Telecom Churn Case Study

Group member:
Mr. Anurag Srivastava
Mr. Sagar Bhattacharya
Mr. Baseeruddin Avez

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.

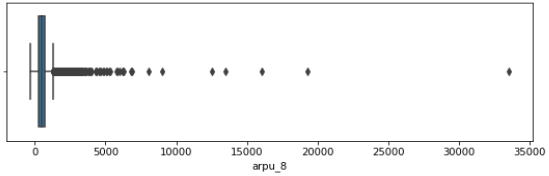
To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

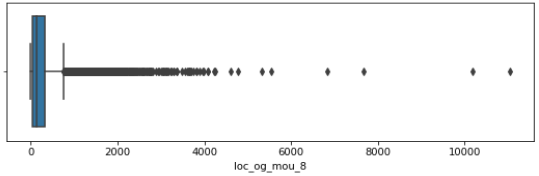
Goal

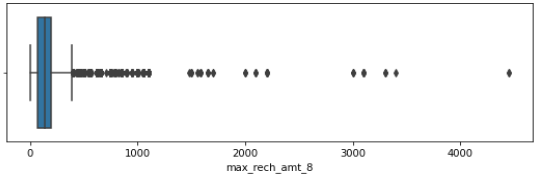
- ➔ predict which customers are at high risk of churn..
- ➔ build predictive models to identify customers at high risk of churn and identify the main indicators of churn.
- ➔ recommend strategies to manage customer churn based on observations.

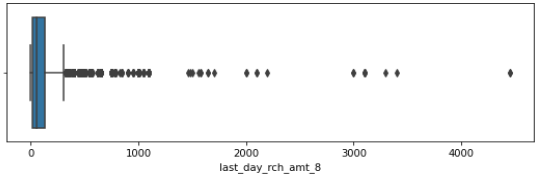
STRATEGY

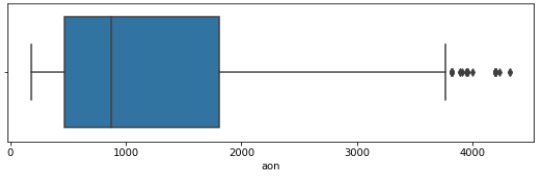
- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Data Preparation : Prepare the data for model building
- Build a logistic regression model
- Build Decision Tree Model
- Build Random Forest Model
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

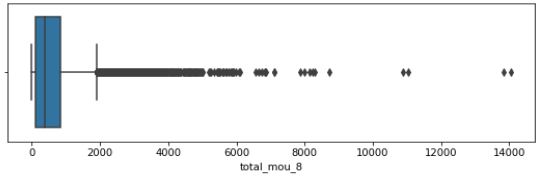


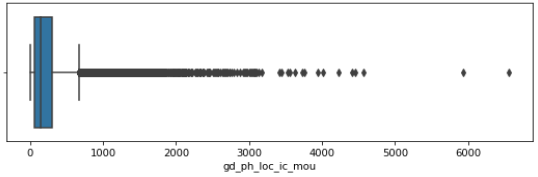


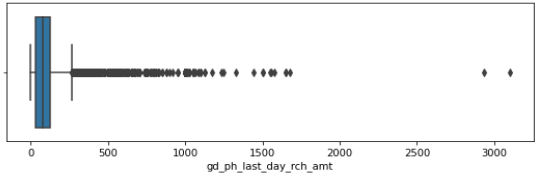


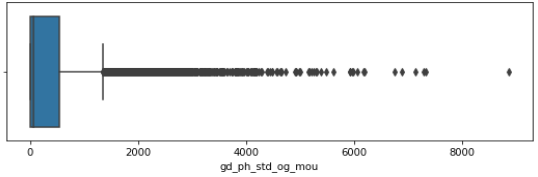


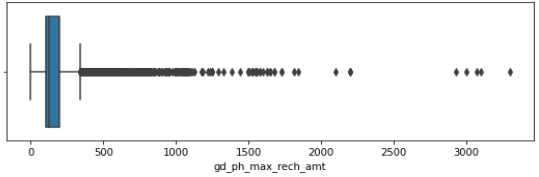


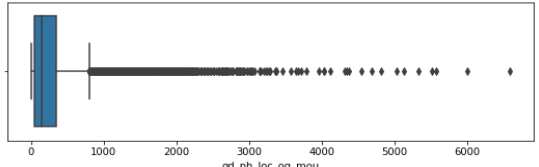


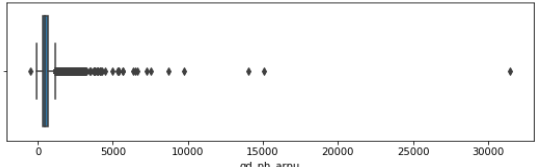












1. We can see almost every columns has some outliers, while most of them are because there are 0.0 as the service was not used some are actual outliers

2. Since we don't have actual businesspeople to check the tactfulness of the data, we will cap those features

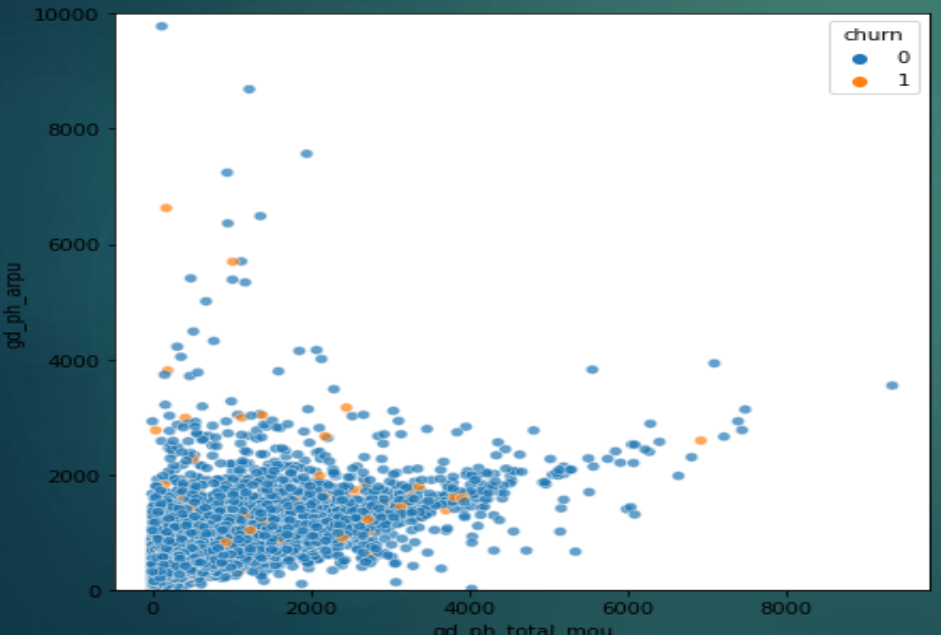
From the above slide plots, we can define following upper limits to the selected variables

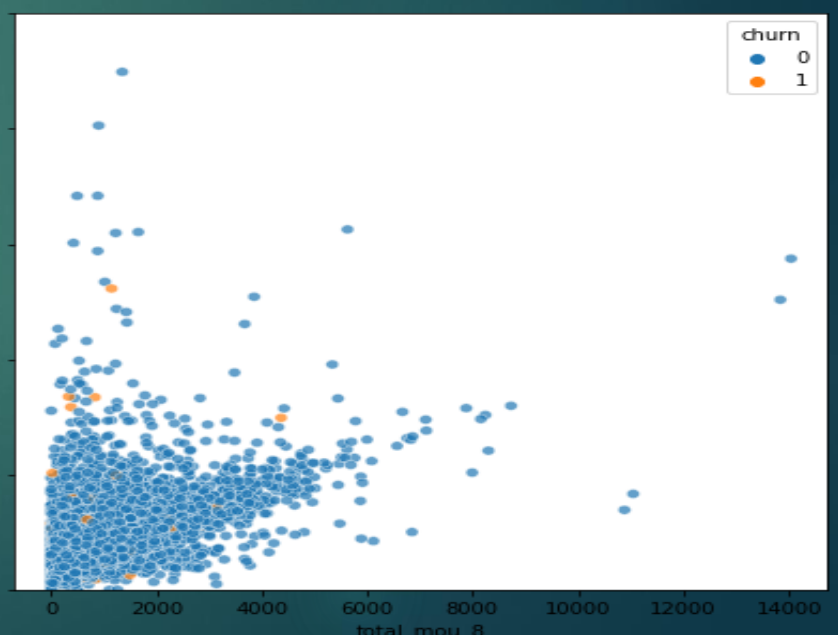
Feature	Value
arpu_8	7000
loc_og_mou_8	4000
max_rech_amt_8	1000
last_day_rch_amt_8	1000
aon	3000
total_mou_8	4000
gd_ph_loc_ic_mou	3000
gd_ph_last_day_rch_amt	1000
gd_ph_std_og_mou	4000
gd_ph_max_rech_amt	1500
gd_ph_loc_og_mou	3000
gd_ph_arpu	7000

VBC effects the revenue :

* We can see that the users who were using very less amount of VBC data and yet were generating high revenue churned

* Yet again we see that the revenue is higher towards the lesser consumption side

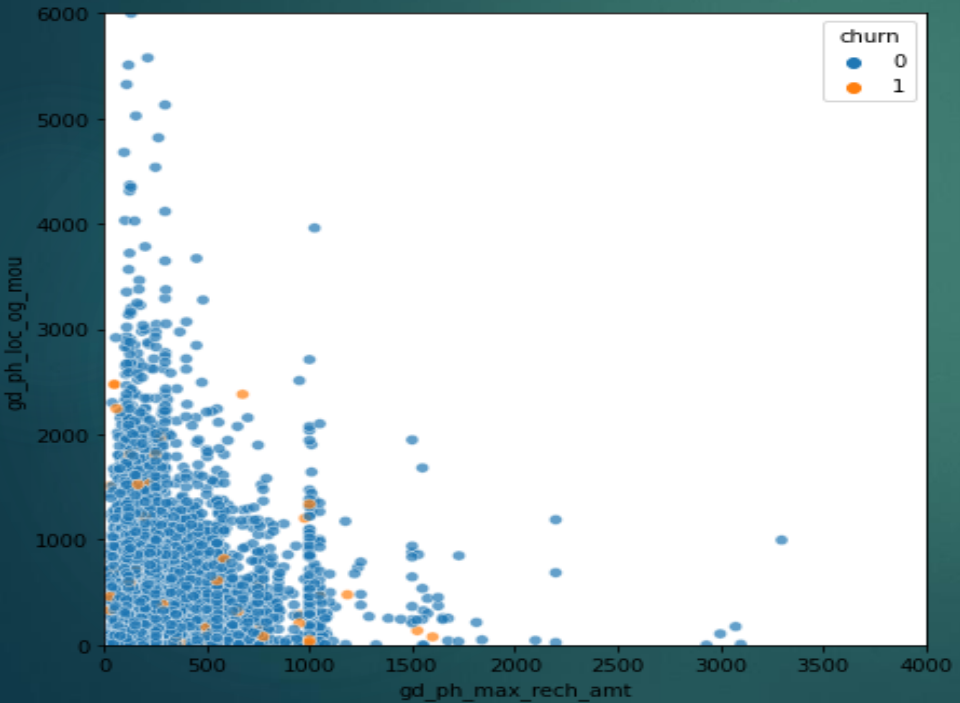


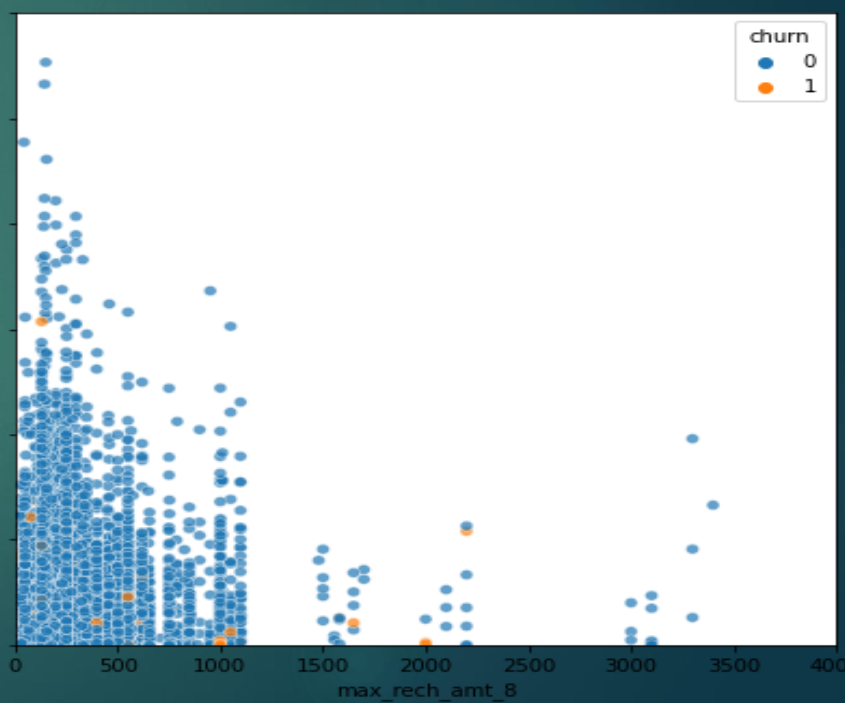


Relation between recharge amount and local outgoing calls :

Users who were recharging with high amounts were using the service for local uses less as compared to user who did lesser amounts of recharge

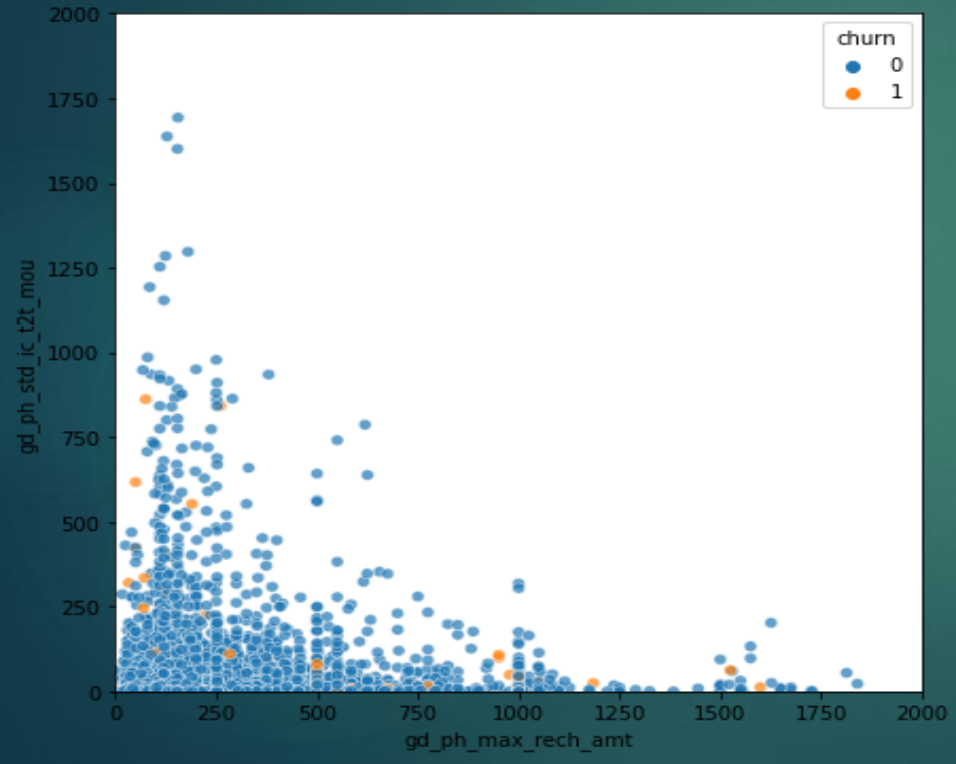
Intuitively people whose max recharge amount as well as local out going were very less even in the good phase churned more

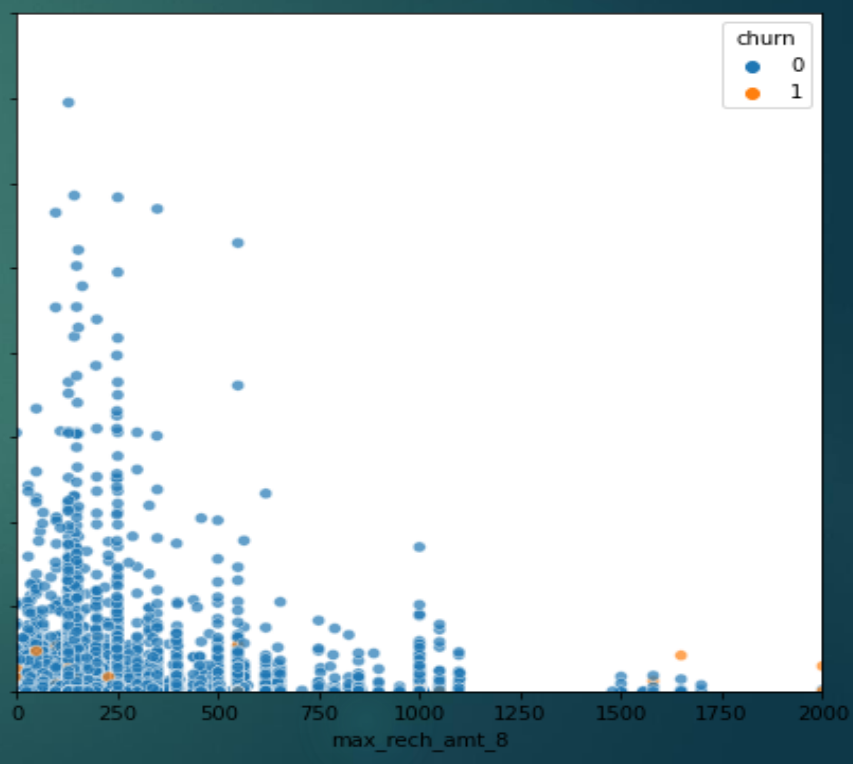




Incoming from the same service provider vs the recharge amount :

Users who have max recharge amount on the higher end and still have low incoming call MOU during the good phase, churned out.

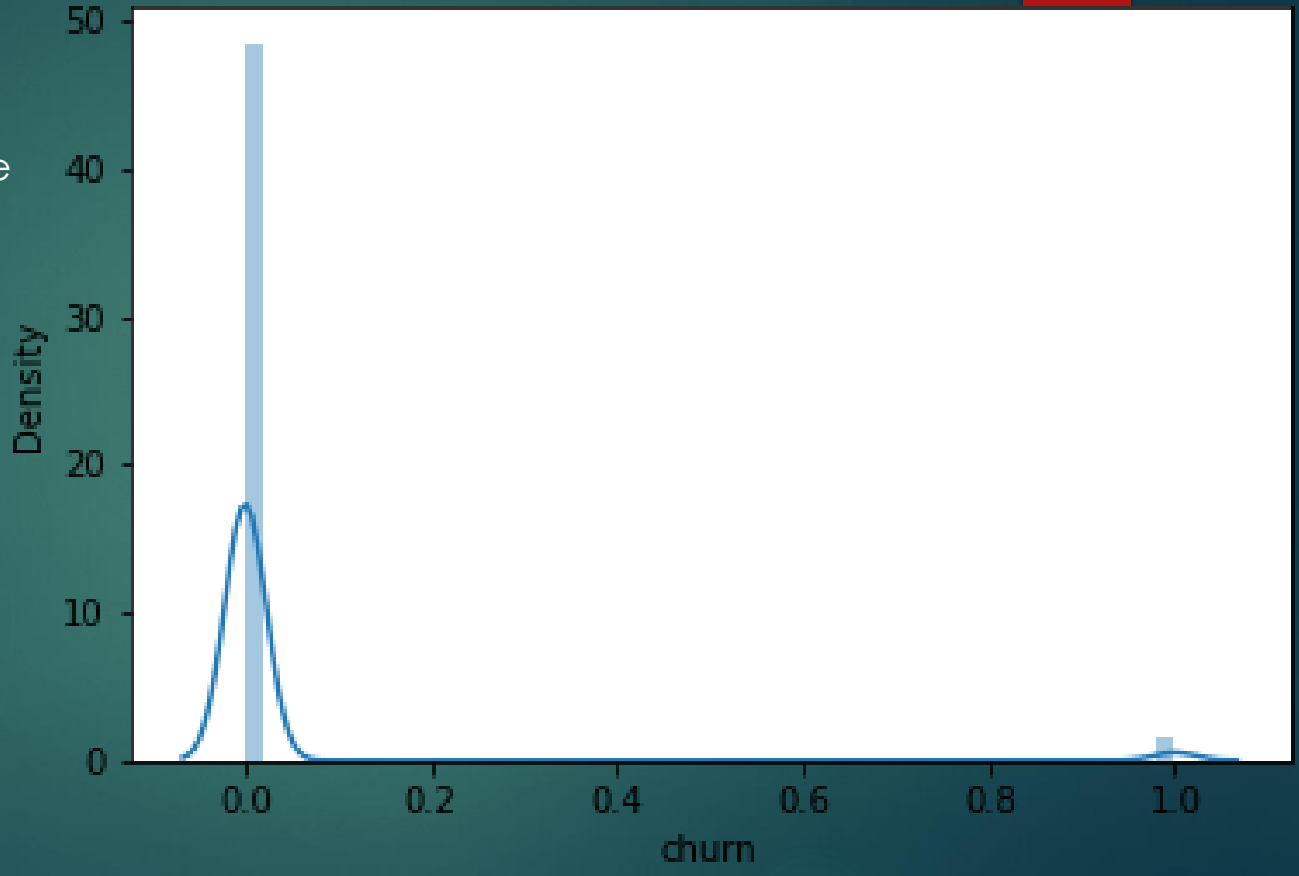




Distribution of target variable :

1. Though the variable is not skewed it is highly imbalanced, the number of non-churners in the dataset is around 94%

2. We will handle this imbalance using SMOTE algorithm



Model Building

- SPLITTING INTO TRAIN AND TEST SET
- SCALE VARIABLES IN TRAIN SET
- BUILD THE FIRST MODEL
- USE RFE TO ELIMINATE LESS RELEVANT VARIABLES
- BUILD THE NEXT MODEL
- ELIMINATE VARIABLES BASED ON HIGH P-VALUES
- CHECK VIF VALUE FOR ALL THE EXISTING COLUMNS
- PREDICT USING TRAIN SET
- EVALUATE ACCURACY AND OTHER METRIC
- PREDICT USING TEST SET
- PRECISION AND RECALL ANALYSIS ON TEST PREDICTIONS

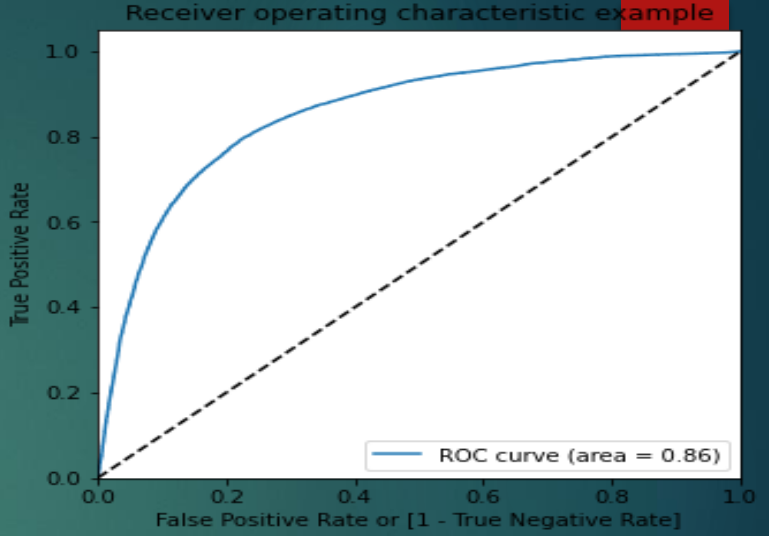
Logistic Regression

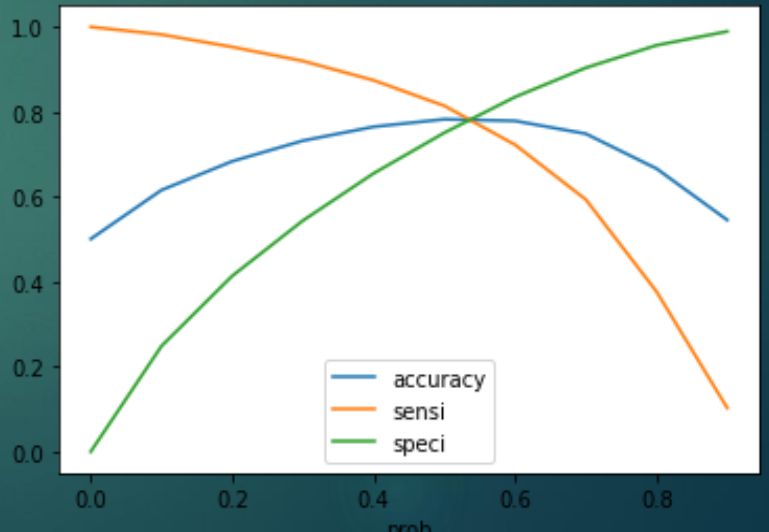
For logistic regression we used the unaltered X and y so that we can use RFE for feature selection instead of PCA, to find out the strong predictor of churn

- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.5
- 78% Accuracy
- 81 %Specificity
- 75 %Sensitivity

So, using Logistic regression we are getting an accuracy of 78.5% on train data and 78.8% on test data

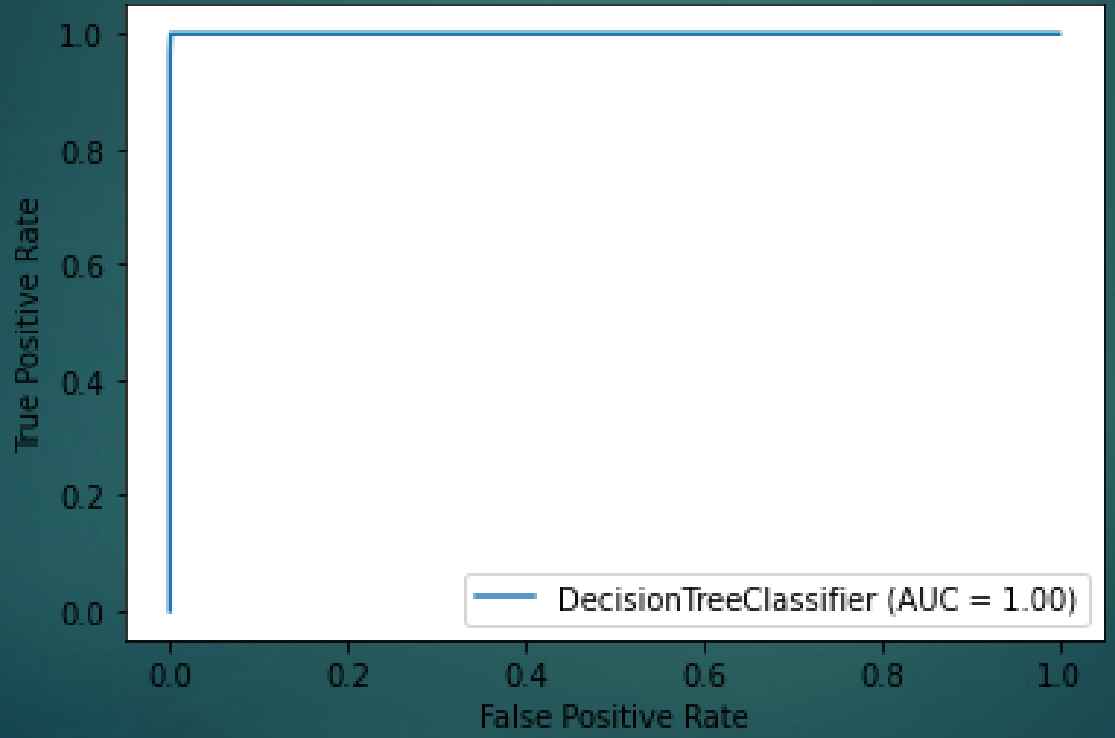
We can clearly see most of the critical features are form the action phase, which is inline with the business understanding that action phase needs more attention





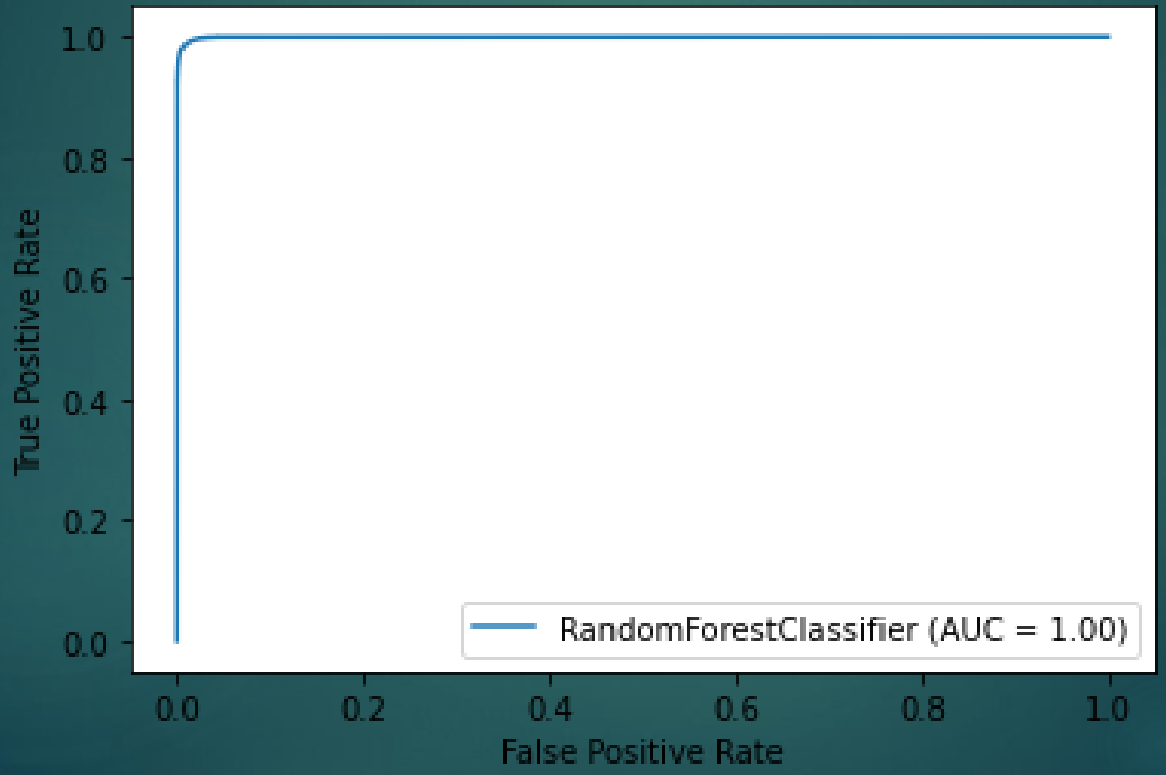
Decision Tree

We are getting an accuracy of 90% on test data, with decision tree



Random Forest

We are getting an accuracy of 95% on test data, with Random forest



The top 10 predictors are :

Features
loc_og_mou_8
total_rech_num_8
monthly_3g_8
monthly_2g_8
gd_ph_loc_og_mou
gd_ph_total_rech_num
last_day_rch_amt_8
std_ic_t2t_mou_8
sachet_2g_8
aon

Conclusion

- Given our business problem, to retain their customers, we need higher recall. As giving an offer to an user not going to churn will cost less as compared to loosing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.

- When we compare the models trained, we can see the tuned random forest and ada boost are performing the best, which is highest accuracy along with highest recall i.e., 95% and 97% respectively. So, we will go with random forest instead of adaboost as that is comparatively simpler model.

- Users whose maximum recharge amount is less than 200 even in the good phase, should have a tag and re-evaluated time to time as they are more likely to churn

Users that have been with the network less than 4 years, should be monitored time to time, as from data we can see that users who have been associated with the network for less than 4 years tend to churn more

MOU is one of the major factors, but data especially VBC if the user is not using a data pack if another factor to look out


Thank You !