Predicting Bank Customer Churn Rate Using Regression and ANN

Contents

[Introduction 1](#_Toc183693666)

[The Data Set 1](#_Toc183693667)

[Data Cleaning and Exploration 1](#_Toc183693668)

[Data Preprocessing 3](#_Toc183693669)

[Building a Regression Model 4](#_Toc183693670)

[Business Insights and Recommendation 4](#_Toc183693671)

[Building the Artificial Neural Network 4](#_Toc183693672)

[Making prediction 4](#_Toc183693673)

[Business Solutions 4](#_Toc183693674)

# Introduction

The joy of every business is to impact people’s everyday lives positively while aiming to maximize profits. The bank is not exempted from this rule and that is why every additional customer to the bank matters a lot and every customer loss is a potential profit loss, which explains why data science plays a crucial role in the banking sector. This report focuses on using data science models to gain insights into bank customer behaviour and suggesting on ways to reduce churn rate while keeping retention high. I discuss the process and methods used in the analysis are discussed below.

The Data Set: The Dataset used in this analysis is titled ‘Credit Card customers’ and was gotten from the Kaggle open-source library. The dataset is about a bank manager who is worried about why people are opting out of the bank’s credit card services and would be more than happy if some can predict a customer is customer who is likely to churn and thereby suggest means and ways to make the client remain with the bank.

Upon the Analysts’ understanding of the business needs and with the dataset at hand, the researcher loaded the dataset using the Pandas library in the Jupyter Notebook where the dataset is to be assessed further to effectively suggest a feasible business models to increase customer retention. The Analyst loaded the dataset into a panda’s data frame to get a glimpse of the data and understand what each column represents.

Data Cleaning and Exploration: To get the best analysis out of our dataset, it is important to work with a properly cleaned datasets. A clean dataset ensures that all missing values are dealt with professionally, duplicated values are deleted and many more. For the credit card dataset, the first thing the Data Scientist did was to drop unwanted columns such as the ClinetNum and the Naïve\_Bayes column. The reason for dropping the columns was that they play no active or passive role in the analysis to be done and they do not affect the outcome of our model or prediction, hence they were dropped.

Additionally, the scientist used the. isna().sum() method to check if there were any missing values in the dataset and can be dealt with accordingly. Per the dataset, there was no missing values. To ensure the dataset is properly cleaned and ready for use, the Analyst check for the data types of every column to ensure that they are all in the right data type, the reason for using this method is to improve the data quality. Upon cleaning the dataset, the analyst proceeded to Exploring the Dataset.

This is a crucial step in understanding the structure of the dataset. The researcher adopted this technique to gain full insights into the data and determine which column is the target variable. For this dataset, the target variable is the Attrition\_Flag and it is distributed as: Existing Customer = 8500, Attrited Customer =1627. With the distribution understood, the task at hand how to use data science to understand the behaviour of these customers and create to model for future prediction of likely attrited customers so that they can be targeted by a customized marketing plan.

Another important method the analysts used is the describe () function, this gives insights into the statistics of the data such as the mean and maximum age of customer, the customer with the highest credit limit and the lowest as well. By conducting Data cleaning and exploration, the researcher has gotten a proper insight of the data set but still not enough to make any inform decision. To get a clearer view the researcher graphically represented the heatmap to understand the relationships between the variables. The allows us to understand which variables are most correlated. See picture below.

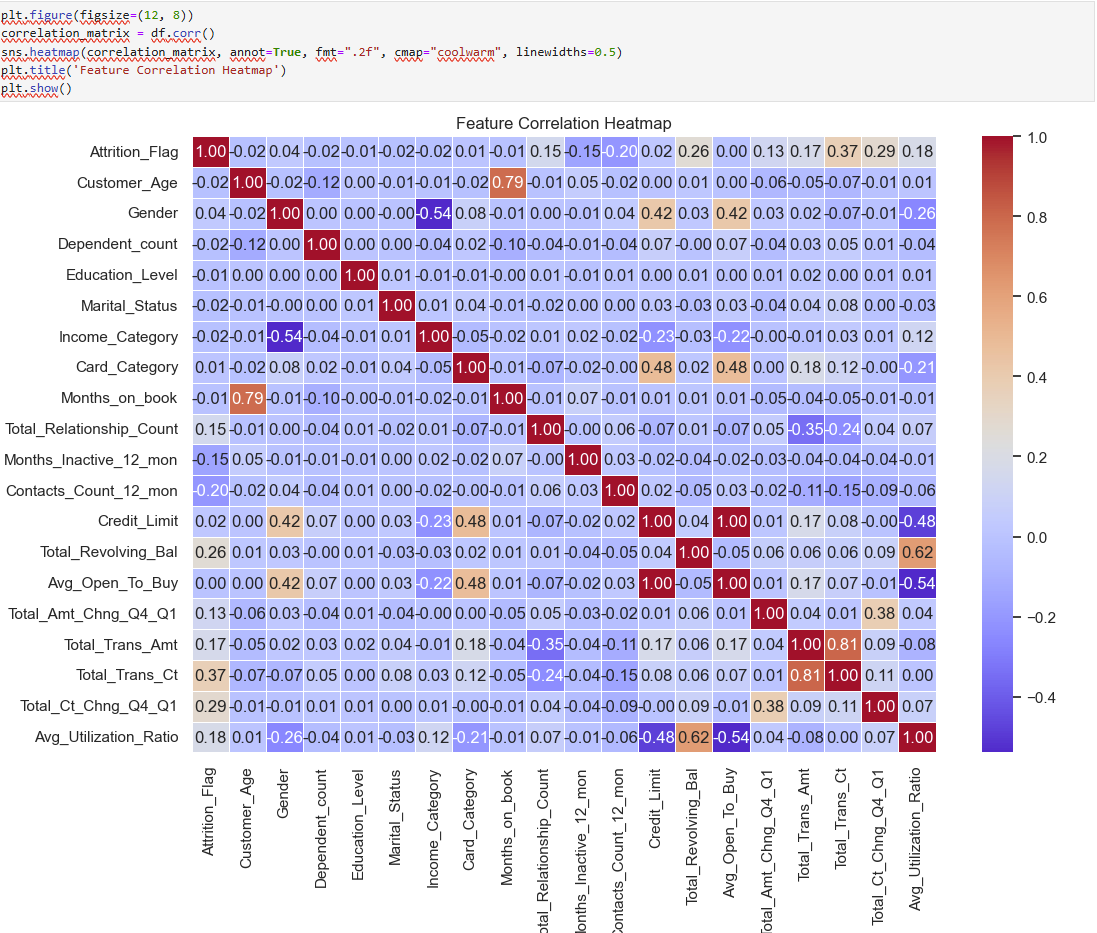


Figure 1Heat Map of the Bank credit card customer

By observing the image above clearly, we can see that

1. Attrition\_Flag Correlations (Target Variable):

Total\_Trans\_Ct: Strongest positive correlation (0.37) with customer churn. This indicates that customers with fewer transactions are more likely to churn.

Total\_Ct\_Chng\_Q4\_Q1: Positive correlation (0.29) suggests that a drop in transaction count across quarters is associated with churn.

Total\_Revolving\_Bal and Avg\_Utilization\_Ratio: Moderate positive correlations (0.26 and 0.18, respectively) indicate that higher balances or utilization ratios may lead to increased churn.

1. Highly Correlated Features:

Avg\_Open\_To\_Buy and Credit\_Limit: High positive correlation (0.48) indicates that credit limit strongly influences the open-to-buy amount.

Total\_Trans\_Ct and Total\_Trans\_Amt: Strong positive correlation (0.62) shows that transaction count is proportional to the transaction amount.

1. Negative Correlations:

Contacts\_Count\_12\_mon and Attrition\_Flag: Negative correlation (-0.20) suggests that higher customer interactions may reduce churn.

Months\_Inactive\_12\_mon: Negative correlation with Total\_Relationship\_Count (-0.35), indicating that inactive customers tend to have fewer relationships with the bank.

1. Independent Features:

* Features like Education\_Level and Marital\_Status show negligible correlations with other features or the target, making them potentially less influential.

# Data Preprocessing

Encoding categorical columns - For our dataset, there are a few columns that needs to be encoded. Columns such as, ('Gender', 'Education\_Level', 'Marital\_Status', 'Income\_Category', 'Card\_Category', 'Attrition\_Flag'). These were encoded using the LabelEncoder class. For example, the Attrition\_Flag was encoded into 0 = Attrited Customer and 1 = Existing customer. This becomes easy for identification and we can easily know what each number stands for

Based on the nature of the Data Set we have and having studied the data carefully, the data scientist believed that the next thing is to Normalize the data. In the context of data science, Normalization is the process of scaling data to fit within a specific range or to have certain statistical properties, like a mean of 0 and a standard deviation of 1. This is a very important step because it will improve the performance of the model that we are about to build and ensures that all our numerical variables are stable. Normalization is only done one numerical value.

The table below shows our data before and after they had been normalized.

|  |  |  |  |
| --- | --- | --- | --- |
| Customer\_Age | Credit\_Limit | Total\_Trans\_Amt | Avg\_Utilization\_Ratio |
| 45 | 15000.0 | 5000 | 0.25 |
| 50 | 12000.0 | 7000 | 0.33 |

Figure 2 Data before Normalization

|  |  |  |  |
| --- | --- | --- | --- |
| Customer\_Age | Credit\_Limit | Total\_Trans\_Amt | Avg\_Utilization\_Ratio |
| 2.25 | 0.60 | 0.45 | 0.20 |
| 0.40 | 0.48 | 0.63 | 0.28 |

Figure 3 Data After Normalization

# Building a Regression Model

Business Insights and Recommendation.

# Building the Artificial Neural Network

# Making prediction

# Business Solutions