Analysis of a Bank’s Marketing Dataset

Springboard DSC Capstone Project I

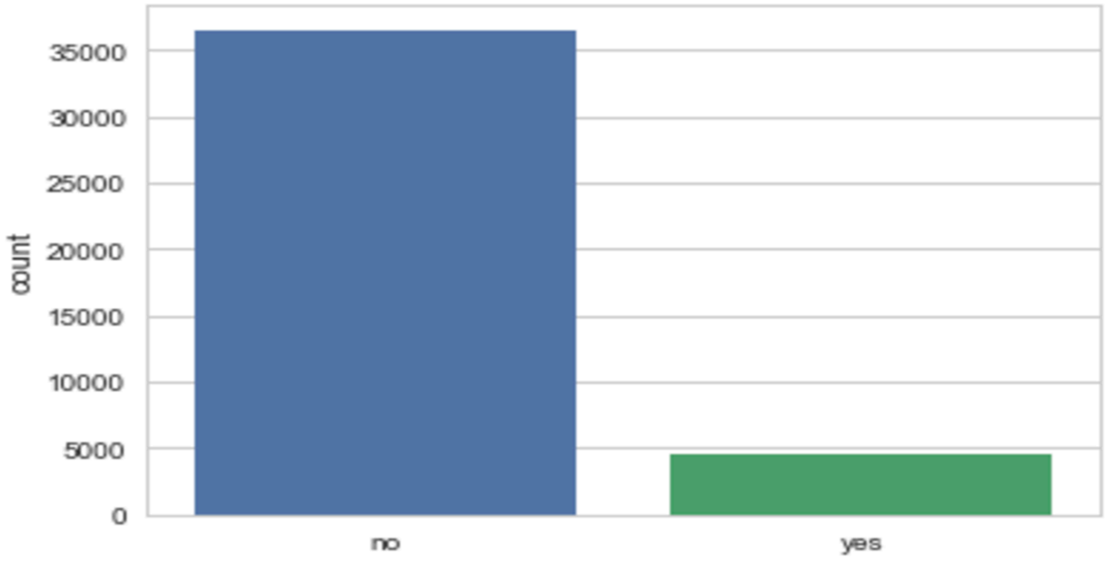
By: Anshul Gupta

Introduction

- Banks can have a huge consumer base and if there is a model which can predict the potential customers of a bank, it will not only save time but less resources of the bank will be used to identify new costumers.

- This is a binary classification problem which will predict the response variable ‘yes’ or ‘no’ based on certain attributes.

- The proportion of positive to negative labels is skewed towards the latter making this a class imbalance problem.



Approach

- Evaluation metric used will be sensitivity aka True Positive Rate as ‘yes’ is very important from business point of view.

- Classification algorithms used as a baseline metric are Logistic Regression, Random Forest, ADABoost, Gradient Boosting and XGBoost.

	Train_AS	Test_AS	Train_TNR	Test_TNR	Train_TPR	Test_TPR
Model						
LR	0.909	0.908	0.975	0.974	0.394	0.391
RFC	1.0	0.907	1.0	0.976	1.0	0.376
ADABoost	0.906	0.901	0.974	0.973	0.365	0.348
GBT	0.928	0.915	0.972	0.967	0.575	0.510
XGBoost	0.962	0.918	0.987	0.964	0.761	0.550

Techniques to improve Sensitivity Of Model

- To improve the sensitivity of the model I used resampling techniques such undersampling and oversampling.
- Undersampling works with sample of majority class to taken so its size is equal to minority class and then train test split is used in a stratified by manner.
- Oversampling works with sample of minority class resulting in equal proportion of majority and minority class.

	ROC	Test_A_S	Test_TNR	Test_TPR
Model				
SMOTE_LR	0.873	0.873	0.856	0.889
UnderSampler _GBT	0.891	0.891	0.863	0.920
UnderSampler _LR	0.861	0.861	0.863	0.859
SMOTE_XGB	0.953	0.953	0.958	0.947
UnderSampler_XGB	0.889	0.889	0.868	0.911

Recommendations to the Client

XGBoost model is the best algorithm to predict the output variable. Using XGBoost with SMOTE gives the best TPR which is about 94.7%, compared to 55% which was the result obtained without using any resampling technique. Even the Specificity is 95.8% which is far better than any other undersampling technique.

In terms of the business problem, these results indicate that the client can be confident that positive and negative cases will be predicted with a high level of confidence.

Future Work

To improve the results, we can use a Neural Networks. We can add multiple hidden layers and make it a deep learning model and train it on a GPU to get high processing power and optimize it by using different optimizers.

Alternatively we can use higher n_estimators in XGBoost with a PC with higher processing power to further improve the TPR.

- Thank You