Alexandru COŞER, PhD Candidate
Economic Cybernetics and Statistics Doctoral School
E-mail: alexandru.coser@ie.ase.ro
Associate Professor Anamaria ALDEA, PhD
The Department of Economic Informatics and Cybernetics
E-mail: anamaria.aldea@csie.ase.ro
Associate Professor Monica Mihaela MAER-MATEI, PhD
The Department of Economic Informatics and Cybernetics
E-mail: matei.monicamihaela@gmail.com
Levida BEŞIR, PhD Candidate
Economic Cybernetics and Statistics Doctoral School
E-mail: levida.besir@gmail.com
The Bucharest University of Economic Studies

PROPENSITY TO CHURN IN BANKING: WHAT MAKES CUSTOMERS CLOSE THE RELATIONSHIP WITH A BANK?

Abstract. *Nowadays, customer analytics plays a fundamental role in the commercial activity of a financial company, which is interested to deliver the best products and services to consumers. However, sometimes things may possibly go wrong, resulting in inconveniences experienced by the clients, who may eventually decide to end the relationship with a bank. This study uses a dataset which contains a bank's customer base who churned, with a series of variables regarding socio-demographic characteristics, member activity, balance, estimated salary and tenure. An issue of general interest in banking field is predicting the probability of churn or attrition, a phenomenon which may lead to significant decreasing of revenues if the company fails to act on time. Therefore, in this paper we examine which are the main characteristics of customers that influence the propensity to churn by means of an exploratory data analysis. Moreover, we employ two machine learning algorithms and use an optimization technique called Grid Search to obtain the optimal predictive model, which can estimate the likelihood of a customer leaving a bank in the future. Ultimately, we use the Area Under the Curve (AUC) as model performance metric to compare the outcomes of Logistic Regression and Random Forest classifiers on the test dataset. Final results show that Random Forest classifier outperformed Logistic Regression model in terms of AUC values and, at the same time,the outcomes emphasize different behaviors between countries.*

Keywords*: churn, attrition, machine learning, grid search, random forest, logistic regression, AUC.*

JEL Classification: C38, C51, C52, C55, C61

DOI: 10.24818/18423264/54.2.20.05

1. Introduction

Recent advances in data science field improved the ability to extract customer intelligence from data by transforming raw information into meaningful and actionable knowledge. Customer analytics plays a major role in the commercial activity of a financial company since the stakeholders are interested to deliver the most relevant products and services to consumers worldwide and improve customer experience and satisfaction. However, sometimes things may possibly go wrong, resulting in poor customer experience. Such cases can turn into inconveniences experienced by the client, who may eventually decide to end the relationship with a bank if his or her issues are not resolved in a timely manner. In order to address such issues, data scientists are able to use data mining techniques to extract knowledge from data and offer insights to decision makers. Thus, they can help the business to make better decisions in the future for revenue growth or loss reduction.

An issue of general interest in banking field is predicting the probability of churn or attrition, a phenomenon which may lead to a significant decrease in profits if the company fails to act on time. Therefore, in this paper we aim to identify which are the main characteristics of customers that significantly influence the propensity to churn. This study is based on a dataset from *Kaggle*[1] which consists of a sample of bank customers. Collected data offer information on variables such as the *churn status*, *socio-demographiccharacteristics*, *member activity*, *balance*, *estimated salary* and *tenure*, to name a few of them. An exploratory data analysis was firstly performed. Secondly, we used two machine learning algorithms and an optimization technique called *Grid Search* to obtain the optimal predictive model, which might estimate the likelihood of a customer to leave a bank in the future. Ultimately, we use the Area Under the Curve (AUC) as model performance metric to assess the accuracy of *Logistic Regression* and *Random Forest* classifiers.

The data analysis is performed using Python programming language version 3.7, with the following libraries: *pandas*,*numpy*,*scikit-learn*,*matplotlib*,*seaborn*,*scipy*.The paper is divided in these main sections: *datadescription*, *methodology*, *exploratorydataanalysis*, *predictive models* and ends with the *conclusions* and *discussion*.

Data mining is a process that seeks to discover hidden patterns in large datasets and aims to explain the behavior and profile of the objects which are analyzed from a specific dataset. A series of examples consist of techniques such as *machine learning, cluster analysis, neural networks, dimensionality reduction* etc.

Banking analytics registers continuous progresses through recent developments in Big Data techniques, which seek to uncover valuable information and knowledge from the large available data of a company's Data Warehouse. The ultimate goal is to reveal intelligence from the data in order to achieve better strategic decisions and improve customer satisfaction (Hassani*et al.*, 2018).

[1]https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

DOI: 10.24818/18423264/54.2.20.05

Recent studies have investigated various challenges in banking area. Churn prediction is a fundamental process in banking field because companies can avoid losing revenues. Churn refers to customers who have an intention to leave the bank, thus the process involves making predictions for clients who have the highest probabilities of closing a relationship with a bank, so it seeks to minimize possible movement of customers from their current bank to another competitor in the market. It is widely known that the cost of acquiring a new customer is between 5 and 25 times higher than retaining existing ones (Gallo, 2014).

With regard to predicting the propensity to churn of banking customers, Sayed *et al.* (2018) conducted a study to compare the performance of two Spark packages (*ML and MLlib*). They assessed the model accuracy dealing with customer transactional data on a dataset from Kaggle. The results showed that the *ML* Apache Spark package has better precision and manages to provide more accurate results than the other *MLlib* package. Frempong and Jayabalan (2017) analyzed a bank direct marketing campaign dataset to predict the customer response to a bank's offer. The outcome of the case study revealed that the best model in predicting the target response variable was the *Random Forest* classifier, which had the greatest predictive power with an accuracy of 87% and an AUC value of 92.7%.

Kumar *et al.* (2019) emphasized that in order to predict the probability to churn for banking customers, they tested two models based on *Decision Tree* classifier and *Artificial Neural Networks*. They claim that Neural Network model had the greatest accuracy of 86.5%, compared to Decision Tree model which obtained only 79.8%.Another study conducted by Wang (2017) focused on comparing the performance of newer machine learning models for predicting bankruptcy. The results show that *Support Vector Machine*, *Neural Network* with dropout and *Autoencoder*out perform older models namely, robust *Logistic Regression*, *inductive learning algorithms* and *genetic algorithms*. The main advantages of using the former methods is the power to obtain the global optimum accuracy, better control of overfitting and higher efficiency on large amounts of data.

In another study conducted by Coşer, Maer-Matei and Albu (2019), a series of machine learning algorithms were applied to estimate the probability of default. The best results were obtained using Random Forest with an AUC value of 89%.

2. Data description and methodology

The research is based on data mining techniques and supervised machine learning used to predict the probability to churn of a customer belonging to a bank. This section aims to present the *dataset* and the main *variables* used in the analysis, continuing with the *summary statistics* and the *methodology* employed for this study.

DOI: 10.24818/18423264/54.2.20.05

This study is based on a series of data from Kaggle[2] regarding the task of predicting which customers of a bank are most likely to churn, also known as *attrition* in banking field. The dataset has 10,000 records and 12 variables. The target variable is *Exited* which refers to customers who closed their relationship with the bank and possibly went to another competitor. The independent variables that were retained in the analysis can be found in Table 1.

Table 1. List of variables from the analyzed database

Variable name	Variable description
CustomerId	Bank customer unique identifier
CreditScore	Credit score of the customer
Geography	Country of residence
Gender	Male or female
Age	Age of the customer
Tenure	Number of years since a customer opened the relationship with the bank
Balance	Balance of products (money deposited in the accounts)
NumOfProducts	Number of products a customer holds
HasCrCard	Binary flag for whether the customer holds a credit card with the bank or not
IsActiveMember	Binary flag for whether the customer is an active member with the bank or not
EstimatedSalary	Estimated salary of the customer in Euros
Exited	Binary variable, 1 if the customer closed the relationship with the bank and 0 if the customer is retained

CustomerId variable is not used because it is only a unique client identifier. The distribution of customers according to geography reveals the following churn rates by country: France **16.15%**, Germany **32.44%** and Spain **16.67%**.

For the purpose of this analysis, we decided to split the data according to the geographic variable, thus we analyze three datasets: customers from *France*(5014 records), customers from *Germany*(2509) and the rest, those from*Spain* (2477). However, the study mainly focuses on customers from France, as it has the lowest churn rate amongst other segments and is similar with the churn rate of Spain. Germany exhibits a churn rate which is twice as the ones for the other countries. Another reason for splitting the dataset was to exclude a heterogeneity from the data with regard to geographical areas. Although, we mainly focus on analyzing the France dataset, we additionally make a comparison with the other data available for Germany and Spain, in order to highlight potential differences in variables distribution, customer profile, as well as changes in model performance.

In table 2 we present the main results concerning France customers for some of the quantitative characteristics.

Table 2. Descriptive statistics on France customers

	Credit Score	Age	Tenure	Balance	Num Of Products	Estimated Salary
mean	649.7	38.5	5.0	62,092.6	1.5	99,899.2

[2]https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

DOI: 10.24818/18423264/54.2.20.05

std	97.0	10.5	2.9	64,133.6	0.6	57,336.3
min	350.0	18.0	0.0	0.0	1.0	90.1
25%	582.0	31.0	2.0	0.0	1.0	51,399.2
50%	653.0	37.0	5.0	62,153.5	1.0	99,132.3
75%	717.0	43.0	7.0	121,444.9	2.0	149,295.4
95%	812.0	60.0	9.0	161,081.0	2.0	190,014.6
99%	850.0	72.0	10.0	186,531.2	3.0	198,052.9
max	850.0	92.0	10.0	238,387.6	4.0	199,929.2

As a data preparation step, *Gender* variable is transformed into a binary one so that it can be used in the model development stage. The new one-hot encoded variable is called *Gender_2* which has a value of 1 for *male* and 0 for *female*. A series of summary descriptive statistics are calculated for the whole dataset, on each country, as well as on each split of the target variable: the customers who churned (*Exited = 1*) and the customers who did not churn (*Exited = 0*).

One can easily notice in table 2 that the average age of the French customers is 38.5 years. However, after splitting according to the target variable, customers who churn have an average age of 45.1 years, compared to an average of just 31 years for those who do not churn. In addition, the bank might lose more valuable customers because those who exited have an average balance of 71,192.8 Euros and an average estimated salary of 103,493.3 Euros, whereas customers who did not leave the bank have a lower average balance of only 60,339.3 Euros and a lower estimated salary of 99,217.1 Euros.

According to Song and Lu (2015), the *classification tree* methodology involves the use of a machine-learning algorithm based on a classification system for predicting a target variable, while minimizing the error rate at each step. The purpose is to classify the analyzed population considering rules to divide nodes at each level; the classification tree starts from the root node and ends with the leaf nodes, where no further optimal splits can be made. The algorithm uses a non-parametric technique and is efficient on large datasets, without the need for a complex parametric structure. Also, if the dataset has many observations, it can be divided into training and validation sets to decide the appropriate size of building or pruning the tree, until the optimal level is reached with the highest accuracy of classifying objects in classes, according to the studied event(*target / non-target*).

Random Forest is an advanced supervised machine learning algorithm which is based on a collection of classification trees. Several trees are built and different predictions are combined through ensemble methods, so that the final prediction is obtained through a majority vote which is more accurate.

Logistic Regression is used to model the relationship between a dependent binary variable and one or several independent variables. This generates the standard coefficients and errors for the significance levels of a formula used to predict a *logit* transformation of the probability of occurrence of the event (Sperandei, 2014). The classifier models the logarithm of the *odds ratio* of a result based on independent variables. Finally, the model predicts a probability of

DOI: 10.24818/18423264/54.2.20.05

occurrence with values in range(0, 1) for each observation. Zhao (2015) explains how logistic regression is used to predict the probability of an event by estimating a distribution with a logistic curve. The model can be expressed mathematically by the following equation:

$$logit(y) = c_0 + c_1x_1 + c_2x_2 + \cdots + c_kx_k$$

where $x_1, x_2, \dots, x_k$ are independent variables and y is the probability of the event, while $logit(y) = \ln\left(\frac{y}{1-y}\right)$. Thus, the previous equation can be written as:

$$y = \frac{1}{1 + e^{-(c_0+c_1x_1+c_2x_2+\cdots+c_kx_k)}}.$$

Regularization in supervised learning is used to create less complex models when the dataset has a large number of features and can deal with over-fitting issues. These methods are *L1 regularization technique*, also called **Lasso Regression** and *L2*, which is called **Ridge Regression**. The methods are distinguished by the fact that the former uses the *"absolute value of magnitude"* of coefficient as penalty term of the loss function, while the latter uses the *"squared magnitude"* of coefficient. The principles are explained using the equations below.

Lasso *regularization cost function:*

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Ridge *regularization cost function:*

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

3. Results

Our aim is to apply an exploratory data analysis that includes evaluating the churn rate amongst the categorical variables and the assessment of major differences in the occurrence of attrition event. Afterwards, we examine the kernel densities plots of the main characteristics such as *age, balance, credit score* and *estimated salary* providing a series of boxplots so that we can recognize the variations in customer behavior for a couple of variables, according to the target variable.

Next step of the analysis consists in implementing a series of classifiers using *machine learning* algorithms and use Python to apply an optimization method called *Grid Search*, aiming to obtain an optimal predictive model. The evaluation metric used to compare the model results is AUC. We performed the above-mentioned methods for all the three datasets.

This section starts with the main insights which reveal several clues about the customer profile and the factors that might influence the propensity to churn. The distribution of France customers who churn, by *gender* variable in figure 1(a) shows the occurrence of churn events for women is **20.34%**, which is significantly higher than the average churn rate of the entire base of **16.2%**. In contrast, men have a much lower churn rate of only **12.71%**.

DOI: 10.24818/18423264/54.2.20.05

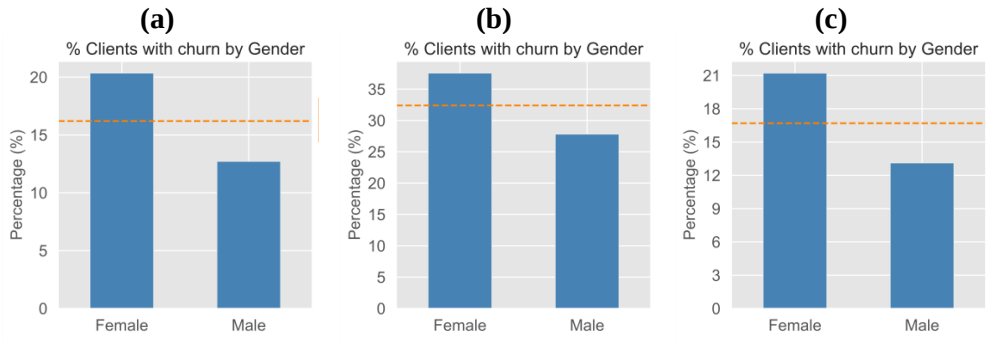

Figure 1. Distribution of customers who churn by *Gender*and *Geography* variables: France (a), Germany (b) and Spain (c)

Source: our own computation

Comparative analysis of churn rates by gender in figures 1 (b) and (c)shows that, for all other countries, women drop out banking services to a greater extent than men. Germany again presents double churn rates for both genders.

Consequently, from the *active memberindicator* perspective -Figure 2 for France consumers, inactive customers (*0 label on x-axis*) have a higher attrition rate of **21.13%**, whereas active clients have a lower probability to churn of just **11.5%**, a common sense that was expected to be found, since customers who are actively involved with a bank and have open products are most likely to remain loyal to their bank. In addition, same findings apply also for the other countries which have likewise distributions for activity.

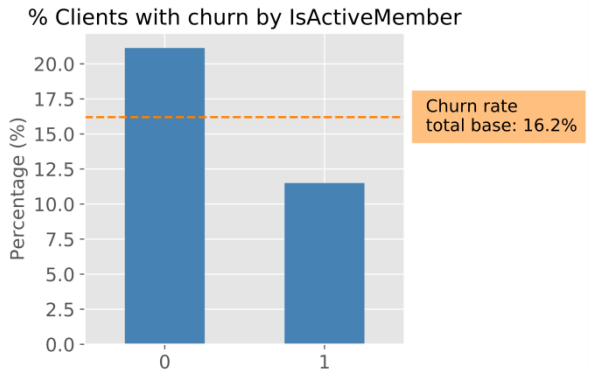

Figure 2. Distribution of customers who churn by *Active Member* variable for France

Source: our own computation

One can notice in figure 3 that *age* distribution of France customers who churned means that customers with a higher age have a greater likelihood of closing the relationship with a bank and consider other competitors' offers, so we expect to find age as a strong discriminator. Similarly, age distribution is analogous to the other two data subsets for Germany and Spain, respectively, which still shows that clients with a higher age, who are more experienced with banking products, are more likely to churn in a future period.

DOI: 10.24818/18423264/54.2.20.05

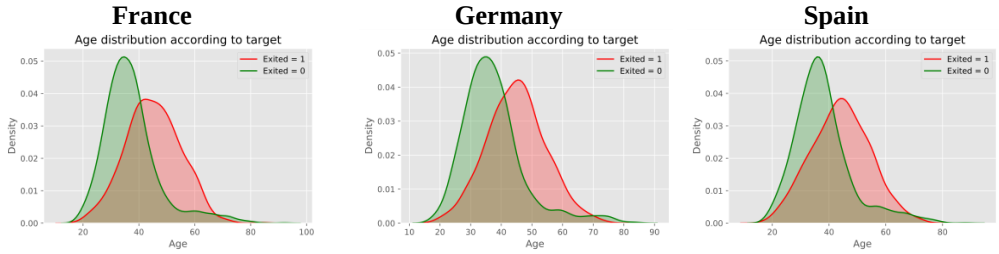

Figure 3. Comparison of Age distribution for all countries according to target variable

Source: our own computation

The *balance* variable for French consumers (Figure 4) displays a *bimodal distribution*: the left side of the plot reveals that there is a reduced occurrence of clients who exited with zero or negative balances in their accounts, compared to others, while more valuable customers exited and closed their accounts with a slightly larger wealth than the others (the right tail of he distribution). This can be a warning indicator because the company could lose revenues and also valuable customers if it does not act on time in the future. Therefore, understanding the drivers that can explain the customer behavior can bring numerous benefits to the firm on the long term to secure its profits, knowing that it is more difficult to acquire new clients than retaining them.

Another thing that we might also highlight from figure 4 is the presence of slightly larger negative balances for customers who exited, compared to the others who did not. This means that clients could have a higher degree of indebtedness, since negative balances are usually related to overdraft products that customers use when they run out of money in their current accounts. This allows them to still use their debit cards for payments, in exchange for an interest paid to the bank.

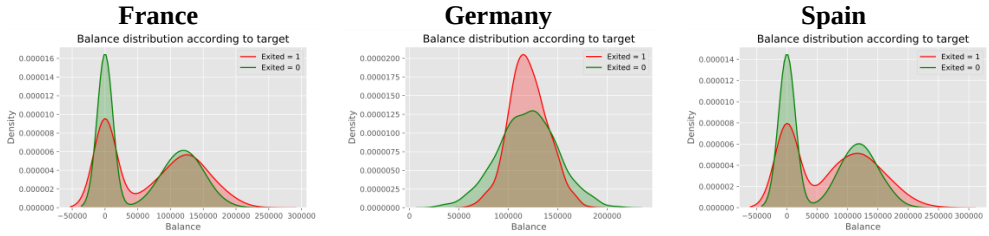

Figure 4. Comparison of Balance distribution for all countries

Source: our own computation

One intriguing aspect one can notice from figure 4is that Germany sets apart from the other countries in terms of *balance* distribution, which, unlike France and Spain, does not show any negative values, while there is a small incidence of zero balances. This could mean that German customers are more careful regarding their budget and personal finances considering they may possibly want to save more money for unexpected expenses. This behavior could relieve them of long-term financial problems without the need of borrowing money. They might also be more financially independent than customers from France and Spain.

Moreover, analyzing the distribution of German customers who exited, we can reason that there are more people who have balances near the central tendency

DOI: 10.24818/18423264/54.2.20.05

indicators, mean and median, the distribution being more homogeneous compared to the other consumers who didn't leave the bank. This behavior clearly distinguishes the German customer profile compared to people in the other countries.

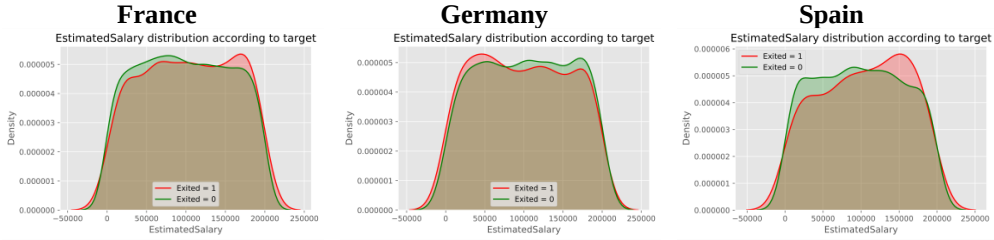

Figure 5. Comparison of Estimated Salary distribution for all countries

Source: our own computation

Considering the *estimated salary* variable, the distribution of France, Germany and Spain clients in figure 5does not emphasize any clear inconsistency between the two classes. However, in figure 6 one can notice a distribution slightly shifted to the left for French customers who churn, while having a marginally lower *credit score*. This suggests that they might be, to a certain extent, moderately riskier clients from creditworthiness perspective, compared to the other segment of people who did not churn. In addition, from figure 6 we can immediately notice that there is no clear differentiation between Credit Score distribution regarding all three countries.

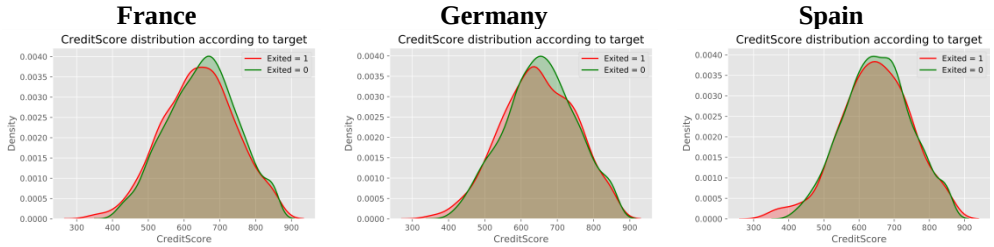

Figure 6. Comparison of Credit Score distribution for all countries

Source: our own computation

Taking into account the distribution of churn rate by the *tenure* variable in figure 7, which refers to the number of years since a customer first opened the relationship with the bank, we can observe distinct patterns among countries. Concerning customers from **France**, people with tenure between 5 and 8 years are less likely to churn, because the related segments in the bar chart plot have values below the dotted orange horizontal line, which denotes the churn rate value for the entire analyzed subset. However, although clients with 7 and 8 years of tenure have the lowest churn rate, they present the *highest risk* of leaving the bank, because after two and three years, respectively, they will reach a tenure level of 10 years, a segment which has the *greatest churn rate* of **19.8%,** much higher than the overall churn rate of **16.2%** for the entire subset with French customers.

DOI: 10.24818/18423264/54.2.20.05

Spain behavior is different to France because the segments with lowest value of seniority (*tenure*) have the highest churn rates, while customers from upper segments have considerably lower attrition rates, customers with 10 years of tenure reaching the lowest rate of only **9.2%**, compared to **16.7%** for the entire subset. The most risky customers are the newest ones, who just opened their relationship with the bank and have less than a year of tenure, with a churn rate of **23.3%**.

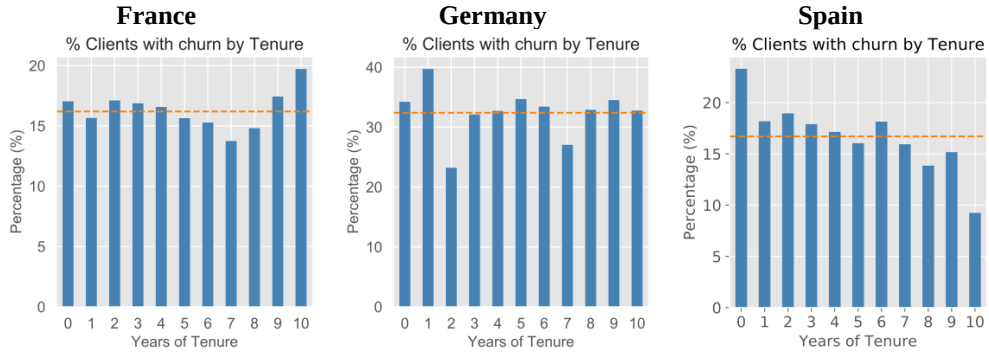

Figure 7. Comparison of Tenure distribution by churn rate for all countries

Source: our own computation

Additionally, the **German** customers with tenure of *one year* have the highest churn rate (**39.8%**), compared to the overall rate for Germany of **32.4%**. Surprisingly, customers with only *two years* of tenure have the *lowest churn* rate of only **23.3%**, and at a short distance, clients with 7 years of seniority have the second churn rate of **27.1%.**

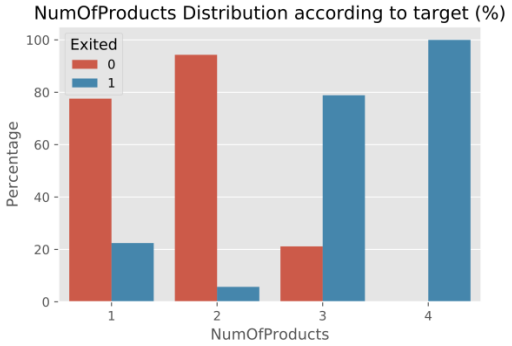

Figure 8. Number of customer's products by target variable for France

Source: our own computation

We can infer that French clients who stick to the bank have fewer products than those who churn, especially when looking at the distribution of customers who have **2 products**, which shows a clear evidence of the primary relationship with the bank. On the contrary, those who have more than 2 products have a higher probability of attrition than people who remain loyal to the bank. This might happen because of large fees and commissions for additional banking products or

DOI: 10.24818/18423264/54.2.20.05

services, which could irritate to a certain degree the unsatisfied clients, who might ultimately close their relationship and turn to more advantageous offers from other competitors.

We decided to check the distribution of the continuous variables in the form of a series of *boxplots* (Figure 9), with splits by target variable, in order to identify particular differences in the profile of French customers from the two classes. The *credit score* series has a lower median value for customers who churned, which presents a series of outlier values. Moreover, from *tenure* perspective, clients who exited have similar levels for the median value; however, the distribution is more widespread than that of the customers who did not close their relationship with the bank, exceeding the boundaries of first and third quartiles.

Regarding *balance*, the distribution of the French people who churned is slightly shifted to the right and has a median equal to 80,376.5 euros, which is 61.2% greater than the median of the other segment (49,853.6 euros). The fact that the *estimated salary* is, to some extent, higher for attrition cases, it reiterates the idea that the bank lost more valuable customers, correlating it with higher balances in their accounts than those who were retained. In contrast, the biggest gap comes from *age* which reveals a median which is 25% higher for clients who churned (45 years), compared to the others (36 years).

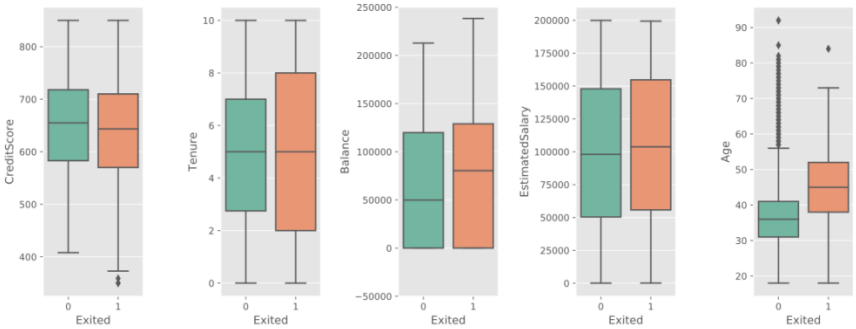

Figure 9. Boxplot diagrams and main characteristics according to target variable for France

Source: our own computation

In order to predict which customers have the highest propensity to churn, we implement two machine learning algorithms based on *Random Forest* and *Logistic Regression* classifiers. The goal is to obtain the optimal model with the highest accuracy in predicting the event of bank attrition, by comparing the most important performance metric called **AUCscore**, which stands for the value of *Area Under Curve* for *Receiver Operating Characteristic* (*ROC*).

The data science task for model building was implemented using the Python programming language, which has an optimization method called **Grid Search**. This generates multiple combinations of models to be fitted, based on an input of parameters carefully chosen by the analyst and provides an output with the

DOI: 10.24818/18423264/54.2.20.05

best recommended model by evaluating a performance metric which is selected before building the models. *Grid Search* uses a cross-validation technique by which the machine learning algorithms can be evaluated on test subsets, in order to decide which the optimal model with the highest stability is, meaning that it leads to comparable results on unseen data.

Random Forest

The following combinations of parameters were tested using **Grid Search**, with 4 alternatives for the *number of estimators* (classification trees) and 6 choices for the *maximum depth* of growing the trees (levels). For each scenario, a *cross-validation* is performed using 5 folds, which means that in total 120 models were trained for the *Random Forest* classifier.The performance was evaluated on both train and test datasets, by means of **AUC values**. We managed to find the **optimal combinations** of *model parameters*, choosing the one which reduces the *over-fitting* behavior. The AUC results are provided in figure 10 below through heatmap graphs developed in Python.

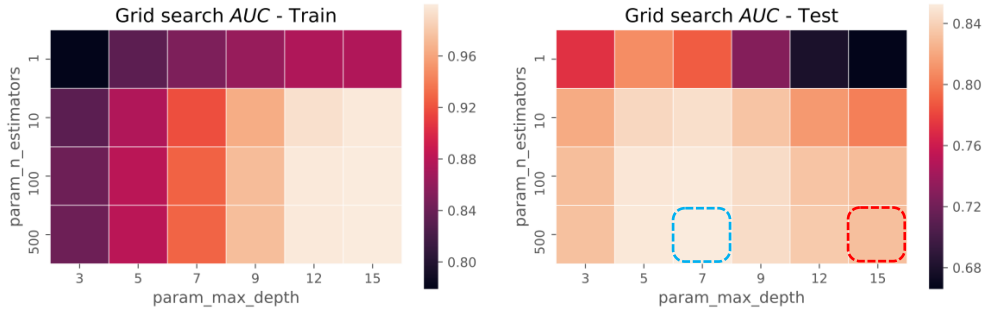

Figure 10. Grid Search heatmap for Random Forest classifier with AUC values on Train and Test datasets for France

Source: our own computation

We can imply the performance of the Random Forest machine learning algorithm on the training dataset is getting better and better as the value of the *maximum depth parameter* increases, reaching a maximum score for a depth of **15levels** and a number of **500estimators** used to ensemble all the models and improve the overall prediction. However, when looking at the performance of the model on the test dataset (*highlighted with red dashed line*), we can understand that it doesn't perform as well; actually the outcome is worse in terms of AUC value when dealing with new data, a behavior that we cannot accept because the model would give poor predictions results when scoring unseen data.

Therefore, we decided to consider the **optimal model** as the one found at the intersection of the following two parameters (*highlighted in blue dashed line*): number of **estimators** equal to **500** and maximum depth of **7 levels** in order to keep a balance between *accuracy* and *stability* of the model in time, so that it can provide consistent results of probability to churn when scoring new customer data.

Consequently, the over-fitting issue is overcome because we choose a model with a high *number of estimators*, which means that many individual classification trees models are developed and the final prediction is averaged by the

DOI: 10.24818/18423264/54.2.20.05

majority votes, resulting in a final predicted class with an increased accuracy. Moreover, the number of *maximum depth levels* trees can grow was chosen in such a way to also deal with over-fitting, because a small number of depth can lead to under-fitting, while a higher depth can make the trees grow too large by capturing every noise from data.

Finally, our decision is also validated with *Grid Search* best model results, which also recommends the same combination of the optimal parameters, when evaluating the model performance metric using 5 folds **cross-validation** technique. Nonetheless, we also wanted to understand how the optimal model performs in comparison with a default Random Forest model, without any modification of its parameters and without performing Grid Search. This is used in the final research stage as a benchmark to compare model performance improvement over the non-optimized standard classifier.

The same Random Forest classifier applied on *Spain* subset reveals the identical choice of the optimal model found by Grid Search method, with the matching number of **500 estimators** and a maximum depth of **7 levels**. On the other hand, the model applied for *Germany* consumers reaches the top performance on the test dataset for a lower depth of only **5 levels**, with the same number of **500 estimators**, a fact clearly noticed in figure 11. Future comparison regarding the actual AUC values on the test sets between all the three countries are presented in the final part of this section.

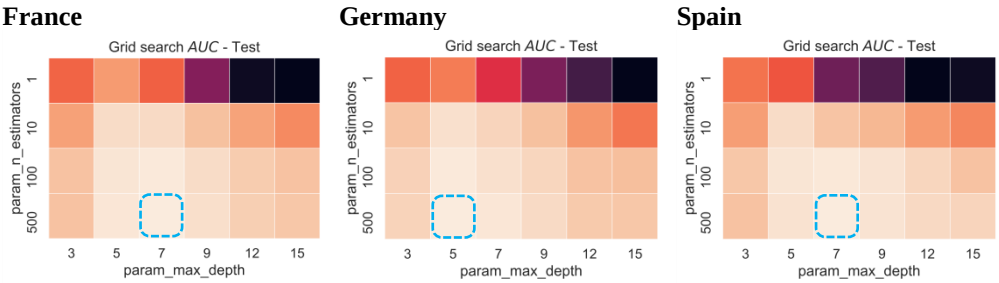

Figure 11. Heatmap with AUC values comparison for all countries on test dataset

Source: our own computation

Logistic Regression

Furthermore, a *Logistic Regression* classifier was also implemented and the best model parameters were obtained through *Grid Search*. The only two possible tested combinations are *L1* (Lasso) and *L2* (Ridge) *regularizations*, with *penalty parameter C* values for which a logarithm in base 10 returns 15 numbers equally spaced on a log scale for a range between $(-4, 0.2)$. The performance was evaluated again on the same train and test datasets using **AUC values**; the **optimal combinations** of *model parameters* are in figure 12.The two heatmaps reveal a much more stable prediction behavior by applying the *Logistic Regression* model, which shows similar conduct between train and test datasets. We can observe that best classification results are obtained using the **L1** (*Lasso*)regularization

DOI: 10.24818/18423264/54.2.20.05

technique, because the other **L2** (*Ridge*) regularization underperforms. In addition, starting with a penalty coefficient of 0.1, the AUC values start to stabilize to the right side reaching a value around **0.73**, until the end of the input parameter range.

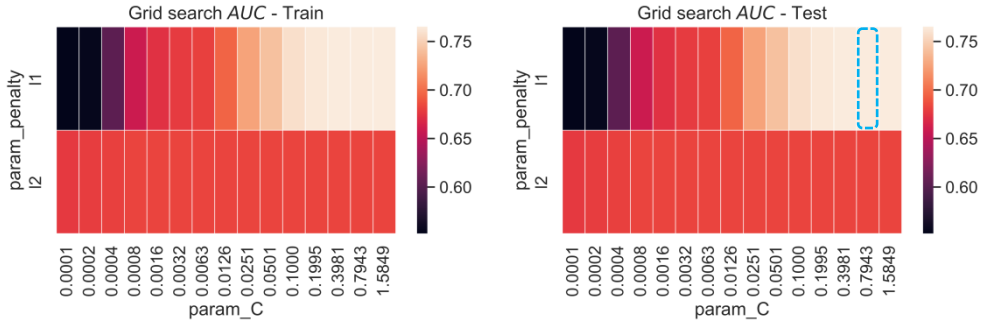

Figure 12. Grid Search heatmap for Logistic Regression classifier with AUC values on Train and Test datasets for France

Source: our own computation

Consequently, the **optimal model** has **L1** *Lasso regularization* with a *penalty coefficient* of **0.7943**, highlighted with the blue dashed line on the test dataset heatmap located in figure 12.Comparing the outcomes obtained on Germany and Spain datasets, the results of Logistic Regression expose identical choice of optimal parameters found by Grid Search method, with **L1** *Lasso regularization* and a *penalty coefficient* equal to **1.5849**. The same behavior of L1 regularization being significantly more accurate than L2 (Ridge) remains also valid, which can be visually inspected in figure 13 with AUC values heatmap applied on test dataset.

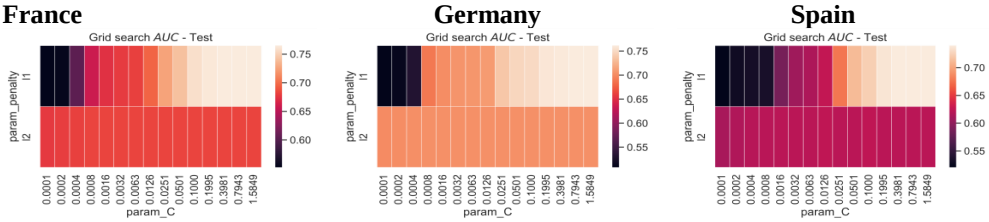

Figure 13. Heatmap with AUC values comparison for all countries on test dataset

Source: our own computation

Finally, the scores are compared with Random Forest to make a decisive call for choosing the best model in predicting the propensity to churn by means of AUC values. Considering the aforementioned aspects regarding the standard classifiers, in figure 14 we can observe the comparison of AUC values for *baseline* benchmark models (plotted in less intense colors), as against best ones obtained through *Grid Search* (represented in brighter colors). Consequently, *Logistic Regression* optimal model for France has an AUC value of **0.73**, a significant improvement over the **standard non-optimized** Logistic Regression model which has an AUC of **0.63**.

In contrast, the best *Random Forest* model for French clients got an AUC of **0.85**, while the baseline model for the same classifier along with default

DOI: 10.24818/18423264/54.2.20.05

parameters obtained an AUC of **0.82**. Although this may not represent a considerable increase, it could have a major impact compared to a weaker model, thus capturing the churn event more effectively. In the end, the optimal model can bring a significant increase in the company's revenues or greatly prevent possible substantial losses.

Therefore, we choose the **final optimal model** as the one obtained through **Random Forest classifier** with *Grid Search optimization*. Such a model could be implemented into production as an early warning system, with the ultimate goal of reducing the significant risk of attrition.

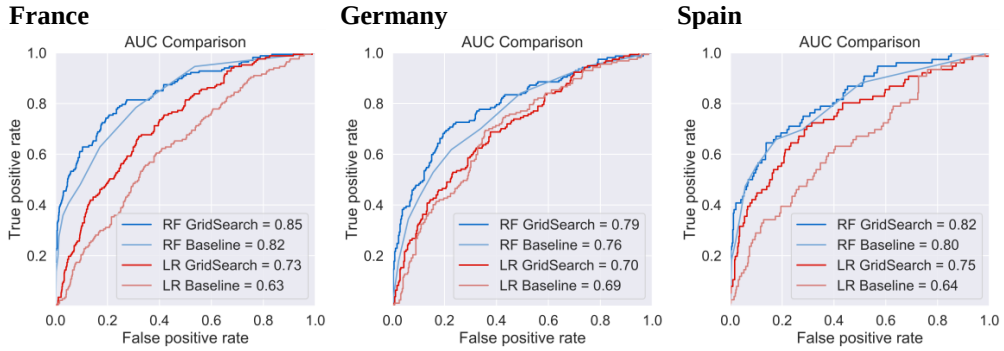

Figure 14. Model comparison with AUC values between baseline and the optimal classifier obtained through Grid Search for all countries on test dataset

Source: our own computation

In figure 14 we immediately recognize substantial differences in model performance when evaluating on test dataset, considering changes in AUC scores from one country to another. The best models are obtained when applied to **France** customer data, Random Forest grid search having the highest value of **0.85**, closely followed by **0.82** on Spain data, and at a greater distance, **0.79** found for Germany records. An interesting aspect resides from the fact that on Germany data, the baseline Logistic Regression model performed best in class with an AUC of **0.69**, compared to **0.63** for France and **0.64** for Spain. However, for Germany, after applying the Grid Search optimization technique, the final AUC value is **0.7**, the lowest of them all, France having a value of **0.73**, while Spain recorded**0.75**.

The most important features of the *Random Forest* model which influence the propensity to churn emerges from figure 15:**age** plays the greatest role in estimating the probability of attrition, which we also found during the data analysis by looking at the kernel densities plots in relation to the target variable, leading to the conclusion that *age* is, indeed, a powerful discriminator.

DOI: 10.24818/18423264/54.2.20.05

Figure 15. Feature importance comparison of Random Forest models between all countries

Source: our own computation

Another explanation might arise from the fact that older clients are more experienced when dealing with banking products, thus knowing how to better defend their customer rights and in the end, if they are not fully satisfied with the current financial services, they could immediately close their accounts and leave the bank. This top characteristic is closely followed by the **number of products** a customer holds. Nevertheless, **member activity**, **credit score** and **estimated salary** impact to a lesser extent the likelihood of churn, while the last variables have a marginal effect in predicting churn event. Considering the top characteristics that were also discovered for the best Random Forest applied on Germany and Spain data, we can examine the variables in decreasing order of importance, represented in figure 15.

The only variation in variable importance appears for the model employed on Germany client base, with changes in top features positions, the *activity indicator* receiving a higher importance than for France and Spain models, and at the same time, the balance is considered slightly less important, being downgraded with one level. In addition, the estimated salary is considered to have a higher impact in predictions, being upgraded in front of credit score, compared to the models implemented for both France and Spain. In the end, the rest of the variables remain unchanged in the top feature importance between all three countries. All in all, this change in behavior shows a clear differentiation of the profile of consumers in Germany with respect to the other two countries.

4. Discussion

Other studies focused on the same issue of predicting the probability to churn in banking using the same dataset from Kaggle. For example, Chowdhury (2019) conducted an analysis on Medium blog by applying a series of machine learning algorithms to predict attrition through a series of classifiers such as *Logistic Regression*, *Decision Tree*, *Random Forest*, *Support Vector Machine*, *Gradient Boosting* and *AdaBoost*. The highest accuracy (86.35%) was obtained using Gradient Boosting and AdaBoost. However, in our opinion, we argue that the author overlooked the performance assessment of the models without measuring the AUC value, which is crucial when evaluating a classification problem. The only performance metric that the author presented in his study is the *accuracy*, which has the main disadvantage of being extremely sensitive to imbalanced datasets. As such, for this analyzed dataset, because the overall churn rate was 20.4%, if a naive classifier would have been implemented, which always predicts

DOI: 10.24818/18423264/54.2.20.05

the majority class of people who didn't churn, then it would still have an accuracy of 79.6%, which is highly misleading.

Additionally, another research was performed by Malik (2019) from VSH Solutions blog, who also analyzes the same dataset. The researcher claims that also Random Forest obtained the best performance with an accuracy of 86.15%, while Logistic Regression obtained a value of only 78.5%, results which also confirm the study employed by us where we obtained likewise Random Forest as the optimal model. Even though, this time also, the only evaluation metric is still accuracy.

Therefore, we strongly suggest that the rigorously analysis that we conducted in this research is a better way of evaluating the real performance of a model on unseen data by evaluating the AUC values on test datasets, a metric which takes into account both the true positive rate, as well as false positive rate, with the goal of finding the optimal model which has the lowest misclassification rate. Also, we can argue that although the data analyzed in this study have been used in other researches, no such comparison has been previously made between the performance of the models for each country, by highlighting the differences and particularities of each subset of the original data, as the one presented in the research employed by us.

5. Conclusions

In this paper we analyzed the main characteristics of customers that influence the propensity to churn, revealing that *age* is a powerful discriminator combined with the *number of products* a client holds, which explains the existence or absence of a primary relationship with the bank. Several models were tested through *Logistic Regression* and *Random Forest* machine learning algorithms. In addition, we obtained improved model performance results through Grid Search optimization technique, compared to the same type of classifiers implemented only with default parameters (*baseline models*).

The other factors of influence on probability of attrition are *activity* indicator, *credit score* and *estimated salary*, but they impact the likelihood of a customer leaving the bank to a lesser extent than the main characteristics.

We found that *Random Forest* classifier offered the best results in terms of AUC with a value of 0.85 for France, which was validated on the test dataset. The outcome is considerable higher than the one obtained through the same Grid Search technique, but applied using the *Logistic Regression* model, which recorded a maximum AUC value of only 0.73 on the same test dataset.The separate comparative analysis on the three countries highlighted different aspects of the influence of the explanatory variables on the decision to churn, as well as different customer behavior between distinct geographical areas.

The limits of our research mainly ensue from the fact that we only took into consideration two machine learning algorithms. Future research might imply testing different classifiers, especially more advanced techniques based on *Gradient Boosting* to assess any improvements on the accuracy of predictions on the same analyzed dataset.

DOI: 10.24818/18423264/54.2.20.05

REFERENCES

[1] **Aldea, A., Maer-Matei, M. (2019),** ***Data Analysis Course Notes*****.** Master of Database for Business Support;The Bucharest University of Economic Studies;
[2] **Chowdhury, M. (2019),** ***Churn Analytics: Data Analysis to Machine Learning,***available at:*https://medium.com/@mchowdhuryca/churn-analytics-from-data-analysis-to-machine-learning-95854d102ed6*;
[3] **Coşer, A., Maer-Matei, M., Albu, C. (2019),** ***Predictive Models for Loan Default Risk Assessment.*** *Economic Computation and Economic Cybernetics Studies and Research, ASE Publishing;*53(2): 149-165;
[4] **Frempong, J., Jayabalan, M. (2017),*****Predicting Customer Response to Bank Direct Telemarketing Campaign.***The International Conference on Engineering Technologies and Technopreneurship, IEEE;
[5] **Gallo, A. (2014),** ***The Value of Keeping the Right Customers***.Harvard Business Review, available at*: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers*;
[6] **Kaggle Bank Churn Dataset**, available at: *https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling*;
[7] **Kumar, G., Tirupathaiah, K., Reddy, B. (2019),** ***Client Churn Prediction of Banking and Fund Industry Utilizing Machine Learning Techniques***. International Journal of Computer Sciences and Engineering, 7(6): 842-846;
[8] **Hassani, H., Huang, X., Silva, E. (2018),** ***Digitalisation and Big Data Mining in Banking.***Big Data and cognitive computing, 2(18);
[9] **Malik, U. (2019),** ***Predicting Customer Churn Using Machine Learning Models,*** available at: *https://www.vshsolutions.com/blogs/predicting-customer-churn-using-machine-learning-models/*;
[10] **Nagpal, A. (2017),** ***L1 and L2 Regularization Methods,*** available at: *https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c*;
[11] **Python Software Foundation,** ***Python Language Reference***, version 3.7, available at:*http://www.python.org*;
[12] **Sayed, H., Fattah, M., Kholief, S. (2018),** ***Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib*** **Packages:** ***A Comparative Study*****.** International Journal of Advanced Computer Science and Applications, 9(11): 674-677;
[13] **Song, Y., Lu, Y. (2015),** ***Decision Tree Methods: Applications for Classification and Prediction***.Shanghai Archives of Psychiatry, 27(2): 130-135, available at:*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/*;
[14] **Sperandei, S. (2014),** ***Understanding Logistic Regression Analysis.*** Biochemia Medica (Zagreb), 24(1): 12-18;
[15] **Wang, N. (2017),** ***Bankruptcy Prediction Using Machine Learning;*** Journal of Mathematical Finance, 7: 908-918;
[16] **Zhao, Y. (2015),** ***R and Data Mining: Examples and Case Studies.*** Elsevier, p.57.

DOI: 10.24818/18423264/54.2.20.05