By: James Tondt, Zach Lees, Kyler Sturgill

Part 1:

The dataset used in this analysis is sourced from Kaggle and contains information about clients with potential credit fraud. It comprises 122 variables, a mix of binomial and categorical, and was last updated four years ago. Our exploration aims to answer: are females at a higher risk of defaulting on loans? Moreover, we investigate whether gender significantly influences the likelihood of defaulting when controlling for other key factors.

Exploratory Data Analysis (EDA):

The response variable (TARGET) represents clients with reported payment difficulties, and it will signify credit fraud if positive. For study purposes TargetYes will determine if the client is determined to have defaulted (yes) or not (no). After exploring the data, the predictor variables are:

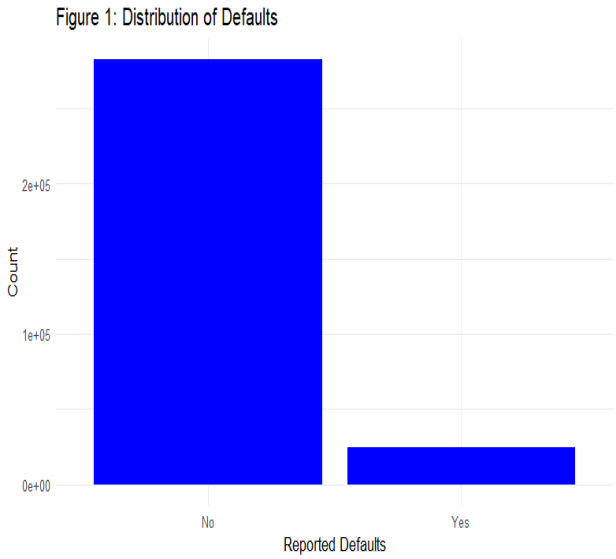
CODE_GENDER (binary): Gender of the client.

AMT_INCOME_TOTAL (quantitative): Income of the client.

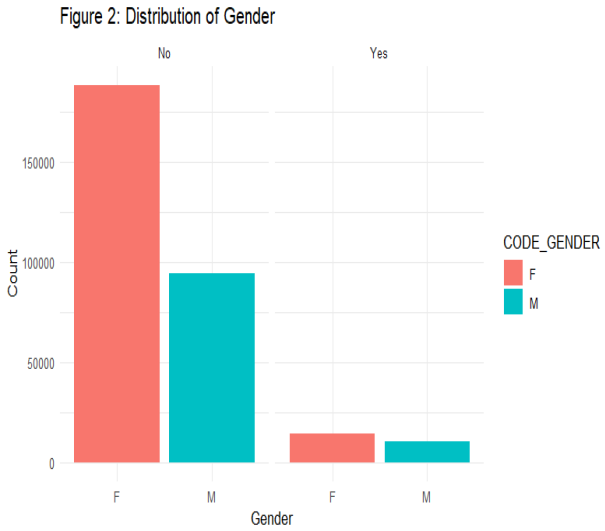
AMT_CREDIT (quantitative): Credit amount of the loan.

NAME_CLIENT_TYPE (binary): Whether the client was old or new when applying for the previous application.

Based upon Figure 1, most of the data collected shows that the vast majority of responses in Target are negative for default detection. This means that most of the clients, based upon the data, collected may not be subject to defaulting and that further analysis utilizing Target will be a good response variable viable. Next, we graph the distribution of Gender (male or female) to observe the split between genders collected and reported defaults. Based upon Figure 2, we can say that about twice the amount of clients reported to be females versus males had no reported default. However, from further observation, we can see that females have a slightly higher count than males for reported default.

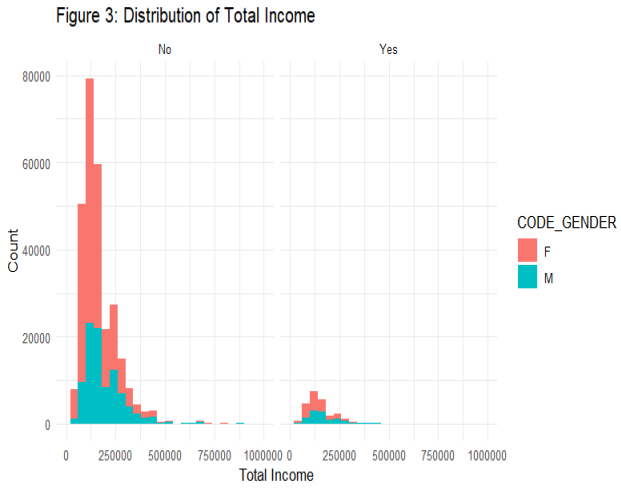


Bar-plot for Credit Fraud distribution.

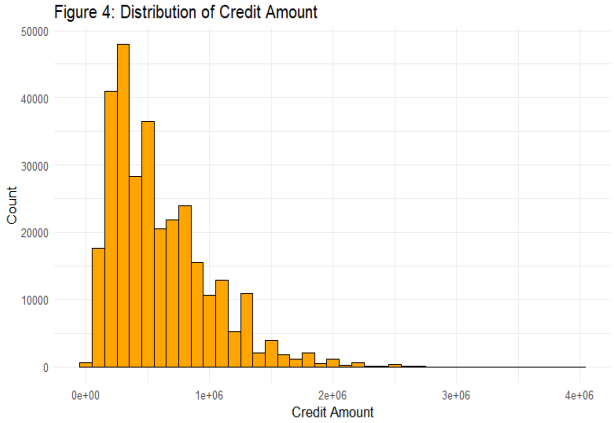


Bar-plot for Gender distribution.

Next, we will observe the relationship between total income and gender. Removing any data exceeding an income of 1,000,000 would keep our data from having errors or excessive outliers. We will set a maximum income value of 1,000,000 to clean our data. As can be seen in Figure 3, the data comparing total income and the amount of credit fraud reported by gender, it is clear that females typically show a pattern of having more reported defaults than males do with the same to similar total incomes. Looking at this graph it is crucial to keep in mind that more females than males were also surveyed, so this graph alone is not enough to provide enough evidence to support our research statement. Comparing the graphs for Total Income(Figure 3) and Credit amount(Figure 4), the graphs seem to be very similar. Thus it is fare to consider the fact that the credit amount of a loan is dependent on the house income and based upon the data is closely related to the same amount.

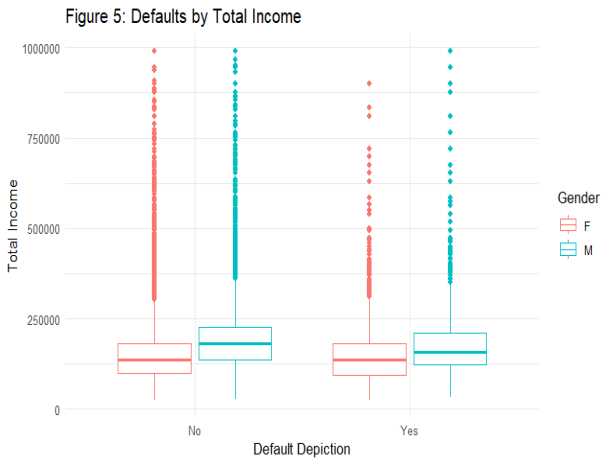


Histogram for Total Income filled by Gender.

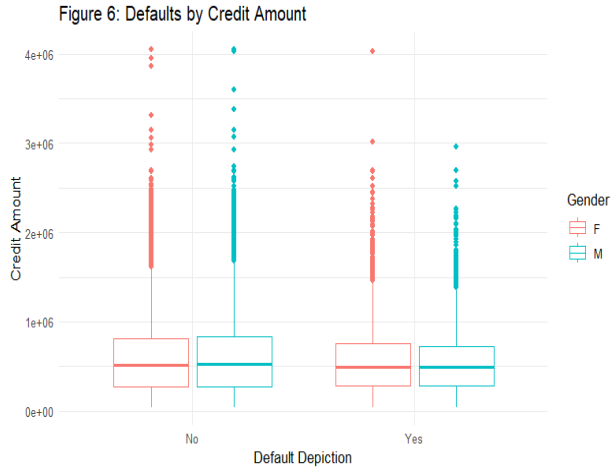


Histogram for Credit Amount.

Upon further analysis, we have found that males have higher income on average in each respective quarterly, as can be seen in Figure 5. Income doesn't appear to be a very strong predictor for determining if a client defaulted or not. The data doesn't appear to be very different upon comparing total income and credit amount (between males and females regarding if they have reported default or not). Credit amount may not be a strong predictor for determining if a client defaulted or not. This could be because the credit amount given is not sex-dependent, so males and females receive credit at the same rate based on their income.



Boxplot for Credit Fraud by Total Income colored by Gender.



Boxplot for Credit Fraud by Credit Amount colored by Gender.

Two logistic regression models were employed in the statistical analysis: the full model and the reduced model. The full model included predictors such as total income (AMT_INCOME_TOTAL), credit amount (AMT_CREDIT), income type (NAME_INCOME_TYPE), and gender (CODE_GENDERM). The reduced model excluded the gender variable, encompassing total income, credit amount, and income type as predictors. The drop-in-deviance test was employed to assess the contribution of the gender variable and a value of 756.44 was calculated. Along with its p-value of < 2.2e-16 there was compelling evidence to reject the null hypothesis. This rejection affirmed the importance of gender in predicting credit defaults and we will use the model with the variable for gender. Additionally, the backward selection method confirmed the significance of all predictors in the model, further reinforcing the comprehensive nature of the chosen variables.

Null Hypothesis: $H_0: \beta_{GENDER} = 0$

Implied model: $log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i,CREDIT} + \beta_2 x_{i,INCOME\_TOTAL} + \beta_3 x_{i,INCOME\_TYPE} + \in_i$

Alternative Hypothesis: $H_0: \beta_{GENDER} \neq 0$

Implied model: $log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i,CREDIT} + \beta_2 x_{i,INCOME\_TOTAL} + \beta_3 x_{i,INCOME\_TYPE} + \beta_4 x_{i,GENDER} + \in_i$

Moving on to model interpretation, the logistic regression output revealed insightful parameter estimates. The coefficient for gender (CODE_GENDERM) was 0.3868 with a standard error of 0.01391. This positive coefficient indicates that being male is associated with an increase in the log odds of credit default. The other predictors, such as total income (AMT_INCOME_TOTAL) and credit amount (AMT_CREDIT), demonstrated negative coefficients, suggesting a negative relationship with the log odds of default. These interpretations align with our expectations: higher income and credit amounts are associated with a lower likelihood of credit defaults.

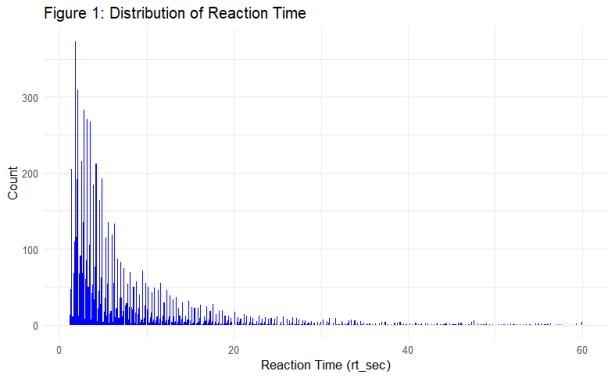
In conclusion, the final logistic regression model, incorporating gender alongside other key predictors, emerges as a robust framework for predicting credit fraud. The variables selected proved to be significant contributors, with gender playing a noteworthy role. Therefore, we can assert that, when controlling for total income, credit amount, and income type, gender remains a crucial factor in determining the likelihood of credit defaults. This analysis contributes valuable insights to the ongoing discourse on credit risk assessment, emphasizing the need to consider gender dynamics in predicting and managing credit fraud.

Part 2:

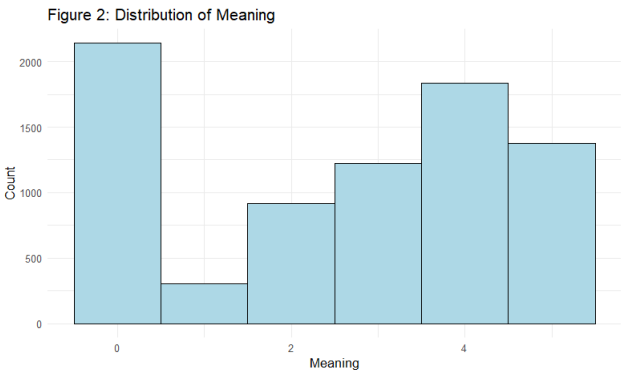
We chose dataset 3 to do our analysis. A pivotal aspect of the analysis revolves around discerning whether the relationship between meaning and reaction time is contingent upon the participants' age. This study employs reaction time (rt_sec) as the response variable, with a multi-level framework incorporating both individual- and observation-level covariates. At the individual level, participants' sex, age, and an age-group indicator (oldage) are considered, while the observation level includes variables related to the meaning and salience of stimuli. Additionally, random effects are introduced at the participant level, acknowledging the inherent variability between individuals. This research aims to shed light on the nuanced interplay between cognitive processes, age-related differences, and contextual meaning, contributing to a comprehensive understanding of the factors influencing reaction time in a dynamic perceptual task. Our goal is to discover whether the association between meaning and reaction time differ based on old age.

Exploratory Data Analysis (EDA):

As can be seen in Figure 1, many of the reaction times are very quick, with in the first 0 to 5 seconds, and it is right skewed. As seen in Figure 2, is a numerical representation of how useful the picture is. It is interesting to note that it is not an equal spread, it has large amounts in 0 and 4.

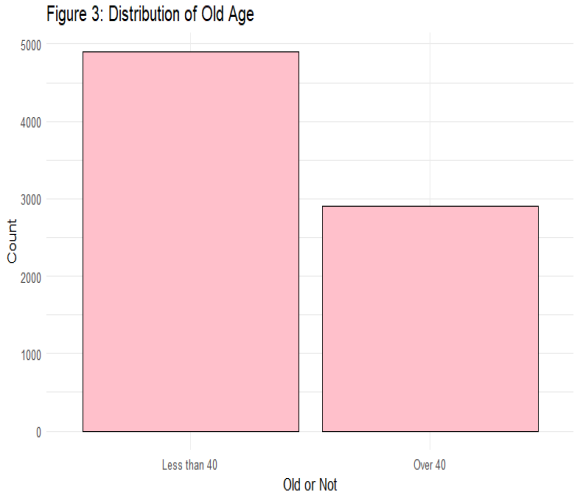


Histogram for Reaction Time Distribution.

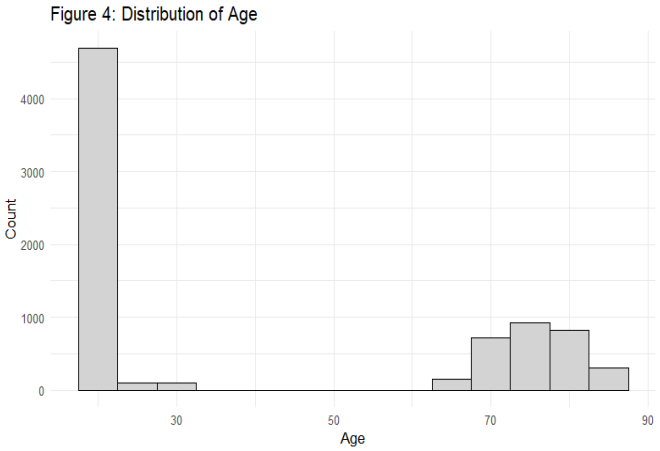


Histogram for Meaning Distribution.

As seen in Figure 3, there are more participants that are under the age of 40 than over the age of 40, almost double the amount. The age of participants is bimodal. As seen in Figure 4, there are many who are very young, newer participants in their teens and twenties and large amounts of participants within the 65 to 80 age range. This was likely done to compare the two age groups and to get large differences by using such a drastic difference in ages.

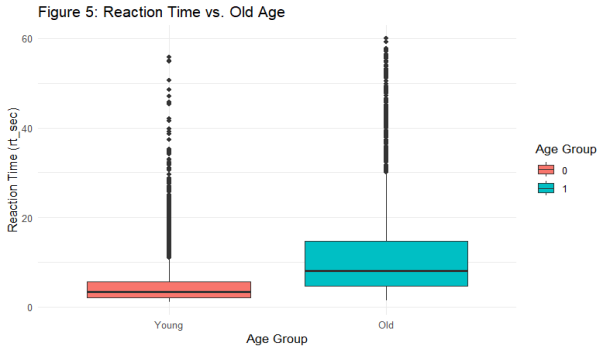


Bar Plot for Distribution of Old Age.

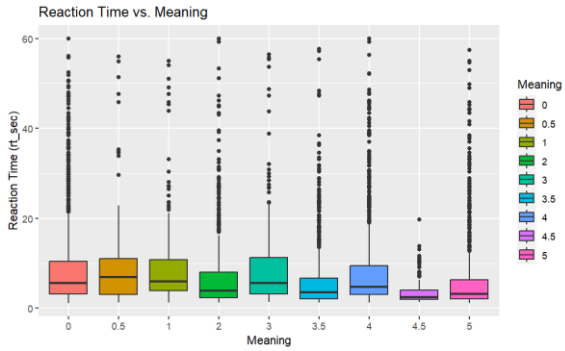


Histogram for Distribution of Age.

When comparing the groups of young and old, we see that each quartile of old has longer reaction times compared to the young box plot. As seen in Figure 5, there is also more variance between the old group compared to the young group. Meaning represents how important the picture is when driving. As we would expect, the least meaningful pictures have some of the highest reaction times, while most of the more important pictures have lower reaction times, in general. You can see this is Figure 6, Reaction time vs. Meaning.



Box plot for Reaction by OldAge filled by OldAge.



Box plot for Reaction by Time with Meaning

In the conducted analysis, we aimed to assess the association between the variables "meaning" and reaction time, considering potential differences based on old age. The null hypothesis (H_0) posited that there is no association between meaning and reaction time that varies based on old age, while the alternative hypothesis (H_1) suggested that such an association does exist. The likelihood ratio test was employed, and the test statistic, a drop-in-deviance of 91.062, resulted in an extremely low p-value (< 2.2e-16). This compelling evidence led to the rejection of the null hypothesis, indicating a statistically significant difference in the association between meaning and reaction time based on old age. We did an intercept-only model and the fixed intercept is 7.4402 with a standard error of 0.2957 and a t-value of 25.16. This is the estimated intercept that represents the average reaction time when there are no other predictors in the model. The estimated variance of within participant is 12.31. The estimated variance for between participants. The Intraclass correlation coefficient is 0.188, meaning, 18.8% of the total variation in rt_sec(reaction) is due to differences between musicians.

In conclusion, our analysis supports the adoption of the full model, which includes a random slope for old age, over the reduced model. This decision is substantiated by the significant drop-in-deviance and p-value, signifying that the association between meaning and reaction time indeed differs based on old age. The inclusion of the random slope allows for a more nuanced understanding of the relationship, acknowledging the variability in reaction time patterns across different age groups. Consequently, our findings emphasize the importance of considering the moderating effect of old age when examining the association between meaning and reaction time in the given dataset.