Did They Stay or Did They Leave

Ryan Fox, rfox2@bellarmine.edu

ABSTRACT

This logistic regression test aimed to predict churn for a ABC Multistate Bank. The term churn describes the rate at which customers leave. In other words, this test was used to predict whether a customer would churn/leave. Using Python, I was able to conduct three different test splits to predict churn.

I. INTRODUCTION

I obtained this dataset from Kaggle.com and used a logistic regression model. The variables in this dataset are customer ID, credit score, country, gender, age, tenure, balance, products number, credit card (yes or no), active member, estimated salary, and churn. Churn is the target value, and the customer ID was not used.

II. BACKGROUND

A. *Data Set Description*

As mentioned before, this dataset was found on Kaggle from a user named Gaurav Topre. I picked this dataset because I find financial data interesting due to being an Accounting and Finance major. The intention of this dataset was to use personal information provided by customers to predict whether they would stay at the bank or leave.

B. *Machine Learning Model*

The Machine Learning Model used was logistic regression. In this model, numerous variables were used to predict one qualitative variable. Categorical variables were turned into binary values of 1 and 0. 1 for yes and 0 for no. After converting these variables, it was able to predict whether a customer would leave the bank or stay a customer.

III. EXPLORATORY ANALYSIS

This dataset contained 12 columns or various kinds of variables that contained 10,000 samples. None of these variables contained missing/null values.

Table 1: Data Types

Variable Name	*Data Type*
Customer ID	Continuous
Credit Score	discrete
Country	Nominal
Gender	Binary
Age	Continuous
Tenure	Continuous
Balance	Discrete
Products number	Discrete
Credit card	Binary
Active member	Binary
Estimated salary	Discrete
Churn	Binary

IV. METHODS

A. *Data Preparation*

This dataset was fairly clean when I found it. However, I removed the column ‘customer ID’ because it is a number that simply does not affect whether a customer will stay at the bank. The only other cleaning process used was checking for null values but there were no null values in the dataset.

B. *Experimental Design*

Table X: Experiment Parameters

Experiment Number	Parameters
1	60/40 split for training and testing sets.
2	70/30 split for training and testing sets.
3	80/20 split for training and testing sets.

C. Tools Used

The following tools were used for this analysis: Python (jupyter notebook) running on Anaconda. In addition to base Python, the following libraries were used: Pandas, Numpy, Matplotlib.pyplot, Seaborn, and SKLearn. Why:

- Pandas was used to read the csv file which allowed me to work on the dataset within python.
- SKlearn was used for the logistic regression model as well as predicting values, a classification report and a confusion matrix.

V. RESULTS

A. Classification Measures

The results of the tests are provided down below in exhibits 1-3.

B. Discussion of Results

The largest amount in all three confusion matrices was true negative which is good because that means the customer did not leave the bank. The reports, however, produced a lot of false negatives which is bad because it would lead the bank to believe that they will retain more customers than they actually will. All three tests had an accuracy score of 80% meaning they were good/decent models at predicting churn.

C. Problems Encountered

One problem I ran into was that there was no variation in the results of the test sizes. Changing my test size did not improve the accuracy of my model. I also ran into a problem with a large amount of false negatives.

D. Limitations of Implementation

This model may not be the best way to represent this data, however I would not change my model. If I were to do this again I would like to add other variables such as marital status or if they moved housing locations. I think these variables would help make my model more accurate.

E. Improvements/Future Work

In the future, I would like to do these same tests but with a different bank. I think comparing large financial company banks to your everyday citizen banks like 5/3 would be interesting. I would still like to use a logistic regression model.

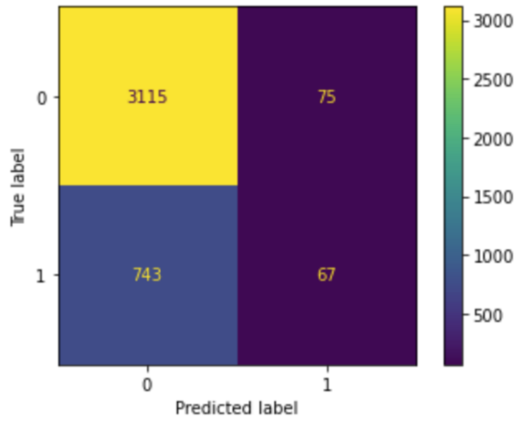
VI. CONCLUSION

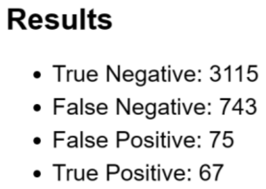
Overall, I think this was a good model. It was 80% accurate at predicting whether a customer would stay or leave the bank. This is much better than relying on no prediction at all and can begin discussions about how to improve customer retention. I think this dataset needs a couple more variables and more samples. In the future I would like to use data for a bank located in the United States because the financial system and liquidity here is much different than overseas.

REFERENCES

Bank Customer Churn Dataset | Kaggle

Exhibit 1: Test 1 split 60/40





Results

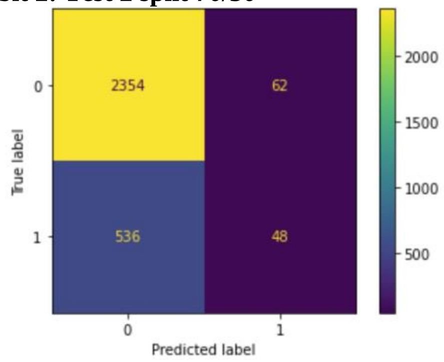
- True Negative: 3115
- False Negative: 743
- False Positive: 75
- True Positive: 67

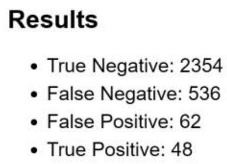
```
              precision    recall  f1-score   support

           0       0.81      0.98      0.88      3190
           1       0.47      0.08      0.14       810

    accuracy                           0.80      4000
   macro avg       0.64      0.53      0.51      4000
weighted avg       0.74      0.80      0.73      4000
```

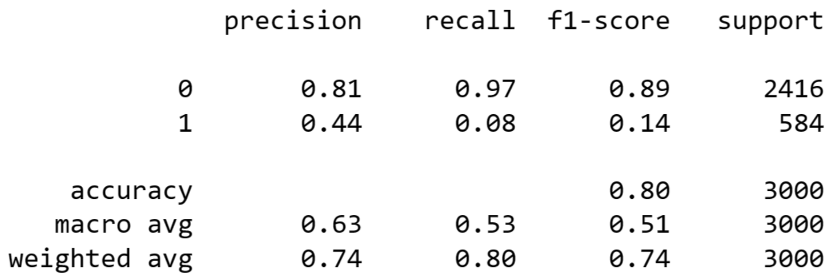
Exhibit 2: Test 2 split 70/30





Results

- True Negative: 2354
- False Negative: 536
- False Positive: 62
- True Positive: 48

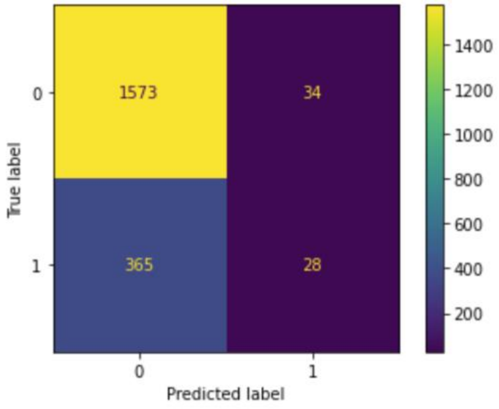


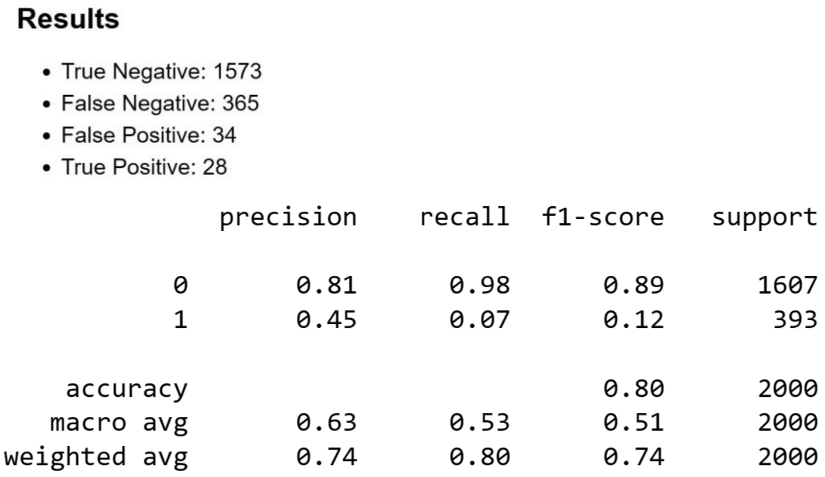
```
              precision    recall  f1-score   support

           0       0.81      0.97      0.89      2416
           1       0.44      0.08      0.14       584

    accuracy                           0.80      3000
   macro avg       0.63      0.53      0.51      3000
weighted avg       0.74      0.80      0.74      3000
```

Exhibit 3: Test 3 split 80/20





Results

- True Negative: 1573
- False Negative: 365
- False Positive: 34
- True Positive: 28

```
              precision    recall  f1-score   support

           0       0.81      0.98      0.89      1607
           1       0.45      0.07      0.12       393

    accuracy                           0.80      2000
   macro avg       0.63      0.53      0.51      2000
weighted avg       0.74      0.80      0.74      2000
```