TELECOM CHURN GROUPCASE STUDY

PRESENTED BY

TOUSEEF ASHRAFI

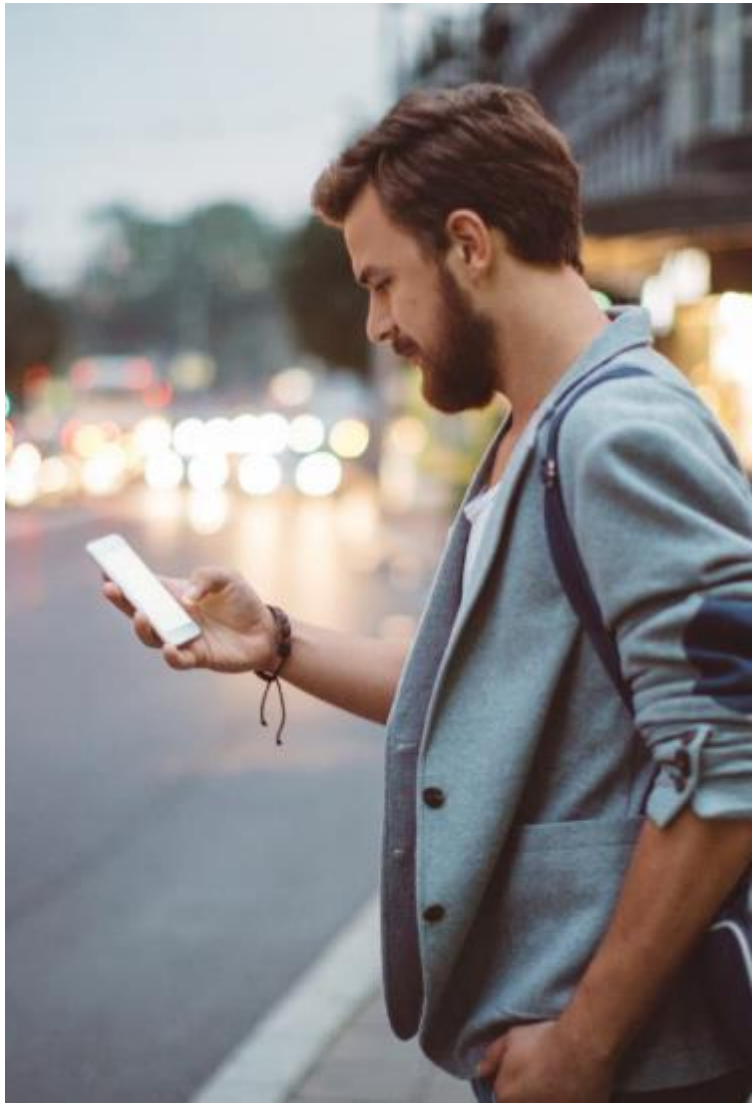
SHAIK GHOUSE MOIN UDDIN

ERWIN JAMES

GROUP CASE STUDY

TOPICS

- Business problem overview
- Derive new features & Analysis
- Filter High Value Customer
- Tag Churners
- DATA Cleaning & EDA
- Recharge amount related Variables
- Plots for Variables & Analysis
- Heat Map & Analysis
- Principle component Analysis
- Applying Logistic Regression & Random Forest
- Parameter Tunning
- Conclusion - Business Insight



OBJECTIVE

Analyze 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.

Business Problem Overview

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.

Derive new features

We can derive more meaningful information

- Total recharge amount
- Total recharge for data
- Maximum recharge amount
- Last date of Recharging the data
- Average recharge amount for data

Analysis:

- We can create new feature as total_rech_amt_data using total_rech_data and av_rech_amt_data to capture amount utilized by customer for data.
- As the minimum value is 1 we can impute the NA values by 0, Considering there were no recharges done by the customer.

```
telecom.loc[:,amt_recharge_columns].describe()
```

	total_rech_amt_6	total_rech_amt_7	total_rech_amt_8	total_rech_amt_9	max_rech_amt_6	max_rech_amt_7	max_rech_amt_8	max_rech_amt_9	total_
count	99999.000000	99999.000000	99999.000000	99999.000000	99999.000000	99999.000000	99999.000000	99999.000000	2
mean	327.514615	322.962970	324.157122	303.345673	104.637486	104.752398	107.728207	101.943889	
std	398.019701	408.114237	416.540455	404.588583	120.614894	124.523970	126.902505	125.375109	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	109.000000	100.000000	90.000000	52.000000	30.000000	30.000000	30.000000	28.000000	
50%	230.000000	220.000000	225.000000	200.000000	110.000000	110.000000	98.000000	61.000000	
75%	437.500000	428.000000	434.500000	415.000000	120.000000	128.000000	144.000000	144.000000	
max	35190.000000	40335.000000	45320.000000	37235.000000	4010.000000	4010.000000	4449.000000	3399.000000	

Filter high-value customers

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	last_d
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
7	7000701601	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
8	7001524846	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
21	7002124215	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
23	7000887461	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	

High-value customers-

Those who have recharged with an amount more than or equal to X, where X is greater than 70th percentile of the average recharge amount in the first two months (the good phase)

Tag churners and remove attributes of the churn phase

- Tag churners and remove attributes of the churn phase
- Now tag the churned customers (churn=1, else 0) based on the fourth month as follows:
- Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase.
- The attributes to use to tag churners are:
  - •total_ic_mou_9
  - •total_og_mou_
  - •vol_2g_mb_9
  - •vol_3g_mb_9

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	last_d
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
7	7000701601	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
8	7001524846	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
21	7002124215	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
23	7000887461	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	arpu
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	197.3
7	7000701601	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	1069.1
8	7001524846	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	378.7
21	7002124215	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	514.4
23	7000887461	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	74.3

After tagging churners, remove all the attributes corresponding to the churn phase (all attributes having ' _9', etc. in their names)

Data with only 1 unique Value

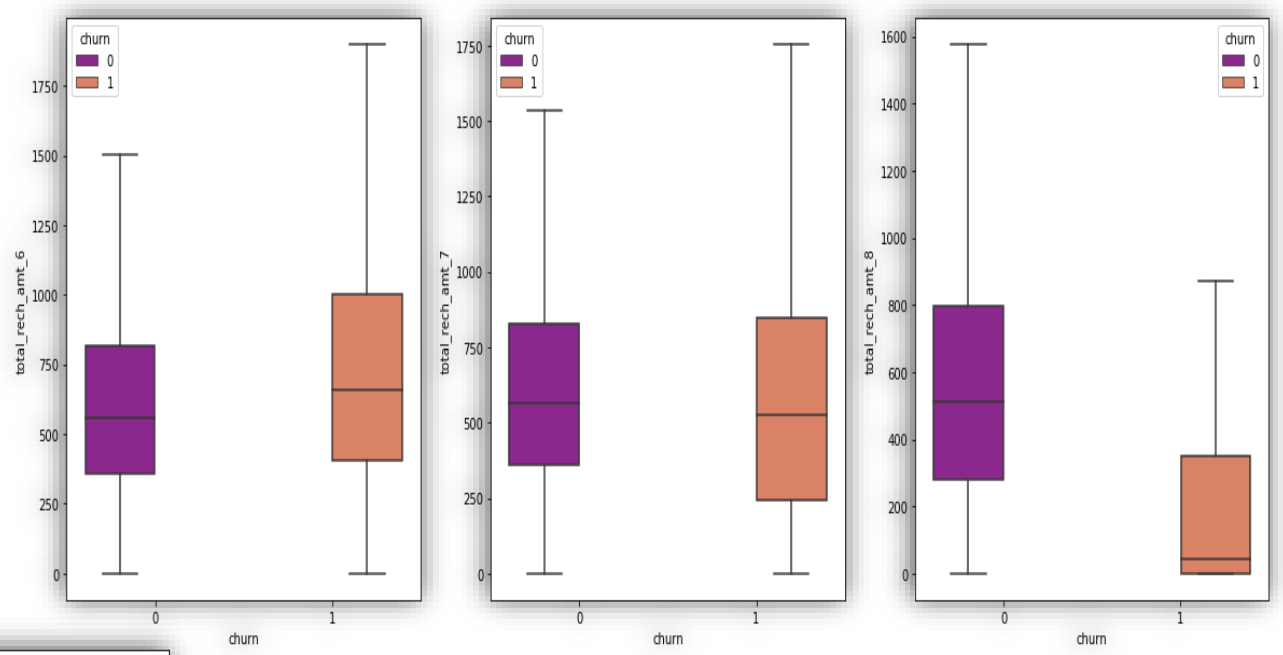
	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	std_og_t2c_mou_6	stc
0	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	NaN	
7	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	
8	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	
21	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	
23	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	

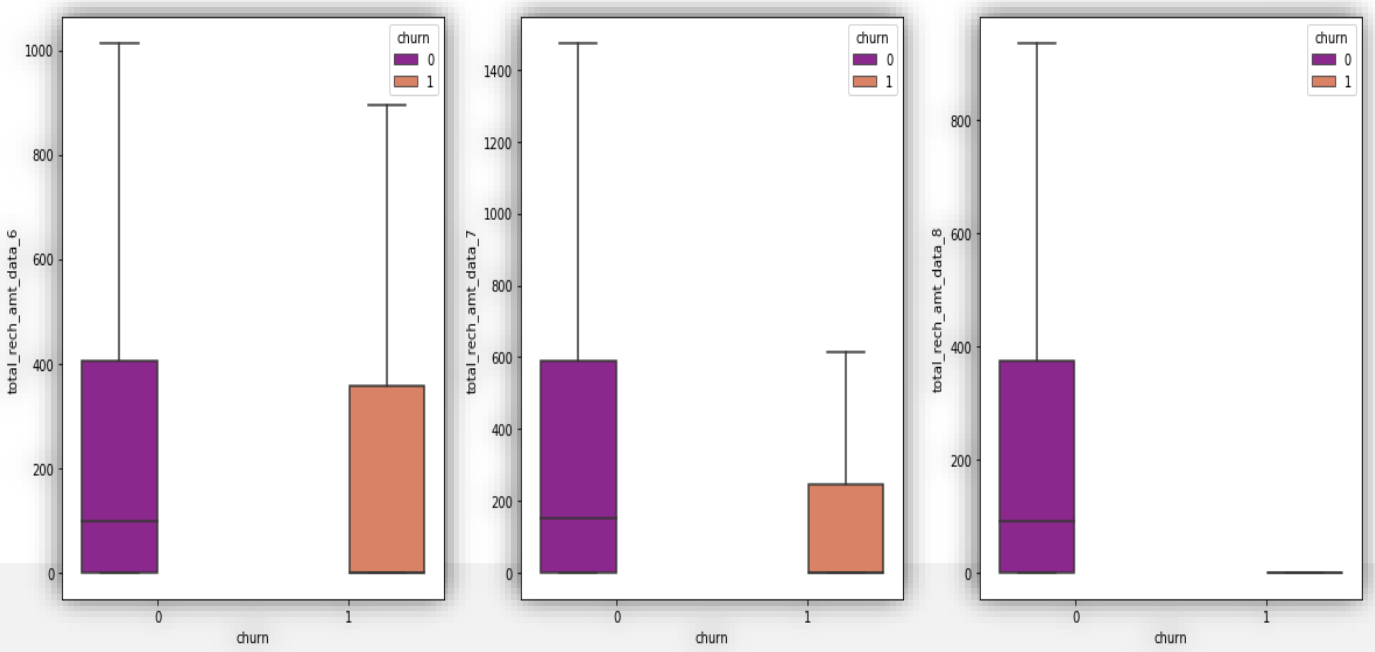
Analysis: Dropping above features with only **one unique** value as they will not add any value to our model building and analysis

Recharge amount related variables

Plotting for total recharge amount:

Analysis: We can see a drop in the total recharge amount for churned customers in the 8th Month (Action Phase)





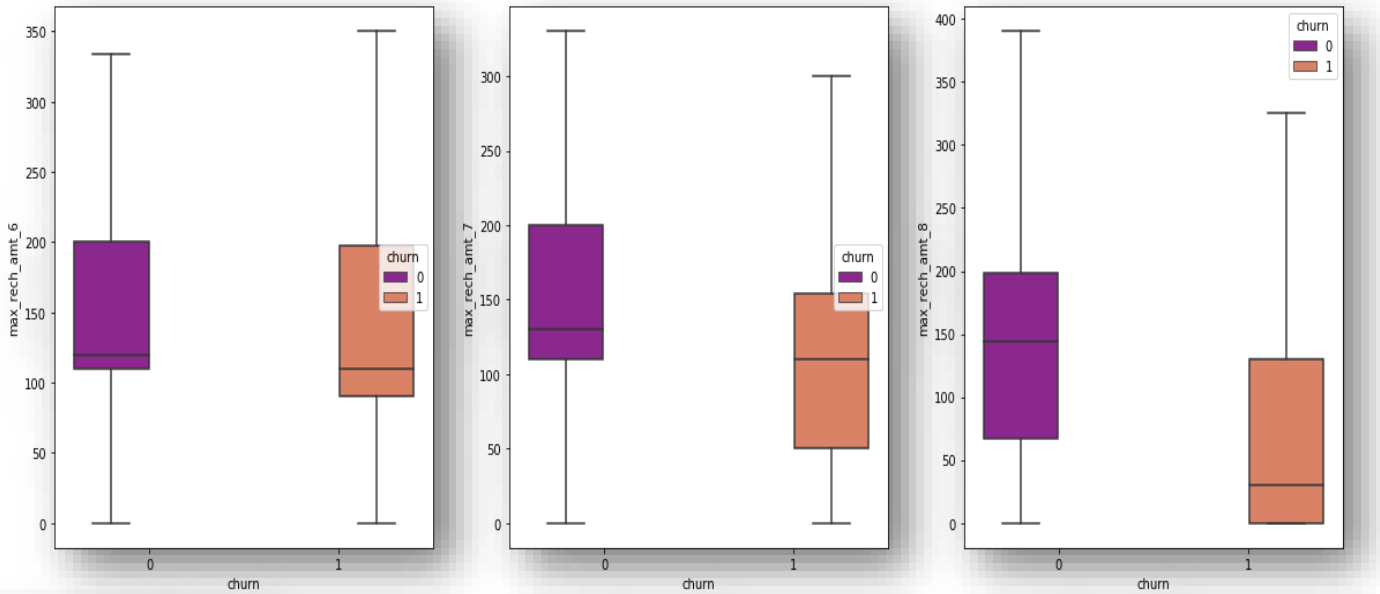
Plotting for total recharge amount for data:

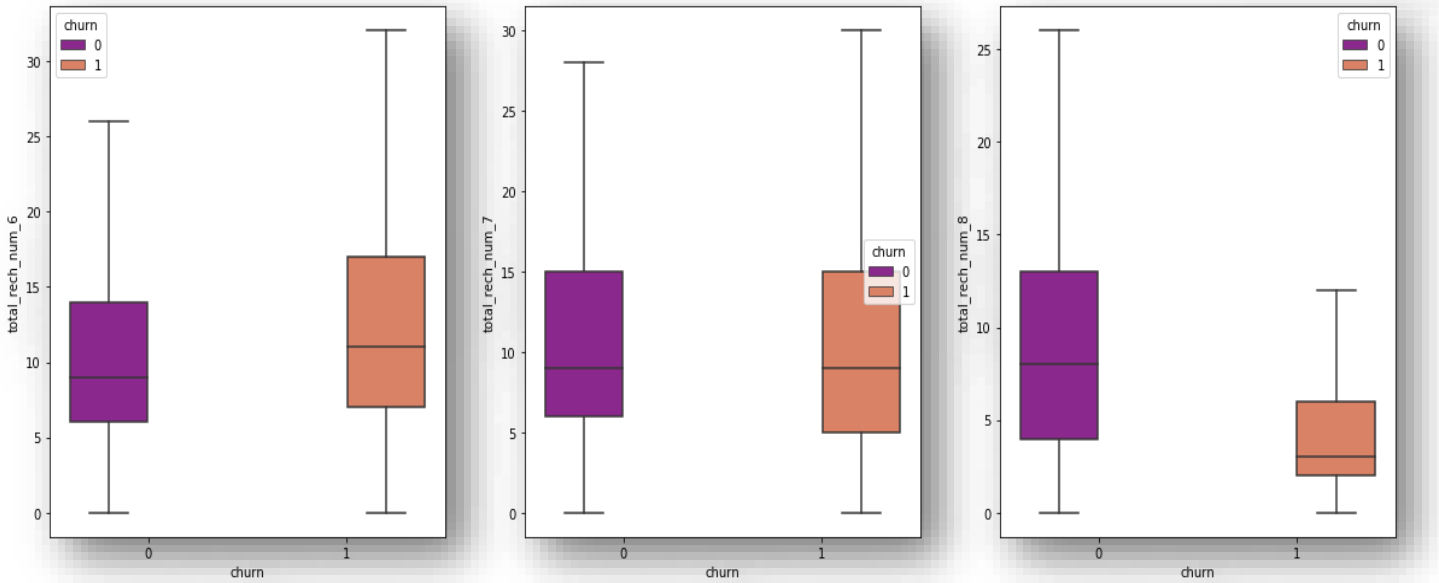
Analysis: We can see that there is a huge drop in total recharge amount for data in the 8th month (action phase) for churned customers.

Recharge amount related variables

Plotting for maximum recharge amount for data:

Analysis: We can see that there is a huge drop in maximum recharge amount for data in the 8th month (action phase) for churned customers.





Plotting for Total recharge for Number:

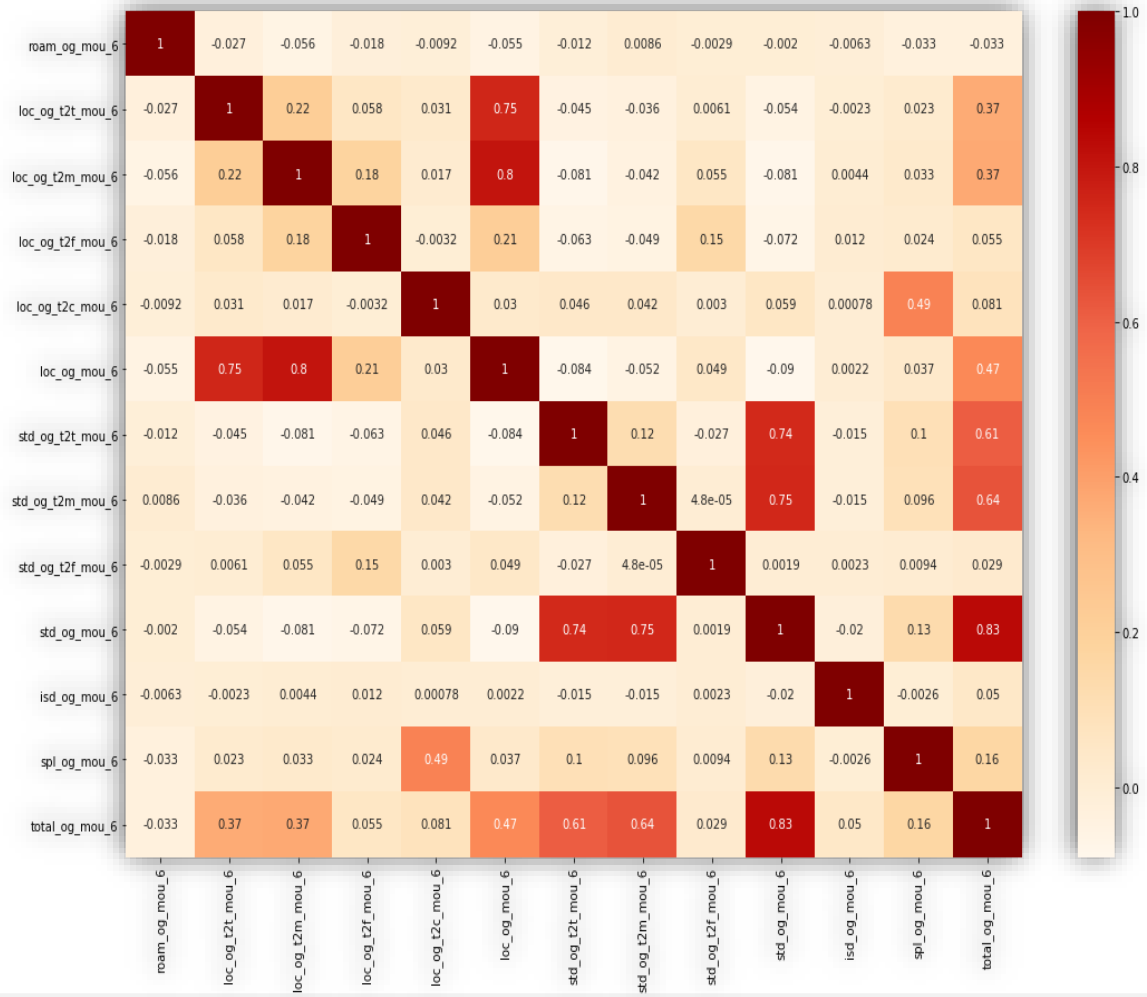
Analysis: We can see that there is a huge drop in total recharge number in the 8th month (action phase) for churned customers.

PLOTTING HEAT MAP

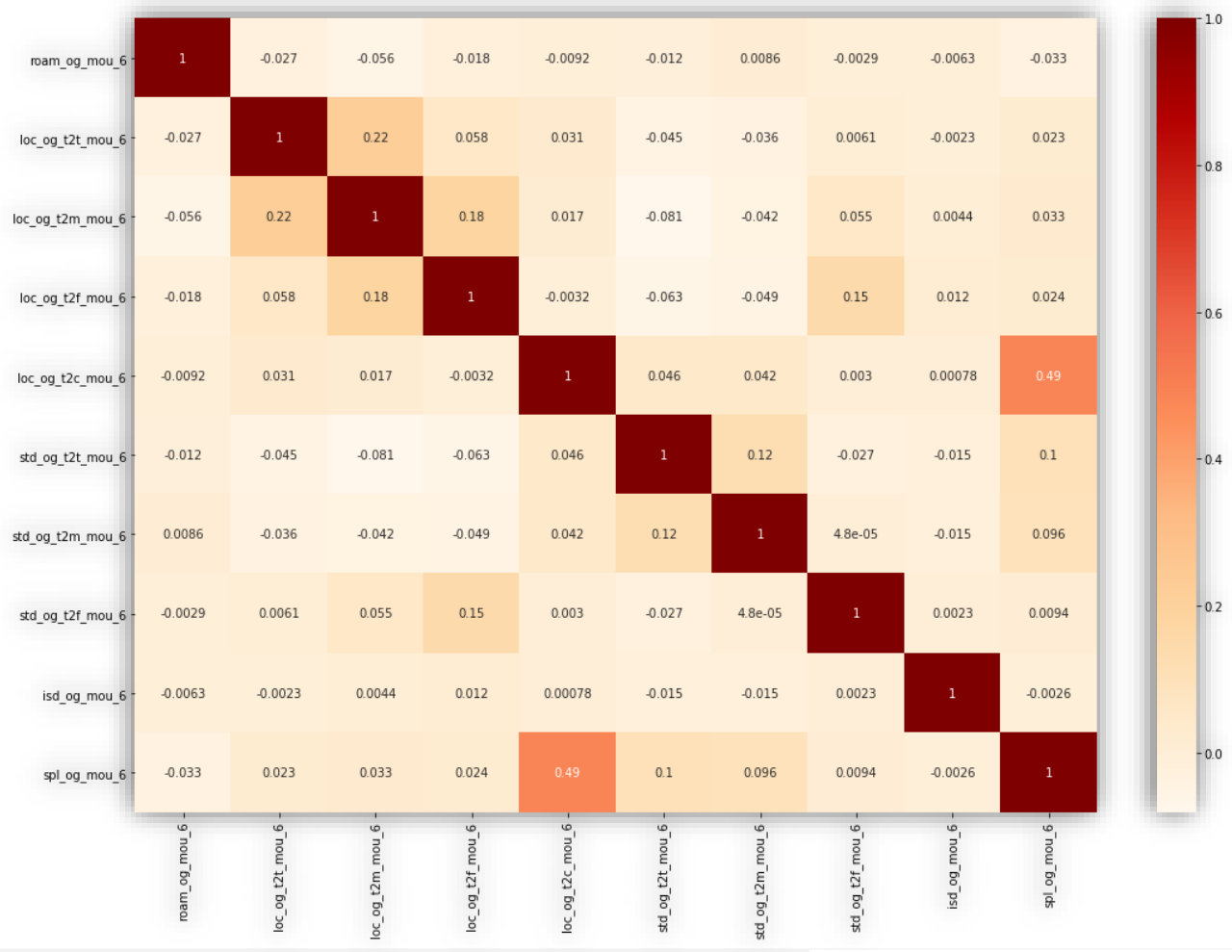
Checking for outgoing & incoming MOU Variables

PLOTTING HEAT MAP

Checking for Outgoing MOU variables



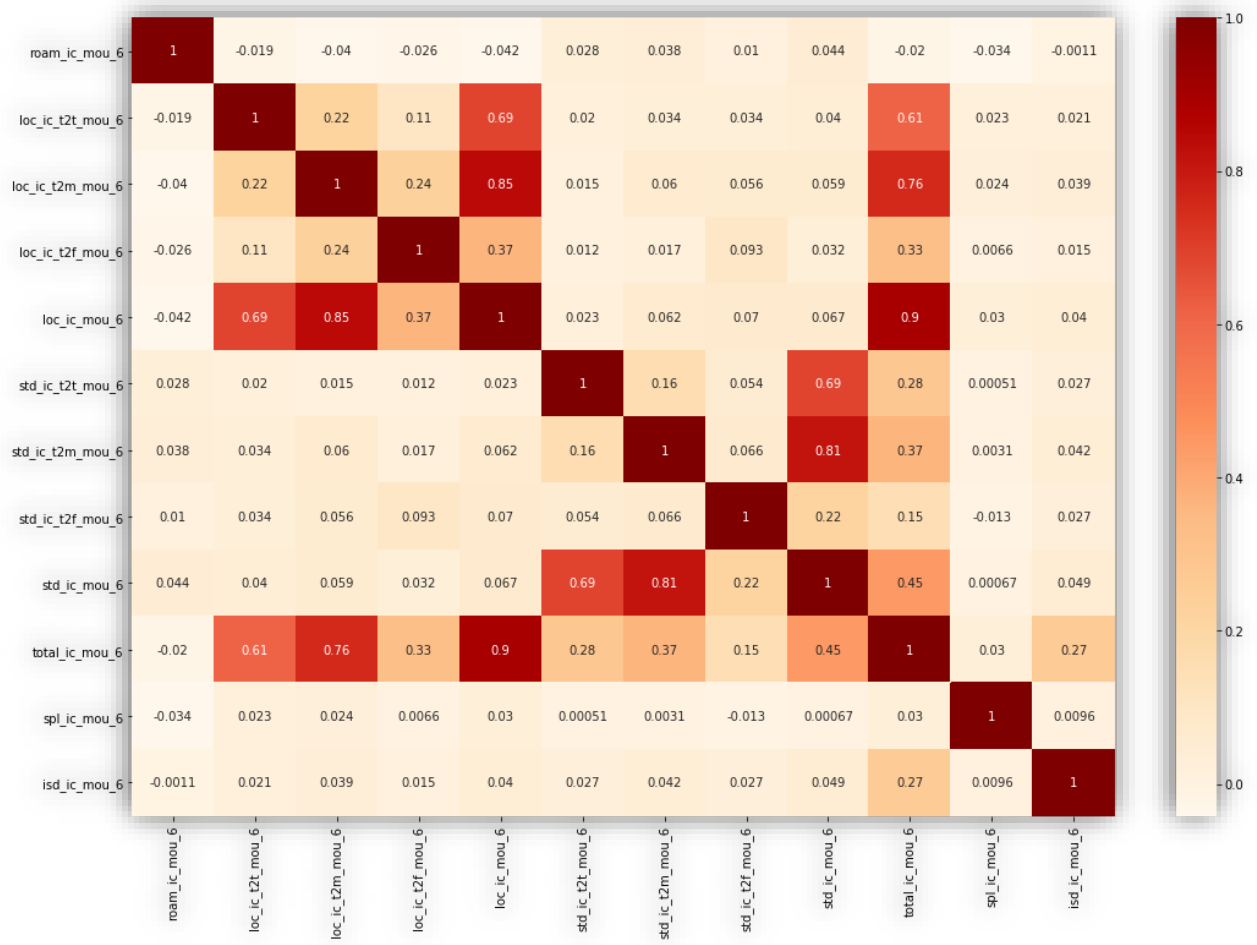
Dropping Highly corelated attributes



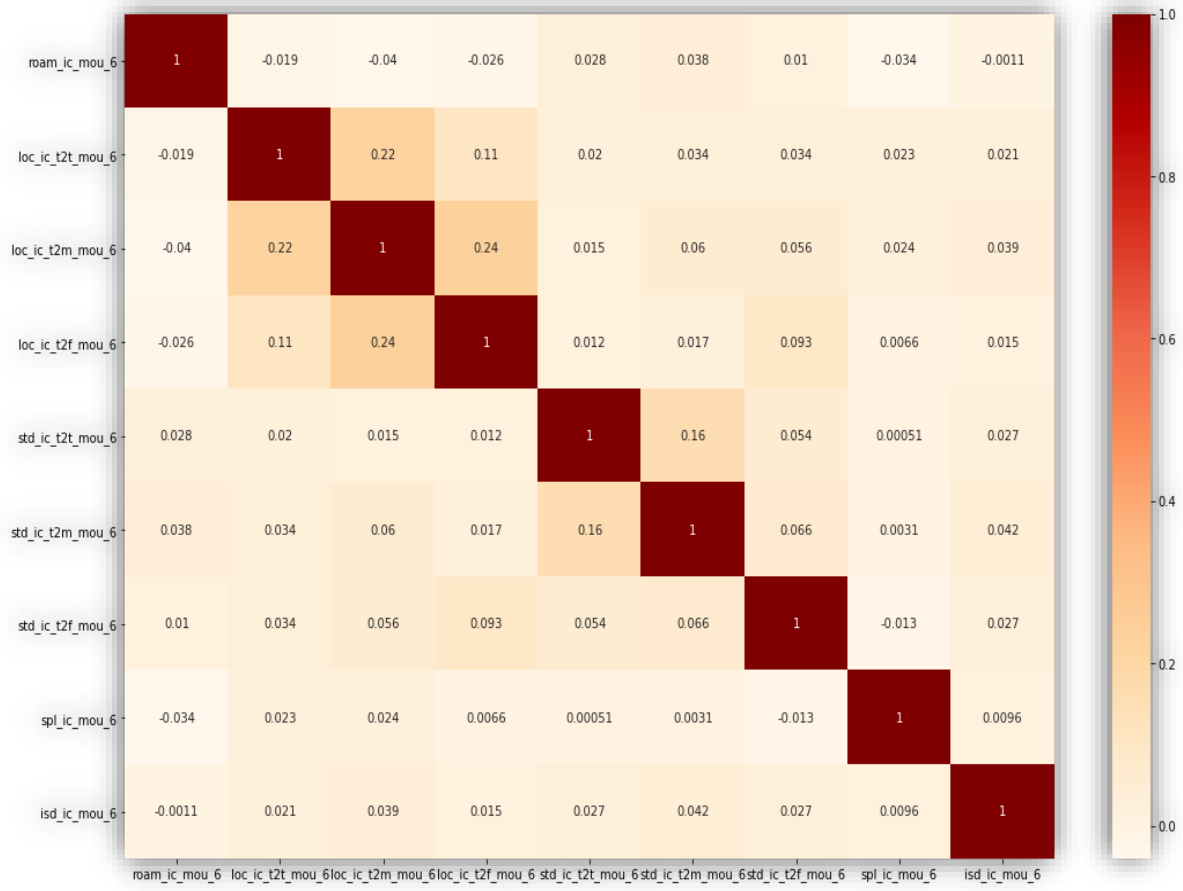
TREY

PLOTTING HEAT MAP

Checking for Incoming MOU variables

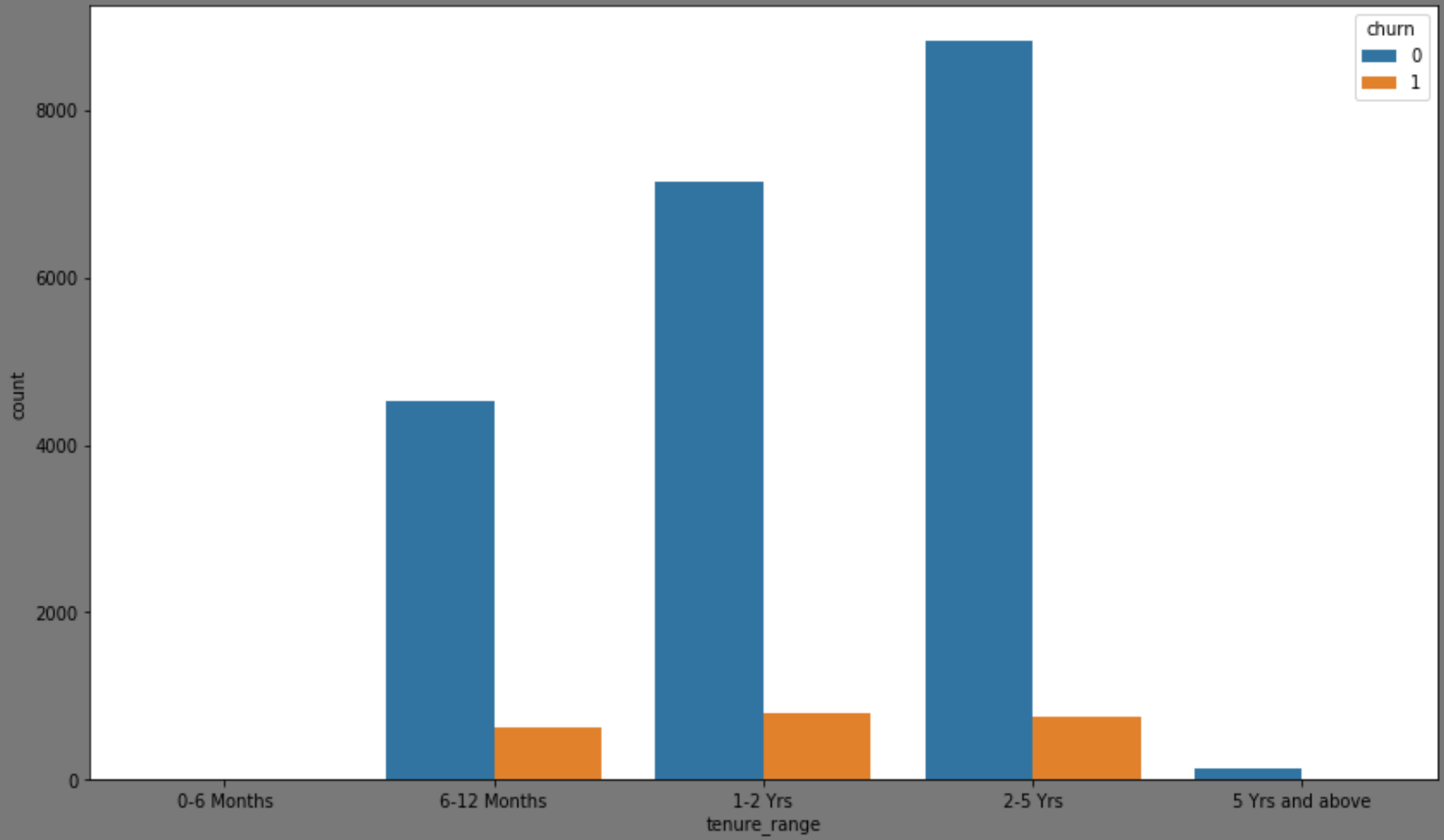


Dropping Highly corelated attributes



TREY

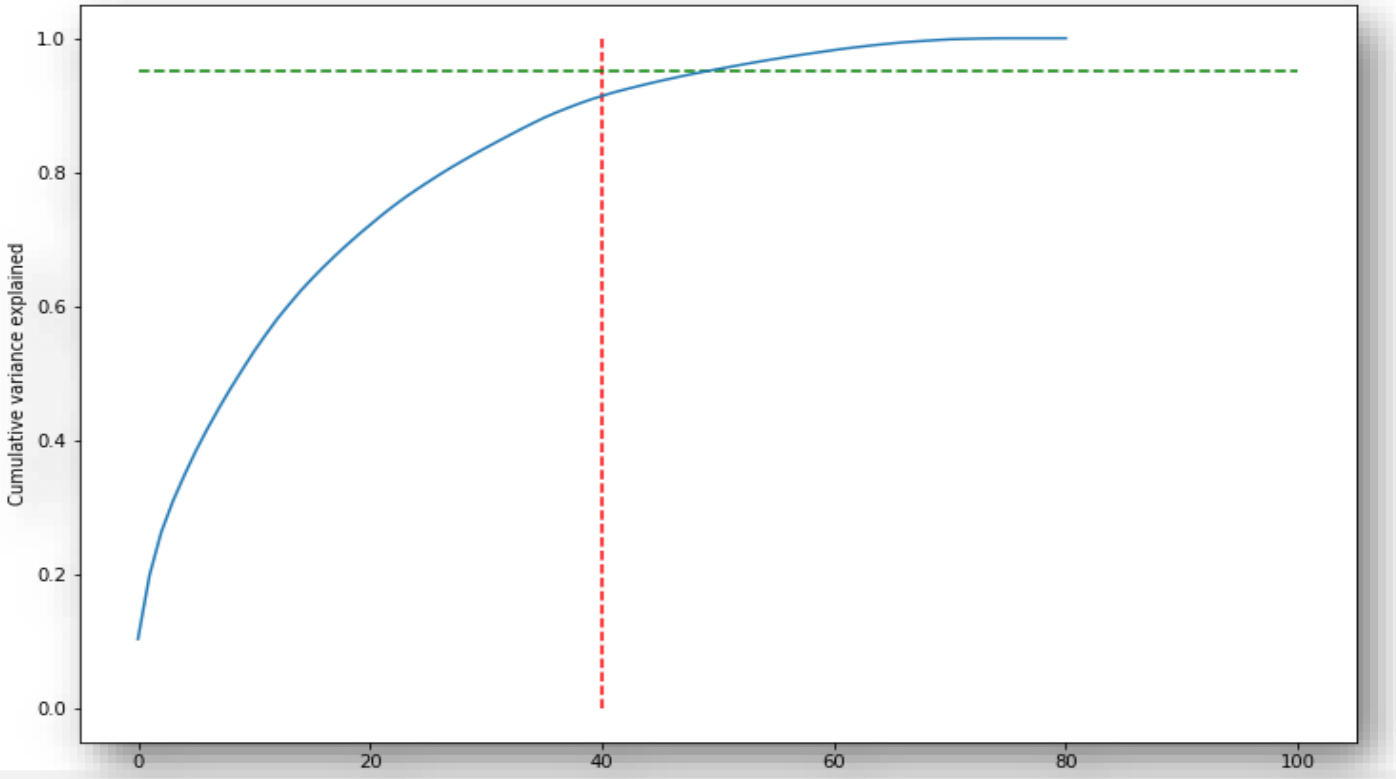
Tenure Analysis for Customers



PCA : Principal Component Analysis

- While computing the principal components, we must not include the entire dataset. Model building is all about doing well on the data we haven't seen yet!

- We'll calculate the PCs using the train data, and apply them later on the test data

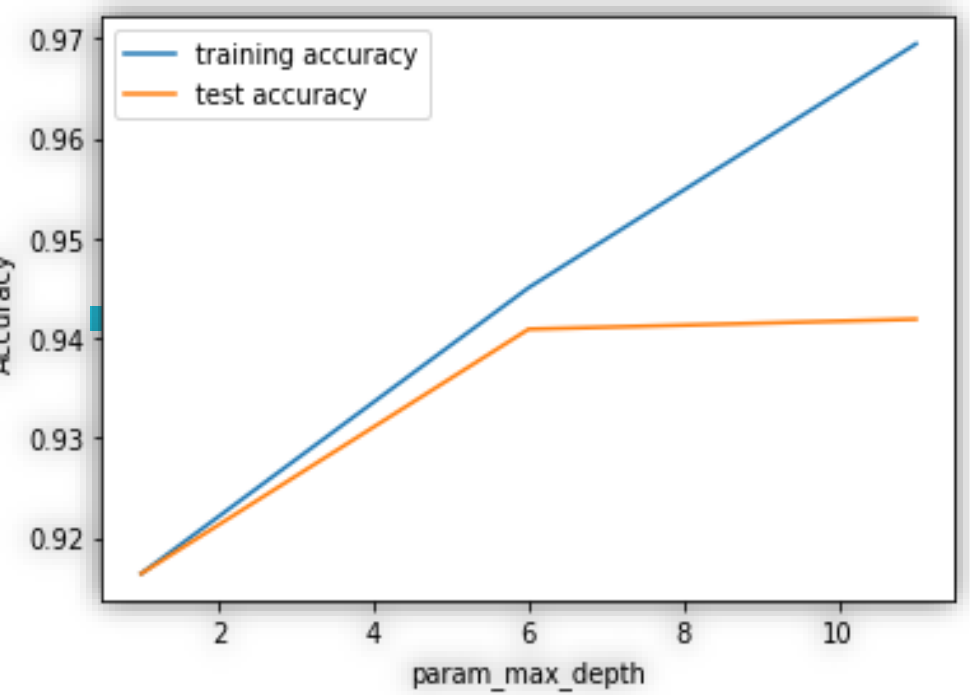


Analysis:

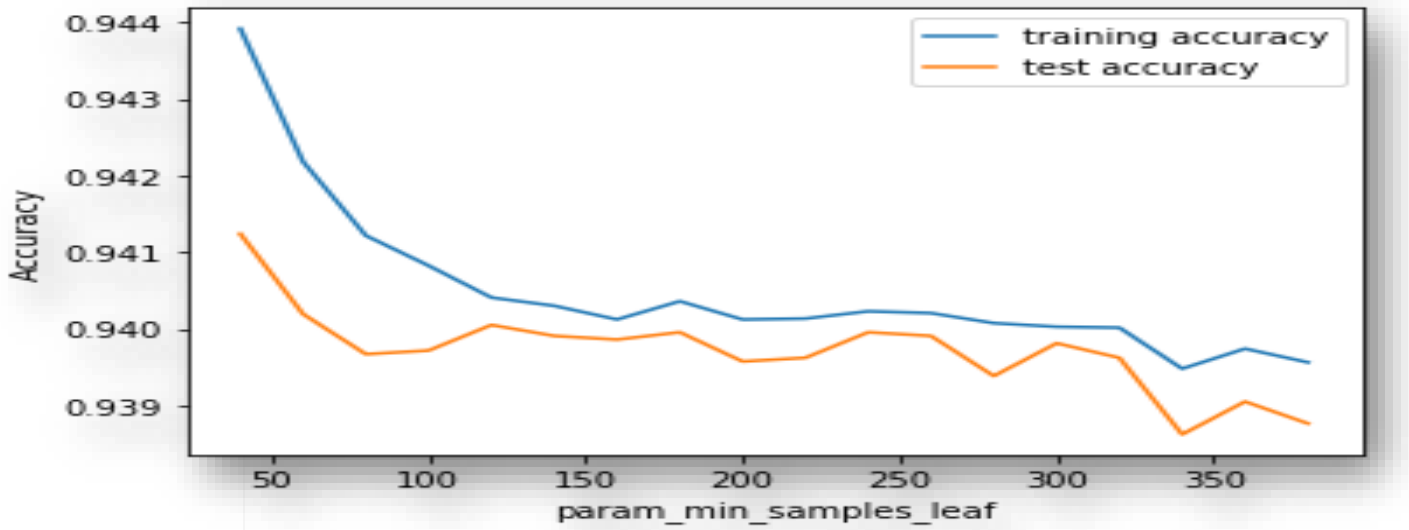
Looks like 45 components are enough to describe 95% of the variance in the dataset.

We'll choose 45 components for our modeling

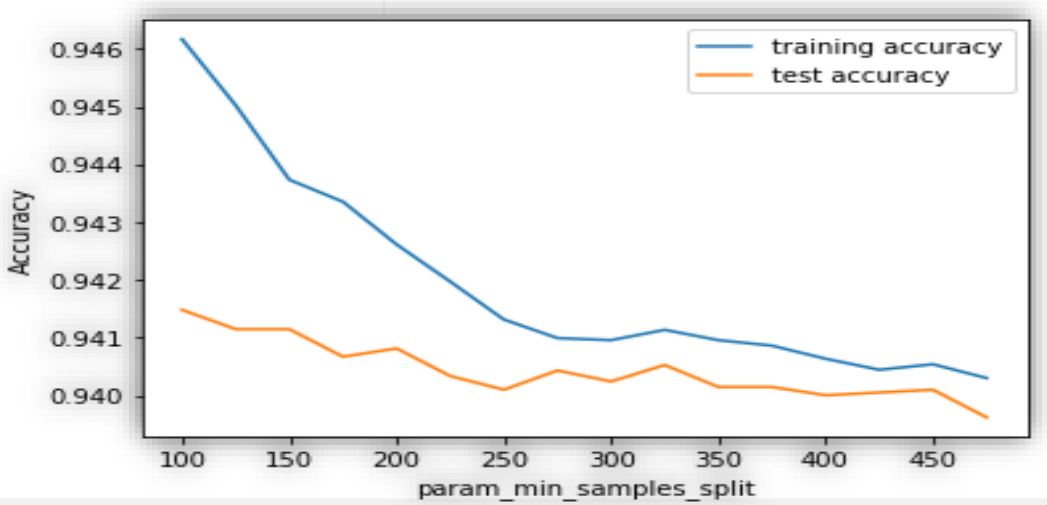
PARAMETER TUNNING



Analysis: We can see that as we increase the value of max_depth, both train and test scores increase till a point, but after that test score become stagnant. The ensemble tries to overfit as we increase the max_depth. Thus, controlling the depth of the constituent trees will help reduce overfitting in the forest. 6 value has peek converges and can be used for grid view search.



Analysis: We can see that the model starts to overfit as value is decrease the value of min_samples_leaf. 100 seems to be a good range and that will be used in grid search



Analysis: Score almost remain the same with very low dip through the range. We will use 1000 for grid view search.

RANDOM FOREST

RF Using Default Hyperparameters

```
# Making predictions
predictions = rfc.predict(X_test)
```

```
# Importing classification report and confusion matrix from sklearn metrics
from sklearn.metrics import classification_report,confusion_matrix, accuracy_score
```

```
# Let's check the report of our default model
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       0.96      0.99      0.97      8305
           1       0.76      0.44      0.56       681

    accuracy                           0.95      8986
   macro avg       0.86      0.72      0.77      8986
weighted avg       0.94      0.95      0.94      8986
```

Applying logistic regression on the data from our Principal components

```
----------------------  Confusion Matrix ---------------------
[[8213   92]
 [ 519  162]]

----------------------  Classification Report ----------------------
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      8305
           1       0.64      0.24      0.35       681

    accuracy                           0.93      8986
   macro avg       0.79      0.61      0.66      8986
weighted avg       0.92      0.93      0.92      8986


----------------------  Accuracy Score ---------------------
0.9320053416425551
```

CONCLUSION

Business Insights

Less number of **high value customer** are churning but for last **6 month** no new high valued customer is onboarded which is concerning, and company should concentrate on that aspect.

Customers with less than **4 years** of tenure are more likely to churn and company should concentrate more on that segment by rolling out new schemes to that group.

Average revenue per user seems to be most important feature in determining churn prediction.

Incoming and Outgoing Calls on romaing for 8th month are strong indicators of churn behavior

ThankYou

LEARNING IS ALWAYS AN ASSURANCE TO GROWTH