# Predicting Customer Churn with Machine Learning-

## In a Bank

Name : Sara Samaninia

Student Number : 501090785

Supervisor’s Name : Tamer Abdou


Ryerson University

# Table of Contents

# Project Abstract

Customer churn or customer attrition happens when the customers stop using the product or service of a business. It directly affects the company profitability. The success rate of selling to an existing customer is usually higher than the success rate of selling to a new customer. Also, it is estimated that acquiring a new customer can cost five times more than retaining an existing one. So, customer churn analysis is essential for any business or industry including banking. Customer churn prediction is one of the challenging issues, but it helps the business to identify the problems. Weather it is the poor quality product/service or wrong target market.

The goal of this project is to predict customer churn for a bank. In this dataset (churn_modeling.csv), we look at the "Exited" column to see if that customer is churned or not. We use the features such as Credit Score, Gender, Age, Tenure, Number of Products and … to predict customer churn.

The dataset includes 10,000 records and 13 columns (excluding row number), combination of numerical and non-numerical features. Fortunately, it doesn’t contain missing entries.

We will initially perform EDA analysis and Data Pre-processing to identify and visualise the features contributing to customer churn. It’s a classification task and we will use ML classifiers such as Logistic Regression, Random Forests, SVM and maybe other algorithms to compare the prediction performance.

The bank in this study has been gathering customer data for a while to identify potential churners. With analysing those customers who have already left the bank, we identify if they have some shared feature or behaviour patterns. Bank needs to identify customers at risk of churn before it is too late to take appropriate actions and optimize their strategic plans. Machine learning algorithms can help us here to resolve the below problems:

- Having the up-to-date list of potential churners, would greatly help sales and marketing to engage with customers differently. For example, customers who are currently at churn risk are not the good candidates to target in marketing campaigns to buy new products. When customers have already showed signs of churn, it is not a great time for sales department as well to reach out about additional services.  Non-churn risk customers are probably better candidates to target at launching new service or product.

- Customer service management can use this study result to take appropriate actions, reach potential churners and understand their issues or pain points and gain back their trust. Customer satisfaction/success managers need this insight to know which customers they should contact. Successful customer interaction and retention strategy is related to speaking with the right customers at the right time.

- Identifying the features that contribute the most in customer churning help them address the specific and common issues the potential churners with those features might have. Implementing these insights is the opportunity to improve the product or service for growth and to reduce customer churn.

# Literature Review

One of the main objectives of Churn prediction is finding the strategies for customer retention. The risk of customer churn in global markets and competition growth is always increasing. Hence, identifying early the churn signals for the customers that may leave voluntarily is becoming more and more necessary. Companies have realized keeping their existing customers is one of their most valuable assets ( Lalwani, Mishra, Chadha & Sethi (2021).

This project aims to predict customers that are most likely to get churned in a bank using machine learning methods. The dataset in this study to create the churn models, is available in Kaggle. The dataset contains 10000 records of bank customers. The target or dependent parameter is a binary variable called "Exited". It reflects whether the client has left this bank or not. Among 10000 customers, 7963 were retained and 2037 were exited. When the customer is retained, the target parameter reflects the binary flag 0 and when the customer has churned, the target parameter reflects the binary flag 1. The dataset includes 13 features or independent variables from customer data and transactions. This study considered various studies in this topic and in various industries with various dataset; however the main focus was on one study that have been conducted by Vasimalla & Rahman in 2020 (ML Based Customer Churn Prediction In Banking) with the same dataset. All attributes with the brief description are listed in the below table.

| Feature Name | Feature Description |
|---|---|
| Row number | Row numbers from 1 to 10000. |
| Customer Id | Unique Ids for bank customer identification. |
| Surname | Customer's last name. |
| Credit Score | Credit score of the customer. |
| Geography | The country from which the customer belongs. |
| Gender | Male or Female. |
| Age | Age of the customer. |
| Tenure | Number of years for which the customer has been with the bank. |
| Balance | Bank balance of the customer. |
| Num of Products | Number of bank products the customer is utilizing(savings account, mobile banking, internet banking etc.). |
| Has Cr Card | Binary flag for whether the customer holds a credit card with the bank or not. |
| Is Active Member | Binary flag for whether the customer is an active member with the bank or not. |
| Estimated Salary | Estimated salary of the customer in Dollars. |
| Exited | Binary flag 1 if the customer closed account with bank and 0 if the customer is retained. |

## Creating a Test Set

At first, will split our dataset into train and test with a function that  implements random sampling.

## EDA Analysis

Exploratory Data Analysis helps us understand our dataset better and get some insightful statistical information (like mean, max, and min) about the features and to perform initial investigations on data to discover patterns, spot anomalies and check assumptions with the help of summary statistics and visualizations. The statistical summary of the numerical features are provided in the below.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| CreditScore | 10000.0 | 650.529 | 96.653 | 350.00 | 584.00 | 652.000 | 718.000 | 850.00 |
| Age | 10000.0 | 38.922 | 10.488 | 18.00 | 32.00 | 37.000 | 44.000 | 92.00 |
| Tenure | 10000.0 | 5.013 | 2.892 | 0.00 | 3.00 | 5.000 | 7.000 | 10.00 |
| Balance | 10000.0 | 76485.889 | 62397.405 | 0.00 | 0.00 | 97198.540 | 127644.240 | 250898.09 |
| NumOfProducts | 10000.0 | 1.530 | 0.582 | 1.00 | 1.00 | 1.000 | 2.000 | 4.00 |
| HasCrCard | 10000.0 | 0.706 | 0.456 | 0.00 | 0.00 | 1.000 | 1.000 | 1.00 |
| IsActiveMember | 10000.0 | 0.515 | 0.500 | 0.00 | 0.00 | 1.000 | 1.000 | 1.00 |
| EstimatedSalary | 10000.0 | 100090.240 | 57510.493 | 11.58 | 51002.11 | 100193.915 | 149388.247 | 199992.48 |
| Exited | 10000.0 | 0.204 | 0.403 | 0.00 | 0.00 | 0.000 | 0.000 | 1.00 |

In this data set, there was not any duplicates and missing (null) values. Continuous variables are Age, CreditScore, Balance, EstimatedSalary and Categorical variables are Geography, Gender, Tenure, NumOfProducts, HasCrCard, IsActiveMember (After dropping the irrelevant data/attributes which is explained in Pre-processing section.)

From conducting the exploratory data analysis some initial insights are reached. For example, we understand only a small percentage leaves within the first year; The bank kept 80% of its clientele and our dataset is skewed/imbalanced since the number of instances in the 'Retained' class outnumbers the number of instances in the 'Churned' class by a lot.

Different visualisation techniques are applied to different types of variables to differentiate between continuous and categorical variables and look at them separately. You can find further details in the links below for EDA analysis:

https://github.com/sara-cloud/Project820/blob/main/EDA_Analysis.ipynb

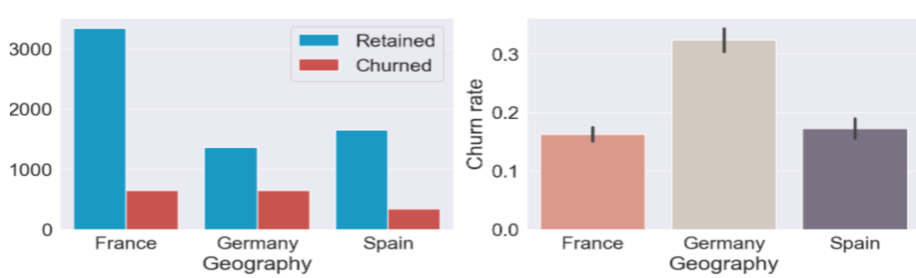
https://github.com/sara-cloud/Project820/blob/main/EDA_Analysis.pdf

## Data Preprocessing

Preprocessing is an important phase to convert raw data into a suitable format for building and training ML models that can guarantee the model can make good prediction. During that, the tasks such as feature selection, data conversions and imbalanced data will be handled.

### Feature Selection

Data or attributes which have no impact on our prediction are **irrelevant** and keeping them may negatively affect the performance of our classification models. Irrelevancy of some attributes including 'RowNumber', 'CustomerId', and 'Surname' are obvious as they are specific to each customer. These attributes can be dropped. In the similar study(Vasimalla & Rahman, 2020), they proposed that Geography too has nothing to do with the prediction and they dropped this in the initial phase. However, based on EDA analysis done, this feature is kept since it is observed that customers in Germany are more likely to churn than customers in the other two countries (the churn rate is almost double compared to Spain and France). So, this feature hasn't been neglected for this study. (Please see the graph in below)



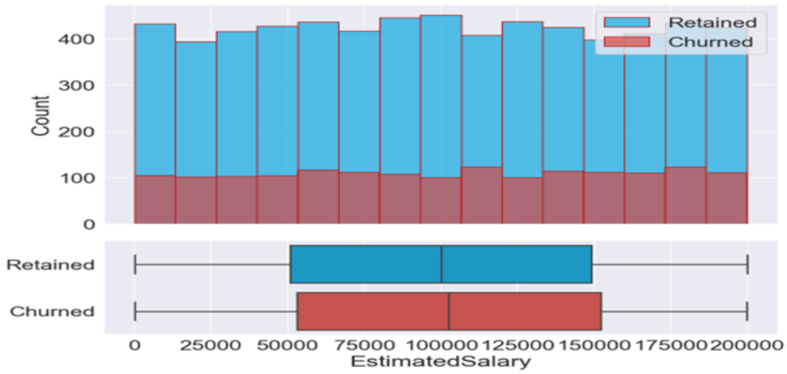
Vasimalla & Rahman used mRMR (Minimum Redundancy Maximum Relevance) and Relief which are both Filter type feature selection methods. The Filter method in feature selection evaluates the significance of feature characteristics, such as the variance of feature and reaction relevance. In mRMR for classification problems, it ranks the features sequentially applying the Minimum Redundancy Maximum Relevance algorithm (Jiang and Li, 2015).

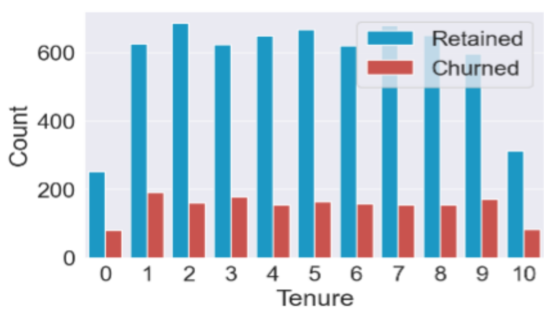
Relief rates features using the Relief algorithm. This method is more suitable to estimate the features significance for distance-based supervised models which apply pair distances between observations to predict the response. It ranks the predictors based on importance of specified numbers of nearest neighbors (Beretta and Santaniello, 2011).

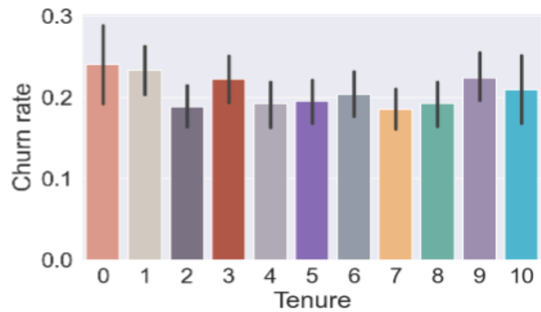
Although the mentioned algorithms will be studied for any addition or advantage of application for this study, other Filter type feature selection methods(tests) such as Chi-square test for categorical variables and Anova for numeric variables can serve the purpose.

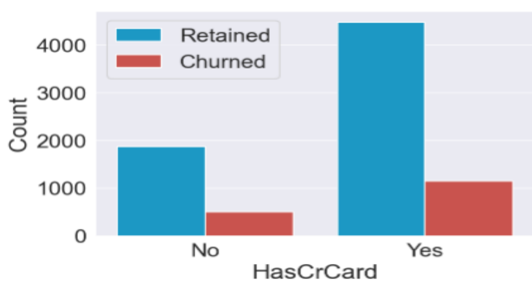
The Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with the best Chi-square scores. ANOVA is used when one variable is numeric and one is categorical, such as numerical input variables and a classification target variable in a classification task.

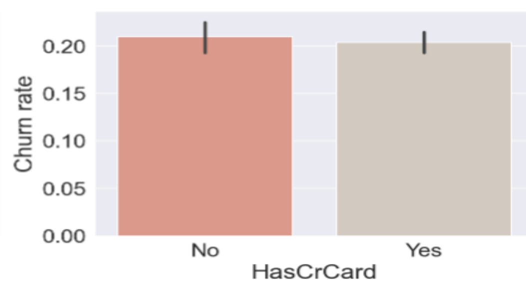
Since EDA already revealed more features that can be dropped as they do not provide any value in predicting the target variable. Chi-square and Anova will be used only to confirm the initial hypothesis. Then we can use the drop method to remove the three variables from the train set. EstimatedSalary in continuous variables for both Churned and Retained shows a uniform distribution. And Tenure and HasCrCard in categorical features deemed redundant as they have a similar churn rate. Please see the graphs in below.




## Encoding Categorical Features

Machine learning algorithms work with numeric features. So, categorical variables require to be transformed (encoded) to numeric values in Preprocessing steps.

In this dataset two variables require encoding requirement, Gender and Geography. For example, Gender can be transformed like Male --> 1 and Female --> 0. And the 3 categorical values in Geography as well will be manually mapped to numeric values. In the previous similar study(Vasimalla & Rahman, 2020), they only encoded Gender and Geograpgy had been removed from the features list.

## Scaling

Scaling is applied to normalise the range of variables in a dataset. Some machine learning algorithms are sensitive to feature scaling such as SVMs, while others like Random Forests are invariant. During this method the features will be standardised using mean and standard deviation. Feature scaling in this dataset will be applied for Age, CreditScore and Balance. Although Vasimalla & Rahman(2020) used SVMs as one of their ML algorithms, they skipped scaling.

## Addressing Class Imbalance

As we have seen previously, the data is highly imbalanced(7963 Retained class and 2037 Churned class). If we apply Classifications on imbalanced data, the result will be biased in favour of the majority class. There are some techniques or strategies such as oversampling and undersampling that can address this problem and configure the class distribution.

Vasimalla & Rahman(2020) used typical random oversampling and they stated that didn't use undersampling because the size of data will decrease and there will not be enough data to build the model. Therefore, they used the random oversampling for the minority class. In this study however, SMOTE technique is considered due to the main disadvantage with oversampling that by making exact copies of existing examples, overfitting is more likely. This technique is suggested in the study of Paulose et al(2021) regarding Effective ML Techniques to Predict Customer Churn. Random oversampling just increases the size of the training data set through repetition of the original examples. Oversampling using SMOTE not only increases the size of the training data set, it also increases the variety by generating synthetic examples rather than by oversampling with replacement (Huang, 2015).

## Building Machine Learning Models

Aljoumaa et al.(2019) applied decision tree, random forest, GBM tree algorithm, and XGBoost algorithms for the customer churn Prediction in telecommunication industry. XGBoost performed superior than others in terms of AUC accuracy. They also suggested accuracy can be improved further with the optimization feature selection algorithms.

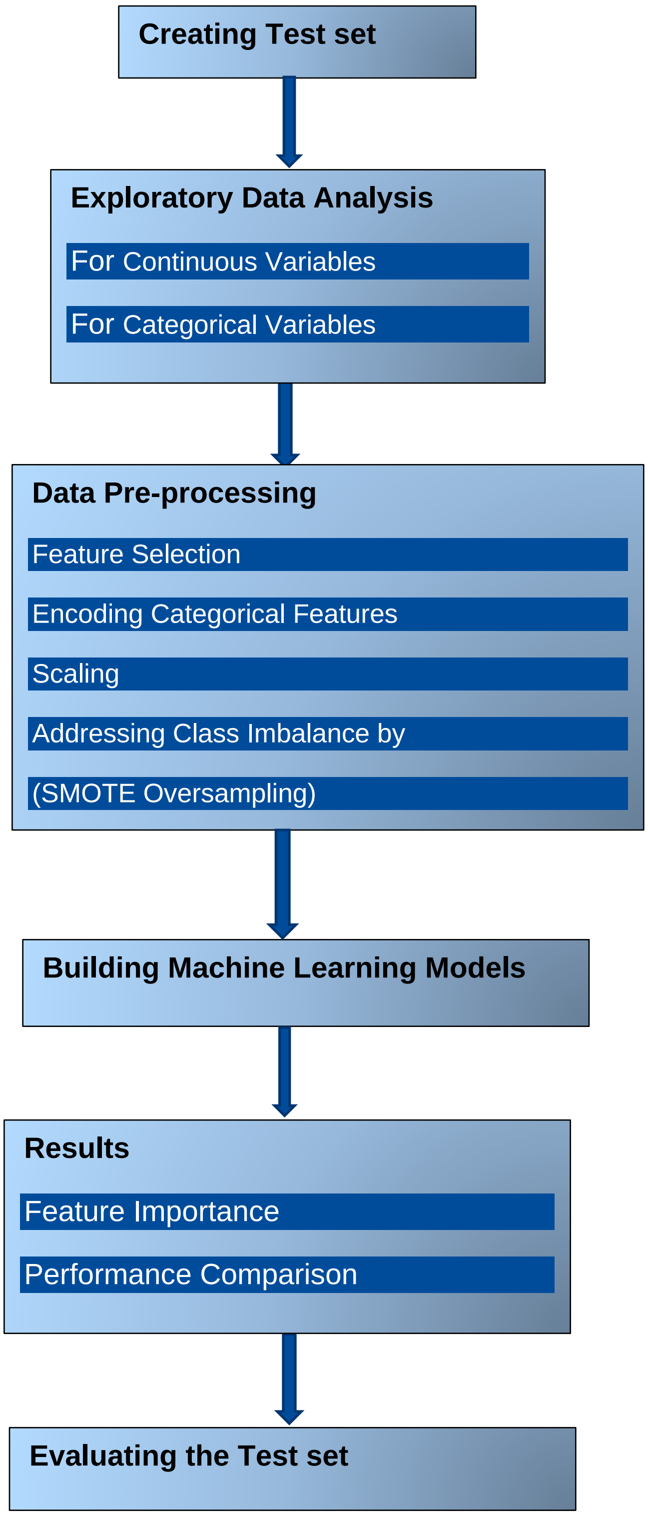
Huang et al.(2015) applied various classifiers for the customer churn Prediction and the results confirmed that random forest gives maximum accuracy compared to others in terms of AUC and PR-AUC analysis. They suggested that accuracy can be further improved applying the feature extraction optimization techniques.

Lalwani et Al. (2021) used the famous machine learning methods such as Logistic Regression, Naïve Bayes, SVM, Decision Trees, Random Forest, XGBoost and CatBoost Classifier, AdaBoost Classifier and Extra tree Classifier. The results reflect that the ensemble learning techniques, Adaboost and XGBoost classifiers performed superior than others in terms of AUC accuracy with the score of 84% for the churn prediction. They also performed superior compared to other algorithms in terms of all the performance measures including accuracy, precision, F-measure, recall and AUC score.

Vasimalla & Rahman (2020) applied KNN, SVM, Decision Tree and Random Forest classifiers and the result in different classifiers were compared over the selected features by various feature selection methods. As previously mentioned they used 2 different geature selection techniques mRMR and Relief. The best result was obtained from RF classifier together with oversampling with the score of 95.74% and Feature selection methods had nothing to do with RF and Decision Tree classifiers. It was observed that feature reduction in feature selection is decreasing the prediction score of tree classifiers. Another result achieved was that unlike other three classifiers, oversampling is decreasing the accuracy score in SVM.

In this study also the most common machine learning methods such as Logistic Regression, Random Forest, Support Vector Machines(SVM), … will be applied and they will be evaluated by performing k-fold cross-validation. Since correctly classifying the customers who will churn or positive class is more critical, so in this project, the focus will be more on recall for optimising our models as the scoring measure. Providing Confusion matrix and learning curves will help us visualise the training size impact on the errors.

## Methodology Summary



## Rererences:

Aljoumaa K, Ahmad AK, Jafar A (2019) Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data 6(1):28

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J.: (2015) Telco churn prediction with big data. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp. 607–618

K. Vasimalla; M. Rahman(2020) Machine Learning Based Customer Churn Prediction In Banking; Fourth International Conference on Electronics, Communication and Aerospace Technology

L. Beretta and A. Santaniello (2011), "Implementing relieff filters to extract meaningful features from genetic lifetime datasets," Journal of biomedical informatics, vol. 44, no. 2, pp. 361–369,

L. Geilerab, S. Affeldtb (2022) An effective strategy for churn prediction and customer profiling, Data & Knowledge Engineering

P. Lalwani, M. Mishra, J. Chadha, P. Sethi (2021), Customer churn prediction system: a machine learning approach

P. J. Huang (2015) Classification of Imbalanced Data Using Synthetic Over-Sampling Techniques, A thesis submitted for the degree Master of Science in Statistics

S. De; P. Prabu; J. Paulose (2021), Effective ML Techniques to Predict Customer Churn,

Third International Conference on Inventive Research in Computing Applications (ICIRCA)

Y. Jiang and C. Li (2015) "mrmr-based feature selection for classification of cotton foreign matter using hyperspectral imaging," Computers and Electronics in Agriculture, vol. 119, pp. 191–200,

https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

https://www.naturalspublishing.com/files/published/yt9r868jnrv116.pdf

https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

https://towardsdatascience.com/intro-to-feature-selection-methods-for-data-science-4cae2178a00a#:~:text=There%20are%20three%20types%20of,with%20examples%20in%20Python%20below.

Dataset Reference

https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers