# **Churn for Bank Customers - Report**

## **Muthukumar Kadhirvel**

## Topic

## Churn for Bank Customers - Predict whether a bank customer will leave the bank or not.

## Background/History

I have always wondered how businesses retain customers and in the case of banks with so many options available, retention of existing customers takes a vital role in keeping the banks profitable. I come from a Financial Sector background and know that retaining customers and increasing their lifetime value is one of the most important aspects of profitability. Churn is defined as a degree of customer inactivity over a given time. So, predicting churn is an important aspect of this when customer feedback is absent. (Sethuraman 2022).

## Business Problem

For a bank to be successful they need to retain existing customers and be on the lookout for any new customers. Acquiring new customers are costlier than retaining existing customers so this is a crucial factor. So, if we can predict that an existing customer might potentially leave the bank, we can develop loyalty programs to retain those customers. I am going to use Data Analytics to build a predictive model that would analyse and predict whether a customer would leave the bank or not based on the available features. (Krook 2022).

## Data Explanation

I have downloaded the churn dataset from the below Kaggle URL that lists ten thousand customers from France, Spain and Germany and the target – Exited variable of 0 (Not left the bank) or 1 (Left the bank).

<https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers>

The dataset has ten thousand rows and fourteen columns out of which one variable is the target – Exited variable. I will use this data to split into training, test data, and develop models to predict the target variable of whether the customer left the bank or not.

As part of data pre-processing, I will drop the attributes of RowNumber, CustomerID and Surname that would not be needed in the model prediction. Since there are no missing or NULL values, we need not handle them and replace them with mean/median etc. Since Geography and Gender variables are categorical, we need to create dummy variables for them. I will then perform EDA on the dataset to visualize the attributes so that we see a trend in them. Also, will create a correlation matrix between the attribute variables to infer the correlation between the variables.

Figure 1: The below pie chart shows the % of customers who exited the bank.

Chart, pie chart

Description automatically generated

We can see that we have around 20% of the surveyed customers exited the bank.

Figure 2: The below histogram shows the relationship between Age and the number of customers.

Chart, bar chart

Description automatically generated

As we can see that customer with age between 50 and 60 exited the bank more than other age classes.

Figure 3: The below histogram shows the relationship between Gender and the number of customers.

A picture containing graphical user interface

Description automatically generated

As we can see that female customers leave the bank more than the male customers.

Figure 4: The below Heatmap shows the relation between variables.

Chart, table

Description automatically generated

We can see from the Correlation Matrix that there are no clear variables that have a good correlation with target variable. Also, there are no variables that have a good correlation with each other either.

## Methods

I will first split the data into training dataset and test data in the ratio of 80% to 20%. Then will train and test various prediction models such as Logistic Regression/KNN – K Nearest Neighbours/Support Vector Machine Model/Random Forest Classifier etc. to find the performance using Accuracy/Confusion Matrix etc. to arrive at the best model to be used.

## Analysis

Since ours is a binary classification problem, accuracy is the best metric to be used. Along with accuracy we will consider Confusion Matrix also as a metric to arrive at the best model.

Initially we started off with the basic Logistic Regression model and got a 79.75 % Accuracy which is good but still can be improved. The confusion matrix also had more False Negatives than what we expected.

Then we tried with KNN model which can be used with both classification and regression problems and the accuracy reduced to 76.4%. The confusion matrix yielded similar False Negatives, but the number of False positives increased by three times.

Then we moved on to Support Vector Machine model and the accuracy reduced drastically to 57.85% and the number of False Positives jumped sixfold compared to KNN. But this is expected as SVM performs well when there are large features but less training data.

Random Forest Classifier model always performs better for classification problems and as expected the accuracy improved to 85.4%. The confusion matrix yielded far less false responses compared to Logistic Regression as well.

Since the various features are of varying ranges, we try applying MinMaxScaler on the Random Forest Classifier model as it would transform features by scaling each feature to a given range. The accuracy increased to 86.6% and the confusion matrix yielded lesser false responses compared to default Random Forest Classifier model.

Figure 5: The below confusion matrix shows the results post the execution of Random Forest Classifier with MinMaxScaler applied.

Chart, waterfall chart

Description automatically generated

Figure 6: The below bar plot shows the top 10 Feature Importance scores against Target Variable.

Chart, bar chart

Description automatically generated

We can see that Age has the most feature importance.

## Conclusion

Since Random Forest Classifier with MinMaxScaler applied has the best accuracy and the confusion matrix, we choose that model and go ahead with implementation.

## Assumptions/Limitations

There could be various other factors that impact Churn for Bank Customers like Accessibility/Digitalization etc. but we do not have those factors in our dataset. But those variables are difficult to quantify and so we go ahead with our data modelling with the available features.

## Challenges

The dataset has only thirteen feature variables out of which three are useless and hence dropped and so we are just left with ten features. The remaining ten features might not be covering all the important aspects of decision making for the customer if they would leave the bank or not. But since the accuracy and the confusion matrix was having reliable results, we look good.

## Future Uses/Additional Applications

We can use this modelling technique with any other datasets that calculate the churn for customers of various institutions as well if the question we are answering is a binary classification problem. For example, Internet companies/Phone companies etc. can use this similar model technique with appropriate features to predict whether their customers would move on to a different provider.

## Implementation Plan/Recommendations

We can implement this model to target customers which the model predicts will exit the bank and give special focus to them. Loyalty programs can be offered to those customers to retain them and make sure the bank’s resources are spent wisely.

## Ethical Assessment

The dataset has rows pertaining to Credit Score/Balance/Estimated Salary etc. that are sensitive information, but the dataset is a sample dataset that is used for data analytics education purposes and so there are no ethical issues with this data.

## Appendix

Reasons for Customer Churn (Lappeman, J, 2022)Chart, sunburst chart

Description automatically generated

## References

Sethuraman, S. (2022, November 2). *Why customers leave & what can banks do?* Tiger Analytics. Retrieved April 16, 2023, from <https://www.tigeranalytics.com/blog/addressing-customer-churn-in-banking/>

Krook, S. (2022, March 7). *What is customer churn prediction and why is it important?* Avaus. Retrieved April 16, 2023, from <https://www.avaus.com/blog/predicting-customer-churn/>

Lappeman, J., Franco, M., Warner, V., & Sierra-Rubia, L. (2022, May 26). *What social media sentiment tells us about why customers churn*. Journal of Consumer Marketing. Retrieved April 30, 2023, from <https://www.emerald.com/insight/content/doi/10.1108/JCM-12-2019-3540/full/html>