**CREDIT CARD DEFAULT CASE STUDY**

**Mikayla Edwards**

I conducted research in an R Script file on Taiwan credit card default data to see if any of the datasets independent variables could be used to predict whether someone would default on their credit card payment. The data source is property of the UCI Machine Learning Repository and can be found here: [https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients](https://archive.ics.uci.edu/ml/datasets/default%2Bof%2Bcredit%2Bcard%2Bclients)

To begin I first familiarized myself with the dataset which had the following variables:

* Default payment (binary variable, Yes = 1, No = 0)
* Amount of the given credit (NT dollar) \* This shows in the data as LIMIT\_BAL
* Gender (1 = male; 2 = female) \* This shows in the data as ‘SEX’
* Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
* Marital status (1 = married; 2 = single; 3 = others) \*This shows in the data as MARRIAGE
* Age (year)
* Repayment status for the past monthly payments from April to September, 2005 (scale: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ...; 8 = payment delay for eight months; 9 = payment delay for nine months and above)
* Amount of bill statement for the past monthly payments from April to September, 2005 (NT dollar)
* Amount of previous payment for the past monthly payments from April to September, 2005 (NT dollar)

**Type of Analysis**

When deciding what I wanted to investigate within this data set I came up with the following:

* The correlation between gender, marital status, and default status of credit card clients in Taiwan. This is due to the historical gender inequality in Taiwan and the government's stated progress towards gender equality. The goal is to see if the recent progress towards gender equality is reflected in the credit card default data, specifically with regards to women taking on debt (Shen).
* The relationship between LIMIT\_BAL and default status as per this dataset's data dictionary says, LIMIT\_BAL includes both the individual consumer credit and their supplementary family credit. I would like to see if there's a correlation between MARRIAGE and LIMIT\_BAL since married individuals would have access to family credit.
* Building and testing a logistic regression model based on a cleaned dataset to see which independent variables are the strongest predictors of a default status of 1.

**Cleaning the Data**

I knew that I would first have to trim down the dataset via data cleaning to then run some correlations and ultimately build and test a logistic regression model's accuracy. The first step was to import, view and create an alias for the data.

* read.csv('/Users/mikayla/Downloads/CreditCardDefaultData.csv')
* View(CreditCardDefaultData
* data <- CreditCardDefaultData

The dataset includes columns of data on monthly payment records from April to September 2005. These records include the history of past payments, the amount on bill statements, and the amount of previous payments. To simplify the analysis, I excluded this data and got rid of any blank data rows to increase efficiency. The leftover columns that I focused my research on were: LIMIT\_BAL (credit limit), SEX, EDUCATION, MARRIAGE, AGE, and Default (default status). The code I produced to complete this is as follows:

* install.packages("dplyr")
* library(dplyr)
* data\_clean <- CreditCardDefaultData %>% select(LIMIT\_BAL, SEX, EDUCATION, MARRIAGE, AGE, Default)

**Checking for Missing Values**

Next, I needed to make sure my data didn’t include any missing values that would through of any further analysis. To double check for any missing values I I ran a sum and ColSums:

* sum(is.na(data\_clean))
* colSums(is.na(data\_clean))

**Creating a Correlation Matrix**

To begin the analysis, I loaded 2 packages: ggplot2 and corrplot packages. The cleaned data was then transformed into a data frame.

* install.packages("ggplot2")
* library(ggplot2)
* install.packages("corrplot")
* library(corrplot)
* df\_clean<- as.data.frame(data\_clean)

The independent variables of interest were added to a new dataset (analysis1) and the correlations between these variables and the response variable were calculated. A visual correlation matrix was also created to further understand the relationships between the variables.

* analysis1 <- data\_clean[c("MARRIAGE", "SEX", "EDUCATION", "LIMIT\_BAL","AGE")]
* cor1 <- cor(analysis1,df\_clean$Default,use = 'everything')
* View(cor1)
* print(x=cor1,y=df\_clean$Default) # This shows the correlation between the independent variables and Default status



* print(cor(data\_clean)) #This shows the correlation between all variables
* corrplot(cor(df\_clean), type = "upper", order = "hclust", tl.col = "black") #This visualization shows the correlation between all variables



The results of the analysis show that there is a low, weak negative correlation between the independent variables of MARRIAGE and SEX with the dependent variable of default payment. There is also a weak positive correlation between the independent variables of education and age with default payment. The only independent variable that had a strong enough correlation with default payment to be visible on the visual correlation matrix was credit limit, LIMIT\_BAL, with a correlation of -0.15, which is not a very strong correlation. However, this makes sense as those that default often are likely to have lower credit limits, but one may think the correlation would be higher.

**Logistic Regression**

**To build the regression model I libraried the glm package.**

* **install.packages("glm")**
* **library(glm)**

**Next, I split the data into 2: testing data and validating data then fit the model to the training set**

* **set.seed(123)**
* **train\_index <- sample(1:nrow(df\_clean), 0.8 \* nrow(df\_clean))**
* **train\_data <- df\_clean[train\_index, ]**
* **test\_data <- df\_clean[-train\_index, ]**

**The model was then built and the predications were tested and converted to binary values to make the results easier to understand. I also found the mean of the model to see the accuracy percentage.**

* **logistic\_model <- glm(Default ~ ., data = train\_data, family = binomial(link = "logit"))**
* **test\_predictions <- predict(logistic\_model, test\_data, type = "response")**
* **test\_predictions <- ifelse(test\_predictions > 0.5, 1, 0)**
* **mean(test\_predictions == test\_data$Default)**

The mean of the model was .785 meaning the model has an accuracy of 78.5% making the model pretty good.

Finally, I ran a summary of the model. This showed that all the independent variables have an impact on the predictability of the default status with some variables like LIMIT\_BAL, SEX, MARRIAGE, and AGE having a higher impact than others.



**Conclusion**

In conclusion, after conducting a correlation matrix and a logistic regression model, it became clear that there are stark differences in the correlation between the independent and dependent variables. This is because a correlation matrix only shows linear relationships. On the other hand, logistic regression models can capture non-linear relationships between independent and dependent variables. As a result, it is crucial to assess correlations in multiple ways, with a preference to regression models to accurately capture the relationship between the variables.

**Resources:**

Shen, L. (2022). Gender Equality and Economic Development in Taiwan. Journal of Asian and African Studies, 57(8), 1007-1017. doi: 10.1007/s12140-022-09392-3

UCI Machine Learning Repository. (n.d.). Default of credit card clients [Data set]. Retrieved from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients