A Work Project, presented as part of the requirements for the Award of a Master Degree in Management from the NOVA – School of Business and Economics.

Sales as a Science: Predictive Model for Upselling Opportunities in the PaaS Industry

SOFIA LEMOS ROCHA, 25827

A Project carried out on the Master in Management Program, under the supervision of:

Qiwei Han
Bruno Silva

January 3rd 2020

Abstract

Nowadays, companies have the ability to generate and continuously update a high volume of data about their customers and at a much faster rate than ever before. At the same time, companies are incorporating Advanced Analytics to develop a successful Customer Experience strategy. Leveraging on Machine Learning solutions in the sales B2B PaaS environment presents an opportunity for companies to create significant business value.

This paper explores a predictive model to gain insights into customers' future behavior as a crucial step towards customer-centricity as well as to uncover upselling opportunities, based on a data-driven approach rather than intuition.

Keywords: Machine Learning, OutSystems, Platform as a Service, Predictive Analytics

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

1. Introduction

a. Context

In today's technology-driven times, data is one of the most valuable assets of a firm. Data is now at the center of major technological advances and business disruptions as companies are progressively producing and consuming more data and at a faster speed than ever before.

Digital transformations have sparked off the importance in gathering data and, more importantly, generating valuable insight from it in order to achieve competitive advantage (Opresnik *et al.*, 2015).

Data Science and other intertwined concepts such as big data and data-driven decision making play a big role in prompting companies to make better and faster-informed decisions, with greater precision and impact. Prostov *et al.* (2013) note that success in today's data-oriented business environment requires being able to think about how these fundamental concepts apply to particular business problems—to "think data-analytically".

Research by Mckinsey & Company (2014) has reported on the impact of customer analytics in corporate performance, showing that "data-centered organizations are twenty-three times more likely to acquire customers, six times as likely to retain those customers, and nineteen times as likely to be profitable as a result." According to Philips-Wren *et al.* (2015), "companies are realizing the potential value of data to gain insight into their customers and the power of analytics to guide decision-making". Leveraging on machine learning and predictive analytics solutions presents an opportunity for companies to create significant business value. Particularly, customer-centric companies that are able to harness these capabilities can unlock the potential to improve the sales cycle in the Business-to-Business (B2B) environment.

Predictive analytics in the sales environment is a powerful tool to uncover opportunities to effectively enhance performance across the entire customer lifecycle - to help predict customer behavior to firm offerings, to successfully promote upselling and cross-selling opportunities and to improve customer satisfaction and retention, foresting a long-term relationship. Adam (2018) suggests that another area of value is applying machine learning to gain customer and opportunity insights. Salespeople need to be fully informed about customers, markets, segments, opportunities, and competitors.

A data-driven approach allows companies to make decisions based on a full range of information and to better and more efficiently use their time and resources rather than relying on personal knowledge and gut feeling about customers.

Hale (2018) summarized why companies should apply predictive analytics: to "go beyond learning what happened and why to discovering insights about the future" in order to "better serve customers today."

b. Cloud Computing and Application Platform as a Service (aPaaS)

Cloud computing is growing interest among organizations around the globe.

As outlined by Schubert *et al.*'s (2010) definition and other work (Zhang *et al.,* 2010) "cloud computing does not refer to a specific technology but rather a concept comprising a set of combined technologies, forming a "new operations model that brings together a set of existing technologies to run business in a different way". It represents a fundamental change in the way "Information Technology services are invented, developed, deployed, scaled, updated, maintained and paid for" (Marston *et al*., 2011).

The three most well-known cloud computing service models are Infrastructure as a Service (IaaS), Software as a service (SaaS) and platform as a service (PaaS).

PaaS users only need to login and start using the platform, usually through a Web browser interface, to develop and manage applications. With cloud service revenue on the rise, the PaaS market has also experienced growth. According to research and advisory firm Gartner (2019), the total market revenue is forecasted to reach $20 billion in 2019 and is expected to exceed $34 billion by 2022.

b. Low-Code Application Platforms (LCAP)

Particularly in recent years since cloud computing entered the public consciousness, independent analyst firms and industry experts have composed other terms to define a specific segment. Back in 2014, analyst Forrester (2014) coined the term low-code to describe a category of new app-delivery thinking, referring to "application platforms that significantly decrease the amount of hand-coding required thus accelerating application delivery", one of the most notable benefits. In that same year, Gartner (2014) defined a subcategory of PaaS with the concept of aPaaS, which stands for application Platform as a Service denoting a "cloud service with the ability to develop, deploy, and execute applications as a service".

Advantages of aPaaS include high scalability, speed-of-delivery and the opportunity to build applications even for those with less development experience (Alexander, 2019). According to Morgan (2019), aPaaS and PaaS are conceptually similar and tightly linked, although the first targeting citizen developers while the latter typically cater professional developers and "both PaaS and aPaaS have been logical steps in companies' cloud migration strategies and they'll become even more relevant as organizations digitally transform into more software-dependent and software-driven organizations." By 2017, Gartner (2017) segmented further its aPaaS category adding high-productivity to application platform as a service (hpaPaaS).

In 2019, Gartner (2019) published its Magic Quadrant for Low-Code Application Platforms creating yet another category, Low-Code Application Platform (LCAP) refers to an "application platform that supports rapid application development, one-step deployment, execution and management using declarative, high-level programming abstractions, such as model-driven and metadata-based programming languages".

As business demands become more complex and application delivery timelines shrink, enterprises are looking for better ways to develop software applications, turning into LCAP at a remarkable rate as a result. According to the latest report by Gartner, it is expected that LCAP will represent "more than 65% of application development activity" by the year 2024.

d. OutSystems Company Overview

Founded in Portugal in 2001, OutSystems is a low-code application development platform that entirely runs off cloud infrastructure and that enables the rapid, agile and continuous development, delivery, and management of applications. OutSystems offers compelling productivity gains by enabling its customers to develop applications at once and proceed to their deployment on any device or platform at a lower cost than traditional technology and at a faster rate that meets the speed the market demands ("OutSystems Evaluation Guide.", n.d.). Analyst firms have named OutSystems as the leader of the low-code, rapid-application delivery market. Just recently, Gartner's 2019 Magic Quadrant for Enterprise Low-Code Application Platforms (LCAP) distinguished OutSystems as leader.

e. Research Question

For the scope of this thesis, the main research question is stated as:

Research Question: *Who are the customers OutSystems should target in order to increase their Annual Recurring Revenue (ARR) in the next six months?*

To thoroughly answer the main research question, three sub-questions are also outlined:

Sub-question 1: *How can predictive analytics be used to classify customers' behavior in the next 6 months and improve the customer experience in a PaaS B2B company?*

Sub-question 2: *What factors may help explain a customer's upcoming behavior?*

Sub-question 3: *What machine learning model is the best suited to gain expansion insights in the compound growth engine phase of the sales cycle?*

The steps required to provide answers to the main research question as well as the sub-questions are:

(1) View the business problem from a data perspective;

(2) Construct a dataset with information regarding the customers and additional information;

(3) Conduct an exploratory data analysis to provide insight into the problem;

(4) Construct a predictive model using Data Mining and Machine Learning techniques to predict the probability of a customer expanding his current plan within the next six months;

(5) Apply the model developed to new data to make predictions and see which fits better to the problem;

(6) Identify the variables that contribute the most towards a client's expansion in the next six months using the SHAP algorithm.

f. Report Organization

The remainder of this report first discusses findings from the literature in Chapter 2, regarding the B2B market, the PaaS bowtie sales methodology and the utilized tools and machine learning algorithms. Then the methods, consisting of the research framework, the evaluation metrics, the oversampling techniques and time series cross validation method are described in Chapter 3. The methodology chapter is followed by the Data Chapter 4 focused on the description of each step of the methodology and relating it to the theory outlined in Chapter 2. The results and overall model performance comparison are presented and discussed afterward in Chapter 5. The thesis ends in Chapter 6 with a conclusion, key findings and description of implications for OutSystems, limitations and suggestions for future research.

2. Literature Review

a. B2B Customer Experience and Customer Journey

Adam (2018) outlined some of the key characteristics of the B2B environment which include: (1) longer sales cycle when compared to B2C; (2) fewer customer relationships but complex ones across multiple stakeholders; (3) fewer transactions but of large value; (4) reduced number of customers, leads, prospects, opportunities, and sales; (5) focus on solution-based selling; (6) extreme importance of reliability (Cáceres *et al.,* 2007). Given the magnified complexity inherent in the B2B environment, it has become important to understand the customer experience as a driver of business success and competitive advantage (Hollyoake, 2009; Lemon *et al.,* 2016).

Customer experience is complex, dynamic and difficult to capture, encompassing customer responses to all the interactions they have with a firm (Homburg *et al.,* 2015). Capturing customer experience in a B2B context is even further hampered since it incorporates the understanding of experiences that emanate from direct and indirect interactions between providers, clients, and end-users (Zolkiewski, 2017). Thus, "outcomes in customer experience are not simply individual perceptions but rather produce, and are a product of interactions, described as touchpoints" (Homburg *et al.,* 2015; Pucinelli *et al.,* 2009).

As such, a major consideration when studying customer experience is tracking it at touchpoints to gain an understanding of the customer journey.

Traditionally, the purchase funnel has been adopted to consider the multiple stages a customer goes through: (1) prepurchase, which covers all aspects of a customer's interactions before a purchase transaction; (2) purchase, encompassing all interactions during the purchase event; and (3) post-purchase, that includes aspects of a customer's experience after purchase, such as usage and consumption (Lemon *et al.,* 2016). (Annex I). Further research (Court *et al.,* 2009) has broadened this process to include the "loyalty loop" (Annex II) as part of the overall customer journey, suggesting that during the post-purchase stage, a "trigger may occur that either leads to customer loyalty through future engagement or repurchase". Customer loyalty has been proven to be a major source of competitive advantage for companies not only important in the B2C context, but in the B2B context as well (Lam *et al.,* 2004) as it ultimately leads to corporate profitability

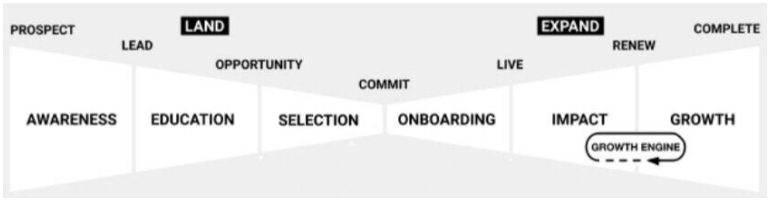
(Chalmeta, 2006). Viveiros (2016) notes that B2B companies focus the majority of their marketing efforts on the acquiring of new customers. However, "once those customers are acquired, the creation of loyalty falls short" (Bardauskaitė, 2014). Cahill *et al.* (2010) conceptualized customer loyalty as "the intention to purchase the same services (retention) and additional services (expansion) from the current provider in the future, as well as recommending a provider to others (referral)". Kumar *et al.* (2018) defined strategic CRM (Customer Relationship Management) as "the strategic process of selecting customers that a firm can most profitably serve and shaping interactions between a company and these customers" with the ultimate objective of optimizing "the current and future value of customers for the company."

b. B2B PaaS Bowtie Sales Methodology

Understanding of B2B relationships and the experiences that emanate from them can also be passed to the sales cycle sphere for companies that sell B2B applications and platforms.

VanderKooij (2019) suggested reconceptualizing the B2B sales methodology motivated by the fact that the traditional B2B sales funnel and its qualification methodologies do not address the needs of a recurring business model. Depicted with a bow tie (Figure I), the new proposed model has additional stages - to achieve recurring "Impact" and "Growth" of impact - which create a loop resulting in a compound "Growth Engine" crucial to the growth of recurring revenue businesses. Continuous impact for customers results in recurring revenue for businesses.

Figure I - Bow Tie Sales Methodology. (Retrieved from VanderKooij, 2019)



A "customer-centric focus is an important facilitator within firms to create stronger customer experiences" (Ramani *et al.,* 2008) which should extend beyond the point of the purchase.

Delivering effective customer experiences during the purchase cycle across multiple touchpoints thus bringing impact for customers occurs, for instance, when a customer has an opportunity that can positively be impacted by the seller, when the seller can help a customer achieve the impact in the timeframe needed or

when there is a growth potential beyond the original impact (VanderKooij, 2019). Furthering this path, Atkins *et al.* (2018) point out that companies are expanding their efforts to include customer success as a growth engine.

c. The Role of Technology in Enhancing the B2B Customer Experience

Technology is transforming how the customer journey is perceived by companies which have now the ability to generate and continuously update a high volume of data about their customers at a much faster rate than it was previously available. To fully explore a successful B2B relationship, companies should consider using journey-based tracking (Conway, 2017) with the strategic purpose of better mapping out and supporting a desirable customer experience. Firms should incorporate advanced analytics into customer success activities to develop successful customer experience strategies. Taking into consideration the various characteristics and complexities inherent in the B2B customer journey, it can be concluded that "being aware of every customer's experience is a crucially important component of every firm's strategies" (Adam, 2018).

Andersson (2017) suggests that by analyzing data along the customer journey, companies can describe and predict where customers are in their journey and individually enhance current product offerings or develop new offerings, promoting upselling and cross-selling opportunities that might accelerate value capture.

Gulati *et al.* (2005) consider that shift focus from past relationships to future behavior is a crucial step towards customer-centricity - firms should strive to gain insights into customers from past behavior to understand future behavior.

A strong customer-success effort in the B2B environment encompasses acknowledging the analytical insights that are crucial in understanding the customers' needs without needing to ask them directly (Opresnik *et al.,* 2015). Consequently, customers can then be managed and treated in different ways, according to their needs. Atkins (2018) highlighted that by seeking "opportunities to deliver more value to customers, companies derive more value in return".

A report by Forrester (2012) suggests that firms rely on various analytical methods to make personalization work across inbound and outbound customer interactions. Another research by McKinsey Analytics (2018)

found that "analytics create value when big data and advanced algorithms are applied to business problems to yield a solution that is measurably better than before".

Adam (2018) argues that opportunities embedded in big data, machine learning and predictive analytics capabilities can unlock the potential to improve the sales cycle in the B2B environment namely regarding cross-selling and upselling opportunity creation, management of customer relationships, efficient use of time and resources and sales effectiveness.

d. Big Data and Predictive Analytics

Lozada *et al.* (2019) considered big data as a phenomenon on which the competitive advantage of companies will be leveraged in the future. Big Data's most important goal is to enable organizations to make better decisions with the potential of improving organizational performance in all areas. (Martin, 2018)

Data-driven decision making refers to the "practice of basing decisions on the analysis of data rather than purely on intuition" (Li *et al.,* 2019). Asamoah *et al.* (2019) further suggest that "the value in big data is created when insights are mined to support business processes". Hollyoake (2009) claims that customers expect organizations to understand their needs. A recent Accenture (2018) report emphasized that organizations must "master big data analytics to anticipate their customer needs".

The core of big data is forecast and predictive analytics according to Mayer-Schönberger *et al.*(2013). Gartner (2012) has outlined the big data application roadmap (Annex III) where predictive analytics corresponds to the third stage of business analytics. Defined by SAS, it refers to "the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data." Typically the output is a score or code indicating the likelihood of future behavior or event. (Leventhal, 2018)

Table I - Application of predictive models examples and the related business questions (Leventhal, 2018).

Predictive Models Applications	Business Question
Customer selection from prospects	Which prospects are most likely to buy?
Cross-sell and up-sell campaigns	Which existing customers of a particular product are the most likely to buy another product or buy more of the particular product?
Next Best Offer	Which product customer is likely to buy next?
Customer retention	Which customers have the highest likelihood of lapsing?
Customer lifecycle management	How long it will take for the customer to likely lapse?
Customer lifetime value	What is the predicted future value of purchases for customers?

e. Machine Learning Models

Predictive analytics and machine learning go hand-in-hand, as predictive models include a machine learning algorithm. As a predictive analytics task, the goal of a classification problem is to predict a categorical target variable from a set of input variables in a labeled data set. There are various machine learning algorithms used in predictive models which will be presented in more detail in this chapter.

Logistic Regression (LR) is used to predict categorical target variables. LR utilizes a sigmoid function that ranges between zero and one. If the estimated probability is lower than a certain threshold, the model predicts the instance belonging to class 0, or else it predicts belonging to class 1.

Decision Trees (DT) may be used for both classification and regression problems, they "are fairly intuitive and their decisions are easy to interpret". (Géron, 2017) In general, the classification of a new example starts at the root node, moving down to a branch corresponding to a particular value of a certain feature, arriving at a new decision node with a new feature. This process is repeated until arrival at the terminal node.

A Random Forest (RF) is an ensemble of decision trees that can be used for either classification or regression problems. RF operates by constructing a multitude of decision trees, each one trained independently, using a random sample of the data. To classify an instance, each individual tree votes and the RF yields the mode of the classes.

The default base learners of Extreme Gradient Boosting (XGBoost) are tree ensembles. The tree ensemble model is a set of classification and regression trees. XGBoost uses leaf-wise growth strategy when growing the decision tree. Trees are grown iteratively, with each one being trained to reduce the misclassification rate of subsequent interactions.

Light Gradient Boosting Machine (Light GBM) is a gradient boosting framework that uses tree based learning algorithms. LightGBM splits the tree leaf-wise, so when growing on the same leaf, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence, yield better accuracy results.

f. Data Mining Process

The Cross Industry Standard Process for Data Mining (CRISP-DM) will be the research framework used in this study. The CRISP-DM is the process of extracting useful knowledge from data to solve business

problems and it can be broken down into reasonably well-defined. (Chapman *et al.,* 1999). Annex IV shows these phases while Annex V further presents an outline of phases accompanied by generic tasks and outputs.

3. Methodology

a. Imbalanced Learning Models

The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of unequal distribution of data between classes. (He *et al.,* 2009)

The class imbalance may either occur due to the relative proportion of examples belonging to each class being low (relative rarity) or the absolute number of examples belonging to each class available for learning being low (absolute rarity). There is also a distinction for between-class imbalance, which refers to imbalance occurring only due to the minority class, and within-class imbalance which refers to rare cases present within either the minority or the majority class. (Japkowicz *et al.,* 2001) When learning from data sets that contain very few instances of the minority class, in many cases the classifier tends to favor the majority class. (Babar, 2015) while the minority class usually represents the most important concept to be learned. (Lopez *et al.,* 2013). For both multiclass and binary classification problems, various data engineering techniques are practiced to handle the imbalanced data (Tanveer, 2019) which can be categorized into four categories: sampling-based methods, cost-based methods, kernel-based methods, and active learning-based methods. (Babar, 2015). Data sampling is the most common method and it occurs when the training instances are modified in such a way to produce a more or less balanced class distribution that allows classifiers to perform in a similar manner to standard classification. (Hu *et al.,* 2009; Babar, 2015)

Oversampling methods add samples to original imbalanced dataset to balance the size of the minority and majority classes. There are two types of oversampling: random oversampling and synthetic oversampling. SMOTE (synthetic minority oversampling technique), proposed by Chawla *et al.* (2003) is one of the most commonly used oversampling methods to solve the imbalance problems. Its graphical representation can be found in Annex VI.

SMOTE manages to handle between-class imbalance, whereas within-class imbalance remains ignored. (Prati *et al*., 2004). Based on the premise that SMOTE may generate synthetic instances in unsuitable

locations, such as overlapping regions and noise regions, He *et al.* (2008) proposed a novel approach adaptive synthetic (ADASYN). While SMOTE provides equal chance of each minority instance to get selected, with ADASYN the selection process is based on the minority class distribution, using the weighted distribution of minority samples according to their level of difficulty in learning. Samples that are harder to classify compared to those minority examples that are easier to learn have a higher weight than others.

According to He *et al.* (2008), "the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples."

b. Time Series Cross-Validation

In the context of predictive models, cross-validation is used to assess the performance of a classifier by obtaining an estimate of the true error. An estimate of the true error is important in practice, as it allows checking if a model generalizes well to unseen data or just memorizes the patterns in the training data, resulting in overfitting. (Neunhoeffer, 2019) With time series data, due to temporal dependencies, particular care must be taken in splitting the data in order to prevent data leakage. In such a context, instead of the common types of cross-validation such as k-fold cross-validation and hold-out cross-validation, a time-series cross-validation should be used. Since chronological ordering matters, data cannot be randomly split but rather it is ordered from the past to the present. The walking forward window cross-validation is an iterative method with multiple splits across the different time periods. The training set expands in each fold indicated by k. Instances that compose the test set come necessarily after the training set and similarly, the validation set is created with instances chronologically after the training subset (Hyndman, 2019).

c. Hyperparameter Optimization

Besides data splitting, another factor that needs to be considered in order to find the best algorithm is the choice of parameters values. According to Bergstra *et al.* (2015), hyperparameter optimization is the act of searching the space of possible configuration variables for a training algorithm in order to find a set of variables that allows the algorithm to achieve more desirable results. Hyperparameters choices generally have a significant effect on the success of machine learning algorithms. Hyperopt is a Python library for Sequential

Model-Based Optimization (SMBO) provided by Scikit-learn designed to perform hyperparameter optimization.

The first step to use Hyperopt is to define the objective function to minimize. Secondly, a configuration space object is defined with variable hyperparameters stochastic expressions which describes the domain over which Hyperopt is allowed to search for the classification algorithm. It is also possible to specify the values of specific hyperparameters which will remain constant during the search.

Optionally, a trials database and a search algorithm can also be defined. Each evaluation during optimization estimates and returns the desired metric on a validation set. At the end of the search, the best configuration is restrained on the whole dataset to produce the classifier that handles subsequent predict calls. (Komer, 2014)

d. Evaluating Classifier Performance

There have been several metrics proposed to measure the performance of a classifier applied on imbalanced data. A confusion matrix as shown in Table II is typically used to evaluate the performance of a machine learning algorithm alongside with the metrics on Table III.

Table II - Two-by-two confusion matrix for a binary classification problem

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

Table III - Common performance metrics based on the confusion matrix (Solokolova *et al.*, 2009)

Measure	Formula	Evaluation Focus
$Accuracy$	$\frac{TP + TN}{TP + FN + FP + TN}$	Overall effectiveness of a classifier
$Precision$	$\frac{TP}{TP + FP}$	Class agreement of the data labels with the positive labels given by the classifier
$Recall\ or\ True\ Positive\ Rate\ (TPR)$	$\frac{TP}{TP + FN}$	Effectiveness of a classifier to identify positive labels
$Specificity\ or$ $True\ Negative\ Rate\ (TNR)$	$\frac{TN}{TN + FP}$	How effectively a classifier identifies negative labels
$F - value$	$2\times \frac{Recall \times Precision}{Recall + Precision}$	Relations between data's positive labels and those given by a classifier

Traditionally, the "accuracy rate has been the most commonly used empirical measure" (Tanveer, 2019) . However, for imbalanced datasets, accuracy is not an appropriate measure as it may give an outstanding

performance level by classifying all the instances as the majority class, having a higher predictive accuracy over the majority class, but poorer predictive accuracy over the minority class. (Hu *et al.,* 2009)

Provost *et al.* (1998) suggest that classification accuracy assumes "equal misclassification costs" (for false positive and false negative errors), although for most real-world problems one type of classification error is much more expensive than another. The authors recommend using the Receiver Operating Characteristics (ROC) as an evaluation framework instead of accuracy for the binary classification problems. ROC curve is a two-dimensional graph depicting the trade-offs between the false positive rate (FPR) and true positive rate (TPR) along the x and y-axis, respectively. The area under the curve (AUC) score is calculated from the ROC curve and its value lies between 0 and 1. The larger the value of area under the curve, the better the model performance is. The Precision-recall (PR) curve is "highly informative about the performance of binary classifiers" (Keilwagen *et al.,* 2014). It is a trade-off between Precision at y-axis and Recall at x-axis for different threshold values. For instance, lowering the threshold value will decrease the Precision value and increase the Recall value.

Additionally, metrics such as precision, recall and F-value have been used to understand the performance of the learning algorithm on the minority class. (Chawla *et al.,* 2003)

Guo *et al.* (2008) points out that "for extremely skewed class distributions the recall of the minority class is often 0, meaning that there are no classification rules generated for the minority class". The main focus of all learning algorithms is to improve the Recall, without sacrificing the Precision. However, "the Recall and Precision goals are often conflicting and attacking them simultaneously may not work well, especially when one class is rare" (Chawla *et al.,* 2003). While ROC curves represent the trade-off between values of FPR and TPR, the F-value is a combination of both Precision and Recall where F-measure is the weighted harmonic mean of Precision and Recall of a classifier and it is a popular evaluation metric for the imbalanced datasets classification.

e. Model Interpretability

As outlined by Molnar (2019), "the higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made". In a business context, this

may hint that if the stakeholders understand the model, their confidence regarding its predictions is higher. Consequently, they may be more compelled to complement their current workflow with the model's output. SHAP (SHapley Additive exPlanations) recently developed by Lundberg *et al.* (2017) is a method to explain individual predictions of any machine learning model. SHAP is based on the game theoretically optimal Shapley Values. Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value. Features with large absolute Shapley values are considered important and can be plotted through the feature importance plot.

Lundberg *et. al* (2018) proposed TreeSHAP, a variant of SHAP for tree-based machine learning models where the Shapley values of a tree ensemble are the weighted average of the Shapley values of the individual trees thanks to the Additivity property of Shapley values. Individual predictions can be visualized with the Python SHAP package, with feature attribution as "forces" that either have a positive or negative impact in the prediction which starts from the baseline as the average of all predictions in the dataset.

f. Tool Selection

In this study, the tool used for the collection from different platforms and integration of the data was SQL. Python was used for examining the data, feature engineering, modeling and evaluation of models. The packages used in Python are described in Annex VII.

4. Data

a. Business Understanding

Selections regarding which customers to target for upselling opportunities based on individual personal knowledge and experience can be considered a baseline for the purpose of this study.

With the goal of identifying new ways of using predictive analytics that can complement already existing methods to improve the decision making process and the relevancy of actions undertaken by the Sales and Customer Success Teams, a predictive model is designed to classify customers' behavior. By highlighting customers who are more likely to expand their plan in the next six months (and therefore increase their Annual Recurring Revenue - ARR) whilst simultaneously understanding their needs, OutSystems may be able to spot more opportunities that might accelerate the value captured. The predictions will be updated on a quarterly basis.

b. Data Understanding

The data used throughout the thesis was provided by OutSystems.

OutSystems uses a centralized analytics database (commonly known as a data lake) aggregating data from multiple sources. Therefore, a significant amount of time was spent collecting and analyzing the quality of the vast volume of data collected and cleaning it.

When it comes to the temporal scope, the time interval defined in the extraction of the data is between January 2016 and August 2019. As for the geographical scope, the study concerns OutSystems' customers in Asia-Pacific (APAC), Europe, the Middle East and Africa (EMEA) as well as North America and South America (Americas).

The target definition was a critical step because of its implications on the type of observations selected. More importantly, the target definition had to take into account the business objectives, meaning that to be considered an expansion, a customer needs to increase his Annual Recurring Revenue (ARR) above a certain threshold measured in Euros. The target is given by the *Is_Exp_Next_6m* binary variable. It indicates whether a customer, identified through the *Acc_Id* feature, expanded his plan in the previous six months (*Is_Exp_Next_6m*=1) or not (*Is_Exp_Next_6m*=0). Most of the customers are not expanding their plan higher than a certain threshold frequently, resulting in an imbalanced response variable.

The individual features that were retrieved to build feature vector for every customer can be divided into three categories: customer demographics, customer information and platform variables. Customer demographics refer to factors that regarding the type, scale and other attributes concerned with customers which are independent of OutSystems. These include the size of the company, the geographical demography and the type of industry the customer belongs to. Customer information encompasses the age of each customer in the organization (in months) historic buying patterns, including the recency of the previous expansions. Platform features concern the measurement of each customer's usage of various metrics across different environments (development and production) accessible through the OutSystems platform.

Owing to the confidential nature of the data used, the name and corresponding data of customers is anonymized. To gain further acquaintance of the data, descriptive statistics were computed and an exploratory

data analysis approach was conducted to "develop an initial idea of possible associations between the attributes and the target variable". (Larose, 2005)

c. Data Preparation

Before proceeding to the modeling phase, data preparation techniques were applied to ensure that data was in the proper format to serve as input for modeling. Feature identification and selection are important steps for supervised machine learning algorithms. Domain expertise and past experience helped in identifying a set of features that are relevant. Regarding the definition of the target variable and the problem context, further understanding needed to be provided. This was accomplished by modifying the way some features were represented and creating new ones by turning them into time-lagged variables. Some customer features were left out since they were considered to be redundant, for instance, territory customer feature was dropped as it is always associated with the country customer feature.

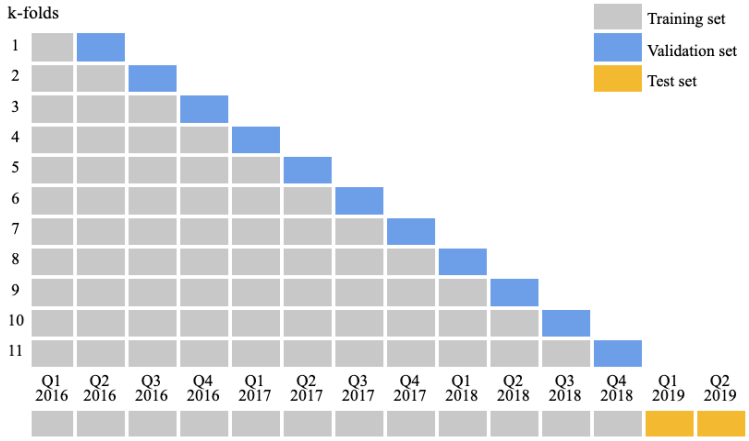
Next to engineering the features, the dataset also required further processing to make it suitable as input for the machine learning models. Missing values were replaced and categorical variables were encoded, assuming the order of the categories is non-existent.

There was a need to handle issues associated with temporal data and to ensure all features were correctly accounting for time-shifts across the various time-windows were used for the train, validation and test sets the various machine learning models. With the final dataset is completely cleaned, in order to feed the features to the prediction model, they were divided into two types: time variant and time invariant.

d. Modeling

The approach followed for partitioning the data into training and validation sets was based on the time series cross-validation method. When dividing the available data into training and validation for the different k-folds, careful preparation was required to represent the time variant features. The final dataset was divided into quarters. For each of the k-folds, the validation set was represented by instances belonging to a new quarter and the training set was composed of instances that occurred in quarters prior to the observation that forms the test set.

Figure II - Time series cross validation



For each of the platform usage features, new features were created to account for the usage during the quarters represented in each of the training sets: (1) the average usage; (2) coefficient of variation, given by the ratio of the standard deviation to the mean, to provide a measure of the variability; (3) percentage difference.

The five classifiers used are Logistic Regression, Decision Tree, Random Forest, XGBoost and LightGBM and two oversampling methods compared are SMOTE and ADASYN.

For each of the classifiers, a space of hyperparameters to explore is defined in hyperopt functions which returns the best possible hyperparameter combination and the metric on the validation set for each of the folds which was chosen to be the Area Under the Receiver Operating Characteristic curve (AUROC).

Again, for all the classifiers, the hyperparameters space was thoroughly explored, running it for 100 rounds by setting max_evals equal to 100. Each of the parameter grids can be found in Annex VIII.

After iterating through all the k folds and running suggested search in the hyperopt algorithm in each of them, the hyper parameters that yield the highest AUROC will be used to predict on the final test set, composed of the last two quarters of the period in analysis, to evaluate the performance of the final classifier.

A single seed value was used for all random factors where applicable in order to reproduce the experiments with the same results

e. Evaluation

The performance results were evaluated separately for the five different classifiers different and two oversampling methods. Additionally, the performance was evaluated separately for the three different binary

targets, which differed on the amount considered to be qualified as a customer expansion - they will be mentioned as "low", "medium" and "high" expansion thresholds.

5. Results

a. Overall Model Comparison

The classification results are determined from the average of a 11-fold time series cross validation. Annex IX shows the overall comparison of the model performance (AUROC) for each of the Algorithms for different expansion thresholds used in this thesis, while Table IV highlights the results for the highest expansion threshold on the test data.

Table IV - Results of all algorithms for the "High" expansion threshold.

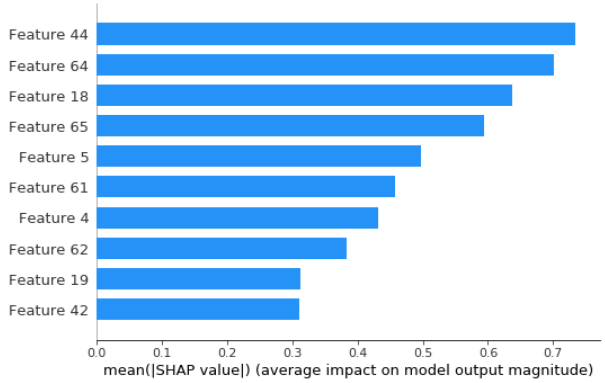
Expansion Threshold	Classifier	ADASYN method AUROC Test	SMOTE method AUROC Test
High	Logistic Regression	*0.573*	*0.561*
High	Decision Tree	*0.555*	*0.562*
High	Random Forest	*0.611*	*0.586*
High	XGBoost Classifier	*0.577*	*0.597*
High	LightGBM	***0.640***	***0.642***

The LightGBM model with the ADASYN oversampling method performed best, with the hyperparameters outlined in the Annex X. After selecting the classifier that yielded the highest AUROC, careful consideration was necessary when varying the output probability threshold of this classifier.

b. Explanation at the Global Level

The LightGBM Python package has a *feature_importance* method which can generate visualizations for two metrics referenced as gain (total gains of splits which use the feature) and spilt (numbers of times the feature is used) (Annex XI). However with these metrics only, it is difficult to draw a definitive conclusion whether these features are contributing positively or negatively to the predictions. As such, depicted by Figure III, the SHAP tree algorithm will be used to analyze which variables had the most significant impact global on the model prediction for class 1, that is, when a customer is predicted to expand in the next six months.

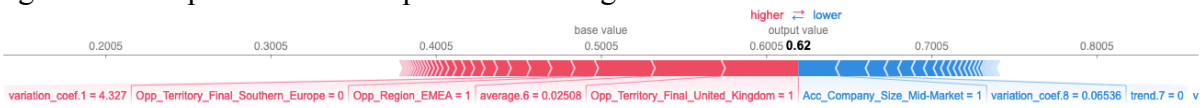
Figure III – LightGBM feature importance for class 1 prediction



c. Explanation at the Local Level

Given the problem context, the global approach might not be representative of each individual customer across the customer base (i.e. a feature that has an important positive impact for one customer might have an opposite effect for another customer). As such, the force plot for a single instance is shown in Figure III, where each SHAP value measures how much each predictor has contributed positively (in red) or negatively (in blue) to an individual predicted expansion opportunity.

Figure IV - force plot for an instance predicted to belong to class 1



6. Conclusion

a. Key Findings

This thesis demonstrated an approach to developing a predictive model to support the compound growth engine of sales cycle at OutSystems by predicting who are the customers that will expand their current plan in the next six months. The first part detailed the theoretical aspects alongside some business specifications for the B2B PaaS environment. The data mining process was presented and related to a predictive model development. Data cleaning and analysis was conducted, with the main emphasis placed on getting a deep understanding of different features and their distribution. Then, the dataset was split into training, validation and test data according to the time series cross validation. Then, the various algorithms with different oversampling methods were applied and their evaluation was reviewed. Additionally, since interpretability is considered a crucial aspect in this context, the SHAP algorithm was applied to the final model to get further insights on which features contribute the most for customers to expand their plan in the next six months both

on a global and on a local level. The LightGBM model with the SMOTE method performed better in comparison to the baseline model, which was considered as the subjective and manual classification of expansion opportunities. It was shown that with 88% of accuracy, 89% of weighted recall and 80% of weighted f-score, it is possible to predict which customers that are going to expand and which are not within the next six months.

b. Limitations

The dataset used in the thesis is anonymized, making it hard to draw conclusions about their meaning. Due to the influence of internal and external factors, the classification of customers' likelihood of expanding in the next six months is complex. For the purpose of this study, it would be helpful to disclose some insights through the data mining process. At the moment, the model does not take into account the tracking of the various touchpoints across the customer experience as well as behavioral and other non-transaction customer data. Additionally, seasonality is not being taken into consideration.

There may be a struggle to convert the analytical insights into action across the sales cycle as end users of the model may resist using it for many reasons (i.e. mistrust on the predictions or lack of training).

c. Further Work

Given the limitations recognized, suggestions for further research can be made. Future work on feature engineering could be a meaningful step towards achieving a model with higher performance. Including seasonality in the model could be valuable as well. There are other approaches to solving the problem of imbalanced data set that can be further investigated. Additional variables could be tracked and measured, specifically regarding touchpoints across customer experience (i.e. how many times and how many months prior a customer was contacted before successfully expanding) and behavioral customer features (i.e. integrating information on customer's engagement with the emails).

Mapping such customer interactions could also be further complemented with the level of upselling customers choose following the expansion opportunity. Future work could consider assessing the SHAP values explanations' implementation and utility in particular scenarios as well as quantifying the model's business value and how it impacts the decision-making process.

7. References

Opresnik, D. & Taisch, M. 2015. "The value of big data in servitization", *International Journal of Production Economics,* Vol. 165, pp. 174–184. 10.1016/j.ijpe.2014.12.036.

Provost, F., & Fawcett, T. 2013. "Data Science and Its Relationship to Big Data and Data-Driven Decision Making." *Big Data* 1, no. 1: 51–59. 10.1089/big.2013.1508.

Bokman, Alec, & Fiedler, L., Perrey, J., & Pickersgill, A. 2014. "Five Facts: How Customer Analytics Boosts Corporate Performance." https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/five-facts-how-customer-analytics-boosts-corporate-performance. Accessed on 2019-12-28.

Phillips-Wren, G., & Hoskisson, A. 2015. "An Analytical Journey towards Big Data." *Journal of Decision Systems* 24, no. 1: 87–102. 10.1080/12460125.2015.994333.

Adam, M. 2018. "Improving complex sale cycles and performance by using machine learning and predictive analytics to understand the customer journey." http://hdl.handle.net/1721.1/118010

Hale, K. 2018. "Predictive Analytics for Marketing: What It Can Do and Why You Should Be Using It. Towards Data Science Blog". https://towardsdatascience.com/predictive-analytics-for-marketing-what-it-can-do-and-why-you-should-be-using-it-afdbde131b36. Accessed: 2019-12-28.

Schubert, L., Jeffery, K., & Neidecker-Lutz, B. 2010. The Future Of Cloud Computing, Opportunities for European. Cloud Computing Beyond. E*uropean Commission Informarion and Society Theme – Expert Group Report*.

Gartner. 2012. Magic Quadrant for BI platforms. Analytics Value Escalator. Accessed on 2019-12-28.

Zhang, Q., and Cheng, L. & Boutaba, R.. 2010. "Cloud Computing: State-of-the-art and Research Challenges." *Journal of Internet Services and Applications*. 1. 7-18. 10.1007/s13174-010-0007-6.

Marston, S., Zhi, L., Bandyopadhyay S., & Ghalsasi, A. "Cloud Computing - The Business Perspective." *2011 44th Hawaii International Conference on System Sciences*, 2011. 10.1109/hicss.2011.102.

Stamford, C. 2019. "Gartner Says Nearly 50 Percent of PaaS Offerings Are Now Cloud-Only". Gartner. https://www.gartner.com/en/newsroom/press-releases/2019-02-27-gartner-says-nearly-50-percent-of-paas-offerings-are-. Accessed on 2019-12-28.

Richardson, C., Rymer, J., Mines, C., Cullen, A., & Whittaker, D. 2014. "New Development Platforms Emerge For Customer-Facing Applications". Forrester. https://www.forrester.com/report/New+Development+Platforms+Emerge+For+CustomerFacing+Applications/-/E-RES113411. Accessed on 2019-12-28.

Smith, D., Iijima, K., Altman, R., Driver, M., Pezzini, M. & Natis, Y. 2014. Gartner. "Magic Quadrant for Enterprise Application Platform as a Service". https://www.gartner.com/en/documents/2645317. Accessed on 2019-12-28.

Morgan, L. "What Is Platform as a Service?" OutSystems. https://www.outsystems.com/blog/posts/what-is-platform-as-a-service/. Accessed on 2019-12-28.

Driver, M, Baker, V., Iijima, K., Natis, Y., Dunie, R., & Vincent, P. 2017. Gartner. "Magic Quadrant for Enterprise High-Productivity Application Platform as a Service". https://www.gartner.com/en/documents/3695317/magic-quadrant-for-enterprise-high-productivity-applicat. Accessed on 2019-12-28.

Vincent, P., Iijima, K., Driver, M., Wong, J., & Natis, Y. 2019. Gartner. https://www.gartner.com/doc/reprints?id=1-1XQ92DO5&ct=191105&st=sb. Accessed on 2019-12-28.

OutSystems, n.d.. "OutSystems Evaluation Guide." https://www.outsystems.com/evaluation-guide/. Accessed on 2019-12-28.

Cáceres, R. & Paparoidamis, N. 2007. "Service Quality, Relationship Satisfaction, Trust, Commitment and Business-to-business Loyalty". *European Journal of Marketing - EUR J MARK*. 41. 836-867. 10.1108/03090560710752429.

Hollyoake, M. 2009. "The Four Pillars: Developing a 'Bonded' Business-to-Business Customer Experience." *Journal of Database Marketing & Customer Strategy Management* 16. 10.1057/dbm.2009.14.

Lemon, K.N., & Verhoef, P.C. 2016. "Understanding Customer Experience Throughout the Customer Journey." *Journal of Marketing* 80, no. 6: 69–96. 10.1509/jm.15.0420.

Homburg, C., Jozić, D., & Kuehnl, C. 2015. "Customer experience management: towards implementing an evolving marketing concept", *Journal of the Academy of Marketing Science*. 10.1007/s11747-015-0460-7.

Zolkiewski, J., Story, V., Burton, J., Chan, P., Gomes, A., Hunter-Jones, P., O'Malley, L., D. Peters, L., Raddats, C., & Robinson, W. 2017. "Strategic B2B Customer Experience Management: the Importance of Outcomes-Based Measures." *Journal of Services Marketing* 31, no. 2. 172–84. 10.1108/jsm-10-2016-0350.

Puccinelli, N.M., Goodstein, R.C., Grewal, D., Price, R., Raghubir, P., & Stewart, D. 2009. "Customer Experience Management in Retailing: Understanding the Buying Process." *Journal of Retailing* 85, no. 1. 15–30. 10.1016/j.jretai.2008.11.003.

Court, D., Elzinga, D., Mulder, S., & Vetvik, O.J.. 2009. "The consumer decision journey". https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey. Accessed on 2019-12-28.

Lam, S.Y., Shankar, V., Erramilli, M.K., & Murthy, B. 2004. "Customer Value, Satisfaction, Loyalty, and Switching Costs: An Illustration From a Business-to-Business Service Context." *Journal of the Academy of Marketing Science* 32. 10.1177/0092070304263330.

Chalmeta, R. 2006. Methodology for customer relationship management. 10.1016/j.jss.2005.10.018.

Viveiros, B.N. 2017. "B2B EVENT SPENDING ON RISE: SURVEY". https://www.chiefmarketer.com/b2b-event-spending-on-rise-survey/. Accessed on 2019-12-28.

Bardauskaite, I. 2014. Loyalty in the Business-to-Business Service Context: A Literature Review and Proposed Framework, *Journal of Relationship Marketing*, 13:1, 28-69. 10.1080/15332667.2014.882628.

Cahill, D.L., Goldsby, T.J., Knemeyer, A.M., & Wallenburg, C.M. 2010. "Customer Loyalty In Logistics Outsourcing Relationships: An Examination Of The Moderating Effects Of Conflict Frequency." *Journal of Business Logistics* 31. 10.1002/j.2158-1592.2010.tb00151.x.

Kumar, V., and Reinartz, W. 2012. "*Customer Relationship Management: Concept, Strategy, and Tools*." Berlin: Springer Berlin.

vanderKooij, J. 2019. "Frameworks That Govern B2B Marketing and Sales and Why SaaS Needs its Own Framework". https://winningbydesign.com/frameworks-that-govern-b2b-marketing-and-sales/. Accessed on 2019-12-28.

Ramani, G., & Kumar, V. 2008. "Interaction Orientation and Firm Performance." *Journal of Marketing* 72. 27–45. 10.1509/jmkg.72.1.027.

Atkins, C., Gupta, S., & Roche, P. 2018. "Introducing customer success 2.0: The new growth engine". https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/introducing-customer-success-2-0-the-new-growth-engine. Accessed on 2019-12-28.

David, C., & Knight, T. 2017. "B2B Customer Experience: Winning in the Moments that Matter". https://customerthinking.kpmg.co.uk/articles/b2b-customer-experience-winning-in-the-moments-that-matter/. Accessed on 2019-12-28.

Andersson, T., Boedeker, M., & Vuori, V. 2017. "Emotion-Gauge: Analyzing Affective Experiences in B2B Customer Journeys." *Strategic Innovative Marketing Springer Proceedings in Business and Economics*. 10.1007/978-3-319-56288-9_5.

Gulati, R. & Oldroyd, J.B. 2005. "The quest for customer focus." *Harvard Business Review*. 92-101, 133.

Opresnik, D. & Taisch, M. 2015." The value of Big Data in servitization." *International Journal of Production Economics*. 165. 10.1016/j.ijpe.2014.12.036.

Atkins, C., and, Gupta, S., & Roche, P. 2018. "Introducing customer success 2.0: The new growth engine". https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/introducing-customer-success-2-0-the-new-growth-engine. Accessed on 2019-12-28.

Sridharan, S.S., and, Frankland, D., & Smith, A. 2012. "Use Customer Analytics To Get Personal". Forrester. https://www.forrester.com/report/Use+Customer+Analytics+To+Get+Personal/-/E-RES61430. Accessed on 2019-12-28.

McKinsey Analytics. 2018. "Analytics comes of age - McKinsey". https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Analytics%20comes%20of%20age/Analytics-comes-of-age.ashx. . Accessed on 2019-12-28.

Lozada, N., Arias-Pérez, J. & Perdomo-Charry, G. 2019. "Big data analytics capability and co-innovation: An empirical study." Heliyon, Volume 5, Issue 10, e02541, ISSN 2405-8440, 10.1016/j.heliyon.2019.e02541.

Martin, V. M.A, Dr. K. David, A.Vignesh. 2018. "Big Data and Its Challenges", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* (IJSRCSEIT), ISSN : 2456-3307, Volume 3, Issue 3, pp.533-538. http://ijsrcseit.com/CSEIT1833169

Li, M., Wu, Y., He, Y., Huang, S., & Nair, A. 2019. "Sparse Inverse Covariance Estimation: A Data Mining Technique to Unravel Holistic Patterns among Business Practices in Firms." *Decision Sciences*. 10.1111/deci.12404.

Asamoah, D.A., & Sharda, R. 2019. "CRISP-eSNeP: Towards a data-driven knowledge discovery process for electronic social networks", *Journal of Decision Systems*, 28:4, 286-308, 10.1080/12460125.2019.1696614

Chakraborty, A., & Vernocchi, M. 2018. Accenture. "Analytics Everywhere Smarter Actions, Happier Customers, Greater Value". https://www.accenture.com/us-en/~/media/PDF-26/Accenture-CMT-Analytics-Everywhere-July18.pdf. Accessed on 2019-12-28.

Mayer-Schönberger, V. & Cukier, K., 2013. Big Data: A Revolution that Will Transform how We Live, Work and Think. 1st ed. New York: Houghton Mifflin Harcourt. 10.3233/ip-140322.

SAS, ed. "Predictive Analytics - What it is and why it matters". https://www.sas.com/en_us/insights/analytics/predictive-analytics.html. Accessed on 2019-12-28.

Leventhal, Barry. 2018. "Predictive Analytics for Marketers: Using Data Mining for Business Advantage." Kogan Page Publishers.

Géron, A. 2017. "Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems". Sebastopol, CA: O'Reilly Media. ISBN: 978-1491962299

Chapman, P., *et al*. 1999. "CRISP-DM 1.0 Step-by-step data mining guide", SPSS.

Japkowicz, N., & Stephen, S. 2002. "The Class Imbalance Problem: A Systematic study1." *Intelligent Data Analysis* 6, no. 5. 429–49. 10.3233/ida-2002-6504.

Varsha, S.B. & Ade, R. 2015. "A Review on Imbalanced Learning Methods." *IJCA Proceedings on National Conference on Advances in Computing* NCAC, 23-27.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. 2013. "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics." *Information Sciences* 250: 113–41. 10.1016/j.ins.2013.07.007.

Tanveer, A. 2019. Churn Prediction Using Customers' Implicit Behavioral Patterns and Deep Learning.

Hu, S., Liang, Y., Ma, L., & He, Y.. 2009. "MSMOTE: Improving Classification Performance When Training Data Is Imbalanced." *2009 Second International Workshop on Computer Science and Engineering*. 10.1109/wcse.2009.756.

Chawla, N.V., Lazarevic, A., Hall, L.O., & Bowyer, K.W. 2003. "SMOTEBoost: Improving Prediction of the Minority Class in Boosting." *Knowledge Discovery in Databases: PKDD 2003 Lecture Notes in Computer Science*. 107–19. 10.1007/978-3-540-39804-2_12.

Xie, W., Liang, G., Dong, Z., Tan, B., & Zhang, B.. 2019. "An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data," *Mathematical Problems in Engineering*, Article ID 3526539, 10.1155/2019/3526539.

Prati, R.C., Batista, G. E. A. P. A. , & Monard, M.C. 2004. "Learning with Class Skews and Small Disjuncts." *Advances in Artificial Intelligence – SBIA 2004 Lecture Notes in Computer Science*, 296–306. 10.1007/978-3-540-28645-5_30.

He, H., Bai, Y., Garcia, E., & Li, S. 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the International Joint Conference on Neural Networks. 1322 - 1328. 10.1109/IJCNN.2008.4633969.

Neunhoeffer, M., & Sternberg, S. 2019. How Cross-Validation Can Go Wrong and What to Do About It. Political Analysis, 27(1), 101-106. 10.1017/pan.2018.39

Hyndman, R. 2016. "Cross-validation for time series". https://robjhyndman.com/hyndsight/tscv/. Accessed on 2019-12-28.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D.D. "Hyperopt: a Python Library for Model Selection and Hyperparameter Optimization." *Computational Science & Discovery* 8, no. 1 (2015): 014008. 10.1088/1749-4699/8/1/014008.

Komer, B., Bergstra, J., & Eliasmith, C. 2014. "Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn." *Proceedings of the 13th Python in Science Conference*. 10.25080/majora-14bd3278-006.

Sokolova, M., & Lapalme, G. 2009."A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing & Management* 45, no. 4: 427–37. 10.1016/j.ipm.2009.03.002.

Provost, Fawcett & Kohavi. 1998. The Case Against Accuracy Estimation for Comparing Induction Algorithms.

Keilwagen, J., Grosse, I., & Grau, J. 2014. "Area under Precision-Recall Curves for Weighted and Unweighted Data." *PLoS ONE* 9, no. 3. 10.1371/journal.pone.0092209.

Guo, X., Yin, Y., Dong, C., Yang, G. & Zhou, G. 2008. On the Class Imbalance Problem. *Fourth International Conference on Natural Computation*, ICNC '08. Vol. 4. 110.1109/ICNC.2008.871.

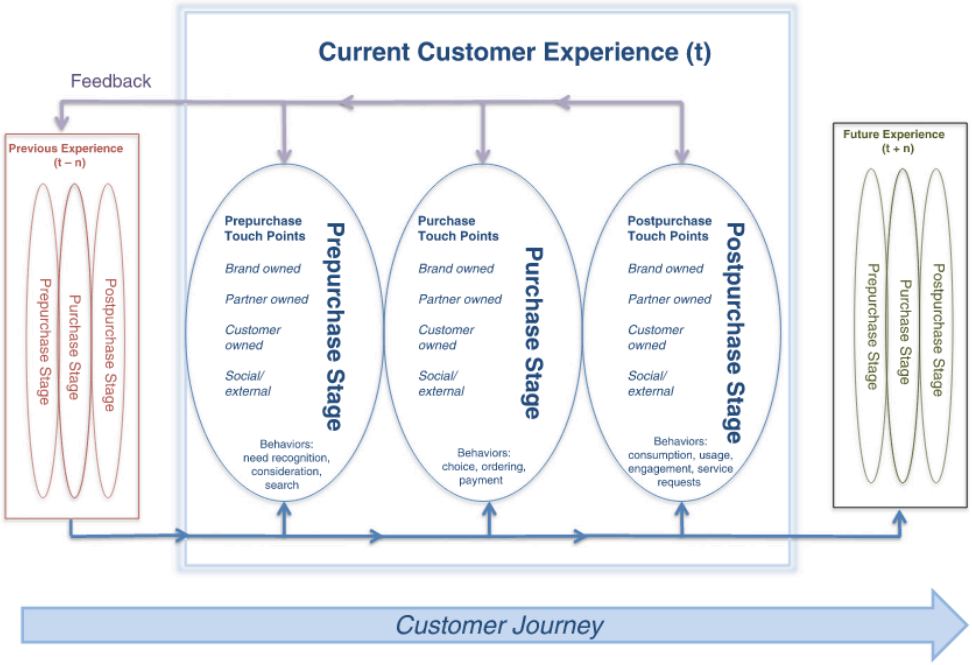
Molnar, C. 2019."Interpretable machine learning. A Guide for Making Black Box Models Explainable". https://christophm.github.io/interpretable-ml-book/. Accessed on 2019-12-28.

Lundberg, S., & Lee, S. 2017. "A Unified Approach to Interpreting Model Predictions." ArXiv abs/1705.07874

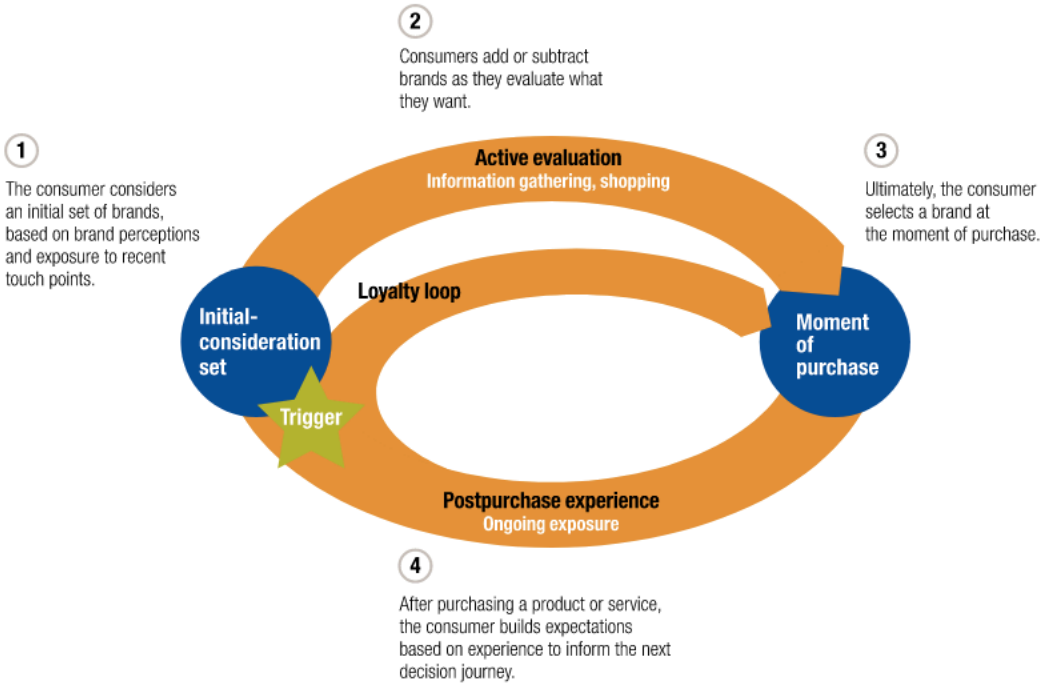
Lundberg, S.M., Erion, G.G., & Lee, S. 2018. "Consistent Individualized Feature Attribution for Tree Ensembles." ArXiv, abs/1802.03888

Appendix

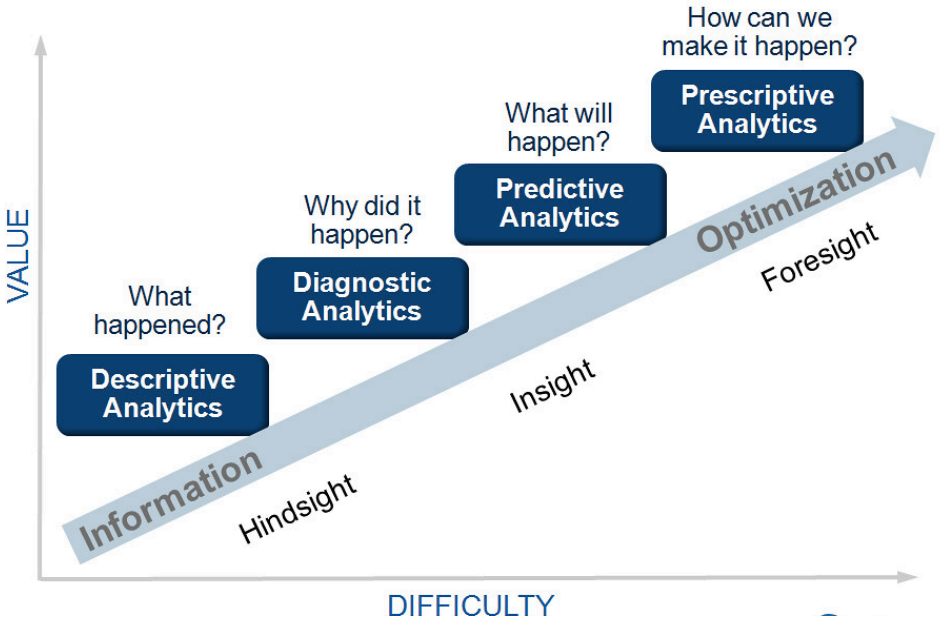
Annex I - Process Model for Customer Journey and Experience (Retrieved from Lemon *et al.*, 2016)



Annex II - Loyalty Loop (Retrieved from Court *et al.*, 2009)

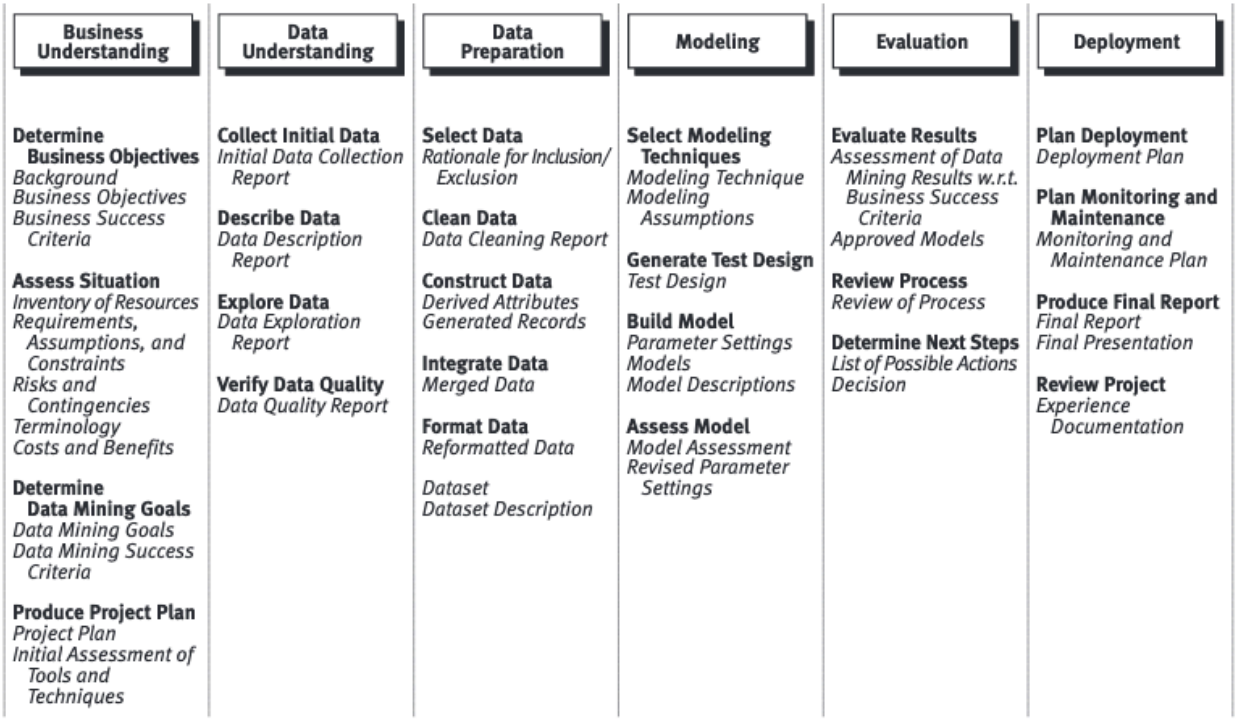


Annex III - Analytic Value Escalator (Retrieved from Gartner, 2012)

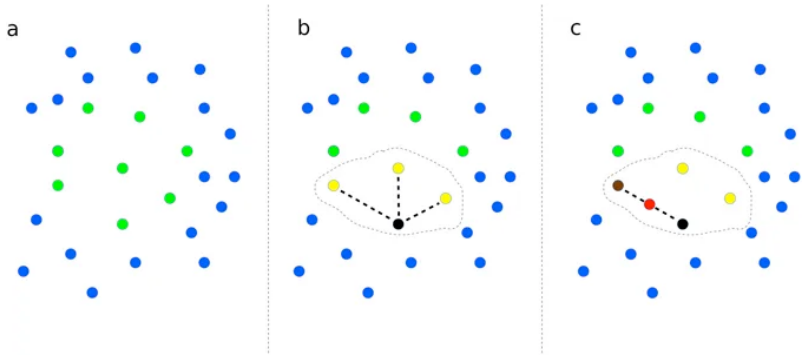


Annex IV - Phases of the CRISP-DM reference mode (Retrieved from Chapman *et al.* 1999)

Annex V - Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model (Retrieved from Chapman *et al.* 1999)



Annex VI - Graphical representation of the process of SMOTE (Retrieved from Schubach *et al.,* 2017)



"(a) SMOTE starts from a set of positive (green points) and negative (blue points) examples; (b) It then selects a positive example (black) and its k nearest neighbors among the positives (yellow points, with k = 3), (c) Finally one of the k nearest neighbours is randomly selected (brown point) and a new synthetic positive example is added, by randomly generating an example (red point) along the straight line that connects the black and brown points. The procedure depicted in (b,c) is repeated for all the positives, by adding each time a new synthetic example similar (in an Euclidean sense) to the other positive examples."

Annex VII - Python packages used in this thesis

Package	Version
numpy	1.17.4
pandas	0.25.3
sklearn	0.21.3
shap	0.34.0
imblearn	0.4.3
lightgbm	2.2.3
xgboost	0.90

Annex VIII - Hyperparameter search space for the different machine learning models in Python code

Annex VIII.A. Logistic Regression

```
space_lr = {'warm_start' : hp.choice('warm_start', [True, False]),
            'fit_intercept' : hp.choice('fit_intercept', [True, False]),
            'tol' : hp.uniform('tol', 0.00001, 0.0001),
            'C' : hp.uniform('C', 0.05, 3),
            'max_iter' : hp.choice('max_iter', range(100,1000)),
            'multi_class' : 'auto',
            'class_weight' : 'balanced'}
```

Annex VIII.B. Decision Tree

```
space_dt = {
    'max_depth': hp.choice('max_depth', range(1,50)),
    'max_features': hp.choice('max_features', range(1,50)),
    'criterion': hp.choice('criterion', ["gini", "entropy"])}
```

Annex VIII.C. Random Forest

```
param_space_rf = {
    'max_depth': hp.choice('max_depth', range(1,20)),
    'max_features': hp.choice('max_features', range(1,50,5)),
    'n_estimators': hp.choice('n_estimators', range(100,500,10)),
    'criterion': hp.choice('criterion', ["gini", "entropy"])}
```

Annex VIII.D. XGBoost

```
space_xgb ={'max_depth': hp.choice('max_depth', range(1,50)),
        'min_child_weight': hp.quniform ('x_min_child', 1, 10, 1),
        'subsample': hp.uniform ('x_subsample', 0.7, 1),
        'gamma' : hp.uniform ('x_gamma', 0.1,0.5),
        'colsample_bytree' : hp.uniform ('x_colsample_bytree', 0.7,1),
        'reg_lambda' : hp.uniform ('x_reg_lambda', 0,1),
        'n_estimators': hp.choice('n_estimators', range(100,300,10))}
```

Annex VIII.E. LightGBM

```
space_lgb = {'bagging_fraction': hp.quniform('bagging_fraction', 0.85, 0.95, 0.005),
             'bagging_freq': 1,
             'cat_l2': hp.quniform('cat_l2', 5, 15, 0.25),
             'cat_smooth': hp.quniform('cat_smooth', 5, 15, 0.05),
             'cegb_tradeoff': hp.quniform('cegb_tradeoff', 0.93, 0.99, 0.01),
             'feature_fraction': hp.quniform('feature_fraction', 0.5, 0.95, 0.005),
             'is_unbalance':'true',
             'learning_rate': hp.quniform('learning_rate', 0.05, 0.15, 0.005),
             'max_depth': -1,
             'metric': 'auc',
             'min_sum_hessian_in_leaf': hp.quniform('min_sum_hessian_in_leaf', 0.05, 0.15, 0.01),
             'min_data_in_leaf': hp.randint('min_data_in_leaf', 25),
             'min_gain_to_split': hp.quniform('min_gain_to_split', 0.05, 0.06, 0.005),
             'n_estimators': hp.choice('n_estimators', np.arange(50, 501, 10, dtype='int')),
             'num_leaves': hp.choice('num_leaves', np.arange(50, 2501, 50, dtype='int')),
             'objective': 'binary',
             'tree_learner': hp.choice('tree_learner',['serial', 'feature', 'data', 'voting'])}
```

Annex IX - Overall model performance comparison

Expansion Threshold	Classifier	ADASYN method AUROC Test	SMOTE method AUROC Test
High	Logistic Regression	*0.573*	*0.561*
High	Decision Tree	*0.555*	*0.562*
High	Random Forest	*0.611*	*0.586*
High	XGBoost Classifier	*0.577*	*0.597*
High	LightGBM	*0.640*	***0.642***
Medium	Logistic Regression	*0.571*	*0.561*
Medium	Decision Tree	*0.559*	*0.572*
Medium	Random Forest	*0.581*	*0.603*
Medium	XGBoost Classifier	*0.577*	*0.585*
Medium	LightGBM	*0.635*	*0.571*
Low	Logistic Regression	*0.553*	*0.560*
Low	Decision Tree	*0.528*	*0.570*
Low	Random Forest	*0.601*	*0.603*
Low	XGBoost Classifier	*0.604*	*0.596*
Low	LightGBM	*0.609*	*0.571*

Annex X - Best hyperparameters for the Final Model (LightGBM model using SMOTE)

Hyperparameter	Value
bagging_fraction	*0.9*
bagging_freq	*1*
cat_l2	*9.0*
cat_smooth	*9.65*
cegb_tradeoff	*0.96*
feature_fraction	*0.915*
is_unbalance	*'true'*
learning_rate	*0.08*
max_depth	*-1*
metric	*'auc'*
min_sum_hessian_in_leaf	*7*
min_gain_to_split	*0.05*
min_sum_hessian_in_leaf	*0.15*
n_estimators	*370*
num_leaves	*1800*
objective	*'binary'*
tree_learner	*'serial'*

Annex XI – Feature Importance according to "split" for the Final Model (LightGBM model using SMOTE)