2018FA_MSDS_422-DL_SEC58

Practical Machine Learning Assignment 2

Prabhat Thakur

1. Problem Definition

A Portuguese bank conducted seventeen telephone marketing campaigns between May 2008 and November 2010 and recorded demographic, previous use of banking services and contact information for each client.

The bank is interested in identifying factors that affect client responses to new term deposit offerings, which are the focus of the marketing campaigns. What kinds of clients are most likely to subscribe to new term deposits?

2. Research Design and methods

The dataset provides the bank customers' information. It includes 4521 records and 17 fields which also includes the customer response (subscribed to term deposit). The target variable (response) has only two possible values Yes or No which makes this a binary classification problem. As per the assignment instructions, I used two classification methods: (1) logistic regression and (2) naïve Bayes classification. Also it has been suggested to use three binary explanatory variables relating to client banking history: default, housing, and loan to predict the binary response variable. I used cross-validation design to evaluate the accuracy of these methods and used the area under the receiver operating characteristic (ROC) curve as an index of classification performance.

Python Scikit Learn package was used for conducting this research.

3. Implementation and Programming:

For this analysis, I used Python programming language and different python packages like numpy, pandas and matplotlib for data preparation and visualization. Also used various methods from sklearn package for building machine learning classification models and their evaluation. The first step is to read the bank customer data from comma delimited txt file into python data-frame object. All three identified predictive variables and the target variables are categorical variables with only two categories so we converted them to binary values 0 and 1.

The starter code reads the customer data from file into data-frame and sliced to use only 3 suggested predictive variables and the target variable. All four variables are then converted to binary integer 0 and 1 and added to 2 dimensional array object so that it can be used in sklearn package. Two binary classifiers Naive_Bayes and Logistic_Regression are initialized and passed through k-fold cross-validation with 10 folds to evaluate classifiers performance using ROC curve AUC. Also for all possible combinations of three predictive variables, predicted probabilities is calculated using logistic regression.

I added some additional code to analyze all other customer data. Calculated mean of all numeric variables with respect to response, job, marital status and education. Also calculated confusion matrix, precision, recall and f1 scores for the logistic regression model. Generated Precision-Recall and Threshold plot, Precision Recall plot and ROC curve. Calculated confusion matrix using threshold greater than -2.26 to get non-zero precision, recall and f1 scores.

4. Findings and recommendations

From the data we can see that percentage of no subscription is 88.476 and percentage of subscription 11.524, this show that this task is an imbalanced classification problem. This is a

scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes. Classes should be balanced for efficient model.

The calculated performance matrices confusion matrix, precision, recall and f1 scores are all zero because of two reasons: target variable class is imbalance and the three predictor variables doesn't seem to be good predictor. **The model does not predicts any customer to subscribe for term deposit.**

If we have to use only default, housing and loan variables as predictor, based on the logistic regression model predictions, and we consider 10% probability as a good criteria to selected potential customer for campaign. Customer with following characteristics who have defaulted but don't have housing or personal loan are the best choice. Other two choices are customer with no defaulted, no housing and personal loan and customer with default, housing loan but no personal loan.

Recommendation: Campaign should target customers with following characteristic:

1. Defaulted, No Housing loan and No Personal loan for 18.8% success chance.
2. Not Defaulted, No Housing loan and No Personal loan for 16.5% success chance.
3. Defaulted, has Housing loan and No Personal Loan for 10.8% success chance.

Other customer characteristics should be further analyzed to explore better predictive variables.