Assignment 2: Evaluating Classification Models by Mimi Trinh

Section 1: Summary and Problem Definition

A Portuguese bank wants its clients to invest in term deposits, so they conduct a study to identify factors that affect client responses to the new term deposit offerings. The analysis uses three predictor variables (default, housing, loan) to predict the binary response variable: Has the client subscribed to a term deposit?

Section 2: Research Design, Measurement, and Statistical Methods

The study first starts with an exploratory data analysis (EDA) to determine the shape of the dataset, whether there's missing value, and the class proportion of each of the four variables. Then, the team employs two classification models: logistic regression and naïve Bayes classification. Within a cross validation design, we use area under the curve (AUC) of the receiver operating characteristic (ROC) as an index of classification performance.

Section 3: Programming Work

Python Scikit Learn is the primary environment for conducting this research. The raw dataset in csv format is loaded into Python. The initial EDA shows that there are 4521 rows or observations and 17 columns or variables in the dataset. Among these 17, we only use four variables: default, housing, loan, and response. Also, the EDA shows that there's no missing value in the dataset. Then we use LogisticRegression() to fit the logistic regression model and BernoulliNB() to fit the naïve Bayes classification model. Three metrics are utilized to evaluate the models: accuracy using score(), confusion matrix using confusion_matrix(), and ROC AUC using roc_auc_score(). Among these, the last metric is the most important one, so we apply cross validation design with it using cross_val_score() automatically and KFold() manually.

Section 4: Results and Recommendations

The EDA results indicate that beside housing, all three other variables have an unequal class proportion (see exhibit 1 in the appendix). Specifically, the response variable only has 11.5% observations belonging to positive class, which may cause an issue later in the analysis.

When we fit the models, both logistic regression and naïve Bayes classification have an accuracy of 88%. However, since the response variable is highly imbalanced, this is not a good metric to evaluate. The confusion matrix for both models is also the same with 100% accurate true negative rate and 0% false negative rate. The models predict all observations as negative class due to the unequal class proportion in the response variable. Therefore, confusion matrix is not a good metric either since it forces the models to have a binary output of 0 (no) or 1 (yes).

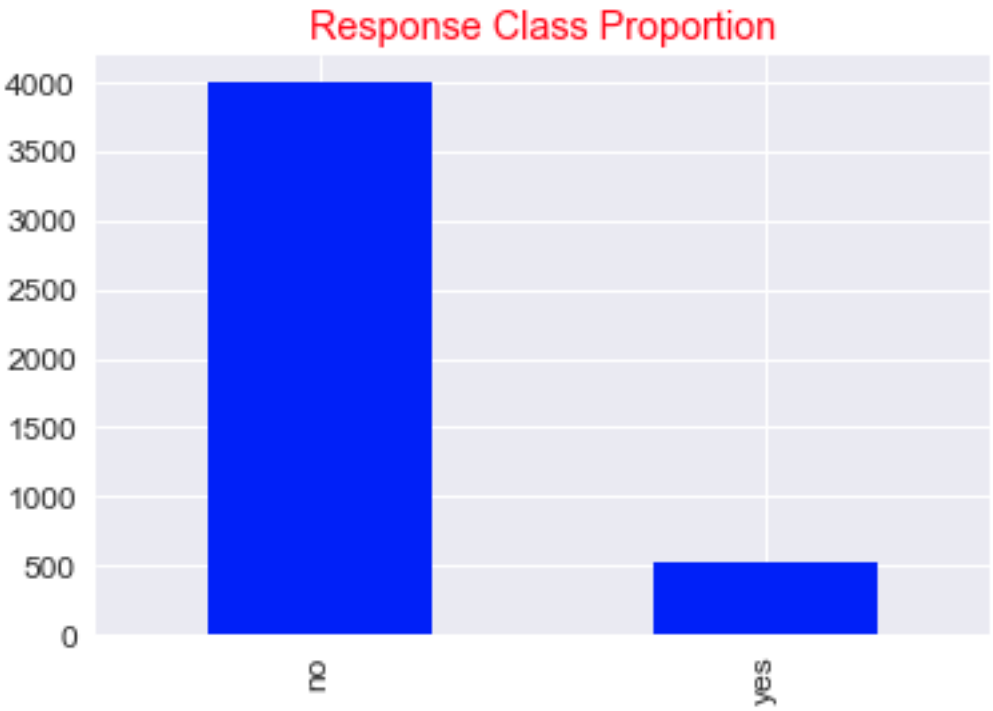
Meanwhile ROC AUC gives probability between 0 and 1, so it's a better index to evaluate model performance. Both models generate ROC AUC of 0.61, which is slightly better than random classifiers with ROC AUC of 0.5 (exhibit 2). Using 10-fold cross validation design both automatically and manually, the mean of 10 evaluation scores is also 0.61 for Logistic Regression and Naïve Bayes Classification, which shows poor performance from two methods.

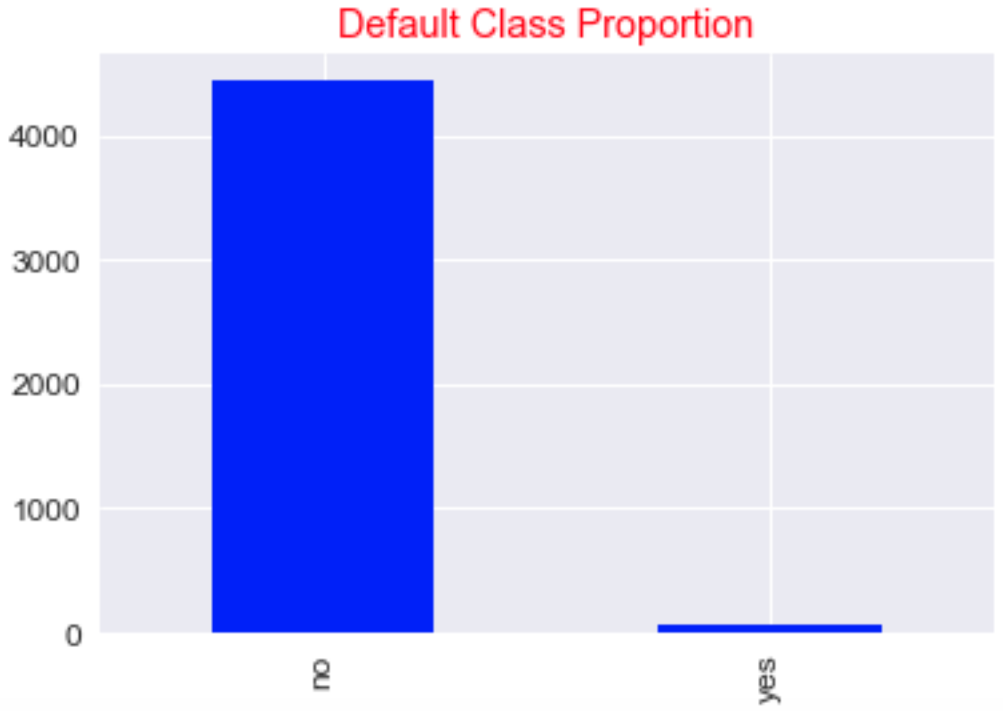
Both models provide the same result, so we can't recommend one over another. We applied the model on a dataset to predict the response variable. Marketers should target clients with predicted probability above 11.5% because that's the positive class proportion in the dataset used in this analysis. However, since ROC AUC shows that the models perform poorly, this study needs to be further developed to generate actionable results using below recommendations.

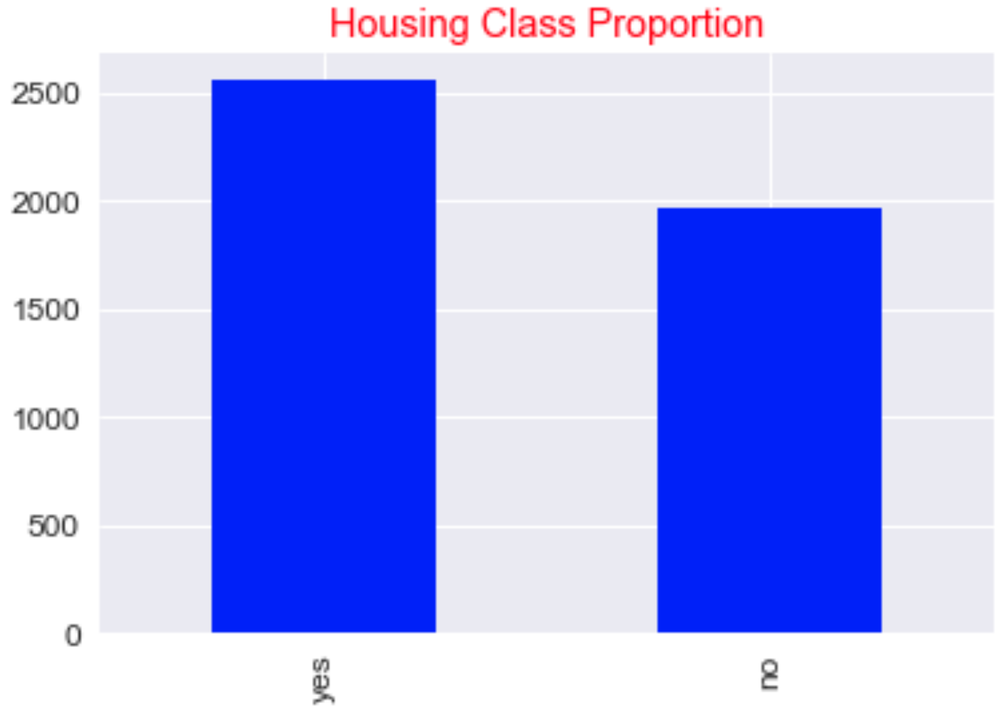
- Apply all 16 predictor variables instead of three (default, housing, loan) in modeling
- Use other statistical models that are more suitable for imbalanced response variable such as zero inflated Poisson and zero inflated negative binomial
- If the same issue still appears, gather more data and/or add more variables to modeling

Appendix

Exhibit 1




Default Class Proportion


Housing Class Proportion

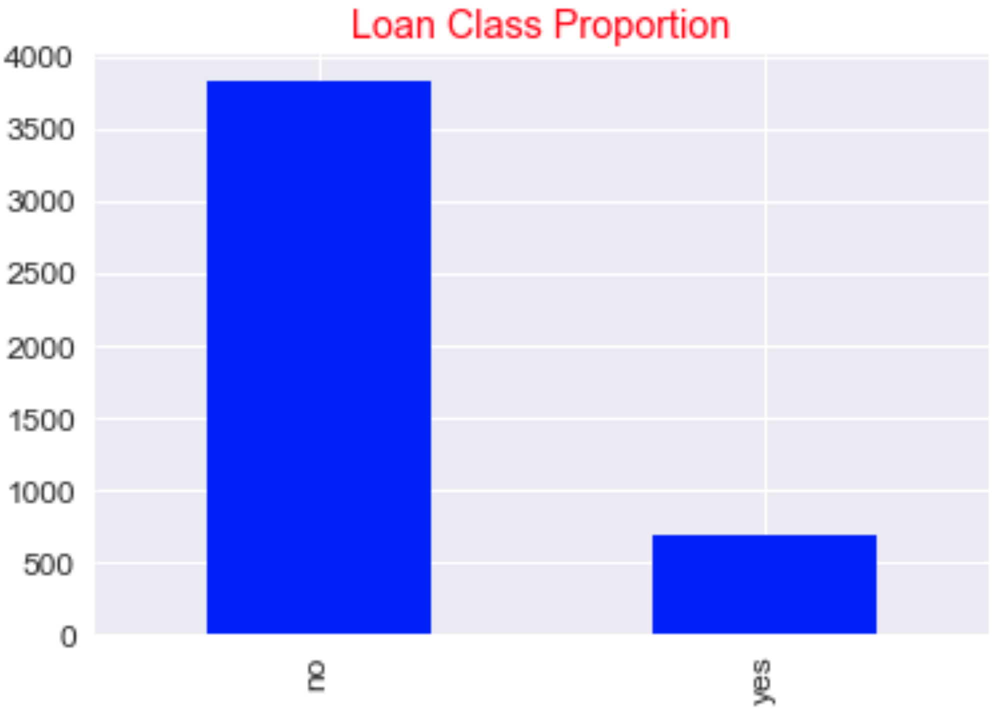
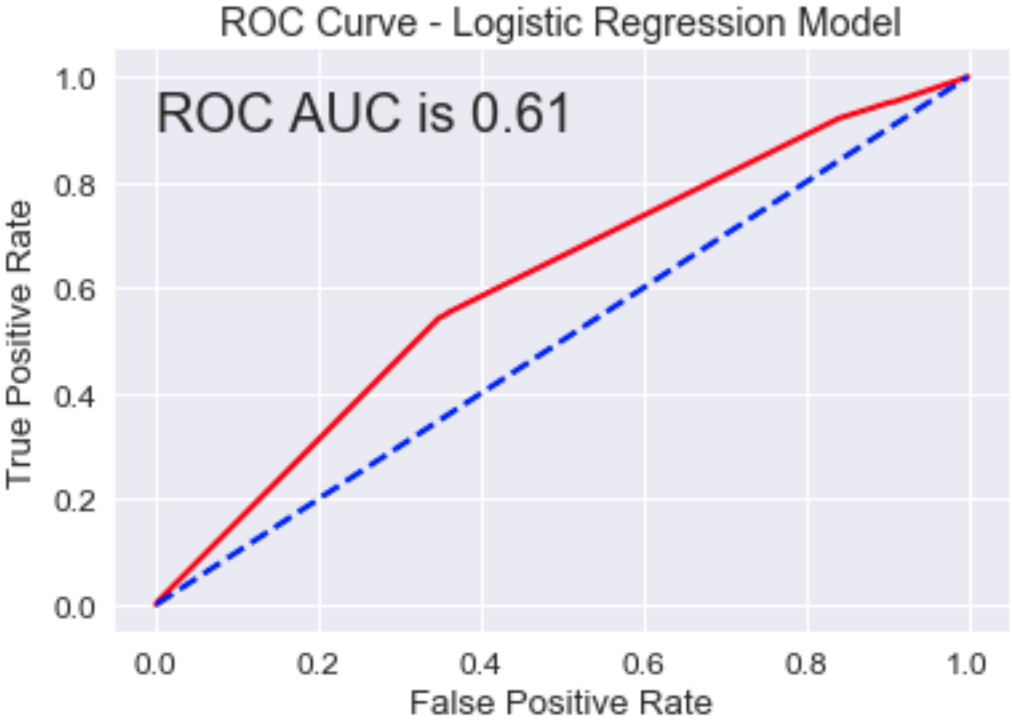
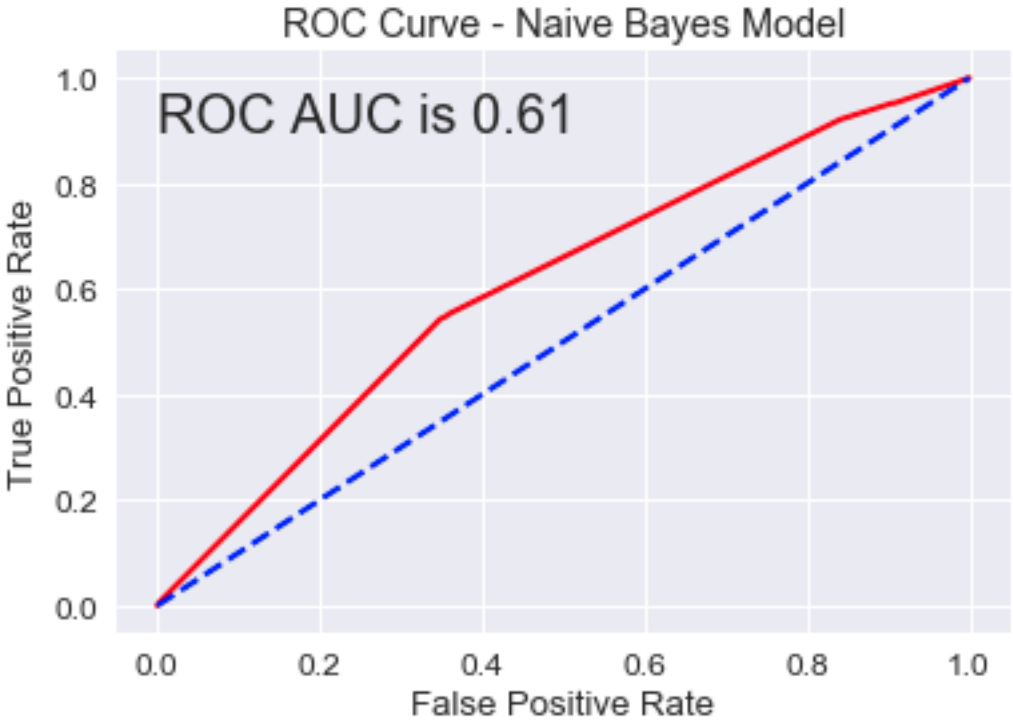

Loan Class Proportion
4000
3500
3000
2500
2000
1500
1000
500
0
no
yes

Exhibit 2


ROC Curve - Logistic Regression Model
ROC AUC is 0.61
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate


ROC Curve - Naive Bayes Model
ROC AUC is 0.61
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0