**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
* Standardize 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
* Perform Hyper-parameterization
* Random Forest classifier was applied to training data
* 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**

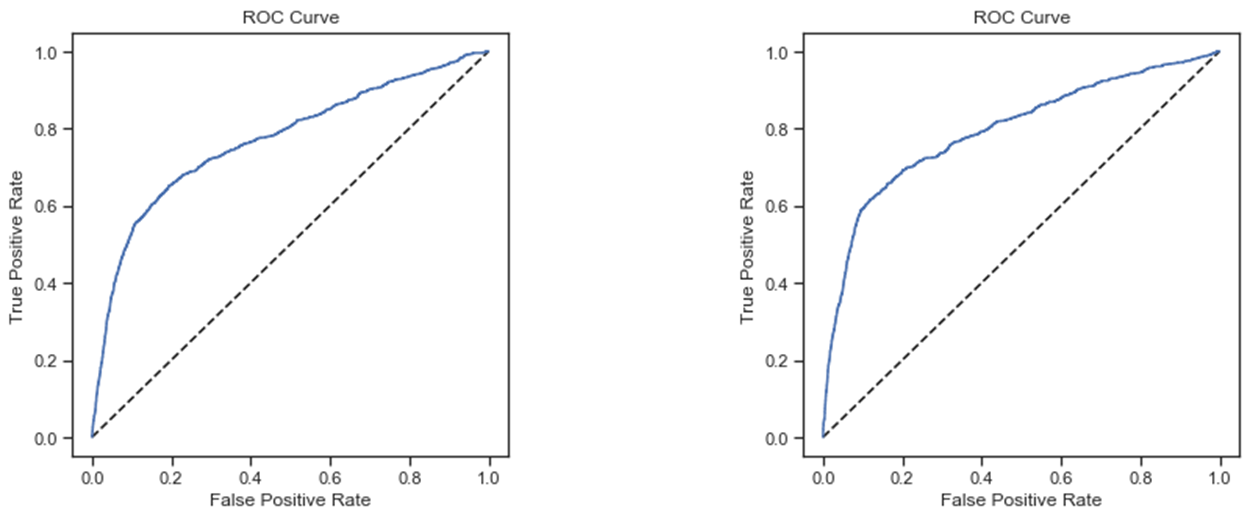
Logistic Regression: [[5588 1677] [310 661]]

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

**Table of Results: Classification Report**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | | | | **Random Forest** | | | |
|  | **precision** | **recall** | **f1-score** | **support** | **precision** | **recall** | **f1-score** | **support** |
| Subscribed: No | 0.95 | 0.77 | 0.85 | 7265 | 0.91 | 0.98 | 0.94 | 7265 |
| Subscribed: Yes | 0.28 | 0.68 | 0.40 | 971 | 0.61 | 0.23 | 0.34 | 971 |
|  |  |  |  |  |  |  |  |  |
| accuracy |  |  | 0.76 | 8236 |  |  | 0.89 | 8236 |
| macro avg | 0.62 | 0.72 | 0.62 | 8236 | 0.76 | 0.61 | 0.64 | 8236 |
| weighted avg | 0.87 | 0.76 | 0.80 | 8236 | 0.87 | 0.89 | 0.87 | 8236 |

Logistic Regression Random Forest



**Results**

Logistic Regression gave us about 73% accuracy on train and 76% accuracy on test data. Confusion matrix shows that the model correctly predicted 5588 no subscription(0), and 661 subscription(1) with 6249 correct prediction in total and we had 1987 incorrect prediction in total (310 false positives and 1677 false negatives). This means that 1677 customers who could have subscribed would have been ignored, while 310 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 91% on train and 89% accuracy both on train and test data. Confusion matrix shows that the model correctly predicted 7118 no subscription(0), and 226 subscription(1) with 7344 correct prediction in total and we had 892 incorrect prediction in total (147 false positives and 745 false negatives). This means that 745 customers who could have subscribed would have been ignored, while 147 customers who would not have subscribed would have been targeted for marketing campaigns, thus any resources allocated would have been a waste.

**Conclusion**

Comparing the F1 scores of Logistic regression and random forest models respectively, Logistics regression predicted Subscribed: Yes at .40, which is quite low, but predicted non-subscribers with a respectable 85 score a . The overall accuracy of the model is a respectable .76, with a macro average of .62, and weighted average of .80

The Logistic Regression had very strong precision score of .95 for non-subscribers, but a very low precision score of 0.28 for subscribers. The Random Forest had very strong precision and recall for subscribers at 0.91 and 0.98 respectively, with a 0.94 f1 score. The precision score for subscribers was 0.61, much better than the 0.28 for logistic regression. However, the recall score for random forest subscribers was 0.23, which is much lower than 0.68 in logistics regression.

The roc\_auc\_score of random forest is 79% which is higher than the 77% reported for logistic regression.

|  |  |  |
| --- | --- | --- |
|  | **Random Forest** | **Logistic Regression** |
| Accuracy | 0.89 | 0.76 |
| F1 subscribed (N) | 0.94 | 0.85 |
| F1 subscribed (Y) | 0.34 | 0.40 |
| Macro avg | 0.64 | 0.62 |
| Weighted Average | 0.87 | 0.80 |
| Roc\_aoc\_score | 0.795 | 0.77 |

This analysis would be further improved with availability of more data. However, considering the data available, random forest is a better predictor of variable ‘y’ than logistic regression on all measures with the exception of the F1 scores for customers who subscribed.