Univ.AI

Santander Customer Satisfaction

4-Folds : Heera Lal, Muhammad Akmal, Harshvardhan Palawat and Ishani Bari

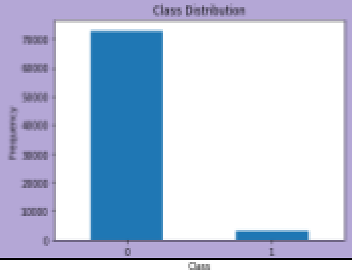
Motivation and About the Project

Santander bank[1] wants to evaluate whether their new customers are happy or not at their early stage relationship. The bank can proactively take measures to retain their customers at their early stage if the customers are classified as **unsatisfied**. Based on the different anonymized predictors, we trained our model to classify whether a customer is an unsatisfied customer.

The data provided is a highly imbalanced, we pre-processed our data by removing constant, quasi-constant and duplicate features and trained different models to get better results. In our point of view tuned XGBoost gives us the best recall scores.

Data and Labels

The dataset has 76020 observations in train dataset with 370 features along with one TARGET variable. From the total observations, there are only 4.11% values are positives cases(1) and rest are negatives (0)in the dataset.



References

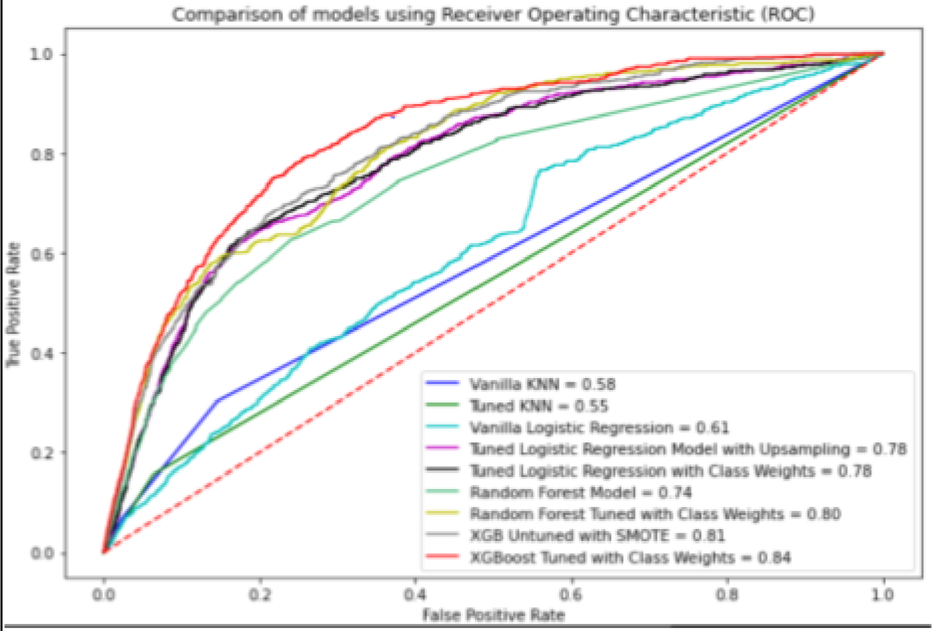
[1].https://www.kaggle.com/c/santander-customer-satisfaction/overview/description

Model

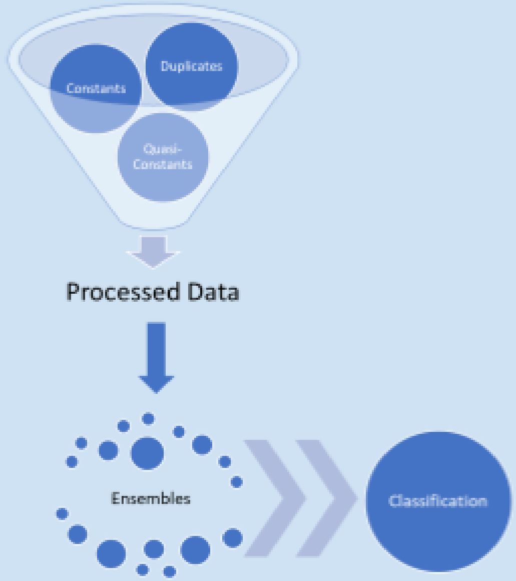
The ensembles of models we choose are the following

1. Vanilla KNN
2. Tuned KNN
3. Vanilla Logistic Regression
4. Tuned Logistic Regression Model with Upsampling
5. Tuned Logistic Regression with Class Weights
6. Random Forest Model
7. Random Forest Tuned with Class Weights
8. XGB Untuned with SMOTE
9. XGBoost Tuned with Class Weights

Results



Methodology



Conclusion

From the trade-off between False Negatives and False positives, we chose to have low False Negatives at the cost of False Positives.

As the customer acquisition cost is high, not to mention the damage to goodwill it causes when an unsatisfied customer leaves due to negligence of the business, we need to minimize false negative rate.

Future Work

Re-design the way we collect the data as a lot of unsatisfied customers might be leaving unnoticed.

Use Neural Networks.