process is costly in computational sense and it might account for a significant share of the wall time.

We firmly believe this work is a meaningful contribution to a set of stakeholders involved in risk assessment in the banking sector:

- National Central Banks (NCBs) can leverage on the findings of this work and use these models to develop early warning systems. These sup-tech initiatives are currently in the limelight , with many projects being developed in this area by the ECB, Bank of International Settlements (BIS) and worldwide NCBs. A decision support system like this would provide an enhanced risk assessment perspective to supervisors.

- Banks and the consulting industry can convey these principles into their own systems. Consultancy companies can further support their clients in implementing their decision support systems using the data owned by the banks themselves. A bank can then proactively monitor and adjust their risk profile and strategy according to the regulatory requirements.

- Academia can use this work to extend and apply this type of ML methodologies to expand its usage on a regulatory perspective. Furthermore, we stress the postulates of using high quality, highly validated relevant data, and adopting an universal methodology for risk assessment, one that standardises how to appraise a bank.

Through this paper, we aim to contribute to the technical understanding of ML that can be applied to sup-tech use cases according to the business needs. Grounded in historical supervisory data, we propose a Sup-Tech tool that improves the European supervisory risk assessment by providing early warnings on several risks.

4.4.1 Limitations and future work

There are several aspects we have identified over the course of this study that would merit revision and improvement.

The dataset we used in this work reflects the Portuguese banking sector. Ideally, expanding to the European level and using data from all central banks in the Euro-area would provide a complete supervisory perspective. Additionally, more diverse data, with more business models would strengthen the ML models presented here.

Each of the risks would also merit from context specific data, in order to enhance the generalisation of each model. This would also allow the supervisors to access more timely decisions. Supervisory data is mostly quarterly which prevents quick reactions to adverse events. By combining it with daily data sources, such as market data, payments systems and credit responsibilities data, we might be able to obtain a daily signal for each aspect of a bank's risk. Confirming this decision path will strengthen the aforementioned models and provide a running risk assessment on which supervisors can rely on.

The ability to explain the reasoning behind each model is of utmost importance, in particular for critical systems such as for crisis detection. Explainable AI benefits hold true not only for experts to validate the decision process carried out by the ML models, but also as common ground language to report any issue to banks. As such, investing in explainable models will deliver a better understanding of the technology, bringing supervisors closer to sup-tech, and will also set forth a clearer communication between institutions and central banks.

Combining our quantitative data with qualitative expert judgement, using Natural Language Processing (NLP), will allow for automated score adjustments based on internal supervisory notes and risk assessment reports.

As a final remark, consolidating the results of each risk model with the relevant qualitative data could provide a single integrated bank score as an additional measure for the SREP exercise.

Chapter 5

Conclusion

The ever-growing amount of data retrieved by organisations has led almost every industry to invest in big data technologies. Machine learning is one of today's top choices to harvest data, and the financial sector has been one of its main drivers. Similarly, National Central Banks also thrive by leveraging on innovation as a key pillar to their mission: keeping prices stable and banks healthy. However, industries are adopting these new technologies at different paces. The much-needed changes on how supervision is accomplished have been delayed due to the financial sector being particularly conservative, the constant updating of the regulatory outlook, and the lack of qualified professionals who can carry out these changes in a sustainable manner. This is the identified gap that triggered our research.

Here, we use machine learning in the supervisory context with the purpose of modelling risk assessment processes. We achieved our objective and contributed to the forementioned research gap in three steps:

1. We first reviewed the literature on risk assessment and machine learning at central banks. Regarding this recent intersection of knowledge areas, we discovered the need for a common risk assessment methodology. Additionally, we gleaned that there are several central banks carrying out sup-tech initiatives, evidencing the potential of ML in the supervisory realm (Broeders and Prenio, 2018; Beerman et al., 2021; Hertig, 2021).

2. We then proceeded to successfully model the liquidity risk of a bank as a classification problem, comparing several ML models. The two key components of this step were the application of ECB's Risk Assessment System (RAS) used across the Euro-area, and the use of real-world supervisory data from the Portuguese banking sector.

3. Finally, we compiled equally valid results when applying the same methodology to the remaining risk perspectives: credit, market, operational and profitability.

The findings we accomplished in the above mentioned steps serve a vast number of stakeholders that can be grouped in three categories:

- Banks and consultancy companies, which can rely on the developed ML models to proactively manage their risk posture. This can positively influence and support how banks approach compliance obligations and also, how they adjust their practices to their business model.

- Central Banks can uplevel their supervisory processes and innovation initiatives with the ML techniques here described for financial data. Early Warning Systems are among the best use cases at supervisory agencies and can already be found and the ECB and BIS (Hertig, 2021).
- Academia will be able to support new investigation on the applications of ML to the supervisory area, and expand the usage of innovative technologies in the regulatory realm.

There are some limitations of the work here developed that we would like to address as future development topics. In a first tier we would improve model robustness by extending the sample to all banks in the Euro-area (approximately 3150 organisations). This could be achieved through a joint project with the ECB, where data from all Central Banks is already retrieved. Afterwards, we would increment accuracy by complementing each ML model with risk specific data. As a final recommendation, we would suggest intersecting the quantitative perspective studied in this thesis with the qualitative information present in internal documents and reports. This would develop an innovative use-case for Natural Language Processing (NLP) in supervision.

Bibliography

Abellán, J., Castellano, J.G., 2017. A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications 73, 1–10. doi:10.1016/j.eswa.2016.12.020.

Acuna, E., Rodriguez, C., 2004. The treatment of missing values and its effect on classifier accuracy.

Adankon, M.M., Cheriet, M., 2009. Encyclopedia of Biometrics - Support Vector Machine. Springer US. URL: https://doi.org/10.1007/978-0-387-73003-5_299, doi:10.1007/978-0-387-73003-5_299.

Ala'raj, M., Abbod, M.F., 2016. A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Systems with Applications 64, 36–55. doi:10.1016/j.eswa.2016.07.017.

Alessi, L., Detken, C., 2018. Identifying excessive credit growth and leverage. Journal of Financial Stability 35, 215–225. doi:10.1016/j.jfs.2017.06.005.

Alonso, A., Carbo, J.M., 2020. Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost. SSRN Electronic Journal doi:10.2139/ssrn.3724374.

Alonso, A., Carbo, J.M., 2021. Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation. SSRN Electronic Journal doi:10.2139/ssrn.3774075.

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23, 589–609. doi:10.1111/j.1540-6261.1968.tb00843.x.

Angelini, E., di Tollo, G., Roli, A., 2008. A neural network approach for credit risk evaluation. Quarterly Review of Economics and Finance 48, 733–755. doi:10.1016/j.qref.2007.04.001.

Antunes, J.A.P., 2021. To supervise or to self-supervise: a machine learning based comparison on credit supervision. Financial Innovation 7. URL: https://doi.org/10.1186/s40854-021-00242-4, doi:10.1186/s40854-021-00242-4.

Authority, E.B., 2013. Eba implementing technical standards (its). URL: http://www.eba.europa.eu/documents/10180/532570/EBA-ITS-2013-12+(Final+draft+ITS+on+Hypothetical+Capital+of+a+CCP).pdf.

Bank, E.C., . Supervisory review and evaluation process. URL: https://www.bankingsupervision.europa.eu/about/ssmexplained/html/srep.en.html.

Beerman, K., Prenio, J., Zamil, R., 2021. Fsi insights no 37: Suptech tools for prudential supervision and their use during the pandemic. FSI Insights on Policy Implementation URL: www.bis.org/emailalerts.htm.

Board, F.S., 2020. The use of supervisory and regulatory technology by authorities and regulated institutions: Market developments and financial stability implications. URL: www.fsb.org/emailalert.

Boyacioglu, M.A., Kara, Y., Ömer Kaan Baykan, 2009. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (sdif) transferred banks in turkey. Expert Systems with Applications 36, 3355–3366. doi:10.1016/j.eswa.2008.01.003.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Broeders, D., Prenio, J., 2018. Fsi insights innovative technology in financial supervision. FSI Insights on policy implementation July 2018, 29. URL: https://www.bis.org/fsi/publ/insights9.pdf.

Burstein, F., Holsapple, C.W., Power, D.J., 2008. Decision Support Systems: A Historical Overview. Springer Berlin Heidelberg. URL: https://doi.org/10.1007/978-3-540-48713-5_7, doi:10.1007/978-3-540-48713-5_7.

Casabianca, E., Catalano, M., Forni, L., Giarda, E., Passeri, S., 2019. An early warning system for banking crises: From regression-based analysis to machine learning techniques.

di Castri, S., Hohl, S., Kulenkampff, A., 2019. Fsi insights on policy implementation no. 19: The suptech generations. Financial Stability Institute 19, 19. URL: https://www.bis.org/fsi/publ/insights19.htm.

Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. URL: https://www.researchgate.net/publication/220320908.

Chakraborty, C., Joseph, A., 2017. Machine learning at central banks. SSRN Electronic Journal doi:10.2139/ssrn.3031796.

Chang, Y.C., Chang, K.H., Wu, G.J., 2018. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Applied Soft Computing Journal 73, 914–920. doi:10.1016/j.asoc.2018.09.029.

Chaudhuri, A., De, K., 2011. Fuzzy support vector machine for bankruptcy prediction. Applied Soft Computing Journal 11, 2472–2486. doi:10.1016/j.asoc.2010.10.003.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, International Conference on Knowledge Discovery and Data Mining. pp. 785–794. doi:10.1145/2939672.2939785.

Climent, F., Momparler, A., Carmona, P., 2019. Anticipating bank distress in the eurozone: An extreme gradient boosting approach. Journal of Business Research 101, 885–896. doi:10.1016/j.jbusres.2018.11.015.

Commission, E., 2015. Single supervisory mechanism. URL: https://ec.europa.eu/info/business-economy-euro/banking-and-finance/banking-union/single-supervisory-mechanism_en.

Consoli, S., Recupero, D., Saisana, M., 2021. Data Science for Economics and Finance. Springer International Publishing. doi:https://doi.org/10.1007/978-3-030-66891-4.

Cox, D.R., 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) 20. doi:https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

Dastile, X., Celik, T., Potsane, M., 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing Journal 91, 106263. URL: https://doi.org/10.1016/j.asoc.2020.106263, doi:10.1016/j.asoc.2020.106263.

Doerr, B.S., Gambacorta, L., Serena, J.M., 2021. How do central banks use big data and machine learning ? The European Money and Finance Forum , 1–6.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Meunier-FitzHugh, K.L., Meunier-FitzHugh, L.C.L., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., Williams, M.D., 2021. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management 57. doi:10.1016/j.ijinfomgt.2019.08.002.

Fawagreh, K., Gaber, M.M., Elyan, E., 2014. Random forests: From early developments to recent advancements. Systems Science and Control Engineering 2, 602–609. doi:10.1080/21642583.2014.956265.

Filippopoulou, C., Galariotis, E., Spyrou, S., 2020. An early warning system for predicting systemic banking crises in the eurozone: A logit regression approach. Journal of Economic Behavior and Organization 172, 344–363. doi:10.1016/j.jebo.2019.12.023.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29, 1189–1232. URL: https://doi.org/10.1214/aos/1013203451.

Galindo, J., Tamayo, P., 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Computational Economics 15, 107–143. doi:10.1023/a:1008699112516.

Glonek, G.F., McCullagh, P., 1995. Multivariate logistic models. Journal of the Royal Statistical Society 57, 533–546.

Gogas, P., Papadimitriou, T., Agrapetidou, A., 2018. Forecasting bank failures and stress testing: A machine learning approach. International Journal of Forecasting 34, 440–455. doi:10.1016/j.ijforecast.2018.01.009.

Guerra, P., Castelli, M., 2021. Machine learning applied to banking supervision a literature review. Risks 9, 136. doi:10.3390/risks9070136.

Guerra, P., Castelli, M., Côrte-Real, N., 2022. Machine learning for liquidity risk modelling: A supervisory perspective. Economic Analysis and Policy 74, 175–187. URL: https://linkinghub.elsevier.com/retrieve/pii/S0313592622000200, doi:10.1016/j.eap.2022.02.001.

Gusenbauer, M., 2019. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 118, 177–214. URL: https://doi.org/10.1007/s11192-018-2958-5, doi:10.1007/s11192-018-2958-5.

Hammer, P.L., Kogan, A., Lejeune, M.A., 2012. A logical analysis of banks' financial strength ratings. Expert Systems with Applications 39, 7808–7821. doi:10.1016/j.eswa.2012.01.087.

Hertig, G., 2021. Using artificial intelligence for financial supervision purposes. Bank of England , 1–29.

Hillegeist, S.A., Keating, E.K., Cram, D.P., Lundstedt, K.G., 2004. Assessing the probability of bankruptcy. Review of Accounting Studies 9, 5–34. doi:10.1023/B:RAST.0000013627.90884.b7.

Huang, S.C., Wu, C.F., Chiou, C.C., Lin, M.C., 2021. Intelligent fintech data mining by advanced deep learning approaches. Computational Economics URL: https://doi.org/10.1007/s10614-021-10118-5, doi:10.1007/s10614-021-10118-5.

Iturriaga, F.J.L., Sanz, I.P., 2015. Bankruptcy visualization and prediction using neural networks: A study of u.s. commercial banks. Expert Systems with Applications 42, 2857–2869. doi:10.1016/j.eswa.2014.11.025.

Jagtiani, J., Wall, L., Vermilyea, T., 2018. The roles of big data and machine learning in bank supervision. Banking Perspectives, Forthcoming , p. 1–11URL: https://www.theclearinghouse.org/banking-perspectives/2018/2018-q1-banking-perspectiv...

Kolari, J.W., López-Iturriaga, F.J., Sanz, I.P., 2019. Predicting european bank stress tests: Survival of the fittest. Global Finance Journal 39, 44–57. doi:10.1016/j.gfj.2018.01.015.

Kou, G., Chao, X., Peng, Y., Alsaadi, F.E., Herrera-Viedma, E., 2019. Machine learning methods for systemic risk analysis in financial sectors. Technological and Economic Development of Economy 25, 716–742. doi:10.3846/tede.2019.8740.

Krizhevsky, A., Sutskever, I., Hinton, G.E., . Imagenet classification with deep convolutional neural networks. URL: http://code.google.com/p/cuda-convnet/.

Kupiec, P.H., 2018. On the accuracy of alternative approaches for calibrating bank stress test models. Journal of Financial Stability 38, 132–146. doi:10.1016/j.jfs.2018.08.001.

Le, H.H., Viviani, J.L., 2018. Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. Research in International Business and Finance 44, 16–25. doi:10.1016/j.ribaf.2017.07.104.

Lee, I., Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons 63, 157–170. doi:10.1016/j.bushor.2019.10.005.

Leo, M., Sharma, S., Maddulety, K., 2019. Machine learning in banking risk management: A literature review. Risks 7. doi:10.3390/risks7010029.

Madley-Dowd, P., Hughes, R., Tilling, K., Heron, J., 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology 110, 63–73. doi:10.1016/j.jclinepi.2019.02.016.

Massaro, P., Vannini, I., Giudice, O., 2020. Institutional sector cassifier, a machine learning approach. SSRN Electronic Journal 548. doi:10.2139/ssrn.3612710.

Milian, E.Z., de M. Spinola, M., Carvalho, M.M., 2019. Fintechs: A literature review and research agenda. Electronic Commerce Research and Applications 34. doi:10.1016/j.elerap.2019.100833.

Min, J.H., Lee, Y.C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28, 603–614. doi:10.1016/j.eswa.2004.12.008.

Moody's, 2021. Moody's datahub. URL: https://datahub.moodys.io/.

Ng, J., 2011. The effect of information quality on liquidity risk. Journal of Accounting and Economics 52, 126–143. URL: http://dx.doi.org/10.1016/j.jacceco.2011.03.004, doi:10.1016/j.jacceco.2011.03.004.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18, 109. doi:10.2307/2490395.

Olson, R.S., Bartley, N., Urbanowicz, R.J., Moore, J.H., 2016. Evaluation of a tree-based pipeline optimization tool for automating data science, pp. 485–492. doi:10.1145/2908812.2908918.

Parliament, E., 2013. Directive 2013/36/eu. URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:176:0338:0436:En:PDF.

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: Machine learning in python. URL: http://scikit-learn.sourceforge.net.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., Klamargias, A., 2018. A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. The use of big data analytics and artificial intelligence in central banking 50, 30–31. URL: https://www.bis.org/ifc/publ/ifcb49_49.pdf.

Pompella, M., Dicanio, A., 2017. Ratings based inference and credit risk: Detecting likely-to-fail banks with the pc-mahalanobis method. Economic Modelling 67, 34–44. doi:10.1016/j.econmod.2016.08.023.

de Portugal, B., 2021. Banco de portugal microdata research laboratory. URL: https://bplim.bportugal.pt/.

Ribeiro, B., Silva, C., Chen, N., Vieira, A., Neves, J.C.D., 2012. Enhanced default risk models with svm+. Expert Systems with Applications 39, 10140–10152. doi:10.1016/j.eswa.2012.02.142.

Rish, I., 2001. An empirical study of the naive bayes classifier.

Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. SN Applied Sciences 2. doi:10.1007/s42452-020-3060-1.

Shah, S.Q.A., Khan, I., Shah, S.S.A., Tahir, M., 2018. Factors affecting liquidity of banks: Empirical evidence from the banking sector of pakistan. Colombo Business Journal 9, 01. doi:10.4038/cbj.v9i1.20.

Soui, M., Gasmi, I., Smiti, S., Ghédira, K., 2019. Rule-based credit risk assessment model using multi-objective evolutionary algorithms. Expert Systems with Applications 126, 144–157. doi:10.1016/j.eswa.2019.01.078.

Stock, J., Watson, M., 2001. Vector autoregressions. Journal of Economic Perspectives 15, 101–115. doi:10.1002/9780470996249.ch33.

Strydom, M., Buckley, S., 2019. AI and Big Data's Potential for Disruptive Innovation. 1st editio ed., IGI Global. doi:10.4018/978-1-5225-9687-5.

Tavana, M., Abtahi, A.R., Caprio, D.D., Poortarigh, M., 2018. An artificial neural network and bayesian network model for liquidity risk assessment in banking. Neurocomputing 275, 2525–2554. doi:10.1016/j.neucom.2017.11.034.

Tharwat, A., 2018. Classification assessment methods. Applied Computing and Informatics 17, 168–192. doi:10.1016/j.aci.2018.08.003.

Vento, G.A., Ganga, P.L., 2009. Bank liquidity risk management and supervision : Which lessons from recent market turmoil ? Journal of Money, Investment and Banking 10, 79–126.

Wang, T., Zhao, S., Zhu, G., Zheng, H., 2021. A machine learning-based early warning system for systemic banking crises. Applied Economics 00, 1–19. URL: https://doi.org/10.1080/00036846.2020.1870657, doi:10.1080/00036846.2020.1870657.

Weston, J., Watkins, C., 1998. Multi-class support vector machines.

Xia, Y., Liu, C., Li, Y.Y., Liu, N., 2017. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications 78, 225–241. doi:10.1016/j.eswa.2017.02.017.

Yang, Y., Wu, M., 2021. Explainable machine learning for improving logistic regression models, Institute of Electrical and Electronics Engineers (IEEE). pp. 1–6. doi:10.1109/indin45523.2021.9557392.

Zopounidis, C., Doumpos, M., Matsatsinis, N.F., 1997. On the use of knowledge-based decision support systems in financial management: A survey. Decision Support Systems 20, 259–277. doi:10.1016/S0167-9236(97)00002-X.

Zöller, M.A., Huber, M.F., 2019. Benchmark and survey of automated machine learning frameworks. arXiv 70, 411–474. doi:10.1613/jair.1.11854.

Appendix A

Tables

Authors	Year	Affiliation	Title	Citations
Abellan et al.	2017	academia	A comparative study on base classifiers in ensemble methods for credit scoring	88
Ala'raj et al.	2016	academia	A new hybrid ensemble credit scoring model based on classifiers consensus system approach	66
Alessi et al.	2018	central bank	Identifying excessive credit growth and leverage	135
Alonso et al.	2020	central bank	Machine Learning in Credit Risk: Measuring the Dilemma Between Prediction and Supervisory Cost	1
	2021	central bank	Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation	0
Angelini et al.	2008	academia	A neural network approach for credit risk evaluation	305
Antunes	2021	central bank	To supervise or to self-supervise: a machine learning based comparison on credit supervision	0

Continued on next page

Table A.1 – continued from previous page

Authors	Year	Affiliation	Title	Citations
Boyacioglu et al.	2009	academia	Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey	272
Broeders et al.	2018	industry	FSI Insights Innovative technology in financial supervision	23
Chakraborty et al.	2017	central bank	Machine Learning at Central Banks	62
Chang et al.	2018	academia	Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions	17
Chaudhuri et al.	2011	academia	Fuzzy Support Vector Machine for bankruptcy prediction	155
Climent et al.	2019	academia	Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach	10
Dastile et al.	2020	academia	Statistical and machine learning models in credit scoring: A systematic literature survey	0
Doerr et al.	2021	industry	How do central banks use big data and machine learning?	0
Dwivedi et al.	2019	academia	Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy	39
Filippopoulou et al.	2020	academia	An early warning system for predicting systemic banking crises in the Eurozone: A logit regression approach	1

Continued on next page

Table A.1 – continued from previous page

Authors	Year	Affiliation	Title	Citations
Galindo et al.	2000	academia	Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications	213
Giudice et al.	2020	central bank	Institutional Sector Classifier, a Machine Learning Approach	0
Gogas et al.	2018	academia	Forecasting bank failures and stress testing: A machine learning approach	20
Hammer et al.	2012	academia	A logical analysis of banks' financial strength ratings	49
Hillegeist et al.	2004	academia	Assessing the probability of bankruptcy	1393
Hohl et al.	2019	industry	FSI Insights on policy implementation The suptech generations	3
Huang et al.	2021	academia	Intelligent FinTech Data Mining by Advanced Deep Learning Approaches	0
Jagtiani et al.	2018	central bank	The Roles of Big Data and Machine Learning in Bank Supervision	4
Kolari et al.	2019	academia	Predicting European bank stress tests: Survival of the fittest	4
Kou et al.	2019	academia	Machine learning methods for systemic risk analysis in financial sectors	47
Kupiec et al.	2018	industry	On the accuracy of alternative approaches for calibrating bank stress test models	5
Le et al.	2018	academia	Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios	24
Lee et al.	2020	academia	Machine learning for enterprises: Applications, algorithm selection, and challenges	7

Continued on next page

Table A.1 – continued from previous page

Authors	Year	Affiliation	Title	Citations
Leo et al.	2019	(blank)	Machine learning in banking risk management: A literature review	11
Lopez Iturriaga et al.	2015	academia	Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks	129
Milian et al.	2019	academia	Fintechs: A literature review and research agenda	31
Min et al.	2005	academia	Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters	866
Petropoulos et al.	2018	central bank	A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting	6
Pompella et al.	2017	academia	Ratings based Inference and Credit Risk: Detecting likely-to-fail Banks with the PC-Mahalanobis Method	5
Ribeiro et al.	2012	academia	Enhanced default risk models with SVM+	57
Soui et al.	2019	academia	Rule-based credit risk assessment model using multi-objective evolutionary algorithms	3
Tavana et al.	2018	academia	An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking	30
Wang et al.	2021	academia	A machine learning-based early warning system for systemic banking crises	2
Xia et al.	2017	academia	A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring	158

Table A.1: List of papers collected through the research query, referenced by author, year of publication, affiliation, and number of citations.

Quartile / Origin	Journal	Number of Papers
A (ERA)	Advances in Neural Information Processing Systems	1
Banca d'Italia	Questioni di Economia e Finanza	1
Banco de España	SSRN Electronic Journal	2
Bank for International Settlements	FSI Insights on policy implementation	2
Bank of England	Bank of England	1
Bank of Greece	Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"	1
Federal Reserve	Banking Perspectives, Forthcoming	1
Q1	Applied Soft Computing Journal	3
	Business Horizons	1
	Electronic Commerce Research and Applications	1
	Expert Systems with Applications	9
	International Journal of Forecasting	1
	International Journal of Information Management	1
	Journal of Business Research	1
	Journal of Economic Behavior and Organization	1
	Journal of Financial Stability	2
	Neurocomputing	1
	Research in International Business and Finance	1
	Review of Accounting Studies	1
	Technological and Economic Development of Economy	1
Q2	Applied Economics	1
	Computational Economics	2
	Economic Modelling	1
	Financial Innovation	1
	Global Finance Journal	1
	Quarterly Review of Economics and Finance	1
	Risks	1
SUERF	SUERF - The European Money and Finance Forum	1

Table A.2: Journals of the selected articles and their quartile. Where the journal is not indexed, the entity responsible for the publishing was included.

Authors	Year	Summary sentence
Galindo et al.	2000	CART decision-trees out-perform statistics for credit risk assessment, using a commercial bank loans dataset
Hillegeist et al.	2004	Black–Scholes–Merton option-pricing model is a better indicator of bankruptcy probability than Z-Score and O-Score.

Continued on next page

Table A.3 – continued from previous page

Authors	Year	Summary sentence
Min et al.	2005	Motivated by the increasing use of machine learning techniques, this paper aims to outperform classical statistics in bankruptcy prediction. An optimised SVM model performs better than MDA, logit and BPN for bankruptcy prediction.
Angelini et al.	2008	Regulation-imposed capital requirements increase the need for precise credit risk assessment systems. This paper shows ANNs' very good results predicting the default tendency of a borrower.
Boyacioglu et al.	2009	Multi-layer perceptrons and learning vector quantization are the most successful models predicting bank failure as a classification problem, in a Turkish case.
Chaudhuri et al.	2011	Fuzzy-SVM satisfies Basel II demands for detecting bankruptcy probability, outperforming other approaches. This algorithm also proved to have more clustering capabilities than PNN.
Hammer et al.	2012	The logical analysis of data (LAD) is able to reverse-engineer Fitch risk ratings of bank, showing better results than support-vector machines and logistic regression when evaluating the creditworthiness of banks.
Ribeiro et al.	2012	This study establishes the limitations of using exclusively quantitative financial data when developing default risk models. The authors propose a new approach that includes contextual knowledge in an SVM model, showing better predictability performance t
Lopez Iturriaga et al.	2015	Profiling distressed banks using self-organising maps and modelling failure detection with multi-layer perceptron outperforms traditional models of bankruptcy prediction. The resulting model detects 96% of failures, up to 3 years before the bankruptcy ev
Ala'raj et al.	2016	The proposed hybrid ensemble model improves predicting capability compared to base classifiers, using 7 real-world datasets. It uses a classifier consensus system to compare this new approach with the traditional combination methods.
Abellan et al.	2017	Selection of the best base classifier in ensemble methods for credit scoring problems. The individual performance of classifiers is not the only criteria for ensemble schemes.
Chakraborty et al.	2017	An overview of the applications of machine learning to financial problems, the most popular modelling approaches, and three case studies of relevant works for central banks. This study also establishes that machine learning models usually outperform tradi
Pompella et al.	2017	An EWS is proposed to detect likely-to-fail banks. This method is compared with risk agencies' rating and detects possibly wrongly rated banks. The authors suggest the adoption of this EWS by regulators.
Xia et al.	2017	The credit scoring problem is addressed using a XGBoost model with Bayesian hyper-parameter optimisation, not only obtaining better accuracy than baseline models, but also providing feature importance and a decision chart for interpretability.

Continued on next page

Table A.3 – continued from previous page

Authors	Year	Summary sentence
Alessi et al.	2018	The use of random forest to predict banking crises secondary to excessive credit growth, using credit and real estate predictors.
Broeders et al.	2018	A survey on the use of innovative technologies in financial supervision, the challenges faced by supervisory agencies and the need for a clear suptech strategy. Additionally, the experience of early adopters is described.
Chang et al.	2018	The development of a credit risk model using XGBoost classifier to address the heterogeneous nature of financial data. An undersampling method is applied to deal with the imbalanced data.
Gogas et al.	2018	Outperforming the Ohlson's score with stress-testing tool based on a support-vector machine model to forecast bank failures. The adopted methodology defines a clear boundary between solvent and insolvent banks.
Jagtiani et al.	2018	The impact of machine learning in banking supervision in terms of new possible analytical solutions and risks involved in those new approaches.
Kupiec et al.	2018	Addressing the need for validation of bank stress test models, by emphasising model forecast accuracy. A Lasso model shows the best forecasting capabilities for determining capital requirements in stressful conditions.
Le et al.	2018	Artificial neural networks and k-nearest neighbour methods are more accurate for predicting bank failure than traditional statistics.
Petropoulos et al.	2018	Predicting the probability of default of Greek banks using data mining techniques to reduce dimensionality, with XGBoost emerging as the best model. The authors aim to fully capture the information within these large datasets to better support the overall
Tavana et al.	2018	Addressing liquidity risk assessment through a model that uses neural networks and Bayesian networks. The models were capable of distinguishing the most critical factors in liquidity risk measurement.
Climent et al.	2019	Using XGBoost to identify the best predictors of bank failure and develop a classification model to label failed and non-failed banks in the Eurozone. The data used in this study is composed of 25 annual financial ratios for commercial banks in the Eurozo
Dwivedi et al.	2019	Expert contributors identify and compile a series of opportunities, impacts and research topics raised by the rapid adoption of AI. The financial sector shows enormous potential in robot advisory and automation, and bankruptcy prediction.
Hohl et al.	2019	A survey of activities within the scope of suptech, classifying the degree of technological development, and the strategies in place to implement them, highlighting the experimental nature of these initiatives and the need for international coordination.

Continued on next page

Table A.3 – continued from previous page

Authors	Year	Summary sentence
Kolari et al.	2019	Successfully undergoing European bank stress-tests depends largely on the risks a bank is exposed to, as opposed to being prepared for specific adverse scenarios. Using Bankscope data, the developed model accurately predicts 90% of the failing banks.
Kou et al.	2019	A survey depicting the most common methodologies to assess systemic risk in the financial system, using machine learning, big data analysis, network analysis and sentiment analysis. The paper showcases current researches on the use of machine learning in
Leo et al.	2019	A literature review evidencing machine learning use for risk management purposes in the banking industry, while also noting the experimental nature of most approaches.
Milian et al.	2019	A literature review aiming to find consensus on a fintech definition, showing how banks and supervisory agencies are using these innovative technologies and dealing with the risks involved.
Soui et al.	2019	Using evolutionary algorithms to address credit risk assessment by considering it as an optimisation (rule-based) search problem: minimise complexity, maximise accuracy and weight (rules importance).
Alonso et al.	2020	Comparing machine learning models from credit default prediction. Necessity for a structured strategy for assessing ML models to increase transparency in the use of these technologies, and promote innovation in the financial industry.
Dastile et al.	2020	A systematic literature review on how statistic and machine learning techniques have been used to address the credit scoring problem. Although machine learning is often incapable of explaining predictions, these models consistently outperform the classic
Filippopoulou et al.	2020	Developing an EWS to detect systemic banking crisis based on the ECB Macroprudential database. Most of the risk indicators used in the dataset are key to forecast a systemic risk crisis 1 to 4 years before the event.
Giudice et al.	2020	Developing an automatic classification system for the sector of economic activity for Italian companies, using a multi-step classifier with gradient boosting and support-vector machine models. The developed model is already being used in a production envi
Lee et al.	2020	A study on types of machine learning applications, exploring the accuracy-interpretability trade-off, and three use cases in financial industry.
Alonso et al.	2021	Predicting credit default probability with machine learning surpasses traditional statistic methods, potentially leading to savings of up to 17% in regulatory capital requirements.
Antunes	2021	Establishing the need for supervisory on-site inspection by comparing the results of two machine learning models, one based on the banks' own risk assessment and the other based on the findings from previous on-site inspections.

Continued on next page

Table A.3 – continued from previous page

Authors	Year	Summary sentence
Doerr et al.	2021	Policy brief showing central banks are relying on big data for daily tasks, and identifying a clear need for specialised knowledge on how to adequately use machine learning, and extract greater value from that data.
Huang et al.	2021	This study is developed under the assumption that the intricate nature of financial data cannot be properly explored through traditional methods. An advanced deep learning model to address the complex and hierarchical features of financial data, that outp
Wang et al.	2021	Random forest based EWS outperforms the classic logit approach as the predictive tool to prevent systemic banking crises. This paper shows an expert voting approach to model the multivariate nature of systemic risk assessment data.

Table A.3: Short summary of each analysed paper, referenced by authors and year.

Authors	ML Methods	Dataset
Abellan et al.	ada-boosting, bagging, random subspace, DECORATE, rotation forest	public: Australian, German, and Japanese datasets obtained from UCI repository of machine learning; Iranian dataset from "A comparison between statistical and data mining methods for credit scoring in case of limited available data. (2007)"; Polish datase
Ala'raj et al.	neural networks, support vector machines, random forests, decision trees, Naive Bayes	public: Australian, German, and Japanese datasets obtained from UCI repository of machine learning; Iranian dataset from "A comparison between statistical and data mining methods for credit scoring in case of limited available data. (2007)"; Polish datase
Alessi et al.	logit, decision trees, random forest	public: crisis dataset by Detken et al. 2014, capturing systemic banking crises related to domestic credit cycle
Alonso et al.	logit, lasso, CART, random forest, xgboost, deep learning	private: anonymized dataset from Banco Santander, containing more than 75000 credit operations
Alonso et al.	logit, lasso, CART, random forest, xgboost, deep learning, RL & ensemble methods	public: kaggle.com "Give me some credit" dataset
Angelini et al.	ann	private: SME loans from na Italian bank

Continued on next page

Table A.4 – continued from previous page

Authors	ML Methods	Dataset
Antunes	random forest	public: Central Bank of Brazil financial series repository
Boyacioglu et al.	Multi-layer perceptron, Competitive learning, Self-organizing map, Learning vector quantization, Support vector machines, Multivariate discriminant analysis, K-means cluster analysis, Logistic regression analysis	public: financial ratios using CAMELS system; annual publication "Banks Association of Turkey"
Broeders et al.	NA	NA
Chakraborty et al.	ann, dt,svm, clustering	NA
Chang et al.	logit, gmdh, svm, xgboost	private: credit data from a financial institution in Taiwan (2009-2016)
Chaudhuri et al.	logit, ann, svm, ga-svm, fuzzy-svm	private: dataset comprising American organizations with capitalization greater than $1 billion that filed for protection (2001-2002).
Climent et al.	xgboost	public: Orbis database (2006-2016)
Dastile et al.	LR (Logistic Regression), NB (Naïve Bayes), LDA (Linear Discriminant Analysis), XGB (XGBoost), EML (Extreme Learning Machines), k-NN (k-Nearest Neighbor), SVM (Support Vector Machine), ANN (Artificial Neural Network), BA (Bagging), BO (Boosting), RF (Rand	NA
Doerr et al.	NA	NA
Dwivedi et al.	evolution	NA
Filippopoulou et al.	logit, ewm	public: Macroprudential Database by the ECB
Galindo et al.	probit, knn, dt, CART	private: loans from a commercial bank provided by Comision Nacional Bancaria y de Valores (Mexico's security exchange and banking commission)
Giudice et al.	svm, xgboost	private: Bank of Italy Entities Register

Continued on next page

Table A.4 – continued from previous page

Authors	ML Methods	Dataset
Gogas et al.	O-score, svm	public: US banks (2007-2013); 481 failed and 962 solvent banks (1443 in total).
Hammer et al.	logit, svm, lad	public: 800 banks rated by Fitch along with 24 explanatory variables (2001).
Hillegeist et al.	logit, classic statistics	public: Moody's Default Risk Services' Corporate Default database and SDC Platinum Corporate Restructurings database (1980-2000)
Hohl et al.	evolution	NA
Huang et al.	deep CCAE, fuzzy rules, fuzzy rough nn, fuzzy nn, random tree, random forest	public: enterprise financial statement information from Taiwan securities market - Taiwan Economic Journal (2008-2013)
Jagtiani et al.	evolution, big data, ml	NA
Kolari et al.	AdaBoost, logit, ann, random forest, svm radial, svm linear	public: Bankscope database (2010, 2011 and 2014); 273 banks where 29 failed at least one stress test
Kou et al.	comparison	NA
Kupiec et al.	comparison; classic methods severely underestimate stress tests	public: quarterly financial data (balance sheet, income statements, etc.) from Federal Reserve Bank of St. Louis FRED economic database (1993-2011)
Le et al.	svm, ann, k-NN, linear discriminant analysis, logit	public: Bankscope database (2010-2016); 3000 US banks, 1438 failed, 1562 active. 31 ratios based on financial statements
Lee et al.	evolution	NA
Leo et al.	evolution	NA
Lopez Iturriaga et al.	mlp, som	public: 32 indicators extracted from financial statements - Federal Deposit Insurance Corporation between 2002 and 2012
Milian et al.	evolution	NA
Min et al.	mda, logit, svm, ann backpropagation	private: a Korean credit guarantee organization (2000-2002); 1888 institutions, 944 failed and 944 non-failed.
Petropoulos et al.	logit, LDA, XGBoost MXNET	private: Bank of Greece corporate loans database (2005-2015).

Continued on next page

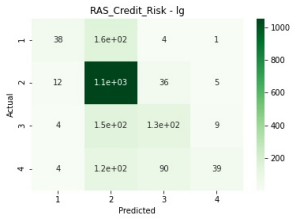
Table A.4 – continued from previous page

Authors	ML Methods	Dataset
Pompella et al.	ewm	public: Bloomberg indicators extracted from balance sheet, income statement and others (solvency, performance, etc.) (2005 to 2014); 482 banks
Ribeiro et al.	svm, svm+, svm+mtl	public: Diane database by COFACE; financial statements of French companies from 2002 to 2006.
Soui et al.	"Non-dominated" Sorting Genetic Algorithm (NSGAII), multi-objective evolutionary algorithm based on decomposition (MOEA/D), multi-objective particle swarm optimisation (SMOPSO), Strength Pareto Evolutionary Algorithm (SPEA2)	public: German (1000 observations, 70% good applicants, 30% bad applicants, 20 features) and Australian (690 observations, 383 good applicants, 307 bad applicants, 14 features) datasets from University of California, Irvine;
Tavana et al.	ANN, bayes	public: monthly reports on loan data provided by a large US bank (2005-2011); 353 observations, 10 features; balance sheet ratios
Wang et al.	logit, svm, adaboost, ann, random forest	public: yearly data for 95 economies with crisis data (1981-2017); 1690 observation of which 210 are crises; 11 features; dataset from Laeven and Valencia 2018, Global Financial Database and IMF International Financial Statistics
Xia et al.	AdaBoost, AdaBoost-NN, Bagging-DT, Bagging-NN, DT, LR, NN, RF, SVM, GBDT, XGBoost-MS, XGBoost-GS, XGBoost-RS, XGBoost-TPE	public: three datasets from UCI machine learning repository (German, Australian and Taiwan); two datasets from P2P lending platforms (Lending Club from the US and We.com from China)

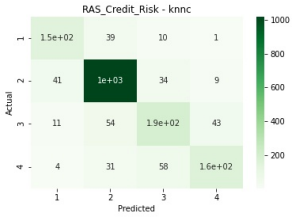
Table A.4: Machine learning methods applied in each paper and the respective dataset, referenced by authors.

Appendix A

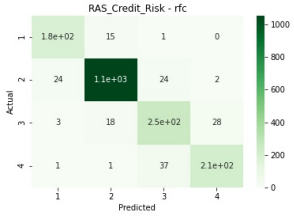
Confusion Matrices



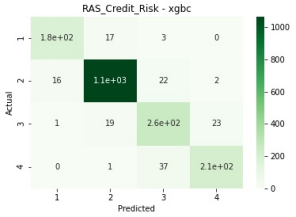
(a) Logistic Regression.



(b) k-Nearest Neighbours classifier.

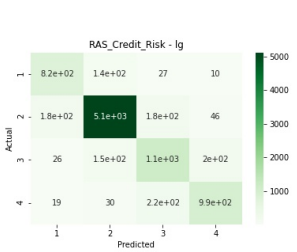


(c) Random Forest classifier.

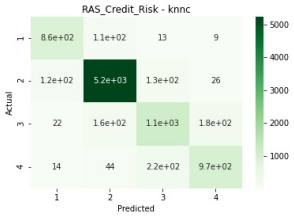


(d) Extreme Gradient Boosting classifier.

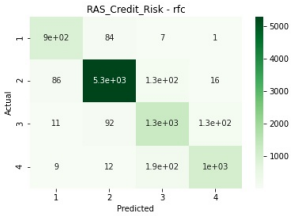
Figure A.1: Credit risk: Confusion matrices generated when evaluating the above mentioned models, using train-test split approach.



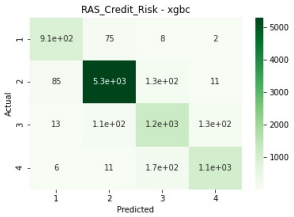
(a) Logistic Regression.



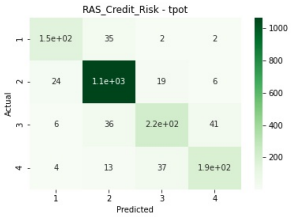
(b) k-Nearest Neighbours classifier.



(c) Random Forest classifier.

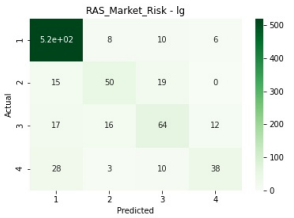


(d) Extreme Gradient Boosting classifier.

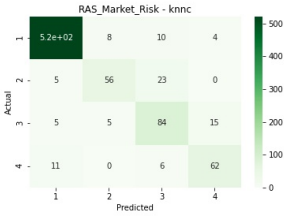


(e) TPOT classifier autoML framework.

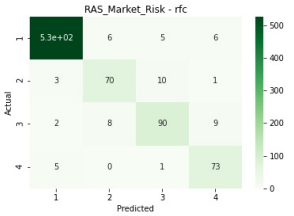
Figure A.2: Credit risk: Confusion matrices generated when evaluating the above mentioned models, using cross-validation approach.



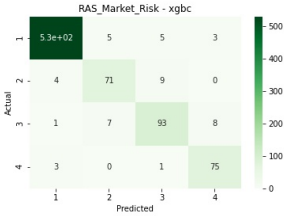
(a) Logistic Regression.



(b) k-Nearest Neighbours classifier.

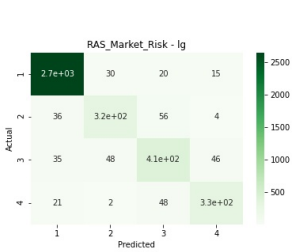


(c) Random Forest classifier.

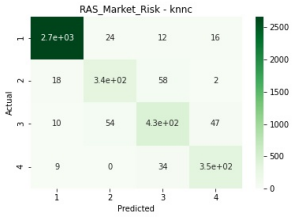


(d) Extreme Gradient Boosting classifier.

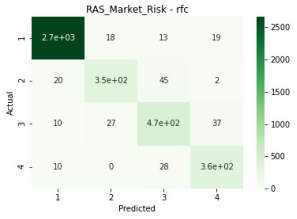
Figure A.3: Market risk: Confusion matrices generated when evaluating the above mentioned models, using train-test split approach.



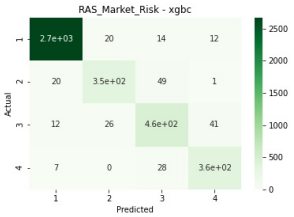
(a) Logistic Regression.



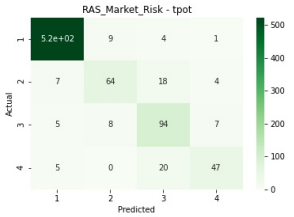
(b) k-Nearest Neighbours classifier.



(c) Random Forest classifier.

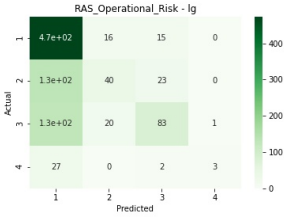


(d) Extreme Gradient Boosting classifier.

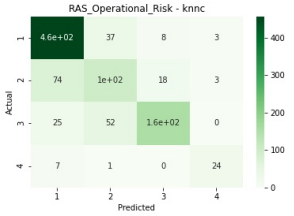


(e) TPOT classifier autoML framework.

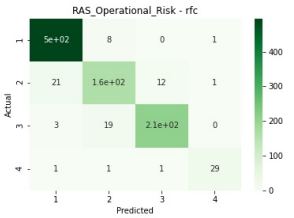
Figure A.4: Market risk: Confusion matrices generated when evaluating the above mentioned models, using cross-validation approach.



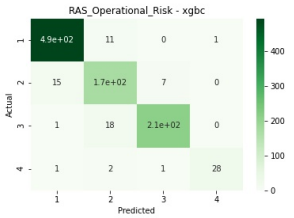
(a) Logistic Regression.



(b) k-Nearest Neighbours classifier.

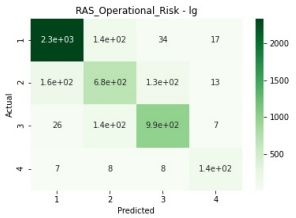


(c) Random Forest classifier.

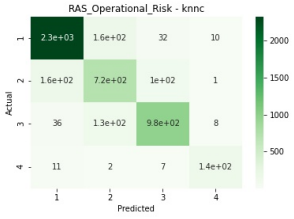


(d) Extreme Gradient Boosting classifier.

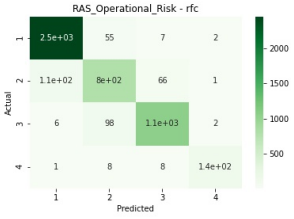
Figure A.5: Operational risk: Confusion matrices generated when evaluating the above mentioned models, using train-test split approach.



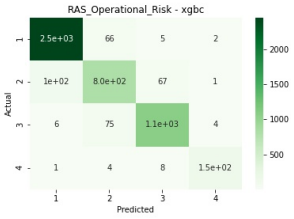
(a) Logistic Regression.



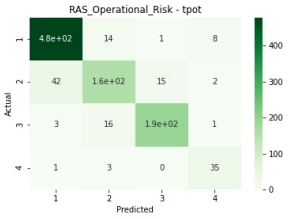
(b) k-Nearest Neighbours classifier.



(c) Random Forest classifier.

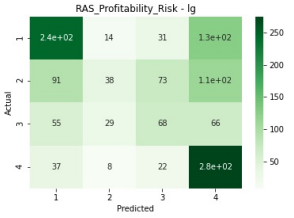


(d) Extreme Gradient Boosting classifier.

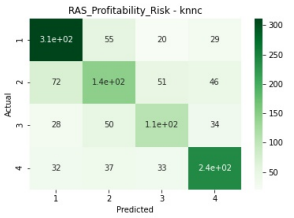


(e) TPOT classifier autoML framework.

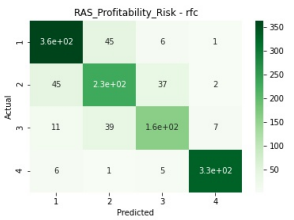
Figure A.6: Operational risk: Confusion matrices generated when evaluating the above mentioned models, using cross-validation approach.



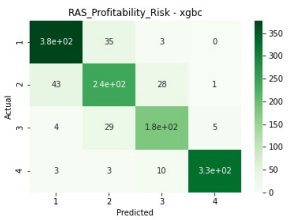
(a) Logistic Regression.



(b) k-Nearest Neighbours classifier.

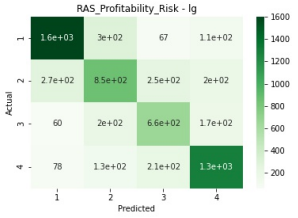


(c) Random Forest classifier.

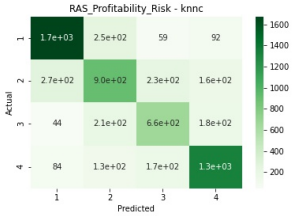


(d) Extreme Gradient Boosting classifier.

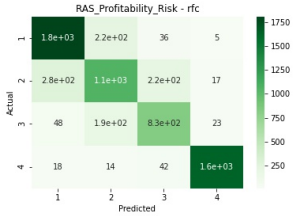
Figure A.7: Profitability risk: Confusion matrices generated when evaluating the above mentioned models, using train-test split approach.



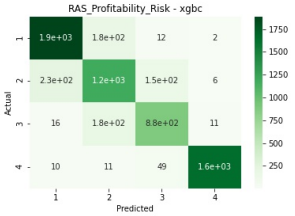
(a) Logistic Regression.



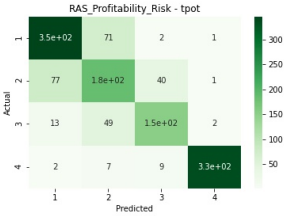
(b) k-Nearest Neighbours classifier.



(c) Random Forest classifier.



(d) Extreme Gradient Boosting classifier.



(e) TPOT classifier autoML framework.

Figure A.8: Profitability risk: Confusion matrices generated when evaluating the above mentioned models, using cross-validation approach.