NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Bank Marketing Study

Ranaa Ansari, Daniel Arenson, Yining Feng, Han Nguyen

Northwestern University

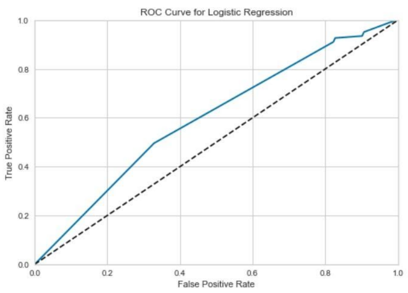
MSDS-422 Assignment 3: Evaluating Classification Models

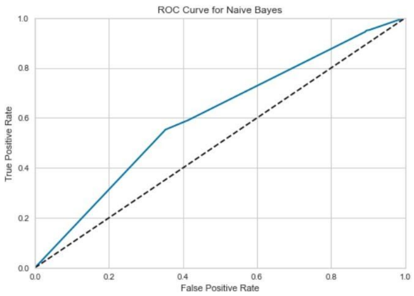
Summary and problem definition for management

We will be exploring the data and implications of the Bank Marketing Study. This data set was originally utilized by a portuguese bank to identify factors that affect client responses to new term deposit offerings, which are the focus of the telephone marketing campaigns. It contains data for 17 different variables that include the client's demographic characteristics, current contact information and previous use of banking services. Our group's focus will be the prediction of the binary response variable indicating whether the client will subscribe to a term deposit, based on three binary explanatory variables relating to client banking history (default, housing, and loan) by employing the logistic regression and naïve Bayes classification methods. For this project, our goal is to evaluate these methods within a cross-validation design using the area under the receiver operating characteristic (ROC) curve as an index of classification performance, and provide a recommendation on which classification method to use.

Research Design

The Bank Marketing dataset consisted of 4,521 client records, with each row containing factors that affect client responses to new term deposit offerings. The objective of the research was to examine 3 binary explanatory variables (default, housing, and loan) that describe client banking history, and apply logistic regression and naïve Bayes classification methods to predict the client's likelihood of subscribing to a term deposit based on these 3 binary explanatory variables.





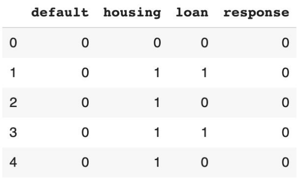
Logistic regression and naive Bayes classification models were conducted on the selected explanatory variables. While the naïve Bayes classification model is based on Bayes Theorem with the assumption of all variables as conditionally independent, the logistic regression model splits feature space linearly, and typically works reasonably well even when some of the variables are correlated. (Ottesen, 2017) We used the confusion matrix to evaluate the prediction accuracy of each model. The confusion matrix informed us that the logistic regression model had 3,194 correct predictions and 422 incorrect predictions in total, whereas the naïve Bayes model had 3,218 correct predictions and 398 incorrect predictions in total.

To evaluate model performance within a cross validation (CV) design, we first used resampling methods to split the data set into training and testing sets. Then we kept aside a sample/portion of the data which was not used to train the model, but was used for testing /validation.

We also selected the area under Receiver Operating Characteristic (ROC) curve as a model performance metric. The ROC curve plots the true positive rate (also called *recall*) against the false positive rate (FPR). The FPR is the ratio of negative instances that are incorrectly classified as positive. It is equal to 1 – true negative rate (TNR), which is the ratio of negative instances that are correctly classified as negative. The TNR is also called specificity. Hence, the ROC curve plots sensitivity (recall) versus 1– specificity. (Koehrsen, 2018)

Programming Work

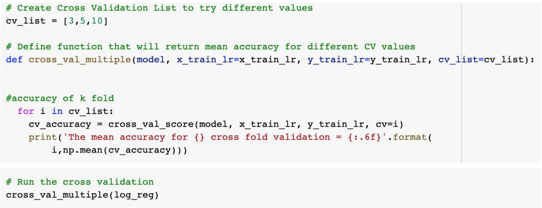
Initial exploratory data analysis was performed by taking the Bank Marketing Study dataset and exploring the data structure, data type, trends, and correlation among the sixteen provided attributes that describe clients' demographics and banking history. From those sixteen attributes, three were used as explanatory variables to predict the response variable and forecast whether a client would subscribe to a term deposit: (1) credit default history, (2) housing loan status, and (3) personal loan status; these three variables were converted to binary variables in order to apply the classification models.



	default	housing	loan	response
0	0	0	0	0
1	0	1	1	0
2	0	1	0	0
3	0	1	1	0
4	0	1	0	0

Logistic regression and naïve Bayes classification models were each trained with 80% (3616 records) of the sample data, and tested with the remaining 20% (905 records) by splitting the data with the `train_test_split` function and training the model with the `LogisticRegression` and `BernoulliNB` functions from the *scikit-learn* library. The performance for each model was evaluated by plotting the respective receiver operating characteristic (ROC) curve and calculating the area under the curve, using it as an index of performance.

Cross validation occurred on 3, 5, and 10 folds to determine the mean accuracy for the cross fold validation. In k fold cross validation, the data is divided into k subsets. The holdout method is repeated k times, such that each time, one of the k subsets is used as the test set and the other subsets are put together to form a training set.



```
# Create Cross Validation List to try different values
cv_list = [3,5,10]

# Define function that will return mean accuracy for different CV values
def cross_val_multiple(model, x_train_lr=x_train_lr, y_train_lr=y_train_lr, cv_list=cv_list):


#accuracy of k fold
  for i in cv_list:
    cv_accuracy = cross_val_score(model, x_train_lr, y_train_lr, cv=i)
    print('The mean accuracy for {} cross fold validation = {:.6f}'.format(
        i,np.mean(cv_accuracy)))
```

```
# Run the cross validation
cross_val_multiple(log_reg)
```

Results and Recommendations

The model ran against test data to view predictions and view probabilities. Finally, the ROC and AUC were evaluated again with the test data to determine the accuracy of the logistic regression classifier on train and train dataset. Both models displayed identical accuracy scores for the training and test sets, as well as identical AUC scores for both training sets. However, naïve bayes revealed a better AUC score, determining this was the better model for classification purposes.

Classification Models	# Cross Validation	**Accuracy**		**AUC**	
		Training	*Testing*	*Training*	*Testing*
Logistic Regression	10	0.88	0.89	0.605	0.467
Naïve Bayes	10	0.88	0.89	0.605	0.625

Although naïve bayes had a better AUC score in comparison to the logistic regression model, it was still a fairly low score. Considering, AUC ranges from 0.0 to 1.0 and both models resulted in 46.7% and 62.5% correct, respectively. This confirms that as far as predictive models go, the chosen variables of credit default history, housing loan status, and personal loan status would not be very useful in predicting whether the client would subscribe to a long-term deposit or not. To help improve the accuracy score on this predictive classifier we would recommend increasing the number of variables into the model. This could potentially help increase performance by training the data with more exposure across multiple variables and boost predictive power to the model.

References

Koehrsen, W. (2018, March 3). Beyond Accuracy: Precision and Recall. Retrieved April 25, 2020, from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Ottesen, C. (2017, October 24). Comparison between Naïve Bayes and Logistic Regression. Retrieved April 25, 2020, from https://dataespresso.com/en/2017/10/24/comparison-between-naive-bayes-and-logistic-regression/