



TELECOM CHURN CASE STUDY

SHYAMLI KUMARI

DEBASISH RATH



PABITRA KUMAR PRADHAN

BUSINESS PROBLEM OVERVIEW

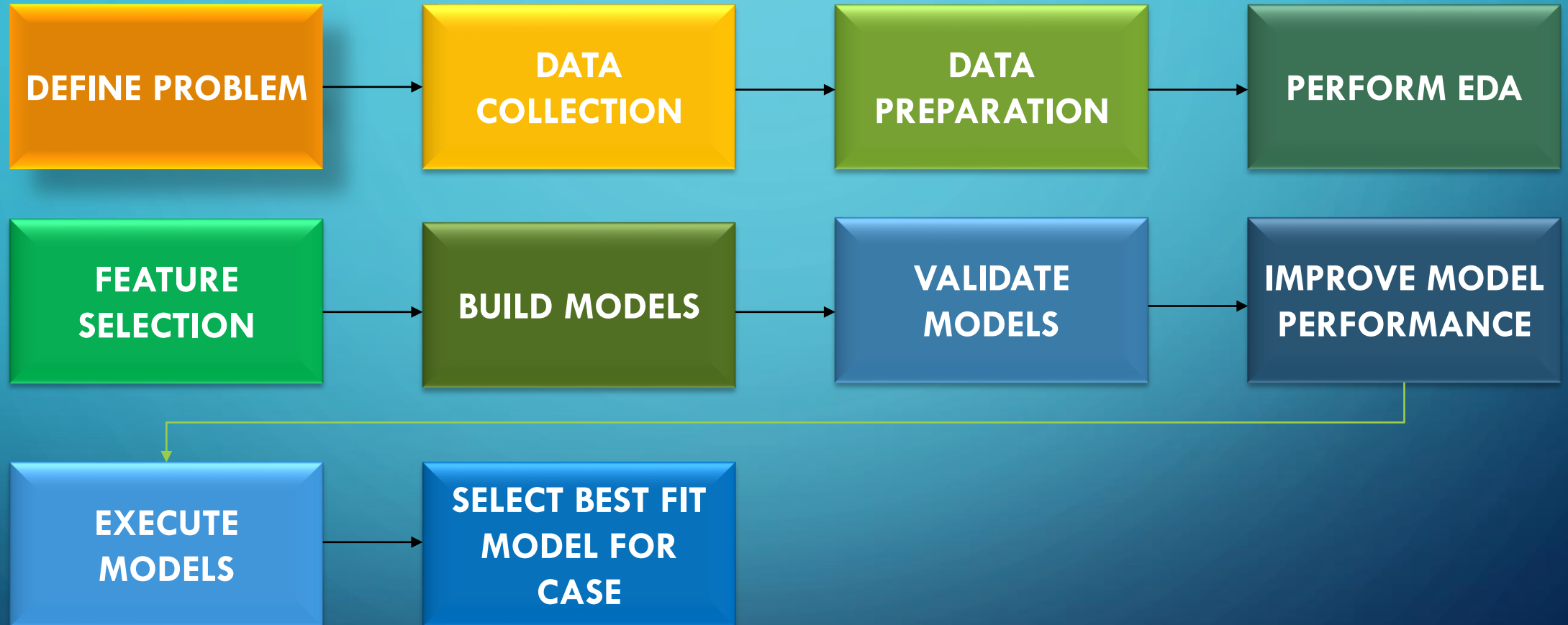
- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more 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.
- In this project, we 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.

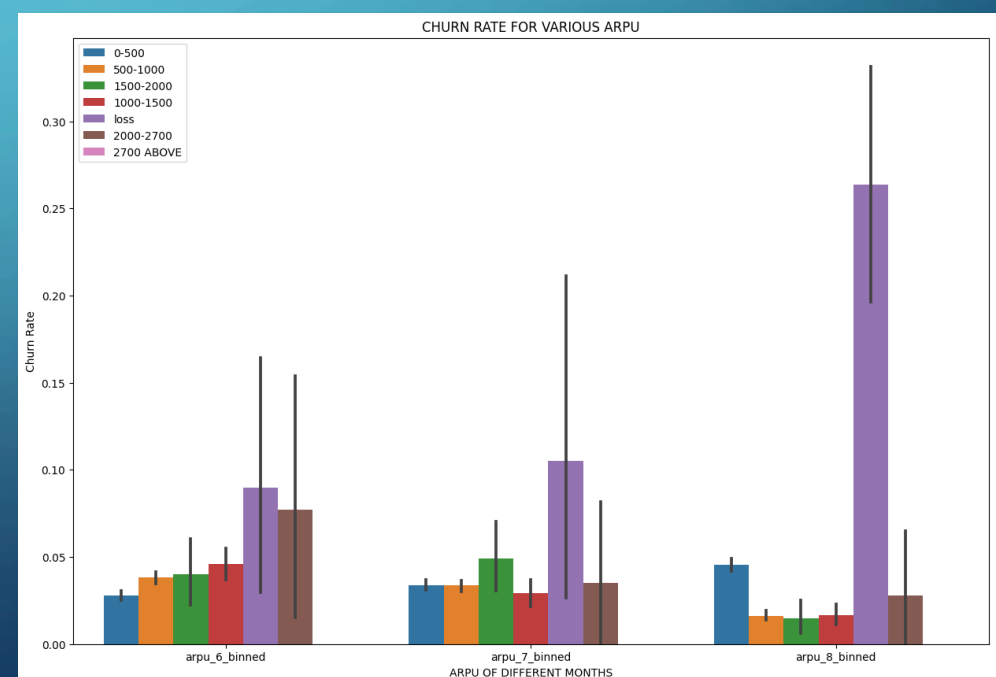
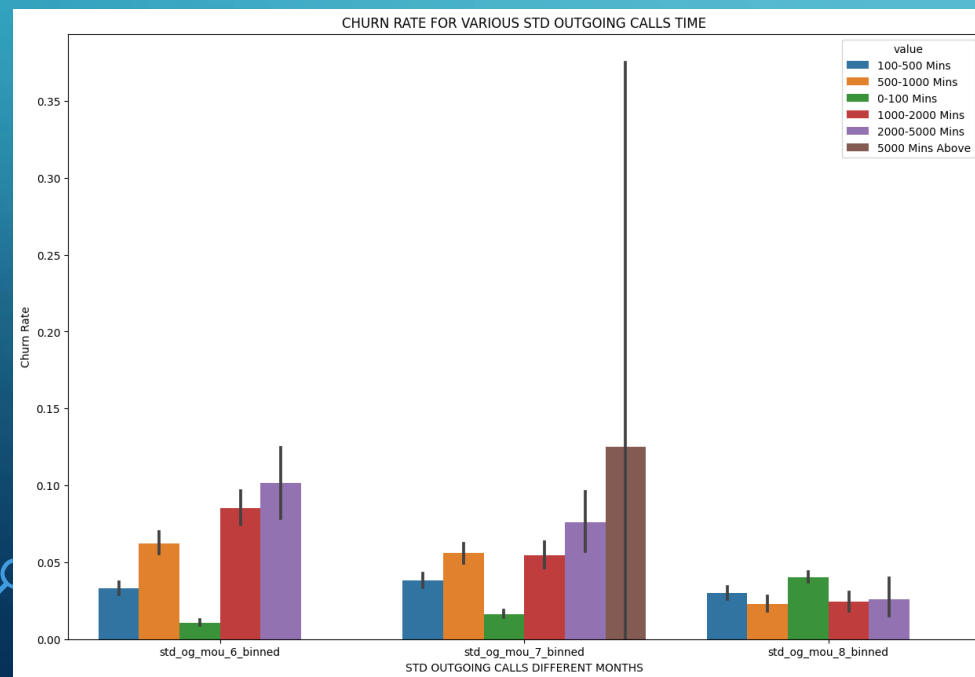
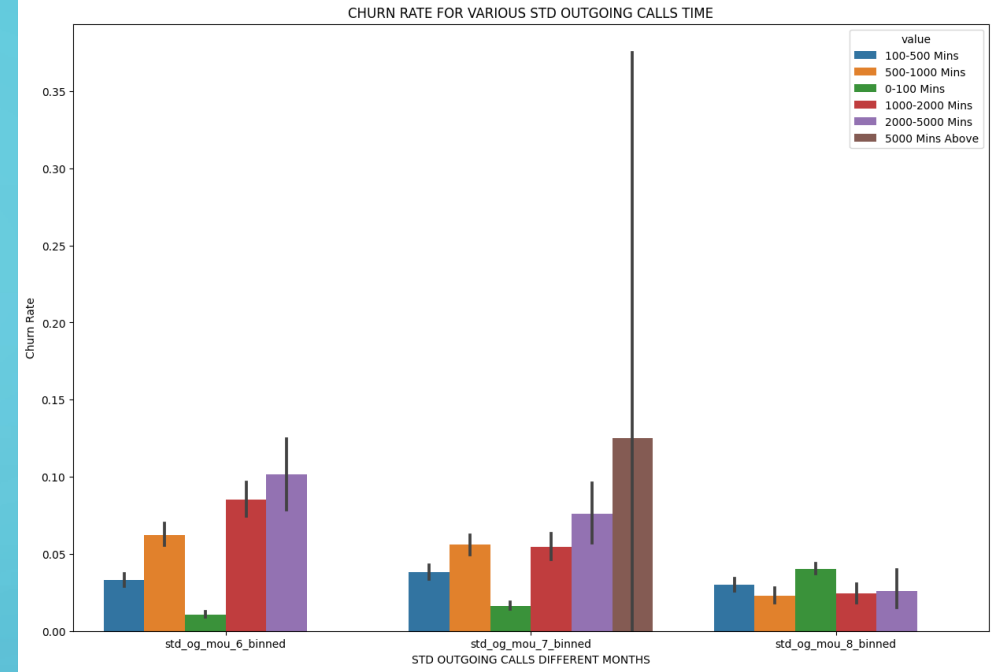
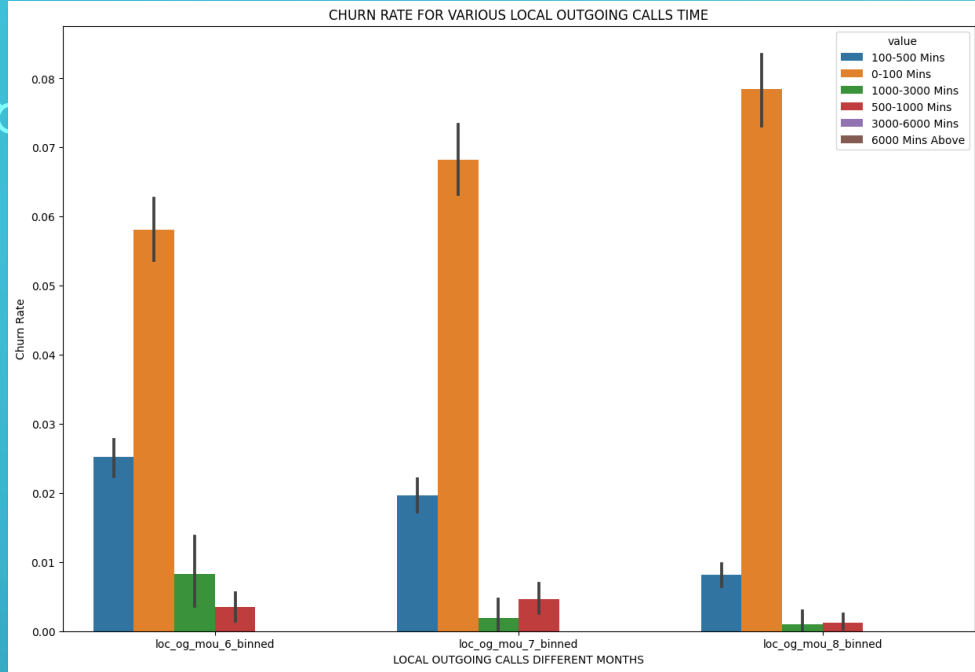


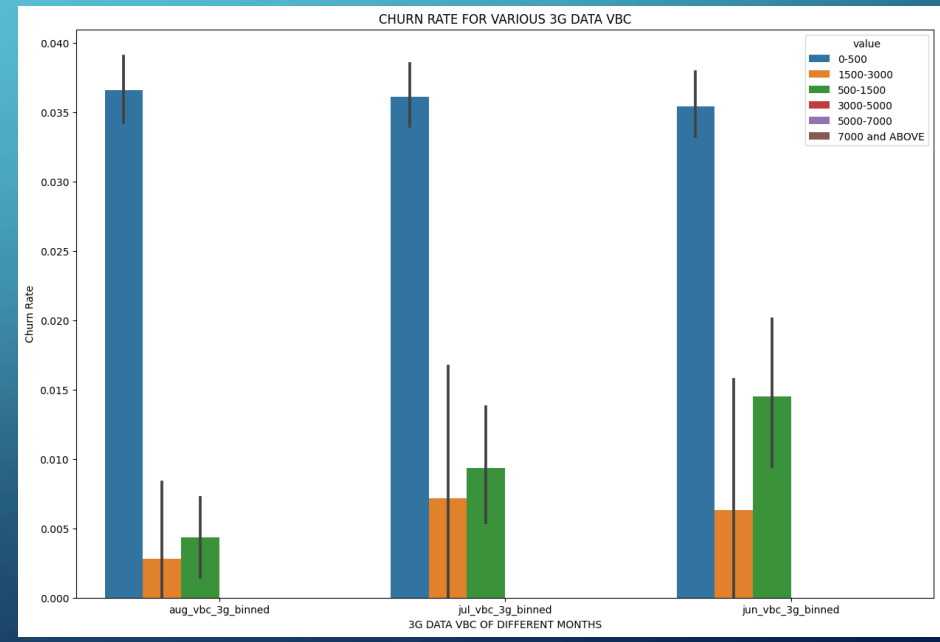
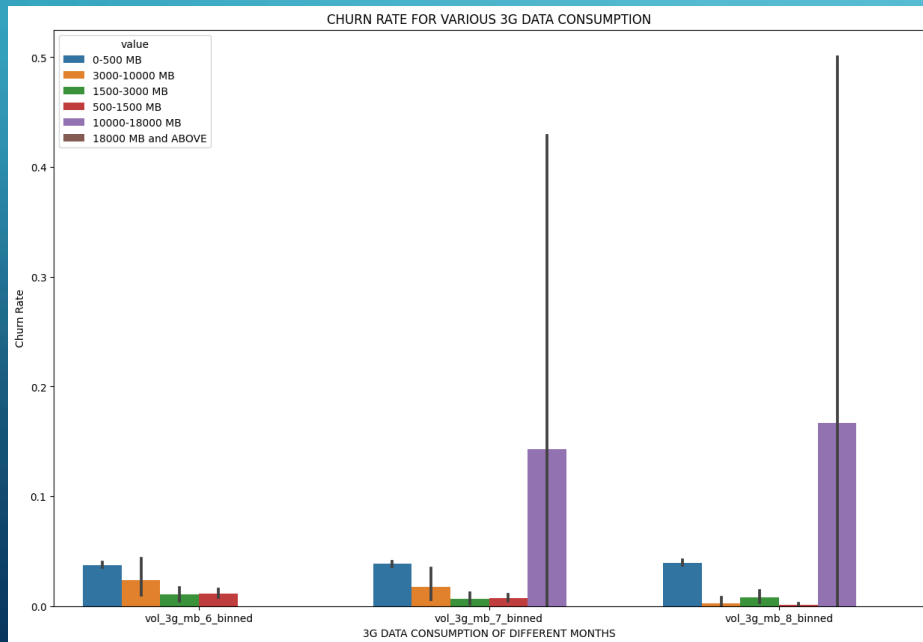
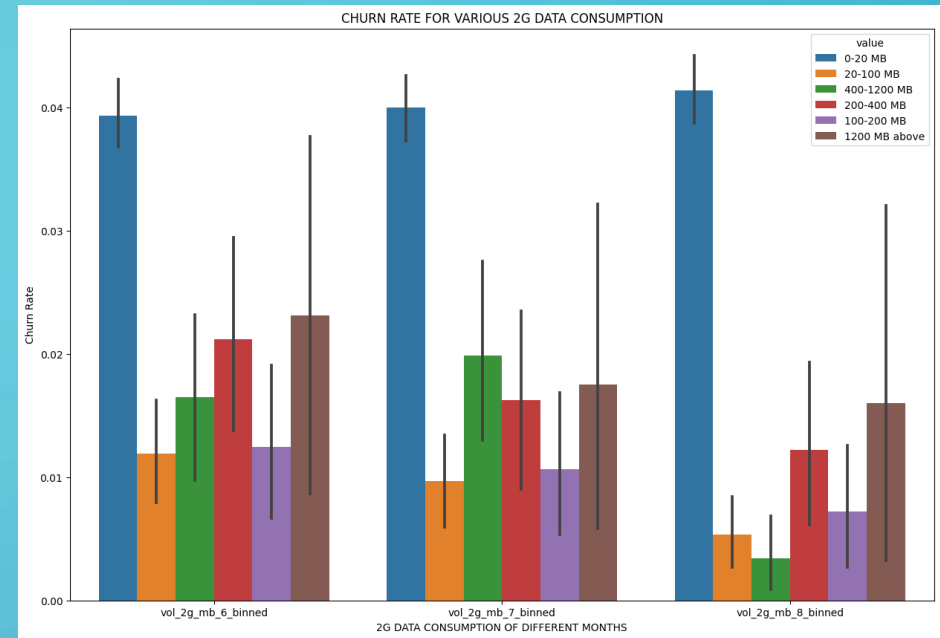
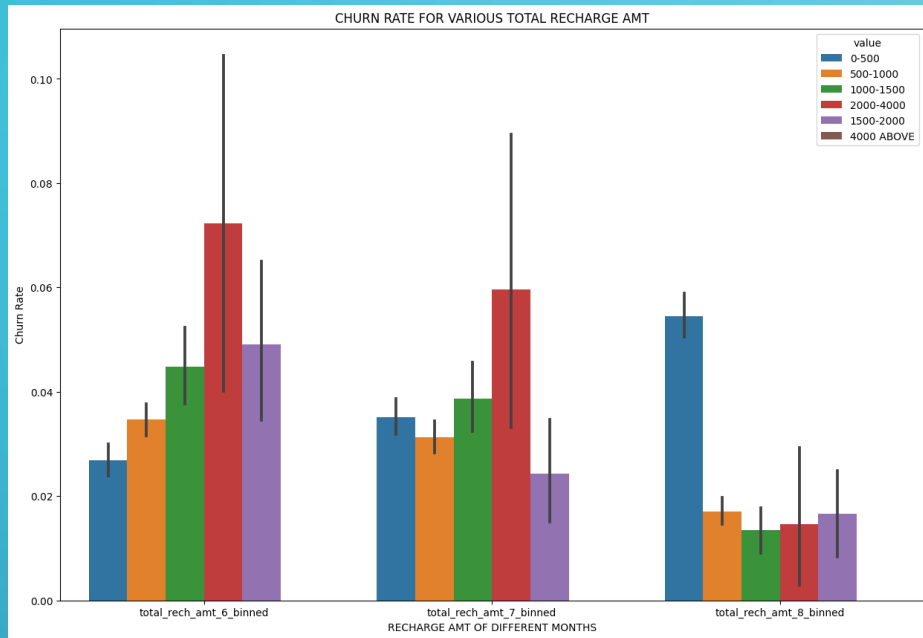
OBJECTIVE

- To predict Customer Churn
 - Highlighting Main Variables or factors influencing Customer Churn.
 - Use Various ML algorithms to build prediction models, evaluate the accuracy and performance of these models.
 - Finding out best model for our business case & providing executive Summary
- 
- 

MODEL BUILDING STEPS

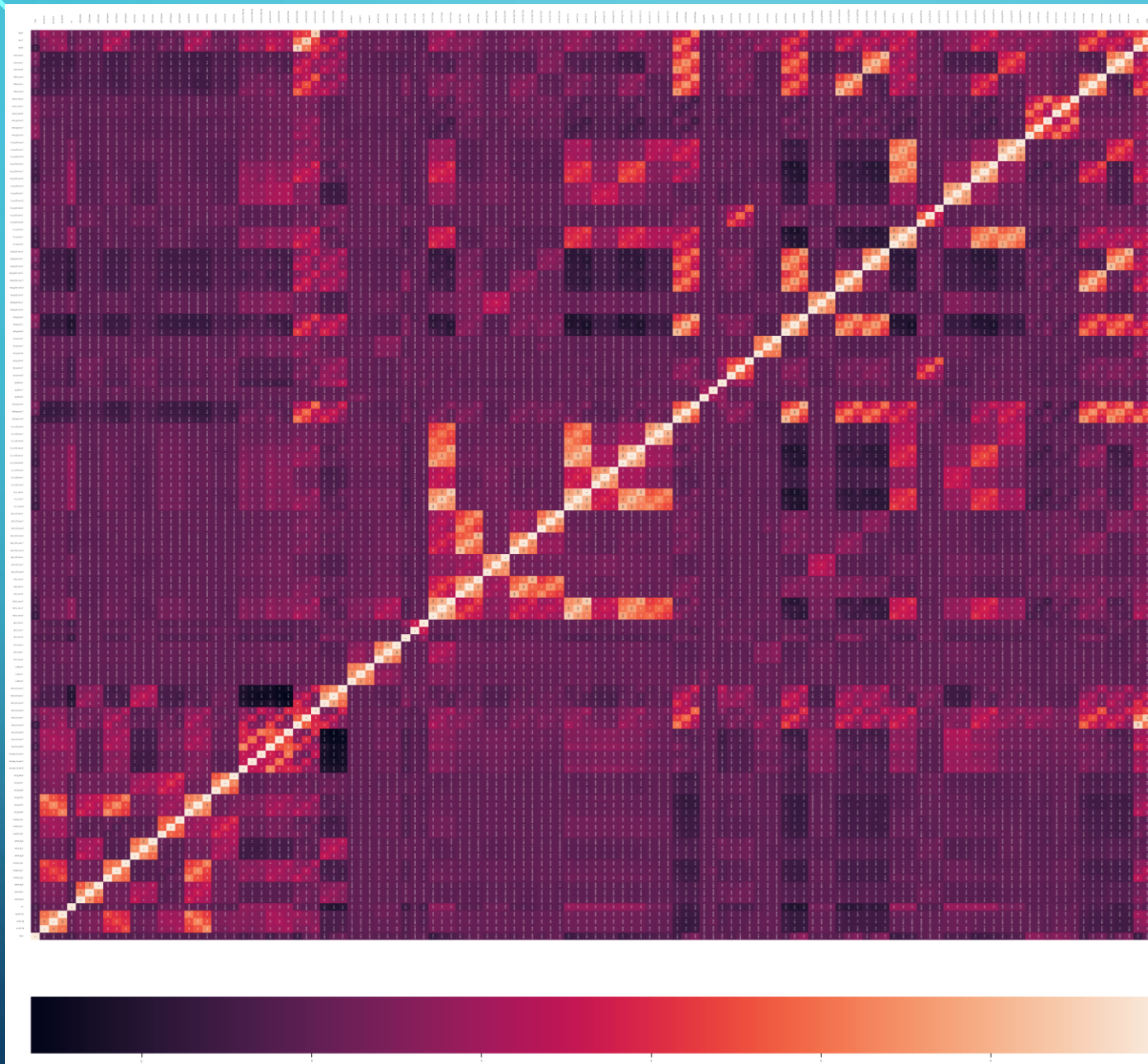






CONCLUSION FROM BIVARIATE ANALYSIS

- Customers using 0-100 Mins LOCAL calls and 10000-18000 MB of 3G data are Highly likely to Churn at 40% average.
- Similarly for Customers using 0-100 Mins STDL calls and 10000-18000 MB of 3G data are Highly likely to Churn at 33% average.
- Customers on whom ARPU is loss and doing STD calls of 2000-2500 Mins are most likely to Churn at average 75%.
- Those from whom ARPU suddenly changed from Loss to 2000-2700 are most likely to Churn.
- Those with ARPU in range of 1500-2000 and using 10000-18000 MBV of 3G data are Most likely to Churn.
- Those who's 3G data consumption increase from 3000-10000 to 10000-18000 MB are most likely to Churn.



DROPPING
HIGHLY
CORRELATED
DUMMY
VARIABLES

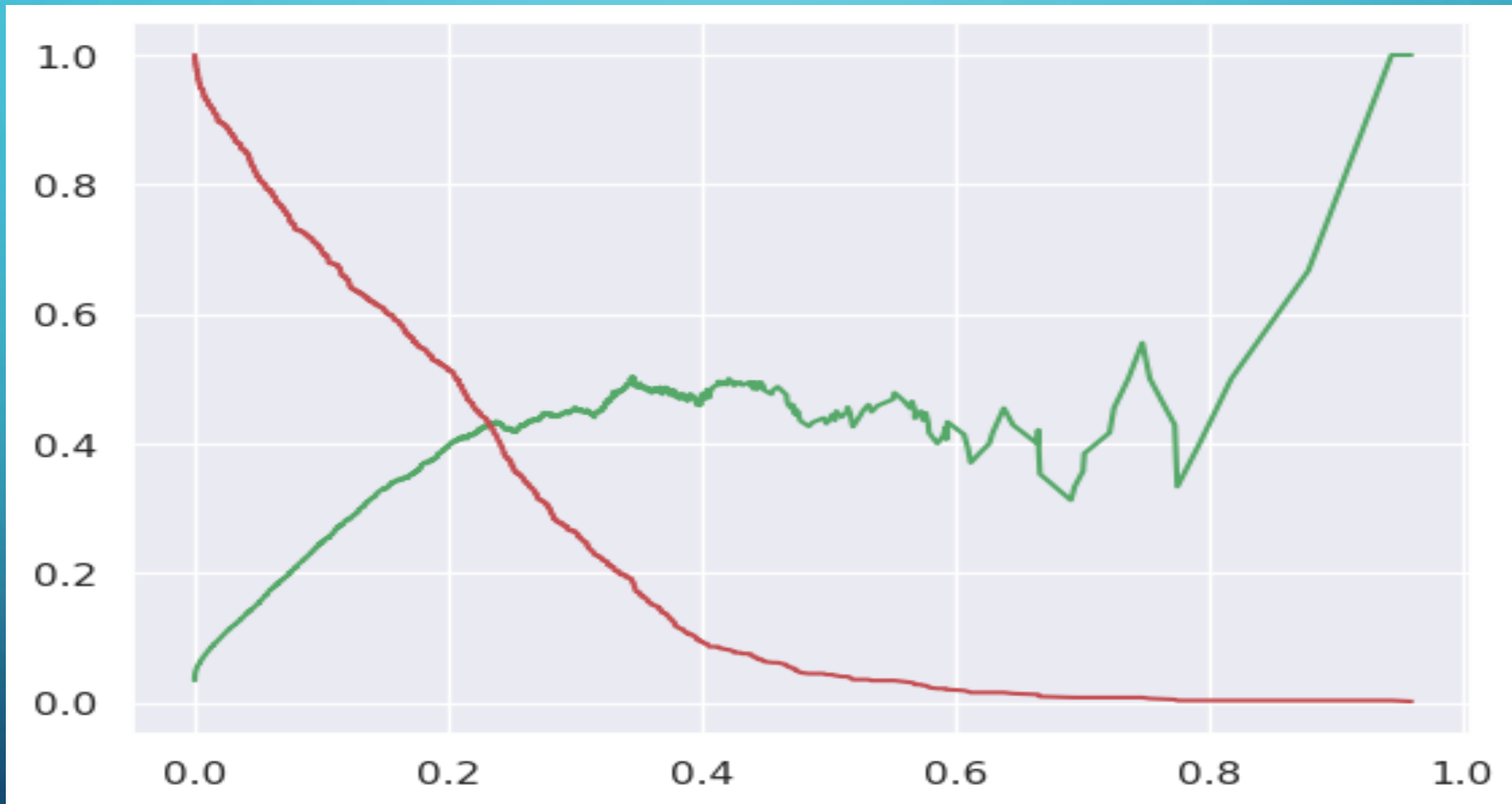
Optimal cutoff probability

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

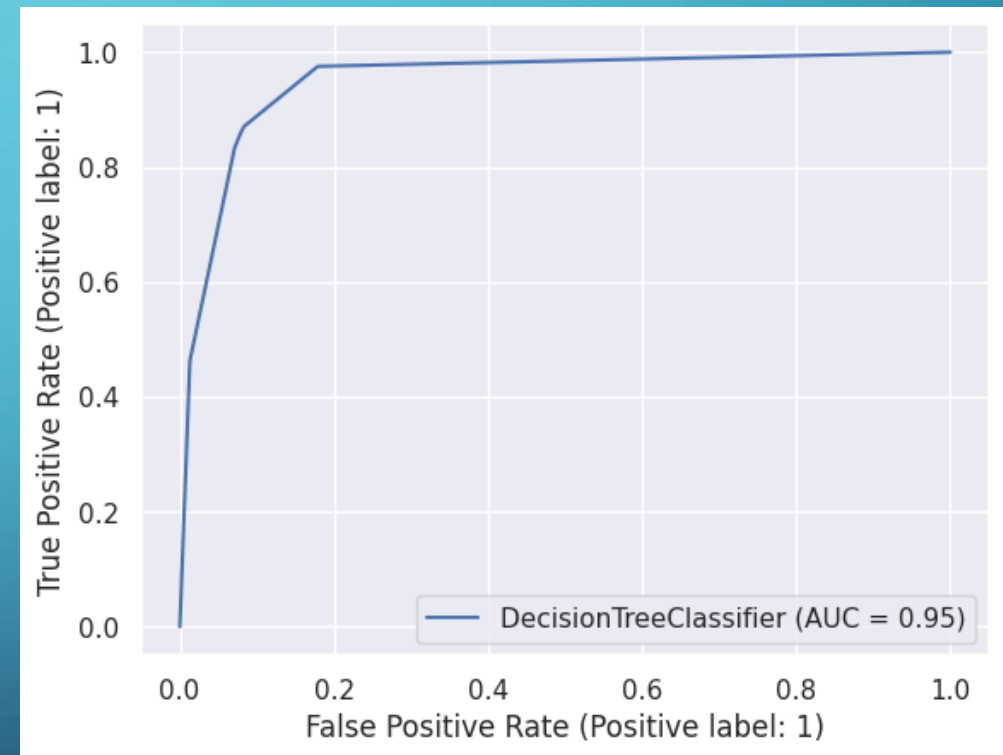
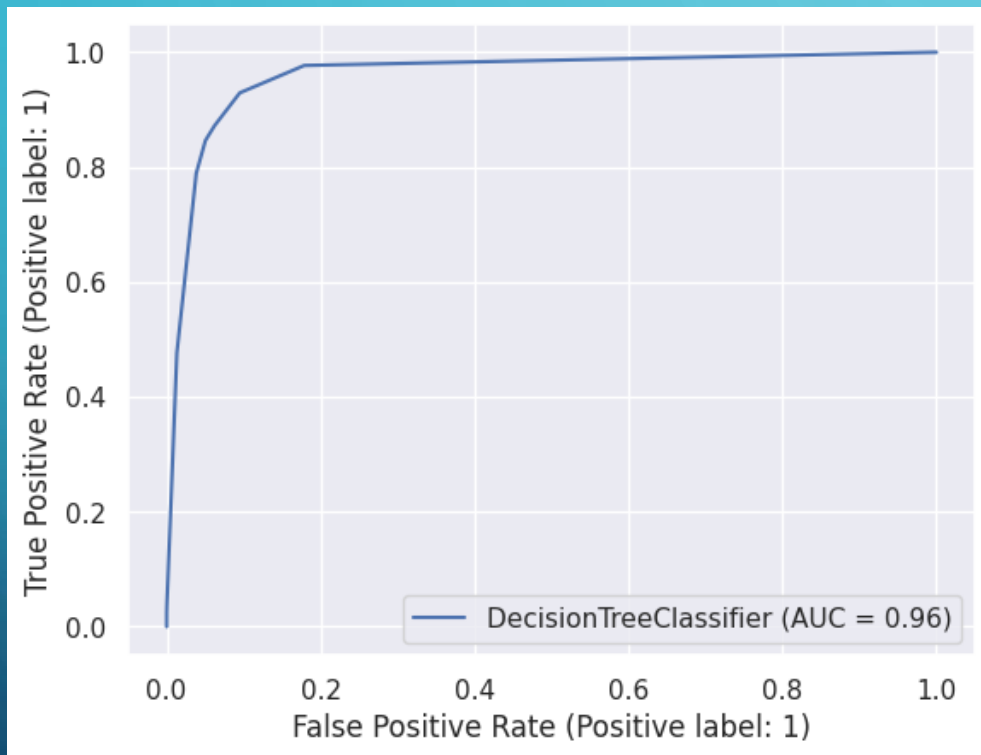


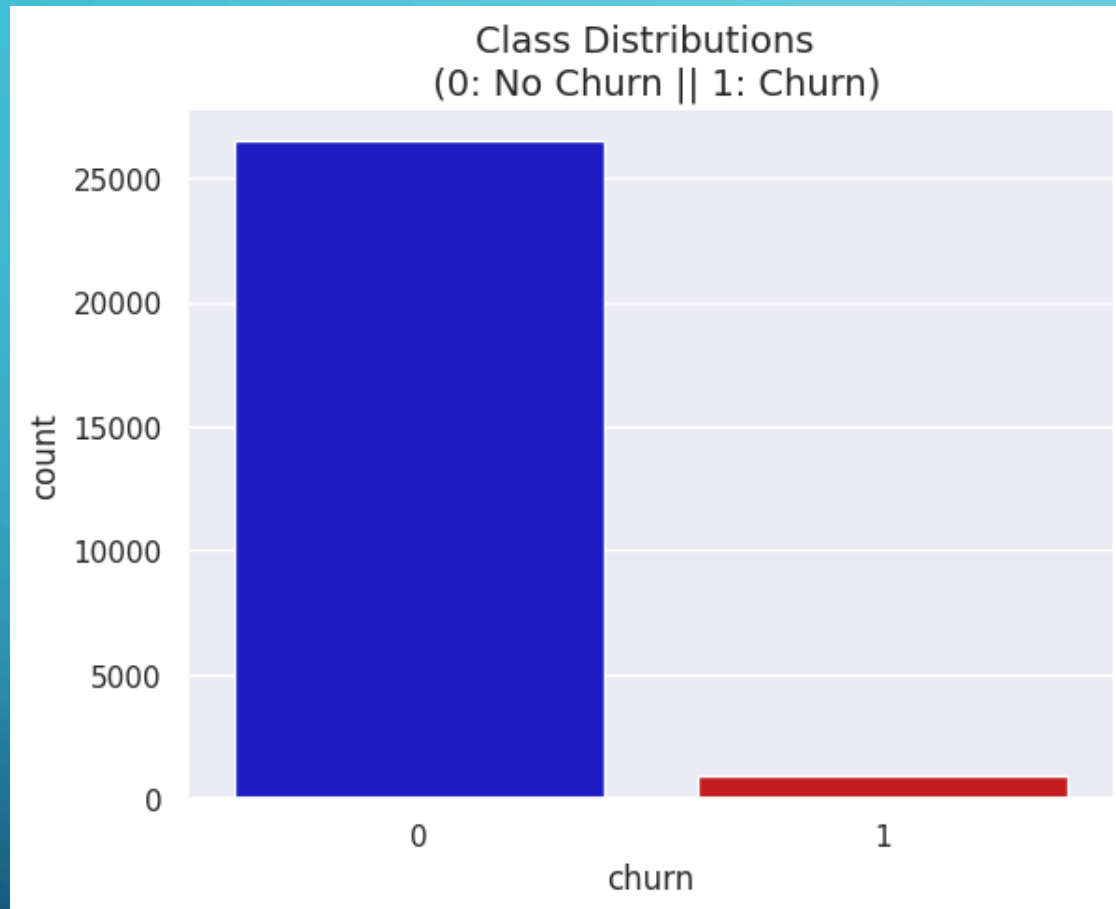
From the curve above, 0.08 is the optimum point to take it as a cutoff probability

PRECISION AND RECALL TRADEOFF



MAKING PREDICTIONS ON THE TEST SET





Class Distribution

```
0    26500  
1      922  
Name: churn, dtype: int64
```

	index	churn
0	0	0.966377
1	1	0.033623

BUSINESS INSIGHT

- Telecom company needs to pay attention to the roaming rates. They need to provide good offers to the customers who are using services from a roaming zone.
- The company needs to focus on the STD and ISD rates. Perhaps, the rates are too high. Provide them with some kind of STD and ISD packages.
- To look into both of the issues stated above, it is desired that the telecom company collects customer query and complaint data and work on their services according to the needs of customers.

MODELS SUMMARY

MODELS MADE WITHOUT HANDLING CLASS IMBALANCE

Method	ACCURACY	SENSITIVITY	POSITIVE PREDICTED VALUE
Gen_Logistic_Regression	0.897317	0.730769	0.209159
Gen_Logistic_Regression_0.22	0.897317	0.518382	0.413490
Gen_Logistic_Regression_0.08	0.897317	0.518382	0.413490
GEN_DECISION_TREE_UNTUNED	0.968883	0.477941	0.532787
GEN_DECISION_TREE_TUNED	0.969369	0.466912	0.542735
GEN_RANDOM_FOREST_UNTUNED	0.967424	0.058824	0.571429
GEN_RANDOM_FOREST_TUNED	0.971314	0.371324	0.608434

MODELS SUMMARY

LOGISTIC REGRESSION MODELS MADE FOR DIFFERENT HANDLING OF CLASS IMBALANCE

Method	ACCURACY	SENSITIVITY	POSITIVE PREDICTED VALUE
IMB_Logistic_Regression	0.966330	0.000000	NaN
IMB_US_Logistic_Regression	0.966330	0.000000	NaN
IMB_TOMEK_Logistic_Regression	0.966330	0.000000	NaN
IMB_OS_Logistic_Regression	0.966330	0.000000	NaN
IMB_SMOTE_Logistic_Regression	0.966330	0.000000	NaN
IMB_ADASYN_Logistic_Regression	0.033670	1.000000	0.033670
IMB_TOMEK_SMOTE_Logistic_Regression	0.966330	0.000000	NaN

MODELS SUMMARY

DECISION TREE MODELS MADE FOR DIFFERENT HANDLING OF CLASS IMBALANCE

Method	ACCURACY	SENSITIVITY	POSITIVE PREDICTED VALUE
IMB_Decision_Tree	0.955877	0.364621	0.350694
IMB_US_Decision_Tree	0.863620	0.866426	0.181132
IMB_Tomek_Decision_Tree	0.958430	0.353791	0.375479
IMB_OS_Decision_Tree	0.960010	0.415162	0.407801
IMB_SMOTE_Decision_Tree	0.938009	0.512635	0.274662
IMB_ADASYN_Decision_Tree	0.941412	0.581227	0.305503
IMB_SMOTE_TOMEK_Decision_Tree	0.941777	0.541516	0.298805

MODELS SUMMARY

RANDOM FOREST MODELS MADE FOR DIFFERENT HANDLING OF CLASS IMBALANCE

Method	ACCURACY	SENSITIVITY	POSITIVE PREDICTED VALUE
IMB_Random_Forest	0.968032	0.267148	0.552239
IMB_US_Random_Forest	0.904826	0.945848	0.254369
IMB_TOMEK_Random_Forest	0.968761	0.303249	0.567568
IMB_OS_Random_Forest	0.969004	0.444043	0.549107
IMB_SMOTE_Random_Forest	0.953446	0.599278	0.378995
IMB_ADASYN_Random_Forest	0.953203	0.613718	0.379464
IMB_SMOTE+TOMEK_Random_Forest	0.952960	0.595668	0.375000

CONCLUSION OF REGRESSION MODELS

- Of all the Models we evaluated, Tuned Random Forest gave the best accuracy at 97.13%.
- After handling Class Imbalance we observe that Logistic Regression made from ADASYN sampling gives highest sensitivity 100%.
- However due to extremely low Accuracy of such model, it's better to consider Random Forest derived from Random Under Sampling of data at 94%.
- As per POSITIVE PREDICTED VALUE, Random Forest with parameters `n_estimators=10`, `max_depth=4`, `max_features=5`, `random_state=100` is the best model at 57.14%

*Thank
You!*