**MACHINE LEARNING REPORT: CAPSTONE 1**

Banking client sought to predict which customers were most likely to subscribe to term deposit, thereby maximizing its marketing campaign by targeting customers most likely to subscribe to term deposit ‘y’. This is a classification problem, therefore Logistic Regression and Random Forest models were compared to determine which model would best predict  customers most likely to subscribe.

**Logistic Regression Steps**

* Create training and test sets: 20% of data was set aside for testing
* Set random sate at 42 to ensure reproducibility
* Normalize data using StandardScaler in order to bring all the values of numeric columns in the dataset to a common scale.
* Preprocess imbalanced data: only 11% data set responded yes, therefore data is imbalanced. method used is SMOTE
* Logistic regression classifier was applied to training data
* Predict the labels of the test set using Logistic Regression.predict()
* Compute predicted probabilities for independent variables test data
* Compute and print the confusion matrix
* Compute and print the classification report
* Generate ROC

**Random Forest Steps**

* Create training and test sets: 20% of data was set aside for testing
* Set random sate at 42 to ensure reproducibility
* Random Forest classifier was applied to training data: (n\_estimators=200, max\_depth = 3, max\_features=17)
* Predict the labels of the test set using RandomForestClassifier.predict()
* Compute predicted probabilities for independent variables test data
* Compute and print the confusion matrix
* Compute and print the classification report
* Generate ROC

**Table of Results: Confusion Matrix**

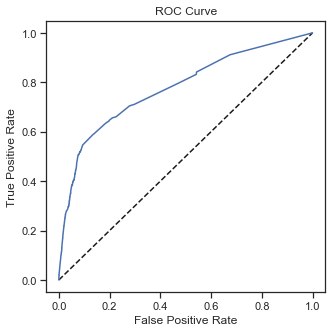
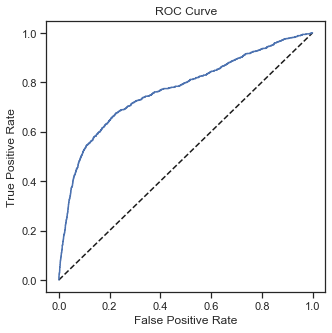
Logistic Regression: [[7015 250] [700 271]]

Random Forest: [[7143 122] [790 181]]

**Table of Results: Classification Report**

Logistic Regression

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | | | | **Random Forest** | | | |
|  | **precision** | **recall** | **f1-score** | **support** | **precision** | **recall** | **f1-score** | **support** |
| Subscribed: No | 0.91 | 0.97 | 0.94 | 7265 | 0.90 | 0.98 | 0.94 | 7265 |
| Subscribed: Yes | 0.52 | 0.28 | 0.36 | 971 | 0.60 | 0.19 | 0.28 | 971 |
|  |  |  |  |  |  |  |  |  |
| accuracy |  |  | 0.88 | 8236 |  |  | 0.89 | 8236 |
| macro avg | 0.71 | 0.62 | 0.65 | 8236 | 0.75 | 0.58 | 0.61 | 8236 |
| weighted avg | 0.86 | 0.88 | 0.87 | 8236 | 0.86 | 0.89 | 0.86 | 8236 |



Logistic Regression

Random Forest

**Results**

Logistic Regression gave us about 89% accuracy both on train and test data. Confusion matrix shows that the model correctly predicted 7015 no subscription(0), and 271 subscription(1) with 7296 correct prediction in total and we had 941 incorrect prediction in total (250 false positives and 700 false negatives). This means that 700 customers who could have subscribed would have been ignored, while 250 customers who would not have subscribed would have been targeted for marketing campaigns, this any resources allocated would have been a waste.

Random Forest gave us about 89% accuracy both on train and test data. Confusion matrix shows that the model correctly predicted 7143 no subscription(0), and 181 subscription(1) with 7324 correct prediction in total and we had 912 incorrect prediction in total (122 false positives and 790 false negatives). This means that 790 customers who could have subscribed would have been ignored, while 122 customers who would not have subscribed would have been targeted for marketing campaigns, this any resources allocated would have been a waste.

**Conclusion**

Comparing the F1 scores of Logistic regression and random forest models respectively, the results are very similar, with the logistics regression having a very slight edge in predicting the Subscribed: Yes at .36 as opposed to .28 from the random forest model.