CAPSTONE PROJECT REPORT

INDEX

APPENDIX FOR TABLES

APPENDIX FOR IMAGES

1.1 INTRODUCTION

A DTH company is facing a horrendous predicament; Customer Churn. It has become extremely difficult for the company to retain the customers. The company wants to develop a model to predict the customer churn. To exacerbate the situation, the company is keeping tracks of the accounts that are churning, however there are more than one user linked to the account. So, for e.g., an account has 5 customers, it would be losing a huge deal of business opportunities. Further, it is required to suggest a campaign in order to retain the customers. However, the campaign should be very unique and carefully planned keeping in mind the financial resources of the company. It would be a loss to the company if money spend to retain a single customer is more than the business that customer provides to the company.

1.2 SPECIFIC GOAL

1. Perform Descriptive analytics to understand the data.
2. Predicting churn beforehand.
3. Identify recommendations to reduce the churn rate.

2.1 UNIVARIATE ANALYSIS

1. Tenure

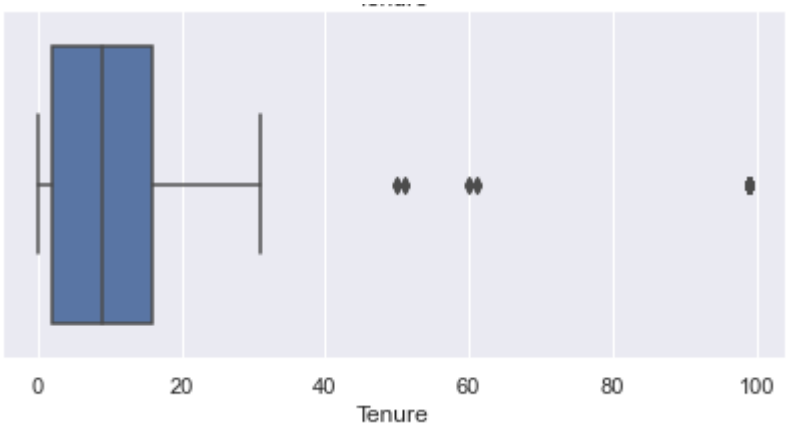


Figure 1: Tenure Boxplot

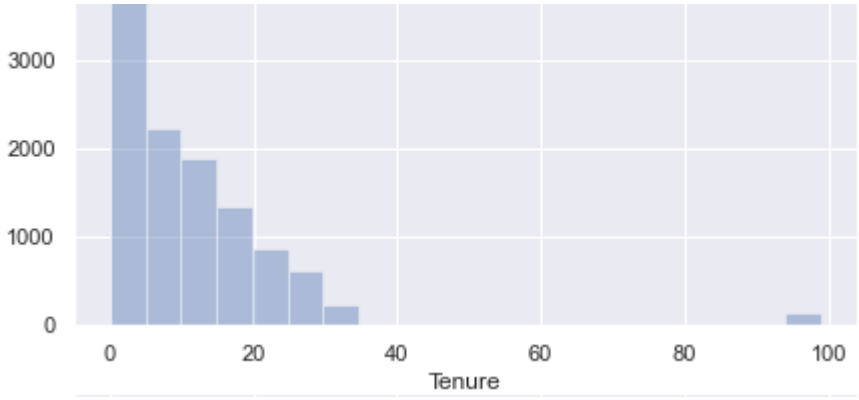


Figure 2: Histogram Tenure

There are a few outliers in the data and the distribution is right skewed. The outliers will be treated later, as it is quite unlikely that customers would have a tenure of more than 50 years in A DTH company

2. Service Score

Figure 3: Count plot of service score

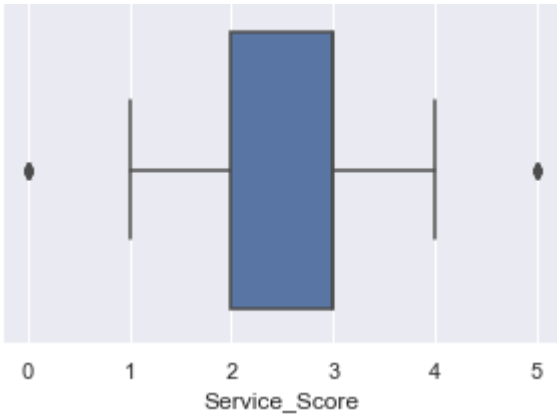


Figure 4: Boxplot Service Score

Churn	No	Yes
Service_Score		
0.0	8	0
1.0	77	0
2.0	2701	550
3.0	4554	936
4.0	1937	394
5.0	5	0

Quite surprising that customers with low Service score have not churned and vice versa.

3. CC Agent Score

Figure 6: Count plot of Agent score

Figure 5: Boxplot of Agent score

Majority customers have rated customer care as 3. Also, there are quite a many customers rated 4 and 5.

4. Complain Ly

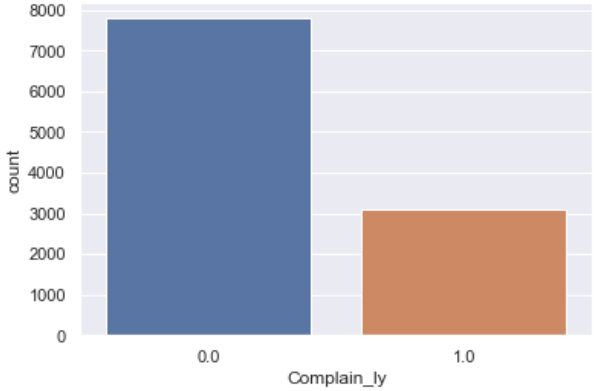


Figure 7: Count plot of complain ly

Complain_ly	0.0	1.0	All
Churn			
No	6942	2123	9065
Yes	850	988	1838
All	7792	3111	10903

Table 1: Churn vs complain count

0's are the ones who did not complain and 1's who complaint. More customers have not complaint. Also, those who churned, have registered complaints in majority.

5. Day Since Last Connect

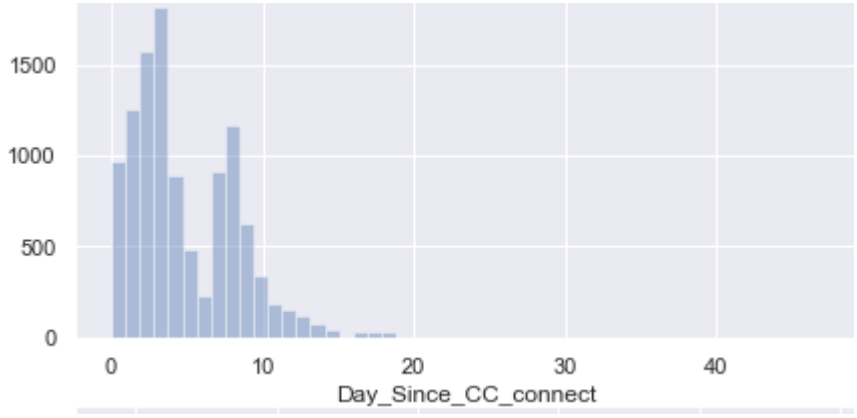


Figure 9: Distribution of last cc connect

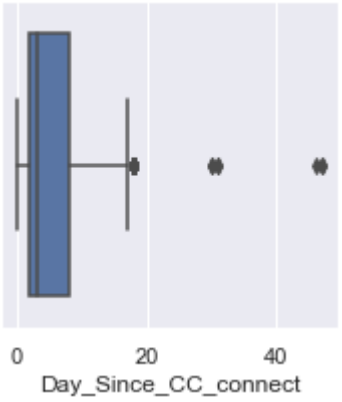


Figure 8: Boxplot last cc connect

More the value of this variable, more is the customer satisfied or he has completely churned.

6. Account user count

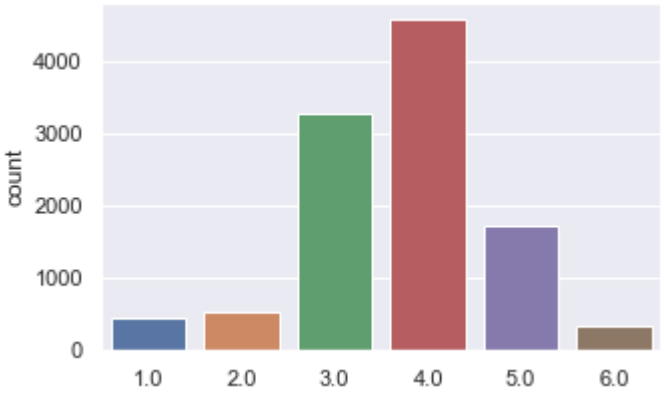


Figure 11: Count plot of account user count

Figure 10: Account user count boxplot

Most users have 4 customers' accounts linked, and these are the accounts providing max business.

2.2 BIVARIATE ANALYSIS

Heat Map: -

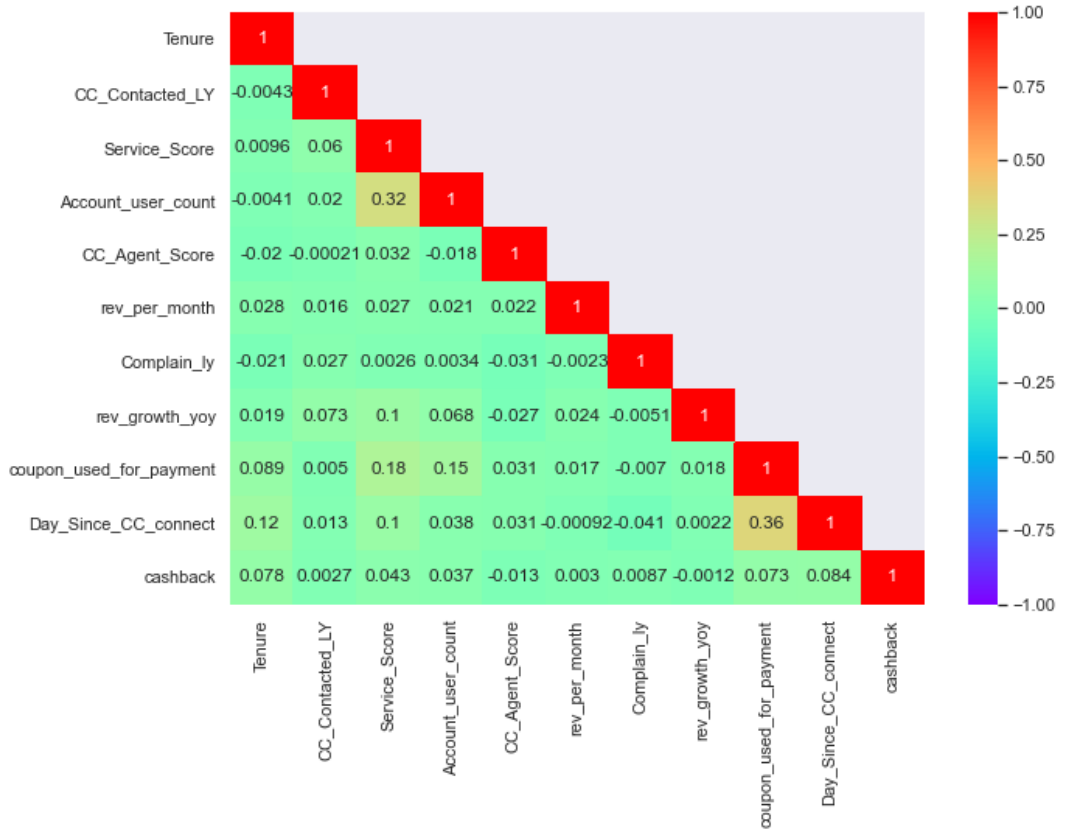


Figure 12: Heatmap

Clearly there is no multicollinearity in the data.

1. Churn vs Tenure

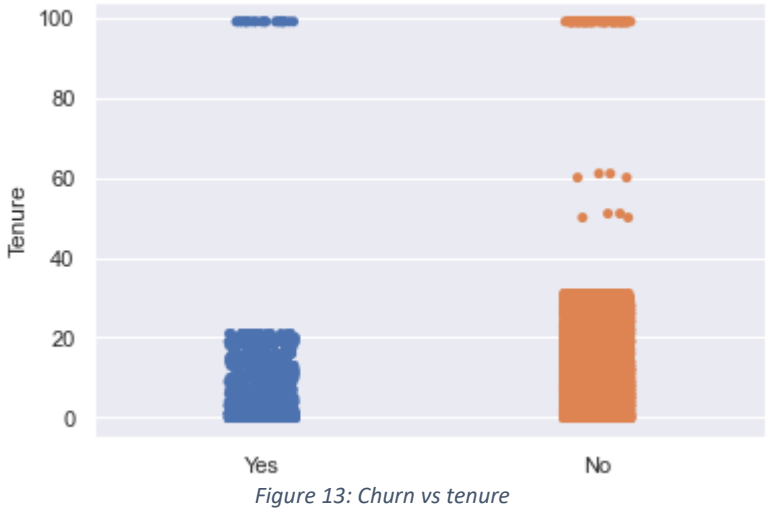


Figure 13: Churn vs tenure

Churn	Tenure
No	12.371329
Yes	4.327381

Table 2 Churn and mean tenure

Those who churned have a mean tenure of 4 years and that is pretty obvious, that they left due to unsatisfaction.

2. Churn vs CC last contacted

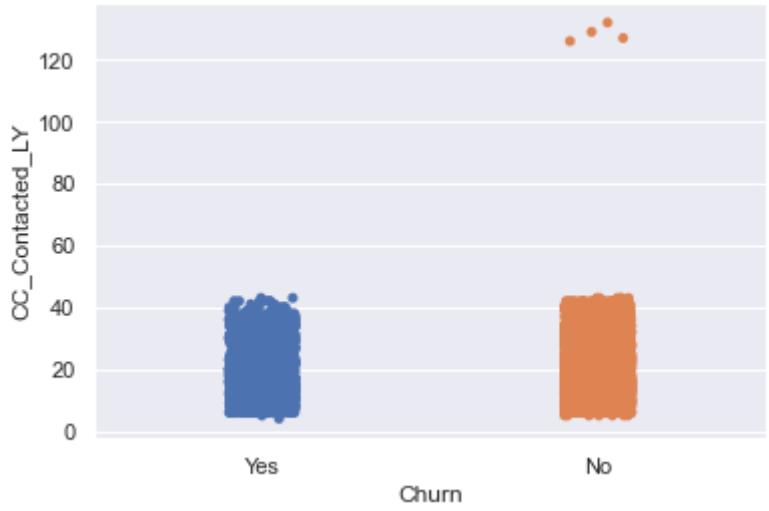


Figure 14: Churn vs CC last contacted

	Churn	CC_Contacted_LY	Service_Score	CC_Agent_Score
1309	No	126.0	2.0	1.0
4124	No	127.0	3.0	1.0
6939	No	132.0	2.0	1.0
9754	No	129.0	3.0	1.0

Table 3: Evidence, those rated low, still did not churn

From the scatterplot, it is quite difficult to distinguish, but from the table we can see those who contacted customer care way too more have not churned.

3. Service score vs Churn

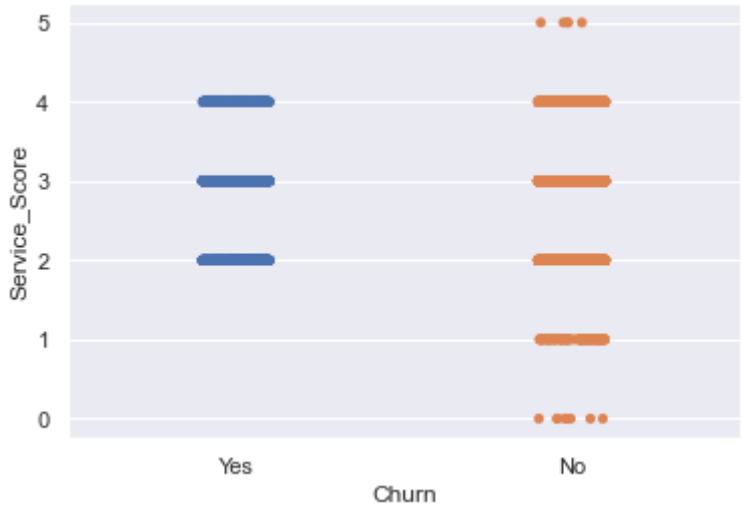


Figure 15: Churn vs service score

It can be seen that, even those customers who gave service score as 3 and 4 churned while those who gave way too less service scores still have not churned.

4. Account user count vs Churn

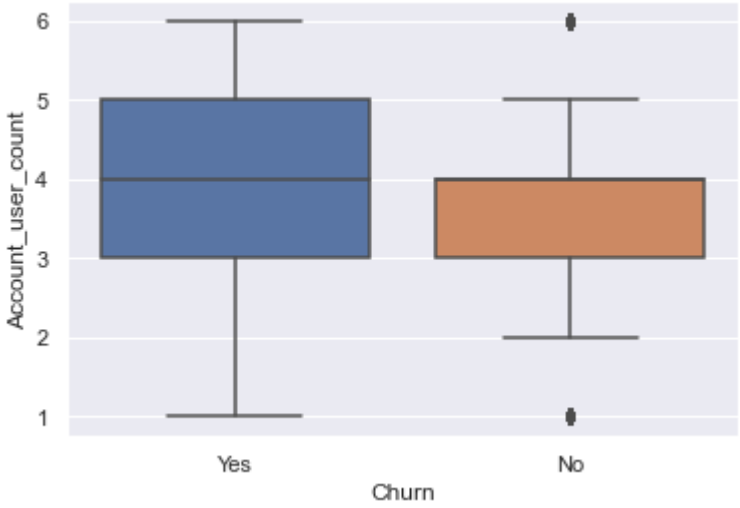


Figure 16: Churn vs account user count

Churn	Account_user_count
No	3.643334
Yes	3.936066

Table 4: Chur vs mean user count

Mean no of user count is more for churned customers which is quite alarming

5. CC Agent Score vs Churn

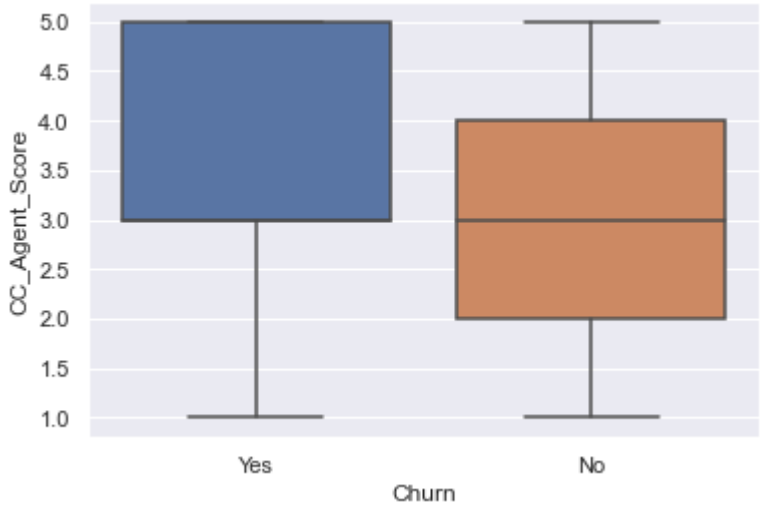


Figure 17: Churn vs CC agent score

Churn	CC_Agent_Score	Service_Score
No	3.000863	2.899591
Yes	3.391142	2.917021

Table 5: Churn vs mean cc and service score

Those who have churned have surprisingly higher mean agent scores and service scores than those who have not.

6. Revenue per month vs Churn

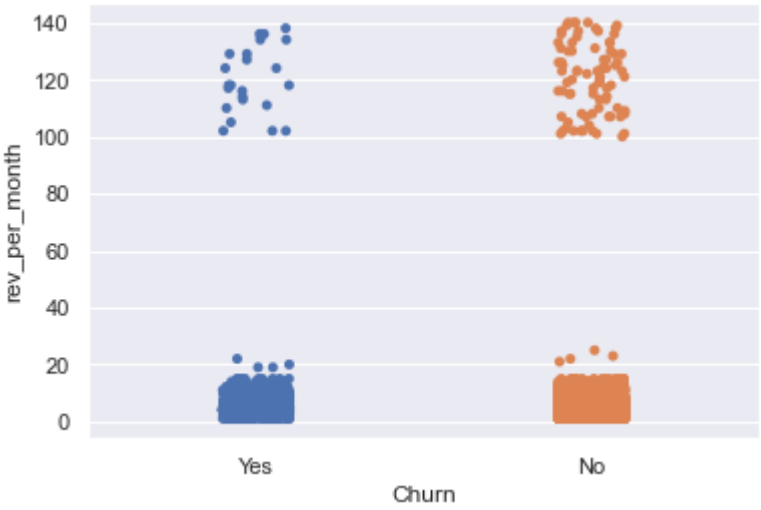


Figure 18: Churn vs rev per month

Churn	rev_per_month
No	6.241316
Yes	6.956620

Table 6: Churn vs mean rev per month

It is quite ameliorating that those with high revenue have retained more than those who have churned. There are only 23 customers in the higher range (>80) who have not churned. The

mean revenue is greater for those who churned; however, it is something not to be worried as more customers in the higher range have sustained.

7. Complaint vs Churn

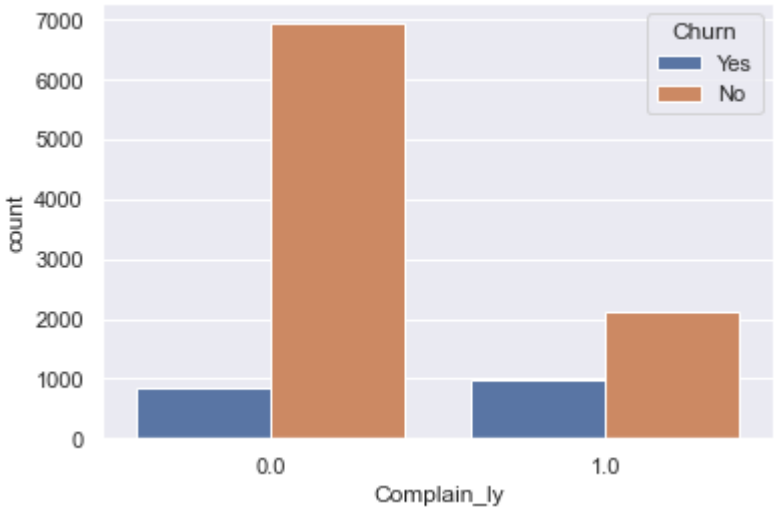


Figure 19: Churn vs Complain_ly

Complain_ly	0.0	1.0	All
Churn			
No	6942	2123	9065
Yes	850	988	1838
All	7792	3111	10903

Table 7: Churn vs complain count

Out of a total 1838, who complaint 988 churned which is 53%. This is pretty aghast for the company

8. Day Since last connect vs Churn

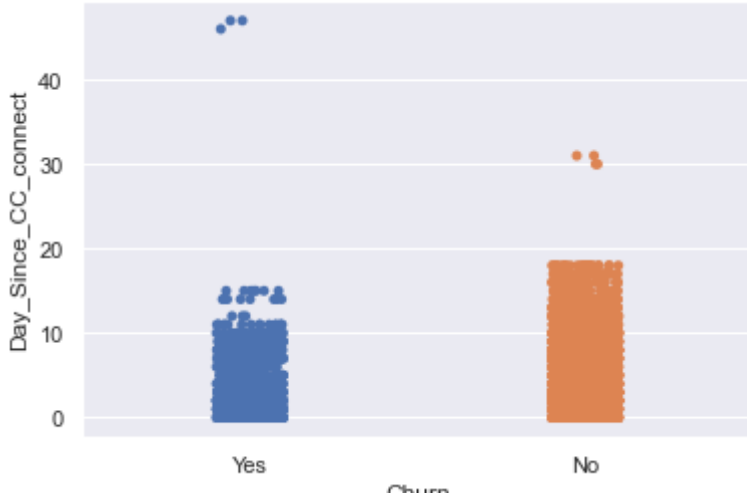


Figure 20: Churn vs day since last cc connect

	Day_Since_CC_connect
Churn	
No	4.879052
Yes	3.415939

Table 8: Churn vs mean last connect

Those who have churned have contacted the customer care on a mean of 3.41 days.

Churn	Day_Since_CC_connect	Service_Score	CC_Agent_Score	Tenure
No	4.879052	2.899591	3.000863	12.371329
Yes	3.415939	2.917021	3.391142	4.327381

Table 9: Churn vs mean of last connect, tenure, agent score and service score

It is quite surprising that those who churned have rated the service and customer care more than the ones who did not churn. In spite of a higher mean rating, they churned.

9. Cashback vs Churn

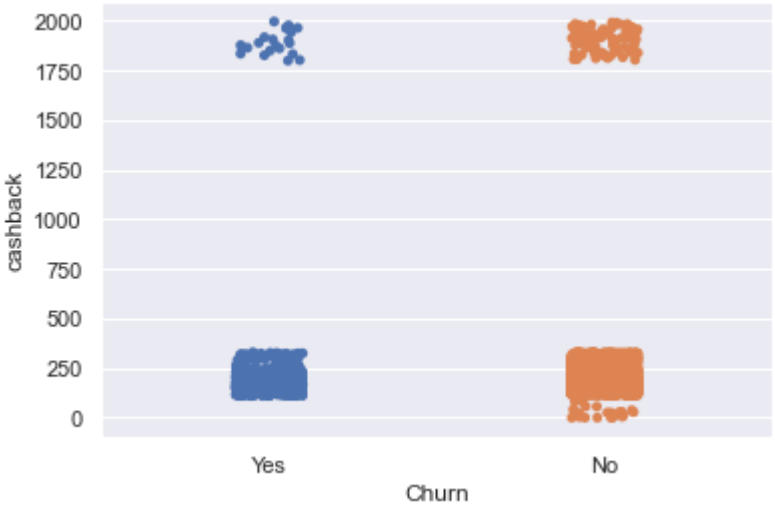


Figure 21: Churn vs cashback

Churn	cashback
No	198.834930
Yes	183.356924

Table 10: Churn vs mean cashback

Quite evident that customers with high cashback are less likely to churn.

10. City Tier vs Churn

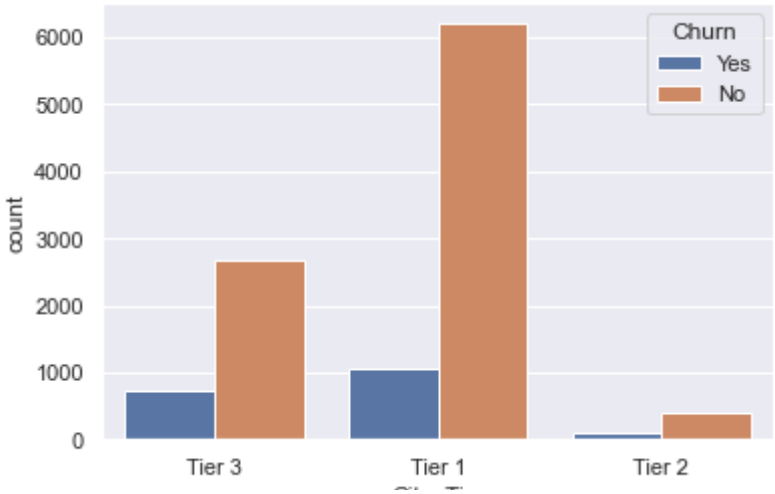


Figure 22: Churn vs city tier

Churn	No	Yes
City_Tier		
Tier 1	6207	1056
Tier 2	384	96
Tier 3	2678	727

Table 11: Churn vs city tier

Maximum churned customers are from Tier 1, 1056. These customers should be paid more attention to.

11. Payment vs Churn

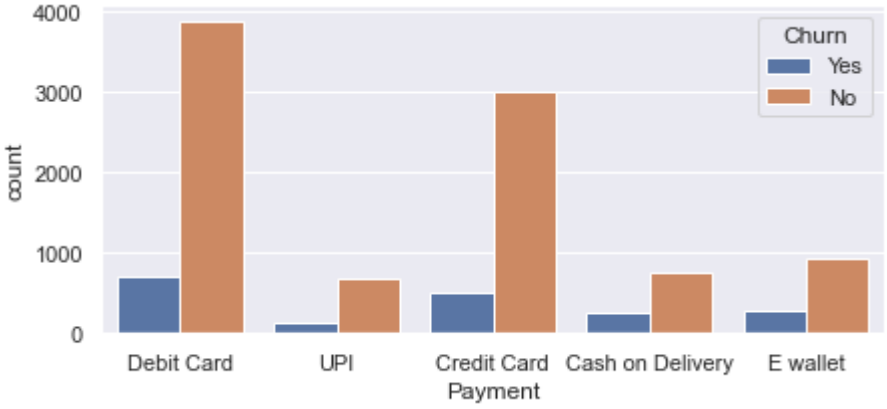


Figure 23: Churn vs Payment

Churn	No	Yes
Payment		
Cash on Delivery	760	254
Credit Card	3012	499
Debit Card	3885	702
E wallet	941	276
UPI	679	143

Table 12: Churn vs payment mode

Customers using Debit card have churned the most. Might be the reason there are not structured payment gateways for the payment of Debit Card

12. Account Segment vs Churn

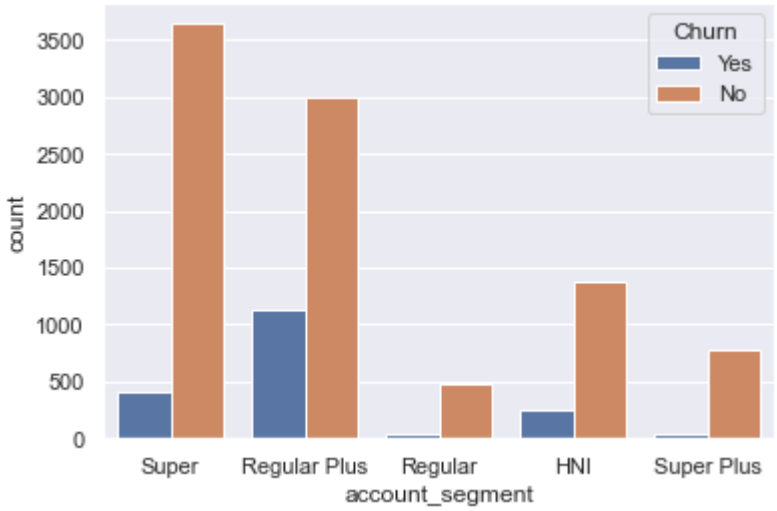


Figure 24: Churn vs Account segment

Churn	No	Yes
account_segment		
HNI	1384	255
Regular	480	40
Regular Plus	2997	1127
Super	3646	416
Super Plus	778	40

Table 13: Churn vs account type

Max % of customer churned from regular plus account.

13. Gender vs Churn

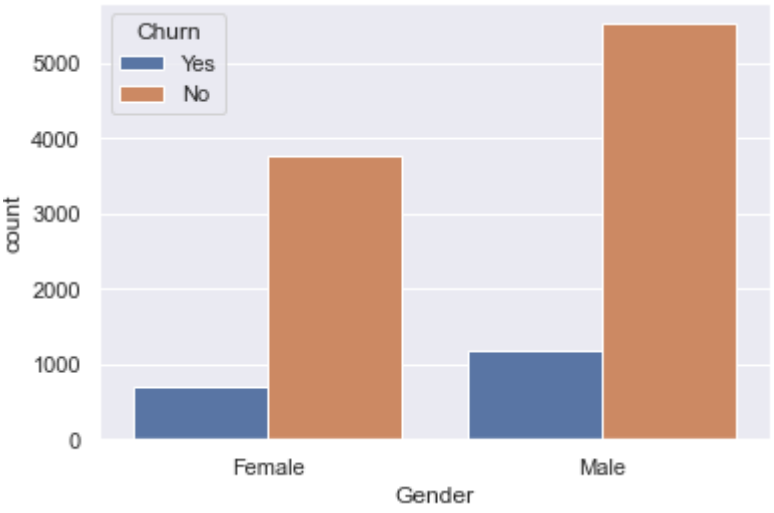


Figure 25: Churn vs gender

Churn	No	Yes
Gender		
Female	3759	689
Male	5519	1185

Table 14: Churn vs gender

Male customers have churned more than the female customers.

14. Login Devices vs Churn

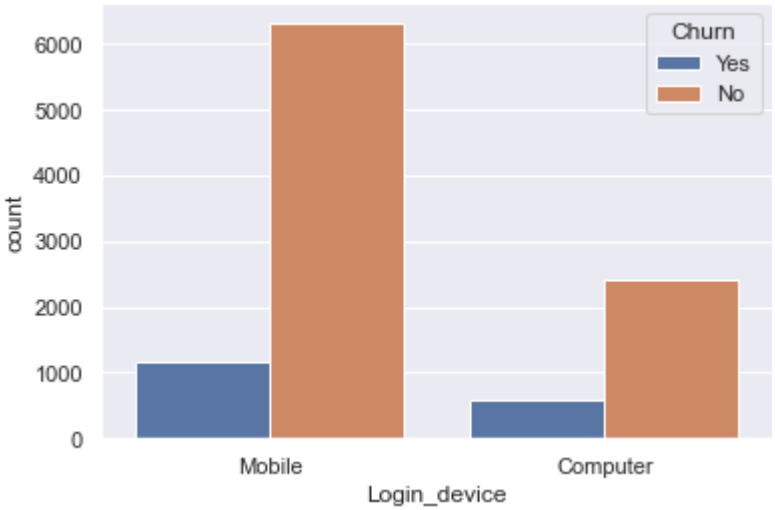


Figure 26: Login device vs churn

Churn	No	Yes
Login_device		
Computer	2421	597
Mobile	6310	1172

Table 15: Churn vs Log in device

Customers with mobile devices have churned more. Might be the reason that, the mobile website of the service is not very user friendly.

3. DATA CLEANING AND PREPROCESSING

3.1. Missing value treatment

	Null Values	Not Null	% Missingg	Data Type
AccountID	0	11260	0.000000	int64
Churn	0	11260	0.000000	int64
Tenure	218	11042	1.974280	float64
City_Tier	112	11148	1.004665	float64
CC_Contacted_LY	102	11158	0.914142	float64
Payment	109	11151	0.977491	object
Gender	108	11152	0.968436	object
Service_Score	98	11162	0.877979	float64
Account_user_count	444	10816	4.105030	float64
account_segment	97	11163	0.868942	object
CC_Agent_Score	116	11144	1.040919	float64
Marital_Status	212	11048	1.918899	object
rev_per_month	791	10469	7.555640	float64
Complain_ly	357	10903	3.274328	float64
rev_growth_yoy	3	11257	0.026650	float64
coupon_used_for_payment	3	11257	0.026650	float64
Day_Since_CC_connect	358	10902	3.283801	float64
cashback	473	10787	4.384908	float64
Login_device	760	10500	7.238095	object

Figure 27: Null values

As per standard rule (taught in the program), if the missing values are well below 15%, they can be dropped, hence the missing values are dropped.

3.2. Checking Duplicates

There are no duplicates in the data (Shown in the python file).

3.3. Treating outliers

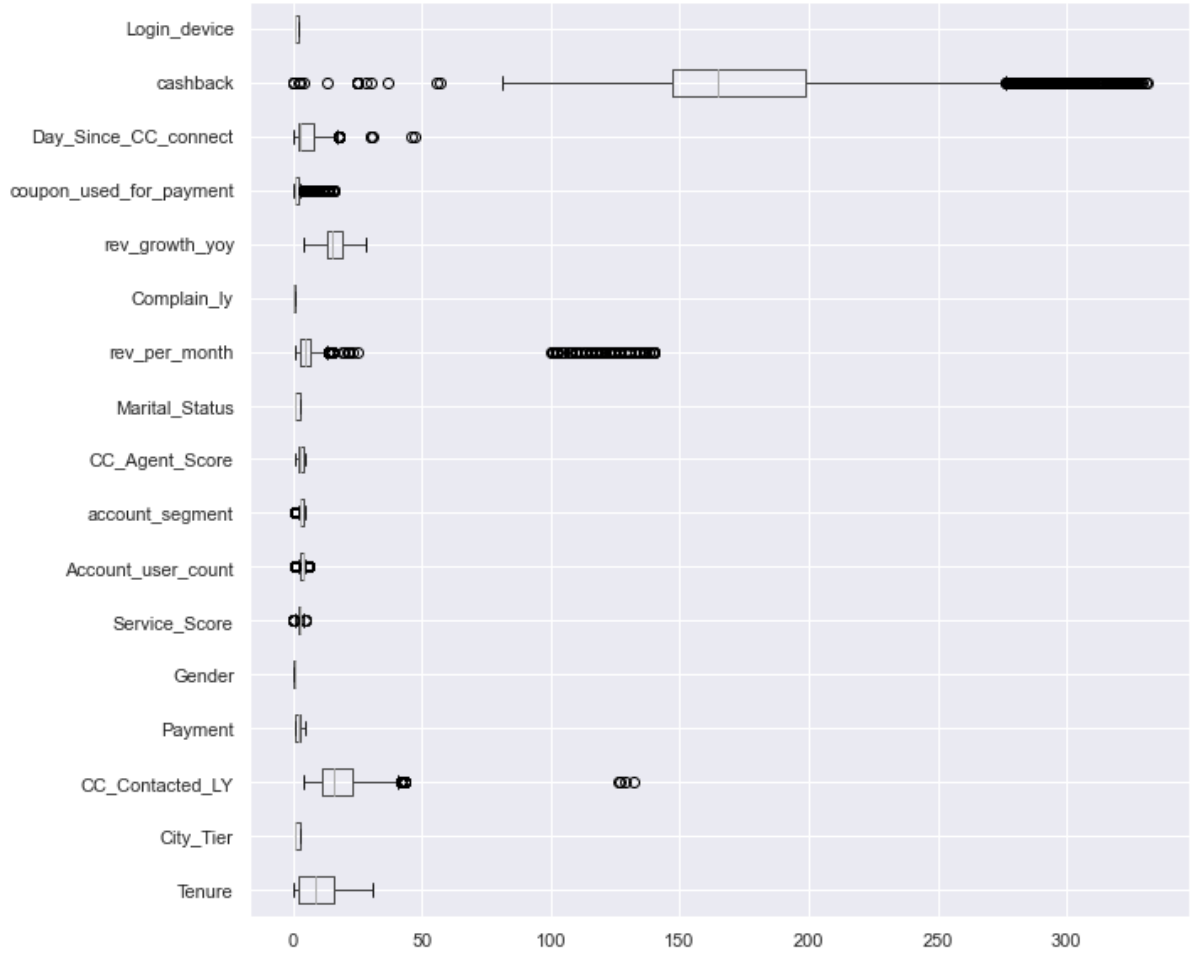


Figure 28: Outliers

There are outliers in the data and they are treated as per standard rule.

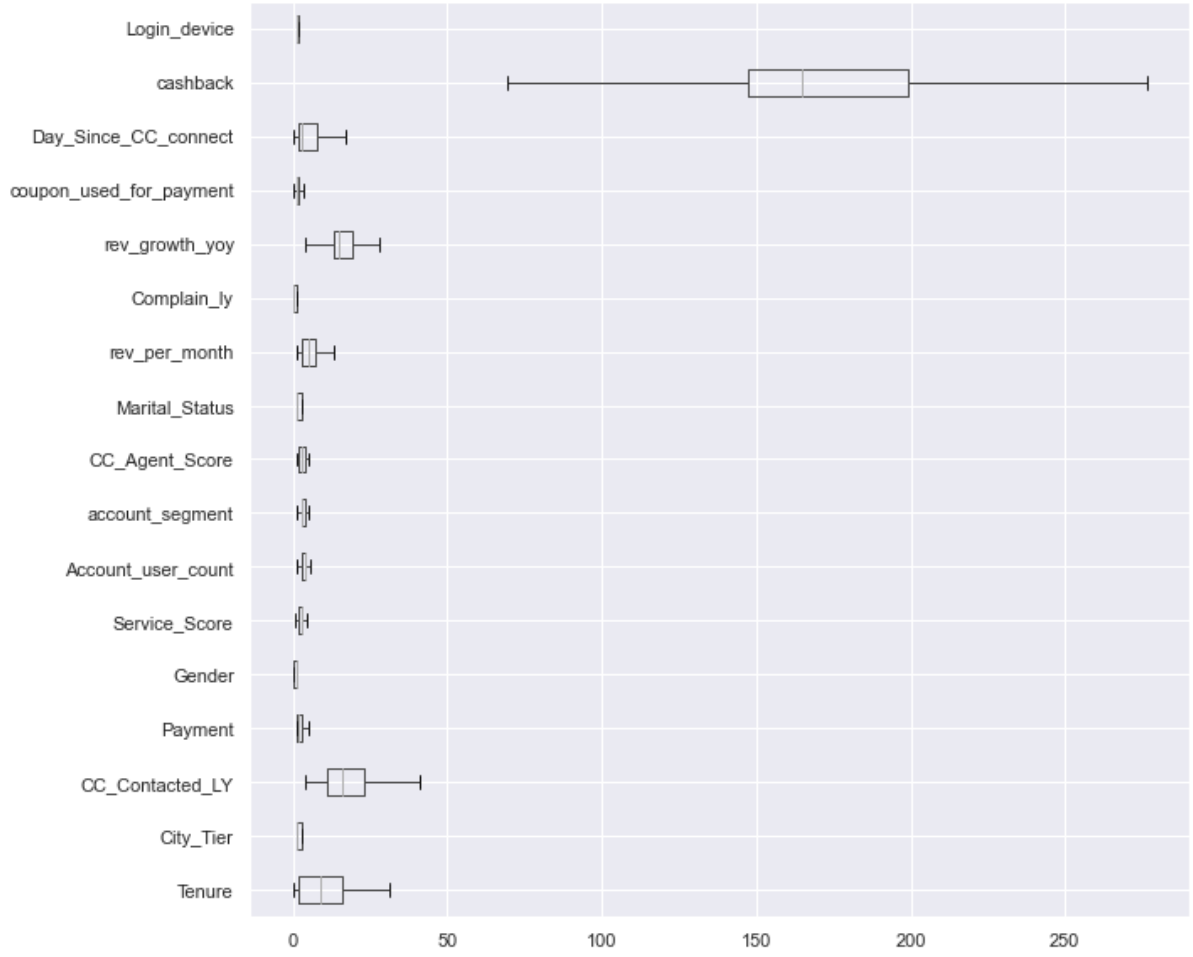


Figure 29: Data after removing outliers

The outliers are treated within the 1.5 x IQR range, and after removing, the box plot is shown as above.

3.4. Label Encoding

ML algorithms can only understand numerical data. It is important to convert the categorical columns to numerical form.

3.5. Scaling

For ML models that are weight based, it is important to scale them. Min Max scaling is used to scale the data.

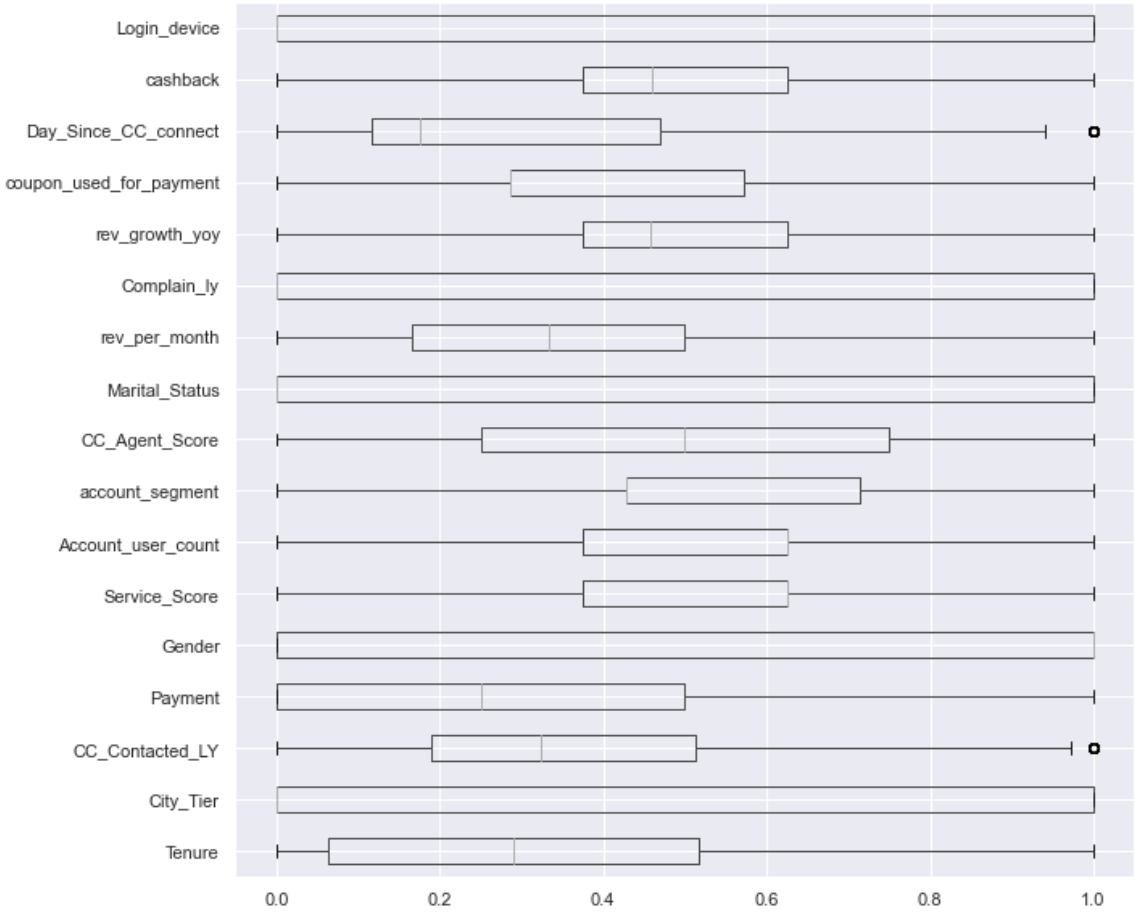


Figure 30: Data after scaling

3.6. SMOTE on Train Data

The target variable is highly unbalanced, which would provide biased results for the ML algorithms. Hence, they are sampled to produce equal distribution of target variable.

Distribution before and after applying SMOTE: -

	% Distribution
No	83.161634
Yes	16.838366

Table 17

	% Target Variable Distb
1	50.0
0	50.0

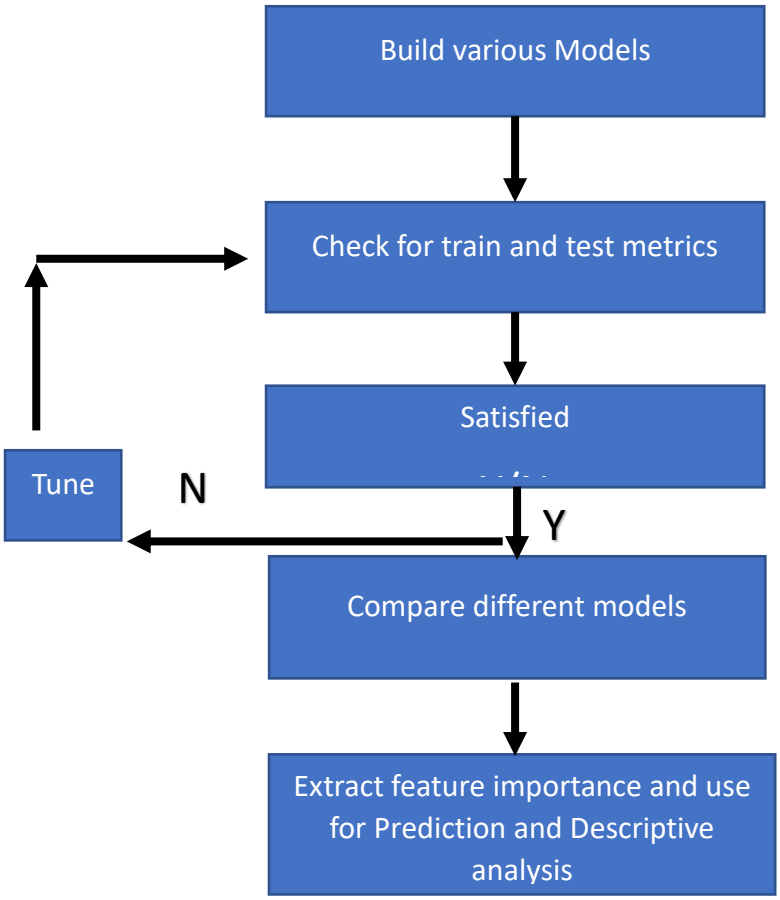
Table 16

3.7 Feature Engineering

There was no need to add/deduct any variable or transform any column. The column Account ID was removed, as it had no significance for the ML Models.

4. MODEL BUILDING

Model building had a very simple approach which is depicted below: -



Several models were built and compared. If the metrics scores were satisfactory, there was no need for further improvement. The results are presented in the table below: -

TRAINING DATA							
MODEL	ACCURACY	PRECISION		RECALL		F1 SCORE	
		0	1	0	1	0	1
LOGISTIC	81.39	0.83	0.8	0.79	0.83	0.81	0.82
LDA	80.76	0.84	0.78	0.77	0.85	0.8	0.82
CART	100	1	1	1	1	1	1
RANDOM FOREST	100	1	1	1	1	1	1
KNN	9787%	0.99	0.7	0.92	0.96	0.95	0.81
ANN	95.25	0.97	0.93	0.93	0.97	0.95	0.95
ADA BOOST	91.025	0.91	0.91	0.91	0.91	0.91	0.91
GRADIENT BOOST	93	0.93	0.94	0.94	0.93	0.93	0.93
BAGGING	100	1	1	1	1	1	1
VOTING CLASSIFIER	100	1	1	1	1	1	1

Table 18: Comparison of models on train data

TEST DATA							
MODEL	ACCURACY	PRECISION		RECALL		F1 SCORE	
		0	1	0	1	0	1
LOGISTIC	79.51	0.95	0.45	0.79	0.81	0.86	0.58
LDA	77	0.96	0.42	0.76	0.83	0.85	0.56
CART	92	0.96	0.75	0.94	0.81	0.95	0.78
RANDOM FOREST	96	0.97	0.91	0.98	0.88	0.98	0.9
KNN	92	0.9	0.7	0.92	0.96	0.95	0.81
ANN	91	0.97	0.68	0.92	0.89	0.94	0.77
ADA BOOST	88	0.94	0.62	0.91	0.73	0.92	0.67
GRADIENT BOOST	90	0.95	0.67	0.92	0.77	0.94	0.72
BAGGING	95	0.97	0.85	0.97	0.88	0.97	0.86
VOTING CLASSIFIER	96	0.97	0.89	0.98	0.88	0.98	0.88

Table 19: Comparison of models on test data

- Logistic and Linear discriminant models usually give decent results; hence they were tried.
- Random Forests are not affected by weight and outliers and overfit the training data, however, by tunning and optimization they can be modelled to give desired results.
- ANN, have several layers, which can give desired results with proper designation of no of layers.

- Ada Boost, Gradient Boost, Bagging and voting classifier are very good enfeebling techniques, that can provide required results.

This was the basic methodology to build the selected models.

(For further details you can have a look at project notes 2)

Model Improvement: RFCL had given the best results, and the results were in the required range, there was literally no scope for improvement. As per the flow chart mentioned, RFCL was selected as the best model.

5. MODEL VALIDATION

In this scenario, False negative or type II error plays a pivotal role. False negative should be as low as possible. A false negative would mean, a prediction where a customer is churning but the prediction says the negation, creating a false narrative for the company, in turn causing a huge loss.

False negative is incorporated in Recall.

$$Recall = \frac{True\ positive}{True\ positive + False\ Negative}$$

Lesser the false negative higher the Recall. Hence Recall is of utmost importance here.

Further we also need to check True positives are higher in number.

True positives are incorporated in both Precision and Recall.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

For a higher precision we require a high true positive but at the same time a low false positive. A high false positive could be dangerous (not as dangerous as false negative) as, the company would go on trying to retrieve a customer and spending resources, when in reality the customer is not going to churn

There is indeed a trade-off between false positives and false negatives, but in this business problem, false negative is more vital.

Although both metrics are important in their own sense, the model with the least trade-off is the most suitable for production.

Observing the models tested on test data, it comes out that, random forests is the best model with the required recall, precision and least trade-off between the metrics.

(NOTE: ROC and AUC Score, of all the models are well within limits and they are not so important to the business problem comparatively, hence their comparison is not presented here)

Best Model: RFCL

This works similar to Decision trees, only difference being this is a collection of several decision trees, and thus a black box algorithm.

Train and test accuracies: -

Train Data: 100%

Test Data: 96.48%

The algorithm overfits the train data however, the difference between train and test accuracies is not much and it proves, it does not overfits the test data

Classification reports: -

Classification Report for Train Data

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4825
1	1.00	1.00	1.00	4825
accuracy			1.00	9650
macro avg	1.00	1.00	1.00	9650
weighted avg	1.00	1.00	1.00	9650

Figure 31: Classification report Train data

Classfication Report for Test Data

	precision	recall	f1-score	support
0	0.97	0.98	0.98	2048
1	0.91	0.88	0.90	427
accuracy			0.96	2475
macro avg	0.94	0.93	0.94	2475
weighted avg	0.96	0.96	0.96	2475

Figure 32: Classification report test data

As can be observed, all the metrics on the test data show extremely well performances and seems they can be used for production.

ROC and AUC Curves: -

```
ROC AUC SCORE 1.0
```

Figure 33: ROC Curve train data

```
ROC AUC SCORE 0.9866025687939111
```

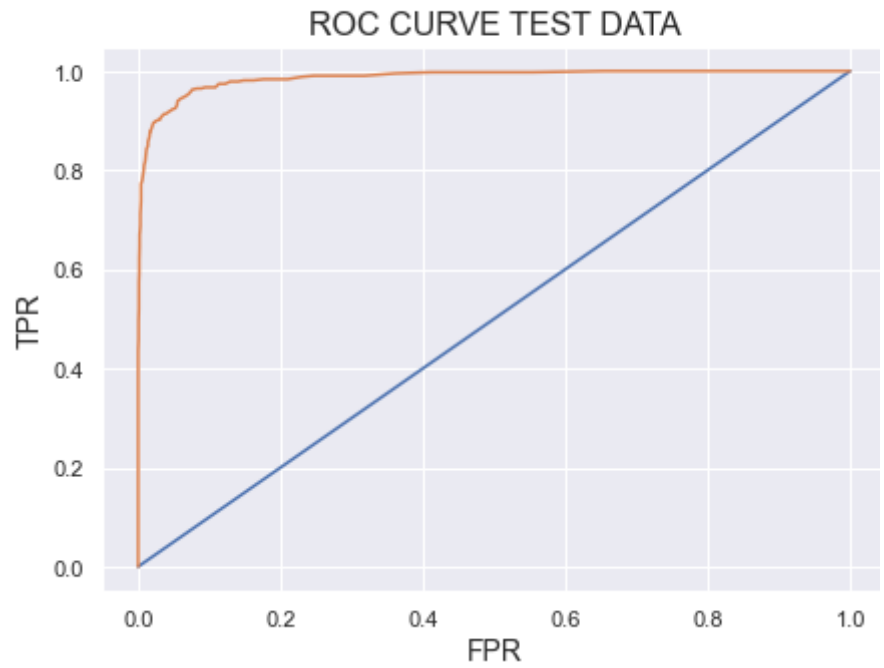


Figure 34

Figure 35 ROC Curve Test Data

6. INSIGHTS AND RECOMMENDATIONS

INSIGHTS: -

1. **Last customer care contacted**: More the frequency more the churn and vice versa.

2**. Service Score**: Mixed Insights. Customers with high scores have churned and low scores have stayed and vice versa

3. **Customer care scor**e: Those churning have given higher median and mean scores to consumer care.

4. **Complaints**: 46% of those who complaint has churned.

5. **Cashback**: Those who stayed received more cashback than those who churned.

6**. City Tier**: Max customers churned from Tier 2 city: 25% churned.

7. **Payment:** 33% of the customers who used COD churned (Highest of all payment methods)

8. **Account Type**: 33% of those who use regular plus subscriptions have churned.

9. **Log in Device**: Those who use computers for log-in have churned more: 24%.

RECOMMENDATIONS: -

1. A thorough market survey: The company must conduct a market survey and judge its market competition. Even the customers who gave high rating, churned at a high rate. If they are satisfied with the service and customer care, then why they churned at first place? There might be other companies, offering the same services at a cheap rate, or better services at the same rate, Company needs to evaluate this and change its pricing based on market competition.

2. Customer survey: Finding the root cause for customer churn is really very important. Customer care should contact a random lot of churned customers and try to understand why they churned. This would help find the root cause of customer churn and thus company could try to eliminate it.

3. Reward system for customer care executives: The customer care executives must be given a rating on every issue they resolve and based on it they should be offered a reward. For instance, for every rating above 4.5, the executive could be offered rewards such as free amazon coupons and awards such as “best employee”.

4. Referral program: The existing customers can refer the service to others and if they join, both of them can be given 10% discount on their plan.

5. As we observed, tier 2 customers churned the most. These customers should be given an additional offer of 15% discount on the high-end plans. This would encourage them to join the DTH service more often than before.

6. The customer using regular plus subscription churned the most. These customers must be offered a few services of the next higher plan for free. For e.g., if these customers could not avail Netflix at this subscription, they can be offered Netflix subscription (a basic std definition plan) for free.