FDS Project 2021/2022
Churn Prediction for Bank Customers

Conti Andrea - 1849300, Cruoglio Antonella - 2025992, Iovino Giuliana - 2017512, Mascolo Davide - 2001991, Napoli Mario - 2015169

26 December 2021

Abstract

This report presents unbalanced binary classification problem using a customer churn dataset on which *Machine Learning* techniques have been applied. In the first part there is an introduction to the problem and its state of the art, then it focuses on the materials and methods used. In particular re-sampling techniques were used to improve the predictive power of the models, considering the unbalancing of this kind of data. The models used are: *Logistic Regression*, *K-Nearest Neighbors*, *Support Vector Machine* and they have been evaluated using *Accuracy*, *F2-Score*, *AUC*, *Sensitivity*, *Specificity.*

1 Introduction

Customer churn is defined as the propensity of customers to cease doing business with a company in a given time period. Considering that acquiring a new customer is more expensive than keeping an existing one, the churn represents one of the principal problems that many companies worldwide are having to face. In particular, banking is one of the highly competitive sectors where customer relations is extremely important. In order to survive in this competitive market, it is strongly needed for commercial banks to improve the capabilities to predict customer churn, taking timely measures to retain customers and preventing other clients from churning. The banks spend a lot of time making statistical predictions that can help make business decisions.Customer churn can be averted by studying the demographic features and history of the customers.

2 Related Works

In pursuance of predicting customer attrition for commercial banks, many scholars carried out the research by using various classification methods. It is interesting how the problem of unbalanced classes has been addressed in [1] using Over-Sampling, Under-Sampling and Synthetic Minority Over-Sampling (SMOTE). We also found interesting the use of *Logistic Regression* in [2]. In [3] the author uses different metric in order to evaluate the models in addiction to *accuracy.* The last article we read, [4], shows the various steps to follow in customer churn classification problems.

3 Proposed Methods

3.1 Re-sampling Methods

The main sampling techniques to conquer the class imbalanced issue are:

- *Over-Sampling* balances the data by replicating the minority class samples. This does not cause any loss of information, but the dataset is prone to over-fitting as the same information is copied;
- *Under-Sampling* adjusts the class distribution of a dataset subsampling the majority class. A limitation of under-sampling is that observations from the majority class are deleted and they could be useful;
- *SMOTE* creates new synthetic samples to balance the dataset so rather than copy an existing sample it creates a new one based on the k-nearest neighbors observations.

3.2 Classification Methods

The classification methods chosen are:

- *Logistic Regression* measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution;
- *K-Nearest Neighbors* finds a predefined number of training samples closest in distance to the new point and predict the label from these;
- *Support Vector Machine* is based on the research of the optimal hyper-plane so that the distance (the margin) from itself to the nearest data point on each side is maximized.

4 Dataset and Benchmark

4.1 Data Acquisition

For our project we have chosen a dataset on Kaggle [5] which contains details of a bank's customers and the target variable is a binary variable reflecting if the customer left the bank or not. The dataset has 10,000 observations and 14 features.

4.2 Exploratory Data Analysis

In exploratory data analysis we observed that the dependent variable *Exited* is strongly unbalanced, in particular about 80% of observations belong to class *0*. In addiction we have seen the behaviour of independent variables with respect to *Exited*[1].

4.3 Data Cleaning and Pre-Processing

From the EDA we observed that there were 3 useless columns: RowNumber, CustomerId and Surname, so we removed them. After that we used One-Hot Encoding for the categorical variables and we standardized the numerical ones in order to put different variables on the same scale.

The dataset has been split randomly in two parts: the 90% form the *Train Set* on which the resampling methods were applied, the remaining 10% constitutes the so called *External Validation Set*.

4.4 Validation Methods

In order to tune the hyperparameters of the models mentioned above and select the best one we have used 10-fold Cross Validation, repeated 20 times. This procedure has been applied over the Train Set and the three re-sampled datasets. The metrics used in CV are Accuracy and F2-score. Over all the cases, the dataset on which the models performed better was the SMOTE one.

5 Experimental Results

The *External Validation Set* was used to evaluate the performance of our models on unseen data. In this case we used Accuracy, F2-score, AUC, Sensitivity and Specificity to assess the models.

The results are shown in the following table:

Table 1: Performance on External Validation Set - SMOTE dataset

	Accuracy	F2-Score	AUC	Sensitivity	Specificity
Logistic Regression	0.776	0.436	0.650	0.425	0.875
K-Nearest Neighbors	0.818	0.645	0.764	0.667	0.860
Support Vector Machine	0.820	0.596	0.740	0.598	0.882

From this table we can observe that the method which performs better is KNN.

[1] You can see the relative plots in the notebook

Although all models have sufficiently large accuracy we note that F2-Score is not as high as we expected (but they are similar to the reference papers ones). Similarly we have an high specificity for all the models but the sensitivity is quite low, considering that we are interested in maximizing True Positive (and minimizing False Negative).

6 Additional Work

Here we can consider achieved the goals we had set for our project, but due to the not very high performances we thought to use an ensemble model taking into account the literature we revised: Random Forest.

Also on this model we did Cross Validation with Hyperparameter Tuning on the four datasets and we discovered that this model performs better on the oversampled one. After that, we applied our "best performing model" on the *External Validation Set* and it returned low performances. So, we tried to apply the models trained on other datasets on the same validation set and we obtained the best performance from the model trained on undersampled data.

We could assume that the models trained on oversampled and SMOTE data are overfitted and they are unable to generalize on unseen data. Another reason may be too many decision trees.

The best performance is obtained training the model on undersampled data. We could also assume that a lower number of decision trees could avoid the overfitting so that the model is able to generalize.

Table 2: Performance on External Validation Set - Undersampled dataset

	Accuracy	F2-Score	AUC	Sensitivity	Specificity
Random Forest	0.724	0.667	0.754	0.804	0.705

7 Conclusion

As you can see in Table 2 the Accuracy is lower than the ones in Table 1 but this metric in unbalanced classification problems is not informative due to *accuracy paradox*. With Random Forest we have marked improvement on sensitivity which is the reference metric in this problem because our goal is to minimize false negative.

Furthermore if we want to consider performances more, we can use RF, instead if we care about the interpretability of models is better to use simpler models like the ones chosen at the beginning of this work (the ones in Table 1) .

8 Future Work

This work does not aim to be an univocal solution to the problem of binary classification, but proposes a number of useful approaches to countering this problem. However, other techniques could be used for future work. For example, if possible, you might want to collect more minority class observations because it can be useful later when we do resampling.

In addition, the penalised classification that imposes an additional cost on the model for having committed misclassification on the minority class during training could be used. These sanctions can push the model to pay more attention to the minority class.

It would be also interesting to examine how a single variable impacts on model choices, in other words the feature importance.

For more insights please see the notebook.

References

[1] Chun Gui: *Analysis of imbalanced data set problem: The case of churn prediction for telecommunication*, (2017).

[2] Teemu Mutanen: *Customer churn analysis – a case study*, (2006).

[3] Prashant Verma: *Churn Prediction for Savings Bank Customers: A Machine Learning Approach*, (2020).

[4] Abbas Keramati1 et al: *Developing a prediction model for customer churn from electronic banking services using data mining*, (2016).

[5] Churn Modeling, classification data set: https://www.kaggle.com/shrutimechlearn/churn-modelling