Final Course Project

Marketing for the Banking System

Denis Thai
501176832

I. Introduction

In this final project for ITM 618, I've decided to take on the project individually. The project is set up so that I learn and apply methods of classification-learning, preprocessing, and cleaning data, with a focus on data exploration, and model testing. In this report, it will be approached by two methods. I will utilize decision trees and logistic regression as my two classification methods, and I will attempt to use labs, and the api libraries to help learn and apply their features to help: classify, model, and clean/preprocess data. Using the ROC AUC verification method, I will then be able to decide the most effective model for the given sets to predictively classify customers into who subscribe and who are not subscribed.

II. Data Exploration

Some precautions I set up in my code was to use debugging prints to highlight important data points. I have learned, through this, that the training data has 29271 samples, while the test data set has 11917 samples.

Each Data Set Represents:

- Customer Demographics
- Financial Attributes
- Employment Indicators
- Campaign-Related Features

Some key Observations:

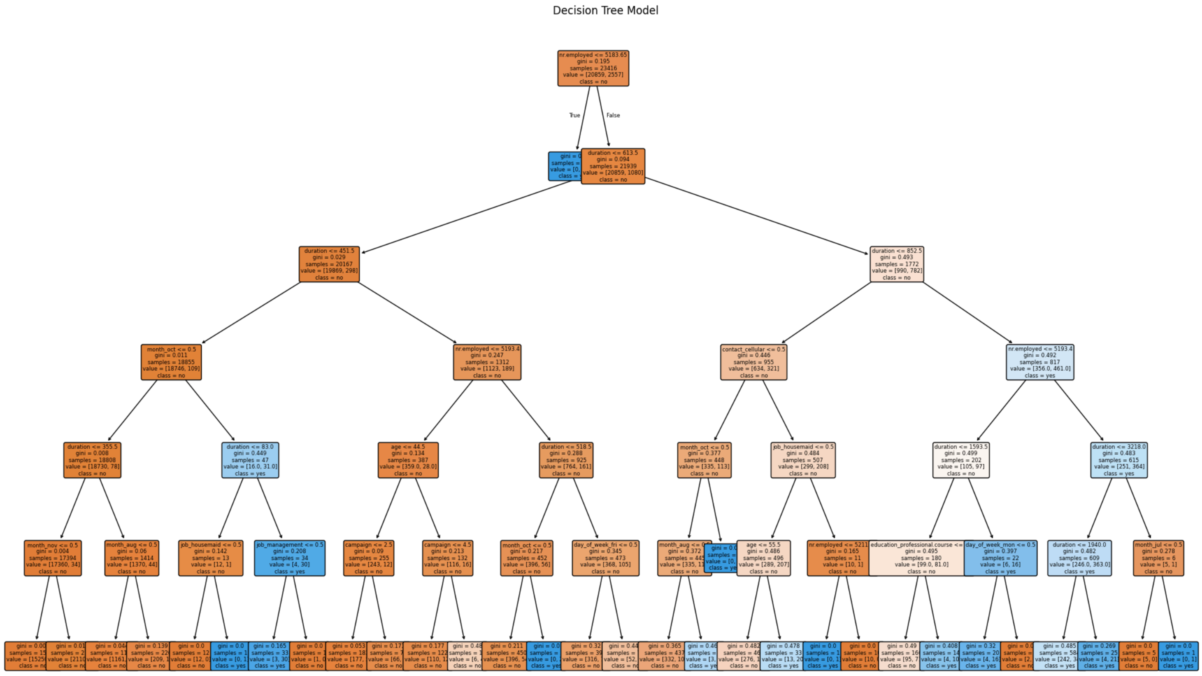
- Our target variable, "Subscribed" has 11% Yes, and 89% no– a very high discrepancy.
    - Given this, we can expect behaviors such as precision and recall being impacted.
- The most common job categories consist of:
    1. Admin
    2. Blue Collar
    3. Technician
- The most common education levels consist of:
    1. University Degree
    2. High School Diploma
- Throughout, there were a lot of categorical variables that were "unknown," therefore requiring a lot of data cleaning and preprocessing.
- In terms of Information Gain, I have found that among all attributes,
    - **nr.employed** attribute produced the highest information gain of approx. **0.234**.
        - This would make it a good attribute for the root split in the decision tree.
    - Duration showed high predictive power because of the pattern where longer calls may indicate higher interest in customers.

III. Learning Methods

a) Decision Tree (Supervised Learning)

I selected Decision Tree due to its “easy to use” graphics, where one just simply needs to follow the line and perform a bool check to decide whether an attribute applies or not, until reaching the end, resulting in a relevant solution.

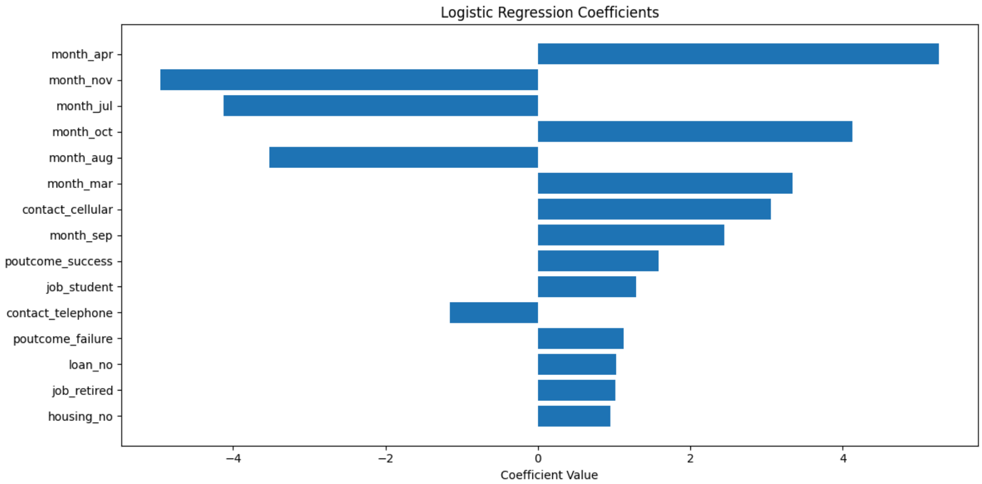
- Helps realize maximum Information Gain.
- Utilized One-Hot Encoding for categorical data.



b) Logistic Regression (Supervised Learning)

Logistic Regression was chosen based on familiarity with mathematically based graphs, and that because of the simplicity of the Decision Tree, the importance of Logistic Regression’s complexity is needed to provide further, and more accurate representation of the predictive data.

- Utilized One-Hot Encoding for categorical data.
- Less prone to overfitting than a deep decision tree.



IV. Evaluation

Evaluating was the most important part of this project. The methods/performance indicators of interest were:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC AUC
- Confusion Matrix

Validation Metrics are important to tell whether a model will perform on new data, and especially at ad-hoc levels of data, while also revealing any overfitting.

Below are the results for each model.

a) **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]

b) **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]

The drastic drop from the validation performance, and the test performance indicates a significant overfit. The collapse is alarming.

c) **Logistic Regression (Validation Performance)**
  - Accuracy: 0.9493
  - Precision: 0.8701
  - Recall: 0.6291
  - F1 Score: 0.7302

Continued

- ROC AUC: 0.8088
- Confusion Matrix:
    - [5156][60]
    - [237][402]

d) 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]

There was still a decline from the Validation Performance to the Test Performance. But because the Logistic Regression's decline wasn't as drastic as the Decision Tree's decline, it is generally the better model to use due to how well it retained the generalization of the data.

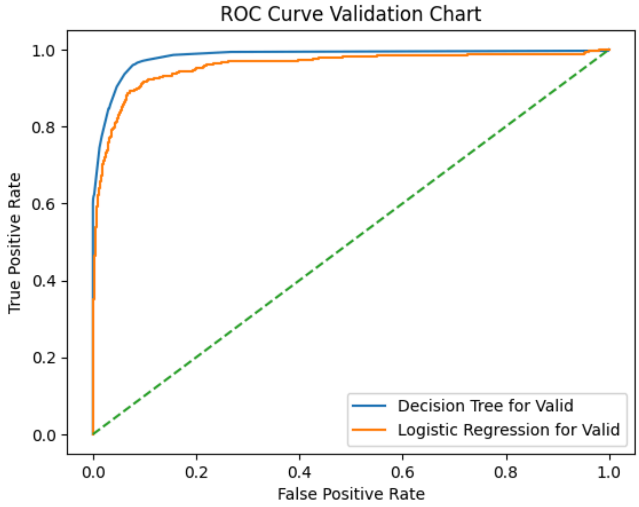
V. Discussion

The results highlight the difference between sampled and out of sampled data.

- On the validation (portion of the train set), there was clear cut superiority with the decision tree.
    - The downside was that it severely under-performed with the test set.
        - This severe under-performance indicates overfitting.
        - The limited tree (depth = 6 in the code) formed too many decision paths.
        - Too much skewing due to the majority of "no" in the responses.
- Logistic Regression generalized better versus the decision tree on the test set.
    - Prevented too much overfitting.
    - In the real-world, Logistic Regression is king.

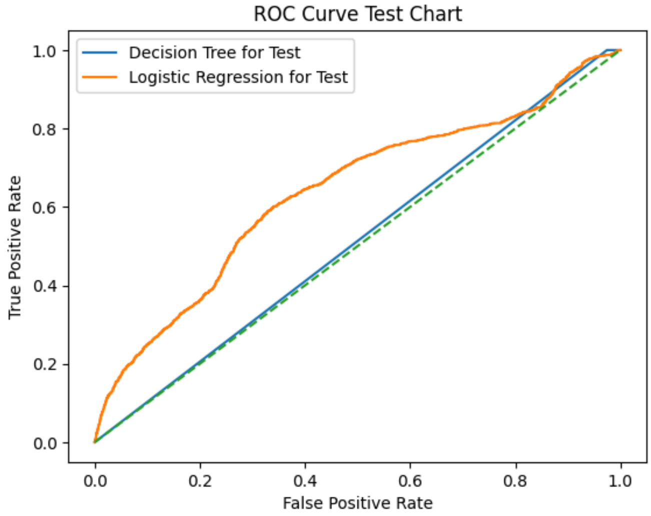
Continued

Utilizing the ROC curves, I was able to visually represent the discrepancies between the two models.



The Validation Chart shows that the higher curve of the Decision Tree indicates the effectiveness of the Decision Tree on the Validation data.

But on the Test Chart, it is observed that the Decision Tree curve isn't as consistent as the Logistic Regression curve. The drastic changes are not reliable predictive measures for a model used to predict.



These curve graphs are used in the project to further reinforce the reliability of the data and observations given prior.

In short,

- Closer to the Diagonal line = bad
- Closer to the Top left = good
- Consistency is key! (or reliability)

VI. Conclusion

Using methods of data analysis in the textbook and online, I've come to the conclusion that the:

- Decision Tree proved excellent on training set data/validation set data, but performed poorly on the test set.
  - Due to the importance and relativity to real-world application, the test set performance holds importance.
- The Logistic Regression method, although not superior in any way, showed true reliable performance across all data sets, achieving the highest ROC AUC score on the test set.

Thus, The **Logistic Regression** model should be deployed in future marketing campaigns, as it fits best for real-world applications/ best prepares for real-world unaccounted for data/outlying.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Chapman & Hall/CRC.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Matplotlib Development Team. (2024). *Matplotlib documentation*. https://matplotlib.org/

NumPy Developers. (2024). *NumPy documentation*. https://numpy.org/doc/

Pandas Development Team. (2024). *Pandas documentation*. https://pandas.pydata.org/docs/

Scikit-learn Developers. (2024). *Scikit-learn documentation*. https://scikit-learn.org/stable/

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media. https://doi.org/10.1007/978-1-4493-6132-7 (ISBN: 978-1449361327)

Sharda, R., Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective* (5th ed.). Pearson. (ISBN: 978-0134633282)

Miller, T. W. (2015). *Modeling techniques in predictive analytics with Python and R*. Pearson. (ISBN: 978-0133892062)

OpenAI. (2025). *ChatGPT* (Version 5.1) [Large language model]. https://chat.openai.com/