STAT6021 – Fall 2022
Project 2 – Group 7

Evanston, Mary – Haas, Tara – Hanley, Grant – Lankalapalli, Omprakash

Executive Summary

The habits and future behavior of credit card users is important to understand for individuals and for banking or credit card companies. Defaulting on a payment can lead to patterns of debt and is costly for banks to cover. Credit card companies want to be able to offer the credit line that a customer will require for their spending habits, ensuring that they have access to the level of credit that they can responsibly handle and that they might need in the future.

In examining data derived from 30,000 credit card users in Taiwan, statistical analysis can be used to predict the ensuing month's bill. Quantitative variables in this dataset proved very significant, particularly the prior month's bill amount, but also the bill amount and payment amount of the preceding two months and the lender's limit balance on their credit card. While helpful in prediction, the qualitative variables of this dataset were less significant in predicting the next month's bill amount. Education level and sex were both included as useful predictors.

This same data set can be used to predict whether or not a credit card user will likely default on their payment during the subsequent month. Different models can be used to predict whether a user will likely default. Picking the best model is a question of what risks a credit card company is willing to take. Are they more worried about falsely predicting that someone will default, or about missing a person who is likely to default? The model for prediction that we ultimately recommend is one that utilizes sex, age, the preceding two months' payment amounts, as well as transformed versions of the previous two months' bill amounts and the predicted bill amount generated by our model for the previous question.

Besides the output of our predictive model acting as a predictor variable for users that are likely to default on a payment, these two questions together could help banks understand their customers – particularly any that might need additional help or services. If a lender knows that an individual's bill amount is growing and that they have a low likelihood of defaulting on their future payments, the

creditor might want to reach out to their customer and offer them a greater credit limit before that customer looks elsewhere for additional credit. The bank can also use this information to identify risky, and therefore costly, customers.

2.0 Data Description

2.1 Where the Data Came From

The data set that we choose is "Default of Credit Card Clients Dataset"[1] from Kaggle.com which contains data related to credit card statements from clients in Taiwan during certain months in 2005. The referenced origin is the Machine Learning repository. The publicly available dataset includes payment and demographic data of credit card holders from an unspecified Taiwanese bank. The raw dataset included 30,000 observations on 24 variables. We will consider The Machine Learning Repository references the first use of the data set by Yeh, I. C., & Lien, C. H. (2009) in comparing techniques for predicting the probability of default of credit card clients.

2.2 Description of the Variables

- LIMIT_BAL: Limit balance refers to the credit limit in 2005 NT dollars for the individual and family. We anticipate that a larger limit balance will result in an increased likelihood of default. We believe this because people carrying higher levels of revolving credit who experience a financial shock would be more likely to default, in theory.
- PAY_AMT1: Payment amount one represents the amount in 2005 NT dollars in the previous month. We anticipate that a decreased payment amount will result in an increased likelihood of default the following month.
- PAY_AMT2: Payment amount two represents the amount in 2005 NT dollars two months ago. We anticipate that a decreased payment amount will result in an increased likelihood of default the following month.
- PAY_AMT3: Payment amount three represents the amount paid in 2005 NT dollars three months ago. We anticipate that a decreased payment amount will result in an increased likelihood of default the following month.

[1] https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

- PAY_AMT4: Payment amount four represents the amount paid in 2005 NT dollars four months ago. We anticipate that a decreased payment amount will result in an increased likelihood of default the following month.
- PAY_AMT5: Payment amount five represents the amount paid in 2005 NT dollars five months ago. We anticipate that a decreased payment amount will result in an increased likelihood of default the following month.
- PAY_AMT6: Payment amount six represents the amount paid in 2005 NT dollars six months ago. We anticipate that a decreased payment amount will result in an increased likelihood of default the following month.
- BILL_AMT1: Bill amount one represents the monthly bill total in the previous month in 2005 NT dollars. We anticipate that an increased bill amount will result in an increased likelihood of default the following month.
- BILL_AMT2: Bill amount two represents the monthly bill total from two months ago in 2005 NT dollars. We anticipate that an increased bill amount will result in an increased likelihood of default the following month.
- BILL_AMT3: Bill amount three represents the monthly bill total from three months ago in 2005 NT dollars. We anticipate that an increased bill amount will result in an increased likelihood of default the following month.
- BILL_AMT4: Bill amount four represents the monthly bill total from four months ago in 2005 NT dollars. We anticipate that an increased bill amount will result in an increased likelihood of default the following month.
- BILL_AMT5: Bill amount five represents the monthly bill total from five months ago in 2005 NT dollars. We anticipate that an increased bill amount will result in an increased likelihood of default the following month.
- BILL_AMT6: Bill amount six represents the monthly bill total from six months ago in 2005 NT dollars. We anticipate that an increased bill amount will result in an increased likelihood of default the following month.
- MARRIAGE: A categorical variable. We believe marriage would influence the likelihood of default one way or the other. We can see how having a joint income could reduce the likelihood of default, but anecdotally understand how marriage can be very costly.
- AGE: A quantitative variable, there is reason to believe that experience and income are correlated with age which may be negatively correlated with default likelihood.

- EDUCATION: A categorical variable, there is reason to believe that education levels are positively correlated with income and would accordingly be negatively correlated with likelihood of defaulting within the next month.
- SEX: A categorical variable, anecdotally we believe there is some differentiation in credit scores due to sex, accordingly we think there may be some observational evidence in credit card defaults.

3.0 Linear Regression

3.1 Introduction

3.1.1 Statement of First Question

How does limit balance, previous monthly bill amounts, payment amounts, and demographic information (Age, Sex, Education, and Marriage status) predict the next month's bill amount?

3.1.2 Why this Question is worth exploring

Predicting a given month's bill amount using the two preceding months' bill and statement amounts as well as demographic information would be of interest to banks trying to understand the extent their customers are utilizing credit and when they might be approaching a bill that they are unable to pay. In January 2021, Bankrate produced a survey indicating that fewer than 40% of Americans could "comfortably cover an unexpected expense of $1,000." With savings low, access to credit becomes ever more important. If banks better understand the payment habits of their customers, they might choose to adjust limit balances in order to keep customers from raising bills above the amount they can comfortably pay. The data within our dataset looks at credit card users in Taiwan, but the importance of credit access, and the risk of credit default, is the same globally. If demographic information is significant in this linear regression, it might indicate insufficient education on credit limits and family financing in different demographic groups.

3.2 Data Visualizations

3.2.1 Data Wrangling Description

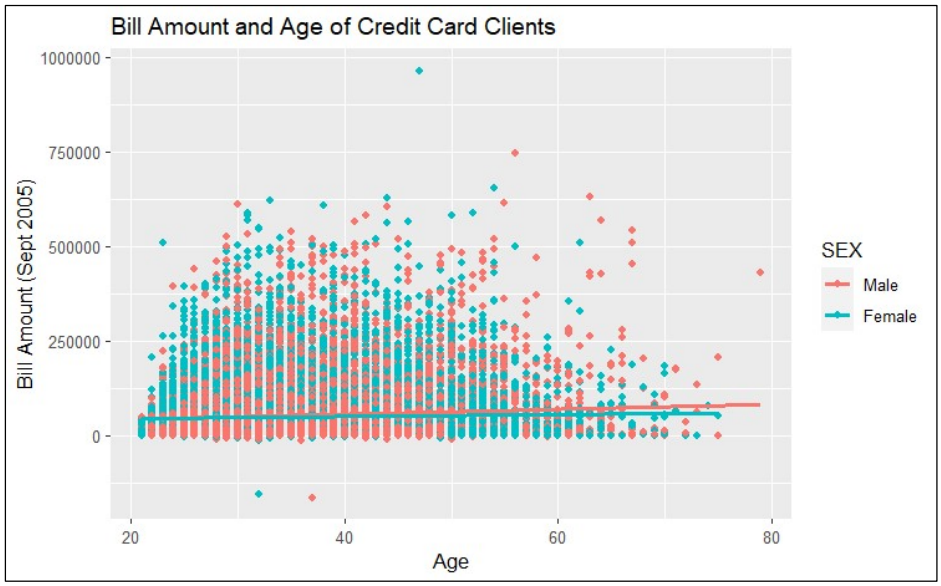
We created a data frame from the raw data at https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/download?datasetVersionNumber=1 and imported it into a dataframe in R called "ccd".

ccdSEX, ccdEDUCATION and ccd$MARRIAGE were all converted from character to factors to aid in the categorical prediction and residual visualizations.

We removed the following factor levels for EDUCATION: None and Unknown. These factors did not exist in the original dataset, therefore we removed them.

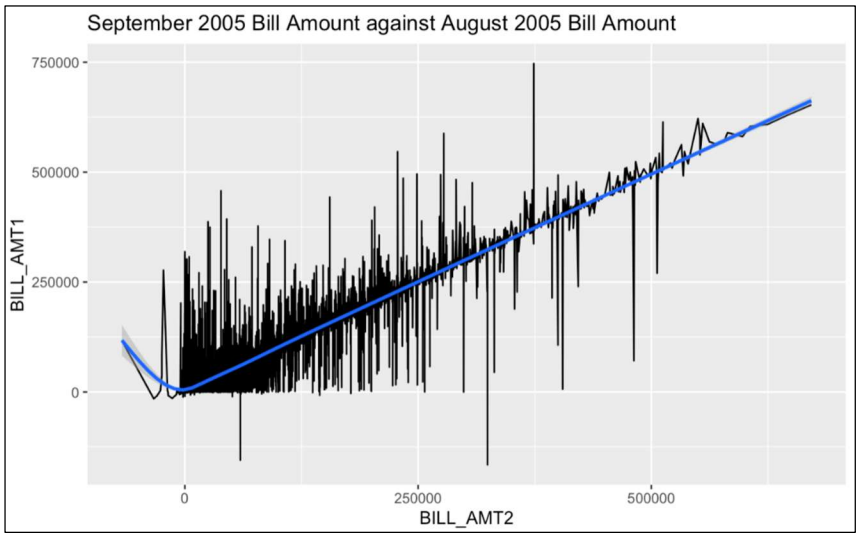
3.2.2 Relevant Data Visualizations and Contextual Interpretations

When reviewing our question of if BILL_AMT1 can be predicted by LIMIT_BAL, BILL_AMT2, BILL_AMT3, PAY_AMT2, PAY_AMT3, and demographic variables AGE, SEX, EDUCATION, and MARRIAGE, we regressed BILL_AMT1 against AGE and grouped by SEX.

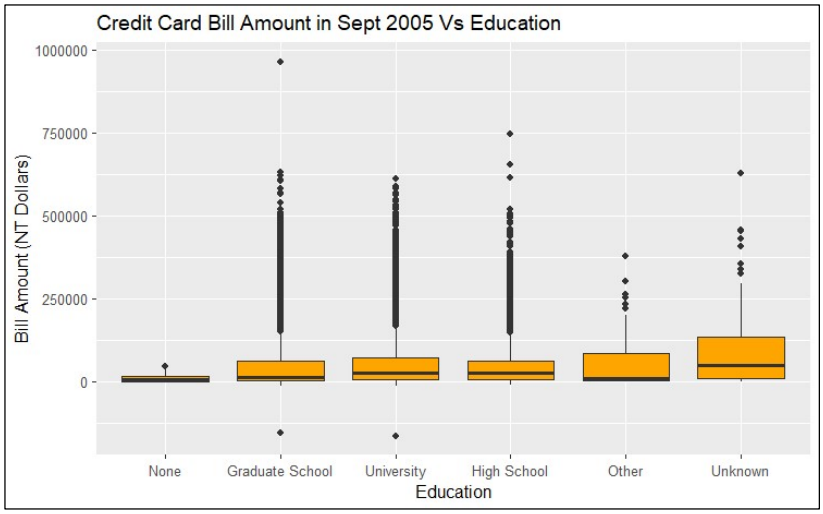


Plotting BILL_AMT1 against AGE, grouped by SEX, appears to have a positive linear relationship for both Male and Female. The visualization appears to have a slightly increasing relationship for Males over Females. We can potentially interpret this as: for Males, as they age, they tend to have slightly higher average bills than Females.

Next, we plotted BILL_AMT2 against BILL_AMT1 and observed the graph below. There appears to be a strong linear relationship between the variables.



Another visualization we utilized was the boxplot of EDUCATION against BILL_AMT1. There appears to be little variation between clients with high school and above education, as shown below.



3.3 Model Building

3.3.1 Linear Model Choice

To choose our linear regression model, we first fit a multiple linear regression model of BILL_AMT1 against all predictors to generate the following result:

```
Call:
lm(formula = BILL_AMT1 ~ LIMIT_BAL + BILL_AMT2 + BILL_AMT3 +
    PAY_AMT2 + PAY_AMT3 + AGE + SEX + EDUCATION + MARRIAGE, data = ccd)

Residuals:
    Min      1Q  Median      3Q     Max
-489733   -4391   -2143     705  414124

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.029e+03  8.335e+02  -1.235  0.21698
LIMIT_BAL            1.442e-02  1.286e-03  11.213  < 2e-16 ***
BILL_AMT2            8.842e-01  6.878e-03 128.545  < 2e-16 ***
BILL_AMT3            1.059e-01  7.361e-03  14.386  < 2e-16 ***
PAY_AMT2            -8.000e-02  8.673e-03  -9.224  < 2e-16 ***
PAY_AMT3             6.934e-02  9.511e-03   7.290 3.19e-13 ***
AGE                  5.287e+00  1.907e+01   0.277  0.78162
SEXMale              6.792e+02  2.997e+02   2.266  0.02346 *
EDUCATIONHigh School 1.485e+03  4.715e+02   3.151  0.00163 **
EDUCATIONOther       6.791e+03  2.308e+03   2.942  0.00326 **
EDUCATIONUniversity  1.673e+03  3.434e+02   4.871 1.12e-06 ***
MARRIAGESingle      -1.845e+00  3.439e+02  -0.005  0.99572
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22590 on 23220 degrees of freedom
Multiple R-squared:  0.9081,    Adjusted R-squared:  0.9081
F-statistic: 2.086e+04 on 11 and 23220 DF,  p-value: < 2.2e-16
```

The T-tests seem to indicate that AGE and MARRIAGESingle are not significant, given the other variables in the model. We performed a hypothesis test against this full and a reduced model removing AGE and MARRIAGE.

```
Call:
lm(formula = BILL_AMT1 ~ LIMIT_BAL + BILL_AMT2 + BILL_AMT3 +
    PAY_AMT2 + PAY_AMT3 + SEX + EDUCATION, data = ccd)

Residuals:
    Min      1Q  Median      3Q     Max
-489738   -4386   -2141     710  414107

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.696e+02  3.880e+02  -2.241 0.025027 *
LIMIT_BAL            1.449e-02  1.258e-03  11.526  < 2e-16 ***
BILL_AMT2            8.842e-01  6.876e-03 128.583  < 2e-16 ***
BILL_AMT3            1.059e-01  7.360e-03  14.390  < 2e-16 ***
PAY_AMT2            -8.003e-02  8.672e-03  -9.228  < 2e-16 ***
PAY_AMT3             6.933e-02  9.509e-03   7.291 3.19e-13 ***
SEXMale              6.890e+02  2.971e+02   2.319 0.020395 *
EDUCATIONHigh School 1.523e+03  4.547e+02   3.349 0.000812 ***
EDUCATIONOther       6.787e+03  2.308e+03   2.941 0.003279 **
EDUCATIONUniversity  1.680e+03  3.393e+02   4.952 7.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22590 on 23222 degrees of freedom
Multiple R-squared:  0.9081,    Adjusted R-squared:  0.9081
F-statistic: 2.55e+04 on 9 and 23222 DF,  p-value: < 2.2e-16
```

We then performed an ANOVA test to determine if we can remove those predictors. The null hypothesis was that all the coefficients were equal to 0, the alternative hypothesis was that at least on coefficient is not zero. The results of the ANOVA test are shown below:

```
Analysis of Variance Table

Model 1: BILL_AMT1 ~ LIMIT_BAL + BILL_AMT2 + BILL_AMT3 + PAY_AMT2 +
PAY_AMT3 +
    SEX + EDUCATION
Model 2: BILL_AMT1 ~ LIMIT_BAL + BILL_AMT2 + BILL_AMT3 + PAY_AMT2 +
PAY_AMT3 +
    AGE + SEX + EDUCATION + MARRIAGE
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1  23222 1.1854e+13
2  23220 1.1854e+13  2  50394338 0.0494 0.9518
```

Interpreting our ANOVA test, our F-statistic is 0.0494 with a p-value of 0.9518. We do not reject the null hypothesis. Our data suggests we can drop the predictors AGE and MARRIAGE and go with the reduced model. The resulting linear model is:

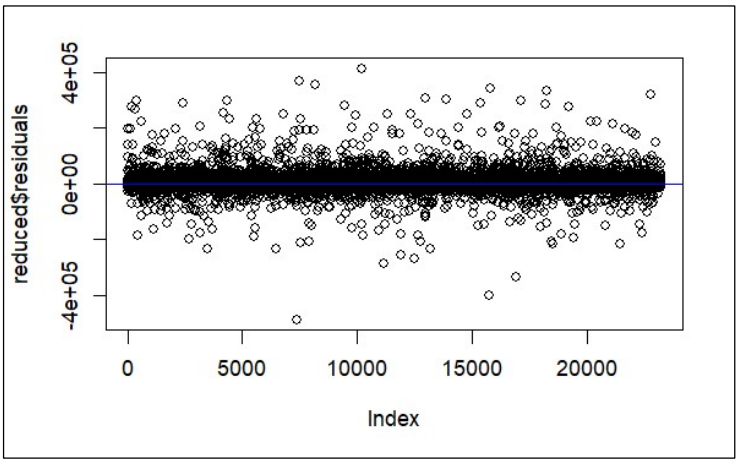
BILL_AMT1 = -869.60 + 0.0145*LIMIT_BAL* + 0.8842*BILL_AMT2* + 0.1509*BILL_AMT3* – 0.08*PAY_AMT2* + 0.0693*PAY_AMT3* + 0.0689*SEXMale* + 152.3*EDUCATIONHighSchool* + 0.0068*EDUCATIONOther* + 0.0017*EDUCATION*

To validate our choice of model, we used the regsubsets() function to run all possible regressions on our original (full) predictor set. We then compared the adjusted $R^2$and found that it is also possible to drop the predictor SEX to increase model accuracy.

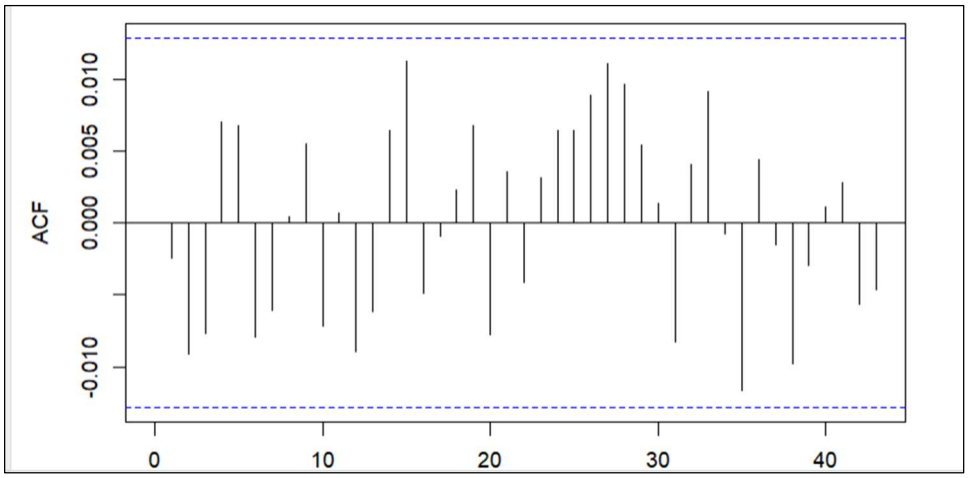
As further confirmation of our choice we ran the Mallow's C_p test (which indicated we drop AGE, MARRIAGE, and SEX) and looked at which model had the lowest BIC. Relying on BIC chose a model that only utilized EDUCATION if it was the category University. Finally, we ran a stepwise regression to see which model would be the best fit and concluded that the stepwise regression ends with the same predictors as our reduced model above. Thus, we chose to only remove AGE and MARRIAGE.

3.3.4 Model Regression Assumptions

To validate the regression assumptions, we generated a residual plot, created an ACF plot and QQ plot, and used boxplots to validate the assumptions of the categorical predictors.

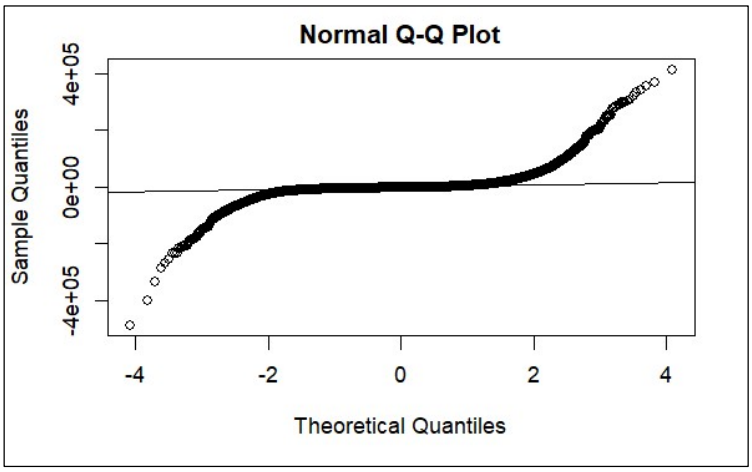


In the residual plot above, the data points appear evenly scattered around the line with no apparent pattern. Therefore, the linear relationship assumption and the mean = 0 assumption appear to be met.

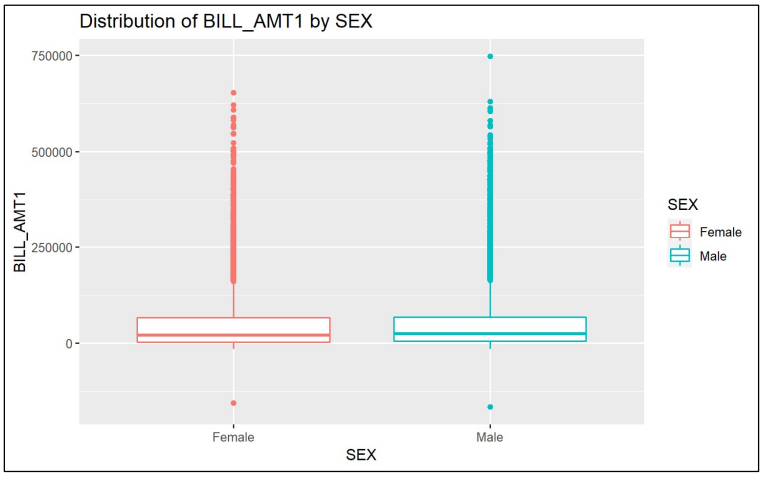


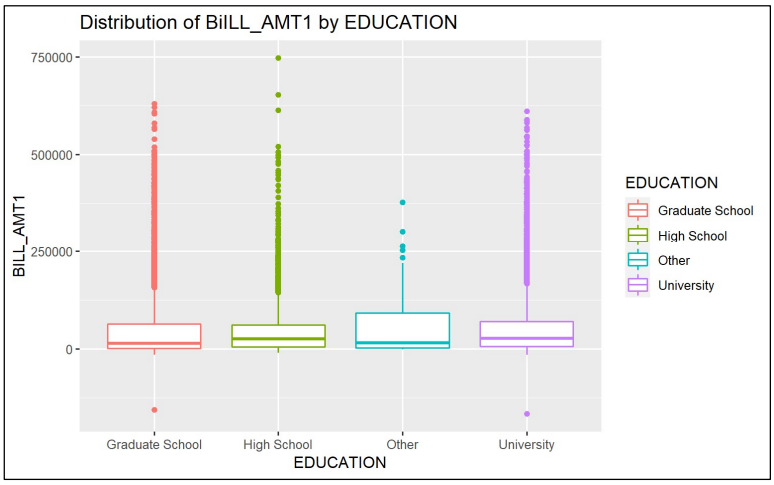
For the Auto Correlation Function (ACF) plot, we removed the Lag=0 on the x-axis to make the horizontal blue lines easier to read. Since all of the lag lines are within the blue dashed lines, we can assume that the error terms are uncorrelated.

Next, we conducted a QQ plot on the residuals of the reduced model. The results are seen below. The QQ plot is intended to test the normality function and is the least important of the assumption tests. We observe that for the majority of the Quantiles, the data points fall on the line. We accept this as acceptable proof of the normality function.



Since there are categorical predictors in the model, we need to conduct a further test on the variance of the response variables between all classes of the categorical predictors. For both SEX and EDUCATION, we conducted box plots of the categorical predictor against BILL_AMT1. In all cases, the spread for each class by category is similar. We can conclude that the variance assumption is met.





3.4 Conclusions

We can conclude that there is a linear relationship between limit balance, previous monthly bill amounts, payment amounts and demographic information when predicting the next month's bill amount. Specifically, the linear regression equation is:

$$BILL\_AMT1 = -869.60 + 0.0145LIMIT\_BAL + 0.8842BILL\_AMT2 + 0.1509BILL\_AMT3 - 0.08PAY\_AMT2 + 0.0693PAY\_AMT3 + 0.0689SEXMale + 152.3EDUCATIONHighSchool + 0.0068EDUCATIONOther + 0.0017EDUCATIONUniversity$$

We can further conclude that there can be a case where one might remove SEX and still retain a viable linear regression model.

3.4.1 How this answered our question of interest

It is important for banks to be able to predict the amount of credit they will be extending each month to their clients in order to obtain a profitable business. Banks make their profits by extending credit and having consumers pay for that privilege. With proper prediction they can cover their own debts and return a profit to their shareholders.

We showed in this analysis that we could predict how much each credit card bill would be (in effect the amount of credit the bank would have to pay if the consumer did not pay the bill) using multiple linear regression and removing predictors that did not significantly contribute to the final model.

3.4.2 Interesting Insights

It was interesting to note that SEX had far less effect on the linear model than we previously predicted. In fact, the logistical regression that we conducted in Section 4 below removed this as a predictor all together. This analysis shows that there is no one "right" model and that the analyst needs to take into account the use of the data when creating models. In this case, we were looking for a predictive linear model and chose to leave SEX in the regression equation.

3.4.3 Challenges Faced

As stated above, when we were decided on which regressors we could remove from the model, we used the Mallow's C_p test and looked at which model had the lowest BIC. These actions indicated that we could possibly drop SEX, and even change how we analyzed EDUCATION. We were challenged to create a model prediction methodology similar to what was conducted in the logistic regression section that categorically stated which model was the best fit.

4 Logistic Regression

4.1 Introduction

Our second question of interest is: What is the likelihood of credit card default provided a variety of demographic factors and recent monthly payment data?

Answering the question of an individual's likelihood of credit card default is a matter of value for banks, lenders, and other providers of credit. The ability to look at someone's recent bank statements while also accounting for demographic factors amounts to small cost when considering that when borrowers default, they often default for large some sums of money. The possibility of being able to prevent or mitigate default for even a few lenders may produce substantial financial or economic significance for banks, lenders, or others making credit-based decisions – which in turn helps to keep the interest rates lower for credit card users. While the dataset may be specific to Taiwan Credit Card Users in the mid 2000's, the application and model may extend to those outside of the original observation population.

4.2 Data Visualization

4.2.1 Data Wrangling Description

ccd$DEFAULT: A categorical factor was converted to factor data type to enable use with R data visualizations and built-in linear modeling tools.

ccd$AVG_UTILIZATION_RATE: Credit utilization, a numerical variable with the average bill amount divided by the limit balance per user.

ccd$OVERSPENDER: Create a logical binary variable for if the average monthly excess balance over the past five months was greater than the limit balance.

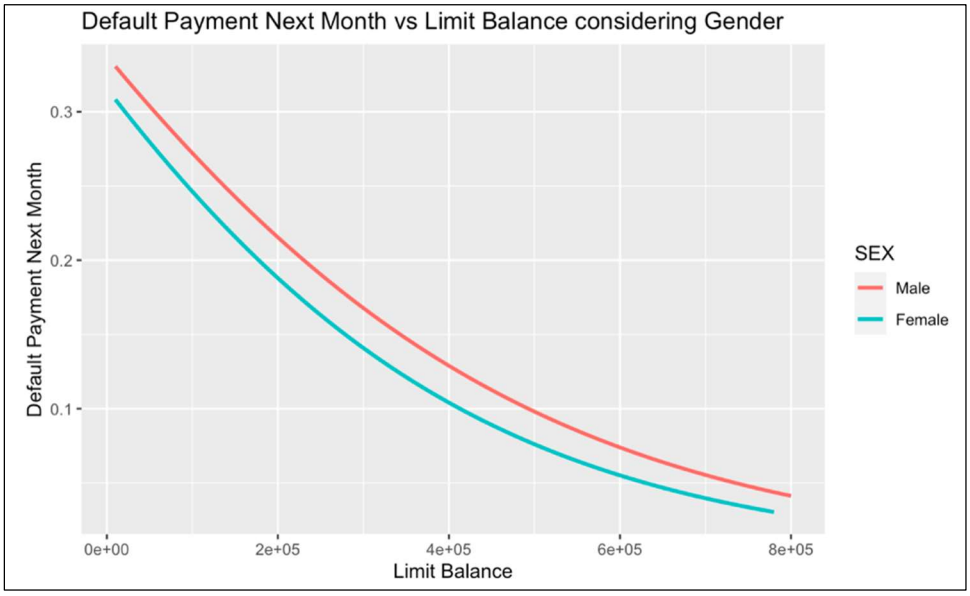
ccd$TOTAL_EXCESS_BALANCE: Create a numerical variable which captures how much excess balance over the past three months is being carried. The idea here is that users who make payments greater than their statement balance will have a negative excess balance average over three months. Those who are having trouble paying or are carrying revolving credit will.

Scaling Techniques: Due to the presence of outliers, we thought we could potentially improve our model performance through scaling. We explored four different scaling techniques: No Scaling, MinMaxScaling, Robust Scaling, and Standard Scaling.

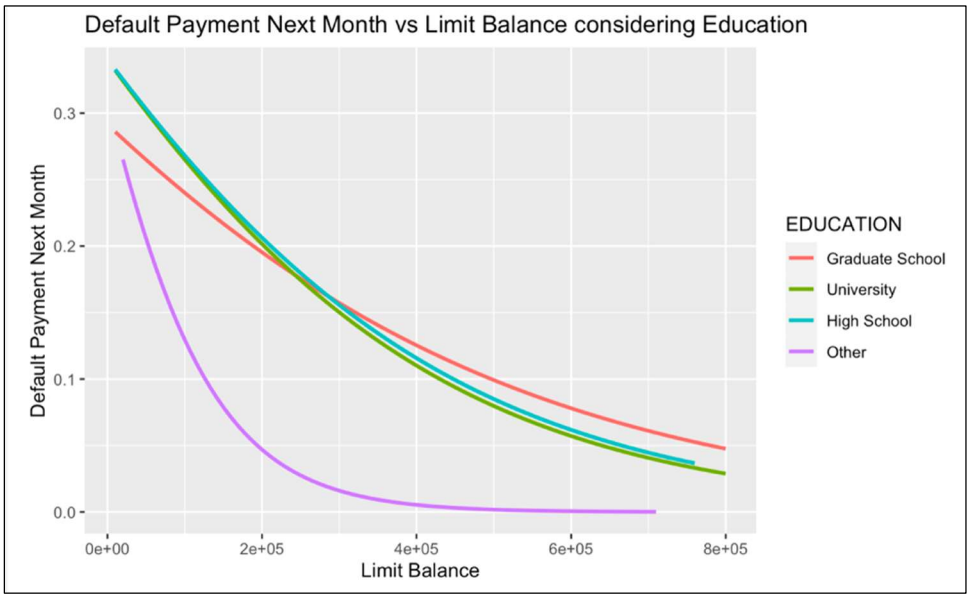
Sampling Techniques: We explored two sampling techniques and compared results with no adjustment to sampling. We utilized random over-sampling and random over-sampling to assist in handling the imbalanced dataset with respect to defaults.

Data types: Depending on the application the logistic regression required us to switch between double and factor data types to apply scaling functions and data visualizations.

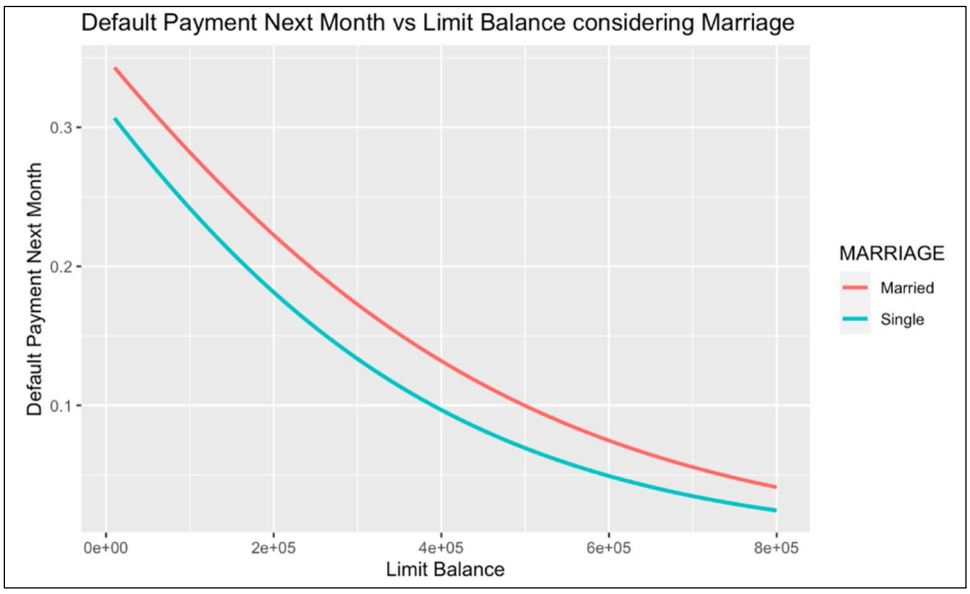
4.2.2 Relevant Data Visualization



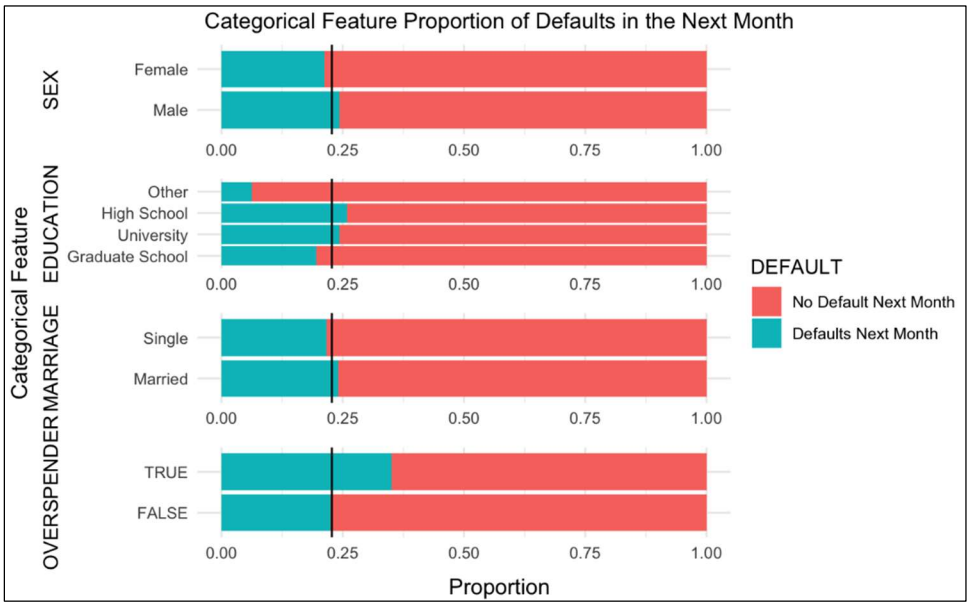
The graph above shows that as the limit balance increases, the probability of a default payment for next month decreases. When split by SEX, it shows that males have a higher rate of defaulting on payment than females. We can use this information to determine if the risk of default is significantly different between males and females with the same limit balance.



The graph above shows that as the limit balance increases, the probability of a default payment for next month decreases, like the previous graph but considering the education of the person. This graph delineates by education level, showing that the Other category for education has the lowest default rate compared to the others. The other three categories are clustered close together.

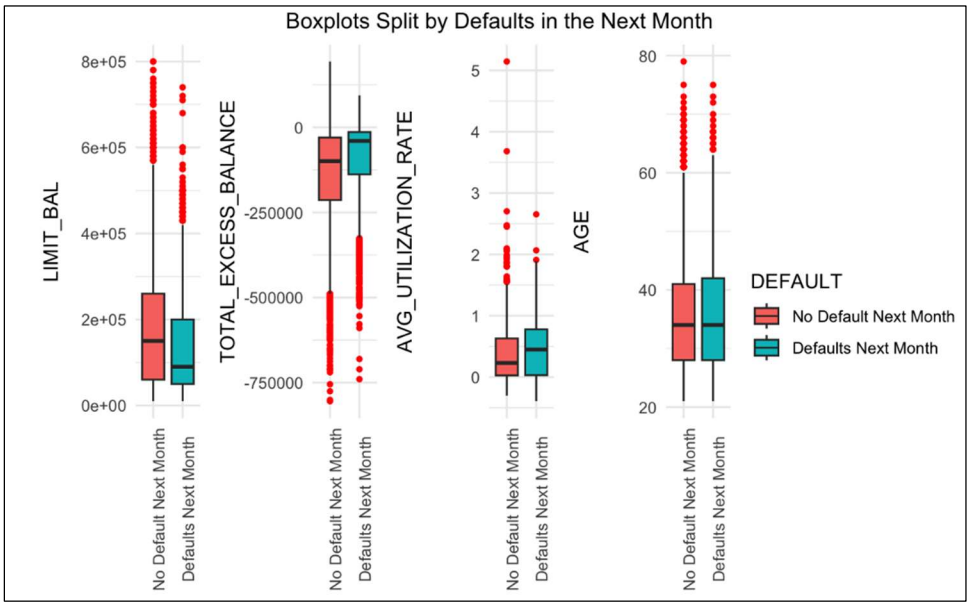


The graph above displays how marital status and limit balance might impact the probability that a person will default on their credit card payment in the following month. It shows that married people have a slightly higher chance of defaulting on payment compared to single people.

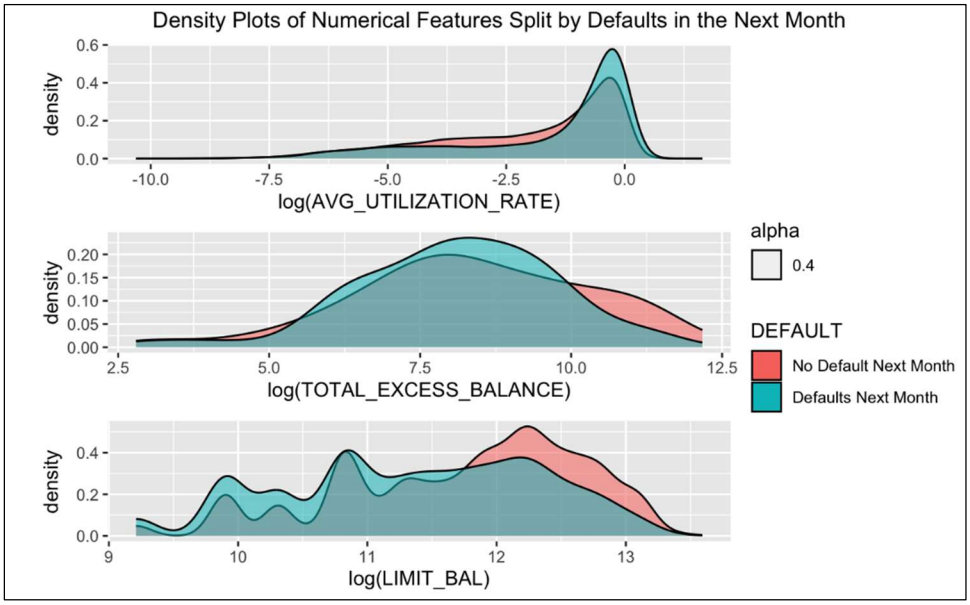


Using proportion bar charts, we explore the categorical features within our data with intent to identify impactful predictors. Splitting by individuals that default, our response variable and what we are attempting to model predictions for with the logistic regression, we note possible differences based on Sex, Education, Marital Status, and the Overspender Flag.

The Overspender Flag and Other Education immediately stick out as features with a relatively large difference in proportion of defaults in comparison to the population average default proportion. Males appear to have a larger proportion of defaults within the data, but the magnitude of the effect is difficult to discern from visualization alone. Married individuals also appear to have a slightly larger share of the population that results in default.



We explore some of our numeric predictor variables using boxplots with outliers depicted in red. Limit Balance on average appears to be larger for those who do not default versus those that default. As we would expect with Total_Excess_Balance, the greater the balance over the limit balance, the more defaults on average. Average Credit Card Utilization Rates appear to have a positive impact on the share of the population subset which defaults within the next month.



4.3 Model Building

4.3.1 Logistic Model Chosen Description

Total 8 Models were made.

Model 1: All variables was our first logistic regression model that simply included every feature available in the dataset.

$$\log\left[\frac{P(DEFAULT=1)}{1-P(DEFAULT=1)}\right] = \alpha + \beta_1(LIMIT\_BAL) + \beta_2(BILL\_AMT1) + \beta_3(BILL\_AMT2) + \beta_4(BILL\_AMT3) + \beta_5(BILL\_AMT4) + \beta_6(BILL\_AMT5) + \beta_7(BILL\_AMT6) + \beta_8(PAY\_AMT1) + \beta_9(PAY\_AMT2) + \beta_{10}(PAY\_AMT3) + \beta_{11}(PAY\_AMT4) + \beta_{12}(PAY\_AMT5) + \beta_{13}(PAY\_AMT6) + \beta_{14}(AVG\_UTILIZATION\_RATE) + \beta_{15}(PAY\_2) + \beta_{16}(PAY\_3) + \beta_{17}(PAY\_4) + \beta_{18}(PAY\_5) + \beta_{19}(PAY\_6) + \beta_{20}(SEX_{Female}) + \beta_{21}(MARRIAGE_{Single}) + \beta_{22}(EDUCATION_{University}) + \beta_{23}(EDUCATION_{High\ School}) + \beta_{24}(EDUCATION_{Other}) + \beta_{25}(AGE)$$

Model 2: Variables with Interactions was a model that was created to investigate variables that could have importance. Since the purpose of our model is prediction, we care about how well the model classifies a user as defaulter.

$$\log\left[\frac{P(DEFAULT=1)}{1-P(DEFAULT=1)}\right] = \alpha + \beta_1(BILL\_AMT1) + \beta_2(BILL\_AMT2) + \beta_3(BILL\_AMT3) + \beta_4(BILL\_AMT4) + \beta_5(BILL\_AMT5) + \beta_6(BILL\_AMT6) + \beta_7(PAY\_AMT1) + \beta_8(PAY\_AMT2) + \beta_9(PAY\_AMT3) + \beta_{10}(PAY\_AMT4) + \beta_{11}(PAY\_AMT5) + \beta_{12}(PAY\_AMT6) + \beta_{13}(PAY\_2) + \beta_{14}(PAY\_3) + \beta_{15}(PAY\_4) + \beta_{16}(PAY\_5) + \beta_{17}(PAY\_6) + \beta_{18}(SEX_{Female}) + \beta_{19}(MARRIAGE_{Single}) + \beta_{20}(EDUCATION_{University}) + \beta_{21}(EDUCATION_{High\ School}) + \beta_{22}(EDUCATION_{Other}) + \beta_{23}(AGE) + \beta_{24}(LIMIT\_BAL) + \beta_{25}(AVG\_UTILIZATION\_RATE) + \beta_{26}(SEX_{Female} \times MARRIAGE_{Single}) + \beta_{27}(EDUCATION_{University} \times AGE) + \beta_{28}(EDUCATION_{High\ School} \times AGE) + \beta_{29}(EDUCATION_{Other} \times AGE) + \beta_{30}(AGE \times LIMIT\_BAL) + \beta_{31}(AGE \times AVG\_UTILIZATION\_RATE)$$

Model 3: Two Stage Model uses the BILL_AMT1 and PAY_AMT1 to predict the likelihood of the person defaulting on their payment. We take the optimal model output from the linear regression refinements and add two categorical variables representing information on how far the credit user is behind on their payments.

First Stage:

$$BILL\_AMT1 = \alpha + \beta_1(LIMIT\_BAL) + \beta_2(BILL\_AMT2) + \beta_3(BILL\_AMT3) + \beta_4(PAY\_AMT2) + \beta_5(PAY\_AMT3) + \beta_6(SEX_{Female}) + \beta_7(EDUCATION_{University}) + \beta_8(EDUCATION_{High\ School}) + \beta_9(EDUCATION_{Other}) + \beta_{10}(AVG\_UTILIZATION\_RATE) + \epsilon$$

$$PAY\_AMT1 = \alpha + \beta_1(LIMIT\_BAL) + \beta_2(BILL\_AMT2) + \beta_3(BILL\_AMT3) + \beta_4(PAY\_AMT2) + \beta_5(PAY\_AMT3) + \beta_6(SEX_{Female}) + \beta_7(EDUCATION_{University}) + \beta_8(EDUCATION_{High\ School}) + \beta_9(EDUCATION_{Other}) + \beta_{10}(AVG\_UTILIZATION\_RATE) + \epsilon$$

Second Stage: Predictions from the first stage are included as predictors in the second stage along with the same controls from the first stage.

$$\log\left[\frac{P(DEFAULT=1)}{1-P(DEFAULT=1)}\right] = \alpha + \beta_1(LIMIT\_BAL) + \beta_2(BILL\_AMT2) + \beta_3(BILL\_AMT3) + \beta_4(PAY\_AMT2) + \beta_5(PAY\_AMT3) + \beta_6(SEX_{Female}) + \beta_7(EDUCATION_{University}) + \beta_8(EDUCATION_{High\ School}) + \beta_9(EDUCATION_{Other}) + \beta_{10}(AVG\_UTILIZATION\_RATE) + \beta_{11}(BILL\_AMT1\_PRED) + \beta_{12}(PAY\_AMT1\_PRED)$$

Model 4: Log Transformation to the Two Stage Model

Our motivation for the log transformations is the spread of features scale. In our data visualizations we identified that the bifurcation between default and no default became more pronounced when we conducted a log transformation.

$$\log\left[\frac{P(DEFAULT=0)}{1-P(DEFAULT=0)}\right] = \alpha + \beta_1(\log(LIMIT\_BAL + 0.01)) + \beta_2(SEX_{Female}) + \beta_3(AGE) + \beta_4(\log(BILL\_AMT1\_PRED + 0.01)) + \beta_5(\log(BILL\_AMT2 + 0.01)) + \beta_6(\log(BILL\_AMT3 + 0.01)) + \beta_7(\log(PAY\_AMT1\_PRED + 0.01)) + \beta_8(\log(PAY\_AMT2 + 0.01)) + \beta_9(\log(PAY\_AMT3 + 0.01)) + \beta_{10}(\log(AVG\_UTILIZATION\_RATE)) + \beta_{11}(PAY\_2) + \beta_{12}(PAY\_3)$$

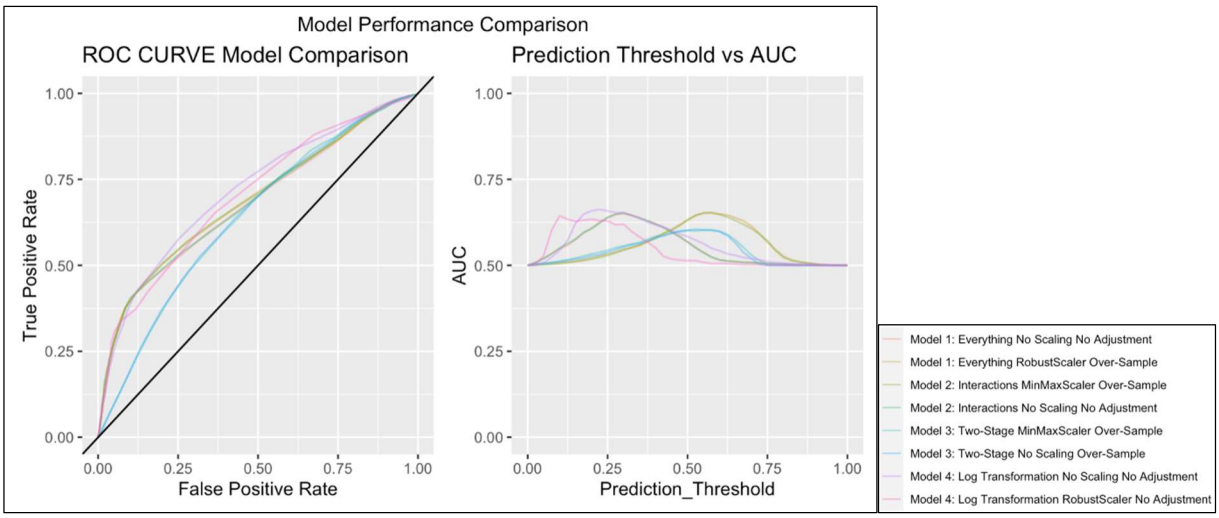
4.3.2 Logistic Model Improvements

We sought model improvements using multiple automated search tools to optimize across specific information criteria. Our feature scaling and sampling techniques were also attempts at model improvement. In initial model development we used the Anova drop method and chi squared test results from the Anova output to inform keeping and retaining variables. We utilized the automated processing tools from the bestglm package and the boruta package to support identifying the most important features to maintain and tune our model.

4.3.3 Model Comparisons

We evaluated the performance of our different logistic regression models utilizing different scaling techniques, sampling techniques, and prediction thresholds. In total we processed the model performance for (48) different models at (50) prediction thresholds. While the different scaling techniques and sampling techniques had an ambiguous impact on model performance in general. There were distinct differences in model performance depending on what performance feature we prioritized. We considered accuracy, precision, recall, F1 Score, Area Under Curve, False Negative and False Positive ratios.

4.3.3.1 ROC Curve



We hypothesized that a credit card company is more likely to desire a model that minimizes false negatives or balances the cost of false negatives with cost of false positives. Because we lack the domain expertise to associate a specific cost with a false negative and a false positive, we caveat our recommendation pending additional information, but established the meta-analysis framework to refine the recommendation provided enhanced domain knowledge

4.3.3.2 AUC

We considered the tradeoffs of the various model performance metrics. Using the dataframe-based analysis framework we were able to pivot through model performance results at a variety of prediction thresholds, arrange by model performance metrics, and filter based upon specific criteria and grouping. Seeking optimal model performance, we preprocessed model performance of the various combinations of scaling, sampling, and model structure.

combined <chr>	max_AUC <dbl>	Average_Accuracy <dbl>
Model 4: Log Transformation No Scaling No Adjustment	0.6626522	0.7143659
Model 1: Everything RobustScaler Over-Sample	0.6529411	0.5566341
Model 2: Interactions No Scaling No Adjustment	0.6524546	0.6957317
Model 2: Interactions MinMaxScaler Over-Sample	0.6524485	0.5550488
Model 1: Everything No Scaling No Adjustment	0.6505972	0.6955854
Model 4: Log Transformation RobustScaler No Adjustment	0.6439588	0.7861220
Model 3: Two-Stage No Scaling Over-Sample	0.6045215	0.5376098
Model 3: Two-Stage MinMaxScaler Over-Sample	0.6020093	0.5403902

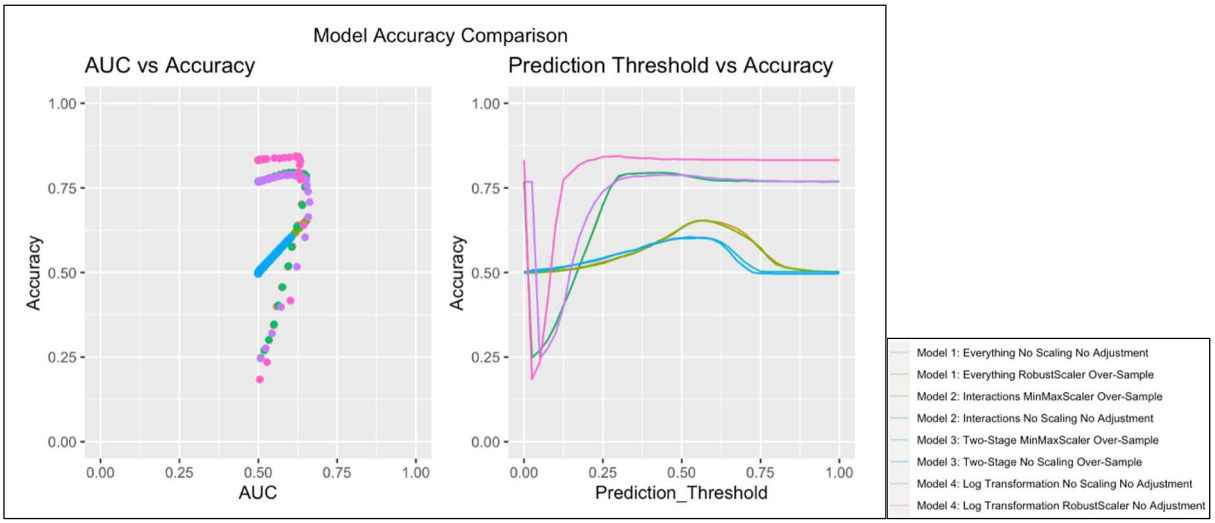
Filtered Data Frame output based on selected criteria.

Name	Accuracy	Precision	Recall	Specificity	F1_Score	FPR	TP	TN	FP	FN	Prediction_Threshold
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Model 1: Everything	0.530	0.516	0.969	0.092	0.673	0.908	8685	830	8151	280	0.25
Model 2: Interactions	0.531	0.516	0.962	0.101	0.672	0.899	8622	903	8078	343	0.25
Model 3: Two-Stage	0.534	0.518	0.955	0.114	0.672	0.886	8564	1027	7954	401	0.25
Model 4: Log Transformation	0.532	0.518	0.935	0.128	0.666	0.872	8063	1107	7512	564	0.25

4.3.3.3 Accuracy and/or Error Rate

Accuracy was evidently sensitive to each specific model and respective prediction threshold level. A separate model we identified with relatively high accuracy was Model 1: Everything with MinMax Scaling and no sampling adjustments at a prediction threshold of .40. The model had an accuracy rate of just over 80%, while also maintaining respectable relative precision and area under the curve. The average accuracy for the respective models, as a function of prediction threshold is captured in the table above.

Name	Scaling	Sampling	Accuracy	Precision	Recall	Specificity	F1_Score	FPR	TP	TN	FP	FN	Prediction_Threshold	AUC
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Model 1: Everything	MinMaxScaler	No Adjustment	0.803	0.634	0.308	0.948	0.414	0.052	810	8517	467	1822	0.4	0.6278847



While Robust Scaling in conjunction with Model 4 appears to perform well here, the predictions are not sufficiently balanced in comparison to the other models. Despite an increased accuracy, the model does not successful classify enough defaults to make it a "better" model than Model 4: Log Transformation No Scaling, No Sampling Adjustment.

4.3.3.4 Other ways of Comparison, if Appropriate

We considered balancing False Positives and False Negatives to hedge prediction error of the model as a part of a hypothetical tailored Credit Card Company intervention upon identification of probable-future defaulters. This would allow some level of reduced uncertainty for the Credit Card Company into the

future regarding defaults. However, the unknown specific application or decision that the prediction model would inform, our model selection and comparison could arrive at different conclusions.

4.3.4 Model Recommendation

Recommend Model 4: log transformation applied to the two-stage model with no scaling and no sampling adjustment at a prediction threshold of .275.

Name	Scaling	Sampling	Accuracy	Precision	Recall	Specificity	F1_Score	FPR	TP	TN	FP	FN	Prediction_Threshold	AUC
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Model 4: Log Transformation	No Scaling	No Adjustment	0.758	0.478	0.459	0.849	0.468	0.151	1188	7295	1299	1403	0.275	0.6536791

At this prediction threshold, the model has an accuracy of 75.8%, balanced precision and recall, and a relatively high AUC when compared to the set of AUCs observed. This model recommendation arbitrarily pursued balanced false positives and false negatives with the intent to hedge model prediction risk potentially associated with planning for future defaults. The model recommendation could be further refined through better domain knowledge of the specific costs of classification error and the costs associated with credit card company interventions.

4.4 Conclusions

4.4.1 How the Questions were Answered?

Our question was answered through meta-analysis of numerous models combined various modeling tuning techniques, resulting in a model selection that we deemed to be acceptable. Our techniques and models provided flexibility to select a final useful in the prediction of the probability of default given specific business decision and costs. Anyone of the models we produced can be used to predict the probability of default to varying degrees of accuracy and with tradeoffs in classification errors.

4.4.2 Interesting Insight into Data

The data shows that limit balance is higher for those who do not default on average. When looking at demographics such as education, sex, and marital status, our data shows that females have a higher limit balance than males and females are less likely to default, according to our graphs. The increase in the balance utilization rate is associated with higher proportions of defaults on average, according to overspending variable. One thing that seemed very interesting in our R output is the ROC Curve Build Model Comparison. This creates a table that provides a confusion matrix for each prediction threshold with 0.05 difference. For example, at 0.3 threshold the false positive rate is 2839 with a log transformation model that has an accuracy of 55.8%. The same model at a 0.5 threshold has a false

positive rate of 572 and accuracy of 67.4%. This shows how increasing the threshold can help increase the accuracy and how the model lowers the false positive rate.

4.4.3 Challenges Faced

There were several challenges faced when dealing with the dataset. Firstly, the imbalanced data set. The unbalanced dataset needed to account for the fewer amount of defaulted personnel. Resampling was required to maximize the number of people that defaulted. The method of down sampling and up sampling were used to help achieve this. The dataset was then split into training and testing datasets. The second challenge that we faced was choosing a threshold. The idea of keeping the threshold at 50% or 0.5 was because we wanted to lower the False Positive Rate. When looking for the threshold, we graphed when a person will likely default and found that average to be around 0.3, but that caused our false positive rate to be higher. If the credit card company wanted to be prepared financially, then it would be better to lower the false positive rate. Another challenge we faced was trying to figure out what model to go with. We lack the applied domain expertise to claim that one model is the fix all. Depending on the credit card company and what they would like to investigate, our models can answer several different questions.