EXECUTIVE SUMMARY

Introduction

This executive summary provides an overview of a customer churn project of DTH Service Provider. The project aims to analyze customer churn, which refers to the loss of customers, in the context of a business operating in various Indian tier cities. By understanding the factors contributing to churn and implementing targeted strategies, the goal is to reduce churn rates and improve customer retention.

Data Collection and Preprocessing

Data pertaining to customer behavior, demographics and support interactions is provided including transactional records & customer surveys. The data is preprocessed to ensure its quality and reliability and relevant features are extracted to capture important aspects of customer behavior.

Exploratory Data Analysis (EDA)

An in-depth analysis of the data is conducted to gain insights into customer churn patterns. Key metrics such as churn rates, customer demographics and customer engagement levels are analyzed to identify trends and correlations. This analysis forms the basis for further investigation and model development.

Feature Inclusion

Based on the exploratory analysis, specific features are added to capture the unique characteristics. These features may include preferences for regional products, customer preferences, etc. By incorporating these features, the models can capture localized causes leading to customer churn.

Model Development and Evaluation

Various machine learning models, such as logistic regression, decision trees, random forests are trained and evaluated using appropriate performance metrics. The models are fine-tuned using cross-validation techniques, with a focus on optimizing metrics such as accuracy, sensitivity, specificity and the F1 score.

Intervention Strategies

Based on the insights gained from the models, targeted intervention strategies are developed to reduce churn rates. These strategies leverage trends and customer preferences to enhance customer experience and loyalty. For example, personalized offers during festive seasons, regional product recommendations or customer support in local languages can be implemented.

Validation and Deployment

The final model is validated on a separate dataset to ensure its performance and generalizability. Once validated, the model and the intervention strategies are deployed into the business operations.

Expected Benefits

Implementing a customer churn analysis project provides several benefits:

1. **Improved Customer Retention**

By understanding the specific factors influencing customer churn in the market, targeted strategies can be implemented to reduce churn rates and improve customer retention.

2. **Enhanced Customer Experience**

Incorporating cultural elements and preferences into intervention strategies can lead to a more personalized and engaging customer experience, increased loyalty and satisfaction.

3. **Increased Revenue**

By retaining more customers, the business can enjoy a higher customer lifetime value, increased revenue and a stronger market position.

Conclusion

Implementing a customer churn project provides an opportunity to understand the distinct characteristics of the business and develop targeted strategies to reduce churn rates. By incorporating trends and preferences into intervention strategies, the project aims to improve customer retention, enhance customer experience, increase revenue and gain a competitive advantage.

TABLE OF CONTENTS

List of Tables.. 4
List of Graphs.. 4
INTRODUCTION AND BACKGROUND..5
1.1) Executive Summary... 5
1.2) Introduction & Background.. 6
1.3) Problem Statement..7
1.4) Objective of Study... 8
1.5) Company and Industry Overview... 9
1.6) Overview of Theoretical Concepts...10
RESEARCH METHODOLOGY...12
2.1) Scope of Study... 12
2.2) Methodology.. 12
2.2.1) Research Design.. 12
2.2.2) Data Collection..13
2.2.3) Sampling Method...13
2.2.4) Data Cleaning.. 13
2.2.4.a) Removal of Unwanted Variables.. 13
2.2.4.b) Missing Value Treatment..14
2.2.4.c) Outlier Treatment...16
2.2.4.d) Variable Transformation...18
2.2.5) Data Analysis Tools...19
2.2.5.1) Univariate Analysis..19
2.2.5.2) Bivariate Analysis.. 24
2.2.5.3) Unbalanced Data and Treatment... 27
2.2.5.4) Business Insights Using Clustering.. 28
2.2.5.5) Any Other Business Insights..30
2.3) Period of Study... 30
DATA ANALYSIS AND INTERPRETATION.. 31
3.a) Building Various Models... 31
3.b) Testing the Model.. 33
3.c) Model Interpretation... 35
3.d) Ensemble Modeling..36
3.e) Model Tuning Measures.. 37
3.f) Interpretation of Optimum Model... 38
FINDINGS, RECOMMENDATIONS AND CONCLUSION...39

List of Tables

Table No.	Table Title	Page No.
1	Treatment of NaN Values for Columns In Report	15
2	Identification of key columns	19
3	Model Comparison Table	38

List of Graphs

Graph No.	Graph Title	Page No.
1	Before Outlier Treatment	17
2	After Outlier Treatment	17
3	Tenure in Months	20
4	Revenue Growth Year on Year	21
5	Payment Mode Representation	22
6	Account Segment Representation	23
7	Account Segment - Churn Distribution	24
8	Account Segment - Revenue Distribution	25
9	Heatmap for Variable Correlation	26
10	Proportion of Customer Churned Vs Retained	27
11	The Elbow Method Graph	28
12	Clusters of Account	29
13	Count Plot Using Churn Data	30
14	Confusion Matrix - Random Forest	34
15	Important Features	36

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1) Executive Summary

In an era where customer loyalty drives business sustainability, our Customer Churn Prediction Project takes center stage. This initiative harnesses the power of data analytics to proactively identify customers at risk of churning and implement counter measures to retain their loyalty.

By leveraging historical customer data, advanced machine learning algorithms and predictive modeling techniques, we aim to anticipate and address customer attrition. This project aligns with our commitment to enhancing customer relationships and sustaining growth through data-driven insights and actionable strategies.

1.2) Introduction & Background

In today's highly competitive market, DTH (Direct-to-Home) service providers are facing the problem of retaining their existing customers. To address this issue, a churn prediction model is essential for identifying potential churners and offering targeted campaigns to ensure they are retained. Also, it is crucial to develop a campaign strategy that is effective and able to achieve the targets with minimum losses for the company's profit margin.

With intense competition in the market, customer churn poses a significant risk for DTH providers. Losing an account means losing a pool of active customers, causing an impact which is exponential. Developing a churn prediction model coupled with a well designed campaign can help identify potential churners and incentivize them for continued association with the provider.

Problem Motivation

1. **Intense Market Competition**

Companies are facing strong competition from attractive offerings of other service providers, resulting in customer retention being vital for sustaining market share and ensuring profitability. Retaining existing customers is a more cost-effective solution than acquiring new ones, highlighting the need to predict churn and implement effective strategies to retain valuable accounts in a proactive approach.

2. **Financial Implications**

Account churn can lead to the loss of multiple customers tied to a single account, increasing the financial impact on the company. This resulted in the need for developing a churn prediction model and offering targeted campaigns to minimize customer attrition, maintain revenue and enhance overall business performance.

1.3) Problem Statement

A DTH provider is facing intense market competition, making customer retention a significant challenge. To address this problem, the company aims to develop a churn prediction model which would identify potential churners among their accounts and offer attractive incentives to mitigate customer churn. Account churn is particularly efficient as a single account can represent multiple active customers, magnifying the loss incurred by the company with each churned account.

A churn problem, also known as customer churn/attrition, refers to the situation where customers or subscribers discontinue their relationship with a company or stop using its products/services. Churn is a common challenge faced by businesses across various industries, including E-commerce, telecommunications, subscription-based services where more than few service providers operate.

The churn problem arises when customers switch to competitors, cancel their subscriptions or stop making purchases, leading to a decline in existing revenue and potential loss of market share. It is crucial for companies to address the churn problem effectively to retain valuable customers, minimize revenue losses and maintain a competitive edge in the market.

To tackle the churn problem, companies often employ churn prediction models that use data analysis and predictive analytics techniques to identify customers who are at the edge of being churned. By identifying potential churners in advance, companies can take proactive measures to retain these customers through targeted retention strategies, personalized offers, improved customer experiences and effective communication best matching the needs of their customers.

Solving the churn problem requires understanding the underlying reasons why customers churn and implementing strategies to mitigate the key factors. It involves analyzing customer behavior, preferences, satisfaction levels, designing effective retention campaigns and loyalty programs to incentivize customers to stay loyal to the company's products or services.

1.4) Objective of Study

The need for studying customer churn in the DTH Industry is crucial for DTH providers for the following reasons:

1. Retaining Subscribers

Customer churn in the DTH industry can result in the loss of subscribers, leading to decreased revenue and market share. By studying customer churn, DTH providers can gain insights into the reasons behind churn and identify specific subscriber segments that are more likely to churn. This knowledge enables them to develop targeted retention strategies, personalized offers, and improved services to reduce churn and retain valuable subscribers.

2. Improving Customer Satisfaction

Understanding customer churn allows DTH providers to identify areas where customer dissatisfaction may be leading to churn. By studying the reasons behind churn, such as service quality issues, pricing concerns, or lack of content relevance, providers can make necessary improvements to enhance the overall customer experience.

3. Optimizing Business Operations

Studying customer churn data helps DTH providers optimize their business operations. By analyzing churn patterns and identifying the factors that contribute to churn, providers can uncover operational inefficiencies, such as installation delays, billing errors, or service disruptions, that may impact customer satisfaction and retention.

4. Building Competitive Advantage

In a competitive DTH market, understanding customer churn provides a competitive advantage. By studying churn rates, analyzing competitor offerings, and identifying areas where competitors may have an edge, DTH providers can differentiate their services and offerings. This includes introducing innovative features, improving customer engagement, offering attractive pricing plans, and providing value-added services to attract and retain subscribers.

Overall, studying customer churn in the DTH industry helps DTH providers retain subscribers, improve customer satisfaction, optimize operations, enhance content strategy, and gain a competitive edge. It enables providers to identify and address the factors that contribute to churn, resulting in improved subscriber retention rates, increased revenue, and long-term business success

1.5) Company and Industry Overview

After successfully solving the customer churn problem for a DTH (Direct-to-Home) company, several business and social opportunities can emerge:

1. **Increased Subscriber Retention**

By accurately predicting churn and implementing targeted retention strategies, the DTH company can boost customer loyalty, reduce acquisition costs, and drive revenue growth.

2. **Enhanced Customer Satisfaction**

Understanding churn reasons enables the company to address service quality, pricing, and content concerns, leading to improved customer experiences and positive brand perception.

3. **Competitive Advantage**

Effective churn management differentiates the DTH company from competitors, positioning it as a customer-centric provider that values loyalty and attracts new subscribers.

4. **Data-Driven Decision Making**

Analyzing customer data for churn prediction empowers the company to make informed decisions on content, packages, and engagement strategies, driving operational efficiency and business growth.

5. **Improved Content Strategy**

Churn insights allow the company to curate personalized content offerings, secure exclusive partnerships, and provide relevant and engaging programming to increase subscriber satisfaction and retention.

6. **Social Impact**

Successful churn reduction sustains employment, supports content creators, and promotes digital connectivity, fostering social and economic development within communities served by the DTH company.

1.6) Overview of Theoretical Concepts

In the scope of predictive analytics, understanding the theoretical concepts is essential for effectively addressing complex challenges like customer churn prediction. Our Project is built upon a foundation of key theoretical concepts that empower us to tackle this classification problem with precision.

1. **Classification Algorithms**

Central to our project are classification algorithms such as logistic regression, decision trees and random forests. These algorithms analyze the patterns within customer data, enabling us to classify customers as either potential churners or loyal ones.

2. **Feature Selection**

Feature selection involves identifying the most relevant attributes that contribute to customer churn. This process enhances model accuracy by focusing on the most influential factors.

3. **Feature Engineering**

This involves transforming and combining attributes to reveal hidden insights, thereby improving the model's predictive capabilities.

4. **Cross-Validation**

Ensuring the model's robustness is achieved through cross-validation techniques. This involves partitioning data into training and validation sets, assessing the model's performance on unseen data and preventing overfitting efficiently.

5. Evaluation Metrics

Precision, recall, accuracy, F1-score and the ROC curve are vital evaluation metrics. These metrics allow us to measure the model's effectiveness, guiding us in optimizing its performance.

6. Imbalanced Data Handling

Given the nature of churn prediction, where positive churn instances are often a minority, techniques like oversampling and undersampling help balance the data for unbiased model training.

7. Confusion Matrix

The confusion matrix aids in visualizing the model's performance by providing insights into true positives, true negatives, false positives and false negatives.

8. Hyperparameter Tuning

Fine-tuning hyperparameters is crucial for optimizing model performance. It involves experimenting with different parameter values to achieve the best results.

9. Interpretability

Interpretability techniques, such as feature importance and decision tree visualization, help us understand the factors driving model predictions.

These theoretical concepts serve as the milestones of our Customer Churn Prediction Project. By inculcating these concepts with real-world data, we aspire to create accurate and robust churn prediction models that empower businesses to retain customers and drive sustainable growth.

CHAPTER 2
RESEARCH METHODOLOGY

2.1) Scope of Study

Studying customer churn within a DTH (Direct-to-Home) company offers important facts into consumer behavior and industry dynamics. The scope includes investigating patterns behind subscriber attrition, identifying factors influencing churn and formulating strategies to enhance customer retention. By analyzing usage patterns, service quality, pricing models and competitive landscape, this study aims to provide actionable insights for reducing churn rates. The findings hold the potential to transform how DTH companies engage with their customers, ensure loyalty and sustain the business growth.

2.2) Methodology

The research methodology for studying customer churn within a DTH (Direct-to-Home) company involves a comprehensive approach. It includes data collection from various sources, such as subscriber demographics, usage patterns, customer interactions, subscription history and related technical details. Employing quantitative analysis techniques, the study explores correlations between churn rates and factors like service quality, pricing and customer engagement. Machine learning models, including logistic regression and decision trees are implemented to predict potential churn. This research methodology aims to uncover actionable insights that guide effective retention strategies and contribute to a deeper understanding of customer behavior within the DTH industry.

2.2.1) Research Design

The research design for examining customer churn within a DTH (Direct-to-Home) company is carefully structured to yield meaningful insights. It encompasses a mix of exploratory and analytical phases, including data collection, processing and analysis. The design involves defining variables, selecting appropriate data sources and implementing statistical tools to examine correlations and trends. By employing both quantitative and qualitative methods, the research design aims to provide a comprehensive understanding of factors influencing churn.

2.2.2) Data Collection

Data has been gathered provided by the DTH company. There is no clarification received on the period at which the data is gathered, tools/mode using which the data is gathered and correlation of provided records with actual no.of customers by the organization at the time of research.

2.2.3) Sampling Method

We have used random sampling and clustering techniques to gather insights over the data provided. Data provided in a structured format which comprises 11261 records and 19 features associated with each record.

2.2.4) Data Cleaning

Data cleaning would involve identifying and removing irrelevant or redundant features that do not contribute much to the analysis of the model. By following these steps, the dataset is transformed into a clean, consistent, and properly formatted representation. This would significantly contribute to the quality and effectiveness of the models and improve their ability to make accurate predictions.

2.2.4.a) Removal of Unwanted Variables

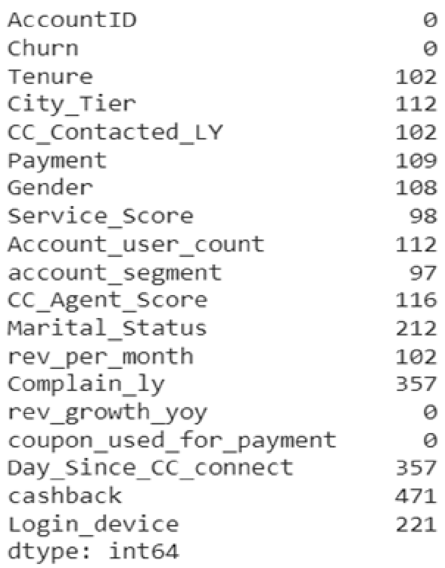
We have skipped the “AccountID” variable/feature from the dataset after verification of duplicate records present in the dataset provided.

2.2.4.b) Missing Value Treatment

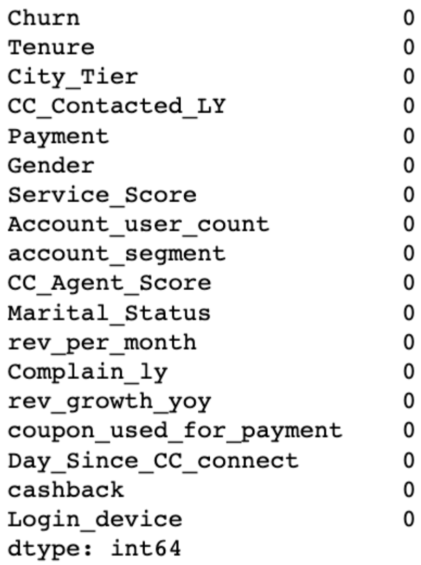
In the data cleaning phase, we check for missing values (NaN / Not A Number), special characters, leading and trailing white spaces. The various methods used are deletion, mean/median/mode imputation, regression imputation, multiple imputation using algorithms that handle missing values.

- **Deletion**: Since the dataset contains less than 15% of missing values, we need to impute them instead of deleting them.
- **Space/Special character replacement**: Presence of leading/trailing white spaces and special characters (#, $, *, + & etc.) were eliminated by programmatic string replacement.
- **Mean/median/mode imputation**: Missing values are replaced with the mean, median or mode of the available values for that variable. Mode is used for categorical column/variable while Mean/Median is conditionally used for numerical column/variable.
- **Duplicate categorical value replacement**: We found duplicate categorical values under "gender" and "account_segment" variables and carried out value replacement to match the unique categorical values.

Evidence of missing values and result after treatment is showcased below:



```
AccountID                  0
Churn                      0
Tenure                   102
City_Tier                112
CC_Contacted_LY          102
Payment                  109
Gender                   108
Service_Score             98
Account_user_count       112
account_segment           97
CC_Agent_Score           116
Marital_Status           212
rev_per_month            102
Complain_ly              357
rev_growth_yoy             0
coupon_used_for_payment    0
Day_Since_CC_connect     357
cashback                 471
Login_device             221
dtype: int64
```



```
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
```

Comprehensive outlook of all the treatments against each column/variable listed in below report:

It is important to note that data-type transformation is carried out to eliminate programmatic limitation on processing few columns expected to contain only numeric values.

Treatment of NaN values:

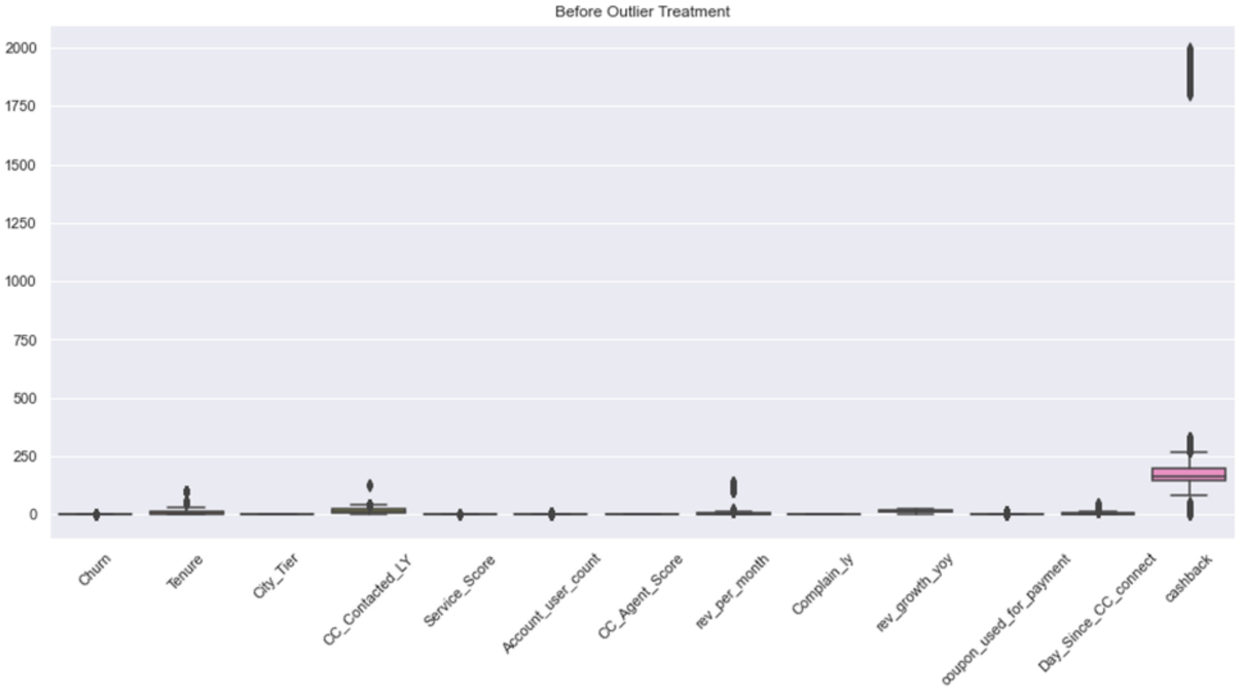
Column/Variable	Space Trim	Special Character	Null Value Replacement	Duplicate Value Replacement	Data-type Transformation
AccountID	Yes		NA		
Churn	Yes		NA		
Tenure	Yes	Yes	Mode		Yes
City_Tier	Yes		Mean		
CC_Contacted_LY	Yes		Mean		
Payment	Yes		Mode		
Gender	Yes		Mode	Yes	
Service_Score	Yes		Mean		
Account_user_count	Yes	Yes	Mode		Yes
account_segment	Yes		Mode	Yes	
CC_Agent_Score	Yes		Mean		
Marital_Status	Yes		Mode		
rev_per_month	Yes	Yes	Mode		Yes
Complain_ly	Yes		Mean		
rev_growth_yoy	Yes	Yes	Mode		Yes
coupon_used_for_payment	Yes	Yes	Mode		Yes
Day_Since_CC_connect	Yes	Yes	Mode		Yes
cashback	Yes	Yes	Mode		Yes
Login_device	Yes	Yes	Mode		

2.2.4.c) Outlier Treatment

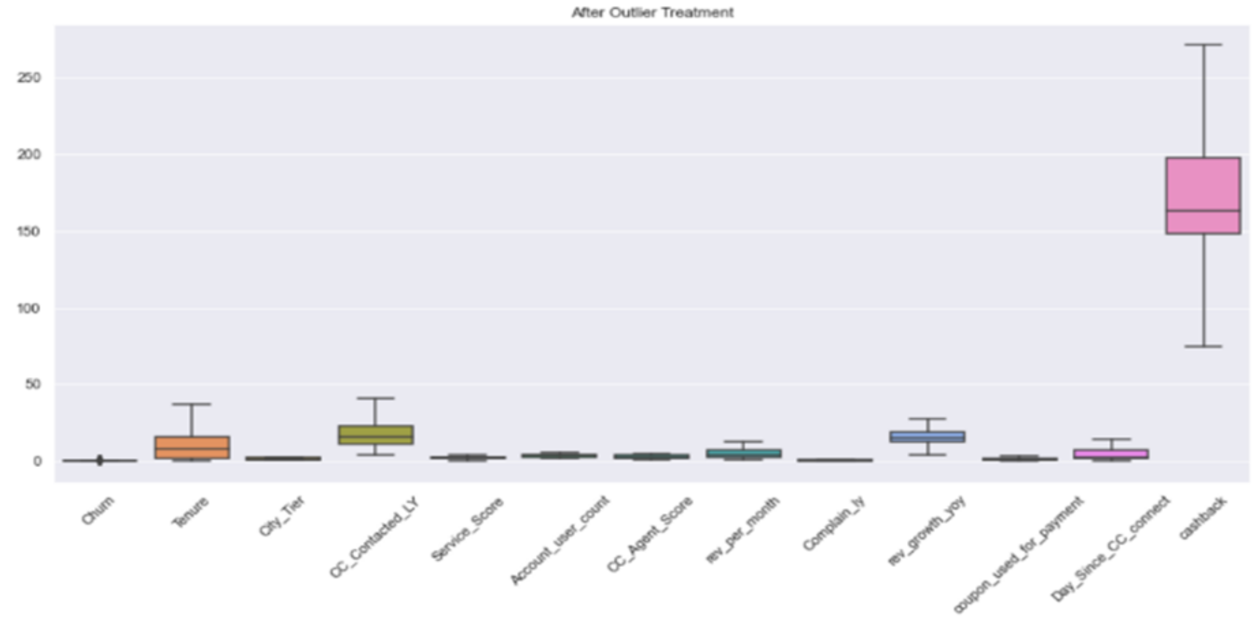
Detecting and treating the outliers gain importance as they can negatively affect the statistical analysis and the training process of a machine learning algorithm resulting in lower accuracy.

- We have used boxplot to detect the outliers and treated them using the Quantile based flooring and capping method.
- We used masking technique where outlier is capped at a certain value above the Upper bound (Q3 + 1.5*IQR) and floored at a factor below the Lower bound (Q1 – 1.5*IQR).

The visualization of boxplot before the outlier treatment:



The visualization of boxplot after the outlier treatment where values are now listed in a range bound fashion. This is evident that outliers which are significantly out of range were eliminated successfully.



2.2.4.d) Variable Transformation

The goal is to transform the variables in a way that makes them more suitable for the model to learn from and make accurate predictions. This can involve several techniques, like scaling, encoding categorical variables, feature engineering, binning, logarithmic or exponential transformations & dimensionality reduction.

1. **Scaling**: Scaling involves transforming the variables to a specific range or distribution. Standard scaler method has helped us to bring our data to a specific range.
2. **Encoding categorical variables**: Categorical variables often need to be transformed into numerical representations for machine learning models. Common encoding techniques label encoding for column/variables named Gender, Payment, Account Segment, Marital Status and Login Device.

During our analysis we carefully performed this step after the train/test data split in order to prevent any form of data leakage (train-test contamination) problems.

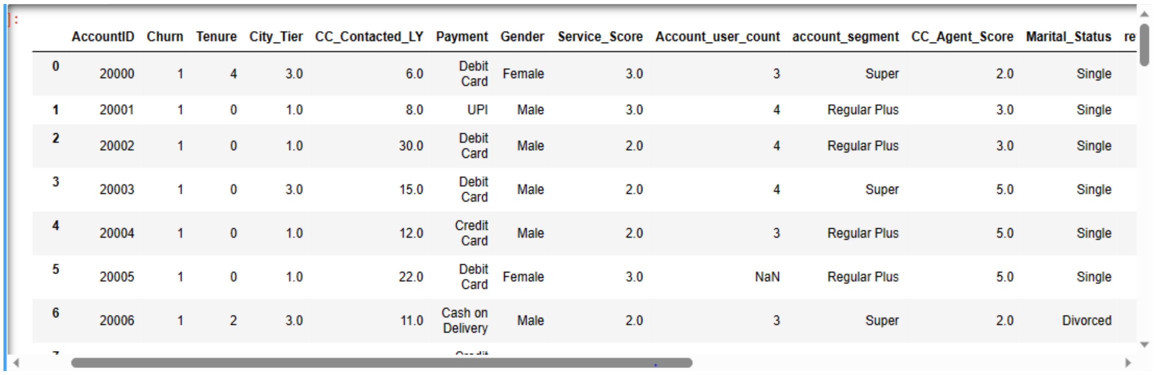
2.2.5) Data Analysis Tools

This section describes the various analysis tools and techniques used to derive meaningful insights out of the dataset.

2.2.5.1) Univariate Analysis

We aim to showcase the structure of our data and categorize the columns/variables upon which we need to perform our analysis. Initially we import the key libraries for data wrangling numpy and pandas, read the csv file and obtain the variables present in the tabular data.

Visual form of input data, resulting in a snapshot as follows:



	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	re
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0	Single	
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0	Single	
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0	Single	
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0	Single	
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0	Single	
5	20005	1	0	1.0	22.0	Debit Card	Female	3.0	NaN	Regular Plus	5.0	Single	
6	20006	1	2	3.0	11.0	Cash on Delivery	Male	2.0	3	Super	2.0	Divorced	

The next step which we need to do is to identify the key columns which would have a bearing on our output variable "Churn" and also categorize the columns into two types:

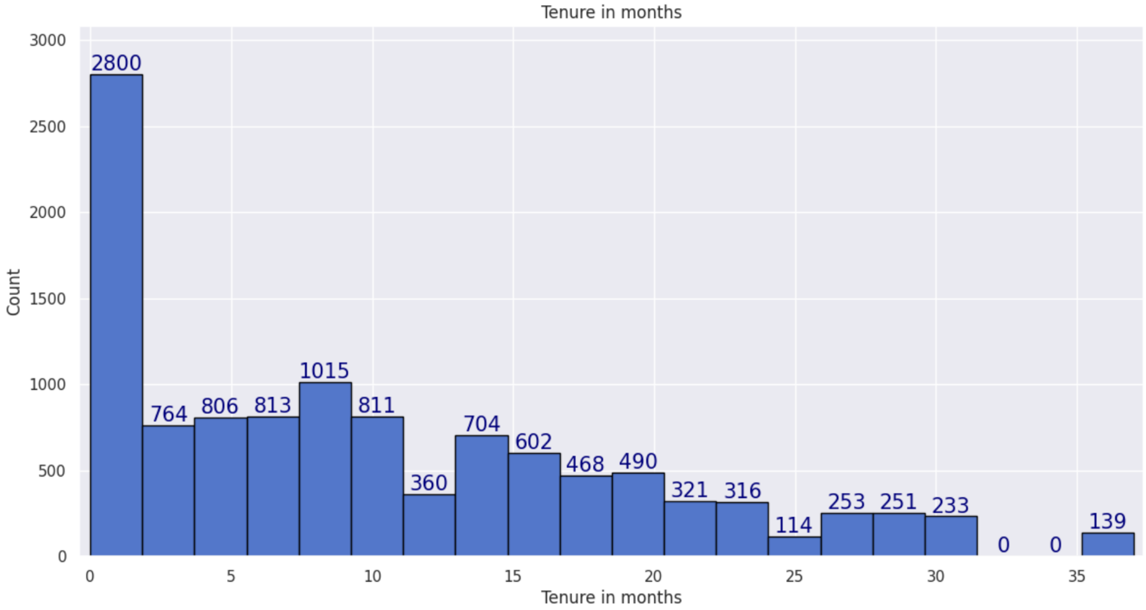
Continuous Columns/Variables	Categorical Columns/Variables
1. Tenure 2. CC_Contacted_LY 3. Service_Score 4. Account_user_count 5. rev_growth_yoy 6. Day_Since_CC_connect 7. cashback	1. Payment 2. Gender 3. account_segment 4. Marital_Status 5. Complain_ly 6. Login_device

For Univariate Analysis, we would be first considering two Continuous columns i.e "Tenure" and "rev_growth_yoy" and two categorical columns i.e "Account segment" and "Payment type" as they seem to be having a bearing on churn.

We can observe that:

- 50% of the customer accounts are 9 months old.
- 25% of the customer accounts are 2 months old.
- 75% of the customer accounts are 16 months old.
- Maximum tenure is 99 months and minimum tenure is 0 months
- Mean of tenure value i.e 11 months is higher than the median i.e 9 months, implying that there are exceptional cases.

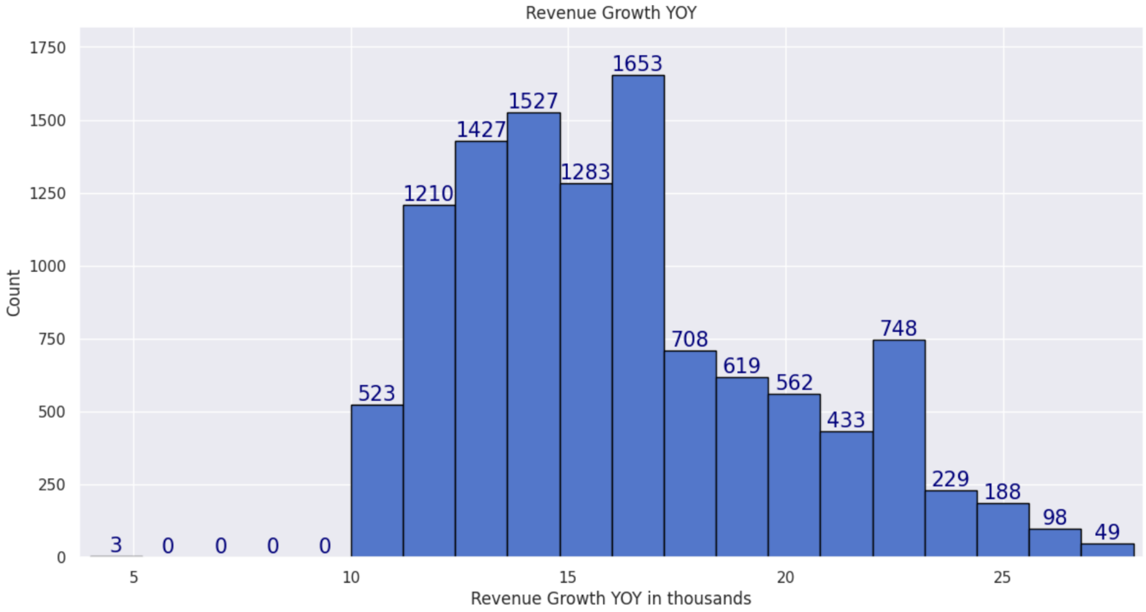
Graphically we can represent it as below:



We have performed a similar analysis of the "Revenue Growth Year on Year" column as follows. We can observe that:

- 50% accounts where Revenue (YoY) is between 4 and 15 thousands.
- 25% accounts where Revenue (YoY) is 13 thousands.
- 75% accounts where Revenue (YoY) is 19 thousands.
- Maximum revenue (YOY) is 28 thousand and minimum revenue is 4 thousand.
- Mean value is 16 thousands which is higher than the median value 15 thousands, implying that there are outliers.

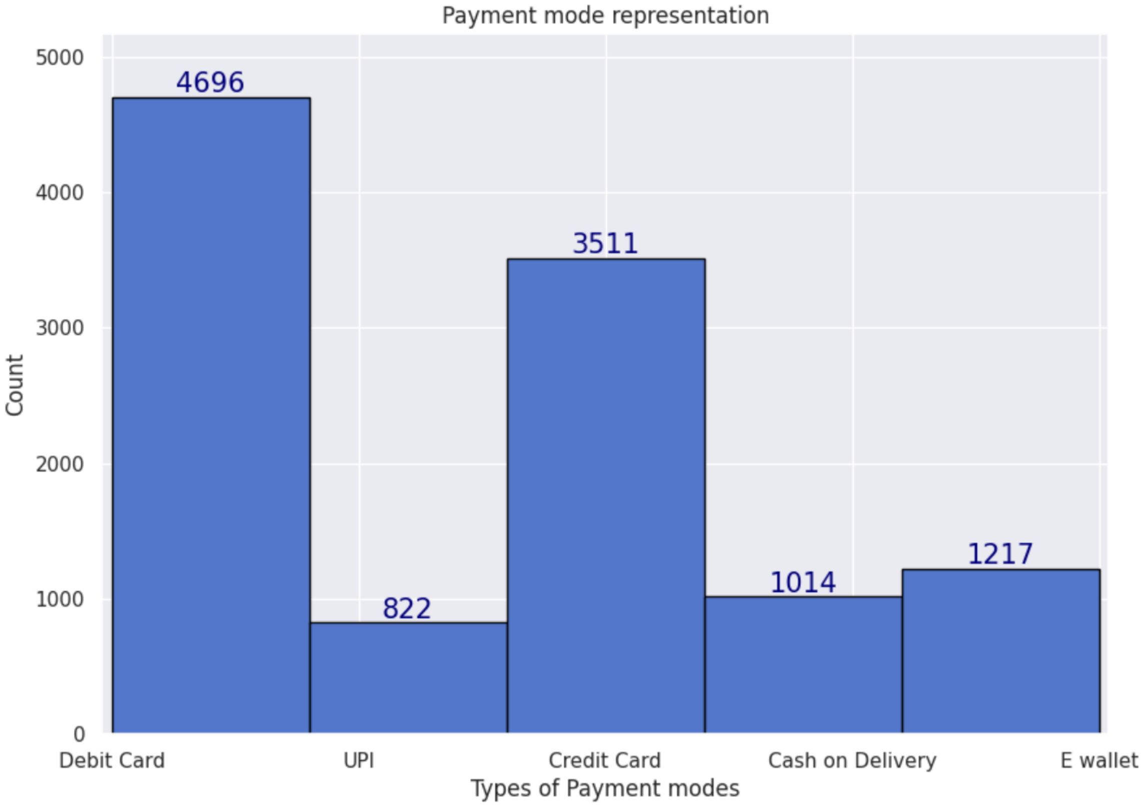
Graphically represented as below:



We have analyzed the categorical columns, we have drawn inferences for Payment type as follows:

- The most used payment type has been "Debit Card" which remains popular among 41.7% of accounts.
- "Credit card" mode remains a close second among 31.35% of accounts.
- Other payment modes have near equal distribution among on no.of accounts.

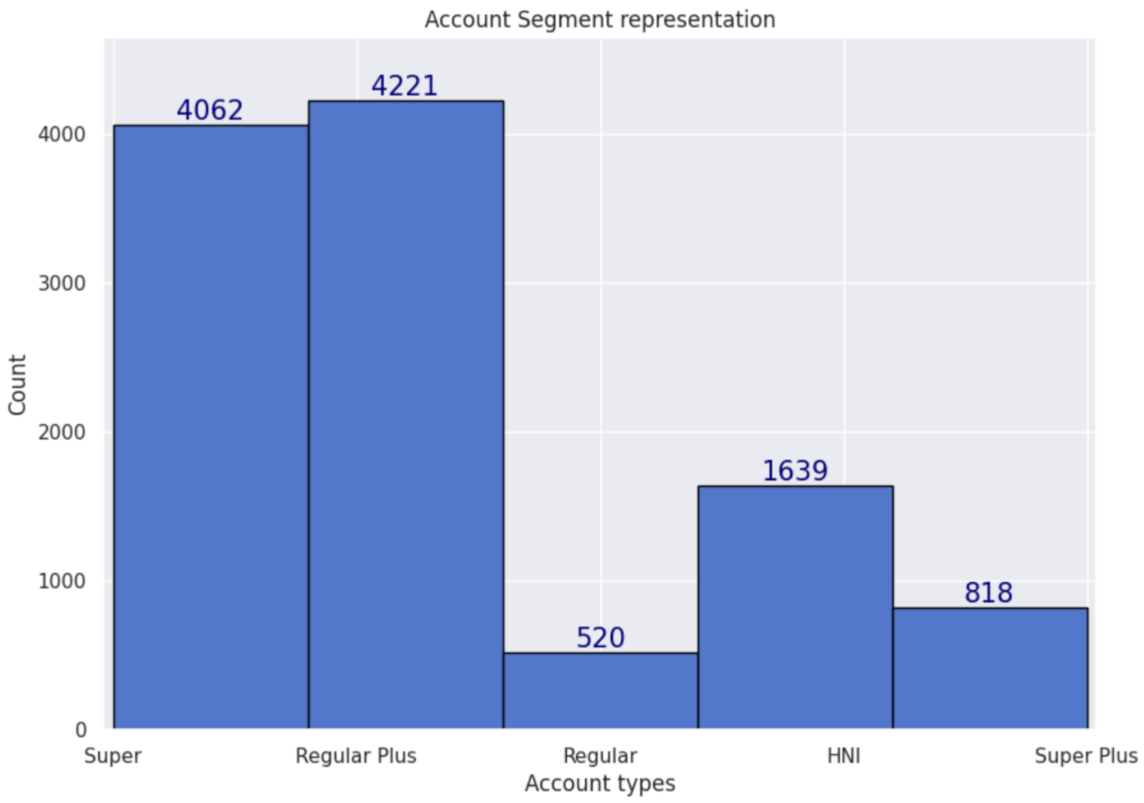
Graphically represented as below:



Outcome of "Account Segment" analysis as follows:

- 37.48% customers are subscribed to Regular Plus
- 36.07% customers are subscribed with Super, which is close second.

Graphically represented as below:



Since most of the customers belong to the Regular Plus Category, we have carried out bi-variate analysis to find answers to the open questions.

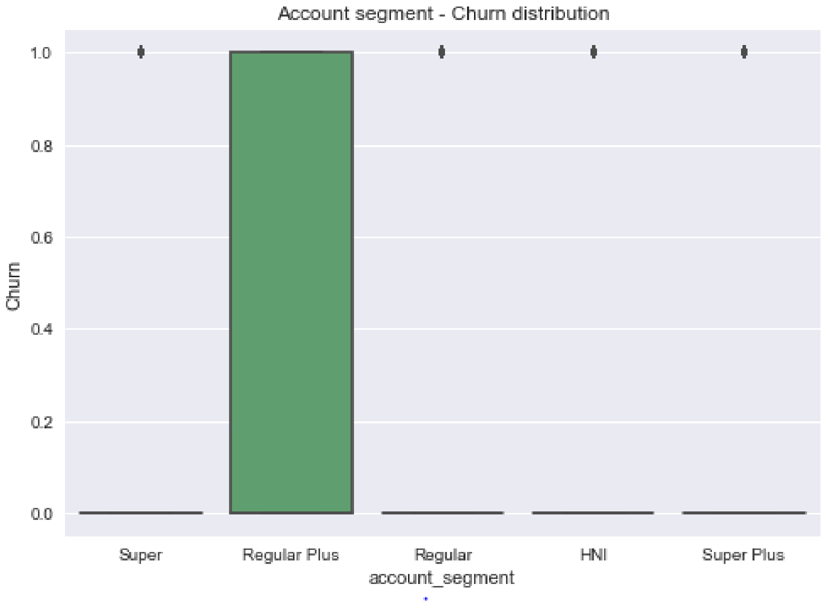
- What is the correlation between churn and listed categories?
- What are the measures which could be taken to bring these customers to at least one segment above, i.e to Super category?

2.2.5.2) Bivariate Analysis

Outcome of analysis done between Churn and Account segment column is showcased in the diagram below.

- 60.39% of churn is recorded from the Regular Plus account segment.

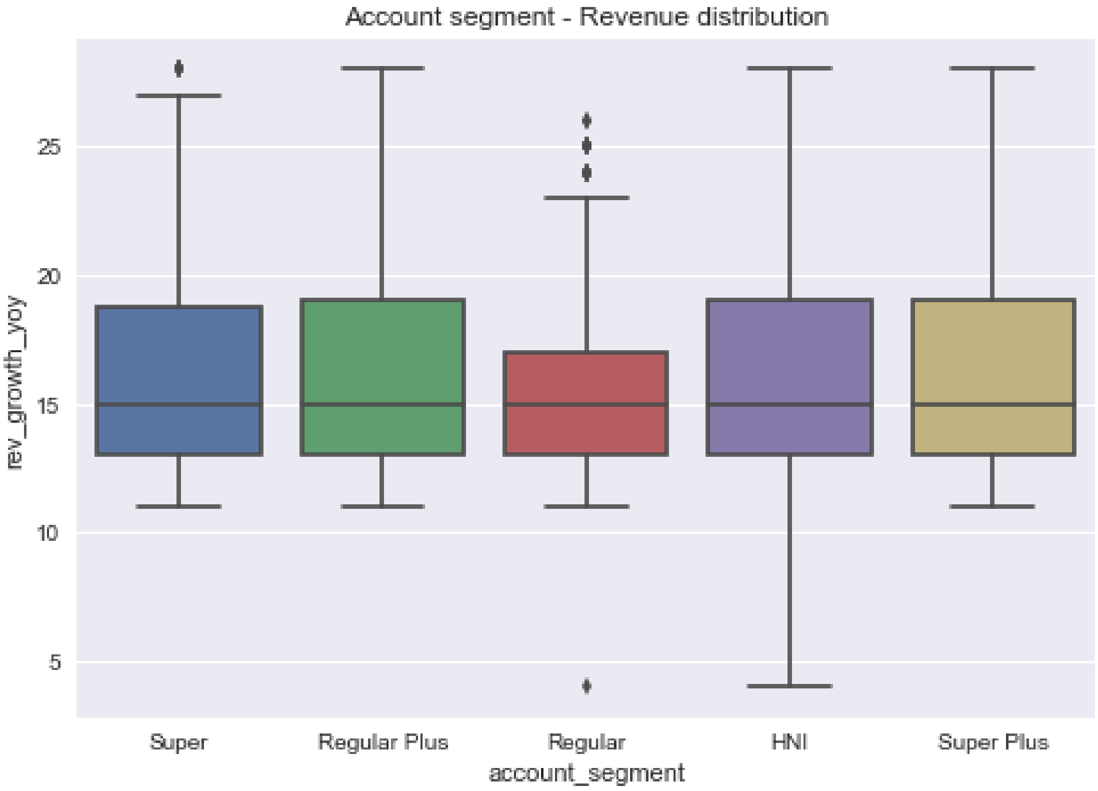
Graphically represented as below:



We have analyzed which account segment has the highest contribution with respect to revenue.

- The Regular Plus Category account segment has got the most contribution to Revenue, also recorded highest churn from this segment concerning.
- The other finding is that the second highest revenue contribution is from the Super category account segment.
- Median for all the segments remains 15 thousand.

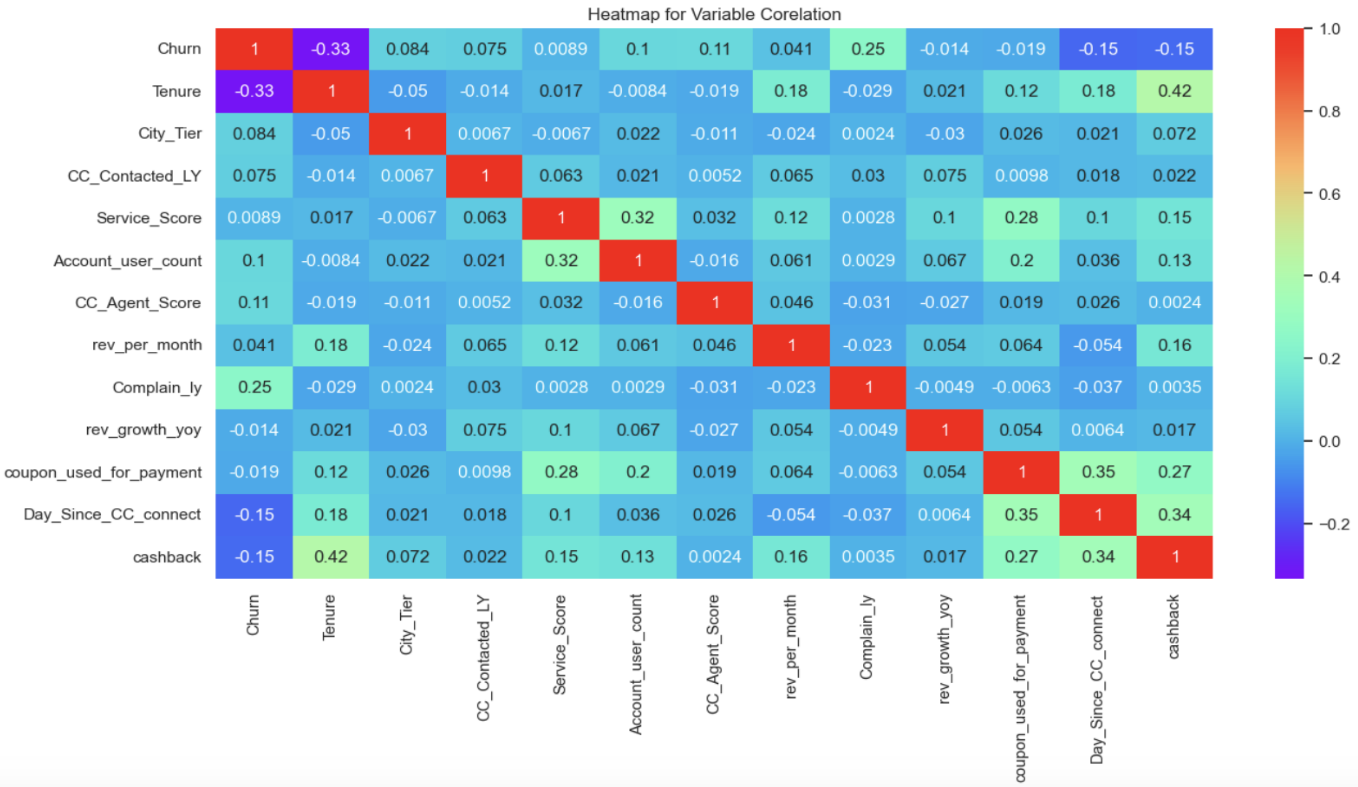
Graphically represented as below:



For correlation, we would be considering the three columns Tenure, Account User Count and Churn. Obtaining the correlation between these columns, results as follows:

- We find that there is a positive correlation between Account user count and Churn.
- The highest correlation is between Account User count and Service score columns.

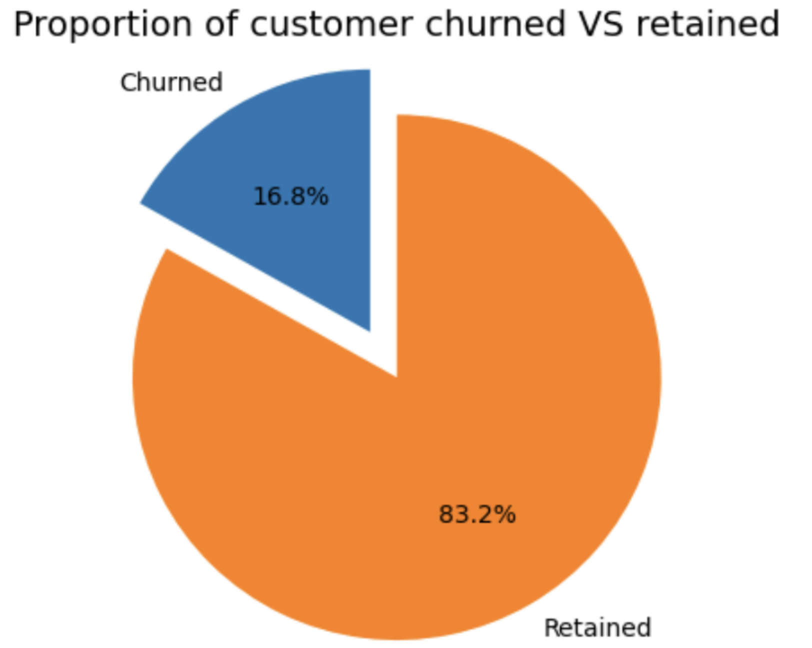
Graphically represented using heatmap as follows:



	Churn	Tenure	City_Tier	CC_Contacted_LY	Service_Score	Account_user_count	CC_Agent_Score	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback
Churn	1	-0.33	0.084	0.075	0.0089	0.1	0.11	0.041	0.25	-0.014	-0.019	-0.15	-0.15
Tenure	-0.33	1	-0.05	-0.014	0.017	-0.0084	-0.019	0.18	-0.029	0.021	0.12	0.18	0.42
City_Tier	0.084	-0.05	1	0.0067	-0.0067	0.022	-0.011	-0.024	0.0024	-0.03	0.026	0.021	0.072
CC_Contacted_LY	0.075	-0.014	0.0067	1	0.063	0.021	0.0052	0.065	0.03	0.075	0.0098	0.018	0.022
Service_Score	0.0089	0.017	-0.0067	0.063	1	0.32	0.032	0.12	0.0028	0.1	0.28	0.1	0.15
Account_user_count	0.1	-0.0084	0.022	0.021	0.32	1	-0.016	0.061	0.0029	0.067	0.2	0.036	0.13
CC_Agent_Score	0.11	-0.019	-0.011	0.0052	0.032	-0.016	1	0.046	-0.031	-0.027	0.019	0.026	0.0024
rev_per_month	0.041	0.18	-0.024	0.065	0.12	0.061	0.046	1	-0.023	0.054	0.064	-0.054	0.16
Complain_ly	0.25	-0.029	0.0024	0.03	0.0028	0.0029	-0.031	-0.023	1	-0.0049	-0.0063	-0.037	0.0035
rev_growth_yoy	-0.014	0.021	-0.03	0.075	0.1	0.067	-0.027	0.054	-0.0049	1	0.054	0.0064	0.017
coupon_used_for_payment	-0.019	0.12	0.026	0.0098	0.28	0.2	0.019	0.064	-0.0063	0.054	1	0.35	0.27
Day_Since_CC_connect	-0.15	0.18	0.021	0.018	0.1	0.036	0.026	-0.054	-0.037	0.0064	0.35	1	0.34
cashback	-0.15	0.42	0.072	0.022	0.15	0.13	0.0024	0.16	0.0035	0.017	0.27	0.34	1

2.2.5.3) Unbalanced Data and Treatment

With a primary observation the churn data is found to be unbalanced. As we are trying to classify churn data into two buckets, the outcome of data distribution is found to be in 83:17 ratio. Evidence of this finding was listed below as shown in the graph:



Accurate distribution count can also be determined by using the shape function over data frames extracted using churn data, listed below.

Customer Churned : 1896 (16.84%)

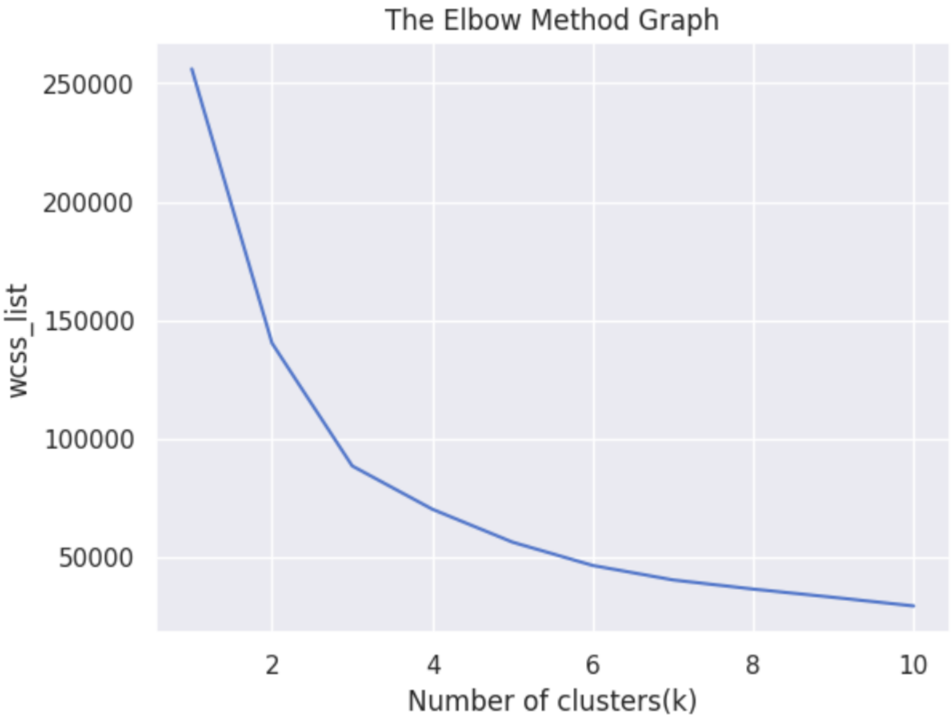
Customer Not-churn : 9364 (83.16%)

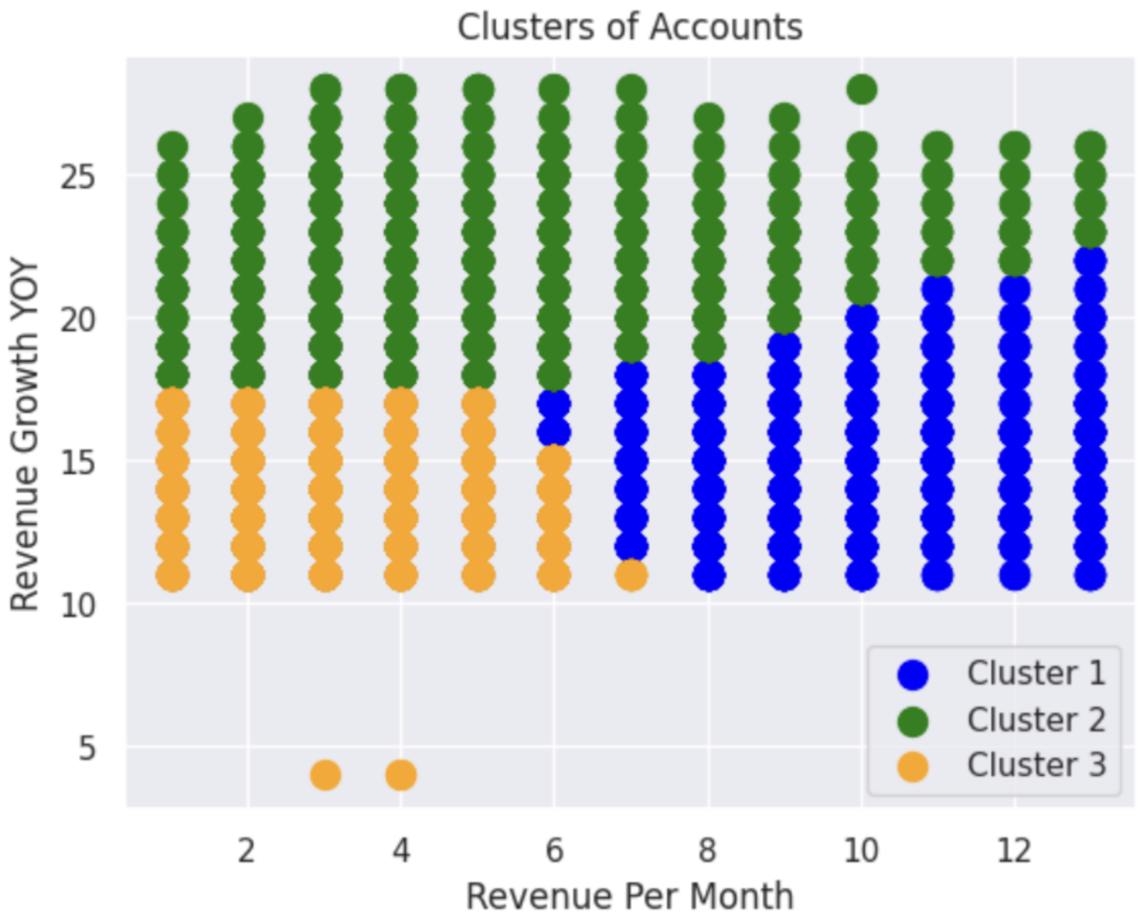
Good ML model recommends data must be balanced so that prediction could be accurate and avoid biased results.

2.2.5.4) Business Insights Using Clustering

We used K-means clustering technique to derive the correlation between two dependent variables "Revenue_Per_Month" and "Revenue_Growth_YOY" in order to derive the insights upon a company's revenue. Steps used are as follows:

- Elbow method is used to derive an elbow point which derives the no.of clusters can be present. This method has resulted in recommendation to have 3 clusters.
- Based on this value, the scatter plot was plotted for 3 clusters.





Insights derived from the same as follows:

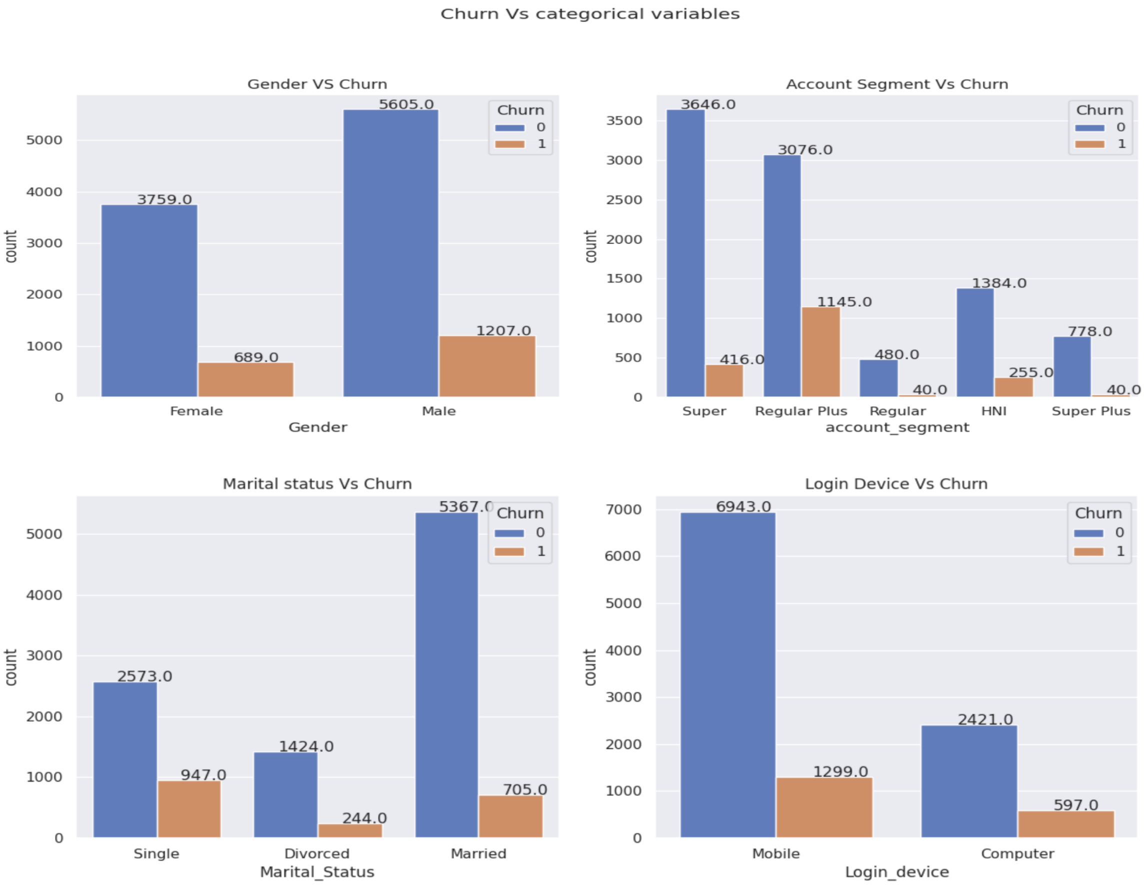
- Cluster 3 showcasing the customers density with lower contribution towards monthly and yearly growth. Organizations must focus on this sector to **increase their business**.
- Cluster 2 is showcasing higher customer density contributing to higher growth year on year, spread across the range of monthly revenue. Organizations must focus on this sector to **build long term relationships** with the customer and look for opportunities to increase the growth and profitability.
- It is interesting to observe cluster 1 which was contributing higher monthly income but lower growth year or year. This is the sector where organizations must focus immediately and this would define the **major scope for customer retention**.

2.2.5.5) Any Other Business Insights

In an attempt to find business insights, we have found a correlation with categorical variables like Gender, LoginDevice associated with Churn variable and outcome as follows:

- 63.66% of the customers who did churn are “**male”**.
- “**Super**” category was popular and resulted in a lower churn of 21.94%.
- 60.39% of churn recorded from the “**Regular Plus**” category which is the highest churn in proportion.
- 49.94% of customers who are “**single**” are being churned.
- Churn remains proportional across other categories.

Evidence of this finding was listed below as shown in countplot using churn data distribution in relation with categorical variables.



2.3) Period of Study

This study was conducted between June 2023 to August 2023 as part of the academic year project.

CHAPTER 3

DATA ANALYSIS AND INTERPRETATION

In order to find the best churn prediction model, we have to build various models and interpret their performance measures. Some of the models that we have built are Decision Tree, K-Nearest-Neighbor/KNN, Random Forest and Logistic Regression.

3.a) Building Various Models

Random Forest:

- Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.
- It leverages an ensemble of multiple decision trees to generate predictions or classifications. By combining the outputs of these trees, the random forest algorithm delivers a consolidated and more accurate result.

Logistic Regression:

- Logistic Regression is a supervised machine learning technique that models the relationship between predictors and the probability of a categorical response. It uses a process known as maximum likelihood estimation (M.L.E) Logistic Regression is most often used to solve classification problems.

K-Nearest-Neighbour:

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- It assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.
- It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

Decision Tree:

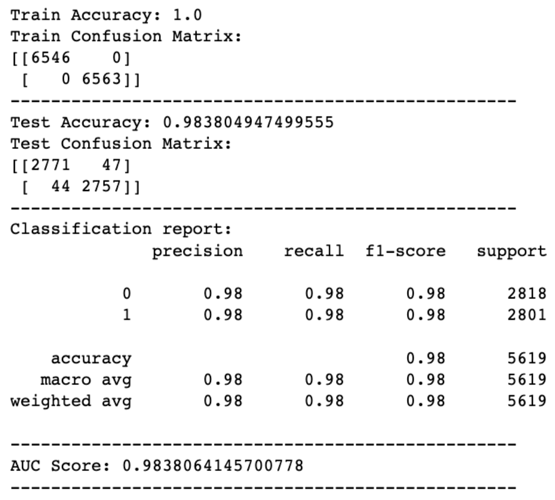
- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- It follows a tree-like model of decisions and their possible consequences. The algorithm works by recursively splitting the data into subsets based on the most significant feature at each node of the tree.

3.b) Testing the Model

In this phase we train the model and the same is evaluated using a testing dataset, detailed explanation is provided for Random Forest as below.

- Split the data:
    - The dataset must be split into two sets i.e Training set and Testing set.
    - The training set is used to train the random forest model, while the testing set is used to evaluate its performance.
    - We have split the data in a 70:30 ratio where 70% of data is considered for training and the rest of the 30% used for testing.
- Train the model:
    - Fit the training data to the random forest model.
    - Make predictions using the testing data.
- Evaluating Model Performance:
    - Compare the predicted values with the actual values from the testing set to assess how well the random forest model performs.
    - The performance metrics we have used are train/test accuracy score, precision, recall, F1-score and area under the ROC curve (AUC-ROC) and Confusion Matrix.
    - Validated the model using the k-fold model validation method where performance metrics from each fold are averaged to estimate the model's generalization performance.

Sample results derived was showcased below:

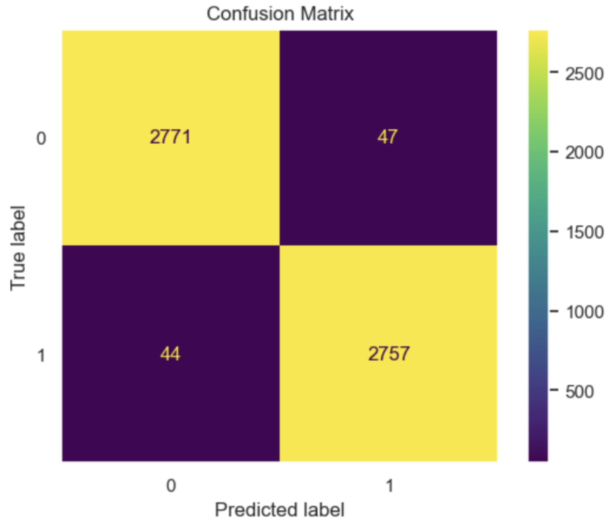


```
Train Accuracy: 1.0
Train Confusion Matrix:
[[6546    0]
 [   0 6563]]
--------------------------------------------------
Test Accuracy: 0.983804947499555
Test Confusion Matrix:
[[2771   47]
 [  44 2757]]
--------------------------------------------------
Classification report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      2818
           1       0.98      0.98      0.98      2801

    accuracy                           0.98      5619
   macro avg       0.98      0.98      0.98      5619
weighted avg       0.98      0.98      0.98      5619

--------------------------------------------------
AUC Score: 0.9838064145700778
--------------------------------------------------
```



Using the confusion matrix we can analyze that the model has predicted 2771+47=2818 accounts as non churner out of which 2771 is predicted to be actual non-churn and 47 accounts predicted to be churn along the way. Similarly, models have also predicted 44+2757=2801 accounts would churn out of which only 44 would not actually churn but a huge number of accounts i.e 2757 would churn eventually.

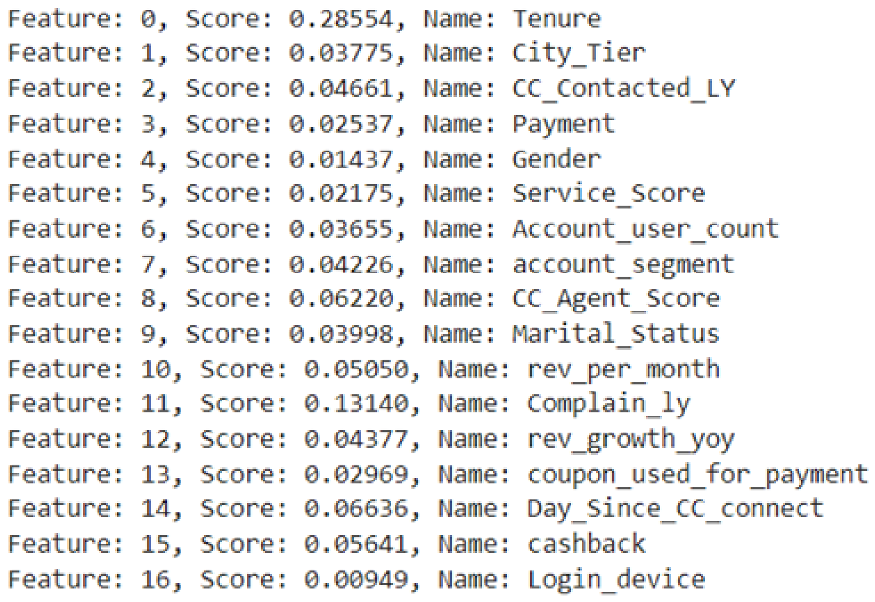
We have started the analysis with 1896 (16.84%) records churned. However after implementing this model prediction, we are finding 47+2757=2804 total accounts would churn thus pointing 10% additional accounts possibly churn as we look forward. This is clear evidence to showcase the impact of the problem is big and needs immediate counter measures.

3.c) Model Interpretation

Interpreting a random forest model involves understanding the importance of features, the impact of individual features on predictions and gaining insights into the decision-making process of the model. We have performed the same steps for all the 4 algorithms where a detailed explanation is provided for Random Forest as below.

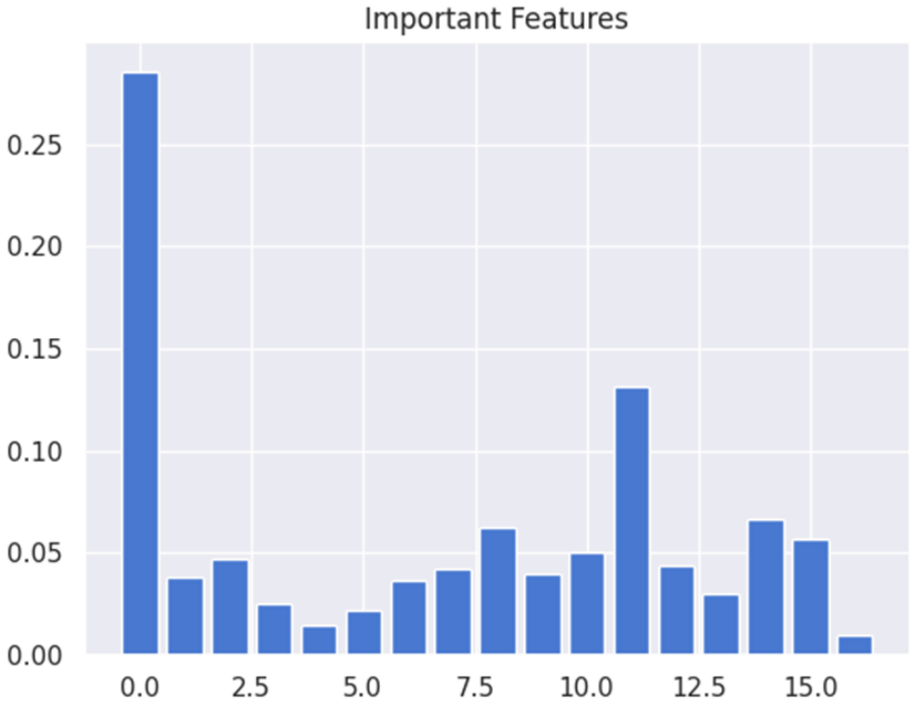
- Random forests provide a measure of feature importance that indicates the relative contribution of each feature to the model's predictive performance. The importance can be computed based on metrics such as mean decrease impurity or mean decrease accuracy. Higher values indicate more influential features.

Here is a list of important features that we interpreted from random forest model:



```
Feature: 0, Score: 0.28554, Name: Tenure
Feature: 1, Score: 0.03775, Name: City_Tier
Feature: 2, Score: 0.04661, Name: CC_Contacted_LY
Feature: 3, Score: 0.02537, Name: Payment
Feature: 4, Score: 0.01437, Name: Gender
Feature: 5, Score: 0.02175, Name: Service_Score
Feature: 6, Score: 0.03655, Name: Account_user_count
Feature: 7, Score: 0.04226, Name: account_segment
Feature: 8, Score: 0.06220, Name: CC_Agent_Score
Feature: 9, Score: 0.03998, Name: Marital_Status
Feature: 10, Score: 0.05050, Name: rev_per_month
Feature: 11, Score: 0.13140, Name: Complain_ly
Feature: 12, Score: 0.04377, Name: rev_growth_yoy
Feature: 13, Score: 0.02969, Name: coupon_used_for_payment
Feature: 14, Score: 0.06636, Name: Day_Since_CC_connect
Feature: 15, Score: 0.05641, Name: cashback
Feature: 16, Score: 0.00949, Name: Login_device
```

Graphical representation is showcased as follows:



3.d) Ensemble Modeling

Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model. Practically we have not performed this step and look forward to working on them during future scope.

3.e) Model Tuning Measures

Classification model tuning is required to improve the performance and effectiveness of the model in making accurate predictions. This would result in performance improvement while also helping in preventing overfitting. Using tuning the model would be well adapted to the dataset and proves best fit with data specific characteristics. It would also help in handling imbalanced data.

The primary objective is to try with various hyperparameter combinations to check the trade-off between precision and recall values. Also, to limit the grid-search range for optimum utilization of computational resources.

We have performed the same steps for all the 4 algorithms where a detailed explanation is provided for Random Forest as below.

- We have formatted a hyper-parameter dictionary for RandomForest including parameters like "n_estimators" and "max_depth".
- When used with the "RandomizedSearchCV" class, it would test various combinations of hyper-parameters train/test data and help in deriving the best combination of parameters resulting in best performance.
- Once derived the best model, the same is tested for train/test data accuracy.
- Following this step, the best combination of hyperparameters is extracted using which model is rebuilt.
- Rebuilt model is trained and tested with sample data, derived with revised performance metrics. We have also compared the results with/without tuning to ensure the performance increment.

3.f) Interpretation of Optimum Model

In this section we collated the results of various models built, interpreted, tuned and tested to derive favorable results.

- Accuracy is a measure of how correctly a ML model predicts the outcome. It is calculated by dividing the number of correct predictions by the total number of predictions made.
- Sensitivity: measures the ability of a model to correctly identify positive instances from all the actual positive instances in the dataset.
- Specificity: measures the ability of a model to correctly identify negative instances from all the actual negative instances in the dataset.
- F-measure: provides a balanced measure of a model's performance by considering both false positives and false negatives.
- 10 Folds Test Validation: method allows for a more robust evaluation of the model's performance, as it utilizes different subsets of the data for training and validation.
- AUC Score: indicates better model performance in distinguishing between the two classes, with a score of 1 representing a perfect classifier.
- Train/Test Accuracy: indicates the performance score <TODO>

Model comparison table :

Algorithm	**Accuracy**	**Sensitivity / Recall**	**Specificity / Precision**	**F measure (F1 Score)**	**10 Folds Test Validation**	**AUC Score**	**Train Accuracy**	**Test Accuracy**	**Train Accuracy (Tuned)**	**Test Accuracy (Tuned)**
Logical Regression	0.80	0.80	0.80	0.80	0.7997	0.8046	0.80	0.79	0.80	0.79
Random Forest	0.98	0.98	0.98	0.98	0.9771	0.9804	1.0	0.97	0.99	0.97
Decision Tree	0.94	0.94	0.94	0.94	0.9318	0.9401	1.0	0.94	0.97	0.93
KNN	0.99	0.99	0.99	0.99	0.9826	0.9886	0.98	0.95	1.0	0.98

CHAPTER 4

FINDINGS, RECOMMENDATIONS AND CONCLUSION

4.1 Findings Based on Observations

- Dataset is unbalanced towards the churn and the distribution is 83:17 ratio.
- Dataset contains 12% missing values, 3 irrelevant features accounting for 22% of total features and 2% garbage values (special characters).
- The dataset has a combination of 6 categorical and 7 numerical data columns.
- Dataset contains duplicate categorical values for 2 columns.
- Dataset contains outliers across a few columns.

4.2 Findings Based on analysis of Data

- 16.8% of the customers have churned overall.
- 50% of the customer accounts tenure is 9 months.
- 15% churn recorded from customers who have contacted customer support in the last 10 days of the tenure.
- 50% accounts where revenue (YoY) is between 4 and 15 thousands.
- 41.7% of accounts used payment type has been "Debit Card".
- "Credit card" mode remains a close second among 31.35% of accounts.
- 37.48% of customers are subscribed to the Regular Plus category.
- 36.07% customers are subscribed with Super, which is close second.
- 60.39% of churn is recorded from the Regular Plus account segment.
- 63.66% of the customers who did churn are **"male"**.
- "**Super**" category was popular and resulted in a lower churn of 21.94%.
- 60.39% of churn recorded from the "**Regular Plus**" category which is the highest churn in proportion.
- 49.94% of customers who are "**single**" are being churned.

4.3 General findings

- The Account with higher Tenure value has a lesser churn rate.
- The Regular Plus Category account segment has got the most contribution to Revenue, also recorded highest churn from this segment concerning.
- The other finding is that the second highest revenue contribution is from the Super category account segment.
- We find that there is a positive correlation between Account user count and Churn.
- The highest correlation is between Account User count and Service score columns.
- One of the clusters found to be contributing higher monthly income but lower growth year or year.
- Higher the tenure resulting in lower the churn as it found strong negative correlation.
- Higher churn observed with higher no.of complaints raised in the last 12 months.

4.4 Recommendation based on findings

- Potential churners who are “male” and using the “Regular Plus” segment can be offered with a free upgrade to “Super” category.
  - Targeted campaigns for “Regular Plus” as a segment can help us prevent the churn by 18%.
- Offer goodies to “credit card” users in order to enhance their loyalty towards the company.
- Establish a specialized customer service team to counter complaints raised from “Regular Plus” Segment.
  - Find the scope of service quality improvement and address the concerns on priority.
  - Offer cashback to customers who have raised repeated complaints in the last 12 months, to ensure their loyalty and thanking them for their patience.

4.5 Suggestions for areas of improvement

- Service quality would need an improvement
  - Better service would result in higher satisfaction & lower the complaints.
- Customer service is a crucial area where organizations must focus.
  - Conduct detailed study on the context of customer problems.
  - Handle customers with repeated escalations on priority.
  - Setup dedicated focus for “Regular Plus” category customers.

4.6 Scope for future research

- Adding new variables.
- Ensemble models or placing them in combination to derive enhanced outcomes.
- Analyzing the outcome post implementation of recommendations with new tools or techniques.
- Collect more data towards the trend of churn so that seasonal factors could be derived. This can help us trigger season campaigns.
- Provide more data towards customer feedback, complaints raised by them and time taken by the support team to resolve the issues.

4.7 Conclusion

In conclusion, the study on customer churn within the DTH (Direct-to-Home) company sheds light on the dynamics that drive customer attrition. Through an effective analysis of subscriber behavior, service quality and pricing strategies, the study highlights the key factors currently influencing customer decisions to churn. The findings highlight the significance of proactive measures to enhance customer engagement, improve service offerings and retention strategies. This research equips DTH companies with actionable insights to mitigate churn and ensure customer loyalty.

As we conclude, we expect the DTH company would implement the actions while collecting the revised feedback from the customers. We would like to revisit the effect of our recommendation in a few months to identify the effect and help the organization with revision required in strategy if any. Finally, we are thankful to the DTH company for helping us with this project and willing to work again in future.