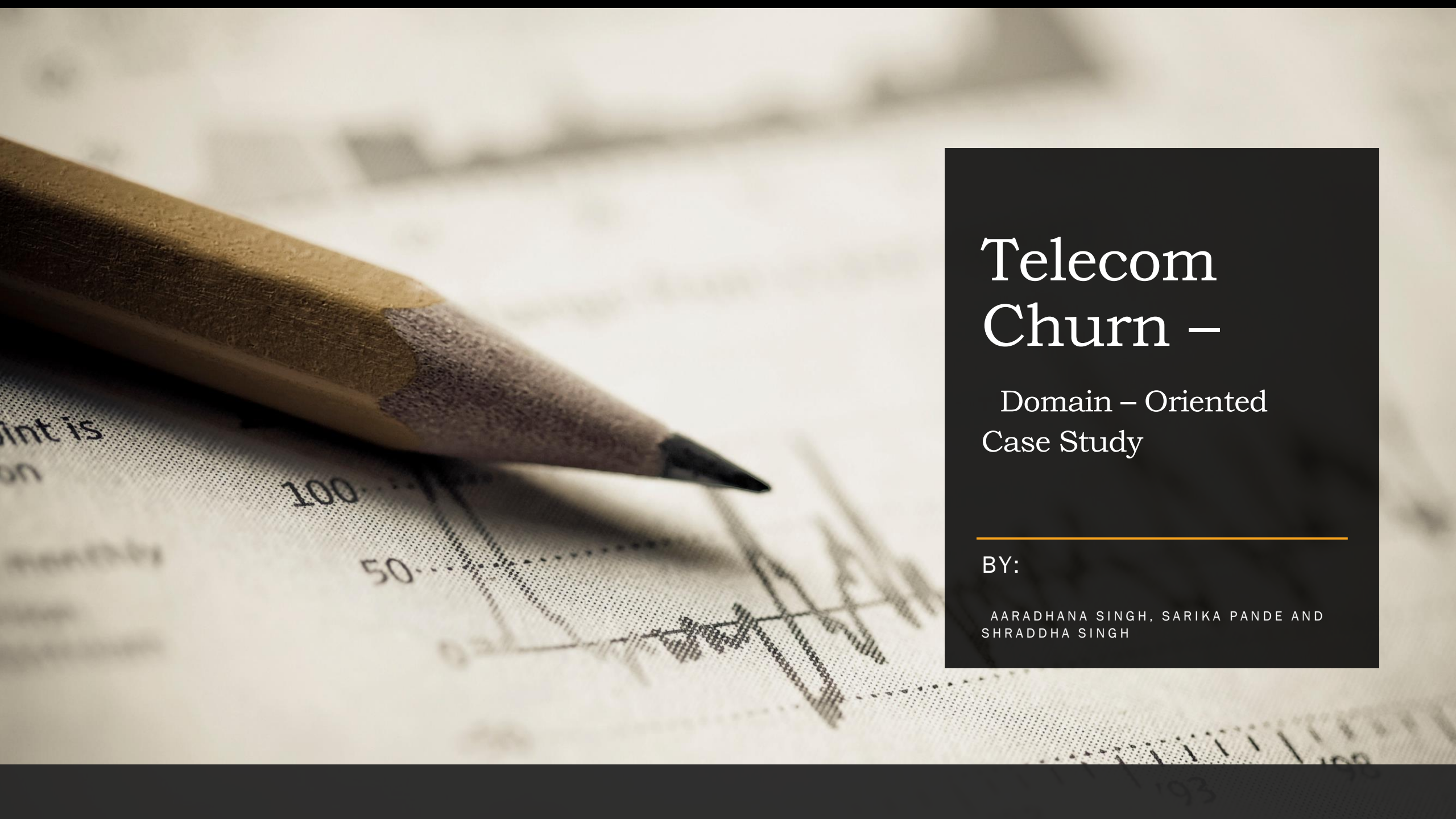


# Telecom Churn –

Domain – Oriented Case Study


BY:

AARADHANA SINGH, SARIKA PANDE AND SHRADDHA SINGH

# 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.

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

# BUSINESS GOALS:

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

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

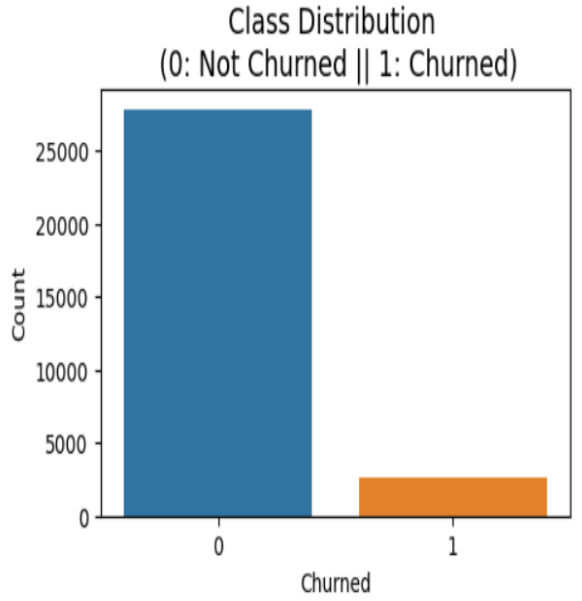
# OVERALL APPROACH

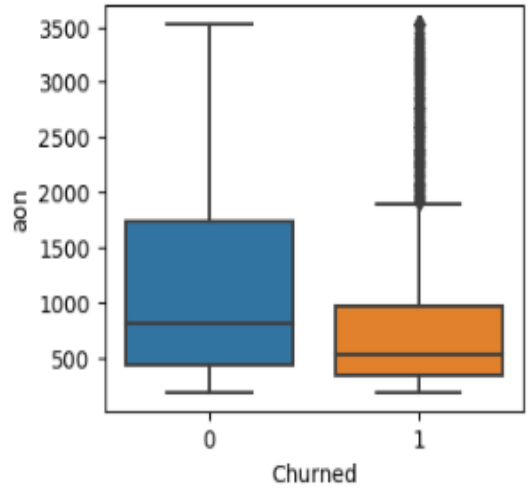
1. DATA CLEANING AND IMPUTING MISSING VALUES
2. FEATURE ENGINEERING AND HANDLING OUTLIERS
3. EXPLORATORY DATA ANALYSIS
4. FEATURE SCALING AND CLASS BALANCING
5. LOGISTIC REGRESSION MODEL BUILDING
6. DECISION TREE WITH GRID SEARCH CV
7. RANDOMFOREST WITH GRID SERACH CV
8. MODEL EVALUATION: ACCURACY, SENSITIVITY AND ROC(AUC)
9. TOP FEATURES THAT INFLUENCE CHURN RATE
10. CONCLUSION AND RECOMMENDATION

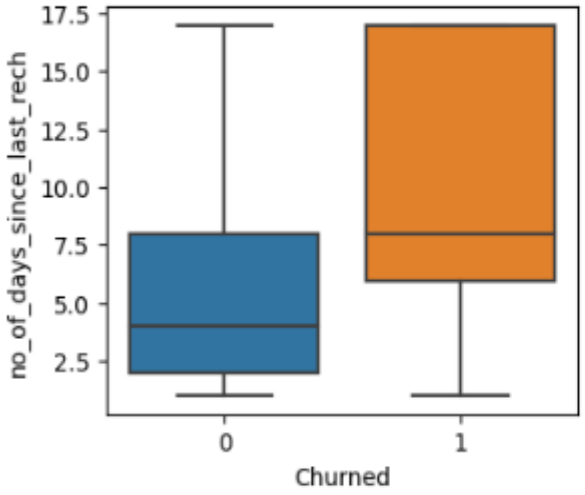
# EXPLORATORY DATA ANALYSIS

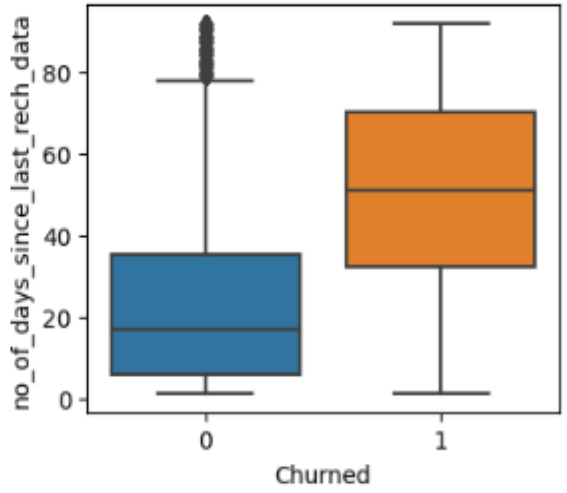
We have around 8% Churn rate data with us. Which is highly imbalance to work with.

- The Customers with “aon” (age on network) < 800 are more likely to churn than others.
- The ”no_of_days_since_lat_rech” and ”no_of_days_since_lat_rech_data” if 50 days or more are most likely to churn.



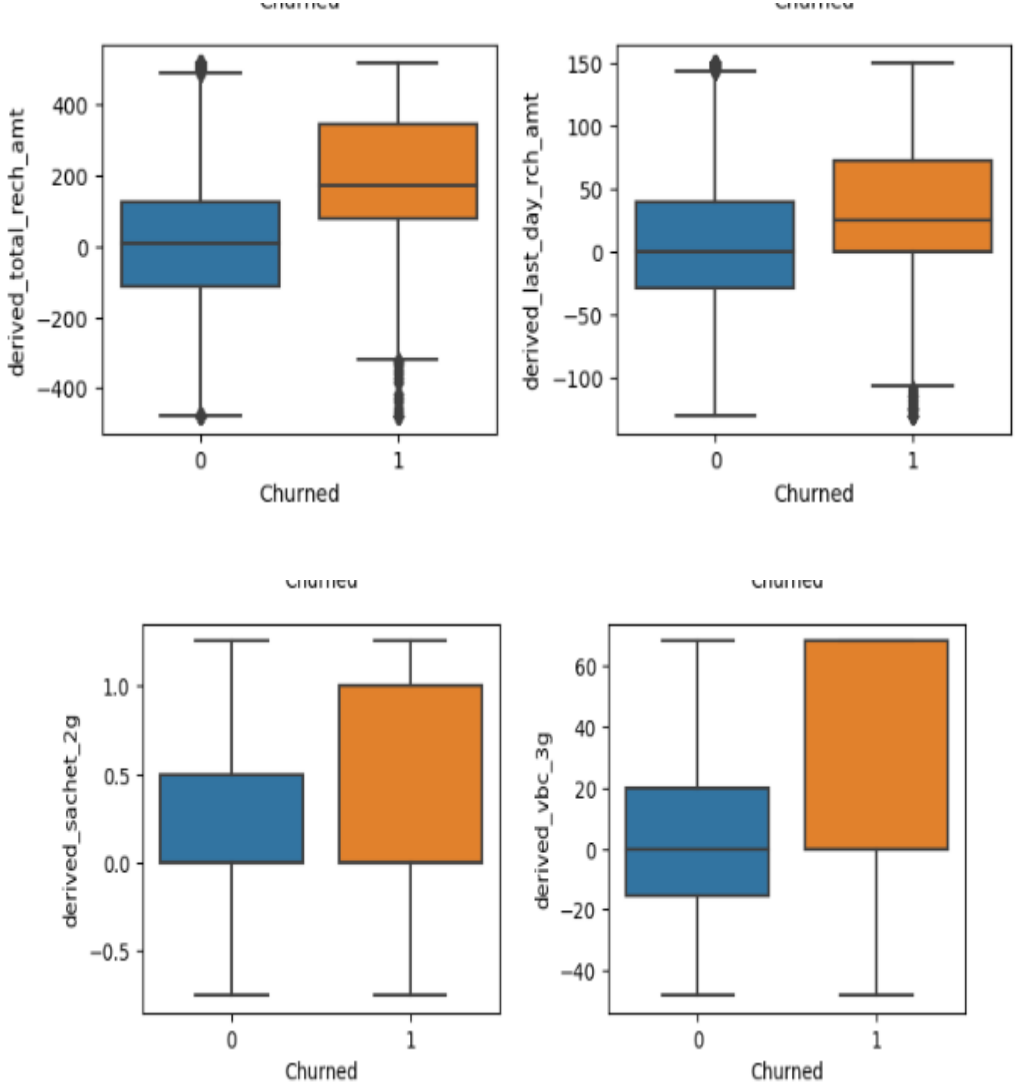



# EXPLORATORY DATA ANALYSIS

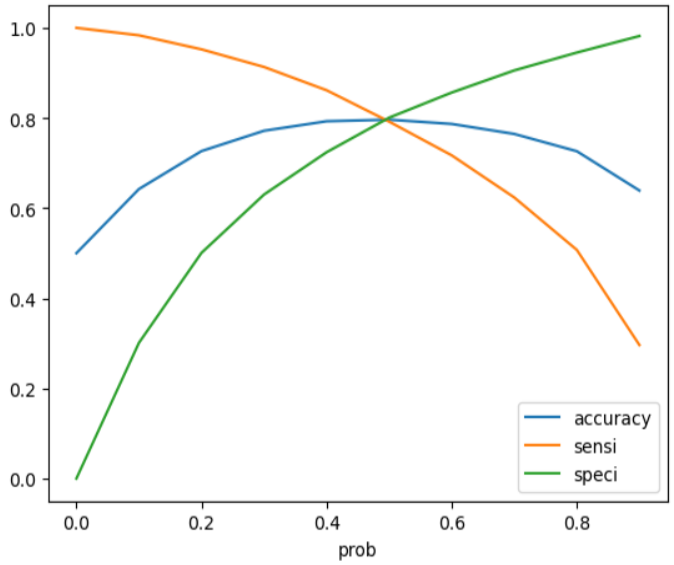
- Customers with high positive values in ‘derived_last_day_rch_amt’ and ‘derived_total_rech_amt’ are going to churn. Sales team may focus on the customers whose derived+last_day_rch_amt > 50.

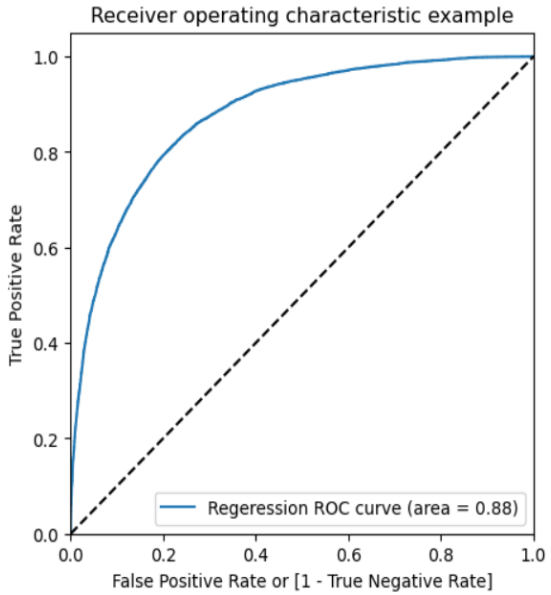
- High positive values of ‘derived_vol_2g_mb’ means for churned customers means huge drop in 2g data usage in action phase for customer who churned. Telecom company may need to revise its 2g/3g data plans in order to retain customers.

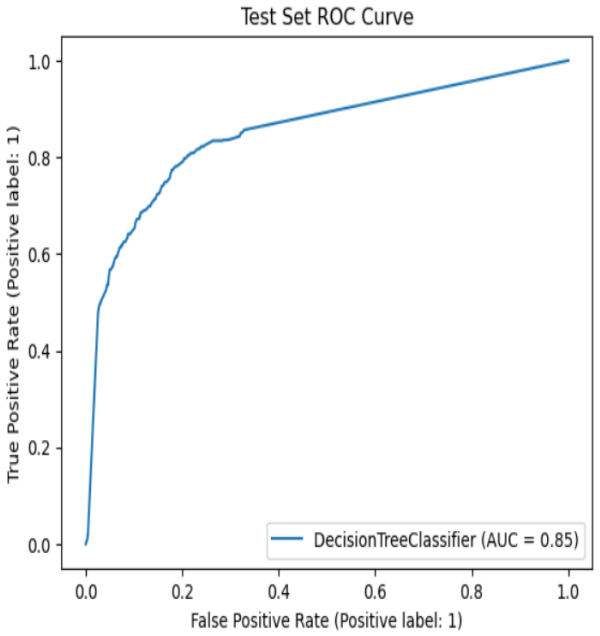


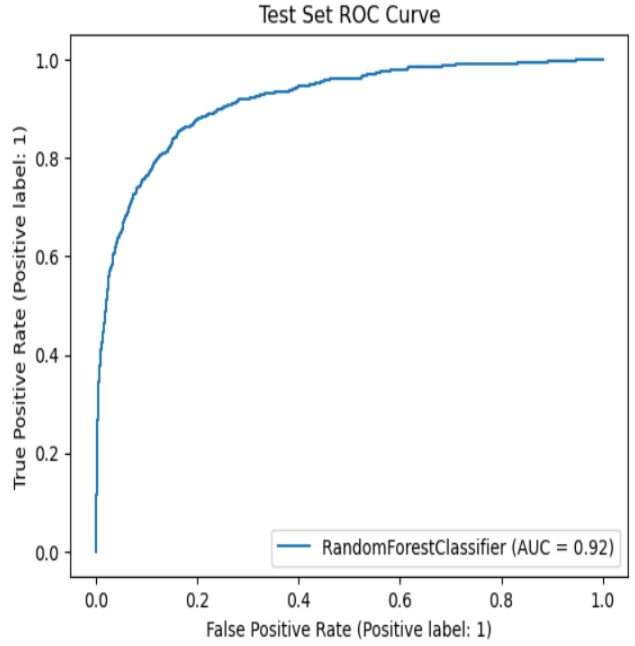
# MODEL BUILDING

- Splitting the data into TEST and TRAIN sets
- For logistic Regression, we used SMOTE to do oversampling of churned cases
- Found the optimal cutoff of 0.5
- Using RFE to choose Top 20 Variables
- Build Model by removing variables whose p-value >.05 and VIF>5
- Used Hyper-Parameters to tune the random Forest and decision tree grid search.
- Plotted ROC curve and visualize AUC



# MODEL EVALUATION AND PREDICTION

Our Logistic Regression gives high True Positive Rate for Test Set.

On other hand, Random Forest gives high accuracy ~92% on Test set

| | Accuracy | Sensitivity |
|---|---|---|
| Logistic Regression(Train) | 0.796 | 0.792 |
| Logistic Regression(Test) | 0.792 | 0.766 |
| Decision Tree(Train) | 0.825 | 0.730 |
| Decision Tree(Test) | 0.900 | 0.642 |
| Random Forest(Train) | 0.940 | 0.931 |
| Random Forest(Test) | 0.923 | 0.662 |

| features | correlation with Target |
|---|---|
| aon | -0.770604 |
| no_of_days_since_last_rech | 0.487370 |
| no_of_days_since_last_rech_data | 0.672177 |
| derived_offnet_mou | -0.245340 |
| derived_loc_og_t2m_mou | 0.285041 |
| derived_loc_og_t2f_mou | -0.187603 |
| derived_loc_og_t2c_mou | -0.135313 |
| derived_loc_og_mou | -0.452352 |
| derived_total_og_mou | 0.642899 |
| derived_loc_ic_t2t_mou | 0.088469 |
| derived_std_ic_mou | 0.064105 |
| derived_total_ic_mou | 0.391171 |
| derived_total_rech_amt | 0.194021 |
| derived_last_day_rch_amt | 0.356739 |
| derived_av_rech_amt_data | -0.161831 |
| derived_vol_2g_mb | 0.144805 |
| derived_sachet_2g | 0.226814 |
| derived_vbc_3g | 0.191138 |

# CONCLUSION

Since reducing Customer churns with attractive offers is less expensive than making ne customers. Sensitivity score is considered for the building the final predictor model.

The final model logistic could predict 76% of the churned customers correctly out of all the churned customers. (based on our test set sensitivity score)

Accuracy score .79 means 79% of the predictions made by the final model are correct out of all the predictions on the test set.

Thank you