Telecom Churn Case Study

By Aanchal Pahwa

Problem Statement

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition.

For many incumbent operators, *retaining high profitable customers is the number one business goal*.

To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn.**

In this project, you will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Understanding and defining churn

There are two main models of payment in the telecom industry - **postpaid** (customers pay a monthly/annual bill after using the services) and **prepaid** (customers pay/recharge with a certain amount in advance and then use the services).

In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.

However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).

Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term 'churn' should be defined carefully. Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe in North America.

This project is based on the Indian and Southeast Asian market.

Understanding Data

Understanding the business objective and the data.

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.

The **business objective** is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behaviour during churn will be helpful.

Steps to be followed

- Data Understanding
- Data Cleaning
- Filtering High Value Customers
- Creating target Variable
- Deriving New Features
- Handling Missing values
- Data Visualization-Univariate Analysis
- Data Visualization- Bivariate Analysis
- Outlier Detection
- Data Preparation
- Data Modeling and Eavlaution
  - Non-Interpretable Models
  - Interpretable Models
- Conclusion

METHODOLOGIES

Model Validation:

-Validate the model to assess its performance and reliability. This may involve techniques such as cross-validation, splitting data into training and testing sets, and evaluating metrics like accuracy, precision, recall, and F1 score.

ASSUMPTIONS

In this case study, columns with missing value percentages exceeding 40% were removed. Imputing such extensive missing data could negatively affect the analysis.

-'SELECT' values in certain columns were replaced with null..

-For numerical columns where the mean and median were identical but the maximum value was far outside the range, null values were imputed with 1.5 times the Interquartile Range (IQR) of the variable.

-Unnecessary columns with heavily skewed data were dropped to prevent adverse impacts on overall model building.

Insights

Average revenue of call per user

We observe that the minimum values in arpu_6, arpu_7, arpu_8 & arpu_9 are negative. This indicates that some customers are making loss to the company. We will retain them in our study as our criteria of high value customer is based on `usage base churn` and not on `revenue based churn`. Removing them might lead to loss of some meaningful insights. Let us observe their importance in the exploratory data aanalysis part and then take a call.

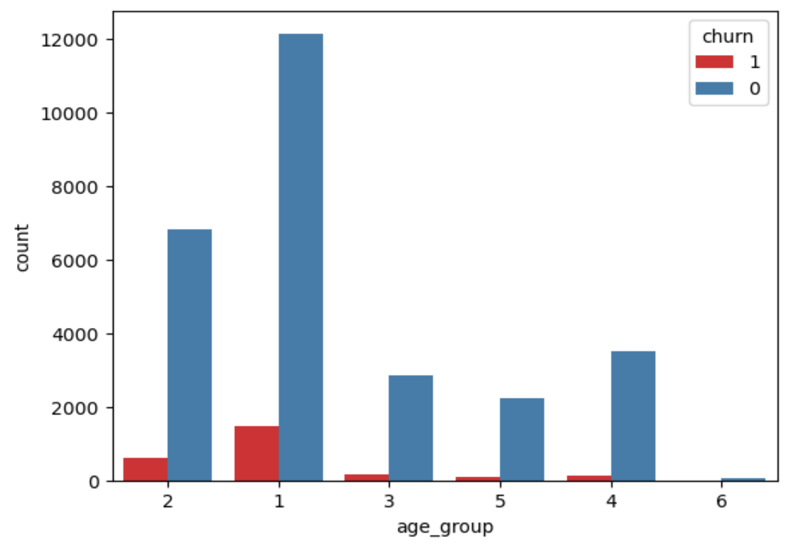
Sachet recharge

`Sachet recharge` are Service schemes with validity smaller than a month. This means that the days of recahrge in sachet recharge should be less than 30. Any service schemes beyond 29 days should mean that the client has done a monyhly recharge or the entry is wrong. Let us `cap the values beyond 29 days` by the highest number of days recharge below 30 days.

Univariate Analysis

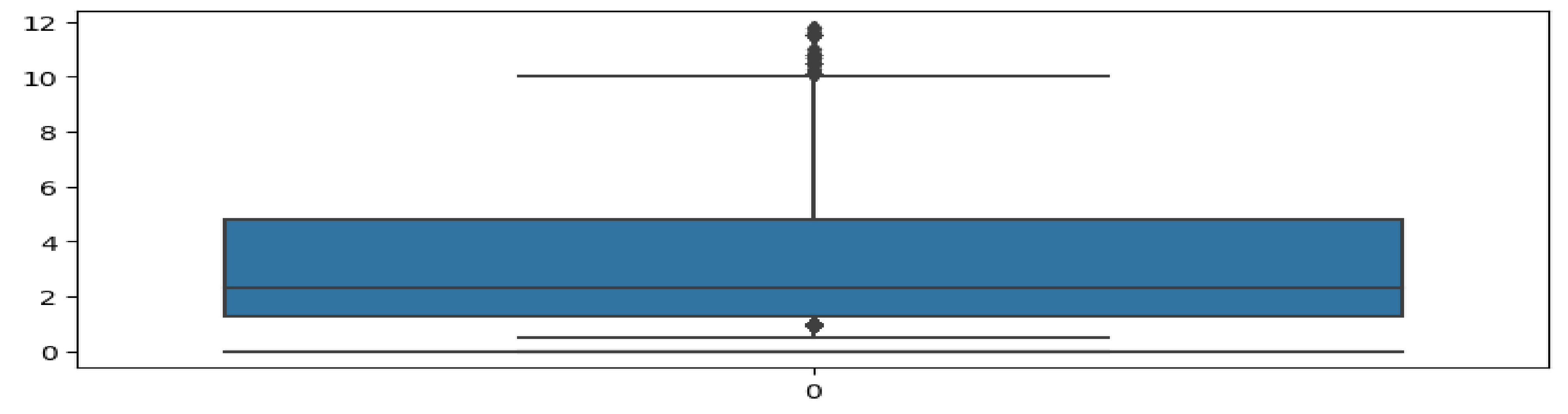
Data Distribution

Let us check the churn data according to the different age group.



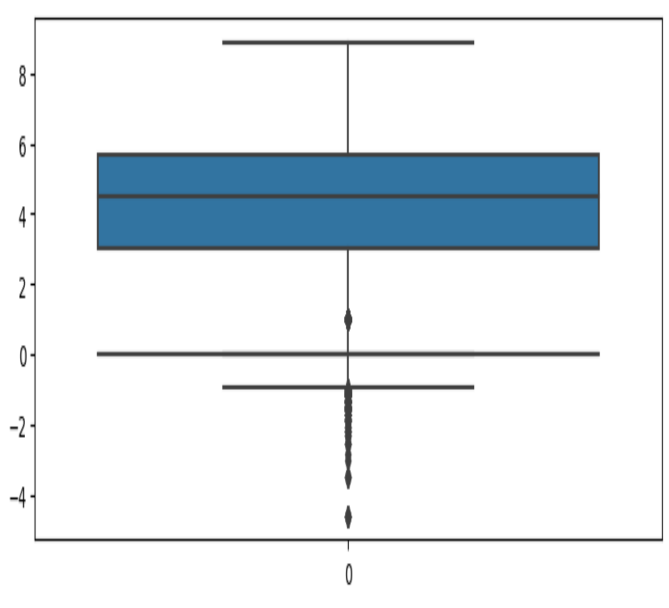
Bivariate Analysis

a.) Analyzing age of network with churn

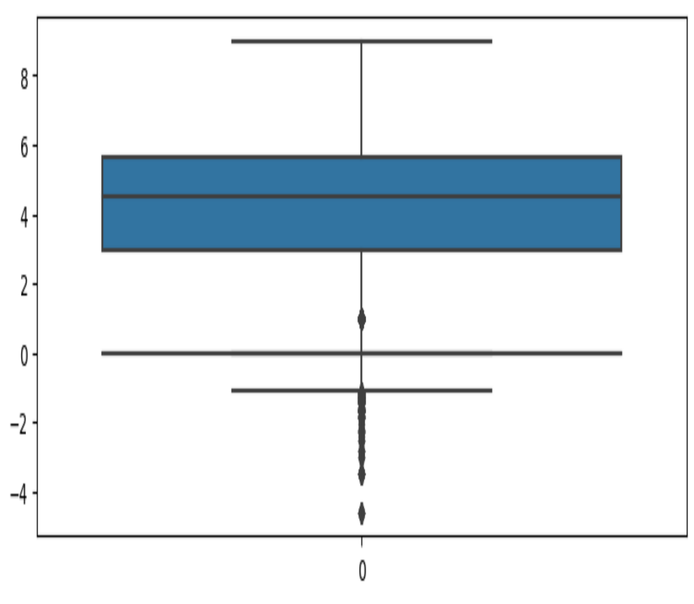


Average number of days customer using the network before churning is less than 1000.

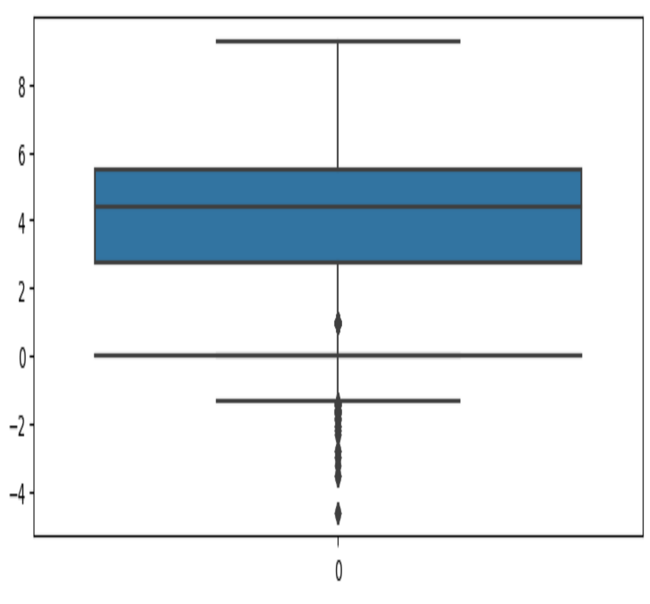
b.) Analyzing onnet with churn



onnet_mou_6



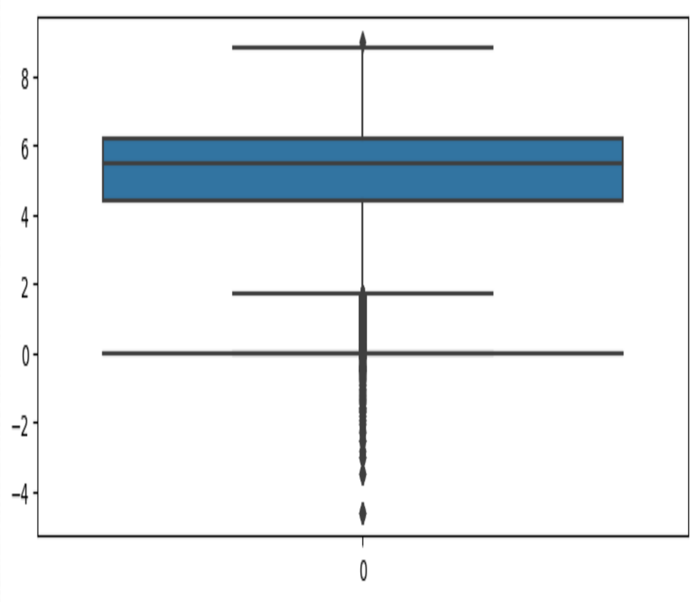
onnet_mou_7



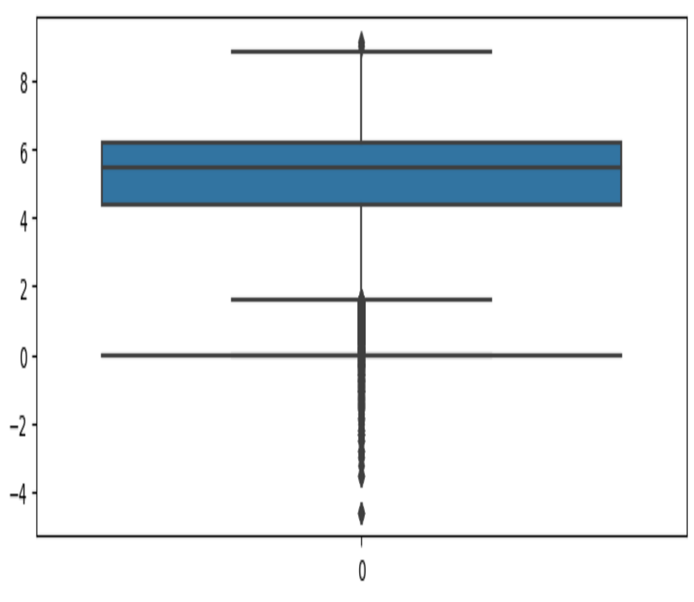
onnet_mou_8

c.) Analyzing offnet with churn

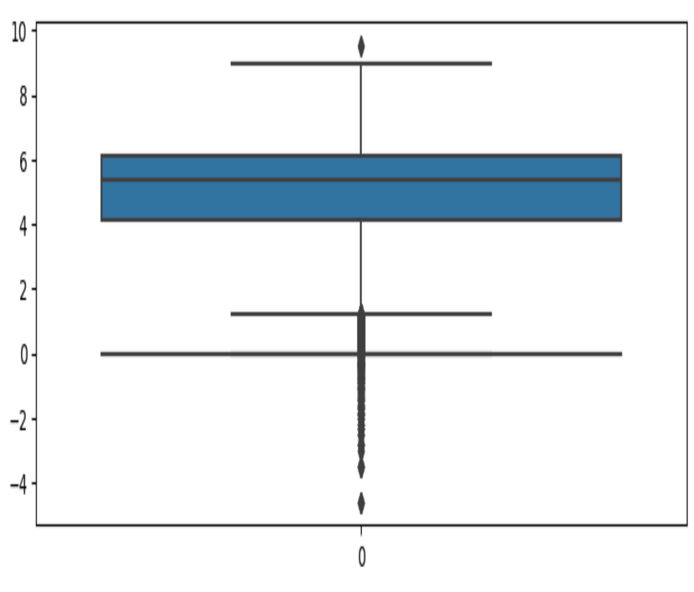
People who have churned, used most of the calling in month of June and July on different network, where as in month of August there calling is reduced and it is indicating that the customers are likely to churn



offnet_mou_6



offnet_mou_7



offnet_mou_8

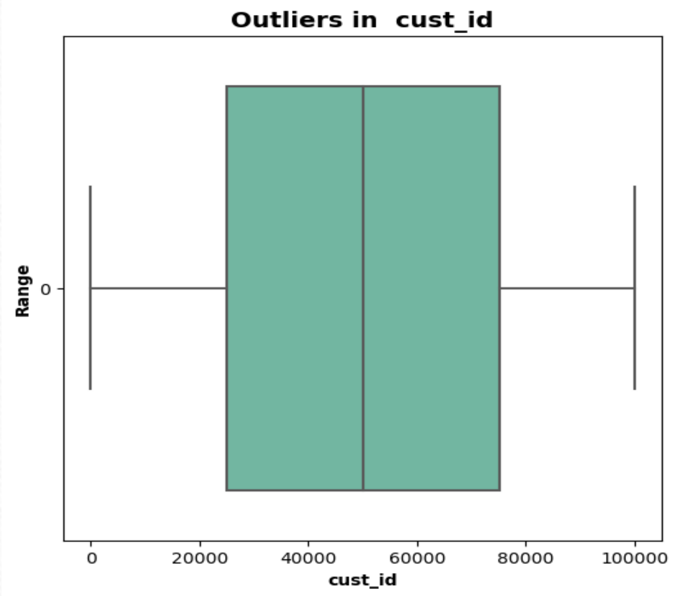
d.) Analyzing night pack user with churn

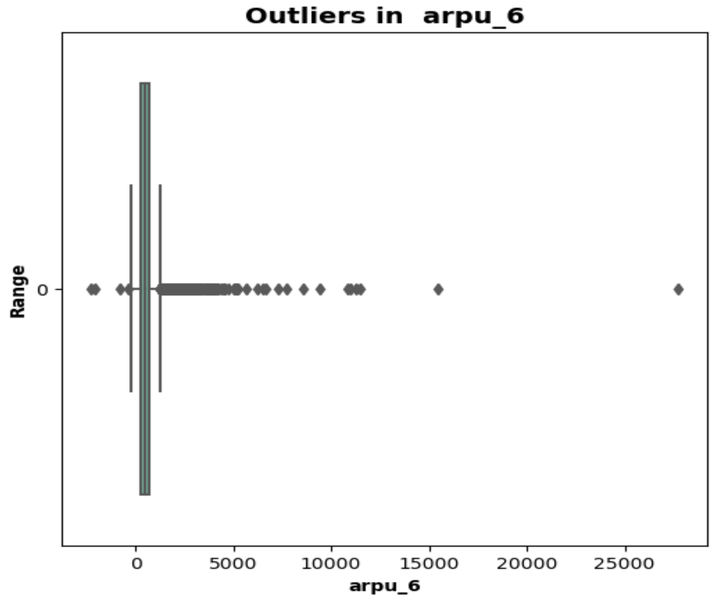
night_pck_user_8	-1.0	0.0	1.0
churn			
0	85.89123	97.117602	97.360704
1	14.10877	2.882398	2.639296

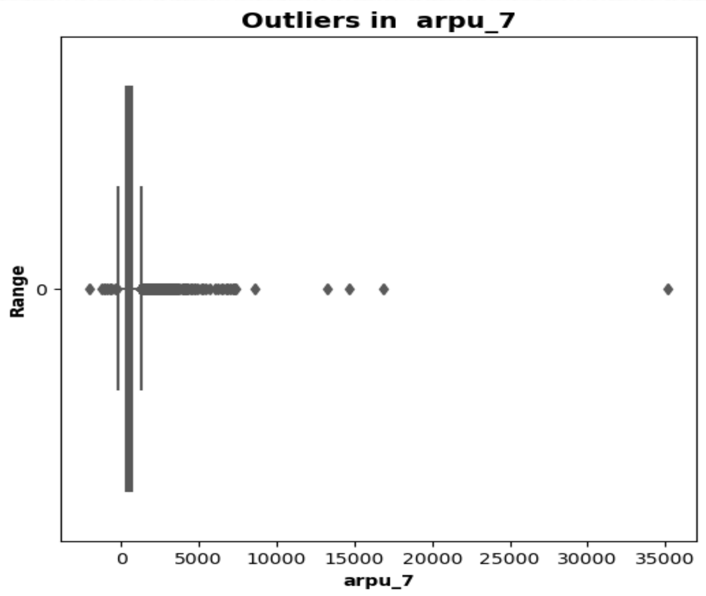
e.) Analyzing Sachet with churn

sachet_3g_8	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	23	25	27	29
churn																										
0	24972	1609	399	184	106	86	43	35	28	19	15	8	11	10	6	6	2	2	3	1	3	3	2	1	1	5
1	2369	48	5	8	4	2	1	0	2	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0

Outlier Analysis

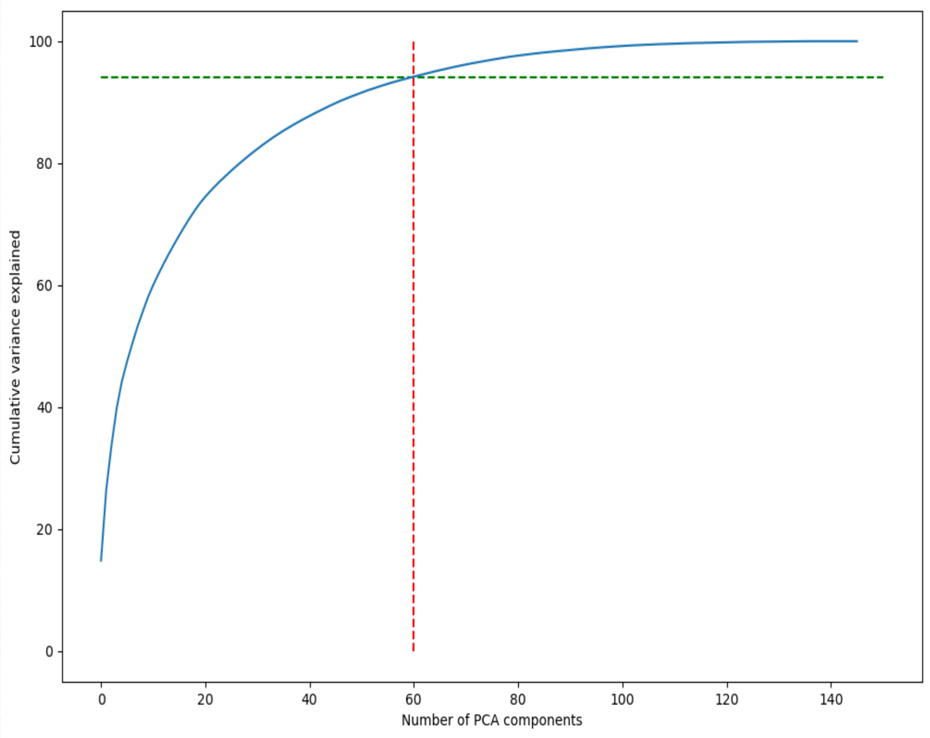






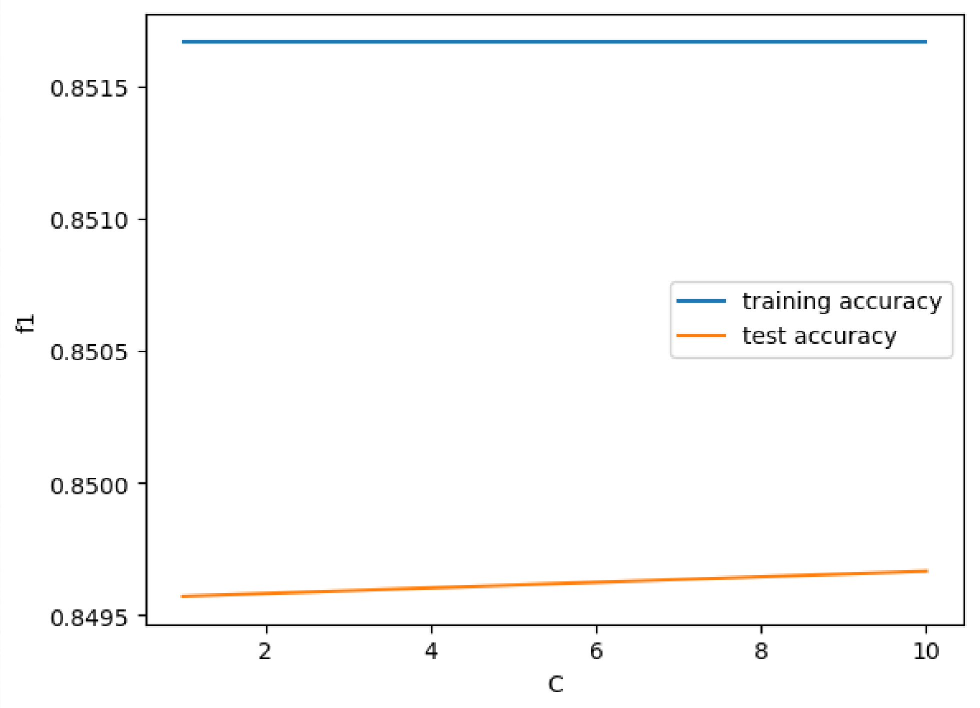
Data Modeling and Model Evaluation

- In the exploratory data analaysis we saw that some columns have significant correlation among themselves. This collinearity can hamper our interpretations. Principal component analysis (PCA) is one of the most commonly used dimensionality reduction techniques in the industry. So let us adopt PCA to solve this problem. PCA will also help in dimensionality reduction.

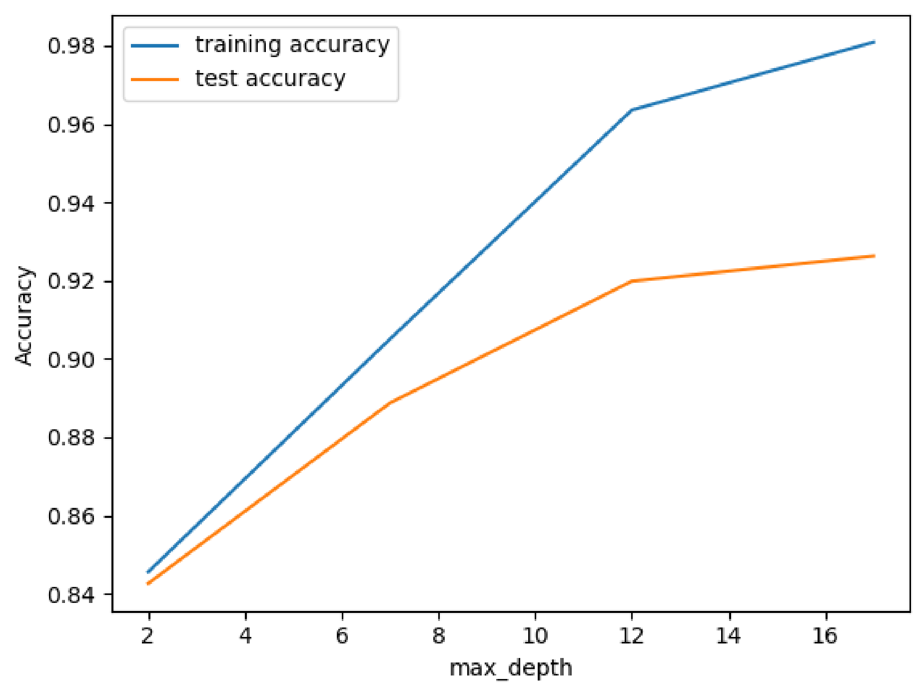


Model Building

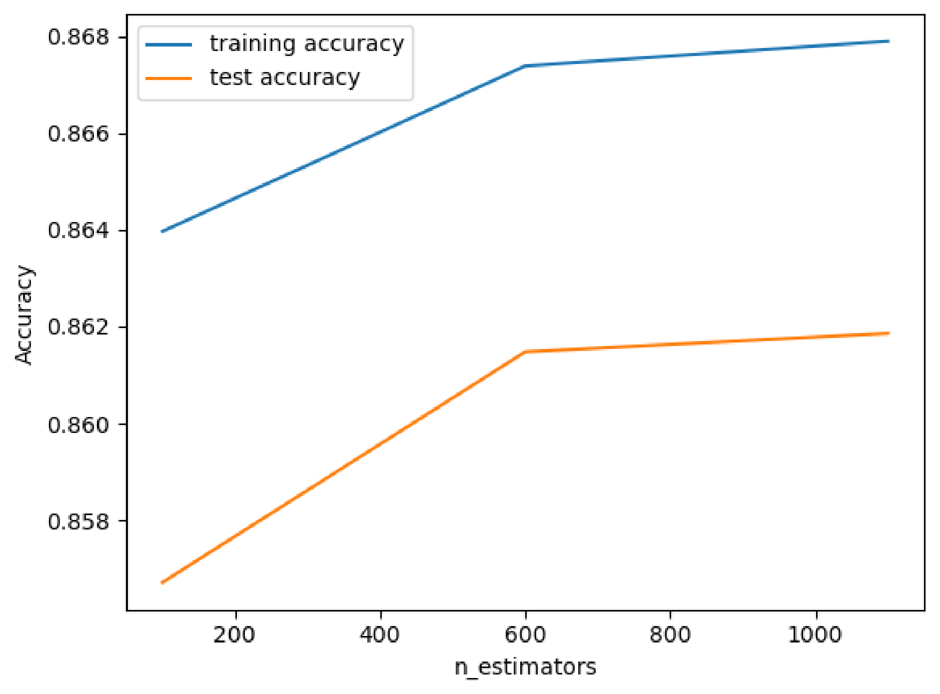
- Let us now build various models on this PCA Transformed dataset to predict churn.
- **Model1- PCA and Logistic Regression**
- **Hyper Parameter tuning - Logistic Regression**
- **Building and Evaluating the Final Model**
- **Model3 PCA and Random Forest**
- **Hyperparameter Tuning**

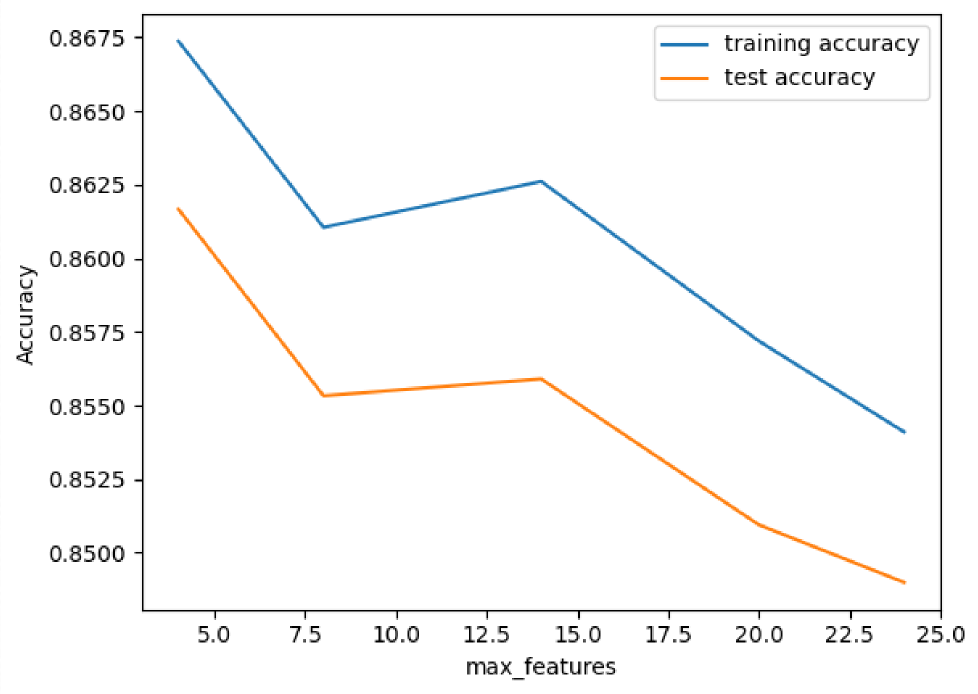


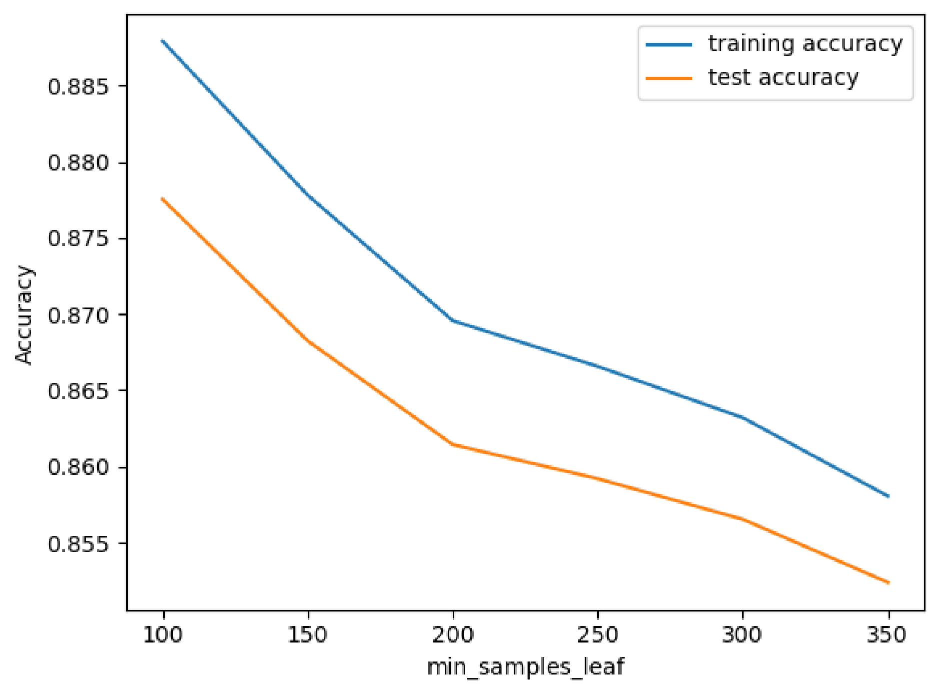
0.8515
0.8510
0.8505
0.8500
0.8495
f1
2
4
6
8
10
C
training accuracy
test accuracy



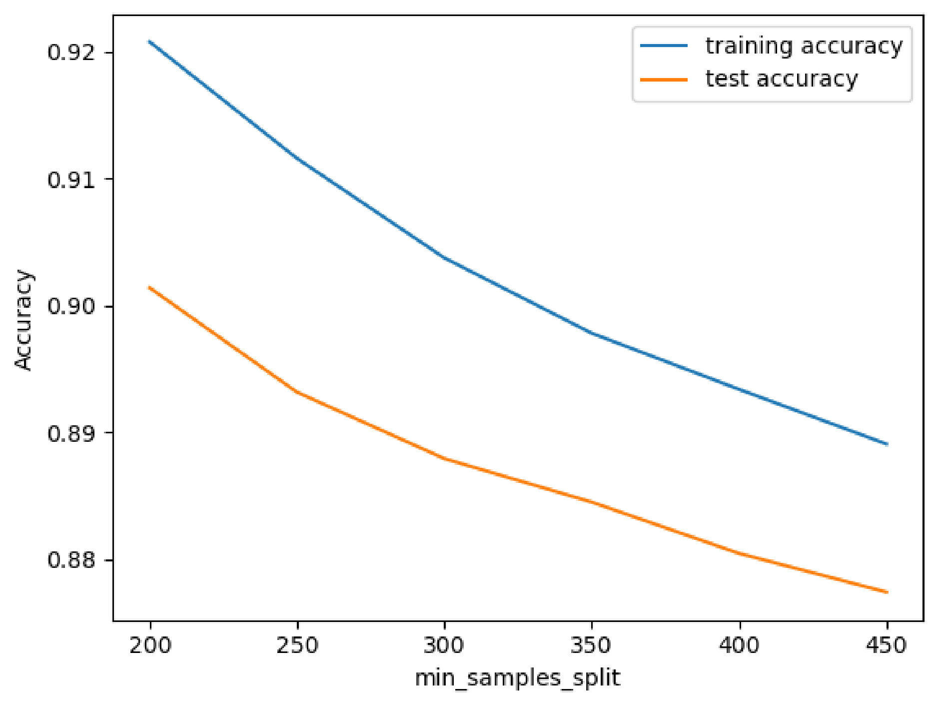
training accuracy
test accuracy
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
Accuracy
2
4
6
8
10
12
14
16
max_depth


training accuracy
test accuracy
0.868
0.866
0.864
0.862
0.860
0.858
Accuracy
200
400
600
800
1000
n_estimators


training accuracy
test accuracy
0.8675
0.8650
0.8625
0.8600
0.8575
0.8550
0.8525
0.8500
Accuracy
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
max_features



training accuracy
test accuracy
Accuracy
0.885
0.880
0.875
0.870
0.865
0.860
0.855
100
150
200
250
300
350
min_samples_leaf



training accuracy
test accuracy
Accuracy
0.92
0.91
0.90
0.89
0.88
200
250
300
350
400
450
min_samples_split

Final Choice of Model

- Recall is the most important business metric for the telecom churn problem. The company would like to identify most customers at risk of churning, even if there are many customers that are misclassified as churn. The cost to the company of churning is much higher than having a few false positives.

Model/Metricity

a) Logistic Regression

- Sensitivity/Recall - .83
- Specificity - .83
- Roc AUC Score - .90

b) SMV

- Sensitivity/Recall - .82
- Specificity - .85
- Roc AUC Score - .90

c) Random Forest

- Sensitivity/Recall - .73
- Specificity - .90
- Roc AUC Score - .90

Strategies to be Incorporated

1. In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate.

2. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.

3. In order to manage High Value Customer Churn, we have predicted customers that are more likely to churn and the factors that influence the high churn.

4. From the exploratory analysis, we observed that there is considerable drop in recharge, call usage and data usage in the 8th month which is the `Action Phase`. From the list of important predictors affecting the churn, this is again evident as follows:

- 'arpu_7', ‘max_rech_amt_6', 'std_og_t2m_mou_8', 'loc_og_t2m_mou_8', 'max_rech_data_8',
- 'last_day_rch_amt_8', 'total_data_rech_8’, 'total_amt_8', 'roam_og_mou_8', 'loc_ic_t2m_mou_8'

Strategies to be Incorporated

5. Average revenue per user in the `7th month` plays a vital role in deciding churn. A sudden drop in it might indicate that the customer might be thinking about churning and appropriate actions should be taken.

6. Local & STD Minutes of usage (incoming & outgoing) are the most affecting features on the customer churn.

7. Lastday of recharge amount in the action phase also plays a crucial role in determing churn.

8. The maximum rcharge for calling data by a client in the 6th Month and 8th Month should be carefully focussed as the 6th month indicates the begining of the good phase and 8th month indicates the action phase.

9. Last day of recharge in the 8th month, the total recharge for data done in the 8th month and the total amount spent on calls and data by clients in 8th month also play crucial role in indicating churn.

10. Outgoing roaming calls made by clients in the 8th month also play key role in indicating churn.

Strategies to be Incorporated

Following strategies can be incorporated :

- Sudden drop in Local & STD Minutes of usage might be because of the unsatisfactory customer service or because of poor network or unsuitable customer schemes/plans. Efforts shall be made to provide better network and focus on customer satisfaction.

- Customized plans should be provided to such customers to stop them from churning.

- Based on the usage / last recharge, routine feedback calls to understand the customer satisfaction regarding services can be made to understand their grievances & expectations. Appropriate action should be taken to avoid them rom churning.

- Various attractive offers can be introduced to customers showing sudden drop in total amount spent on calls & data recharge in the action phase to lure them.

- Promotional offers can also be very helpful .