Liya Zhang, Zhu Lyu

MA 346

Professor Carter

12/03/2020

Final Project Documentation

For this project, our team is trying to predict if a client will subscribe to a deposit due to marketing effort. In this case, we are looking to predict the result of campaigns of a Portuguese banking institution (y), if the client will subscribe to a term deposit. The logistic regression model we came up with is composed of only three variables documented by the bank: last contact duration (duration), euribor 3 month rate (euribor3m), and number of days that passed by after the client was last contacted from a previous campaign (pdays). The process that we took is being published on https://github.com/LiyaZhang-ziqing/Bank. Below are steps we take to predict the variable 'y'.

The source of this bank marketing dataset is from UCI machine learning repository. We are using this dataset by splitting into training and validation datasets, so that training dataset is used to build the logistic regression model. And the validation dataset would be used to access our model, which would be further used in real-life if the model has a good performance. The model we build is looking to target marketing to the population that are most likely to accept deposit subscriptions in order to save resources from banks.

We took several steps in getting the data ready. After we load the dataset, we first drop NA fields inside the dataframe. Then, we dropped fields like jobs that are categorical values and cannot be turned into numerical or boolean values, because those are the acceptable forms for values in logistic regression modeling. Afterwards, we replaced the rest categorical values into boolean numeric or numerical values. The data is therefore ready to create the model.

Because we need to train and validate our model, we split the data into 60%(training) and 40%(validation). Secondly, we create a model using the training dataset using the scikit-learn's logistic regression tool. That is achieved by building a function, so that if the dataset changes, a new model can be built quickly using the same logic. We also build another function to test the model using the validation dataset. The values returned are the F1 score from the training dataset and validation dataset. F1 score represents how good our model is by using the precision and recall values. We can see that both F1 scores for training(0.5037) and validation(0.5112) datasets are relatively low. Possible reason is that we have a high number of predictor variables. It's likely that some predictors do not predict our response variable well.

Consequently, we try to create a better model and aim to achieve higher F1 scores or to utilize less variables to build the model. We first fit the model like we did in the last step. Then, we created a list of coefficients sorted in order of their importance to predict the model. With the list, we start using the most important variables and adding the rest one by one to see if the F1 scores would improve. After sets of trial and error, we came up with the best set of variables in predicting the variable y: 'duration', 'euribor3m', and 'pdays'. Although its F1 did not exceed the original model, with F1 score about 3% lower, it only uses 3 variables in predicting the response variable. We decided to only keep these 3 variables to minimize correlation between predictors while maintaining a relatively good F1 score. By reducing 17 variables to 3, this will significantly reduce the resources needed to gather the data.

Lastly we decide to create a heatmap showing the visualization of the correlation between subscription status and its predictors. This visualization is being published on https://ancient-forest-16543.herokuapp.com/. The part that we would like to focus on is the last column of the heatmap. It's showing that none of the predictors have strong correlation with the response variable, which potentially explain the relatively low F1 score from the logistic regression modeling. Moreover, the correlations between each predictors are weak. This helps avoid the situation where similar variables are counting multiple times during the modeling process.

By taking both the bank's time and resources and the F1 score into consideration, we decided to keep: last contact duration (duration), euribor 3 month rate (euribor3m), and number of days that passed by after the client was last contacted from a previous campaign (pdays), in our logistic regression modeling. Although it doesn't have a strong performance, it does help the Portuguese banking institution to narrow down the predictor information that they need to gather to help predicting subscription status. This significantly reduces the amount of time and money that the bank needs to input by sacrificing a little bit of accuracy. There are areas for improvement once there are more strong predictors available in the dataset.