MBA

Semester - IV

Research Project - Final Report

Name	Abhilash B J
Project	Unveiling Customer Churn
Date of Submission	16th June 2024

A study on "Unveiling Customer Churn"

Research Project submitted to Jain Online (Deemed-to-be University)

In partial fulfillment of the requirements for the award of:

Master of Business Administration

Submitted by:

Abhilash B J

USN:

221VMBR00047

Under the guidance of:

Mr. Hrushikesha Shastry B S

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2023 - 24

DECLARATION

I Abhilash B J, hereby declare that the Research Project Report titled "Customer Churn" has been prepared by me under the guidance of the Mr. Hrushikesha Shastry B S. I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Eight Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/ diploma from any other University / Institution.

Place: Bengaluru

Date: 12-06-24

_____

Abhilash B J
221VMBR00047

TABLE OF CONTENTS

List of Tables

List of Graphs		
Graph No.	**Graph Title**	**Page No.**

List of Graphs		
Graph No.	**Graph Title**	**Page No.**

CHAPTER 1

INTRODUCTION AND BACKGROUND

INTRODUCTION AND BACKGROUND

1.1 EXECUTIVE SUMMARY

In today's fiercely competitive market, retaining existing customers has become a paramount challenge for E-Commerce companies and DTH providers alike. Recognizing the critical importance of customer retention, our project aims to develop a robust churn prediction model to anticipate and mitigate customer attrition. The primary objective is to enable the company to identify potential churners proactively and design segmented offers tailored to their needs, thereby enhancing customer retention rates and ensuring sustained business growth.

The project centers around leveraging advanced analytics and machine learning techniques to analyze a comprehensive dataset comprising various customer attributes and behaviours. At the core of our analysis lies the 'Churn' variable, which serves as the target variable for our predictive modeling. Each entry in the dataset is uniquely identified by the 'AccountID', which allows us to track churn at the account level, crucial for understanding the broader impact of customer attrition on the company's revenue and customer base.

Key features included in the dataset encompass a wide array of customer-related attributes, ranging from demographic information such as gender and marital status to behavioural indicators such as service satisfaction scores, complaint history, and usage patterns. These features provide valuable insights into customer behaviour, preferences, and engagement with the company's products or services, forming the basis for predictive modeling and campaign recommendation strategies.

Exploratory Data Analysis (EDA) plays a pivotal role in unraveling the underlying patterns, trends, and relationships within the dataset, thereby providing invaluable insights to guide subsequent modeling and decision-making processes. By meticulously examining the distribution, summary statistics, and visual representations of the dataset's variables, EDA allows us to grasp the fundamental characteristics of the data, identify potential outliers or anomalies, and discern any discernible patterns or trends. Through techniques such as scatter plots, histograms, and correlation analysis, EDA enables us to uncover hidden relationships between variables, ascertain the impact of individual features on the target variable (in this case, churn), and gain a holistic understanding of the data landscape. Moreover, EDA facilitates the identification of data preprocessing requirements, such as handling missing values, encoding categorical variables, and standardizing numerical features, thereby ensuring the robustness and reliability of subsequent predictive modeling efforts. Ultimately, by illuminating the nuances and intricacies of the dataset, EDA empowers us to make informed decisions, formulate effective strategies, and derive actionable insights to address the challenge of customer churn prediction effectively.

Our predictive modeling approach involves training machine learning algorithms on historical data to predict future churn probabilities for each account. Utilizing techniques such as logistic regression, decision trees, or ensemble methods, we aim to build a predictive model capable of accurately identifying accounts at risk of churn. By analyzing factors such as tenure, service satisfaction scores, payment preferences, and past interactions with customer care, the model can generate actionable insights to guide targeted retention efforts.

Furthermore, to ensure the viability and effectiveness of our campaign recommendations, we emphasize the importance of aligning proposed offers with the company's revenue objectives and constraints. While our goal is to reduce churn and enhance customer loyalty, we must also consider the financial implications of our recommendations. Therefore, our campaign suggestions prioritize targeted, value-driven offers that incentivize customer retention without compromising the company's profitability.

In conclusion, the Customer Churn Prediction project represents a proactive and data-driven approach to address the challenge of customer attrition in a competitive market landscape. By leveraging predictive analytics and tailored campaign strategies, we aim to empower the company to retain its valuable customer base, drive sustainable revenue growth, and maintain a competitive edge in the industry.

1.2 INTRODUCTION AND BACKGROUND

In the dynamic landscape of modern commerce, customer retention stands as a cornerstone of sustainable business growth. For E-Commerce companies and DTH providers grappling with intensifying competition, the ability to anticipate and mitigate customer churn has emerged as a strategic imperative. Customer churn, defined as the loss of customers over a specified period, not only impacts revenue streams but also erodes market share and undermines brand loyalty. Recognizing the critical importance of retaining existing customers, our project endeavours to develop a robust churn prediction model and formulate targeted retention strategies tailored to the unique needs and preferences of customers. By harnessing the power of data analytics and machine learning, we aim to equip the company with the tools and insights necessary to proactively identify potential churners, mitigate attrition, and foster long-term customer relationships.

The genesis of this project lies in the evolving landscape of consumer behaviour and market dynamics. Rapid technological advancements, coupled with the proliferation of digital platforms and alternative service providers, have ushered in an era of unprecedented choice and convenience for consumers. Against this backdrop, customer loyalty has become increasingly elusive, with customers exhibiting greater propensity to switch providers in pursuit of better deals,

superior service, or enhanced value propositions. In such a fiercely competitive environment, understanding the drivers of customer churn and devising effective retention strategies are paramount for companies seeking to maintain a competitive edge and sustain long-term growth.

Central to our project is the utilization of data-driven methodologies to unravel the underlying patterns and behaviours driving customer churn. By leveraging a rich dataset encompassing a diverse array of customer attributes, behaviours, and interactions, we seek to gain deeper insights into the factors influencing churn propensity and customer engagement. From demographic variables such as gender and marital status to transactional metrics such as payment preferences and service usage patterns, each facet of the dataset offers a window into the complex interplay of factors shaping customer decisions and behaviours. Through rigorous exploratory data analysis (EDA) and predictive modeling techniques, we aim to distill these multifaceted insights into actionable intelligence, enabling the company to anticipate churn risks, personalize retention efforts, and maximize customer lifetime value.

Furthermore, the success of our endeavour hinges not only on the technical prowess of our predictive models but also on the alignment of our recommendations with the broader business objectives and constraints of the company. By adhering to a data-driven, customer-centric approach, we endeavour to empower the company to navigate the complexities of customer churn, foster enduring customer relationships, and thrive in an ever-evolving marketplace.

1.3 PROBLEM STATEMENT

An E Commerce company or DTH (you can choose either of these two domains) provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because 1 account can have multiple customers. hence by losing one account the company might be losing more than one customer.

You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign. Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation.

Hence be very careful while providing campaign recommendation.

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_LY	How many times all the customers of the account has contacted customer care in last 12 months
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_ly	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 months)
coupon_used_for_payment	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

1.4 OBJECTIVE OF STUDY

- **Develop a robust churn prediction model:** Design and implement machine learning algorithms capable of accurately predicting customer churn based on historical data and relevant customer attributes.

- **Identify key drivers of customer churn:** Conduct exploratory data analysis (EDA) to uncover the underlying patterns, trends, and relationships within the dataset, with a specific focus on identifying the primary factors influencing customer churn.

- **Provide actionable insights for targeted retention efforts:** Leverage the findings from the churn prediction model and EDA to derive actionable intelligence, enabling the company to personalize retention strategies and mitigate churn risks effectively.

- **Formulate segmented offers based on churn propensity:** Segment customers based on their likelihood of churn and devise tailored offers and incentives aimed at retaining high-risk customers and maximizing customer lifetime value.

- **Ensure alignment with business objectives and constraints:** Prioritize campaign recommendations that strike a balance between reducing churn rates and optimizing revenue streams, taking into account the financial implications and constraints of the company.

- **Enhance customer satisfaction and loyalty:** By proactively addressing churn risks and offering personalized incentives, aim to enhance overall customer satisfaction, foster long-term loyalty, and strengthen the company's competitive position in the market.

- **Enable data-driven decision-making:** Establish a framework for ongoing monitoring and evaluation of churn prediction performance, enabling iterative refinement of predictive models and campaign strategies based on real-time feedback and insights.

- **Facilitate continuous improvement and innovation:** Foster a culture of data-driven innovation and continuous improvement within the organization by leveraging insights gained from the churn prediction project to inform product development, marketing strategies, and customer engagement initiatives.

These objectives collectively aim to empower the company to proactively manage customer churn, optimize customer retention efforts, and drive sustained business growth in an increasingly competitive marketplace.

1.5 COMPANY AND INDUSTRY OVERVIEW

The company at the heart of this project operates in the dynamic and competitive landscape of E-Commerce or Direct-to-Home (DTH) services, where customer retention holds paramount importance for sustained business success. As a prominent player in either sector, the company faces intense competition from a myriad of rivals vying for market share and customer loyalty. In the realm of E-Commerce, the company operates within a rapidly evolving ecosystem characterized by shifting consumer preferences, technological innovations, and disruptive market entrants. Likewise, in the realm of DTH services, the company contends with a diverse array of competitors offering alternative entertainment solutions and subscription-based services. Against this backdrop, understanding customer behaviour, anticipating churn risks, and devising effective retention strategies are critical imperatives for the company's continued growth and viability.

In the broader industry context, the E-Commerce and DTH sectors have witnessed exponential growth and transformation fueled by advancements in technology, changing consumer habits, and evolving regulatory landscapes. The proliferation of smartphones, widespread internet penetration, and the advent of digital payment solutions have catalyzed the expansion of E-Commerce platforms, enabling seamless online transactions and enhancing customer convenience. Similarly, in the realm of DTH services, the transition from traditional cable television to satellite-based broadcasting has revolutionized the way consumers access and consume entertainment content, offering a diverse array of channels, on-demand programming, and interactive features. Amidst these transformative shifts, companies operating in these sectors must navigate a myriad of challenges, including customer churn, pricing pressures, and regulatory compliance, while also capitalizing on emerging opportunities for innovation and growth.

Given the fiercely competitive nature of the industry and the ever-changing dynamics of consumer behaviour, the company recognizes the imperative to harness data-driven insights and predictive analytics to gain a competitive edge. By leveraging advanced analytics and machine learning techniques, the company seeks to proactively identify churn risks, personalize customer engagement, and optimize retention efforts, thereby ensuring long-term profitability and market leadership. In embarking on this endeavour, the company underscores its commitment to customer-centricity, innovation, and continuous improvement, positioning itself for sustained success in the dynamic landscape of E-Commerce or DTH services.

1.6 OVERVIEW OF THEORETICAL CONCEPTS

1. **Customer Churn:** Customer churn refers to the phenomenon of customers discontinuing their relationship with a company or brand. In this project, churn is a critical metric that serves as the target variable for predictive modeling, aiming to identify accounts at risk of churn and implement targeted retention strategies.

2. **Predictive Modeling:** Predictive modeling involves the use of statistical algorithms and machine learning techniques to predict future outcomes based on historical data. In the context of this project, predictive modeling techniques such as logistic regression, decision trees, or ensemble methods are employed to forecast customer churn probabilities and anticipate churn risks.

3. **Exploratory Data Analysis (EDA):** EDA is a fundamental data analysis technique used to explore and summarize the main characteristics of a dataset. By examining the distribution, summary statistics, and visual representations of variables, EDA provides insights into data patterns, trends, and relationships, laying the groundwork for subsequent modeling efforts.

4. **Feature Engineering:** Feature engineering involves selecting, transforming, or creating new features from raw data to improve the performance of machine learning models. In this project, feature engineering techniques such as encoding categorical variables, scaling numerical features, and creating derived features are employed to enhance the predictive power of the churn prediction model.

5. **Segmentation Analysis:** Segmentation analysis involves dividing customers into distinct groups based on shared characteristics or behaviours. In the context of churn prediction, segmentation analysis enables the company to identify different customer segments with varying churn propensities and tailor retention strategies accordingly.

6. **Customer Lifetime Value (CLV):** CLV is a metric that represents the total expected revenue generated by a customer over their entire relationship with the company. Understanding CLV is crucial for prioritizing retention efforts and allocating resources effectively, as high CLV customers may warrant more personalized retention incentives.

7. **Campaign Optimization:** Campaign optimization involves designing and implementing marketing campaigns aimed at maximizing the effectiveness of customer retention efforts. In this project, campaign optimization strategies focus on aligning targeted offers with customer preferences, optimizing the timing and delivery of retention incentives, and measuring the impact of campaigns on churn reduction and revenue optimization.

8. **Model Evaluation and Validation:** Model evaluation and validation are critical steps in assessing the performance and reliability of predictive models. Techniques such as cross-validation, ROC curves, and confusion matrices are used to measure the accuracy, precision, recall, and overall predictive power of the churn prediction model, ensuring its effectiveness in real-world applications.

These theoretical concepts collectively form the foundation for the Customer Churn Prediction project, enabling the company to leverage data-driven insights and predictive analytics to mitigate churn risks, enhance customer retention, and drive sustainable business growth.

CHAPTER 2

RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

2.1 SCOPE OF THE STUDY

The scope of this study encompasses the development and implementation of a churn prediction model for an E-Commerce company or DTH provider. The primary objective is to leverage predictive analytics and machine learning techniques to anticipate and mitigate customer churn, thereby enhancing customer retention rates and driving sustained business growth. The study focuses on analyzing a comprehensive dataset comprising various customer attributes and behaviours to identify key drivers of churn and formulate targeted retention strategies. Additionally, the study explores the utility of research findings in optimizing marketing campaigns, allocating resources effectively, and fostering long-term customer relationships. The research design involves a combination of exploratory data analysis (EDA) and predictive modeling to uncover patterns, trends, and relationships within the dataset.

2.2 METHODOLOGY

2.2.1 Research Design

This study adopts a quantitative research design, leveraging statistical analysis and machine learning algorithms to analyze data and derive actionable insights. The research design involves a combination of exploratory data analysis (EDA), predictive modeling, and segmentation analysis to uncover patterns, trends, and relationships within the dataset and develop a robust churn prediction model.

2.2.2 Data Collection

The data collection process involves obtaining a comprehensive dataset from the E-Commerce company or DTH provider, encompassing various customer attributes and behaviours relevant to churn prediction. Since the data is not collected directly by the company, it is considered as secondary data collection. Utilize industry reports, academic journals, and case studies to supplement understanding of customer churn dynamics and best practices in retention strategies The dataset is curated and preprocessed to ensure data quality and consistency before analysis.

2.2.3 Sampling Method

In cases where sampling is required, a stratified sampling method may be employed to ensure representation from different customer segments or geographical regions. However, since the dataset is likely to include historical records of all accounts, sampling may not be necessary for this study.

2.2.4 Data Analysis Tools

The data analysis process utilizes a combination of statistical software and machine learning libraries, such as Python's pandas, NumPy, scikit-learn, matplotlib and Seaborn. These tools enable data exploration, preprocessing, model development, visualization and evaluation, facilitating a comprehensive analysis of churn prediction factors and the generation of actionable insights.

Summary of Approach to EDA and Pre-processing

EDA begins with an exploration of the dataset's structure, including variable types and missing values. Visualizations such as bar plots, pie plots, scatter plots, histograms, box plots, and correlation matrices are used to identify patterns and relationships between variables. Pre-processing involves handling missing values, encoding categorical variables, and scaling numerical features.

Insightful Visualizations:

- A histogram of tenure distribution reveals a bimodal pattern, indicating distinct groups of long-term and short-term customers.
- A correlation matrix highlights strong positive correlations between service score and customer satisfaction, indicating a relationship between service quality and retention.
- A box plot of churn by payment method suggests that customers preferring certain payment modes may be more prone to churn.

Meaningful Features:

- Service score and service satisfaction emerge as significant predictors of churn, indicating the importance of service quality in retaining customers.
- Customer tenure and complaint history also exhibit strong associations with churn, suggesting that long-term customers with unresolved complaints may be at higher risk.

2.3 PERIOD OF STUDY

The period of study encompasses historical data spanning a specified timeframe, typically ranging from the last few years to the present. The exact duration may vary depending on data availability and business requirements but generally covers a sufficient timeframe to capture meaningful trends and patterns in customer behaviour and churn dynamics. As far as 'Customer Churn' data, we don't have time frame that from when data is collected but we do have sufficient data to preprocess and get some valuable insights.

2.4 UTILITY OF RESEARCH

The research findings inform strategic decision-making, marketing campaign optimization, and resource allocation, driving sustained business growth and competitive advantage. The churn prediction model and actionable insights enable the company to proactively identify at-risk customers, personalize retention strategies, and enhance overall customer satisfaction and loyalty.

CHAPTER 3

DATA ANALYSIS AND INTERPRETATION

DATA ANALYSIS AND INTERPRETATION

3.1 Loading the Data

Loaded the required packages, set the work directory and load the datafile.
Data set has 11,260 number of observations and 19 variables (18 independent and 1 dependent or target variable).

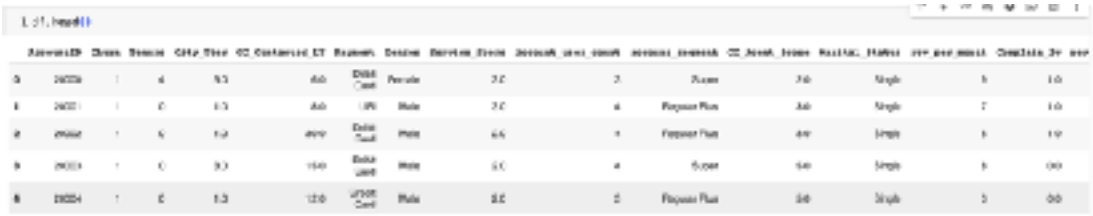

Table 1 - glimpse of dataset

Understanding how data was collected in terms of time, frequency and methodology

- Data has been collected for random 11,260 unique account ID, across gender and marital status.
- Looking at variables "CC_Contacted_L12m", “rev_per_month”, “Complain_112m”, “rev_growth_yoy”, “coupon _used_112m”, “Day_Since_CC_connect” and “cashback_112m” we can conclude that the data has been collected for last 12 month.
- Data has 19 variables, 18 independent and 1 dependent or the target variable, which shows if customer churned or not.
- The data is the combination of services customers are using along with their payment option and also then basic individuals details as well.
- Data is mixed of categorical as well as continuous variables.

Visual inspection of data (rows, columns, descriptive details)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   AccountID                    11260 non-null  int64
 1   Churn                        11260 non-null  int64
 2   Tenure                       11158 non-null  object
 3   City_Tier                    11148 non-null  float64
 4   CC_Contacted_LY              11158 non-null  float64
 5   Payment                      11151 non-null  object
 6   Gender                       11152 non-null  object
 7   Service_Score                11162 non-null  float64
 8   Account_user_count           11148 non-null  object
 9   account_segment              11163 non-null  object
 10  CC_Agent_Score               11144 non-null  float64
 11  Marital_Status               11048 non-null  object
 12  rev_per_month                11158 non-null  object
 13  Complain_ly                  10903 non-null  float64
 14  rev_growth_yoy               11260 non-null  object
 15  coupon_used_for_payment      11260 non-null  object
 16  Day_Since_CC_connect         10903 non-null  object
 17  cashback                     10789 non-null  object
 18  Login_device                 11039 non-null  object
dtypes: float64(5), int64(2), object(12)
memory usage: 1.6+ MB
```

Table 2 - Dataset information

Statistical analysis of numeric columns on the dataset

	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626960	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.0	0.00	1.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.0	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.0	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.0	3.00	5.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.0	4.00	5.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.0	1.00	1.0

Table 3 - Statistical information of numeric variables

Statistical analysis of categorical columns on the dataset

	count	unique	top	freq
Tenure	11158	38	1	1351
Payment	11151	5	Debit Card	4587
Gender	11152	4	Male	6328
Account_user_count	11148	7	4	4569
account_segment	11163	7	Super	4062
Marital_Status	11048	3	Married	5860
rev_per_month	11158	59	3	1746
rev_growth_yoy	11260	20	14	1624
coupon_used_for_payment	11260	20	1	4373
Day_Since_CC_connect	10903	24	3	1816
cashback	10789.0	5693.0	155.62	10.0
Login_device	11039	3	Mobile	7482

Table 4 - Statistical information of categorical variables

3.2 Data Cleaning

Treating 'Tenure' column

- As you can see there is '#' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '#s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
Tenure: 38
[4 0 2 13 11 '#' 9 99 19 20 14 8 26 18 5 30 7 1 23 3 29 6 28
24 25 16 10 15 22 nan 27 12 21 17 50 60 31 51 61]
```

After treatment

```
array([ 4.,  0.,  2., 13., 11., nan,  9., 99., 19., 20.,
14.,  8., 26., 18.,  5., 30.,  7.,  1., 23.,  3., 29.,  6.,
28., 24., 25., 16., 10., 15., 22., 27., 12., 21., 17., 50.,
60., 31., 51., 61.])
```

Treating 'Gender' column

- As we can see 'F' and 'M' are also known as 'Female' and 'Male'.

Before treatment

```
Gender: 4
['Female' 'Male' 'F' nan 'M']
```

After treatment

```
array(['Female', 'Male', nan], dtype=object)
```

Treating 'Account_user_count' column

- As you can see there is '@' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '@s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
Account_user_count: 7
[3 4 nan 5 2 '@' 1 6]
```

After treatment

```
array([ 3.,  4., nan,  5.,  2.,  1.,  6.])
```

Treating 'account_segment' column

- As you can see there is 'Regular +' and 'Super +' symbol need to be treated as 'Regular Plus' and 'Super Plus'.
- Since 'Regular +' is nothing but 'Regular Plus' and same in case of 'Super +'.

Before treatment

```
account_segment: 7
['Super' 'Regular Plus' 'Regular' 'HNI' 'Regular +' nan
'Super Plus' 'Super +']
```

After treatment

```
array(['Super', 'Regular Plus', 'Regular', 'HNI', nan,
'Super Plus'], dtype=object)
```

Treating 'rev_per_month' column

- As you can see there is '+' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '+s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
rev_per_month: 59
[9 7 6 8 3 2 4 10 1 5 '+' 130 nan 19 139 102 120 138 127 123
124 116 21 126 134 113 114 108 140 133 129 107 118 11 105 20
119 121 137 110 22 101 136 125 14 13 12 115 23 122 117 131
104 15 25 135 111 109 100 103]
```

After treatment

```
array([  9.,   7.,   6.,   8.,   3.,   2.,   4.,  10.,   1.,
5.,  nan, 130.,  19., 139., 102., 120., 138., 127., 123.,
124., 116.,  21., 126., 134., 113., 114., 108., 140., 133.,
129., 107., 118.,  11., 105.,  20., 119., 121., 137., 110.,
22., 101., 136., 125.,  14., 13.,  12., 115.,  23., 122.,
117., 131., 104.,  15.,  25., 135., 111., 109., 100., 103.])
```

Treating 'rev_growth_yoy' column

- As you can see there is '$' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '$s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
rev_growth_yoy: 20
[11 15 14 23 22 16 12 13 17 18 24 19 20 21 25 26 '$' 4 27
28]
```

After treatment

```
array([11., 15., 14., 23., 22., 16., 12., 13., 17., 18.,
24., 19., 20., 21., 25., 26., nan,  4., 27., 28.])
```

Treating 'coupon_used_for_payment' column

- As you can see there is '#', '$' and '*' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '#s', '$s' and '*s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
coupon_used_for_payment: 20
[1 0 4 2 9 6 11 7 12 10 5 3 13 15 8 '#' '$' 14 '∗' 16]
```

After treatment

```
array([ 1.,  0.,  4.,  2.,  9.,  6., 11.,  7., 12., 10.,
5.,  3., 13., 15.,  8., nan, 14., 16.])
```

Treating 'Day_Since_CC_connect' column

- As you can see there is '$' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '$s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
Day_Since_CC_connect: 24
[5 0 3 7 2 1 8 6 4 15 nan 11 10 9 13 12 17 16 14 30 '$' 46
18 31 47]
```

After treatment

```
array([ 5.,  0.,  3.,  7.,  2.,  1.,  8.,  6.,  4., 15.,
nan, 11., 10., 9., 13., 12., 17., 16., 14., 30., 46., 18.,
31., 47.])
```

Treating 'cashback' column

- First need to know what is the special character.
- As you can see there is '$' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '$s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
cashback: 5693
[159.93 120.9 nan ... 227.36 226.91 191.42]
```

After treatment

```
array([159.93, 120.9 ,    nan, ..., 227.36, 226.91, 191.42])
```

Treating 'Login_device' column

- As you can see there is '&&&&' symbol need to be treated as 'np.nan' value and we can perform imputation in future.
- Once the '&&&&s' are converted to 'np.nan', this column will be converted to numerical column.

Before treatment

```
Login_device: 3
['Mobile' 'Computer' '&&&&' nan]
```

After treatment

```
array(['Mobile', 'Computer', nan], dtype=object)
```

3.3 Exploratory Data Analysis (EDA)

3.3.1 Univariate Analysis

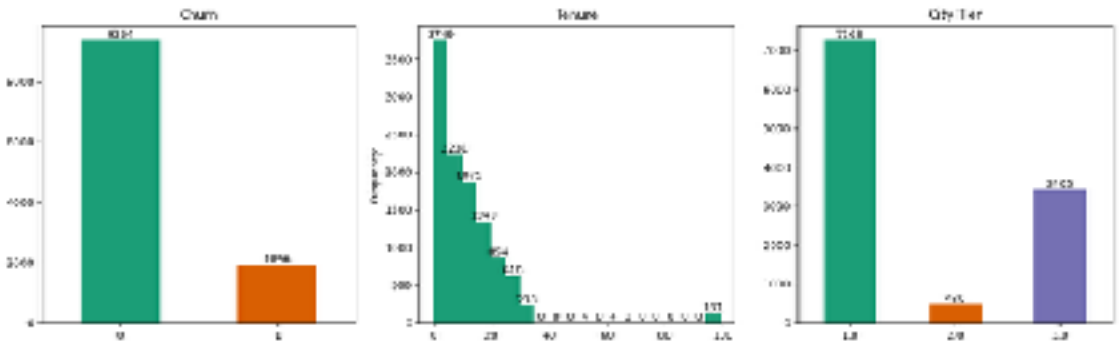

Fig 1.1 - Distribution of Churn, Tenure, City Tier

Data Insights:

- **Churn** - Imbalance dataset, as Churn samples are less.
- **Tenure** - More number of customers are falling under less Tenure.
- **City_Tier** - More number of customers are from City Tier 1 followed by City Tier 3 and followed by City Tier 2

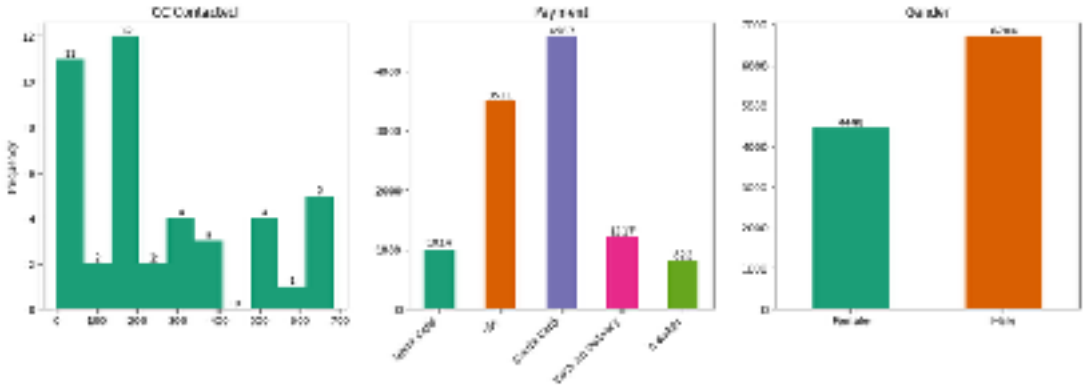

Fig 1.2 - Distribution of CC_Contacted, Payment, Gender

Data Insights:

- **CC_Contacted** - CC_Contacted doesn't follow any distribution.
  Payment - Most of the customers preferred Payment using credit card followed by UPI.
  Gender - Most of the customers are Male.

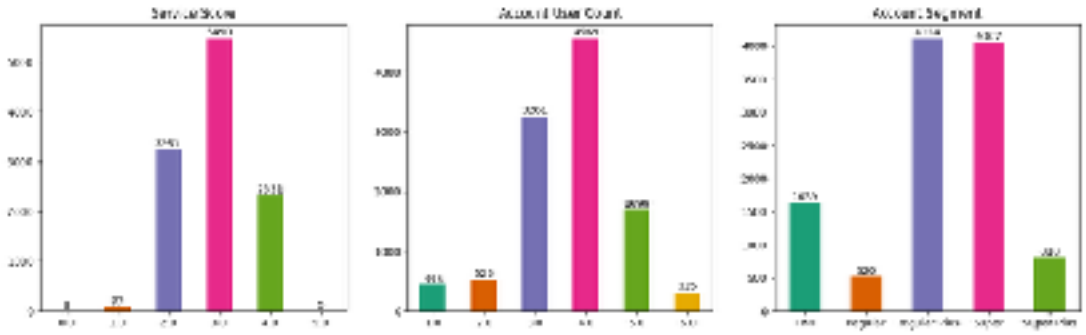

Fig 1.3 - Distribution of Service_Score, Account_user_Count, Account_Segment

Data Insights:

- **Service_Score** - Most of the customers provided Service Score as 3.
- **Account_User_Count** - Most of the customers using 4 accounts.
- **Account_Segment** - Most of the customers are under Regular Plus and Super Account Segment.

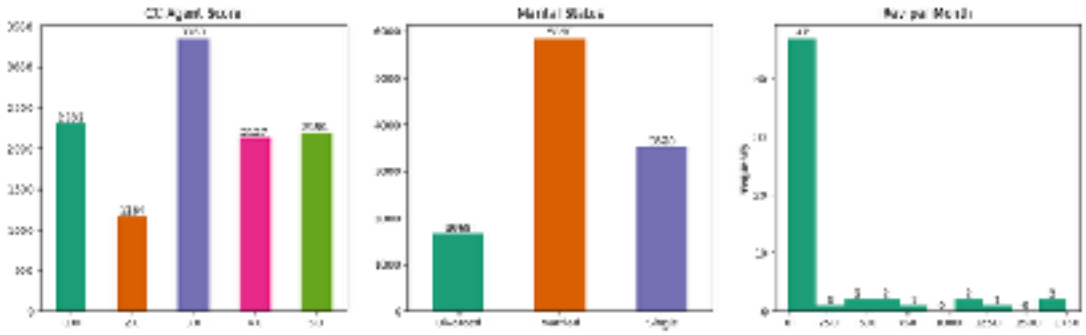

Fig 1.4 - Distribution of CC_Agent_Score, Marital_Status, Rev_per_Month

Data Insights:

- **CC_Agent_Score** - Most of the customers provided score as 3 for the CC Agent.
  Marital_Status - Most of the customers are under married and followed by Single users.
  Rev_per_Month - Most of the customers are under Regular Plus and Super Account Segment.

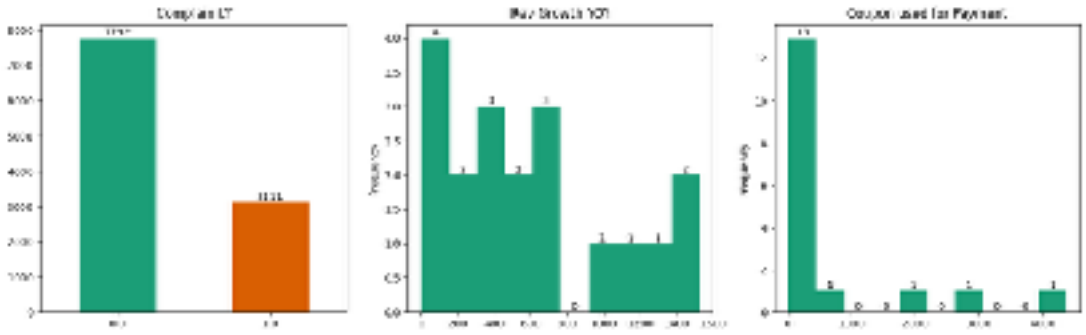

Fig 1.5 - Distribution of Complain_LY, Rev_Growth_YOY, Coupon_used_for_Payment

Data Insights:

- **Complain_LY** - Complain_ly has less complains.
- **Rev_Growth_YOY** - Rev_growth_yoy does seems like normal distribution.
- **Coupon_used_for_Payment** - Most of the customers used Coupon for payment between 0 and 500.

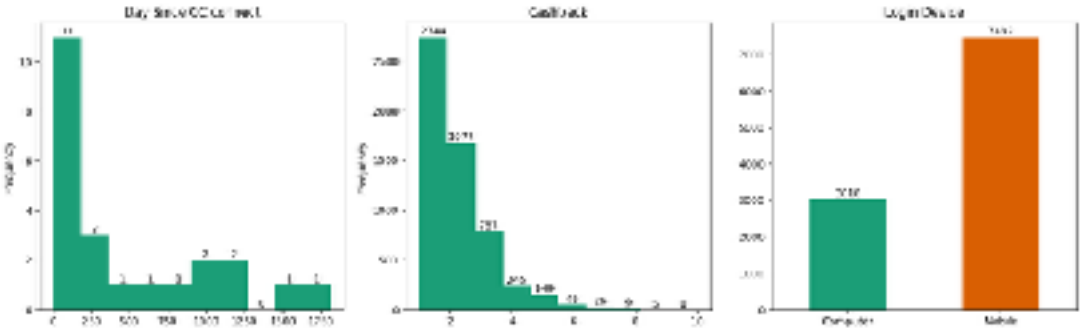

Fig 1.6 - Distribution of Day_Since_CC_Connect, Cashback, Login_Device

Data Insights:

- **Day_Since_CC_Connect** - Most of the customers has not contacted CC connect.
  Cashback - Most of the customers got Cashback which is less in amount.
  Login_Device - Most of the customers Login Device is Mobile compared to Computer.

3.3.2 Bivariate Analysis

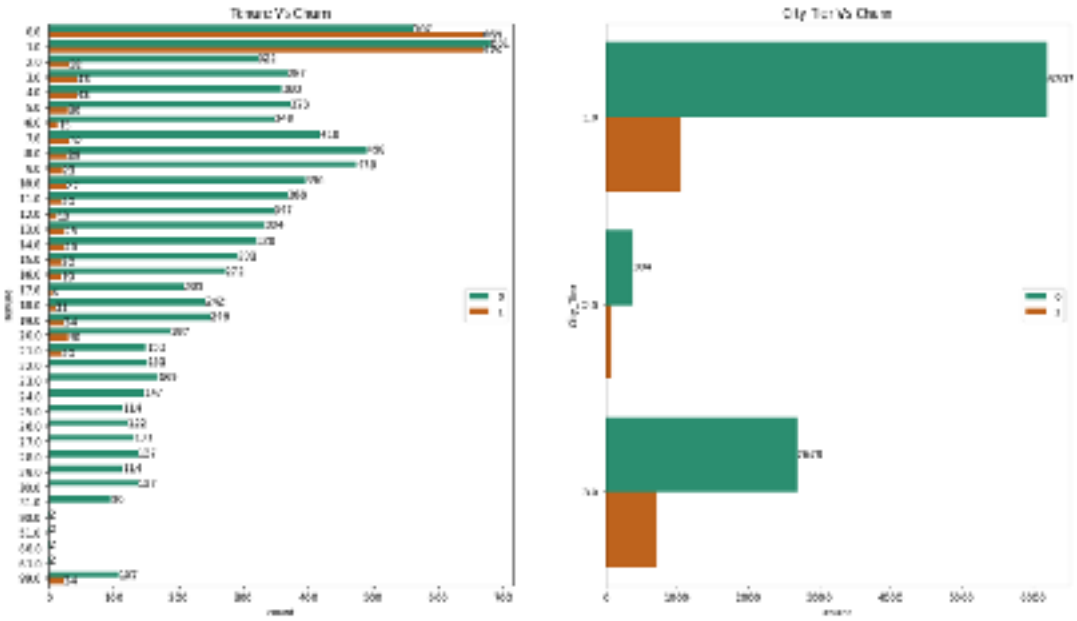

Fig 2.1 - Tenure Vs Churn and City_Tier Vs Churn

Data Insights:

- **Tenure Vs Churn** - If the tenure is less, the Churn is more.
- **City_Tier Vs Churn** - City_Tier 2 and 3 has comparably more number of Churn rate.

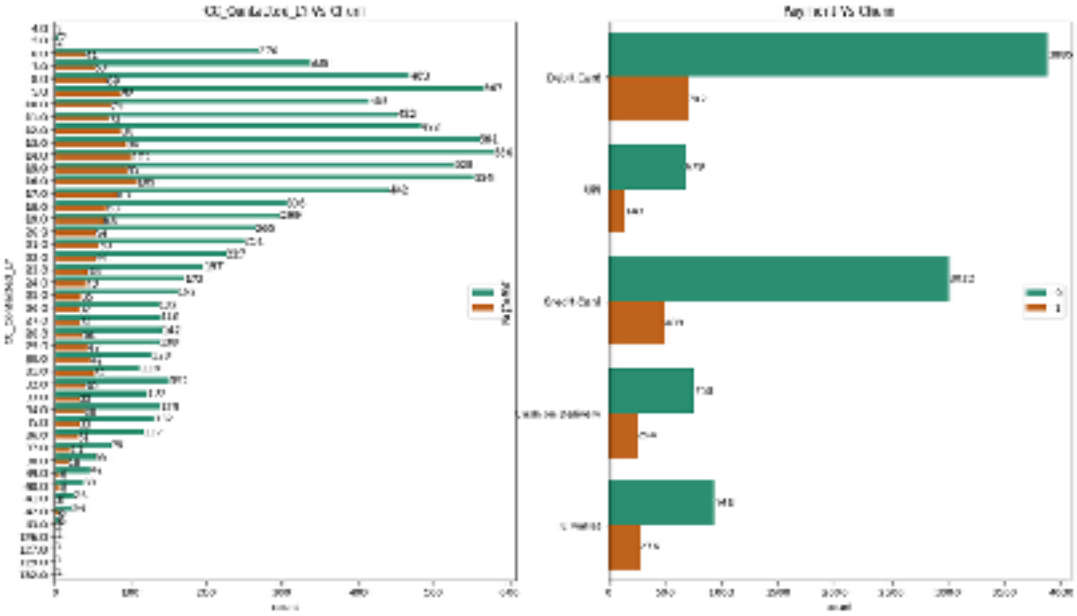

Fig 2.2 - CC_Contacted_LY Vs Churn and Payment Vs Churn

Data Insights:

- **CC_Contacted_LY Vs Churn** - If the CC_Contacted_LY is more, the Churn is also more.
- **Payment Vs Churn** - Except Debit and Credit card all other payment methods are having high Churn rate.

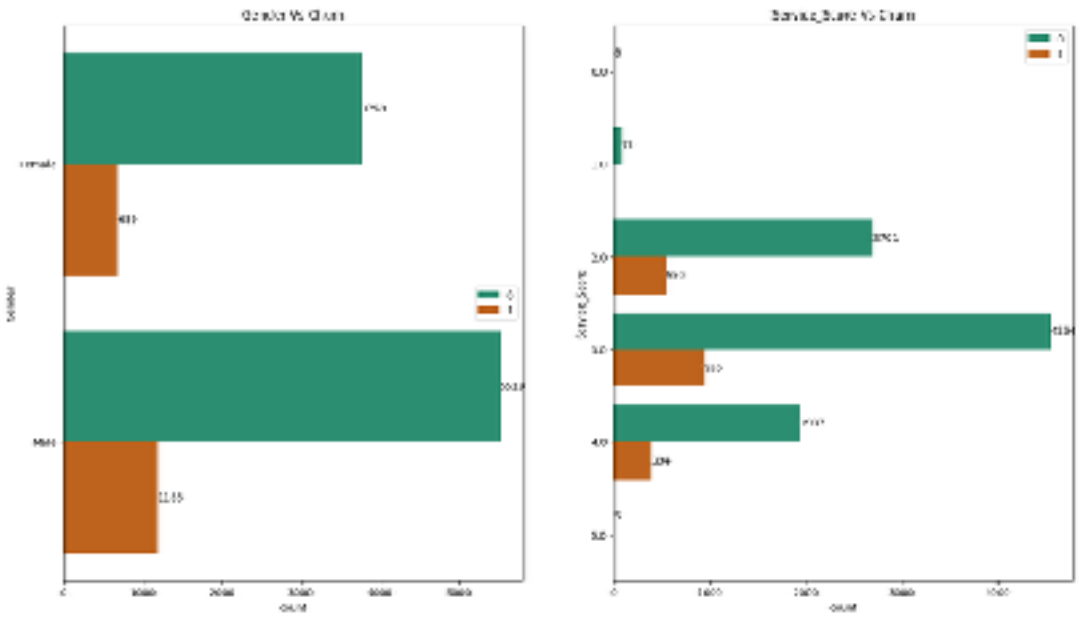

Fig 2.3 - Gender Vs Churn and Service_Score Vs Churn

Data Insights:

- **Gender Vs Churn** - Gender doesn't show any dependency over Churn.
- **Service_Score Vs Churn** - It seems like Service_Score has a positive relation with Churn.

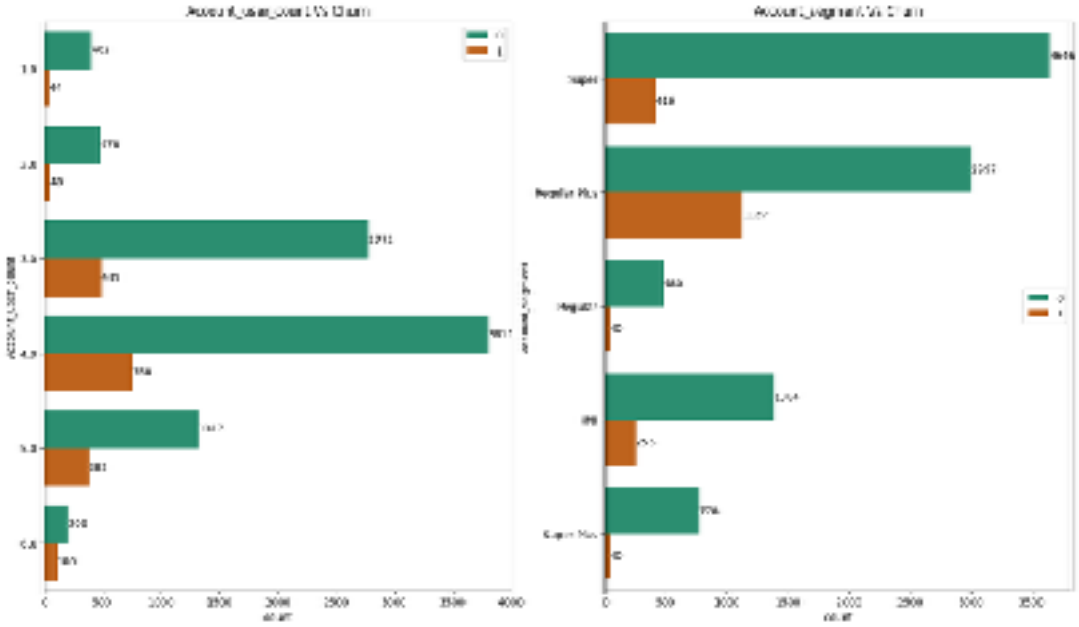

Fig 2.4 - Account_user_count Vs Churn and Account_Segment Vs Churn

Data Insights:

- **Account_user_count Vs Churn** - Account_user_count shows positive relation with Churn rate.
- **Account_Segment Vs Churn** - It seems like Regular Plus Segment user are more likely to Churn.

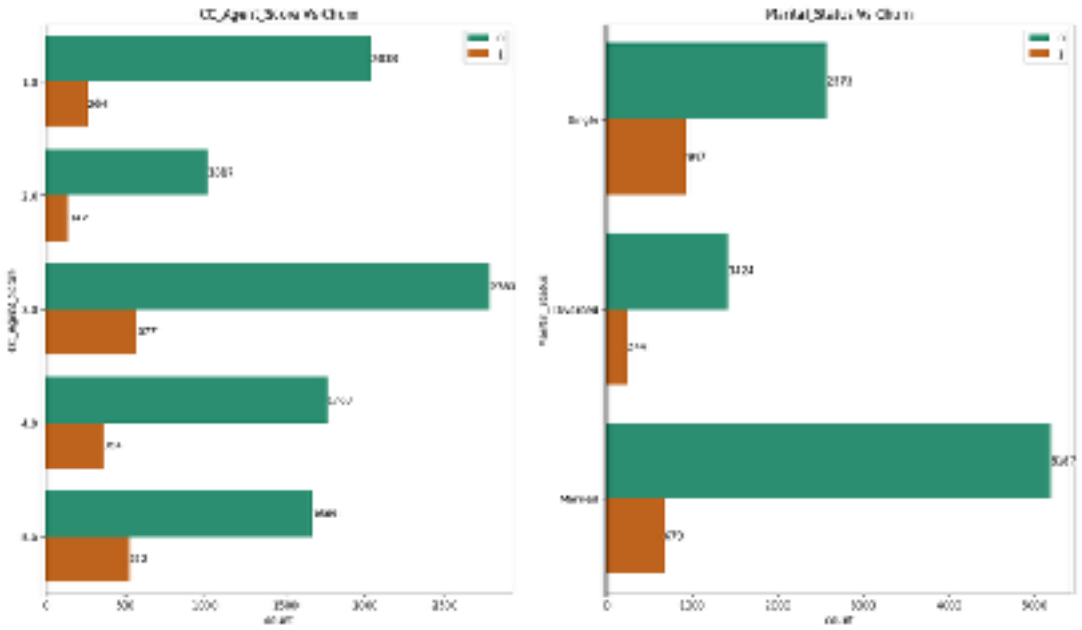

Fig 2.5 - CC_Agent_Score Vs Churn and Marital_Status Vs Churn

Data Insights:

- **CC_Agent_Score Vs Churn** - CC_Agent_Score shows no relation with Churn rate.
- **Marital_Status Vs Churn** - It seems like Single Marital_status user are more likely to Churn.

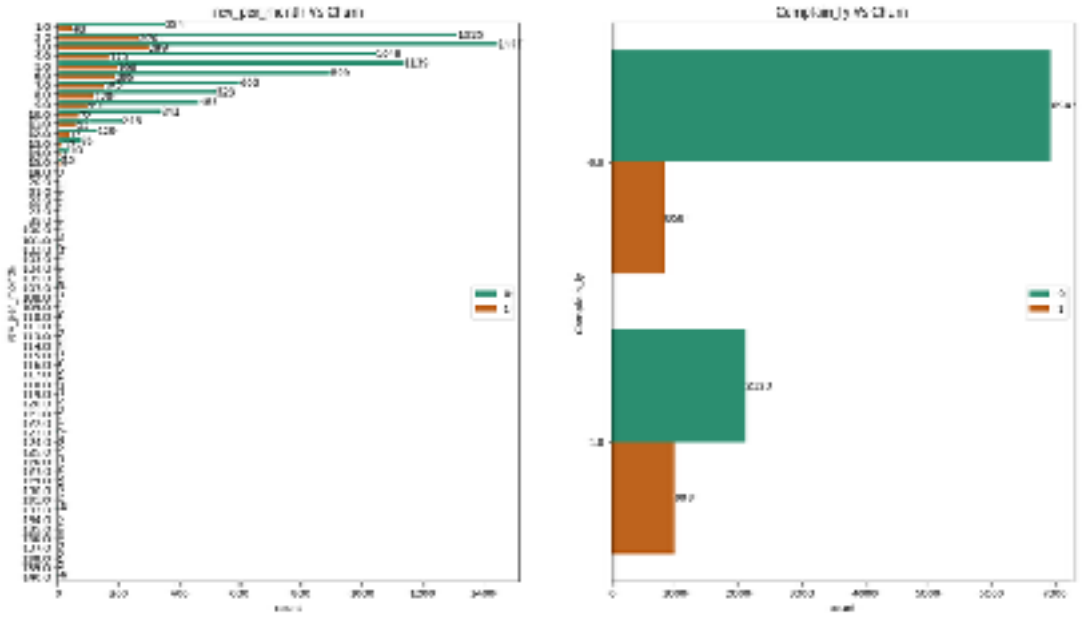

Fig 2.6 - Rev_per_Month Vs Churn and Company_LY Vs Churn

Data Insights:

- **Rev_per_Month Vs Churn** - rev_per_month shows positive relation with Churn rate.
- **Company_LY Vs Churn** - Positive Complain_ly has more likely to Churn.

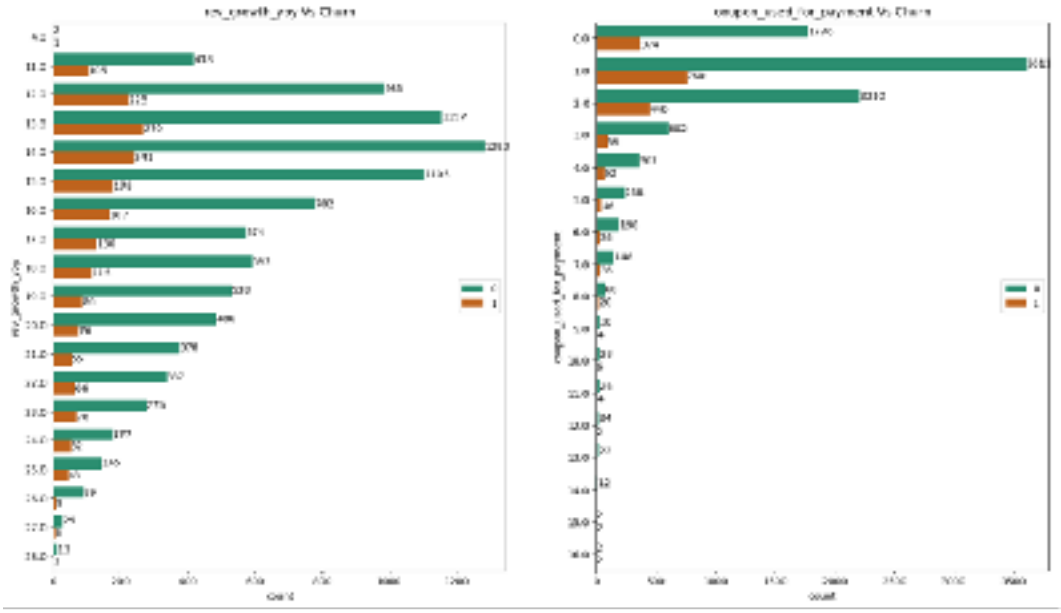

Fig 2.7 - rev_growth_yoy Vs Churn and coupon_used_for_payment Vs Churn

Data Insights:

- **rev_growth_yoy Vs Churn** - rev_growth_yoy shows positive relation with Churn rate.
- **coupon_used_for_payment Vs Churn** - coupon_used_for_payment shows positive relation with Churn rate.

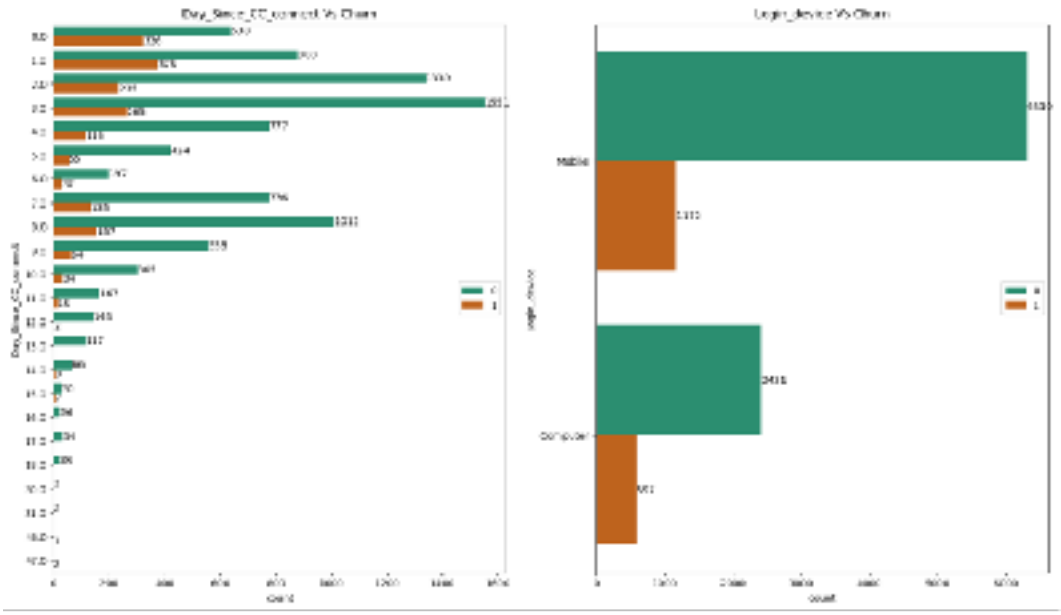

Fig 2.8 - Day_since_CC_connect Vs Churn and Logic_device Vs Churn

Data Insights:

- **Day_since_CC_connect Vs Churn** - Day_Since_CC_connect shows positive relation with Churn rate.
- **Logic_device Vs Churn** - Login_device shows not much of the relation with Churn rate.

3.3.3 Multivariate Analysis

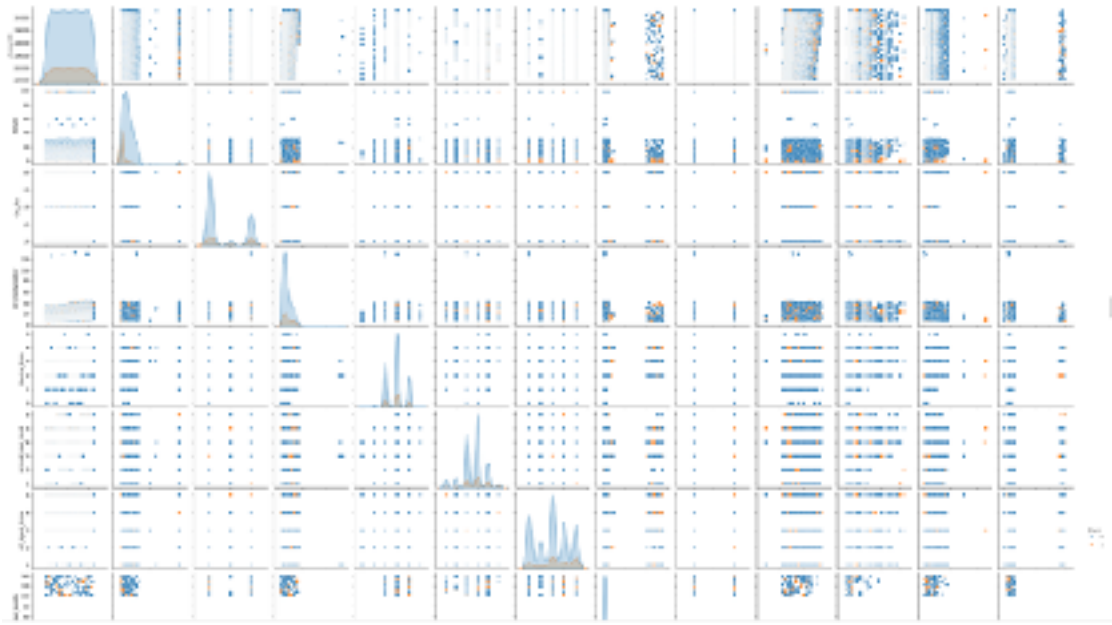

Fig 3 - Pair plot to get the relation between two variables against Churn

Data Insights:

- The pair-plot shown above indicated that the independent variable are week or poor predictors of target variable as we the density of independent variable overlaps with the density of target variable.

3.3.4 Correlation Matrix

Data Insights:

- The Data doesn't shows much of the correlation with each other except few columns (i.e., Day_Since_CC_connect with coupon_used_for_payment, Account_user_count with Service_Score and negative correlation with Tenure and Churn along with Day_Since_CC_connect.

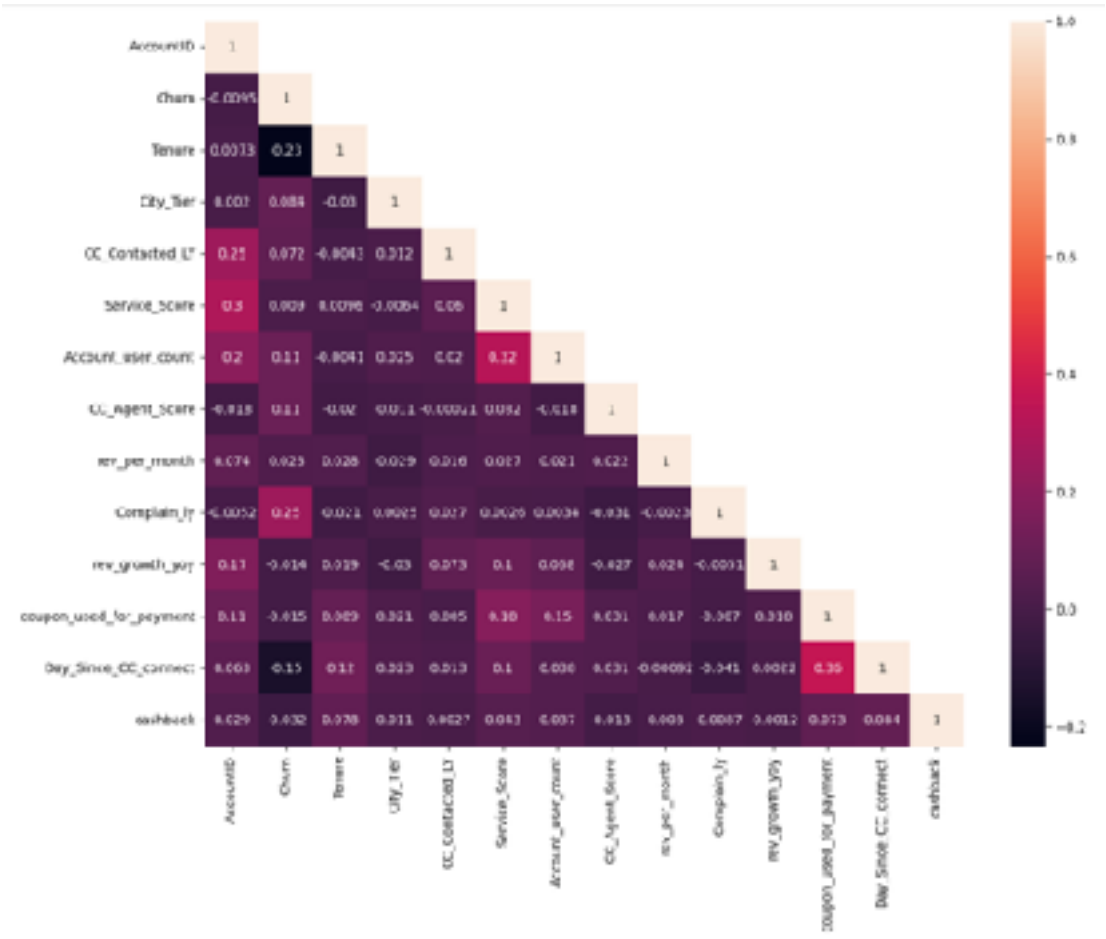

Fig 4 - Correlation Matrix

3.3.5 Handling Missing Values

Before Treatment:

- Out of 19 variables we have null values in 17 variables.

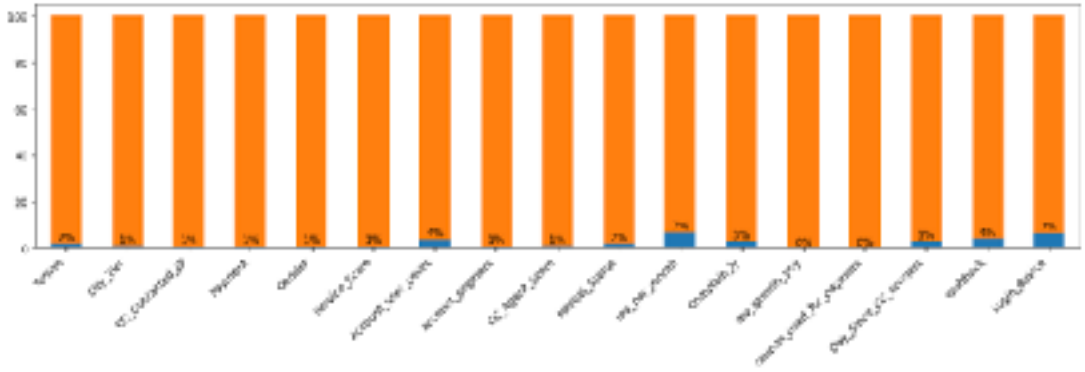

Fig 5.1 - Missing value before treatment

After Treatment:

- Using "Median" to impute null values where variable is continuous in nature because Median is less prone to outliers when compared with mean.
- Using "Mode: to impute null values where variables are categorical in nature. We have treated null values variable by variable as each and every variable is unique in its nature.

```
Churn                        0
Tenure                       0
City_Tier                    0
CC_Contacted_LY              0
Payment                      0
Gender                       0
Service_Score                0
Account_user_count           0
account_segment              0
CC_Agent_Score               0
Marital_Status               0
rev_per_month                0
Complain_ly                  0
rev_growth_yoy               0
coupon_used_for_payment      0
Day_Since_CC_connect         0
cashback                     0
Login_device                 0
dtype: int64
```

Fig 5.2 - Missing value after treatment

3.3.6 Outliers Treatment

Tenure

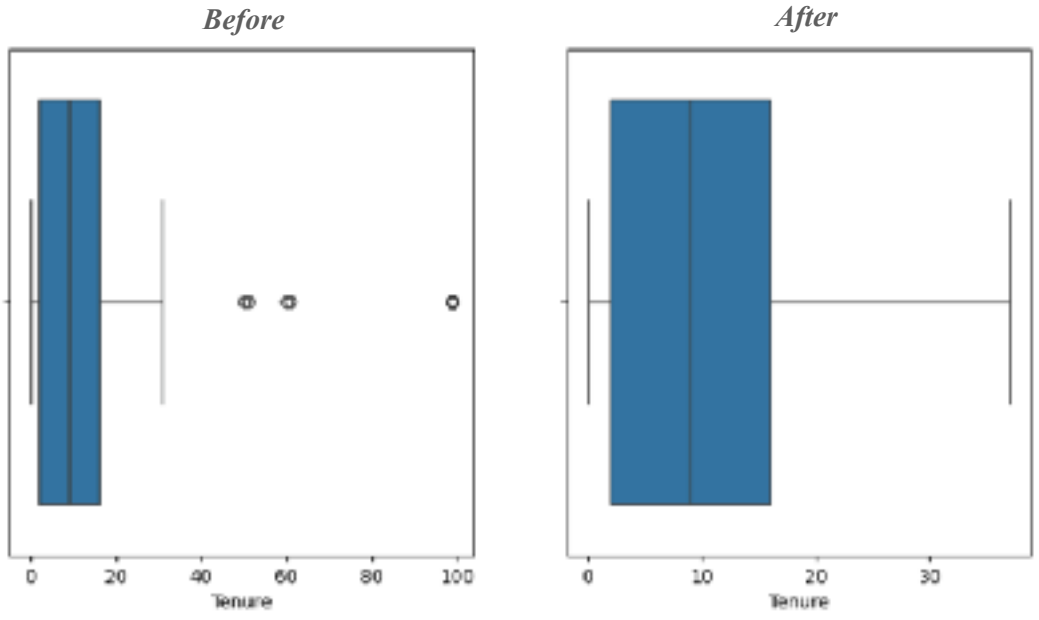

Fig 6.1 - Tenure Outlier Treatment

CC_Contacted_LY

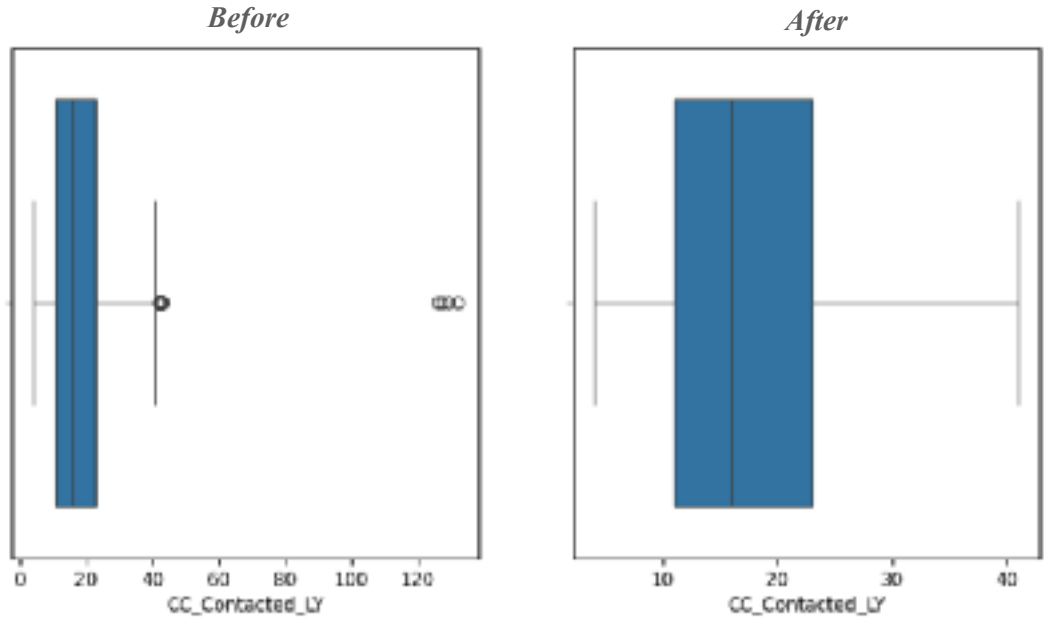

Fig 6.2 - CC_Contacted_LY Outlier Treatment

Account_user_count

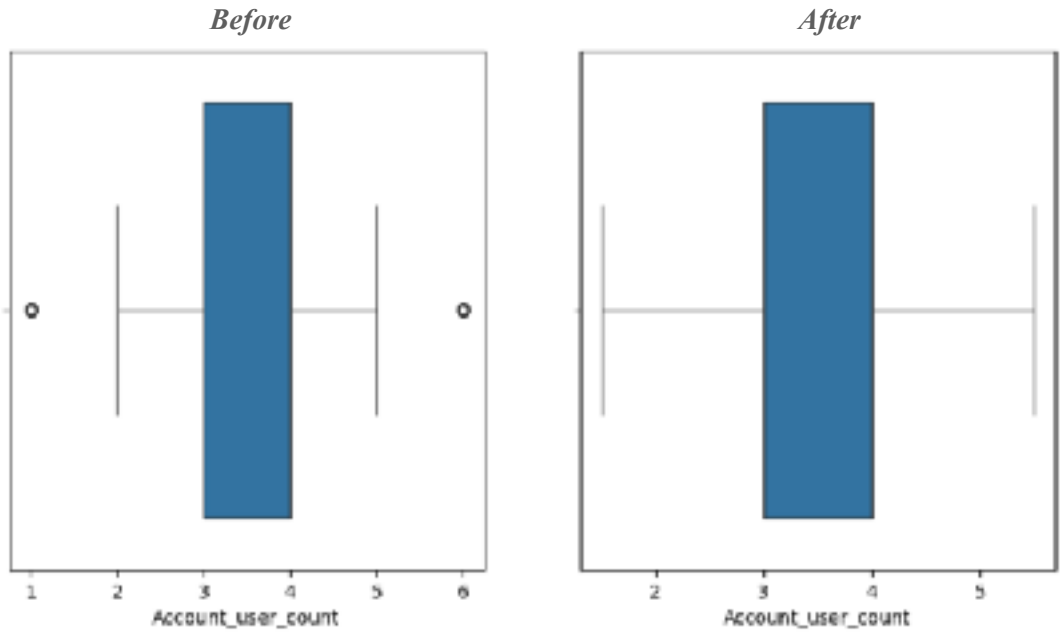

Fig 6.3 - Account_user_count Outlier Treatment

rev_per_month

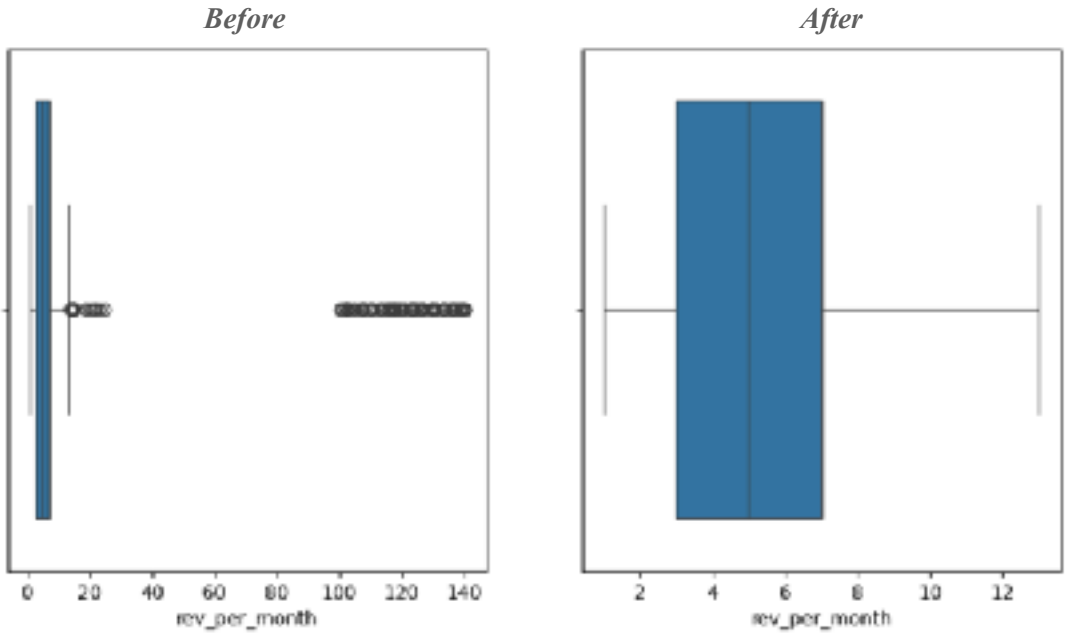

Fig 6.4 - rev_per_month Outlier Treatment

rev_growth_yoy

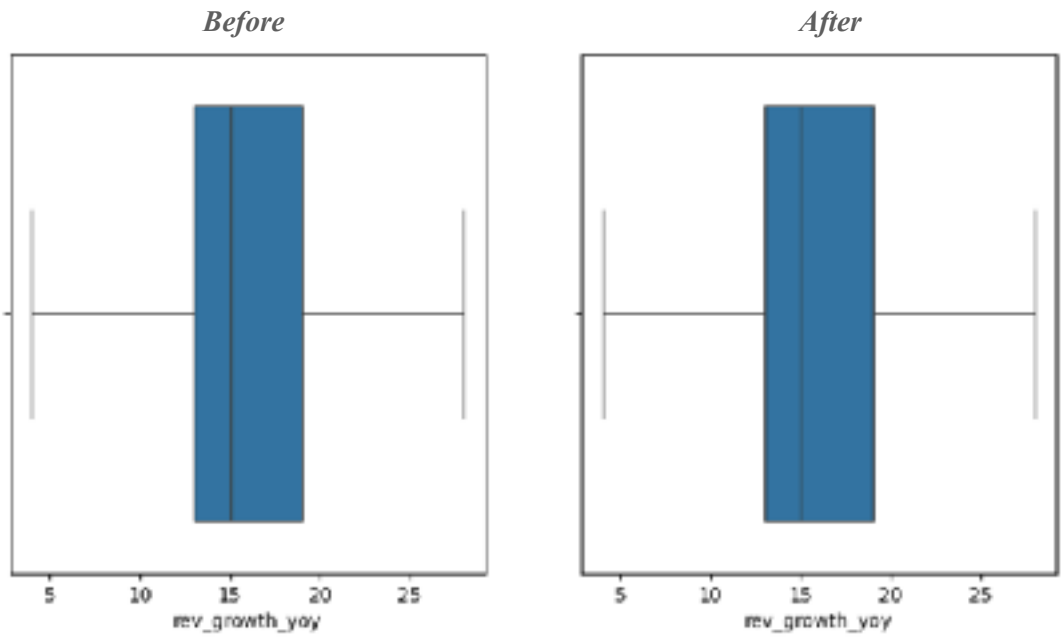

Fig 6.5 - rev_growth_yoy Outlier Treatment

coupon_used_for_payment

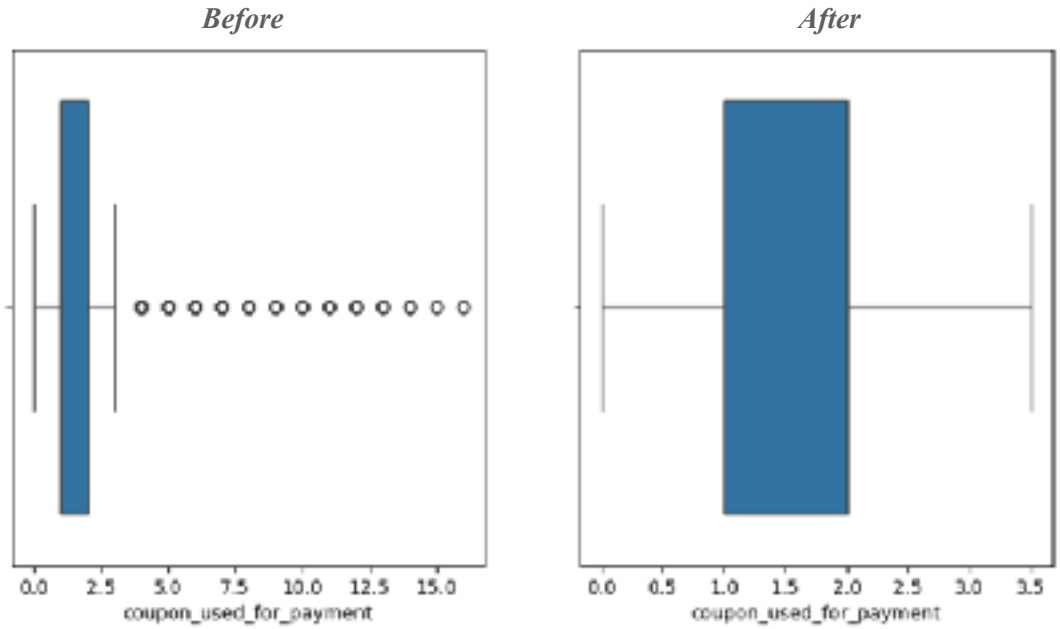

Fig 6.6 - coupon_used_for_payment Outlier Treatment

Day_Since_CC_connect

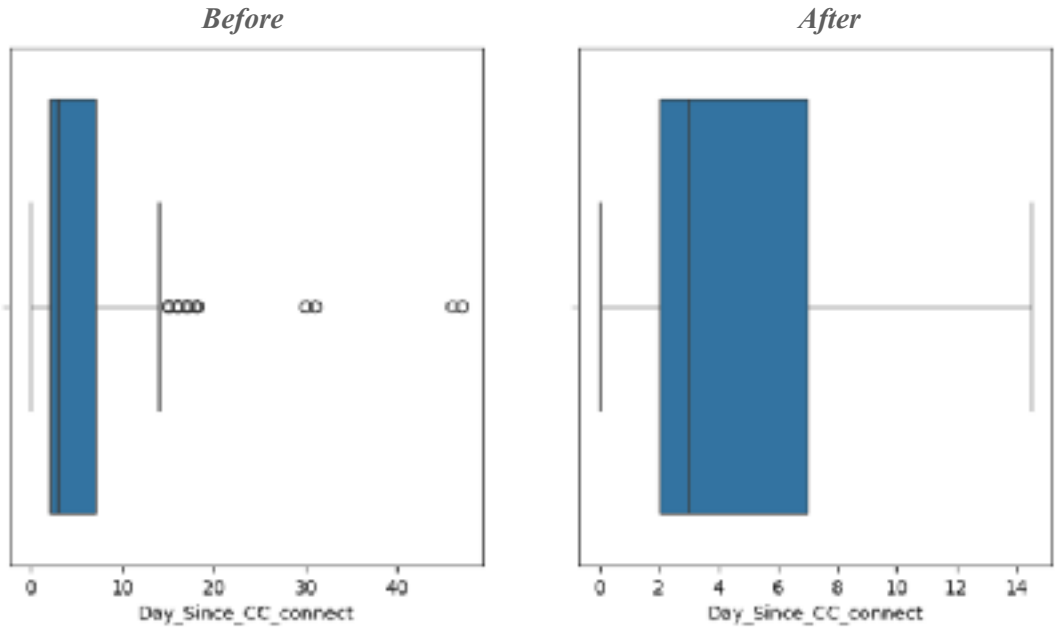

Fig 6.7 - Day_Since_CC_connect Outlier Treatment

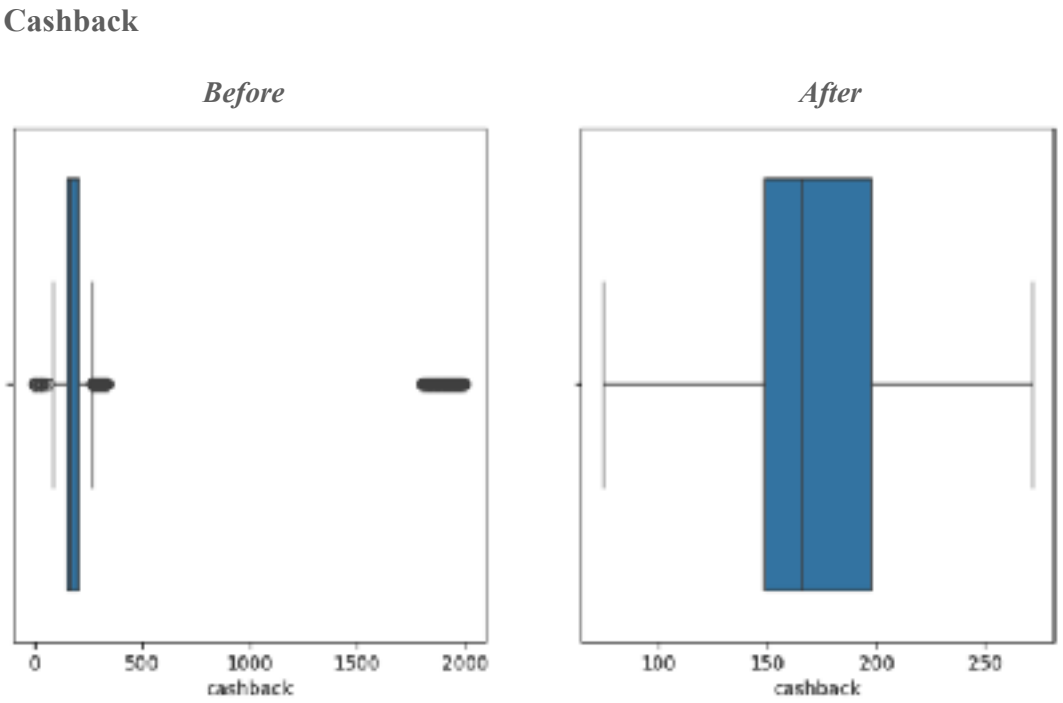

Fig 6.8 - Cashback Outlier Treatment

3.4 Data Preprocessing

3.4.1 Splitting the dataset

```
[ ]  1 X = df.drop('Churn', axis=1)
     2 y = df[['Churn']]
```

Dividing the dataset into X and y. Where X containing independent variables and y containing dependent variable.

```
[ ]  1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
     2
     3 X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((9800, 17), (2252, 17), (9800, 1), (2252, 1))
```

Splitting the dataset into 80% of train and 20% of test dataset.

Fig 7 - Dataset splitting

3.4.2 Encoding Categorical Variables

Table 5 - Encoded the dataset

Selecting all the columns which are of object type such as (Payment, Gender, Marital_Status, etc.,) and converting into numerical data using LabelEncoder from scikit-learn library.

3.4.3 Scaling the Dataset

Table 6 - Scaling the dataset

Data Insights:

Since the data doesn't follow uniform distribution or bell curve shape, we need to use normalization instead of standardization for better results.

3.4.4 Treating Imbalanced Dataset

Before Sampling

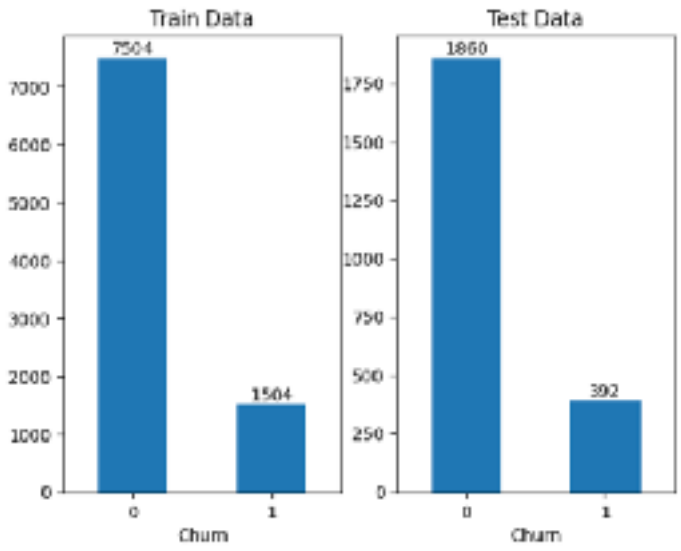

Fig 8.1 - Imbalance before sampling

After Sampling

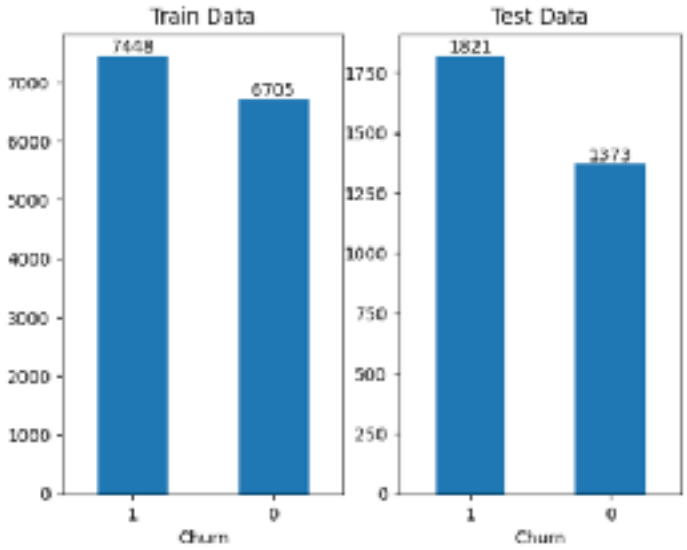

Fig 8.2 - Imbalance After sampling

Data Insights:

Since the dataset is imbalance, it will be hard for the model to learn about the minority class. So it is preferred to balance the dataset using the techniques available. I chose to use 'SMOTEENN' for balancing the dataset.

3.4.5 Clustering the Dataset

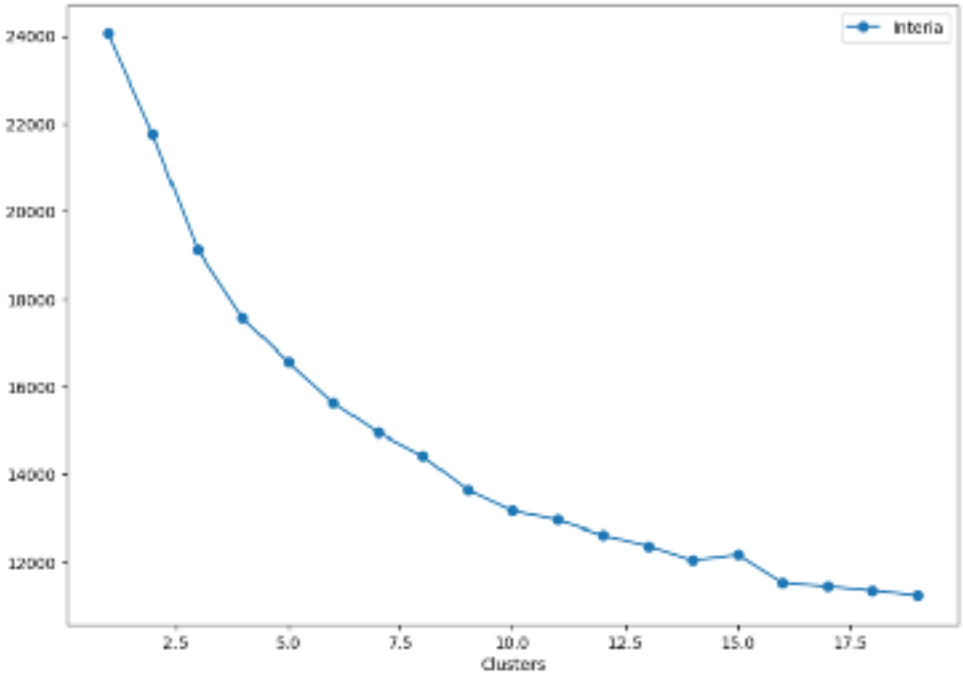

Fig 9.1 - Clustering Interia

Data Insights:

By the graph of ‘interia’ we can see the ‘L’ bow curve points around 7. Using 7 clusters to group the data using k-means clustering technique.

Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device	Clusters
0.0	0.466533	1.000000	0.827586	0.882848	1.0	2
1.0	0.500000	0.857143	0.206897	0.544759	1.0	0
0.0	0.916667	0.571429	0.620690	1.000000	0.0	5
0.0	0.375000	0.285714	0.206897	0.377985	0.0	2
1.0	0.291667	0.285714	0.068966	0.383004	0.0	5

Table 7 - After clustering

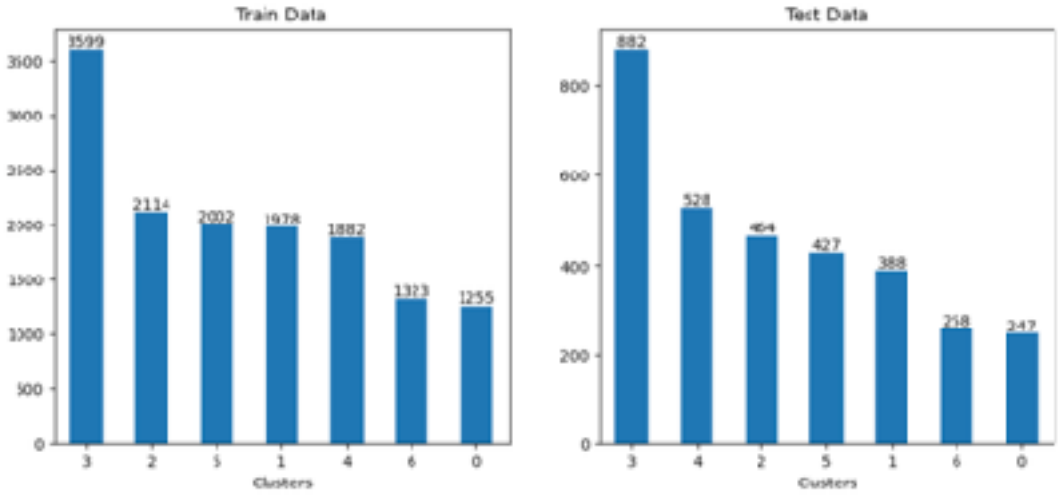

Fig 9.2 - Clusters group

Data Insights:

We can observe that in the training data, most of the data are grouped under cluster ‘3’ followed by ‘2’ and then followed by ‘5’ and so on.

Similar to training data, even test data has group most under cluster ‘3’ followed by ‘4’ and then followed by ‘2’ and so on.

CHAPTER 4

TRAINING THE MODEL

4. Training the model

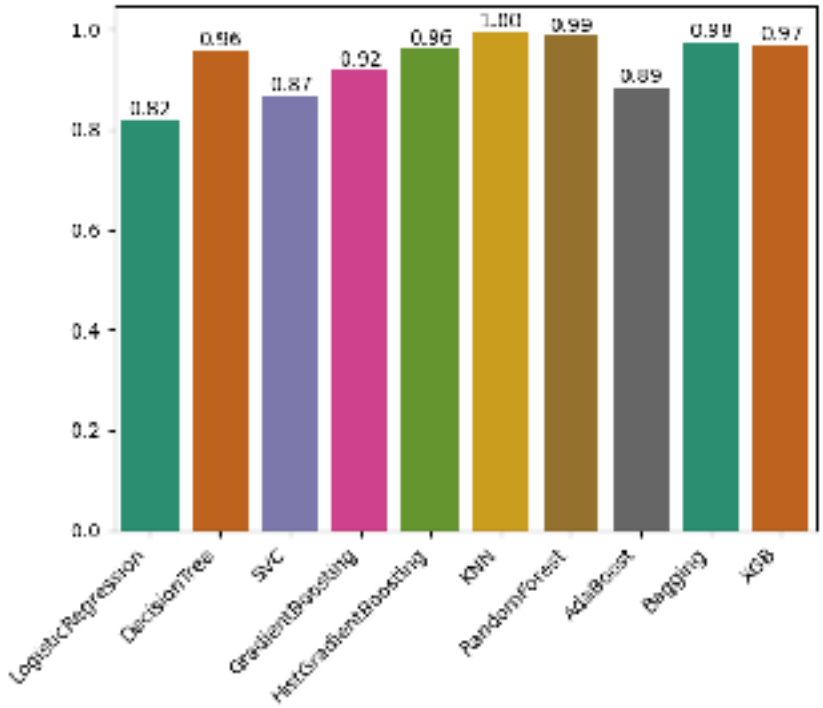

Fig 10.1 - Cross_val_score on train data results

Data Insights:

By the 'cross_val_score' observation, we can see that 'DecisionTree', 'HistGradientBoosting', 'KNN', 'RandomForest', 'Bagging' and 'XGB' produced best results on the training data.

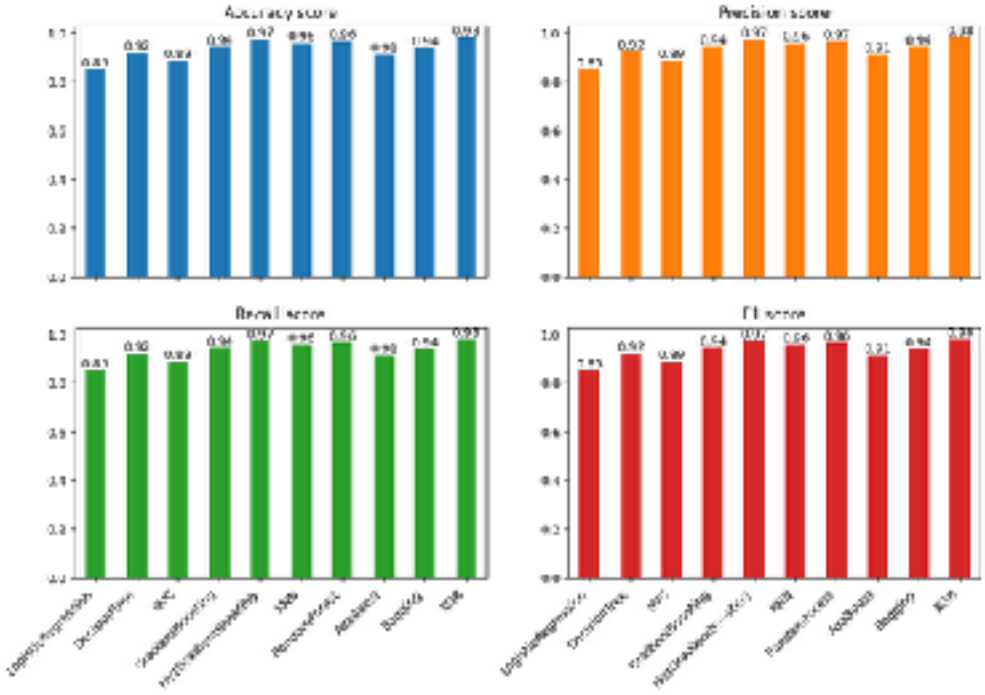

Fig 10.2 - Test data results

Data Insights:

Test Results are as above and 'XGB' produces best results on all the metrics (Accuracy, Precision, Recall and F1-score).

4.1 Training all the best models

1. DecisionTree

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	DecisionTree	1.0	1.0	1.0	1.0

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	DecisionTree	0.91891	0.924086	0.91891	0.919277

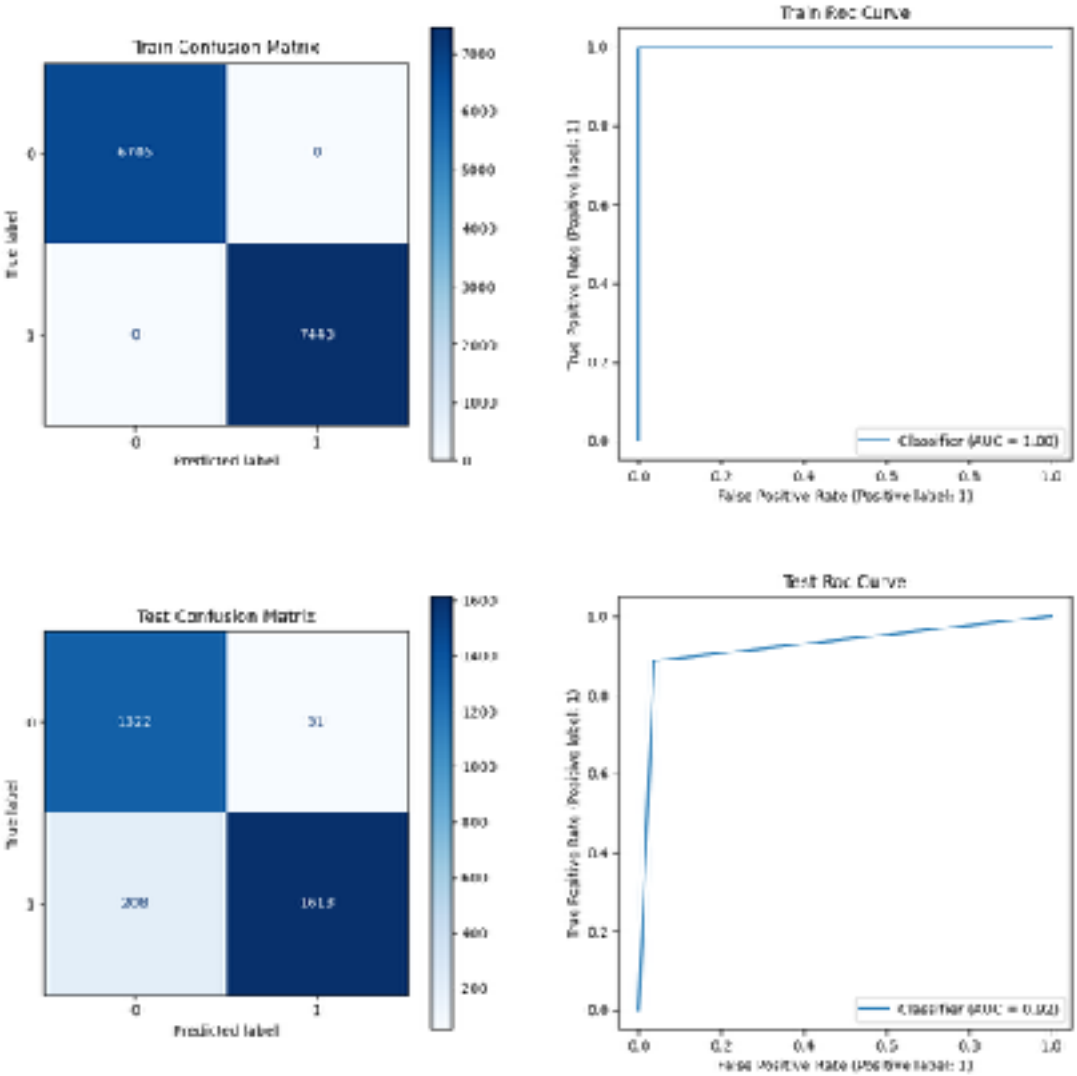

Fig 11.1 - DecisionTree results

2. GradientBoosting

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	GradientBoosting	0.933018	0.933192	0.933018	0.933042

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	GradientBoosting	0.943331	0.94414	0.943331	0.943353

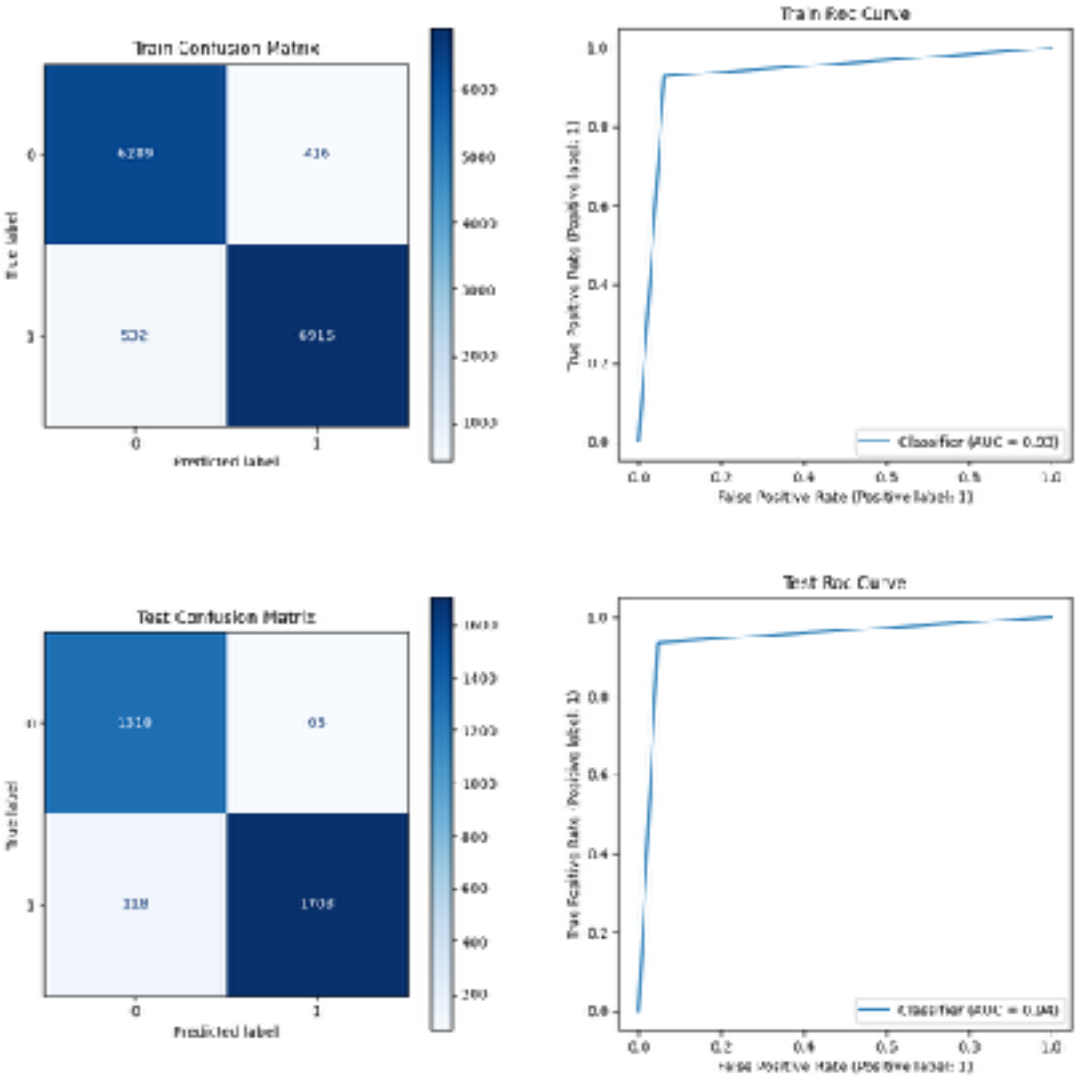

Fig 11.2 - GradientBoosting results

3. HistGradientBoosting

```
              Model Name  Accuracy score  Precision score  Recall score  \
0  HistGradientBoosting        0.989967         0.989967      0.989967

   F1 score
0  0.989967
              Model Name  Accuracy score  Precision score  Recall score  \
0  HistGradientBoosting        0.970883          0.97109      0.970883

   F1 score
0  0.970916
```

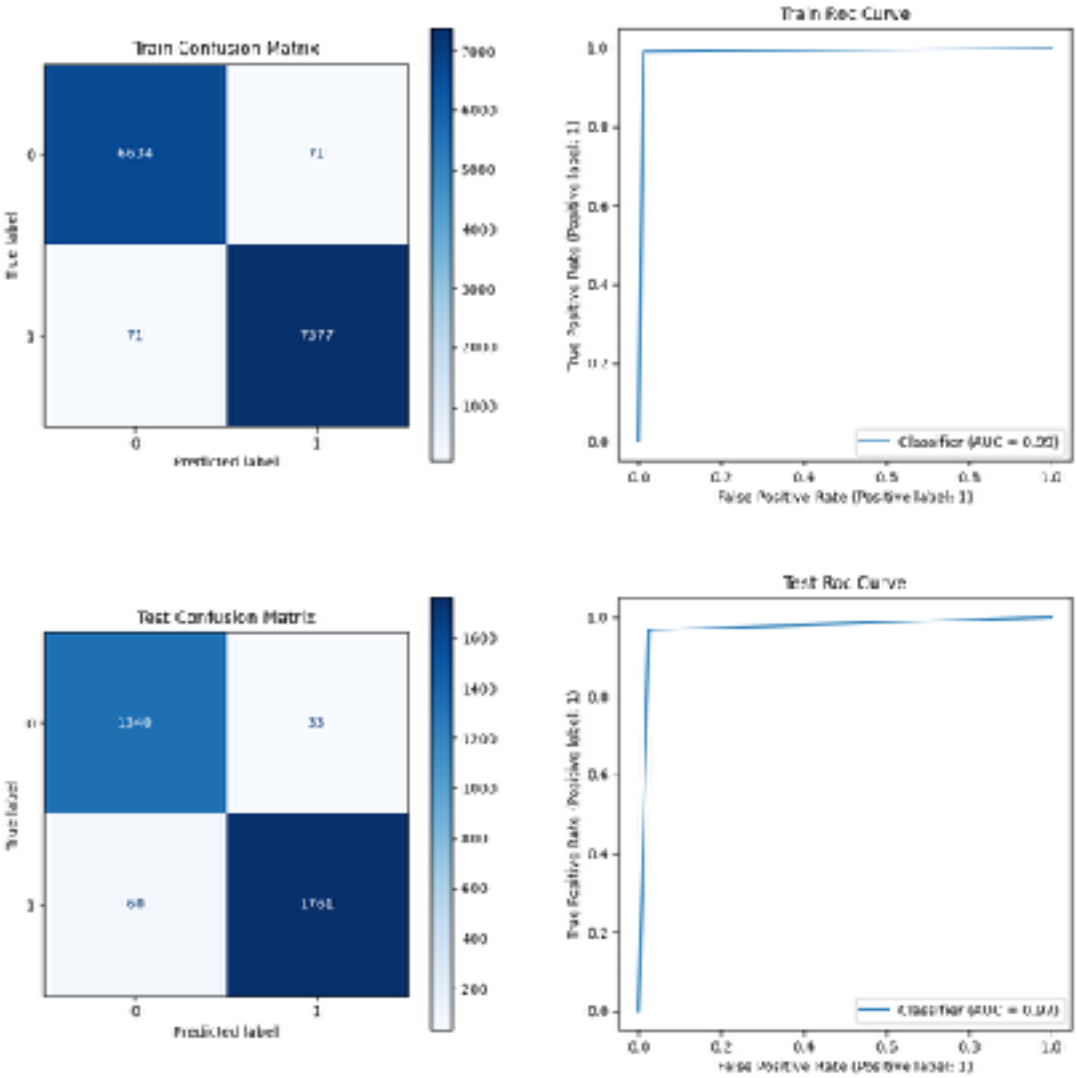

Fig 11.3 - HistGradientBoosting results

4. KNN

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	KNN	0.999929	0.999929	0.999929	0.999929

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	KNN	0.956481	0.957105	0.956481	0.956556

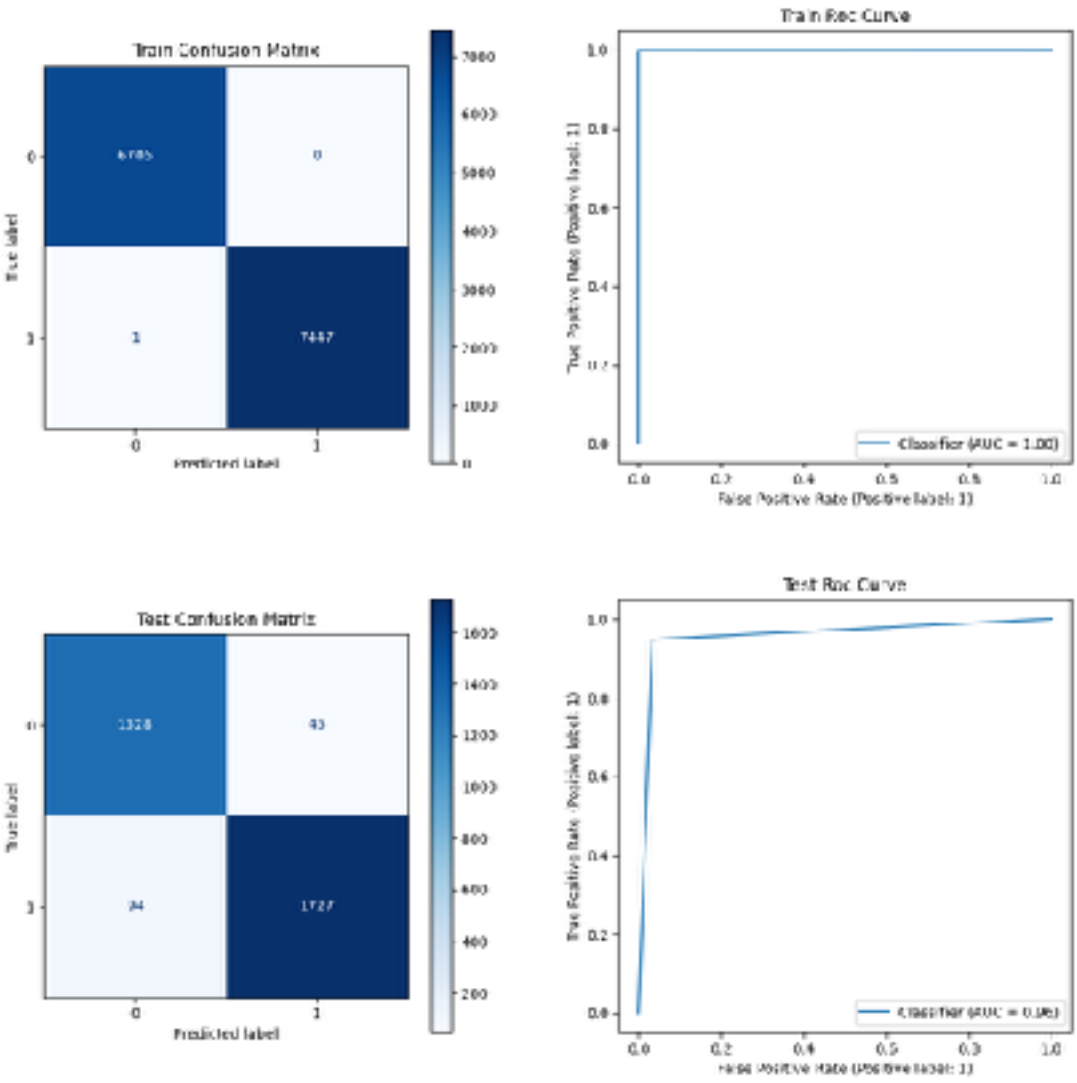

Fig 11.4 - KNN results

5. RandomForest

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	RandomForest	1.0	1.0	1.0	1.0

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	RandomForest	0.964621	0.965657	0.964621	0.964711

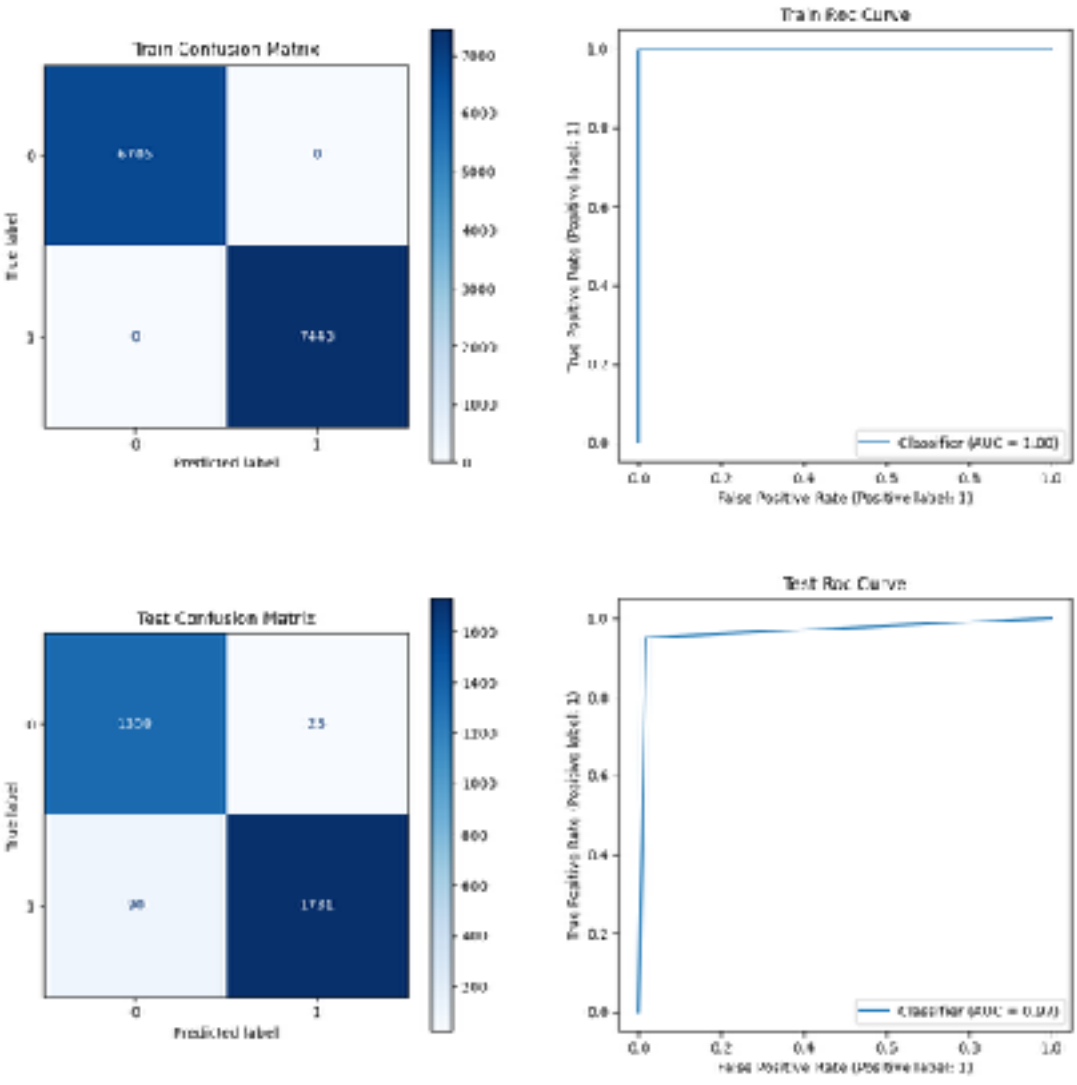

Fig 11.5 - RandomForest results

6. Bagging

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	Bagging	0.999505	0.999506	0.999505	0.999505

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	Bagging	0.939261	0.943722	0.939261	0.939526

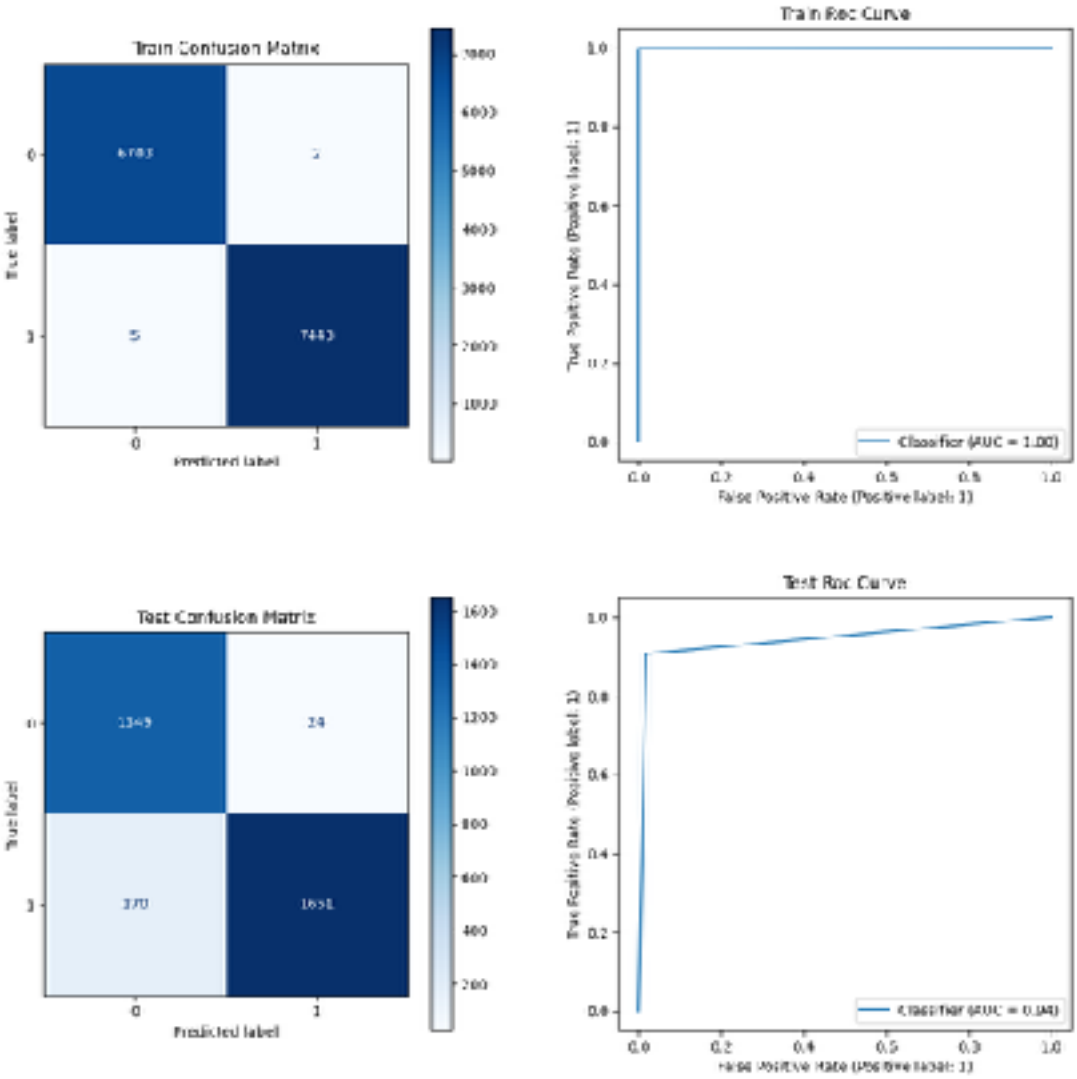

Fig 11.6 - Bagging results

7. XGB

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	XGB	0.999859	0.999859	0.999859	0.999859

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	XGB	0.980589	0.980719	0.980589	0.980607

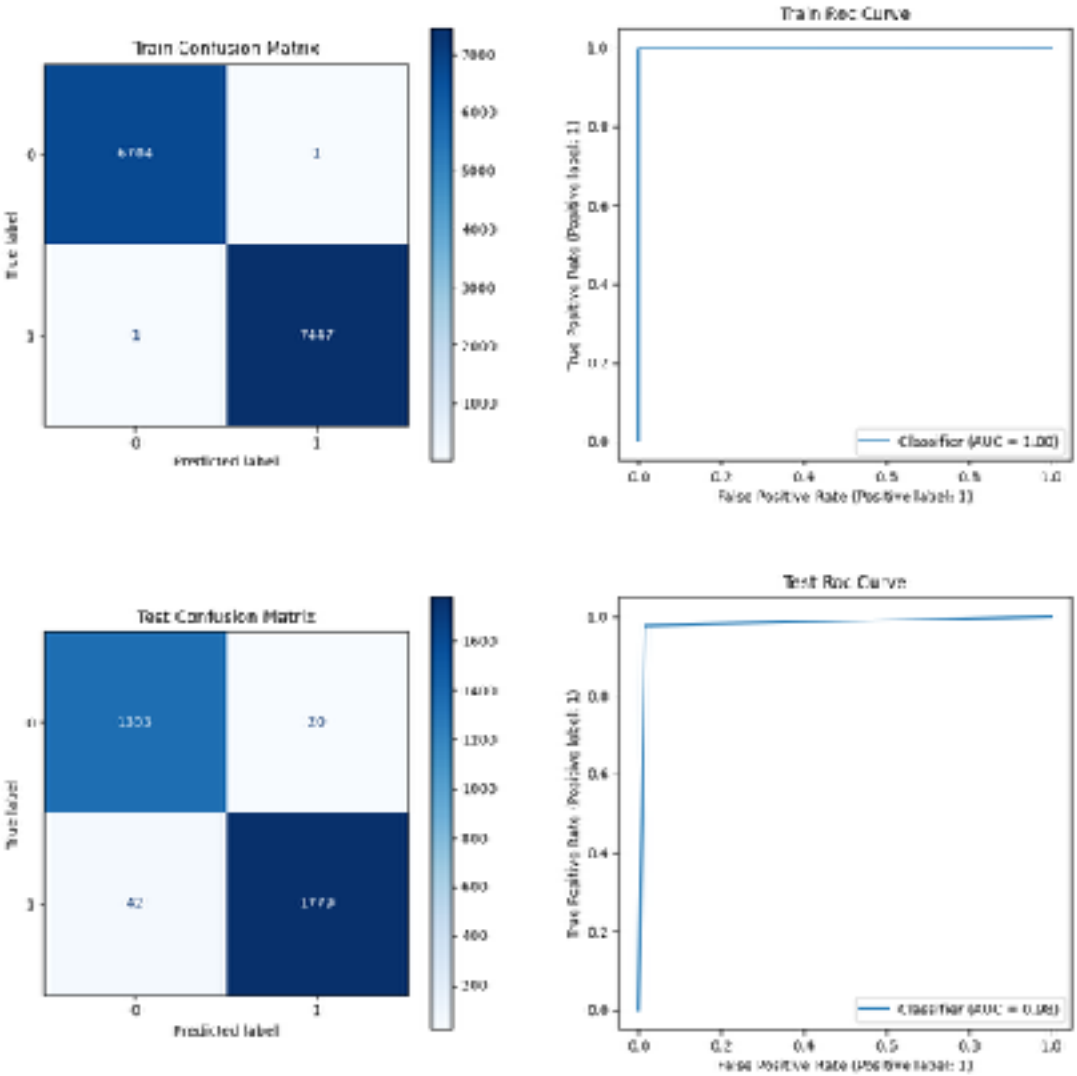

Fig 11.7 - XGB results

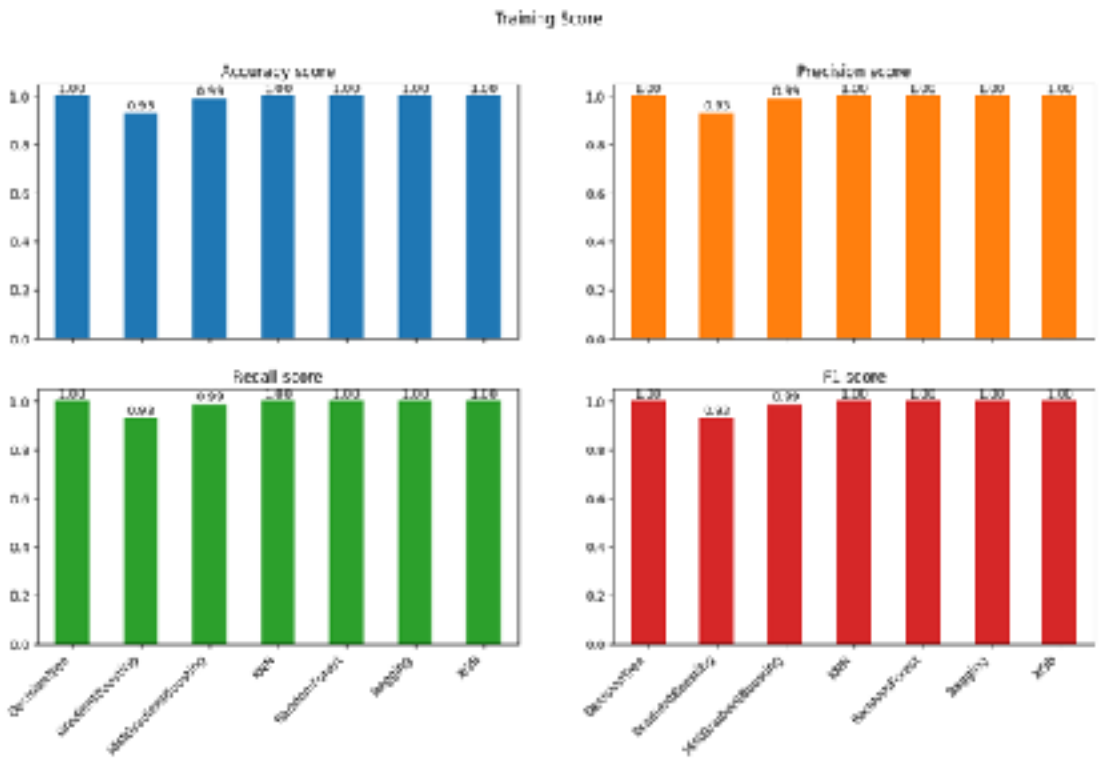

Fig 11.8 - All model train data results

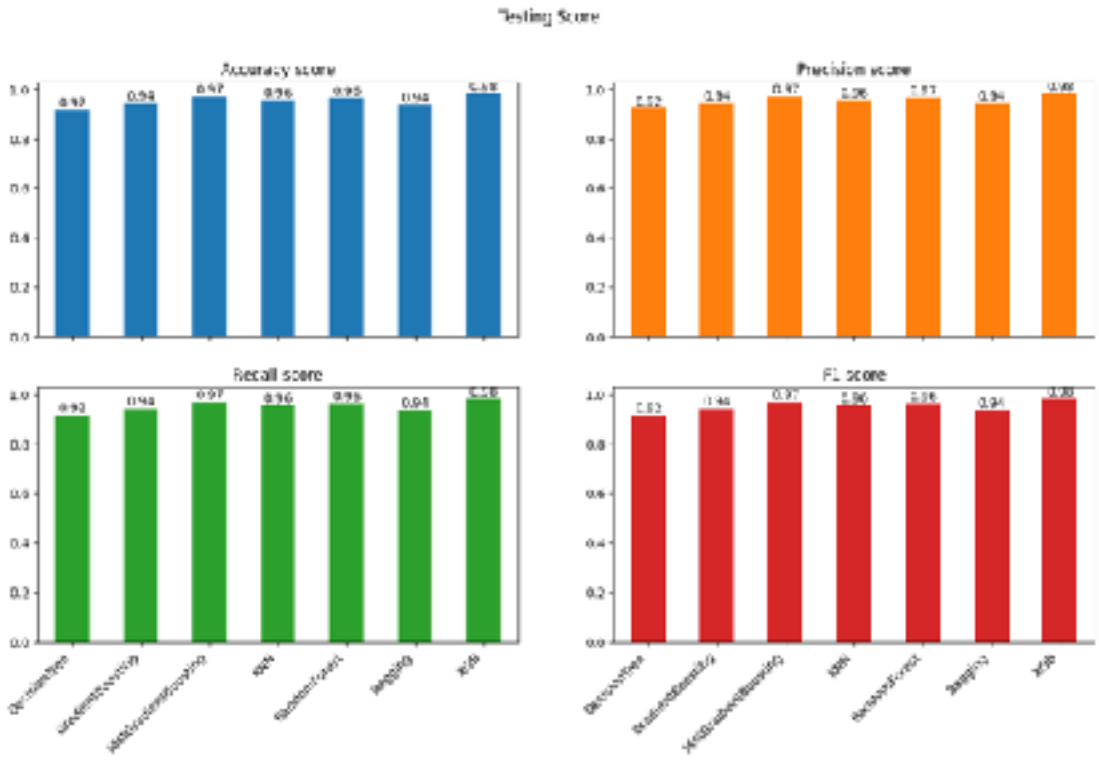

Fig 11.9 - All model test data results

CHAPTER 5

TUNING THE MODEL

5. Tuning the model

Data Insights:

Since, almost all the models produced best results on the training dataset, it's unnecessary to tune the model to get the best results except 'GradientBoosting'.

```
                Model Name  Accuracy score  Precision score  Recall score  \
0  GradientBoostingTuned        0.982265         0.982271      0.982265

   F1 score
0  0.982266
                Model Name  Accuracy score  Precision score  Recall score  \
0  GradientBoostingTuned        0.960551          0.96106      0.960551

   F1 score
0  0.960621
```

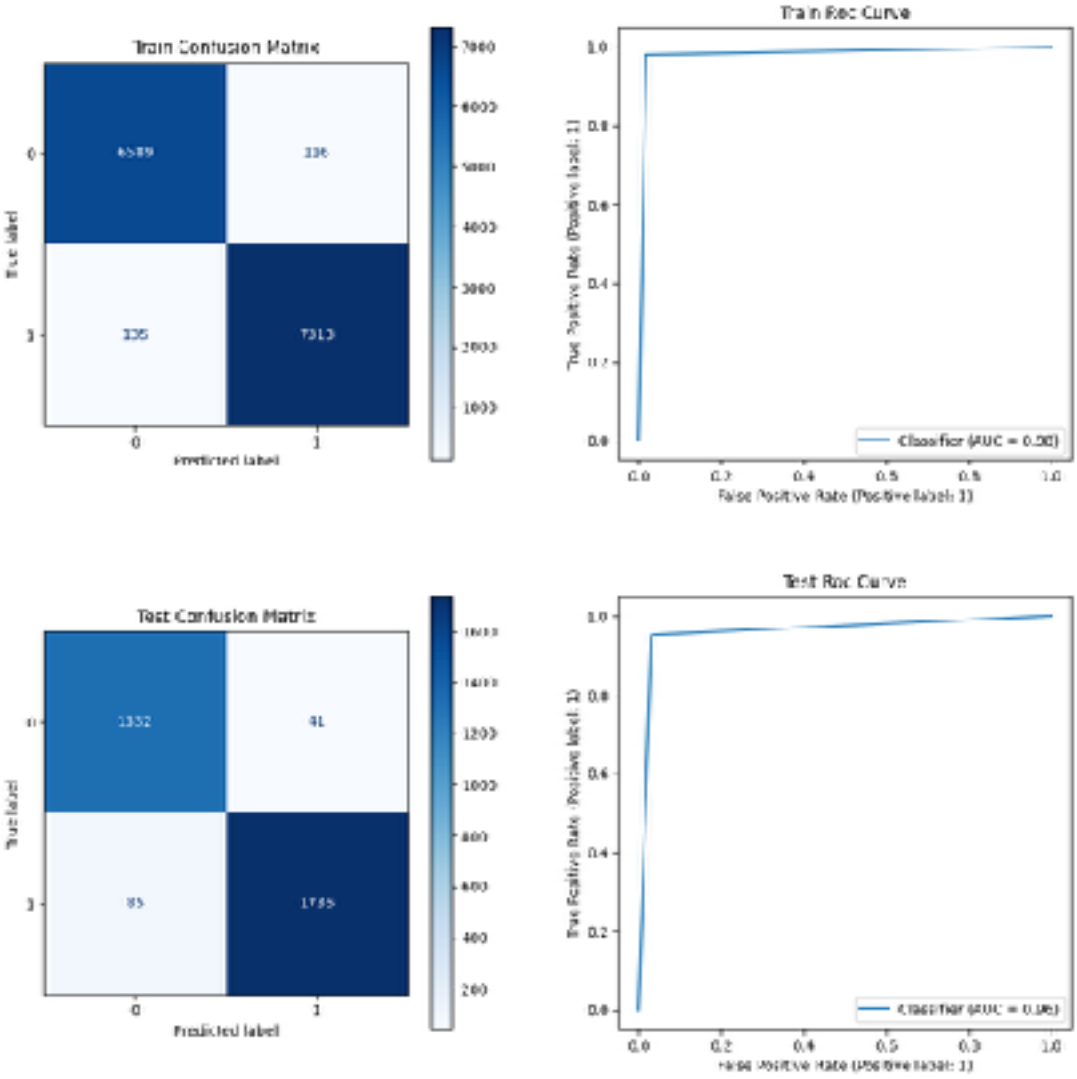

Fig 11.10 - GradientBoosting tuned results

CHAPTER 6

FINAL RESULTS, CONCLUSION AND IMPROVISATION

6.1 Final Results

	Model Name	Accuracy score	Precision score	Recall score	F1 score
6	XGB	0.980589	0.980719	0.980589	0.980607
2	HistGradientBoosting	0.970883	0.971090	0.970883	0.970916
4	RandomForest	0.964821	0.965657	0.964821	0.964711
3	KNN	0.956481	0.957105	0.956481	0.956586
1	GradientBoosting	0.943331	0.944140	0.943331	0.943453
5	Bagging	0.939261	0.943722	0.939261	0.939526
0	DecisionTree	0.918910	0.924086	0.918910	0.919277

Table 8 - Final results data frame

Data Insights:

It's found the best model so far is XGB. We got 100% results on training dataset on all the metrics (Accuracy, Precision, Recall, F1-score) and around 98% on testing dataset on all the metrics. The above DataFrame is sorted w.r.t 'F1-score'.

6.2 Conclusion

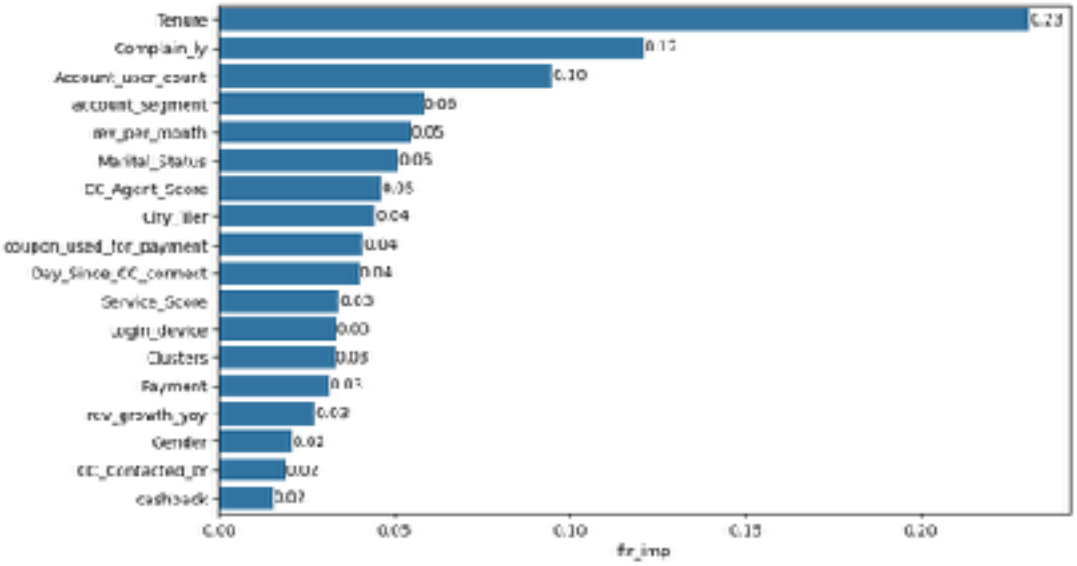

Fig 12 - Feature Importance

Data Insights:

It is found that the top most important features of the dataset which is contributing most to the model are 'Tenure', 'Complain_ly' and 'Account_user_count'.

6.3 Improvisation

Data Insights:

We can improve our best model 'XGB' by altering the threshold of the ROC curve and we can increase the 'Recall Score'. Since we are dealing with Churn, recall score played a vital role. We don't want our model to predict wrong when the customers are planning/decided to churn by their data behaviour that we have collected.

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	XGB	0.999859	0.999859	0.999859	0.999859

	Model Name	Accuracy score	Precision score	Recall score	F1 score
0	XGB	0.980589	0.980719	0.980589	0.980607

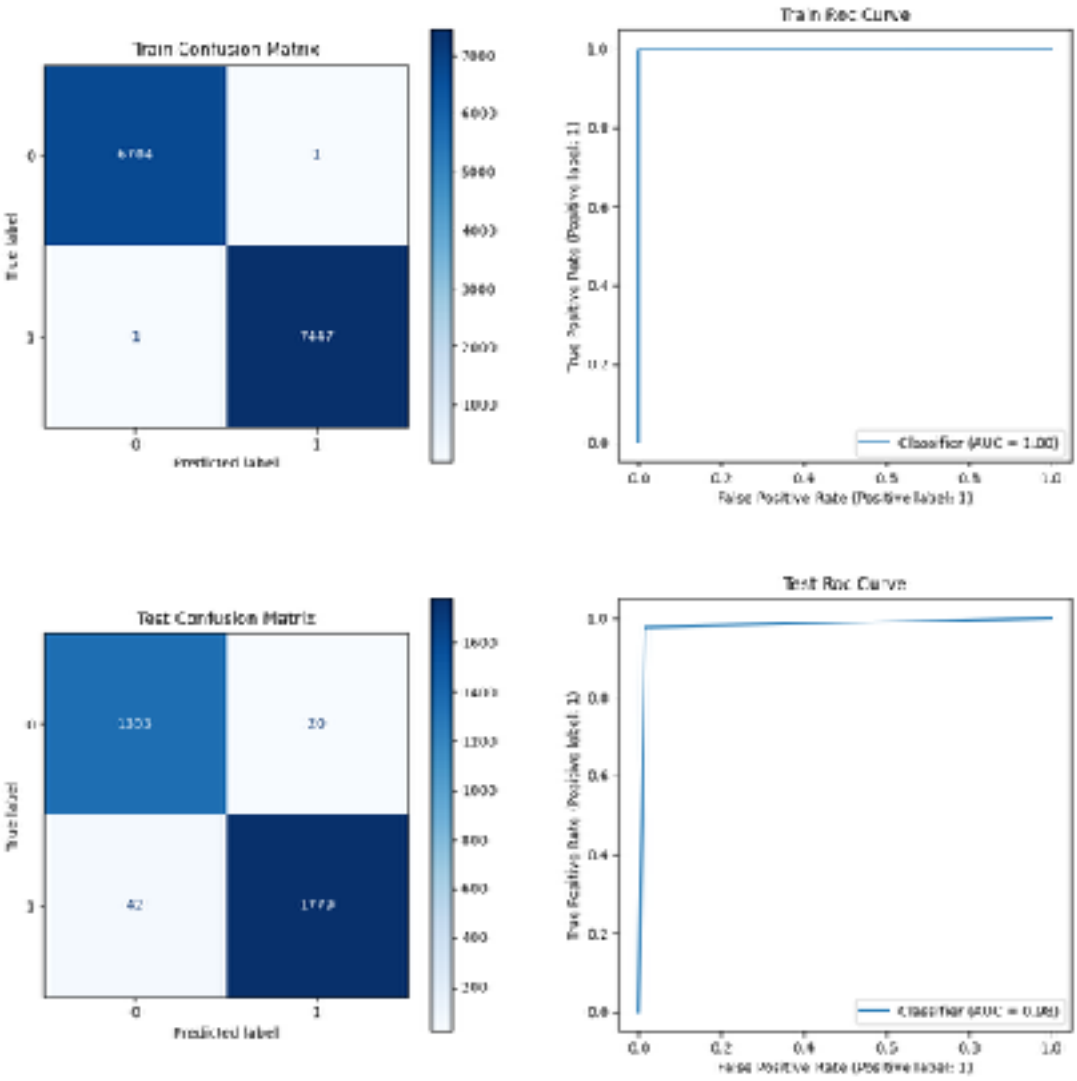

Fig 13 - XGB results improvisation

THANK YOU