Section 3 Predictive Modelling

**Preliminary**

Unlike using all of the variables for predictive modelling, we have decided to use EFA (Exploratory Factor Analysis) scores to create our predictive modelling. EFA is simply PCA (Principal Component Analysis) but the components have rotated in the multivariate space. This allows us to identify the fundamental relationship between variables.

Like other predictive modelling process, the data needs to split into training (the sample of data to fit the model) and test (the sample of data to assess the performance of the model). In this instance, the data has split into 70% training 30% testing data. Additionally, the process will be randomly initialised 42 times to ensure validity.

**Predictive Models**

There are 6 independent variables (as we have 6 factors in our EFA) and 1 target variable (which is creditability) in our predictive model. A series of predictive models will be used to determine which model yields the highest F1-score. F1-score is used to evaluate the performance of the model as it is a combined matrix for both Precision and Recall. Additionally, it is important to note that we are only interested in the cases where the model has correctly predicted the approval of the loan (which our target = 1).

|  |  |
| --- | --- |
| Prediction Algorithm | F1-Score |
| Logistic Regression | 0.852 |
| Naïve Bayes | 0.844 |
| Random Forest | 0.821 |
| Decision Trees | 0.734 |

Figure \_\_ The results of each of the predictive models

**Interpretation**

Based on the figure above, it shows that logistic regression yields the highest F1-Score.

However, if we look at the confusion matrices of each of the predictive models. It tells us a different story.

|  |  |  |  |
| --- | --- | --- | --- |
| Cost Matrix | | Predicted | |
| Good | Bad |
| Actual | Good | 0 | 1 |
| Bad | 5 | 0 |

Table \_\_ Cost Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression Confusion Matrix | | Predicted | |
| Good | Bad |
| Actual | Good | 190 | 48 |
| Bad | 18 | 44 |

Table \_\_ Logistic Regression Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Decision Tree Confusion Matrix | | Predicted | |
| Good | Bad |
| Actual | Good | 149 | 59 |
| Bad | 49 | 43 |

Table \_\_ Decision Tree Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest Confusion Matrix | | Predicted | |
| Good | Bad |
| Actual | Good | 179 | 29 |
| Bad | 49 | 43 |

Table \_\_ Random Forest Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Naïve Bayes Confusion Matrix | | Predicted | |
| Good | Bad |
| Actual | Good | 186 | 22 |
| Bad | 47 | 45 |

Table \_\_ Naïve Bayes Confusion Matrix

Cost Calculation:

Logistic Regression = (0\*190) + (1\*48) + (5\*18) + (0\*44) = 138 Units

Decision Tree = (0\*149) + (1\*59) + (5\*49) + (0\*43) = 304 Units

Random Forest = (0\*179) + (1\*29) + (5\*49) + (0\*43) = 274 Units

Naïve Bayes = (0\*186) + (1\*22) + (5\*47) + (0\*45) = 257 Units

As a result, it appears that Logistic Regression is the best model to reduce cost for the bank.