Customer Churn Prediction

-Aakanksha Kadam

PGP DSBA capstone

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Business Problem: A DTH 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.

Churn rate, also referred to as attrition rate, measures the number of individuals or units leaving a group over a specified time. The term is used in many contexts, including in business, human resources, and IT. Most notably, churn rate is referred to as the proportion of contractual (or subscribed) customers who terminate their contractual relationships/subscriptions with a company in each timeframe. In this context, the term is primarily associated with companies operating on a subscription basis.

We must predict future churn rates, because it will help the business to gain a better understanding of future expected revenue. In addition, when we can use churn prediction to forecast the potential churn rate of a particular customer, it allows us to target that individual to prevent them from discontinuing their subscription with the company. And, since the cost of acquiring a new customer as per research is much higher, 5 to 6 times more to acquire new customers than keeping an existing one, there's plenty of revenue-based reason to do everything in our power to keep those existing customers. This emphasizes the importance of managing churn by organisation. To decrease the churn rate, companies utilize different methods and strategies. Generally, the strategies are focused on improvements in customer retention and satisfaction by establishing proactive communication with customers, obtaining constant customer feedback on the company's performance, and improving the company's operations.

Business Objective:

The business objective on which we will focus are:

- Building a churn prediction model for the DTH company to identify and control subscriber churn. To build such a model most important step would be to identify important variables/factors from provided dataset which are influencing customer churn.
- Then develop Machine Learning model using these variables and evaluate their accuracy and performance.
- Finally, find the best model having best scores and provide business insights and recommendations to the DTH company

2. EDA & BUSINESS IMPLICATIONS

The DTH company has collected data for the purchases of the customer for various account segments. The data has information about the Tenure of the selected plan, the city tier where the plan was selected, the payment method used, gender demographic, marital status of the customer, revenue generated per month, type of login device used for the account and other factors. Below table summarises all the attributes present for the data collected.

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in last 12months
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_l12m	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

Table 1: Dataset attributes

	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status
0	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0	Single
1	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0	Single
2	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0	Single
3	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0	Single
4	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0	Single

Fig 1: The dataset

Observations:

1. Initially above dataset had 11260 Rows and 19 columns, we have removed Account ID column since it is not of much use.
2. The data has 5Float variables, 1 Integer variable and 12 object variables.
3. We have renamed values for columns which have naming inconsistency- renamed the values for Gender and Account segment variable.
4. We have kept same naming for Male and Female data, earlier we had Male, Female, F and M values for this column. So converted F and M to Female and Male respectively.
5. We have renamed values for Account segment variable, to deal with naming inconsistencies. Regular + to Regular_Plus and Super + to Super_Plus
6. There are duplicate records in the dataset, around 259, we have removed them.
7. There are also a lot of special characters in every column as well as missing values around 3616, we have replaced special characters with null value, and to compute null value we will see the distribution of numerical variables first and decide on appropriate method for imputation of null values.

Univariate Analysis:

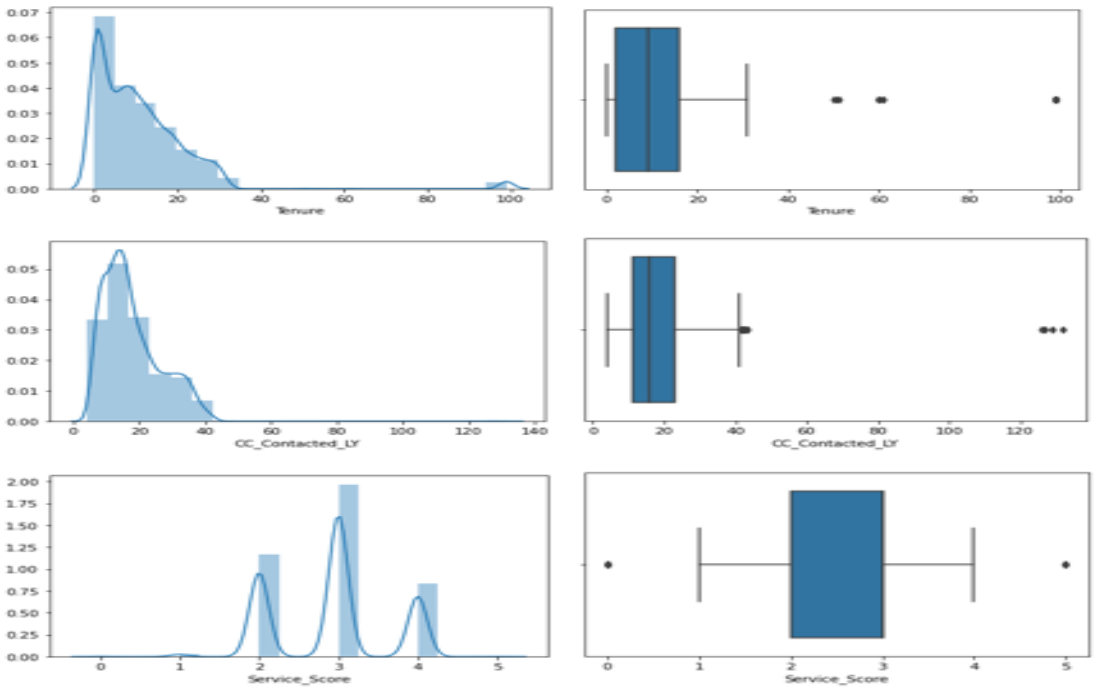


Fig 2: Distribution plot for numerical variables 1

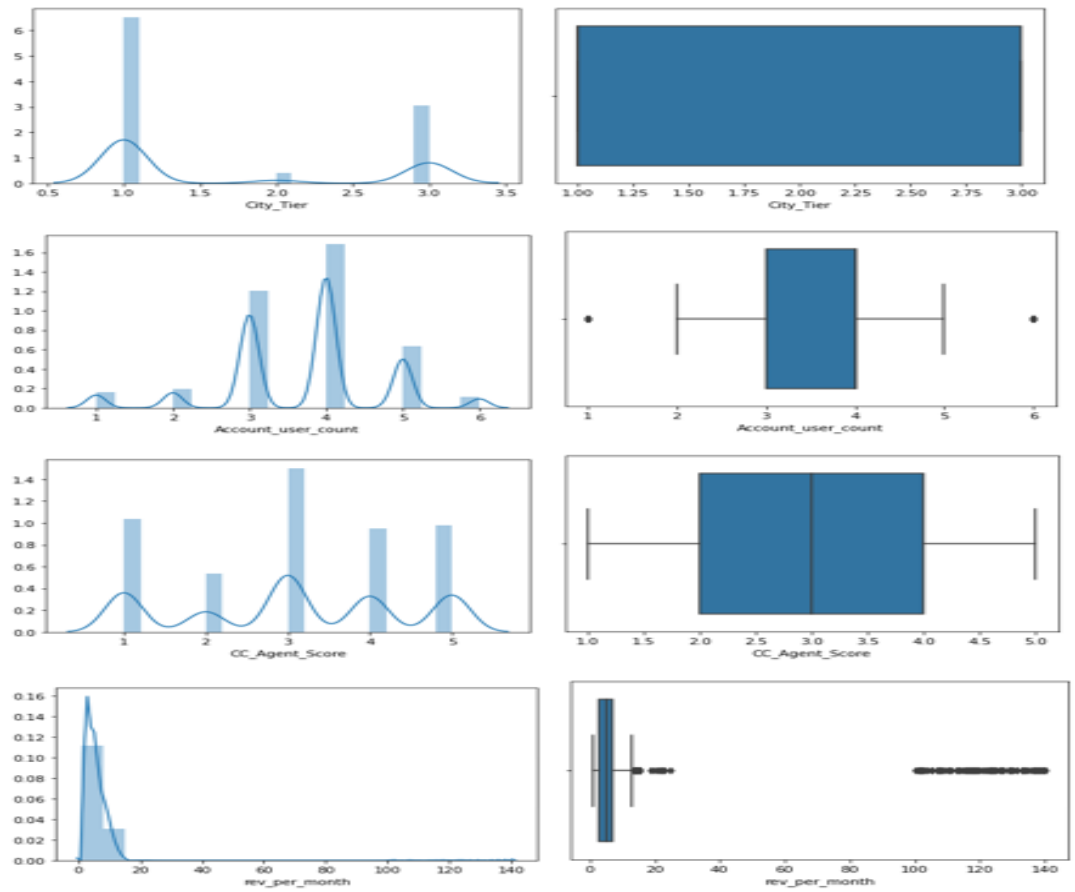


Fig 3: Distribution plot for Numerical variables 2

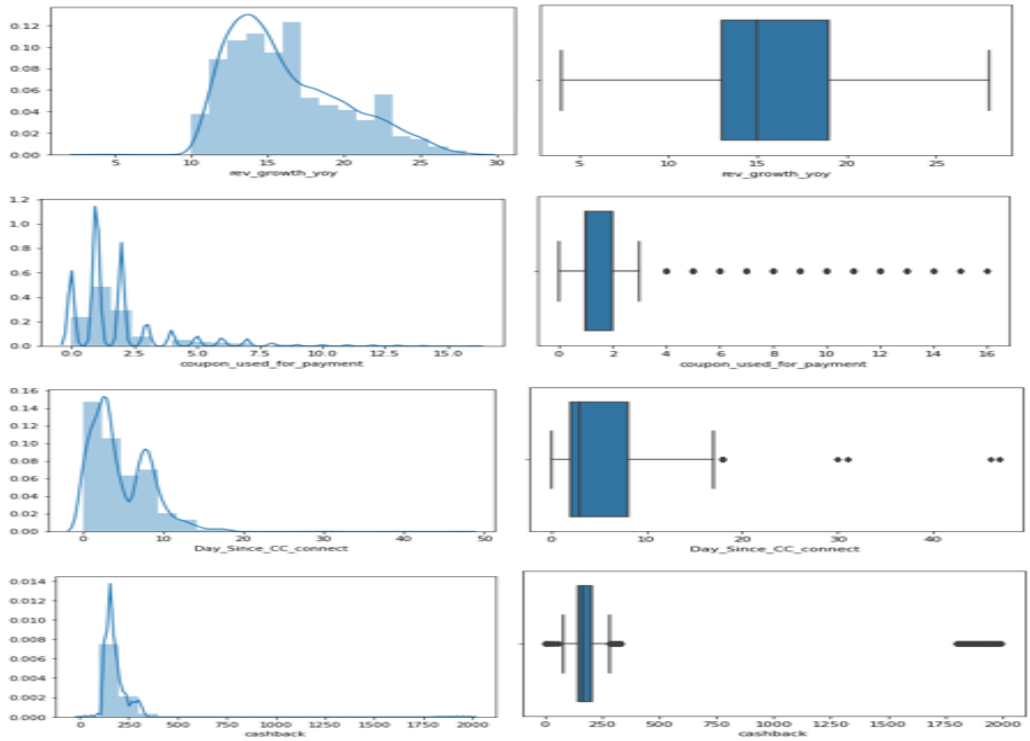


Fig 4: Distribution plot for Numerical variables 3

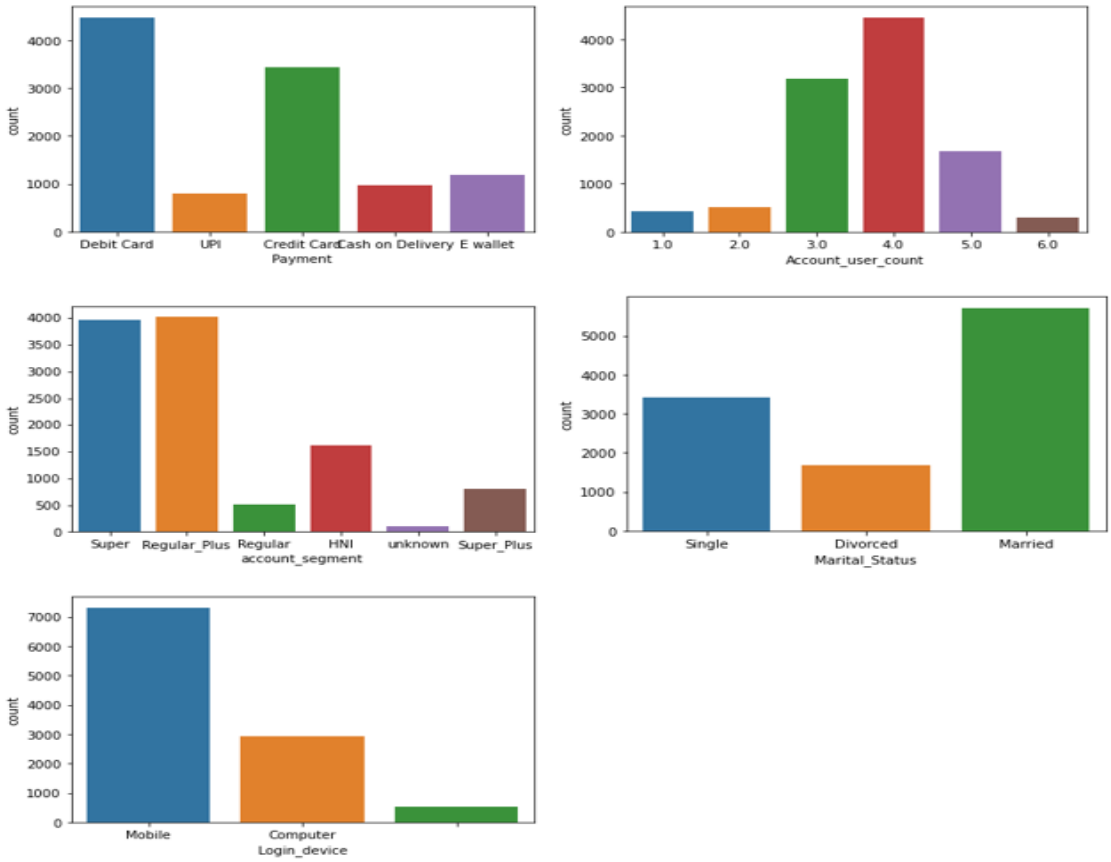


Fig 5: Count plot for categorical variables

Observations:

Univariate analysis of variables has given the below results:

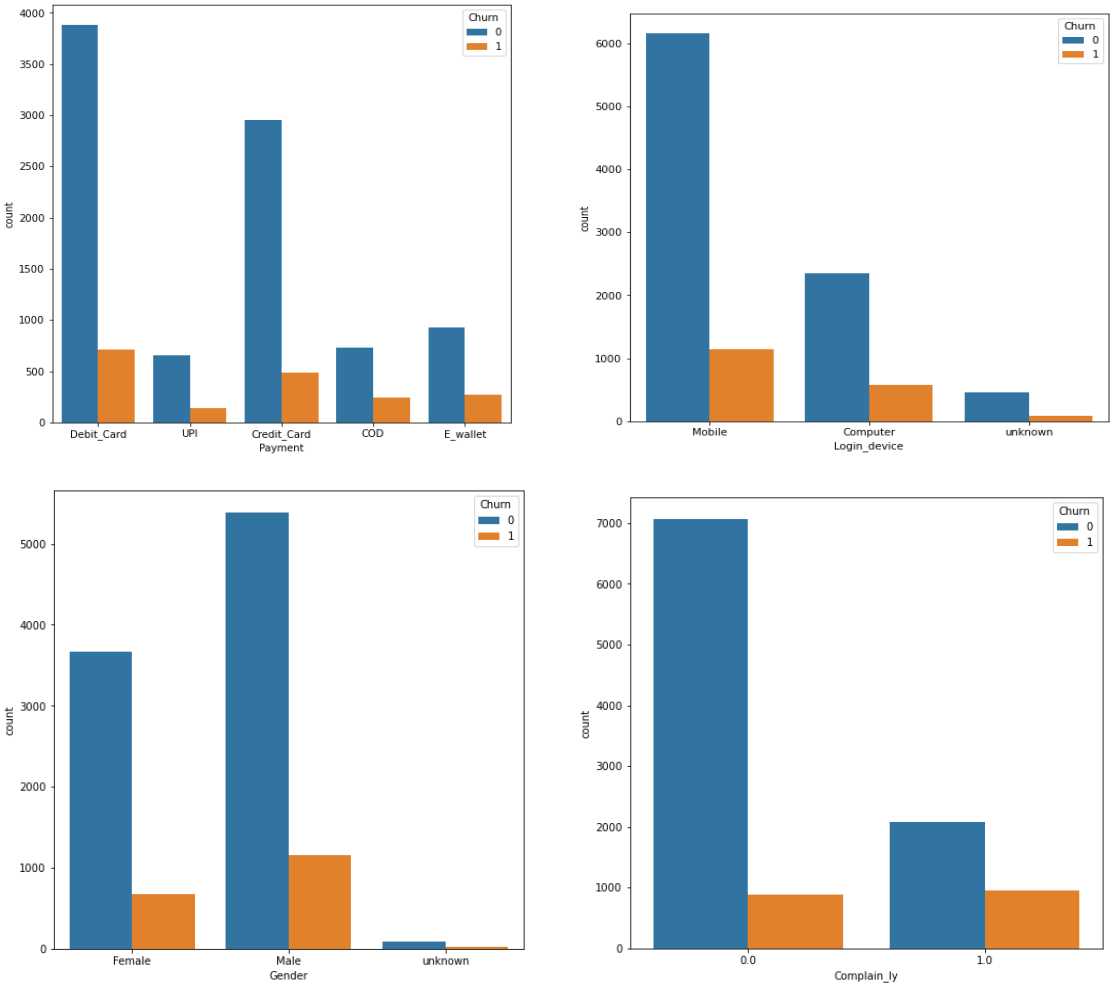
- **Numerical variables** like Tenure, CC_Contacted_LY, Rev_growth_yoy, Rev_per_month, Days_since_cc_connect, Cashback have shown a right skewed distribution. Box plot for these also have shown a lot of outliers at the extreme right end.
- Average Tenure for an account is 11 years, average number of times a customer contacted customer care (CC_Contacted_LY) is 17 times. Average Revenue for a month is 6.4, Average revenue_growth_yoy is 16. Average number of days since the customer contacted customer care is 4.6 days. Average Cashback generated is 196.
- Service_score, City_tier, Account_user_count, CC_Agent_score is represented as numerical variables, but we can assume them to be as categorial variables. Service score value ranges from 0 to 5, most common value being 3 least is 1. City_tier values range from 1 to 3, 1 has many accounts followed by City tier 3. Account_user_count value ranges from 1 to 6, the greatest number of users tagged to an account are 4, least is 1. CC_Agent score most common value is 3. Customers might have rated as 3 since it can be considered as somewhat middle score, since 1 is lowest and 5 is highest.
- For **categorical variables** like Payment mode, Account Segment, Marital status, and Login device, Debit card is the most preferred payment mode followed by Credit card, least preferred is UPI. Most popular account segment is Regular plus followed by Super. Married people are the greatest number of users of the DTH service. Mobile device is most preferred login device used for DTH service. We can now plot bivariate plots and study the effect of Churn variable on other variables.

```
0    0.831652
1    0.168348
```

Fig 6: Distribution of Target variable ‘Churn’

- 0 means a customer is retained, 1 means customer has churned
- The DTH company kept 83% of its users. Since the data is skewed, the number of instances in the 'Retained' class outnumbers the number of instances in the 'Churned' class by a lot.
- But since Industry rate of churning is 14 to 16% for DTH companies, this distribution does not need oversampling. If needed, we can decide it later by building models and analysing their performance metrics.

Bivariate Analysis:



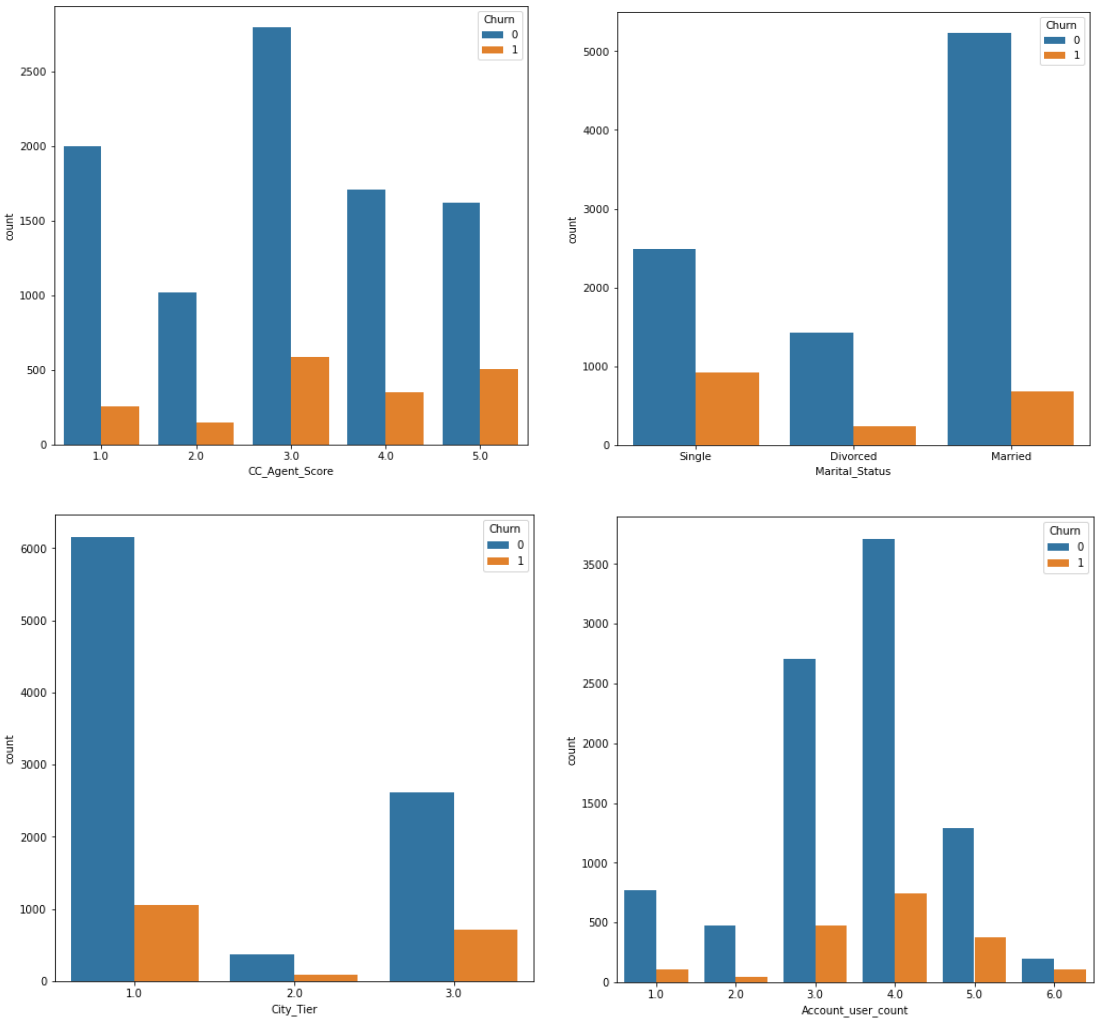


Fig 7: Bivariate analysis for Churn and other variables

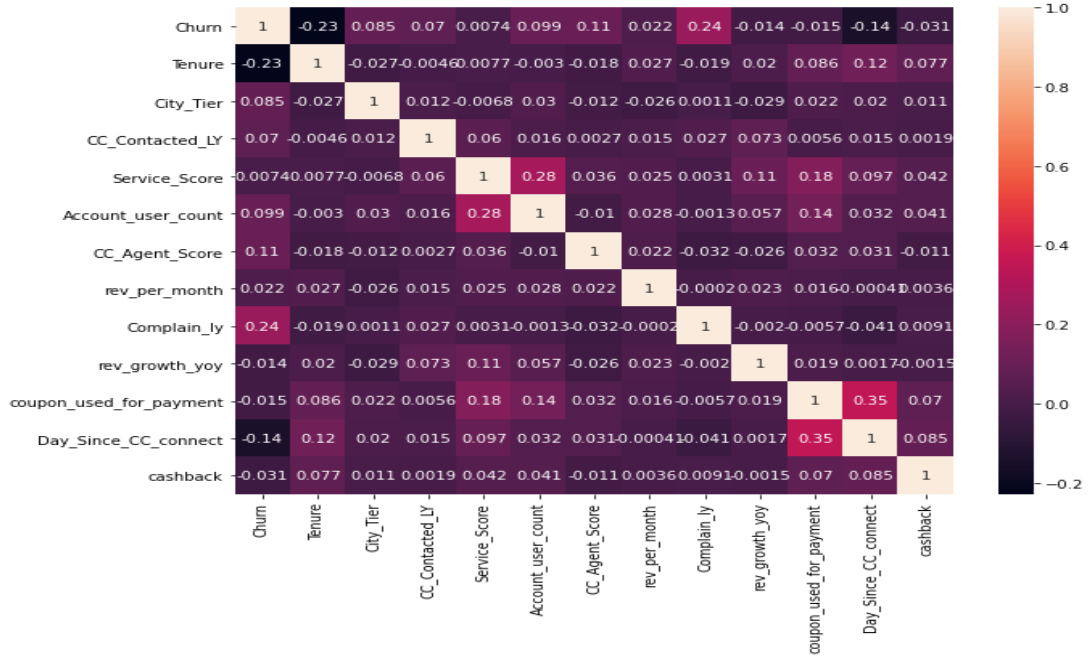


Fig 8: Heatmap for all variables

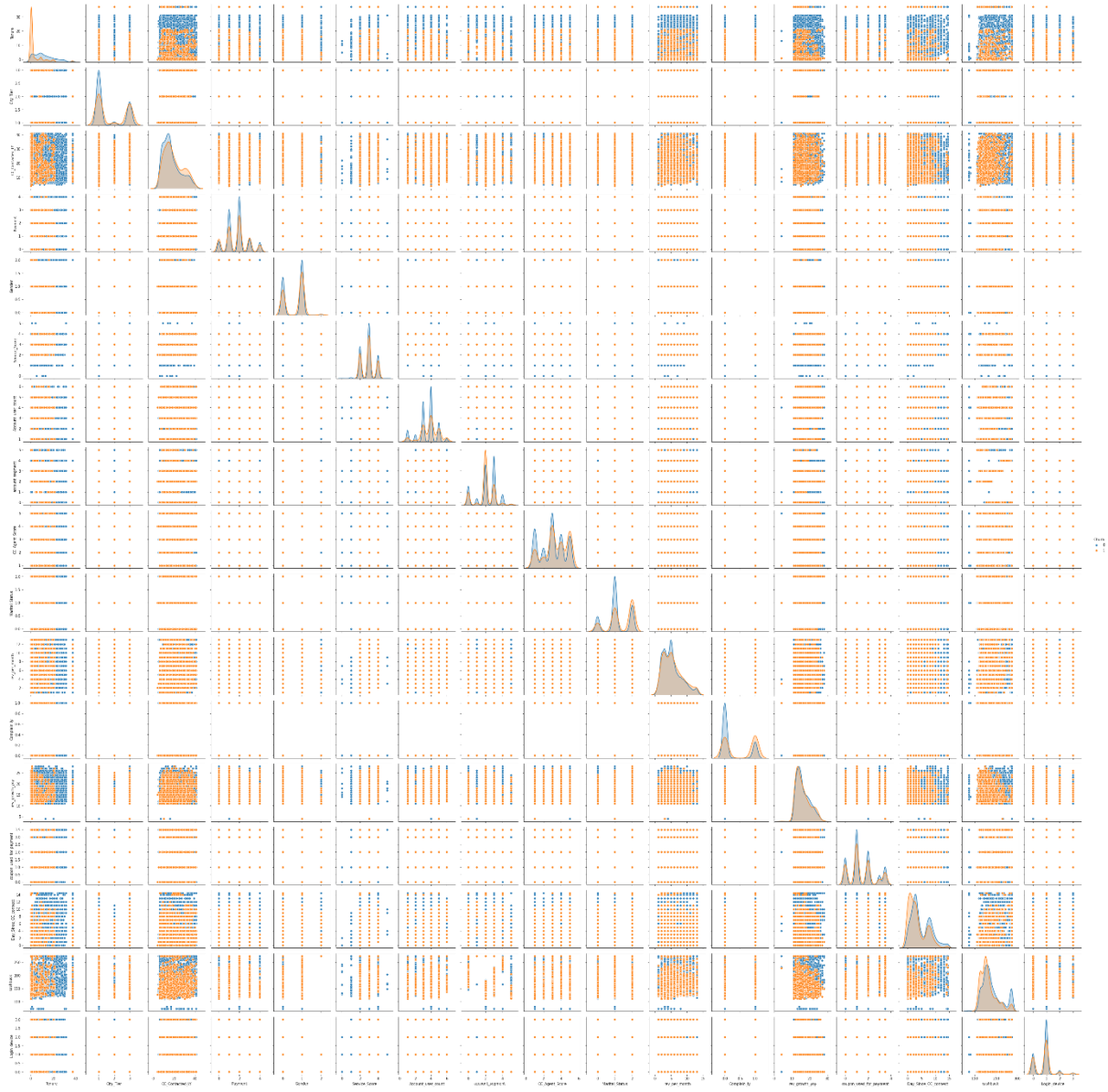


Fig 9: Pairplot for the variables

Observations:

- Tenure: Tenure variable does not seem to have significant effect on churn rate, average Tenure is 11 years.
- Payment method: Most preferred payment method is Debit card since the number of Debit card users is highest around 3800 and a greater number of users that have churned were using Debit card. Least preferred method is UPI, having around 520 users out of which 200 have churned. This number is more than Debit card users that have churned. So, customers using UPI are more highly likely to churn.
- Login device: Mobile users are the highest number of DTH users around 6000 from which 1000 have churned. Similar can be said for users using Computer. Using a particular type of device does not seem to affect churn rate that much.
- Gender: Male users are highest around 5000 and up, and that have churned are around 1000, we can say that male users are highest users that have churned. But this variable does not

have a significant impact on churn rate, since number of male users continuing the service is higher.

- Complain_ly: Customers that have contacted the customer care highest number of times are more likely to churn
- CC_Agent_score: Customers that have given low rating to the Agents are most likely to churn since they might be dissatisfied with the service of the agent, thus resulting to churn from the dth service.
- Marital_status: There are a greater number of users that are married around 5000 and up followed by Single users around 2500. The number of Single users that have churned is highest around 1000. So, we can say that Marital_status is affecting churn rate.
- City_Tier: Users from city tier 1 are highest around 6000 and around 1000 have churned from tier 1 city. City tier does not seem to have greater effect on churn rate.
- Account_user_count: Highest number of users tagged to an account is 4, and they are around 3500 and above out of which 550 customers have churned.
- There is no significant intercorrelation between our features, so we do not have to worry about multicollinearity.

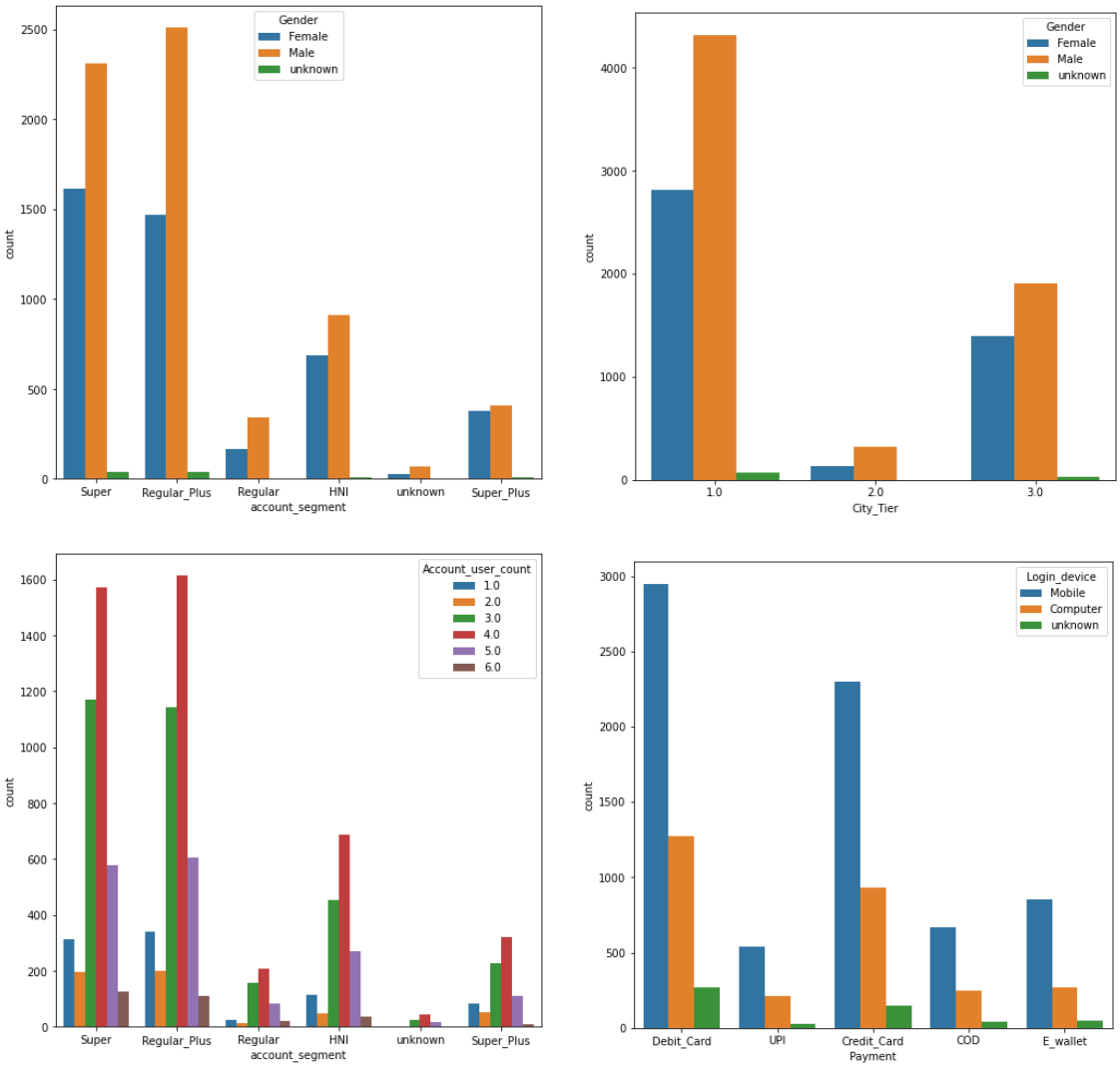


Fig 10: Bivariate analysis of other variables

Observations:

- Male users are the highest users for Super and Regular_plus account segment, around 2000 to 2500 users. Female users also seem to prefer these two account segments.
- City Tier 1 has highest number of Male and Female users, City tier 2 has least amount of users. More number of customers are likely to churn from this city tier.
- For both Super and Regular plus account segments, highest number of users tagged to a single account is 4, least tagged is 6. Regular account segment is not at all preffered by the customers. So Regular account segment users are more likel to churn.
- Number of users preferring Debit card payment and using Mobie devices are highest. UPI and COD are less preferred.

3. DATA CLEANING & PRE-PROCESSING

Imputing Special characters

Tenure	#	116
Account_user_count	@	332
Rev_per_month	+	689
Coupon_used_for_payment	$,*,#	3
Rev_growth_yoy	$	3
Day_since_cc_connect	#,*,$	1
Login_device	&	539

Table 2: special characters for each attribute

- We have removed special characters by replacing them with blank values

Imputing Missing values

```
Churn                         0
Tenure                      218
City_Tier                   112
CC_Contacted_LY             102
Payment                     109
Gender                        0
Service_Score                98
Account_user_count          444
account_segment               0
CC_Agent_Score              116
Marital_Status              211
rev_per_month               791
Complain_ly                 357
rev_growth_yoy                3
coupon_used_for_payment       3
Day_Since_CC_connect        358
cashback                    473
Login_device                221
dtype: int64
```

Fig 11: Missing values for each attribute

- Before treating missing values, we convert some of the object variables into numerical variables to plot their distribution. After looking into the distribution, we can decide on appropriate missing value treatment method.

- We have checked the distribution plot earlier in the report.
- For right-skewed distribution we use Median to replace the missing value of the attribute. (Tenure, CC_Contacted_LY, rev_per_month, rev_growth_yoy, coupon_used_for_payment, Day_Since_CC_connect, cashback).
- For distributions that have outliers or are skewed, the median is often the preferred measure of central tendency because the median is more resistant to outliers than the mean. The mean is pulled in the direction of the skewness; hence we have used median here.
- For categorical variables like Gender and account segment we replace missing values with a new value 'Unkown'. And for some variables we have used frequently occurring variable to replace missing value.

Imputing Outliers

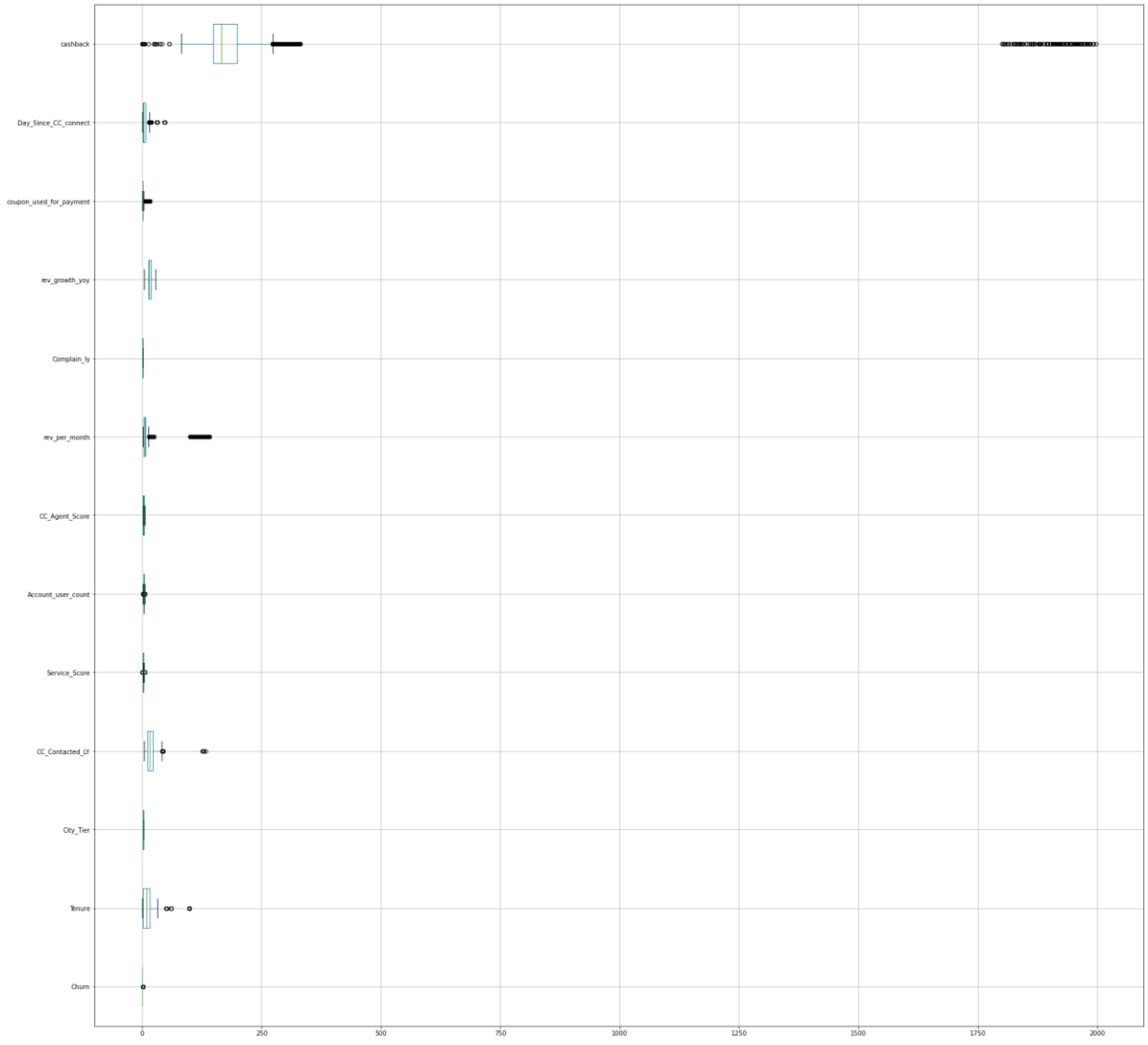


Fig 12: Box plot to check outliers

- We will treat outliers for cashback, days_since_cc_connect, coupon_used_for_payment, rev_per_month, CC_Contacted_LY, Tenure. We will not treat outliers for account user count and and service score as they are ratings.

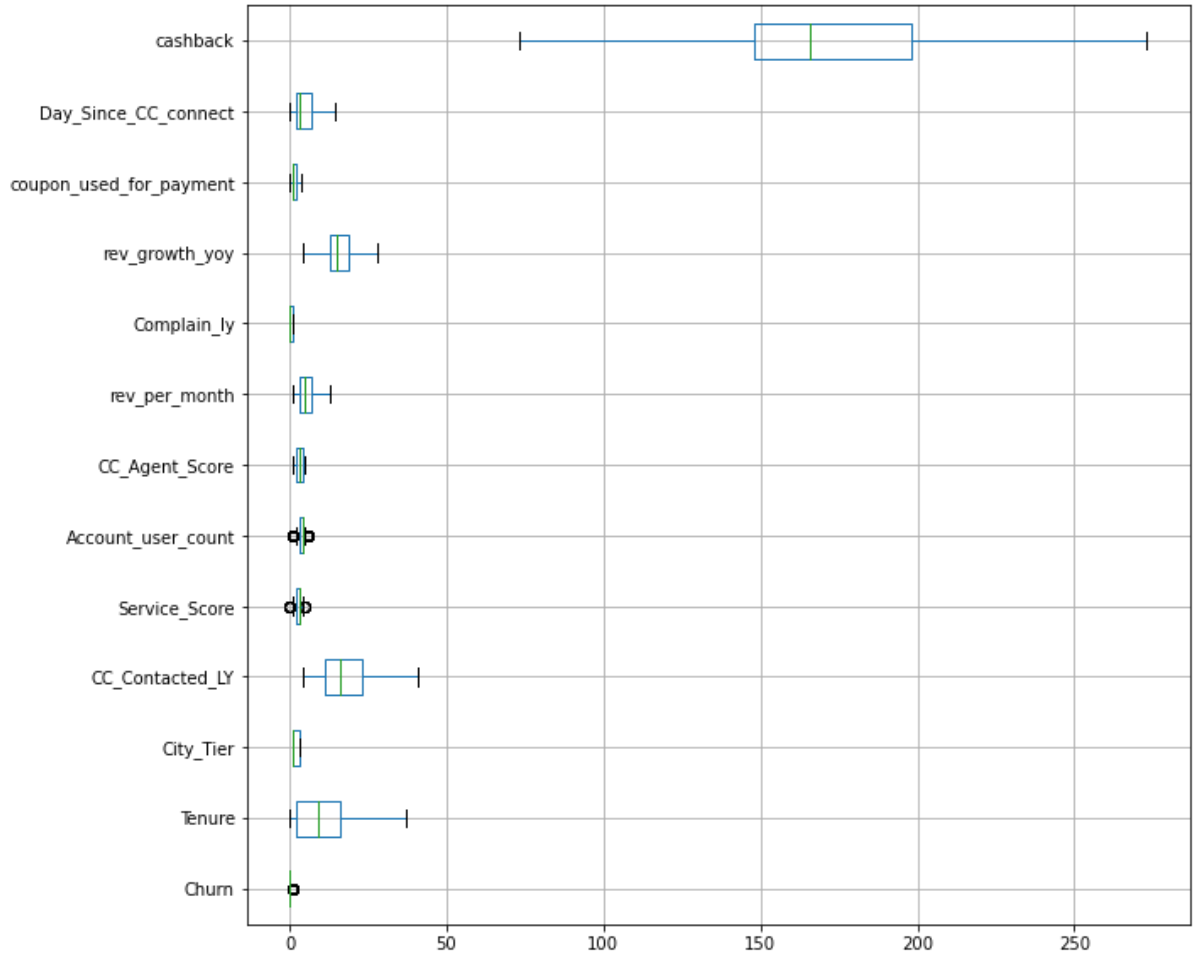


Fig 13: Box plot after outlier treatment

Variable Transformation

- We use label encoder to transform some categorical variables into numeric variables
- Assigning numerical values and storing in another column for below attributes:
- Payment, Gender, account_segment, Marital_Status and Login_device
- We have used Label encoding since most of the categorical variables are ordinal and number of categories is large.

4. MODEL BUILDING

Models Build:

- Applied Logistic regression
- Applied LDA
- Applied KNN
- Applied Naïve Bayes
- Applied Random Forest
- Applied ensemble methods

Before applying machine learning models on the dataset, we split it into Train and Test data in 70:30 ratio.

We have scaled the data using z score, since the range for some numerical attributes are quite high and some machine learning algorithms like KNN are biased towards variables with high magnitude.

Initially we have chosen not to oversample the target variable since percentage of churn in this dataset is 0.16. And the standard Industry rate of churn for DTH companies is 14 to 16%. Therefore, we first build models with the data as it is.

	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status
0	-0.704189	1.479979	-1.380526	0.236140	-1.214159	0.135956	-0.515055	0.700766	-0.770012	1.273428
1	-1.153115	-0.709243	-1.147274	2.226773	0.760555	0.135956	0.366969	-0.166988	-0.041708	1.273428
2	-1.153115	-0.709243	1.418500	0.236140	0.760555	-1.247623	0.366969	-0.166988	-0.041708	1.273428
3	-1.153115	1.479979	-0.330891	0.236140	0.760555	-1.247623	0.366969	0.700766	1.414899	1.273428
4	-1.153115	-0.709243	-0.680769	-0.759177	0.760555	-1.247623	-0.515055	-0.166988	1.414899	1.273428

Fig 14: dataset after scaling

1) Logistic Regression

Feature importance calculated for Logistic Regression:

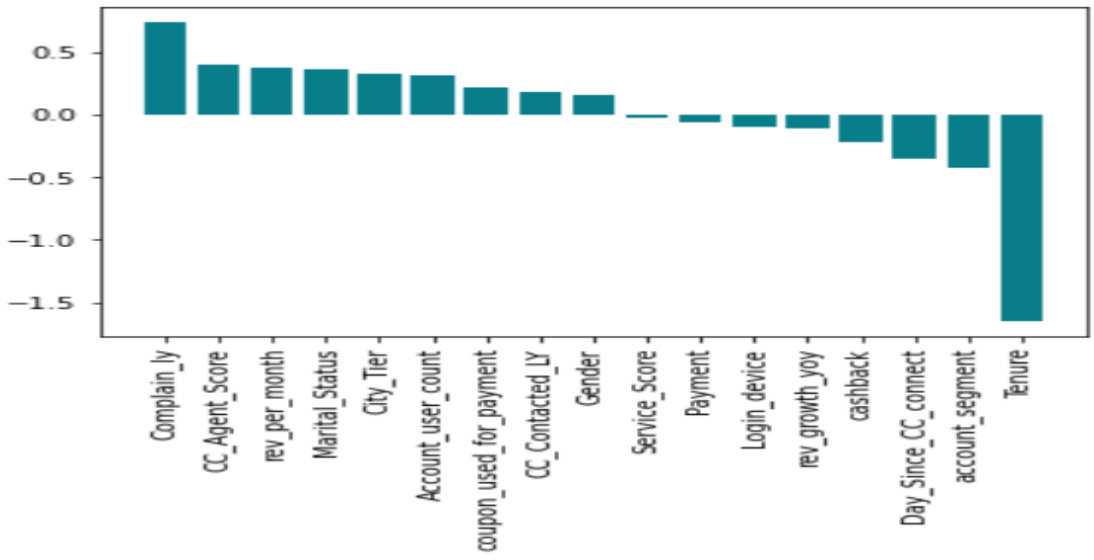


Fig 15: Feature Importance

Observations:

- Attributes with positive value for co-efficient are predictors of churn: Complaint_ly, CC_Agent_score, Rev_per_month, Marital_status, city_tier, account user count, coupon used for payment, cc contacted ly are the variables that are giving rise to customer churn.
- Attributes with negative value for co-efficient: Tenure, Account_segment, Days_since_cc_connect, Cashback, rev_growth_yoy, Login_device, Payment method, service score are negative predictors of churn. These are the attributes that prevent customer from churning.
- We have chosen not to delete any columns from the dataset.

Classification Report on Training and Test data for Logistic Regression:

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.90	0.97	0.93	6406
1	0.74	0.45	0.56	1294
accuracy			0.88	7700
macro avg	0.82	0.71	0.75	7700
weighted avg	0.87	0.88	0.87	7700

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	2743
1	0.73	0.48	0.58	558
accuracy			0.88	3301
macro avg	0.82	0.72	0.76	3301
weighted avg	0.87	0.88	0.87	3301

Fig 16: Classification report on Train and Test data (logistic regression)

Logistic regression performs slightly better on Test data

2) LDA (Linear Discriminant Analysis)

We have chosen LDA model since the output variable is categorical and it is a classification problem.

Classification Report on Training and Test data for LDA:

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.89	0.97	0.93	6406
1	0.74	0.38	0.51	1294
accuracy			0.87	7700
macro avg	0.81	0.68	0.72	7700
weighted avg	0.86	0.87	0.86	7700

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.89	0.97	0.93	2743
1	0.72	0.41	0.52	558
accuracy			0.87	3301
macro avg	0.80	0.69	0.72	3301
weighted avg	0.86	0.87	0.86	3301

Fig 17: Classification report on Train and Test data (LDA)

LDA performs slightly better on Test data

3) KNN

Classification Report on Training and Test data for KNN:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	6406
1	0.96	0.89	0.93	1294
accuracy			0.98	7700
macro avg	0.97	0.94	0.96	7700
weighted avg	0.98	0.98	0.98	7700

	precision	recall	f1-score	support
0	0.96	0.98	0.97	2743
1	0.91	0.80	0.85	558
accuracy			0.95	3301
macro avg	0.93	0.89	0.91	3301
weighted avg	0.95	0.95	0.95	3301

Fig 18: Classification report on Train and Test data (KNN)

KNN model does not perform good on Test data, but the scores are relatively higher than the previous models

4) Naïve Bayes

	precision	recall	f1-score	support
0	0.91	0.94	0.93	6406
1	0.64	0.57	0.60	1294
accuracy			0.87	7700
macro avg	0.78	0.75	0.76	7700
weighted avg	0.87	0.87	0.87	7700

	precision	recall	f1-score	support
0	0.91	0.93	0.92	2743
1	0.61	0.54	0.57	558
accuracy			0.86	3301
macro avg	0.76	0.73	0.74	3301
weighted avg	0.86	0.86	0.86	3301

Fig 19: Classification report on Train and Test data (Naïve Bayes)

Naïve Bayes does not perform well on Test data

5) Random Forest

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6406
1	1.00	1.00	1.00	1294
accuracy			1.00	7700
macro avg	1.00	1.00	1.00	7700
weighted avg	1.00	1.00	1.00	7700

	precision	recall	f1-score	support
0	0.97	0.99	0.98	2743
1	0.95	0.84	0.89	558
accuracy			0.96	3301
macro avg	0.96	0.91	0.93	3301
weighted avg	0.96	0.96	0.96	3301

Fig 20: Classification report on Train and Test data (Random Forest)

Random Forest overfits on Train data, it performs good on Test data and the overall scores for this model are greater than other models

6) Ensemble Models- Bagging, Boosting

Bagging:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6406
1	1.00	1.00	1.00	1294
accuracy			1.00	7700
macro avg	1.00	1.00	1.00	7700
weighted avg	1.00	1.00	1.00	7700

	precision	recall	f1-score	support
0	0.97	0.99	0.98	2743
1	0.93	0.86	0.90	558
accuracy			0.97	3301
macro avg	0.95	0.93	0.94	3301
weighted avg	0.97	0.97	0.97	3301

Fig 21: Classification report on Train and Test data (Bagging)

Bagging overfits on Train data, it performs best on Test data and the overall scores for this model are greatest than other models.

Boosting:

	precision	recall	f1-score	support
0	0.92	0.96	0.94	6406
1	0.76	0.58	0.66	1294
accuracy			0.90	7700
macro avg	0.84	0.77	0.80	7700
weighted avg	0.89	0.90	0.89	7700

	precision	recall	f1-score	support
0	0.92	0.96	0.94	2743
1	0.74	0.59	0.66	558
accuracy			0.90	3301
macro avg	0.83	0.77	0.80	3301
weighted avg	0.89	0.90	0.89	3301

Fig 22: Classification report on Train and Test data (Boosting)

Boosting performs the same on Train and Test data

Overall Model scores:

Model	AUC		Recall		Accuracy	
	Train	Test	Train	Test	Train	Test
Logistic Regression	0.87	0.85	0.45	0.48	0.88	0.88
LDA	0.86	0.84	0.38	0.41	0.87	0.87
Naive Bayes	0.833	0.891	0.57	0.54	0.87	0.86
KNN	0.995	0.973	0.89	0.80	0.97	0.95
Random Forest	1	0.991	1	0.84	1	0.96
Bagging	0.950	0.933	1	0.86	1	0.97
Boosting	0.923	0.911	0.58	0.59	0.90	0.90

Table 3: Model scores

Observation:

- From EDA we know that Churn data was imbalanced, instances in retained class outnumbered instances in churned class by a lot. Therefore, Accuracy is probably not the best metric in this case for evaluating model performance and since we are trying to predict churning probability, **we will focus on reducing possibility of False negative. So primary criteria for evaluation will be Recall first and then Accuracy**
- These are the model scores before applying any oversampling technique or model tuning.
- The scores on Test data are very low for all the models except KNN, Random Forest and Bagging.
- Random Forest and Bagging are overfit on Train data for the score Recall and Accuracy.
- KNN performs good on Train and Test data.

Now we use SMOTE on Churn data and then build the same models using oversampled data

1) Logistic Regression after SMOTE

Classification report on Train and Test data

	precision	recall	f1-score	support
0	0.81	0.77	0.79	6378
1	0.78	0.82	0.80	6430
accuracy			0.80	12808
macro avg	0.80	0.80	0.80	12808
weighted avg	0.80	0.80	0.80	12808

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.81	0.77	0.79	2771
1	0.78	0.82	0.80	2719
accuracy			0.80	5490
macro avg	0.80	0.80	0.80	5490
weighted avg	0.80	0.80	0.80	5490

Fig 23: Classification report on Train and Test data (Logistic regression after smote)

Logistic regression after oversampling performs the same on Train and Test data

2) LDA after SMOTE

Classification report on Train and Test data

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.82	0.75	0.78	6378
1	0.77	0.83	0.80	6430
accuracy			0.79	12808
macro avg	0.79	0.79	0.79	12808
weighted avg	0.79	0.79	0.79	12808

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.82	0.75	0.78	2771
1	0.76	0.83	0.80	2719
accuracy			0.79	5490
macro avg	0.79	0.79	0.79	5490
weighted avg	0.79	0.79	0.79	5490

Fig 24: Classification report on Train and Test data (LDA after smote)

LDA after oversampling performs the same on Train and Test data

3) KNN after SMOTE

	precision	recall	f1-score	support
0	1.00	0.95	0.97	6378
1	0.95	1.00	0.97	6430
accuracy			0.97	12808
macro avg	0.97	0.97	0.97	12808
weighted avg	0.97	0.97	0.97	12808

	precision	recall	f1-score	support
0	0.99	0.92	0.95	2771
1	0.92	1.00	0.96	2719
accuracy			0.96	5490
macro avg	0.96	0.96	0.96	5490
weighted avg	0.96	0.96	0.96	5490

Fig 25: Classification report on Train and Test data (KNN after smote)

KNN performs slightly less good on Test data. It is overfit on Recall for both Train and Test data

4) Naïve Bayes after SMOTE

	precision	recall	f1-score	support
0	0.68	0.93	0.79	6378
1	0.90	0.56	0.69	6430
accuracy			0.75	12808
macro avg	0.79	0.75	0.74	12808
weighted avg	0.79	0.75	0.74	12808

	precision	recall	f1-score	support
0	0.68	0.94	0.79	2771
1	0.89	0.56	0.69	2719
accuracy			0.75	5490
macro avg	0.79	0.75	0.74	5490
weighted avg	0.79	0.75	0.74	5490

Fig 26: Classification report on Train and Test data (Naïve Bayes after smote)

Performance for Naïve Bayes has not improved after oversampling

5) Random Forest

	precision	recall	f1-score	support
0	0.92	1.00	0.96	6378
1	1.00	0.92	0.96	6430
accuracy			0.96	12808
macro avg	0.96	0.96	0.96	12808
weighted avg	0.96	0.96	0.96	12808

	precision	recall	f1-score	support
0	0.92	1.00	0.96	2771
1	1.00	0.91	0.95	2719
accuracy			0.95	5490
macro avg	0.96	0.95	0.95	5490
weighted avg	0.96	0.95	0.95	5490

Fig 27: Classification report on Train and Test data (Random Forest after smote)

Random Forest performs best between all models that were built after oversampling

6) Bagging Model after SMOTE

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6378
1	1.00	1.00	1.00	6430
accuracy			1.00	12808
macro avg	1.00	1.00	1.00	12808
weighted avg	1.00	1.00	1.00	12808

	precision	recall	f1-score	support
0	0.98	0.97	0.97	2771
1	0.97	0.98	0.97	2719
accuracy			0.97	5490
macro avg	0.97	0.97	0.97	5490
weighted avg	0.97	0.97	0.97	5490

Fig 28: Classification report on Train and Test data (Bagging after smote)

Bagging is overfit on Train data but performs well on Test data although scores have decreased on Test data

7) Boosting Model after SMOTE

	precision	recall	f1-score	support
0	0.89	0.90	0.90	6378
1	0.90	0.89	0.90	6430
accuracy			0.90	12808
macro avg	0.90	0.90	0.90	12808
weighted avg	0.90	0.90	0.90	12808

	precision	recall	f1-score	support
0	0.89	0.91	0.90	2771
1	0.90	0.88	0.89	2719
accuracy			0.90	5490
macro avg	0.90	0.90	0.90	5490
weighted avg	0.90	0.90	0.90	5490

Fig 29: Classification report on Train and Test data (Boosting after smote)

Boosting performs well on Test data after oversampling

5. MODEL VALIDATION

We will focus on reducing possibility of False negative. So primary criteria for evaluation will be Recall first and then Accuracy

Using single metrics is not the only way of comparing the predictive performance of classification models. The ROC curve (Receiver Operating Characteristic curve) is a graph showing the performance of a classifier at different classification thresholds. It plots the true positive rate (another name for recall) against the false positive rate.

Model	AUC		Recall		Accuracy	
	Train	Test	Train	Test	Train	Test
Logistic Regression	0.872	0.873	0.820	0.820	0.800	0.800
LDA	1.000	0.995	0.830	0.830	0.790	0.790
Naïve Bayes	0.834	0.831	0.810	0.820	0.740	0.740
KNN	1.000	0.995	1.000	1.000	0.970	0.960
Random Forest	0.999	0.997	0.920	0.910	0.960	0.950
Bagging	1.000	0.970	1.000	0.980	1.000	0.970
Boosting	0.964	0.960	0.890	0.880	0.900	0.900

Table 4: Model scores after oversampling

Observations:

- After oversampling the data, the scores of all the models have increased a lot.
- KNN, Random Forest and Bagging have best overall scores from all the models.
- But KNN has overfit on Recall on Train data as well as Test data
- Bagging has overfit on Train data on Recall as well as Accuracy
- Random Forest shows better scores on Recall as well as Accuracy for both Train and Test set, so we can choose best model as Random Forest as it has best overall scores.

6. INTERPRETATION & RECOMMENDATIONS

Interpretation:

- Tenure, Account_segment, Days_since_cc_connect, Cashback, rev_growth_yoy, Login_device, Payment method, service score are negative predictors of churn. These are the attributes that prevent customer from churning.
- Tenure variable does not seem to have significant effect on churn rate, average Tenure is 11 years. So, it is obvious that a customer who has stayed with DTH service for more year's greater than 11 years, is less likely to churn than customer who has a less duration service.
- Days since cc connect represents the least no of days after which customer has contacted customer care, average days being 4. This is also negative predictor for churn because if customer has not called customer care in a smaller number of days means he is satisfied with the service and is less likely to churn from DTH provider.
- Customer receiving more cashback, average 196 is less likely to churn.
- Customers belonging to higher Account segment, 'Regukar plus' and 'Super plus' are spending more and are more involved with services provided by DTH company, and hence less likely to churn.
- Attributes that are positive predictors for churn: Complaint_ly, CC_Agent_score, Rev_per_month, Marital_status, city_tier, account user count, coupon used for payment, cc contacted ly are the variables that are giving rise to customer churn.
- Complain_ly: Customers that have contacted the customer care highest number of times are more likely to churn
- CC_Agent_score: Customers that have given low rating to the Agents are most likely to churn since they might be dissatisfied with the service of the agent, thus resulting to churn from the dth service.
- Marital_status: There are a greater number of users that are married around 5000 and up followed by Single users around 2500. The number of Single users that have churned is highest around 1000. So, we can say that Marital_status is affecting churn rate.
- City_Tier: Users from city tier 1 are highest around 6000 and around 1000 have churned from tier 1 city. City tier does not seem to have greater effect on churn rate.
- Account_user_count: Highest number of users tagged to an account is 4, and they are around 3500 and above out of which 550 customers have churned.

Recommendations:

- While customer churn clearly is bad for business, organizations can implement strategies to all but eliminate customer churn.
- Improving customer service is at the top of the list. Agents should be provided with proper training to resolve customer issues, so that customer is less likely to give a bad feedback to the agent which in turn can reduce customer churn from the company.
- Try to address customer issue in least possible time to gain more customer satisfaction rate.
- Another way to avoid customer churn is to build customer loyalty through relevant experiences and personalized service. Another advisable tip is to offer incentives, such as discounts and special offers, to those customers who were identified as likely to defect. - Users in Top city tier can be provided with more services that are relevant to that city tier.

Male users can also be provided with discount. In this way, company is identifying customers who are at risk of churning and can work to improve their satisfaction.

- The company needs to give special attention to their most valuable customers because they bring the biggest revenue.
- Asking customers why they're cancelling is one of the easiest ways to get valuable insights that can save DTH business.
- Company can Build a relationship with them before they cancel.