A MINOR PROJECT REPORT

ON

# CUSTOMER CHURN PREDICTION

SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF DEGREE OF

**BACHELOR OF TECHNOLOGY**

**IN**

**ELECTRONICS AND COMMUNICATION ENGINEERING**


**Submitted By:**

**AMANN ANAND (9918102150)**

**Under the Guidance Of**

**MR. VINAY ANAND TIKKIWAL**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**

**JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA (U.P.)**

**May, 2021**

# CERTIFICATE

This is to certify that the minor project report entitled, "**Customer Churn Prediction**" submitted by **Amann Anand** in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in **Electronics and Communication Engineering** of the Jaypee Institute of Information Technology, Noida is an authentic work carried out by them under my supervision and guidance. The matter embodied in this report is original and has not been submitted for the award of any other degree.

**Signature of Supervisor:**
**Name of the Supervisor: Mr. Vinay Anand Tikkiwal**
**ECE Department,**
**JIIT, Sec-128,**
**Noida-201304**

**Dated: 03/05/2021**

# DECLARATION

We hereby declare that this written submission represents our own ideas in our own words and where others' ideas or words have been included, have been adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission.

Place:

Date: 03/05/2021

Name: Amann Anand

Enrollment: 9918102150

# ACKNOWLEDGEMENT

On the very outset of this report, we would like to express our sincere and heartfelt obligations to all the personages who have helped us in this endeavour. Without their active guidance, help, cooperation, and encouragement, we would not have made headway in this project.

We are ineffably indebted to **Mr. Vinay Anand Tikkiwal** for his constant guidance and encouragement to accomplish this assignment. We are extremely thankful to him for his valuable guidance and support for the completion of this project.

We extend our gratitude to the **Department of Electronics and Communication, Jaypee Institute of Information Technology** for giving us this opportunity. We also acknowledge with a deep sense of reverence, our gratitude towards our parents, members of family, and friends who have always supported us morally.

# TABLE OF CONTENTS

# List of Figures and Tables

# Abstract

Churn prediction means detecting which customers are likely to cancel a subscription to a service, based on how they use the service. Churn rate is an indicator for businesses whose customers are subscribers and paying for services on a recurring basis. How to keep customer loyalty and prevent customer churn is an important problem for businesses like banking as different customers exhibit different behaviours and preferences, so they cancel their subscriptions for various reasons. It is a critical prediction for many businesses because acquiring new clients often costs more than retaining existing ones. This study utilizes the Kaggle dataset which contains many features of Customer demographics, Transaction data, and Bank branch data required for predicting customer churn. Exploratory data analysis (EDA) was performed to select the best features, further models were trained using logistic regression and evaluated using different evaluation metrics. Our RFE model performed the best with high precision, recall, and AUC score which can help in identifying those customers that are at risk of cancelling, so the required marketing action can be taken for each customer to maximize the chances of retaining the customers.

# 1. Introduction

Customer churn occurs whenever a customer stops taking the services of service providing company. It is a big issue for many businesses because acquiring new customers often costs more than retaining existing ones. As a result, tools for developing and implementing customer retention models (churn models) are essential Business Intelligence applications. Churning can occur as a result of low customer satisfaction, aggressive competitive strategies, new products, regulations, and other factors in today's dynamic market environment. Churn models are designed to detect early churn signals and identify customers who are more likely to leave voluntarily and transfer their businesses to a competitor.

Customer churn here can be defined as the movement of people from one bank to another bank. The main reasons for churn are dissatisfaction with the customer service, high costs, unattractive plans, bad support. It is an expensive problem in many sectors since acquiring new customer costs five to six times more than retaining existing ones [1]. The ability to predict that a specific customer is at a high risk of churning represents a huge additional potential revenue source for every business. In addition to the direct loss of revenue that results from a customer leaving the business, the costs of initially acquiring that customer may not have been covered by the customer's spending to date. The aim of customer churn prediction is to detect customers with high tendency to leave a company. In order to retain the existing customers, the banking industry need to know the reasons of churn, which can be found through the knowledge extracted from gathered data.

Due to the customer churn prediction model's great utility, many sectors are using this model to analyze their shortcomings in their customer service. One of its application is in the telecom industry. Nowadays, because of a highly competitive market and a wide range of products/services many big telecom firms are already utilizing machine learning for reducing churn rate. Today even smaller telecom companies and start-ups try AI applications right after their services enter the market. Customer churn also happens when a client stops buying a retailer's products, avoids visiting a particular retail store, and prefers switching to the competitor. Therefore, big retail firms are also putting this technique to minimize the churning rate of their customers. It is also finding great use in subscription-based services like Netflix, Amazon Prime Video, etc.

Customer churn has become a big issue in many banks [2] because it costs a lot more to acquire a new customer than retaining existing ones. With the use of a customer churn prediction model possible churners in a bank can be identified, and as a result the bank can take some action to prevent them from leaving. For example, UBS, the largest Swiss bank, utilized ML algorithms to avoid client churn after HSBC, its main

European competitor, launched several ads campaigns aimed at billionaires who were already with UBS for a long time.

Most of the churn prediction modelling methods rely on quantifying risk based on static data and metrics, i.e., customer information if it exists right now. These methods have some value and can identify a certain percentage of at-risk customers, but they are relatively inaccurate and result in money being lost.

For our project we have taken the dataset of a bank with 225,000 customers, which is facing the problem of stagnating customer balances since the last 3 months Through customer churn prediction we need to identify what customer segments are likely to churn balances in the next quarter by at least 50% considering the current quarter. For this purpose, we do extensive Exploratory data analysis (EDA) to select the best features for implementing logistic regression on our models. Finally, the models are evaluated using different evaluation metrics.

The remaining contents of this paper are organized as follows. Section 3 summarizes the related work done in customer churn prediction. Section 4 gives the description of the methodology used. In Section 5 Discussion is done on the results obtained, while Section 6 concludes the work presented in the paper.

# 2. Literature Survey

Michel Ballings and Dirk Van den Poel [3] carried out research on customer event history to predict customer churn. In their work they used logistic regression, classification trees and classification trees in combination with bagging to study the relation between the length of customer event history and classification performance, with AUC metric to compare model's performance. Finally, concluded that the length of the predictors period is logarithmically related to classification performance.

Chen et al. [4] proposed a novel data mining technique called hierarchical multiple kernel support vector machine (H-MK-SVM) for customer churn prediction using longitudinal behavioural data. The AUC of the H-MK-SVM, the MK SVM, the SVM and the LS-SVM were all larger than 90% for = 1, 2, 5 and 15 with H-MK-SVM as the proposed model having the highest AUC among the four models.

Statistically based performance measures are typically used to evaluate churn prediction models, resulting in inefficient model selection. Verbeke et al. [5] conducted research on this subject/area and developed a novel, profit centric performance measure that is developed, by calculating the maximum profit that can be generated. The findings showed that applying the maximum profit criterion to a retention campaign can have a significant impact on the profits generated.

Nowadays, gambling companies are searching for new alternative ways of building and retaining customer relationships. Churn prediction modelling is one of these alternatives, and it ranks customers from most to least likely to churn based on the allocation of a churn probability to each customer in the customer database. Coussement K. and De Bock K.W [6] carried out research to show the beneficial impact of ensemble algorithms over single prediction models in this customer churn prediction setting. They used various techniques and tools such as decision trees, random forests, generalized additive models for this research. The findings showed that ensemble algorithms RF and GAMens, are more robust and had better prediction performance, than the single algorithms, CART and GAM.

Moeyersomes J. and Martens D. [7] carried out research on including high cardinality attributes in predictive models used for determining customer churn. A unique data set of a large energy company with more than 1 million customers is used to prove the value of using high cardinal attributes for building predictive models. The various techniques and tools used are logistic regression, decision trees and SVM. The results in a churn prediction setting demonstrate that including such data can lead to substantial improvements in predictive performance.

Coussement K. et al. [8] performed A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction by using a case study in the telecom industry. Data preparation is a process that aims to convert independent (categorical and continuous) variables into a form appropriate for further analysis. This research examines data-preparation alternatives to enhance the prediction performance for the commonly-used logit models. For this research a dataset of Telecom company with 30104 customers is used. The various techniques and tools used are Logistic regression, decision trees, SVM, bagging, Bayesian network, Naïve Bayes and random forests. The analysts concluded that the data-preparation technique they choose actually affects churn prediction performance.

# 3. PROPOSED METHODOLOGY

After collection of data, we identified and type casted the variables and then performed exploratory data analysis (EDA) and collected insights from it to make our models. Further, we implemented logistic regression and compared the performance of our models using different evaluation metrics. The proposed workflow is represented in Figure 1.

This module is divided into 5 parts: 3.1) Univariate analysis, 3.2) Bi-Variate Analysis, 3.3) Multivariate analysis, 3.4) Logistic Regression, 3.5) Evaluation Metrics.

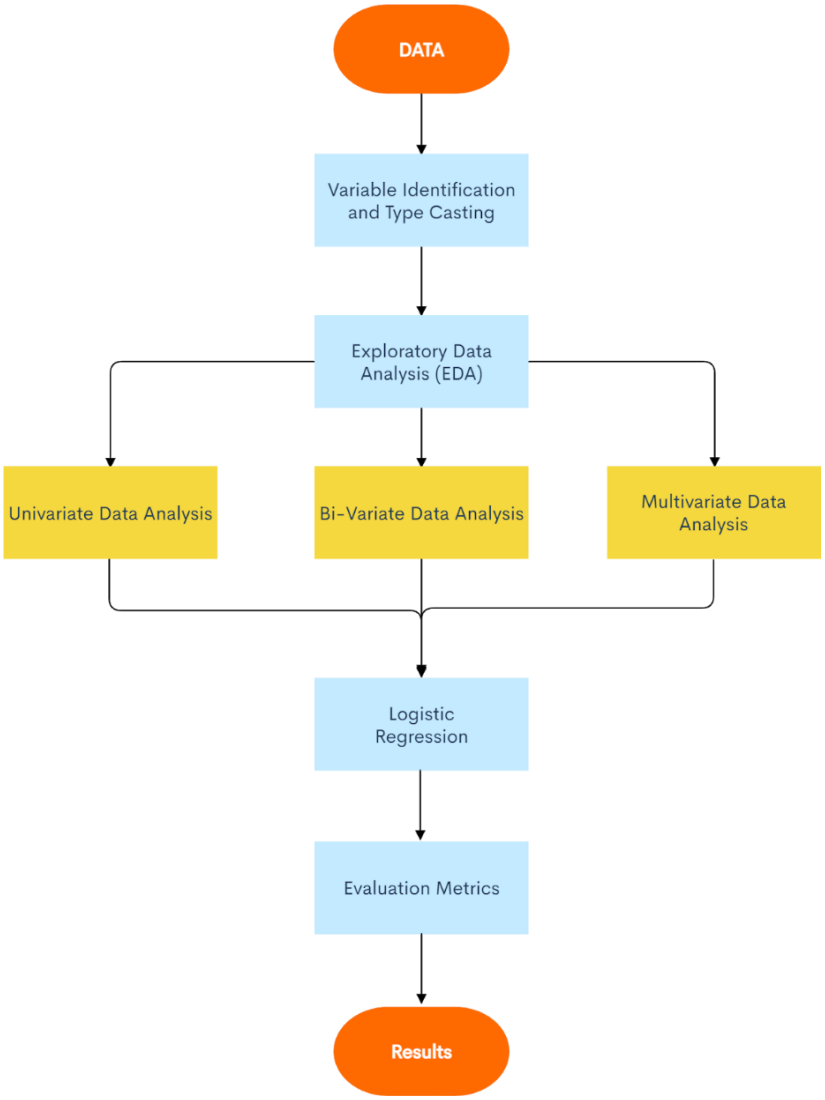


**Figure 1- Proposed Methodology Flowchart**

## 3.1 Univariate analysis

It is the most basic type of data analysis. "Uni" means "one", so our data contains only one variable. It does not deal with causes or relationships (unlike regression), and its primary goal is to extract insights from data based on a single variable.

## 3.2 Bi-Variate analysis

It is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. Our aim is to get insights from the relation of two variables. Data types can be broadly classified as categorical which contains nominal and ordinal data types and numerical variables which contain discrete and continuous data types. In our work, we will be getting insights from various relationships of features such as:

### 3.2.1 CONTINUOUS-CONTINUOUS VARIABLES

A continuous variable is a variable whose value is obtained by measuring, i.e., one which can take on an unlimited set of values. We will be exploring the relationship between the various continuous variables and hypothesis of whether transaction variables like credit/debit have correlation among themselves or balance variables for that matter.

### 3.2.2 CONTINUOUS-CATEGORICAL VARIABLES

Here, we will be exploring the relationship between continuous and categorical variables. A categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property. Some hypothesis we will be exploring are a) Are vintage customers less likely to churn? b) Are customers with higher average balance less likely to churn? c) Are customers with higher balance during the previous two quarters are likely to churn?

### 3.2.3 CATEGORICAL-CATEGORICAL VARIABLES

Here we will explore the relationship among categorical-categorical variables and various hypothesis like a)Are females less likely to churn than males?, b) Are young customers more likely to churn?, c) Customers

from low income bracket more likely to churn, d) Are customers with dependent(s) less likely to churn?, e) Customers whose last transaction was more than 6 months ago, do they have a higher churn rate?

## 3.3 Multivariate analysis

It is used to study more complex sets of data than what univariate analysis methods can handle. Here we will explore the relationship between various variables together and hypothesis like, 1) Gender, occupation and customer_nw_category with churning status. 2)Gender, occupation and current balance with churning status. 3) Age, occupation and churning status

## 3.4 Logistic Regression

Regression analysis is a statistical method for estimating the relationships between dependent and independent variables. It assumes that a relationship exists between the variables and finds the line of best fit. Regression analysis is not commonly used in consumer churning because linear regression models are only useful for predicting continuous values. Logistic Regression, also known as Logit Regression analysis (LR), is a type of probabilistic statistical classification model. The logistic regression is essentially an extension of a linear regression, only the predicted outcome value is between [0, 1]. It is also used to generate a binary prediction of a categorical variable (e.g., consumer churn) that is dependent on one or more predictor variables (e.g., customer characteristics).

The threshold value is commonly set to 0.5, so if the prediction is greater than 0.5, the output is classified as class 1, otherwise it is classified as class 0 (here class 0 refers to non-churners and 1 refers to churners). Logistic regression is accomplished by calculating the log odds of

$(\mathbf{p}_i/1 - \mathbf{p}_i)$ where P is the likelihood or probability of churning or not churning.

In the churn prediction problem, LR is typically used after proper data transformation on the initial data, and it performs very well **[9]**.

**Equations of Logistic Regression:**

$$Z=\beta x + b \quad (1)$$

$$y = \frac{1}{1+e^{-(\beta x+b)}} \quad (2)$$

$$Z = \log\left(\frac{p}{1-p}\right) \quad (3)$$

## 3.5 Evaluation Metrics

We use different measures to evaluate classifier performance in churn prediction for various schemes with their respective parameters.

Here, we look mainly at the recall value here because a customer falsely marked as churn would not be as bad as a customer who was not detected as a churning customer and appropriate measures were not taken by the bank to stop him/her from churning.

Our main metric here would be Recall values, while AUC ROC Score would take care of how well predicted probabilities are able to differentiate between the 2 classes while other metrics will be useful in Exploratory data analysis (EDA).

### 3.5.1 Precision:

Precision is a metric that measures out of all the positive predictions, how many are positive predictions. It is calculated by dividing the number of true positives by the total number of true positives and false positives.

Precision = TruePositives / (TruePositives + FalsePositives)

The result is a number ranging from 0.0 (no precision) to 1.0 (perfect precision).

### 3.5.2 Recall:

Recall is a metric that measures out of all actual positive, how many are predicted positive. The number of true positives divided by the total number of true positives and false negatives yields the recall value (e.g. it is the true positive rate).

Recall = TruePositives / (TruePositives + FalseNegatives)

The result is a value between 0.0 for no recall and 1.0 for full or perfect recall.

### 3.5.3 Confusion matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. In the field of machine learning and specifically the problem of statistical classification, a confusion matrix is also known as an error matrix. It is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix

represents the instances in a predicted class, while each column represents the instances in an actual class (or vice versa).

### 3.5.4 AUC-ROC curve:

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR (true positive ratios) against FPR (false positive ratio) at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

### 3.5.5 Two-sample t test:

The two-sample *t*-test (also known as the independent samples *t*-test) is a method used to test whether the unknown population means of two groups are equal or not. The t test tells us how significant the differences between groups are; In other words it lets you know if those differences (measured in means) could have happened by chance. This test can be used to for continuous-categorical variables.

T-tests are best performed when an experiment has a small sample size, less than 30. Also, t-tests assume the standard deviation is unknown.

$$t = \frac{m-\mu}{s/\sqrt{n}} \qquad (4)$$

Where, t= student t test, m=mean, μ=theoretical value, s=standard deviation, n=variable set size.

### 3.5.6 Two-sample z test:

A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. The test statistic is assumed to have a normal distribution, and parameters such as standard deviation should also be known in order for an accurate z-test to be performed. This test can be used to for continuous-categorical variables.

The z-test is best used for greater-than-30 samples because, under the central limit theorem, as the number of samples gets larger, the samples are considered to be approximately normally distributed.

$$z = \frac{(\bar{x_1}-\bar{x_2})-(u_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \qquad (5)$$

### 3.5.7 Chi Square Test

The chi square test is a useful test to conduct to help gauge the unexpectedness or expectedness of outcomes in data. The test, in particular, measures the expected distribution of data points into various labels against the actual distribution of those data points. This test can be used to for categorical-categorical variables. In the Chi-Square goodness of fit test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (6)$$

Where, $\chi$ = Chi-Square goodness of fit test O = observed value E= expected value

### 3.5.8 Spearman Correlation

Spearman's correlation measures the strength and direction of monotonic association between two variables. Monotonicity is "less restrictive" than that of a linear relationship.

$$\rho = 1 - \frac{\sum d_i^2}{n(n^2-1)} \qquad (7)$$

Where ρ= Spearman rank correlation, di= the difference between the ranks of corresponding variables, n= number of observations

### 3.5.9 K-fold Cross validation

In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness.

# 4. Discussion on results obtained

This project aims at predicting whether a customer will churn or not using logistic regression technique from a given customer dataset of a bank that was taken from Kaggle.

## 4.1 Univariate analysis

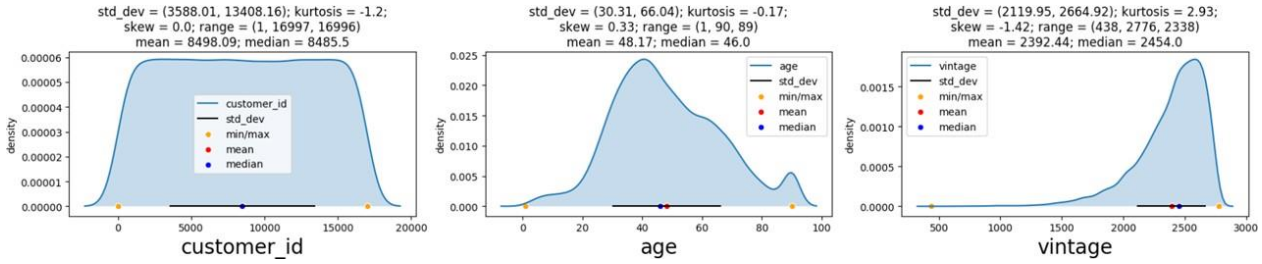


**Figure 2**

In Figure 2, customer _id has a uniform distribution as each customer is unique. As they don't contribute to any information this variable can be eliminated.

age variable analysis gave a median age of 46. Most of the customers were aged between 30-66 years. Its skewness was positive i.e., age of customer is negligibly biased towards younger age. A negative kurtosis meant its less likely to have extreme values.

vintage: most of the customers joined between 2100 and 2650 days from data extraction. It is left skewed, this variable is biases towards longer customer association. A positive Kurtosis indicated the presence of outliers.

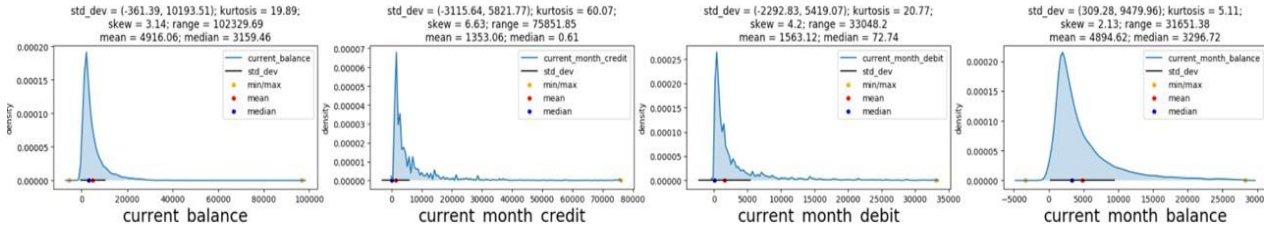


**Figure 3**

As shown in Figure 3, all the plots are right skewed which means that there are presence of outliers/extreme values.

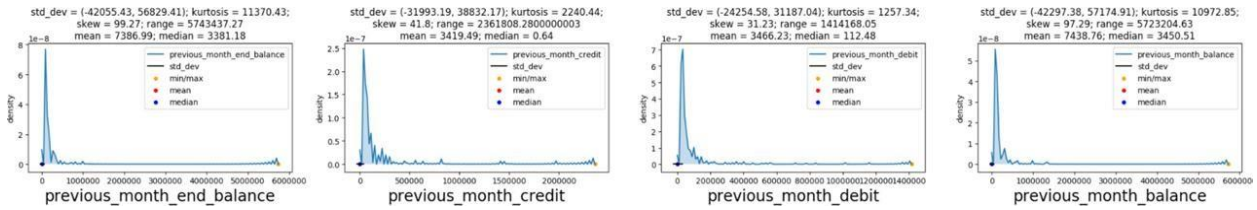


**Figure 4**

Figure 4 looks very similar to current_month. Most of the customers perform low amount transactions.

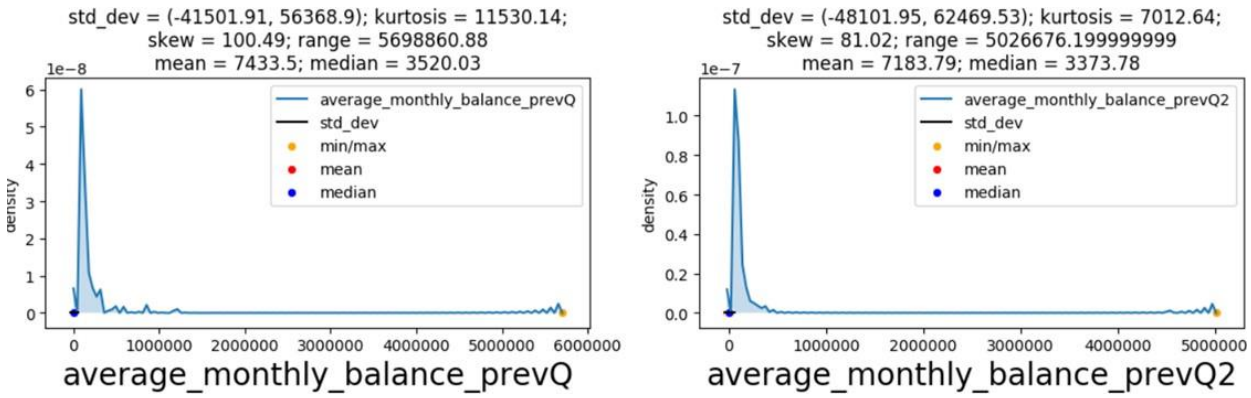


**Figure 5**

The plots are right skewed in Figure 5, hence presence of outliers.

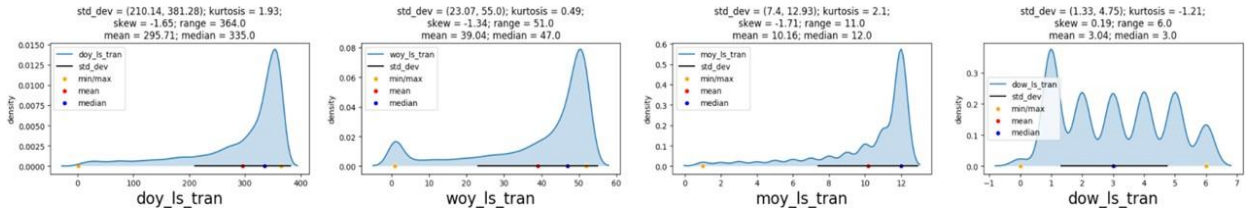


**Figure 6**

In Figure 6, day_of_year indicated most of the transactions were made in last 60 days of data extraction. There are even transactions which were made more than a year ago. Week_of_year and Month_of_year variables validate the findings from the day_of_year.

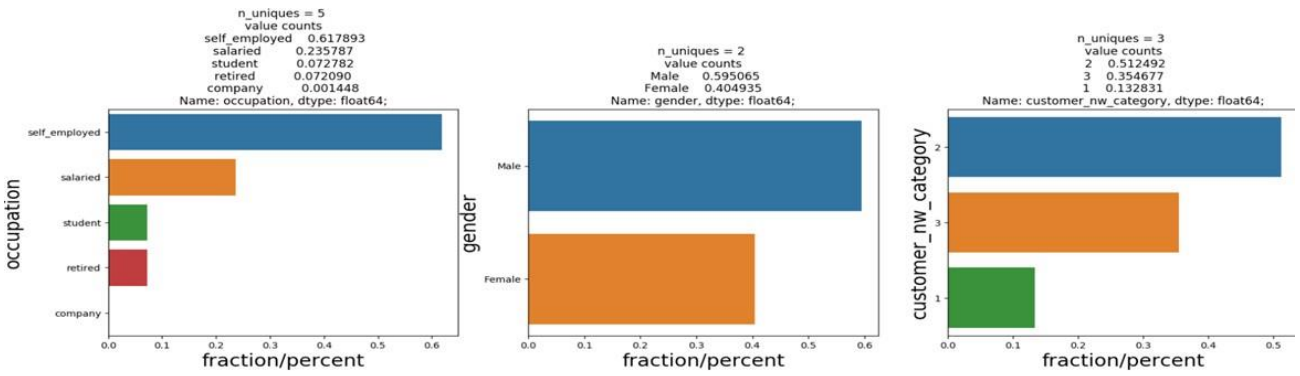


**Figure 7**

Figure 7 shows, that in occupation-the majority of people are self_employed. There are extremely few Company Accounts. Gender- Males accounts are 1.5 times in number than Female Accounts. customer_nw_category: Half of all the accounts belong to the 3rd net worth category. Less than 15% belong to the highest net worth category.

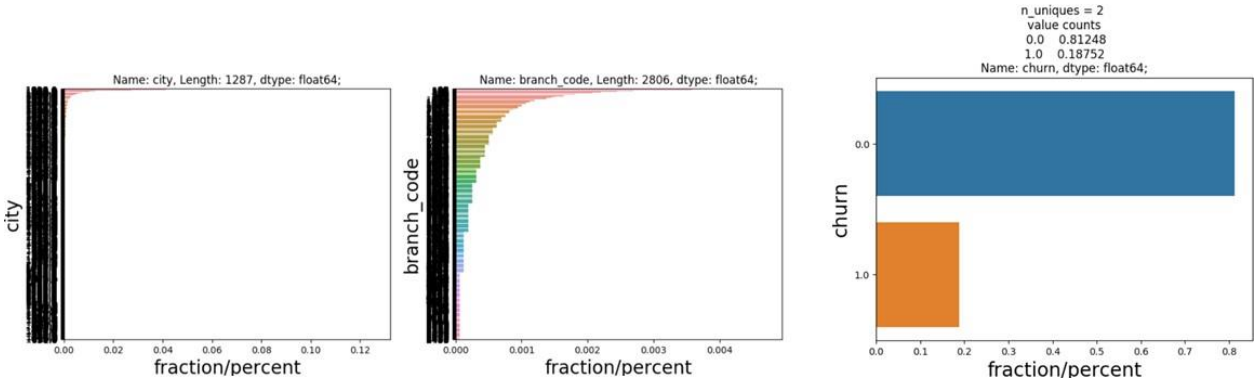


**Figure 8**

In figure 8, for both variable "city" and "branch_code", there are too many categories. The number of people who churned are 1/4 times of the people who did not churn in the given data.

## 4.2 BIVARIATE ANALYSIS

### 4.2.1 Continuous-Continuous Variable

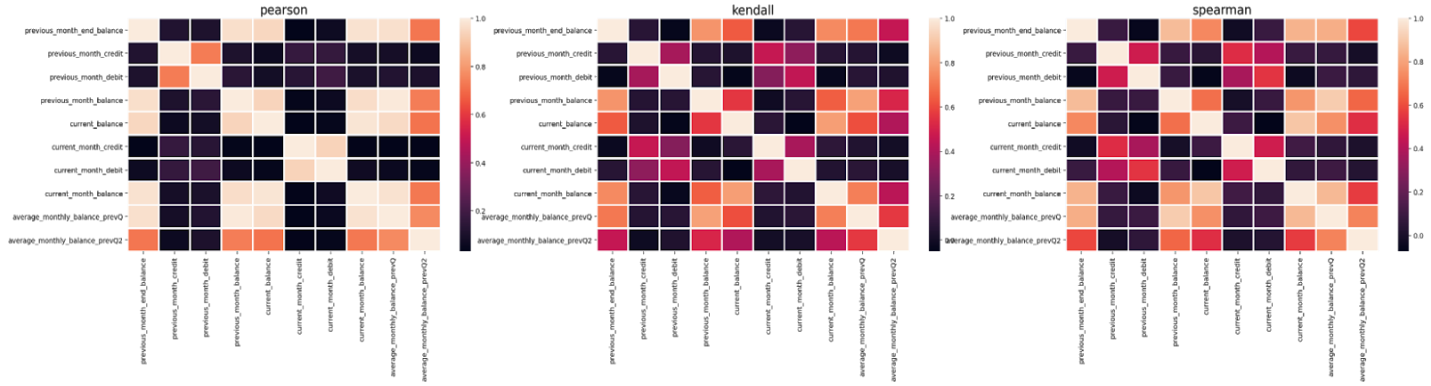


**Figure 9- Correlation Heat Map**

Figure 9, shows transaction variables like credit/debit have a strong correlation among themselves. Also, balance variables have strong correlation among themselves. Transaction variables like credit/debit have insignificant or no correlation with the Balance variables.

## 4.2.2 Continuous-Categorical

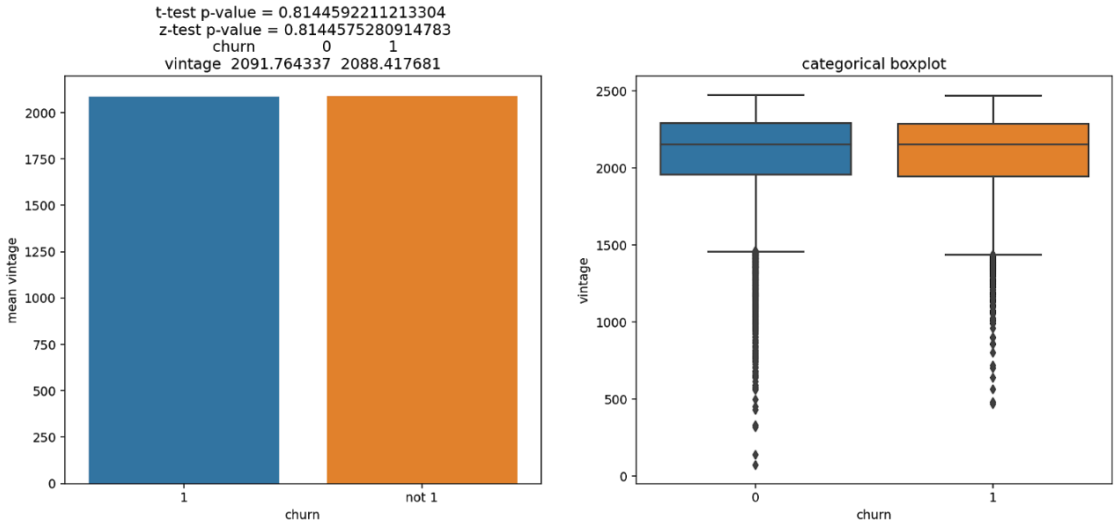


**Figure 10**

Figure 10, shows vintage customers churned more, but results are not significantly different. Boxplot shows very similar distribution with outliers on the lower end. So, we can safely reject the hypothesis that vintage customers are more likely to churn.

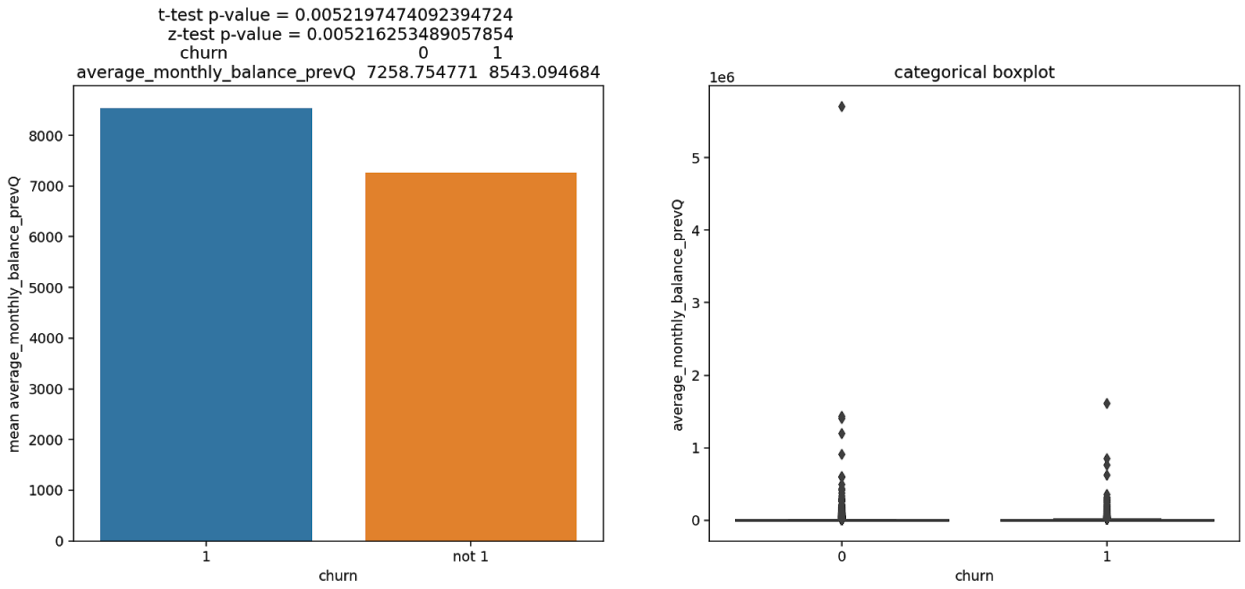


**Figure 11**

In Figure 11, we can see that customers who churned have significantly higher balance during the immediately preceding quarter.

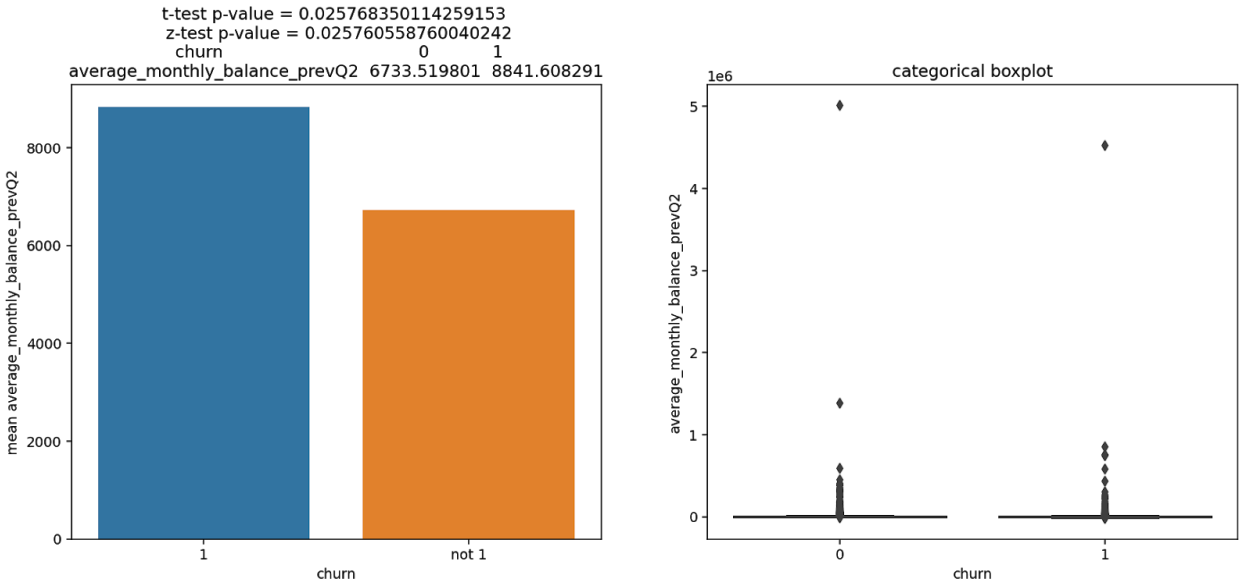


**Figure 12**

In Figure 12, we can see that people who churned actually had significantly higher balance during their previous two quarters.

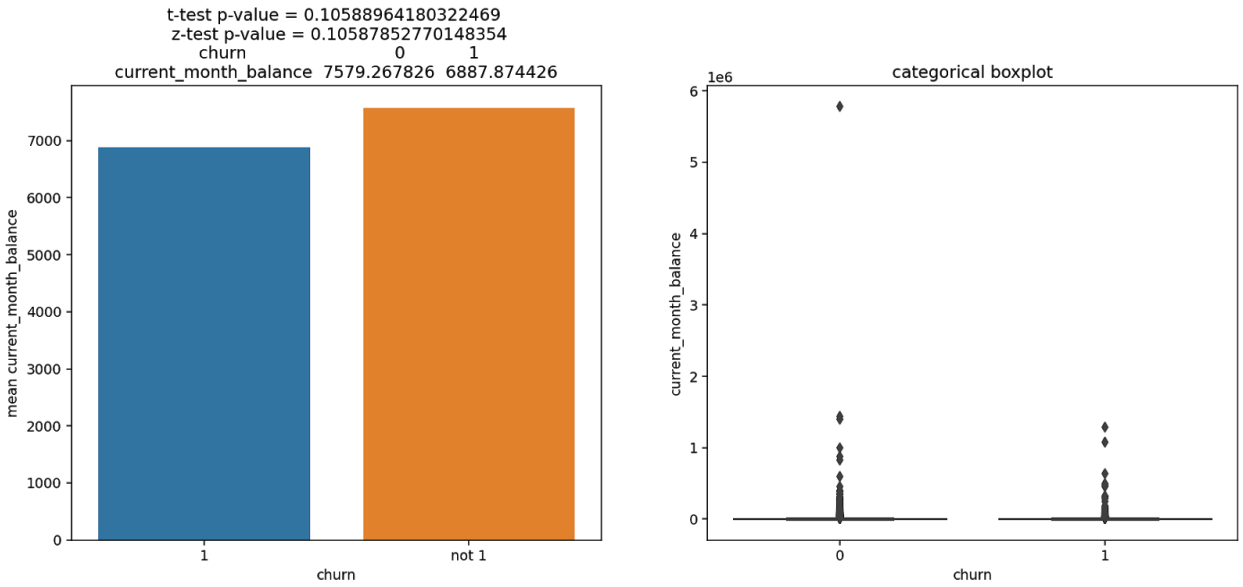


**Figure 13**

In Figure 13, we can see that the customers who churned had significantly high balance throughout the previous two quarters and previous month. But their average balance reduced significantly in the current month.

### 4.2.3 Categorical-categorical variables

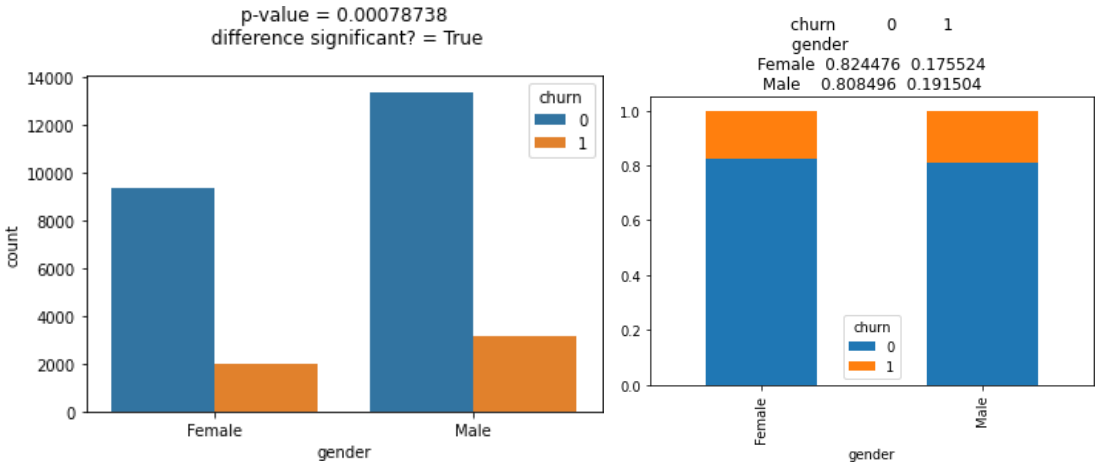


**Figure 14**

In Figure 14, we can see the difference between the males and females customer churning is significant.

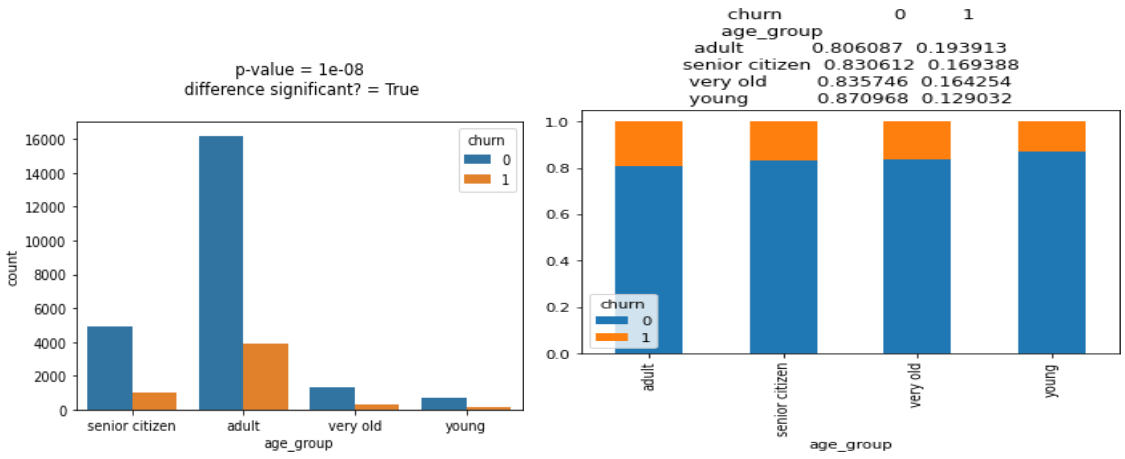


**Figure 15**

In Figure 15, we can see that the age group has significant effect on the churning rate.

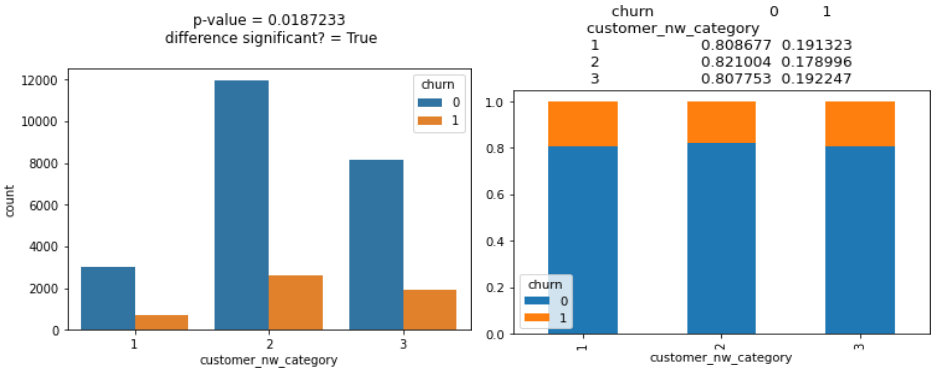


**Figure 16**

In Figure 16, we can see that the different income brackets have significant effect on the churn rate.

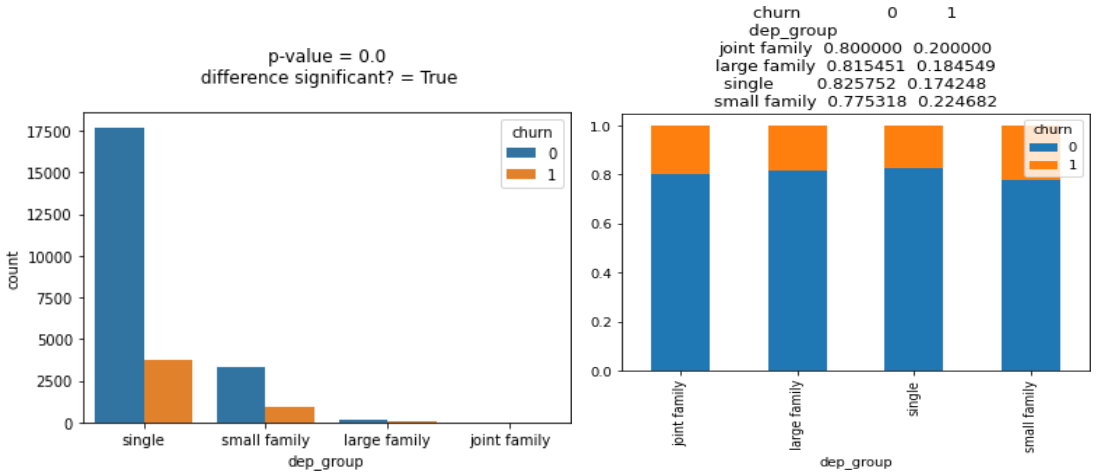


**Figure 17**

In Figure 17, we can see that the number of dependents also play a significant role in churning.

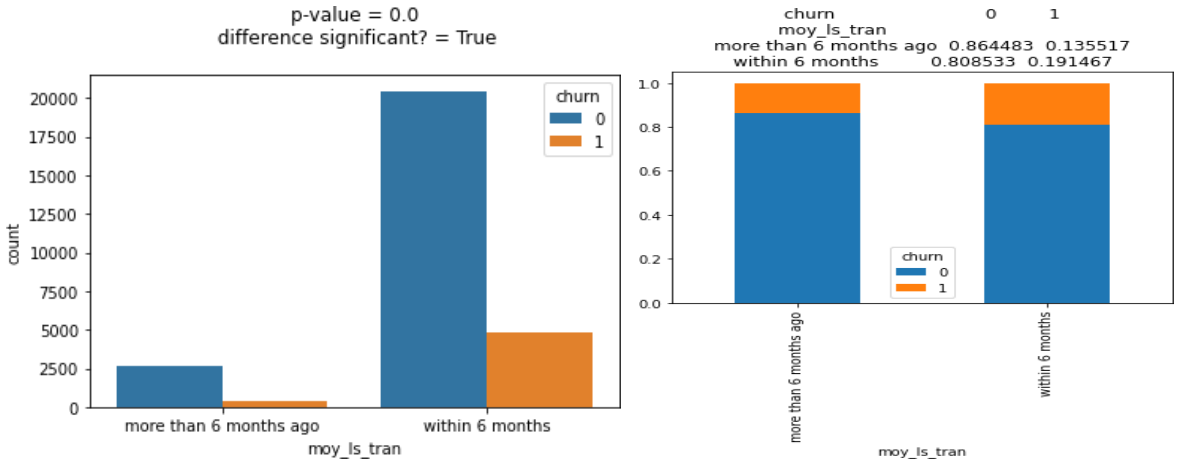


**Figure 18**

In Figure 18, we can see that there is a significant difference between the people who made their last transaction in the last 6 months and the customers who had their last transaction more than 6 months ago.

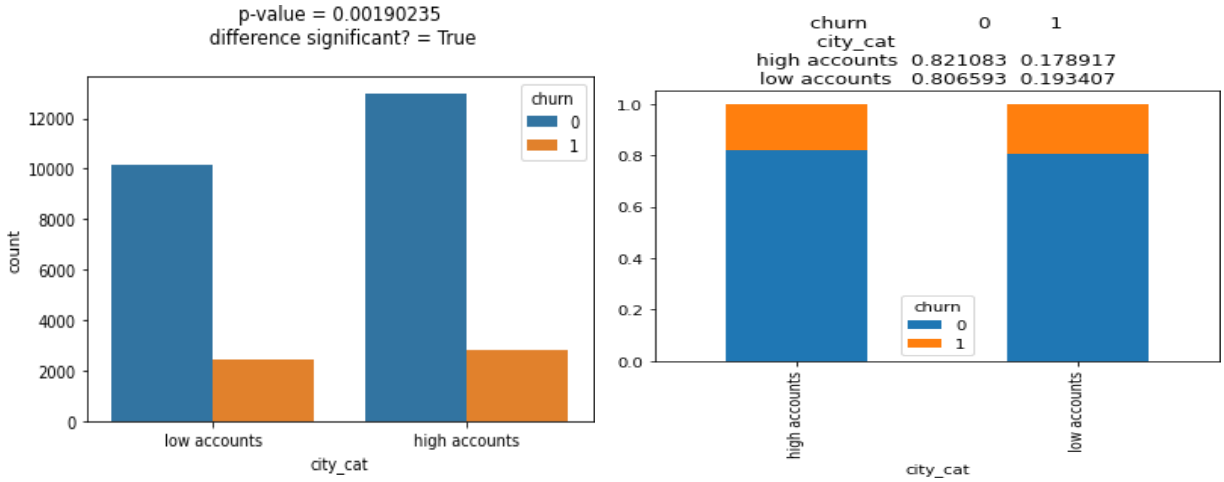


**Figure 19**

In Figure 19, we can see that the cities having less than 1 percent of the total have significantly different churn rates as compared to the cities with more accounts.

## 4.3 Multivariate Analysis

In multivariate analysis, we constructed a pivot table with various features and obtained the following results.

The Highest number of churning customers are those Male Customers who lie in 2 net worth category and belong to Self-employed profession. Proportion wise for net worth category 1, Approximately 22% Male customers who belong to the Self-employed profession are churning. Proportion wise for net worth category 2, 20% Male customers who belong to the Self-employed profession are churning. For net worth category 3, Approximately 21% of Male customers who belong to the Self-employed profession are churning. In all the cases of Customer net worth category, Self-employed Male customers are more likely to churn. It is visible at first look that for low current balance more number of customers are churning. Young Male Self-employed customers are churning more than young female self-employed customers.

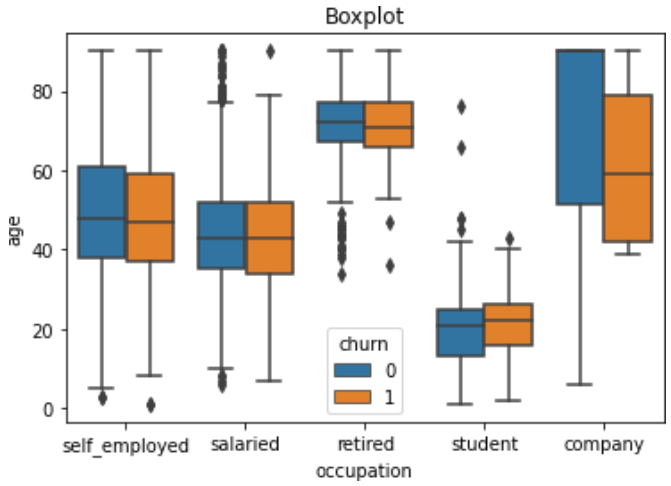


**Figure 20**

According to figure 20, For Self-employed profession churning customers are slightly younger than non-churning customers. In the retired occupation for non-churning customers, there are many outliers that indicate young people who retire early are not churning.

## 4.4 Logistic Regression

After implementing logistic regression on training dataset following AUC-ROC curve and confusion matrix are obtained.

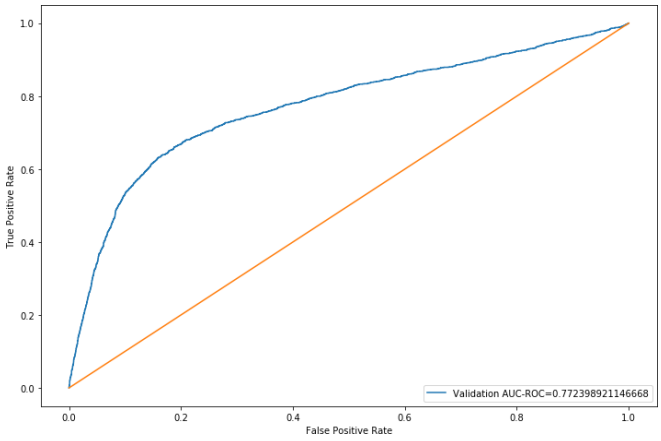


**Figure 21 - AUC-ROC Curve**

An excellent model has AUC near to the 1 which means it has a good measure of separability.

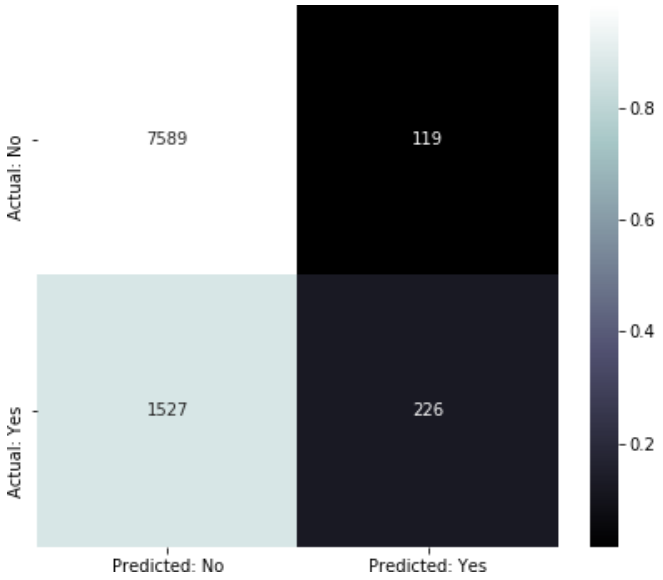


**Figure 22 -Confusion Matrix**

According to the confusion matrix in figure 22,

True Positive (TP=226), False Positive (FP=119), True Negative (TN=7589), False Negative (FN=1527).

## 4.5 Comparison of different models

After performing EDA, we made a baseline model, one all features model and another model using the top 10 features using RFE function from sklearn library and evaluated using 5-fold cross validation. Table 1-3, shows the AUC-ROC score, precision and recall score for 5-fold cross validation on different models.

| ALL FEATURES MODEL SCORES | | | |
|---|---|---|---|
| K Fold | RUC AUC Score | Recall Score | Precision score |
| 1 | 0.7939 | 0.2158 | 0.7492 |
| 2 | 0.8058 | 0.2405 | 0.7270 |
| 3 | 0.7768 | 0.1911 | 0.7153 |
| 4 | 0.8026 | 0.2101 | 0.7016 |
| 5 | 0.7894 | 0.2006 | 0.7326 |

**Table 1 -All features model scores**

| BASELINE COLUMNS MODEL SCORES | | | |
|---|---|---|---|
| K Fold | RUC AUC Score | Recall Score | Precision score |
| 1 | 0.7625 | 0.1359 | 0.6560 |
| 2 | 0.7822 | 0.1236 | 0.6533 |
| 3 | 0.7558 | 0.1283 | 0.6081 |
| 4 | 0.7688 | 0.1217 | 0.6632 |
| 5 | 0.7628 | 0.1331 | 0.6422 |

**Table 2 -Baseline model scores**

| RFE MODEL SCORES | | | |
|---|---|---|---|
| K Fold | RUC AUC Score | Recall Score | Precision score |
| 1 | 0.7923 | 0.2272 | 0.7469 |
| 2 | 0.8118 | 0.2253 | 0.7248 |
| 3 | 0.7891 | 0.2091 | 0.7120 |
| 4 | 0.8034 | 0.2129 | 0.7226 |
| 5 | 0.7963 | 0.2063 | 0.7138 |

**Table 3 -RFE model scores**

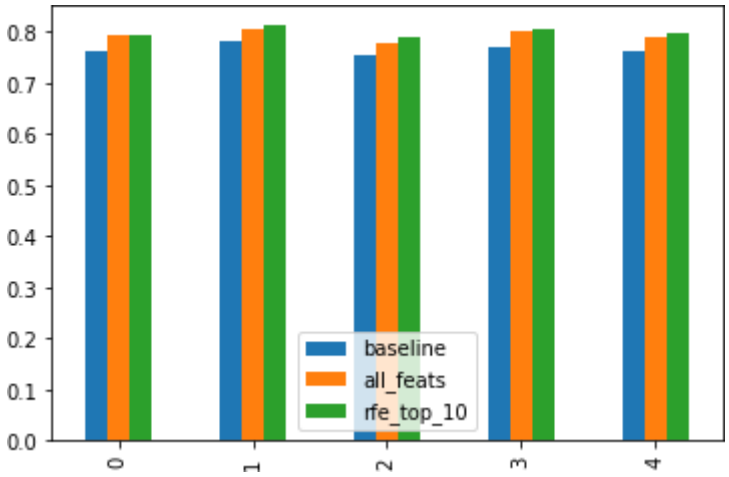


**Figure 23 -Comparison of different models**

Here, we can see that the model based on RFE is giving the best result for each fold and also our EDA analysis is very similar to the features selected by the RFE function.

# 5. Conclusion and Future Scope

Modern businesses require innovative prediction models because there is a critical need for a defensive marketing strategy that prevents customers from switching service providers. Customer churn reduces revenue and has a negative impact on corporate operations and profits. As a result, service providers require a churn prediction model because it will assist them in identifying the critical factors that cause churn.

We used dataset of a bank with various features and took insights from it after performing extensive Exploratory Data analysis (EDA) and then selecting the best features for building our models and implementing Logistic Regression. We then used k-fold cross validation strategy to evaluate the best models.

Our RFE model performed the best with high precision, recall and AUC score. This model can help the banks to identify potential churning customers beforehand and thus minimising losses. The model could perform better with better feature selection and different classification techniques which we will try to implement in the future.

# 6. References

[1] T.Vafeiadis, K.I. Diamantaras, G.Sarigiannidis,K.Chatzisavvas “Customer churn prediction in telecommunications”, Simulation Modelling: Practice and Theory 55 (2015) 1-9.

[2] Mutanen, Teemu & Nousiainen, Sami & Ahola, Jussi. (2010). Customer churn prediction - A case study in retail banking. 218. 10.3233/978-1-60750-633-1-77.

[3] Kristof Coussement, Koen W. De Bock, Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning, Journal of Business Research, Volume 66, Issue 9, 2013, Pages 1629-1636

[4] Julie Moeyersoms, David Martens, Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector, Decision Support Systems, Volume 72, 2015, Pages 72-81

[5] Kristof Coussement, Stefan Lessmann, Geert Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry, Decision Support Systems, Volume 95, 2017, Pages 27-36

[6] Michel Ballings, Dirk Van den Poel, Customer event history for churn prediction: How long is long enough? Expert Systems with Applications, Volume 39, Issue 18, 2012, Pages 13517-13522

[7] Zhen-Yu Chen, Zhi-Ping Fan, Minghe Sun, A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, European Journal of Operational Research, Volume 223, Issue 2, 2012, Pages 461-472

[8] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, Bart Baesens, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, European Journal of Operational Research, Volume 218, Issue 1, 2012, Pages 211-229

[9] S. Gürsoy, U. Tug˘ba, Customer churn analysis in telecommunication sector, J. School Bus. Admin. Istanbul Univ. 39 (1) (2010) 35–49.

# 7. APPENDIX

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import  LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import KFold, StratifiedKFold, train_test_split
from sklearn.metrics import roc_auc_score, accuracy_score, confusion_matrix, roc_cur
ve, precision_score, recall_score, precision_recall_curve
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.simplefilter(action='ignore', category=UserWarning)
df = pd.read_csv('churn_predictioan.csv')
pd.isnull(df).sum()
df['gender'].value_counts()
#Convert Gender
dict_gender = {'Male': 1, 'Female':0}
df.replace({'gender': dict_gender}, inplace = True)

df['gender'] = df['gender'].fillna(-1)
df['dependents'].value_counts()
df['occupation'].value_counts()
df['dependents'] = df['dependents'].fillna(0)
df['occupation'] = df['occupation'].fillna('self_employed')
df['city'] = df['city'].fillna(1020)
df['days_since_last_transaction'] = df['days_since_last_transaction'].fillna(999)
# Convert occupation to one hot encoded features
df = pd.concat([df,pd.get_dummies(df['occupation'],prefix = str('occupation'),prefix
_sep='_')],axis = 1)
num_cols = ['customer_nw_category',  'current_balance',
            'previous_month_end_balance', 'average_monthly_balance_prevQ2', 'average
_monthly_balance_prevQ',
            'current_month_credit','previous_month_credit', 'current_month_debit',
            'previous_month_debit','current_month_balance', 'previous_month_balance'
]
for i in num_cols:
    df[i] = np.log(df[i] + 17000)

std = StandardScaler()
scaled = std.fit_transform(df[num_cols])
scaled = pd.DataFrame(scaled,columns=num_cols)
df_df_og = df.copy()
df = df.drop(columns = num_cols,axis = 1)
df = df.merge(scaled,left_index=True,right_index=True,how = "left")
y_all = df.churn
df = df.drop(['churn','customer_id','occupation'],axis = 1)
baseline_cols = ['current_month_debit', 'previous_month_debit','current_balance','pr
evious_month_end_balance','vintage'
```

```
                ,'occupation_retired', 'occupation_salaried','occupation_self_emplo
yed', 'occupation_student']
df_baseline = df[baseline_cols]
# Splitting the data into Train and Validation set
xtrain, xtest, ytrain, ytest = train_test_split(df_baseline,y_all,test_size=1/3, ran
dom_state=11, stratify = y_all)
model = LogisticRegression()
model.fit(xtrain,ytrain)
pred = model.predict_proba(xtest)[:,1]
from sklearn.metrics import roc_curve
fpr, tpr, _ = roc_curve(ytest,pred)
auc = roc_auc_score(ytest, pred)
plt.figure(figsize=(12,8))
plt.plot(fpr,tpr,label="Validation AUC-ROC="+str(auc))
x = np.linspace(0, 1, 1000)
plt.plot(x, x, linestyle='-')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc=4)
plt.show()
# Confusion Matrix
pred_val = model.predict(xtest)
label_preds = pred_val

cm = confusion_matrix(ytest,label_preds)


def plot_confusion_matrix(cm, normalized=True, cmap='bone'):
    plt.figure(figsize=[7, 6])
    norm_cm = cm
    if normalized:
        norm_cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        sns.heatmap(norm_cm, annot=cm, fmt='g', xticklabels=['Predicted: No','Predic
ted: Yes'], yticklabels=['Actual: No','Actual: Yes'], cmap=cmap)

plot_confusion_matrix(cm, ['No', 'Yes'])
# Recall Score
recall_score(ytest,pred_val)
def cv_score(ml_model, rstate = 12, thres = 0.5, cols = df.columns):
    i = 1
    cv_scores = []
    df1 = df.copy()
    df1 = df[cols]

    # 5 Fold cross validation stratified on the basis of target
    kf = StratifiedKFold(n_splits=5,random_state=rstate,shuffle=True)
    for df_index,test_index in kf.split(df1,y_all):
        print('\n{} of kfold {}'.format(i,kf.n_splits))
        xtr,xvl = df1.loc[df_index],df1.loc[test_index]
        ytr,yvl = y_all.loc[df_index],y_all.loc[test_index]

        # Define model for fitting on the training set for each fold
        model = ml_model
```

```
        model.fit(xtr, ytr)
        pred_probs = model.predict_proba(xvl)
        pp = []

        # Use threshold to define the classes based on probability values
        for j in pred_probs[:,1]:
            if j>thres:
                pp.append(1)
            else:
                pp.append(0)

        # Calculate scores for each fold and print
        pred_val = pp
        roc_score = roc_auc_score(yvl,pred_probs[:,1])
        recall = recall_score(yvl,pred_val)
        precision = precision_score(yvl,pred_val)
        sufix = ""
        msg = ""
        msg += "ROC AUC Score: {}, Recall Score: {:.4f}, Precision Score: {:.4f} ".f
ormat(roc_score, recall,precision)
        print("{}".format(msg))

         # Save scores
        cv_scores.append(roc_score)
        i+=1
    return cv_scores
baseline_scores = cv_score(LogisticRegression(), cols = baseline_cols)
all_feat_scores = cv_score(LogisticRegression())
from sklearn.feature_selection import RFE
import matplotlib.pyplot as plt

# Create the RFE object and rank each feature
model = LogisticRegression()
rfe = RFE(estimator=model, n_features_to_select=1, step=1)
rfe.fit(df, y_all)
ranking_df = pd.DataFrame()
ranking_df['Feature_name'] = df.columns
ranking_df['Rank'] = rfe.ranking_
ranked = ranking_df.sort_values(by=['Rank'])
ranked
rfe_top_10_scores = cv_score(LogisticRegression(), cols = ranked['Feature_name'][:10
].values)
cv_score(LogisticRegression(), cols = ranked['Feature_name'][:10].values, thres=0.14
)
results_df = pd.DataFrame({'baseline':baseline_scores, 'all_feats': all_feat_scores,
 'rfe_top_10': rfe_top_10_scores})
results_df.plot(y=["baseline", "all_feats", "rfe_top_10"], kind="bar")
```