MBA
Semester – IV
Capstone Project

Name	Savita Ganapati Hima Bindu Darshan Rahul Ravish Kumar
Project	Customer Churn Prediction Model
Group	6
Date of Submission	18th August 2023

A study on

"Customer Churn Prediction"

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:

Student Name	USN
Savita Ganapati	211VMBR03991
Darshan	211VMBR01115
Rahul	211VMBR03448
Hima Bindu	211VMBR01643
Ravish Kumar	211VMBR03579

Under the guidance of:

Hrushikesha Shastry

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2022-23

DECLARATION

We, ***Savita Ganapati, Hima Bindu, Ravish Kumar, Rahul, Darshan*** hereby declare that the Research Project Report titled *"**Customer Churn Prediction Model**" has been* prepared by me under the guidance of ***Hrushikesha Shastry.*** We 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. We have undergone a project for a period of Eight Weeks. We further declare that this Project is based on the original study undertaken by us and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Bengaluru

Date:18th August 2023

Student Name	USN
Savita Ganapati	211VMBR03991
Darshan	211VMBR01115
Rahul	211VMBR03448
Hima Bindu	211VMBR01643
Ravish Kumar	211VMBR03579

EXECUTIVE SUMMARY

In today's fiercely competitive market, companies operating in the e-commerce sector face significant challenges in retaining existing customers. This is particularly crucial for companies like E-commerce or Direct-to-Home (DTH) service providers, where the loss of one account could mean losing multiple customers. To address this challenge, the company in focus aims to develop a churn prediction model that will enable them to identify potential churners and provide targeted offers to retain them. This paper presents a case study that explores the application of churn prediction and segmented offers for customer retention in the chosen domain.

Industry Overview:

The e-commerce industry has witnessed unprecedented growth and development, driven by technological advancements and changing consumer preferences. However, this rapid expansion has also led to intense competition among companies within the sector (Zhang, 2015). In such a dynamic landscape, customer retention has become a critical success factor, necessitating the implementation of effective strategies to maintain a loyal customer base.

Importance of Churn Prediction:

Churn prediction plays a vital role in addressing customer retention challenges. By accurately identifying potential churners, companies can proactively engage with these customers, offering personalized solutions to prevent them from switching to competitors. For the chosen company, where one account can have multiple customers, the impact of losing an account is magnified, making churn prediction an even more crucial aspect of their business strategy.

Account Churn and Customer Impact:

In the case of the chosen company, account churn represents a significant concern. Losing a single account not only means losing one customer but potentially multiple customers associated with that account. This has a direct impact on the company's revenue and market position. Therefore, it becomes imperative to develop a comprehensive understanding of customer behavior and preferences to tackle churn effectively.

Churn Prediction and Segmented Offers:

To address the challenges posed by customer churn, the company aims to develop a churn prediction model tailored to their specific needs. This model will leverage advanced data analysis techniques and machine learning algorithms to identify customers at risk of churning. By analyzing historical customer data, including purchase patterns, browsing

behavior, and customer feedback, the model will generate predictive insights to anticipate churn likelihood.

Once potential churners are identified, the company can implement segmented offers that cater to each customer's unique needs and preferences. By personalizing the offers based on customer segments, the company can increase the likelihood of customer retention and loyalty. These segmented offers may include targeted discounts, exclusive promotions, personalized recommendations, or enhanced customer support.

Benefits and Implications:

Implementing an effective churn prediction and segmented offers strategy can yield several benefits for the company. Firstly, it allows the company to focus its resources on high-risk customers, optimizing retention efforts and reducing overall churn rate. Secondly, personalized offers and tailored experiences enhance customer satisfaction and loyalty, fostering long-term relationships. Finally, the company can gain a competitive edge by leveraging data-driven insights to proactively address churn and improve customer retention rates.

Conclusion:

In conclusion, in the highly competitive e-commerce industry, customer retention is a critical factor for success. For companies like the one in focus, where losing a single account can have a substantial impact, churn prediction and segmented offers are essential strategies for mitigating customer churn. By leveraging advanced analytics and machine learning techniques, the company can accurately identify potential churners and tailor offers to meet their individual needs. By adopting this proactive approach, the company can enhance customer loyalty, reduce churn, and ultimately achieve sustainable growth in the market.

Keywords: E-commerce, DTH, churn prediction, customer retention, segmented offers, personalized marketing, data analytics, machine learning.

TABLE OF CONTENTS

Title	Page Nos.
Executive Summary	i
List of Tables	ii
List of Graphs	iii
Chapter 1: Introduction and Background	1-3
Chapter 2: Research Methodology	4-9
Chapter 3: Data Analysis and Interpretation	10-15
Chapter 4: Findings, recommendations and conclusion	
References	
Annexures	

Table 1: Variables 7
Graph 1 : Account vs Churn 9
Table 2: Null Values 9
Table 3: Data Summary 10
Graph 2: Missing Value 12
Table 4: Null Value Count 13
Image 1 : Null Analysis 13
Image 2: Imputation 1 14
Image 3: Imputation 2 14
Graph 3: Outliers 1 14
Graph 4: Outliers 2 15
Graph 5: Outliers Plot Analysis 16
Graph 6: Multivariate Analysis 16
Image 4: Hot encoding 17
Image 5: Feature Scaling 18
Image 6: Feature Selection 18
Image 7: VIF 19
Image 8: Model Training 19
Table 5: Model Result 20
Image 9: Training 2 20
Table 6: Training Result 2 21
Image 10: Recursive Feature 22
Graph 7 : Univariate Analysis 23
Graph 8: Bivariate Analysis 23
Image 11 : Logistic Regression Model 24
Image 12 : Logistic Regression Model Performance 25
Table 7 : Logistic Regression Model AUC 26
Table 8 : Logistic Regression Model Performance 26
Image 13 : Decision Tree Model 26
Table 9 : Decision tree Model Result 27
Table 10 : Decision tree Model Performance 27
Table 11 : Decision tree Model AUC 28
Table 12 : Decision tree Model Performance 28
Table 13 : Random Forest Model Prediction 29
Table 14 : Random Forest Model Result 29
Table 15 : Random Forest Model AUC 30
Table 16 : Random Forest Model Performance 30
Table 17 : Model Accuracy on Training Data 30
Table 17 : Model Accuracy on Test Data 30
Table 17 : Model Accuracy Chart 31
Graph 9 : SHAP Result 32

CHAPTER 1

INTRODUCTION AND BACKGROUND

INTRODUCTION AND BACKGROUND

1.1 Executive Summary

This project focuses on developing a churn prediction model for an E-commerce company or Direct-to-Home (DTH) provider facing intense competition in the current market. Retaining existing customers has become a significant challenge, as losing one account can result in the loss of multiple customers. The objective is to develop a model that accurately predicts churn and enables the company to provide segmented offers to potential churners, thereby increasing customer retention.

The project team is tasked with developing the churn prediction model and providing strategic business recommendations for a unique campaign. However, it is crucial to balance the campaign's attractiveness with financial sustainability. The recommendations need to be clear, effective, and avoid excessive giveaways or subsidies that could lead to losses for the company.

1.2 Introduction and Background

Customer retention poses a significant challenge for businesses in the highly competitive e-commerce sector. Retaining existing customers is more crucial and cost-effective than acquiring new ones. This project aims to develop a churn prediction model specifically for the e-commerce industry, enabling companies to identify key attributes leading to customer churn.

The study will employ data analysis techniques to examine historical customer data and identify patterns and factors contributing to churn. By leveraging advanced analytics and machine learning algorithms, the project team will develop a robust churn prediction model that accurately predicts customer churn.

The key focus is on understanding the key attributes and behaviors that contribute to customer churn in the e-commerce sector. By identifying these factors, businesses can proactively engage with customers at risk of churn and implement targeted strategies to retain them.

The developed churn prediction model will facilitate personalized marketing campaigns and segmented offers, ensuring tailored approaches to address the unique needs and preferences of potential churners. This approach will increase the effectiveness of customer retention efforts and enhance customer loyalty.

The project team will also consider the financial implications of the recommended campaign strategies. The proposed recommendations will strike a balance between attractive offers and financial sustainability to gain approval from the revenue assurance team.

Ultimately, the churn prediction model and the associated campaign recommendations will empower e-commerce companies to reduce customer churn, enhance customer satisfaction, and achieve long-term business growth in the highly competitive market.

1.3 Problem Statement

The E-commerce sector faces a significant challenge in retaining existing customers due to intense competition and the high cost of acquiring new customers. Customer churn, or the loss of customers, has a profound impact on businesses, particularly in cases where one account can represent multiple customers. The company under consideration aims to address this issue by developing a churn prediction model that can accurately identify potential churners and provide targeted offers to retain them. The challenge lies in creating a unique campaign recommendation that effectively reduces customer churn without compromising the company's financial sustainability. The recommendation should strike a balance between attractive offers and avoiding excessive giveaways or subsidies that could result in financial losses. The project seeks to develop a robust churn prediction model and provide clear and viable campaign recommendations that enhance customer retention and drive long-term growth in the e-commerce sector.

1.4 Objective of Study

- **Develop a Churn Prediction Model:**

The primary objective of the study is to develop a robust churn prediction model specifically tailored to the e-commerce sector. This model will utilize advanced data analysis techniques and machine learning algorithms to accurately identify potential churners based on historical customer data.

- **Identify Key Attributes Leading to Churn:**

Through the churn prediction model, the study aims to identify the key attributes and behaviors that contribute to customer churn in the e-commerce sector. Understanding these factors will provide valuable insights into customer preferences, enabling businesses to address potential churn risks effectively.

- **Provide Segmented Offers:**

The study aims to develop personalized and segmented offers for potential churners identified by the churn prediction model. By tailoring offers to specific customer segments, businesses can increase the effectiveness of retention efforts and enhance customer loyalty.

- **Ensure Financial Viability of Campaign Recommendations:**

The study will carefully consider the financial implications of the recommended campaign strategies. The objective is to provide unique campaign recommendations that strike a balance between attractiveness and financial sustainability, ensuring approval from the revenue assurance team.

- **Enhance Customer Retention and Loyalty:**

Ultimately, the study seeks to enhance customer retention and loyalty in the e-commerce sector. By leveraging the churn prediction model and implementing targeted strategies, businesses aim to reduce churn rates, improve customer satisfaction, and foster long-term relationships with their customers.

- **Drive Sustainable Growth:**

The study aims to contribute to the sustainable growth of businesses in the e-commerce sector. By effectively addressing customer churn, companies can gain a competitive edge, increase market share, and achieve long-term business growth.

1.5 Company and industry overview

The DTH (Direct-to-Home) industry in India has been facing a decline in subscriber numbers in recent years. In 2022, the industry lost 2.3 million subscribers, bringing the total number of subscribers down to 69.2 million.

There are a number of factors that have contributed to the decline in DTH subscribers in India. These factors include:

- The rise of OTT (Over-the-Top) platforms, such as Netflix, Amazon Prime Video, and Disney+ Hotstar. These platforms offer a wide variety of content at a relatively low cost, which has made them attractive to many consumers.
- The increasing popularity of streaming services has led to a decline in the demand for traditional cable and satellite TV services.
- The rising cost of DTH subscriptions has also made them less affordable for some consumers.

The decline in DTH subscribers has put a strain on the industry. DTH providers have been forced to cut costs and lay off employees in order to remain profitable. The industry is also facing increasing competition from OTT platforms.

In order to address the decline in subscribers, DTH providers need to find ways to differentiate themselves from OTT platforms. They need to offer a wider variety of content, improve their customer service, and reduce their prices.

If DTH providers are unable to address the decline in subscribers, they could face further losses in the future. This could lead to the consolidation of the industry, as smaller providers are forced to merge with larger providers in order to survive.

Here are some additional thoughts on the DTH industry and customer churn problem:

- The decline in DTH subscribers is a global trend, not just a problem in India.
- The rise of OTT platforms is one of the main drivers of this decline.

- DTH providers need to find ways to compete with OTT platforms in order to survive.
- This could include offering a wider variety of content, improving customer service, and reducing prices.
- The future of the DTH industry is uncertain, but it is clear that the industry is facing some challenges.

1.6 Overview of Theoretical Concepts

The proposed research aims to build a churn prediction model using the CRISP-DM(Huber 2019) methodology, which provides a structured approach to the data mining process. This methodology encompasses several stages starting with understanding the business problem and identifying the opportunity it presents. Data understanding involves gathering data from multiple sources, followed by data preparation, which includes cleaning the data by addressing missing values, outliers, and irrelevant columns.

The subsequent stages of the methodology are modeling, evaluation, and deployment. In these stages, the research will implement and test various models to identify churn, contributing to effective customer retention strategies.

The chosen methodology, CRISP-DM, was selected for its cross-industry applicability, providing a uniform framework for planning and managing data mining projects. It offers a roadmap for researchers and has been proven to be time and cost-effective.

By applying the CRISP-DM methodology and utilizing the identified variables specific to the churn prediction problem, the research project aims to guide the knowledge discovery process in a structured manner, ensuring efficient data understanding, preparation, modeling, evaluation, and implementation.

CHAPTER 2

Research Methodology

RESEARCH METHODOLOGY

2.1 Scope of the Study

The scope of study is to develop a customer churn model based on the given data set and provide business recommendations on the campaign.

The scope as given in the assignment is as below:

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(Feng 2018). 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.

2.2 Methodology

2.2.1 Research Design

2.2.2 Data Collection

The dataset about 11261 accounts are provided, where one account may contain more than one customer. The metadata of given dataset is as below:

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_L12m	How many times all the customers of the account has contacted customer care in last 12months
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_l12m	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 month)
coupon_used_l12m	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_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

Table 1: Variables

2.2.3 Sampling Method (if applicable)

As the given dataset is not very complex and huge, no sampling is required, and hence we will be using the complete dataset in our analysis.

2.2.4 Data Analysis Tools

To analyze the data, we will be using python programming along with respective libraries which are as below:

- Basic EDA
  - numpy
  - pandas
  - matplotlib
  - seaborn
- Model Preparation
  - sklearn
  - statsmodels
- Model Building
  - Logistic Regression
  - Decision Tree

    - RandomForestClassifier
- Model Performance
    - sklearn.metrices(roc_auc_score,roc_curve,classification_report,confusion_matrix)
- Model Validation
    - StandardScaler
    - Kfold
    - Cross Validation

2.3 Period of Study

The study period is not defined in this project. So we are assuming this as real time data of current customers, which needs to be analyzed for churn predictions and to minimize the churn.

2.4 Utility of Research

To explore and visualize the data in Python, the initial step involves importing several libraries such as pandas, numpy, and matplotlib. The next step is to analyze both numerical and categorical columns, while also identifying any missing data. In this particular dataset, the outcome variable is "Churn," and fortunately, there are no missing values in this column. However, it should be noted that the outcome variables are imbalanced, with a higher number of retained customers compared to churned customers, as indicated in the table below.

Graph 1 : Account vs Churn

There are 11260 records of accounts, where 9364 accounts records have no churn, and 1896 have churn which respectively has 9272 and 7203 users associated.

There were null values in the given data sets:

```
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
```

Table 2: Null Values

The given data set summary is as below:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Churn	11260.0	NaN	NaN	NaN	0.168384	0.374223	0.0	0.0	0.0	0.0	1.0
Tenure	11158	38	1	1351	NaN	NaN	NaN	NaN	NaN	NaN	NaN
City_Tier	11148.0	NaN	NaN	NaN	1.653929	0.915015	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	11158.0	NaN	NaN	NaN	17.867091	8.853269	4.0	11.0	16.0	23.0	132.0
Payment	11151	5	Debit Card	4587	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	11152	4	Male	6328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	11162.0	NaN	NaN	NaN	2.902526	0.725584	0.0	2.0	3.0	3.0	5.0
Account_user_count	11148	7	4	4569	NaN	NaN	NaN	NaN	NaN	NaN	NaN
account_segment	11163	7	Super	4062	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	11144.0	NaN	NaN	NaN	3.066493	1.379772	1.0	2.0	3.0	4.0	5.0
Marital_Status	11048	3	Married	5860	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	11158	59	3	1746	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Complain_ly	10903.0	NaN	NaN	NaN	0.285334	0.451594	0.0	0.0	0.0	1.0	1.0
rev_growth_yoy	11260	20	14	1524	NaN	NaN	NaN	NaN	NaN	NaN	NaN
coupon_used_for_payment	11260	20	1	4373	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Day_Since_CC_connect	10903	24	3	1816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cashback	10789	321	152	208	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Login_device	11039	3	Mobile	7482	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 3: Data Summary

In the next section, we'll correct the null values and missing values and develop the model for churn prediction.

Statistical Observation:-

From the above statistical information, we have found that there are 5 unique payment modes,7 unique account_segments and 3 login device types which are unique. It means we have 7 types of subscription plans for our customers.Most of the account_segment are "Super" and payment done by customers is through debit cards in terms of majority.Top login devices is "Mobile".

In terms of Service overall 3 star rating exists.17-18 average times customers have contacted the Customer care in last years. Given data indicates that 75% of the customers are from Tier 3 cities which shows the majority.

We will consider these valuable key points for our further analysis and based on that we would be able to provide campaign recommendation or segmented offers.

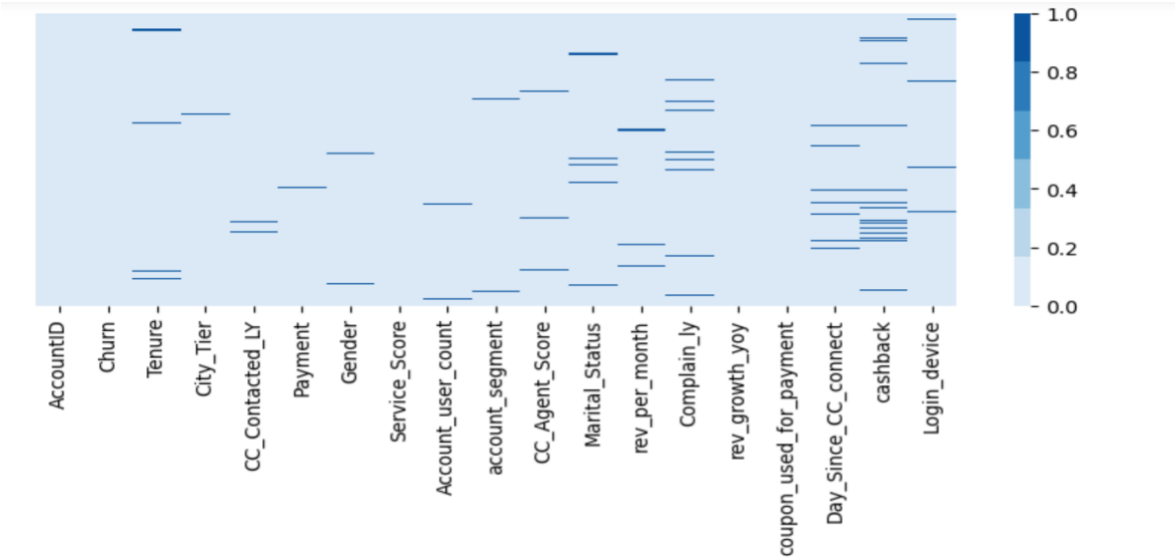
CHAPTER 3

DATA ANALYSIS AND INTERPRETATION

DATA ANALYSIS AND INTERPRETATION

Data Cleaning

- **Missing Values Treatment:-** First we started with missing value treatment using below codes:
  Missing Values(**In Visual**):-



Graph 2: Missing Value

"AccountId" columns is a unique identifier to keep records unique.For the processing & analysis its not going to make a sense.It doesn't have a statistical importance that'why i have dropped the column.

```
Churn_df.isnull().sum()
```

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	

Table 4: Null Value Count

We found that most of the columns contain missing values but we know simply deleting the null rows is not an option. Some features could have null values.After carefully dealing with columns,we found inconsistency in data

Multiple columns contains symbol (**'#','@','+','&&&'**) which has no meaning in the record.so first we will replace with NaN but we can see except "Tenure" most of the columns contains literals at the same time so its better to delete those rows.Deletion would not be affect or imbalanced the dataset anyhow as we have thousands of records.

```
Churn_df.loc[Churn_df['Tenure'] == '#', 'Tenure'] = np.NaN
Churn_df.loc[Churn_df['Account_user_count']=='@','Account_user_count'] = np.NaN
Churn_df.loc[Churn_df['rev_per_month']=='+','rev_per_month'] = np.NaN
Churn_df.loc[Churn_df['Login_device']=='&&&&','Login_device'] = np.NaN
```

```
Churn_df.dropna(subset=['rev_per_month','Account_user_count','CC_Contacted_LY','Complain_ly','Day_Since_CC_connect','rev_growth_y
```

Image 1 : Null Analysis

It's better to drop null records for 'Account_user_count' feature because every account has multiple users associated with it and this is a defined set. We cannot impute directly by mean,mode,median or anything. It could be possible that one account is used by 1 person and there may be no other users associated with it.

Same with ["CC_Contacted_LY","Complain_ly","Day_Since_CC_connect"] It could be possible some users haven't contacted CC since the last 12 months as they wouldn't have faced any issue.

Now we can impute some numerical columns with mean as they represent feedback score and cashback which can be more or less in terms of numbers.

```
# Impute numerical column with average

numerical_col=['Service_Score','CC_Agent_Score','coupon_used_for_payment','cashback']
for col in numerical_col:
    Churn_df[col]=pd.to_numeric(Churn_df[col],errors='coerce')
    Churn_df[col].fillna(Churn_df[col].mean(),inplace=True)
```

Image 2: Imputation 1

Now we can see Payment,Login device and City_tier are categorical in nature and could be impute with most frequent value.It shouldn't impact further analysis.so we impute with mode

```
# Imputation with mode

Churn_df['Login_device'].fillna(Churn_df['Login_device'].mode()[0],inplace=True)
```

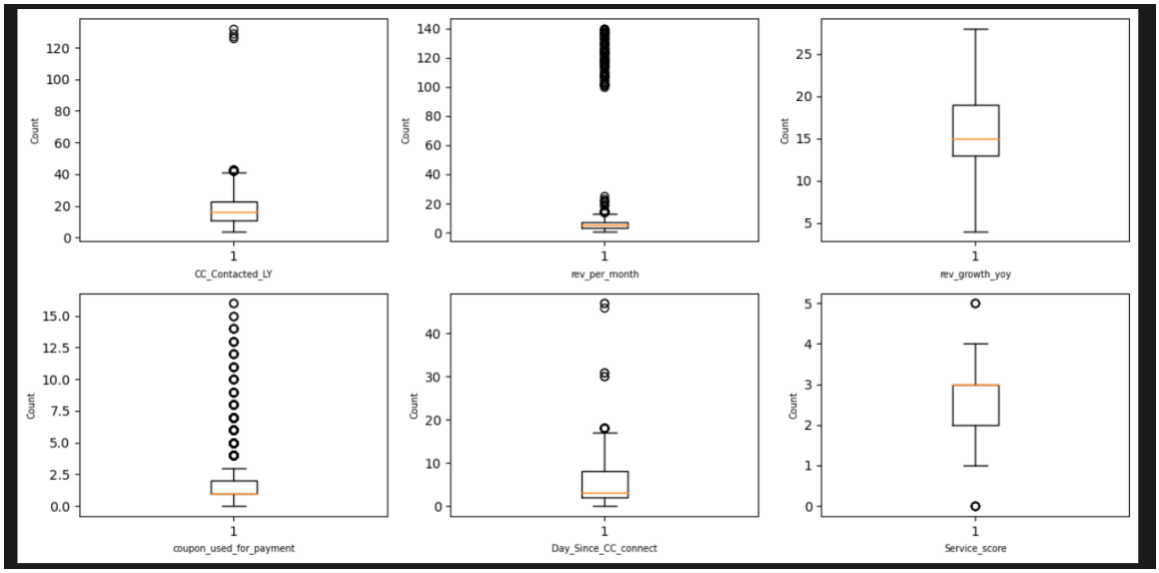
```
Churn_df['Payment'].fillna(Churn_df['Payment'].mode()[0],inplace= True)
```

```
Churn_df['City_Tier'].fillna(Churn_df['City_Tier'].mode()[0],inplace= True)
```

```
Churn_df['Marital_Status'] = Churn_df['Marital_Status'].fillna('Not Available')
Churn_df['account_segment'] = Churn_df['account_segment'].fillna('Not Available')
Churn_df['Gender'] = Churn_df['Gender'].fillna('Not Available')
```

Image 3: Outlier Detection

Post missing value treatment we tested for outliers in the data:

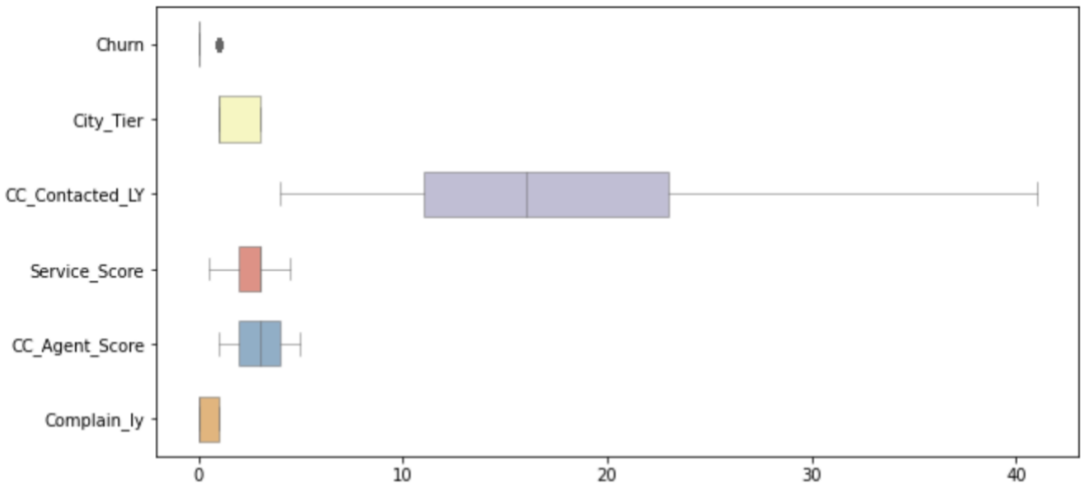


CC_Contacted_LY we can see pretty large number of times customer have contacted the care since last 12 months.It could be possible that they might have faced multiple issues in terms of services and all.

rev_per_month have lot of values which is out of the distribution means data is highly skewed for this feature.Huge variation shown in the coupon_used_for_payment.cashback has lot of values which out of the box. We will consider this facts to handle outlier as well.

Graph 3: Outliers Removal

Further we did outlier treatment for respective column:

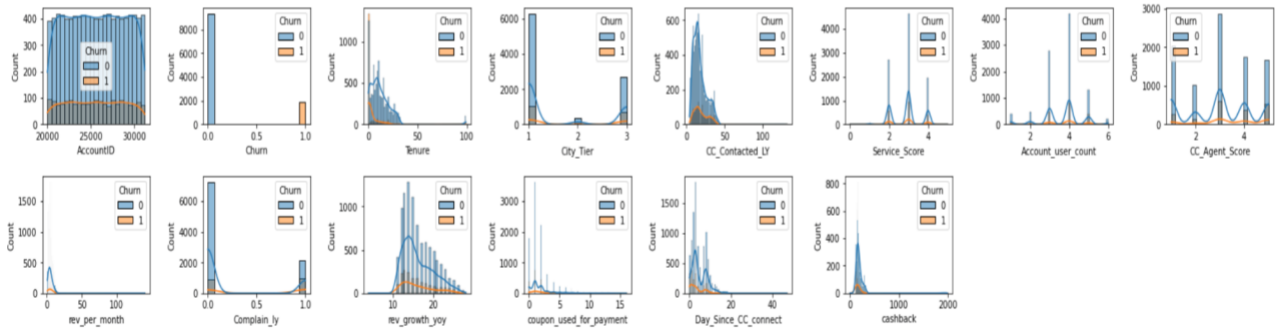


Graph 4: Univariate Analysis

Post outlier treatment, we identify relationship with churn as target variable:

```
# KDE plot highlighting Churn as the target variable
plt. figure(figsize=(24, 5))
for i in range(0, len(nums)):
    plt. subplot(2, 8, i+1)
    sns.histplot(data=df2, x=df2[nums[i]], hue="Churn", kde="True")
    plt. tight_layout(pad = 2)
```

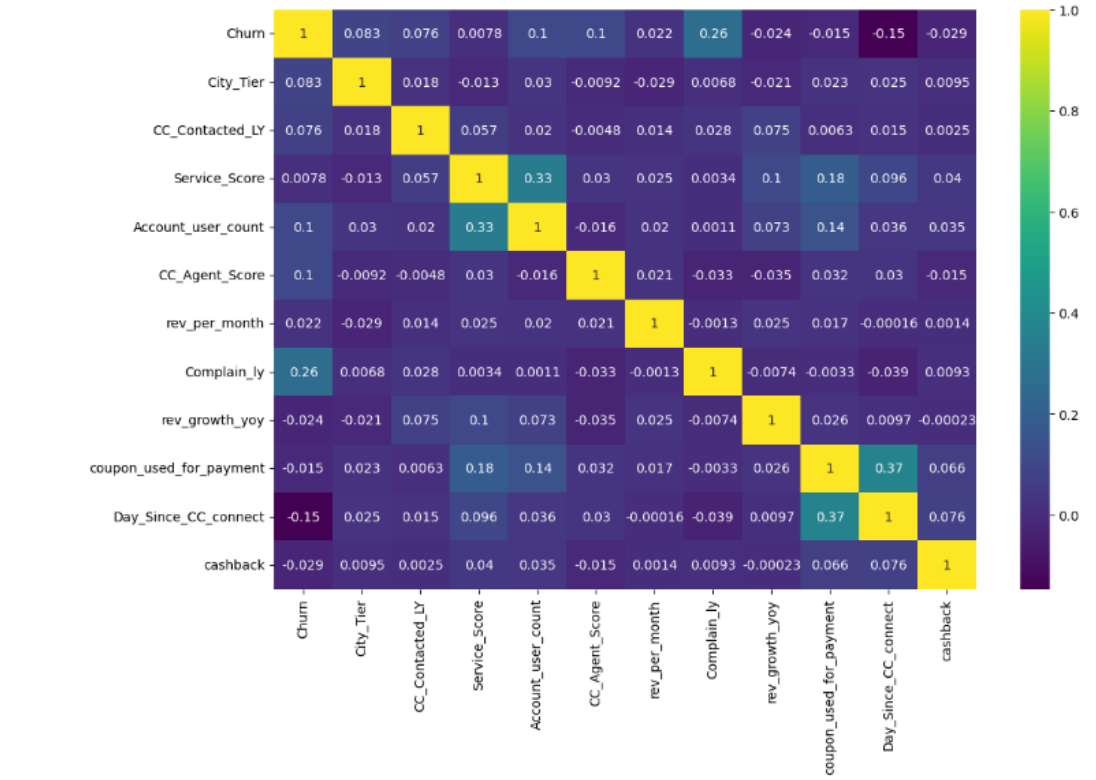
Python



Graph 5: Heat Map

Now we run a multivariate heatmap analysis to see the relationship:

```
In [50]: plt.figure(figsize=(12,8))
         sns.heatmap(Churn_df.corr(),cmap='viridis',annot=True)
Out[50]: <Axes: >
```



Univariate Analysis:

In all above study the key findings are:

- Customers from Tier 1 city have a major volume of not churned as compared to others.We need to draw our attention to Tier 2 cities because the ratio between churn and not churn is pretty close.
- The more who made payment through Debit card or credit card are not likely to be churned.In the COD(Cash on Delivery) or E wallet,Ratio between churners and not churners are close as compared to others.It means If COD or E-wallet happens customers are more likely to churn.We should recommend customers to purchase online or we can provide discounted offers to increase the engagement of online transaction.
- On the gender basis mens are not churned majorly because of the volume here but ratio seems the same for both around ~4. We cannot make conclusions here based on the gender standalone.
- An account who has 6 more users associated with it,they are more likely to churn but accounts who have 3-4 users are not churned majorly.

- If we talk about the account segment, we have "super" and regular plus" subscriber who do not churned majorly as compared to others.
- Customers who are married or couples take more subscriptions not churned majorly in terms of others segment.The rate of churners are high in Singles.
- Those who didn't complain last year haven't churned majorly.
- Observation clearly indicates that based on the tenure_period those customers who just signup or activated the account,Means customers who haven't completed months are likely to churn.Churners are high for 0 months.We need to target new customers.customers with a tenure of 0 are valuable as they represent new business opportunities and potential long-term subscribers. It is crucial for the service provider to ensure a smooth onboarding process, provide timely support, and offer attractive benefits to engage and retain these new customers.

Bivariate Analysis:

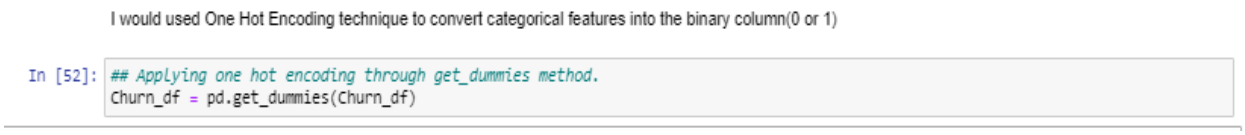
In all above study the key findings are:

- From the heatmap above, it can be obtained that there is a feature correlation with the target, where the target is churn.
- The feature that has a high correlation with the churn target is Tenure -0.23.
- This includes a negative correlation, meaning that the greater the tenure value, the lower the churn rate.
- In addition, the correlation of other features is quite large, namely between churn and complaint_ly of 0.25. Where this includes a positive correlation, meaning that the greater the value of the complaint, the greater the churn rate.
- Thus, features that are relevant and must be maintained are tenure and complain_ly features.

Feature Encoding

We used One Hot Encoding technique to convert categorical features into the binary column(0 or 1).

Image 4: Hot encoding



Feature Scaling

Feature scaling is a very important step to scale the data values in the same scale so that it will not impact the overall analysis because every column's value has a different unit.

```
In [55]: # Feature Scaling
         from sklearn.preprocessing import MinMaxScaler

         Scalr =  MinMaxScaler()

         columns = ['CC_Contacted_LY','rev_per_month','rev_growth_yoy','coupon_used_for_payment','cashback','Day_Since_CC_connect']
         Churn_df[columns]=Scalr.fit_transform(Churn_df[columns])
```

Image 5: Feature Scaling

Feature Selection

```
#VIF Dataframe
vif_df = pd.DataFrame()
vif_df["feature"] = X_train.columns
```

+ Code + Markdown

```
# Calculate VIF for each feature
vif_df["VIF"] = [variance_inflation_factor(X_train.values,i)
                          for i in range(len(X_train.columns))]
print(vif_df)
```

	feature	VIF
0	Tenure	2.406536
1	City_Tier	4.386839
2	CC_Contacted_LY	5.163291
3	Payment	3.388428
4	Gender	2.414068
5	Service_Score	18.029489
6	Account_user_count	6.763558
7	account_segment	5.007305
8	CC_Agent_Score	5.461601
9	Marital_Status	2.798227
10	rev_per_month	6.357670
11	Complain_ly	1.416356
12	rev_growth_yoy	3.750396
13	coupon_used_for_payment	4.061597
14	Day_Since_CC_connect	3.933805
15	cashback	2.557695
16	Login_device	7.962963

Image 6: Feature Selection

Dropping VIF>5

```
X_train = X_train.drop(["rev_per_month","Login_device","Service_Score","Account_user_count"], axis=1)
X_test = X_test.drop(["rev_per_month","Login_device","Service_Score","Account_user_count"], axis=1)
print('X_train',X_train.shape)
print('X_test',X_test.shape)
```

```
X_train (7882, 13)
X_test (3378, 13)
```

```
#VIF Dataframe
vif_df = pd.DataFrame()
vif_df["feature"] = X_train.columns
#
# Calculate VIF for each feature
vif_df["VIF"] = [variance_inflation_factor(X_train.values,i)
                          for i in range(len(X_train.columns))]
print(vif_df)
```

```
                    feature       VIF
0                    Tenure  2.349900
1                 City_Tier  4.263793
2            CC_Contacted_LY  4.805932
3                   Payment  3.341835
4                    Gender  2.348363
5            account_segment  4.640596
6             CC_Agent_Score  4.940548
7             Marital_Status  2.679771
8                Complain_ly  1.400877
9             rev_growth_yoy  3.545252
10  coupon_used_for_payment  3.681908
11     Day_Since_CC_connect  3.818128
12                 cashback  2.456347
```

Image 7: VIF

Now training the model:

```
col = X_train.columns
X_train_sm = sm.add_constant(X_train[col])
logm2 = sm.Logit(train_labels.astype(float),X_train_sm.astype(float), family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Image 8: Model Training

Result:

Logit Regression Results

Dep. Variable:	Churn	**No. Observations:**	7882
Model:	Logit	**Df Residuals:**	7868
Method:	MLE	**Df Model:**	13
Date:	Sun, 16 Jul 2023	**Pseudo R-squ.:**	0.2513
Time:	18:16:34	**Log-Likelihood:**	-2673.7
converged:	True	**LL-Null:**	-3571.1
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6288	0.210	-12.499	0.000	-3.041	-2.217
Tenure	-0.0762	0.004	-20.478	0.000	-0.084	-0.069
City_Tier	0.3470	0.039	8.973	0.000	0.271	0.423
CC_Contacted_LY	0.0292	0.004	7.174	0.000	0.021	0.037
Payment	-0.0584	0.029	-1.998	0.046	-0.116	-0.001
Gender	0.2763	0.070	3.940	0.000	0.139	0.414
account_segment	-0.1540	0.021	-7.306	0.000	-0.195	-0.113
CC_Agent_Score	0.2711	0.026	10.502	0.000	0.221	0.322
Marital_Status	0.4003	0.032	12.572	0.000	0.338	0.463
Complain_ly	1.5346	0.074	20.871	0.000	1.390	1.679
rev_growth_yoy	-0.0136	0.009	-1.440	0.150	-0.032	0.005
coupon_used_for_payment	0.0265	0.008	3.411	0.001	0.011	0.042
Day_Since_CC_connect	-0.0450	0.005	-9.728	0.000	-0.054	-0.036
cashback	-0.0042	0.000	-8.806	0.000	-0.005	-0.003

Table 5: Model Result

2nd Iteration of training:

Dropping the features having 'p-value' > 0.05

```
X_train = X_train.drop(['rev_growth_yoy'],axis=1)
X_test = X_test.drop(['rev_growth_yoy'],axis=1)
```

Iteration-2

```
col = X_train.columns
X_train_sm = sm.add_constant(X_train[col])
logm2 = sm.Logit(train_labels.astype(float),X_train_sm.astype(float), family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Image 9: Training 2

Result:

Logit Regression Results

Dep. Variable:	Churn	**No. Observations:**	7882
Model:	Logit	**Df Residuals:**	7869
Method:	MLE	**Df Model:**	12
Date:	Sun, 16 Jul 2023	**Pseudo R-squ.:**	0.2510
Time:	18:52:47	**Log-Likelihood:**	-2674.7
converged:	True	**LL-Null:**	-3571.1
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7056	0.204	-13.279	0.000	-3.105	-2.306
Tenure	-0.0761	0.004	-20.452	0.000	-0.083	-0.069
City_Tier	0.3484	0.039	9.016	0.000	0.273	0.424
CC_Contacted_LY	0.0288	0.004	7.087	0.000	0.021	0.037
Payment	-0.0584	0.029	-1.998	0.046	-0.116	-0.001
Gender	0.2792	0.070	3.985	0.000	0.142	0.417
account_segment	-0.1539	0.021	-7.297	0.000	-0.195	-0.113
CC_Agent_Score	0.2718	0.026	10.534	0.000	0.221	0.322
Marital_Status	0.3995	0.032	12.548	0.000	0.337	0.462
Complain_ly	1.5337	0.074	20.865	0.000	1.390	1.678
coupon_used_for_payment	0.0257	0.008	3.318	0.001	0.011	0.041
Day_Since_CC_connect	-0.0450	0.005	-9.720	0.000	-0.054	-0.036
cashback	-0.0042	0.000	-8.801	0.000	-0.005	-0.003

Table 6: Training Result 2

Recursive feature elimination:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
#
chi2_features = SelectKBest(score_func = chi2, k = "all")
X_train_kbest = chi2_features.fit(X_train, train_labels)
#
X_train_scores = pd.DataFrame(X_train_kbest.scores_,columns=["Score"])
X_train_columns = pd.DataFrame(X_train.columns)
#
X_train_features_rank = pd.concat([X_train_columns,X_train_scores],axis=1)
#
X_train_features_rank.columns = ['Features','Score']
X_train_features_rank
```

	Features	Score
0	Tenure	7017.771325
1	City_Tier	25.047843
2	CC_Contacted_LY	176.852914
3	Payment	0.017068
4	Gender	2.033894
5	account_segment	16.054867
6	CC_Agent_Score	51.596115
7	Marital_Status	172.396081
8	Complain_ly	330.026546
9	coupon_used_for_payment	8.320010
10	Day_Since_CC_connect	623.096140
11	cashback	8807.235499

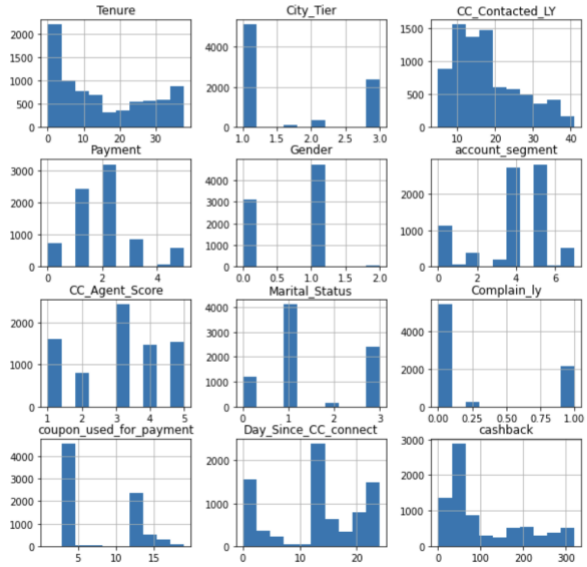
Image 10: Recursive Feature

Univariate analysis diagram based on significant feature

Univariate Analysis Diagram (based on significant features)

- It can shows the variability and distribution of the values of each features with respect to the timelines.

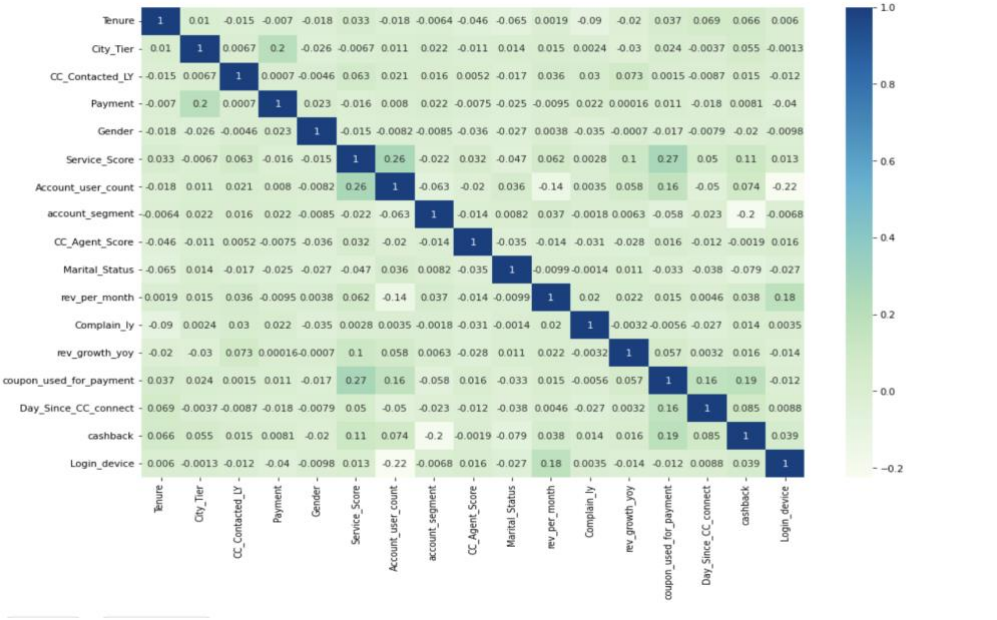
```
X_train.hist(bins=10, figsize=(10,10))
plt.show()
```



Graph 7 : Univariate Analysis

Bivariate analysis based on significant features:

```
f, ax = plt.subplots(figsize=(15,10))
ax = sns.heatmap(df.corr(), cmap = 'GnBu', annot = True)
```



Graph 8: Bivariate Analysis

Testing and Deploying Model

We have prepared the following model procedures to analyze and review the dataset and get the performance and importance of the features available on the dataset which can gather more information about the subjects.

Following is the list of model building procedure can used in this project :

Models

The subsequent phase involves a series of sequential actions, commencing with the division of the dataset into training and testing subsets. Subsequently, three distinct machine learning models will be constructed to facilitate a comparative evaluation of their respective accuracies. A more detailed overview of the chosen algorithms is provided below:

Decision Tree:

Decision trees represent a type of unsupervised machine learning algorithms that find utility in both classification and regression tasks. Their fundamental objective is to generate a predictive model by learning uncomplicated decision rules inferred from the attributes present in the data (Decision Trees, 2022).

Logistic Regression:

Logistic regression entails the modeling of the probability associated with discrete outcomes, primarily applied in scenarios involving binary classification. This technique finds widespread employment in classification tasks, particularly when the objective is to ascertain whether a sample appropriately belongs to a particular class. It holds a prominent position as one of the fundamental analytical algorithms (Thomas W. Edgarm, 2017).

Random Forest:

Random forest stands as a machine learning algorithm that amalgamates the outputs of multiple decision trees into a consolidated outcome. This approach mitigates concerns such as overfitting and bias that are commonly associated with individual decision trees. Notably, random forest yields accurate predictions, particularly when the constituent trees exhibit minimal correlation among themselves (Random Forest, 2020).

Logistics Regression

(a) Model Prediction

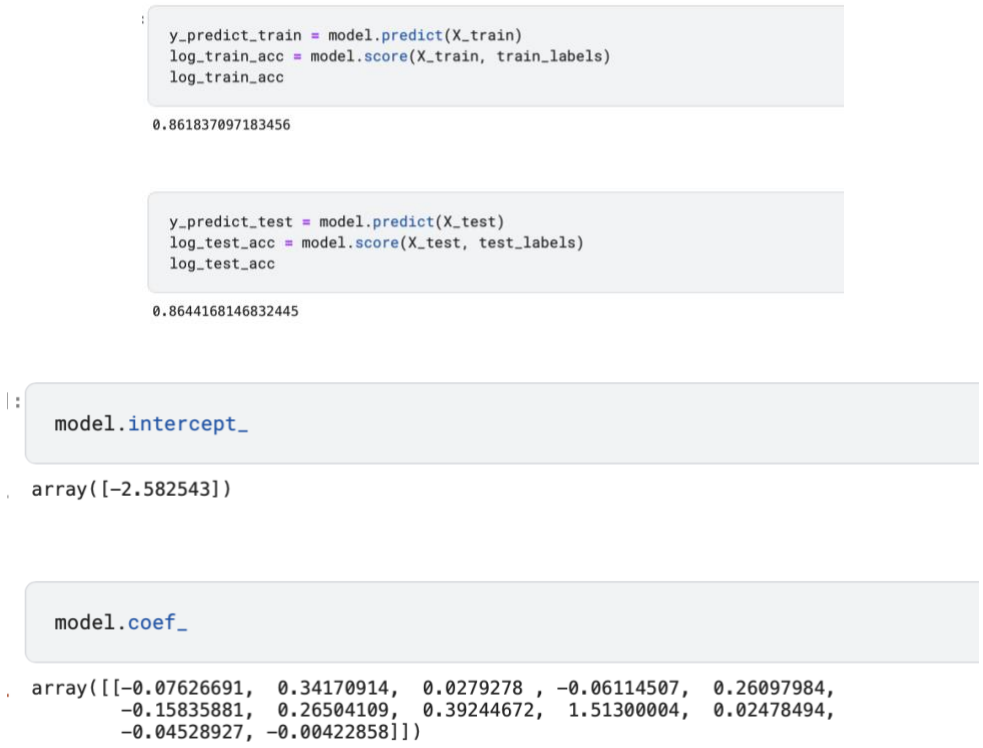


Image 11 : Logistic Regression Model

(b) Model Performance:

```
confusion_matrix(train_labels, y_predict_train)
```

```
array([[6355,  201],
       [ 888,  438]])
```

```
print(classification_report(train_labels, y_predict_train))
```

```
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      6556
           1       0.69      0.33      0.45      1326

    accuracy                           0.86      7882
   macro avg       0.78      0.65      0.68      7882
weighted avg       0.85      0.86      0.84      7882
```

```
confusion_matrix(test_labels, y_predict_test)
```

```
array([[2711,   97],
       [ 361,  209]])
```

+ Code + Markdown

```
print(classification_report(test_labels, y_predict_test))
```

```
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      2808
           1       0.68      0.37      0.48       570

    accuracy                           0.86      3378
   macro avg       0.78      0.67      0.70      3378
weighted avg       0.85      0.86      0.85      3378
```

Image 12 : Logistic Regression Model Performance

(c) ROC-AUC Graph

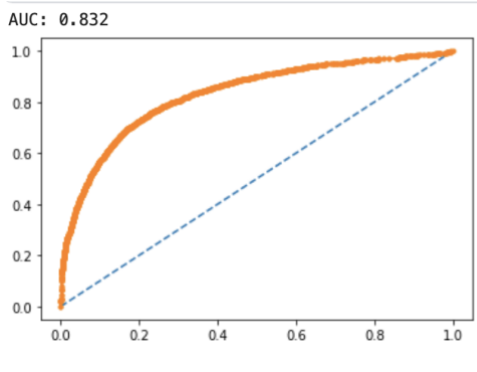
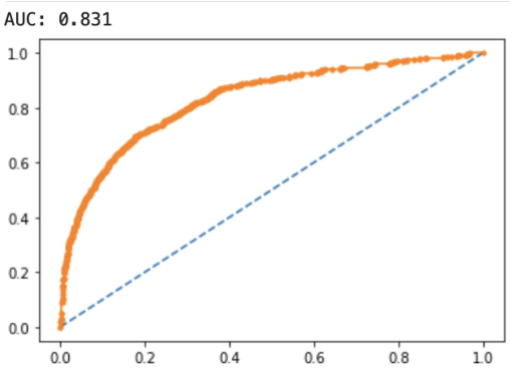
Training Data	Test Data
	

Table 7 : Logistic Regression Model AUC

(d) Model Performance Metrics:

Training Data	Test Data
log_train_precision 0.69 log_train_recall 0.33 log_train_f1 0.45	log_test_precision 0.68 log_test_recall 0.37 log_test_f1 0.48

Table 8 : Logistic Regression Model Performance

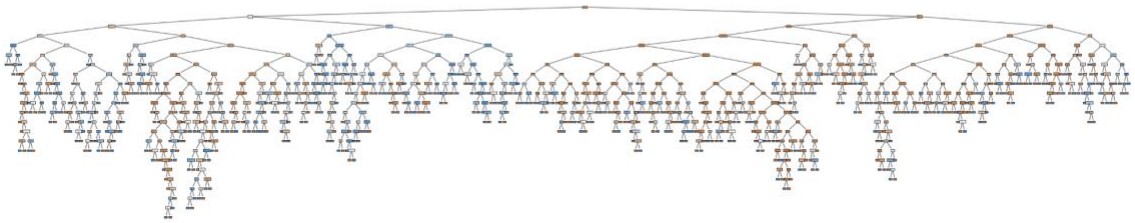
Decision Tree

(a) Model Prediction

```
gs_dtcl.best_score_
```

```
0.8963461916058572
```

```
dot_data1 = tree.export_graphviz(dtcl, out_file=None , feature_names = list(X_train))
fig = plt.figure(figsize=(50,10))
_ = tree.plot_tree(dtcl,
                   feature_names=list(X_train),
                   filled=True)
```



+ Code + Markdown

```
dt_imp = pd.DataFrame(dtcl.feature_importances_, columns = ["Imp"], index = X_train.columns)
print(dt_imp)
```

```
                          Imp
Tenure                0.315236
City_Tier             0.045080
CC_Contacted_LY       0.089020
Payment               0.074856
Gender                0.025927
account_segment       0.048399
CC_Agent_Score        0.086387
Marital_Status        0.046451
Complain_ly           0.062495
coupon_used_for_payment  0.030827
Day_Since_CC_connect  0.082239
cashback              0.093082
```

Image 13 : Decision Tree Model

Training Data	Test Data

	0	1
0	0.978814	0.021186
1	0.984496	0.015504
2	0.920000	0.080000
3	0.984496	0.015504
4	0.352941	0.647059

	0	1
0	0.000000	1.000000
1	0.285714	0.714286
2	0.918367	0.081633
3	1.000000	0.000000
4	0.984496	0.015504

Table 9 : Decision tree Model Result

(b) Model Performance

Training Data	Test Data

	precision	recall	f1-score	support
0	0.91	0.96	0.94	6556
1	0.72	0.56	0.63	1326
accuracy			0.89	7882
macro avg	0.82	0.76	0.78	7882
weighted avg	0.88	0.89	0.88	7882

	precision	recall	f1-score	support
0	0.91	0.95	0.93	2808
1	0.69	0.56	0.62	570
accuracy			0.88	3378
macro avg	0.80	0.76	0.78	3378
weighted avg	0.88	0.88	0.88	3378

Table 10 : Decision tree Model Performance

(c) ROC-AUC Graph

Training Data	Test Data

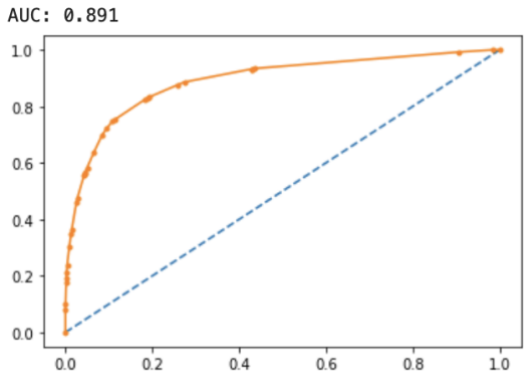
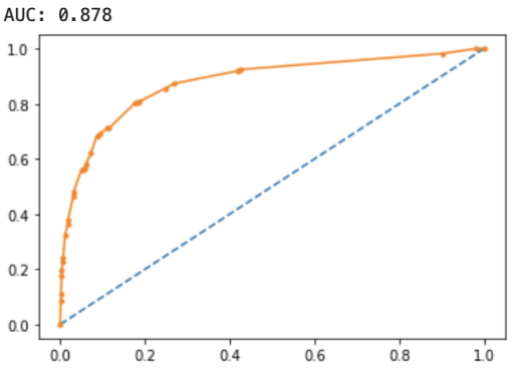
	

Table 11 : Decision tree Model AUC

(d) Model Performance Metrics

Training Data	Test Data
cart_train_precision 0.72 cart_train_recall 0.56 cart_train_f1 0.63	cart_test_precision 0.69 cart_test_recall 0.56 cart_test_f1 0.62

Table 12 : Decision tree Model Performance

Random Forest

(a) Model Prediction

Training Data	Test Data

	0	1
0	0.991965	0.008035
1	0.993107	0.006893
2	0.868914	0.131086
3	0.987959	0.012041
4	0.524457	0.475543

	0	1
0	0.072611	0.927389
1	0.320965	0.679035
2	0.965016	0.034984
3	0.573265	0.426735
4	0.968302	0.031698

Table 13 : Random Forest Model Prediction

(b) Model Performance

Training Data

	precision	recall	f1-score	support
0	0.93	0.97	0.95	6556
1	0.83	0.66	0.74	1326
accuracy			0.92	7882
macro avg	0.88	0.82	0.84	7882
weighted avg	0.92	0.92	0.92	7882

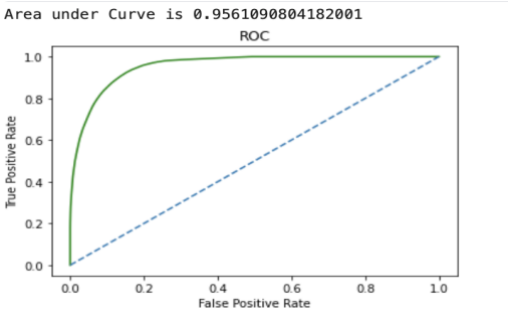
Test Data

	precision	recall	f1-score	support
0	0.93	0.97	0.95	2808
1	0.80	0.63	0.71	570
accuracy			0.91	3378
macro avg	0.86	0.80	0.83	3378
weighted avg	0.91	0.91	0.91	3378

Table 14 : Random Forest Model Result

(c) ROC-AUC Graph

Training Data



Test Data

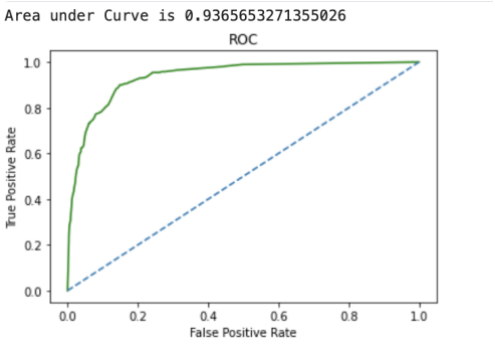


Table 15 : Random Forest Model AUC

(d) Model Performance Metrics

Training Data	Test Data
rf_train_precision 0.83 rf_train_recall 0.66 rf_train_f1 0.74	rf_test_precision 0.8 rf_test_recall 0.63 rf_test_f1 0.71

Table 16 : Random Forest Model Performance

Model Accuracy Comparison

Training Data

Train	Decision Tree	Random Forest	Logistic Regression
Accuracy	0.91	0.91	0.86
AUC	0.89	0.96	0.83
Recall	0.56	0.66	0.33
Precision	0.72	0.83	0.69
F1 Score	0.63	0.74	0.45

Table 17 : Model Accuracy on Training Data

Test Data

Test	Decision Tree	Random Forest	Logistic Regression
Accuracy	0.9	0.9	0.86
AUC	0.88	0.94	0.83
Recall	0.56	0.63	0.37
Precision	0.69	0.8	0.68
F1 Score	0.62	0.71	0.48

Table 17 : Model Accuracy on Test Data

DT	RF	LR

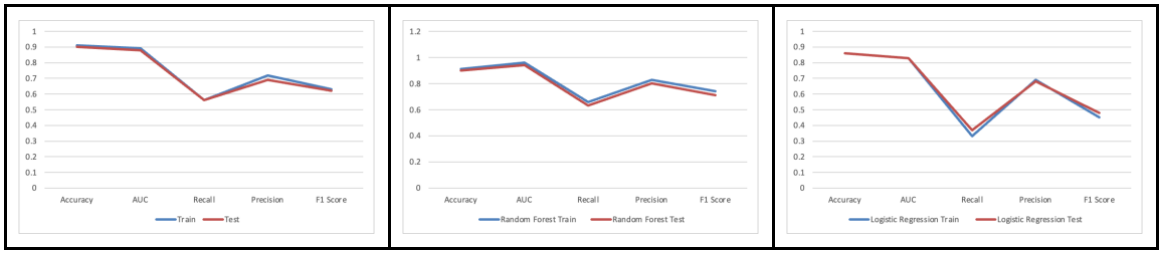


Table 17 : Model Accuracy Chart

Model Selection Reason:-

To avoid false positive churn, we decided to choose a model with the smallest ROC-AUC gap and the highest precision.

From several model test results, Random Forest is the most suitable algorithm model.

Model Validation:

We have validated the model(**Random Forest Classifier**) by using cross validation(**K-Fold)** before tuning the model because Model validation is a critical step in the process of building and training machine learning models. It involves assessing the performance of a trained model on data that it has never seen before.

Model validation is crucial for tuning hyperparameters. Hyperparameters are settings that are not learned from the data, such as the learning rate or the number of layers in a neural network.

Outcome:

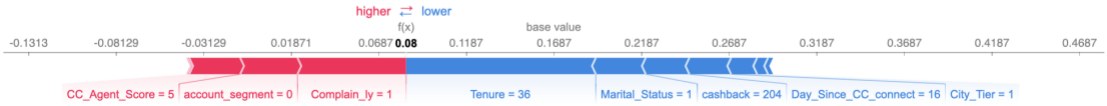
After k-fold cross validation, the model is best-fit, as can be seen from the train scores and test scores, which have very small differences.

Feature Importance:

```
import shap  # package used to calculate Shap values
row_to_show = 5
data_for_prediction = X_train.iloc[row_to_show]
explainer = shap.TreeExplainer(best_grid_rf) # Create object that can calculate shap values
shap_values = explainer.shap_values(data_for_prediction) # Calculate Shap values
#
shap.initjs()
shap.force_plot(explainer.expected_value[1], shap_values[1], data_for_prediction)
```

+ Code + Markdown

```
# use Kernel SHAP to explain test set predictions
k_explainer = shap.KernelExplainer(best_grid_rf.predict_proba, X_train)
k_shap_values = k_explainer.shap_values(data_for_prediction)
shap.force_plot(k_explainer.expected_value[1], k_shap_values[1], data_for_prediction)
```



Graph 9 : SHAP Result

CHAPTER 4

FINDINGS, RECOMMENDATIONS AND CONCLUSION

4.1 Findings Based on Observations

From a Direct-to-Home (DTH) perspective, customer churn can be defined as the phenomenon where initial patrons of a DTH enterprise discontinue purchasing its goods or availing its services and instead opt for the services provided by competitors (Wu et al., 2017). Churn rate prediction holds significant importance, particularly within the telecommunication sector. In the context of DTH services, customer churn takes on a specific form, signifying customers' departure from the enterprise, its products, or services due to factors like subpar quality or delays in service delivery. This type of churn operates within a non-contractual relationship, which poses challenges for businesses in proactively detecting such instances (Shao, 2016).

Hence when we analyzed data, against churn for multiple variables, some observations are made, which are given below:

1. Lower mean and median Cashback.
1. Higher mean CityTier, but equal median CityTier to non-churned customers.
2. Higher mean and median Complain.
3. Lower mean CouponUsed, but equal median.
4. Higher mean Login_device, but equal median.
5. Greatly lower mean and median Tenure.

4.2 Findings Based on analysis of Data

Post observations we made series on analysis in chapter 3 of this documents where It is observed that churned customers are associated with:

1. Male gender.
2. Single marital status.
3. Mobile is preferred login device
4. Cash on Delivery is preferred payment mode.
5. Approximately 54% of customers exhibit a tenure of 1 year, indicating either active service usage (0) or discontinued services (1).
6. 11% of customers possess a tenure of 0, indicative of recent subscription initiation or DTH service enrollment.
7. A tenure of 0 suggests new subscribers who may have just installed equipment and activated their subscription (0) or potentially canceled their subscription shortly after joining (1).

4.3 General findings

The objective here is to distinguishing between churned and retained customers while identifying associated attributes for churn. The general findings are:

1. Slightly higher probability of churn among single male customers.
2. Customer churn linked to the "Mobile preferred" order category.
3. Churned customers more likely to use phone/mobile phone for login (potential user experience influence).
4. Higher complain rate, city tier, number of addresses, and registered devices for churned customers.
5. Surprisingly, churned customers exhibit higher satisfaction scores.
6. Lower tenure and count of orders for churned customers, which is expected.

4.4. Recommendation based on findings

Based on above finding some of the recommendations are:

1. Customers with a tenure of 0 signify promising business prospects and potential long-term subscribers.
2. Ensuring a seamless onboarding process, offering timely support, and presenting enticing incentives becomes pivotal to engage and retain these new customers.
3. In order to foster customer loyalty and increase customer tenure, our DTH company should consider implementing loyalty programs or special pricing offers exclusively for our loyal customers.
4. Complaint management should be treated with utmost care, as it ranks second in terms of importance. It is imperative that our organization ensures our customer service team is well-trained and equipped to handle complaints professionally and effectively.
5. To eliminate the root causes of complaints, we need to focus on enhancing the overall customer experience. Conducting regular surveys to gather customer feedback will provide valuable insights and help us make necessary improvements.
6. By conducting A/B testing, we can optimize the user experience and user interface of our DTH platform. This, in turn, will have a positive impact on the conversion rate, ultimately leading to higher customer satisfaction and engagement.

4.5 Suggestions for areas of improvement

As this project is on data analysis in model building area of improvement also revolve around data only. There were some limitations which might hindered the project, few points are listed below:

1. Lack of similar previous projects as most of customer churn projects are directed towards the telecommunication/ecommerce sector.
2. Difficulty to get a rich dataset.

Hence the primary area of improvement would be:

1. Continuous collection of data
2. Enriching demographic and geographic data
3. As we have noticed mobile app is important factor, application usage and other app behaviour data will be a good addition.

4.6 Scope for future research

In our forthcoming initiatives, we envision the creation of a real-time analysis framework tailored for a DTH platform, coupled with seamless integration into mail marketing software. This strategic amalgamation holds the potential to revolutionize customer engagement and retention efforts. By harnessing this synergy, businesses can dynamically automate personalized offers to specifically target individuals showing signs of churn. The anticipated outcome is a substantial reduction in customer attrition rates, underpinning heightened customer satisfaction and loyalty within the realm of DTH services. Furthermore, we can also take initiative to delving deeper into the behavioral intricacies of our retained customers. This exploration will encompass an exhaustive examination of their preferences, inclinations, and favored product categories. Such an in-depth study of retained customer behavior is poised to yield invaluable insights, profoundly impacting the company's revenue streams. By uncovering the underlying drivers behind the preferences of our steadfast patrons, we can meticulously tailor our offerings, optimize resource allocation, and further bolster competitive advantage within the DTH industry.

4.7 Conclusion

In the realm of Direct-to-Home (DTH) services, the focus has been on optimizing customer acquisition strategies, often involving substantial investments. However, the longevity of customer relationships within the DTH sector is influenced by a multitude of variables. This project, carried out under the auspices of JGI Institution, was dedicated to the construction of a robust customer churn prediction model tailored to the unique landscape of DTH services.

The dataset utilized in this project is provided by JGI Institution, with no details of customer. Our project commenced with an in-depth exploratory analysis and the creation of data visualizations, a crucial step that enhanced our comprehension of customer churn dynamics specific to the DTH context. Notably, discernible patterns emerged, revealing an association between churned customers and characteristics such as male gender and single marital status.

The heart of our study revolved around the application of three distinct machine learning algorithms – Decision Tree, Logistic Regression, and Random Forest. These methodologies were employed to prognosticate customer churn within the DTH domain. Through rigorous experimentation and evaluation, we observed that the Random Forest algorithm demonstrated the

most promising outcomes. Notably, it exhibited a remarkable accuracy rate of 93.5%, complemented by a robust kappa score of 0.75. This highlights its potential efficacy in forecasting customer churn and subsequently empowering DTH service providers with actionable insights to bolster customer retention efforts.

References

1. Zhang, D. (2015). Establishment and application of customer churn prediction model. Beijing Institute of Technology.
2. Wu, X. J., & Meng, S. S. (2017). Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost. Industrial Engineering, 20(02), 99- 107.
3. Shao, D. (2016). Analysis and prediction of insurance company's customer lossbased on BP neural network. Lanzhou University
4. Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp, 79, 403-408.
5. Feng, X., Wang, C., Liu, Y., Yang, Y., & An, H. G. (2018). Research on customer churn prediction based on comment emotional tendency and neural network. Journal of China Academy of Electronics Science, 13(03), 340-345
6. Decision Trees. (2022, 04 22). Retrieved from Sickit learn: https://scikit-learn.org/stable/modules/tree.html#:~:text=Decision%20Trees%20(DTs)%20are%20 a,as%20a%20piecewise%20constant%20approximation.
7. Thomas W. Edgarm, D. O. (2017). Research Methods of Cyber Security.
8. Random Forest. (2020, December 7). Retrieved from IBM: https://www.ibm.com/cloud/learn/random- forest

—---------- END OF DOCUMENT —----------