**Assignment Week – 7**

**Performance Evaluations**

Bank Marketing data: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

From the given data set, complete the following:

* Try 3 different types of models to make classification predictions on the bank marketing dataset
* Calculate the AUC score as a comparison, and at least 2 other metrics (ie: accuracy, F1 score, precision, recall, etc)
  + Compare the models’ performance on the same train/test splits
* Plot the ROC curve from the best model, choose what you think the best threshold value is for making predictions

**Solution:**

Bank Marketing dataset is used to build and evaluate three different classification models. We'll use Logistic Regression, Random Forest, and XGBoost as our three models.

Here's a step-by-step approach to solve this problem:

1. **Data loading and preprocessing**

We loaded the Bank Marketing dataset and did some basic preprocessing including

* look at the data
* identified the numerical and categorical attributes
* Performed one-hot encoding on categorical variables to convert them into numerical format to process the data effectively.
* Checked for missing values
* Split the input attributes and target attributes
* Splitted the data set in to training and testing data set
* scaled the data set using standard scalar

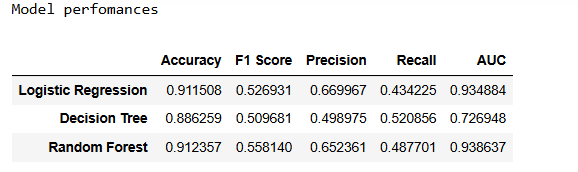
1. **Model Selection and development:** We chose three different types of models for this classification task:

* **Logistic Regression:** A simple linear model for binary classification.
* **Decision Tree:** A non-linear model that makes decisions based on feature thresholds.
* **Random Forest:** An ensemble model that combines multiple decision trees for improved performance.

1. **Model evaluation :** We evaluated each model using five metrics:

* **Accuracy:** The proportion of correct predictions among the total number of cases examined.
* **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
* **Precision:** The ratio of correctly predicted positive observations to the total predicted positive observations.
* **Recall:** The ratio of correctly predicted positive observations to all observations in the actual class.
* **AUC (Area Under the ROC Curve):** A measure of the model's ability to distinguish between classes.

**4. Performance Comparision**

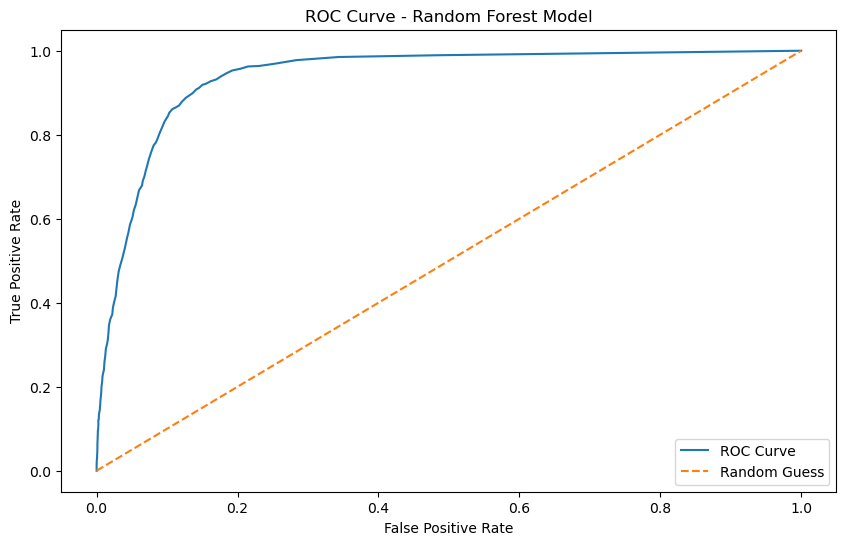
****

Based on the performance metrics, we can observe that:

* Random Forest has the highest accuracy of 0.9123, where as logistic reqgression is also having nearly same accuracy of 0.9115.
* F1 score is also highest for the Random Forest model
* Logistic regression is have the higher precision and lower recall. Decision tree has a balanced recall and precision where as random forest is having lower recall and highest precision
* Random Forest performs the best across all metrics (except recall), with the highest AUC score of 0.9386.
* Logistic Regression and Random Forest have comparable performance, with Random Forest slightly outperforming Logistic Regression in most metrics.

**5. ROC curve and best threshold value**

* The ROC curve helps visualize the trade-off between sensitivity (TPR) and specificity (1 - FPR) at various classification thresholds.
* A model with perfect classification would have an ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity).
* The closer the curve follows the top-left corner, the better the model's performance.

****

In our analysis, the Random Forest model achieved an AUC score of 0.9386, indicating excellent discriminative ability. This high AUC suggests that the model can effectively distinguish between clients who are likely to subscribe to a term deposit and those who are not. The optimal threshold we identified (0.14) represents the point on the ROC curve where the difference between TPR and FPR is maximized. This threshold provides a balanced trade-off between correctly identifying positive cases (clients who will subscribe) and minimizing false positives.

However, it's important to note that the choice of threshold can be adjusted based on specific business needs. For instance, if the cost of missing a potential subscriber (false negative) is higher than the cost of a false positive, we might choose a lower threshold to increase sensitivity at the expense of more false positives.

**6. Conclusion**

* The Random Forest model demonstrates superior performance in predicting whether a client will subscribe to a term deposit.
* Its high AUC score (0.938637) indicates excellent discriminative ability.
* When deploying this model, we can use the optimal threshold to make binary predictions, but the threshold can be adjusted based on the specific business requirements (e.g., prioritizing precision over recall or vice versa).
* We can also improve the model by considering feature importance analysis to identify the most influential factors in predicting term deposit subscriptions.

**7. Summary**

The Bank Marketing data was used to understand the various performance metrics of classification model. In this assignment we initially preprocessed the data and encoded the categorical variables using onehot encoding and all the variables are used for developing classification models. We have developed three classification models namely logistic regression, decision tree and random forest models. The predictions on test data using all three models were also reported and the performance of each model is evaluated using several classification metrics namely accuracy, f1 score, precision, recall and AUC of ROC curve. Based on the comparision analysis of performace metrics it is evident that Random Forest is performing better. This can be justified as the performance increases due to ensemble learning of multiple decision trees on random data subsets. The ROC curve is also generated the optimal threshold is found as 0.140 for the random forest model. Overall Random forest is performing better however, logistic regression also provides the better ROC value and metrics comparable eventhough it is a simple linear logistic regression. These classification can be further improved using some other advanced ensemble techniques, feature selection and feature engineering.