1. Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and analysis of An E-Commerce Company or DTH, 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. We are using the best module to find religion for customer churn.

Customer Churn is very expensive for any business or organization. A high Churn Rate requires a company to deal with the stress of doubling down to bring in new customers; just to stay afloat. Even a minuscule single-digit increase in the Churn Rate can seriously impede a company's growth rate and what's worse is that high Churn Rates are more likely to compound over time.

This is necessary for every organization to retain their customer to sustain in the market. Customer Churn can lead to financial loss and so we need to study this project.

- **Expose Product Weaknesses:** Churn Analysis plays a crucial role in revealing the patterns that indicate the common motivators for customers to part ways with your company. These could be anything from price productivity to poor product adoption. It is also instrumental in demonstrating the exact way customers engage with your product throughout their lifecycle. You can use this to maximize what your customers already love, and improve upon everything they don't.
- **Unearth Customer Opportunities:** Customer experience improvement comes with a need to understand the customer requirements and expectations at every stage of their journey and moulding your product accordingly. Churn Analysis depicts trends in customer behaviour at every touch point. Personalized engagement preferred by your customers allows you to make your customers feel valued and appreciated.

Executive Summary

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 recommendations.

Data Description

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

Tab_1.a

Conclusion | insight: There is always an event that serves as the trigger that urges you to start analyzing churn. For example, you might see a sudden jump in your churn rate in a given month, or perhaps your churn has been on a gradual increase for the past 8-9 months.

These are indications of red flags, which means that you need to dig in a little deeper to find out what's happening with your data. The trigger doesn't necessarily have to be negative. For instance, you might have just rolled out a product on boarding email series for a couple of new customers a few weeks ago, and you wish to see if it had any impact on your churn rate. This also falls under the purview of Churn Analysis.

It is suggested that you should turn to Churn Analysis whenever you notice something out of the ordinary, irrespective of whether that is good or bad or if you can make a change that could potentially impact churn.

Data Report

Conclusion | insight:

Data collection is defined as the procedure of collecting, measuring, and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research.

Data Collection Methods: Phone vs. Online vs. In-Person Interviews

Essentially there are four choices for data collection – in-person interviews, mail, phone and online. There are pros and cons to each of these modes.

In-Person Interviews

Pros: In-depth and a high degree of confidence on the data
Cons: Time consuming, expensive and can be dismissed as anecdotal

Mail Surveys

Pros: Can reach anyone and everyone – no barrier
Cons: Expensive, data collection errors, lag time

Phone Surveys

Pros: High degree of confidence in the data collected, reach almost anyone
Cons: Expensive, cannot self-administer, need to hire an agency

Web/Online Surveys

Pros: Cheap, can self-administer, very low probability of data errors
Cons: Not all your customers might have an email address/be on the internet, customers may be wary of divulging information online.

In-person interviews always are better, but the big drawback is the trap you might fall into if you don't do them regularly. It is expensive to regularly conduct interviews and not conducting enough interviews might give you false positives. Validating your research is almost as important as designing and conducting it. We've seen many instances where after the research is conducted – if the results do not match up with the "gut-feel" of upper management, it has been dismissed as anecdotal and a "one-time" phenomenon. To avoid such traps, we strongly recommend that data collection be done on an "ongoing and regular" basis. This will help you in comparing and analysing the change in perceptions according to marketing done for your products/services. The other issue here is the sample size. To be confident with your research you have to interview enough people to weed out the fringe elements.

- **Visual inspection variables:-**

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
AccountID	20000	20001	20002	20003	20004	20005	20006	20007	20008	20009	20010	20011	20012	20013	2001
Churn	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Tenure	4	0	0	0	0	0	2	0	13	0	4	11	#	0	
City_Tier	3	1	1	3	1	1	3	1	3	1	NaN	1	1	1	
CC_Contacted_LY	6	8	30	15	12	22	11	6	9	31	18	6	11	15	1
Payment	Debit Card	UPI	Debit Card	Debit Card	Credit Card	Debit Card	Cash on Delivery	Credit Card	E wallet	Debit Card	Cash on Delivery	Debit Card	Cash on Delivery	Credit Card	Credi Car
Gender	Female	Male	Male	Male	Male	Female	Male	Male	Male	Male	F	Male	Male	Male	Mal
Service_Score	3	3	2	2	2	3	2	3	2	2	2	3	2	3	
Account_user_count	3	4	4	4	3	NaN	3	3	4	5	3	4	3	4	
account_segment	Super	Regular Plus	Regular Plus	Super	Regular Plus	Regular Plus	Super	Regular Plus	Regular Plus	Regular Plus	Regular	HNI	Regular Plus	Regular Plus	HN
CC_Agent_Score	2	3	3	5	5	5	2	2	3	3	3	3	3	3	
Marital_Status	Single	Single	Single	Single	Single	Single	Divorced	Divorced	Divorced	Single	Divorced	Single	Single	Divorced	Singl
rev_per_month	9	7	6	8	3	2	4	3	2	2	2	10	2	1	
Complain_ly	1	1	1	0	0	1	0	1	1	0	0	1	1	1	
rev_growth_yoy	11	15	14	23	11	22	14	16	14	12	12	13	13	17	1
coupon_used_for_payment	1	0	0	0	1	4	0	2	0	1	9	0	2	0	
Day_Since_CC_connect	5	0	3	3	3	7	0	0	2	1	8	0	2	0	
cashback	159.93	120.9	NaN	134.07	129.6	139.19	120.86	122.93	126.83	122.93	295.45	153.81	134.41	133.88	196.1
Login_device	Mobile	Mobile	Mobile	Mobile	Mobile	Computer	Mobile	Mobile	Mobile	Mobile	Mobile	Mobile	Mobile	Mobile	Mobil

fig_1.b

- **Row:-**

```
AccountID                11260
Churn                    11260
Tenure                   11158
City_Tier                11148
CC_Contacted_LY          11158
Payment                  11151
Gender                   11152
Service_Score            11162
Account_user_count       11148
account_segment          11163
CC_Agent_Score           11144
Marital_Status           11048
rev_per_month            11158
Complain_ly              10903
rev_growth_yoy           11260
coupon_used_for_payment  11260
Day_Since_CC_connect     10903
cashback                 10789
Login_device             11039
```

- **Columns: - 19**

```
11260, 19
```

```
['Churn', 'Tenure', 'City_Tier', 'CC_Contacted_LY', 'Payment', 'Gender',
 'Service_Score', 'Account_user_count', 'account_segment',
 'CC_Agent_Score', 'Marital_Status', 'rev_per_month', 'Complain_ly',
 'rev_growth_yoy', 'coupon_used_for_payment', 'Day_Since_CC_connect',
 'cashback', 'Login_device'],
dtype='object')
```

- **Descriptive details:-**

	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626350	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.0	0.00	1.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.0	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.0	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.0	3.00	5.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.0	4.00	5.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.0	1.00	1.0

fig_1.c

Conclusion | insight: -

1. There are a total of 11260 rows and 19 feature columns available in the data set.

2. CC_Contacted_LY is having the highest mean and standard deviation which can be observed from descriptive details.

Understanding of attributes (variable info, renaming if required)

- **Variable info:-**

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

fig_1.d

- **Objects detail :-**

i.) Login_device

```
Login_device
Mobile      7482
Computer    3018
&&&&         539
dtype: int64
```

ii.) Account_segment

```
account_segment
Super          4062
Regular Plus   3862
HNI            1639
Super Plus      771
Regular         520
Regular +       262
Super +          47
```

iii.) Gender

Gender	
Male	6328
Female	4178
M	376
F	270

iv.) Payment

Payment	
Debit Card	4587
Credit Card	3511
E wallet	1217
Cash on Delivery	1014
UPI	822

v) Marital_Status

Marital_Status	
Married	5860
Single	3520
Divorced	1668

fig_1.e

Conclusion | insight: No columns Name changes required. All detail is mentioned in variable info and all attributes logic_devies, are three attributes two mobile & computer error &&&& are shown 2nd account_segment all are ok and Gender detail is same represents to MALE & FEMALE are show 2 time M& F . Payment detail are ok. Then merital_status data variables are ok.

1. All details are mentioned in variable info. As we can see there are 12 features which are of object data type and the rest are either int64 of float64 type data.
2. We have found the count for different marital statuses, count for different payment methods, and count of different gender, count of different account segments, and count of different login devices.
3. We can see that maximum people are logged in a mobile device, maximum people are having super account segment and most people are doing payment through debit card. These are the observations.

2. EDA and Business Implication

Univariate analysis

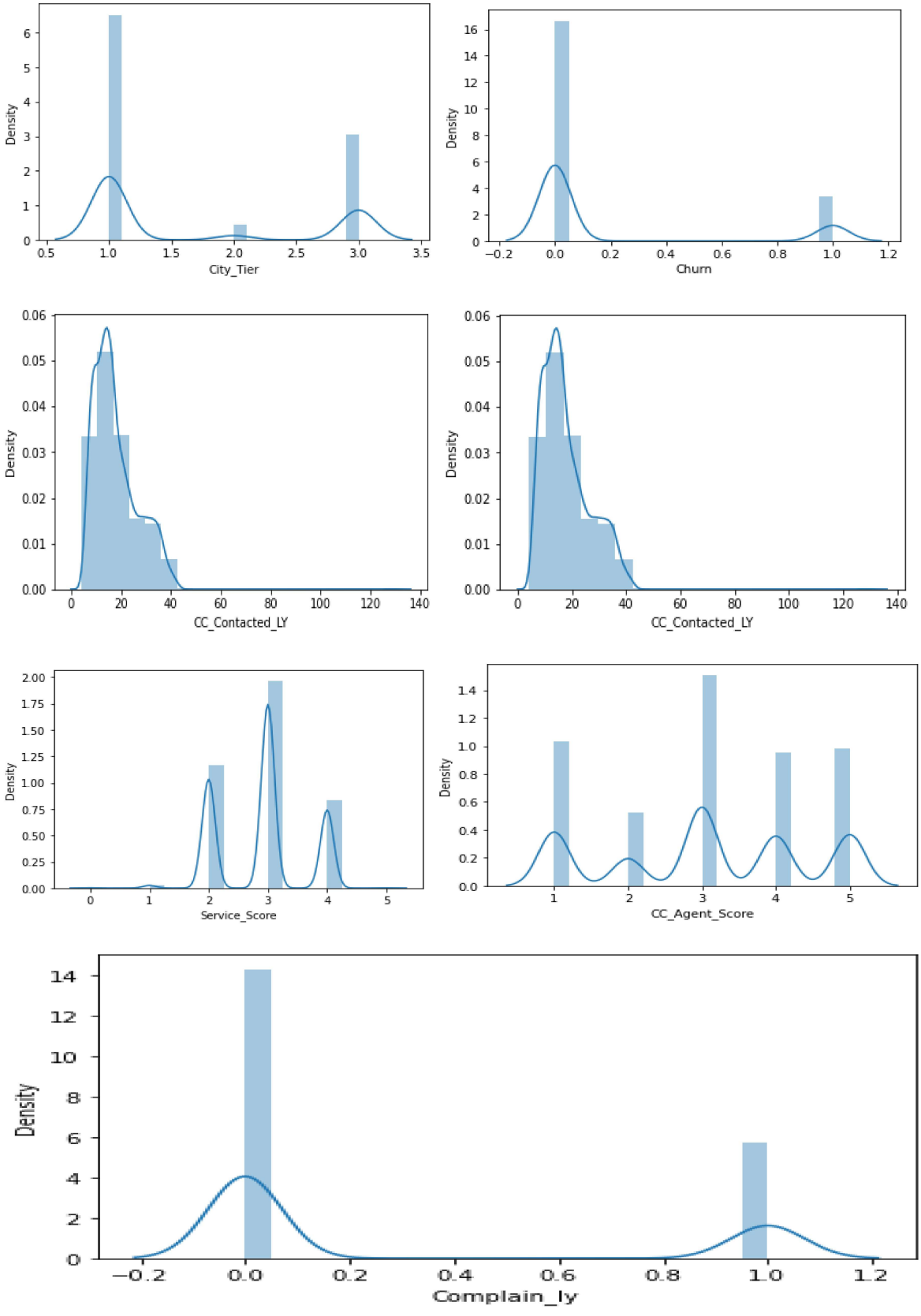

fig_2.a

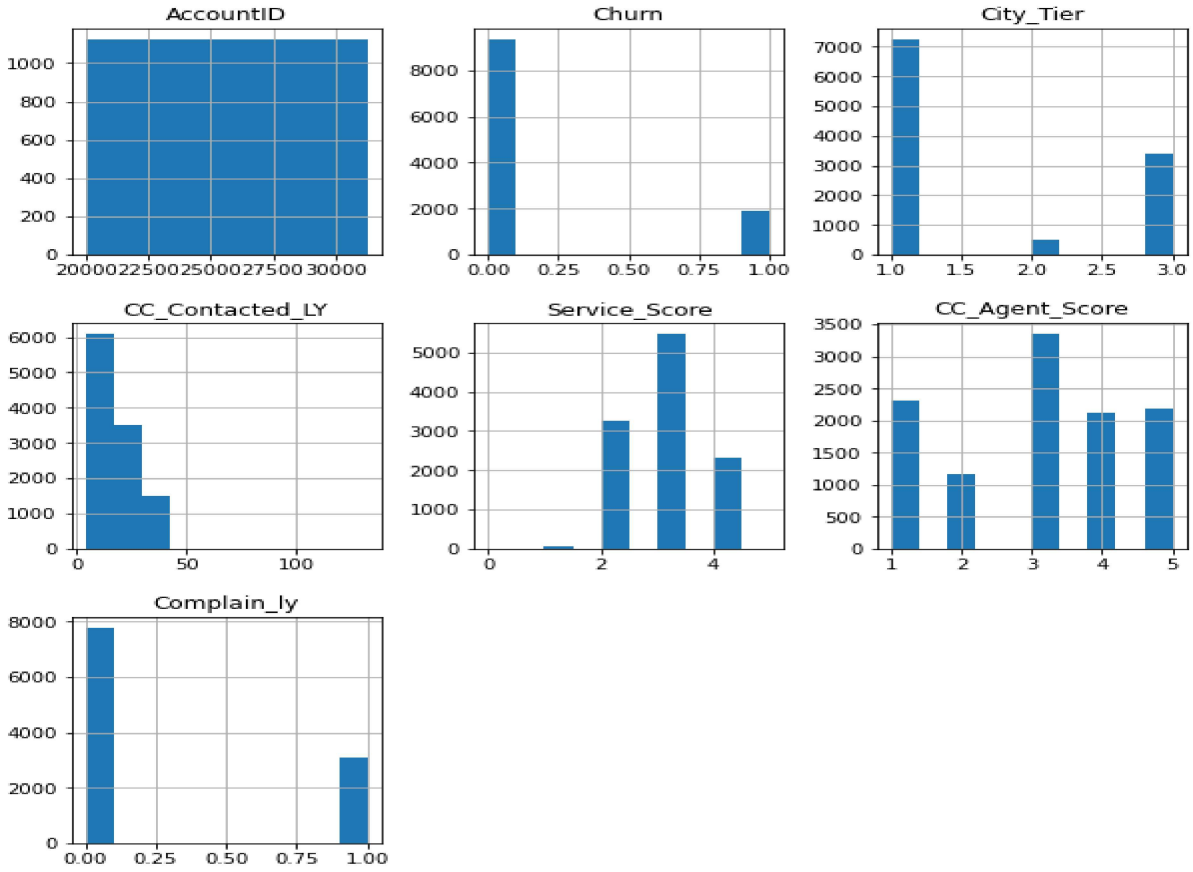

Fig_2.b

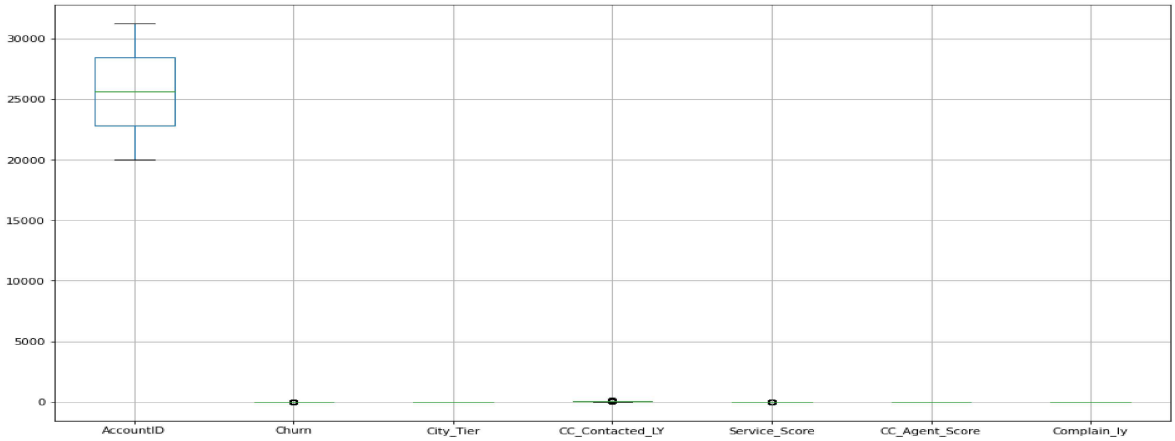

fig_2.c

Conclusion | insight: All columns data single behaviour show in Fig_2.a & Fig_2.b, Fig_2.C.

1. We have a count plot for our target variable churn and we can observe a high no. of churn as compared to non-churn.
2. It is found that customer churn is higher in Tier 1 cities and lowest in tier 2 cities. Tier 1 customer needs more focus to be retained with the organization.
3. It is found that the count for customer care contacted is very high who churn.

Bivariate Analysis

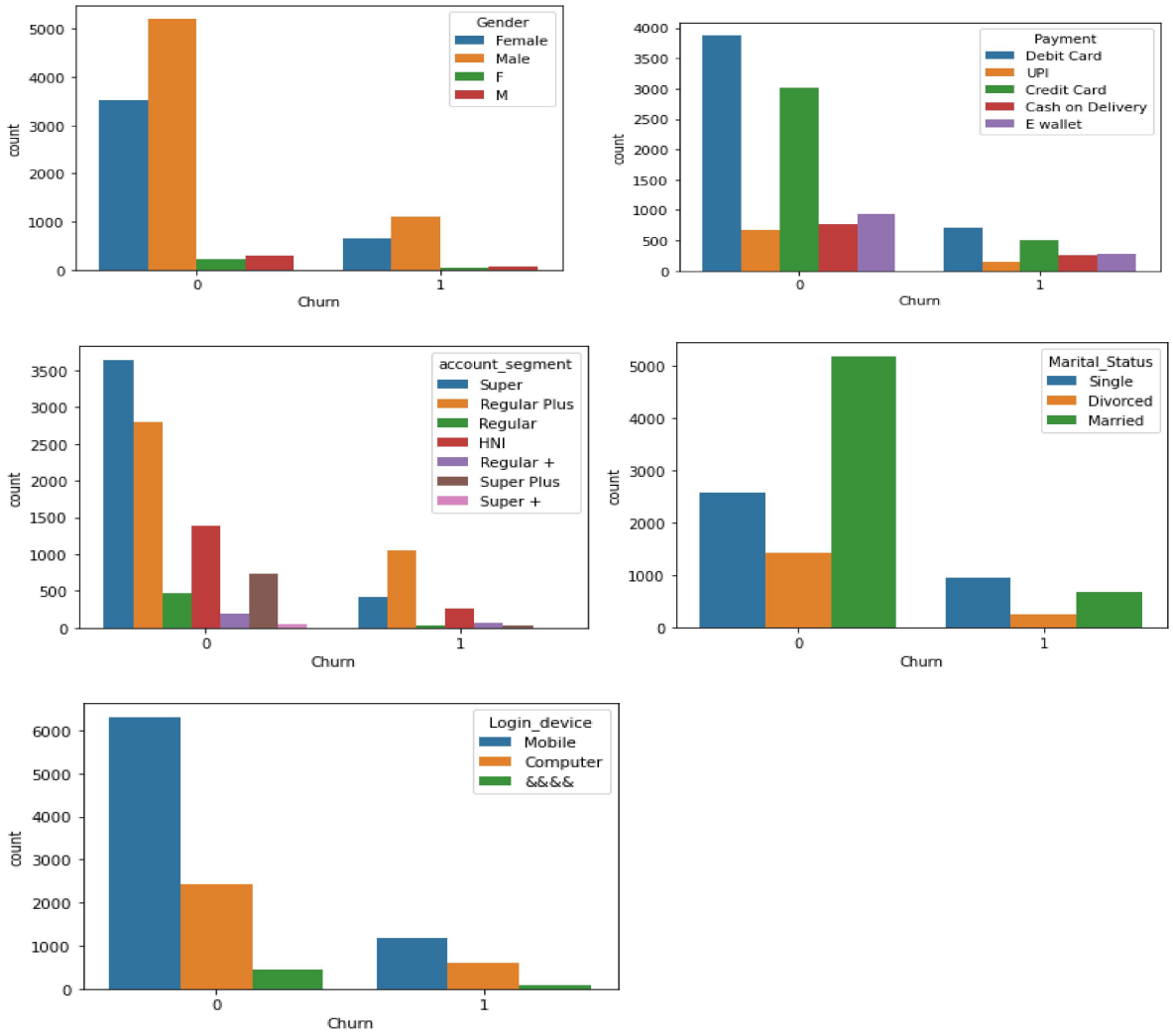

Fig_3.f

Conclusion | insight:

1. Here in Bivariate Analysis, we have obtained churn count with respect to Gender (Male & Female), Payment method (Debit card, UPI, Credit card, Cash on Delivery and E wallet),

account segment and marital status (single, Divorced and Married).

2. We can clearly see that churn count is higher for male in Gender and churn count is higher for married people in marital status as compared to other categories.

If we look for different account segment then churn rate is maximum for super account segment and it is minimum in super+ plan. So we can say that Super plan is most disliked and Super + plan is most liked by customer and we should spread it more across.

Multi-variate analysis

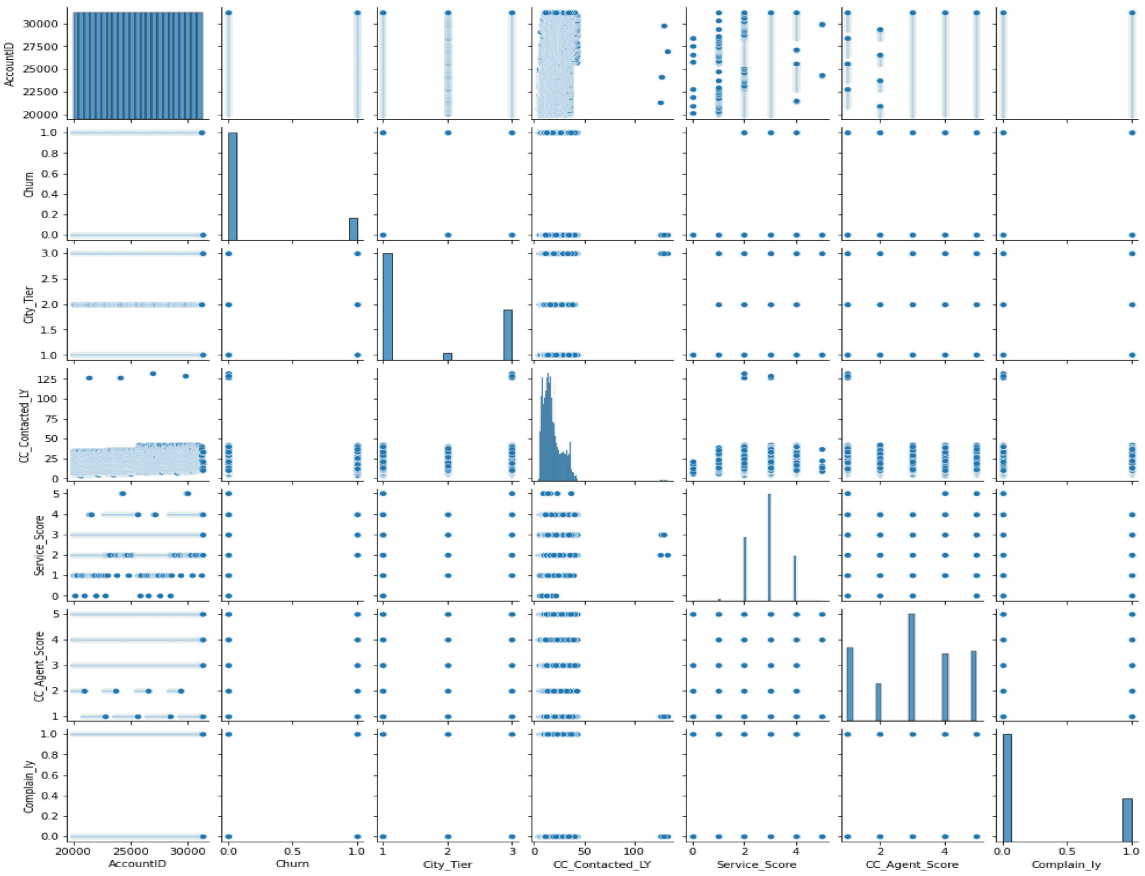

Fig_3.p

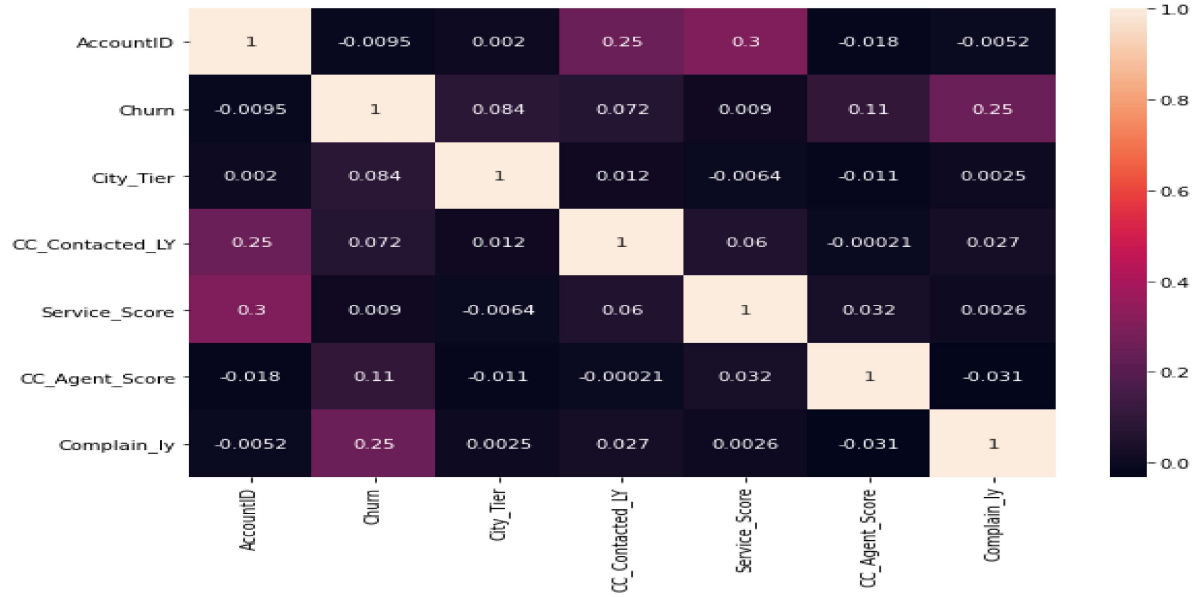

Fig_3.k

Conclusion | insight: Correlation of all two data behaviour's a shown in fig_3.f fig_3.p, fig_3.k. The Churn rate is higher for customers who logged in with mobile and also a maximum number of customers are logged in with mobile as compared to the rest lodevicesvice.

3. Data Cleaning and Pre-processing

Removal of unwanted variables

- **Before Removal column**

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
AccountID	20000	20001	20002	20003	20004	20005	20006	20007	20008	20009	20010	20011	20012	20013	2001
Churn	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Tenure	4	0	0	0	0	0	2	0	13	0	4	11	#	0	9
City_Tier	3	1	1	3	1	1	3	1	3	1	NaN	1	1	1	3
CC_Contacted_LY	6	8	30	15	12	22	11	6	9	31	18	6	11	15	1
Payment	Debit Card	UPI	Debit Card	Debit Card	Credit Card	Debit Card	Cash on Delivery	Credit Card	E wallet	Debit Card	Cash on Delivery	Debit Card	Cash on Delivery	Credit Card	Credi Car
Gender	Female	Male	Male	Male	Male	Female	Male	Male	Male	Male	F	Male	Male	Male	Mal
Service_Score	3	3	2	2	2	3	2	3	2	2	2	3	2	3	3
Account_user_count	3	4	4	4	3	NaN	3	3	4	5	3	4	3	4	4
account_segment	Super	Regular Plus	Regular Plus	Super	Regular Plus	Regular Plus	Super	Regular Plus	Regular Plus	Regular Plus	Regular	HNI	Regular Plus	Regular Plus	HN
CC_Agent_Score	2	3	3	5	5	5	2	2	3	3	3	3	3	3	2
Marital_Status	Single	Single	Single	Single	Single	Single	Divorced	Divorced	Divorced	Single	Divorced	Single	Single	Divorced	Singl
rev_per_month	9	7	6	8	3	2	4	3	2	2	2	10	2	1	2
Complain_ly	1	1	1	0	0	1	0	1	1	0	0	1	1	1	(
rev_growth_yoy	11	15	14	23	11	22	14	16	14	12	12	13	13	17	1
coupon_used_for_payment	1	0	0	0	1	4	0	2	0	1	9	0	2	0	(
Day_Since_CC_connect	5	0	3	3	3	7	0	0	2	1	8	0	2	0	7
cashback	159.93	120.9	NaN	134.07	129.6	139.19	120.86	122.93	126.83	122.93	295.45	153.81	134.41	133.88	196.1
Login_device	Mobile	Mobile	Mobile	Mobile	Mobile	Computer	Mobile	Mobile	Mobile	Mobile	Mobile	Mobile	Mobile	Mobile	Mobil

Fig_3.v

- **After Removal column**

	0	1	2	3	4
Churn	1	1	1	1	1
Tenure	4	0	0	0	0
City_Tier	3	1	1	3	1
CC_Contacted_LY	6	8	30	15	12
Payment	Debit Card	UPI	Debit Card	Debit Card	Credit Card
Gender	Female	Male	Male	Male	Male
Service_Score	3	3	2	2	2
Account_user_count	3	4	4	4	3
account_segment	Super	Regular Plus	Regular Plus	Super	Regular Plus
CC_Agent_Score	2	3	3	5	5
Marital_Status	Single	Single	Single	Single	Single
rev_per_month	9	7	6	8	3
Complain_ly	1	1	1	0	0
rev_growth_yoy	11	15	14	23	11
coupon_used_for_payment	1	0	0	0	1
Day_Since_CC_connect	5	0	3	3	3
cashback	159.93	120.9	NaN	134.07	129.6
Login_device	Mobile	Mobile	Mobile	Mobile	Mobile

Fig_3.l

Conclusion | insight: - Only one column are deleted “AccountID” shown in fig_3.l

Missing Value treatment

All missing value	Percentage of missing value /Per columns

	count	percentage
Churn	0	0.00
Tenure	102	0.91
City_Tier	112	0.99
CC_Contacted_LY	102	0.91
Payment	109	0.97
Gender	108	0.96
Service_Score	98	0.87
Account_user_count	112	0.99
account_segment	97	0.86
CC_Agent_Score	116	1.03
Marital_Status	212	1.88
rev_per_month	102	0.91
Complain_ly	357	3.17
rev_growth_yoy	0	0.00
coupon_used_for_payment	0	0.00
Day_Since_CC_connect	357	3.17
cashback	471	4.18
Login_device	221	1.96

Fig_4.a

Percentage of missing value pic data:-

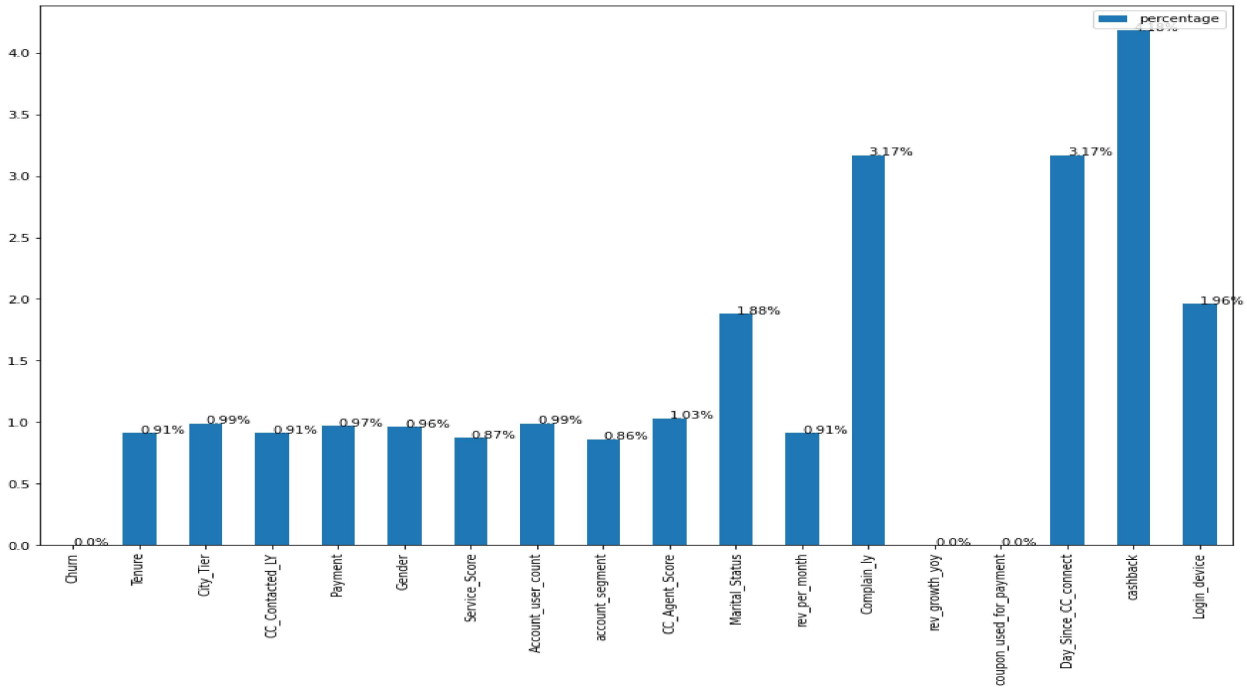

Fig_4.b

- After Treatment Missing Values

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

Fig_4.d

Conclusion | insight:- all missing value are deleted shown in fig_4.d

Outlier treatment

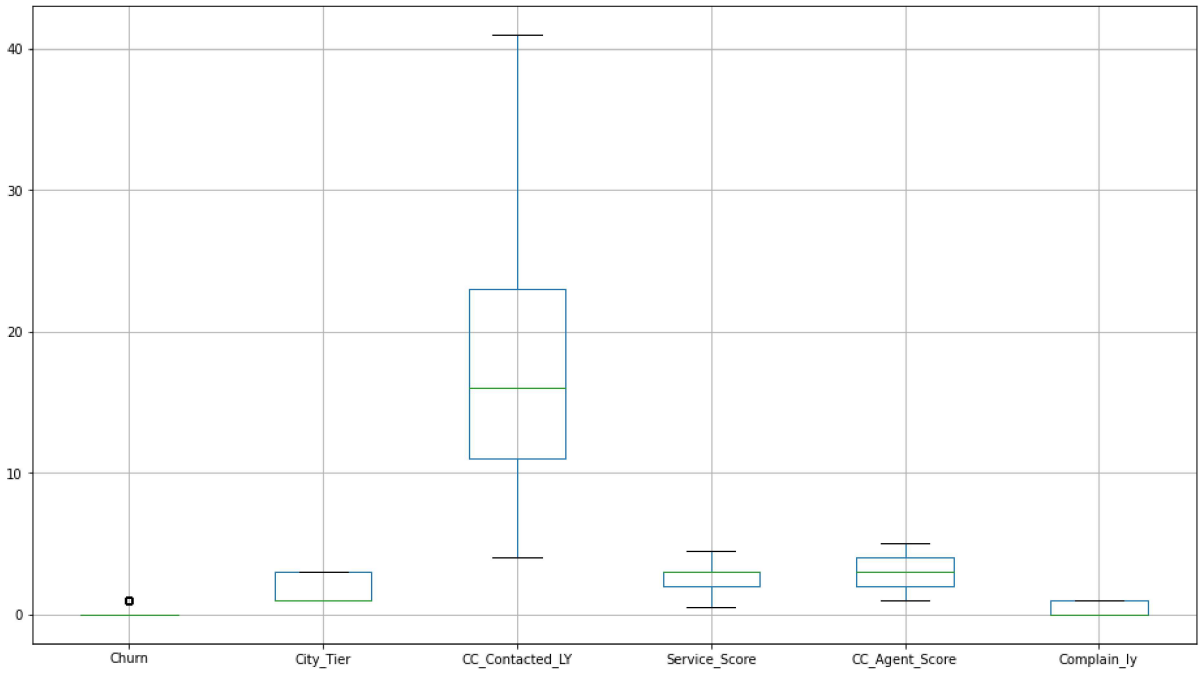

Fig_4.m

Conclusion | insight:- All outlier treated now no any outliner available show in fig_4.m

Variable transformation

feature: Tenure
[4, 0, 2, 13, 11, ..., 50, 60, 31, 51, 61]
Length: 39
Categories (39, object): [0, 1, 2, 3, ..., 61, 99, '#', 'No_info']
[4 0 2 13 11 37 9 36 19 20 14 8 26 18 5 30 7 1 23 3 29 6 28 24
25 16 10 15 22 38 27 12 21 17 32 34 31 33 35]
feature: Payment
['Debit Card', 'UPI', 'Credit Card', 'Cash on Delivery', 'E wallet', 'No_info']
Categories (6, object): ['Cash on Delivery', 'Credit Card', 'Debit Card', 'E wallet', 'No_info', 'UPI']
[2 5 1 0 3 4]
feature: Gender
['Female', 'Male', 'No_info']
Categories (3, object): ['Female', 'Male', 'No_info']
[0 1 2]
feature: Account_user_count
[3, 4, 'No_info', 5, 2, '@', 1, 6]
Categories (8, object): [1, 2, 3, 4, 5, 6, '@', 'No_info']
[2 3 7 4 1 6 0 5]
feature: account_segment
['Super', 'Regular Plus', 'Regular', 'HNI', 'Regular +', 'No_info', 'Super Plus', 'Super +']
Categories (8, object): ['HNI', 'No_info', 'Regular', 'Regular +', 'Regular Plus', 'Super', 'Super +', 'Super Plus']
[5 4 2 0 3 1 7 6]
feature: Marital_Status
['Single', 'Divorced', 'Married', 'No_info']
Categories (4, object): ['Divorced', 'Married', 'No_info', 'Single']
[3 0 1 2]
feature: rev_per_month
[9, 7, 6, 8, 3, ..., 135, 111, 109, 100, 103]
Length: 60
Categories (60, object): [1, 2, 3, 4, ..., 139, 140, '+', 'No_info']
[8 6 5 7 2 1 3 9 0 4 58 48 59 15 56 23 39 55 46 42 43 35 17 45
51 32 33 28 57 50 47 27 37 10 26 16 38 40 54 30 18 22 53 44 13 12 11 34
19 41 36 49 25 14 20 52 31 29 21 24]
feature: rev_growth_yoy
[11, 15, 14, 23, 22, ..., 26, '$', 4, 27, 28]
Length: 20
Categories (20, object): [4, 11, 12, 13, ..., 26, 27, 28, '$']
[1 5 4 13 12 6 2 3 7 8 14 9 10 11 15 16 19 0 17 18]
feature: coupon_used_for_payment
[1, 0, 4, 2, 9, ..., '#', '$', 14, '*', 16]
Length: 20
Categories (20, object): [0, 1, 2, 3, ..., 16, '#', '$', '*']
[1 0 4 2 9 6 11 7 12 10 5 3 13 15 8 17 18 14 19 16]
feature: Day_Since_CC_connect
[5, 0, 3, 7, 2, ..., '$', 46, 18, 31, 47]
Length: 25
Categories (25, object): [0, 1, 2, 3, ..., 46, 47, '$', 'No info']

[5 0 3 7 2 1 8 6 4 15 24 11 10 9 13 12 17 16 14 19 23 21 18 20
22]
feature: cashback
[159.93, 120.9, 'No_info', 134.07, 129.6, ..., 240.54, 252.31, 227.36, 226.91, 191.42]
Length: 5694
Categories (5694, object): [0, 2, 3, 4, ..., 1992, 1997, '$', 'No_info']
[2222 109 5693 ... 4488 4481 3680]
feature: Login_device
['Mobile', 'Computer', '&&&&', 'No_info']
Categories (4, object): ['&&&&', 'Computer', 'Mobile', 'No_info']
[2 1 0 3]

Tab_1.b

	0	1	2	3	4
Churn	1.0	1.0	1.0	1.0	1.0
Tenure	4.0	0.0	0.0	0.0	0.0
City_Tier	3.0	1.0	1.0	3.0	1.0
CC_Contacted_LY	6.0	8.0	30.0	15.0	12.0
Payment	2.0	5.0	2.0	2.0	1.0
Gender	0.0	1.0	1.0	1.0	1.0
Service_Score	3.0	3.0	2.0	2.0	2.0
Account_user_count	2.0	3.0	3.0	3.0	2.0
account_segment	5.0	4.0	4.0	5.0	4.0
CC_Agent_Score	2.0	3.0	3.0	5.0	5.0
Marital_Status	3.0	3.0	3.0	3.0	3.0
rev_per_month	8.0	6.0	5.0	7.0	2.0
Complain_ly	1.0	1.0	1.0	0.0	0.0
rev_growth_yoy	1.0	5.0	4.0	13.0	1.0
coupon_used_for_payment	1.0	0.0	0.0	0.0	1.0
Day_Since_CC_connect	5.0	0.0	3.0	3.0	3.0
cashback	2222.0	109.0	5693.0	748.0	552.0
Login_device	2.0	2.0	2.0	2.0	2.0

Fig_4.j

Conclusion | insight:- All Row data converted shown in table Tab_1.b & fig_4.j .

4. Model building

Train your predictive model against the test set using various appropriate performance metrics.

Conclusion | insight:

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 to gather more information about the subjects. Following is the list of model building procedures that can be used in this project -

(1) Logistics Regression (2) Decision Tree (3) Random Forest (4) Linear Discriminant Analysis (5) K Nearest Neighbours (6) Naive Bayes (7) Gradient Boosting (8) Extreme Gradient Boosting (9) Extra Tree Classifier

For every model building procure, we have gone through the following steps to get outputs generated by the algorithm –

(a) Model Prediction

(b) Model Performance

(c) ROC-AUC Graph

(d) Model Performance Metrics

(1) Logistics Regression

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[6165,  235],
 [ 600,  697]],
```

classification_report

	precision	recall	f1-score	support
0	0.91	0.96	0.94	6400
1	0.75	0.54	0.63	1297
accuracy			0.89	7697
macro avg	0.83	0.75	0.78	7697
weighted avg	0.88	0.89	0.88	7697

ROC-AUC Graph

```
AUC: 0.891
```

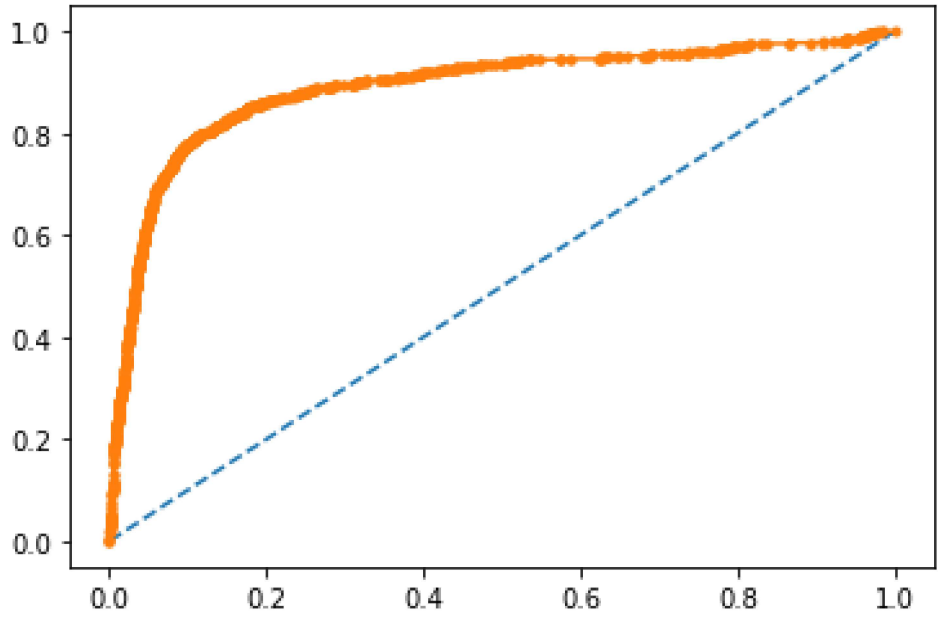

Fig-1.a.1

(2) Decision Tree

Train:- Graphviz

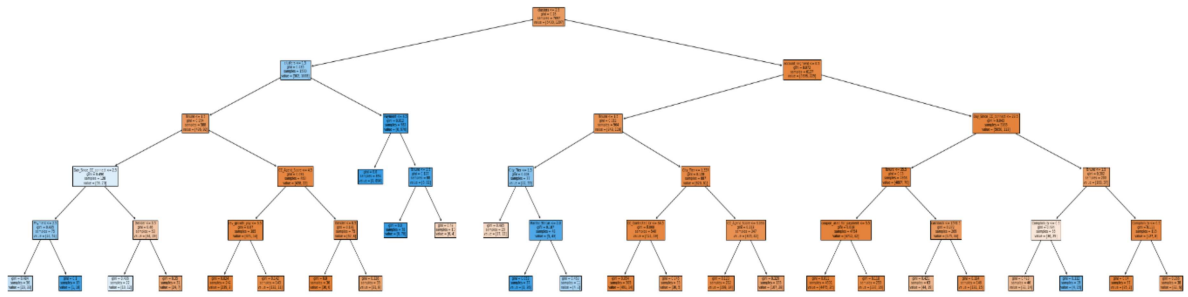

Fig-1.a.2

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[6357,   43],
 [ 204, 1093]],
```

classification_report

	precision	recall	f1-score	support
0	0.97	0.99	0.98	6400
1	0.96	0.84	0.90	1297
accuracy			0.97	7697
macro avg	0.97	0.92	0.94	7697
weighted avg	0.97	0.97	0.97	7697

ROC-AUC Graph

AUC: 0.977

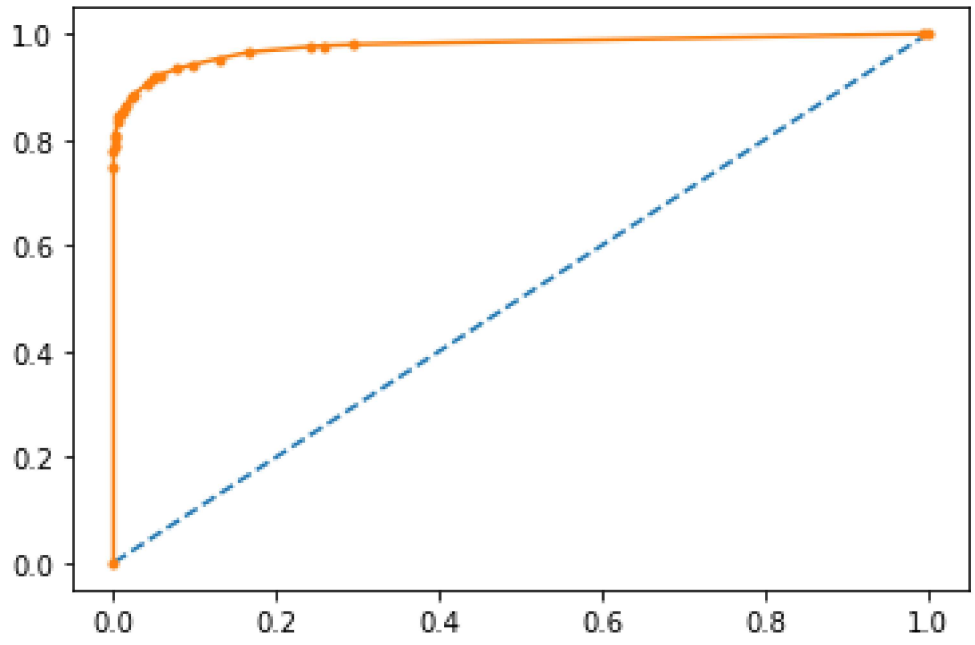

Fig-1.a.3

(3) Random Forest

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[6357,   43],
 [ 204, 1093]],
```

classification_report

	precision	recall	f1-score	support
0	0.97	0.99	0.98	6400
1	0.96	0.84	0.90	1297
accuracy			0.97	7697
macro avg	0.97	0.92	0.94	7697
weighted avg	0.97	0.97	0.97	7697

ROC-AUC Graph

AUC: 0.9931

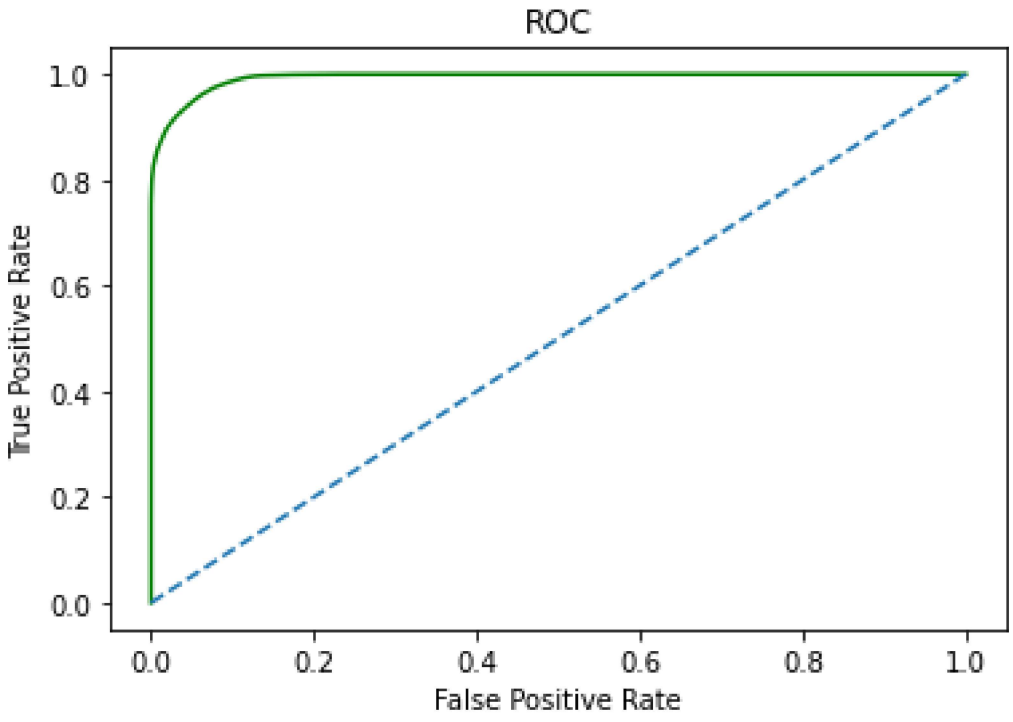

Fig-1.a

(4) Linear Discriminant Analysis

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[5966,  434],
 [ 495,  802]]
```

classification_report

	precision	recall	f1-score	support
0	0.92	0.93	0.93	6400
1	0.65	0.62	0.63	1297
accuracy			0.88	7697
macro avg	0.79	0.78	0.78	7697
weighted avg	0.88	0.88	0.88	7697

ROC-AUC Graph

AUC: 0.977

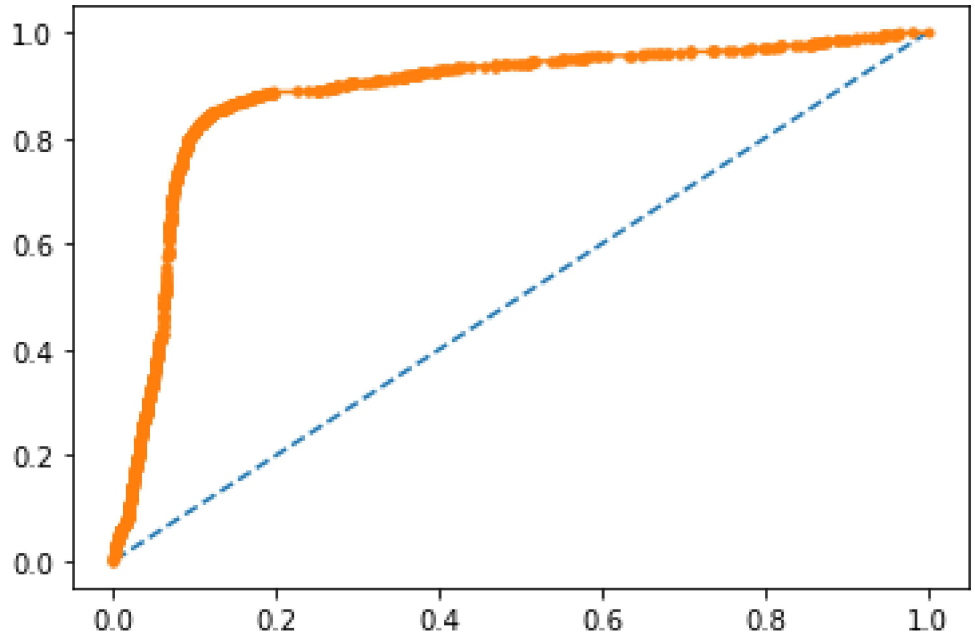

Fig-1.a.5

(5) K Nearest Neighbours

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[6310,   90]
[1105,  192]
```

classification_report

	precision	recall	f1-score	support
0	0.84	0.98	0.91	2745
1	0.45	0.06	0.11	554
accuracy			0.83	3299
macro avg	0.65	0.52	0.51	3299
weighted avg	0.77	0.83	0.77	3299

ROC-AUC Graph

AUC: 0.8406

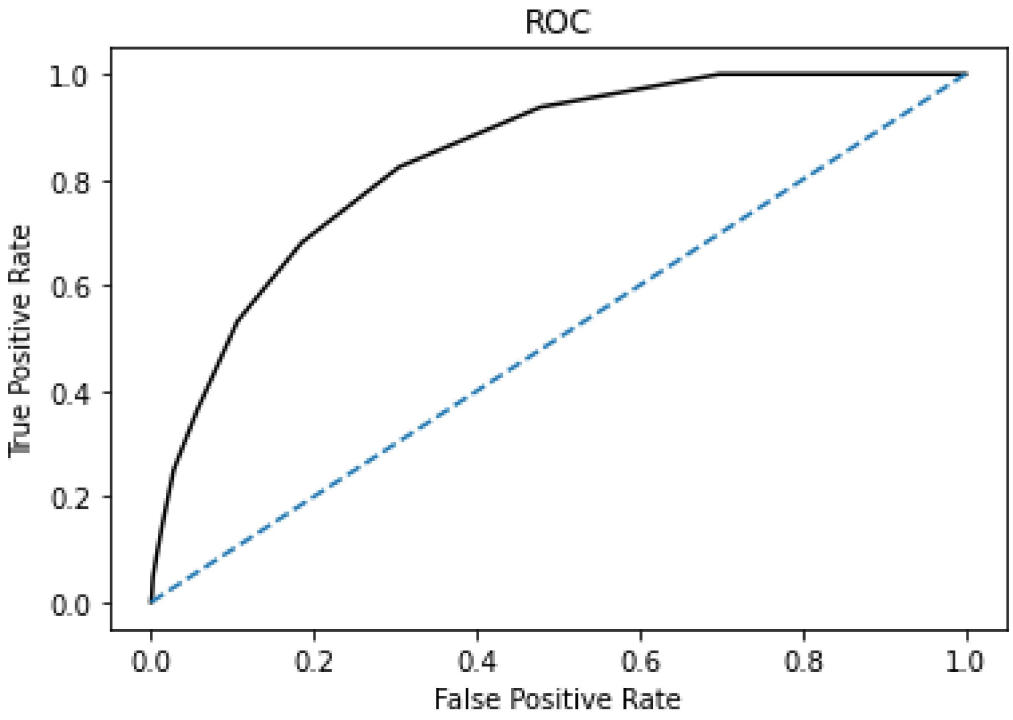

Fig-1.a.6

(6) Naive Bayes

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[6014,  386],
 [ 398,  899]],
```

classification_report

	precision	recall	f1-score	support
0	0.94	0.94	0.94	6400
1	0.70	0.69	0.70	1297
accuracy			0.90	7697
macro avg	0.82	0.82	0.82	7697
weighted avg	0.90	0.90	0.90	7697

ROC-AUC Graph

AUC: 0.868

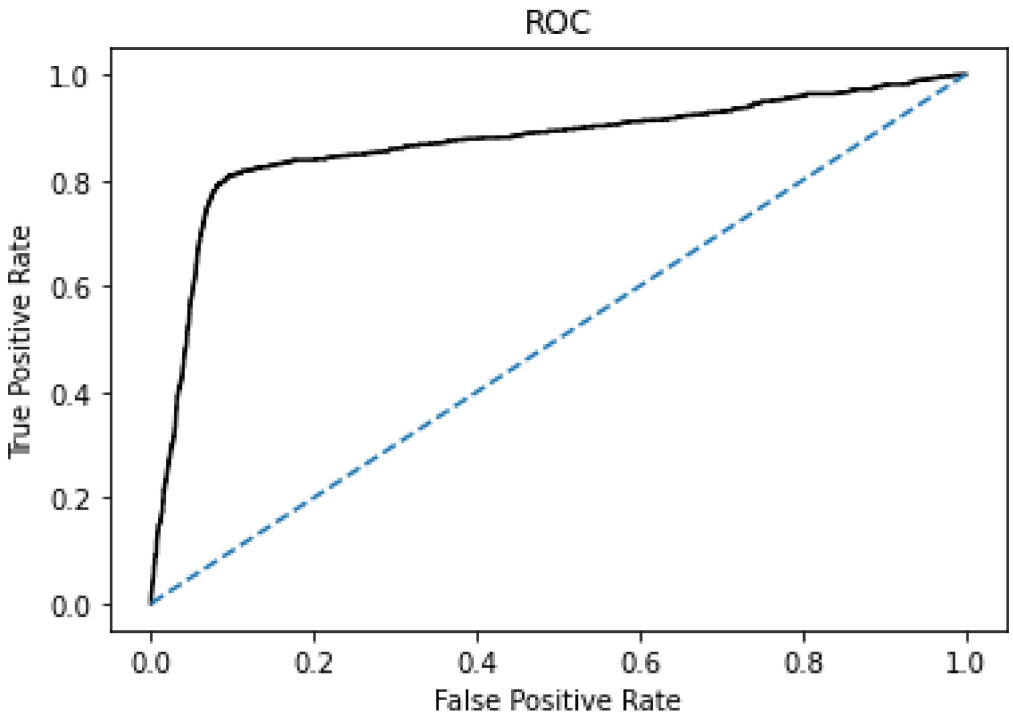

Fig-1.a.7

(7) Gradient Boosting

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[6400,    0],
[1297,    0]]
```

classification_report

	precision	recall	f1-score	support
0	0.83	1.00	0.91	6400
1	0.00	0.00	0.00	1297
accuracy			0.83	7697
macro avg	0.42	0.50	0.45	7697
weighted avg	0.69	0.83	0.75	7697

ROC-AUC Graph

AUC: 0.5

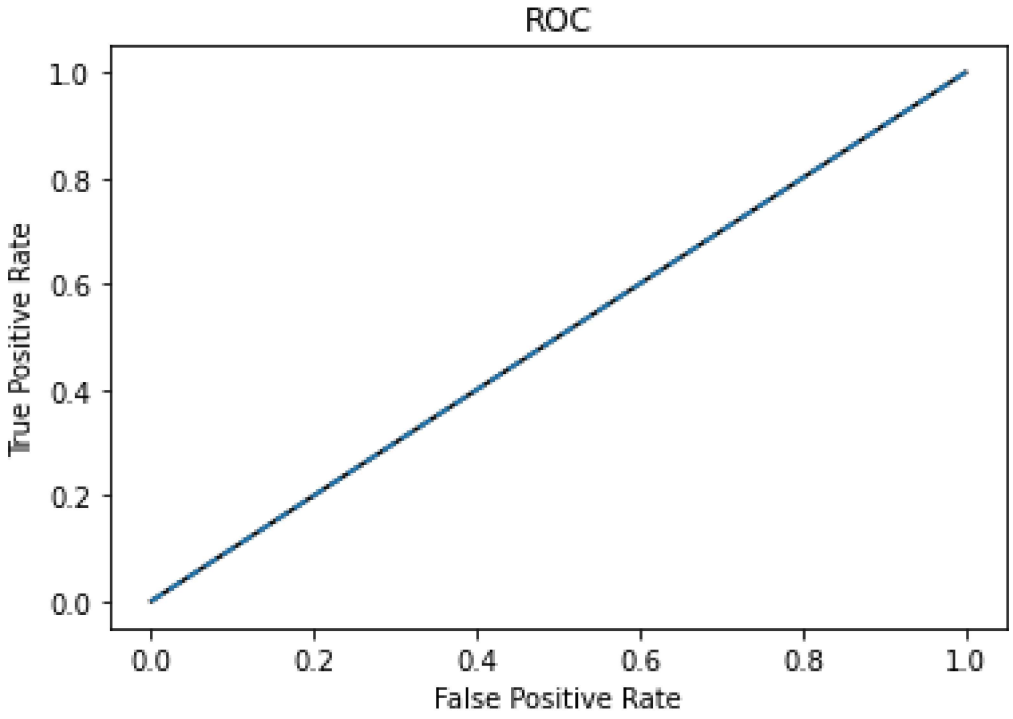

Fig-1.a.8

(8) Extreme Gradient Boosting

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[6400,    0],
 [  18, 1279]]
```

classification_report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6400
1	1.00	0.99	0.99	1297
accuracy			1.00	7697
macro avg	1.00	0.99	1.00	7697
weighted avg	1.00	1.00	1.00	7697

ROC-AUC Graph

AUC: 0.999

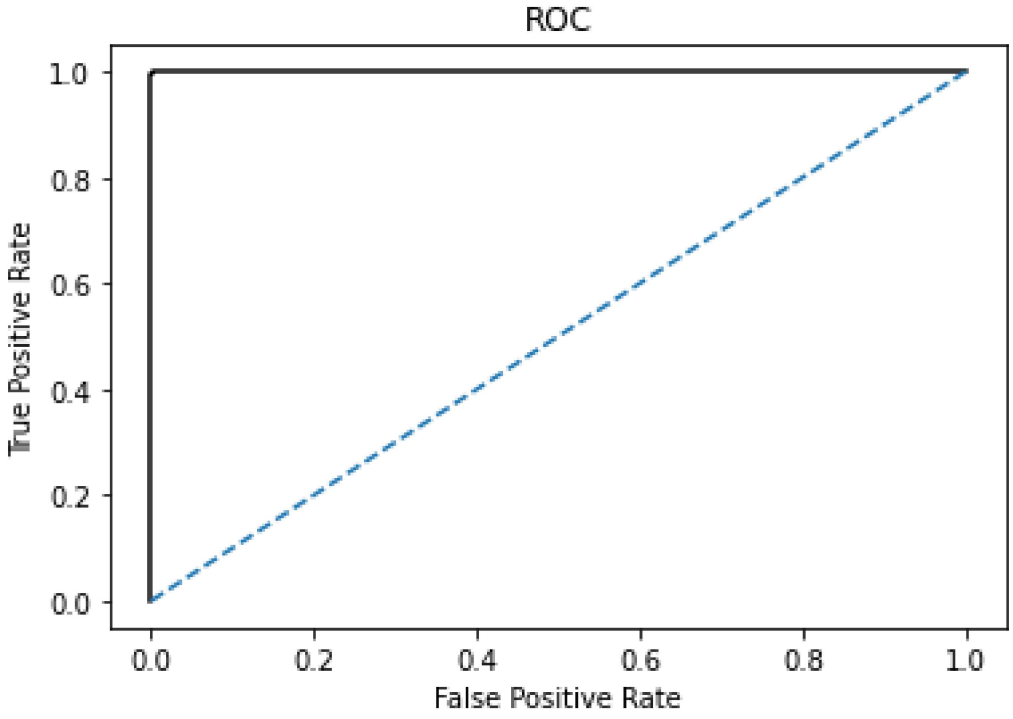

Fig-1.a.9

(9) Extra Tree Classifier

Train:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[6400,    0],
 [1297,    0]]
```

classification_report

	precision	recall	f1-score	support
0	0.83	1.00	0.91	6400
1	0.00	0.00	0.00	1297
accuracy			0.83	7697
macro avg	0.42	0.50	0.45	7697
weighted avg	0.69	0.83	0.75	7697

ROC-AUC Graph

AUC: 0.5

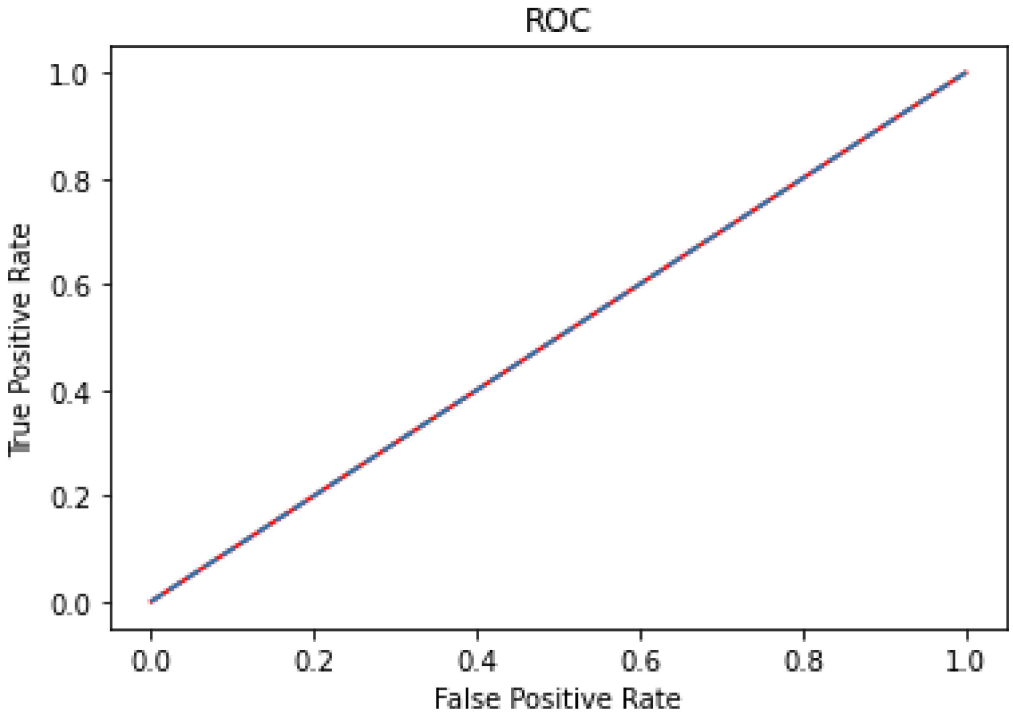

Fig-1.a.10

Conclusion | insight: all modal names and train modal confusion matrix and classification and accuracy are mentioned in all models.

1. The most important feature for determining customer churn for the random forest model is tenure i.e. for how long the customer is attached with the organization.

Why Random Forest chosen ?

Random Forest was chosen because this model outperforms rest all the models in all the performance metrics either it is accuracy, precision, f1-score, recall or AUC Score.

That's why the Random Forest model is chosen for the given dataset.

Efforts to improve model performance

1. We performed cross-validation 5 to find out the best parameters for the model as shown above to improve model performance.
2. We used the best parameters to build models and get the best performance metrics for the same.

Q 5. Model validation

Test your predictive model against the test set using various appropriate performance metrics.

Conclusion | insight:

(1) Logistics Regression

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2645,  100]
[ 270,  284]
```

classification_report

	precision	recall	f1-score	support
0	0.91	0.96	0.93	2745
1	0.74	0.51	0.61	554
accuracy			0.89	3299
macro avg	0.82	0.74	0.77	3299
weighted avg	0.88	0.89	0.88	3299

ROC-AUC Graph

AUC: 0.878

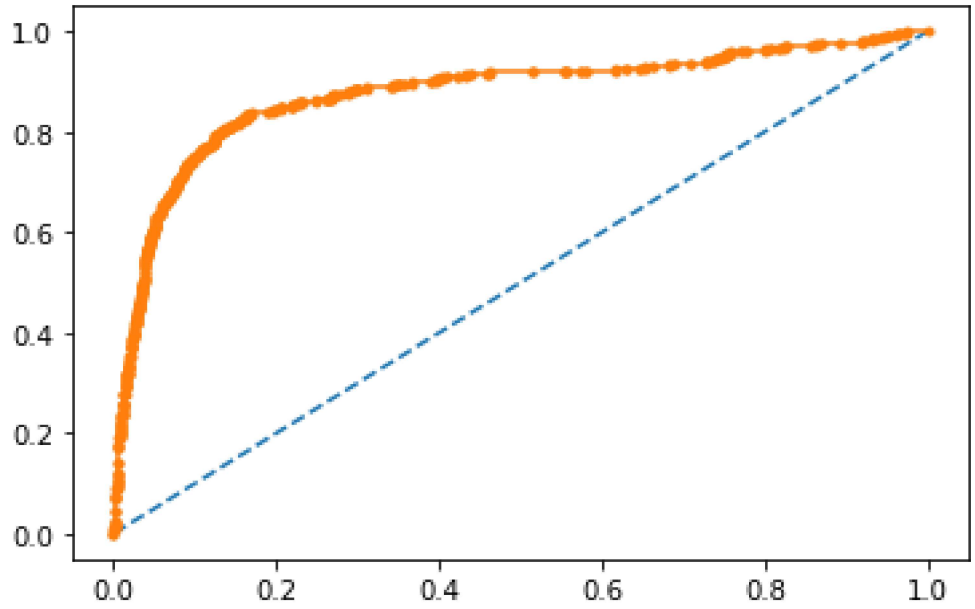

Fig-1.b

(2) Decision Tree

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2719,   26],
[ 106,  448]]
```

classification_report

	precision	recall	f1-score	support
0	0.96	0.99	0.98	2745
1	0.95	0.81	0.87	554
accuracy			0.96	3299
macro avg	0.95	0.90	0.92	3299
weighted avg	0.96	0.96	0.96	3299

ROC-AUC Graph

AUC: 0.971

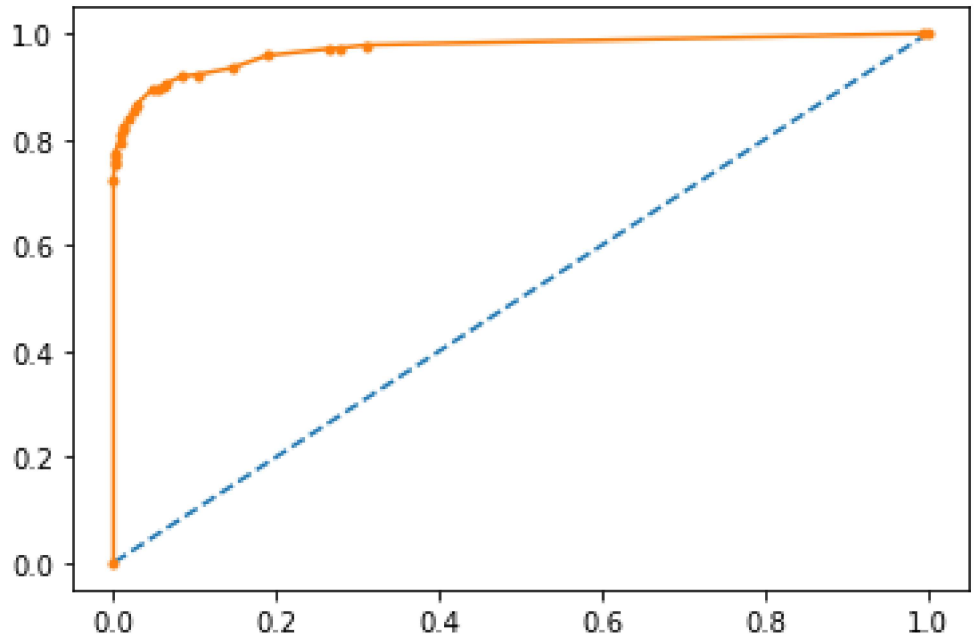

Fig-1.b.2

(3) Random Forest

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2719,   26],
[ 106,  448]]
```

classification_report

	precision	recall	f1-score	support
0	0.96	0.99	0.98	2745
1	0.95	0.81	0.87	554
accuracy			0.96	3299
macro avg	0.95	0.90	0.92	3299
weighted avg	0.96	0.96	0.96	3299

ROC-AUC Graph

AUC: 0.978

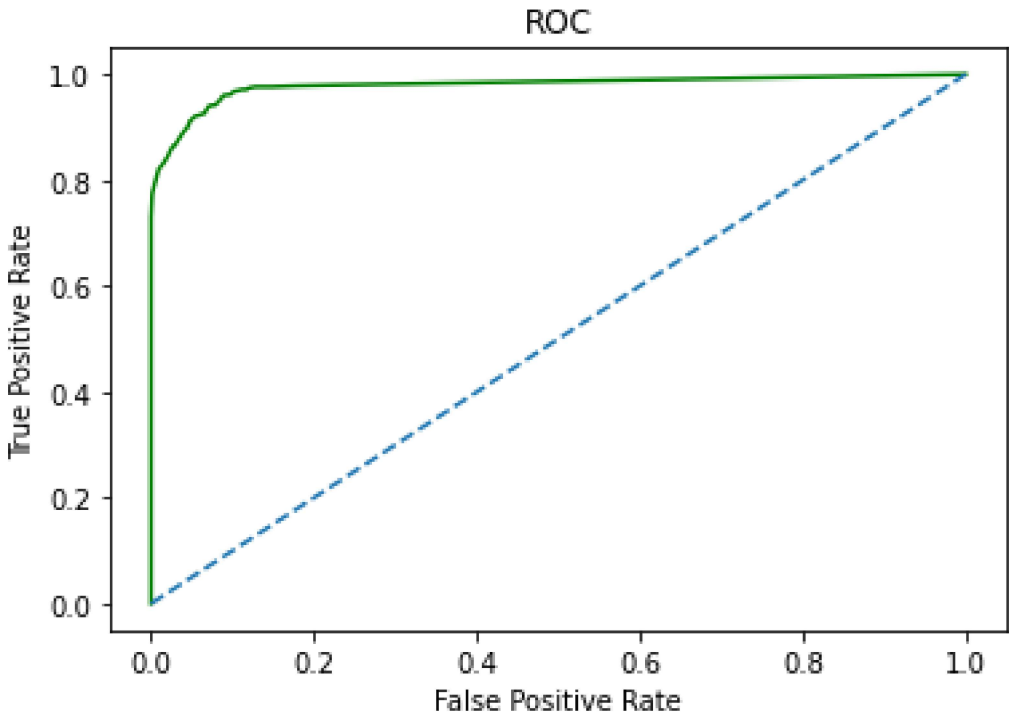

Fig-1.b.3

(4) Linear Discriminant Analysis

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[2567,  178],
 [ 232,  322]]
```

classification_report

	precision	recall	f1-score	support
0	0.92	0.94	0.93	2745
1	0.64	0.58	0.61	554
accuracy			0.88	3299
macro avg	0.78	0.76	0.77	3299
weighted avg	0.87	0.88	0.87	3299

ROC-AUC Graph

AUC: 0.971

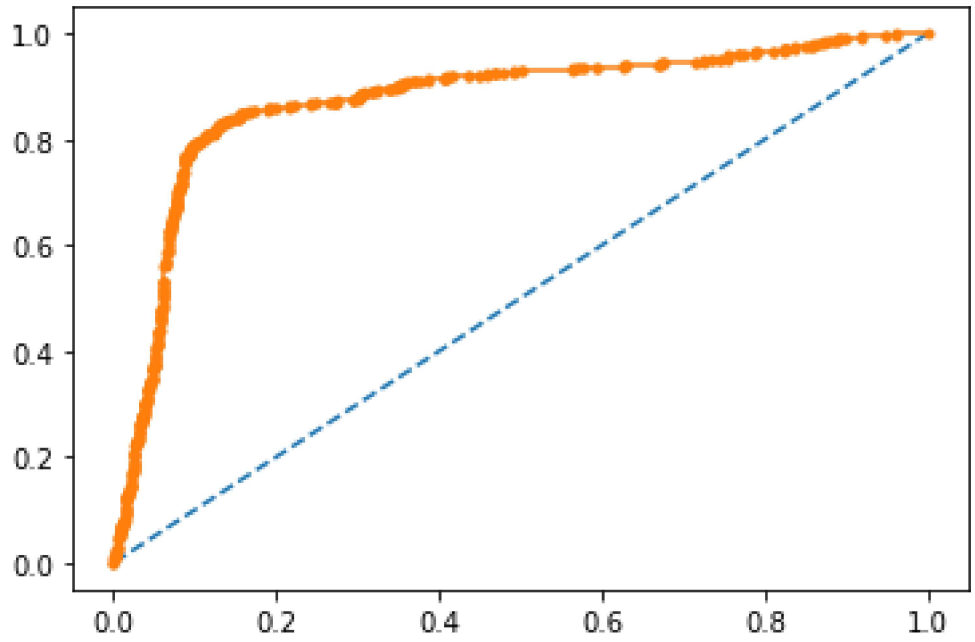

Fig-1.b.4

(5) K Nearest Neighbours

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[2703,   42],
 [ 519,   35]]
```

classification_report

	precision	recall	f1-score	support
0	0.84	0.98	0.91	2745
1	0.45	0.06	0.11	554
accuracy			0.83	3299
macro avg	0.65	0.52	0.51	3299
weighted avg	0.77	0.83	0.77	3299

ROC-AUC Graph

AUC: 0.711

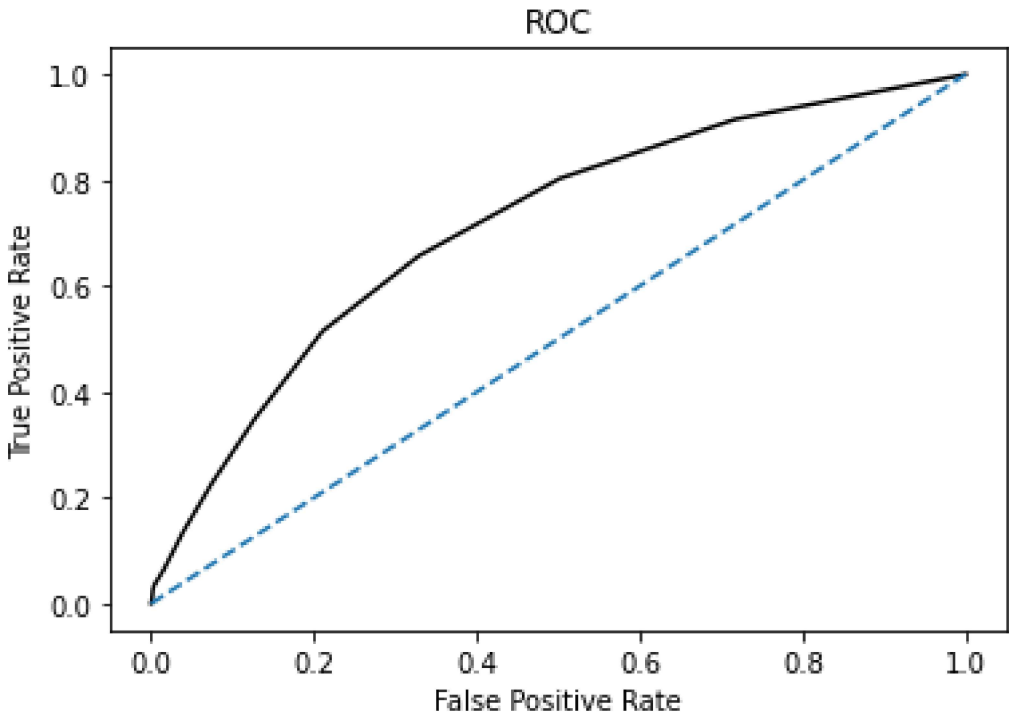

Fig-1.b.5

(6) Naive Bayes

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2597,  148],
 [ 183,  371]],
```

classification_report

	precision	recall	f1-score	support
0	0.93	0.95	0.94	2745
1	0.71	0.67	0.69	554
accuracy			0.90	3299
macro avg	0.82	0.81	0.82	3299
weighted avg	0.90	0.90	0.90	3299

ROC-AUC Graph

AUC: 0.8504

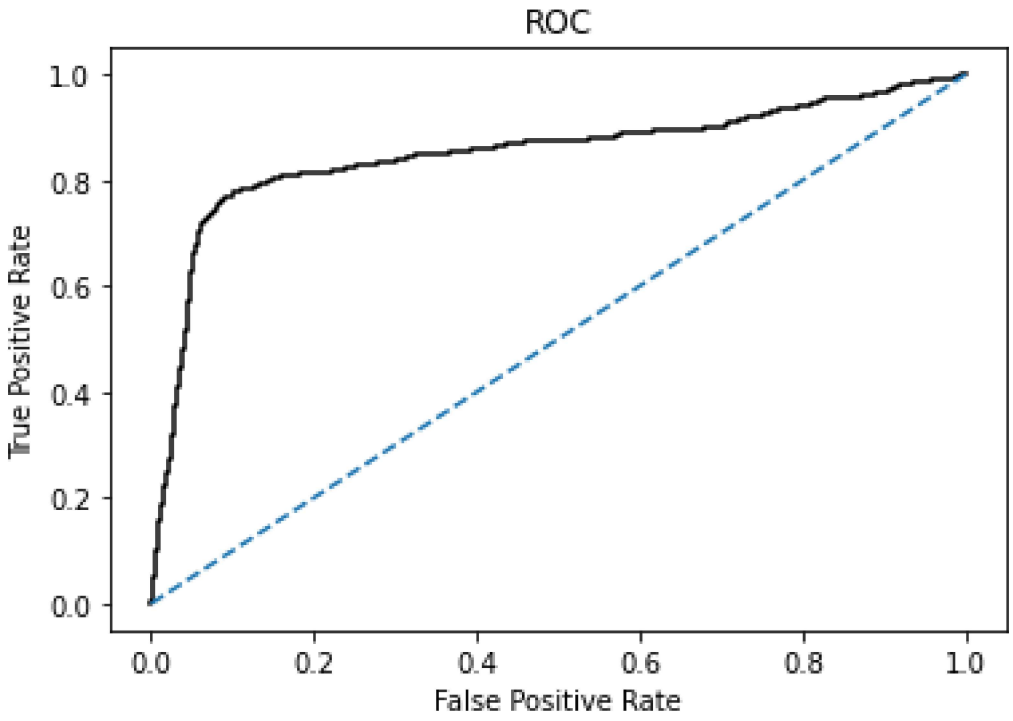

Fig-1.b.6

(7) Gradient Boosting

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2745,    0],
[ 554,    0]]
```

classification_report

	precision	recall	f1-score	support
0	0.83	1.00	0.91	2745
1	0.00	0.00	0.00	554
accuracy			0.83	3299
macro avg	0.42	0.50	0.45	3299
weighted avg	0.69	0.83	0.76	3299

ROC-AUC Graph

AUC: 0.5

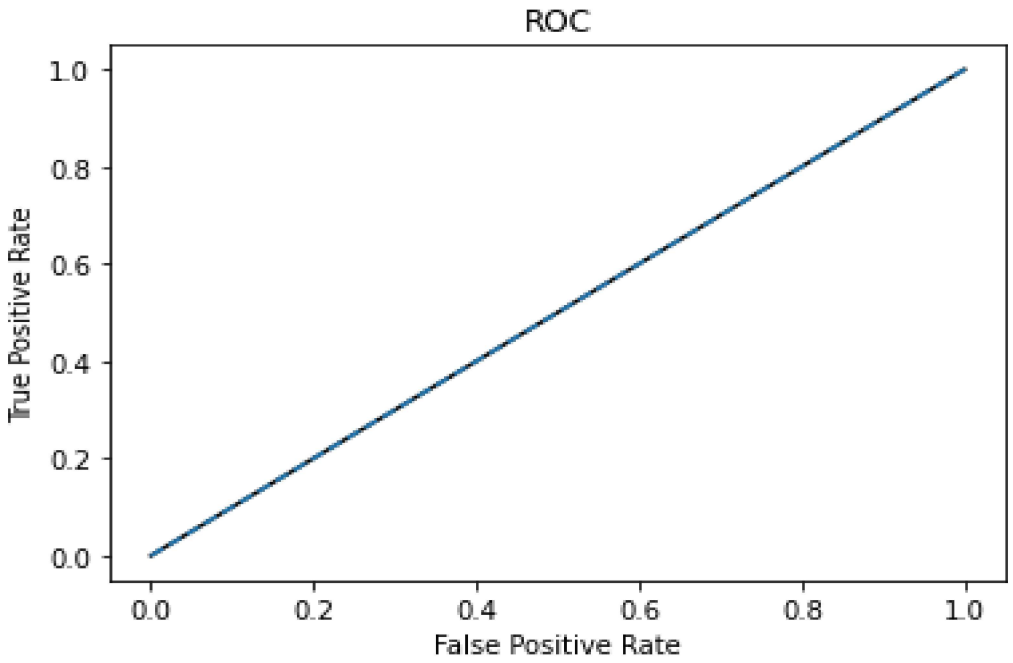

Fig-1.b.7

(8) Extreme Gradient Boosting

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2726,   19],
[  58,  496]]
```

classification_report

	precision	recall	f1-score	support
0	0.98	0.99	0.99	2745
1	0.96	0.90	0.93	554
accuracy			0.98	3299
macro avg	0.97	0.94	0.96	3299
weighted avg	0.98	0.98	0.98	3299

ROC-AUC Graph

AUC: 0.995

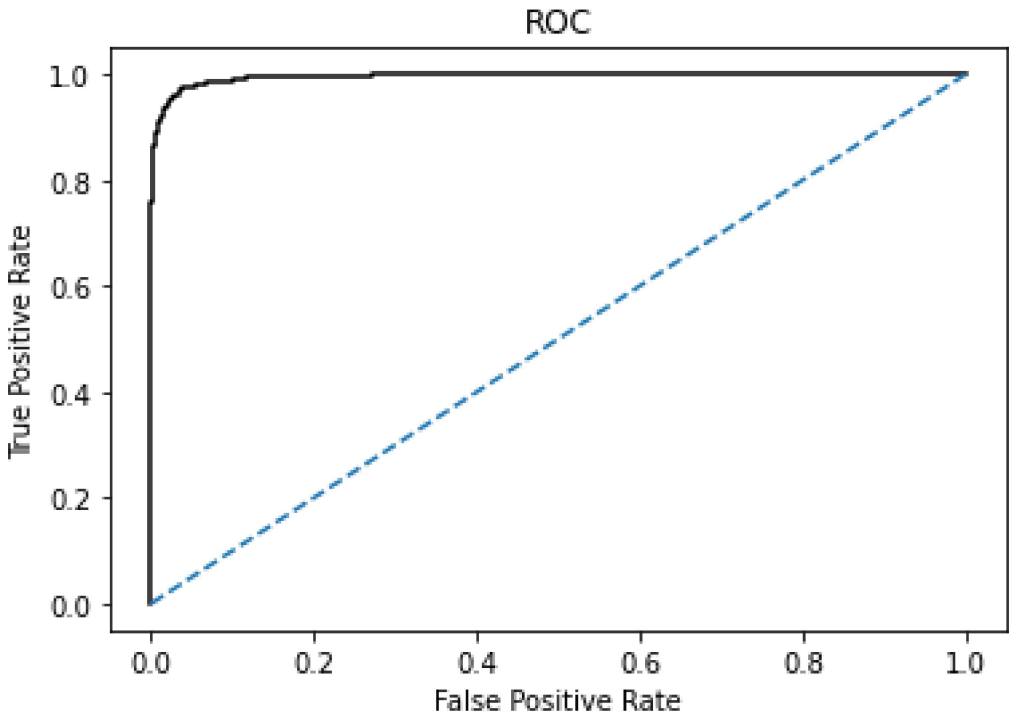

Fig-1.b.8

(9) Extra Tree Classifier

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[[2745,    0]
 [ 554,    0]
```

classification_report

	precision	recall	f1-score	support
0	0.83	1.00	0.91	2745
1	0.00	0.00	0.00	554
accuracy			0.83	3299
macro avg	0.42	0.50	0.45	3299
weighted avg	0.69	0.83	0.76	3299

ROC-AUC Graph

```
AUC: 0.5
```

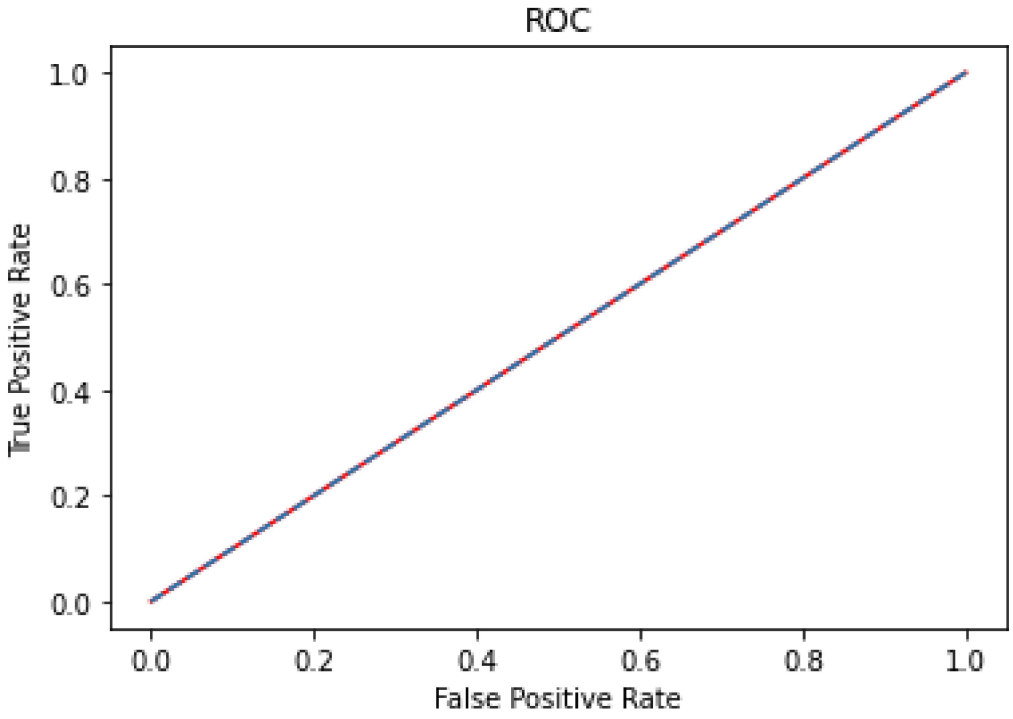

Fig-1.b.9

Conclusion | insight: All test modal are performed best modal results are RF and CART.

Extreme Gradient Boosting

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2726,   19],
[  58,  496]]
```

classification_report

	precision	recall	f1-score	support
0	0.98	0.99	0.99	2745
1	0.96	0.90	0.93	554
accuracy			0.98	3299
macro avg	0.97	0.94	0.96	3299
weighted avg	0.98	0.98	0.98	3299

ROC-AUC Graph

AUC: 0.995

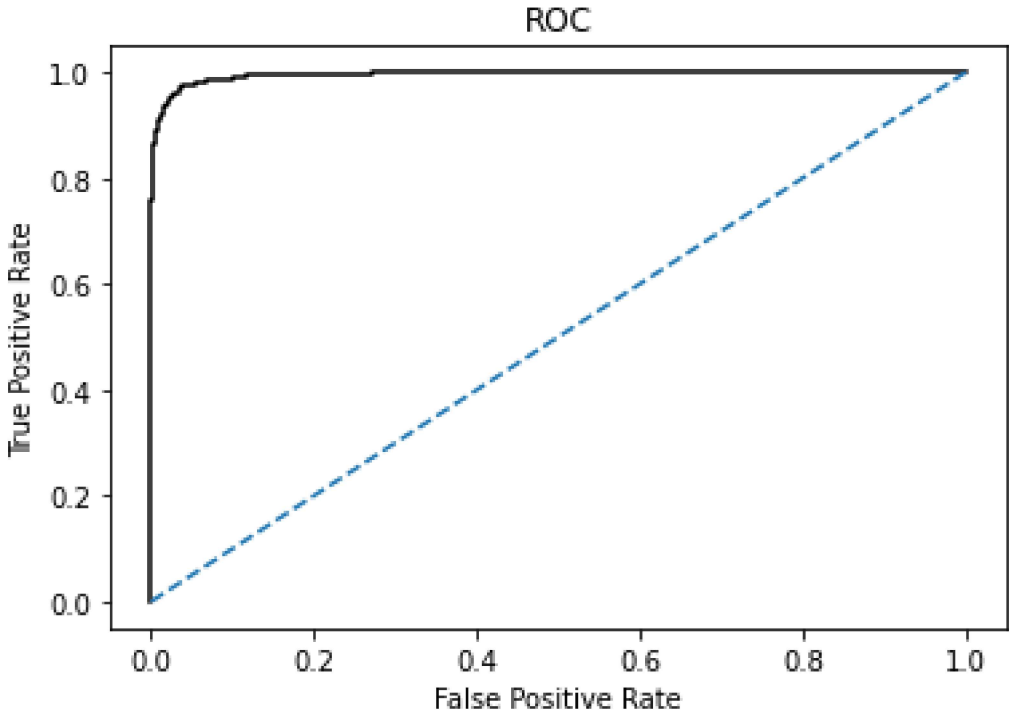

Fig-2.a.1

Gradient Boosting

Test:- confusion_matrix & classification_report , ROC-AUC Graph

confusion_matrix

```
[2745,    0],
[ 554,    0]]
```

classification_report

	precision	recall	f1-score	support
0	0.83	1.00	0.91	2745
1	0.00	0.00	0.00	554
accuracy			0.83	3299
macro avg	0.42	0.50	0.45	3299
weighted avg	0.69	0.83	0.76	3299

ROC-AUC Graph

```
AUC: 0.5
```

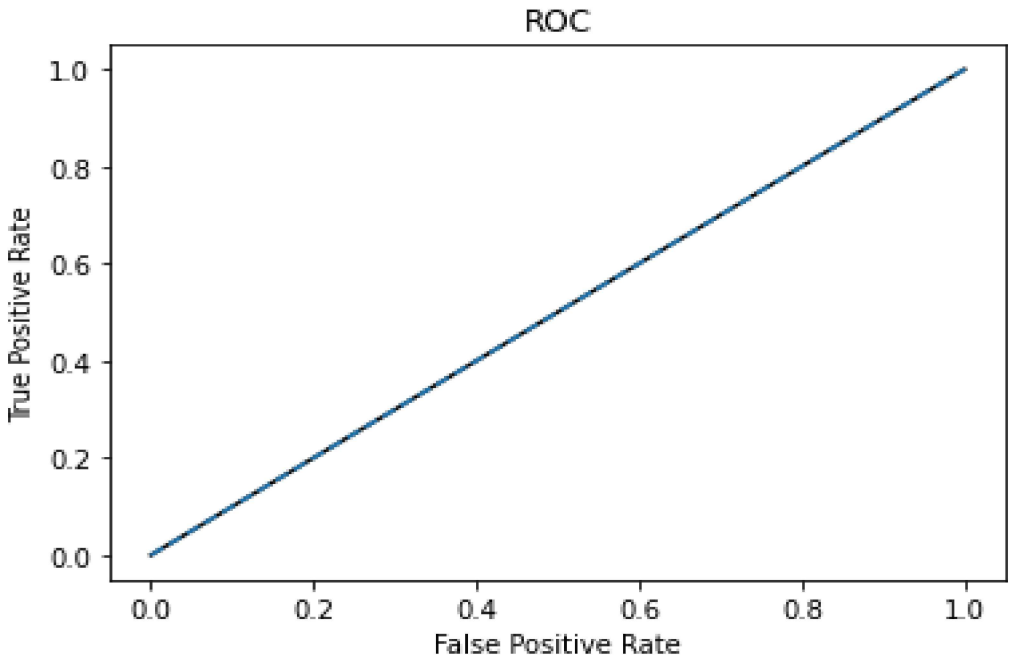

Fig-2.a.2

Conclusion | insight:

Model validation is done on the test dataset which is 30% of the complete data set. Not just accuracy but all the performance metrics are kept in mind before concluding the Random Forest as Best Model. Using Gradient Boosting & Extreme Gradient Boosting both modals are performed but accuracy is low as camper rest another model so not the best solution using boosting techniques.

Q6. Final interpretation / recommendation

Interpretation of the most optimum model and its implication on the business
Conclusion | insight:

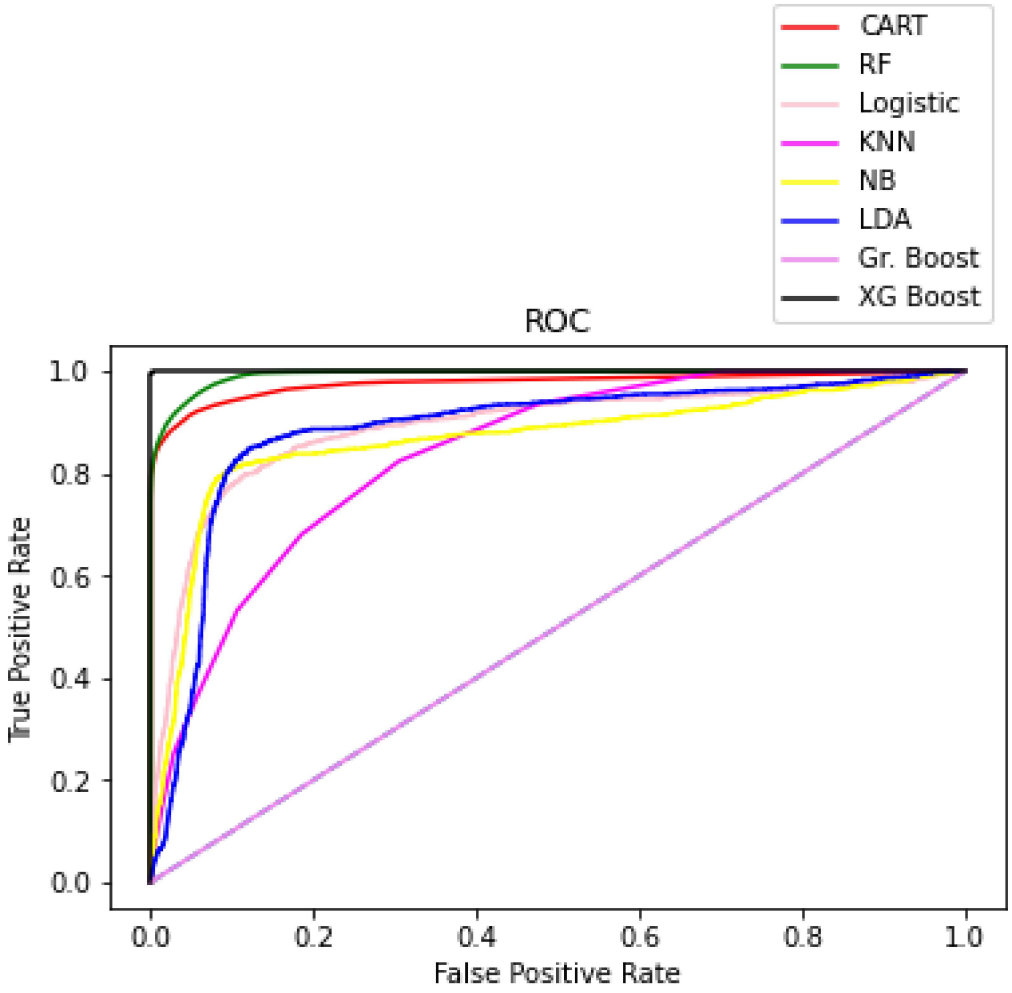

Fig-6.a

These predictions all Train Eight models need to achieve high AUC values. The sample data is divided into 70% for training and 30% for testing. We have applied feature engineering, effective feature transformation, and selection approach to make the features ready for machine learning algorithms. Best XG Boost, RF & CART AUC curve is best shown in campier all models.

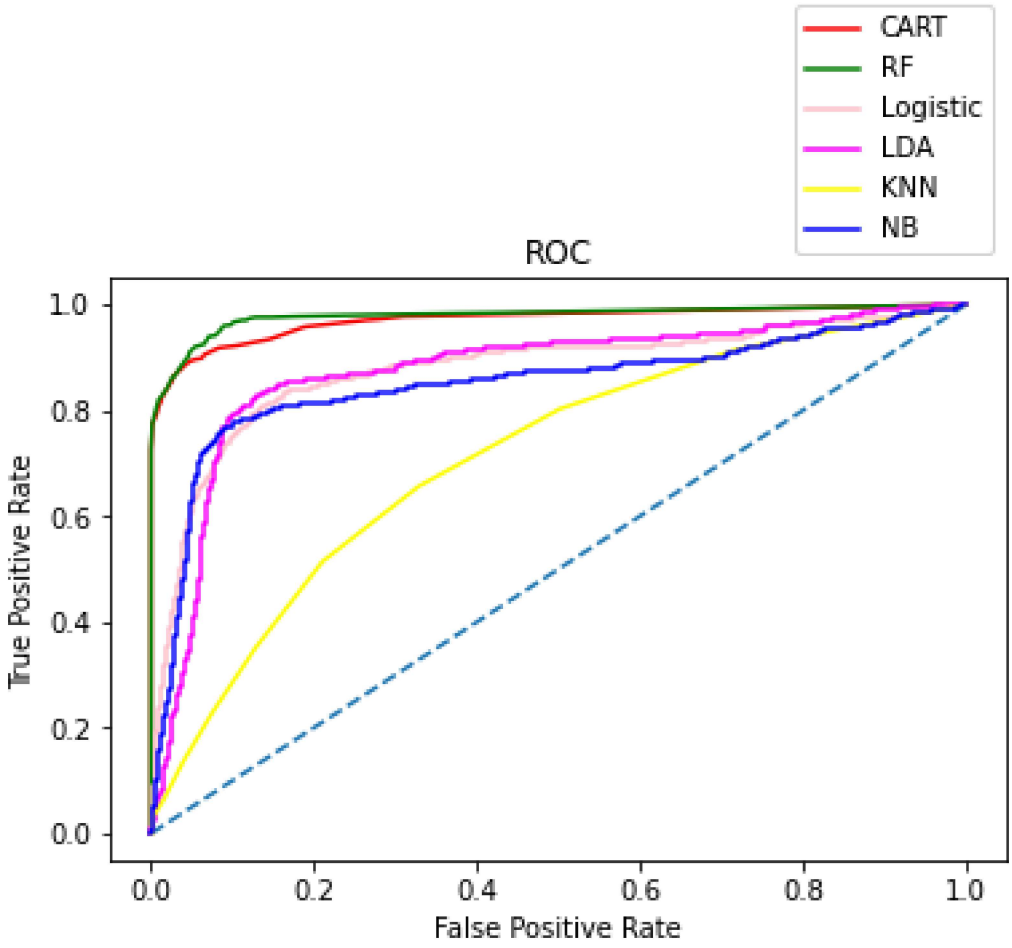

Fig-6.b

This prediction all Test six models need to achieve high AUC values. The sample data is divided into 70% for training and 30% for testing. We have applied feature engineering, effective feature transformation, and selection approach to make the features ready for machine learning algorithms. Best Model RF & CART AUC curves are best shown in campier all models.

Train

	CART Train	RF Train	Log Train	LDA Train	KNN Train	NB Train	Gr.Boost Train	XG Boost Train	Extra Tree Train
Accuracy	0.97	0.97	0.89	0.88	0.84	0.90	0.83	1.00	0.83
AUC	0.98	0.99	0.89	0.89	0.84	0.87	0.50	1.00	0.50
Recall	0.84	0.84	0.54	0.62	0.15	0.69	0.00	0.99	0.00
Precision	0.96	0.96	0.75	0.65	0.68	0.70	0.00	1.00	0.00
F1 Score	0.90	0.90	0.63	0.63	0.24	0.70	0.00	0.99	0.00

Fig-6.c

Test

	CART Test	RF Test	Log Test	LDA Test	KNN Test	NB Test	Gr.Boost Test	XG Boost Test	Extra Tree Test
Accuracy	0.96	0.96	0.89	0.88	0.83	0.90	0.83	0.98	0.83
AUC	0.97	0.98	0.88	0.87	0.71	0.85	0.50	1.00	0.50
Recall	0.81	0.81	0.51	0.94	0.06	0.67	0.00	0.90	0.00
Precision	0.95	0.95	0.74	0.92	0.45	0.71	0.00	0.96	0.00
F1 Score	0.87	0.87	0.61	0.93	0.11	0.69	0.00	0.93	0.00

Fig-6.d

These algorithms are Logistics Regression, Decision Tree, Random Forest, Linear Discriminant Analysis, K Nearest Neighbours, Naive Bayes, Gradient Boosting, Extreme Gradient Boosting, and Extra Tree Classifier. The method of preparation and selection of features and entering the EDTH features had the biggest impact on the success of this model since the value of AUC reached 96.0%. Random Forest & CART model achieved the best results in all measurements. The AUC value was 96.0%. The comes in second place and the LDA & KNN and Decision Tree came third and fourth regarding AUC values. We have evaluated the models by fitting a new dataset related to different periods and without any proactive action from marketing, RF also gave the best result with 96.0% AUC. The decrease in result could be due to the non-stationary data model phenomenon, so the model needs training each period of time.

What we have understood from assessing the tenure is that there are quite a lot of customers who are new to the

Platform. This is a good opportunity for the company to bring in new strategies and work upon increasing Customer retention.

The number of people that have contacted customer care is high. The number of times CC has been contacted by a single account is high as 132, this is an alarming high. We have noted that comparatively 3 to 4 users per account are higher in number. This is helping us to understand. That the products and the offers we need to plan to offer should be crafted keeping in mind that it is a for widest age distribution. Better offers around less-priced subscription plans will help us to retain new customers, increase customer numbers. The lower-priced subscription plans are more favoured by the customers. The revenue growth is seeming bright with mean and median around 15%.

Impurity-based Importance train value

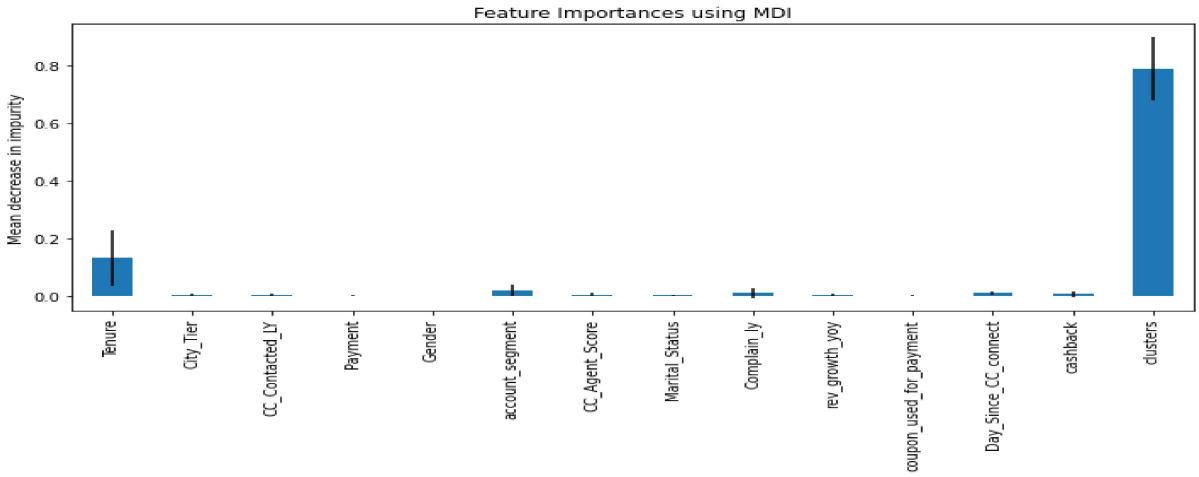

Fig-2.c.1

Permutation-based Importance test value

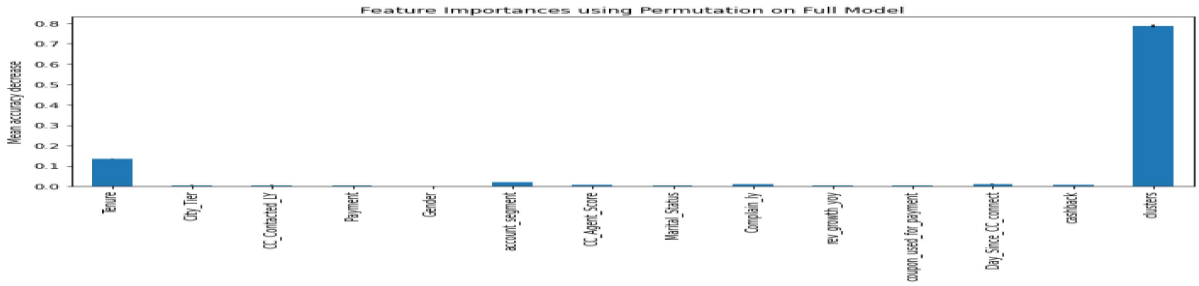

Fig-2.c.2

SHAP Value-based Importance

Fig-2.c.3

ELI5 Feature Importance

Weight	Feature
0.7865 ± 0.2209	x13
0.1334 ± 0.1917	x0
0.0198 ± 0.0370	x5
0.0119 ± 0.0124	x11
0.0112 ± 0.0335	x8
0.0071 ± 0.0170	x12
0.0065 ± 0.0103	x6
0.0054 ± 0.0065	x2
0.0050 ± 0.0084	x1
0.0039 ± 0.0061	x9
0.0033 ± 0.0068	x7
0.0027 ± 0.0057	x3
0.0025 ± 0.0047	x10
0.0009 ± 0.0023	x4

y=0 (probability **0.934**) top features

Contribution?	Feature
+0.832	<BIAS>
+0.158	x0
+0.010	x8
+0.008	x2
+0.006	x6
+0.005	x1
+0.004	x11
+0.002	x3
+0.001	x10
-0.000	x7
-0.004	x5
-0.006	x12
-0.036	x13
-0.043	x9

y=0 (probability **0.934**) top features

Contribution?	Feature	Value
+0.832	<BIAS>	1.000
+0.158	Tenure	32.000
+0.010	Complain_ly	0.000
+0.008	CC_Contacted_LY	8.000
+0.006	CC_Agent_Score	3.000
+0.005	City_Tier	1.000
+0.004	Day_Since_CC_connect	8.000
+0.002	Payment	2.000
+0.001	coupon_used_for_payment	1.000
-0.000	Marital_Status	0.000
-0.004	account_segment	5.000
-0.006	cashback	1913.000
-0.036	clusters	1.000
-0.043	rev_growth_yoy	16.000

y=0 (probability **1.000**) top features

Contribution?	Feature	Value
+0.832	<BIAS>	1.000
+0.116	clusters	3.000
+0.024	Tenure	11.000
+0.016	account_segment	4.000
+0.006	Day_Since_CC_connect	2.000
+0.002	cashback	398.000
+0.002	Complain_ly	0.000
+0.002	coupon_used_for_payment	0.000
+0.001	Gender	1.000
+0.001	Payment	1.000
+0.000	rev_growth_yoy	7.000
+0.000	CC_Agent_Score	3.000
+0.000	City_Tier	1.000
+0.000	CC_Contacted_LY	16.000
-0.000	Marital_Status	0.000

Conclusion | insight:

Fig-2.c.4

Recommendations

We can suggest the management segment customers based on 'City-Tiers'. The reasons for this suggestion are, Clear demographic divide, Very distinct behavior within each tier across multiple parameters. Each tier can be approached with local and custom offers. Marketing can be done both on online and offline platforms across clearly defined tiers.

Insights based on EDA

The importance of this type of research in the telecom market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income for telecom companies. Hence, this research aimed to build a system that predicts