Machine Learning ICA

Machine learning application for

Bank customer churn prediction

By

Name:

Student Id:

Email:

ABSTRACT

Bank Customer Churn Prediction system is a project aim at developing a machine learning application that can predict customers that might churn. In this report, matplotlib scripting layer (pyplot) was used to identify and visualize which factors or columns contribute to customer churn. To achieve this, the statistical and ML abilities of the programming language Python were used to explore a dataset consisting of 10,000 rows with 14 columns in jupyter notebook. The dataset was gotten from here. The algorithm used in this project are Logistic regression in the primal space, Random Forest, Ensemble models (XGBoost), and KNN classifier. Features were engineered and evaluated using the Sklearn packages. The result yielded after model evaluation on test dataset show that out of the various machine learning models, XGBClassifier model which predicts the churn with 60% f1 score, is the most powerful model for this scenario.

1 INTRODUCTION

Currently, Commercial banks are undergoing prodigious and tangled changes. Also, the entire industry faces many difficulties related to the growth of various products and services. Information technology has been grown rapidly using cloud computing, machine learning technologies in the last few decades. Also, financial regulation specialized in capital regulation is increasingly strengthened, and the process of financial disintermediation and interest rate marketization is gradually accelerating, resulting in a sharply narrowed interest margin for banks.

The banking customers pay more attention to their experience, customized services, assortment, and agility, which further intensifies the competitiveness among commercial banks. Customer retention and expanding the revenues from existing customers by up/cross-selling is a much more profitable strategy for growth in comparison to new customer acquisition. In order to maximize the profit, commercial banks must increase the customer base by incrementing sales while decreasing the number of churners. Furthermore, it is common knowledge that retaining a customer is about five times to six times less expensive than acquiring a new one, while suggests acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing customers. In addition, positive word-of-mouth from existing customers leads to low-cost or almost free customer acquisition.

One of the paramount competitions among commercial banks is customer retention, especially for high-value customers. As customers are directly related to profits, commercial banks must avoid the loss of customers while acquiring new customers. Believes that by reducing the customer defection rate by 5%, companies can increase profits by 25% to 95%, while Business Week

thought the profits would increase by 140%. As can be seen, reducing customer attrition has a significant impact not only on increasing profits for commercial banks but also on enhancing their core competitiveness. Therefore, it is strongly needed for commercial banks to improve the capabilities to predict customer churn, thereby taking timely measures to retain customers.

Churn is a critical area in which the banking domain can make or lose their customers. Hence, the business spends a lot of time making statistical predictions, which successively helps to make the necessary business conclusions. The customer churn can be averted by studying the demographic features and history of the customers, especially the transaction patterns using machine learning models.

1.1 LITERATURE REVIEW

New technology, regulation and change in demand has caused a rise of Fintech companies challenging banks dominant position in society (The Economist, 2019). In times of intensified competition, customer turnover can pose a real threat for existing companies (De Caigny, et al., 2018). Customer turnover, also referred to as customer churn, is when a customer leaves or ends an engagement with a company during a given time period (Colgate, et al., 1996). As a result of increasing competition, it is important for banks to maintain existing customers, as this is more cost-effective than acquiring new ones, in order to ensure their position in society.

In addition, new technology has increased banks access to data, and thus data driven customer churn analysis is workable. Taken together, there is a growing demand for customer churn analysis which studies a set of characteristics in order to predict customer churn (De Caigny, et al., 2018). This has intensified the demand for predictive modelling built on for example statistical learning methods. Since, if a bank can predict customer churn, targeted marketing campaigns can be used to persuade customers to keep their engagement (Ganesh, et al., 2000).

2 DATA EXPLORATION

Data exploration helps us to understand the data and is an essential part of machine learning. This is due to machine learning being all about making prediction. On the other hand, data exploration is all about providing elementary information and identifying potential association between variables in a dataset.

Data shape, data type, statistics, null values, data unique and the general pictorial format of the dataset are explored in the notebook section. The dataset contain 10,000 rows and 14 variables, the data types are all in correct format, luckily there is no null value.

The correlation between independent and dependent variables are also explored by

visualizing each independent variables against the target variable. There are several useful Python libraries for visualizing data, but the two used in this project are Matplotlib and Seaborn. The visuals created are available in the notebook section, and were exported to the folder/[project visuals] with unique and informative file names. The percentage of the churn and retain customer in the dataset was also explored which shows that, out of 100% record, 79.63% are retained while 20.37% are churned. Below is the chart.

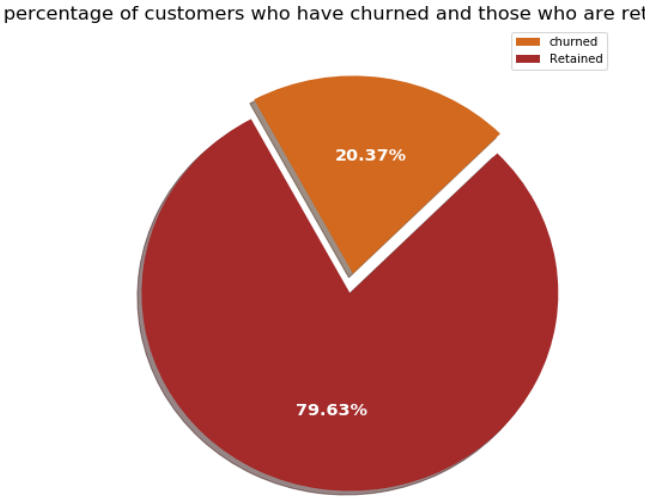


Fig 1 percentage of churn and retain customer

2.1 FEATURES SELECTION

Initially, total 20 variables including the variables obtained through feature engineering, were considered as independent and dependent variables for model building. Out of these variables, 3 have been removed because they are specific to customer, 17 have been identified for model execution.

3 EXPERIMENTS

The suitability of ML models depends on the relationship between the variables, the data structure, and preprocessing. The imported data are mainly continuous and categorical which are encoded during preprocessing. Labels are included, so the data sets are suitable for supervised learning and classification methods. After data preparation and cleaning, the whole dataset translated into 10,000 rows and 17 columns.

In this project, the dataset features were fit to several models, which are:

- KNN classifier
- Logistic regression
- Random Forest
- Ensemble models (XGBoost)

The dataset is partitioned into training and testing datasets in the ratio of 80:20

3.1 EVALUATION METRICS

The model evaluation metric used in this project is classification report, it is one of the performance evaluation metrics of a classification-based machine learning model. Below is the explanation to most of used information in classification report.

3.1.1 Precision

It is defined as the ratio of true positives to the sum of true and false positives.

3.1.2 Recall

It is defined as the ratio of true positives to the sum of true positives and false negatives.

3.1.3 The F1-score

It is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

3.1.4 Support

It is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

3.1.5 Accuracy

It is the fraction of predictions our model got right.

4 RESULTS

In this section the performance for each model calculated on the training and testing dataset is presented. Firstly, the results for KNN classifier is presented, followed by the result for Logistic regression, Random Forest, and lastly for Ensemble models (XGBoost).

4.1 KNN CLASSIFIER

	Train dataset		Test dataset	
	F1-score	accuracy	F1-score	accuracy
0	91	85	90	82
1	53		25	

Table 1 KNN model result on train and test dataset.

4.2 LOGISTIC REGRESSION

	Train dataset		Test dataset	
	F1-score	accuracy	F1-score	accuracy
0	96	94	92	86
1	83		58	

Table 2 logistics regression model result on train and test dataset.

4.3 RANDOM FOREST

	Train dataset		Test dataset	
	F1-score	accuracy	F1-score	accuracy
0	94	90	92	87
1	71		60	

Table 3 random forest model result on train and test dataset.

4.4 ENSEMBLE MODELS (XGBOOST)

	Train dataset		Test dataset	
	F1-score	accuracy	F1-score	accuracy
0	91	85	86	78
1	52		40	

Table 4 Xgboost model result on train and test dataset.

5 DISCUSSION

5.1 Justification of algorithms

Justification of the best performing Algorithms used will briefly be discussed.

5.1.1 KNN CLASSIFIER

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It is a non-parametric algorithm which means it does not make any assumption on underlying data.

5.1.2 LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1**,** it gives the probabilistic values which lies between 0 and 1.

5.1.3 RANDOM FOREST

As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

5.1.4 ENSEMBLE MODELS (XGBOOST)

Xgboost is a popular and efficient open-source implementation of the gradient boosted trees algorithm which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage.

6 CONCLUSION

From the results gotten after model evaluation, the best model that gives the highest overall accuracy with a decent balance of the recall and precision is the XGBoost ensembles method where according to the fit on the testing set, with the overall accuracy of 87% and a precision score on 1's of 0.77, out of all customers that the model thinks will churn, 77% do actually churn and with the recall score of 0.49 on the 1's, the model is able to highlight 49% of all those who churned. In as much as most models has a high accuracy, our aim is to predict those customers that might churn, Therefore XGBClassifier model which predicts the churn with 60% f1 score will be the most powerful model for this scenario.

FUTURE WORK

The f1-score of the model on test dataset is slightly higher with regard to predicting 1's i.e. those customers that churn. However, in as much as the model has a high accuracy, it still misses 40% of those who end up churning whereas our main aim is to predict the customers that will possibly churn so they can be put in some sort of scheme to prevent it. This could be improved by retraining the model with more data over time while in the meantime working with the model to save the 60% that would have churned.

REFRENCES

Aman Kharwal (2021). Classification Report in Machine Learning. URL: https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/

Chicco, D. & Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21(1), pp. 6-13.*

Colgate, M., Stewart, K. & Kinsella, R., 1996. Customer Defection: A study of the student market in Ireland. *International Journal of Bank Marketing, 14(3), pp. 23-29.*

Ganesh, J., Arnold, M. J. & Reynolds, K. E., 2000. Understanding the Customer Base of Service
Providers: An Examination of the Differences between Switchers and Stayers. *Journal of Marketing, 64(3), pp. 65-87.*

Guest_blog (2018). An End-to-End Guide to Understand the Math behind XGBoost. URL: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

He, B., Shi, Y., Wan, Q., Zhao, O. Prediction of Customer Attrition of Commercial Banks based on SVM Model. *Procedia Computer Science, 31, 423-430, 2014.* https://doi.org/10.1016/j.procs.2014.05.286.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X., 2013. *Applied logistic regression*, 3rd ed. New Jersey, NJ: Wiley.

Mutanen, T., Ahola, J. and Nousiainen, S. Customer Churn Prediction-A Case Study in Retail Banking.
ECML/PKDD2006 Workshop, 13-18, 2006.

The Economist, 2019. A Whole New World: *How technology is driving the evolution of intelligent banking*, London: The Economist Intelligence Unit (EIU).

Verbeke, W. et al., 2012. New insights into churn prediction in the Telecommunication Sector: A
profit driven data mining approach. *European Journal of Operational Research, 218(1), pp. 211-*
229.

https://www.google.com/search?client=firefox-b-d&q=classification+report+in+machine+learning

https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

https://www.javatpoint.com/logistic-regression-in-machine-learning

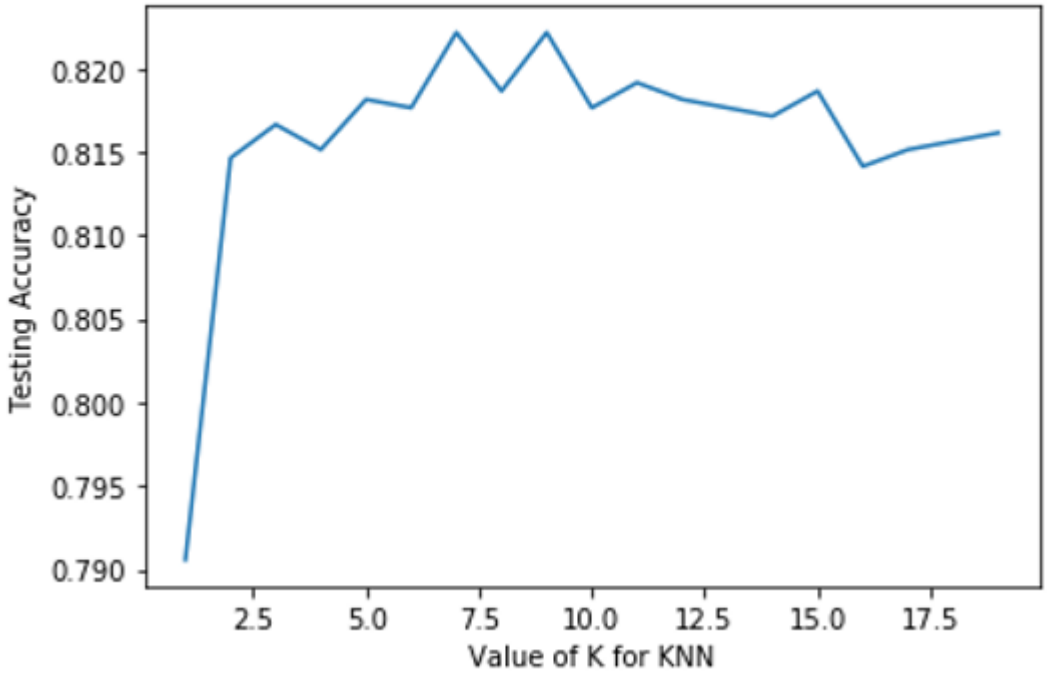
https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

https://www.javatpoint.com/bagging-vs-boosting

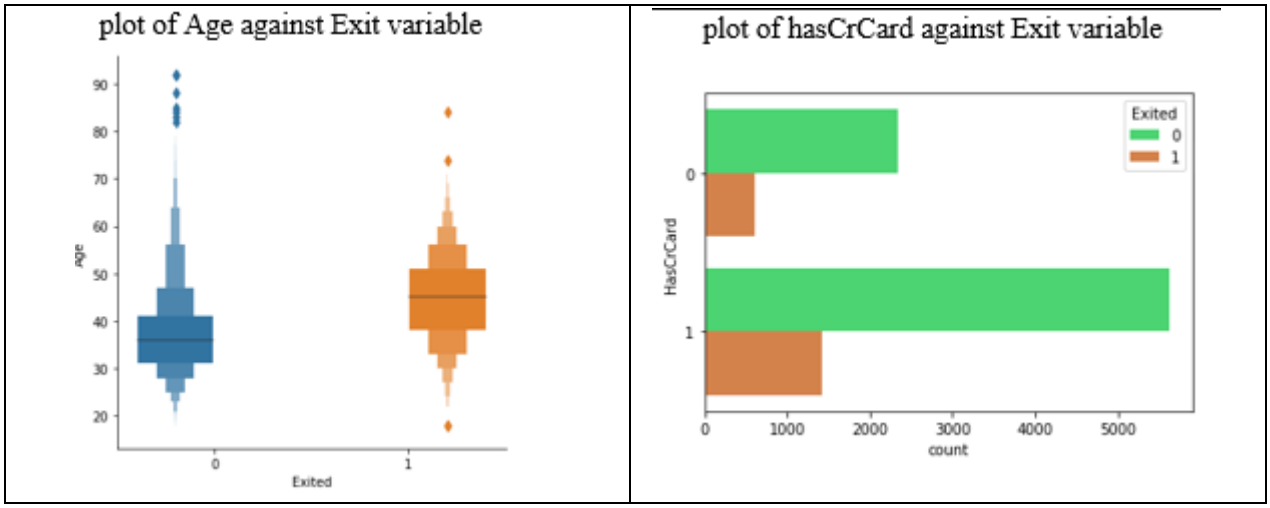
Appendix

Screenshots

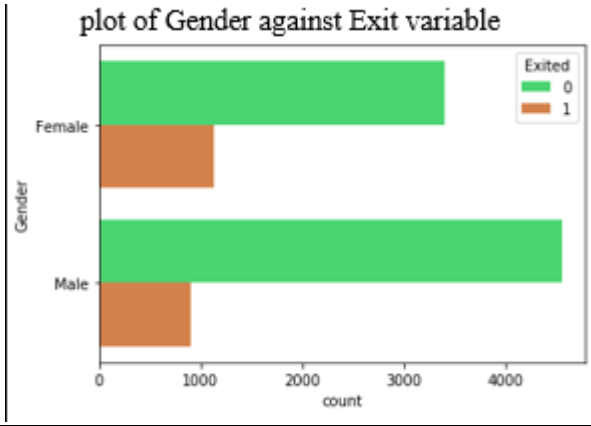
Performing Plot to get the best value of K in order to have a higher accuracy score.



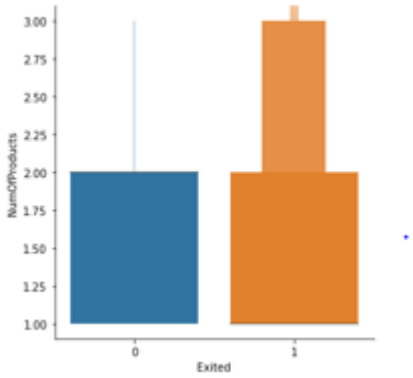
The following images shows the plot Results from Interesting Features data and feature distribution. This can be found in the project visuals folder.



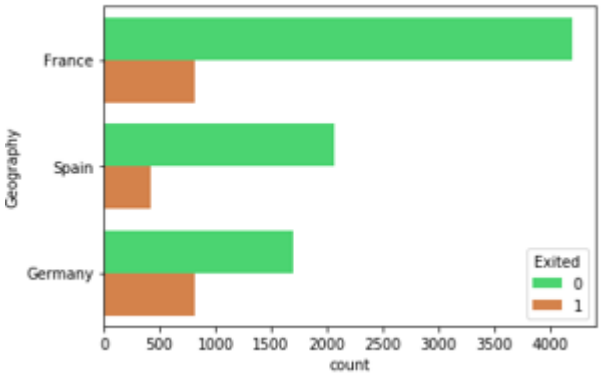
plot of Gender against Exit variable



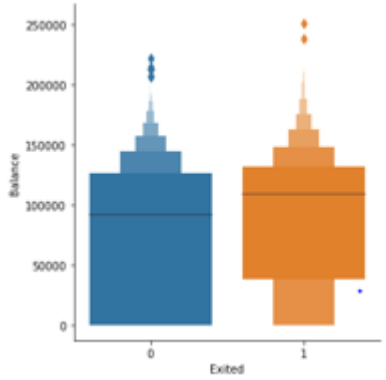
plot of Number of products against Exit variable



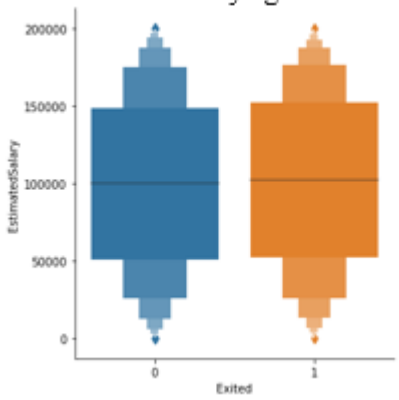
plot of Geography against Exit variable



plot of Balance against Exit variable



plot of Estimated Salary against Exit variable



plot of IsActiveMember against Exit variable