

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/381188797>

Churn detection on bank customers

Article · June 2024

CITATIONS
0

READS
27

3 authors:




Luis Ricardo García Oyervides

Tecnológico de Monterrey

1 PUBLICATION 0 CITATIONS

SEE PROFILE




Erick David Ramírez Martínez

Tecnológico de Monterrey

2 PUBLICATIONS 0 CITATIONS

SEE PROFILE



Jossian Abimelec García Quijano

Tecnológico de Monterrey

2 PUBLICATIONS 0 CITATIONS

SEE PROFILE

Churn detection on bank customers

Ramírez Martínez, Erick David
School of Engineering and Sciences.
Tecnológico de Monterrey
Estado de México, México
A01748155@tec.mx

García Oyervides, Luis Ricardo
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, México
a01088776@tec.mx

García Quijano, Jossian Abimelec
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, México
A01746745@tec.mx

Abstract—Churn is a growing problem that affects several industries, specially with the availability of competitors and when important factors such as money are considered. But it is not limited to a financial approach, as it can also be studied from several other perspectives such as school dropout or show engagement. Nevertheless, modern applications and algorithms have been widely used to study and detect churn in several areas, and it has proven to be effective. In this paper we present a comparison between several different ML models to perform churn detection in an anonymous bank.

Index Terms—Machine learning, churn, Kaggle, data analytics, retail

I. INTRODUCTION

Churn is the intent of a customer to leave a goods or service provider. It is a widely studied phenomenon that may happen to all the people that use a kind of service or consume a product. The concept of churn may be extended to any area or service, such as likeliness of a student to dropout due to varying reasons. In several areas, including telecommunications, churn is often approached as service switching instead of dropout because customers do not stop to consume a service, they change their provider [1], [2]. Additionally it has been increasing in complexity as it not only involves direct competitors and their prices, it also includes other factors such as online availability, fast transportation, delivery to site, amongst others. Moreover, banks also need to take into account benefits, software characteristics, as site design, interaction degree, ease of use and response time are something that prevents or encourages transactions. Causes for churn are large but include:

- Product or service pricing
- Product or service quality
- Product or service availability
- Product or service design
- Product or service price satisfaction
- Number of customer's complaints
- Customer's relationship with the brands providing the service or product

Companies and institutions try to avoid churn as it presents income losses which favors the competition. Nevertheless, churn's effects are not limited to the directly involved parties, it also affects other less related people. For instance, teacher churn not only affects a school work or scheduling force, it also affects students directly [3].

On the other hand, several ways to detect and reduce churn involve deep analysis of consumer behaviour. Nowadays, machine learning (ML) and deep learning (DL) models are popular and have proven to be successful in churn detection [4]–[6]. The main problem to implement these or any type or analysis is the availability of data. Nevertheless, this same reason has encouraged companies and institutions to make a great effort to collect all the available information possible to find ways to reduce churn.

Our objective is to predict churn on customers from an anonymous bank, using the popular and widely used machine learning models which have proven to be useful in a large range of applications. In consequence, to identify which customer profile is the most likely to cause churn.

The research questions are:

- Among our available dataset, what features are the most significant for churn rate prediction in an online retail store scenario?
- How feasible is to obtain this data?
- Which machine learning models perform better using the given features from the dataset?
- What is the performance of our proposed model and how does its performance compare to the state of the art solutions?
- What changes have to be made so that the proposed model can be applied to churn in other areas, industries or perspectives?

Our main contribution is to study the performance of 3 popular prediction methods in predicting churn in the bank sector. Even though churn is widely study across several areas, there is little work on the bank sector. We claim that one of the main reasons for a lack of available work in bank churn is due to the sensitive information used in predicting churn.

Finally, the structure of the document is as follows. Section 2 describes the method and data used to obtain the results. Section 3 describes the results obtained of churn prediction using the selected models. Section 4 discusses the obtained results. Section 5 includes our conclusion and future work.

II. RELATED WORK

Several works in literature focus on developing or optimizing models for churn detection across several areas. They tend to use datasets that provide customer profiles and their

relationship within the providers to gather information and give it to detection models.

Additionally, these works perform exploratory data analysis before training models due to all the data gathered. As churn is a complex process, the more data gathered, the better understanding of the profiles will be. However, some variables may provide enough information to allow the different models to predict churn, and thus data analysis is necessary to manipulate or even discard data.

For instance, C. Rao [7] upgrades a previous model for churn detection by optimizing the different stages of model training and prediction. Additionally, the training stage is optimized in order to mitigate the negative effects caused by class imbalance, which occurs in several datasets as customers who commit churn are not the majority. The optimized model performs better at several metrics compared to the literature and the original models in which optimization was done.

New methods and their results are being explored. For instance Szlag M. [8] applies monotonic rules to predict churn. These rules are sensitive to class imbalance and thus the predictor uses a limited balanced dataset derived from the original stated. However, even though the results are not as high as other models in literature, the monotonic rules are based on complex data analysis that provide insight on how much information can be acquired from a dataset.

However, even though churn detection is performed accurately, it may be necessary to develop an interactive system to predict churn in real time. In this endeavor, Singh P. [9] not only proposes detection models, but also develops an application for predicting churn, based on fields that are filled with customer data. The app allows to select a predictor and returns if the customer is prone to churn.

There is a lot of work done regarding churn, however, as churn can be applied to several endless areas with different models or outcomes, the literature is vast but it is not entirely focused on bank churn.

III. METHOD AND DATA

We make use of a Kaggle dataset [10] for model training and testing. This dataset contains information about an anonymous bank with several features describing the customers and their interaction with the bank. As an initial analysis of the data, class distribution is unbalanced with 80% labeled as 0 and 20% as 1.

The dataset features can be found in Table I, where our target feature is Exited. Most features have an integer nature, as they can be represented as numerical values. More complex features such as Gender or Balance change to float or string types, in order to better allocate the variables and the information. The dataset contains a total of 10,000 records. There is class imbalance, as the customers committing churn represent almost a fifth of the total dataset. However, a deeper data analysis will reveal if class balance strategies are necessary for this dataset. Additionally, the dataset contains entirely complete records, none of the columns or rows present missing or null values so cleaning is unnecessary. Nevertheless, some of the



Fig. 1. Correlation matrix containing only the numerical features in the dataset

columns present no value for prediction such as CustomerId and thus will be discarded.

The correlation matrix for the several features can be observed in Fig. 1. There is little correlation between the features except for Balance and NumOfProducts, which the main cause is that customers with balance in their accounts are able to purchase products unlike the ones with no balance. However, Complain has a high correlation with our target feature Exited, this reflects that if a customer makes a complaint, it is also almost guaranteed that will also commit churn. Churn is often caused by dissatisfaction or a lack of quality from the providers, which often result in a complaint and if not solved, it leads to churn. This is backed with the fact that Complain as a feature does not correlate with Satisfaction Score, which means that complaints have random outcomes and most of them are not addressed satisfactorily.

The importance of each feature can be observed in Fig. 2. Clearly the most important feature is Complain and brings the most information compared to all the features from the dataset. This result is expected as the correlation matrix in Fig. 1 shows that Complain feature has a high correlation with the target feature and thus brings more information than most of the other features.

We will train an AdaBoost classifier, a Random Forest classifier and apply logistic regression as our models for predicting churn. We will apply normalization to the dataset in order to increase the performance of our logistic regression model. We expect to obtain outstanding results due to the high correlation between Complain feature and our target feature. Nevertheless, we will also train the models without considering Complain feature in order to compare results and the impact of a highly correlated feature.

IV. RESULTS

We first trained our models with features containing the Complain feature and then without it.

TABLE I
FEATURES FOUND IN THE DATASET ALONG WITH THEIR DESCRIPTION AND DATA TYPE

| Variable | Type | Description |
|--------------------|---------|--|
| RowNumber | Integer | Sample number |
| CustomerId | Integer | Customer identification within the bank |
| Surname | String | Customer surname |
| CreditScore | Integer | The credit score of the customer |
| Geography | String | The living place of a customer |
| Gender | String | The gender of the customer |
| Age | Integer | The customer age |
| Tenure | Integer | Years of the customer being a client of the bank |
| Balance | Float | The balance in the bank of a customer |
| NumOfProducts | Integer | Products bought by the customer through the bank |
| HasCrCard | Integer | If the customer has a credit card |
| IsActiveMember | Integer | If the customer is actively using the bank |
| EstimatedSalary | Float | The estimated salary of the customer |
| Exited | Integer | If the customer has committed churn |
| Complain | Integer | Whether a customer has a complaint or not |
| Satisfaction Score | Integer | The score given by a customer after the complaint has been addressed |
| Card Type | String | The credit card type of the customer |
| Point Earned | Integer | Points earned by the customer for using a card |

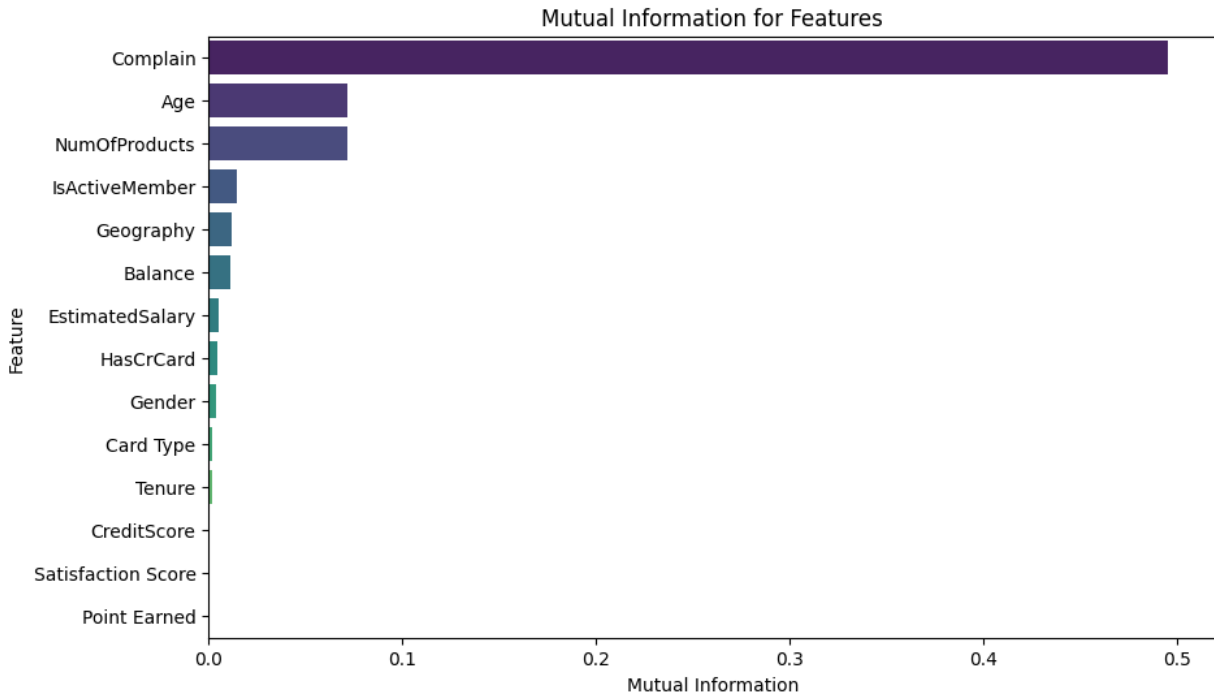


Fig. 2. Feature importance

TABLE II
RESULTS OF THE DIFFERENT MODELS FITTED TO THE DATASET. THE PERFORMANCE IS MEASURED IN SEVERAL METRICS

| Model | Without Complaints | | | With complaints | | |
|---------------------|--------------------|------------|--------|-----------------|------------|--------|
| | Accuracy | F1 - Score | Recall | Accuracy | F1 - Score | Recall |
| Adaboost | 0.835 | 0.69 | 0.67 | 1 | 1 | 1 |
| Random Forest | 0.821 | 0.69 | 0.68 | 1 | 1 | 1 |
| Logistic Regression | 0.8035 | 0.56 | 0.56 | 1 | 1 | 1 |

The results can be observed in Table II. The table contains the Accuracy, F1-Score and Recall of the 3 models trained with and without the Complain feature.

The models perform perfectly when the data contains the Complain feature, which is correct due to the high correlation to the target feature. The high correlation may be an error or a result of dataset manipulation and thus we omit the Complain feature in a further training.

When discarding the Complain feature, the performance decreases but remains high enough for the models to be useful for churn detection. The recall is lower than the accuracy, which means that the models present a high number of missclassifications.

V. DISCUSSION

The perfect performance of the models with the Complain feature was expected in consequence of the high correlation of the feature with the target. Nevertheless, this correlation may be caused by an error or dataset manipulation, rendering the feature unusable.

However, the Complain feature itself is expected to have a high impact in prediction. A complaint from a customer is caused by an dissatisfaction and dissatisfaction may lead to churn. This information may be increased by other measures such as Satisfaction Score, which, in this case, measures a score given by the customer after addressing the complaint. Nevertheless, the Satisfaction Score may also be flawed, as the score is given by the customer that did not follow a procedure for an objective score. Instead, the Satisfaction score may be measured with a metric that dictates if the complaint was resolved as the customer desired or not, even the most minimal difference in the solution may lead to a negative score due to the customer expectations.

On the other hand, the performance is worse when the models are fitted to the data without the Complain feature, which was expected. The models obtain an accuracy above 80% by using the next 5 most significant features but a recall and F1-Score below 70%. This is caused by a high number of false negatives.

For this specific case, the best model depends on a trade off between accuracy and recall, as Adaboost has the highest accuracy but Random Forest has better recall. Nevertheless, the difference between the metrics is below 2% and thus the gain or loss is minimal. The models are still suitable to be deployed to detect churn as it is able to detect churn and lower the costs of losing a customer.

VI. CONCLUSION

Results presented show the viability and performance that can be achieved in a real situation specially at companies with a high regard to customers such as banks, or other service providers. Churn is a very expensive practice, in most of the cases it is more affordable to retain clients instead of bringing new ones. The work presented helps not only to adequately predict churning, but also evaluates the contribution of the independent variables over the target. Results presented

indicate that tree-based classifiers (Adaboost and Random Forest) provide the best accuracy and also a high interpretability useful in business environments to take Data-Driven decisions. Finally, churn modeling in real companies should involve more analysis depending on the business context and the use of frameworks such as expected value.

ACKNOWLEDGMENT

We would like to thank the community at Kaggle that releases to the public domain their datasets, such as the one used in this project, for researchers and students to use in their projects freely.

REFERENCES

- [1] H. Ribeiro, B. Barbosa, A. C. Moreira, and R. G. Rodrigues, "Determinants of churn in telecommunication services: a systematic literature review," Feb. 2023. [Online]. Available: <http://dx.doi.org/10.1007/s11301-023-00335-7>
- [2] H. Triyafabrianda and N. A. Windasari, "Factors influence customer churn on internet service provider in indonesia," *TIJAB (The International Journal of Applied Business)*, vol. 6, pp. 134–144, 11 2022.
- [3] L. Menzies, "Continuity and churn: understanding and responding to the impact of teacher turnover," 2023. [Online]. Available: <http://dx.doi.org/10.14324/LRE.21.1.20>
- [4] E. Shaaban, Y. Helmy, A. Khedr, and M. Nasr, "A proposed churn prediction model," *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, pp. 693–697, 01 2012.
- [5] S. De, P. P, and J. Paulose, "Effective ml techniques to predict customer churn," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 895–902.
- [6] H. Guliyev and F. Yerdelen Tatoğlu, "Customer churn analysis in banking sector: Evidence from explainable machine learning models," *JOURNAL OF APPLIED MICROECONOMETRICS*, vol. 1, no. 2, p. 85–99, Dec. 2021. [Online]. Available: <https://journals.gen.tr/index.php/jame/article/view/1677>
- [7] C. Rao, Y. Xu, X. Xiao, F. Hu, and M. Goh, "Imbalanced customer churn classification using a new multi-strategy collaborative processing method," *Expert Systems with Applications*, vol. 247, p. 123251, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417424001167>
- [8] M. Szelag and R. Słowiński, "Explaining and predicting customer churn by monotonic rules induced from ordinal data," *European Journal of Operational Research*, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0377221723007440>
- [9] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, and E. Hossain, "Investigating customer churn in banking: a machine learning approach and visualization app for data science and management," *Data Science and Management*, vol. 7, no. 1, pp. 7–16, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666764923000401>
- [10] R. Kollipara, "Bank customer churn," <https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn?resource=download>, 2024, accessed: March 5th 2024.