Predictive Targeting for Personal Loan Acquisition: A Logistic Regression Approach for Universal Bank

Introduction

We were tasked with creating a smarter campaign with better target marketing for Universal Bank. We decided it best to proceed by using the bank's dataset on 5000 customers to construct a Logistic Regression model to determine which factors contribute most to converting customers to personal loan customers. The bank's dataset included data on the customers such as the age,zipcode, number of family members of the client, number of years of professional experience, their self-reported level of education, annual income, credit card average per month, value of house mortgage if applicable, and the type of account they held at Universal Bank if any.

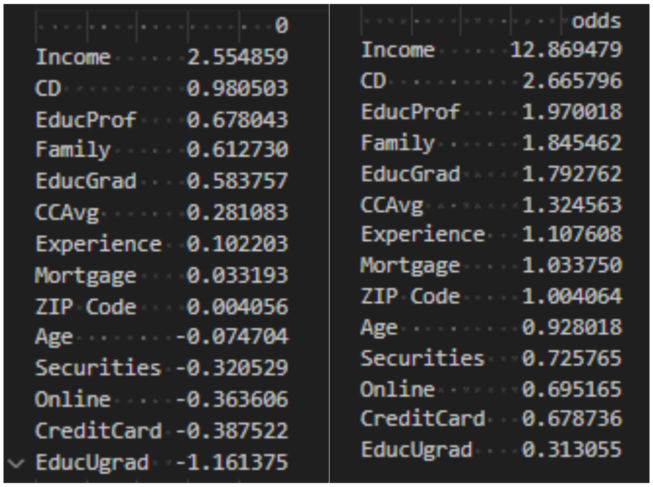
We first proceeded sanitizing and processing the 5000 customers data into the table included in our deliverables known as 'Better Table.xlsx'. We ran an SQL query on the original dataset to include all of the records of all customers. The SQL query is contained under the 'Better Table' query within the 'Final Project Modified'.accdb file. After this we then moved onto creating our Logistic Regression model.

Model Description

For the choice of model, we opted for a Logistic Regression model, as we were trying to predict the outcome of a binary outcome: Would this individual be converted into a Personal Loan customer or not. This type of question is best answered with a Logistic Regression model when considering many compounding and interactive factors. We used 'Better Table.xlsx' as our dataframe by reading it in using Python and performed a 50/50 split for the training set and test/validation set. This model is designed to predict the probability of a customer being a personal loan customer based on the information we have on them. The only change from the 'Better Table.xlsx' file and the data frame we created using Pandas was the dropping of the CustomerID column as it contributed nothing to the model. The details of the Logistic Regression model can be found in the file 'FinalProject.ipynb'.

The aforementioned dataframe contains both numerical and categorical variables (represented in our data as 1/0) for each of the customers within the original Universal Bank dataset. The numerical variables consist of variables such as Age, Family, Income, Average credit card spending. The categorical variables consist of variables such as Zip Code, EducUgrad, CD, etc. Before training the model with this data, we need to conduct data preprocessing, where we will treat the two types of data differently. For numerical variables, we applied standard scaling to normalize them. For the Categorical variables, they were already converted from their categorical strings in the original dataset to binary values via our SQL query from the 'Final Project Modified'.accdb file.

The model's coefficients tell us the relative importance of these variables. If a coefficient is higher (absolute value), it means that that variable has a stronger influence on the outcome. In the second screenshot to the right we can see that the odds ratios for the variables Income, CD, EducProf, Family, EducGrad, CCAvg, and Experience were all more than 1. This practically means that as any of these variables increase, the likelihood of a customer being a Personal Loan customer increases. For this reason, we can say that these appeared to be strong predictors to whether or not someone was more or less likely to be a Personal Loan customer.



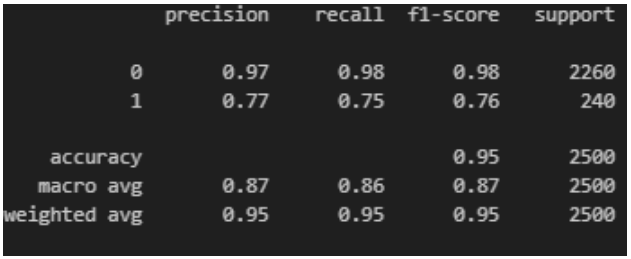
```
                   0
Income      2.554859
CD          0.980503
EducProf    0.678043
Family      0.612730
EducGrad    0.583757
CCAvg       0.281083
Experience  0.102203
Mortgage    0.033193
ZIP Code    0.004056
Age        -0.074704
Securities -0.320529
Online     -0.363606
CreditCard -0.387522
EducUgrad  -1.161375
```

```
                 odds
Income      12.869479
CD           2.665796
EducProf     1.970018
Family       1.845462
EducGrad     1.792762
CCAvg        1.324563
Experience   1.107608
Mortgage     1.033750
ZIP Code     1.004064
Age          0.928018
Securities   0.725765
Online       0.695165
CreditCard   0.678736
EducUgrad    0.313055
```

Validation Results

We first had to figure out how to optimize the performance of our model. We trained our model on the aforementioned 50% training set and then proceeded to use a standard precision recall curve to determine the optimal threshold for our model to maximize both the ‘accuracy’, ‘precision’, and ‘recall’. This terminology comes from a confusion matrix where we considered 4 outcomes, true positives, false negatives, false positives and true negatives. Accuracy was defined as the proportion of predicted results among the total number of observations, precision was defined as the proportion of true positives to all the predicted positives, and lastly recall was defined as the proportion of true positives to all the actual positives. We found from the precision recall curve contained in ‘FinalProject.ipynb’ the threshold for our model was approximately 0.35.

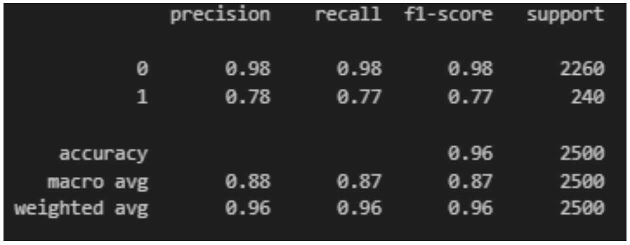
We then analyzed our model’s performance, first against the training data set (top screenshot on the right), and then against the validation data set (bottom screenshot on the right). In both of the data sets our model is rather accurate with 95 and 96% accuracy with its predictions for both non-personal loan customers and personal loan customers. We can see in both cases our model had little trouble predicting correctly non-personal loan customers, but our model only performed above average for predicting actual personal loan customers properly. However with relatively high precision of 0.77, 0.75 recall and a high accuracy, our model overall is performant enough to be considered competent when predicting our target of customers being a personal loan customer.



```
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      2260
           1       0.77      0.75      0.76       240

    accuracy                           0.95      2500
   macro avg       0.87      0.86      0.87      2500
weighted avg       0.95      0.95      0.95      2500
```



```
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      2260
           1       0.78      0.77      0.77       240

    accuracy                           0.96      2500
   macro avg       0.88      0.87      0.87      2500
weighted avg       0.96      0.96      0.96      2500
```

Conclusion

The purpose of this model was to determine what information about customers contribute significantly to whether or not a customer is a personal loan customer. By knowing this information Universal Bank would be able to properly market themselves to the right individuals to accumulate the most personal loan customers. Through the above analysis it is clear from our model’s performance and the factors outlined above as seemingly important we can suggest that Universal Bank should spend its time focusing on affluent customers primarily. Ideally these affluent customers would also either have an existing CD account at Universal Bank, have a professional level of education, have a high number of family members, have a graduate level education, have a high credit card average per month, have more professional experience, and have a mortgage. More succinctly: Universal Bank should focus on affluent, educated, professional customers with an existing cd account for their personal loan program. To us, this conclusion makes sense, as individuals that are going to participate in a personal loan program are those that are going to be able to pay it back with a stable and steady financial basis. The flexibility of a personal loan mirrors the financial flexibility of the customer that may take advantage of it, and thus it seems reasonable to us for Universal Bank to target these individuals outlined by our model and this report.