Data Mining and Machine Learning
Assignment 1

Rajas Vaidya (MDS202131), Rohan Dharmadhikari (MDS202137)

1. Comparison of Performances

Performance Measure	Decision Tree	Naïve Bayes	Random Forest
Accuracy	83.04%	81.04%	85.97%
Precision	34.93%	32.69%	41.16%
Recall	60.69%	55.6%	60.69%
F1 score	44.34	41.17	49.05
Time(s)	2.14s	0.38s	14.14s

Table 1. Performance Matrix

2. Data Preparation

- As explained in the description of the data by UCI, the column 'default' has been dropped from the data. It makes sense to drop the column as the duration of the call is not an attribute we can know before the call is made and at the end of the call, the result if already known, so the attribute is useless.
- The data is really skewed towards 'no' value in the target ('y') column. Thus, the parameter 'class_weight' in Decision Tree and Random Forest classifiers has been assigned the value 'balanced'. Similarly, in Naïve Bayes classifier, the data has been padded with oversampling using Synthetic Minority Oversampling Technique (SMOTE) to achieve similar results.
- The numeric variables in the data have been standardised to ensure that variables measured at different scales do not contribute differently in the analysis.
- The column 'pdays' has been divided in to two columns viz. 'pdays' and 'pdays2'. 'pdays' represents whether the client was previously contacted and 'pdays2' represents how many days back has the client been contacted, if contacted.
- Samples with unknown values for all the attributes in {'education', 'marital', 'housing', 'job', 'loan'} have been removed as without these attributes, a new person will essentially be a blank slate and almost impossible to be classified.
- All ordinal variables and 'education' have been encoded using a pre-decided set of ordered values. All other nominal variables have been encoded using label_encoder.
- About 30% of the samples have been selected randomly and set aside as the test samples.

3. Classifiers

3.1. Decision Tree

- A Decision Tree Classifier has been defined for every depth from 1 to 20 (total number of columns) with minimum number of samples at leaf set to 5.
- It was observed that maximum recall from the best models among all, was around 60-65%. Thus, the best classifier was selected by choosing the classifier with the best precision among the ones with recall value > 60%.
- This strategy was decided on the basis that a comparatively high recall value with very low precision will actually result in the bank making many more unnecessary calls than required. Thus, maximising precision for a recall value comparable to the max recall would reduce the manual efforts.

3.2. Naïve Bayes

- As the data has very few samples of minority class (yes), synthetic minority sampling technique (SMOTE) has been used, where duplicate entries of minority class were created making the no. of minority entries equal to the majority class and then the undersampling of this new data was done so that the majority class will have 50% more entries than the minority class.
- It is observed that even though the accuracy of Naïve Bayes classifier is good, its recall and precision for the given data is low.
- Among the three machine learning models, Naïve Bayes has the least recall.

3.3. Random Forest

- A Random Forest Classifier has been defined for every depth from 1 to 20 (total number of columns) with 100 estimator trees, minimum number of samples at leaf set to 5, bootstrap and oob_score set to True.
- It was observed that maximum recall from the best models among all, was around 60-65%. Thus, the best classifier was selected by choosing the classifier with the best precision among the ones with recall value > 60%.
- This strategy was decided on the basis that a comparatively high recall value with very low precision will actually result in the bank making many more unnecessary calls than required. Thus, maximising precision for a recall value comparable to the max recall would reduce the manual efforts.

4. Observations and Conclusion

- The aim of the analysis is to predict whether a particular client will subscribe to term deposit or not, so as to decide whether calling that client would be worth the time. Hence, it was decided to consider borderline 'No's as 'Yes', so as to not miss out on the potential subscribers.

- The approach was to choose the model which can give the maximum recall (for decision tree and random forest classifiers) without sacrificing precision as much as possible.
- As much as 88% of the original target variable indicates 'No'. Hence, the analysis of accuracy level would not be significant as the target variable is highly skewed towards 'No'.
- In such problems, the minority class has very low representation and often times it's the minority class which is of importance. Here too, the clients with y=yes are of importance to us. Hence, we use oversampling so that minority class also gets enough representation in the train data in Naïve Bayes and set the class weight parameter in Decision Tree and Random Forest to achieve similar result.
- The classifier parameter min_samples_leaf was set at 5 as this was the value that achieved better results than any other value in the range (0,15).
- After executing all three ML models, it was observed that the random forest is better than other two models in every criterion (given a threshold value of recall). The drawback is random forest takes slightly more time than other models.