Santander Customer Satisfaction

By Team - **4 Folds**

•••

Team Members

Harshvardhan Palawat

Heera Lal

Ishani Bari

Muhammad Akmal

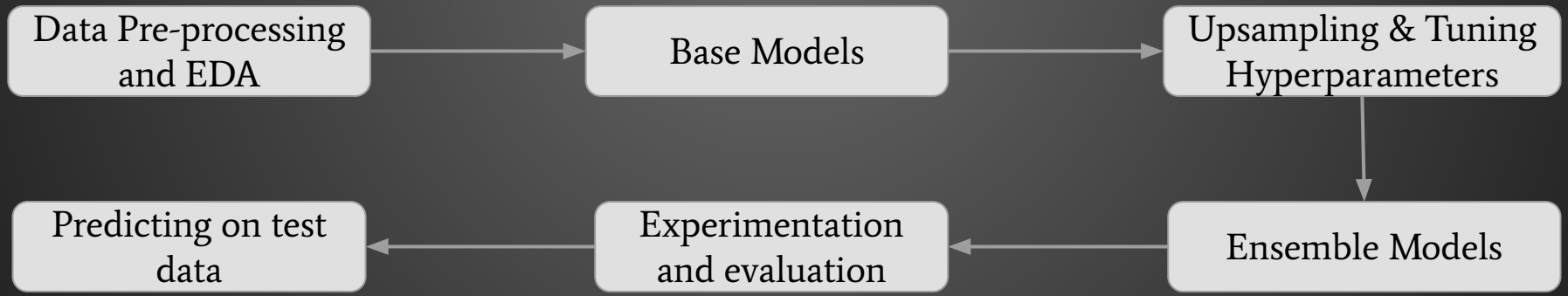
Outline

- Introduction
- Methodology
- Results
- Conclusion
- Future Scope

Introduction -

Problem Statement - We need to identify unsatisfied customers for a business so their experience can be improved. It is a binary classification problem with the target denoting Unsatisfied Customers with '1' and Satisfied Customers with '0'. Predictors have been masked for maybe privacy reasons and don't denote anything specific.

Approach towards the problem statement -



Introduction

Metric - Our focus is to achieve a high recall and minimize false negatives.

Algorithms used - As our TARGET feature has binary classes following algorithms were used:

- **KNN** - It is a non-parametric base model. It works on majority votes from nearest k neighbours.
- **Logistic Regression** - It is a parametric base model and log odds can be converted into class probabilities.
- **Random Forest Classifier** - It is a tree based, non-linear classification model.
- **XGBoost Classifier** - XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that may be useful to solve our problem statement.
- **Stacking** - Stacking multiple models from above mentioned and choosing the final classification based on - Majority Vote, Average with given threshold and Logistic Regression

Methodology

Dataset Description - The shape of original train data is 76020 rows and 370 masked predictors whereas the shape of original test data is 75818 rows and 370 masked columns. The TARGET feature is highly imbalanced as seen in the bar chart

Data Preprocessing done :

- Dropping the constant features (Features which only one value)
- Checking for the Quasi-constant features and dropping them.
- Checking the Duplicate features and dropping one of them.

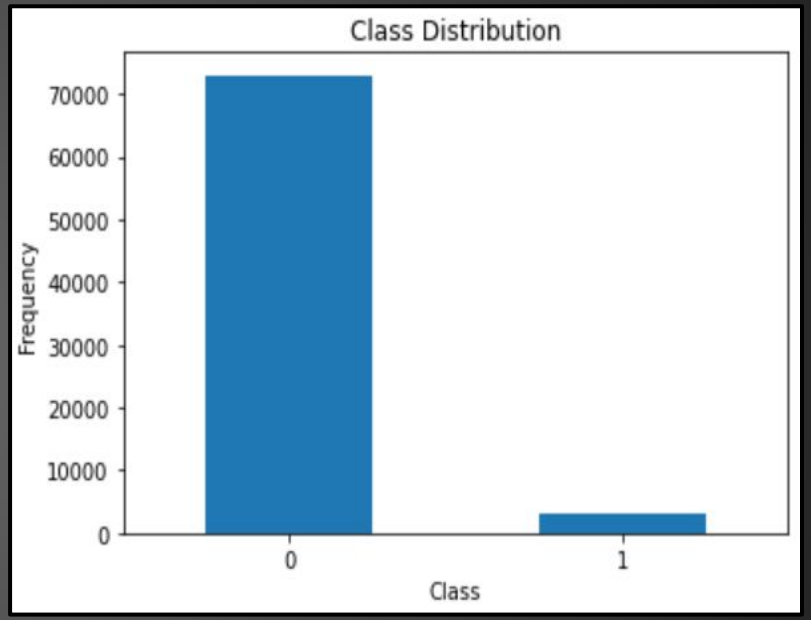


Fig 1. - Bar chart showing the TARGET classes frequency

After pre-processing, we are left with 184 predictors in the data .

Methodology

Packages used

Data Preprocessing and EDA	Pandas, NumPy, Matplotlib, Seaborn, Feature engine library
Resampling	Imbalanced-learn (SMOTE)
Model selection and model evaluation	Sklearn and XGBoost

Results

Comparison of models

The best model is **XGBoost Tuned with Class Weights** with an AUC score of 84% as compared to all the other models and experiments done and as seen from the ROC curves for various models tried.

Comparing the models based on Recall and F1-score as well, XGBoost is performing best.

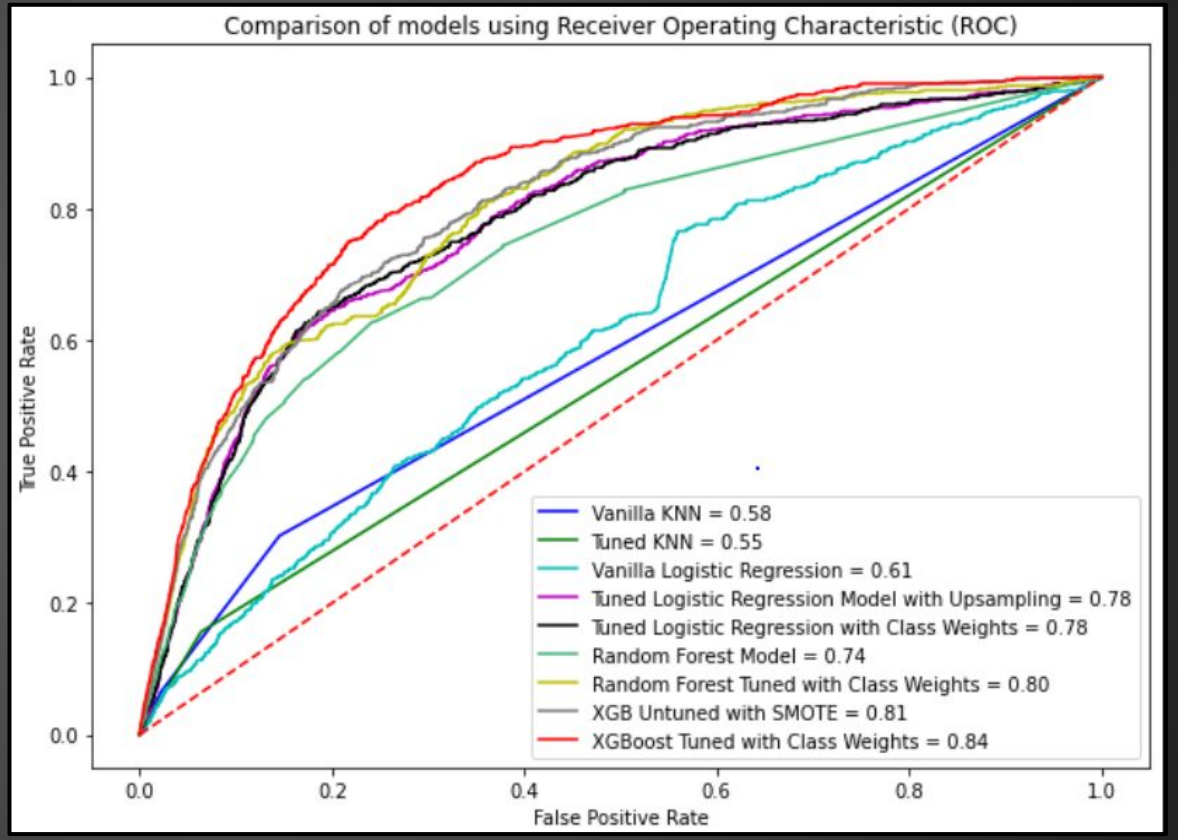


Fig 2 - Comparison of models using ROC

Results

Comparison of models using Recall as a metric

Model	Recall_Overall (in %)	F1-Score_Overall (in%)
Vanilla KNN	50	50
Tuned KNN	50	50
Vanilla Logistic Regression	50	49
Tuned Logistic Regression Model with Upsampling	71	49
Tuned Logistic Regression with Class Weights	71	48
Random Forest Model	52	52
Random Forest Tuned with Class Weights	70	50
XGB Untuned with SMOTE	70	40
XGBoost Tuned with Class Weights	76	53

Results

Ensemble Modeling

Combined different models with different strengths in order to achieve better results. Stacked multiple models to achieve the final classification based on the below 3 strategies. The performance of each strategy is as below:

Strategy 1
Majority Vote

	precision	recall	f1-score	support
0	0.99	0.81	0.89	14602
1	0.13	0.70	0.23	602
accuracy			0.81	15204
macro avg	0.56	0.76	0.56	15204
weighted avg	0.95	0.81	0.86	15204

Strategy 2
Average with given threshold

	precision	recall	f1-score	support
0	0.99	0.65	0.79	14602
1	0.09	0.83	0.16	602
accuracy			0.66	15204
macro avg	0.54	0.74	0.47	15204
weighted avg	0.95	0.66	0.76	15204

Strategy 3
Logistic Regression

	precision	recall	f1-score	support
0	0.97	0.90	0.94	14602
1	0.14	0.38	0.20	602
accuracy			0.88	15204
macro avg	0.56	0.64	0.57	15204
weighted avg	0.94	0.88	0.91	15204

The performance didn't improve much as compared to the best model '**XGBoost Tuned with Class Weights**' got previously. **So the best Model we have is XGB Tuned with Class weights**.

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.
- On the other hand, it'll not hurt much to waste a little budget in order to talk to customers who were misclassified as Unsatisfied.

Future Work

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