LECTURER: DOORUJ RAMBACCUSSING

BUS51039 PREDICTIVE AND PRESCRIPTIVE ASSESSMENT

NWAKAMMA LOVETH OGECHI

INTRODUCTION:

Banking is a fiercely competitive industry, and the Portuguese retail bank is always looking for new and innovative ways to keep their current clientele while also luring in new ones. To create a predictive model that can precisely anticipate if a customer will subscribe to a new line of products is the aim of this study in this context. By using this approach to target those who are more likely to subscribe, the Portuguese retail bank will be able to increase the efficiency of its marketing operations. The development of the bank and its profitability depend on how well this model works. The bank can more effectively identify interested individuals by creating a predictive system, preventing the waste of its resources on consumers who are less likely to use its services. This will assist the bank in increasing the effectiveness of its marketing campaigns, raising the conversion rate of prospective clients, and conserving critical resources.

The Main Problems at Hand

The major challenge is creating a predictive algorithm that can correctly forecast whether a person will subscribe or not. The goal is to create a system that uses client demographics and historical activity to forecast the possibility of a subscription.

The Algorithms/Methods the Report Uses

Three machine learning algorithms will be used to accomplish this goal. (Decision tree, Random Forest, and Logistics regression). These have been used extensively to forecast consumer behaviour and have produced encouraging findings in research of a similar nature. Based on the banks' marketing data that we have been given, we will train each algorithm and assess its effectiveness. The final prediction model will then be chosen using the method that offers the best accuracy and precision, in this case, the Random Forest Model.

METHODS:

The models utilised in this study are the following: Logistic Regression, Decision Trees, and Random Forest. This report will be analysing three distinct machine learning models and apply these models on the train data on RStudio. We start by getting a summary of the training data that is provided. From this, we can see that the customers who were most likely to subscribe to the package totalled **3728** successful subscribers and **27920** unsuccessful subscribers, indicating that only **11.78%** of the records are related to success. On the training data, we then proceed to execute the machine learning models (Logistic regression, Decision Tree, Random Forest) to the train dataset, and after implementing all three models on the test data we choose the Random Forest Model because it delivers the best Kappa, sensitivity, and specificity.

The Model features and Why They Work well.

NO	FEATURES	REASONS
1	**AGE**	younger customers may be less likely to subscribe than older ones due to lower disposable income
2	**JOB**	Higher-paid employees could have greater disposable income than low paid employees.
3	**POUTCOME**	If customers response to the prior campaign's was favourable, they are more likely to sign up for a subsequent campaign, and vice versa.
4	**CAMPAIGN**	Someone may be less likely to subscribe in the future if they have been approached often without responding in the past.
5	**DEFAULT**	Individuals with history of default are less likely to take on subscription than those without.

6	**EDUCATION**	Individuals with higher education may be more financially literate to take up a banking subscription than others less educated
7	**LOAN**	Individuals with an outstanding loan are less likely to subscribe than those without any loan
8	**CONTACT**	Method used to contact clients play a role in subscription or non-subscription
9	**BALANCE**	Individuals with high current account balance are more likely to subscribe than those with low balances
10	**DURATION**	How long the contact with a client last is more likely to determine if the client takes up a subscription or not

Other features listed in the Data do not function well because they are irrelevant to determining an individual's ability to subscribe or not subscribe, and they are irrelevant whether they are included as variables in various models or excluded because they have no positive bearing on the outcomes of the results produced.

Logistics Regression Methods and why: For the following reasons, the choice of logistics regression in a bank's marketing data can be a useful tool for estimating the likelihood that a consumer will subscribe to a good or service:

1. Binary outcome: When the dependent variable only has two possible values, it is appropriate to predict binary outcomes. Either way.

2. When the dependent variable and the independent variable have a non-linear relationship.

3. It offers odds that a consumer will subscribe to a good or service, which is helpful in making decisions.

4. They are simple to grasp and interpret.

5. It is resistant to outliers that can appear in actual data sets.

Decision tree and why: The decision tree was chosen because it can manage complicated data sets, including both category and numerical data, and it can handle missing values and outliers in the data, which are crucial for accurate predictions. Furthermore, it is simple to interpret because stakeholders who are not technically savvy can understand its visualised tree structure.

Random Forest and Why: The random forest model was chosen because it manages complex data, such as the data from our bank's marketing campaign, which has many attributes and intricate interactions between them. It also handles missing data in real-world datasets without the need for imputation. Additionally, it increases generalisation since it chooses the optimum subset of features to split on and builds several trees on the bootstrapped data samples, which decreases overfitting. It explains the value of the features, which is useful when choosing features and evaluating the model.

Measures of forecast evaluation used to derive the best model:

The confusion Matrix is used to evaluate the Binary classification forecast accuracy in this study. Confusion matrix provides a useful measure of forecast for random forest algorithm, particularly in classification problems by allowing an evaluation of the model’s performance on the test set and calculate various performance matric, and with the use of the Receiver Operating Characteristics (ROC) curve we get a more appropriate evaluation, which lead to the identification of the best model Random Forest.

RESULTS:

Logistic Regression Model:

The confusion matrix shows that there are **11,269** true negatives **844** true positives, **717** false negatives, and **733** false positive. The overall accuracy is **0.8931**, meaning that it correctly classifies 89.31% of the data. The confidence interval (CI) for accuracy is from 0.8849 to 0.8982, The Kappa statistic measures the agreement between the model's prediction and the actual values, and with a value of **0.4775** indicates a moderate agreement. The sensitivity of the model is **0.9389**, meaning it correctly identifies 93.89% of the positive cases (Class 1). The specificity is **0.5407** indicating it has a high false positive rate and it is not good at identifying negative cases (Class 0). The positive predicative value (PPV) is **0.9402** meaning when it predicts a positive result it is correct 94.02% of the time, The negative predictive value (NPV) is **0.5352** indicting it predicts negative results correct 53.52% of the time. The prevalence of the positive class is 0.8849 meaning 88.49% of the cases belong to the positive class. Overall, the model has high sensitivity but a low specificity, indicating that it may be good at identifying positive cases (Class 1) but not good at identifying negative cases (Class 0). The kappa statistic indicates the model performance is acceptable.

Decision Tree Model:

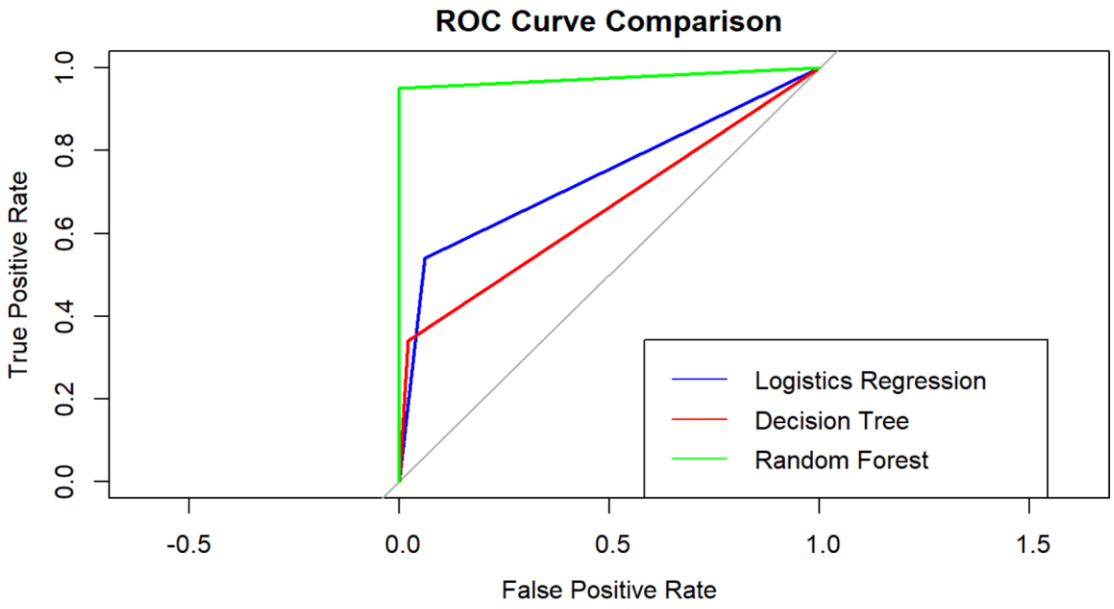
The Confusion Matrix shows that there are **11,760** true negatives, **533** true positives, **1029** false negatives and **242** false positives. The overall accuracy of the model is **0.9063** meaning the model classifies correctly 90.63%. The accuracy's confidence interval (CI) ranges from 0.9013 to 0.9111, The kappa statistic measures the agreement between the model's prediction and actual values and a value of **0.4107** indicates fair agreement. The sensitivity of the model is 0.798 meaning it correctly identifies almost all the positive cases. However, the specificity is only **0.3408** indicating the model has a high false positive rate and is not good at identifying

negative cases. The positive value (PPV) is **0.9195** meaning when it predicts a positive result it is correct 91.95% of the time. The negative predictive value (NPV) is **0.6873** meaning when it predicts negative results it is correct 68.73% of the time. The prevalence of the positive class is 0.8849, meaning 88.49% of the cases belong to the positive class. Overall, the model has high sensitivity but low specificity, indicating it may be good at identifying positive cases but not good at identifying negative cases. The Kapp statistic is also low indicating the model performance is only fair.

Random Forest Model:

The confusion matrix shows that there are **12001** true negatives, **1481** true positives, **80** false negatives and **1** false positive. The overall accuracy of the model is **0.994** meaning it correctly classifies data 99.4% of the time. The Kappa statistic measures the agreement between the model predictions and the actual values with a **0.97** indicating high agreement. The confidence interval (CI) for accuracy is from 0.9926 to 0.9953, The sensitivity of the model is **0.9999** meaning it correctly identifies almost all positive cases 99.99% of the time. The specificity is 0.9488 indicating it correctly identifies a high proportion of negative cases. The positive Predictive Value (PPV) is **0.9934**, meaning the models predicts a positive result correctly 99.34% of the time. The negative predictive Value (NPV) (TN/TN+FN) is **0.9993** meaning it predicts negative results correctly 99.93% of the time. The prevalence of the positive class is **0.8849,** meaning that about 88.49% of the cases belong to the positive class, The detection prevalence of **0.8907**, which is somewhat higher than the prevalence, is the percentage of cases that the model detected as positive (TP+FP)/TP+FP+TN+FN).Overall the model appears to perform very well with high accuracy, sensitivity, specificity and PPV. Making it the model of choice.

Comparison of Random Forest Results Against, Decision Tree, and Logistic Regression Model.



Evaluating the outcomes of the three models as well as the results from the Receiver Operating Characteristics (ROC) curve, The Random Forest model outperforms Logistics Regression, and Decision Tree, also in terms of accuracy, sensitivity, specificity, positive and negative predictive values, Kappa Value, and balanced accuracy. It is therefore the most effective model out of the three.

Corporate Purpose and How these numbers translate to revenue making operation for the bank.

To estimate the possibility that clients will subscribe or not subscribe to the Portuguese bank services the model that made the most accurate predictions was the Random Forest model. The output of the result suggests that there were 1485 accurate predictions, which means that if the bank made £10 for each new subscription. The bank's revenue will be [£10 * 1485 = 14,850] if the marketing staff successfully follows up and sign up the 1485 clients for the product. The

overall fixed cost is £4,000, with the cost of promoting the product being £2. the total cost of the campaign to each customer would be (£1485 * £2) + £4000 = £ 6970. The bank can still make a profit of (£14,850 - £6,970) = £7,880.

A profit certainty of £7,880 might be very important to the stakeholders when determining whether the campaign is a wise business decision, even though 1485 is a small portion of their observations, which total over 31,000.

Limitations of the study:

Because the study only examines data from the years 2008 to 2013, it has several limitations. The most accurate prediction of potential customer behaviour for a campaign in a more recent year cannot always be made using this older data because many socioeconomic factors, the effects of COVID 19 on marketing and customer earnings in recent years, and other changes that have impacted and influenced consumer behaviour in recent years must also be considered. To provide a more informative and pertinent forecast, it is advised that more recent data be gathered and used with the Random Forest Model.