Project Report

CHURN PREDICTION

DEPARTMENT OF INFORMATION TECHNOLOGY ENGINEERING

PANJAB UNIVERSITY SWAMI SARVANAND GIRI REGIONAL CENTRE, HOSHIARPUR

SUBMITTED TO:

Mr Gurpinder Singh

SUBMITTED BY:

VIDYA SHARMA

SG18821

IT-7th Semester

ACKNOWLEDGEMENT

I would like to thank my teacher **Mr Gurpinder Singh**, who gave me this opportunity to pursue this training. I got to learn about Machine Learning.

The constant guidance and encouragement received from rest of the faculty at the Department of Information Technology, has been of great helping in carrying my present work and helped me in completing the project report with success.

At last, I would like to extend my heartfelt thanks to my parents because without their help this project would not have been successful. Finally, I would like to thank my dear friends who have been with me all the time. I am also thankful to all the members for their intellectual support throughout the course of their work.

SUBMITTED TO:

Mr Gurpinder Singh

SUBMITTED BY:

VIDYA SHARMA

SG18821

IT-7th Semester

TABLE OF CONTENTS

1.1 ABOUT CUSTOMER CHURN

The phenomenon where the customer leaves the organization is referred to as customer churn. Identifying which customers are likely to leave the bank, in advance can help them to take measures in order to reduce customer churn.

Customer churn has become a big issue in many banks 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.
Customer satisfaction, happiness, and loyalty can be achieved to a certain degree, but churn will always be a part of the business. Churn can happen because of:

- Bad customer service (poor service quality, response rate, or overall customer experience),
- Finance issues (fees and rates),
- Customer needs change,
- Dissatisfaction (your service failed to meet expectations),
- Customers don't see the value,
- Customers switch to competitors,
- Long-time customers don't feel appreciated.

0% churn rate is impossible. The trick is to keep the churn rate as low as possible at all times.

The impact of the churn rate is clear, so we need strategies to reduce it. Predicting churn is a good way to create proactive marketing campaigns targeted at the customers that are about to churn.

During churn prediction, we are also:

- Identifying at-risk customers,
- Identifying customer pain points,
- Identifying strategy/methods to lower churn and increase customer retention.

1.2 THE PROBLEM STATEMENT

A Bank wants to take care of customer retention for its product: savings accounts. The bank wants to identify customers likely to churn balances below the minimum balance. We have the customers information such as age, gender, demographics along with their transactions with the bank.

My task as a data scientist would be to predict the propensity to churn for each customer.

Data Dictionary

There are multiple variables in the dataset which can be cleanly divided into 3 categories:

I. Demographic information about customers

- **customer_id** - Customer id
- **vintage** - Vintage of the customer with the bank in a number of days
- **age** - Age of customer
- **gender** - Gender of customer
- **dependents** - Number of dependents
- **occupation** - Occupation of the customer
- **city** - City of the customer (anonymized)

II. Customer Bank Relationship

- **customer_nw_category** - Net worth of customer (3: Low 2: Medium 1: High)
- **branch_code** - Branch Code for a customer account
- **days_since_last_transaction** – No. of Days Since Last Credit in Last 1 year

III. Transactional Information

- **current_balance** - Balance as of today
- **previous_month_end_balance** - End of Month Balance of previous month
- **average_monthly_balance_prevQ** - Average monthly balances (AMB) in Previous Quarter
- **average_monthly_balance_prevQ2** - Average monthly balances (AMB) in previous to the previous quarter
- **current_month_credit** - Total Credit Amount current month
- **previous_month_credit** - Total Credit Amount previous month
- **current_month_debit** - Total Debit Amount current month
- **previous_month_debit** - Total Debit Amount previous month
- **current_month_balance** - Average Balance of current month
- **previous_month_balance** - Average Balance of previous month
- **churn** - Average balance of customer falls below minimum balance in the next quarter

2.1 SOFTWARE USED: JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document. We can use Jupyter Notebooks for all sorts of data science tasks including data cleaning and transformation, numerical simulation, exploratory data analysis, data visualization, statistical modeling, machine learning, deep learning, and much more.

A Jupyter Notebook provides you with an easy-to-use, interactive data science environment that doesn't only work as an integrated development environment (IDE), but also as a presentation or educational tool. Jupyter is a way of working with Python inside a virtual "notebook" and is growing in popularity with data scientists in large part due to its flexibility. It gives you a way to combine code, images, plots, comments, etc., in alignment with the step of the "data science process." Further, it is a form of interactive computing, an environment in which users execute code, see what happens, modify, and repeat in a kind of iterative conversation between the data scientist and data. Data scientists can also use notebooks to create tutorials or interactive manuals for their software.

A Jupyter notebook has two components. First, data scientists enter programming code or text in rectangular "cells" in a front-end web page. The browser then passes the code to a back-end "kernel" which runs the code and returns the results. Many Jupyter kernels have been created, supporting dozens of programming languages. The kernels need not reside on the data scientist's computer. Notebooks can also run in the cloud such as Google's Collaboratory project. We can even run Jupyter without network access right on our own computer and perform our work locally.

2.2 DATA SOURCE: KAGGLE

For Customer Churn Prediction data source is taken from Kaggle. **Kaggle**, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

3.1 DATA DESCRIPTION

Data set has various columns such as:
customer_id, vintage, age, gender, dependents, occupation, city, customer_nw_category , branch_code, days_since_last_transaction, current balance, previous_month_end_balance, average_monthly_balance_prevQ, average_monthly_balance_prevQ2, current_month_credit, previous_month_credit, current_month_debit, previous_month_debit, current_month_balance

3.2 TECHNOLOGIES USED:

Python for data scrapping, pre-processing, visualization etc
Libraries used:

- Pandas: It is primarily used for data frame manipulation.
- NumPy: Raise manipulation
- Matplotlib, seaborn: Basic data visualization
- Plotly: It is primarily used for interactive visualization.
- Sklearn: It is used to import couple of metrics.

3.3 EXPLORATORY DATA ANALYSIS

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. In this part of the project, I have simply exploredthe dataset by looking at the following things in the data:

- Importing the necessary libraries and the dataset
- Performing Data Preprocessing (Exploratory Data Analysis and Data Manipulation)
- Performing Prediction and Visualization
- Modelling using Logistic Regression

4.1 STEPS TO PERFORM

STEPS PERFORMED FOR CUSTOMER CHURN PREDICTION ARE :-

1) Reading Files into Python
2) Variable Identification and Typecasting
3) Converting variable to required category
4) datetime Data Type
    - Breaking down the date variable into these granular information will help us in understand when the last transaction was done from different perspectives. Now that we have extracted the essentials from the last_transaction variables, we will drop it from the dataset.
5) Univariate Analysis: Numerical Variables

➢ **Summary of Customer_Information:**

- **customer_id:**
    - variable is unique for every customer, Hence uniform distribution.
    - This variable does not contribute any information
    - Can be eliminated from data
- **age:**
    - Median Age = 46
    - Most customers age between 30 to 66
    - skewness +0.33 : customer age is negligibly biased towards younger age
    - kurtosis = -0.17; very less likely to have extreme/outlier values.
- **vintage:**
    - Most customers joined between 2100 and 2650 days from the day of data extraction.
    - skewness -1.42 : this is left skewed, vintage variable is significantly biased towards longer association of customers.
    - Kurtosis = 2.93: Extreme values and Outliers are very likely to be present in vintage.

Things to Investigate Further down the road:

- The batch of high number of very Old Age customers in age variable.
- Considering the kurtosis and skewness value for all 4 of these plots. Outliers/Extreme values are obvious.
- Need to Remove Outliers to visualise these plots.

➢ **Summary of current_month**

- After Removing extreme/outliers, plots are still very skewed.

Things to investigate further down :

1. Is there any common trait/relation between the customers who are performing high transaction credit/debits?
2. Customers who are performing high amount of transactions, are they doing it every month?

- **Summary of previous month**
  - Looks very similar to current_month. Most of the customers perform low amount transactions.

- **Summary of previous_quarters**
  - The general trend still follows, it is crucial that we find the out if there is any common trait between the customers doing high amount of transactions.

- **Summary of transaction date :**
  - ❖ Day_of_Year:
  - Most of the last transactions were made in the last 60 days of the extraction of data.
  - There are transactions which were made also an year ago.
  - ❖ Week_of_year and Month_of_year:
  - These variable validate the findings from the day_of_year.
  - ❖ Day_of_Week:
  - Tuesdays are often the favoured day relative to others.

Things to investigate further Down :

• Customers whose last transaction was 6 months ago, did all of them churn?

6) Univariate Analysis : Categorical Variables

Grouping Variables

- **customer_info:** gender, occupation, customer_nw_category
- **account_info:** city, branch_code
- **churn**

- **customer info**

Summary :

- Occupation
  - Majority of people are self_employed.
  - There are extremely few Company Accounts. Might explain Outlier/Extreme values in credit/debit.
- Gender:

    - Males accounts are 1.5 times more 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.

Things to investigate further down:

- Possibility: Company accounts are the reason behind the outlier transactions.
- Possibility: customers belonging to the highest net worth category may explain the skewness of the transactions.

➢ **account info**

Summary
For both variable "city" and "branch_code", there are too many categories. There is clear relation that some branches and cities are more popular with customers and and this trend decreases rapidly.

Things to investigate further down:

- Popular cities and branch code might be able to explain the skewness and outliers of credit/debit variables.
- Possibility that cities and branch code with very few accounts may lead to churning.

➢ **churn**

- Number of people who churned are 1/4 times of the people who did not churn in the given data.

7) Univariate: Missing Values

Things to investigate further down:

- Gender: Do the customers with missing gender values have some common behaviour
    - churn: do missing values have any relation with churn?
- Dependents:
    - Missing values might be similar to zero dependents
    - churn: do missing values have any relation with churn?
- Occupation:
    - Do missing values have similar behaviour to any other occupation
    - do they have some relation with churn?
- city:
    - the respective cities can be found using branch_code
- last_transaction:
    - checking their previous month and current month and previous_quarter activity might give insight on their last transaction.

- For almost all the above:
    - vintage: might be recording errors from same period of joining
    - branch_code: might be recording error from certain branch

8) Univariate Analysis: Outliers

- We suspected outliers in current_month and previous_month variable groups. We will verify that using box plots.

➢ **current_month and previous_month**

Summary:

- If we look at corresponding plots in the outputs above, there seems to be a strong relation between the corresponding plots of previous_month and current_month variables.
- Outliers are significant in number and very similar in number between corresponding plots. Which indicates some inherent undiscovered behaviour of Outliers.

➢ **previous quarters**

Summary:

- Outliers in previous two quarters are very similar but significantly large in number.

❖ **Investigation directions from Univariate Analysis :**

1. customer_id variable can be dropped.
2. Is there there any common trait/relation between the customers who are performing high transaction credit/debits?
    - customer_nw_category might explain that.
    - Occupation = Company might explain them
    - popular cities might explain this
3. Customers whose last transaction was 6 months ago, did all of them churn?
4. Possibility that cities and branch code with very few accounts may lead to churning.

9) Bivariate Analysis : Numerical-Numerical
10) Correlation Matrix
11) Heatmap

Summary :

- Kendall and Spearman correlation seem to have very similar pattern between them, except the slight variation in magnitude of correlation.
- Too many variables with insignificant correlation.
- Major correlation lies between the transaction variables and balance variables.

Inferences:

- Transaction variables like credit/debit have a strong correlation among themselves.
- Balance variables have strong correlation among themselves.
- Transaction variables like credit/debit have insignificant or no correlation with the Balance variables.

12) Scatterplot

- The scatter plot is not meaningful due to the presence of outliers

13) Removing Outliers

Inferences:

1. This validates the high correlation between the two previous quarters
2. This high correlation can be used for feature engineering during the later stages.

14) Multivariate Analysis

Pivot Table

We are using Pivot table to comply with the objective of identifying the Churning Customers Profile using multiple categorical features. First, Let's use Gender, Occupation and Customer Net worth category and derive insights from the Pivot Table

- **Gender, Occupation, Customer Net worth category with Churn**

- **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% **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
- This would be interesting to dig deeper and find out if the **"Self-employed Male"** Customers are churning more

- **Gender, Age, Occupation with Churning Status**

- We have created three bins for the age variable dividing age into 3 groups 0-25, 25-50 and 50-100

- Highest number of Customers are churning from **Male category** who belong to the age group of **(25,50)** and are professionally **self employed**
- Highest Proportion of Customers are churning from **Male category** who belong to the age group of **(0,25)** and are professionally **self employed**
- Here also **Self Employed Male customers** are churning more than any other combination of categories

➢ **Gender,Age,Occupation and Current Balance with Churning Status**

- Current balance is divided into 3 quantiles
- It is visible at first look that for **low current balance** more number of customers are churning
- For the first quantile of current balance, More than **18%** (overall average churning) of customers are churning and for second and third quantile percentage of churning customers is less than 18%
- In first quantile of current balance, for **self employed profession** as the age increases for customers, their churning proportion decreases. This means that **Young Self employed Customers** are more prone to churn.
- There is a visible gap in proportion of Self employed females who lie in the age group of (0,25) and Self employed Males who lie in the same group. **Young Male Self employed customers** are churning more than young female self employed customers

15) **Box Plot**

Now in order to comply with our objective of identifying churning customers profile we will use grouped Box plot.

➢ **Age, Occupation, Churn**

- 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**

➢ **Vintage, Gender, Churn**

- There is no visible difference in the vintage feature for genderwise churning and non churning customers

16) **Pair Plot**

➢ **Churn vs Current & Previous month balances**

Here I have included the following:

- Log of current balance & previous month end balance
- Log of average monthly balance of current and previous month
- Churn is represented by the colour here (Orange - Churn, Blue - Not Churn)

The distribution for these features look similar. We can make the following conclusions from this:

- There is high correlation between the previous and current month balances which is expected
- The distribution for churn and not churn is slightly different for both the cases

➢ **Credit and Debits for current and previous months**

Both credit and debit patterns show significant difference in distributions for churned and non churned customers.

- Bimodal distribution/Double Bell Curve shows that there are 2 different types of customers with 2 brackets of credit and debit. Now, during the modeling phase, these could be considered as a separate set of customers
- For debit values, we see that there is a significant difference in the distribution for churn and non churn and it might turn out to be an important feature.

17) Encoding

- Encoding the variables using get dummies pandas function so every variable has numerical value attached to it.

18) Missing Values with Mode

- Filling missing values with the mode of the data

19) Segregating variables: Independent and Dependent Variables

- Insignification variable drop
- Creating features and target variable

20) Splitting the data into train set and the test set

- Splitting the entire data into train and test set

21) Normalising using min_max_scaler

- Scaling the data so model doesn't has bias for high valued features

22) Model Building

- Logistic regression works pretty well and able to provide accuracy more than 80% for both training and test set (approx. 82%)

23) Using Regularization

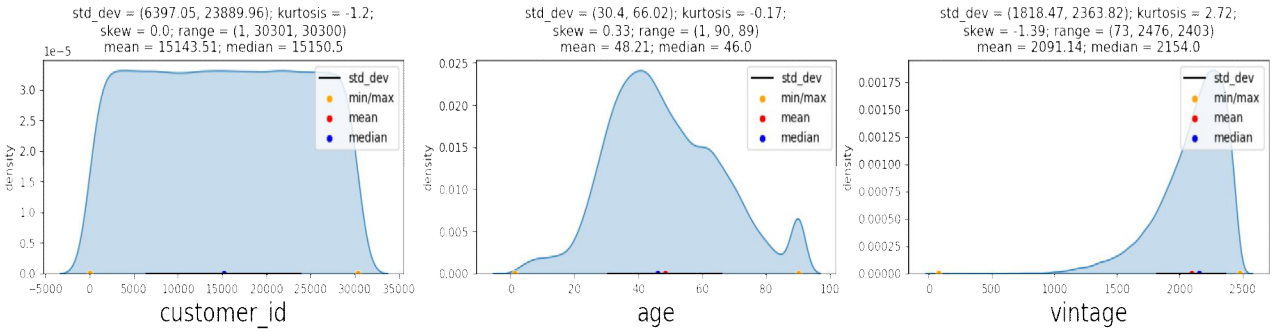
- As the value of C increases, Regularization constant decreases (C is inverse of Regularization Constant)
- So most of the features associate themselves with 0 coefficient value
- This can be a feature selection technique as well while building model

5.1 DATA VISUALIZATION

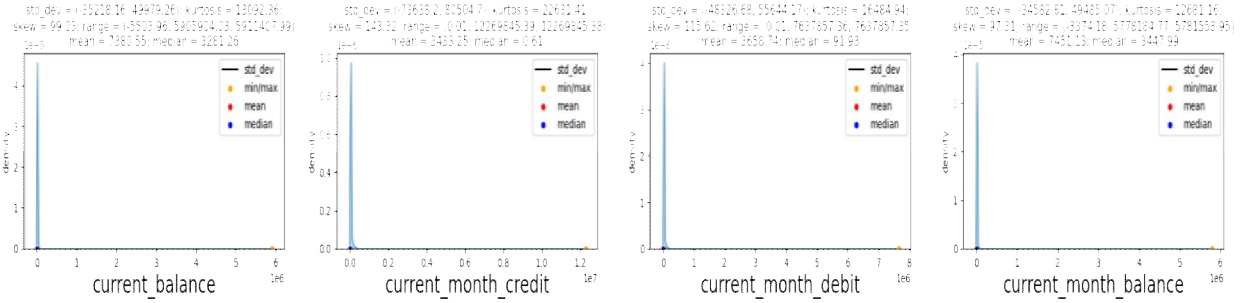
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

A complete visualization of data is shown in my Jupyter notebook.

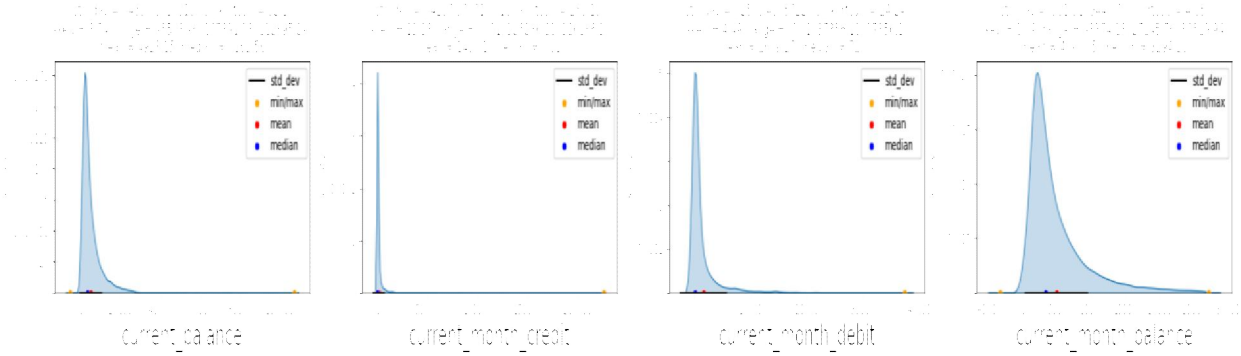
- **For Univariate Analysis: Numerical Variables :-**
  - For **customer_information**



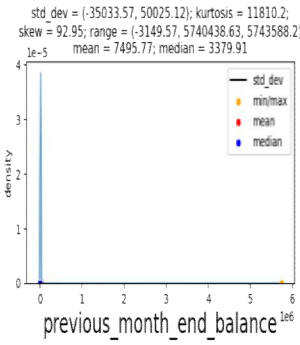
  - For **current_month**

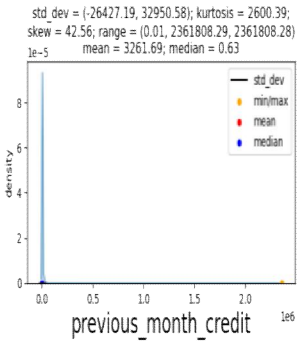


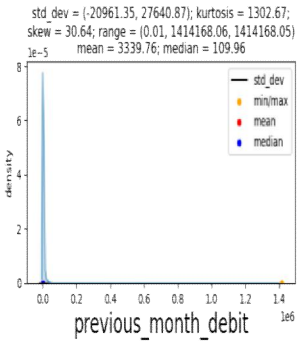
  - Removing outliers from **current_month**

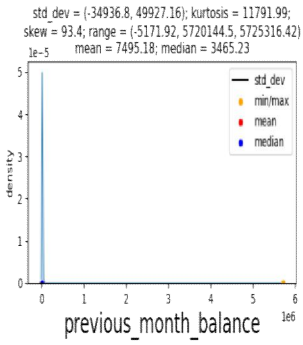


- For **previous_month**

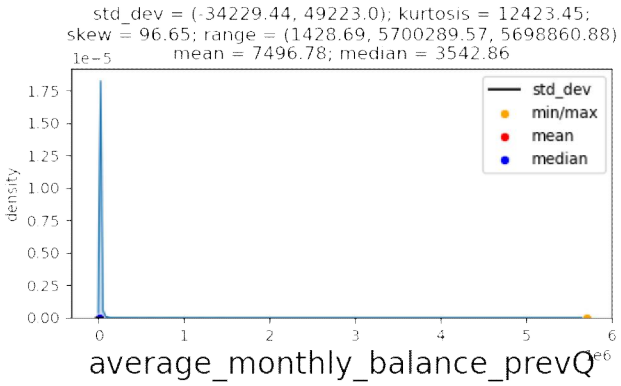


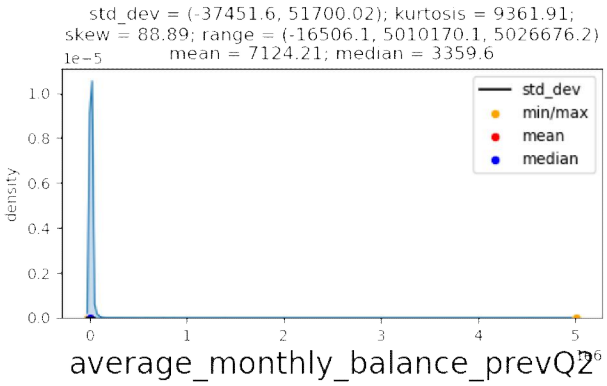




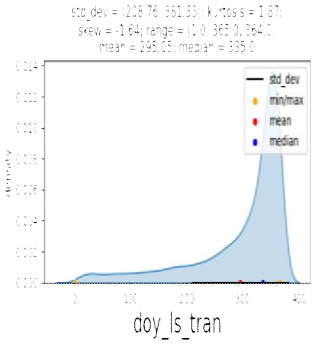


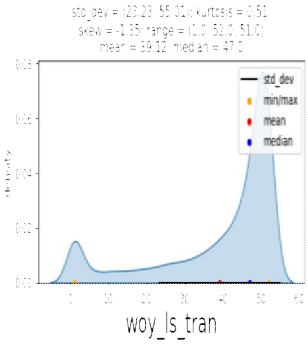
- For **previous_quarters**

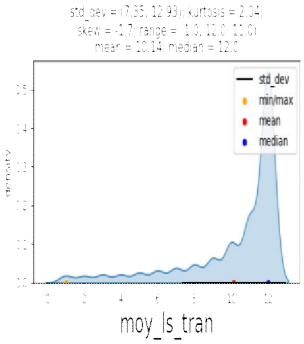


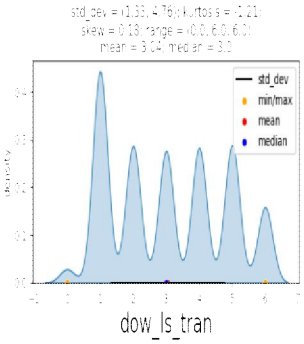


- For **transaction_date**



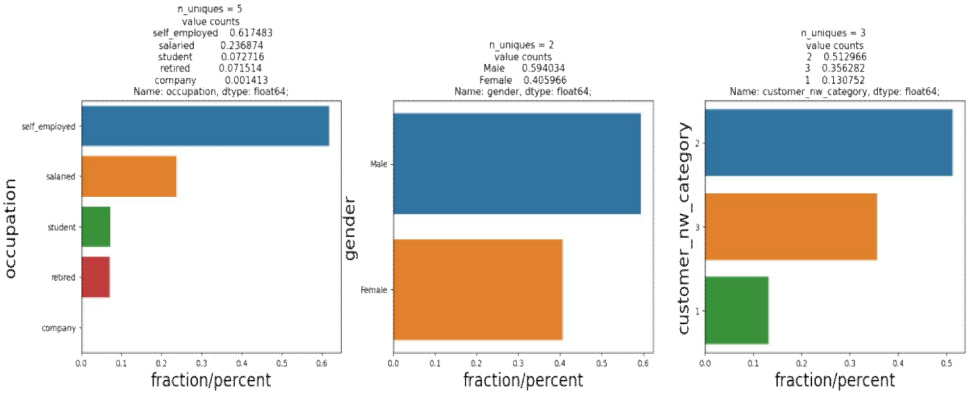




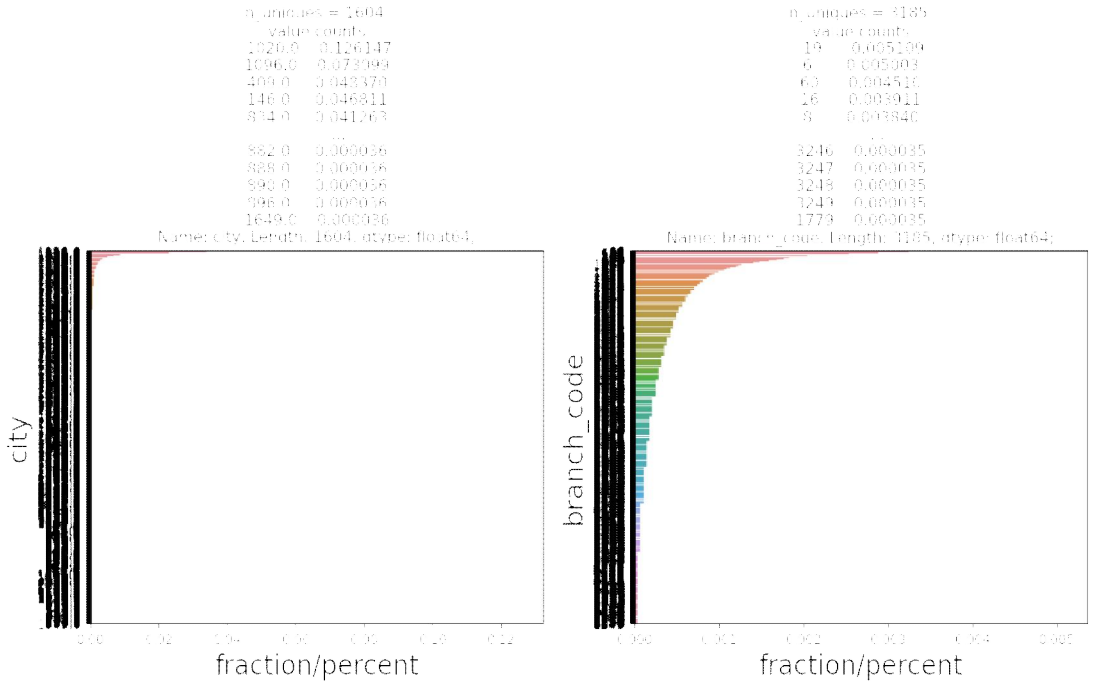


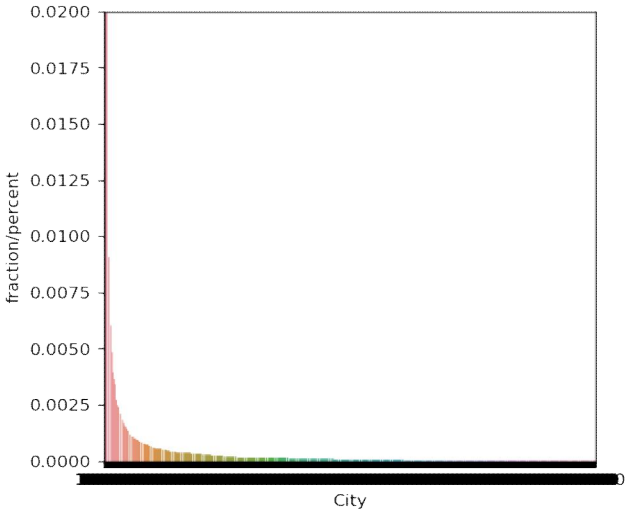
For Univariate Analysis: Categorical Variables :-

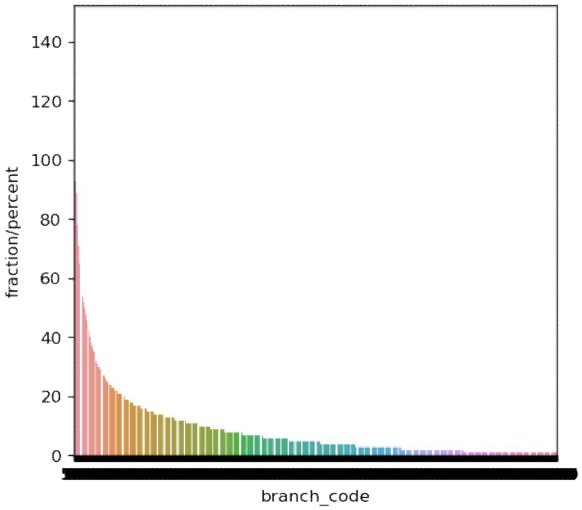
- For **customer_info**



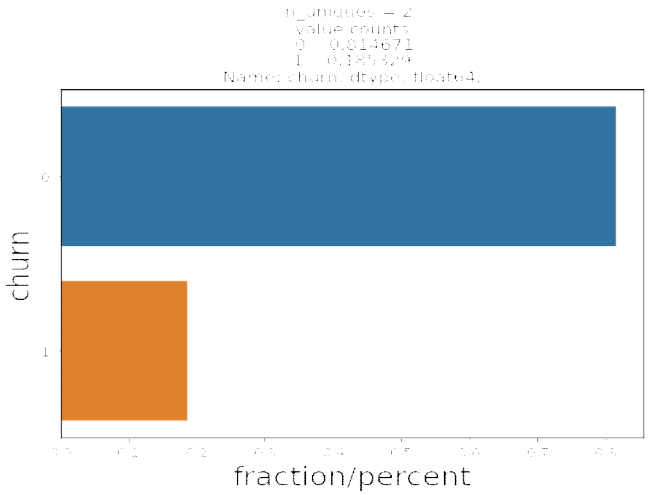
- For **account_info**




0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
fraction/percent
City

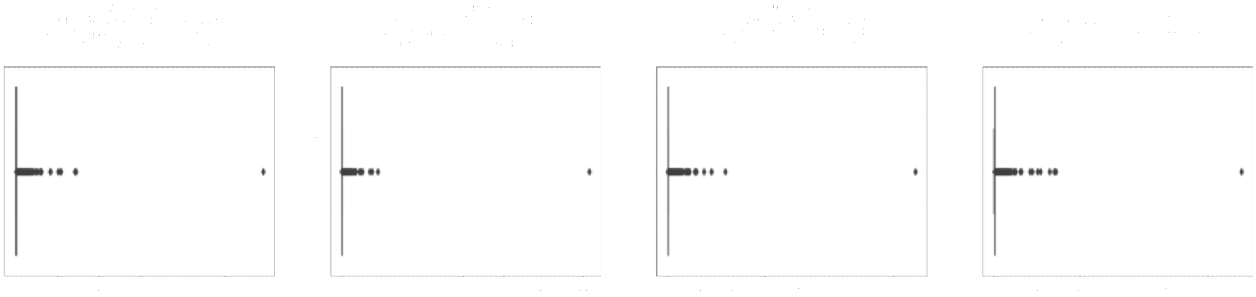

140
120
100
80
60
40
20
0
fraction/percent
branch_code

- For **churn**

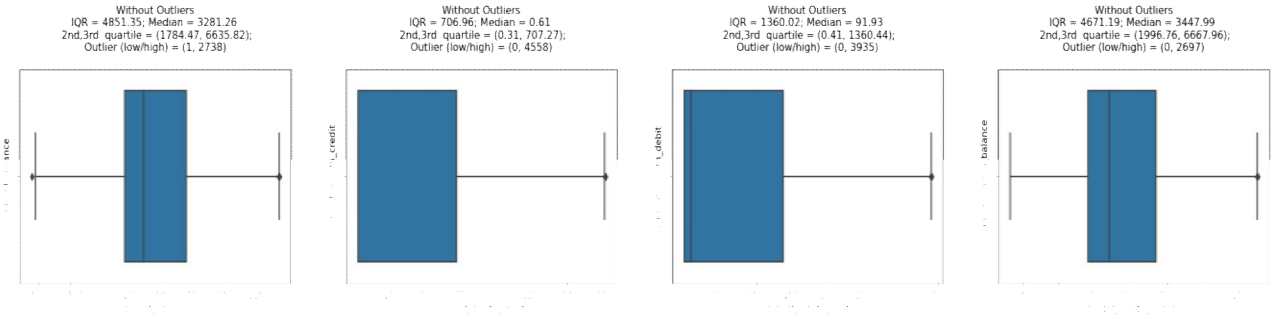

churn
fraction/percent

For Univariate Analysis: Outliers:-

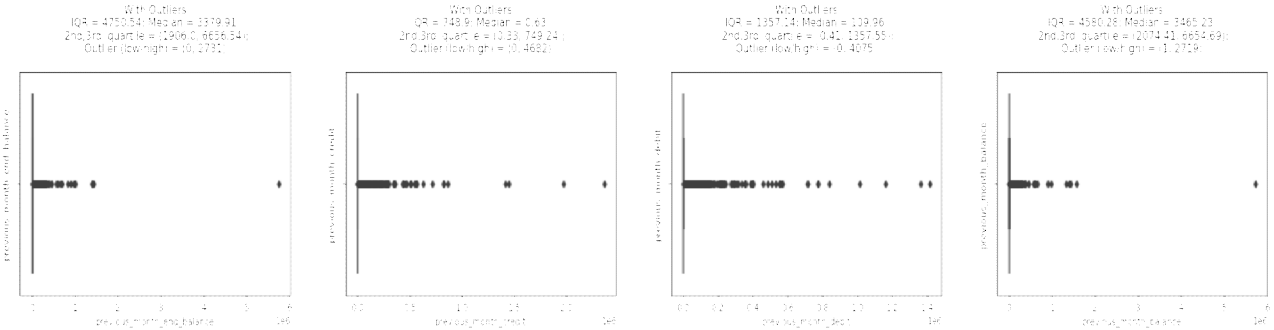
- For **current_month** with outliers



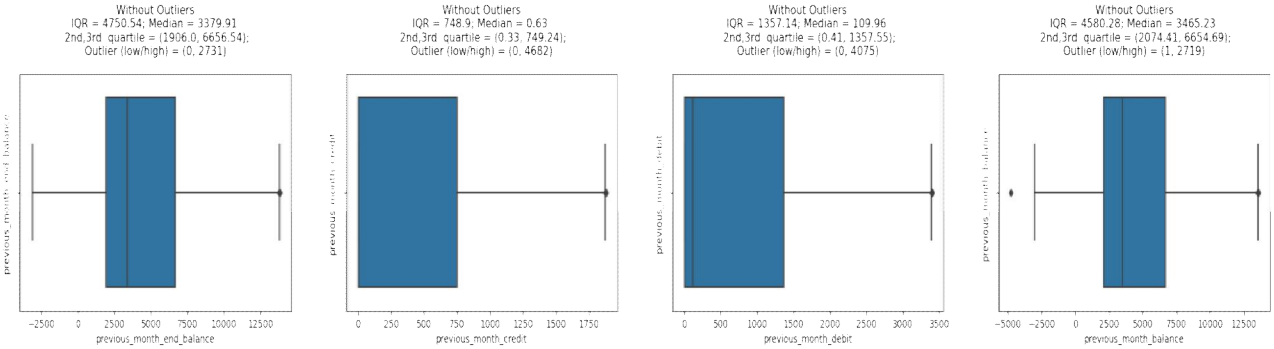
- For **current_month** without outliers



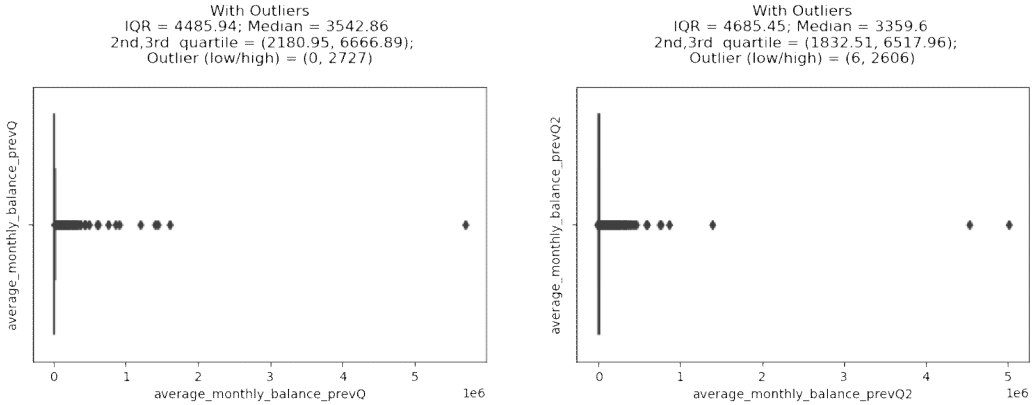
- For **previous_month** with outliers



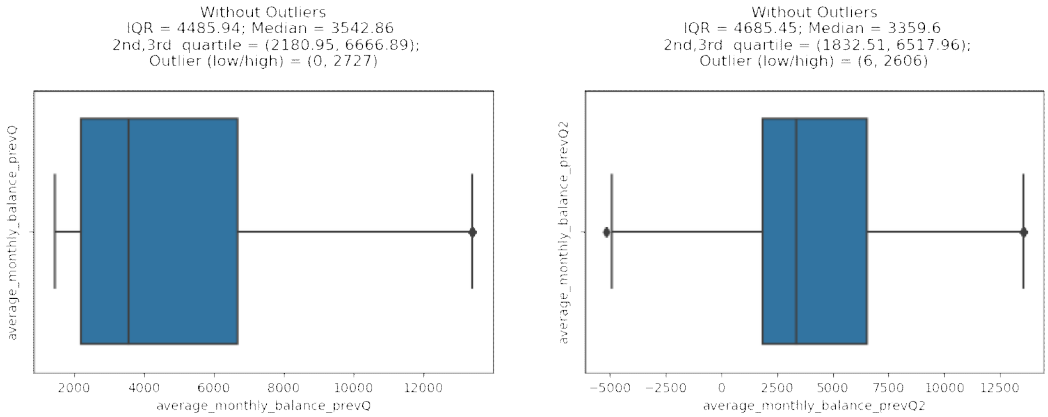
- For **previous_month** without outliers



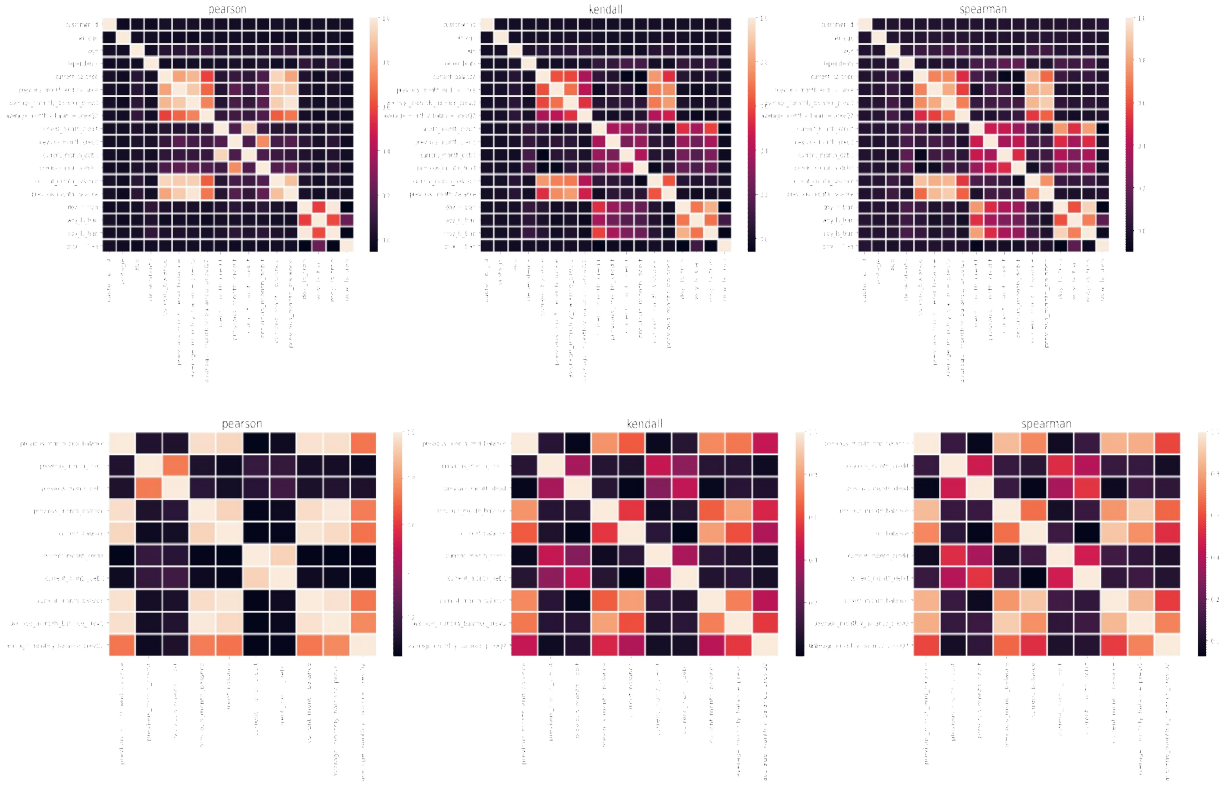
- For **previous_quarters** with outliers



- For **previous_quarters** without outliers

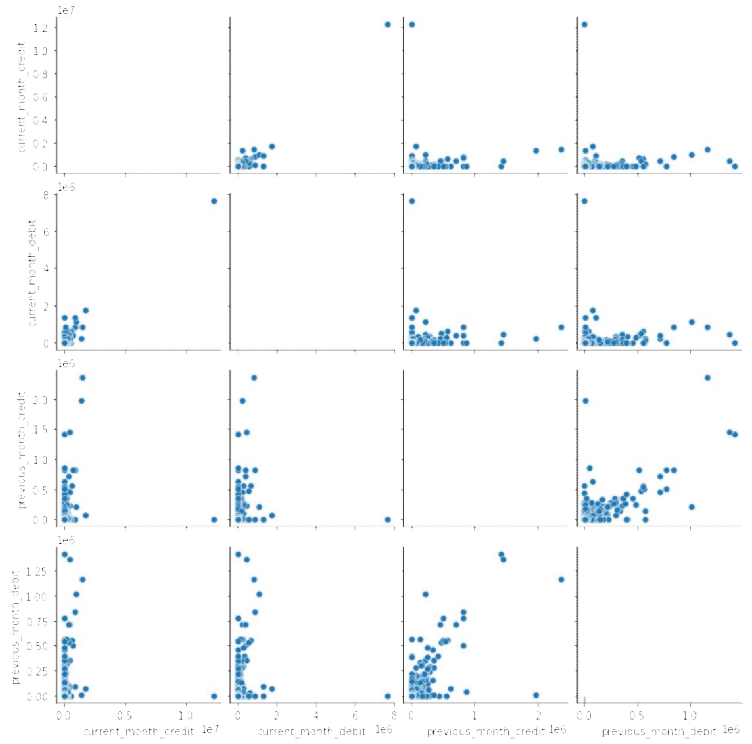


Heatmap :-



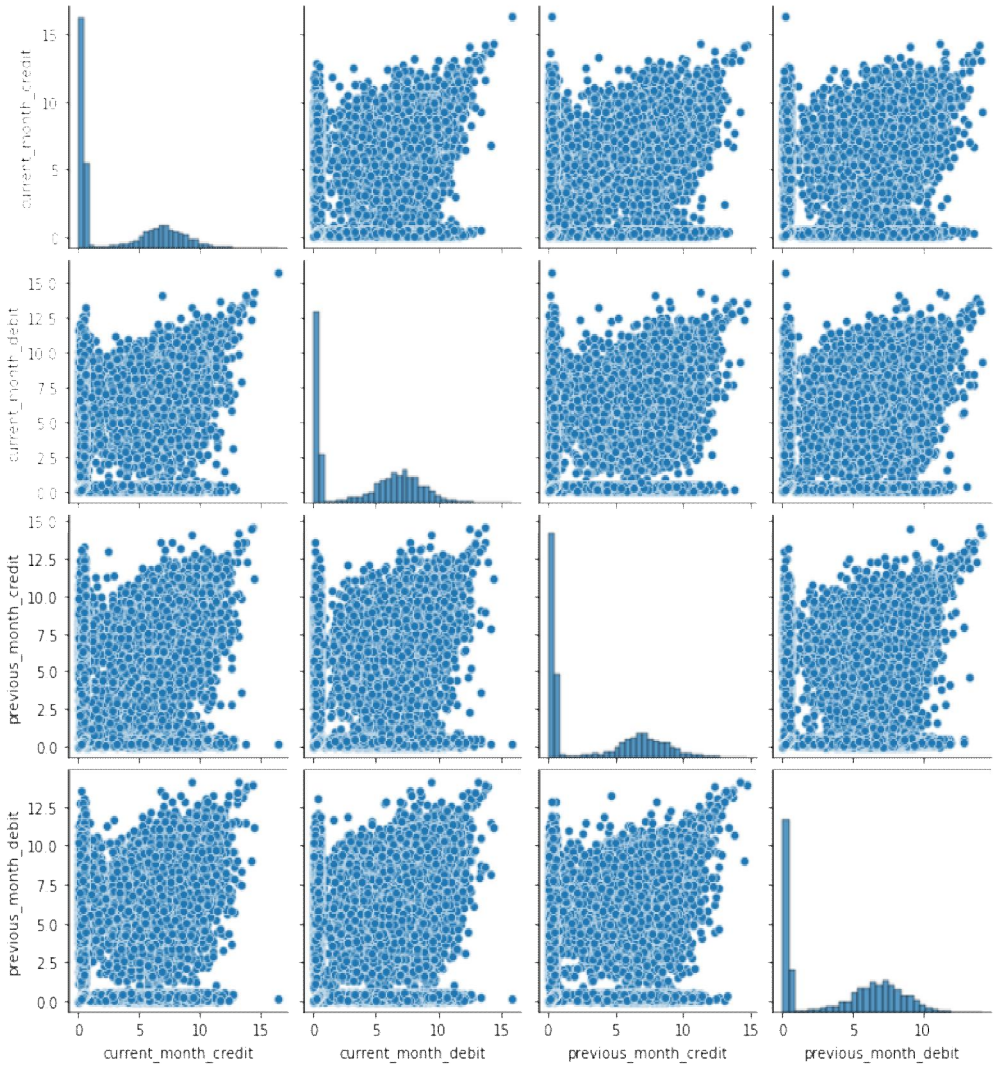
Scatterplot :-

- Scatter plot for transactional variables

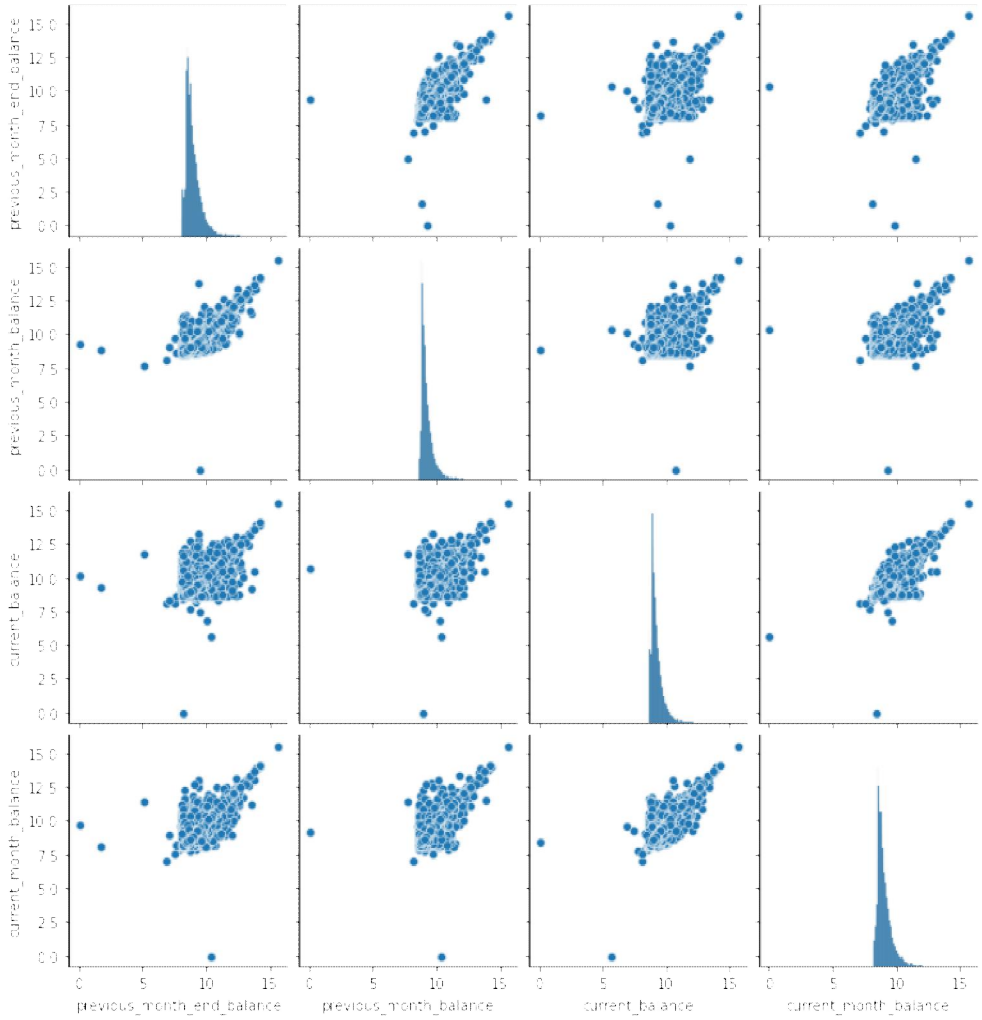


- **The scatter plot is not meaningful due to the presence of outliers.**

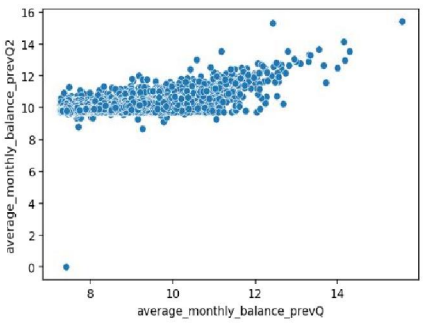
➢ Scatter plot for transactional variable



- Scatter plot for balance variable

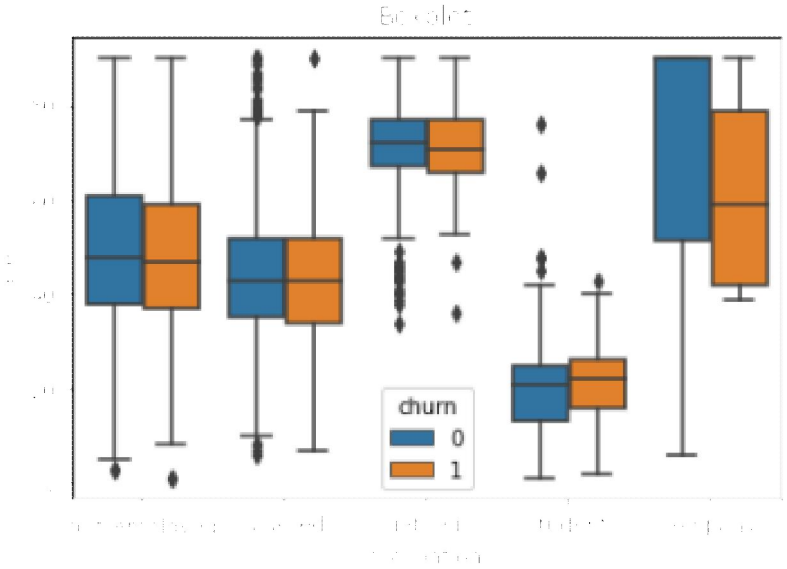


- Scatter plot for previous quarters

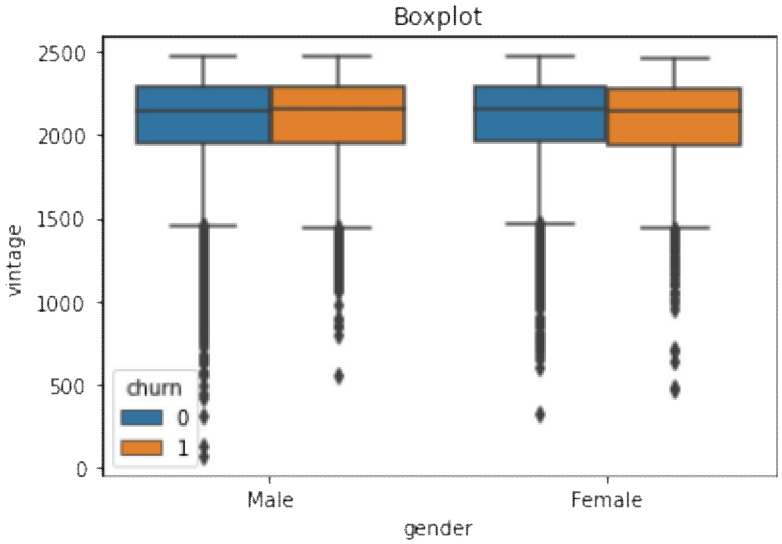


Box Plot :-

- Age, Occupation, Churn

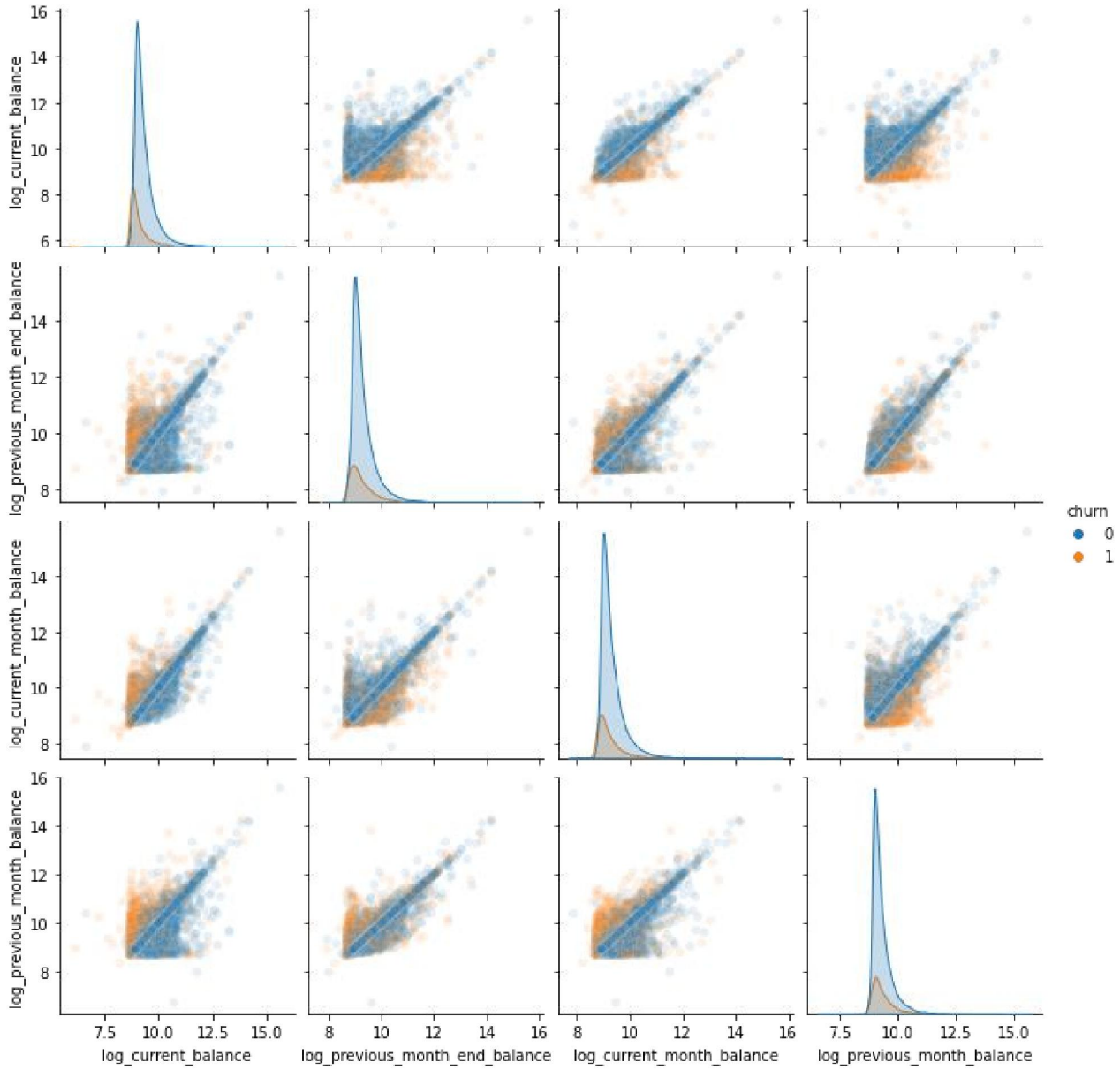


- Vintage, Gender, Churn

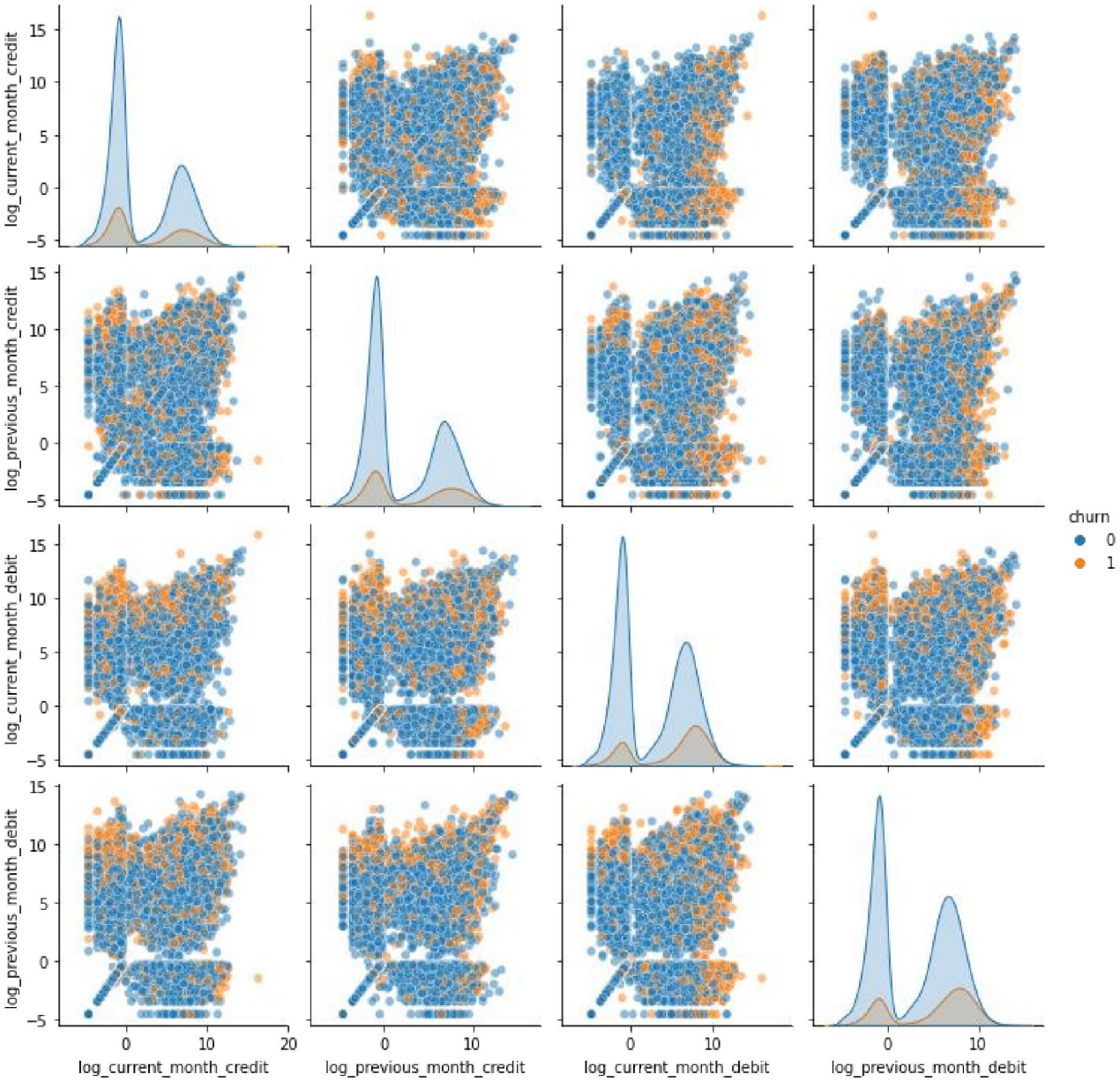


Pair Plot :-

- Churn vs Current & Previous month balances



- Credit and Debits for current and previous months



6.1 CONCLUSION

Churn prediction is one of the most effective strategies used in banking sector to retain existing customers. It leads directly to improved cost allocation in customer relationship management activities, retaining revenue and profits in future. It also has several positive indirect impacts such as increasing customer's loyalty, lowering customer's sensitivity to competitors marketing activities, and helps to build positive image through satisfied customers.

The results predicted by the Logistic Regression algorithm were the most efficient with an accuracy of 82%. Therefore, companies that want to prevent customer churn should utilize this algorithm and remove features like long term contracts and instead replace them with monthly or short term contracts, thereby giving them more flexibility. Providing additional services such as device protection and multiple phone lines proves to be of little value to customer attrition. Lastly, focusing on enhancing the experience of loyal customers who have stayed with the company for long will prove worthwhile, ensuring their retention. The ability to identify customers that aren't happy with provided solutions allows businesses to learn about product or pricing plan weak points, operation issues, as well as customer preferences and expectations to proactively reduce reasons for churn.

6.2 FUTURE ENHANCEMENTS

Churn rate has many different factors and it's up to us to keep track of them. Depending on the industry, product, company, and every other element we can think of, there are many different ways to decrease churn.

Here are a few of the suggestions :

- Address churn as it happens

Churn is inevitable. No matter who you are and what sort of business you run, churn will happen. As churn is calculated, it is vital that you use what you learn to prevent customers from churning for the same reasons in the future.

- Invest in your best customers

Every business has big-ticket clients. These clients are considered more valuable than the rest. It is worth it to spend extra time with them, and make sure they have everything they need.

- Ask for feedback

Feedback is very important in any situation, but especially when talking about churn. The best way to solve any potential issues is to address them as soon as possible. Asking for feedback from customers can really shine some light on why other customers are leaving.

For example, you may be able to determine how long it takes the average customer to churn from the time they first log in. Sending out a feedback survey before that time arrives would be most ideal.

- Communicate often

Transparency is vital for the longevity of a company. Staying open and communicating with you customers is one of the best ways to let them know you care.

In order to communicate on a broad spectrum, you will need to create interesting and engaging content, interact with customers via social media, and keeping every updated via email.

Communicating openly and honestly lets customers know that they can trust you. That means they will most likely continue to buy your product.

- Make it easy for new customers

Making sure your new customers have everything they could possibly need to transition into your brand is important. In a way, this point is a combination of everything else in this list.

Creating a well-rounded onboarding process is the easiest way to make new customers feel welcome. You'd be surprised how far a simple welcome email will go. Taking it a step further with tutorials and educational materials will really help new customers navigate and feel comfortable with your product or service.

- Be competitive

No matter what business you're in, there will be competition. You will have to stay on top of pricing, features, incentives, and quality in order to make sure your customers are getting the best bang for their buck and you maintain your competitive advantage.

6.3 ADVANTAGES AND DISADVANTAGES OF CHURN RATE

❖ **ADVANTAGES**

- Provides clarity on the business quality
- Identify whether customers are satisfied or dissatisfied with the service or product
- Compare with competitors to gauge an acceptable level of churn rate
- Easy calculate pattern

❖ **DISADVANTAGES**

- Doesn't provide clarity on the types of customers leaving which means you couldn't find out which one left, the new or old customer
- Doesn't differentiate companies between industry types

7.1 REFERENCES

- **https://medium.com/@ODSC/why-you-should-be-using-jupyter-notebooks-ea2e568c59f2**
- **https://airfocus.com/glossary/what-is-churn/**
- **https://www.leadmine.net/glossary/churn-rate/**
- **https://en.wikipedia.org/wiki/Kaggle**