**Abstract**

# **Customer Churn Prediction In Banking Using Machine Learning**

Asmita Bhardwaj Joji John Amalraj Irudayasamy

IIIT Delhi IIIT Delhi IIIT Delhi

asmita.opd21005@iiitd.ac.in joji.opd21009@iiitd.ac.in amalraj.opd21025@iiitd.ac.in

Customer churn predicition is an important research discipline as part of customer relationship management because it is a lot more profitable to retain and satisfy existing customers than to attract new customers for several reasons. Successful companies have long term relationships with their customers which allow them to focus on their customer needs instead of looking for new and potentially not very profitable customers who are typically characterized by a higher attrition rateThe project promotes the exploration of the likelihood of churn by analyzing customer behavior. In this study, we have used Adaboost, XG Boost, GridSearch Cv, Random Search CV, Light GBM to compare and find which algorithm gived best accuracy based on various matices. Also, some feature selection methods have been done to find the more relevant features and to verify system performance. We use the dataset from Kaggle for this study. The results are compared to find an appropriate model with higher precision and predictability. As a result we find that Grid Search CV, Random Search CV, and Light GBM gives best results.

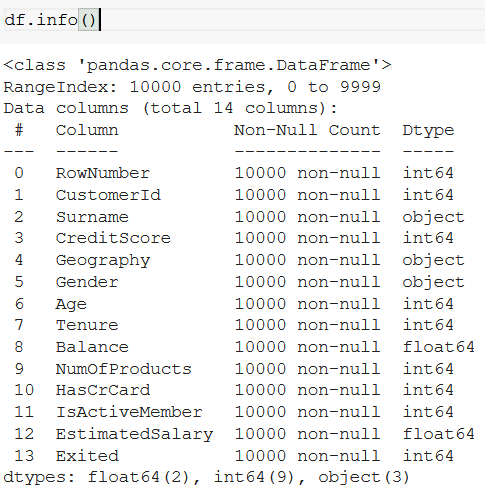
# **Introduction**

# Customer churn is one of the significant challenges that businesses face today, especially those offering finance or account holder-based services. Account holder churn (or customer downfall) is essentially customer unhappiness caused by a change in taste, need for legitimate customer relationship methodology, change of home, and a few other reasons. If companies can effectively predict customer churn, they can segment those customers who are more likely to churn and provide them with better services. Therefore, this project focuses on Churn Prediction model that uses Machine Learning techniques such as Logistic Regression, Decision Trees, K-Nearest Neighbors, and Support Vector Machine algorithms to help companies predict customers who are most likely to churn. In this way, they can achieve a high retention rate and maximize their sales.

# **Motivation & Literature review**

There are numerous things that businesses can make mistakes, from complicated onboarding when customers aren't provided with easy-to-understand information about product usage and their capabilities to poor communication and lack of feedback criticism or delayed responses to inquiries. Long-time customers can feel undervalued because they don't receive as many bonuses or rewards as the unused ones. In general, the overall customer experience defines brand perception and influences how customers perceive the value for money of the products or services they use. Thus, this expands the customer churn prediction models to anticipate customer churn so that any over-deficiencies in potential customers can be overcome by assessing their propensity to churn.

# **Dataset description & Preprocessing**



*Table 1*

**CreditScore**—can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.

**Geography**—a customer’s location can affect their decision to leave the bank.

**Gender**—it’s interesting to explore whether gender plays a role in a customer leaving the bank.

**Age**—this is certainly relevant, since older customers are less likely to leave their bank than younger ones.

**Tenure**—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.

**Balance**—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.

**NumOfProducts**—refers to the number of products that a customer has purchased through the bank.

**HasCrCard**—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.

**IsActiveMember**—active customers are less likely to leave the bank.

**EstimatedSalary**—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.

**Exited**—whether or not the customer left the bank.

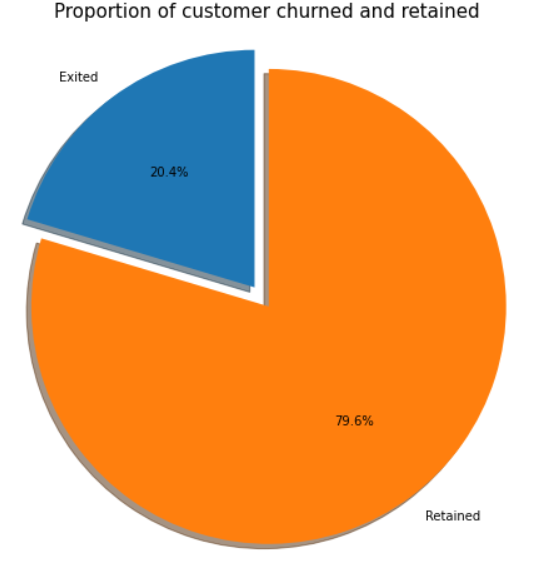
****

Fig 1

Given 20% is a small number, 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.

As a part of Feature selection we have dropped 3 features based on our analysis using a technique called high variance inflation factor (VIF) . To reconfirm this finding, we did the feature selection with RFE. For this ,the data needs to be split into training and testing sets and undergo normalization. RFE is used to identify the *n* features that should be selected. The RFE uses linear regression model to select relevent features from many features. A valuable feature of the RFE is that, it can show the order in which the eliminated features are removed. This can be accomplished with the ‘.ranking\_’ method. A feature with a rank of 1 is a selected feature. Any other ranking represents an eliminated feature.

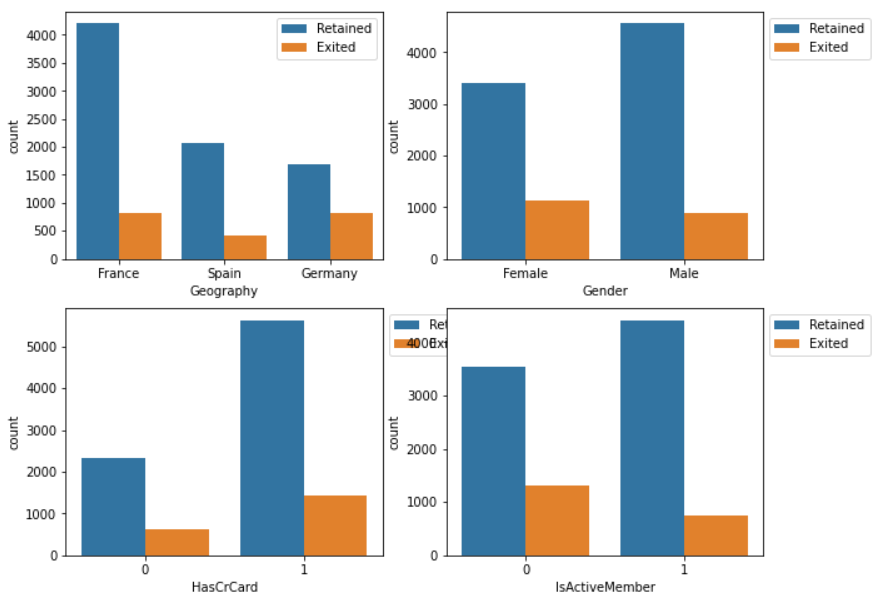


Fig 2

# **Methodology**

To measure the comprehensibility of a classification algorithm the size and the classification output type are important determinants. Classification output types define the format in which the results of the model are returned to the researcher. Commonly used classification output types are linear models, non-linear models, rule-based models and tree based models. The size depends on the output types; the size of linear and non-linear output types is measured by the number of terms while the size of rule-based and tree based output types is given respectively by the number of rules and the number of leaves. By consequence customer churn prediction is a complex process which requires informed decisions from a researcher at various stages. It is important that these decisions are data-driven whenever it is possible. The main decision point for a researcher in the modeling step is which of the classification algorithms one should use for which he or she faces a trade-off between comprehensibility and predictive performance. This explains why xG Boost and Light GBM are popular techniques in customer churn prediction as they combine good predictive performance with good comprehensibility. Moreover, dynamic classifier selection methods are known to be computationally expensive in the scoring phase.

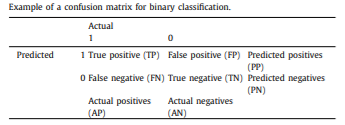


Table 2

# **Individual tasks**

**1.** Joji John - Visualising Dataset, Logistic Regression, kNN, Random Forest, Light GBM

**2.** Asmita Bhardwaj - Loading dataset, Label Encoding, SVM, xG Boost , Random Search CV, Grid Search CV

**3.** Amalraj Irudayasamy - Data Cleaning, Splitting data, Naive Bayes, Decision tree

# **Results and Analysis**

The area under the Receiver Operating Characteristics curve (AUC), Recall Score and Kappa Score are used to measure the classification performance. The Kappa scores and Recall casore of Lght GBM and Random Search CV are significantly better than Grid Search CV or Ada Boost and performs at least as well as more advanced XG Boost models. It was interesting for us to find that SVM, a really powerful classifier, to perform badly with this dataset.

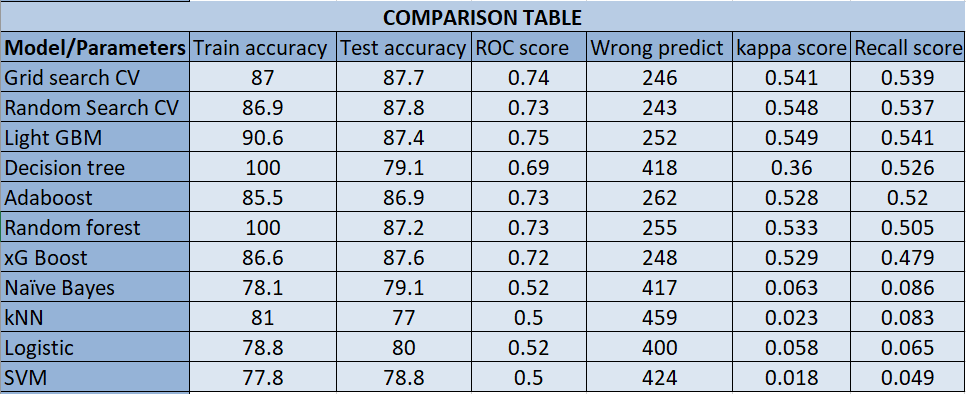


Table 3

The predictive performance of the different classifiers is assessed by the area under the Receiver Operating Characteristics curve (AUC). The AUC can be derived from the confusion matrix and it is used to evaluate the predictive performance of a binary classification system such as Customer Churn Prediction with a simple one-figure score.

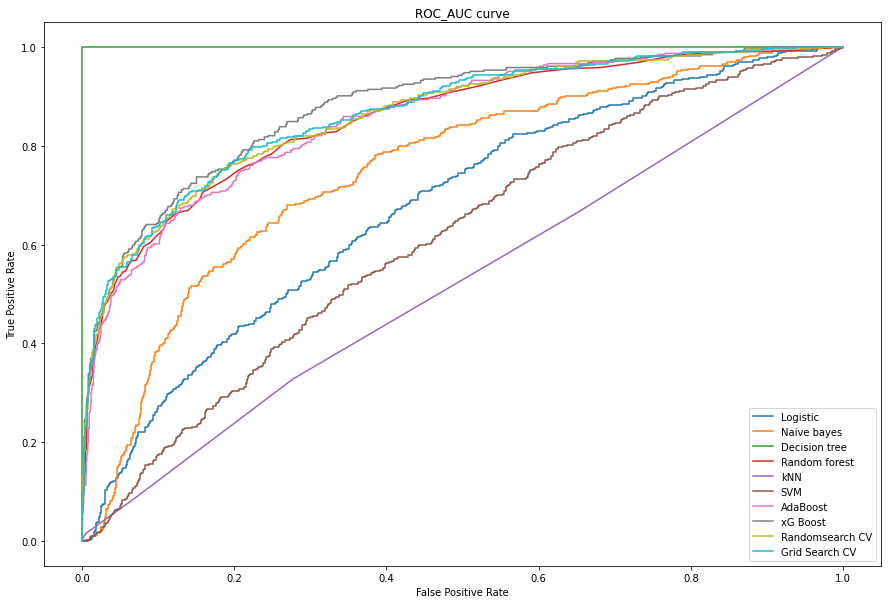


Fig 3

**Future Research**

AUC and Kappa Score are used to assess the predictive performance of the model. These methods have their advantages but they lack a direct profit criterion. The recently introduced maxim profit (MP) criterion (Verbeke et al., 2012) provides a new evaluation metric that incorporates profit. This evaluation metric requires additional information to calculate the customer lifetime value which was not possible to gather for all datasets in this benchmark study. Therefore further research is needed to verify the performance of the models used here by using this maximum profit (MP) criterion.

# **References**

De Caigny, A., Cousement, K., De Bock, K.W (2018)., “A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees”, European . Journal page- 760–772, Volume 1, Issue 1.

Hadden J, Tiwaria A, Roy R, Ruta D (2005) Computer assisted customer churn management: State-of-the-art and future trends. Comput Oper Res 34:2902–2917

Miguéis VL, Van den Poel D, Camanho AS, Falcão e Cunha J (2012) Modeling partial customer churn: On the value of first product-category purchase sequences. Expert System Application 39:11250–11256

Nie G, Rowe W, Zhang L, Tian Y, Shi Y (2011) Credit card churn forecasting by logistic regression and decision tree. Expert System Application 38:15273–15285.

Praveen Asthana (2018), "A comparison of machine learning techniques for customer churn prediction ", International Journal of Pure and Applied Mathematics, Page-1149-1169, Volume 119, Issue 10.

Risselada H, Verhoef PC, Bijmolt THA (2010) Staying power of churn prediction models. J Interact Mark 24:198–208

Seng J-L, Chen TC (2010) An analytic approach to select data mining for business decision. Expert System Application 37:8042–8057

Sharma A, Panigrahi PK (2011) A neural network based approach for predicting customer churn in cellular network services. Int. J Computer Application 27(11):26–31

Tsai C-F, Lu Y-H (2009) Customer churn prediction by hybrid neural networks. Expert Syst Appl 36:12547–12553

V. Kavitha, G. Hemanth Kumar, M. Harish (2020), “Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms”, International Journal of Engineering Research & Technology (IJERT), Volume 9.