A comparison of Random Forest (RF) and Naïve Bayes (NB) on predicting term deposit

Description and motivation

We will solve the binary classification problem of predicting if the customers would subscribe the term deposit which is based on certain demographics, campaigns, social and economic attributes. We aim to employ Random Forest and Naïve Bayes to predict the subscription of term deposit. After then results will be compared to infer which model performed efficiently among them.

Exploratory Analysis

- Dataset: Bank Marketing data set of Portuguese banking institution, from May 2008 to November 2010 from UCI repository.
- The dataset contains 41188 rows and 20 inputs, data is from direct marketing through phone calls. One client was contacted more than one time to know if the client is interested in subscribing the term deposit.
- There were 4 different datasets, but we choose the latest and with maximum features to achieve maximum results from the predictive analysis. The dataset we choose was very close to the data analyzed by S. Moro, P. Cortez and P. Rita. (2014) [1].
- The data cleaning process was carried out by discarding outliers and null values from the features of the data set.
- Feature engineering was performed on categorical features such as age into age groups, encoding (Label Encoder) was applied on default, housing, poutcome and loan. While one hot encoding was employed on job, marital, education, contact, age group, day of week and month.
- Feature selection was conducted through chi square test that aims to select the features that are highly related to the target variable. Feature having P-value less than 0.05 were selected as shown in figure 2 as they possess higher relation with target variable. Few of these processes were conducted in S. Moro, P. Cortez and P. Rita. (2014) [1].
- The cleaned dataset consists of 36584 rows, 57 features and 1 target column.
- The target column contains a 1 if a term deposit was subscribed and 0 otherwise.
- A class imbalance problem exists as shown in the figure 1 as 88% of clients were not interested in investing in term deposit, whereas only 12% of clients were attracted towards investing in term deposit.
- After Exploratory analysis, we got to know that in the figure 3 month of may was most effective for proposal acceptance. People who tend to be professional in the figure 4 were seen to accept the proposal in higher rate. Clients in the figure 5 who were adult, and the ones who were contacted more often accepted the proposal in greater amounts.

Figure 1

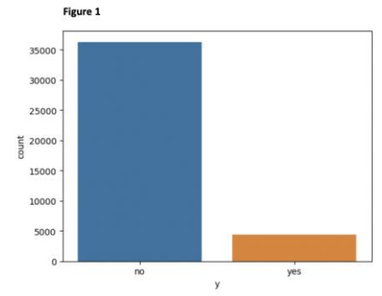


Figure 2

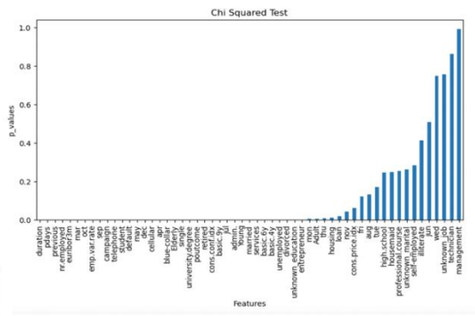


Figure 3

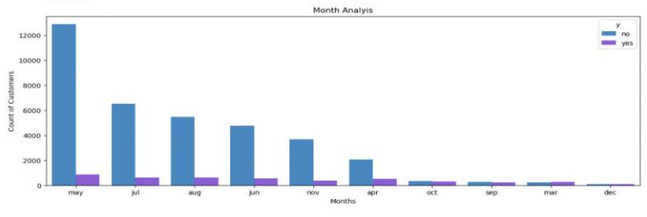


Figure 4

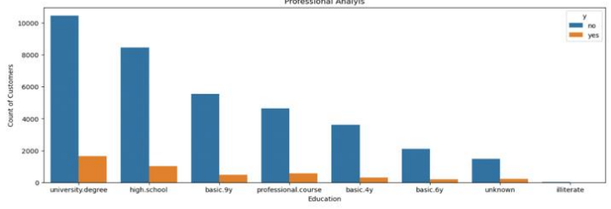
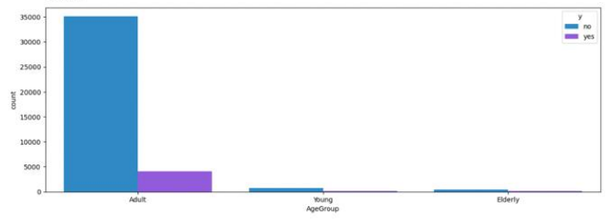


Figure 5



Random Forest

- Random Forest is an ensemble of unpruned classification or regression trees developed through bootstrap sampling of the training data using the random selection of input variables to determine the split point [3].
- In Bagging, it predicts based on maximum voting in classification or aggregation in regression.

Pros

- Bagging technique reduce the variance of model.
- Random Forest can handle missing values.
- It has ability to estimate the relevance of variables.
- It can handle high dimensional data.

Cons

- Random Forest takes a lot of time to train.
- RF is a complex algorithm as it creates a lot of trees thus, requiring a lot of computational resources.

Naïve Bayes

- Naïve bayes (NB) algorithm is a supervised algorithm for classification.
- Naïve Bayes is called 'Naïve' due to the reason of presuming that the features are independent of each other, while given a target label.

Pros

- It is scalable with considerably large datasets
- Deals well with high dimensional data.
- It is not sensitive to features that are irrelevant.

Cons

- It assumes that the features in the dataset are mutually independent.
- If there was category in a class label that was not in the training set, the predictive model would assign it to 0.

Hypothesis Statement

- Random Forest and Naïve Bayes supervised models are expected to perform It is expected that both models will perform well. NB shall not perform better than Random Forest due to the reason of its assumption of independence of features.
- The minority class 'yes' is likely to have highest cost of misclassification.
- We first assumed that there were several irrelevant input attributes that difficult the DM algorithm learning process (e.g., by increase of noise) [2].

Methodology

- The Data is split into 70% train data and 30% test data through holdout method that randomly splits the data.
- Optimization of the model by feature selection, feature engineering, handling class imbalance and hyperparameter tuning.
- Model will be trained by applying 5 - fold cross validation on the training data. A DM model that is fed with training set data using as inputs all relevant features of the first confirmed factor and then AUC is computed over the validation set [1].
- Assess and estimate according to metrics that which model is optimal.
- After training predict the test data and calculate precision, recall, F1 score and confusion matrix to infer results.
- Hyperparameter tuning will be employed to avoid generalization errors for both the models.

Experimental results, parameter choices and feature selection:

Naïve Bayes

- Precision and recall were improved after balancing the target variable.
- The best model was identified as the model which used feature selection, under-sampling to balance target classes.

Choice of Parameters:

- 0.024 width was selected for kernel

Random Forest

- The model was fitted by bagging or bootstrapping algorithm with random features selected at each split.
- The Grid search was used for hyperparameter optimization with maximum number of objective function evaluations at 150, after then hyperparameter optimization options were set to optimize 4 parameters such as the number of predictors to select at random for each split, number of ensemble learning cycles which is log scaled from 10 to 500. Minimum leaf size and maximum number of decision split were also set.

Choice of Parameters:

- Maximum number of splits are 94.
- The number of predictors sampled are 28.
- The number of trees or learning cycles are 37.
- Minimum Leaf size is 5

Figure 10: Prediction Scores

AVG	RF	NB
Precision	**0.9176**	**0.8788**
Recall	**0.9187**	**0.8787**
F1 Score	**0.9181**	**0.8788**

Figure 6: Random Forest Confusion Matrix

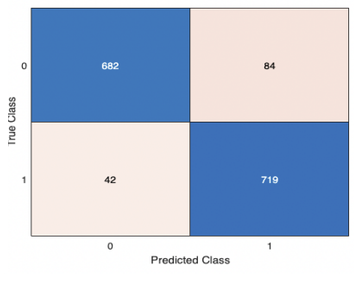


Figure 7: Naïve Bayes Confusion Matrix

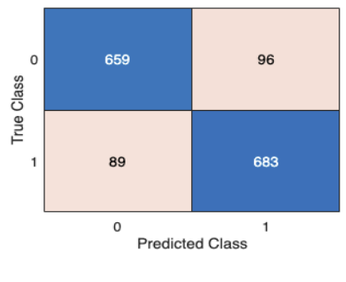


Figure 8: RF ROC Curve

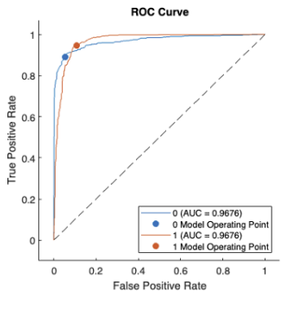
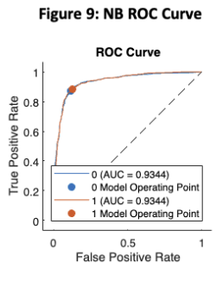


Figure 9: NB ROC Curve



Analysis and Evaluation of results

- Random Forest had performed better in terms of accuracy than Naïve Bayes as shown in fig 6 and fig 7. As it is an ensemble learning and by bagging technique it takes majority voting to select the class thus contributing to attain more accuracy. While Naïve Bayes is a probabilistic classification model that assumes that the features are independent regardless of the any correlation among them thus its accuracy depends on this assumption.
- The Feature Selection also played vital role in improving the model. There were redundant features as well as non-relational features with target variable which were necessary to remove before training the model to achieve optimal results. For this purpose, we used Chi-square test for feature selection where higher chi-squared values mean that the features have strong relation with class labels.
- Due to class imbalance, the precision and recall of minority class was low as compared to majority class so it outperformed. It required improvement so we carried out under sampling by random under sampler which balanced the majority class with the minority class. This increased the average training accuracy score as precision and recall increased.
- For hyperparameter tuning, RF takes a lot of time approximately 20 minutes by employing grid search technique to search in the range of 150 function evaluation in our coursework. But NB took much less time approximately 4 mins to train. For testing the time taken by RF was drastically decreased as compared to training, moreover it outperformed NB as AUC in figure 8 of RF was considerably higher.
- Random Forest, after grid search was applied found the best set of hyperparameter when 92 function evaluation passed. Random Forest was optimized on 4 hyperparameters such as maximum number of decision splits, number of learning cycles, minimum leaf size and number of predictors sampled. For bagging, min leaf size is extremely small and maximum number of splits are n(observations)-1 that means that it grows deeper trees thus, contributes to perform well.
- RF grows deeper trees that doesn't contribute to generalize more, thus in order to cater this problem 5-fold cross validation is applied to make the model generalize and eliminate overfitting.
- For hyperparameter tuning of NB, Bayesian optimization approach was opted in order to find the minimum and maximum of objective function for the purpose to generalize the model.
- The Kernel distribution instead of normal distribution enhanced the average training accuracy as features such as age and duration were skewed. After the hyperparameter tuning we get to know that 0.024 is the best estimated width.
- The AUC of RF is slightly higher than that of NB as shown in the figure 8 and figure 9. Moreover, confusion matrix of fig 6 infers that RF handles the misclassification better than NB. RF also overpowers NB in terms of specifying the correct positive predictions (precision), moreover high recall is also observed as shown in fig 10.

Lessons Learnt

- Feature Engineering and Feature selection plays a vital role in optimizing the model. Chi-square test helped us to select optimal features for the model to train.
- Imbalance class labels leads to inefficient model in terms of precision, recall.

Future Work

- For further work, we could use Lasso or Ridge regression for feature selection instead of Chi-square test as used in this course work.
- We could apply SMOTE (Synthetic Minority Oversampling Technique) to assess the effect of oversampling on the models.

References

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.

[3] Khalilia, M., Chakraborty, S. & Popescu, M. (2011) 'Predicting disease risks from highly imbalanced data using random forest'. BMC medical informatics and decision making, 11 (1), pp.1.