Logo

Description automatically generated

ALY6020 Module 3 Midweek Project

**Authored by: Prateek Singh**

Course: ALY6020 – Predictive Analytics

Submission Date: 04/29/2022

**Instructor’s name: Kristen Drobnis**

# Introduction

Logistic Regression also referred as Binary models is a statistical model that predicts the probability of any event occurring or not. Hence, the output of the model will be binary with only two possible outcomes. In this week’s midweek assignment, we are given a task to predict to which customer a personal loan should be given based on the previous years data where the customers have either accepted or rejected the loan offered. In other words if the customer has accepted the loan in the previous campaign there are more probability of alike customers accepting the loan this year and hence our Target variable becomes “Personal Loan” which has binary values in it signifying if the customer has accepted the loan or not. In the entire dataset, 1 means True/Yes and 0 means False/No.

Even though our dataset has completely numeric values, we must be thoughtful of the fact that the binary categorical values have been converted and must be treated likewise. In a nutshell, we have a dataset of 5000 observations 14 variables. Out of these 14 variables, ID and Zip codes are nominal variables, Family and Education are ordinal categorical variables, Age, Experience, Income, CCAVG and Mortgage are interval variables, & CD Account, Security Account, Online Account, Credit Card and Personal Loan are our binary categorical variables.

# Analysis

The dataset doesn’t have any missing values present; however, our target variable was in the center of the file and to make our analysis easier I moved it to the end.

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Fig 1: Dataset Layout

A picture containing text

Description automatically generated

Fig 2: Null Value Count

We found during our data analysis that we have experience value in negative which is not right and must be corrected. Also, in the statistical analysis, Age and Experience variables look normally distributed with same mean and median values. Average age of the population is 45 years generally having 20 years of working experience. Rest of the variables data looks clean.

Table

Description automatically generated

Fig 3: Statistical Dataset Analysis of Numeric Series

Next step I wanted to visually see the relations between the variables and the distribution. As predicted above age and experience are normally distributed. Most customers age is in between 25 to 65 years. There is a very strong correlation between age and experience. Income, CCAvg and Mortgage are positively skewed with most customers average monthly credit card expenditure is between $700 to $2500. Majority people in the dataset have very less mortgage as well.

A picture containing shoji, building

Description automatically generated

Fig 4: Pair plot between all the variables

As we saw earlier, we have experience value documented as negative, hence I replaced it with the median values. To do so I first categorized the dataset into two portions. First with positive experience another with negative experience. Then I decided to check the same age and education group of people from either of the dataset to replace the negative values from the median values of those dataset. I believe education has a role in the experience level hence it was included in the equation.

Table

Description automatically generated

Fig 5: Cleaned Data Summary

Visually also the same outcome were received about the variables which were either normally distributed or skewed proving our thesis. We also saw that there are more number of customers from undergrad education level when compared with graduate and advanced level. More customers have moved to online banking however, maximum customers do not have credit cards. Most of the customers also do not securities account with the bank.

In our dataset of the previous campaign result we can see that 9.6% of the customers accepted the personal loan offered to them and we are interested to look what were the factors for these customers which drove them to accept the loan because this will help us understand the potential customers which can be targeted for banks next campaign cycle.

A picture containing text

Description automatically generated

Fig 6: Percentage of Customers accepted the loan

We can clearly see here that families having three and more children and higher income are more likely to accept loans.

A picture containing text

Description automatically generated

Fig 7: Family vs Income Plot

People from the same salary range are taking loans irrespective of the education level.

Chart, box and whisker chart

Description automatically generated

Fig 8: Education vs Income Plot

Customers with CD accounts mostly accept the personal loans.

A picture containing chart

Description automatically generated

Fig 9: CD Account holders vs Loans

Customers with credit card and high monthly average expenditure are more likely to accept the personal loan.

Chart, box and whisker chart

Description automatically generated

Fig 10: Credit Card Custmers vs AvgCC

Customers with high average credit card expenditure are more likely to accept the loans.

Chart, histogram

Description automatically generated

Fig 11: Distribution Relation between CCAvg & Personal Loan

Customer with higher incomes tends to accept personal loans when compared with low earning customers.

Chart, histogram

Description automatically generated

Fig 12: Distribution Relation between Income & Personal Loan

Correlation heatmap suggests a very high correlation between age and experience which we observed earlier as well. Income and CCAvg are moderately related (65%). Mortgages have a moderate correlation with income of about 21%. Personal loan has a maximum correlation with Income, CCAvg, CD Account, Mortgage, and Education. We can conclude by seeing the heatmap that Income does have an effect on CCAvg, Personal loan, Mortgage and CD Account variable.

A screenshot of a computer

Description automatically generated with medium confidence

Fig 13: Correlation Heat Map between variables

Finally, after analyzing the dataset statistically and visually, we can go forward to build our logistic regression model. We saw it earlier that ID and Zip Code does not have any impact on the model and hence can be dropped. We also predicted a high multicollinearity between the variables Age and Experience earlier and VIF score confirmed the same. However, the model accuracy wasn’t getting affected and hence both the variables were considered in the model.

Table

Description automatically generated

Fig 14: VIF Score to check multi-collinearity between the variables

We went ahead to fit our logistic regression model and found that almost all but Mortgage and Securities Account variables are our insignificant variables with p-value higher than 0.05.

Table

Description automatically generated

Fig 15: Logistic Regression Model

Model’s Confusion Matrix suggests that the model is quite accurate with successfully predicting results almost 94% times. Model correctly predicted positive scenarios 74 times when the customer is likely to accept the loan and negative scenarios 1337 where customer will not accept the loan offered. Our model inaccurately predicted instances where a customer is likely to accept the loan 30 times and not likely to accept the loans 59 times which when compared to True Positive and True Negative results is comparatively very small hence indicating good accuracy of our model.

Text, letter

Description automatically generated

Fig 16: Confusion Matrix and Accuracy of the Model

# Conclusion

* Type I and Type II errors in our model are relatively very less hence the model efficiency is good.
* Our model is 94.06% accurate.
* Three most significant variables in my model comes out to be CD Account, Education and Credit Card. Impact of CD Account and Education is positive however Credit Card adversely affects the case where customers are likely to accept an education loan.
* Customers with CD Account are 3.5 times more likely to accept the personal loans than the customers who do not hold CD Accounts with the bank.
* With increase in Education Level, we see those chances of customer accepting the loan increases by 1.7 times at each level.
* The most negatively impacting variable in our model, Credit Card, affects the rate of customers not accepting the loan by 1.17 times if they are a credit card user. In another words non credit card users are more likely to accept loans.

# Reference

1. CHOUDHARY, P. R. I. T. A. M. (2020, March 13). Bank personal loan modelling. Kaggle. Retrieved April 29, 2022, from <https://www.kaggle.com/code/pritech/bank-personal-loan-modelling/notebook>
2. cosine. (2021, September 14). Logistic regression using Statsmodels. GeeksforGeeks. Retrieved April 29, 2022, from <https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/>
3. Get the percentage of a column in pandas python. DataScience Made Simple. (2019, November 13). Retrieved April 29, 2022, from <https://www.datasciencemadesimple.com/get-percentage-column-pandas-python-2/>
4. Grosz, J. (n.d.). Implementing Logistics Regression. Canvas. Retrieved April 29, 2022, from <https://northeastern.instructure.com/courses/105748/pages/lesson-3-4-implementing-logistics-regression?module_item_id=7143789>

# Appendix

Note: Python Code is attached separately as Jupyter Python Notebook.