Course Project
ITM 618

Denis Thai
501176832

OBJECTIVES

In this project, we seek to:

- Predict whether a customer will subscribe or not to a term deposit.
  - Using the target class “Subscribe”
- Compare Two Classification Models
  - Decision Tree Classification
  - Logistic Regression Classification
- Identify which model is most suitable for the task
  - Which model is more reliable?
- In the grand scheme of things, we seek to support our marketing decisions using the data provided.

The Data Set

Featuring Bank telemarketing data, the models are fed using a test data set, and a training data set.

- Training set consists of about 29,271 records
- Test set consists of about 11,917 records

Both sets feature attributes such as:

- Demographics (job, marital, education)
- Contact Details (duration, date)
- Economic indicators such as *nr.employed*

The FOCUS: Target Variable

The Target Variable for the project is: “Subscribed”

- Yields either 1 or 0, Yes or No.
- Highly indicative of imbalanced nature due to:
  - 89% No
  - 11% Yes
- Due to imbalance, expect impacts to precision, recall, and overall model behaviour.
  - To evaluate, ROC AUC & visual curve chart will be used to amplify findings.

Preparing the Data

Performed various data cleaning and preprocessing tasks such as:

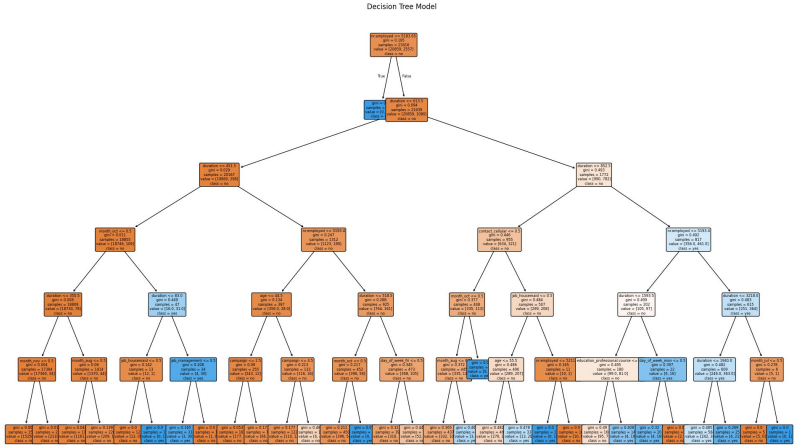
- Replacing “Unknown” values using mode
- One-Hot-Encoding used for categorical attributes
- Encoded the target variables, changing from { yes | no } to { 1 | 0 }
- Split data into three parts:
    1. Training
    2. Validation
    3. Test Set

Validation is needed to determine model reliability using ROC AUC.

Decision Tree

Key Features:

- nr.employed (IG is about 0.234 (The highest))
- Used max depth of 6.
- Used because it was an “Easy to interpret” model.
- Job categories

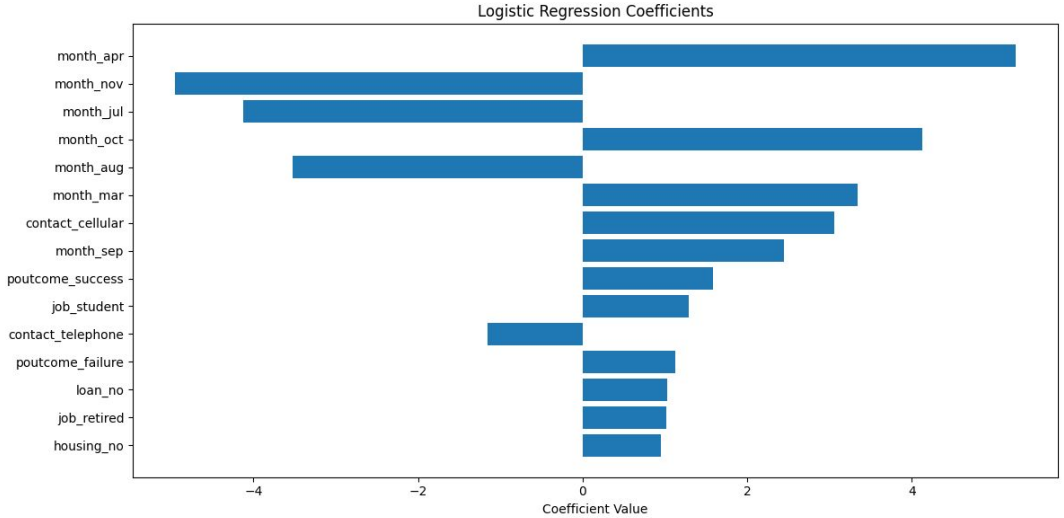


- **Decision Tree (Validation Performance)**
  - Accuracy: 0.9609
  - Precision: 0.8782
  - Recall: 0.7449
  - F1 Score: 0.8061
  - ROC AUC: 0.8661
  - Confusion Matrix:
    - [5150][66]
    - [163][476]

- **Decision Tree (Test Performance)**
  - Accuracy: 0.1427
  - Precision: 0.1238
  - Recall: 1
  - F1 Score: 0.2204
  - ROC AUC: 0.5123
  - Confusion Matrix:
    - [257][10216]
    - [0][1444]

Logistic Regression Model

- One-Hot Encoding
- Model used to find reliable predictive measures as compared to the decision tree.
- Strong Reliability metrics due to ROC AUC Comparison



a) Logistic Regression (Validation Performance)

- Accuracy: 0.9493
- Precision: 0.8701
- Recall: 0.6291
- F1 Score: 0.7302
- ROC AUC: 0.8088
- Confusion Matrix:
  - [5156][60]
  - [237][402]

b) Logistic Regression (Performance Performance)

- Accuracy: 0.4208
- Precision: 0.1452
- Recall: 0.7735
- F1 Score: 0.2445
- ROC AUC: 0.5729
- Confusion Matrix:
  - [3898][6575]
  - [327][1117]

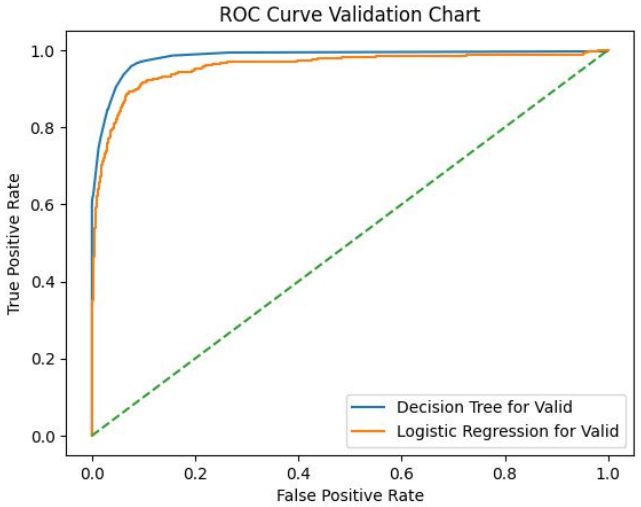
Evaluating The Models

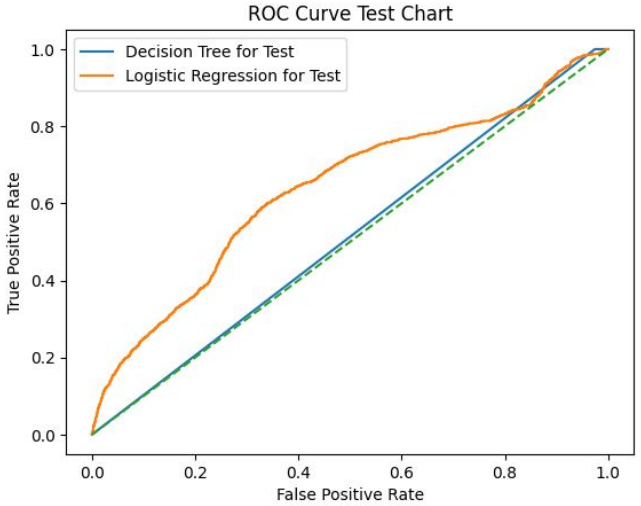
The statistics are given prior, but in this slide, it will highlight the results of the ROC AUC Curve test.

A key guideline: Closer to the top left is more ideal, and the closer to the diagonal line is less ideal.

The test data set (chart) is the most important as it represents real-world reliability.

It can be observed that the Logistic Regression chart is better than the decision tree in the charts in terms of reliability and consistency in predictive modelling for the test set.





Discussion

- Because the Decision Tree shows higher variance, it is indicative to the overfitting of training data.
    - Had a massive drop in performance from validation to test.
    - Predicted almost all “yes” on the test set meaning it was unrealistic.
- Logistic Regression provided more consistent and reliable results.

Key takeaway

- Performance hikes or drops between validation vs test sets highlight key factors in a model.
    - Such as reliability and application effectiveness.
    - The Regression although better than the tree model is not as ideal for the data set.
        - It also had slight overfitting.

Conclusion

Using ROC AUC to verify integrity of the models, cleaning data, tree and regression analysis, It was indicated that the stronger model to incorporate was logistic regression.

- In the future, it is much more wise to incorporate a more comprehensive data leaning method and model to support the data set provided.
- The Curve Charts provided a visual representation of which model is most effective.