Judd Daniel Ang DEC 130 Section X Homework # 2

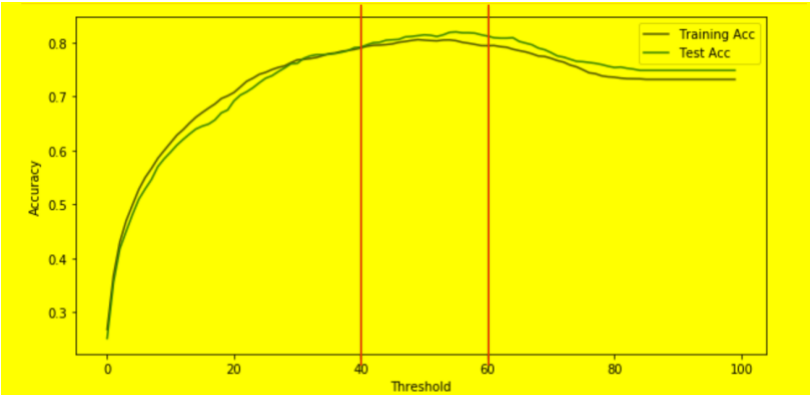
Preliminary Steps

Before running a logistic regression analysis on the training data set, I had to reclass the categorical variables into dummy variables. I opted to do this through Microsoft Excel because reclassing through Python yields columns with the data elements themselves as labels; this could be quite confusing as there are several categorical variables with “Yeses” and “Noes” for data elements. For most of the features (e.g. OnlineSecurity, OnlineBackup, and InternetService), I used “No” as the baseline for consistency and convenience. After obtaining the dummy variables, I then proceeded to split the training data set into 80% for training the model and 20% for testing the model. This was the crucial step that I missed out in the first homework which resulted in a low accuracy of churning predictions.

Fitting the Model

After going through the preliminary steps, the data was then subjected to a LR anal ysis using statsmodel in Python. The resulting slopes for each feature will be presented late r. From the results, it can be concluded that the presence or absence of fiber optic internet s ervice (i.e. InternetServiceDum2) has a significant effect on the odds of churning with an e^{Bi} of 7.809666, while the opposite can be said for two-year contracts (i.e. ContractDum2) with an e^{Bi} of 0.223614.

Furthermore, the cost of false negatives (i.e. Predicting that one will not churn, whe n in fact, the subscriber will most likely churn) is greater as this would mean a missed opport unity for the telecom company to intensify its promotional strategies targeted towards subscri ber retention. I set the threshold at 0.55 as this yielded training and testing accuracies 0.808 4943562272814 and 0.7903596021423106 respectively. By graphing the accuracies vs. thre sholds, it can also be seen that the greatest accuracies can be realized at values of the thres hold between [0.40-0.60]. At threshold = 0.55, the false negative rate is 0.07795907796 or a pproximately 8%.



Predictions

After fitting and testing the logistic regression model, I used the “result.predict()” function of statsmodels to predict the probabilities of churning; save for 2 customers (namely 4075-WKNIU, and 2775-SEFEE) whose data on total charges were missing, hence I was not able to obtain the probability predictions for them. Then with the help of some functions in Microsoft Excel, I was able to translate these probabilities into 1’s and 0’s to denote “Yes” and “No” respectively. The results of which can be seen through the attached .xlsx file.