Classification Models for Bank Marketing Data Set

Shiuli Subhra Ghosh (MDS202035), Suman Roy (MDS202041)

June 10, 2021

1 Comparison of Measures

Performance Measure	Decision Tree	Random Forest	Naive Bayes
Accuracy	89.56%	90.39%	86.65%
Precision	52%	55 %	42 %
Recall	82%	79%	51%
Time(s)	3.7	6.5	4.54

Table 1: Performance Matrix

1.1 Report & Observation

- We dropped **'default'** as it contained high number of unknown samples.
- from the other categorical variables we also removed the samples with unknown values.
- We observed that the data is **highly imbalanced**.
- We used **random over sampling** in Random Forest to negate the effect of imbalance.
- We used **class_weight** as a parameter for the Decision Tree Classifier placing grater weight on the **'yes'** target variable.
- **max_depth** is chosen for both the tree models in accordance with the performances of these models with different max_depths.
- **Clearly, as expected in terms of accuracy Random Forest is performing better than the other models.**
- We chose to maximize the recall for 'yes' target variable. But we did it in the way so that the Precision and Accuracy wouldn't decrease much.
- Naive Bayes Classifier is not performing up to the mark. We have used some user defined transforms on the outliers of **'duration' , 'campaign'** and it improved performance a bit.

- We have implemented the **pipeline** architecture for our model to avoid data leakage.

2 Points of Interest

- We did not drop the 'duration' attribute because in the feature importance list, this was the most importance feature showing the maximum correlation with target variable.
- For Decision Tree and Random Forest we had to perform precision-recall trade-off. We tried to maximise recall without drastic decrease in the precision or the accuracy. In line with this idea, we found that a recall value around 80 and a precision value around 55 is producing the optimum output.
- There are 19 attributes and many of those are not contributing highly towards the model performance. We tried to fit the model with various number of features(by dropping the less important features) but the output is not showing great deal of improvement.
- If we drop all the categorical features for Naive Bayes Classifier, the performance is changing significantly.
- We haven't removed the outliers from the Decision Tree and Random Forest Classifiers, yet those models performed well substantially.

3 Link for Outputs

https://drive.google.com/drive/folders/1nzIXbXsH57IyBHA–YUsmvTXDyVAAU-r?usp=sharing