Christopher Rico
MSDS 422
Assignment 2
10/6/19

Introduction

In order to better understand the types of customers who subscribe to term deposits, a bank has undertaken the analysis of customer demographic data with regard to a recent telemarketing campaign. This will allow the bank to develop and validate a predictive model that will help target and market to future customers who are most likely to subscribe to a term deposit.

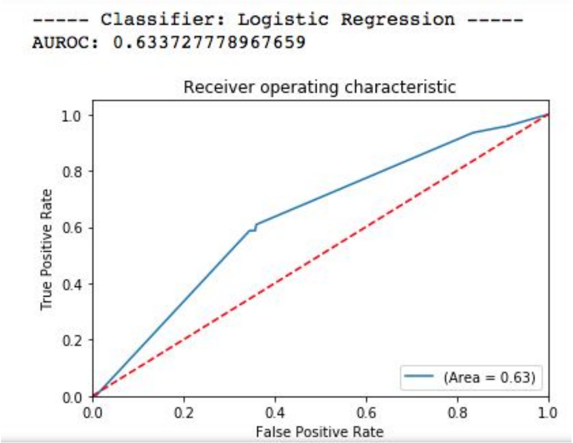
Discussion: Research Design and Statistical Methods

The dataset provided is a collection of about 4500 customers' demographic data from a bank telemarketing campaign. Each line represents one customer, with 17 variables that describe various attributes about the customer: their account balance, which day they were contacted, their vocation, etc. The dataset centers around the 'response' attribute, which measures whether or not the client has subscribed to a term deposit.

Unfortunately, 88% of customers in the dataset have not subscribed to a term deposit, so the objective of this analysis is to understand which attributes had the greatest ability to predict whether a customer will subscribe or not.

Programming Work

To run the analysis, the data were used to train different machine learning models from the Scikit-Learn library. The three models compared were gaussian naive bayes, bernoulli naive bayes, and logistic regression. Each model was then tested for its ability to predict the customer response and validated using a multiple k-folds method. Area under the ROC curve was used as a metric to compare the accuracy of the models.

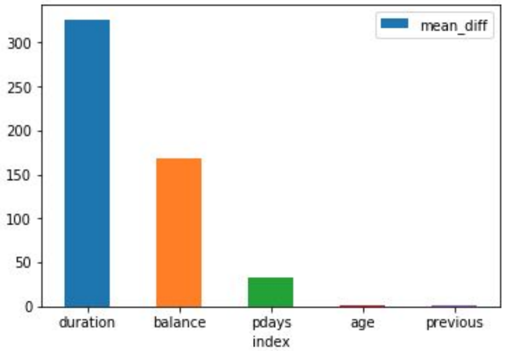


```
----- Classifier: Logistic Regression -----
AUROC: 0.633727778967659
```

10-fold cross validation outcome for suggested variables default, housing, and loan.

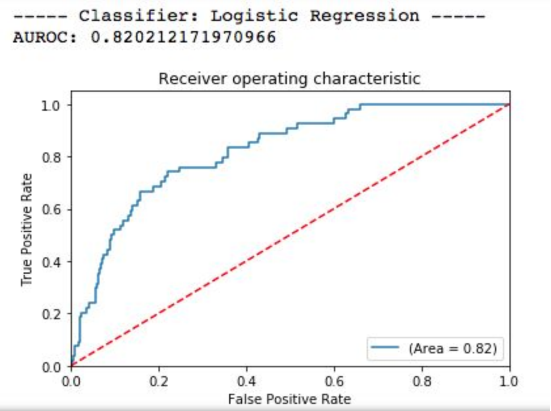
The model comparison was run twice: once using three suggested explanatory variables (*‘default*,’ *‘housing*,’ and *‘loan’*), although the exploratory data analysis showed that these three variables each had a low difference between means when grouped by response. It comes as no surprise, then, that the average AUROC for models trained with these variables as a predictor was fairly low. All three models score 0.59 - 0.61 on average.

Top 5 Variables with greatest mean diff in response



However, I was able to identify three variables (*‘duration*,’ *‘balance*,’ and *‘pdays’*) that have the highest difference between response means. This was an indicator that these three variables might play a more significant role in predicting customer response rates. After training the three models using these three variables as predictors, the cross-validation results suggest the same. The logistic regression model and the gaussian naive bayes models performed the best, with average AUROC values of 0.82 and 0.78, respectively. The bernoulli naive bayes model did not perform better using these predictive variables than it did with the other three variables, suggesting it is a poor fit in general for this data set.

10-fold CV outcome for vars duration, balance, pdays



```
----- Classifier: Logistic Regression -----
AUROC: 0.820212171970966
```

Results and Recommendation

My recommendation to the bank is to use a logistic regression model with the most highly predictive variables: duration of last contact, average yearly balance, and number of days since customer was last contacted. Based on this analysis, it would suggest that these variables paired with this model will allow for the bank to best predict customer responses.

A shared version of the interactive Jupyter notebook used to run this analysis can be found at:

https://colab.research.google.com/drive/1MPFKQ8sJXuy5xMTZjERjo-HLY6tfjBID