NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

IMPROVING THE MATCHING OF REGISTERED UNEMPLOYED TO JOB OFFERS THROUGH MACHINE LEARNING ALGORITHMS

Paula Isabel Moura Meireles Cruz

Dissertation presented as a partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

IMPROVING THE MATCHING OF REGISTERED UNEMPLOYED TO JOB OFFERS THROUGH MACHINE LEARNING ALGORITHMS

by

Paula Isabel Moura Meireles Cruz

Dissertation presented as a partial requirement for obtaining the Master's degree in Information Management

Supervisor: Professor Roberto Henriques, PhD

November 2016

ACKNOWLEDGEMENTS

I would like to thank my supervisor for the valuable comments and remarks. I would also like to extend my gratitude to IEFP, IP, for providing all the necessary data and to SAS, for the valuable technical support and advices. Finally, I would like to thank to all of those, namely family, friends and colleagues, who have given me the necessary emotional support throughout the entire process.

ABSTRACT

Due to the existence of a double-sided asymmetric information problem on the labour market characterized by a mutual lack of trust by employers and unemployed people, not enough job matches are facilitated by public employment services (PES), which seem to be caught in a low-end equilibrium. In order to act as a reliable third party, PES need to build a good and solid reputation among their main clients by offering better and less time consuming pre-selection services. The use of machine-learning, data-driven relevancy algorithms that calculate the viability of a specific candidate for a particular job opening is becoming increasingly popular in this field. Based on the Portuguese PES databases (CVs, vacancies, pre-selection and matching results), complemented by relevant external data published by Statistics Portugal and the European Classification of Skills/Competences, Qualifications and Occupations (ESCO), the current thesis evaluates the potential application of models such as Random Forests, Gradient Boosting, Support Vector Machines, Neural Networks Ensembles and other tree-based ensembles to the job matching activities that are carried out by the Portuguese PES, in order to understand the extent to which the latter can be improved through the adoption of automated processes. The obtained results seem promising and point to the possible use of robust algorithms such as Random Forests within the pre-selection of suitable candidates, due to their advantages at various levels, namely in terms of accuracy, capacity to handle large datasets with thousands of variables, including badly unbalanced ones, as well as extensive missing values and many-valued categorical variables.

KEYWORDS

Job Matching; Registered Unemployed with PES; Machine Learning Algorithms; Social Data; ESCO

TABLE OF CONTENTS

INDEX OF PICTURES

INDEX OF TABLES

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

The belief that "finding the right job should be easier than splitting an atom" (as cited in Bort, 2014) has been consistently behind some of the most innovative approaches to the development of increasingly powerful search engines based on artificial intelligence, which ultimate goal is to efficiently match job seekers to vacancies best suited to their qualifications, skills and experience and help employers to identify and hire the most qualified candidates.

The Bright Score - "a machine-learning, data-driven relevancy algorithm that calculates the viability of a specific candidate for a particular job opening" - is one of such examples, having as one of its key components the enhancement of the limited information which is normally available in job offers and CV's descriptions through the use of social media profiles and other publicly available data (Bollinger, Street, & Francisco, 2012).

Other implementations of automated applicant ranking systems based on machine learning algorithms can be found in the research literature, ranging from neural networks (Akinyede & Daramola, 2013) and other neuro-fuzzy techniques (Drigas, Kouremenos, Vrettos, Vrettaros, & Kouremenos, 2004) to regression trees, support vector machines (Faliagka et al., 2013) and analytical hierarchy process (Tzimas, 2012).

It can be stated that all of the above mentioned approaches share the same major concerns, namely: enhancing the productivity of the human resource personnel; reducing time wastages in collecting and sorting of vacancies postings and applications from job seekers; tackling unstructured and missing data. An important difference in the design of such automated systems should be noted, however: the expected degree of human intervention in the final ranking of results and the possibility or not of using recruiting agents' decisions in the training of the system.

The Portuguese Public Employment Services (PES) comprises 5 regional offices, 30 employment and training Centres, 23 job centres and a training and vocational rehabilitation centre, alongside the central services, which provide technical, administrative and financial support to the former. With the economic crisis in 2008, it started facing not only a rise in the unemployment rates, but also a shortage in the number of human resources that would be necessary to respond in an efficient manner to the needs of their most important clients - the registered unemployed.

In order to mitigate this problem, an increasing number of online services has been made available (and successively considered among the top best in Europe, according to E-Government Benchmark), alongside the more recent adoption, in 2012, of a new performance model, called "Intervention Model for Matching", comprising a set of elements, namely:

a) Profiling system:

Implementation of a profiling system (based on a logistic regression) to assess the risk of long-term unemployment in order to promote personalized interventions and to stipulate the frequency of contacts between the employment services and each type of unemployed jobseeker. According to this model, unemployed jobseekers are segmented into three categories: (1) high risk group, including intensive assistance clients; (2) moderate risk group, including clients for counselling and qualification; (3) low risk group, including market clients and clients for counselling and activation.

b) Different levels of job vacancies handling:

- Level 1 – PES is responsible for advertising the job vacancies, being that PES and the employer are jointly responsible for the recruitment and selection processes (a PES officer will be present at all the job interviews)
- Level 2 – PES is responsible for advertising the job vacancies and for the recruitment process, being that the employer is responsible for the selection process;
- Level 3 – the employer is responsible for the recruitment and selection processes, being that PES is only responsible for online advertisement of the job opportunity.

c) Matching system:

Gradual improvement of the matching system, through the implementation, in a first phase, of the possibility of using the information recorded in open fields, as well as assigning different weights to the most relevant variables, and, in a second phase, of a fuzzy matching system avoiding the immediate exclusion of vacancies and CVs which do not completely fulfil predefined selection criteria.

d) Career manger:

Creation of the career manager role, who consists of the officer responsible for agreeing the integration pathway with each unemployed jobseeker and the respective follow-up, in order to ensure and monitor their timely integration into sustainable jobs/active employment measures.

e) Job vacancies manager:

Enhancement of the job vacancies manager role, the officer responsible for the mediation between the employers' recruitment needs and the unemployed jobseekers.

Since 2012, there has been some improvement in the performance of the Portuguese PES as highlighted below (table 1.1), namely in terms of reported job vacancies and placements:

	2009	2010	2011	2012	2013	2014	2015
Total PES expenditure (in million Euros)	810	786	707	642	884	944	931
Total PES staff	3689	3 547	3 254	3 193	3 186	3 282	3 268
Annual unemployment rate in %	9.4	10.8	12.7	15.5	16.2	13.9	12.4
Total registered jobseekers	751 223	706 558	739 558	801 088	798 413	755 529	759 331
Total registered unemployed	717 588	669 449	704 633	764 670	766 966	723 406	713 719
Long term unemployed	175 417	226 280	228 891	292 755	322 985	294 879	260 039
Number of new vacancies reported during the year	123 078	129 123	103 109	94 059	140 228	165 762	182 449
Number of placed job seekers	63 115	69 102	62 346	58 835	84 440	105 504	124 895

Source: IEFP, IP; INE, IP

Table 1.1 - Portuguese PES main performance indicators

However, it should be noted that around 50% of vacancies reported within the last four years are associated with financial incentives and not necessarily to an effective improvement of the matching process. In fact, the new operating model's main goals - such as facilitating and improving the interaction with the unemployed jobseekers, as well as with the employers and maximizing the opportunities for job matching - are yet to be fully achieved, namely in what concerns the optimal resource allocation based on the estimated risk of long term unemployment of a certain individual and the utilization of a non-linear matching system, which haven't still been adopted.

On the other hand, the market penetration of the online services is still very low, due not only to info-exclusion and digital literacy issues, but also to the need of a face-to-face contact that persists among the clients of the majority of European PES, as stated in the last E-government Benchmark (Capgemini, IDC, Sogeti, IS-practice and Indigov, 2012).

It should also be noted, as a general finding across Europe, that not enough job matches are facilitated by PES, due to the existence of a double-sided asymmetric information problem on the labour market[1] which those services have yet to overcome as a reliable third party, contributing to an actual reduction of search costs. In face of the lack of trust both by employers and the unemployed, PES seem to be caught in a low-end equilibrium, functioning as a last resort job brokerage service or as a source of subsidized low-wage jobs (Larsen & Vesan, 2012).

Radical solutions have been pointed to this apparently unsolvable problem (ranging from the establishing of a public monopoly on job-brokering to freeing PES from the task of helping the weakest workers), alongside more reasonable ones, from which the following are worth noting (Larsen & Vesan, 2012):

- Continuing on focusing on re-qualifying disadvantaged workers and giving employers wage subsidies for a limited period if they hire a person from a disadvantaged group;
- Providing guidance in using informal channels of recruitment and enhancing the use of online-services on an autonomous way or through a network of intermediaries;
- Building a good and solid reputation among employers and unemployed people by offering better and less time consuming pre-selection services.

The most challenging of these options consists of the last one, requiring novel and small budget approaches, especially if one takes into consideration that some European countries are already using expensive and powerful software platforms for smart searching, matching and analysis.

As highlighted in the case studies "A 'Virtual Labour Market Platform' for the Public Employment Service in Germany" (GHK Consulting Ltd, 2011) and "Bundesagentur für Arbeit" (WCC, n.d.), the implementation of a bi-directional and multidimensional matching system (considering over 40 criteria) has helped the German PES (BA) to achieve a leadership position in the field of online job portals. It also enabled BA to position itself as a modern service provider for all groups active on the labour market, bringing job seekers and employers together more quickly, with improved search results, more precise comparison between vacancies and applicants and an enhanced transparency nationwide, alongside more simplified and efficient consulting and job placement procedures and additional services to companies.

[1] The double-sided asymmetric information problem can be stated in the following terms: employers try to avoid the worst employees but this is difficult because the worker is better informed about his or her own capabilities; at the same time, employees try to avoid the worst employers, but this is also difficult because the employer is better informed about the real work conditions they offer (Larsen & Vesan, 2012).

1.2. JOB MATCHING PROCEDURES AT JOB CENTRES

For clarity of understanding, this section aims at providing a brief but (hopefully) clear description of the main phases and procedures of the job matching activities carried out by and at PES job centres, namely: identification of potential candidates; pre-screening interview at the job centre; referral to a job offer; follow-up of job referrals.

Identification of potential candidates

The main methods to identify potential candidates to registered job offers consist of the following ones:

- Manual consultation and analysis of available curricula vitae;
- Identification of potential applicants among job candidates present at the job centre, through face-to-face contact;
- Use of the already available automated matching tool, based on parameters such as: intended (and previous) occupation; experience; educational attainment; study and or training areas; qualification level; professional qualifications; driving, language and soft skills; profiling category; being a single parent; being part on an unemployed couple.

As far as the last method is concerned, it should also be noted that for every 50 potential candidates resulting from the performed queries, the following thresholds shall apply, as a rule of thumb:

- 10 to 15 applicants per job offer, on average (until a maximum of 20 to 30, when necessary), in the case of job offers marked as level 1;
- 8 to 10 applicants per job offer, on average (until a maximum of 20 to 25, when necessary), in the case of job offers marked as level 2.

Pre-screening interview at the job centre

Following the identification of potential applicants in the terms previously described, job centres proceed to the summoning of relevant candidates for a pre-screening interview aiming at giving information about the job offer conditions, as well as collecting additional information on the candidate in order to assess his or her suitability to the vacancy at hand and willingness to accept it.

The following thresholds shall apply, as a rule of thumb:

- 15 to 25 applicants per vacancy, on average (until a maximum of 20 to 30, when necessary), in the case of job offers marked as level 1;
- 8 to 10 applicants per vacancy, on average (until a maximum of 20 to 25, when necessary), in the case of job offers marked as level 2.

Referral to a job offer

Candidates that attend and perform successfully at the pre-screening interview conducted at the job centre will, in principle, receive a job referral aiming at an interview with the employer. At this stage, the following thresholds shall apply (always as a rule of thumb):

- 3 to 8 applicants per vacancy, on average, in the case of job offers marked as level 1;
- 3 to 5 applicants per vacancy, on average, in the case of job offers marked as level 2.

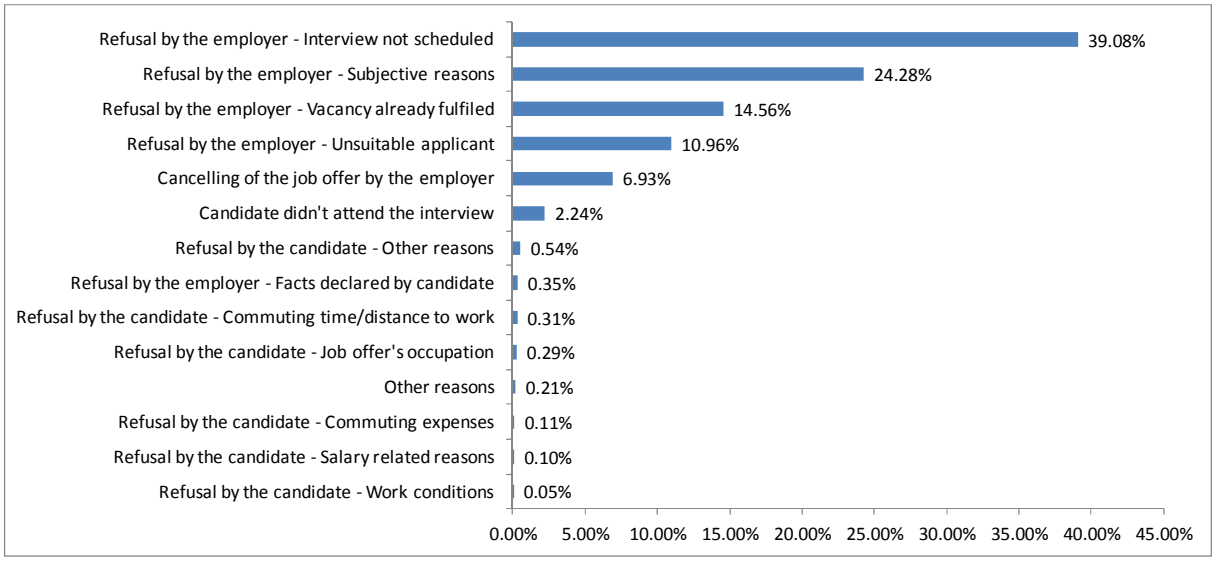
In the case of job offers marked as level 1, it is also mandatory that a PES counsellor (namely the one responsible for its management and follow-up) shall be present; in the case of job offers marked as level 2, that will be up to the responsible counsellor to decide in face of the situation at hand. The main objective behind the participation of a PES job counsellor at the interview(s) with the employer is to ensure a greater effectiveness of the matching process, through the close monitoring and evaluation of potential refusals either by the candidate or the employer.

Follow-up of job referrals

This stage of the matching process aims at determining whether the candidate attended the interview with the employer and, if so, whether he or she declined the job offer and why (when applicable). Job offers refusals can be classified into one of the following categories:

- Refusal by the employer - Vacancy already fulfilled
- Refusal by the employer - Unsuitable applicant
- Refusal by the employer - Selection of another candidate
- Refusal by the employer - Interview not scheduled
- Refusal by the employer - Facts declared by candidate
- Refusal by the employer - Subjective reasons
- Job offer cancelled by the employer
- Refusal by the candidate - Salary related reasons
- Refusal by the candidate - Work conditions
- Refusal by the candidate - Job offer's occupation
- Refusal by the candidate - Commuting time/distance to work
- Refusal by the candidate - Commuting expenses
- Refusal by the candidate - Other reasons
- Candidate didn't attend the interview
- Other reasons
- Placed at another job vacancy
- Cancelled referral
- Candidate on medical leave
- Cancelled job registration - Unavailability because of medical leave

In the graph below (picture 1.1), the relative importance of the relevant types of refusals for the period 2012-2015 is highlighted, for a clearer picture of the context in which job referrals and placements take place:



Source: IEFP, IP

Picture 1.1 - Relative importance of job offers' refusals by categories

As it would be expected, refusals by employers exceed, by far, the ones presented by candidates, who are subject to penalties when it is considered that they are not engaging in a facilitated job match. In 39.1% of the cases, the scheduling of the interview by the employer doesn't even take place, followed by subjective reasons (24.3%) and the fulfilment of the reported vacancy through other channels (14.6%). When employers and candidates do meet, around 11% of the referred applicants are considered unsuitable by the former, in terms of the required skills for the job (other refusals after an effective job interview may be included in the rather unclear category of candidates declined by subjective reasons).

All in all, these figures appear to confirm the employers' generalized lack of trust in the job brokering services provided by PES combined with the dominance of informal channels, such as already employed workers who are able to "provide trustworthy information about the new worker" (Larsen & Vesan, 2012). The greater preponderance of refusals by employers also seems to point to the lesser importance of building a good reputation among the unemployed (Larsen & Vesan, 2012).

The need for an effective screening of suitable applicants and on-time delivery of results thus continues to present itself as an important goal to be achieved by the Portuguese PES, notwithstanding the efforts already undertaken and the improvements obtained in the last years, as previously mentioned.

1.3. STUDY OBJECTIVES

As suggested by its title, the current research is focused on improving job matching services by the Portuguese PES, through the application of machine learning algorithms and can be captured by the following research question:

To what extent can job matching services provided by the Portuguese PES be improved through the application of machine learning algorithms and what is the best approach to the automation of the recruitment process?

The following points summarize the specific objectives that were pursued in order to achieve the above mentioned main goal:

- Extensive review of relevant literature and case studies in order to find the novelist and the most adequate and feasible machine learning algorithm(s);
- Study and application of the most relevant and recent algorithms available in the software package SAS Enterprise Miner;
- Identification of the most relevant variables to be used as inputs;
- Evaluation of the algorithm's performance, based on how effective it is in assigning consistent relevance scores to the candidates, compared to the ones assigned by human recruiters;
- Evaluation of the algorithm's performance against the applicant's risk of long term unemployment as estimated by the previously mentioned logistic regression, among other relevant input variables;
- Evaluation of the feasibility and importance of incorporating external data, by measuring the algorithm's effectiveness with and without those elements.

1.4. STUDY RELEVANCE AND IMPORTANCE

As mentioned in the paper by Strohmeier and Piazza (2013), human resources management (HRM) constitutes a fairly new domain of data mining research and in spite of the existence on an extensive collection of application examples clearly demonstrating the relevance and importance of this field, there is still great room for improvement, namely in what concerns functional relevance and success.

From the literature reviewed, only the paper by Drigas et al. (2004), which is also the least recent one, has a specific focus on registered unemployed, with a relatively restricted approach in what respects to the studied domains and variables, especially if one takes in consideration some of the factors that hinder PES performance in comparison to private sector companies, such as the administration of unemployment benefits and a particularly low skilled applicants' pool.

On the other hand, the first public version of the European Classification of Skills/Competences, Qualifications and Occupations (ESCO) - a standard and multilingual terminology supporting the automated analysis and interpretation of semi-structured and unstructured data, such as CVs and vacancies - has been recently released, providing an important basis for further and novel research in this field.

As already mentioned, an additional and important driver is the growing need to compensate for an insufficient number of employment counsellors (even in a more conservative scenario of semi-automated screening and selection procedures), in order to tackle the needs of the registered unemployed and improve the image and reputation of the Portuguese PES.

It can thus be stated that the present thesis aims not only at extending and updating existing knowledge, by analysing the application of machine learning algorithms within the specific and current context of PES, but also at providing a practical tool for the improvement of the Portuguese PES' technical matching system, through the incorporation of the ESCO database and relevant data mining results in the underlying information systems.

1.5. OUTLINE

The organization of the current thesis is as follows. In the next chapter, a literature review of the work that has been conducted on the topic at hand is provided, namely in what relates to the most adequate classification models and sampling techniques. In chapter 3, the methodological approach and the research design/strategy are presented, including a thorough description of the data sources, sampling and analysis methods that were used as well as of the various steps that were performed during the investigation. In chapter 4, the results obtained within each classification model are provided and discussed, as well as compared with the ones presented in analogous studies. The last chapter (5) presents the key findings arising from the discussion of the results and how they answer to the research question and established objectives. The limitations of the study and its main contributions to present knowledge are also provided, alongside recommendations for future research work. Finally, all relevant materials that may be considered useful for the comprehension of the work and analysis presented in the main body of the thesis will be included in the appendices.

2. CRITICAL LITERATURE REVIEW

According to Strohmeier & Piazza (2013), the following dimensions should be taken in consideration in what relates to the application of data mining techniques to human resources management (HRM):

- Functional dimension, comprising the following criteria: functional domain (HR problem being treated), functional relevance (whether and how the HR problem's relevance is justified) and functional success (evaluation of the success of data mining in solving the HR problem);
- Method dimension, comprising the following criteria: methodical category (mining methods that are employed) and methodical adjustment (whether these data mining techniques are general or domain-specific customized/developed);
- Data dimension: ensuring data availability and suitability;
- Information systems (IS) dimension: whether and which IS are provided;
- User dimension: intended and supported HR user-related tasks;
- Ethical and legal awareness: whether and how these issues are being considered.

Given the relevance of its main findings and implications for future research, the above mentioned article offers a good structured approach to the organization of the present chapter.

2.1. FUNCTIONAL DIMENSION

All of the publications reviewed so far have a clear focus on employee selection, namely on automated pre-screening and ranking of applicants, based on their quality and fitness for a certain job. Personality mining is also taken in consideration (Faliagka et al., 2013) and the paper *Efficient Multifaceted Screening of Job Applicants* (Mehta, Pimplikar, Singh, Varshney, & Visweswariah, 2013) presents an even broader approach by including the following sequential hiring stages:

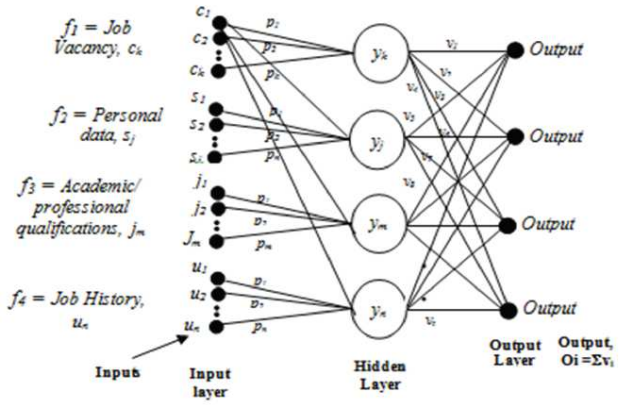
- Technical match, based on text fields (such as skills) extracted from the candidate's (unstructured) CV matched to the job description;
- Quality ranking, based on a dataset comprising past decisions by human screeners (pass or fail);
- Likelihood of a candidate accepting a job offer (onboard);
- Prediction of the risk of leaving the organization (attrition), at the pre-hire stage of the human resource lifecycle.

The individual ranking methods obtained along the above mentioned dimensions are then merged into a single ranked list, for consistency and simplicity reasons. Human experts' judgments are also incorporated in previous hiring decisions used as training data, in the scoring of candidates' relevance as a means to comparatively evaluate the results produced by the machine learning algorithms and, in some cases (Tzimas, 2012), in the adjustment of the weights of the selection criteria. However, resorting to internal recruiters is not always guaranteed, which may lead to worse conclusions (Bollinger et al., 2012).

2.2. METHOD DIMENSION

From a broad range of data mining techniques, classification methods are the most frequently employed ones, which can be explained by the specific character of the addressed problems (discriminating suitable and unsuitable applicants).

In the model proposed by Akinyede & Daramola (2013), a feed forward neural network is considered (picture 2.1), where Ck consists of the jobs applicants applied for; Un and Jm correspond to the job history, qualification and experience of the applicant Sj; Pn represents the weight (i.e., relative importance) of each field; and the output variable measures how suitable the applicant is for a certain job. With a view to improve generalization performance and achieve the best classification, the multilayer perceptron with structural learning is employed and receiver operating characteristic (ROC) is used in order to provide the percentage of detections correctly classified and the non-detections incorrectly classified.



Source: Akinyede & Daramola, 2013

Picture 2.1 - Example of an artificial neural network model

In the paper by Drigas et al. (2004), an expert system for the evaluation of the unemployed at certain offered posts, based on neuro-fuzzy techniques, is presented. There is a fuzzy rule for every criteria of the type: "Candidate's X matches X Criterion", where X belongs to a fuzzy set (training, education, experience, language and computer knowledge, for instance). These criteria can be satisfied in a binary way (yes or no, 0 or 1) or through a membership function (in which case the satisfaction of a certain criterion is scored between 0 and 1). The training of the system ultimately aims at finding a weight vector "w" that minimizes the following constrained linear least-squares problem (picture 2.2):

$$E(w) = \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} s_j^i \cdot w_j - d_i\right)^2 = \min$$

s.t.

$$\mathrm{lb}_j \le w_j \le \mathrm{ub}_j$$

Source: Drigas et al, 2004

Picture 2.2 – Minimization of constrained linear least-squares problem

where:

- Si,j corresponds to the criteria ("j") satisfaction for training case "i";
- Di takes the value 1 if the proposed post for case "i" was accepted and 0 otherwise;
- and Lbj is the lower bound of weight "j" and Ubj is the upper bound (constrained in the interval [0-1]).

The authors manage to demonstrate that learning weights lead to an average 10% increase in job matching.

In the paper by Faliagka et al. (2013), four different machine learning models are considered, namely: Linear Regression (LR), M5' model tree (M5'), REP Tree decision tree (REP), and Support Vector Regression (SVR) with two non-linear kernels (i.e. polynomial kernel and PUK universal kernel). The Tree models and the SVR model with a PUK kernel produce the best results, contrarily to linear regression, which performs poorly. The authors further conclude that the algorithm generated models presented high accuracy except for the jobs that required special skills.

In the model proposed by Mehta et al. (2013), a bipartite ranking through minimization of a standard univariate loss function is applied, based on random forest classifiers (of 100 trees) in order to better handle a great number of categorical features. After an experimental validation of the developed tool, the authors argue that its usage would have led to a "dramatic 14.6% increase in hiring yield over business as usual". An important issue that is also addressed in the paper at hand is that of imbalanced training sets, with many more negative samples than positive ones, due to its implications at various levels, namely in what relates to the most appropriate performance measures that should be taken in consideration.

In fact and according to the paper *"Data mining for imbalanced datasets: An overview"* (Chawla, 2005), predictive accuracy might not be appropriate when dealing with this type of cases and specific sampling techniques are required in order to reach more balanced datasets. In what relates to adequate performance evaluation, the following measures are proposed:

- ROC curves and AUC

  The receiver operating characteristic curve (ROC) consists of a standard technique for summarizing classifier performance based on a two-dimensional graphical illustration of the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity), without having to take into consideration the class distribution or misclassification cost (Iain Brown & Mues, 2012). Ideally, all negative and positive examples should be classified correctly, leading to a %FP and a %TP of 0 and 100, respectively and a perfect coincidence of the curve with the left top corner of the graph depicting it. The Area Under the Curve (AUC) is a generally accepted performance metric for comparing and establishing a dominance relationship between the ROC curves of different classifiers, being similar to the Gini coefficient which is equal to 2 x (AUC - 0.5). The Friedman's test - which is based on the average ranked (AR) performances of the classification techniques on each data set - can be used to compare the AUCs of the different classifiers (Iain Brown & Mues, 2012).

- Precision and recall

  The expressions for precision and recall can be derived from the following formulas:

  Precision = TP / (TP + FP)
  Recall = TP / (TP + FN)

  Although the main goal of learning from imbalanced datasets is to improve the recall without sacrificing precision, these two performance measures are often conflicting, since the increasing of the true positive for the minority class may result in an increase in the number of false positives. A metric that combines the trade-offs of precision and recall, that is, among different values of TP, FP, and FN, is the F-value, which expression is as follows:

  *F-Value = (1 + B^2)*recall*precision / (B^2*recall + precision),* where *B* corresponds to the relative importance of precision versus recall and is usually set to 1.

- Cost sensitive measures (cost matrix and cost curves)

  The main idea behind cost-sensitive measures is that, based, for instance, on a matrix defining the penalties incurred in false positives and false negatives, it is possible to search for a decision that minimizes the expected overall cost.

As for sampling strategies, over and or under-sampling techniques are addressed (Chawla, 2005) as well as their shortcomings, in particular, the potential loss of important information, in the case of under-sampling, and overfitting on the multiple copies of the minority class examples, in the case of oversampling. In order to tackle these potential disadvantages, a specific oversampling technique is presented (SMOTE), according to which rare cases are synthetically generated based on the less frequent cases and their nearest neighbours in an effort to enlarge the model's decision boundary. In general, oversampling methods tend to provide more accurate results (Iain Brown & Mues, 2012).

As far as classification models are concerned, a general finding among the researched literature is the superiority of the combination of classifiers in the presence of imbalanced datasets as a means to improve prediction accuracy. Chawla (2005) points to boosting, defining it as a very popular combining technique in which "the classifiers in the ensemble are trained serially, with the weights on the training instances adjusted adaptively according to the performance of the previous classifiers", enabling the classification algorithm to concentrate on instances that are difficult to learn. Brown and Mues (2012) also support the usage of boosting, namely of gradient boosting - "an ensemble algorithm that improves the accuracy of a predictive function through incremental minimisation of the error term" - when dealing with samples characterized by a large class imbalance, alongside random forest classifiers, which are defined as "a group of unpruned classification or regression trees, trained on bootstrap samples of the training data using random feature selection in the process of tree generation". After the generation of a large number of trees, each tree votes for the most popular class, with these voting procedures being collectively defined as random forests. In a more comprehensive study (Fernández-Delgado, Cernadas, Barro, Amorim, & Amorim Fernández-Delgado, 2014), 179 classifiers arising from 17 families ("discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting,

bagging, stacking, random forests and other ensembles, generalized linear models, nearest neighbours, partial least squares and principal component regression, logistic and multinomial regression" and other methods) are evaluated and random forests are considered to be the best family of classifiers, followed by SVM, neural networks and boosting ensembles. A last study worth mentioning is that of Lessmann, Baesens, Seow and Thomas (2015), where 41 different classification algorithms are compared, without any resampling, however. According to these authors, heterogeneous ensembles classifiers also perform well.

In conclusion and as pointed out in several articles (López, Fernández, & Herrera, 2014; Krawczyk, Wózniak, & Schaefer, 2014), there are three main types of approaches to deal with the class imbalance problem:

- Data level solutions: through the rebalancing of the class distribution by sampling the data space in order to diminish the effect caused by class imbalance, in what can be described as an external approach.
- Algorithmic level solutions: through the adaptation of several classification algorithms to strengthen the learning in favour of the positive class, in what can be described as an internal approach that creates new algorithms or modifies existing ones in order to tackle the problem at hand;
- Cost-sensitive solutions: which incorporate approaches at the data level and or at the algorithm level, in order to minimize cost errors, by considering higher costs for the misclassification of examples of the positive class in relation to the negative class.

2.3. OTHER RELEVANT DIMENSIONS

Real-world data is used in the majority of cases, with conventional (semi)structured data retrieved from HR Information Systems (HRIS) being the most frequent one, followed by the less frequent text data (acquired from documents produced by applicants or employees) and web content data (gathered from social networks). It should also be noted that some of the reviewed research only addresses specific occupations (Drigas et al., 2004; Faliagka et al., 2013) or organizations (Mehta et al., 2013). However, in order to better support selection decisions and enable the discovery of unexpected patterns, data should be as broad as possible, covering a large pool of heterogeneous applicants and job offers.

On the other hand, the majority of the systems presented in the research contributions consist in stand-alone applications that don't require data mining expertise and possess a user-friendly interface, as demonstrated in Faliagka et al. (2013), Mehta et al. (2013) and Akinyede and Daramola (2013).

End-user related issues should also be taken in consideration, by minimizing the need to perform data mining tasks, alongside the lack of transparency of the underlying models. The adoption of an adequate strategy for the embracing of analytics-based solutions is another critical issue to address at this level, as highlighted in Mehta et al. (2013).

Lastly, it should be noted that the need to comply with ethical and legal standards is not always considered in current research, with equality of treatment (i.e., avoidance of discrimination) and protection of privacy being the most discussed issues (Strohmeier & Piazza, 2013).

3. METHODOLOGY

This chapter will start with the presentation of the methodological approach and research strategy, alongside a brief description of the main data sources and variables that have been used and of the corresponding data collection and storage procedures. It will then proceed with a description of the sampling and analysis methods, including the classification models and instruments that have been taken under consideration. In general, and in more technical terms, the SEMMA methodology (which stands for Sample, Explore, Modify, Model, and Assess) has been followed closely:

- *Sample the data*, through the creation of one or more data tables, big enough to contain the most relevant information, without sacrificing, however, capacity and speed of processing (a trade-off that must be handled carefully);
- *Explore the data*, in order to understand it to the fullest, gain knowledge and detect possible anomalies;
- *Modify the data*, through the selection and transformation of existing variables, as well as the creation of new ones in order to enhance model performance;
- *Model the data*, by choosing the most appropriate algorithm among the most relevant ones in order to produce reliable predictions of the desired outcome;
- *Assess the data*, through the evaluation of the obtained results in terms of usefulness and reliability.

3.1. DATA SOURCES DESCRIPTION

The main data sources that have been used in the current thesis consist of the following ones:

Input variables

- Internal data (from Portuguese PES databases): anonymised applications and job vacancies
- External data:
  - Relevant statistical information available for public or research use;
  - ESCO database, freely available for download.

Target variable

- Internal data (from Portuguese PES databases): matching results (i.e., if the applicant got the job or not).

Internal data, reporting to the period of 2012 to 2015, were completely provided by the Portuguese PES in March 2016 and are supported by the raw datasets briefly described in tables 3.1 and 3.2, below:

Data set	Nr. inputs	Size	Main variables
Registrations of job seekers	168	677 853 registered job seekers at stock in 2011 2 978 690 new registrations in 2012-2015 2 573 832 annulations in 2012-2015	• Applicant's personal data (masked id number, age, sex, nationality, marital status, type of existing disability, number of dependents, parish of residence) • Applicant's academic/professional qualification (highest level of qualification, major subject, training certificates, work experience, ...) • Applicant's job history (date employed, date disengaged, job code, status, employer, last-salary, condition for leaving, name, min year) • Data relating to the registration (year, month, type of movement, date and reason for registration, preferred job or jobs, employment status, responsible career manager and job centre, geographical mobility profile, probability of becoming a long term unemployed) • Administrative data relating to unemployment and social benefits (past benefits and new ones being claimed)

Source: IEFP, IP

Table 3.1 – Raw datasets

Data set	Nr. inputs	Size	Main variables
Registered job offers	99	3 897 job offers at stock in 2011 398 151 new vacancies in 2012-2015	• Organization (masked id number, line of trade, number of workers) • Job requirement (job offer code, occupation, required qualifications and skills, place of work, number of vacancies, required experience, salary and other working conditions) • Data relating to employment incentives associated to the job offer • Data relating to the registration of the job offer (year, month, type of movement, date, responsible job centre and manager, status of the job offer, handling level)
Summons for job interviews by PES	106	2 169 170	• "Snapshot" of the applicant's characteristics at the time of the call for a job interview • Data relating to the registration and management of the summon for a job interview (date of the summon, applicant's id number, responsible job centre, outcome and date of the outcome)
Referrals to job offers by PES	135	2 547 923	• "Snapshot" of the applicant's and job offer's characteristics at the time of the referral • Data relating to the registration and management of the referral (date of the referral, applicant's id number, responsible managers and job centres, outcome and date of the outcome)

Source: IEFP, IP

Table 3.2 – Raw datasets (cont.)

It would also be relevant to take in consideration data from Social Security (such as information on the complete previous experience of the unemployed and length of successful job placements) and (private) information available in social media profiles. However, the use of such elements is dependent on explicit consent, which prior studies have demonstrated to be very difficult to obtain (Bollinger et al., 2012).

As for relevant external statistical information, it consists of data freely available on Statistics Portugal online portal, namely:

- Proportion of purchasing power by geographic location;
- Unemployed, employed and active population by geographic location;
- Population commuting patterns;
- Average monthly earnings by occupation;
- Expected evolution of employment over the following three months;
- Quarterly unemployment rate;
- Persons employed at enterprise births by geographic localization and economic activity;
- Demography of enterprises (number of births, deaths and survival rates of enterprises born two years before) by geographic location and economic activity.

The finality of these external data is twofold: to enrich the core data and to support the conversion of nominal and categorical data to numerical data, namely in what concerns occupations, economic activities and geographical classifications.

The first public release of the European Classification of Skills/Competences, Qualifications and Occupations (ESCO v0, last updated in 06/08/2014), has also been used, on an experimental basis, as a means to establish a correlation between occupations based on common (hard) skills. This multilingual classification identifies and categorises skills, competences, qualifications and occupations relevant for the EU labour market, containing around 4.800 occupations and more than 5.000 skills. The system is freely available for use by everyone through an online portal (https://ec.europa.eu/esco/portal/download) and has been developed in an open IT format.

Lastly, it should be noted that Microsoft SQL Server 2012, more precisely SQL Server Management Studio has been used to store and transform the above mentioned data, as well as to create views supporting the subsequent creation of SAS tables through an ODBC connection defined in SAS Enterprise Guide.

3.2. PRE-PROCESSING, SAMPLING AND PARTITIONING

PRE-PROCESSING

For analysis and modelling purposes, four distinct datasets were created based on the previously presented ones:

- One mixing nominal and numerical variables containing only internal data;
- One containing only internal data where all nominal variables are converted to numbers;
- Two additional versions of the former data sets enriched with external data (including ESCO).

This distinction derives from the fact that some classification algorithms are well suited to work with data of mixed scaling level (e.g., classification trees and Bayes classifiers), whereas others (e.g., ANNs and SVMs) benefit from encoding nominal variables (Lessmann et al., 2015).

All of these datasets are based on a table resulting from the joining of the master table containing referrals to job offers with the master table containing data relating to job offers. Since each job offer is related to more than one job referral, the most approximate registers based on the date of the referral and on the date of the creation or last update to the job offer were considered for that effect.

In order to reduce the original number of variables, the following options were taken:

- Elimination of variables with more than 50% of missing values, as well as of variables with a purely administrative nature;
- Aggregation of various dummy variables into a unique binary variable describing a common feature (such as, for instance, whether a job offer is associated with a financial incentive instead of considering each type of incentive as a distinct input);
- Substitution of mirror variables (such as the qualifications of the applicant vs. the minimum required qualifications specified in the job offer) by a variable quantifying whether a match between the two exists, ranging from 0 to 1.

As for the conversion of categorical variables to numerical ones, the following strategies were taken in consideration, depending on each type of situation:

- Substitution of nominal binary classes (of the type "yes" or "no") by numerical binary classes (0 or 1);
- Substitution of categorical variables with more than two classes into a single dummy variable (such as, for instance, whether an applicant is of Portuguese nationality instead of creating a dummy variable for each different possible value of the listed countries of origin);
- Substitution of a categorical variable by one or more numerical variables providing associated relevant statistics, from external sources (for instance, the purchasing power or average qualifications within a certain county or parish instead of regional classification codes);

- Application of the Weight of Evidence (WOE) method, a "technique which converts a nominal input into an interval input by using a function of the distribution of a target variable for each level in the nominal input variable" (Zdravevski, Lameski, & Kulakov, 2013) and (Cathie, Chakraborty, & Garla, 2013).

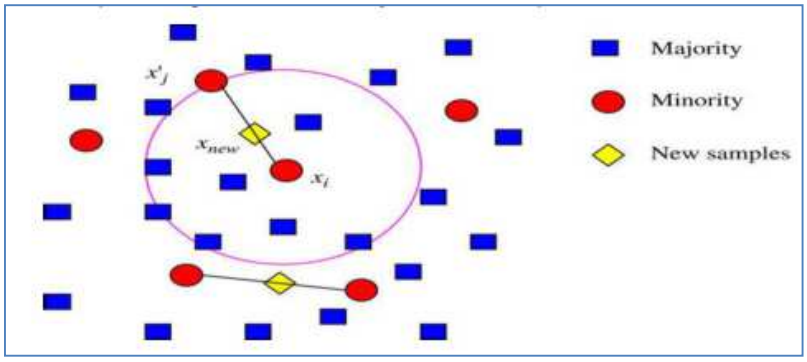
The imputation of missing values (using a median/mode replacement for numeric/nominal attributes, amongst other techniques) and data transformation for normalization purposes will be described in the "Results" section.

SAMPLING

As previously mentioned, there are three common sampling approaches in what concerns unbalanced datasets (Damodaran, Kumar, Raj, Jagan, & State, 2016):

- Under-sampling: According to this technique, the apparent sampling frequency of the majority samples is reduced by randomly removing observations;
- Over-sampling: According to this technique, the apparent sampling frequency of the minority samples is increased by randomly repeating each observation;
- SMOTE: According to this (oversampling) technique, the apparent sampling frequency of the minority samples is increased by creating new synthetic observations using a specific algorithm.

Within the present thesis, the first and last techniques have been applied. The SMOTE approach, in particular, is addressed in several papers by SAS Institute (Damodaran et al., 2016; Wang, Lee, & Wei, 2015) and can be illustrated as follows (picture 3.1):



Source: Damodaran et al., 2016

Picture 3.1 - SMOTE sampling technique

It should also be noted that the oversampling technique in SAS Miner may be mistaken with under-sampling in the sense that it increases the frequency of the minority class by reducing the absolute number of observations of the majority class, which will effectively result in a data set containing all the positive events and a random sample of the negative events.

Another important issue to address in this context is related to the size of the dataset to be processed, especially when the volume of information is big and there is enough processing capacity. Should the entire database be processed or only a sample of it? According to SAS Institute (Milley, Seabolt, & Williams, 1998), processing the entire database may be advantageous in scenarios such as the following ones: when there are more variables than records; when the process underlying the generation of the data is rapidly changing; in exception reporting systems; if the problem's solution is dependent on a few records. In most cases, however, this alternative poses problems at various levels, namely: inference/generalization, since using all of the data leaves no space to test the model's explanatory power on new events or to validate findings on data unseen by the model; quality of the findings, since exhaustive methods may reveal spurious relationships; speed and efficiency of processing. Exploring a sample that reflects and preserves the most important characteristics of the underlying data is, generally, easier and more efficient than processing the entire database, without loss of accuracy.

DATA PARTITIONING

In the presence of a large number of observations, a straightforward approach can be adopted in what concerns data partitioning. In this type of context, the available data is normally divided into a train, validation and test set, except in the case of classification models containing out-of-bag data (such as Random Forests), where a test set won't be necessary.

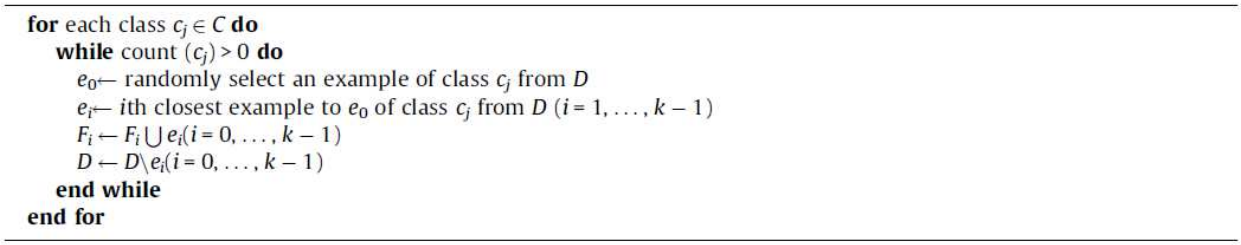
We recall that the main objective of partitioning is to avoid over or under-fitting. The training dataset is used for model fitting, in a preliminary phase, through the pairing of the input with the output; the validation dataset's main goal is to monitor and tune the model, being also used for model assessment; the test partition main goal is to evaluate how the model will work with new data, that has never been presented to the model. In SAS Enterprise Miner there is the possibility of using random sampling, stratified random sampling, or a user-defined partition to create these datasets with the following default proportions: 40%/30%/30%. Within the present thesis, these have been changed to 50%/25%/25% or to 60%/20%/20% whenever appropriate or to 70%/30%, when there is no need or enough data to create a test set.

In the article "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed" (López et al., 2014), a problem known as dataset shift is addressed. It is described as consisting of a different data distribution between the training and test partitions and is also presented as being more severe in classification with imbalanced datasets. In order to prevent this situation, the authors propose a specific validation technique for the partitioning of data, known as ''Distribution optimally balanced stratified cross-validation'' (DOB–SCV), which basically tries to assure that each partition contains enough representatives of every region, by placing close-by samples on different folds. According to the same authors, there are three possible types of dataset shift:

- Prior probability shift: It takes place when the class distribution is different between the training and test sets and, in an extreme case, could result in the training set not having a single example of a class. This kind of problem is normally prevented by a applying a SCV technique.

- Covariate shift: It happens when input attribute values have different distributions between the training and test sets.
- Concept shift/drift: This problem occurs when the relationship between the input and class variables changes, presenting itself as the hardest challenge among the different types of dataset shift.

The DOB–SCV technique attempts to alleviate the problem of covariate shift, preventing, at the same time, prior probability shift. Its pseudo-code is shown in the figure below (picture 3.2):



Source: López et al., 2014

Picture 3.2 - DOB–SCV partitioning method

When working with small datasets and in order to obtain more robust results (as well as to ensure computational feasibility), Lessmann et al. (2015) propose the Nx2-fold cross-validation technique, which involves the following steps, where the N parameter is set depending on dataset size: “i) randomly splitting a data set in half, (ii) using the first and second half for model building and evaluation, respectively, (iii) switching the roles of the two partitions, and (iv) repeating the two-fold validation N times.”

In the paper ”Data Mining and the Case for Sampling - Solving Business Problems Using SAS® Enterprise Miner™ Software” (Milley et al., 1998), cross-validation is also considered better than partitioned sample validation. Bootstrapping (which consists of repeatedly analyzing sub-samples of the data) is also pointed out as an appropriate technique when dealing with small samples.

3.3. MODEL BUILDING

For the reasons presented earlier, in the present thesis the following families of classifiers have been taken in consideration:

- Random forests
- Gradient Boosting
- Support Vector Machines
- Ensembles of neural networks and or decision trees

It should be noted, however, that a bigger emphasis has been placed on random forests due to their advantages at various levels, namely in terms of accuracy, capacity to handle large datasets with thousands of variables, including badly unbalanced ones, as well as extensive missing values and many-valued categorical variables.

In fact (Breiman, 2001), the classifier at hand is considered to perform very well in comparison to many others, including support vector machines and neural networks due to a strategy that may appear to be counterintuitive but turns out to be quite effective (Liaw & Wiener, 2002): contrary to standard trees, where the best split among all variables is used to split each node, in a random forest, the best split among a subset of predictors randomly chosen at each node is used for that effect, yielding a low correlation. On the other hand, as the number of trees in the forest is increased, the generalization error converges a.s. to a limit, minimizing or even eliminating overfitting.

In comparison to other powerful tree-based modelling techniques, such as stochastic gradient boosting (Freeman, Moisen, Coulston, & Wilson, 2016), the random forest classifier is considered to be more user friendly, less prone to overfitting and less sensitive to parameter tuning. RF also performs better in the presence of correlated predictor variables, as it tends to spread importance among more variables than SGB. Another important advantage over SGB, especially when small datasets are involved, is the possibility of not needing to set aside an independent test set due to the out-of-bag option for model evaluation.

It is also worth noting the papers "Leveraging Ensemble Models in SAS" (Maldonado, Dean, Czika, & Haller, 2014) and “The More Trees, the Better! Scaling Up Performance Using Random Forest in SAS® Enterprise Miner (Panneerselvam, 2015)”, which have served as a guide for the development of the models that have been implemented in the present thesis. In the first one, different types of tree based ensemble models, such as boosting, bagging, and model averaging are presented, based on existing SAS nodes, including HP Forest, Model Comparison, Start and End Groups, as well as on custom coding. The second one provides a simple, clear and pragmatic approach to the implementation of a prediction model based on the RF algorithm available in SAS 9.4 (HP Forest).

3.4. BRIEF OVERVIEW OF THE UNDERLYING SAS NODES

For clarity of understanding, this section aims at providing a brief description of the SAS nodes supporting the models previously presented, based on SAS® Enterprise Miner® 14.1 Reference Help and a SAS paper by Brown & Mues (2012).

The **HP Forest** node produces a predictive model known as forest, which consists of several decision trees differing from each other in the following ways: the training data for a tree consists of a sample without replacement within all available observations; the input variables that are used to split a node are randomly selected from all available inputs. Regarding other aspects, the training of trees in a forest is similar to the one applied to standard trees. For that effect, the three main available options comprise the number of trees, the number of inputs for a node and the sampling strategy, which can be fine tuned through the following node properties, respectively: Maximum Number of Trees, Number vars to consider in split search and Proportion of obs. in each sample. It should also be noted that this algorithm differs from Leo Breiman's bagging algorithm (Breiman, 2001), since it samples the original data without replacement, in order to provide more variability between the trees, especially when larger training sets are involved. Another requirement to take in consideration relates to the fact that HP Forest does not generate DATA Step score code, making it necessary to use a special procedure (PROC HP4SCORE) to access HP Forest score code.

A **support vector machine (SVM)** consists of a supervised machine-learning method used to perform classification and regression analysis, having as its basic principle the construction of a maximum-margin hyperplane in a transformed feature space. The exact transformation doesn't need to be specified, though, since SVM models recur to the principle of kernel substitution in order to turn them into a general (nonlinear) model. Only binary classification problems (including polynomial, radial basis function, and sigmoid nonlinear kernels) are supported by the HP SVM node, which does not perform multi-class problems or support vector regression.

The **Gradient Boosting** node uses the algorithms described in "A Gradient Boosting Machine" and "Stochastic Gradient Boosting" by Jerome Friedman. It basically consists of an ensemble algorithm that improves the accuracy of a predictive function by incrementally minimising the error term. The base learner is most commonly a tree and, as such, makes no assumptions about the distribution of the data. However and in comparison to a single tree, is less prone to overfit.

Neural networks (NN) consist of mathematical representations which try to mimic the functioning of the human brain, being highly flexible in modelling non-linear associations between input and target variables. Of the various possible architectures, the most widely used one is the Multilayer Perceptron (MLP), which normally comprises an input layer (obtaining the values of input variables), a hidden layer (providing the required nonlinearity) and an output layer (corresponding to one neuron, in the case of binary target variables). The inputs are processed by each neuron and the resulting output is transmitted to the subsequent layer neurons. During the training process, a weight is assigned to each of these connections and the output of a hidden neuron is computed through the application of an activation function (such as a logistic function, for instance) to the weighted inputs and the associated bias term. During the estimation process and after a random initialization, the network weights are iteratively adjusted in order to minimise an objective function such as the sum of squared errors.

Decisions trees consist of classification and estimation tools, based on algorithms that split the data into smaller branch-like segments and assign a class to each observation, through a series of leaf nodes with a root at the top containing the entire dataset. Two of the major advantages of this modelling technique over algorithms such as Neural Networks, for instance, relate, on the one side, to the production of output that can be represented by easily interpretable rules (written in a language such as sql, for instance) and, on the other side, to the treatment of missing data, which are used as inputs, based on surrogate rules whenever necessary. In SAS Enterprise Miner, the options to choose the splitting criteria and to the determine the tree construction method include popular features such as CHAID (Chi-square automatic interaction detection), as well as those described in Classification and Regression Trees (Breiman, Friedman, Stone, & Olshen, 1984).

The **Ensemble node** supports the creation of new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple preceding models, through one of the following methods: average, maximum and voting (available for categorical targets only). Some requirements have to be taken under consideration, though: this node supports only one target variable and prior probabilities have to be specified after the modelling nodes, in order to obtain correct fit statistics for the combined unadjusted posterior probabilities.

4. RESULTS

The main objective of this chapter is to present and discuss the results that were obtained throughout the study and their implications, within a comparative analysis with previous studies. For that effect, the main results will be presented according to each of the considered datasets and models, in order to better evidence differences that may arise between them and analyse whether they are supported by relevant literature.

4.1. DATASETS AND SAMPLES DESCRIPTION

As mentioned earlier, the following datasets have been considered:

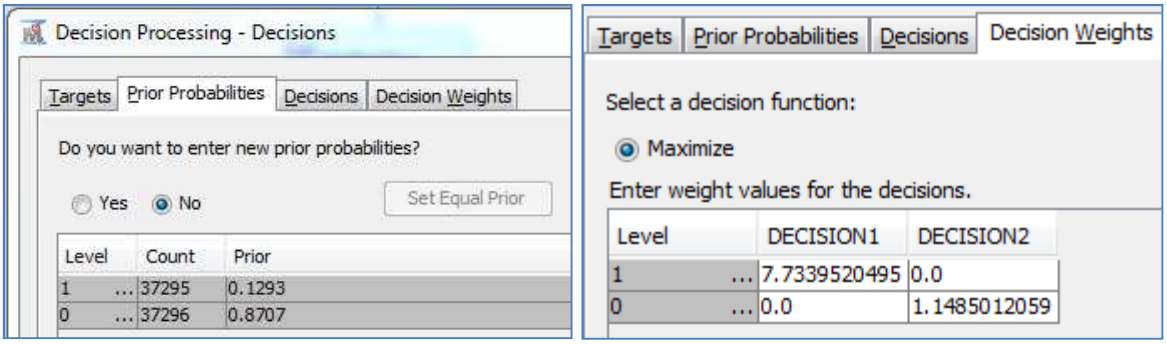
- One mixing nominal and numerical variables containing only internal data ("Internal_mixed");
- One containing only internal data where all nominal variables are converted to numbers, through the WOE technique ("Internal_num");
- Two additional versions of the former data sets enriched with external data (including ESCO), namely: "Complete_mixed" and "Complete_num".

The datasets containing both internal and external data comprise 42 input variables, of which 6 were taken from sources external to the Portuguese PES operational system. The number of observations is common to all datasets, totalling 1,062,651. However, due to performance issues and the need to obtain more balanced datasets, only 10% of the observations have been taken in consideration, containing equal proportions of positive and negative events. For that effect, the following settings have been considered in SAS Sample node (picture 4.1):

Property	Value
Train	
Variables	
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No

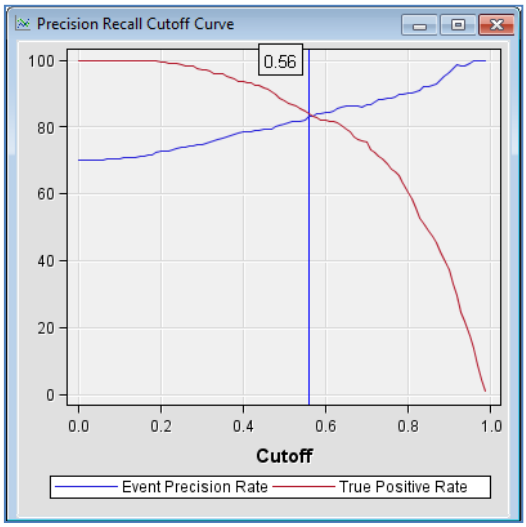
Picture 4.1 - Sample Node Settings

After data partitioning (normally, 70% for training and 30% for validation), the posterior probabilities of the models had to be adjusted, namely through the inversion of prior probabilities, as illustrated below (picture 4.2):



Picture 4.2 - Decisions Node Settings

In order to determine a good cut-off for the predicted probabilities, SAS cut-off node has also been used, namely the option which finds the intersection between the Event Precision Rate and the True Positive, which in, the example below, equals 0.56 (picture 4.3).



Picture 4.3 – Cut-off node

4.2. VARIABLE SELECTION

As highlighted in the paper "Identifying and Overcoming Common Data Mining Mistakes" (Wielenga, 2007), variable selection should not be restricted to just one method in order to avoid missing potentially important predictors, especially when a large number of variables is involved. A safe strategy thus consists of creating a pool of predictors based on the variables that were selected by any of the methods. SAS EM provides a variety of nodes for that effect, namely: (HP) Variable Selection, HP Forest, (HP) Regression, Stat Explore and Interactive Grouping (WOE). This step should be undertaken after data partition, in order to avoid overfitting and after missing values imputation, in the case of using regression nodes as a variable selection method.

In tables 4.1 to 4.3, below, the complete set of variables under consideration (including internal and external data), is presented, alongside the results obtained within the variable selection step, after sampling and data partition. The variables rejected by each of the methods were signalled in red. For that effect and in the case of the HP Forest node, variables with an out-of-bag margin reduction less than or equal to zero were rejected; as far as the HP Regression Node is concerned, non-null parameter estimates and p-values <0.05 were considered; lastly, default settings have been taken under consideration in the case of the remaining nodes.

Variable Name	Role	Level	Description	HP Forest	HP Variable Selection	Stat Explore	WOE	HP Regression
AP_FREGUESIA	INPUT	NOMINAL	Job seeker's parish of residence	10	37	1	8	S
AP_CPP_OFERTA	INPUT	NOMINAL	Job offer's occupation	2	11	2	2	S
AP_CC	INPUT	NOMINAL	Job seeker's municipality of residence	4	31	3	1	S
AP_APOIADA	INPUT	BINARY	Whether job offer benefits from financial incentives	1	2	4	4	S
AP_NUT3	INPUT	NOMINAL	Level 3 of the nomenclature of territorial units for statistics	5	3	5	3	S
ofa_CAE2	INPUT	NOMINAL	Job offer's sector of activity at 2 digits level	3	1	6	5	S
AP_CPP_PRETENDIDA	INPUT	NOMINAL	Job seeker's intended occupation	22	38	7	13	S
AP_CPP_ANTERIOR	INPUT	NOMINAL	Job applicant's previous occupation	19	30	8	12	S
Nascimentos_nr	INPUT	INTERVAL	Births of enterprises (external variable - INE)	20	26	9	11	S
ISDR	INPUT	INTERVAL	Regional development composite index (Overall index) by Geographic localization (NUTS - 2013)	12	7	10	6	S
PPC	INPUT	INTERVAL	Proportion of purchasing power (external variable - INE)	9	12	11	7	S
AP_GOE	INPUT	BINARY	Job offer's manager	6	32	12	10	N
AP_DISTRITO	INPUT	NOMINAL	Job seeker's district of residence	7	9	13	9	N
Taxa_sobrev_2antes	INPUT	INTERVAL	Survival rate of enterprises 2 years before	16	20	14	14	N

Table 4.1 - Variable selection

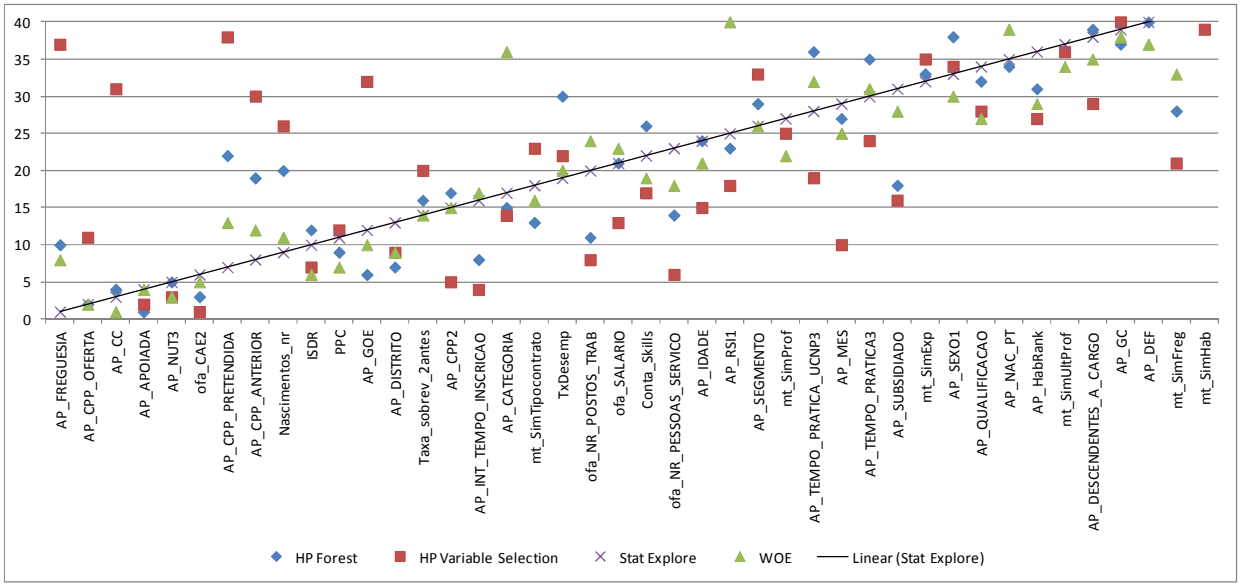
Variable Name	Role	Level	Description	HP Forest	HP Variable Selection	Stat Explore	WOE	HP Regression
AP_CPP2	INPUT	NOMINAL	Job seeker's intended occupation at 2 digit level	17	5	15	15	N
AP_INT_TEMPO_INSCRICAO	INPUT	INTERVAL	Registration period as a job seeker with PES	8	4	16	17	S
AP_CATEGORIA	INPUT	NOMINAL	Category of job applicant (unemployed or employed; searching for a first or a new job; engaged in an active employment measure; unavailable)	15	14	17	36	N
mt_SimTipocontrato	INPUT	INTERVAL	Similarity between job offer's demanded conditions and job seeker's profile in what relates to type of work contract (fixed term...)	13	23	18	16	S
TxDesemp	INPUT	INTERVAL	Quartely unemployment rate	30	22	19	20	N
ofa_NR_POSTOS_TRAB	INPUT	INTERVAL	Number of vacancies contained in the job offer	11	8	20	24	S
ofa_SALARIO	INPUT	INTERVAL	Job offer's (monthly) wage	21	13	21	23	N
Conta_Skills	INPUT	INTERVAL	Number of common skills between job offer's occupation and job seeker's intended occupation (external variable - ESCO)	26	17	22	19	S
ofa_NR_PESSOAS_SERVICO	INPUT	INTERVAL	Number of people working in the company responsible for the job offer	14	6	23	18	S
AP_IDADE	INPUT	INTERVAL	Job seeker's age	24	15	24	21	S
AP_RSI1	INPUT	BINARY	Whether job seeker benefits from social benefits	23	18	25	40	N
AP_SEGMENTO	INPUT	NOMINAL	Job seeker's long term unemployment risk profile	29	33	26	26	S
mt_SimProf	INPUT	INTERVAL	Similarity between job offer's demanded conditions and job seeker's profile in what relates to occupation	25	25	27	22	S
AP_TEMPO_PRATICA_UCNP3	INPUT	INTERVAL	Job seeker's experience at last job (nr. months)	36	19	28	32	S
AP_MES	INPUT	NOMINAL	Month of the job referral	27	10	29	25	S
AP_TEMPO_PRATICA3	INPUT	INTERVAL	Job seeker's experience at intended occupation (nr. months)	35	24	30	31	N
AP_SUBSIDIADO	INPUT	BINARY	Whether job seeker is receiving unemployment benefits	18	16	31	28	S

Table 4.2 - Variable selection (cont.)

Variable Name	Role	Level	Description	HP Forest	HP Variable Selection	Stat Explore	WOE	HP Regression
mt_SimExp	INPUT	INTERVAL	Similarity between job offer's demanded conditions and job seeker's profile in what relates to experience	33	35	32	41	N
AP_SEXO1	INPUT	BINARY	Job seeker's gender	38	34	33	30	S
AP_QUALIFICACAO	INPUT	NOMINAL	Job applicant's level of qualification	32	28	34	27	N
AP_NAC_PT	INPUT	BINARY	Whether job seeker has portuguese nationality	34	41	35	39	N
AP_HabRank	INPUT	NOMINAL	Job seeker's qualifications	31	27	36	29	S
mt_SimUltProf	INPUT	INTERVAL	Similarity between job offer's demanded conditions and job seeker's profile in what relates to last occupation	42	36	37	34	S
AP_DESCENDENTES_A_CARGO	INPUT	INTERVAL	Number of people at care of job seeker	39	29	38	35	S
AP_GC	INPUT	BINARY	Job seeker's career manager	37	40	39	38	N
AP_DEF	INPUT	BINARY	Whether the job seeker has a disability	40	42	40	37	N
mt_SimFreg	INPUT	INTERVAL	Similarity between job offer's demanded conditions and job seeker's profile in what relates to parish of work/residence	28	21	41	33	S
mt_SimHab	INPUT	INTERVAL	Similarity between job offer's demanded conditions and job seeker's profile in what relates to minimum education	41	39	42	42	N
AP_ID_APRESENTACAO	ID	INTERVAL	Observation id	--	--	--	--	--
AP_COLOCADO	TARGET	BINARY	Job referral's sucess (whether job seeker was placed at the vacancy)	--	--	--	--	--
AP_DATA_APRES	TIMEID	INTERVAL	Date of the job referral	--	--	--	--	--

Table 4.3 - Variable selection (cont.)

From the analysis of the table, in which the variable worth obtained from the Stat Explore Node is also presented in a decreasing order of importance, it is possible to conclude that all variables are selected by one of the five methods at hand, with the HP Regression Node rejecting a greater number of variables than the remaining ones (namely, fifteen, half of which were also rejected by the WOE method, used before missing values imputation) and the HP Variable Selection (based on a decision tree) presenting a higher dispersion, as depicted in the graph below (picture 4.4):



Picture 4.4 - Comparison of variable selection methods

Within the 26 variables that are common to all of the considered methods, only three, however, belong to a common top ten, namely: whether job offer benefits from financial incentives (AP_APOIADA), job offer's sector of activity at 2 digits level (CCAE2) and level 3 of the nomenclature of territorial units for statistics (NUTS III). On the other hand, variables which are normally perceived as having a greater importance are ranked in a middle or even low position in terms of importance. It's the case, for instance, of job seeker's age (AP_IDADE) and gender (AP_SEXO1), as well as of the variable measuring the risk of becoming a long term unemployed (AP_SEGMENTO) and the attribute capturing the common skills between the job offer's occupation and the one intended by the job seeker (Conta_Skills).

It should also be noted that the initial set of candidate variables comprised 28 additional inputs (from internal and external sources), that were rejected for the following reasons: high correlation with other input variables (superior to 80%, in absolute terms and according to the Spearman correlation coefficient); more than 50% of missing values; not contributing to an improved performance by some of the most robust algorithms, such as random forests. Within these rejected cases, it is possible to find some of the variables that are currently being used in the previously mentioned matching tool (such as being a single parent or being part of an unemployed couple).

For clarity of understanding, the main variable statistics are presented below (tables 4.4 to 4.8), based on a filtered balanced sample taken from the most complete dataset (statistics for the filtered, non sampled dataset are included in the appendix to this chapter). As previously mentioned, missing values and many-valued categorical variables were maintained within the datasets meant to be used by robust algorithms such as random forests, gradient boosting and decision trees ensembles. In the case of the remaining models, various transformations were applied to the relevant data, through techniques such as weight of evidence and imputation of missing values.

Before data partition, transformation and imputation

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	AP_APOIADA	INPUT	2	0	0	61.16	1	38.84
TRAIN	AP_CATEGORIA	INPUT	4	0	2	78.81	1	8.5
TRAIN	AP_CC	INPUT	236	0	1111	3.94	1317	3.13
TRAIN	AP_CPP2	INPUT	43	0	93	9.21	52	9.03
TRAIN	AP_CPP_ANTERIOR	INPUT	513	1950	9.94	93290	5.25	
TRAIN	AP_CPP_OFERTA	INPUT	513	0	93290	5.77	94120	5.21
TRAIN	AP_CPP_PRETENDIDA	INPUT	513	0	41100	6.75	93290	4.8
TRAIN	AP_DEF	INPUT	2	0	0	98.57	1	1.43
TRAIN	AP_DISTRITO	INPUT	16	0	13	17.51	11	15.37
TRAIN	AP_FREGUESIA	INPUT	513	0	131315	1.47	21121	1.39
TRAIN	AP_GC	INPUT	2	0	0	98.43	1	1.57
TRAIN	AP_GOE	INPUT	2	0	0	56.12	1	43.88
TRAIN	AP_HabRank	INPUT	7	0	7	33.42	5	25.78
TRAIN	AP_MES	INPUT	12	0	10	9.75	5	9.23
TRAIN	AP_NAC_PT	INPUT	2	0	1	94.44	0	5.56
TRAIN	AP_NUT3	INPUT	19	0	170	18.18	11A	17.8
TRAIN	AP_QUALIFICACAO	INPUT	9	1287	2	27.81	3	22.89
TRAIN	AP_RSI1	INPUT	2	0	0	94.48	1	5.52
TRAIN	AP_SEGMENTO	INPUT	4	13626	RB	41.4	RM	32.66
TRAIN	AP_SEXO1	INPUT	2	0	0	53.53	1	46.47
TRAIN	AP_SUBSIDIADO	INPUT	2	0	0	71.02	1	28.98
TRAIN	ofa_CAE2	INPUT	84	0	78	12.59	56	10.54
TRAIN	AP_COLOCADO	TARGET	2	0	1	50.98	0	49.02

Table 4.4 - Class variable summary statistics (before transformation and imputation)

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	AP_COLOCADO	TARGET	1	42262	50.9832
TRAIN	AP_COLOCADO	TARGET	0	40632	49.0168

Table 4.5 - Distribution of class target and segment variables (before transformation and imputation)

Variable Name	Data Role	Mean	Standard Deviation	Non Missing	Missing	Median	Skewness	Kurtosis
AP_DESCENDENTES_A_CARGO	INPUT	0.665162	0.853113	82416	478	0	1.002472	-0.07623
AP_IDADE	INPUT	35.80815	10.68242	82894	0	35	0.311923	-0.83422
AP_INT_TEMPO_INSCRICAO	INPUT	10.88635	11.97822	82894	0	6	1.382296	1.299317
AP_TEMPO_PRATICA3	INPUT	65.9869	78.46699	82473	421	36	1.421291	1.259539
AP_TEMPO_PRATICA_UCNP3	INPUT	64.21807	76.87358	81657	1237	32	1.503192	1.532389
Conta_Skills	INPUT	16.60737	24.63321	82894	0	4	1.85705	3.077118
ISDR	INPUT	99.68157	4.2906	82894	0	98.52	0.375025	-0.97189
Nascimentos_nr	INPUT	506.4369	857.0675	81248	1646	132	2.369415	5.254308
PPC	INPUT	0.960087	0.959002	82894	0	0.562	1.391532	1.135116
Taxa_sobrev_2antes	INPUT	59.82203	20.37629	81118	1776	60.85	-0.98499	2.646859
TxDesemp	INPUT	12.67599	2.121291	82894	0	13.4	-0.34969	-1.07005
mt_SimExp	INPUT	0.882151	0.322432	82894	0	1	-2.37049	3.619297
mt_SimFreg	INPUT	0.536217	0.279104	82894	0	0.67	-0.07392	-0.61296
mt_SimHab	INPUT	0.988359	0.052677	82894	0	1	-4.30403	16.52504
mt_SimProf	INPUT	0.370661	0.442224	82894	0	0	0.584892	-1.49779
mt_SimTipocontrato	INPUT	0.266666	0.442219	82894	0	0	1.055313	-0.88633
mt_SimUltProf	INPUT	0.613442	0.469135	82894	0	1	-0.44759	-1.73616
ofa_NR_PESSOAS_SERVICO	INPUT	85.44536	313.1728	80612	2282	7	6.649352	49.9851
ofa_NR_POSTOS_TRAB	INPUT	2.30124	3.668079	82894	0	1	4.321485	21.94012
ofa_SALARIO	INPUT	562.0756	158.4739	82894	0	505	2.82523	22.43773

Table 4.6 - Interval variable summary statistics (before transformation and imputation)

After data partition, transformation (WOE) and imputation of missing values (median)

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	AP_COLOCADO	TARGET	1	29582	50.9832
TRAIN	AP_COLOCADO	TARGET	0	28441	49.0168

Table 4.7 - Distribution of class target and segment variables

Variable Name	Standard Role	Mean	Deviation	Non missing	Missing	Median	Skewness	Kurtosis
WOE_AP_APOIADA	INPUT	-0.01035	0.566024	58023	0	0.437979	-0.47043	-1.77876
WOE_AP_CC	INPUT	-0.00895	0.60416	58023	0	0.125982	-0.31466	-1.1738
WOE_AP_CPP2	INPUT	-0.0012	0.263994	58023	0	0.031434	-0.33939	-0.97847
WOE_AP_CPP_ANTERIOR	INPUT	-0.00293	0.339598	58023	0	-0.07567	-0.56768	-0.86134
WOE_AP_CPP_OFERTA	INPUT	-0.01769	0.604397	58023	0	0.145703	-0.82731	-0.70869
WOE_AP_CPP_PRETENDIDA	INPUT	-0.00351	0.32793	58023	0	0.052551	-0.85937	-0.61558
WOE_AP_DESCENDENTES_A_CARGO	INPUT	-0.00003	0.038542	58023	0	-0.01093	-3.49348	35.99212
WOE_AP_DISTRITO	INPUT	-0.00159	0.391034	58023	0	0.053257	-0.0201	-0.70829
WOE_AP_FREGUESIA	INPUT	-0.00894	0.440014	58023	0	0.317846	-1.03001	-0.55541
WOE_AP_GOE	INPUT	-0.00302	0.403284	58023	0	0.35222	-0.25439	-1.93535
WOE_AP_HabRank	INPUT	-0.00012	0.094623	58023	0	-0.00953	-0.51826	-1.15106
WOE_AP_IDADE	INPUT	-0.00049	0.162024	58023	0	0.091176	-0.66247	-1.26223
WOE_AP_INT_TEMPO_INSCRICAO	INPUT	-0.00048	0.189677	58023	0	0.040499	-0.23254	-0.87539
WOE_AP_MES	INPUT	-0.00017	0.131757	58023	0	-0.02441	0.019914	-1.19822
WOE_AP_NUT3	INPUT	-0.00469	0.502103	58023	0	0.029981	-0.22544	-1.05384
WOE_AP_QUALIFICACAO	INPUT	-0.00014	0.118925	58023	0	-0.02294	-0.0425	-0.82797
WOE_AP_SEGMENTO	INPUT	2.77E-06	0.14306	58023	0	0.025895	0.838105	-0.35635
WOE_AP_SEXO1	INPUT	-0.00008	0.095849	58023	0	-0.08962	0.136369	-1.98147
WOE_AP_SUBSIDIADO	INPUT	2.30E-06	0.128787	58023	0	-0.08215	0.929837	-1.13544
WOE_AP_TEMPO_PRATICA3	INPUT	-0.00008	0.066487	58023	0	0.029209	-1.48157	0.279918
WOE_AP_TEMPO_PRATICA_UCNP3	INPUT	-0.00002	0.035593	58023	0	0.001655	-0.98349	-0.40151
WOE_Conta_Skills	INPUT	0.000077	0.186255	58023	0	-0.07025	0.778727	-0.77762
WOE_ISDR	INPUT	0.0018	0.41277	58023	0	-0.17464	0.600189	-1.18946
WOE_Nascimentos_nr	INPUT	-0.00329	0.396927	58023	0	0.065963	-0.3526	-0.20629
WOE_PPC	INPUT	-0.00104	0.397728	58023	0	0.044758	0.092738	-0.82417
WOE_Taxa_sobrev_2antes	INPUT	0.00016	0.275328	58023	0	0.071379	0.527908	-1.12238
WOE_TxDesemp	INPUT	0.000255	0.198529	58023	0	-0.11358	0.986518	-0.93224
WOE_mt_SimFreg	INPUT	-0.00001	0.033673	58023	0	0.03024	-0.2147	-1.95397
WOE_mt_SimProf	INPUT	-0.00002	0.157451	58023	0	-0.12934	0.676337	-1.36394
WOE_mt_SimTipocontrato	INPUT	-0.00202	0.25198	58023	0	0.149814	-1.05699	-0.8828
WOE_mt_SimUltProf	INPUT	-9.76E-06	0.029024	58023	0	0.020354	-0.72372	-1.47627
WOE_ofa_CAE2	INPUT	-0.00379	0.428611	58023	0	0.197067	-0.30696	-1.37362
WOE_ofa_NR_PESSOAS_SERVICO	INPUT	-0.00081	0.195959	58023	0	0.106084	-0.68685	-1.11397
WOE_ofa_NR_POSTOS_TRAB	INPUT	-0.00158	0.201654	58023	0	0.089728	-1.75577	1.082751
WOE_ofa_SALARIO	INPUT	5.06E-06	0.148272	58023	0	-0.10058	0.816085	-1.01017

Table 4.8 - Interval variable summary statistics

4.3. MODEL COMPARISON RESULTS

For the comparison of the main results obtained from the different models at hand, summarized in table 4.9, below, the metrics F-value and AUC (Area Under the ROC Curve) have been taken in consideration, alongside the model selection criteria used by the SAS model comparison node. It should also be noted that two unbalanced datasets were considered in addition to the ones previously presented, in order to evaluate the robustness of the different classifiers at this level.

Dataset	Sampling	Model	Accuracy	Error	Precision	Recall	F-Measure	AUC	Best*	Cut-off
Complete_mixed	10%, balanced	HP Forest 50	0.755	**0.245**	0.763	0.752	**0.758**	**0.834**		
Complete_mixed	10%, balanced	Gradient Boosting 50	0.711	0.289	0.714	0.725	0.719	0.780		
Complete_mixed	10%, balanced	Decision Trees Ensemble	0.748	0.252	0.757	0.751	0.754	0.817	Y	0.500
Complete_num	10%, balanced	HP Forest 20	0.742	**0.258**	0.739	0.766	**0.752**	**0.818**		
Complete_num	10%, balanced	HP SVM 25	0.726	0.274	0.726	0.742	0.734	0.795		
Complete_num	10%, balanced	Gradient Boosting 50	0.673	0.327	0.649	0.783	0.710	0.708		
Complete_num	10%, balanced	NN Ensemble (10-50)	0.740	0.260	0.753	0.728	0.740	0.816	Y	0.480
Internal_mixed	10%, balanced	HP Forest 50	0.756	**0.244**	0.762	0.744	**0.753**	**0.837**	Y	0.490
Internal_mixed	10%, balanced	Gradient Boosting 50	0.710	0.290	0.708	0.714	0.711	0.777		
Internal_mixed	10%, balanced	Decision Trees Ensemble	0.745	0.255	0.749	0.736	0.743	0.812		
Internal_num	10%, balanced	HP Forest 20	0.739	0.261	0.733	0.750	**0.741**	0.812		
Internal_num	10%, balanced	HP SVM 25	0.723	0.277	0.716	0.738	0.727	0.794		
Internal_num	10%, balanced	Gradient Boosting 50	0.713	0.287	0.709	0.719	0.714	0.780		
Internal_num	10%, balanced	NN Ensemble (30+)	0.742	0.258	0.748	0.729	0.738	**0.815**	Y	0.480
Complete_mixed	10%, unbalanced	HP Forest 50	0.883	**0.117**	0.836	0.092	0.166	**0.817**	Y	0.230
Complete_mixed	10%, unbalanced	Gradient Boosting 50	--	--	--	--	--	--		
Complete_num	10%, unbalanced	HP Forest 20	0.883	**0.117**	0.602	0.227	0.330	0.768	Y	0.400
Complete_num	10%, unbalanced	HP SVM 25	0.873	0.127	--	0.000	--	0.782		
Complete_num	10%, unbalanced	Gradient Boosting 50	0.873	0.127	--	0.000	--	0.682		
Complete_num	10%, unbalanced	NN Ensemble (10-50 HU)	0.881	0.119	0.589	0.203	0.302	**0.800**		

*According to SAS Model Comparison node's selection criteria: average profit in the case of balanced samples and missclassification rate, in the case on unbalanced samples.

Table 4.9 - Main models comparison

SAS Enterprise Miner default settings were considered in the great majority of the algorithms, with the following exceptions:

- HP Forest: number of trees (20 or 50, instead of the predefined 100);
- Neural Networks: number of hidden units (10 to 50).

Based on the AUC measure, it seems possible to conclude that the random forest classifier presents the best performance within four of the six datasets under consideration, followed closely by the neural network ensembles that dominate in the remaining two, namely in the unbalanced numerical dataset (presenting very low F-values) and the balanced dataset containing only internal numerical data. However, according to SAS model comparison results, the random forest classifier would only be chosen as the best model within the unbalanced datasets and the internal dataset containing categorical and numerical data. In the case of the remaining numerical and mixed datasets, the neural networks and decision trees ensembles stand as the chosen models, respectively.

Another conclusion that is possible to draw from the obtained results consists of the apparently very close proximity between the best performing classification models regardless of the inclusion of external data. In face of the favourable results presented in relevant research literature at this level

(Bollinger et al., 2012), this finding was further investigated through the calculation of the Friedman statistical test (Demšar, 2006), designed for the comparison of algorithms on multiple datasets (picture 4.5):

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_{j=1}^{K} AR_j^2 - \frac{K(K+1)^2}{4}\right] \text{ where } AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j$$

Source: I. Brown, 2012

Picture 4.5 - Friedman test statistic

where:

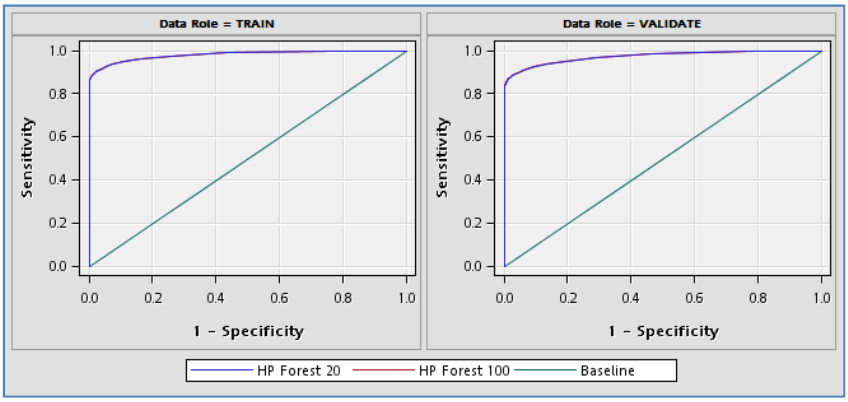
- *AR* corresponds to the average rank of the different classifiers on each dataset;
- r_i^j is the rank of classifier *j* on data set *i*;
- *D* represents the number of datasets used;
- *K* consists of the number of classifiers;
- X_F^2 follows a Chi-square distribution with *k-1* degrees of freedom (when its value is large enough, the null hypothesis of equality between the techniques can be rejected).

Since the number of models within the datasets mixing numerical and categorical variables was too small for that effect, the Friedman test statistic was only calculated for the numerical datasets. After ranking the different algorithms based on the AUC and average profit measures, as illustrated in table 4.10, below, test statistics of 5.4 (0.07) and 6 (0.05) were obtained for the complete_num and internal_num, respectively. Although the obtained p-values should be treated with caution (due to the small number of datasets and classifiers under consideration), it seems possible to conclude that the models yield very similar performances within the two types of datasets being analysed.

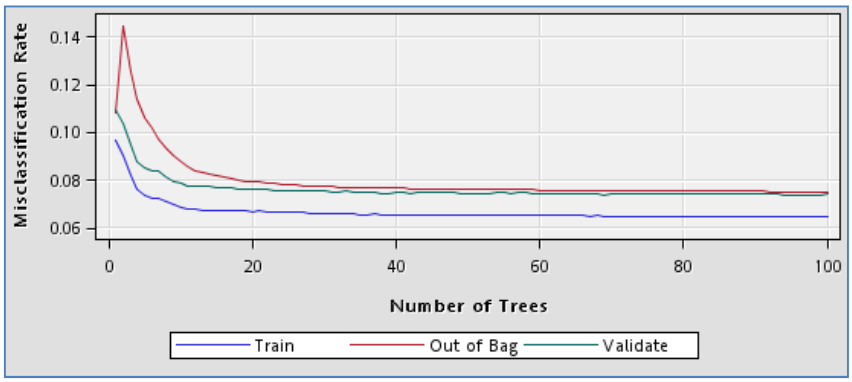
Dataset	Model	AUC	Rank_AUC	Aver_Prof (AvP)	Rank_AvP
Complete_num	HP Forest 20	0.818	1	3.954	2
Complete_num	HP SVM 25	0.795	3	3.946	3
Complete_num	Gradient Boosting 50	0.708	4	3.943	4
Complete_num	NN Ensemble (10-50)	0.816	2	3.971	1
Internal_num	HP Forest 20	0.812	2	3.869	2
Internal_num	HP SVM 25	0.794	3	3.868	3
Internal_num	Gradient Boosting 50	0.780	4	3.864	4
Internal_num	NN Ensemble (30+)	0.815	1	3.886	1
Friedman Test			5.4 (0.06721)		6 (0.04979)

Table 4.10 - Friedman test statistic results for the numerical datasets

An attempt to implement a SMOTE code approach was also undertaken, based on the algorithm provided by the R package "DMwR" (Package, Functions, & Torgo, 2015). However, the very favourable results obtained in terms of the model performance as measured by AUC and the misclassification rate (pictures 4.6 and 4.7) are not supported by the ones obtained from a new dataset, pointing to a possible overfitting issue, as further explained, in section 4.6.



Picture 4.6 - ROC Curves (SMOTED dataset)



Picture 4.7 - Misclassification Rate (SMOTED dataset)

Lastly and for illustrative purposes, the remaining ROC curves' charts are included in the appendix to this chapter, alongside the diagram flows for the different models under consideration.

4.4. MODEL STABILITY EVALUATION

In order to test for the stability of the models' performances, three additional samples (generated with different random seeds) were considered for the HP Forest and Gradient Boosting models, within the complete_mixed dataset, together with a balanced dataset using all the available events (100%) combined with an equal proportion of the non-events. The AUC values thus obtained are presented in table 4.11, below, and point to a stable performance, except for the Gradient Boosting model within the sample containing 100% of the events (which performs only one of the pre-defined 50 iterations, most probably due to the large number of observations involved, namely 195 628):

Sample	Model	AUC
Robustness sample 1 (10%)	HP Forest	0.837
	Gradient Boosting	0.78
Robustness sample 2 (10%)	HP Forest	0.836
	Gradient Boosting	0.78
Robustness sample 3 (10%)	HP Forest	0.835
	Gradient Boosting	0.781
Robustness sample 4 (100%)	HP Forest	0.846
	Gradient Boosting	0.5

Table 4.11 - Stability of results on various samples

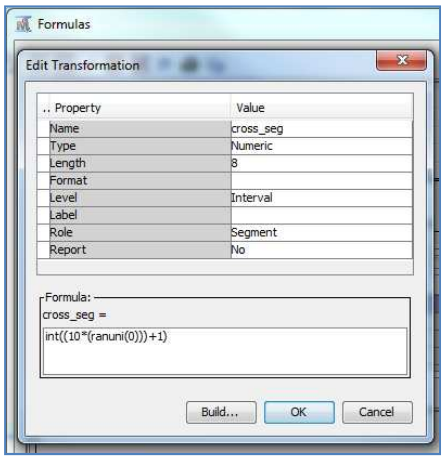
A more robust technique for testing the stability of predictive models consists of the previously mentioned f-fold cross validation (Milley et al., 1998), supported by the group processing facility in SAS Enterprise Miner, through the use of a transformation before entering the looping process (Schubert, 2010), as illustrated in the figure below[2] (picture 4.8):

Source: Schubert, 2010

Picture 4.8 - Use of a transformation node for f-fold cross validation

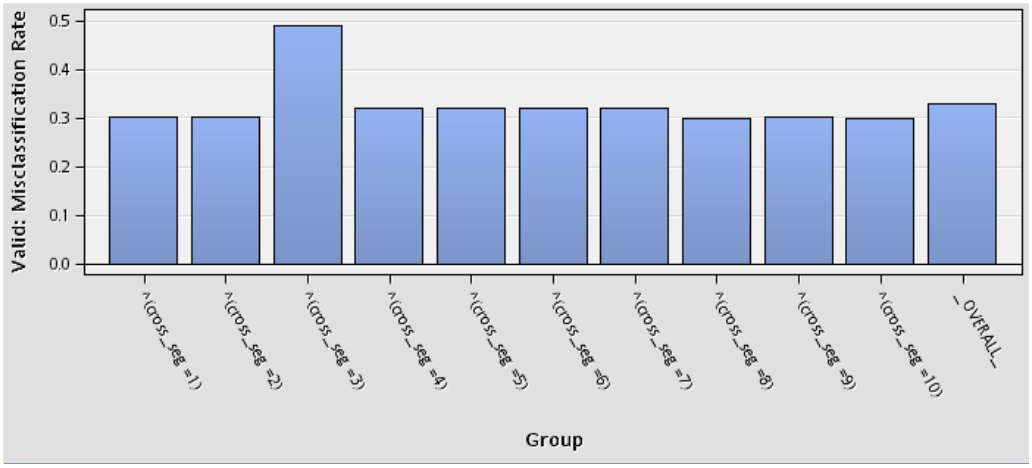
For that effect, the formula *int((f*(ranuni(0)))+1)*, where f is the number of segments (or folds), is created through the transform variables node, with the role "Segment", in order to allow for its immediate use (picture 4.9):

[2] A decision tree is used since the HP Forest is node not supported. All other High Performance Data Mining nodes are supported, with the additional exception of the HP SVM node with the Optimization Method property set to Active.



Picture 4.9 - Cross validation segment id creation

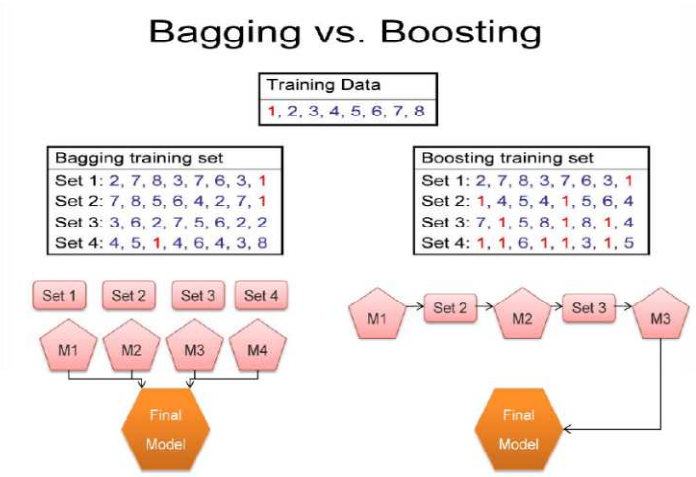
In the figure below (picture 4.10), the misclassification rate resulting from each run is presented, being possible to conclude that the majority of the segments is relatively in line with the model's overall performance, with the exception of the third one:



Picture 4.10 - Overall and segments' misclassification rate

SAS Enterprise Miner group processing facility provides two other ensemble algorithms that can be used to improve predictive models' accuracy and stability - bagging and boosting, which differ with respect to the sampling approach (Schubert, 2010). In the first case, an unweighted resampling with replacement is adopted. Being independent of each other (since each observation has the same chance to be drawn into the training set), the loops can be run in parallel and the final model output is obtained through the averaging of the probabilities generated by each iteration. In the second case, a weighed resampling is performed in order to improve the accuracy of the model by giving a bigger weight to observations that are more difficult to predict or classify and a lower one to correctly identified cases. Thus in each subsequent iteration a sample with a proportion of more misclassified observations in the previous run will be drawn, within a sequential processing of the

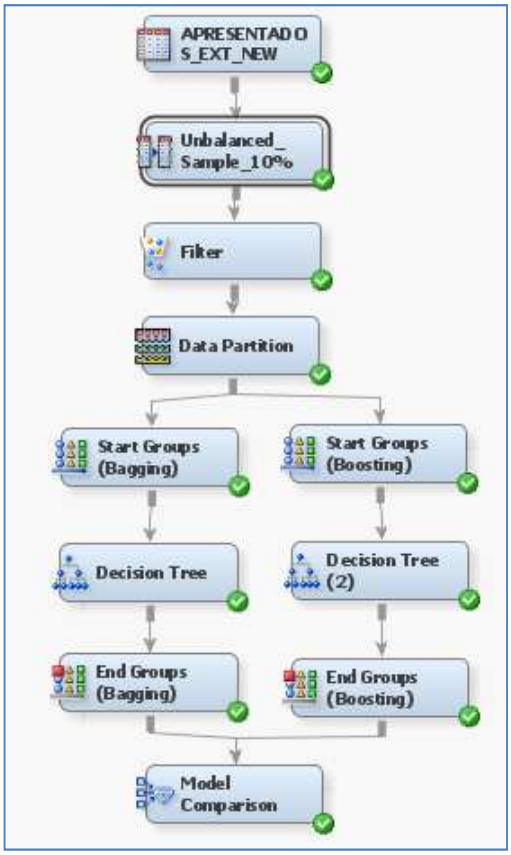
algorithm, where the final model outcome is obtained through a weighted majority voting method (picture 4.11).



Source: Schubert, 2010

Picture 4.11 - Bagging vs. Boosting

In order to evaluate the performance of these models within the present thesis, the complete_mixed unbalanced dataset was considered, according to the flow depicted below (picture 4.12):



Picture 4.12 - Bagging and boosting flows (unbalanced dataset)

The obtained results are summarized in table 4.12, below, from the analysis of which is possible to conclude that the bagging model is considered the best one, in spite of the low values presented by the Recall and F-Measure:

Dataset	Model	Accuracy	Error	Precision	Recall	F-Measure	AUC	Best*
Unbalanced	Bagging	0.881	0.119	0.640	0.138	0.227	0.767	Y
	Boosting	0.760	0.240	0.162	0.214	0.184	0.794	

*SAS selection criteria: missclassification rate

Table 4.12 - Bagging and boosting models' comparison

Another important step towards obtaining the most accurate and generalizable model consists of the fine tuning of the algorithm's most important available parameters. Being one of the best performing and robust models, the HP Forest algorithm was chosen for this effect.

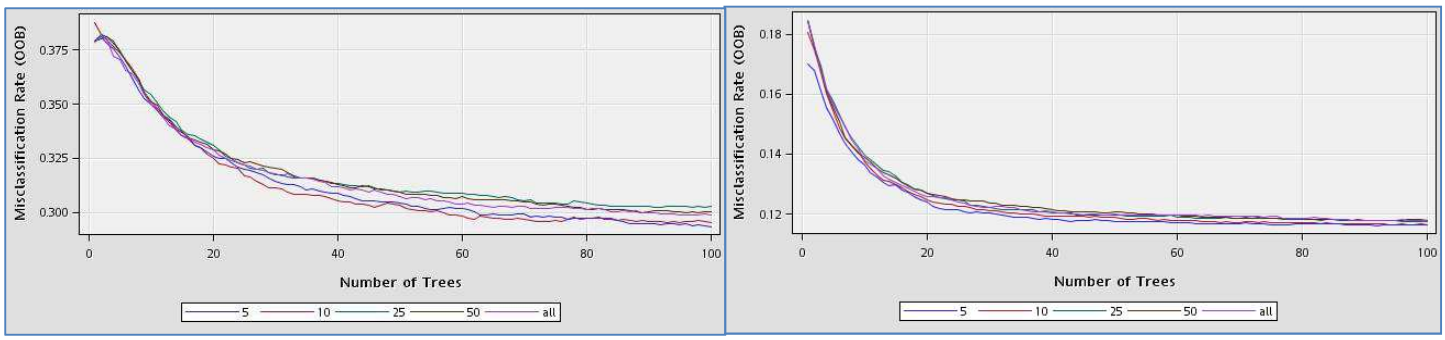
As previously mentioned and with the exception of the number of trees (due to processing capacity reasons), default settings (depicted in picture 4.13) have been considered within the analysis that was carried out based on the algorithm at hand.

Property	Value
General	
Node ID	HPDMForest
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of obs in each samp	0.6
Number obs in each sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of variables to conside	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest percentage of obs in r	1.0E-5
Smallest number of obs in node	1
Split Size	.
Score	
Variable Selection	Yes
Variable Importance Method	Loss Reduction
Number of Variables to Conside	25
Cutoff Fraction	0.01

Picture 4.13 - HP Forest default settings

In order to evaluate the most beneficial tuning of some of the most important parameters to consider (such as the number of variables and the leaf size), an experiment was carried out, based on the macros provided by (Wujek, 2015). The results thus obtained are depicted in pictures 4.14 and 4.15, alongside the base code (tables 4.13 and 4.14).

As it would be expected, the convergence towards the minimum misclassification rate varies with the type of sampling approach – balanced or unbalanced – being faster in the case of the later. During the successive runs, the number of variables to consider remains relatively stable at the interval 5-10, which is compatible with the default value (the square root of the number of inputs, which in the case at hand is around 7 – sqrt(42)). As far as the leaf size is concerned, the optimal minimum is within the range 1-5 in the case of the unbalanced dataset and around 3 in the case of the balanced one.



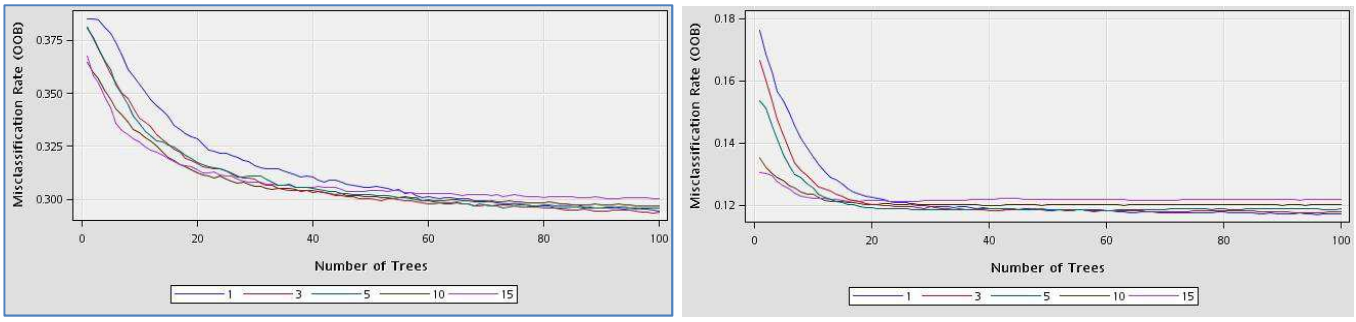
Picture 4.14 - Optimal number of variables (balanced dataset at the left; unbalanced dataset at the right)

```
%macro hpforestStudy (nVarsList=10,maxTrees=100);

 %let nTries = %sysfunc(countw(&nVarsList.));
  /* Loop over all specified number of variables to try */
 %do i = 1 %to &nTries.;
  %let thisTry = %sysfunc(scan(&nVarsList.,&i));

  /* Run HP Forest for this number of variables */
  proc hpforest data=&em_import_data maxtrees=&maxTrees. vars_to_try=&thisTry.;
    input %EM_INTERVAL_INPUT /level=interval;
    target %EM_TARGET / level=binary;
    ods output fitstatistics=fitstats_vars&thisTry. ;
  run;
  /* Add the value of varsToTry for these fit stats */
  data fitstats_vars&thisTry.;
    length varsToTry $ 8;
    set fitstats_vars&thisTry.;
    varsToTry = "&thisTry.";
  run;

  /* Append to the single cumulative fit statistics table */
  proc append base=fitStats data=fitstats_vars&thisTry.;
  run;
 %end;
```

```
%mend hpforestStudy;

%hpforestStudy(nVarsList=5 10 25 50 all,maxTrees=100);
/* Register the data set for use in the em_report reporting macro */
%em_register(type=Data,key=fitStats);
data &em_user_fitStats;
   set fitStats;
run;
%em_report(viewType=data,key=fitStats,autodisplay=y);
%em_report(viewType=lineplot,key=fitStats,x=nTrees,y=miscOOB,group=varsToTry,description=Out of Bag Misclassification Rate,autodisplay=y);
```

Source: Wujek, 2015

Table 4.13 - SAS macro for generating number of variables evaluation



Picture 4.15 - Minimum leaf size (balanced dataset at the left; unbalanced dataset at the right)

```
%macro hpforestStudy (leafsizeList=5,maxTrees=100);

 %let nTries = %sysfunc(countw(&leafsizeList.));
 /* Loop over specified leafsizelist to try */
 %do i = 1 %to &nTries.;
  %let thisTry = %sysfunc(scan(&leafsizeList.,&i));

   /* Run HP Forest for this leafsizelist */
   proc hpforest data=&em_import_data maxtrees=&maxTrees. leafsize=&thisTry.;
      input %EM_INTERVAL_INPUT /level=interval;
      target %EM_TARGET / level=binary;
      ods output fitstatistics=fitstats_vars&thisTry. ;
   run;

   /* Add the value of varsToTry for these fit stats */
   data fitstats_vars&thisTry.;
      length leafsize $ 8;
      set fitstats_vars&thisTry.;
      leafsize = "&thisTry.";
   run;

   /* Append to the single cumulative fit statistics table */
   proc append base=fitStats data=fitstats_vars&thisTry.;
   run;
 %end;
```

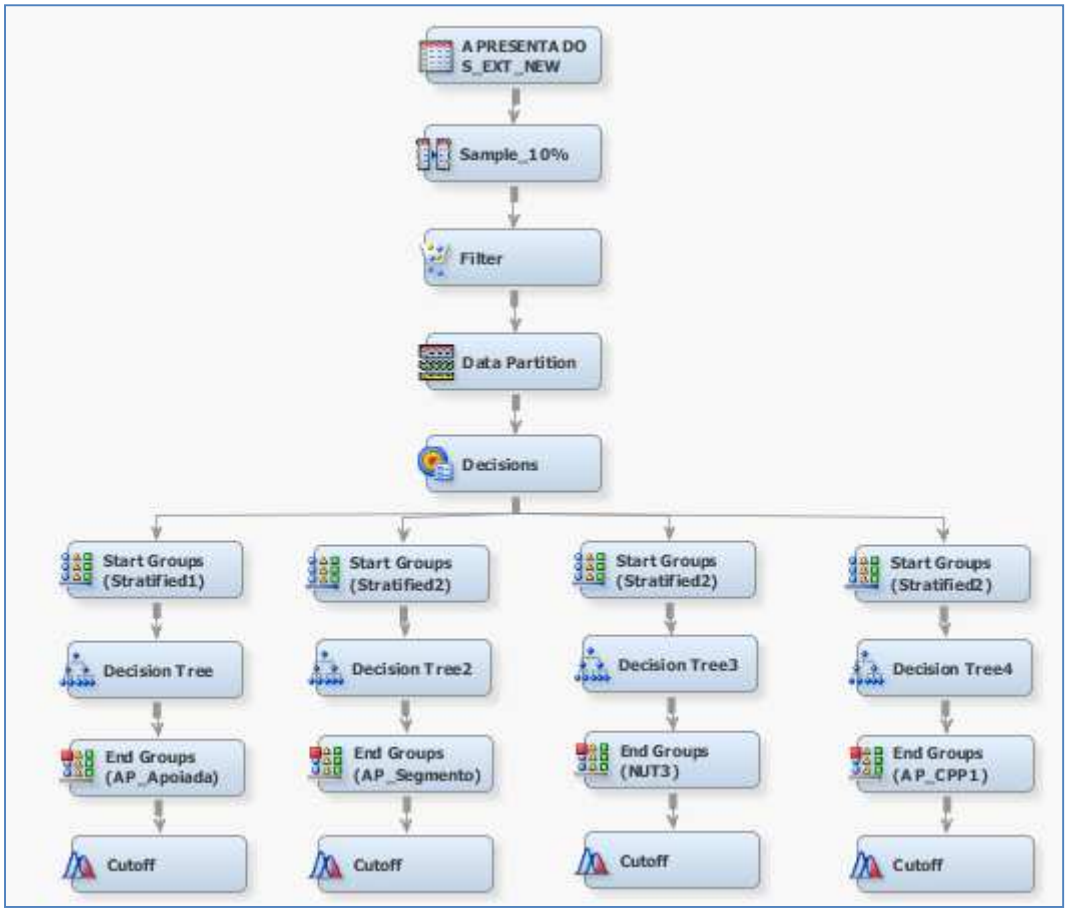
```
%mend hpforestStudy;

%hpforestStudy(leafsizeList=1 3 5 10 15,maxTrees=100);
/* Register the data set for use in the em_report reporting macro */
%em_register(type=Data,key=fitStats);
data &em_user_fitStats;
   set fitStats;
run;
%em_report(viewType=data,key=fitStats,autodisplay=y);
%em_report(viewType=lineplot,key=fitStats,x=nTrees,y=miscOOB,group=leafsize,description=Out of Bag
Misclassification Rate,autodisplay=y);
```

Source: Wujek, 2015

Table 4.14 - SAS macro for generating minimum leaf size evaluation

4.5. SEPARATE MODELS PERFORMANCE

One of the present thesis main objectives consists of evaluating the algorithms' performance against some of the features describing the job offers or the job applicants, such as the risk of long term unemployment, for instance. In addition to the variable importance analysis previously presented, this type of evaluation can be also achieved through the group processing facility available in SAS Enterprise Miner, according to a flow such as the one depicted below[3] (picture 4.16):



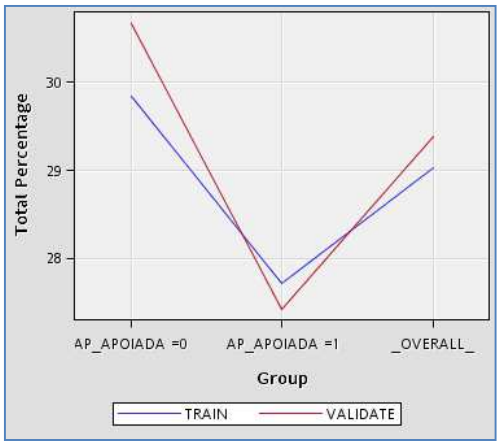
Picture 4.16 - Stratified models' flows

For the analysis at hand and in addition to the risk of long term unemployment, three other input variables were considered, in account of their importance within the various models: whether the job offer benefits from financial incentives (AP_Apoiada), level 3 of the nomenclature of territorial units for statistics associated with the matching register (NUTS III) and the job applicant's intended occupation, aggregated at level 1, for processing reasons (AP_CPP1).

[3] In order to obtain the desired stratification, one shall choose the mode "Stratify" in the properties dialog box of the Start Groups node and set the grouping role of the train variables to use for that effect as "Stratification".

Based on the results thus obtained (of an exploratory nature, however), it seems possible to reach the following conclusions:

- As it would be expected, the matching results of job offers that benefit from financial incentives present a significantly lower misclassification error (picture 4.17 and table 4.15), due to the inherent matching procedures, in which the employer has already pre-selected the most suitable candidate within the financial support application.

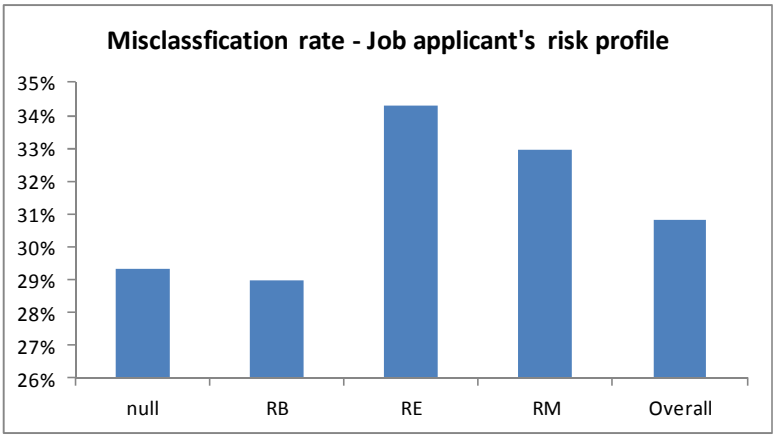


Picture 4.17 - Stratified model assessment chart - Job offers and financial incentives

Code	Explanation
0	The job offer doesn't benefit from financial incentives
1	The job offer benefits from financial incentives

Table 4.15 - Codification of the binary variable AP_Apoiada

- As far as the risk of long term unemployment is concerned, the placement of applicants presenting a moderate to high risk is harder to predict than that of those who are not being followed at this level (null category) or are classified as being of low risk (picture 4.18 and table 4.16).

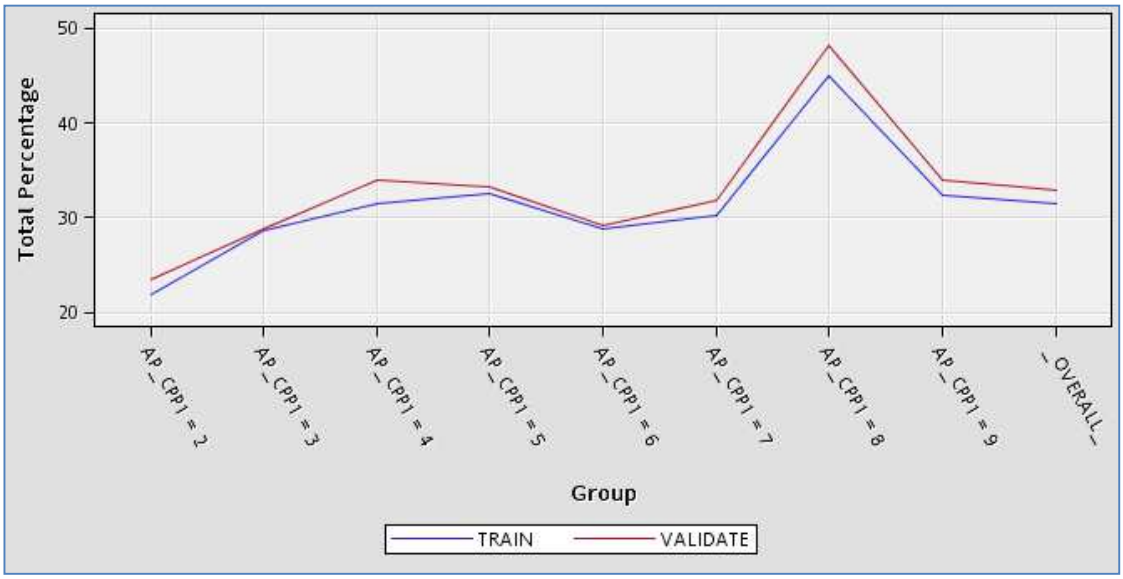


Picture 4.18 - Stratified Model Assessment - Risk of long term unemployment profile

Code	Explanation
null	The job applicant is not classified in one the possible segments
RB	The job applicant presents a low risk of becoming a long term unemployed person
RM	The job applicant presents a moderate risk of becoming a long term unemployed person
RE	The job applicant presents a high risk of becoming a long term unemployed person

Table 4.16 - Codification of the binary variable AP_Segmento

- In what regards the job applicant's intended occupation, the matching results which are the least harder to predict belong to group 2 (intellectual and scientific activities specialists) and the ones harder to predict belong to group 8 (plant and machine operators, and assemblers). The remaining groups present misclassification rates which are in line with the model's overall performance (picture 4.19 and table 4.17).

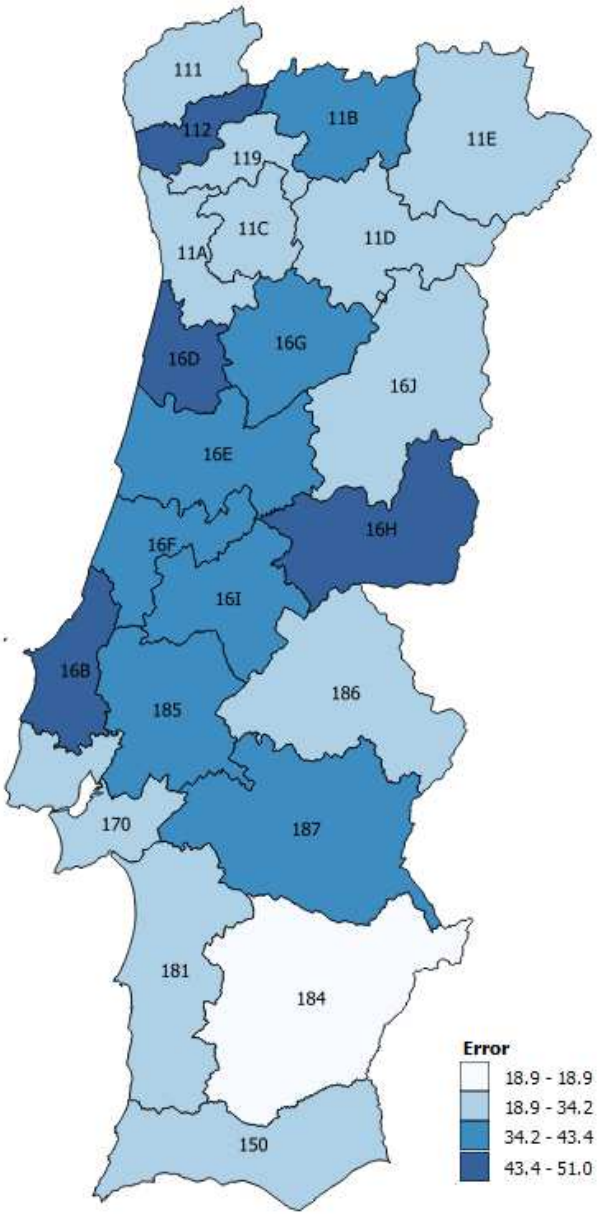


Picture 4.19 - Stratified Model Assessment Chart - Job offer's occupation (level 1)

Code	Designation
2	Intellectual and scientific activities specialists
3	Technicians and associate professionals
4	Clerical support workers
5	Personal service, protection and safety workers and salespersons
6	Farmers and skilled agricultural, fishery and forestry workers
7	Industry and construction skilled workers and craftsman
8	Plant and machine operators, and assemblers
9	Not skilled workers

Table 4.17 - Code and designation of occupations (level 1)

- In terms of the NUTS 3 region associated with the matching register, it isn't possible to identify a clear pattern, being possible to say, however, that the countries' two metropolitan areas (Porto and Lisboa) present misclassification errors which are close but lower than the one presented by the overall model (29% vs. 34.2%). The sub region of Baixo Alentejo presents the lowest misclassification rate, contrasting with the sub regions of Cávado (112), Aveiro (16D), Oeste (16B) and Beira Baixa (16H) which yield the highest rates (picure 4.20 and table 4.18).



Code	Designation
111	Alto Minho
112	Cávado
119	Ave
11A	Área Metropolitana do Porto
11B	Alto Tâmega
11C	Tâmega e Sousa
11D	Douro
11E	Terras de Trás-os-Montes
16B	Oeste
16D	Região de Aveiro
16E	Região de Coimbra
16F	Região de Leiria
16G	Viseu Dão Lafões
16H	Beira Baixa
16I	Médio Tejo
16J	Beiras e Serra da Estrela
170	Área Metropolitana de Lisboa
181	Alentejo Litoral
184	Baixo Alentejo
185	Lezíria do Tejo
186	Alto Alentejo
187	Alentejo Central
150	Algarve

Table 4.18 - NUTS 3 codification

Picture 4.20 - Stratified model misclassification rate chart by NUTS 3

4.6. SCORING OF NEW DATA

In order to assess the generalization capability of the chosen model, the performance of the HP Forest algorithm, trained on balanced and unbalanced datasets, was evaluated on a set of different samples, containing new, unseen data, alongside other tree based algorithms. For that effect and based on the article "Using Random Forest to Learn Imbalanced Data" (Chen, Liaw, & Breiman, 2004), the following comprehensive set of metrics was taken in consideration (picture 4.21):

$$\text{True Negative Rate } (Acc^{-}) = \frac{TN}{TN+FP}$$

$$\text{True Positive Rate } (Acc^{+}) = \frac{TP}{TP+FN}$$

$$G\text{-mean} = (Acc^{-} \times Acc^{+})^{1/2}$$

$$\text{Weighted Accuracy} = \beta Acc^{+} + (1-\beta) Acc^{-}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN} = Acc^{+}$$

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

	Predicted Positive Class	Predicted Negative Class
Actual Positive class	TP (True Positive)	FN (False Negative)
Actual Negative class	FP (False Positive)	TN (True Negative)

Source: Chen et al., 2004

Picture 4.21 - Performance measurement main metrics

These include two additional metrics in relation to the ones previously presented, namely:

- The Weighted Accuracy, which tries to measure the ability of the classifier to attain a high prediction accuracy over the minority class (Acc+), while maintaining reasonable accuracy in what concerns the majority class (Acc−). Weights can be adjusted for that effect, although equal proportions are normally considered (i.e., β equals 0.5).

- The Geometric Mean (G-mean), consisting of an alternative version of the weighted accuracy.

The Kolmogorov-Smirnov (KS) statistic, which measures the discriminatory power of a predictive model, was also taken in consideration. It is calculated by splitting the predicted probability into deciles and then computing the difference between the cumulative % of events and non-events in each part. The KS statistic corresponds to the maximum difference, as illustrated in table 4.19, below:

Decil	Prob.	Nr. Events	Nr. Non Events	Cum. % Events	Cum. % Non Events	Cum. % Dif.
1	15.01%	6	766	0.61%	11.36%	10.75%
2	18.84%	17	755	2.36%	22.56%	20.20%
3	23.05%	36	736	6.05%	33.47%	27.43%
4	26.51%	49	723	11.07%	44.19%	33.13%
5	30.15%	43	729	15.47%	55.01%	39.53%
6	36.09%	50	721	20.59%	65.70%	**45.10%**
7	44.31%	126	646	33.50%	75.28%	41.77%
8	54.51%	147	625	48.57%	84.55%	35.98%
9	66.45%	175	597	66.50%	93.40%	26.90%
10	91.42%	327	445	100.00%	100.00%	0.00%
	Totals	**976**	**6743**			

Table 4.19 - Calculation of the KS statistic (example)

The value of the KS statistic can, theoretically, fall between 0 and 100, with 0 corresponding to a model with no discriminatory power (i.e., to a model that selects cases randomly from the population). Acceptable values, however, will range from 20 to 70 (G. Wang, Peters, Skowron, & Yao, 2006).

The scoring of the random forests was based on the code and procedure presented in table 4.20, and, in the case of the remaining models, SAS score node was used.

```
LIBNAME SQL ODBC DSN='bd_tese';
RUN;
PROC SQL;
CREATE TABLE unbalanced_sample_HPF_ext_new AS SELECT * FROM SQL.unbalanced_sample_HPF_ext_new;
RUN;

proc hp4score
data=unbalanced_sample_HPF_ext_new;
id AP_COLOCADO;
score file=
"C:\Users\PAULA\Documents\Docs\Tese\Tese_matching_hp\Workspaces\EMWS7\HPDMForest\OUTMDLFILE.bin"
out=scored;
run;
```

Table 4.20 - Scoring of new data with the HP4Score SAS procedure

In order to better evaluate the models' performance in relation to the new data, the results obtained on validation datasets are previously presented in table 4.21:

Algorithm	*Recall (Acc+)*	*Acc-*	*Precision*	*F-Measure*	*G-Mean*	*Wt. Accuracy*	*KS*	*Nr. of obs.*	*% of events*
HP Forest trained on a balanced sample									
(cut-off=.5)	74.2	79.0	78.0	76.1	76.7	76.7	53.4	59,922	50.0
HP Forest trained on an unbalanced sample									
(cut-off=.5)	9.2	99.7	83.6	16.6	30.3	54.5	47.9	23,159	12.7
HP Forest trained on a smoted sample									
(cut-off=.5)	73.3	1	1	84.6	85.6	86.7	85.6	48,712	53.9
Gradient Boosting									
(cut-off=.5)	72.4	69.7	71.3	71.9	71.1	71.1	42.2	24,871	51.0
Decision Trees Ensemble									
(cut-off=.5)	75.1	74.2	75.6	75.3	74.7	74.7	49.9	25,427	51.5

Table 4.21 - Results on validation datasets

As for the new data, artificially balanced and unbalanced samples were considered for that effect, in order to evaluate possible biases or dataset shifts at this level. The results thus obtained are as follows (table 4.22):

Sampling	*Recall (Acc+)*	*Acc-*	*Precision*	*F-Measure*	*G-Mean*	*Wt. Accuracy*	*KS*	*Nr. of obs.*	*% of events*
HP Forest trained on a balanced dataset									
Balanced (cut-off=.5)	60.7	80.6	75.8	67.3	69.9	70.6	44.0	7,720	50.0
Unbalanced (cut-off=.5)	58.0	80.8	30.4	40.0	68.5	69.5	45.1	7,719	12.6
HP Forest trained on an unbalanced dataset									
Balanced (cut-off=.23)	20.8	97.8	90.6	33.8	45.1	59.3	42.1	7,720	50.0
Unbalanced (cut-off=.23)	19.9	97.8	56.2	29.3	44.8	58.8	41.6	7,719	12.6

Sampling	*Recall (Acc+)*	*Acc−*	*Precision*	*F-Measure*	*G-Mean*	*Wt. Accuracy*	*KS*	*Nr. of obs.*	*% of events*
HP Forest 3 trained on a balanced smoted dataset									
Balanced smoted (cut-off=.5)	86.7	95.7	96.0	91.1	91.1	91.2	81.8	18,041	53.9
Unbalanced smoted (cut-off=.5)	87.1	95.7	75.6	81.0	91.4	91.5	80.6	9,587	13.2
Unbalanced non-smoted (cut-off=.24)	59.1	79.5	30.6	40.3	68.6	69.3	43.8	4,483	13.3
Gradient Boosting (trained on a balanced dataset)									
Unbalanced (cut-off=.5)	70.0	71.6	25.9	37.8	70.8	70.8	41.9	7,750	12.5
Decision Trees Ensemble (trained on a balanced dataset)									
Unbalanced (cut-off=.5)	71.1	74.7	28.5	40.7	72.9	72.9	46.1	7,750	12.5

Table 4.22 - Results on new data

From the analysis of the new data scoring results based on the performance metrics presented above, it is possible to observe the following:

- In what concerns the precision metric, all the models perform better on balanced samples of new data, with the differences being rather large in most of the cases, except for the results obtained with the smoted data.
- There is also a significant difference between the performances of the different models with respect to the F-Measure.
- In the case of the remaining metrics, the results are relatively similar, except for the ones obtained with the non-smoted data.
- The weighted accuracy and KS statistics appear to be the most robust metrics, presenting values that may be considered reasonable in the context at hand.

4.7. DISCUSSION OF RESULTS

The variable importance analysis which was carried out in section 4.2, according to five different techniques, points to the existence of at least 26 significant features, namely the following ones, presented by decreasing order of importance based on the results provided by the random forest model:

- Whether job offer benefits from financial incentives
- Job offer's occupation
- Job offer's sector of activity at 2 digits level
- Job seeker's municipality of residence
- Level 3 of the nomenclature of territorial units for statistics
- Registration period as a job seeker with PES
- Proportion of purchasing power
- Job seeker's parish of residence
- Number of vacancies contained in the job offer
- Regional development composite index (overall index) by geographic localization (NUTS - 2013)
- Similarity between job offer's demanded conditions and job seeker's profile in what relates to type of work contract (fixed term...)
- Number of people working in the company responsible for the job offer
- Whether job seeker is receiving unemployment benefits
- Job applicant's previous occupation
- Births of enterprises
- Job seeker's intended occupation
- Job seeker's age
- Similarity between job offer's demanded conditions and job seeker's profile in what relates to occupation
- Number of common skills between job offer's occupation and job seeker's intended occupation
- Month of the job referral
- Similarity between job offer's demanded conditions and job seeker's profile in what relates to parish of work/residence
- Job seeker's long term unemployment risk profile
- Job seeker's qualifications
- Job seeker's experience at last job (nr. months)
- Job seeker's gender
- Number of people at care of job seeker

Variables such as the job offer’s wage, enterprises’ survival rate, the quarterly unemployment rate, job seeker's experience at intended occupation and management of the job offer by a dedicated PES counsellor are also considered important by four of the selection techniques.

These findings are in line with common sense and with the results that can be found in studies carried out in the field of the labour market adjustment, namely:

- The importance of the creation of business units as well as of public policies supporting the increase of employment (Escária, 2003);
- The role played by small establishments, educational attainment, skills, gender and type of contract in the creation and destruction of employment (Escária, 2003);
- The effect of unemployment benefits generosity in the decision to match or not (Centeno, 2004);
- The importance of environmental variables in the determination of the matching process efficiency of employment centres (Agovino, Massimiliano, Menezes, António Gomes, Sciulli, 2012).

On the other hand, the nationality of the job seeker, his or her situation in terms of social benefits and the management of the job seeker's registration by a dedicated PES counsellor are only deemed significant by three of the methods and the feature indicating whether the job seeker has a disability only by two of them. It is also possible to find some of the variables that are currently being used in the matching tool at the disposal of job centres (such as being a single parent or being part of an unemployed couple) within the cases that were not included in the final set of input variables. In spite of the exploratory nature of the current study, these results deserve some further attention, namely in what concerns the collection procedures, quality and completeness of the underlying data, as well as the candidate pre-selection methods that are currently being carried out at the job centres.

Based on the selected variables, the performance results of the chosen predictive models can be considered very reasonable, with the random forest, neural network and decision tree ensemble models presenting AUC values superior to 80% in almost all of the considered datasets, followed closely by the SVM and gradient boosting algorithms. The lowest performances were obtained within the unbalanced datasets, especially when the F-Measure is considered. These results thus seem to confirm the main conclusions found in the research literature, with the only caveat being the apparently very close proximity between the best performing models regardless of the inclusion of external data, an aspect that should be further investigated.

As far as the generalization capability of the models at hand is concerned and based on the results obtained with samples of new data, it is possible to observe that:

- The precision and F-Measure metrics present significantly lower values than the remaining ones when the sample is unbalanced and the algorithm was trained on balanced data;
- The recall and F-Measure metrics present significantly lower values than the remaining ones when the algorithm was trained on unbalanced data (regardless of the sample balancing method).

Theoretically, the matching process would make it difficult to differentiate between true positives and false negatives, if one takes in consideration that more than one candidate is referred to each job vacancy and that the same candidate may be referred to two or more similar job offers in the same time period, only accepting one of them. However and based on the results obtained with the models trained on balanced data, the recall measure (TP/(TP + FN)) doesn't seem to point in that way. The KS statistic, presenting an average value of 43.5% (not considering the results obtained with the smoted data), also seems to sustain a reasonable discrimination capacity between events and non-events.

In general, when the rare class if of a greater interest (in this case, the candidates that are effectively placed on the referred job offer), a good classifier will be one that gives a high prediction accuracy in relation to the minority class, without sacrificing a reasonable accuracy for the majority class (Chen et al., 2004). This also appears to be sustained by the study at hand, since the related weighted accuracy metric presents an average value of 71%, based on the results provided by the models trained on balanced (non-smoted) data, regardless of the sampling method.

Lastly, it should also be noted that, according to some authors (Powers, 2015), the F-Measure should be used with some caution since "it focuses on one class only; is biased to the majority class; as a probability assumes the Real and Prediction distributions are identical; doesn't in general take into account the True Negatives; gives different optima from other approaches and tradeoffs". In order to deal with these shortcomings, an adjusted F-Measure (AGF) is proposed by Maratea, Petrosino, & Manzo (2014), as illustrated below (picture 4.22):

$$AGF = \sqrt{F_2 \cdot InvF_{0.5}}$$

Source: Maratea et al., 2014

Picture 4.22 - Adjusted F-Measure (AGF)

where:

- F2 consists of the traditional F-Measure with the *B* parameter equal to 2 (weighting recall more than precision and strengthening the false negative values);
- InvF0.5 is calculated with the help of a new confusion matrix in which positive samples become negative and vice versa and the *B* parameter is set to 0.5.

In this way, all elements of the original confusion matrix are accounted for and more weight is given to patterns correctly classified in the minority class (the positive class). After computing AGF for the two of the models which were evaluated within the new data samples, namely Gradient Boosting and Decisions Trees Ensemble, the initial F-Measure values rise from 37.8 and 40.7 to 67.2 and 69.7, respectively, thus resulting in significant differences.

5. CONCLUSIONS

The current thesis was mainly focused on evaluating the extent to which the job matching services provided by the Portuguese PES could be improved through the application of machine learning algorithms and, if so, on determining the best approach to a possible automation of the recruitment process. For that effect, the following objectives have been established and fully achieved:

- Extensive review of relevant literature and case studies in order to find the novelist and most adequate and feasible machine learning algorithms (Chapter 2);
- Study and application of the most relevant and recent algorithms available in the software package SAS Enterprise Miner (Chapters 3 and 4);
- Identification of the most relevant variables to be used as inputs (Section 4.2);
- Evaluation of the algorithm's performance, based on how effective it is in assigning consistent relevance scores to the candidates, compared to the ones assigned by human recruiters (Sections 4.3, 4.4 and 4.6);
- Evaluation of the algorithm's performance against the applicant's risk of long term unemployment, among other relevant input variables (Section 4.5);
- Evaluation of the feasibility and importance of incorporating external data, by measuring the algorithm's effectiveness with and without those elements (Section 4.3).

After a thorough performance evaluation of robust algorithms such as Random Forests, Gradient Boosting, Support Vector Machines, Neural Networks Ensembles and other tree-based ensemble models, the obtained results (as discussed in Section 4.7) seem to point to the possible improvement of the current pre-selection tools at the disposal of the job centres, bringing potential efficiency and quality gains to the matching procedures of the Portuguese PES.

It can thus be stated that the present thesis contributes not only to the extension and update of existing knowledge, by analysing the application of machine learning algorithms within the specific and current context of PES, based on a large and heterogeneous pool of applicants and job offers, but also to the provision of a practical tool for the improvement of the Portuguese PES' technical matching system, through the incorporation of relevant external input variables (such as the ESCO database) and the scoring code of the best performing predictive models in the underlying information systems.

The lack of transparency of the evaluated algorithms can be pointed out, however, as a limitation that may hinder their practical implementation and that needs to be balanced with the robustness and remaining advantages of these types of models. On the other hand, the results obtained with unseen data and the implications of the chosen sampling methods are in need of a deeper analysis.

Lastly, suggestions for further studies may include the adoption of alternative target variables (such as a nominal variable including three levels: placed, refused by job applicant, refused by employer); the extraction and use of information concerning hard and soft skills required by the job offer and or detained by the job applicant, alongside social media data; the analysis of the risk of leaving the organization based on new registrations with the job centres and or social security information; the implementation of more advanced ensemble models such as the ones recently presented by SAS Institute (Czika, Maldonado, & Liu, 2016).

6. BIBLIOGRAPHY

Agovino, Massimiliano, Menezes, António Gomes, Sciulli, D. (2012). *The Efficiency of Matching in Portuguese Public Employment Service* (Vol. 7595).

Akinyede, R. O., & Daramola, O. A. (2013). Neural Network Web-Based Human Resource Management System Model (NNWBHRMSM), *1*(3), 75–87.

Bollinger, J., Street, B., & Francisco, S. (2012). Using Social Data for Resume Job Matching. *Proceedings of the 2012 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media*, 27–30. http://doi.org/10.1145/2390131.2390143

Bort, J. (2014). LinkedIn Buys Bright. *Business Insider*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees.

Brown, I. (2012). An experimental comparison of classification techniques for imbalanced credit scoring data sets using SAS ® Enterprise Miner ™. *Data Mining and Text Analytics*.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. http://doi.org/10.1016/j.eswa.2011.09.033

Capgemini, IDC, Sogeti, IS-practice and Indigov, R. E. and the D. T. I. (2012). *Public Services Online : Assessing User Centric eGovernment performance in Europe – eGovernment Benchmark 2012*. *Digital Agenda for Europe*.

Cathie, A., Chakraborty, G., & Garla, S. (2013). SAS Global Forum 2013 Applications Development Extension Node to the Rescue of the Curse of Dimensionality via Weight of Evidence (WOE) Recoding Satish Garla , SAS Institute Inc ., Cary , NC Goutam Chakraborty , Oklahoma State University , Stillwater , , 1–9.

Centeno, M. (2004). The match quality gains from unemployment insurance. *Journal of Human Resources*, *39*(3), 839–863. http://doi.org/10.3368/jhr.XXXIX.3.839

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*, 853–867. http://doi.org/10.1007/0-387-25465-X_40

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, (1999), 1–12. http://doi.org/ley.edu/sites/default/files/tech-reports/666.pdf

Czika, W., Maldonado, M., & Liu, Y. (2016). Ensemble Modeling: Recent Advances and Applications, 1–20.

Damodaran, R., Kumar, G., Raj, K., Jagan, M., & State, O. (2016). Predicting Rare Events Using Specialized Sampling Techniques in SAS ® OVER-SAMPLING TECHNIQUE, 1–7.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30. http://doi.org/10.1016/j.jecp.2010.03.005

Drigas, a., Kouremenos, S., Vrettos, S., Vrettaros, J., & Kouremenos, D. (2004). An expert system for job matching of the unemployed. *Expert Systems with Applications*, *26*(2), 217–224. http://doi.org/10.1016/S0957-4174(03)00136-2

Escária, V. (2003). *Analysis of Labour Market Adjustment Using Matched Employer-Employee Data for Portugal*. University of York.

Faliagka, E., Iliadis, L., Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A., & Tzimas, G. (2013). On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV. *Artificial Intelligence Review*, 1–14. http://doi.org/10.1007/s10462-013-9414-y

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Amorim Fernández-Delgado, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Freeman, E. a., Moisen, G. G., Coulston, J. W., & Wilson, B. T. (2016). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance [1]. *Canadian Journal of Forest Research*, *46*(3), 323–339. http://doi.org/10.1139/cjfr-2014-0562

GHK Consulting Ltd. Case study : A " Virtual Labour Market Platform " for the Public Employment Service in Germany (2011).

Larsen, C. A., & Vesan, P. (2012). Why public employment services always fail. Double-sided asymmetric information and the placement of low-skill workers in six european countries. *Public Administration*, *90*(2), 466–479. http://doi.org/10.1111/j.1467-9299.2011.02000.x

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. http://doi.org/10.1016/j.ejor.2015.05.030

Liaw, a, & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22. http://doi.org/10.1177/154405910408300516

López, V., Fernández, A., & Herrera, F. (2014). On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, *257*, 1–13. http://doi.org/10.1016/j.ins.2013.09.038

Maldonado, M., Dean, J., Czika, W., & Haller, S. (2014). Leveraging Ensemble Models in SAS ® Enterprise Miner ™, 1–15. Retrieved from http://support.sas.com/resources/papers/proceedings11/160-2011.pdf

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, *257*, 331–341. http://doi.org/10.1016/j.ins.2013.04.016

Mehta, S., Pimplikar, R., Singh, A., Varshney, L. R., & Visweswariah, K. (2013). Efficient multifaceted screening of job applicants. *EDBT'13 - Proceedings of the 16th International Conference on Extending Database Technology*, 661–671. http://doi.org/10.1145/2452376.2452453

Milley, A. H., Seabolt, J. D., & Williams, J. S. (1998). Data Mining and the Case for Sampling. A SAS Institute Best Practices, 1–36. Retrieved from http://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf

Package, T., Functions, T., & Torgo, A. L. (2015). *Package " DMwR ."*

Panneerselvam, N. D. (2015). The More Trees , the Better ! Scaling Up Performance Using Random Forest in SAS ® Enterprise Miner ™. *Sgf2015*, 1–10.

Powers, D. M. W. (2015). What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes, 19. http://doi.org/KIT-14-001

Schubert, S. (2010). The Power of the Group Processing Facility in SAS Enterprise Miner. *Transformation*, 1–13. Retrieved from http://support.sas.com/resources/papers/proceedings10/123-2010.pdf

Strohmeier, S., & Piazza, F. (2013). Domain driven data mining in human resource management: A

review of current research. *Expert Systems with Applications*, *40*(7), 2410–2420. http://doi.org/10.1016/j.eswa.2012.10.059

Tzimas, E. F. A. T. G. (2012). An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research*, *22*(5), 551–568. http://doi.org/10.1108/10662241211271545

Wang, G., Peters, J. F., Skowron, A., & Yao, Y. (2006). Rough Sets and Knowledge Technology.

Wang, R., Lee, N., & Wei, Y. (2015). A Case Study : Improve Classification of Rare Events with SAS ® Enterprise. *SAS Paper*, 1–12.

WCC. Case Study Bundesagentur für Arbeit One click away from a new job.

Wielenga, D. (2007). Identifying and Overcoming Common Data Mining Mistakes. *Forum American Bar Association*, 1–20.

Wujek, B. (2015). Tip_ Getting the Most from your Random Forest - SAS Support Communities. Retrieved from https://communities.sas.com/t5/SAS-Communities-Library/Tip-Getting-the-Most-from-your-Random-Forest/ta-p/223949

Zdravevski, E., Lameski, P., & Kulakov, A. (2013). Advanced Transformations for Nominal and Categorical Data into Numeric Data in Supervised Learning Problems. *The 10th Conference for Informatics and Information Technology*, *7*(Ciit), 142–146. Retrieved from http://ciit.finki.ukim.mk/data/papers/10CiiT/10CiiT-31.pdf

7. APPENDIX TO CHAPTER 4 (SAS OUTPUTS)

Variable Levels Summary

Variable	Role	Frequency Count
AP_COLOCADO	TARGET	2
AP_ID_APRESENTACAO	ID	771728

Class Variable Summary Statistics

Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
AP_APOIADA	INPUT	2	0	0	70.63	1	29.37
AP_CATEGORIA	INPUT	4	0	2	77.2	5	9.46
AP_CC	INPUT	208	0	1111	5.17	1317	3.69
AP_CPP2	INPUT	43	0	52	9.39	41	8.13
AP_CPP_ANTERIOR	INPUT	513	2259	9.39	91120	5.18	
AP_CPP_OFERTA	INPUT	513	0	91120	6.37	51310	5.72
AP_CPP_PRETENDIDA	INPUT	513	0	41100	7.27	91120	5.42
AP_DEF	INPUT	2	0	0	98.28	1	1.72
AP_DISTRITO	INPUT	15	0	11	20.15	13	19.13
AP_FREGUESIA	INPUT	513	0	131315	2.78	131727	2.59
AP_GC	INPUT	2	0	0	97.98	1	2.02
AP_GOE	INPUT	2	0	0	62.54	1	37.46
AP_HabRank	INPUT	7	0	7	34.95	5	25.5
AP_MES	INPUT	12	0	10	10.24	11	9.22
AP_NAC_PT	INPUT	2	0	1	93.51	0	6.49
AP_NUT3	INPUT	17	0	170	24.23	11A	21.43
AP_QUALIFICACAO	INPUT	9	11946	2	27.3	3	24.72
AP_RSI1	INPUT	2	0	0	92.59	1	7.41
AP_SEGMENTO	INPUT	4	143842	RB	38.31	RM	32.31
AP_SEXO1	INPUT	2	0	0	51.57	1	48.43
AP_SUBSIDIADO	INPUT	2	0	0	68.56	1	31.44
ofa_CAE2	INPUT	84	0	56	13.18	47	9.2
AP_COLOCADO	TARGET	2	0	0	87.35	1	12.65

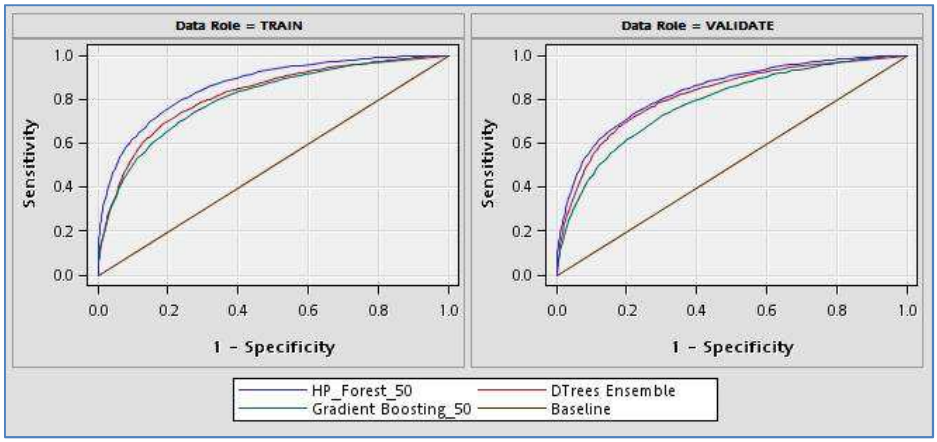
Distribution of Class Target and Segment Variables

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	AP_COLOCADO	TARGET	0	674300	87.3526
TRAIN	AP_COLOCADO	TARGET	1	97629	12.6474

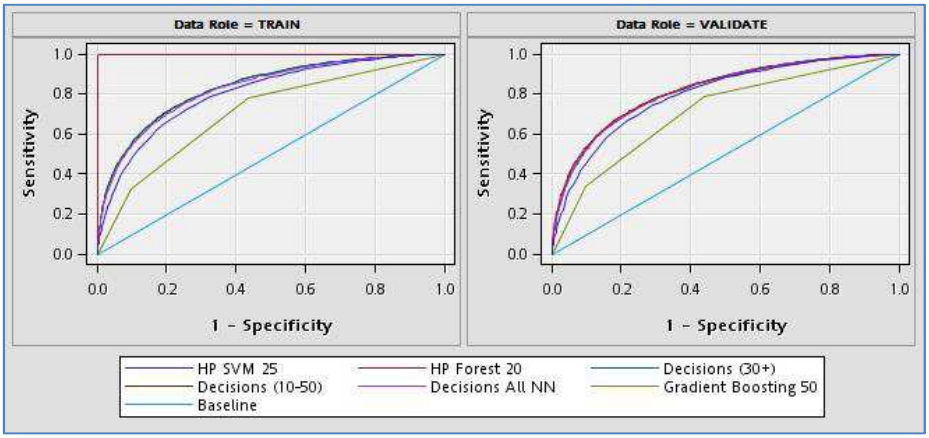
Interval Input Variable Summary Statistics

Variable Name	Mean	Standard Deviation	Non Missing	Missing	Median	Skewness	Kurtosis
AP_DESCENDENTES_A_CARGO	0.67	0.86	768229	3700	0	1.00	-0.06
AP_IDADE	36.38	10.60	771929	0	36	0.26	-0.82
AP_INT_TEMPO_INSCRICAO	11.79	12.75	771929	0	7	1.38	1.28
AP_TEMPO_PRATICA3	67.74	79.15	767646	4283	36	1.38	1.12
AP_TEMPO_PRATICA_UCNP3	65.87	77.71	761003	10926	36	1.46	1.37
Conta_Skills	18.25	25.64	771929	0	5	1.71	2.41
ISDR	100.47	4.44	771929	0	100.74	0.13	-1.19
Nascimentos_nr	633.53	956.57	752772	19157	205	1.95	3.07
PPC	1.12	1.01	771929	0	0.744	1.13	0.34
Taxa_sobrev_2antes	62.53	12.94	752551	19378	60.87	0.87	1.67
TxDesemp	12.76	2.07	771929	0	13.85	-0.45	-0.95
mt_SimExp	0.86	0.34	771929	0	1	-2.12	2.48
mt_SimFreg	0.54	0.28	771929	0	0.67	-0.06	-0.59
mt_SimHab	0.99	0.05	771929	0	1	-4.32	16.66
mt_SimProf	0.40	0.45	771929	0	0.2	0.47	-1.63
mt_SimTipocontrato	0.23	0.42	771929	0	0	1.28	-0.35
mt_SimUltProf	0.62	0.47	771929	0	1	-0.48	-1.70
ofa_NR_PESSOAS_SERVICO	80.38	338.13	752782	19147	7	7.62	65.96
ofa_NR_POSTOS_TRAB	1.89	2.69	771929	0	1	4.61	25.20
ofa_SALARIO	563.63	160.94	771929	0	505	4.02	49.71

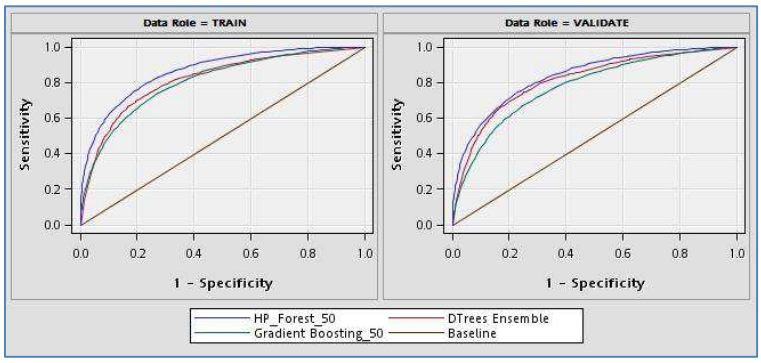
Table 7.1 - Filtered raw dataset summary statistics



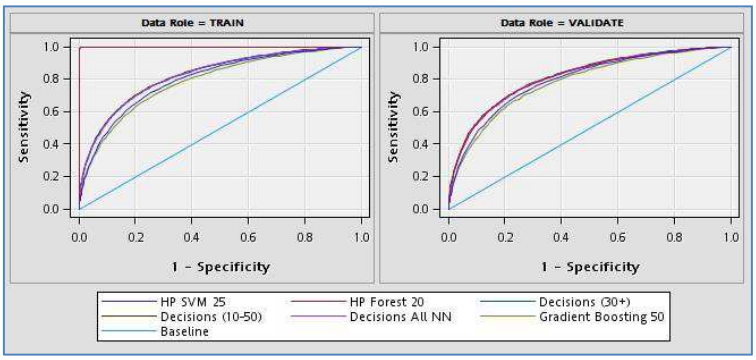
Picture 7.1 - Complete_mixed balanced dataset ROC Curves



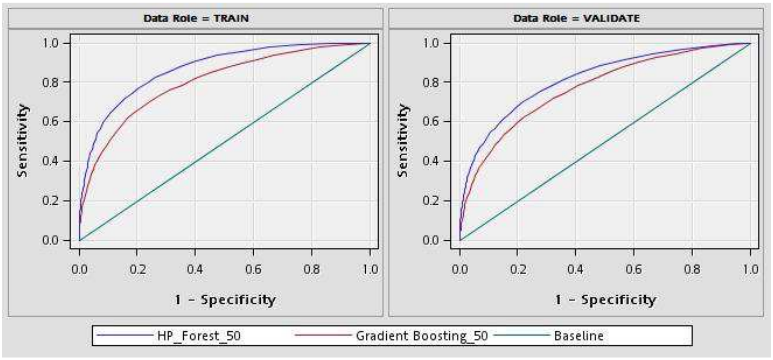
Picture 7.2 - Complete_num balanced dataset ROC Curves



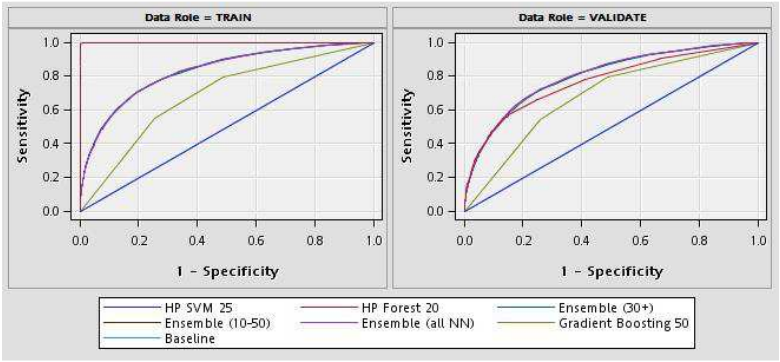
Picture 7.3 - Internal_mixed balanced dataset ROC Curves



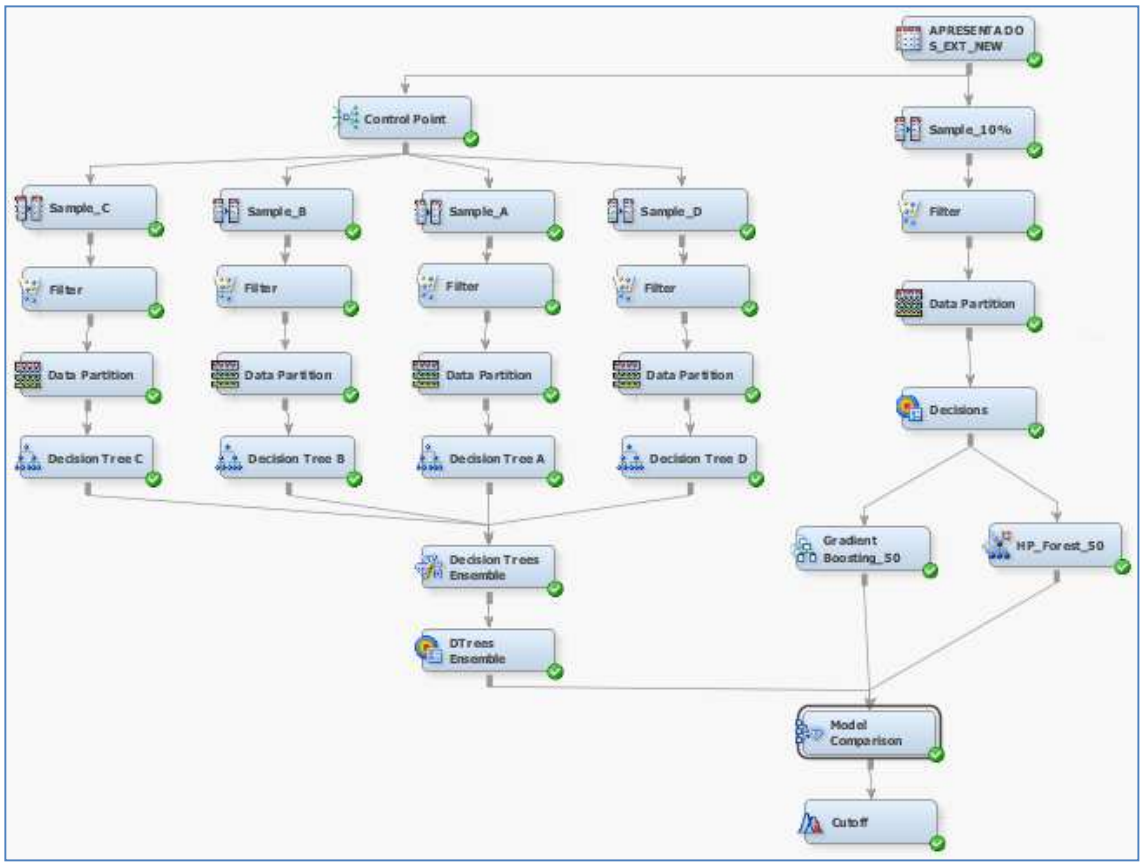
Picture 7.4 - Internal_num balanced dataset ROC Curves



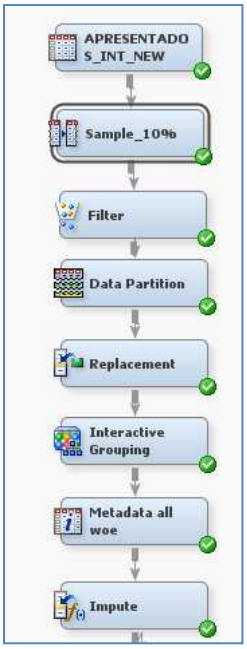
Picture 7.5 - Complete_mixed unbalanced dataset ROC Curves



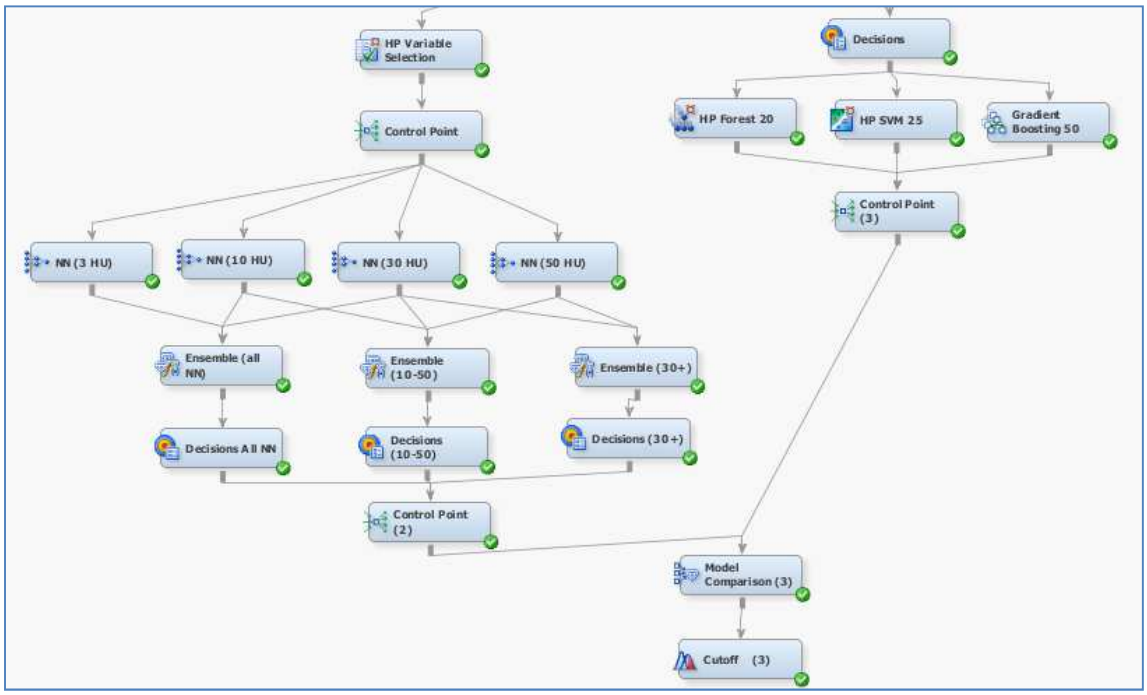
Picture 7.6 - Complete_num unbalanced dataset ROC Curves



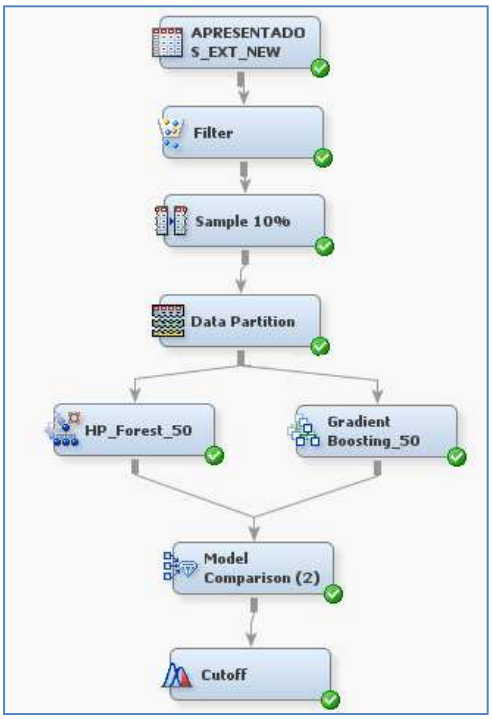
Picture 7.7 - HP Forest, Gradient Boosting and Decision Trees Ensemble (mixed balanced dataset)



Picture 7.8 - Numeric dataset supporting nodes



Picture 7.9 - Numeric balanced dataset models' flow



Picture 7.10 - HP Forest and Gradient Boosting flow on the unbalanced dataset