Santander Customer Satisfaction

The Lazy Learners (Ratna Sambhav, Aashruti Agrawal, Joyson Chacko George, Binit Sengupta)

Motivation and About the Project

From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around. What's more, unhappy customers rarely voice their dissatisfaction before leaving.

In this project, main goal is to see which set of feature values decide whether a customer is satisfied or dissatisfied. Knowing those feature values, the Santander bank can modify their policies or regulations, to decrease further the number of customers dissatisfied with their services.

Data and Labels

Dataset contains ***371*** features out of which 1 column has index numbers, 369 has information about the customers which can be used as predictors and 1 column has the target variable. Feature names have been anonymized to keep the privacy of choice and information about the customers.

Our response variable is target, having two class labels, where 0 represents satisfied customers and 1, dissatisfied customers.

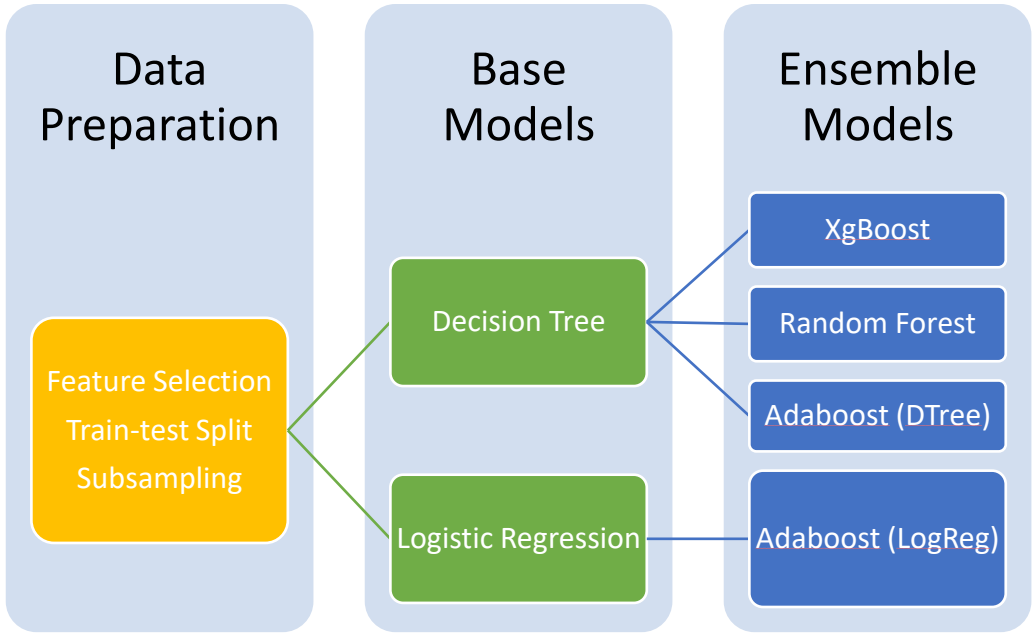
Unique features of our dataset are:

1. **Large feature set**: Presence of large number of features makes the possibility of unnecessary features more certain, and this made us to focus more on the feature selection.
2. **High Unbalance in the response variable**: Since unhappy customers rarely voice their dissatisfaction before leaving, data is highly skewed towards the details of customers who are already satisfied. Satisfied customers constitute ~96.1% of the dataset while dissatisfied customers only ~3.9%.

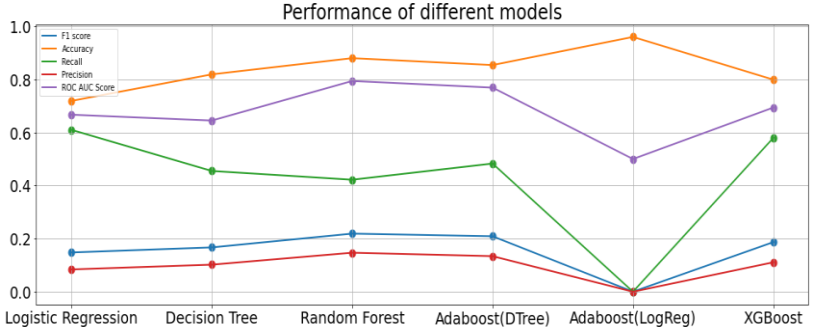
References

1. Santander Customer Satisfaction- A self-case study using Python:https://towardsdatascience.com/santander-customer-satisfaction-a-self-case-study-using-python-5776d3f8b060
2. Under sampling algorithms for imbalanced classification:https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

Model



Results



- ROC AUC score of **Random Forest was 0.8023** which was the best among all models.
- As the dataset is unbalanced the precision scores were bad for all models.
- Adaboost with Logistic Regression as base estimator was the worst model with ROC AUC and precision score of 0, but its accuracy was 93%.

Conclusion and Future Work

Data Preparation Strategy: We need more robust data preparation strategy by developing a deeper understanding of the dataset. Precision scores were consistently low for all of our models, which is a direct result of a poor data preparation. A more suitable subsampling method is needed to remove imbalance. Outlier detection and elimination also might be an important step to get better results.
Handling Features: Identifying features is critical since it will enable us to categorize them in different sets. Then, trying to understand how a particular set is correlated to target variable, can give us huge insight into the data. For this, a literature survey is needed to be done.