Capstone Project – Project Notes -1 Submission

NAME	DEEPAK SINGH
PROJECT	CUSTOMER CHURN
GROUP	PGP-DSBA
SUBMISSION DATE	21ST JANUARY 2024

Problem Statement –

In the competitive landscape of the DTH industry, acquiring and retaining customers is a significant challenge and a key business goal. The Customer Engagement team is committed to re-engaging users who have terminated their subscriptions in a bid to minimize churn rates. Currently, they use a traditional analysis approach using Excel and develop strategies according to analysis which require more human intervention and can be erroneous. They propose an AI-integrated prediction model that can automate and eliminate maximum human intervention in terms of data analysis and churn prediction. They wanted to use the model to predict the potential churners to develop customer retention strategies and campaigns. They have collected historical records and provided for the project. The dataset collected by the company has both the metadata in the first tab and historical records in the second tab.

Table 1: Attribute name and its 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

OBJECTIVE OF THE STUDY AND NEED FOR A PROJECT

BUSINESS OBJECTIVE

By identifying patterns and trends in the data, companies can easily focus on the features inclined to customer churn problems. The organization can act quickly based on the insight from the identified patterns and trends. By building a machine learning model the DTH Service

Provider can easily target the probable churners and devise strategies to retain them by providing offers, targeting personalized campaigns, and conducting market study.

DEMAND AND NEED OF THE PROJECT

The customer is the king who has the power of purchase and choice. Companies compete to add new features to their products to retain and satisfy their customers. Novel technologies and advertisements can entice the customer to purchase their products. The never-ending wants and the desire to acquire the latest technology can lead to customer churn. So, it is inevitable for organizations to have a project to identify and study the pattern of losing customers and predict the possibility of future customer churns. It helps them to formulate strategies to attract and retain their valuable customers. It helps them to add value to their products. So, the demand for a project has become a necessity rather than a need.

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

This a case study of an e-commerce company where in they have certain number of customers assigned with unique account ID and a single account ID can hold many customers (like family plan) across gender and marital status, customers get flexibility in terms of mode of payment they want to opt for. Customers are again be segmented across various types of plans they opt for as per their usage which is also based on the device they use (computer or mobile) moreover they earn cashbacks on bill payment. According to the business statement the overall business runs in customer's loyalty and stickiness which in-turn comes from providing quality and value-added services. Also, running various promotional, bundles, discounts and festivals offers may help organization in getting new customers and also retaining the old one which have left the company. We can conclude that a customer retained is a regular and probably a permanent income for organization, a customer added is a new income for organization and the customers lost will be a negative impact as a single account ID holds multiple number of customers i.e.; closure of one account ID means losing multiple customers.

DATA ANALYSIS AND INTERPRETATION

The data collection for this study on predicting customer churn spanned about 99 months has been collected by the organization and the same has been received in an Excel file with two sheets in it one with attribute description and the other with the actual customer data. This time frame was carefully selected to ensure a comprehensive understanding of long-term customer engagement and potential churn patterns. The specific data contains information and activities for each month, including details such as product usage, customer service interactions,

subscription renewals, complaints, etc. Even though the description was given by the organization, a deep dig was required to understand each feature and to get more data insight. We used the Dataframe.head () function to identify quantitative and qualitative features. It is given in Understanding the Data in the Jupyter Notebook.

Figure 2: First glance of data

A glance of the data

```
In [9]: df.head()
```

Out[9]:

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_S
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0	
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0	
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0	
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0	
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0	

While comparing the data against the data types in “Understanding the Data reference code (DU2)”, some numerical fields such as Tenure, Account_user_count, and Revenue per month, rev_growth_yoy, coupon_used_for_payment, Days_since_connect and cashback are fetched as objects. It represents the presence of non-numerical values in the numerical fields. So, the data has been further drilled down by filtering it as a unique value in Understanding the Data reference code (DU4).

The Below Figure depicts the presence of invalid characters and the need to fix the data inconsistencies.

Figure 3: Invalid character identification

To identify the unique values in the dataset

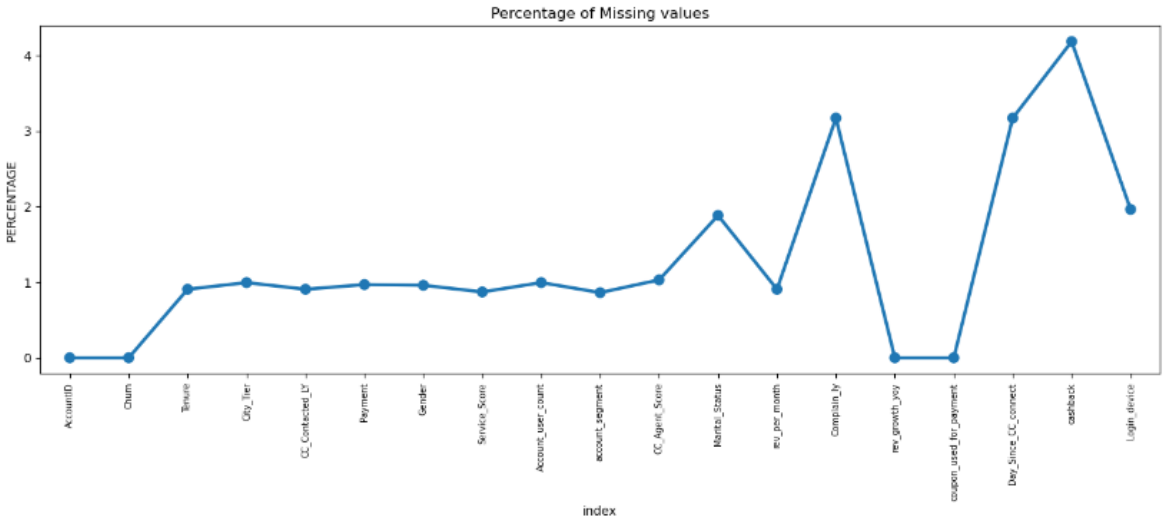
```
In [10]: for col, ro in df.items():          # iterate each columns and rows through the dataset get the column label and row values
           print("Feature Name :", col, "         Data Type :", df[col].dtype, "\n")       # print column lable, its datatype and a new l
           print(df[col].unique(), "\n")          # print unique values
           print("_____________________________________________________________________________")
```

```
Feature Name : AccountID          Data Type : int64

[20000 20001 20002 ... 31257 31258 31259]

_____________________________________________________________________________
Feature Name : Churn          Data Type : int64

[1 0]

_____________________________________________________________________________
Feature Name : Tenure          Data Type : object

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

_____________________________________________________________________________
Feature Name : City_Tier          Data Type : float64

[ 3.  1. nan  2.]

_____________________________________________________________________________
Feature Name : CC_Contacted_LY          Data Type : float64

[  6.   8.  30.  15.  12.  22.  11.   9.  31.  18.  13.  20.  29.  28.
  26.  14.  10.  25.  27.  17.  23.  33.  19.  35.  24.  16.  32.  21.
  nan  34.   5.   4. 126.   7.  36. 127.  42.  38.  37.  39.  40.  41.
 132.  43. 129.]

_____________________________________________________________________________
Feature Name : Payment          Data Type : object

['Debit Card' 'UPI' 'Credit Card' 'Cash on Delivery' 'E wallet' nan]
```

Missing Value Detection

As there were some null values found in the section in Understanding the Data reference code (DU2), a question of feature consideration popped up. That resulted in the identification of the percentage of null values.

Figure 4: Percentage of missing data



To identify the percentage of missing values the code in the section Understanding the Data (DU6) has been used. The code calculates the percentage of missing values and plots it in a line graph with points (point plot). Even though there is no standard percentage to consider a feature in the analyses. But a percentage of 5 to 10 is acceptable. In our case, we can consider all the features as it has less than 6 per cent of null values.

Feature Removal

AccountID is a unique identification number given for each record. It doesn't add value, so the AccountId field is removed. After removing the AccountID, we have 18 features. The Shape of the data for analysis became (11260, 18) – Coded under Feature Engineering reference code (FE1).

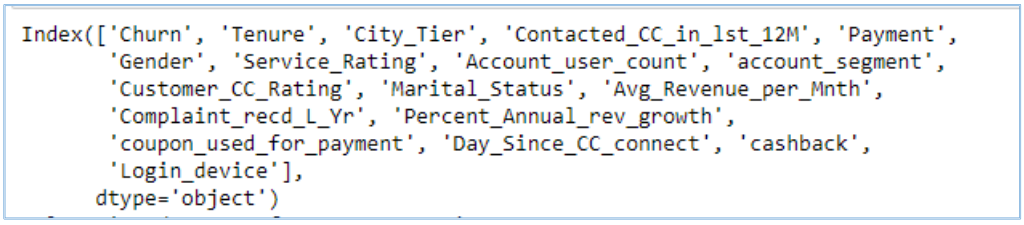
Data Cleaning

While analyzing the data, there was a presence of invalid characters and inconsistencies in the dataset. The raw data contained invalid symbols like '*','&', '$', '+', '@', and '#' in various columns. They have been replaced with Null values. The '&&&&' symbol in the Login_device column has been replaced with "Others". The program under the section Feature engineering reference code "Data Cleaning (FE2)" removed all the invalid characters. After removing the invalid characters. We have a clean dataset of numerical fields and categorical fields.

Feature Name Correction

For clarity in the Feature names, a few labels are renamed for better understanding as below. The same can be seen in the Feature Engineering reference code FE3 section of the Feature Engineering.

Figure 5: Feature name correction



After renaming the features, they are categorized into two types of fields. The program snipper under the section in Feature Engineering reference code "Data Cleaning (FE2)" will create lists of categorical (cat_fld) and numerical fields (num_fld) for target operations.

Missing Value Treatment

The Null values are treated with basic statistical measures. All the null values in numerical fields are replaced with mean and categorical fields are replaced with mode. The output can be found in the coding section in the Feature Engineering reference code “Missing value treatment (FE4)”.

Figure 6: Missing value counts after treatment

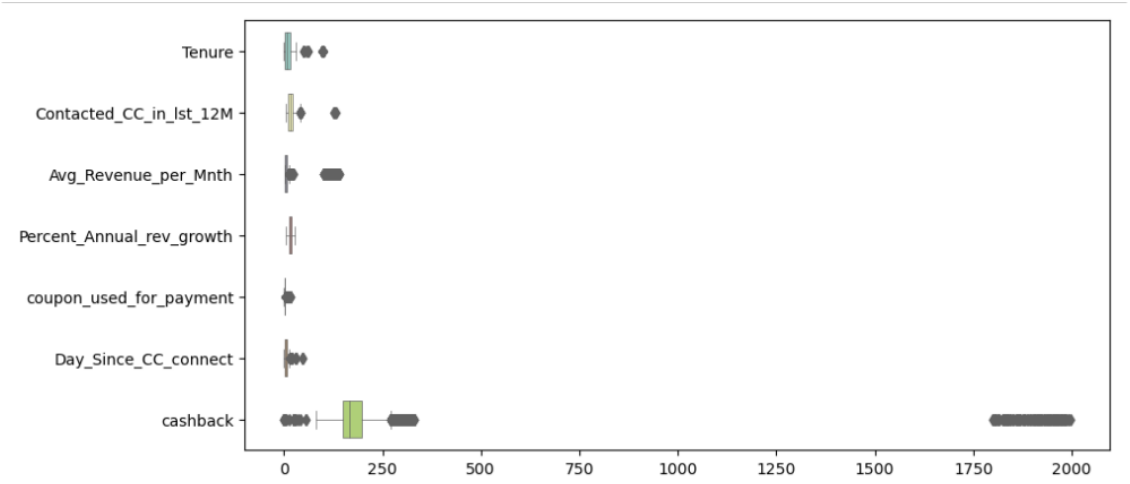
```
Categorical fields
____________________
Payment                0
Gender                 0
account_segment        0
Marital_Status         0
Login_device           0
City_Tier              0
Service_Rating         0
Account_user_count     0
Customer_CC_Rating     0
Complaint_recd_L_Yr    0
Churn                  0
dtype: int64

Numerical fields
____________________
Tenure                          0
Contacted_CC_in_lst_12M         0
Avg_Revenue_per_Mnth            0
Percent_Annual_rev_growth       0
coupon_used_for_payment         0
Day_Since_CC_connect            0
cashback                        0
dtype: int64
```

Outlier Detection & Removal

Outliers are the extreme values present in the numerical fields.

The above Plot depicts the presence of Outliers in the cleaned data. Among all the attributes, the CC_Contacted_Ly, rev_per_month, and cashback fields contain outliers. Outliers concerning the Cashback field contain the majority of the outliers and are the most extreme

Figure 7: Plot showing the Outliers before treating



In the below plot, all the outliers that are above the Maximum and Minimum values of the box plot are replaced by the Maximum and Minimum values respectively for each box plot.

Figure 8: Plot showing the Outliers after treating

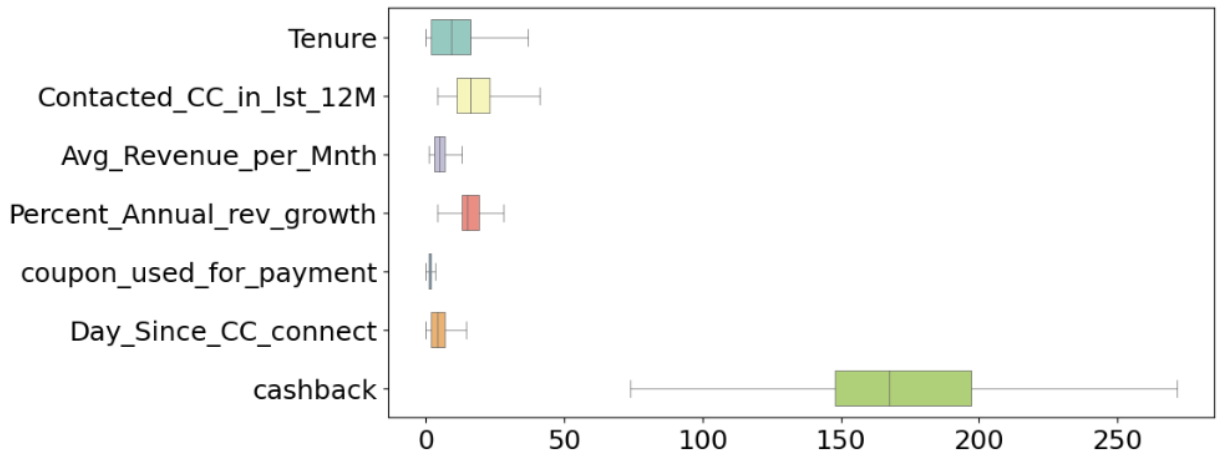


Figure 9: Basic statistics after data cleaning

	count	mean	std	min	25%	50%	75%	max
Churn	11260.0	0.168384	0.374223	0.00	0.00	0.00	0.00	1.00
Tenure	11260.0	10.290142	8.887725	0.00	2.00	9.00	16.00	37.00
City_Tier	11260.0	1.647425	0.912763	1.00	1.00	1.00	3.00	3.00
Contacted_CC_in_lst_12M	11260.0	17.833126	8.562396	4.00	11.00	16.00	23.00	41.00
Service_Rating	11260.0	2.903375	0.722476	0.00	2.00	3.00	3.00	5.00
Account_user_count	11260.0	3.704973	1.004383	1.00	3.00	4.00	4.00	6.00
Customer_CC_Rating	11260.0	3.065808	1.372663	1.00	2.00	3.00	4.00	5.00
Avg_Revenue_per_Mnth	11260.0	5.321048	2.884834	1.00	3.00	5.00	7.00	13.00
Complaint_recd_L_Yr	11260.0	0.276288	0.447181	0.00	0.00	0.00	1.00	1.00
Percent_Annual_rev_growth	11260.0	16.193339	3.757222	4.00	13.00	15.00	19.00	28.00
coupon_used_for_payment	11260.0	1.475577	1.102254	0.00	1.00	1.00	2.00	3.50
Day_Since_CC_connect	11260.0	4.609858	3.482954	0.00	2.00	4.00	7.00	14.50
cashback	11260.0	178.575979	43.653108	73.76	147.89	167.49	197.31	271.44

Now the data is cleaned, removing inconsistencies and outliers. To have a look at the distribution run the “df.describe.T” code under the section Feature Engineering reference code (FE6). The above figure shows the output. The distribution seems to be clean. The data is now ready to go ahead with Exploratory Data Analysis (EDA).

EXPLORATORY DATA ANALYSIS

EDA is one of the foremost in any data analysis to understand if there are any outliers present in the data and how different variables are related to each other and helps in designing statistical analysis that produces meaningful results.

It is a method used to investigate, analyze and summarize data sets and their main characteristics, often employing data visualization methods. It helps in making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

It is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

It can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including Machine Learning.

Various statistical functions and techniques that can be performed with EDA tools include but are not limited to

● Clustering and dimension reduction techniques, help create graphical displays of high dimensional data containing many variables.

● Univariate visualization of each field in the raw dataset, with summary statistics.

● Bivariate visualizations and summary statistics allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.

● Multivariate visualizations, for mapping and understanding interactions between different fields in the data.

● K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e., the number of clusters

● Predictive models, such as linear regression, use statistics and data to predict outcomes.

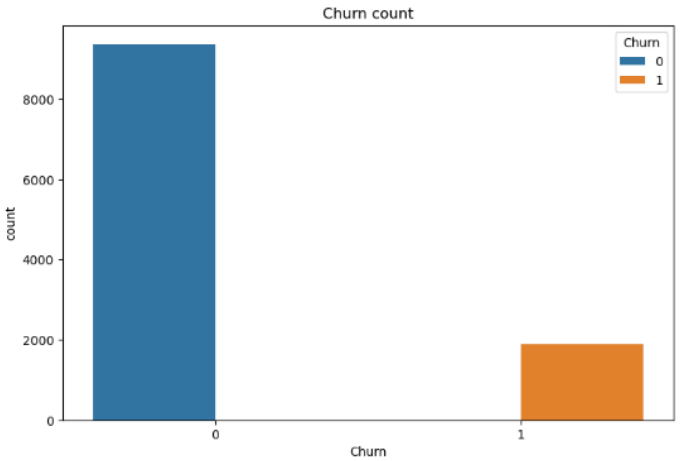
EDA is also used to:

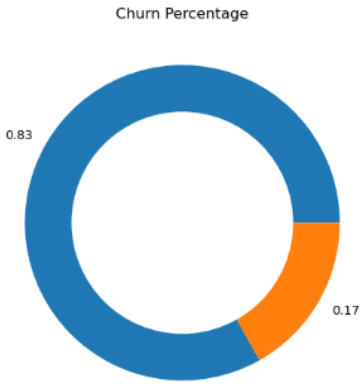
1, Identify variable distribution

2. Plot various graphs like Histograms, box plots, scatter plots, etc.

3. Analysis of the Correlation of different variables in the problem statement.

Univariate Analysis (EDA1)

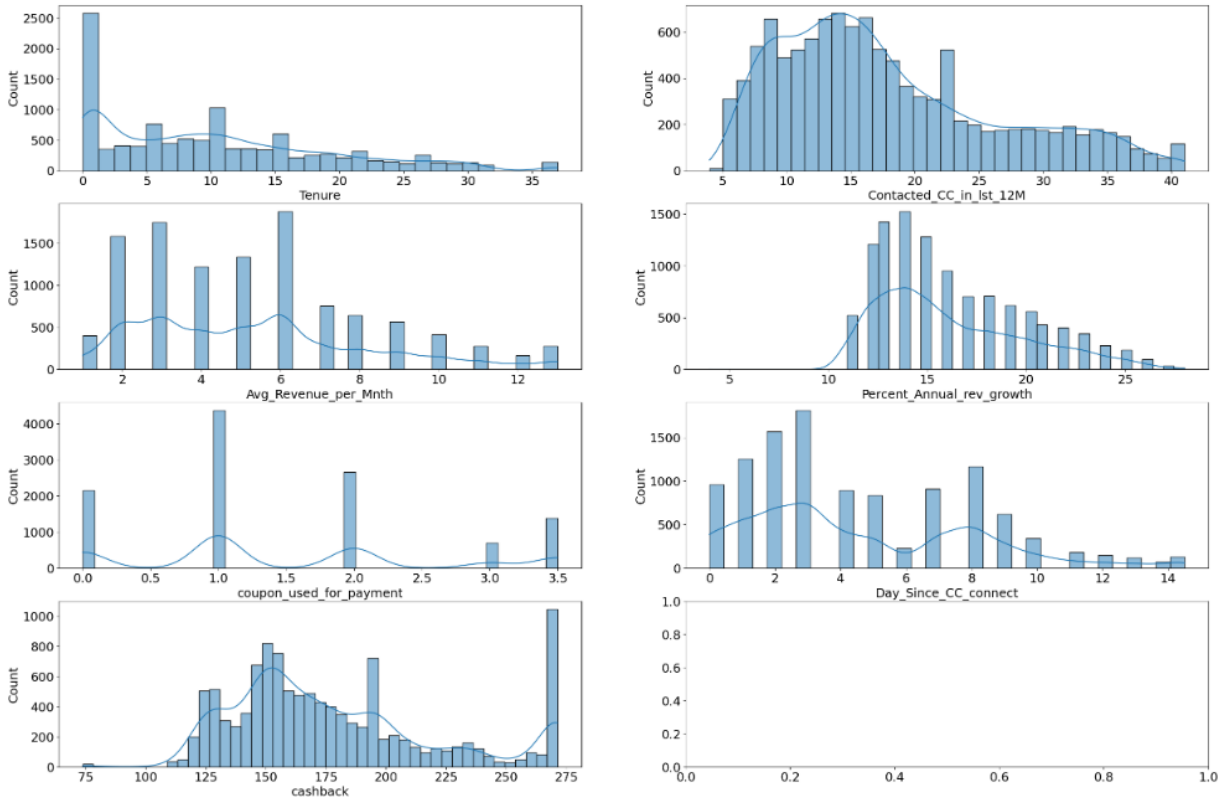




Among all the users in the company (11260) about 17% of the users churned.

Numerical Field Analysis (EDA2)

Figure 11: Plot showing the distribution of numerical attributes

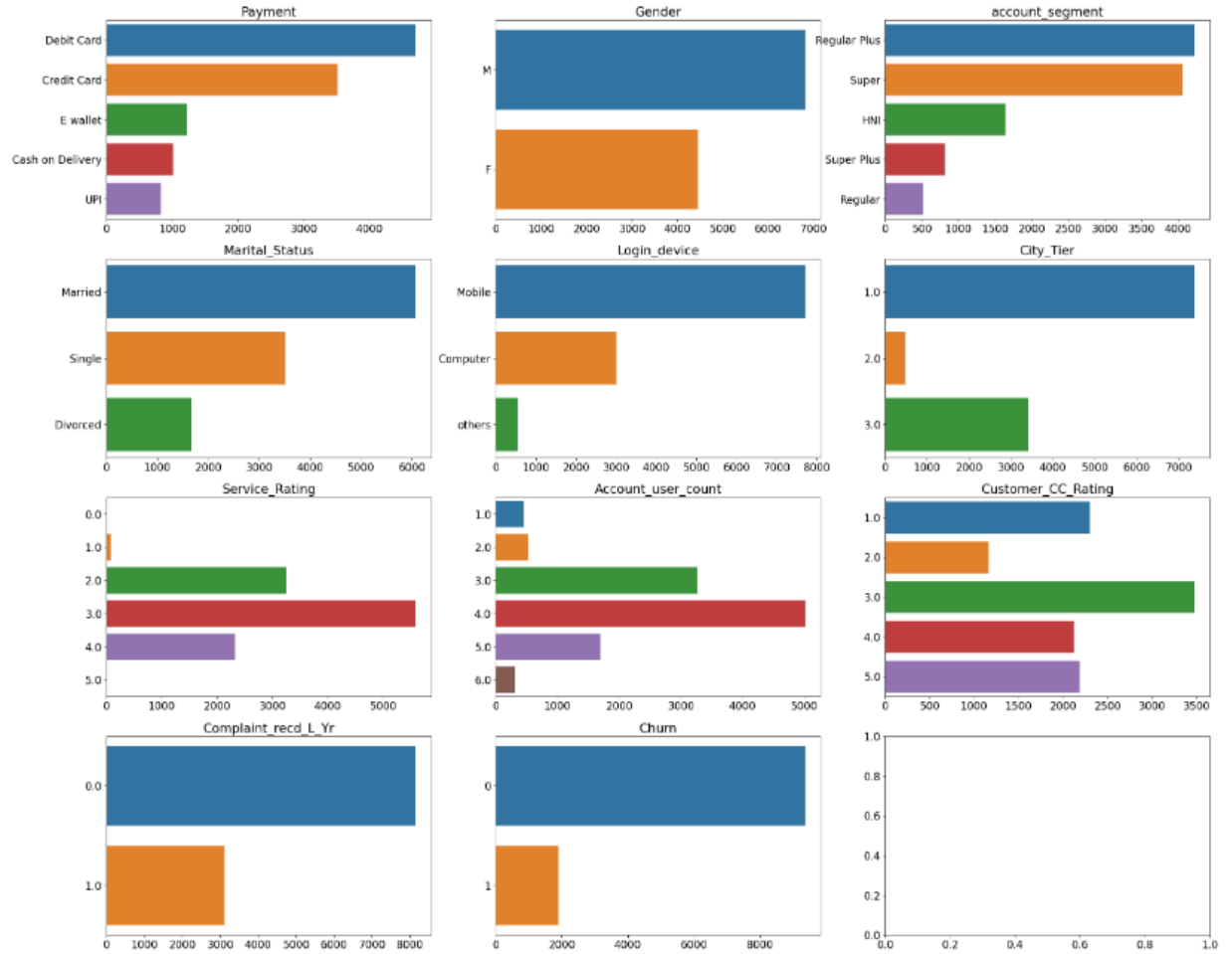


From the above bar chart, it is observed that:

- The majority of Customers have tenure in the range of 1-20 Months.
- The majority of customers contacted Customer Care about 5-25 times in the last 12 months
- The average revenue growth per customer is about 15
- The average Cashback obtained by customers is about 150

Categorical Field Analysis (EDA3)

Figure 12: Plot showing the categorical representation of the data after cleaning



From the above bar charts, it is observed that:

- The majority of users are choosing a Debit Card and Credit Card as their payment mode.
- Significantly more no. of male users are there when compared to female users.
- Among all the Account segments available most of the Users belong to the “Regular Plus” and “Super”
- About 50% of the users are married.
- A large portion of the users are using “Mobile” as their login device
- About 99% of the users belong to Tier 1 and Tier 3 Cities
- The Majority of Users rated the service score as 3/5 which signifies “Average”.
- The majority of user accounts are being used by 4 members.

Categorical Field Analysis (EDA5)

Figure 13: Results of Univariate Analysis of Churned Users

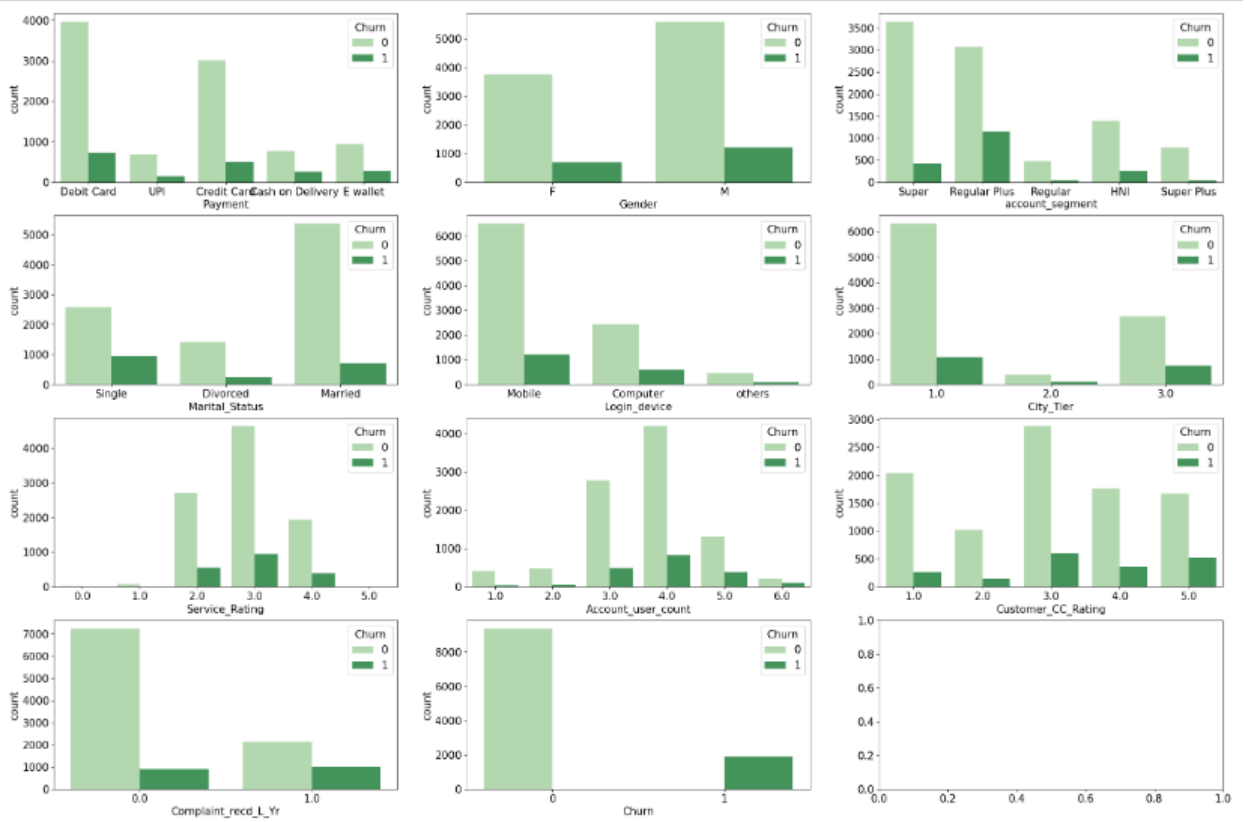
The above plot represents a few observations about the Churned users:

- Even though all the payment modes are used by users, the maximum number of users used Cash on delivery as their payment mode (26%)
- The majority belong to account segments “Regular Plus” (41%) and “HNI” (24%)
- About 51% have their Marital Status as “Single”
- The login devices used by churned users are Computer (56%) and Mobile (44%)
- Tier 1 City (26%), Tier 2 City (36%), Tier 3 City (38%)

Bivariate Analysis

Numerical Field Analysis (EDA4)

Figure 14: Plot that gives insights into Bivariate Analysis of numerical fields of churned user

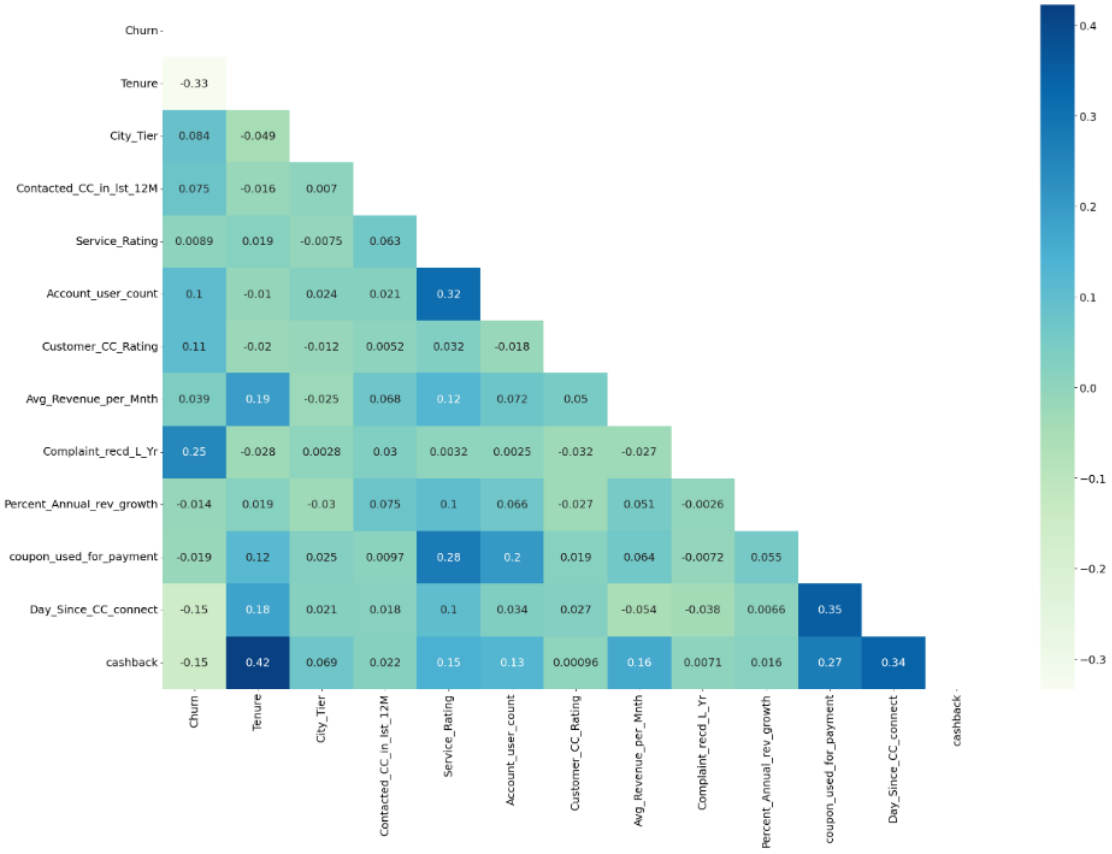


The above plot shows the details of the Churned users:

- The majority of the Churned users belong to the account segment of “Regular Plus” and “Super”
- The majority of Churned users used Mobile as their login device
- The majority of Churned users have given a service score of 3/5
- The majority of Churned user accounts have 4 users per account.

Multivariate Analysis (EDA6)

Figure 15: Correlation Plot of various attributes



The above correlation plot established a Positive Correlation between:

1. Churn – Complain_ly (0.25)

2. Tenure – Cashback (0.42)

3. Service_Score – Account_User_Count (0.32)

4. Coupon_Used_for_payment – Day_Since_CC_connect (0.35)

5. Day_Since_CC_Connect – (0.34)

And a Negative Correlation between:

1. Churn – Tenure (-0.33)

2. Churn – Day_Since_CC_Connect (-0.15)

3. Churn – Cashback (-0.15)

MODEL BUILDING AND INTERPRETATION

The thorough Exploratory data analysis derived valuable insights that show the data is cleaned and ready to undergo machine learning application. As a standard machine learning approach data preprocessing should be conducted before applying any machine learning model. Linear Discriminant Analysis (LDA), Logistic Regression, K-Nearest Neighbor classifier (KNN), Support Vector Machine, Random Forest, and Extreme Gradient Boosting (XGBoost) have been taken for the study. Both are supervised learning classification algorithms. As a standard procedure, before applying machine learning data preprocessing has been done to encode the categorical fields to numeric and standardize the data for better sailing.