Pranavi Lingamaneni

AA 5300: Advanced Analytics

Final Project

**INTRODUCTION:**

**1. Overview of the dataset:  
a) Contextual information:  
i. Source of the data.**

The bank marketing dataset is taken from UCI Machine Learning Repository. The source of the dataset is <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing> **ii. A brief description of objectives behind the collection of the data.**

The bank marketing dataset is related with direct marketing campaigns of a Portuguese banking Institution. The campaigns were based on the phone calls. The primary objective of this dataset is to predict whether the customers will subscribe to a term deposit or not.

The gathering of this data intends to aid researchers and analysts in better comprehending the variables that affect consumers' decisions to sign up for a term deposit and in the future development of more successful marketing strategies. **iii. The entity that collected the data.**

The entity that collected the data is the UCI Machine Learning Repository. **iv. Questions that audience interested in the dataset and its analyses might seek to see answered, etc.**

1. What is the overall success rate of the marketing campaigns in terms of customer subscription to term deposits?

2. What are the most important factors that affect the customer’s preferences in subscription to the term deposit?

3. What is the impact of different contact methods on the success rate of the marketing campaigns?

**b) Variables present:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Measurement type** | **Description** | **Role of the Variables** |
| Age | Numerical/Continuous | Age of the client | Predictor |
| Job | Categorical | Type of job | Predictor |
| Marital | Categorical | Marital Status | Predictor |
| Education | Categorical | Education | Predictor |
| Default | Categorical | Has credit in default? | Predictor |
| Housing | Categorical | Has housing loan? | Predictor |
| Loan | Categorical | Has personal loan? | Predictor |
| Contact | Categorical | Contact communication type | Predictor |
| Month | Categorical | Last contact month of year | Predictor |
| Day of Week | Categorical | Last contact day of the week | Predictor |
| Duration | Numerical/Continuous | Last contact duration | Predictor |
| Campaign | Numeric | Number of contacts performed during this campaign and for this client | Predictor |
| pdays | Numeric | Number of days that passed by after the client was last contacted from a previous campaign | Predictor |
| Previous | Numeric | Number of contacts performed before this campaign and for this client | Predictor |
| poutcome | Categorical | Outcome of the previous marketing campaign | Predictor |
| Emp.var.rate | Numerical/Continuous | Employment Variance Rate (Quarterly Indicator) | Predictor |
| Cons.price.index | Numerical/Continuous | Consumer Price Index (Monthly Indicator) | Predictor |
| Cons.conf.index | Numeric | Consumer confidence Index (Monthly Indicator) | Predictor |
| Euribor3m | Numeric | Euribor 3-month rate (Daily Indicator) | Predictor |
| nr.employed | Numeric | Number of employees (Quarterly Indicator) | Predictor |
| Y | Binary (Categorical) | Has the client subscribed a term deposit? | Outcome |

**2. Type of analyses:**

**a) A brief explanation of which analytical techniques are applicable for regression and why.**

The regression techniques may not be useful for the bank dataset as the outcome variable in this dataset is binary and regression techniques are generally useful to predict the continuous variables. This dataset contains a mix of categorical and numerical variables, and regression techniques are mainly useful for handling the numerical data. Using the regression techniques in this dataset may lead to overfitting since the dataset has large number of features. Therefore instead of regression techniques, classification models may be more useful for the bank marketing dataset.

**b) A brief explanation of which analytical techniques are applicable for classification and why.**

Classification techniques are especially beneficial for mixed data types. Due to the binary result and mixture of numerical and categorical variables in this dataset, classification models like decision trees, random forests, and support vector machines are helpful for this dataset. Also, when working with large dimensional datasets, classification approaches are more beneficial since they are less prone to overfitting than regression techniques. Moreover, non-linear correlations between the predictor variables and the outcome can be captured using classification approaches.

**ANALYSES**

**1. Overview:**

**In a table with three columns and one row per method, describe in sufficient detail:**

**a) Methods of analyses that are applicable.**

**b) For each method, an explanation of whether you intend, or not, to use the method.**

**c) Present concisely the rationale behind using or not using the method, within the context of your dataset, and what you know about the method’s strengths and weaknesses.**

|  |  |  |
| --- | --- | --- |
| **Method** | **An explanation for using or not using the method** | **Rationale for using or not using the method** |
| Decision Trees (Random Forest Model, Bagged Tree Model, Boosted Tree Model) | Yes, decision trees are useful for the bank marketing dataset as they perform well on large datasets. | Decision trees are useful for capturing the non-linear relationship between the predictor variables and the outcome. They can be an excellent place to start when creating more complicated models since they can offer insightful information about the variables that affect customer behavior.  **Strengths:** Interpretability, Handling the non-linear relationships.  **Weakness:** Overfitting of the data, can be unstable even if there are small changes. |
| Logistic Regression | Yes, Logistic Regression is useful for the bank marketing dataset as it can handle both categorical and continuous variables. | Logistic Regression indicates the strengths of the relationships between the predictors and outcome of the bank dataset. It can model the probability of the binary outcome based on one or more predictor variables.  **Strengths:** Simple and straightforward to use, it is useful for wide range of applications.  **Weakness:** Assumes the relationship between the predictors and the binary outcome to be linear, it can overfit the data. |
| Support Vector Machines | Yes, Support Vector Machines useful for the bank marketing dataset as it tries to find out the optimal decision for binary outcome. | Support Vector Machines performs well in high-dimensional spaces and in this instance, there are around 45,000 instances. SVM is also useful for modeling non-linear relationships between the predictors and the binary outcome.  **Strengths:** Effective in high-dimensional instances, it is suitable for large amount of data.  **Weakness:** Sensitive to parameter tuning, SVM is difficult to interpret. |

**d) If you have used clustering or dimensionality reduction, explain in what way this/these technique/techniques aided the model building process. If you have not used either of these approaches, explain why these methods were not used.**

I have used both clustering and dimensionality reduction for the dataset, but the results suggest that the dimensionality reduction is useful for the dataset, whereas clustering is not useful for the dataset.

**Clustering:**

From the results derived from clustering, we can observe that the WSS/TSS ratio is 1. The results explains that the clustering technique did not split the data effectively and the clusters are not separated distinctively. This shows that the clustering outcome might not be very instructive or useful for additional analysis or prediction.

**Dimensionality Reduction:**

A technique for minimizing the number of dimensions in data while retaining the bulk of its information is principal component analysis (PCA). Overfitting is thereby prevented, and computing complexity is reduced. If the original continuous variables and primary components are included, the same dataset may have issues with multicollinearity. Hence, I think the PCA is useful for my dataset.

**e) If you have used subsampling to obtain a reduced (rows) version of your dataset in order to achieve model fit in a reasonable amount of time, explain the details of the approach you have used. If you did not use subsampling, explain why that was not needed.**

The bank dataset comprises over 45,000 instances, making it undesirable to select a small sample of the data, which may prevent the application of subsampling techniques in this dataset. Because the dataset is uneven and only a tiny portion of clients subscribe to a term deposit, subsampling may also result in a sample that is biased and does not adequately reflect the underlying distribution of the data. The primary objective of this dataset is to identify the customers most likely to sign up for a term deposit, but by employing sub-sampling, information may be lost since certain situations might not be included in the research.

**2. Summary of results:**

**Create a table for each method that you have used (that is, if you have used three modeling techniques, you will include three individual tables, one per each technique), where you present:**

**a) Details of the validation method used (k-fold CV, preferably with repetitions, using Caret or hand-written k-fold CV code).**

**b) Model formulas of the various models you have fit using the particular method.**

**c) An explanation, using appropriate evidence, of model selection and evaluation measures used for identifying the best model, and determining the range of its applicability.**

**Modeling Technique 1: Logistic Regression**

|  |  |  |
| --- | --- | --- |
| **Validation method** | **Model Formula** | **Results** |
| 5-fold repeated k-fold CV using Caret | logisticfit <- train(y ~ .,  data = trainTransformed,  method = "glm",  #family = "family",  trControl = fitControl) | Accuracy: 0.8409  Sensitivity: 0.20000  Specificity: 0.92308 |

The evaluation metrics are Accuracy, Sensitivity and Specificity. Accuracy measures the overall correctness of the model’s predictions. Accuracy can be calculated by dividing the correctly predicted observations to the total number of observations. Sensitivity indicates the model’s capacity to recognize positive instances, which is calculated as the proportion of true positives out of the total positive instances. Specificity indicates the model's capacity to recognize negative cases, which is calculated as the ratio of accurately predicted negative instances to the total number of actual negative cases. All of the evaluation metrics needs to be higher for a model to be the best performing model. In this instance, accuracy(84.09%) and specificity(0.92308) are higher but the sensitivity is only 0.20000.

**Modeling Technique 2: SVM with Linear basis**

|  |  |  |
| --- | --- | --- |
| **Validation method** | **Model Formula** | **Results** |
| 5-fold repeated k-fold CV using Caret | svmlinearfit <- train(y ~ .,  data = trainTransformed,  method = "svmLinear",  trControl = fitControl,  verbose = FALSE, tuneGrid = grid) | Accuracy: 0.8636  Sensitivity: 0.00000  Specificity: 0.97436 |

The evaluation metrics are Accuracy, Sensitivity and Specificity. All of the evaluation metrics needs to be higher for a model to be the best performing model. In this instance, accuracy(86.36%) and specificity(0.97436) are higher but the sensitivity is only 0.0000.

**Modeling Technique 3: SVM with Radial basis**

|  |  |  |
| --- | --- | --- |
| **Validation method** | **Model Formula** | **Results** |
| 5-fold repeated k-fold CV using Caret | svmradialfit <- train(y ~ .,  data = trainTransformed,  method = "svmRadial",  trControl = fitControl,  verbose = FALSE, tuneGrid = grid) | Accuracy: 0.8864  Sensitivity: 0.0000  Specificity: 1.0000 |

The evaluation metrics are Accuracy, Sensitivity and Specificity. All of the evaluation metrics needs to be higher for a model to be the best performing model. In this instance, accuracy(88.64%) and specificity(1.0000) are higher but the sensitivity is only 0.0000.

**Modeling Technique 4: Random Forest Model**

|  |  |  |
| --- | --- | --- |
| **Validation method** | **Model Formula** | **Results** |
| 5-fold repeated k-fold CV using Caret | forestfit <- train(y ~ .,  data = trainTransformed,  method = "rf",  trControl = fitControl,  verbose = FALSE,  tuneGrid = grid) | Accuracy: 0.9027  Sensitivity: 0.32692  Specificity: 0.97750 |

The evaluation metrics are Accuracy, Sensitivity and Specificity. All of the evaluation metrics needs to be higher for a model to be the best performing model. In this instance, accuracy(90.27%) and specificity(0.32692) are higher but the sensitivity is only 0.32692.

**Modeling Technique 5: Bagged Tree Model**

|  |  |  |
| --- | --- | --- |
| **Validation method** | **Model Formula** | **Results** |
| 5-fold repeated k-fold CV using Caret | baggedfit <- train(y ~ .,  data = trainTransformed,  method = "rf",  trControl = fitControl,  verbose = FALSE,  tuneGrid = grid) | Accuracy: 0.9004  Sensitivity: 0.39423  Specificity: 0.96625 |

The evaluation metrics are Accuracy, Sensitivity and Specificity. All of the evaluation metrics needs to be higher for a model to be the best performing model. In this instance, accuracy(90.04%) and specificity(0.96625) are higher but the sensitivity is only 0.39423.

**Modeling Technique 6: Boosted Tree Model**

|  |  |  |
| --- | --- | --- |
| **Validation method** | **Model Formula** | **Results** |
| 5-fold repeated k-fold CV using Caret | boostedfit <- train(y ~ .,  data = trainTransformed,  method = "gbm",  trControl = fitControl,  verbose = FALSE,  tuneGrid = grid) | Accuracy: 0.9049  Sensitivity: 0.9625  Specificity: 0.4615 |

The evaluation metrics are Accuracy, Sensitivity and Specificity. All of the evaluation metrics needs to be higher for a model to be the best performing model. In this instance, accuracy(90.49%) and sensitivity(0.9625) are higher but the specificity is only 0.4615.

**CONCLUSIONS**

**1. Based on a comparison of the results from the modeling techniques you have employed, and the results of the associated “best” models, explain which modeling technique performed the best.**

All the models performed good but based on the 6 modeling techniques listed above, the best modeling technique would be Boosted Tree Model as the accuracy is 90.49%, sensitivity is 0.9625 and the specificity is 0.4615. Specificity is lower as compared to other 5 modeling techniques but the sensitivity is more as compared to other models. Accuracy is the key factor to find the best performing model. Comparing all the models accuracy, Boosted Tree Model accuracy is higher with 90.49%. Sensitivity is more rational as compared to specificity as we need to find true positives for a clearer picture. So, even the specificity is more for other models and less in the Boosted Tree Model, I preferred this model as the sensitivity is higher as compared to other models i.e., 0.9625.

**2. Provide a description of the results of the best model. Explain them within the context of your dataset, taking into account the assumptions and theory associated with the modeling technique. For example, if you found that a random forest model out- performed all other models built using several modeling approaches, explain why you think that is. Then, explain what the importance statistics/variation in parameter estimates associated with the model imply to a decision-maker.**

The best performing model is Boosted Tree Model. Boosted tree models are renowned for their excellent accuracy and capacity to handle complicated datasets and are frequently used for regression and classification problems. The approach is particularly helpful when the dataset has a lot of variables since the model can choose the most crucial variables for prediction automatically. Accuracy is the key factor to find the best performing model in the bank dataset as the cost of misclassification is relatively equal for both positive and negative cases. Comparing all the models accuracy, Boosted Tree Model accuracy is higher with 90.49%. Sensitivity is more rational as compared to specificity as we need to find the customer who will subscribe to the term deposit than the customers who do not. So, even the specificity is more for other models and less in the Boosted Tree Model, I preferred this model as the sensitivity is higher as compared to other models i.e., 0.9625. A model's statistics and variation in parameter estimations offer important clues about the model's precision and dependability. Understanding these data will enable you to weigh the model's usefulness and the accuracy of its forecasts with greater knowledge as a decision-maker. From the best performing model, a decision maker needs to take the values of accuracy and sensitivity for comparing the model's usefulness.

**3. Based on your understanding of the dataset and your analysis of it, what future work do you think will provider deeper insights into how the dataset can help a decision-maker who is associated with the context within which the dataset was collected?**

Some of the future works may include:

**1. Incorporating External Data Sources:** There are many variables that provide insights for the customer analysis but there are some external factors such as economic indicators, social media data, that could provide additional insights into customer preferences. Further research after including these factors can provide better analysis for the bank marketing dataset.

**2. Developing Predictive models with higher accuracy:** Even the boosted tree model provide the best possible results, there is always room for improvement. Further research can help us in the exploration of other machine learning techniques such as neural networks, to develop predictive models with higher accuracy.

**3. Exploring the relationships between different variables in the dataset:** Even though the bank dataset is extensively analyzed, there may still be some undiscovered relationships between the variables. Further research can help in identifying these relationships and their implications by the use of advanced statistical techniques.