1. (4 pts) Import the bank loan default data set that has been provided in D2L. You will also find a Word doc that contains the description of the variables in the data set. In this exercise we will build a model to explain the probability of bank loan defaults.

2. (8 pts) Calculate summary statistics for all the numerical variables in the data set.

3. (8 pts) Tabulate all the categorical variables in the data set.

4. (4 pts) Does there seem to be enough variation in the categorical variables to build a reliable model for loan defaults?

Yes, almost all the categorical variables have good variance except the home loan variable

5. (6 pts) Estimate a multiple linear regression model for loan defaults using all the variables in the data set.

6. (4 pts) Export the regression results to MS Word using the stargazer library.

multiple regression using all variables to determine default rates

================================================

Dependent variable:

---------------------------

Default

------------------------------------------------

Checking\_amount -0.0003\*\*\*

(0.00003)

Term 0.014\*\*\*

(0.003)

Credit\_score -0.001\*\*\*

(0.0001)

GenderMale 0.008

(0.027)

Marital\_statusSingle 0.042

(0.026)

Car\_loan -0.079

(0.122)

Personal\_loan -0.147

(0.122)

Home\_loan -0.215\*

(0.126)

Education\_loan 0.043

(0.124)

Emp\_statusunemployed 0.052\*\*\*

(0.020)

Amount 0.0001\*\*

(0.00003)

Saving\_amount -0.0003\*\*\*

(0.00003)

Emp\_duration -0.0002

(0.0002)

Age -0.044\*\*\*

(0.002)

No\_of\_credit\_acc -0.010\*

(0.005)

Constant 3.423\*\*\*

(0.186)

------------------------------------------------

Observations 1,000

R2 0.660

Adjusted R2 0.655

Residual Std. Error 0.269 (df = 984)

F Statistic 127.320\*\*\* (df = 15; 984)

================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

7. (8 pts) Are there independent variables that exhibit a high degree of multicollinearity? Utilize the techniques that you learned in previous assignments (like correlation matrix, or VIF) to examine multicollinearity.





I didn’t find any Independent variables that exhibit high degree of multicollinearity.

8. (6 pts) Remove all the variables that are insignificant and that are problematic due to collinearity and estimate your final multiple linear regression model with the remaining variables.

9. (4 pts) Export the regression results to MS Word using the stargazer library.

Final Model

===============================================

Dependent variable:

---------------------------

Default

-----------------------------------------------

Checking\_amount -0.0003\*\*\*

(0.00003)

Term 0.016\*\*\*

(0.003)

Credit\_score -0.001\*\*\*

(0.0001)

emp -0.028

(0.019)

Saving\_amount -0.0003\*\*\*

(0.00003)

Age -0.047\*\*\*

(0.002)

Constant 3.583\*\*\*

(0.139)

-----------------------------------------------

Observations 1,000

R2 0.636

Adjusted R2 0.634

Residual Std. Error 0.277 (df = 993)

F Statistic 289.577\*\*\* (df = 6; 993)

===============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

10. (6 pts) Can the multiple linear regression model be used as a model of loan default probabilities? Illustrate the limitations of the linear regression model by making a few predictions that result in unexpected probability values.

If we are giving the model extreme values (scenario 3) and High credit score (scenario2) then the prediction is becoming negative. In scenarios 4,5,6 it is showing probability more than one.

If the predicted probability is less than 0 or greater than 1, it highlights the limitation of using linear regression for probability estimation.

QUESTION 3

1. (6 pts) Estimate a multiple logistic regression model for bank loan defaults, using the same variables that you used in your final multiple linear regression model in the previous question.

2. (4 pts) Export the regression results to MS Word using the stargazer library.

=============================================

Dependent variable:

---------------------------

Default

---------------------------------------------

Checking\_amount -0.005\*\*\*

(0.001)

Term 0.175\*\*\*

(0.047)

Credit\_score -0.012\*\*\*

(0.002)

emp -0.507\*

(0.306)

Saving\_amount -0.005\*\*\*

(0.001)

Age -0.632\*\*\*

(0.059)

Constant 39.544\*\*\*

(3.606)

---------------------------------------------

Observations 1,000

Log Likelihood -168.581

Akaike Inf. Crit. 351.162

=============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

=============================================

Multiple Logistic Regression for Loan Default

---------------------------------------------

3. (6 pts) Remove any insignificant variables to arrive at your final logistic regression model. Estimate your final logistic regression model to answer the questions below.

4. (4 pts) Export the final logistic regression model results to MS Word using the stargazer library.

============================================

Dependent variable:

---------------------------

Default

---------------------------------------------

Checking\_amount -0.005\*\*\*

(0.001)

Term 0.175\*\*\*

(0.047)

Credit\_score -0.011\*\*\*

(0.002)

Saving\_amount -0.005\*\*\*

(0.001)

Age -0.629\*\*\*

(0.059)

Constant 38.848\*\*\*

(3.511)

---------------------------------------------

Observations 1,000

Log Likelihood -169.985

Akaike Inf. Crit. 351.970

=============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

===================================================

Final Multiple Logistic Regression for Loan Default

---------------------------------------------------

5. (6 pts) How does the employment status of an individual impact the probability of their loan default?

Employment indicates a weak level of statistical significance. The P-value is less than 0.1, indicating that there is less than 10% probability that the observed data could have occurred by random chance.

6. (6 pts) What is the difference in the probability of a loan default for an individual with a 600 credit score vs. an otherwise similar individual but with an 800 credit score?

IT IS DIFFERENT IN DIFFERENT SCENARIO, SOMETIMES THE VALUE ARE MORE THAN 1, SOME ARE HUGELY DIFFERENT AND SOME ARE VERY CLOSE.

Question 4 (36 pts)

1. (8 pts) Using R, split the data set randomly into two parts: a training data set, consisting of 70% of the observations, and a testing data set, consisting of 30% of the observations.

2. (6 pts) Estimate a logistic regression model with the training data set using the same variables from your final logistic regression model in the question above.

3. (4 pts) Export the training set regression results to MS Word using the stargazer library.

=============================================

Dependent variable:

---------------------------

Default

---------------------------------------------

Checking\_amount -0.004\*\*\*

(0.001)

Term 0.219\*\*\*

(0.059)

Credit\_score -0.011\*\*\*

(0.003)

Saving\_amount -0.004\*\*\*

(0.001)

Age -0.651\*\*\*

(0.076)

Constant 37.542\*\*\*

(4.203)

---------------------------------------------

Observations 647

Log Likelihood -111.811

Akaike Inf. Crit. 235.622

=============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

=====================================

Logistic Regression for Training data

-------------------------------------

4. (6 pts) Use the model you estimated to predict the probabilities of default for the individuals in the testing data set.

5. (6 pts) Assuming a cutoff probability at 70% (i.e. if the predicted probability is greater than or equal to 70% the loan will be considered as default) create a classification table for your model.

6. (6 pts) Using the classification table above, calculate the accuracy rate of your model as the ratio of correctly predicted outcomes over the total possible outcomes.

0.9235127

Question 5 (16 pts)

1. (8 pts) What was wrong with the model that Amazon used to determine where they are going to offer Prime Same Day Shipping service? Briefly discuss based on the article posted in D2L.

* Amazon uses a data-driven approach to determine where to offer same-day delivery, focusing on areas with high concentrations of Prime members and proximity to fulfillment centers.
* This approach, while cost-effective, inadvertently reinforces long-standing racial and economic inequalities in access to retail services.
* The model overlooks racial demographics, leading to the prioritization of wealthier, predominantly white neighborhoods, while Black neighborhoods are often excluded from same-day delivery options despite paying the same $99 Prime membership fee. It perpetuates the system bias.
* In response to criticism from local leaders and community members, Amazon expanded its service to cover all ZIP codes in major cities like Boston, New York City, and Chicago.
* This situation highlights the need for decision-making processes that consider social and racial contexts, as the initial exclusions reveal how reliance on certain data can perpetuate systemic biases.

2. (8 pts) What are some problems associated with using ML algorithms to assist with decision making in the justice system? Briefly discuss based on the article posted in D2L.

* Algorithms can reinforce existing biases found in their training data, resulting in discriminatory outcomes. For instance, the Northpointe algorithm was discovered to flag Black defendants as potential future criminals nearly twice as often as white defendants
* It can be difficult for defendants and the general public to understand how risk scores are calculated because many algorithms do not disclose the methodologies used in their computation. This lack of transparency threatens the fundamentals of an open and equitable legal system.
* The inaccuracy of the algorithms in predicting future crime is also a concern. For instance, the Northpointe algorithm was only 61% accurate in predicting recidivism, and it was particularly unreliable in forecasting violent crime, with only 20% of those predicted to commit violent crimes actually doing so.
* Overreliance on these algorithms can lead to unfair sentencing decisions by Judges who rely on these models.
* These models are not backed by sufficient studies and evidences regarding their efficiency. Many are adopted without thorough testing, which can lead to systemic injustices.

The above are some of the concerns on using the ML algorithms to assist with decision making in the justice system.