Assignment_2_Wanat

A telephone marketing campaign was conducted between May 2008 and November 2010 by a Portuguese bank. During these calls, client characteristics were recorded and include demographic details, banking history, and marketing contact information. The bank would like to identify clients that are more likely to invest in term deposits offered by the bank.

The study variables were collected into a comma separated value file (.csv). The .csv file was opened in Python and saved into a data frame called bank. The data frame contains information from 4521 individuals on 17 variables. Three variables relating to client banking history were used to create a model to determine which group of clients are best targeted for direct marketing efforts for term deposits. The three variables used were: (a) default – does the client have credit in default; (b) housing – does the client have a housing loan; and (c) loan – does the client have a personal loan. The three variables, along with the response (has the client subscribed to a term deposit) were assembled into a new data frame called model_data_df. The 'yes' or 'no' answers recorded for these variables were changed to zero (0) for 'no' or one (1) for 'yes'. The data was split into 80:20 train:test sets.

Two classification methods were used to create models for comparison: (1) logistic regression and (2) naïve Bayes classification. Both are similar to each with slight differences in training efficiency and regularization. Regularization is used to constrain a model to limit the coefficient estimates used to create a simpler model and permit better generalization towards new data. The logistic regression model uses an L2 regularization by default, which constrains coefficient estimates to be as small as possible or be close to zero. Logistic regression also has a hyperparameter (C) that may be specified to control the strength of regularization. The higher the value of C, the less the model is regularized. The naïve Bayes classification will use a BernoulliNB classifier which is designed for binary (0 and 1) features. The BernoulliNB classifier has a smoothing parameter, alpha, which controls model complexity.

Models were evaluated for accuracy, sensitivity, specificity, precision, F1 score, class error, and false positive rate. A confusion matrix was generated, which counts the number of actual and predicted

'yes' and 'no' values for the response variable. The true positive rate (sensitivity or recall) versus the false positive rate (1 – specificity) is plotted in a receiver operating characteristic (ROC) curve. The area under the curve (AUC) allows a comparison of the classifiers to be made. The higher the AUC, the better the classifier performance. Cross-validation of the training data was used and scored by ROC-AUC.

A logistic regression model was created utilizing the three variables of default, housing, and loan. The C hyperparameter was set to a value of 100 to minimize regularization. The model was used to create predictions of the response variable and this was compared to the true values for evaluation. Due to the imbalance in observations in the response variable, the low rate of subscription to a term deposit (521 out of 4521 clients), the model performs poorly and is unable to predict a response value of 1 when the default classification threshold is set to 0.5. It is recommended that the classification threshold is lowered to a value of 0.1 in order to achieve a sensitivity for a response value of 1. Another logistic regression model was created with the C hyperparameter set to 1000, but this obtained the same results with the C set to 100 and was not used. A naïve Bayes classification model with the BernoulliNB classifier was created with the alpha smoothing parameter set to the default value of 1.

Based on cross-validation results, the logistic regression model had a slightly better AUC (0.604) compared to the naïve Bayes classification mode AUC (0.603). The logistic regression model is recommended for use based on the higher AUC. Clients with credit in default appear to be the best target for direct marketing efforts for term deposits. Clients with credit in default have a 2% increase in having a term deposit over clients with no credit in default. Clients with a housing loan have a 40% decrease in having a term deposit and clients with a personal loan have a 9% decrease in having a term deposit compared to clients with no housing or personal loan.