A Technical Report of Two Machine Learning Models for Predicting Churning

By

Sans Basnet
Dated: May 2, 2021

EXPLORATORY DATA ANALYSIS

The dataset provided had `10` columns: 2 categorical variables and the rest integers and floats. A prelim analysis showed that column 5 i.e. the calls columns had `1092` rows of missing data. I used mean values to impute the column. I checked the class imbalance which was about `8000` for 0 and `2000` for 1. Despite the imbalanced-class in the dataset, I proceeded further to let the performance metrics be the judge and follow-up. In frequency count, I noted CA in 'states' column had the highest frequency whilst Home & Home improvement had the highest frequency in 'business category' within the two categorical variables. These were checked to give an idea of the uniqueness of the observations.

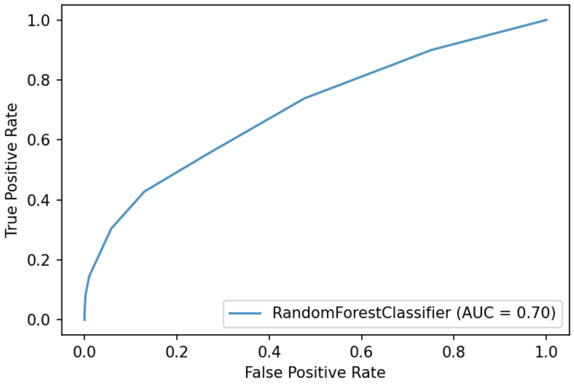
PRE-PROCESSING

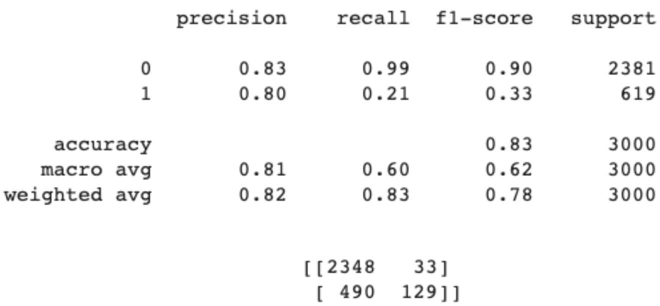
I removed the 'business category' column after visualization. It appeared inessential as well as maximized the column space when transforming any categorical columns into dummy variables which was, however, performed for the 'states' column. The dataset was split with `test_size=0.30`.

Pre-process: Pandas, numpy, csv
Visualization: Matplotlib, plotly, seaborn
Machine Learning: Sci-kit Learn, RandomForestClassifier, DecisionTreeClassifier

MODEL & EVALUATION

I used Random Forest and Decision Tree Classifiers in this experiment given the literature tells us that these models have higher interpretability and decent accuracy. For Random Forest, `n_estimators=10` or the number of trees in the forest. Given an imbalanced dataset, accuracy could cause false assumptions regarding the classifier's performance, in that case it is better to rely on precision and recall. RF performed better than DT comparing model performance metrics where weighted average precision of 82% for RF and 75% for DT tells us how accurate the positive predictions were. Recall or sensitivity tells us the coverage of actual positive sample. Weighted average Recall score of RF was 83% and 75% for DT. Macro average gives each prediction similar weight while calculating loss but given imbalanced-class one needs to give importance to some prediction more (based on their proportion), hence the weighted average is emphasized.



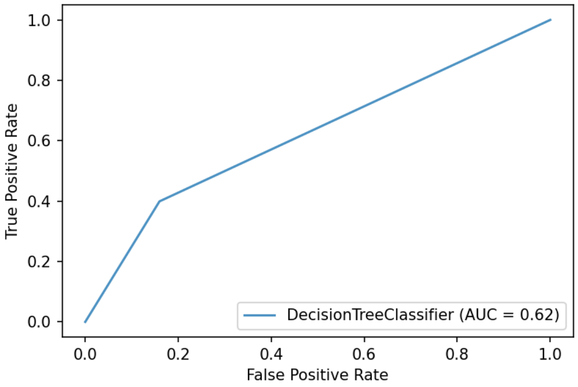


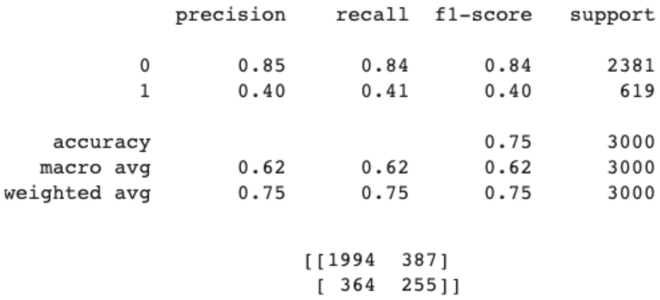
```
              precision    recall  f1-score   support

           0       0.83      0.99      0.90      2381
           1       0.80      0.21      0.33       619

    accuracy                           0.83      3000
   macro avg       0.81      0.60      0.62      3000
weighted avg       0.82      0.83      0.78      3000

                    [[2348   33]
                     [ 490  129]]
```





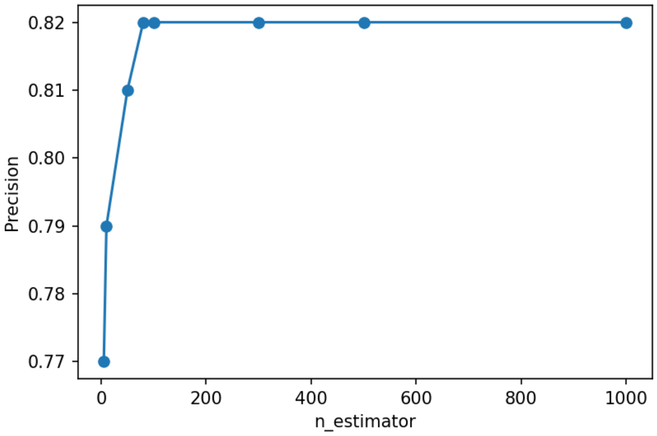
```
              precision    recall  f1-score   support

           0       0.85      0.84      0.84      2381
           1       0.40      0.41      0.40       619

    accuracy                           0.75      3000
   macro avg       0.62      0.62      0.62      3000
weighted avg       0.75      0.75      0.75      3000

                    [[1994  387]
                     [ 364  255]]
```

AUC value and ROC in figures above and the matrix below above proves that the RF is clearly better than the DT. Figure below shows that precision levels at 82% and has no effect as n_estimator is increased past 200-400 trees in the forest.



RECOMMENDATIONS

These models represent good interpretability in understanding the dataset and the training process. However, if we are more interested in accuracy we can opt for a more sophisticated model such as Kernel based methods to optimize the bias-variance tradeoff. Neural Networks are not recommended because 10000 observations in the dataset is insufficient for training a NN from my past experiences.

Cross-validation is recommended for validation of the algorithm performance and balancing the predicted features' classes for this imbalanced dataset.