Customer Churn Prediction Analysis

Submitted in Partial Fulfilment of requirements for the Award of certificate of

Post Graduate Program in Data Science and Business Analytics

Capstone Project Report

Submitted to

Submitted by:

1. Steffin John
2. Vijayaprabakaran L

Under the guidance of

Udaya Kumar Devaraj

Batch: 2022-23

Year of Completion: 2023

CERTIFICATE OF COMPLETION SIGNED BY MENTOR

This is to certify that the participants **Steffin John and Vijayaprabakaran L** who are the students of Great Learning, have successfully completed their project on **Customer Churn Prediction**

This project is the record of authentic work carried out by them during the academic year 2022 - 2023

Mentor **Udhaya Kumar Devaraj**

Date: 05-03-2023

Place:

udayakumardevaraj@gmail.com

ACKNOWLEDGEMENTS

On behalf of Group 7 Members Mr. Steffin John and Mr Vijayaprabakaran L., We would like to extend our heartfelt gratitude to Mr. Udhaya Kumar Devaraj for his mentorship and guidance throughout our capstone project. Your expertise, support, and encouragement have been invaluable to us, and we could not have achieved our goals without your help.

We are also grateful to Great Learning for providing us with the necessary resources and support to carry out our research. The experience has taught us valuable skills in data analysis, statistical modeling, and machine learning, and we are confident that these skills will serve us well in our future endeavors.

Our Program Head Mr. Hemant Verma has also been an integral part of our success, and we have learned so much from you. Your insights, feedback, and advice have helped us stay focused and motivated, and we are grateful for the time and effort you invested in us.

We hope that this project will be a stepping stone for us to make a positive impact in the field of Data Science, and we promise to uphold the values and standards of excellence that you have instilled in us.

Once again, thank you for your unwavering support and guidance, and we look forward to keeping in touch with you.

Contents

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

1. Problem Statement

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

You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.

Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation. Hence be very careful while providing campaign recommendation.

2. Data Description

Below is the list of 19 variables/columns used for modeling as per provided by company:

AccountID	Payment	CC_Agent_Score	coupon_used_l12m
Churn	Gender	Marital_Status	Day_Since_CC_connect
Tenure	Service_Score	rev_per_month	cashback_l12m
City_Tier	Account_user_count	Complain_l12m	Login_device
CC_Contacted_L12m	account_segment	rev_growth_yoy	

For the modeling we have create 3 new additional variables/columns those are:

Noise Index	cashback_percen	Cashback_Benefits

***Note**: Check appendix (1) for full Data Description of each column*

3. Main Results

Out of 9 machine learning algorithm Extra Tree classifier turn out to be the most balanced. After exploring the data company must focus on some important variables like Complain_ly, Cashback , Nosie_Index, Account_segment, Cashback_Benefits, Account_Segment, Day_since_cc_connect, Service_Score.

4. Recommendations

Below are the list of recommendations the company can look into which may help in their campaigns:

a) Create Personalized Video from the company for New Customers login like welcome video, thank you video and what's upcoming.
b) Finding out the top 3 requirements of the customers during the set-up or installation or product
c) Personalized package or bundles of programs using the top 3 requirements and also using Artificial Intelligence (AI) like Netflix use to give personalised suggestions.
d) Aggressive segmentation of the customers as each and every customer is unique
e) High Quality Customer Support Training to the Agents
f) Using of Net Promoter Scores (NPS) regularly as Core just like HDFC uses to keep a track on customers.
g) Creating DIY videos which will educate customers about the product
h) Creating Customer Experience Videos and also giving some awards to long term customers which should also been shown in the website. For eg: Youtube give its creator Silver, Golden and Dimond Play Button upon crossing a threshold of subscriber for each
i) Customers support must be tech savvy as Gen-Z customers does not like call they prefer they problems to solved either on Whatsapp or Live Chat
j) Other metrics like First Contact Resolution (FCR) metric which calculates the percentage of cases solved by agents on 1st attempt and Customer satisfaction Index can be used
k) Creating of Loyal Programs which should be gamified and referral programs will keep the customers from churning
l) Excusive content for the customers.

Section 1: Introduction

Need of the Study:

With e-commerce, it turns out that acquiring new clients is far more expensive for a business than keeping its existing clientele. For this particular reason, it is essential for businesses to know in advance which customers will leave so that they can make offers or reduce consumption of their goods or services in a way that is relevant to increase customer retention, foster positive customer relationships, and reduce acquisition costs. There are a huge number of items and services available in today's cutthroat market. As a result, the majority of consumers have grown accustomed to freely moving between suppliers and brands in search of the good or service that best matches their need. E-commerce Customer "churn" has been causing businesses problems for some time. Hence, instead than concentrating on keeping their current clients, they frequently expend a great deal of work and money on luring new ones.

In response to the above problem, existing research on customer churn mainly includes predictions related to traditional statistics, artificial intelligence, predictions based on statistical learning theory, predictions based on combined classifiers. This paper proposes research on e-commerce customer churn prediction based on customer segmentation, using improved SMOTE for data balance, and then using 9 different machine learning algorithms for prediction out of which top 3 models are been narrowed down out of which final model *known as Extra Tree Classifier* is been selected gives the best result. Finally, predictors importance is identified to help decision makers in choosing the proper decisions on behalf of the organization

Objectives:

The study is critical because churn prediction and analysis will let e-commerce businesses know which customers are likely to migrate and when. In order to implement the appropriate retention steps to lessen or prevent their migration, it will be helpful to understand the true worth of the prospective loss of such clients, as predicted in e-commerce. The objective of this case study are as follows:

- Prediction if customer churn using Machine learning Models
- Reduce Churn Rate,
- Find out why customers are churning
- Best fit Machine Learning model with a high F1-score
- Give some recommendation how to reduce churning of customers

Reference to a few Articles:

- How to keep OTT and DTH or Cable TV customers engaged in these uncertain times (https://brandequity.economictimes.indiatimes.com/news/media/how-to-keep-ott-and-dth-or-cable-tv-customers-engaged-in-these-uncertain-times/75770804)\
- Breaking the Back of Customer Churn (https://www.bain.com/insights/breaking-the-back-of-customer-churn/)

Section 2: Literature Review

Establishment of E-Commerce Customer Churn Prediction Model

Customer churn refers to the rate at which customers stop doing business with a company or stop using its products or services. In the case of a telecom company, this could refer to customers switching to a different service provider, while in e-commerce, it could refer to customers no longer making purchases from a particular online store.

For a telecom company, customer churn can have a significant impact on revenue, as losing customers means lost monthly subscription fees. Additionally, acquiring new customers can be expensive, so retaining existing customers can be more cost-effective. Analyzing customer churn can help a telecom company identify the reasons why customers are leaving, such as poor network coverage or high prices, and take corrective action to improve the customer experience and retain more customers.

In the case of an e-commerce company, customer churn can also have a significant impact on revenue, as customers who don't return to the site to make additional purchases can reduce overall sales. Analyzing customer churn can help an e-commerce company identify the reasons why customers aren't returning, such as slow shipping times, poor customer service, or a confusing website interface, and take corrective action to improve the customer experience and encourage repeat purchases.

Extra Tree Classier:
The Extra Trees Classifier is a type of computer program that can be trained to predict if something belongs to a certain category or not. It works by creating many different "decision trees" that each look at different aspects of the data. Each tree decides if something belongs to the category based on the information it is looking at. Then, the program combines the results of all the trees to make a final prediction. The Extra Trees Classifier is good at handling messy data and not getting too attached to one particular way of looking at the data, which can make it more accurate.

XGBoost:
XGBoost is a computer program that predicts things by combining many simple models to make a more accurate prediction. It's good at handling large amounts of data and finding important patterns.

KNN (K-Nearest Neighbors):
KNN (K-Nearest Neighbors) is a machine learning algorithm that can be used for classification and regression tasks. It works by finding the K nearest data points to a new data point and assigning the class or value of the majority of those K data points to the new data point.

SHAP (SHapley Additive exPlanations):

SHAP (SHapley Additive exPlanations) values are a way of explaining the output of a machine learning model in terms of how much each feature contributes to the final prediction. For example, if a model is used to predict the price of a house based on its size, location, and number of bedrooms, SHAP values can be used to show how much each of these features affects the predicted price. SHAP values take into account the interactions between features and the overall impact of each feature on the prediction. They can be used to gain insights into how the model works and to identify which features are most important for making accurate predictions.

Sections 3: EDA and Insights

3.1 Exploratory Data Analysis(EDA)

The first step in data exploration is to import several libraries in python to explore and visualize the data. Then the numerical and categorical columns will be explored in addition to identification of missing data. The outcome of our data set is Churn, and there are no missing values in "churn" column. However, the outcomes variables are imbalanced due to the high number of retained customers in comparison to churned customers as shown in the table below. Where 0 = not churned and 1 = churned.

0	8934
1	1808

Average Overall Churn Rate: 16.838365896980463

The data type of each column is been observed below:

#	Column	Non-Null Count	Dtype
0	churn	10742 non-null	int64
1	tenure	10742 non-null	float64
2	city_tier	10742 non-null	float64
3	cc_contacted_ly	10742 non-null	float64
4	payment	10742 non-null	object
5	gender	10742 non-null	object
6	service_score	10742 non-null	float64
7	account_user_count	10742 non-null	float64
8	account_segment	10742 non-null	object
9	cc_agent_score	10742 non-null	float64
10	marital_status	10742 non-null	object
11	rev_per_month	10742 non-null	float64
12	complain_ly	10742 non-null	float64
13	rev_growth_yoy	10742 non-null	float64
14	coupon_used_for_payment	10742 non-null	float64
15	day_since_cc_connect	10742 non-null	float64
16	cashback	10742 non-null	float64
17	login_device	10742 non-null	object
18	tenure_bin	10742 non-null	category
19	cc_contacted_ly_bin	10741 non-null	category
20	noise_index	10742 non-null	object
21	rev_per_month_bin	10738 non-null	category
22	coupon_used_for_payment_bin	10738 non-null	category
23	day_since_cc_connect_bin	10742 non-null	category
24	cashback_percen	10742 non-null	object
25	cashback_benefits	10742 non-null	object

Table 1: Information of all Columns

The shape for the data set is as follows:

```
Total nos. of Rows: 11260
Total nos. of Columns: 18
```

The following table contains the summary statistics of all numeric columns in our data. any column that has n=10742 shows that there are no missing values (after cleaning) as we have 10742 unique customer IDs.

	count	mean	std	min	25%	50%	75%	max
churn	10742.000000	0.168311	0.374160	0.000000	0.000000	0.000000	0.000000	1.000000
tenure	10742.000000	11.069508	12.940544	0.010000	2.000000	9.000000	16.000000	99.000000
city_tier	10742.000000	1.648483	0.914493	1.000000	1.000000	1.000000	3.000000	3.000000
cc_contacted_ly	10742.000000	17.897319	8.847684	4.000000	11.000000	16.000000	23.000000	132.000000
service_score	10742.000000	2.902253	0.718785	1.000000	2.000000	3.000000	3.000000	5.000000
account_user_count	10742.000000	3.705083	1.003935	1.000000	3.000000	4.000000	4.000000	6.000000
cc_agent_score	10742.000000	3.048315	1.373534	1.000000	2.000000	3.000000	4.000000	5.000000
rev_per_month	10742.000000	6.426364	11.735007	1.000000	3.000000	5.000000	7.000000	140.000000
complain_ly	10742.000000	0.276764	0.447420	0.000000	0.000000	0.000000	1.000000	1.000000
rev_growth_yoy	10742.000000	16.221467	3.761982	4.000000	13.000000	15.000000	19.000000	28.000000
coupon_used_for_payment	10742.000000	1.811953	1.984488	0.000000	1.000000	1.000000	2.000000	16.000000
day_since_cc_connect	10742.000000	4.640570	3.643059	0.000000	2.000000	4.000000	7.000000	47.000000
cashback	10742.000000	197.653044	178.653948	0.000000	148.000000	168.000000	199.000000	1997.000000

Table 2: Summary Statistics of Numerical Columns

The following table studies the relationship between each the frequency of each object attribute

	count	unique	top	freq
payment	10742	5	Debit Card	4490
gender	10742	2	Male	6500
account_segment	10742	3	Super	4745
marital_status	10742	2	Married	5770
login_device	10742	2	Mobile	7894
noise_index	10742	3	medium_noise	6640
cashback_percen	10742	3	low_discount	8337
cashback_benefits	10742	3	low_benefit	9563

Table 3: Summary Statistics of Categorical Columns

Follows are the observations from table 2:

- **Payments:** Debit card is highly used mode of payment
- **Gender:** Male has the highest frequency
- **account_segment :** Super has the higheset frequency
- **marital_status :** Married Customers has the highest count
- **login_device :** Modile is most preferred mode of login.
- **noise_index :** Medium Noise has the highest count
- **cashback_percen :** low discount for cashback has the highest count
- **cashback_benefits :** Low benefits has the highest count

Below table showcase missing values in each of the columns which are been imputed during the process of Data Cleaning. Hence, all rows containing missing values have been imputed to avoid generating errors when training or testing data.

churn	0.000000
tenure	0.905861
city_tier	0.994671
cc_contacted_ly	0.905861
payment	0.968028
gender	0.959147
service_score	0.870337
account_user_count	0.994671
account_segment	0.861456
cc_agent_score	1.030195
marital_status	1.882771
rev_per_month	0.905861
complain_ly	3.170515
rev_growth_yoy	0.000000
coupon_used_for_payment	0.000000
day_since_cc_connect	3.170515
cashback	4.182948
login_device	1.962700
dtype: float64	

Table 4: Missing Values Percentage (before Data Cleaning) of each column

3.2 Data Cleaning

Some steps were performed after data exploration which will help in creating more accurate machine learning models and eliminate bias. In ever columns missing values are imputed using mean, median and mode accordingly. Columns containing special values like ***, # , %,*** **+** etc are also been cleaned and imputed.

In Columns like:

- Gender 'Female' and 'Male' is replaced with 'F' and 'M',
- In account_segment, "Regular Plus & Regular + "is replaced with "Regular" and "Super Plus & Super +" replaced with "Super"
- In Marital_Status we have replaced "Divorced" with "Single"

We also removed the duplicated values from the dataset.

3.3 Statistical Significance

Statistical Significance of checks relationship each column with our target variable (Churn) as statistical test like T-test, Analysis of Variance (ANOVA) and Chi-Square Test.

In order to check statistical significance Null Hypothesis (H0) and alternative Hypothesis(H1) is needed:

Null Hypothesis (H0) = There is no relationship between churn and compared columns

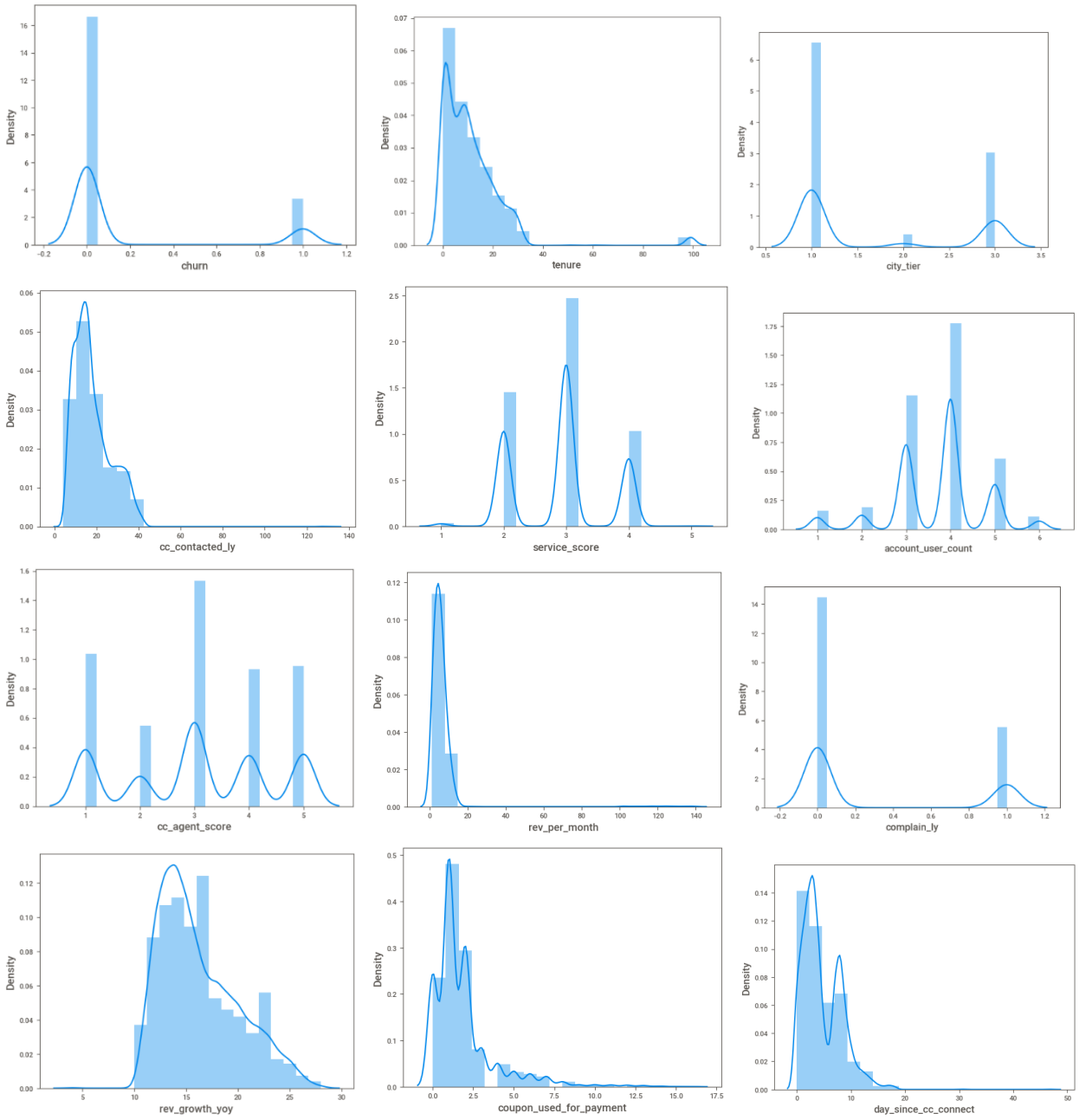
Alternate Hypothesis (H1) = There is relationship between churn and compared columns

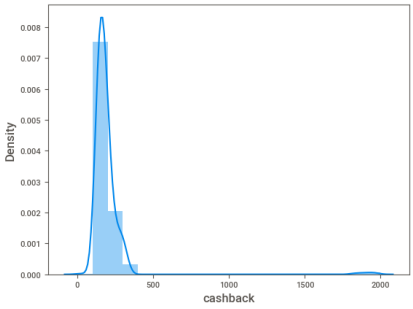
Note: Results of Statistical Significance of each column is in the 3.5 SUMMARY.

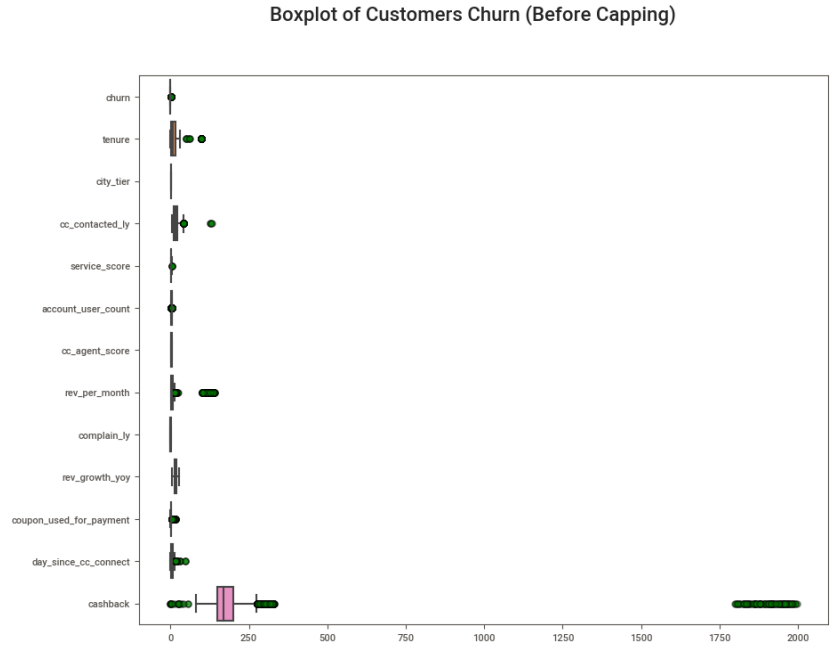
3.4 Data Visualization:

Uni-variate Analysis (Numerical)

In the bar charts below, we illustrated all the numerical attributes in our dataset to study their distribution:



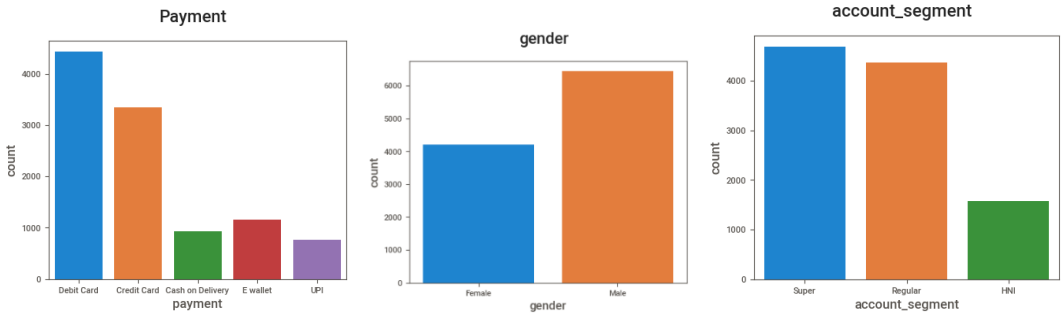


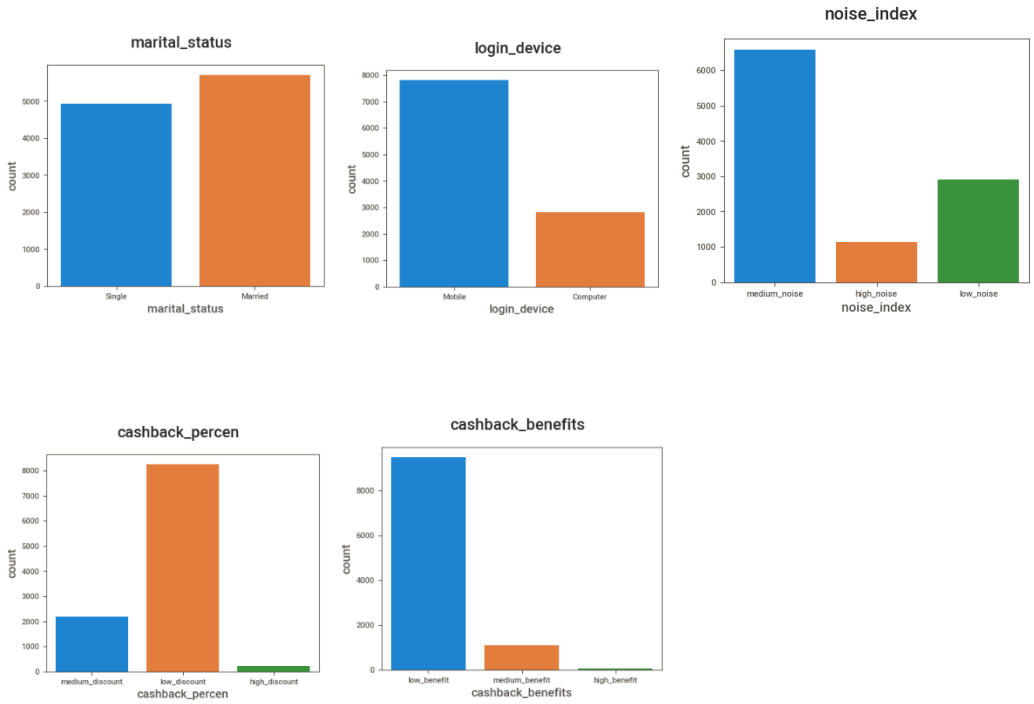


Graph 1: Density plot and Boxplot Uni-Variate (Numerical)

Uni-variate Analysis (Categorical)

In the bar charts below, we illustrated all the categorical attributes in our dataset to study their distribution:

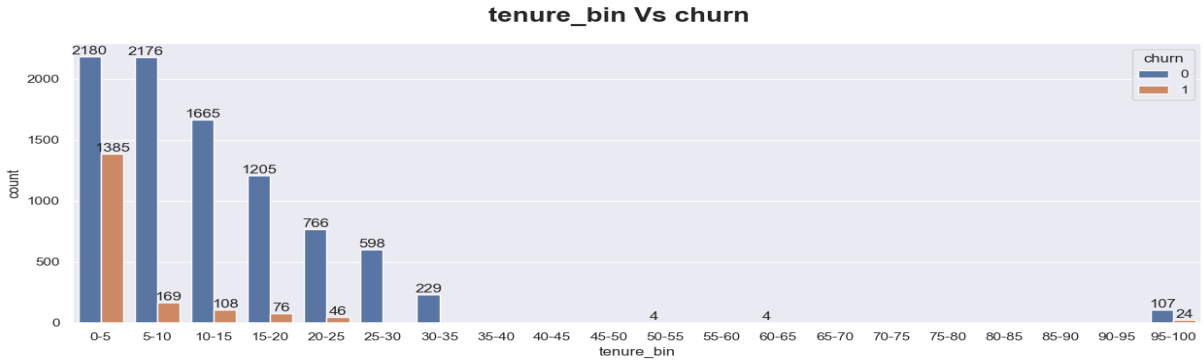


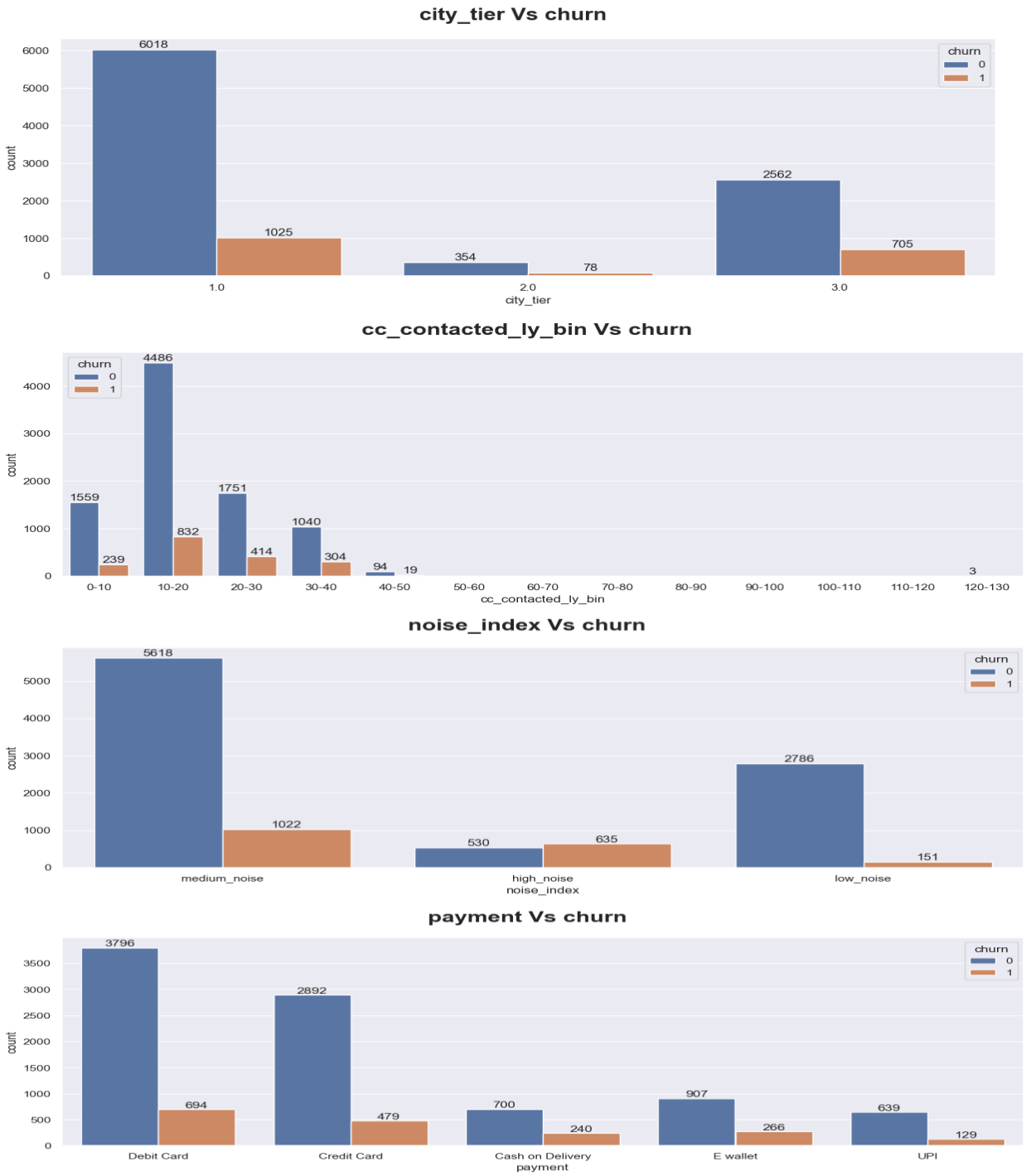


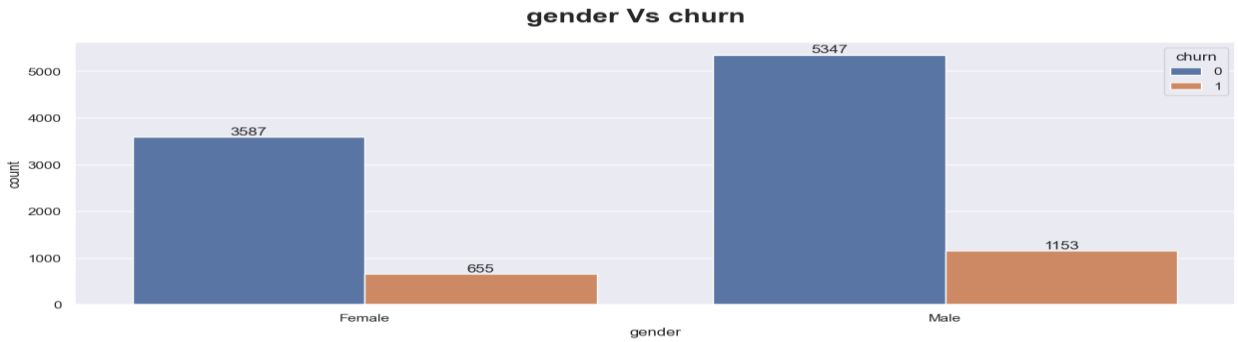
Graph 2: Barplot Univariate (Categorical)

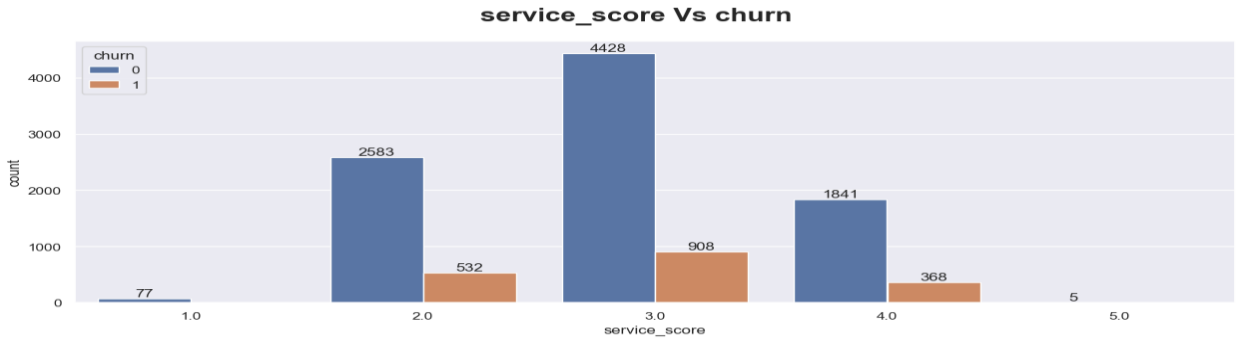
Bivariate Analysis (Churn Vs Each Column)

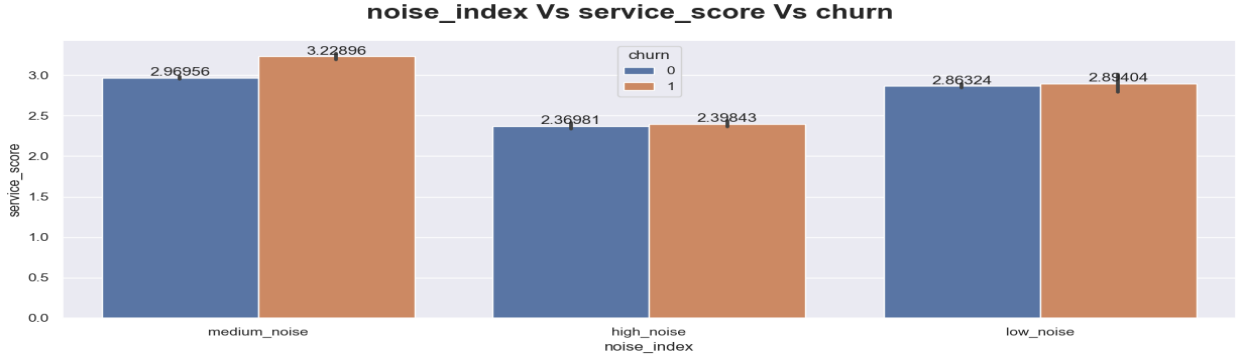
In the bar charts below, we illustrated Churn Vs Each Column in our dataset to study their distribution:

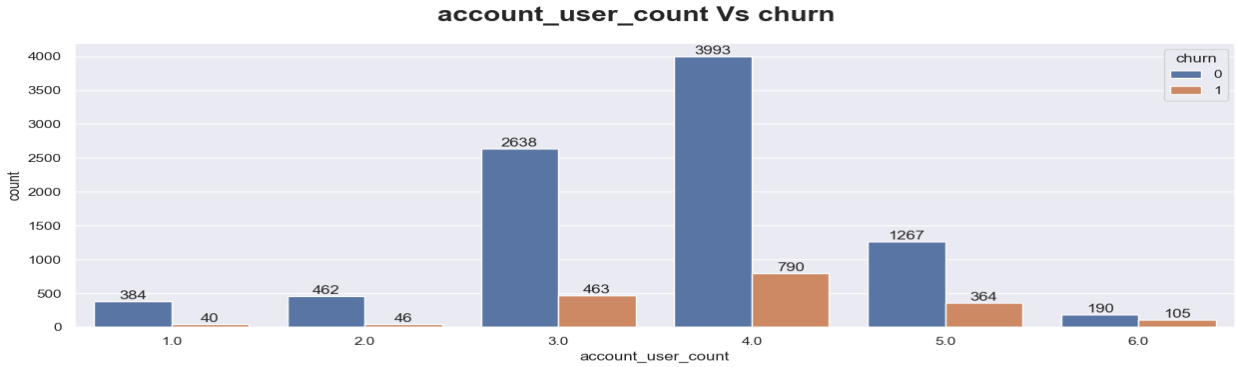


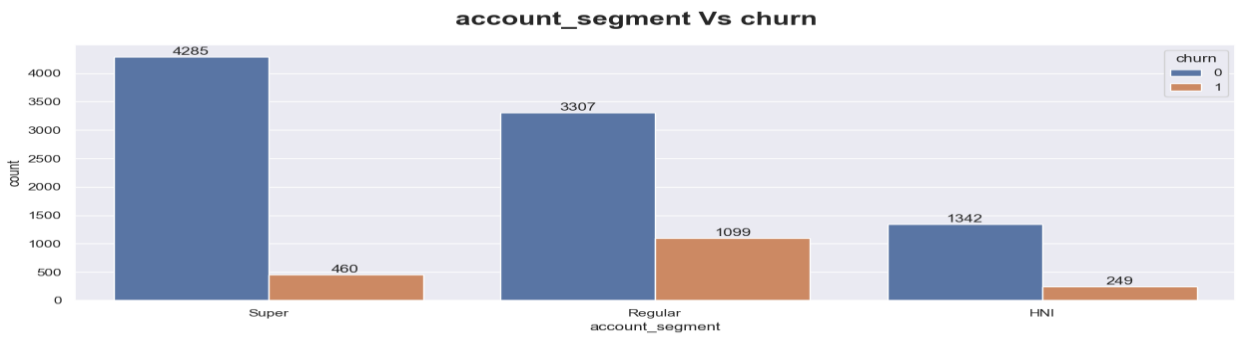

city_tier Vs churn
6018
1025
354
78
2562
705
churn
0
1
count
1.0
2.0
3.0
city_tier
cc_contacted_ly_bin Vs churn
1559
239
4486
832
1751
414
1040
304
94
19
3
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
100-110
110-120
120-130
cc_contacted_ly_bin
noise_index Vs churn
5618
1022
530
635
2786
151
medium_noise
high_noise
low_noise
noise_index
payment Vs churn
3796
694
2892
479
700
240
907
266
639
129
Debit Card
Credit Card
Cash on Delivery
E wallet
UPI
payment

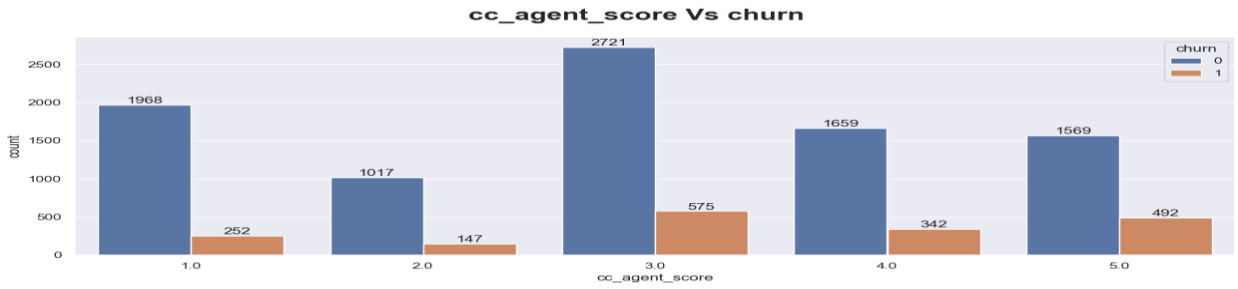

gender Vs churn
churn
0
1
count
gender
Female
Male
3587
655
5347
1153
0
1000
2000
3000
4000
5000

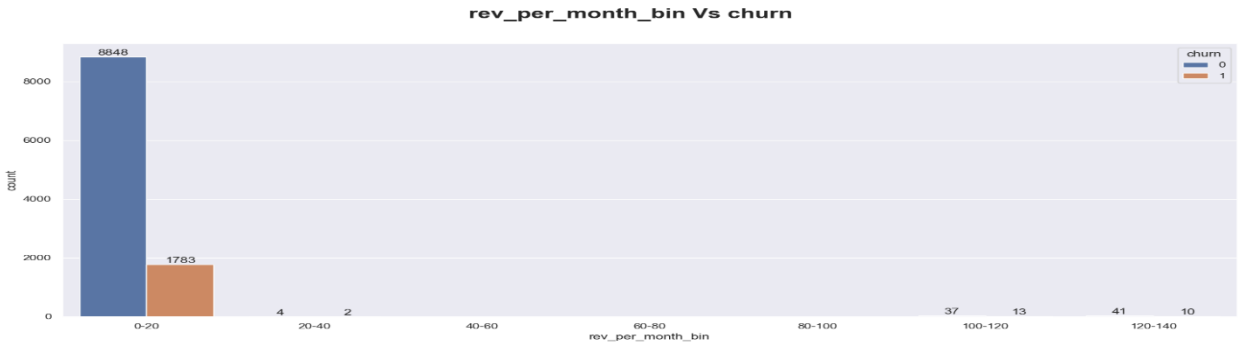

service_score Vs churn
churn
0
1
count
service_score
1.0
2.0
3.0
4.0
5.0
77
2583
532
4428
908
1841
368
5
0
1000
2000
3000
4000

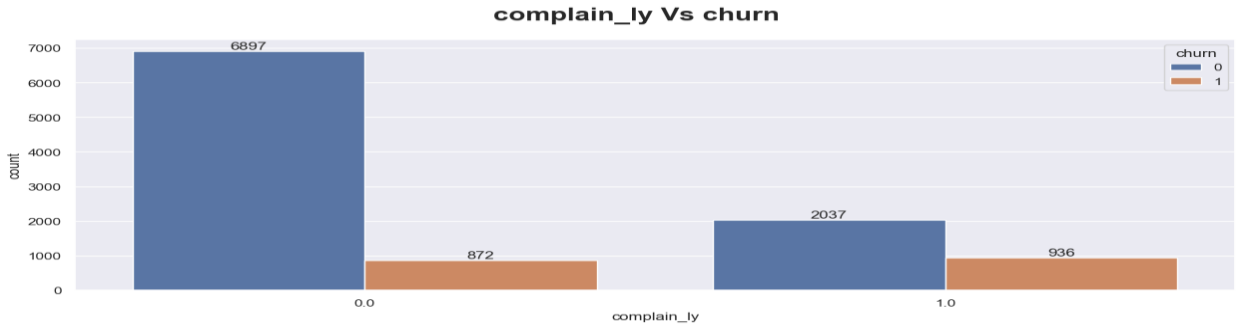

noise_index Vs service_score Vs churn
churn
0
1
service_score
noise_index
medium_noise
high_noise
low_noise
2.96956
3.22896
2.36981
2.39843
2.86324
2.89404
0.0
0.5
1.0
1.5
2.0
2.5
3.0

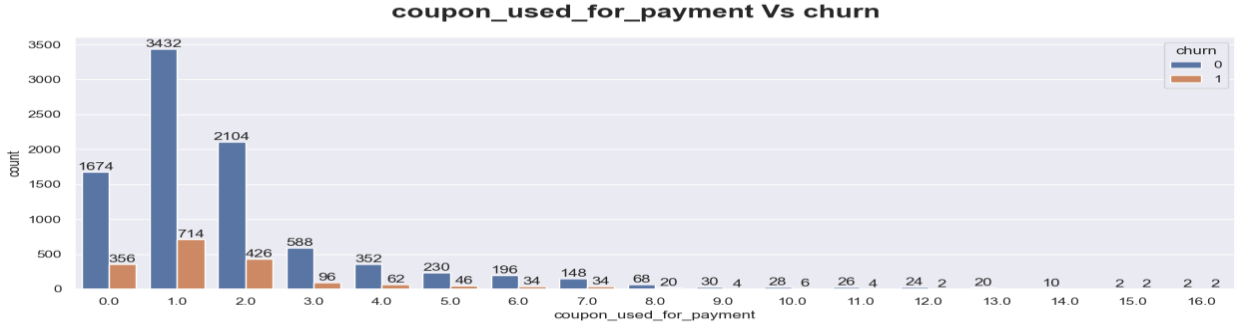

account_user_count Vs churn
churn
0
1
count
account_user_count
1.0
2.0
3.0
4.0
5.0
6.0
384
40
462
46
2638
463
3993
790
1267
364
190
105
0
500
1000
1500
2000
2500
3000
3500
4000

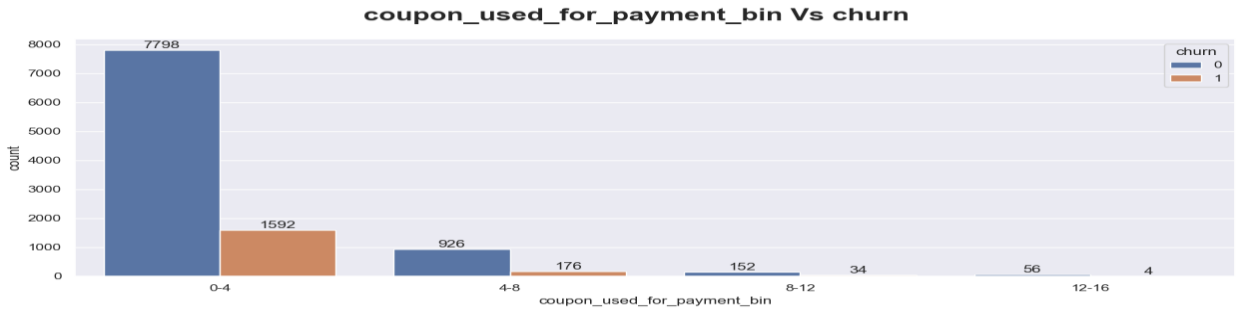

account_segment Vs churn
churn
0
1
count
4285
460
3307
1099
1342
249
Super
Regular
HNI
account_segment

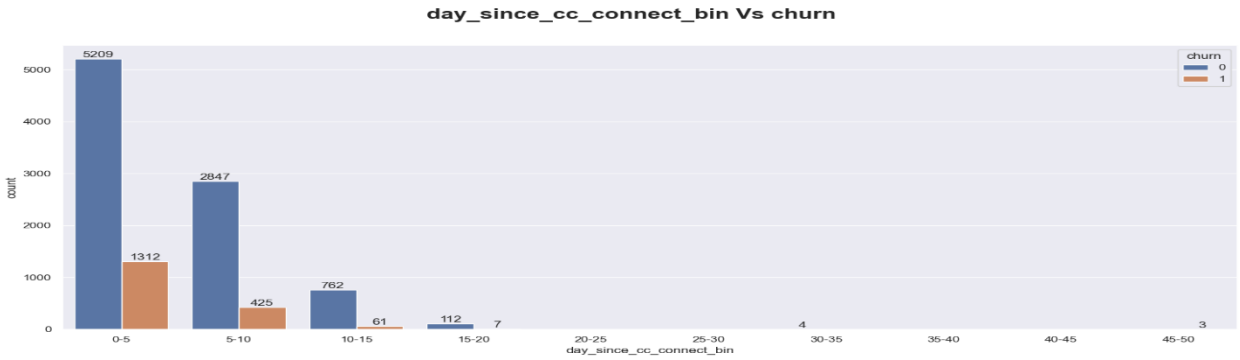

cc_agent_score Vs churn
churn
0
1
count
1968
252
1017
147
2721
575
1659
342
1569
492
1.0
2.0
3.0
4.0
5.0
cc_agent_score

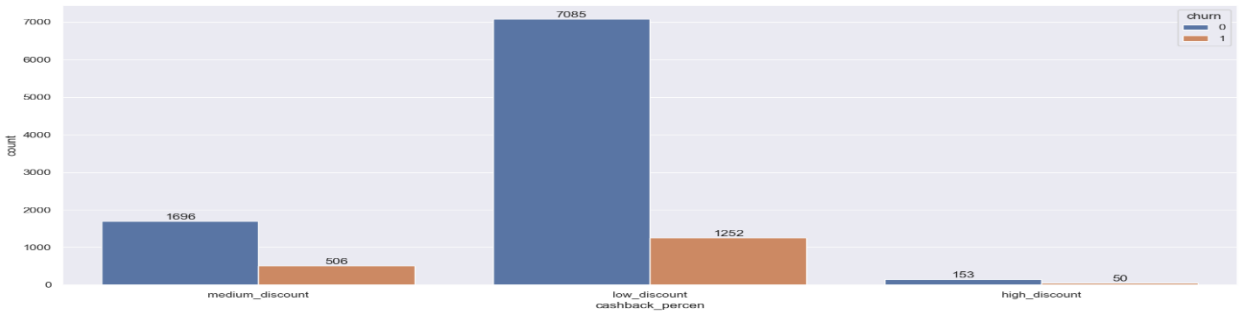

rev_per_month_bin Vs churn
churn
0
1
count
8848
1783
4
2
37
13
41
10
0-20
20-40
40-60
60-80
80-100
100-120
120-140
rev_per_month_bin

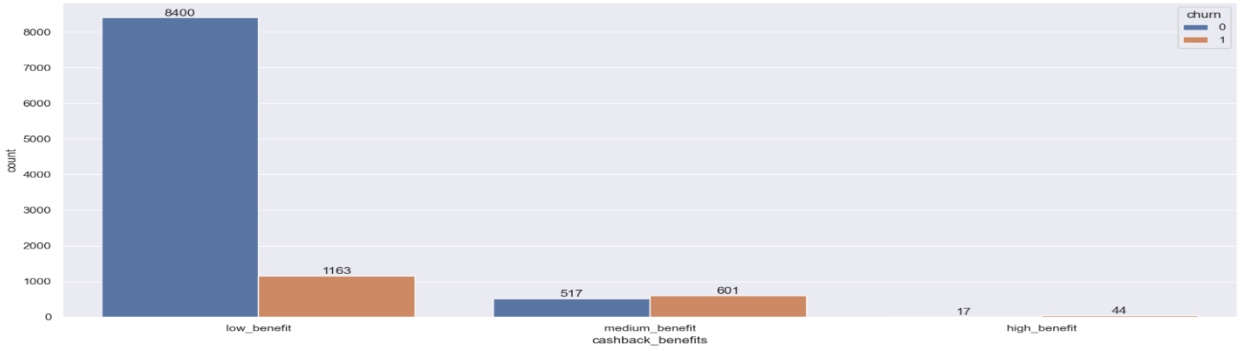

complain_ly Vs churn
churn
0
1
count
6897
872
2037
936
0.0
1.0
complain_ly

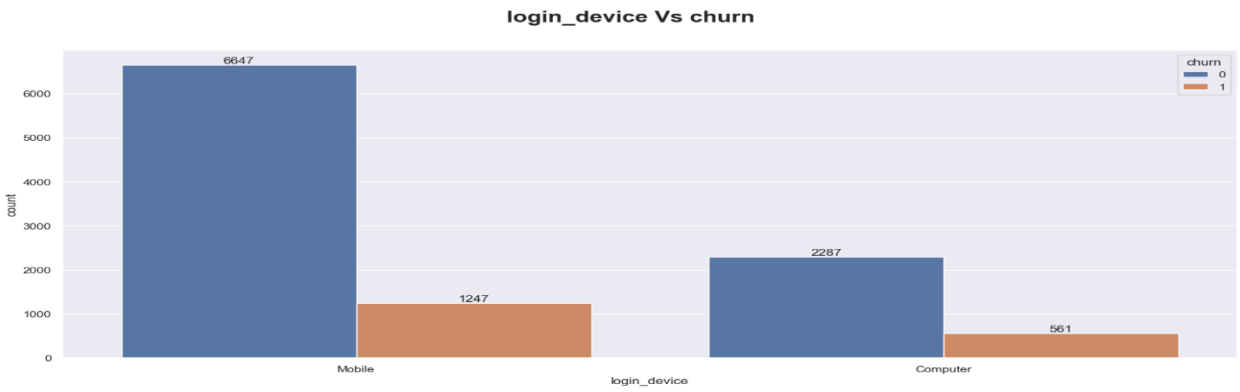

coupon_used_for_payment Vs churn
churn
0
1
count
1674
356
3432
714
2104
426
588
96
352
62
230
46
196
34
148
34
68
20
30
4
28
6
26
4
24
2
20
10
2
2
2
2
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
coupon_used_for_payment


coupon_used_for_payment_bin Vs churn
churn
0
1
count
coupon_used_for_payment_bin
0-4
4-8
8-12
12-16
7798
1592
926
176
152
34
56
4


day_since_cc_connect_bin Vs churn
churn
0
1
count
day_since_cc_connect_bin
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
5209
1312
2847
425
762
61
112
7
4
3


cashback_percen Vs churn
churn
0
1
count
cashback_percen
medium_discount
low_discount
high_discount
1696
506
7085
1252
153
50


cashback_benefits Vs churn
churn
0
1
count
cashback_benefits
low_benefit
medium_benefit
high_benefit
8400
1163
517
601
17
44



Graph 3: Bar plot Biivariate

3.5 Summary of Univariate and Bivariate Analysis and Statistical Significance

(1) Churn (Target Variable):
* Churn = 1 and Not-Churned = 0
* Average Overall Churn Rate: 16.838365896980463
* Imbalance in dataset

(2) Tenure:
* Both Anova Test andT Test says there is a relationship between churn and tenure.

(3) Tenure_bin:
* We can see a very huge spike customer churning within the first 6 Months of Tenure which is a problem
* Chi square Test says there is a relationship between churn and tenure_bin.

(4) City_Tier:
* Highest Churn rate can be seen in TIer-3 City i.e. 22% after that Tier-2 city
* Chi-Square Test says there is a relationship between churn and city_tier.

(5) cc_contacted_ly:
* T- Test and Anova Test says Their is a relationship between churn and cc_contacted_ly.

(6) cc_contacted_ly_bin:
* Churn rate i.e. 22% is High for the customers who have contacted between 30-40 times in the last 12 months
* Chi square Test says There is a relationship between churn and cc_contacted_ly_bin.

(7) Noise Index:

* If the index is < 1 then customers is calling once or less than once a month if index > 1 then customers in calling every month. And if no. is more than 100 then customer reaches support quite frequently
* High Churn rate can be seen in High Nosie i.e. 54%

* Chi square Test says There is a relationship between churn and noise_index.

(8) Payment:
* We can clearly see that Cash on Delivery has the highest churn rate i.e. 25% followed by E-wallet.
* Chi square Test says There is a relationship between churn and payment.

(9) Gender:
* Highest churn rate can be seen in Male. i.e. 17%
* Chi square Test says There is a relationship between churn and gender.

(10) Service_Score:
* High churn rate i.e. 17% can seen for those customers whose satisfaction rating 3 on the services provided by the company
* Chi square Test says Their is a relationship between churn and service_score.

(11) Noise Index vs Service Score vs Churn:
* Medium Noise customers(calls the company for 2 or 3 times a month) who service score rating is more than 3 for the company are churning at high pace which shows these customers are Unhappy Customers

(12) Account_user_count:
* Account User Count having 6 customers has the highest churn rate i.e. 34%
* Chi square Test says Their is a relationship between churn and account_user_count.

(13) account_segment:
* Regular Account Segment Customers has the highest Churn Rate i.e. 25%
* Chi square Test says Their is a relationship between churn and account_segment.

(14) CC_Agent_Score:
* Score HIGH means Good customer. Score LOW means Bad customer according to Agent.
* Customers whose score rate given by Agents is 5 has high Churn rate i.e. 24% means a good quality/loyal customers are going out of the company.
* Chi square Test says Their is a relationship between churn and cc_agent_score.

(15) rev_per_month:
* T- Test says Their is a relationship between churn and rev_per_month.

(16) rev_per_month_bin:
* Revenue per Month earned between 100-120 dollars has the 26% Churn rate which very high means we are losing money every months as customers are leaving the company.
* Chi square Test says Their is no relationship between churn and rev_per_month_bin.

(17) complain_ly:
* We can clearly see here customers who complains a lot usually churn faster. i.e 31%
* Chi square Test says Their is a relationship between churn and complain_ly.

(18) rev_growth_yoy:
* T- Test says Their is no relationship between churn and rev_growth_yoy.

(19) coupon_used_for_payment:
* Customers using coupons once in 12 months for payments are churning followed by customers using coupons twice.
* T- Test says Their is no relationship between churn and coupon_used_for_payment.

(20) coupon_used_for_payment_bin:

* Customers using coupons for payments between 8-12 times have high churn rate. i.e. 18%
* Chi square Test says Their is no relationship between churn and coupon_used_for_payment_bin.

(21) Day_Since_CC_connect:
* T- Test says Their is a relationship between churn and day_since_cc_connect.

(22) day_since_cc_connect_bin:
* Number of days since no customers in the account has contacted the customer care between 45-50 days has highest churn rate i.e 100 % followed by 0-5 days means within 5 days 20% customers connected with customer care they start churning.
* Chi square Test says Their is a relationship between churn and day_since_cc_connect_bin.

(23) cashback:
* T- Test says There is a relationship between churn and cashback.

(24) cashback_percen:
* Convert rev_per_month into a yearly revenue and check the cashback percentage and if it is more than 1 then customers are getting more discount and if less than 1 then they are not getting enough discount
* In Cashback Percentage, High Discount customers has 25% churn rate followed by medium discounts which is also a loss in revenue for the company.
* Chi square Test says There is a relationship between churn and cashback_percen.

(25) Cashback_Benefits:

* These are the benefits received as long as customer stay with the company.

* If the value the very high customer has received a lot of benefits from company and check % of churned customers
* Chi square Test says There is a relationship between churn and cashback_benefits.

(26) Login_device:
* Customers using Computers availing companies servies are churning more. i.e. 20%
* Chi square Test says There is a relationship between churn and login_device.

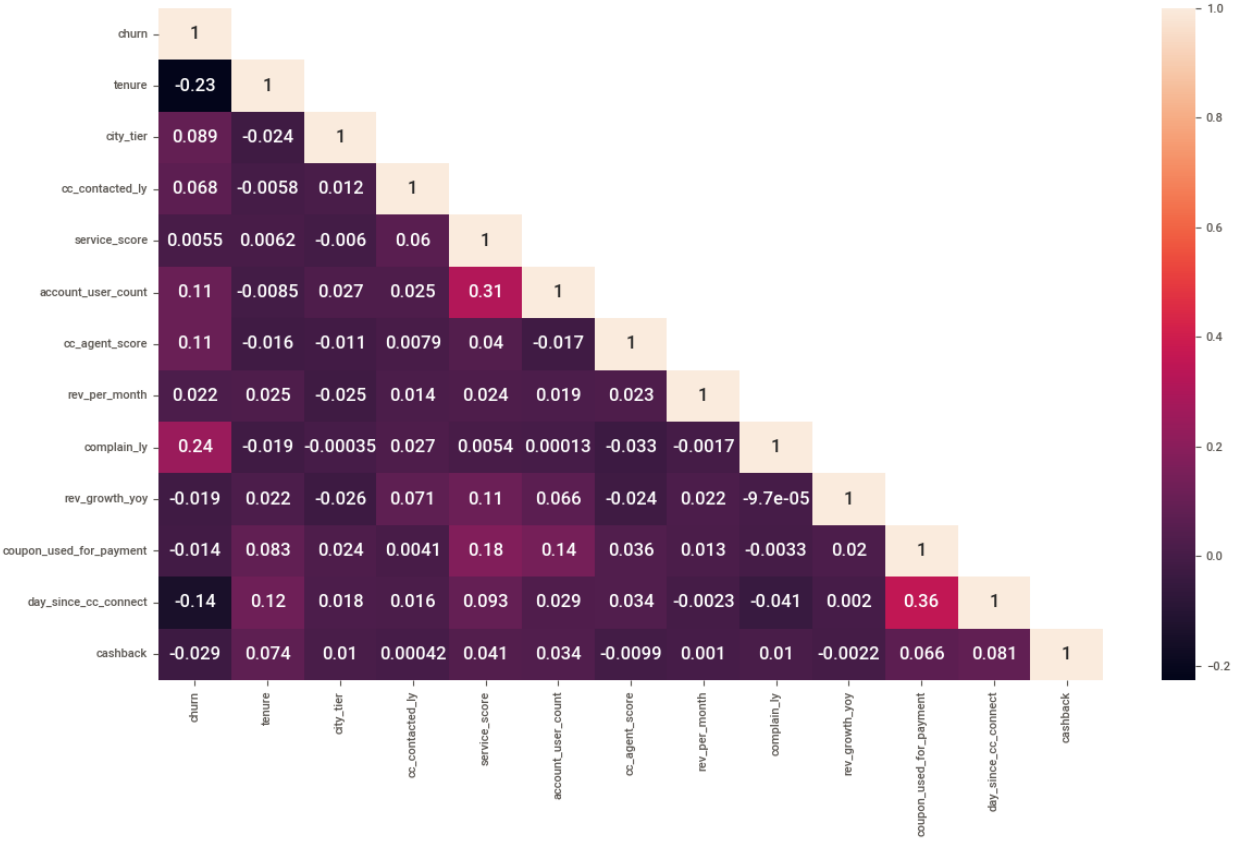
(27) Married Status:

* We can clearly see that Single churn rate is Higher i.e 23% as compared to Married
* Chi square Test says There is a relationship between churn and marital_status.

3.6 Multi – Variate Analysis and Correlation

Graph 4: Pair plot Multi-Variate

Observation : we cannot much see correlation among the columns



Graph 5: HeatMap Correlation

Observation :

- A high **Negative correlation** between **Churn and Tenure**.
- A high **Negative correlation** between **Churn and Day_Since_CC_connect**.
- A high **Positive correlation** between **Churn and complain_ly**.
- A high **Positive correlation** between **Account_user_count and Service_Score**.
- A high **Positive correlation** between **coupon_used_for_payment and Day_Since_CC_connect**.

3.7 Business Insights from the EDA

Based on the summary provided, here are some potential business insights for an e-commerce company:

- The company has a high overall churn rate and needs to take steps to address this issue, such as improving customer service or product offerings.
- Customers who have been with the company for less than six months are churning at a high rate, indicating a need to focus on improving retention in the early stages of the customer journey.
- Customers in Tier-3 cities are churning at a higher rate than those in Tier-1 or Tier-2 cities, suggesting the need for targeted marketing and customer engagement strategies in these areas.
- The company needs to pay attention to the number of customer service calls received, as customers who call frequently (i.e. have a high "noise index") are churning at a high rate.
- Cash on delivery is the payment method with the highest churn rate, indicating a need to explore alternative payment options or improve the cash on delivery experience for customers.
- Male customers are churning at a higher rate than female customers, suggesting the need for targeted marketing and retention strategies for male customers.
- Customers with a service score of 3 are churning at a high rate, indicating a need to focus on improving the quality of service provided by the company.
- Customers who complain frequently are churning at a high rate, suggesting the need to improve the customer service experience and address customer concerns in a timely and effective manner.
- Customers who receive high cashback benefits are churning at a high rate, indicating a need to explore alternative loyalty and retention strategies

Sections 4: Model Development

The following stage consists of several steps that must be taken starting with splitting the data into training and testing data. Then, a base model is created then base model is compare with 9 machine learning models are going to be built to compare their F1 score's.

Data Pre-processing:

Before going deep into building the models, certain steps must be followed to make sure that models are built to give the best performance. As mentioned earlier, the dataset consists of 10626 rows following data cleaning step. Therefore, data must be split into training data (70%) and test data (30%). The outcome of the models will be churn, Account ID column will be removed from predictors column. Furthermore, all categorical predictors will be converted to numerical values. Also, all predicters with zero or low variability will be eliminated.

Few extra steps are going to be carried out exclusively for model which are:

- Data normalization for all predictors.
- Highly correlated variables in all predictors are going to be removed.
- One-hot encoding is a technique used to convert categorical data into a format that can be used by machine learning models. It creates binary columns for each category.
- Synthetic minority oversampling technique (SMOTE) will be applied to balance the levels of Churn column by generating new samples of the minority class (Churned).

Base Model (Logistic Regression)

We are using logistic regression as our base model were we have check the VIF (Variance Inflation Factor), Area Under the Curve (AUC) and ROC (Receiver Operating Characteristic), Confusion Matrix and Classification Report for base Accuracy, Precision, Recall & F1 score.

In addition to the above we have also used Feature Selection- Information Gain and SHAP values to determine which variable is most important according to the model.

Here in the base model we have not included variables like Noise_index,cashback_benefits and cashback_percen

***Note**: Check appendix for definitions*

VIF values:

We can consider a rule of thumb that if vif is greater than 5, we can choose to drop the variable as there can be a problem of multicollinearity. This essentially means that we can choose to drop a predictor variable whose 80% variation is being explained by the other predictor variables.

No multicollinearity can been seen any of the variables from below table:

```
tenure VIF =  1.29
city_tier VIF =  1.45
cc_contacted_ly VIF =  1.03
service_score VIF =  1.21
account_user_count VIF =  1.15
cc_agent_score VIF =  1.02
rev_per_month VIF =  1.09
complain_ly VIF =  1.01
rev_growth_yoy VIF =  1.02
coupon_used_for_payment VIF =  1.26
day_since_cc_connect VIF =  1.31
cashback VIF =  1.58
payment_Credit_Card VIF =  3.17
payment_Debit_Card VIF =  3.38
payment_E_wallet VIF =  2.34
payment_UPI VIF =  1.7
gender_Male VIF =  1.01
account_segment_Regular VIF =  2.76
account_segment_Super VIF =  2.34
marital_status_Single VIF =  1.02
login_device_Mobile VIF =  1.01
```

Table 5: VIF values base model

Below is the finals base model output of the Logistic regression which states how much each variable has the contributes towards model building.

Relation of each variable with the model is as follows:

```
(-3.87) * Intercept + (-0.17) * tenure + (0.38) * city_tier + (0.03) * cc_
contacted_ly + (-0.18) * service_score + (0.38) * account_user_count + (0.
28) * cc_agent_score + (0.15) * rev_per_month + (1.67) * complain_ly + (-0
.03) * rev_growth_yoy + (0.16) * coupon_used_for_payment + (-0.07) * day_s
ince_cc_connect + (-0.59) * payment_Credit_Card + (-0.4) * payment_Debit_C
ard + (-0.68) * payment_UPI + (0.25) * gender_Male + (0.21) * account_segm
ent_Regular + (-1.02) * account_segment_Super + (0.8) * marital_status_Sin
gle + (-0.47) * login_device_Mobile +
```

Dep. Variable:	churn	No. Observations:	10626
Model:	Logit	Df Residuals:	10606
Method:	MLE	Df Model:	19
Date:	Sun, 05 Mar 2023	Pseudo R-squ.:	0.3413
Time:	08:54:57	Log-Likelihood:	-3169.3
converged:	True	LL-Null:	-4811.7
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.8691	0.292	-13.273	0.000	-4.440	-3.298
tenure	-0.1687	0.006	-26.931	0.000	-0.181	-0.156
city_tier	0.3814	0.039	9.661	0.000	0.304	0.459
cc_contacted_ly	0.0279	0.004	7.347	0.000	0.020	0.035
service_score	-0.1755	0.050	-3.531	0.000	-0.273	-0.078
account_user_count	0.3804	0.039	9.839	0.000	0.305	0.456
cc_agent_score	0.2782	0.024	11.400	0.000	0.230	0.326
rev_per_month	0.1467	0.011	12.882	0.000	0.124	0.169
complain_ly	1.6690	0.068	24.659	0.000	1.536	1.802
rev_growth_yoy	-0.0318	0.009	-3.585	0.000	-0.049	-0.014
coupon_used_for_payment	0.1629	0.034	4.744	0.000	0.096	0.230
day_since_cc_connect	-0.0692	0.012	-5.765	0.000	-0.093	-0.046
payment_Credit_Card	-0.5881	0.095	-6.194	0.000	-0.774	-0.402
payment_Debit_Card	-0.4048	0.088	-4.595	0.000	-0.578	-0.232
payment_UPI	-0.6773	0.141	-4.806	0.000	-0.953	-0.401
gender_Male	0.2527	0.067	3.786	0.000	0.122	0.383
account_segment_Regular	0.2125	0.104	2.036	0.042	0.008	0.417
account_segment_Super	-1.0195	0.103	-9.864	0.000	-1.222	-0.817
marital_status_Single	0.8033	0.065	12.287	0.000	0.675	0.931
login_device_Mobile	-0.4732	0.071	-6.687	0.000	-0.612	-0.335

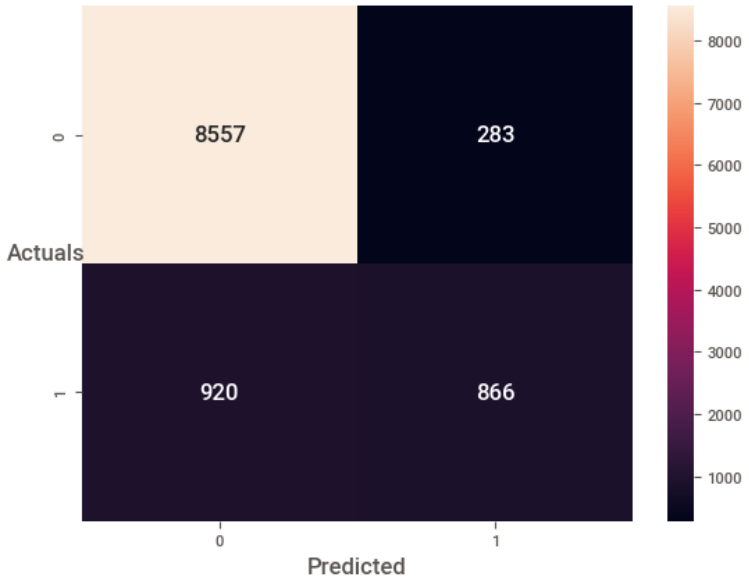
Table 6: Final variables of base model

Classification report of base model:

	precision	recall	f1-score	support
0	0.903	0.968	0.934	8840
1	0.754	0.485	0.590	1786
accuracy			0.887	10626
macro avg	0.828	0.726	0.762	10626
weighted avg	0.878	0.887	0.876	10626

- Accuracy = 0.887
- precision = 0.754
- recall = 0.485
- f1-score = 0.590

Confusion matrix:

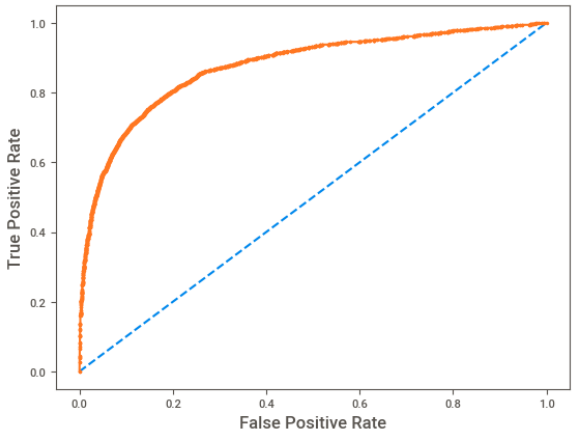


Graph 6: Confusion matrix base model

True Positive (TP) = 8557 are the True predicted positive values by model
True Negative (TN) = 866 are the True predicted negative values by model
False Positive (FP) = 283 are the False predicted positive values by model
False Negative (FN) = 920 are the False predicted negative values by model

Thus, we can see here model is able to capture high TP but less TN. Therefore, we need better model.

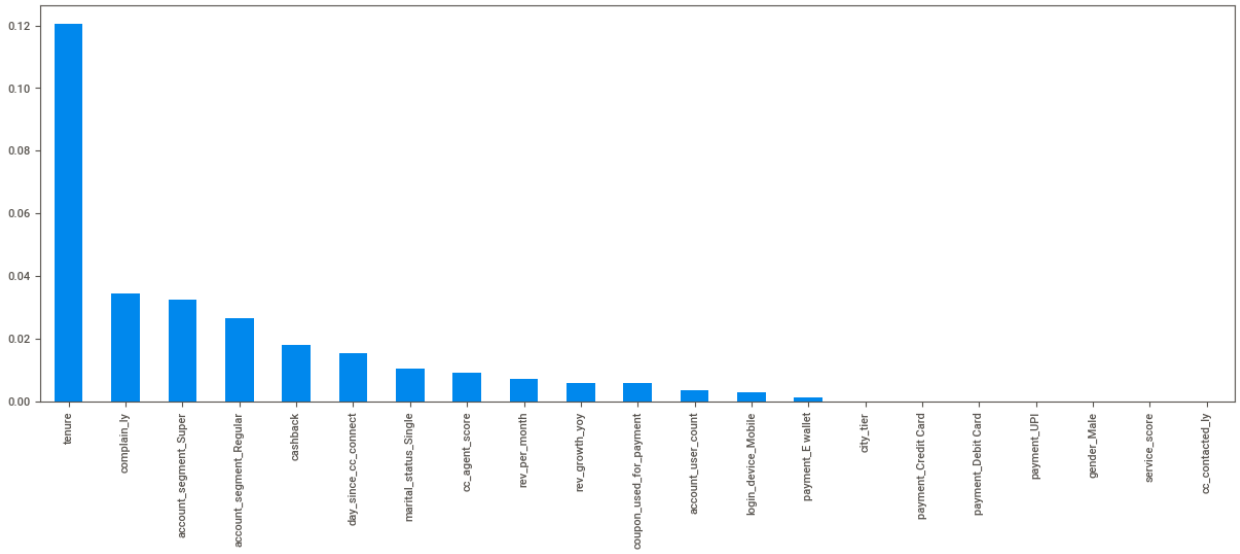
AUC-ROC

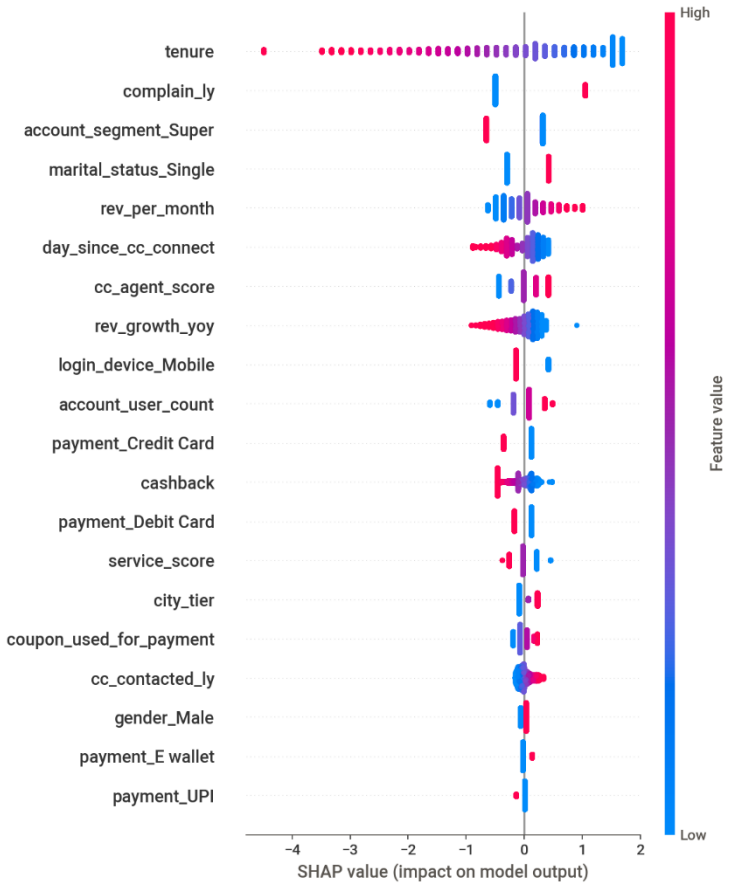


Graph 7: AUC ROC curve base model

AUC - 0.876. A high AUC value indicates a good classifier that has a high true positive rate and a low false positive rate.
Thus, from the above we have got our base values which we can use for comparing with other models

Information Gain and Shap Values:





Graph 8: Information Gain and SHAP values base model

Information Gain and SHAP values base model clearly tell top 5 variables that contributes the most and should be focused on by the company. Those are as follows:
Tenure, Complain_ly, account_Segment, martial_status, rev_per_month and day_since_CC_connect

9 Model Building

We have chosen total 9 classification supervised Machine learning model those are:

- Logistic Regression
- K-nearest neighbour (KNN)
- Bagging – Random Forest
- XGBoost
- Support Vector Machine
- Naive bayes
- Decision Tree
- Extra tree classifier
- Linear Discriminant Analysis

***Note**: Check appendix for definitions*

These 9 models are been trained, tested and hyper tuned so as to derive best scores out of it. We have also used SMOTE in order to create synthetic churn (1) customers so that model can learn and accurately predict F-1 score which is our main objective.

Let's check the score of each model Before and After Hyper tuning with SMOTE:

	Train Accuracy	Test Accuracy	Train AUC	Test AUC	Train Recall	Test Recall	Train precision	Test precision	Train f1	Test f1
Logistic_Regression	0.84	0.84	0.88	0.83	0.70	0.57	0.79	0.52	0.74	0.54
KNN	0.96	0.93	1.00	0.96	0.96	0.84	0.94	0.75	0.95	0.79
Bagging	1.00	0.94	1.00	0.98	1.00	0.77	1.00	0.88	1.00	0.82
XGBoost	1.00	0.96	1.00	0.98	1.00	0.83	1.00	0.91	1.00	0.87
Support_Vector_Machine	0.73	0.84	0.81	0.81	0.27	0.26	0.77	0.55	0.40	0.35
Naive_Bayes	0.81	0.83	0.84	0.80	0.66	0.56	0.74	0.49	0.70	0.52
Decision_Tree	1.00	0.93	1.00	0.88	1.00	0.79	1.00	0.80	1.00	0.80
Extra_Trees_Classifier	1.00	0.97	1.00	0.99	1.00	0.85	1.00	0.96	1.00	0.90
Linear_Discriminant_Analysis	0.84	0.86	0.90	0.85	0.68	0.54	0.81	0.58	0.74	0.56

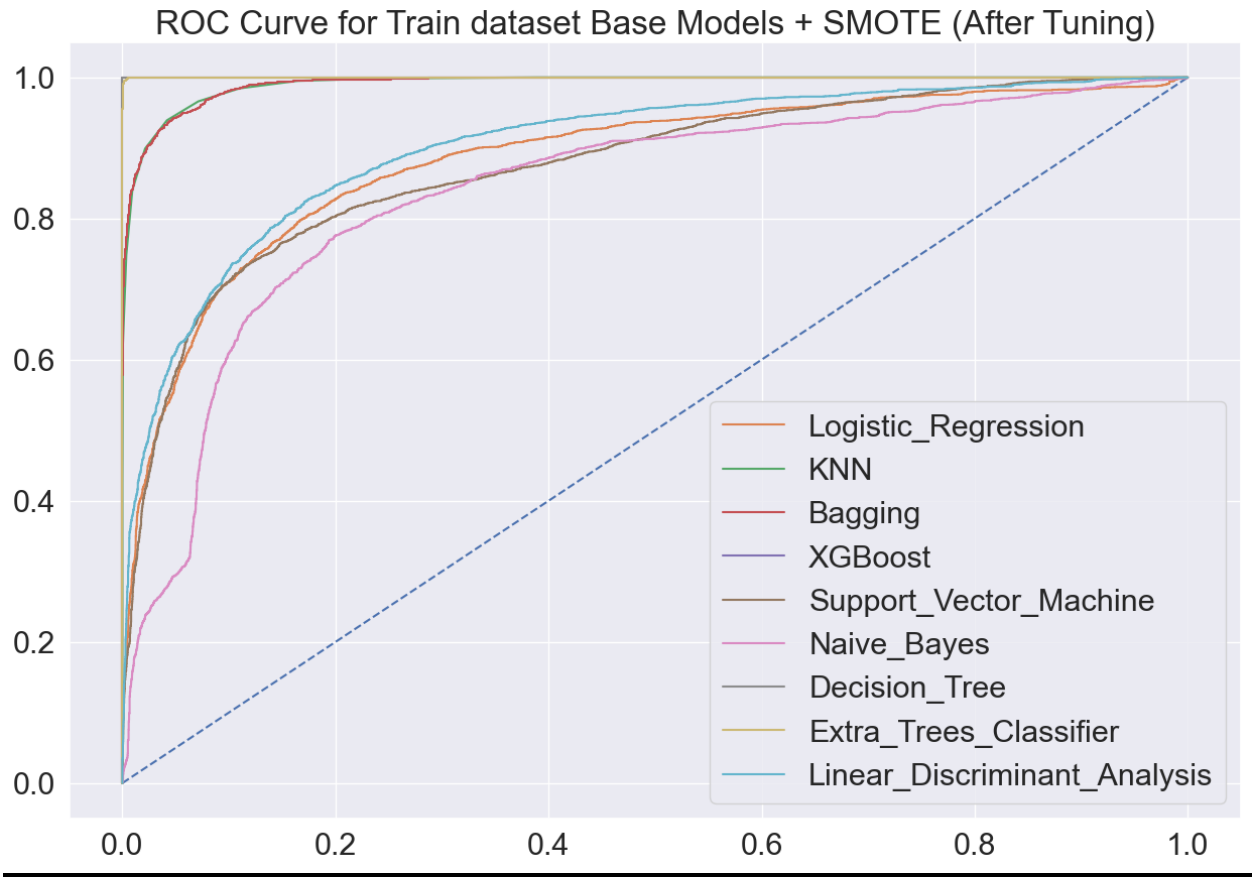
Table 7.1: The 9 Models Scores before Hyper Tuning

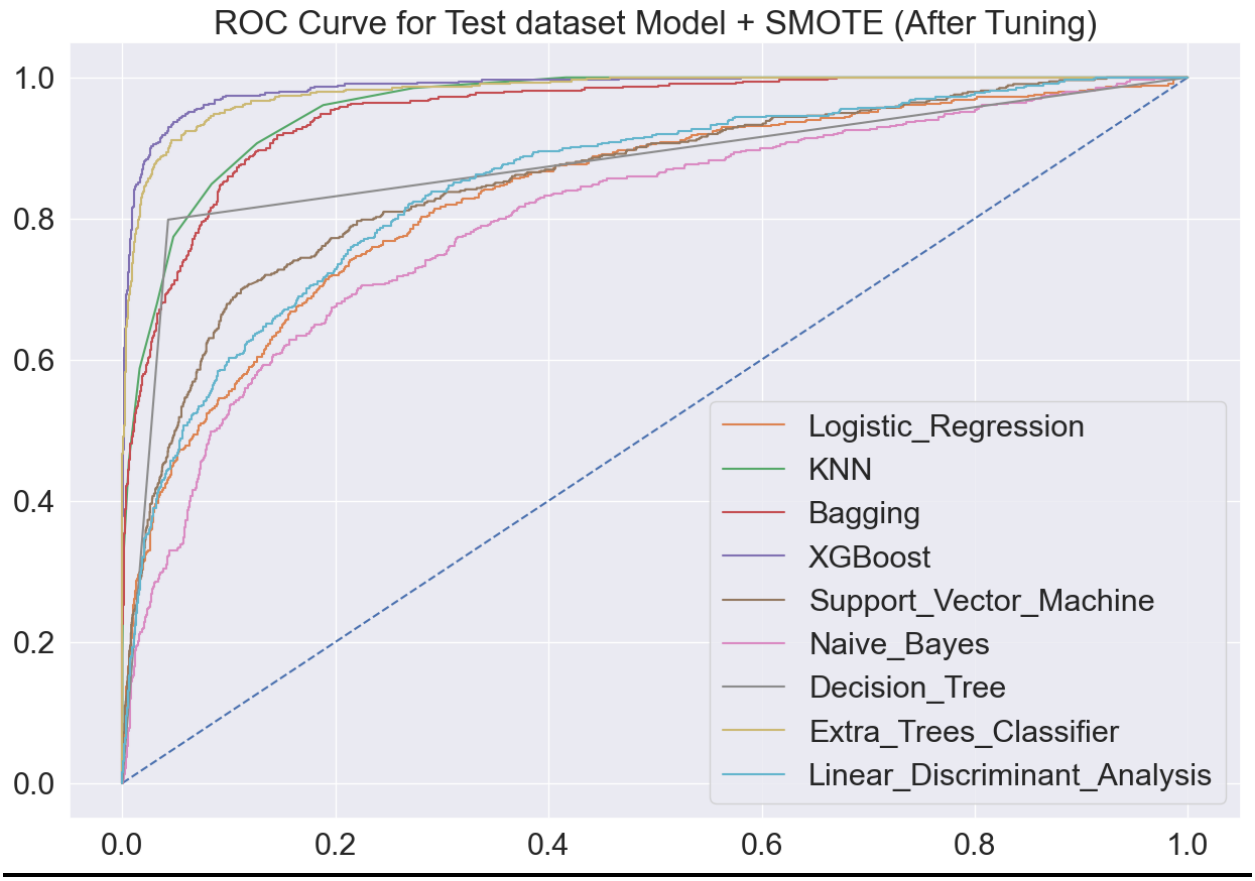
	Train Accuracy	Test Accuracy	Train AUC	Test AUC	Train Recall	Test Recall	Train precision	Test precision	Train f1	Test f1
Logistic_Regression	0.84	0.84	0.88	0.83	0.70	0.57	0.79	0.52	0.74	0.54
KNN	0.95	0.92	0.99	0.96	0.94	0.77	0.92	0.77	0.93	0.77
Bagging	0.95	0.91	0.99	0.95	0.87	0.65	0.97	0.80	0.92	0.72
XGBoost	1.00	0.96	1.00	0.99	1.00	0.85	1.00	0.92	1.00	0.88
Support_Vector_Machine	0.83	0.84	0.87	0.86	0.74	0.71	0.75	0.53	0.74	0.61
Naive_Bayes	0.81	0.83	0.84	0.80	0.66	0.56	0.74	0.49	0.70	0.52
Decision_Tree	1.00	0.93	1.00	0.88	1.00	0.80	1.00	0.79	1.00	0.79
Extra_Trees_Classifier	1.00	0.95	1.00	0.98	0.99	0.79	1.00	0.91	0.99	0.85
Linear_Discriminant_Analysis	0.84	0.86	0.90	0.85	0.69	0.55	0.81	0.58	0.74	0.56

Table 7.2: The 9 Models Scores after Hyper Tuning

From the above table we can clearly see the Extra Tree Classifier ,XGBoost and KNN have given the best F1-scores along with Accuracy, precision and recall scores.

AUC-ROC Curve for all the 9 models





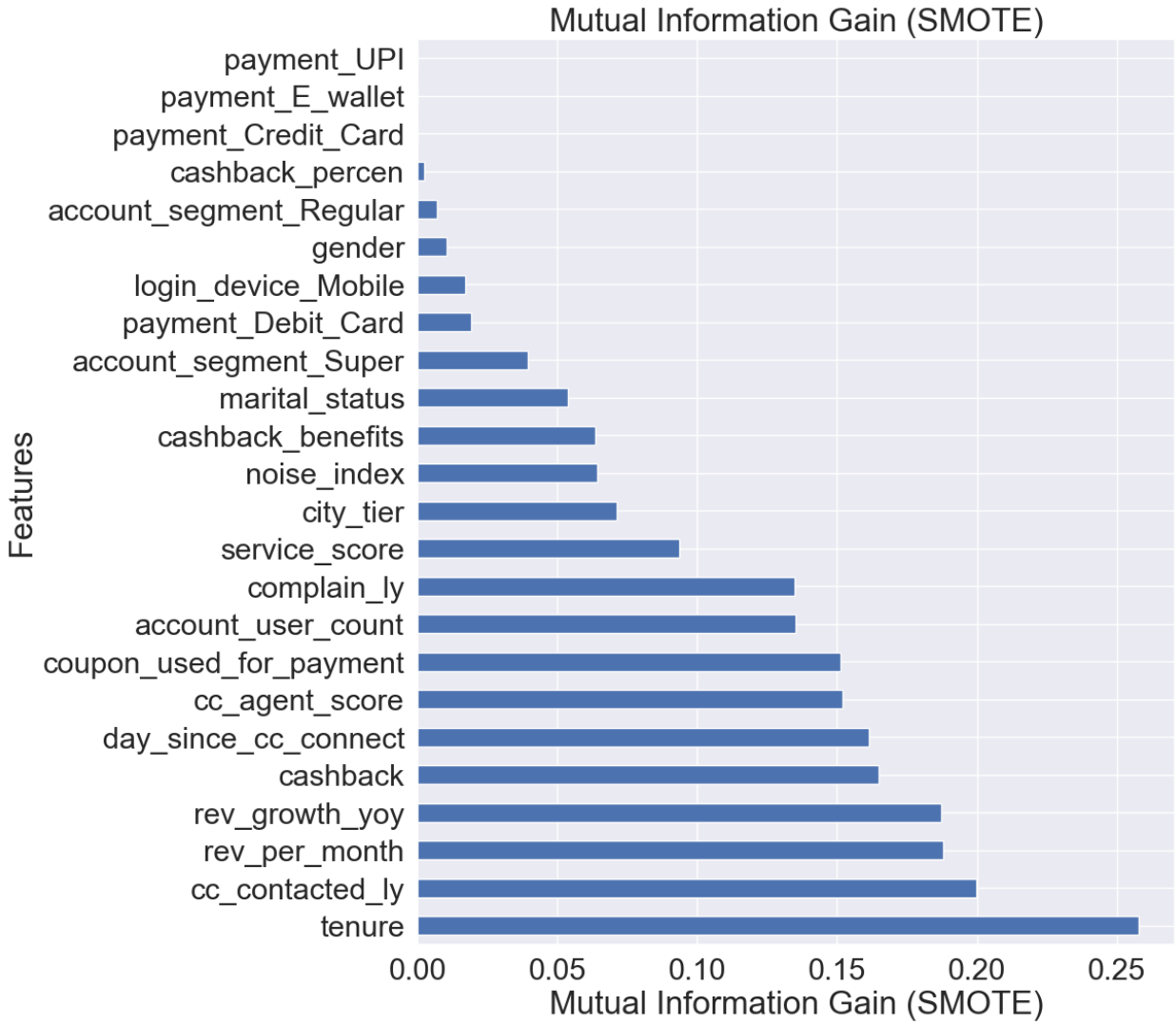
Graph 9: AUC-ROC curve 9 Models (Train and Test)

AUC-ROC Training		AUC-ROC Testing	
AUC for Logistic_Regression is:	0.88	AUC for Logistic_Regression is:	0.83
AUC for KNN is:	0.99	AUC for KNN is:	0.96
AUC for Bagging is:	0.99	AUC for Bagging is:	0.95
AUC for XGBoost is:	1.0	AUC for XGBoost is:	0.99
AUC for Support_Vector_Machine is:	0.87	AUC for Support_Vector_Machine is:	0.86
AUC for Naive_Bayes is:	0.84	AUC for Naive_Bayes is:	0.8
AUC for Decision_Tree is:	1.0	AUC for Decision_Tree is:	0.88
AUC for Extra_Trees_Classifier is:	1.0	AUC for Extra_Trees_Classifier is:	0.98
AUC for Linear_Discriminant_Analysis is:	0.9	AUC for Linear_Discriminant_Analysis is:	0.85

AUC-ROC curve states that how much models are able to learn and after learning how much data points are they able to capture.
Thus, from the above Graph 9 and table we can clearly see Extra Tree classifier, KNN and XGBoost is been able to capture most of the datapoint.

Lets check out the most important variable that companies need to focus on using Mututal Information Gain and SHAP values of all our 3 models

Mutual Information Gain:



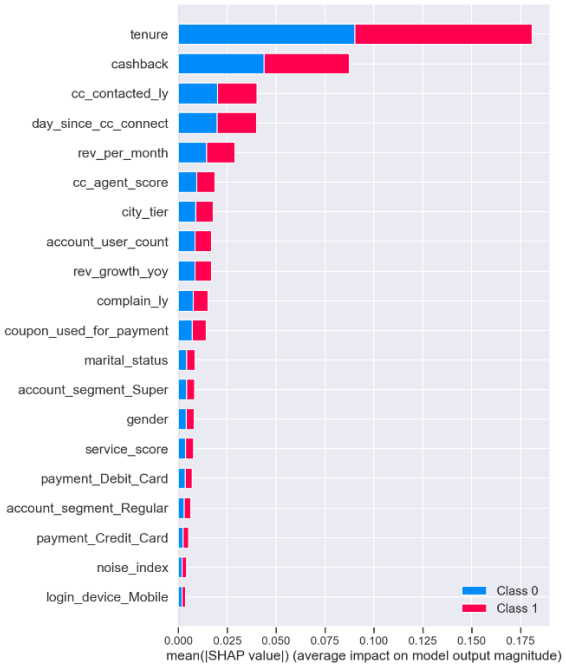
Graph 10: Mutual Information Gain

According to mutual information gain Company must focus on tenure, cc_contacted_ly, rev_per_month, cashback, day_since_cc_connected.

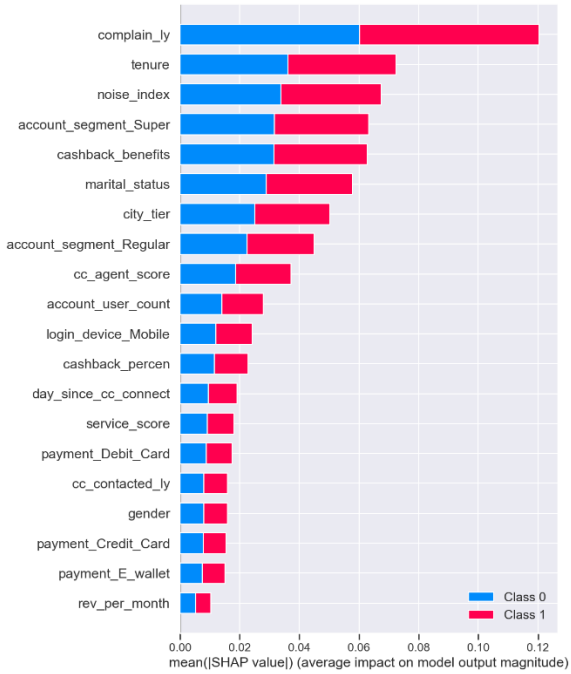
SHAP values (Top 3 models)

Following are the top 5 variable according the SHAP:

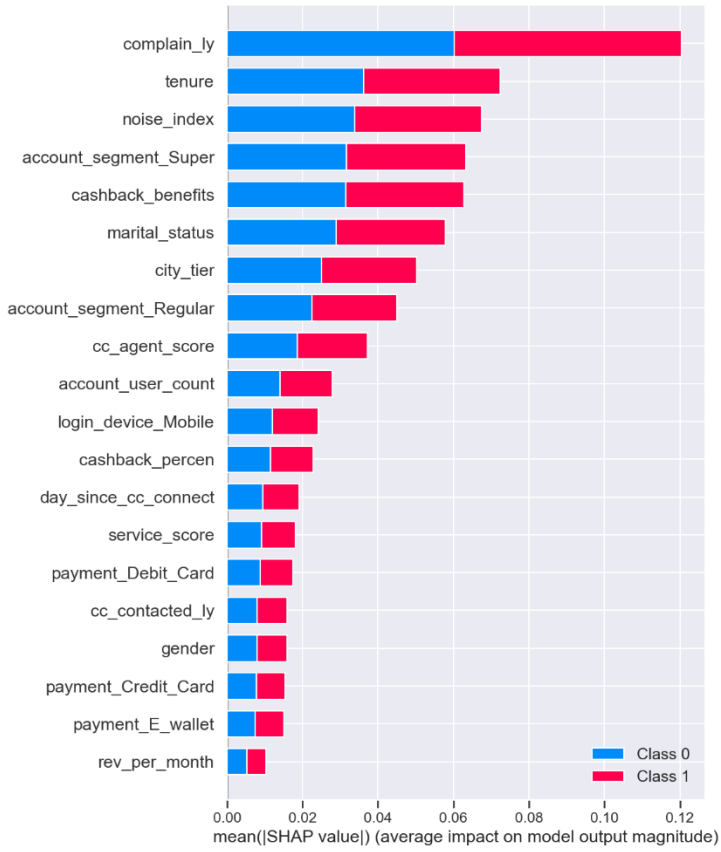
- Cashback
- Complain_ly
- Noise_Index
- Account_segment
- Day_since_cc_connect



KNN



XGBoost



Extra Tree Classifier

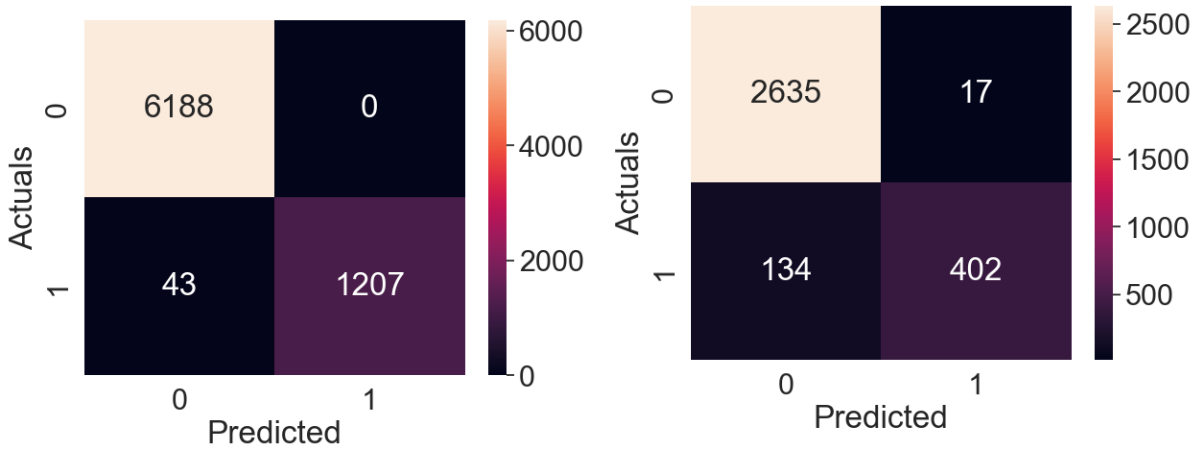
Graph 11: Top 3 models Shap Values

Final Model Selection (Extra Tree Classifier):

Out of all 9 models Extra Tree Classifier has shown the best results. Extra Tree Classifier shown better balance than KNN and XGBoost in terms of accuracy, recall, precision and F1-Score i.e . Moreover, Extra Tree Classifier had AUC test score 99% which is higher than both KNN and XGBoost.

```
Confusion matrix Train :
[[6188    0]
 [  43 1207]]

Classification Report Train :
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      6188
           1       1.00      0.97      0.98      1250

    accuracy                           0.99      7438
   macro avg       1.00      0.98      0.99      7438
weighted avg       0.99      0.99      0.99      7438


Accuracy Score Train :  0.9942188760419468

Confusion matrix Train :
[[6188    0]
 [  43 1207]]

Classification Report Train :
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      6188
           1       1.00      0.97      0.98      1250

    accuracy                           0.99      7438
   macro avg       1.00      0.98      0.99      7438
weighted avg       0.99      0.99      0.99      7438
```

Let’s check the confusion matrix:



Graph 12: Extra Tree Classifier Confusion Matrix

True Positive (TP) = 2635 are the True predicted positive values by model
True Negative (TN) = 402 are the True predicted negative values by model
False Positive (FP) = 17 are the False predicted positive values by model
False Negative (FN) = 134 are the False predicted negative values by model

Thus, Extra Tree Classifier is the Best Model.

Conclusion

E-commerce businesses are allocating huge amount of money to acquire new customers. However, customers lifetime depends on a lot of variables and this study was about building customer churn prediction model for e-commerce businesses The study started with exploratory analysis and data visualisations to increase our understanding to churned customers. It was noticed that churned customers associated with male gender, single marital status. Then, 9 different machine algorithms were applied to predict customer churn. It was found that Extra Tree Classifier has the best F1- score i.e 85% and AUC-ROC score is 99%.

Sections 5: Final Recommendation

After checking out the all the models and graphs, as Data Scientist Following are the variable company should focus on most:

- Complain_ly – Customer are call more than 5 times every month
- Cashback – To much cashback to regular customers is loss for company
- Nosie_Index – High Noise customers must be focus on
- Account_segment - Regular segment must been attended most as these are customer from which we generate revenue every month
- Cashback_Benefits : Regular customers are getting higher discounts and they are not satisfied with the customer care services thus they are Churning
- If company is giving discounts, then give High discounts to your Super and High Net Worth (HNI) customers as they are your loyal and long-term customers seeing such high discounts your regular customers may move to Super or HNI
- Account_Segment: Companies Monthly revenue rotation is usually generated from regular customers as Super and HNI uses yearly or 6 months plans. Rather than giving discount we can provide some add-on to our regular customers
- Day_since_cc_connect: Your regular customers when gets connected with agents and within 5 days they churn this is because the customers is unsatisfied with the service provided by the Agents. Thus, need to train their agents to give quality service to their customers.
- Service_Score: Its almost satisfactory for the customers and shows lack of customers relationship.

Bibliography

- How to keep OTT and DTH or Cable TV customers engaged in these uncertain times (https://brandequity.economictimes.indiatimes.com/news/media/how-to-keep-ott-and-dth-or-cable-tv-customers-engaged-in-these-uncertain-times/75770804)\
- Breaking the Back of Customer Churn (https://www.bain.com/insights/breaking-the-back-of-customer-churn/)
- Frank Ceballos (2019) An Intuitive Explanation of Random Forest and Extra Trees Classifiers (https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b)
- Jason Brownlee (April 22, 2020) How to Develop an Extra Trees Ensemble with Python (https://machinelearningmastery.com/extra-trees-ensemble-with-python/)
- Vinícius Trevisan (Jan 18, 2022) Using SHAP Values to Explain How Your Machine Learning Model Works (https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137)
- XGBoost: A Deep Dive into Boosting (https://medium.com/sfu-cspmp/xgboost-a-deep-dive-into-boosting-f06c9c41349)
- Machine Learning Basics with the K-Nearest Neighbors Algorithm (https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761)

Appendix :

1) Data Dictionary

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target) 0= Not churned, 1 = Churned
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 Agents 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
Noise Index	**Formula = CC_Contacted_LY/Tenure** # noise_index <= 1 in a month : Low noise # noise_index <=100 in a month : medium noise # noise_index > 100 in a month : high noise
cashback_percen	Formula = **rev_per_month * 12/cashback** # cashback_percen <= 0.5 : Low discount # cashback_percen <=1 : medium discount # cashback_percen > 1 : high discount

Cashback_Benefits	Formula = **Cashback/Tenure** These are the benefits received as long as customer stay with the company. If the value the very high customer has received a lot of benefits from company and check % of churned customers Low Benefit, medium benefit and High Benefit

2) Machine Learning Models:

- **Logistic Regression:** Logistic Regression is a machine learning algorithm that is used to predict the probability of an event occurring. It works by analyzing the relationship between a set of input variables (known as predictors) and a binary output variable (0 or 1). Logistic Regression is widely used in many industries, including finance, healthcare, and marketing.
- **KNN:** KNN (K-Nearest Neighbors) is a machine learning algorithm that can be used for classification and regression tasks. It works by finding the K nearest data points to a new data point and assigning the class or value of the majority of those K data points to the new data point.
- **Bagging (Random Forest)**: Bagging is a machine learning technique that involves creating multiple models and combining their predictions to improve accuracy. Random Forest is a type of bagging that creates many decision trees and randomly selects a subset of features for each tree. The final prediction is made by averaging the predictions of all the trees.
- **XGBoost**: XGBoost is a computer program that predicts things by combining many simple models to make a more accurate prediction. It's good at handling large amounts of data and finding important patterns. It's used in finance, healthcare, and marketing.
- **Support Vector Machine (SVM):** It is a machine learning algorithm that can be used for classification and regression tasks. It works by finding the best line or curve that separates two classes in a high-dimensional space. SVM is known for its ability to handle complex data and find the most important features for making accurate predictions.
- **Naive Bayes:** It is a machine learning algorithm that can be used for classification tasks. It works by assuming that the features are independent of each other and using Bayes' theorem to calculate the probability of each class given the features. Naive Bayes is known for its simplicity, efficiency, and effectiveness, especially in natural language processing tasks.
- **Decision Tree:** It is a machine learning algorithm that can be used for classification and regression tasks. It works by creating a tree-like model of decisions and their possible consequences based on a set of input features. Decision Tree is known for being easy to understand and interpret, and for its ability to handle both categorical and numerical data.
- **Extra Tree Classifier**: It is a machine learning algorithm that can be used for classification tasks. It works by creating many different decision trees and randomly selecting a subset of features for each tree. The final prediction is made by averaging the predictions of all the trees. Extra Tree Classifier is known for its ability to handle noisy data and reduce overfitting.
- **Linear Discriminant Analysis (LDA):** It is a machine learning algorithm that can be used for classification tasks. It works by finding the best linear combination of input features that separates two or more classes. LDA is known for its simplicity, effectiveness, and ability to handle high-dimensional data.

3) AUC-ROC Curve

The AUC (Area Under the Curve) and ROC (Receiver Operating Characteristic) Curve are tools for evaluating the performance of a machine learning classifier. The ROC Curve is a graph that shows the trade-off between true positive rate and false positive rate for different classification thresholds. The AUC is a measure of the overall performance of the classifier, with higher values indicating better performance. A perfect classifier would have an AUC of 1.0, while a random classifier would have an AUC of 0.5. The AUC and ROC Curve are commonly used in many applications, including medical diagnosis, fraud detection, and credit scoring.

4) Confusion Matrix

A Confusion Matrix is a table used to evaluate the performance of a machine learning classifier. It shows the number of true positive, true negative, false positive, and false negative predictions made by the classifier on a set of test data. By analyzing the Confusion Matrix, we can calculate metrics such as accuracy, precision, recall, and F1 score, which can provide insights into the performance of the classifier and help identify areas for improvement.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

True positive (TP) is when the classifier predicts a positive instance correctly,
True negative (TN) is when the classifier predicts a negative instance correctly,
False positive (FP) is when the classifier predicts a positive instance but it is actually negative
False negative (FN) is when the classifier predicts a negative instance but it is actually positive.

5) Accuracy, Precision, Recall and F1-Score

Accuracy, recall, precision, and F1-Score are all metrics used to evaluate the performance of a machine learning model.

Accuracy measures the proportion of correctly classified instances among all instances. It is calculated as the number of correct predictions divided by the total number of predictions. It is a good measure when the classes are balanced.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Recall measures the proportion of positive instances that are correctly identified by the model. It is calculated as the number of true positives divided by the sum of true positives and false negatives. It is a good measure when the cost of missing a positive instance is high.

$$Recall = \frac{TP}{TP + FN}$$

Precision measures the proportion of correctly classified positive instances among all instances predicted as positive. It is calculated as the number of true positives divided by the sum of true positives and false positives. It is a good measure when the cost of false positives is high.

$$Precision = \frac{TP}{TP + FP}$$

F1-Score is the harmonic mean of precision and recall. It is calculated as 2 * (precision * recall) / (precision + recall). It is a good measure when there is an imbalance in the number of positive and negative instances.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

6) VIF(Variance Inflation Factor)

A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

A rule of thumb for interpreting the variance inflation factor:

* 1 = not correlated.
* Between 1 and 5 = moderately correlated.
* Greater than 5 = highly correlated.