CODER HOUSE

BANK CHURN PREDICTION

Testing classification models

March,
2022

TABLE OF CONTENT

BANK CHURN PREDICTION

Presentation

My name is Carolina Swoboda, I am 37 years old and I am from Córdoba, Argentina. I have a Master of Business Administration (MBA) and a Bachelor of Economics.
Regarding the choice of dataset, it took me many weeks of searching and reviewing public databases. After having looked at several web pages, and decided on a particular topic on which I wanted to work with, I focused the search a little better and ended up choosing this dataset, "Churn for Bank Customers". The dataset, which I obtained from Kaggle (License CC0: Public Domain), has 10,000 records and 14 columns, and each record represents a different client.

Objectives

There are two objectives that I aim to achieve:

1. on one hand, being able to identify the factors that make a bank customer churn (or not),
2. and on the other hand, to find the best model that predicts which customers are likely to churn.

Consequently, in this work I will focus on:

1. the interpretation of the data, in order to find the largest number of insights that explain the reason why a client may or may not leave the bank. For this reason, in the first part of the work I will make an exhaustive exploratory data analysis of each one of the variables and the relation between them.
2. the prediction, to estimate which customers are going to leave and which are not.

Introduction

In this dataset there is a target variable, "Exited", which when it's equal to 0, it indicates that the client did not churn and when it's equal to 1, it indicates that he left the company (he or she churned). The rest of the variables give information about the client (country of residence, gender, age, estimated salary, etc).
It is a supervised model (the target variable is known) and it is of the classification type.

Some insights

From the exploratory analysis, this are the main insights I got:

- No column with date data is indicated in the dataset, so it cannot be concluded that the data is from a specific day or the average of a month, or some other date. As a consequence, for example, as the balance is for a given date, that gives a lot of questions: 1. what date is it, 2. of what relevance is this date, 3. is it an average of a month or an specific date.
- There are clients who have churned, but still have money deposited in the account. That could probably mean that they are going to retire that amount of money (185 millions, that represents almost the 25% of all the money deposited in the bank).
- The same thing happens with the number of products: there are customers who still have products in the bank, but who are no longer customers. Will be deactivated soon?
- About 20% of the customers have churned. As the target feature is 'Exited', the baseline model could be to predict that 20% of the customers will churn. Given that 20% is a small number (it's an unbalanced dataset), we need to ensure that the chosen model does predict with great accuracy this 20% as it is of interest to the bank to identify and keep this bunch as opposed to accurately predicting the customers that are retained.
- 75% of the clients are under 44 years old, and most of them are between 30 and 40 years old.
- Almost 35% of the customers have no money in their account.
- Most of the customers have at least one or two products and very few customers have more than two products.
- There is no strong correlation between the variables (except for balance and number of products, that have a negative correlation (-0.3)).
- Age is the variable that has the highest correlation with the target variable. This correlation is positive, which indicates that when age increases, the churn rate also increases.
- The next features with a relatively high correlation with 'Exited' are IsActiveMember and Balance. The first has a negative correlation, that is, the more active the customer, the less churn rate. The second one has a positive correlation, which indicates that the higher the amount of money in the account, the higher the churn rate.
- The variables 'Tenure', 'HasCrCard' and 'EstimatedSalary' could be removed from the model, as they have no correlation with the target.
- In terms of their credit score, there is no big difference between the clients who churned and those who didn't.

- Older customers have a higher churn rate. In this sense, and given that it is the feature with the highest correlation, the bank should take action in this regard: that is, carry out customer retention campaigns according to age groups, focusing especially on those over 40 years of age.
- Regarding the years of permanence, although the mean and median of the clients who stay and those who leave is the same (5 years in both cases), the clients who churn, do so or more at the beginning, or rather after the 7 years passed. They should also carry out retention campaigns in order to keep their long-standing clients.
- Customers with larger accounts leave the bank the most. This is an important warning for the bank, since the most important clients are leaving
- There are no differences between clients who churn in terms of estimated salary.
- While the largest number of clients come from France, the largest number of customers who churn are from Germany.
- Women churn in a greater proportion than men.
- Whether or not they have a card does not change the proportion of customers who churn (it is 20% in both cases). However, in absolute terms, more customers who have a credit card left the bank.
- A more active member is less likely to churn than a less active one.

Feature engineer

Since there is no strong correlation with the target (in almost any of the features), I created some new features that are likely to have an impact on the probability of churning, given the correlation between them:

- Tenure by age: is higher in those who didn't churn, which is reasonable
- Credit score, given the age: is lower in those who have churned
- Balance, given the number of products: is higher in those who churned. This is a warning.
- Has credit card, given the age
- Estimated salary by age: is higher for those who didn't churn.

As 'CreditScore', 'Tenure', 'HasCrCard' and 'EstimatedSalary' have a very low correlation (almost next to zero) with the target, I'll drop them. Besides, this features are already included in the features mentioned above.

Modeling

For the prediction, I tested several models, and among them, the model with the highest ROC_AUC is the one that will be chosen, so in all cases I will try to maximize that metric.

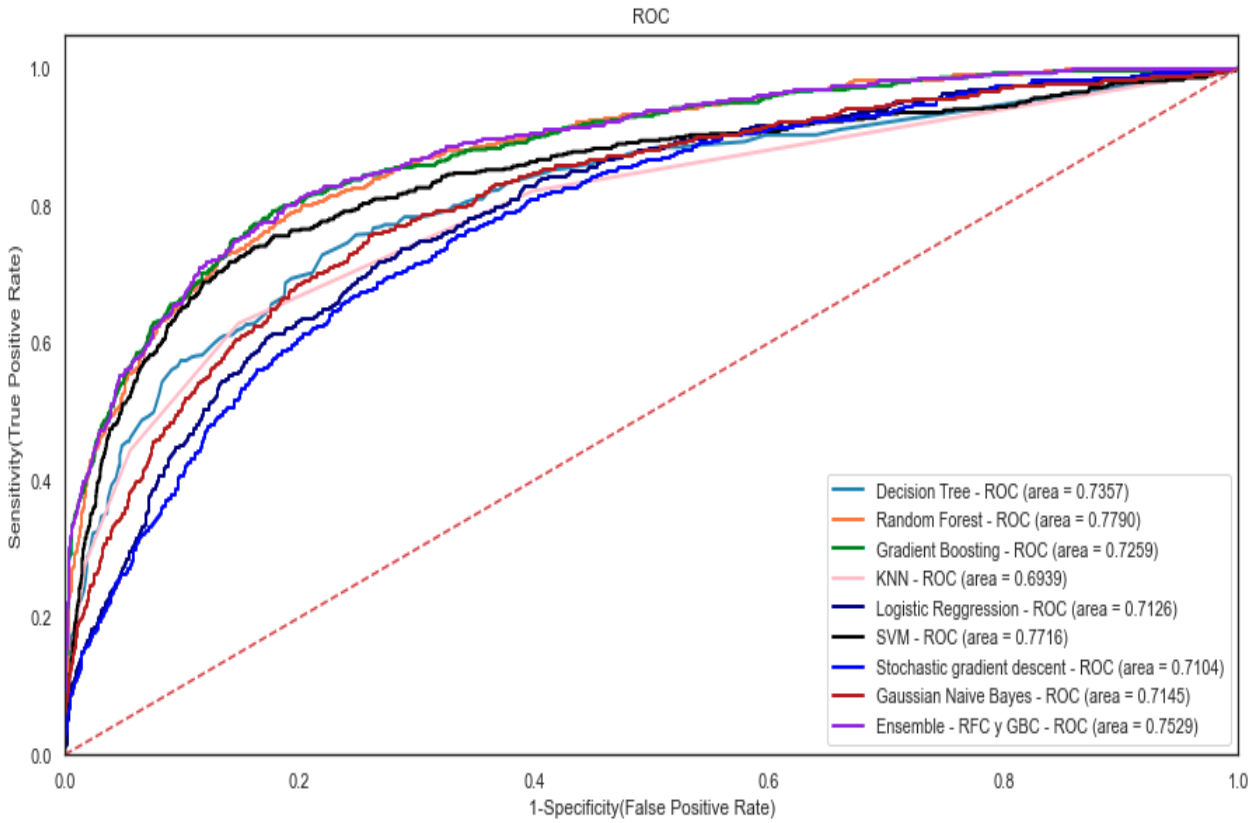
All the models tested where optimized (using GridSearch CV, in order to get better results) and a last model was created, by ensembling the two most performing models (Random Forest and Gradient Boosting). This lead into a model which had the highest ROC_AUC of all.

Overview of the models

This table shows the results of all the models tested, ordered by the highest score.

	Model	Accuracy Test	Accuracy Train	Precision Test	Precision Train	Recall Test	Recall Train	F1 Score Test	F1 Score Train	AUC
8	Ensemble - RFC y GBC	0.8673	0.8814	0.7261	0.7709	0.5597	0.5947	0.6322	0.6714	0.8816
2	Gradient Boosting	0.8690	0.8761	0.7914	0.8239	0.4845	0.4986	0.6010	0.6212	0.8795
1	Random Forest	0.8517	0.8829	0.6305	0.7067	0.6563	0.7265	0.6431	0.7165	0.8773
5	SVM	0.8507	0.8741	0.6321	0.6868	0.6383	0.7027	0.6352	0.6946	0.8418
0	Decision Tree	0.8283	0.8643	0.5784	0.6705	0.5794	0.6564	0.5789	0.6634	0.8137
7	Gaussian Naive Bayes	0.8140	0.7953	0.5431	0.4975	0.5466	0.4881	0.5449	0.4927	0.8109
3	KNN	0.8423	0.8776	0.6708	0.7973	0.4435	0.5351	0.5340	0.6404	0.7956
4	Logistic Reggression	0.7567	0.7430	0.4338	0.4117	0.6383	0.6094	0.5166	0.4914	0.7921
6	Stochastic gradient descent	0.7377	0.7210	0.4109	0.3878	0.6645	0.6388	0.5078	0.4826	0.7785

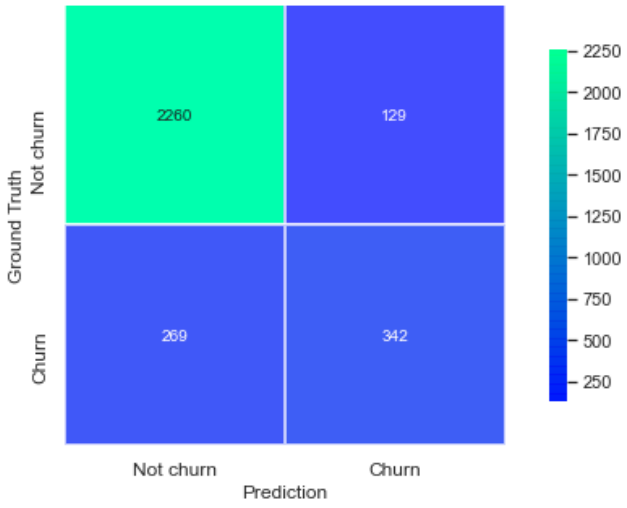
As one can see in the graph below, the ROC curves of the top three models are very similar



Conclusions

The model chosen for the prediction is the one which was ensembled, mixing the best qualities of the Random Forest and Gradient Boosting models. The ROC_AUC score of this ensembled model is 88,16%, with 86.73% of accuracy.

This is the confusion matrix for the chosen model: "Ensemble – RFC and GBC"



The model predicts correctly that 2260 customers will not churn and that 342 will churn, and fails to predict that 269 clients will not churn, when in real life they will, and that 129 will churn, when they will actually not churn.

Recommendations

- The country with the highest churn rate is Germany, almost duplicating the rate of Spain and France.
  This is something to consider, it's important to know why this is happening. Maybe because of the kind of services offering in Germany, perhaps is not as good as in France

- As women churn in a greater proportion than man, some actions must be taken in this area too, in order to let women engage with the bank as well as men.

- There's also recommended to take action regarding the "older" customers, as the are more likely to churn.

- One thing that's really important to take care, is the customers with largest accounts, as they are churning faster than the ones with less money in their account, and this are the most important clients of all.

Glossary

Dataset content

In total there are 10,000 records and 14 columns, and there are no missing values (all fields are complete) and no duplicates.

The features are the following:

- RowNumber: index that comes with the dataset, so it can be deleted. It is of type int64.
- CustomerId: it is an internal number to identify a customer, it does not influence in whether the customer churns or not. It is of type int64.
- Surname: nor does it influence the customer's decision to leave the bank or not. It is of type object.
- CreditScore: is a customer's credit score. It can be used to predict whether a client is going to churn or not, since it is likely that the higher the score, the more intentions they have of continuing to be a customer. This variable is of type int64.
- Geography: where the client lives can influence their decision to remain a client or not. It is of type object.
- Gender: although it shouldn't, the customer's gender can influence their decision to leave the bank. It is of type object.
- Age: this is a relevant attribute, since it is likely that the older the customer, the more reluctant to change then bank. This variable is of type int64.
- Tenure: it is also a relevant feature, since it indicates the number of years that the person has been a bank customer, and in this sense, the more years he or she has been a bank customer, the less likely he or she wants to go to another bank. It is of type int64.
- Balance: it may be another important feature to predict whether or not a customer will churn, since people with more money in the bank account are less likely to change banks. It is of type float64.
- NumOfProducts: it indicates the number of bank products that the customer has. It could also be a good indicator for prediction. It is of type int64.
- HasCrCard: it indicates whether a customer has a credit card or not. This variable is also a good indicator for prediction, since it is less likely that a customer will churn if they have at least one credit card from the bank. It is of type int64.
- IsActiveMember: it may be another relevant feature in the model, since one could say that the most active clients are less likely to leave the bank. It is of type int64.
- EstimatedSalary: it is an important feature when it comes to predicting whether a customer will churn or not, since people with higher salaries are more likely to remain customers. It is of type float64.
- Exited: is the target variable, and indicates whether or not a customer churned. It is of type int64. It is a variable that is also categorical, it can take the value of 1 (the client churned) or 0 (he or she is still a client).