Bank Customer Churning Logistic Regression

Liezel Dela Cruz, ldelacruz@bellarmine.edu

ABSTRACT
In this project, I obtained a dataset based on bank customer churning. I then performed logistic regression and exploratory data analysis using Python. The logistic regression analysis that was performed created a model that could be used to predict results, given specific input criteria. The value we want to predict is whether churning occurs or not. The validity of this model will be measured by calculating the MSE, RMSE, and the R-Square value. It will also be assessed through three experiments, a confusion matrix, accuracy score, and classification report. Overall, the model created using python gave a much larger mean square error than the model created using R.

I. INTRODUCTION

ABC Multistate Bank has a variety of customers who come and go. The dataset used for this logistic regression study contains information about this bank's customers. The aim is to predict whether a customer will churn given certain inputs. With relation to business, churning refers to the rate at which customers stops doing business with a company over a given period of time.

II. BACKGROUND

A. Data Set Description

The dataset, extracted from Kaggle, provides information about a proportion of customers from ABC Multistate Bank. Some of the data it exhibits are the customer's id number, credit score, gender, age, balance, and estimated salary. More importantly, it shows whether the customer churned or not; this is the variable we are trying to predict. I decided to work with this dataset because I have a personal interest in the banking industry, and I thought some data analysis on this would provide an increased insight as to what variables could affect customer satisfaction.

B. Machine Learning Model

Logistic regression is a type of statistical model often used to classify and predict variables from a dataset. Essentially, it predicts the likelihood of an event occurring based on a given dataset of independent variables. In this case, we are using 11 variables to predict 1 variable, churning or not. The model requires that the dataset is split into two – the independent and dependent variable. The model works by using the sklearn module to create a logistic regression object. The object has a method called fit() that takes the independent and dependent values as parameters and fills the regression object with data that describes the relationship.

III. EXPLORATORY ANALYSIS

The dataset contains 1000 row samples with 12 columns of various data type and some missing values. A complete listing is show below in **Table 1**.

Table 1: Data Types

Variable Name	*Data Type*
Customer ID	Int64
Credit Score	Int64
Country	Object
Gender	Object
Age	Int64
Tenure	Int64
Balance	Float64
Product Number	Int64
Credit Card	Int64
Active Member	Int64
Estimated Salary	Float64
Churn	Int64

IV. METHODS

Now, we will discuss how we prepared the data for our model and how we performed multiple experiments using different parameters for the model.

A. Data Preparation

To prepare the data using Python, I removed two columns that I thought did not serve many purposes for the logistic regression model. In this case, I dropped the customer ID and country column because I believe these two variables would not affect a customer's likelihood of churning. I also removed rows that contained null values in order to make the model as accurate as possible. Moreover, I encoded the categorical variable columns to make the cells numerical.

B. Experimental Design

Table X: Experiment Parameters

Experiment Number	**Parameters**
1	All features with 80/20 split for train and test
2	All features with 70/30 split for train and test
3	All features with 90/10 split for train and test

C. Tools Used

Python running in Anaconda environment was used for analyzing this data, particularly Pandas, matplotlib, seaborn, sklearn were employed to perform most of the operations. Pandas is a freely available Python package that is heavily used for data analysis and machine learning. Matplotlib is a python library that is used for data visualization in Python as its built in functions make the process much easier. Seaborn is another Python data visualization library that is actually based on matplotlib, but creates a higher-level interface leading to more finished looking statistical graphs. Finally, SKlearn is a Python data analysis library that is considered one of the best to use for machine learning.

V. RESULTS

A. Classification Measures

The classification measures used for each experiment were through a confusion matrix, classification report, and accuracy score. A confusion matrix describes the performance of the model based on a set of the test set, for which the data values are known. It provides a combination measurement of the predicted and actual values (true/false negative/positive). A classification report provides information on precision (the percentage of correct positive predictions relative to total predicted positives), recall (the percentage of correct positive predictions relative to actual positives), and F1 score (the closer to 1, the better). Finally, the accuracy score measures how often the model correctly predicts.

B. Discussion of Results

The third experiment (90/10 split) had the best accuracy score out of the three. The next highest accuracy score was the second experiment (70/30 split), therefore leaving the first experiment (80/20 split) with the lowest accuracy score. However, each experiment gave an accuracy score above 80%, which is not only ideal but shows that the model is realistic/consistent with industry standards. Moreover, according to my results, the third experiment also had the lowest MSE and RMSE, with 0.19 and 0.44 respectively. Alongside this, the third experiment had the highest F1 score with 0.89, which indicates this model was good at predicting whether or not customers would churn, since its value is very close to 1.

C. Problems Encountered

The main problem I faced when obtaining the data was trying to find a dataset that had a predictor variable that indicated yes/no or true/false because this was what I needed for a logistic regression model. A lot of the datasets I came across were more suitable for a linear/multiple linear regression model. Another issue faced was trying to find a good ratio to split the train and test sets, in order to get the best possible accuracy score.

D. Limitations of Implementation

The biggest limitation of our model are the variables that were used in the dataset. Arguably there could be other variables that have a greater impact on churning than any of the variables in our dataset. Therefore, models using other such variables could be more accurate in their modeling and prediction.

E. Improvements/Future Work

To improve our model in future work, there are several things that could be done. More experiments could be performed to see if a better model could be reached. A particular future course of action that would be very interesting to see would be add and/or remove variables. This could allow the model to become narrowed down to the most essential variables, possibly leading to a better model that could be very accurate at predicting the price. It would also be interesting to find a different dataset that seeks to predict the same thing, and then see it is a better model than the one we used.

VI. CONCLUSION

In this project, I sought to create a model to predict whether a customer churns or not, given several input variables. The results from the experiments indicated that the more data used to train a model the more accurate it will be. Overall, experiment three (70/30 split) had the best model, with the highest accuracy and F1 score.

REFERENCES

https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/#:~:text=There%20are%20many%20ways%20for,used%20metrics%20for%20classification%20problems.