TELECOM CHURN CASE STUDY

by Akash Gupta

PROJECT PROBLEM STATEMENT

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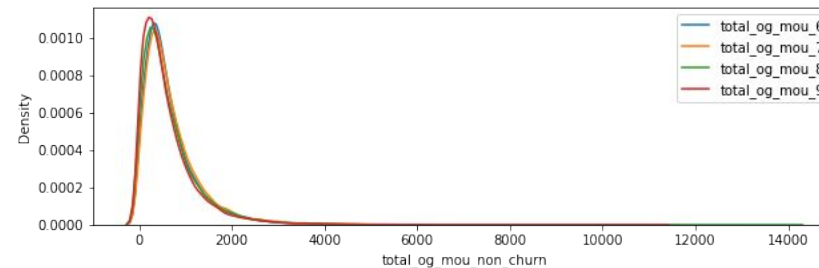
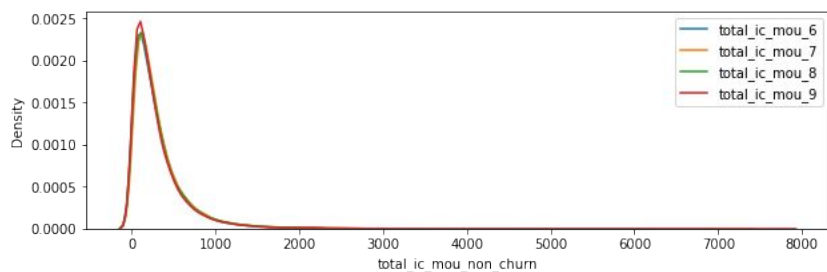The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behaviour during churn will be helpful.

EXPLORATORY DATA ANALYSIS

INCOMING AND OUTGOING CALLS ANALYSIS FOR CHURNS
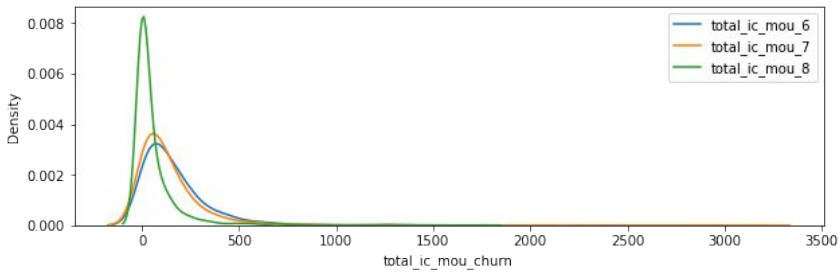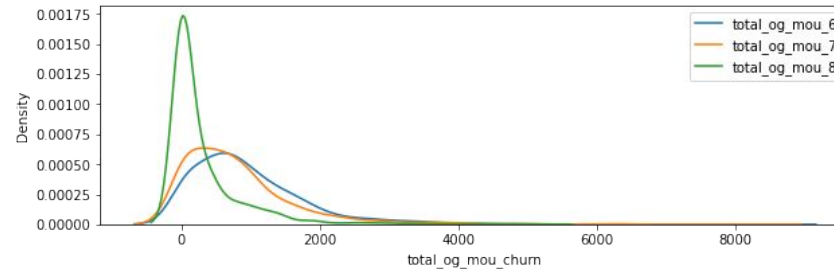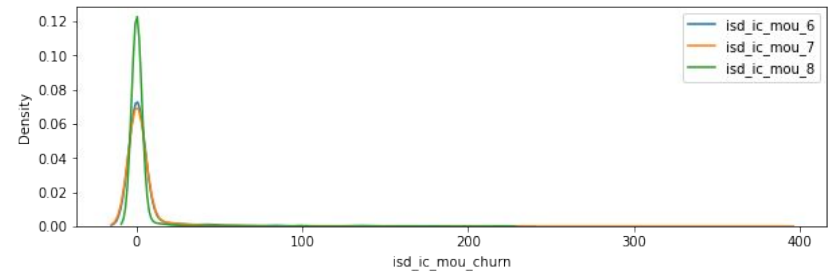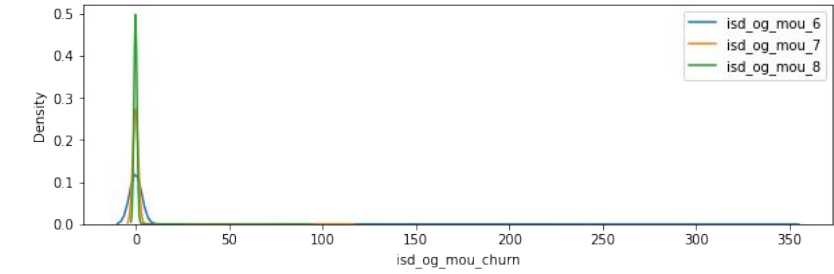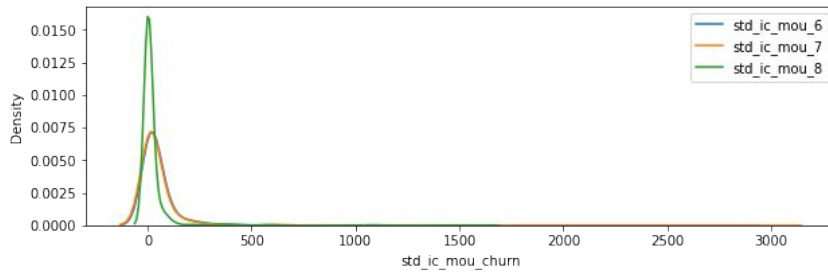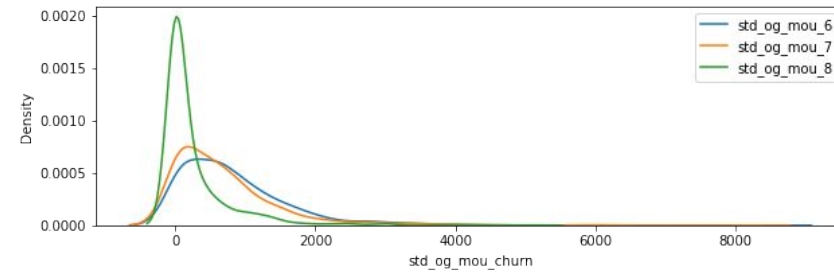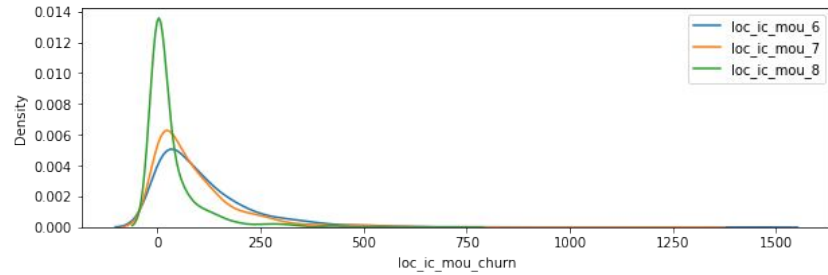
INCOMING AND OUTGOING CALLS ANALYSIS FOR NON-CHURNS
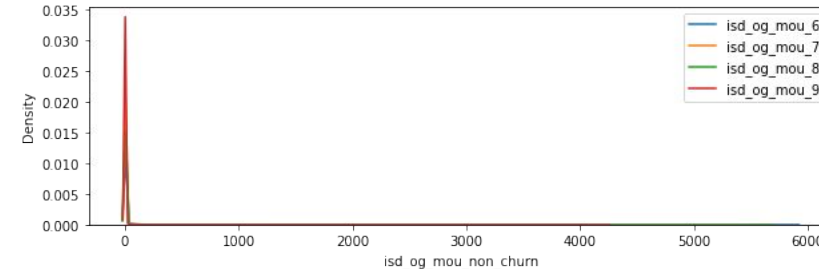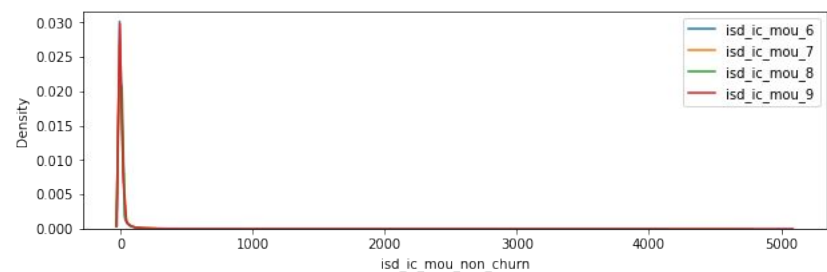
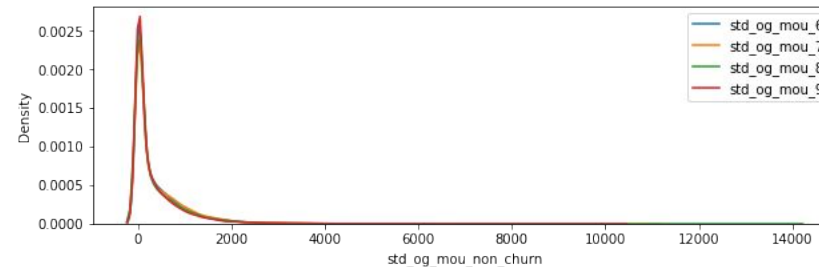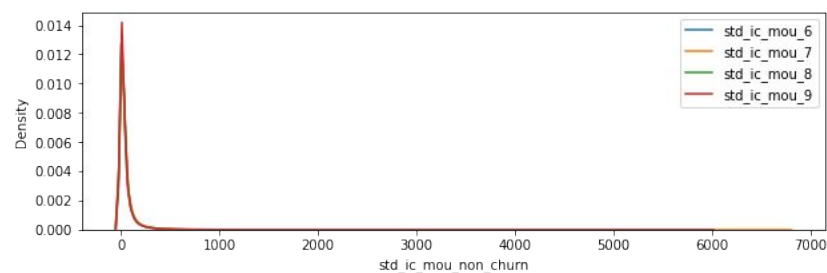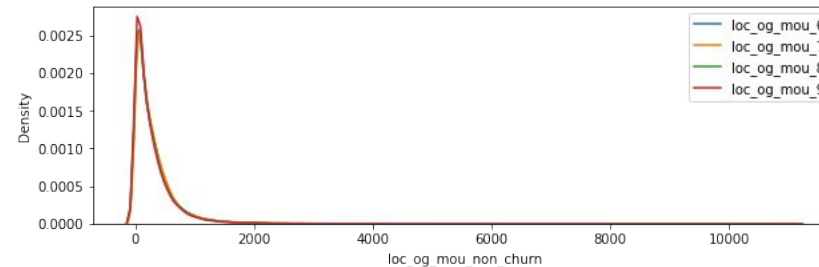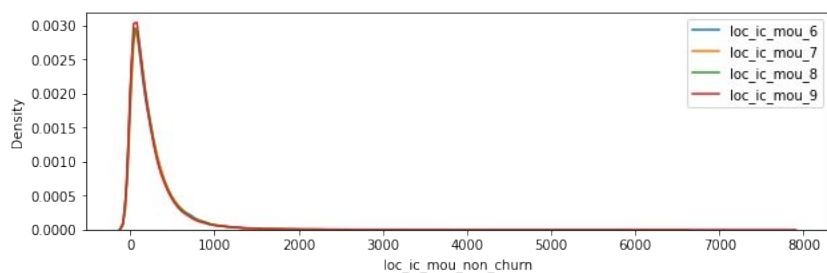2G & 3G VOLUME ANALYSIS FOR CHURNS
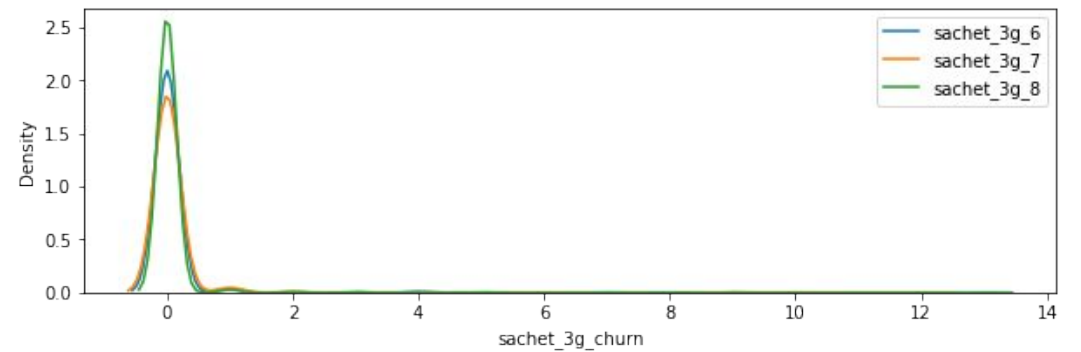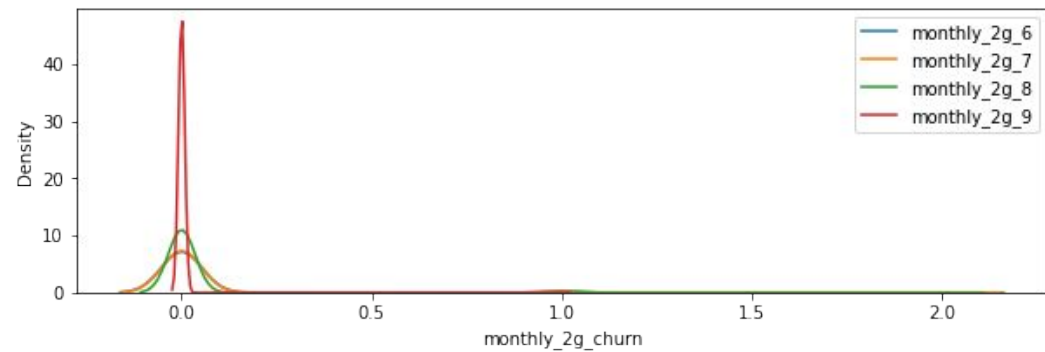
2G & 3G VOLUME ANALYSIS FOR NON-CHURNS
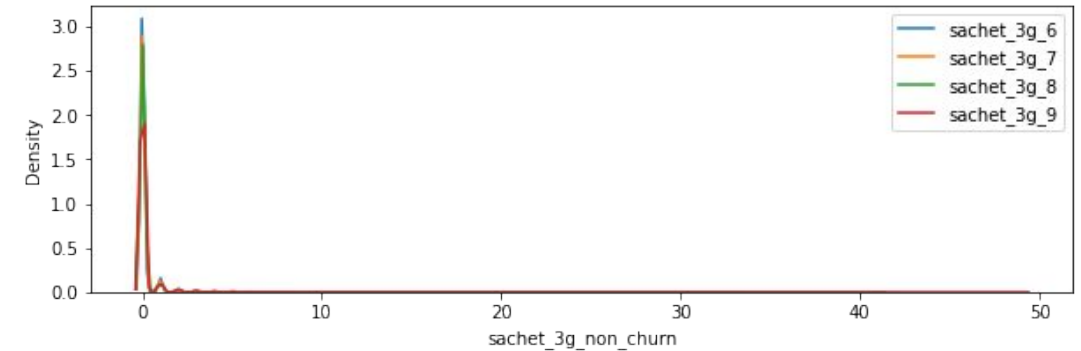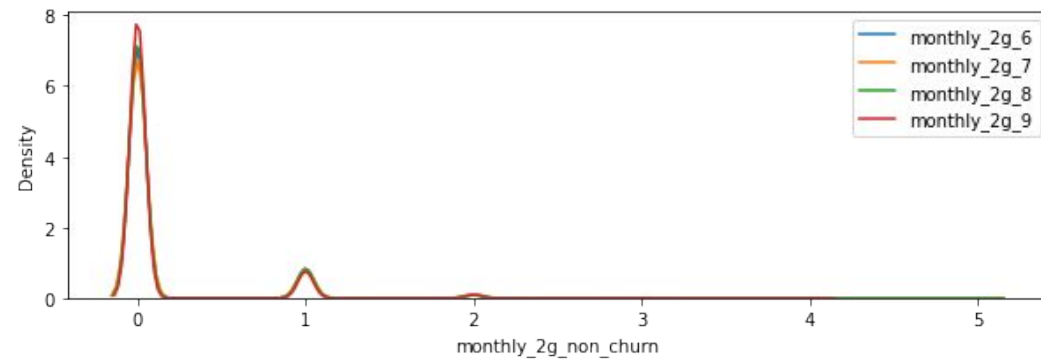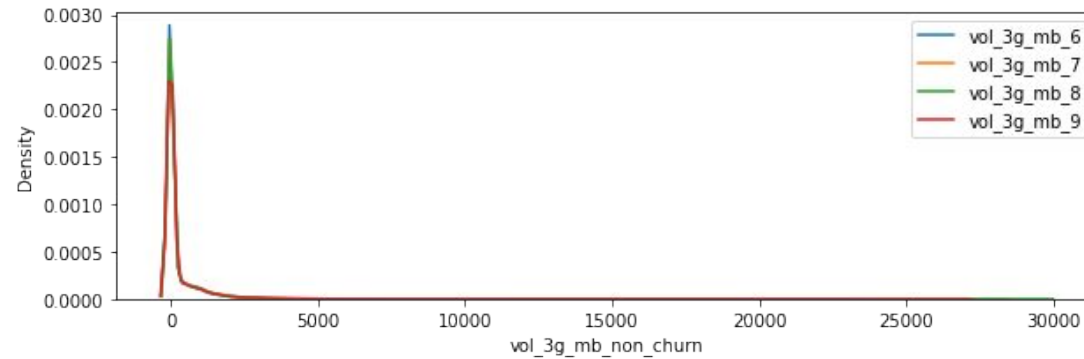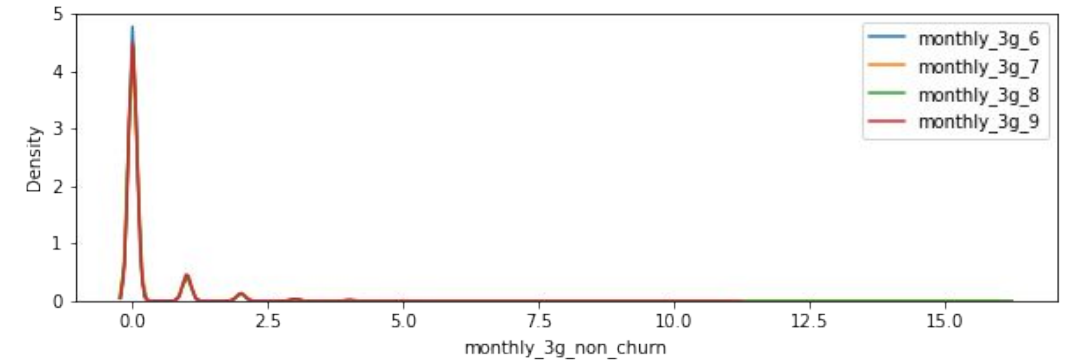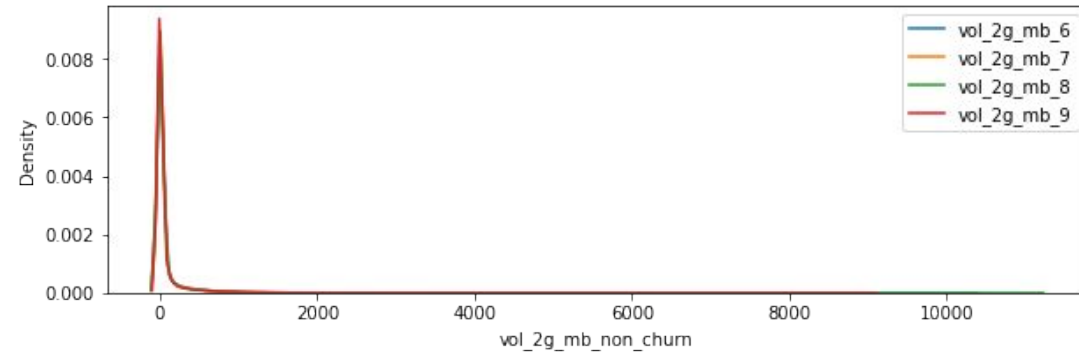
RECHARGE AMOUNT ANALYSIS FOR CHURNS
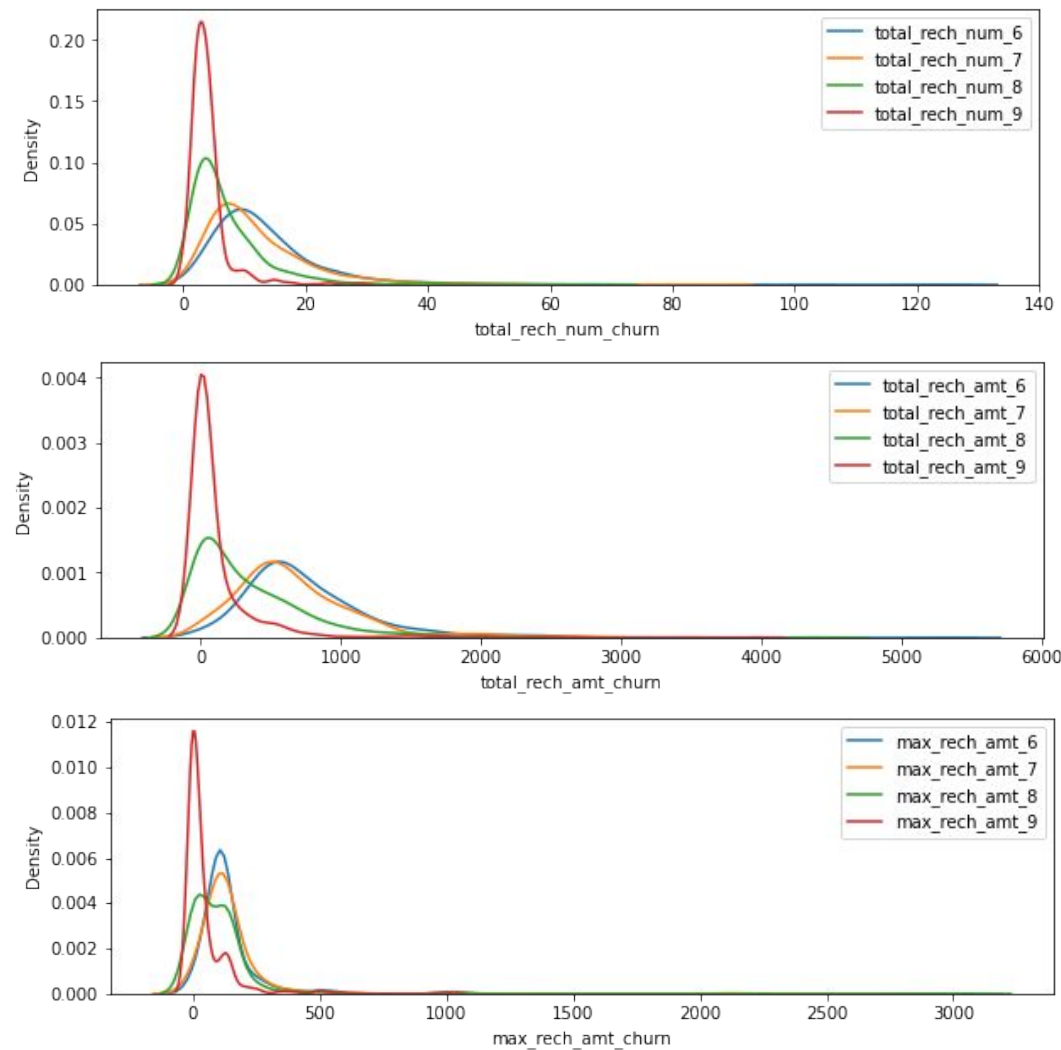
RECHARGE AMOUNT ANALYSIS FOR NON_CHURNS
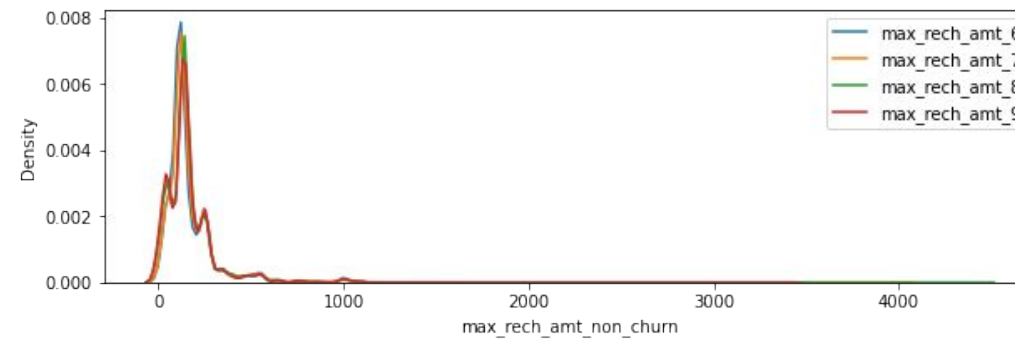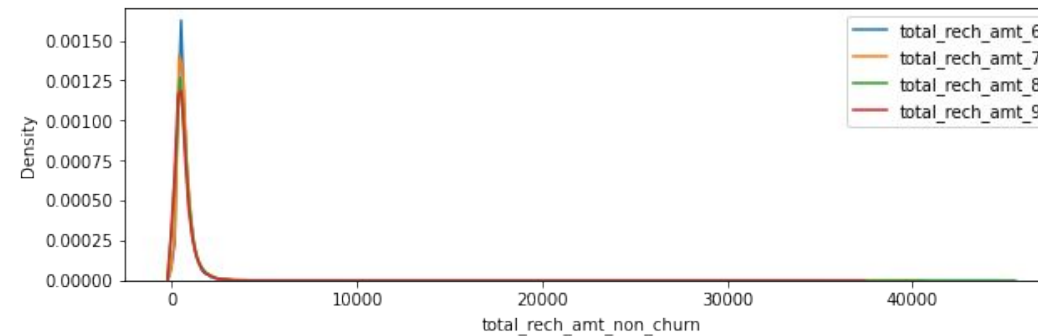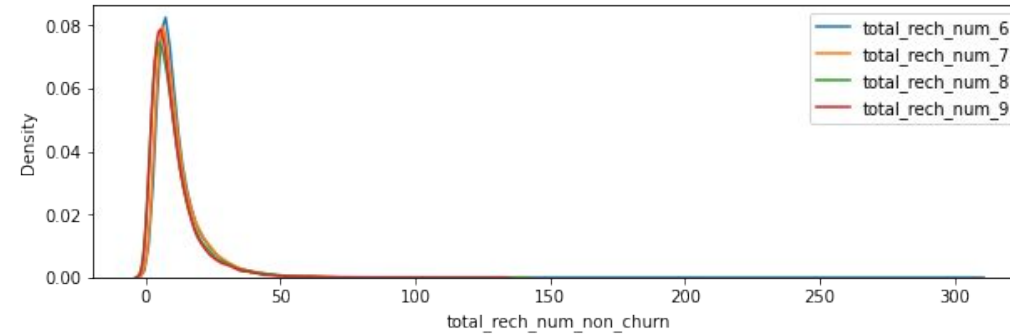
ARPU FOR CHURNS VS NON CHURNS
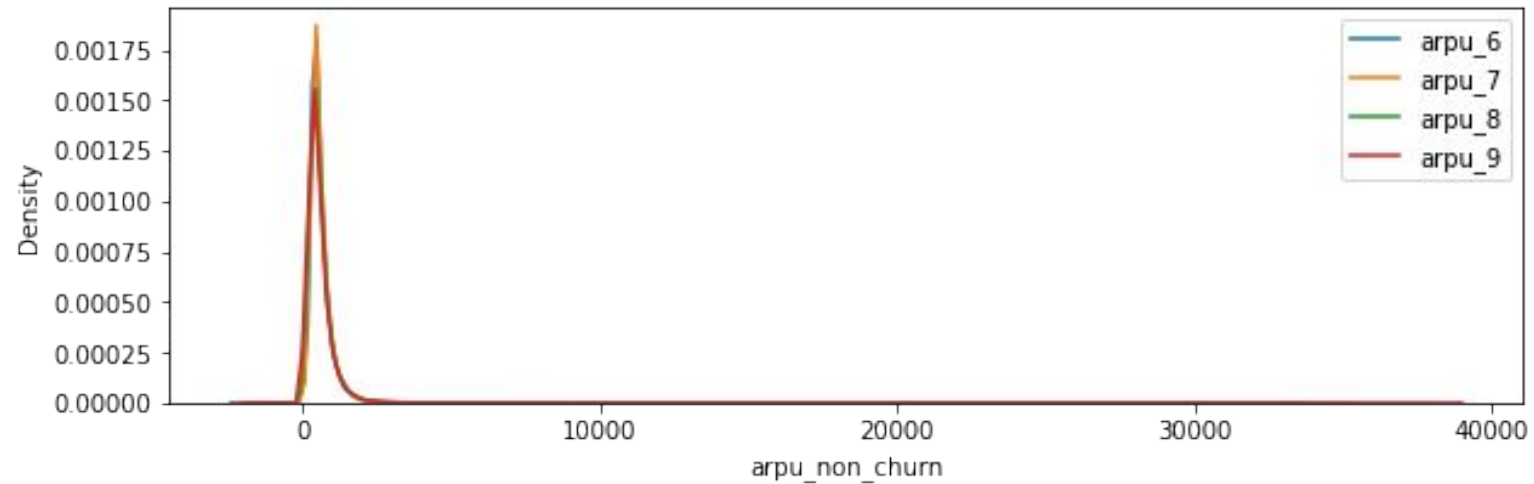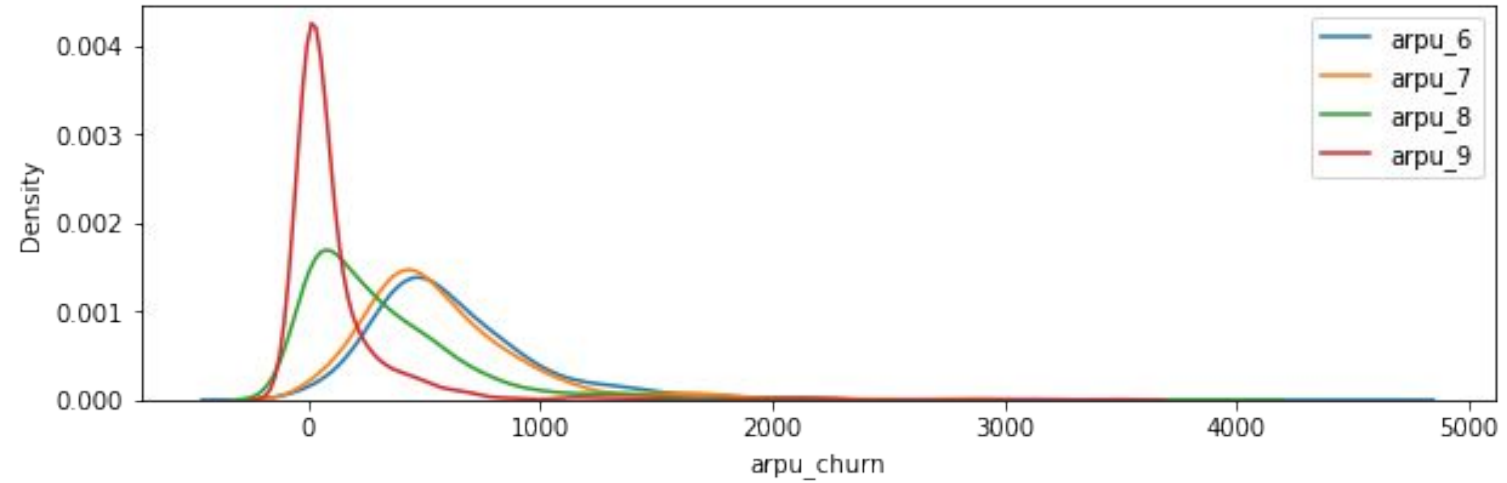
CLASS IMBALANCE
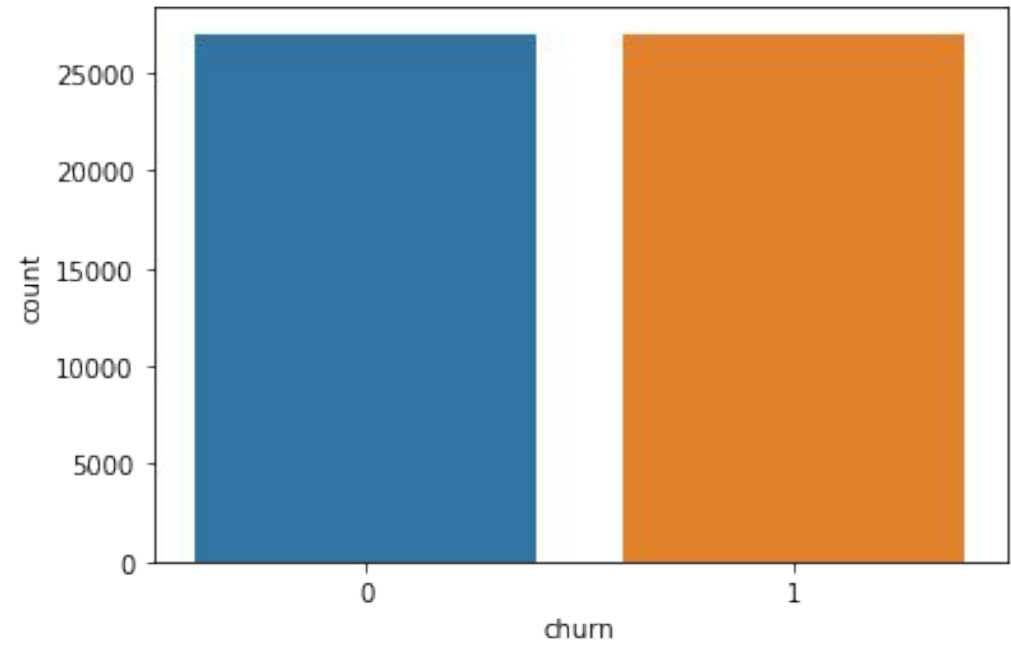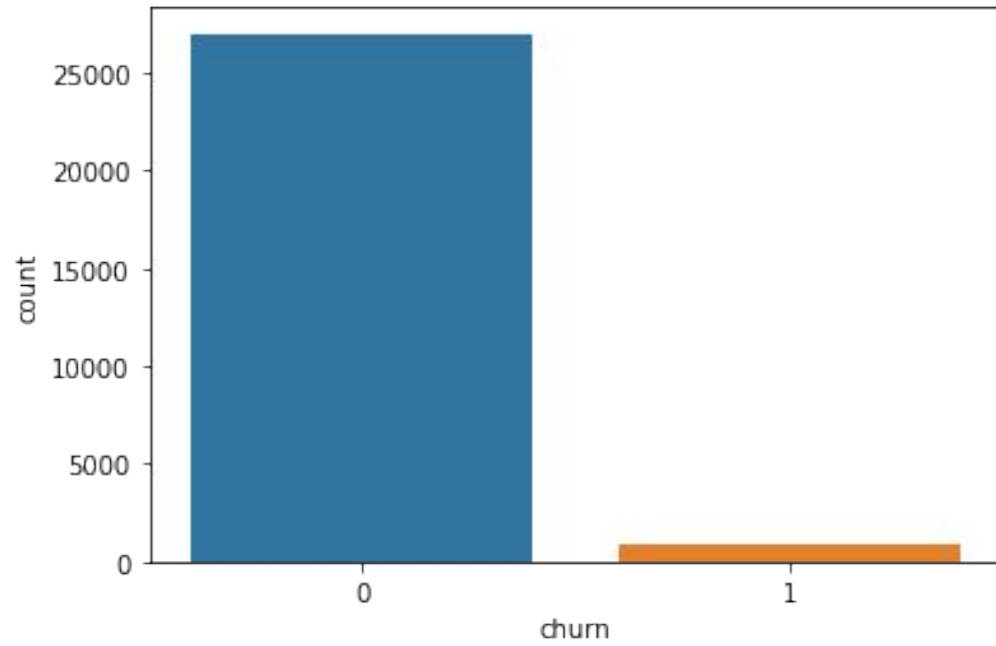
After applying SMOTE Method

FEATURE SELECTION USING RFE

Inference

- Customer is less likely to churn if their recharge amount and number of recharges is high.

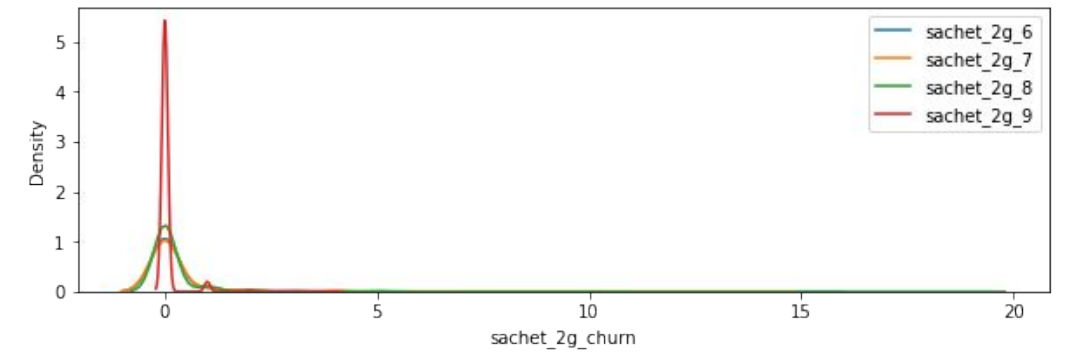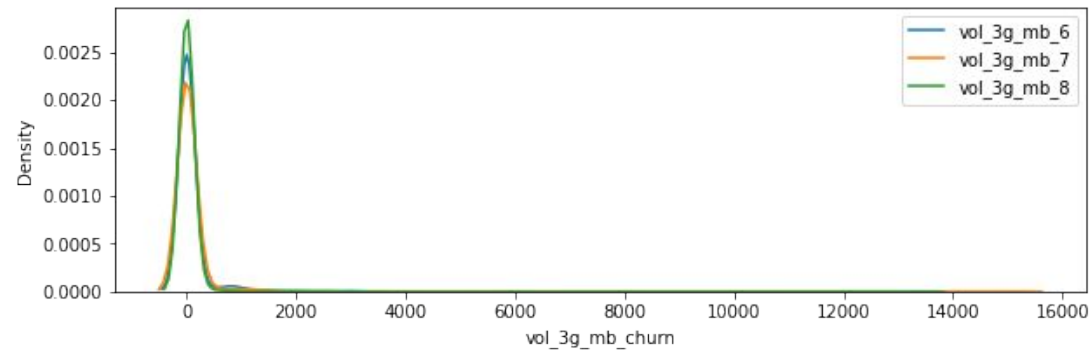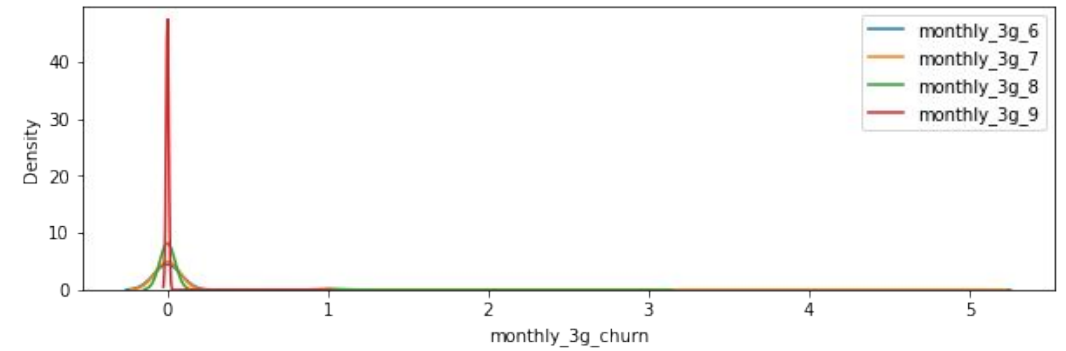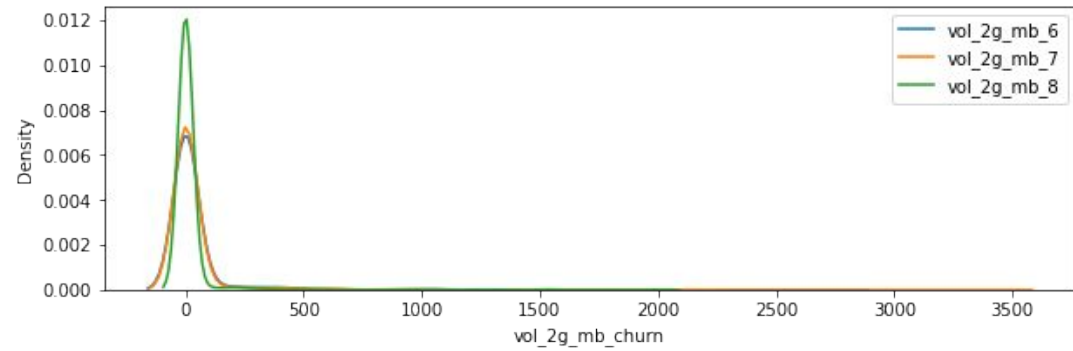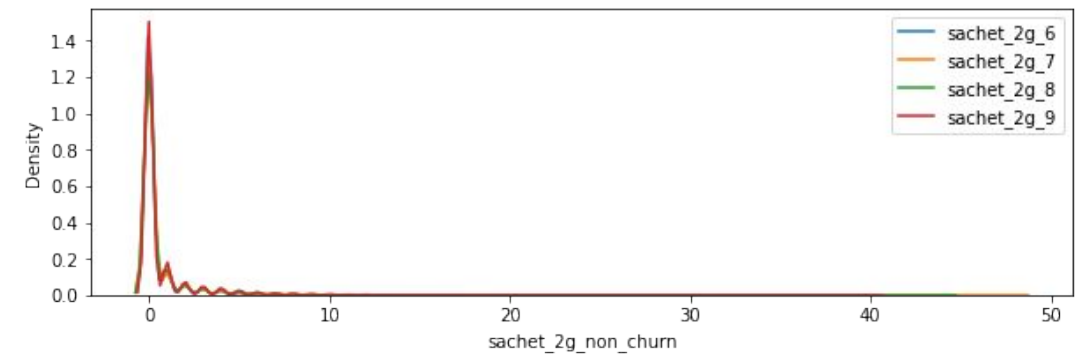- Customer is less likely to churn if they have been on the network for a while (aon or age on network is high).

- Total recharge amount and number of recharges increase if the customer uses Local network.

- As the number of Local incoming and outgoing calls increase the customer is less likely to churn.

- We can also see that columns change_total_rech_num, change_max_rech_amt, change_total_rech_amt, change_arpu are highly correlated.
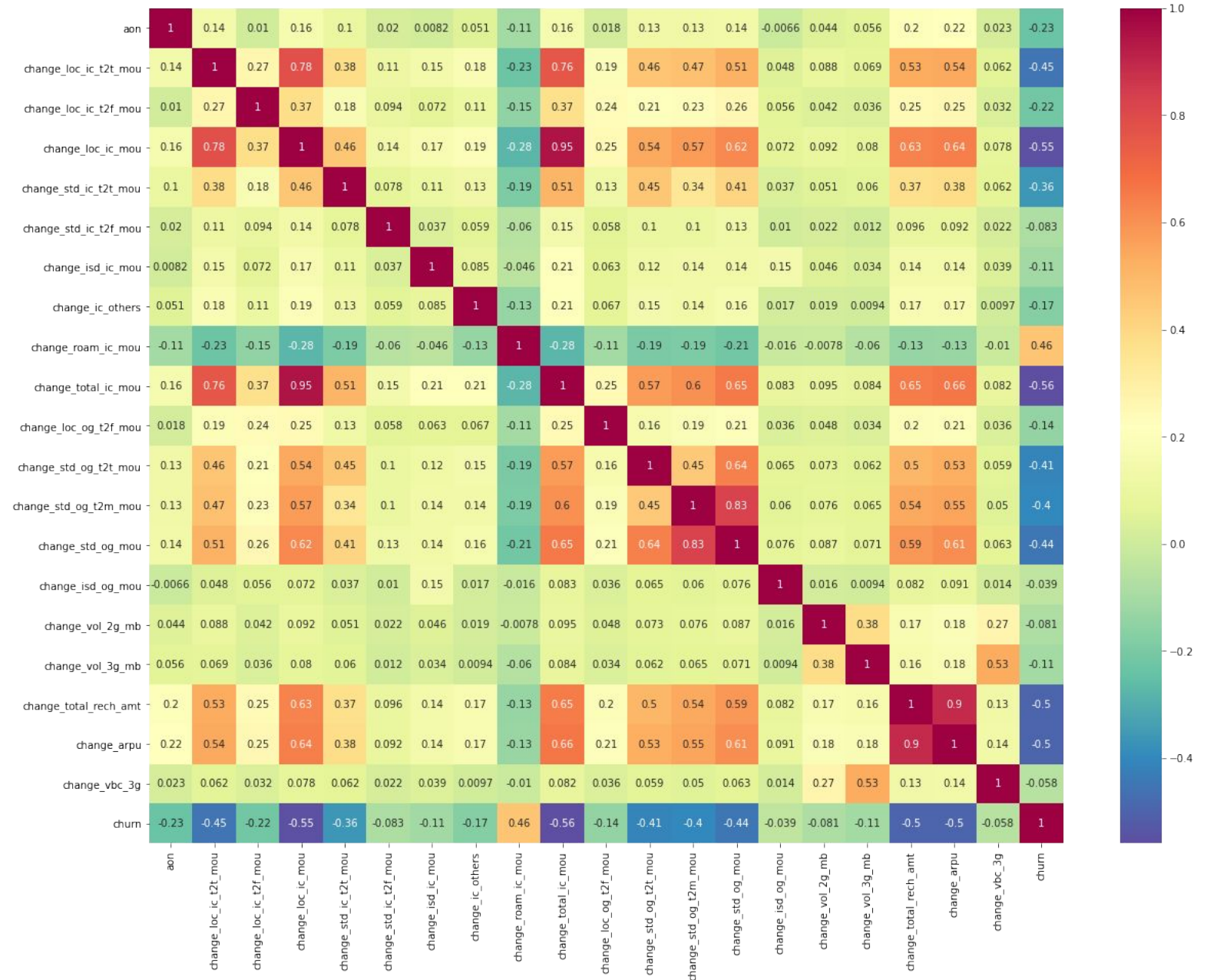
FEATURE REDUCTION USING PCA
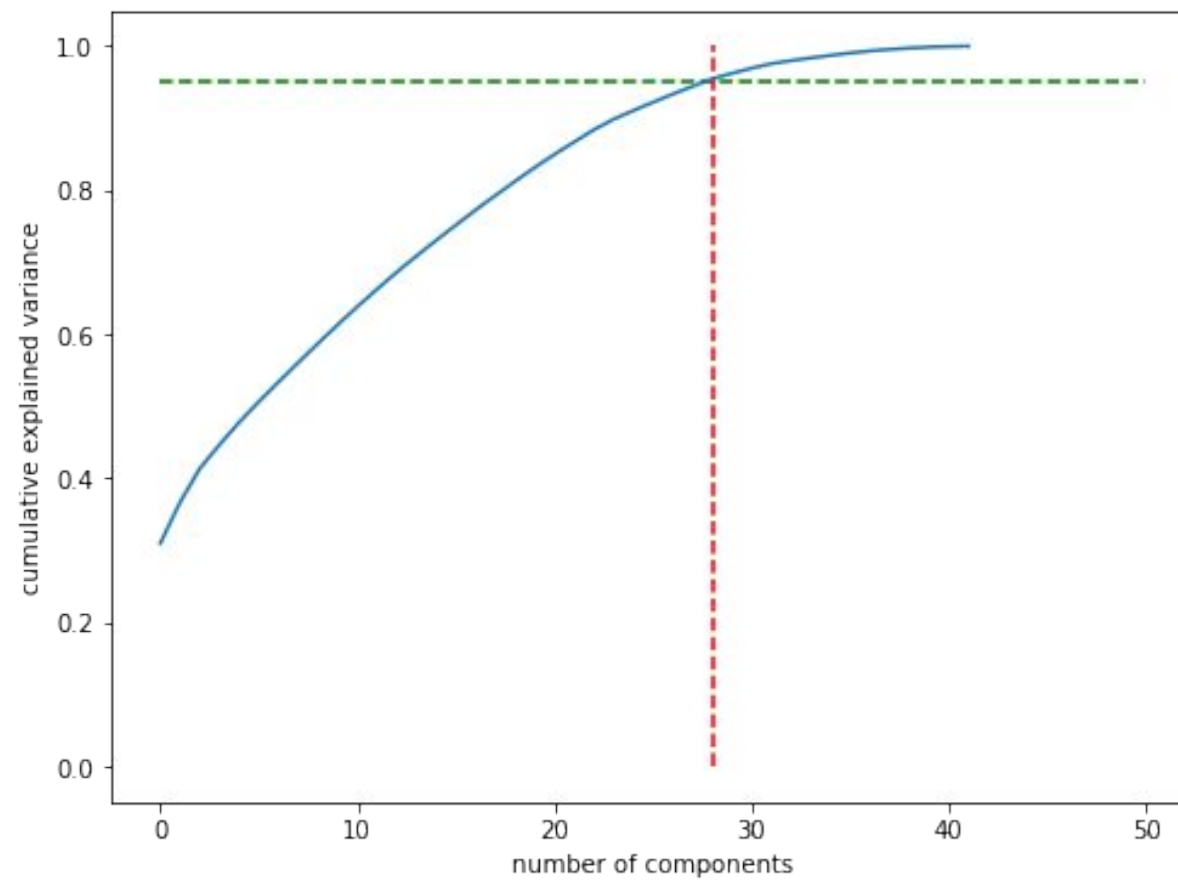
TRAINING MODELS TO FIND ACCURATE PREDICTION

TESTING THREE MODELS FOR ACCURACY

We used the cleaned data to train 3 models

- Logistical Regression

- Random Forrest

- Decision Tree

LOGISTIC REGRESSION
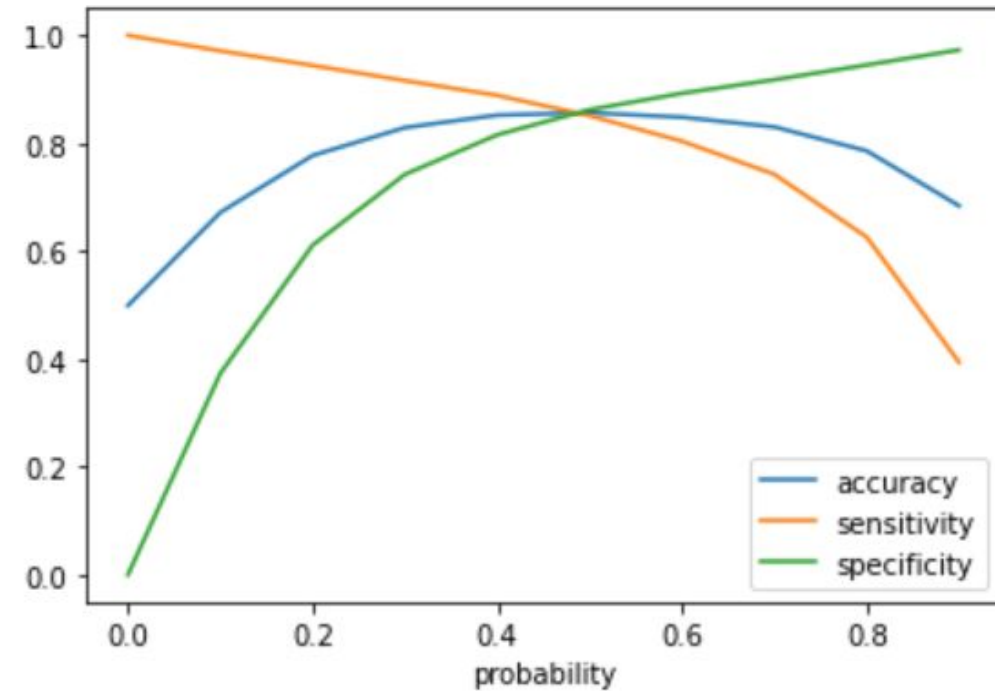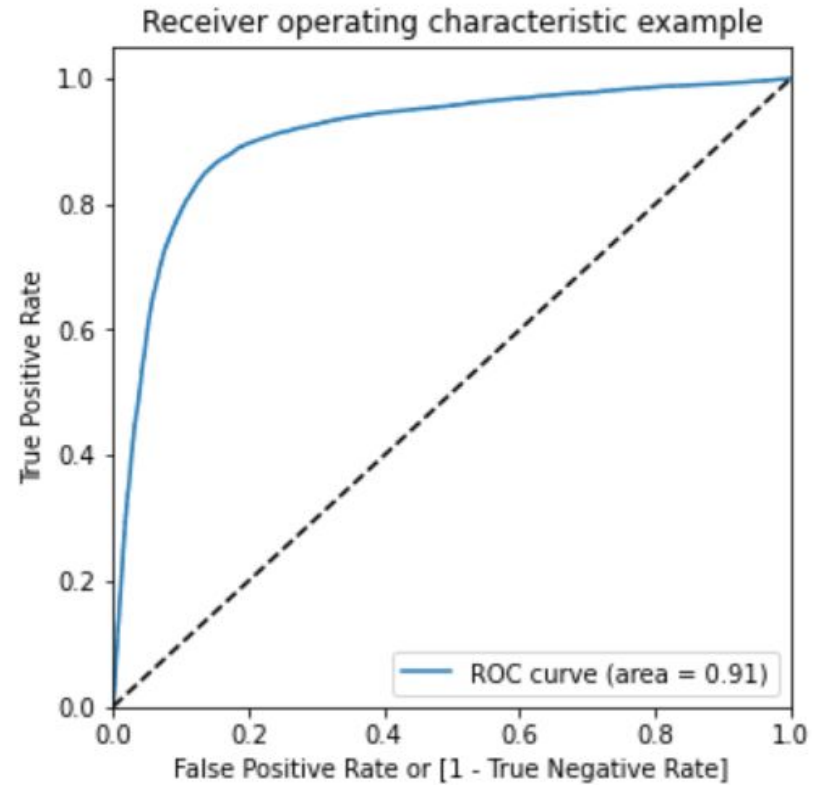
RANDOM FORREST

- Fitting 4 folds for each of 180 candidates, totalling 720 fits

- RandomForestClassifier(max_depth =12, min_samples_leaf=50, min_samples_split=50, n_estimators=70, random_state=34)
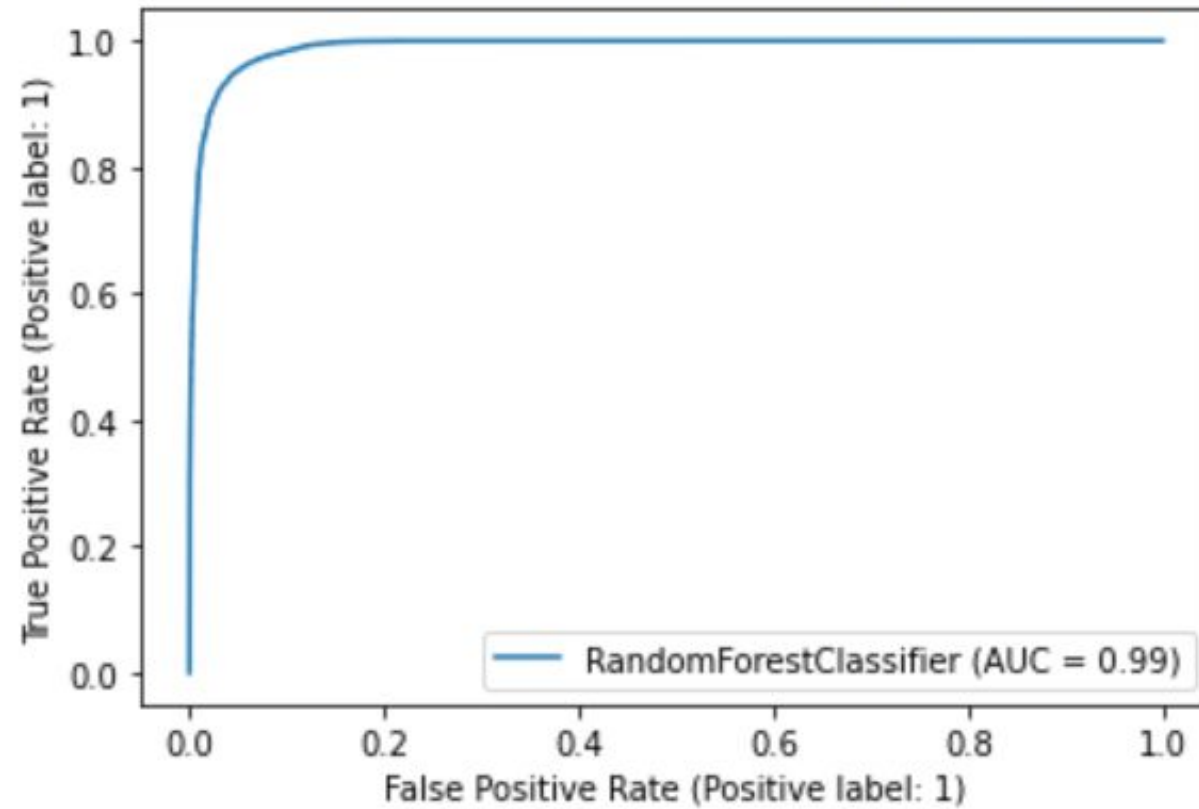
DECISION TREE
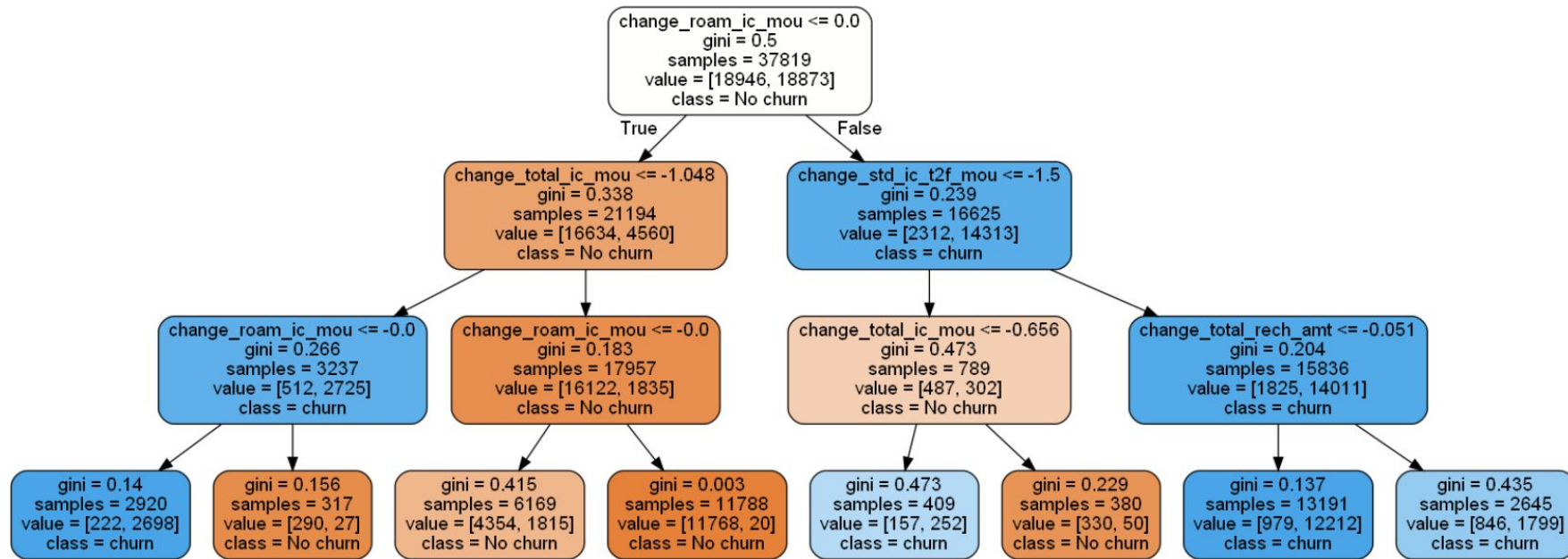
Plotting various hyperparameters to see how accuracy is affected
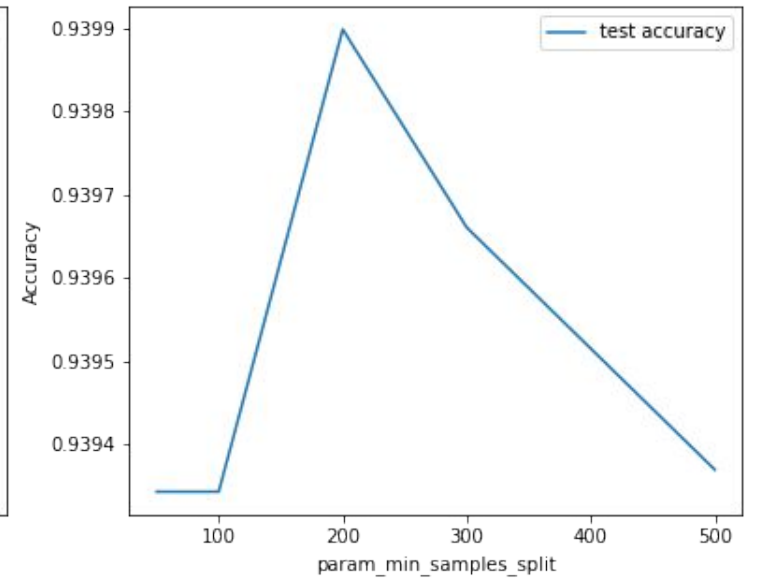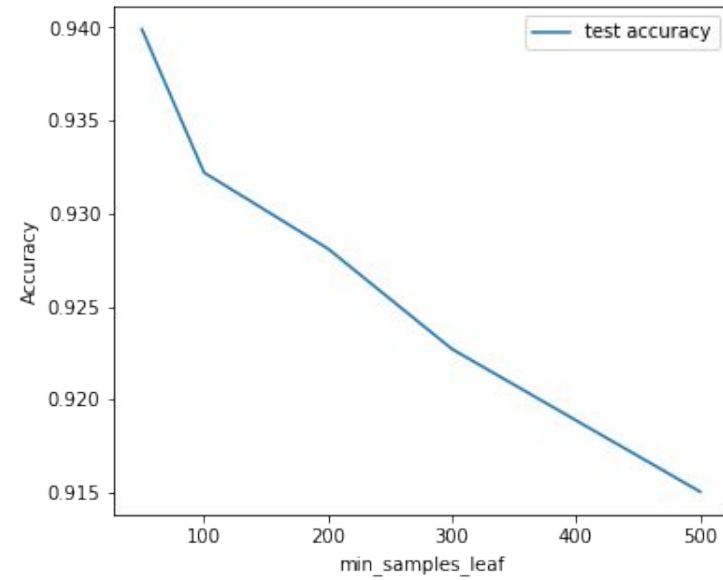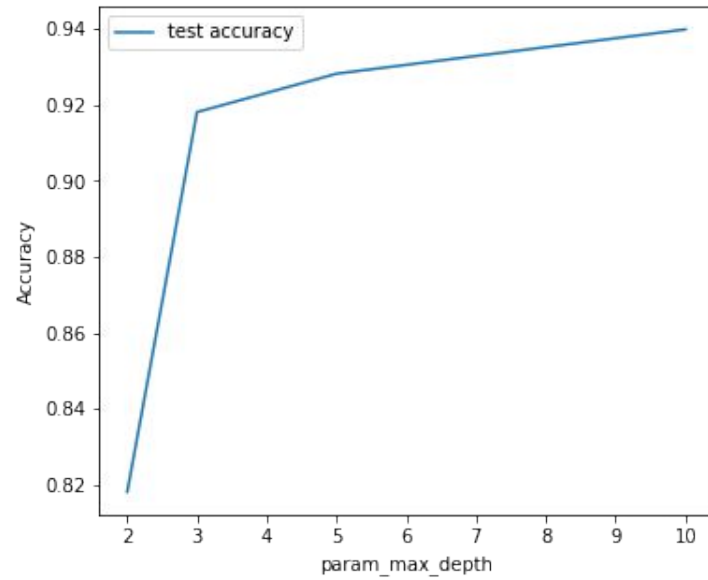
Based on what we observed above, we now tweak the hyperparameters around the values we took for the top model
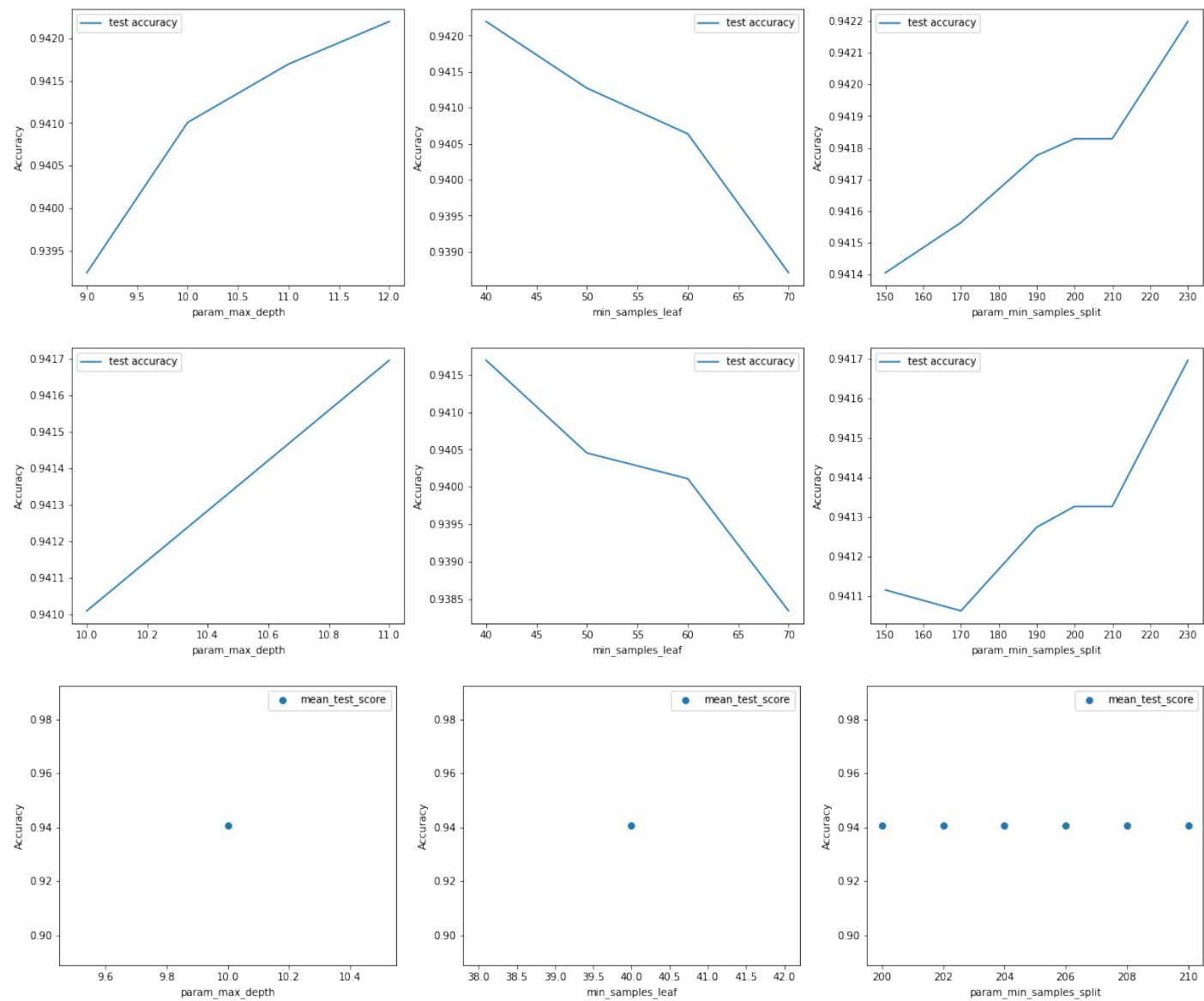
COMPARING ALL THREE MODELS

OBSERVATIONS

- Decision Tree and Random Forest have much better accuracy than Logistic Regression

- Random Forest and Decision Tree have almost the same accuracy

- We choose **Random Forest** because although it has almost the same accuracy as Decision Tree, it is a collection of Decision Trees, so we can trust it better

|	Model	Train Accuracy	Test Accuracy
0	Logistic Regression	0.856527	0.853538
1	Decision Tree	0.947566	0.942563
2	Random Forest	0.949787	0.946018

CONCLUSIONS & SUGGESTIONS

CONCLUSION OF ANALYSIS

- Age on Network (aon) to be considered for a customer to understand their behaviour

- We basically need to understand the change in a customer's service usage in the action phase, (i.e. reduction/addition in recharges, recharge amounts, local, std and roaming calls, internet usage etc) in order to understand their behaviour

- Also, for every user, we need to understand the average revenue they are bringing to the company and if there is a change in it, in the action phase

- Customers will churn out due to some or all of the following reasons:
  - Poor network coverage
  - High call/recharge rates
  - Better offers by the competition
  - Better offers on high value services such as ISD calls
  - Sometimes there are family pack kind of offers, so a user will switch to the provider which has those kind of offers

THANK YOU.