Credit Card Default Analysis

Honor Code: ``The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment:'' (ADD: names of persons or web resources, if any, excluding the instructor, TAs, and materials posted on course website)

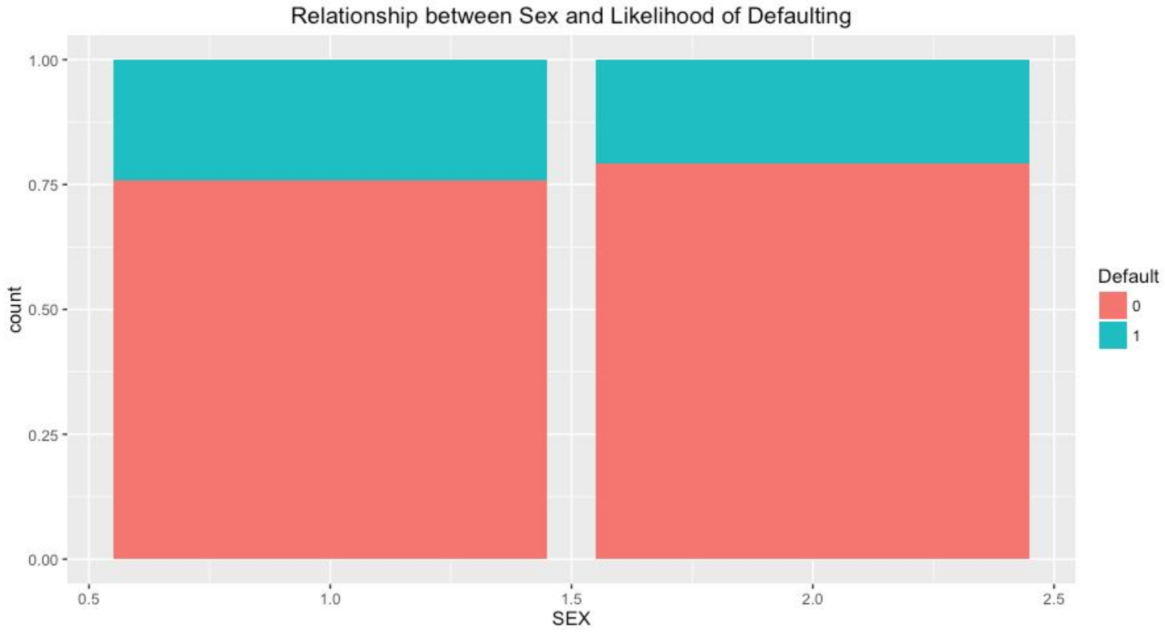
I. Introduction

We are using the UCI Default of Credit Card Clients Dataset from Taiwan. This dataset spans 6 months,from April 2005 to September 2005 and includes limit balance, owe amount, pay amount, age, as well as categorical variables like marital status, sex,education, and whether they defaulted the next month. The data set also provides the unique id's for all the clients. For sex, the data assigns a 1 to males and 2 to females. For marital status, the data assigns 1 to married, 2 to single, and 3 to others[1]. For education, the data assigns 1 to graduate degree,2 to university degree, 3 high school, and 4 to others.[2] For default next month the data assigns 1 for yes and 2 for no.

II. Analysis

We begin our analysis by examining the effects of sex on default rates.

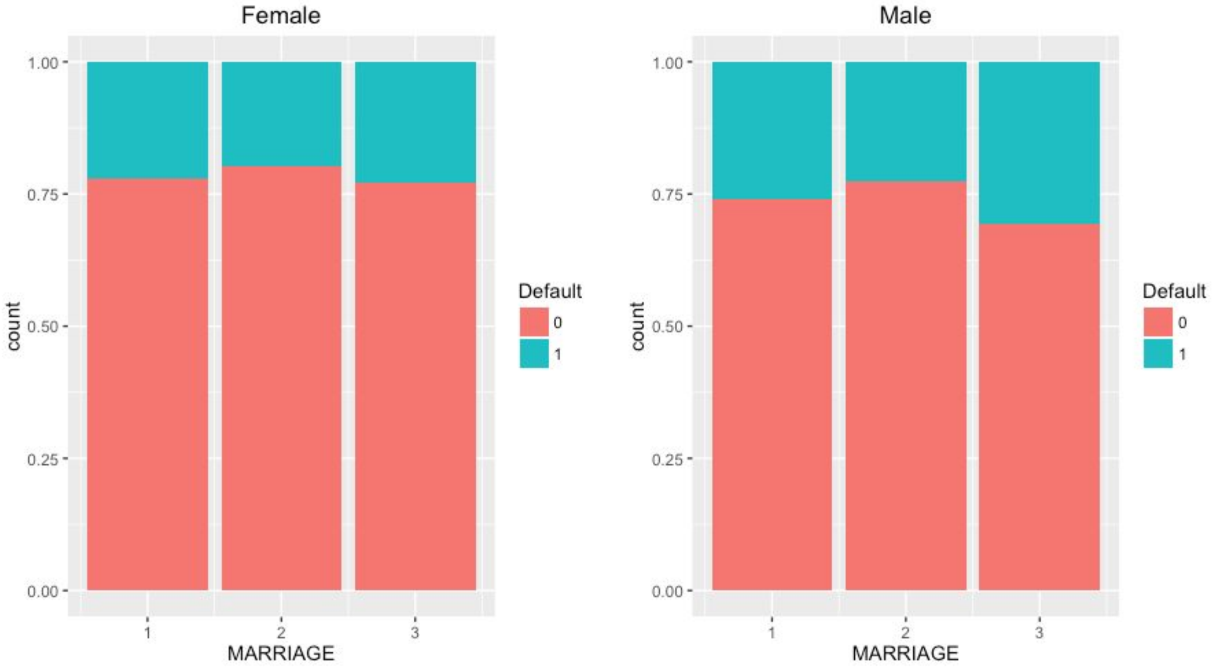
Figure 1: Relationship between Sex and Likelihood of Defaulting



With default being 1 and not defaulting being 0, Figure 1 shows that males are more likely to default as compared to females. After performing a chi-squared test, this data is seen to be strongly significant with a p-value of $4.945e^{-12}$.

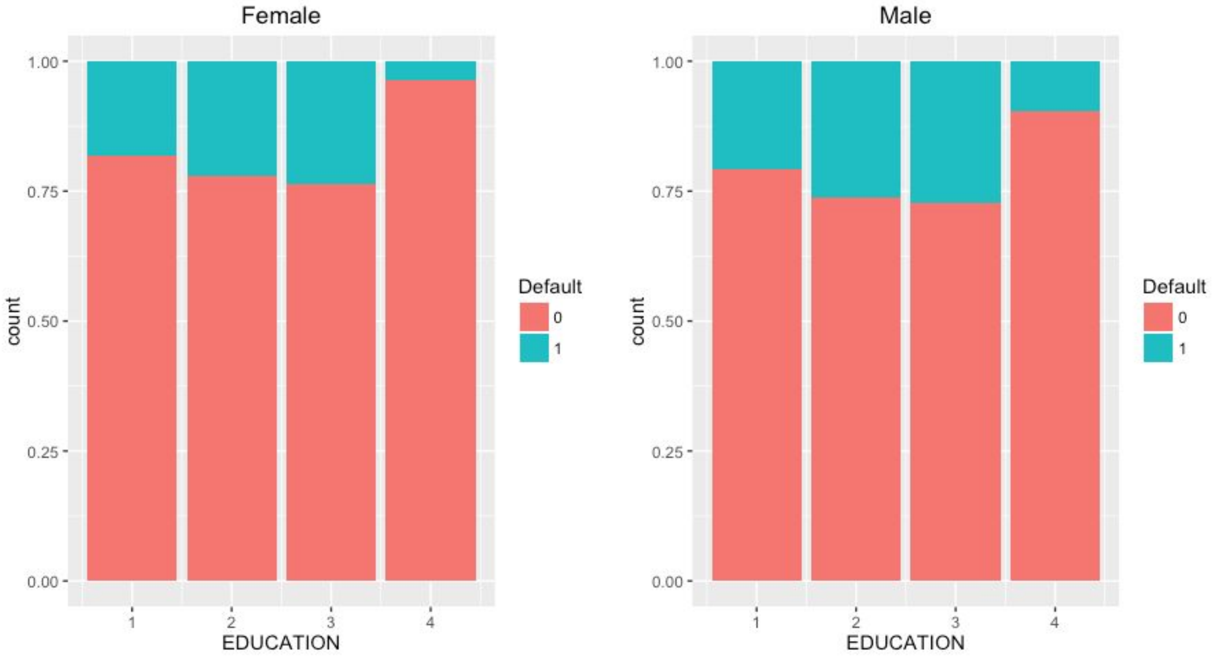
[1] 0 is unknown, we omitted them from our analysis.
[2] 5,6 are both unknown, we removed them for analysis.

Figure 2: Relationship between Sex and Likelihood of Defaulting (separated by marital status)



After separating by marital status, we see a trend where married clients tend to default more than single clients. This could be because married couples have more assets, and it may make sense for them to default.[3]

Figure 3: Relationship between Sex and Likelihood of Defaulting (separated by education)



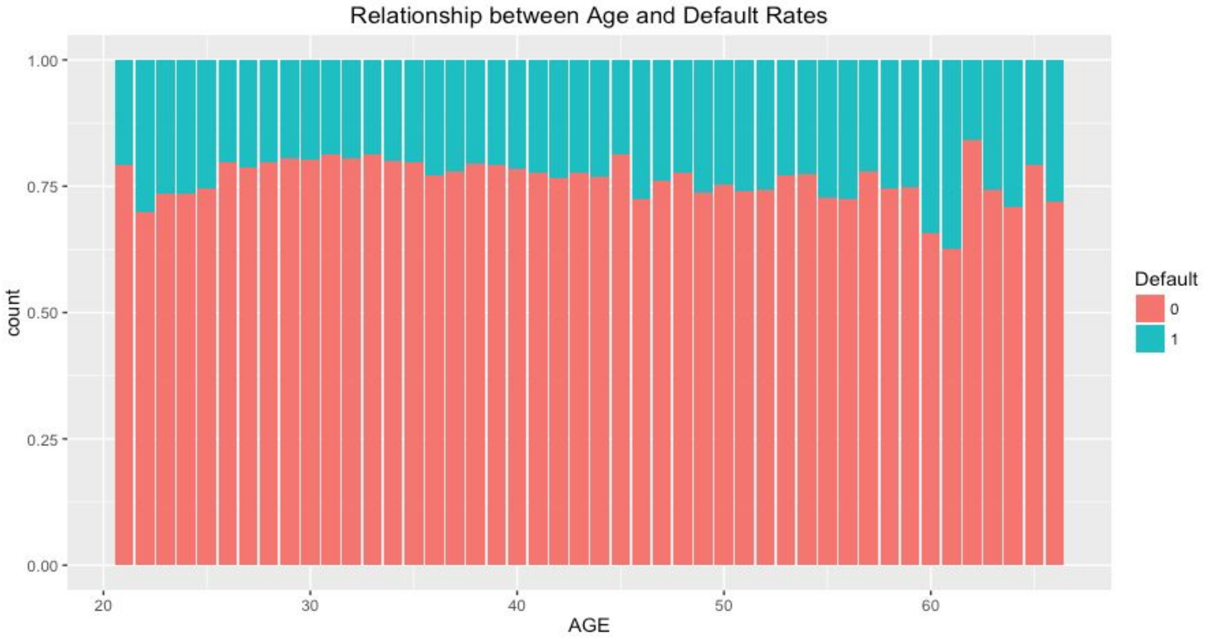
[3] http://www.theatlantic.com/business/archive/2015/07/mortgage-default-finances/398051/

Separating by education shows us that less educated one is, regardless of sex, the more likely one is going to default. A reason for this is because, according to studies[4], 90% of students who default are those who did not graduate and earn degrees. Furthermore, graduate students tend to earn more money, which increases their likelihood of repaying their debt and lowers their likelihood of defaulting on loans.

In sum, regardless of separating through education or marriage, the main trend seen in Figure 1 still holds true in that males tend to default more than females in both Figures 2 and 3. Studies[5] have shown that men, on average, carries 4.3% more debt than women and also take 4.9% more home loans than women.

Another significant fact we found from our analysis is that age does not account for much when it comes to predicting default rates. This is going to be explained through figures 4 through 6.

Figure 4: Relationship between Age and Default Rates

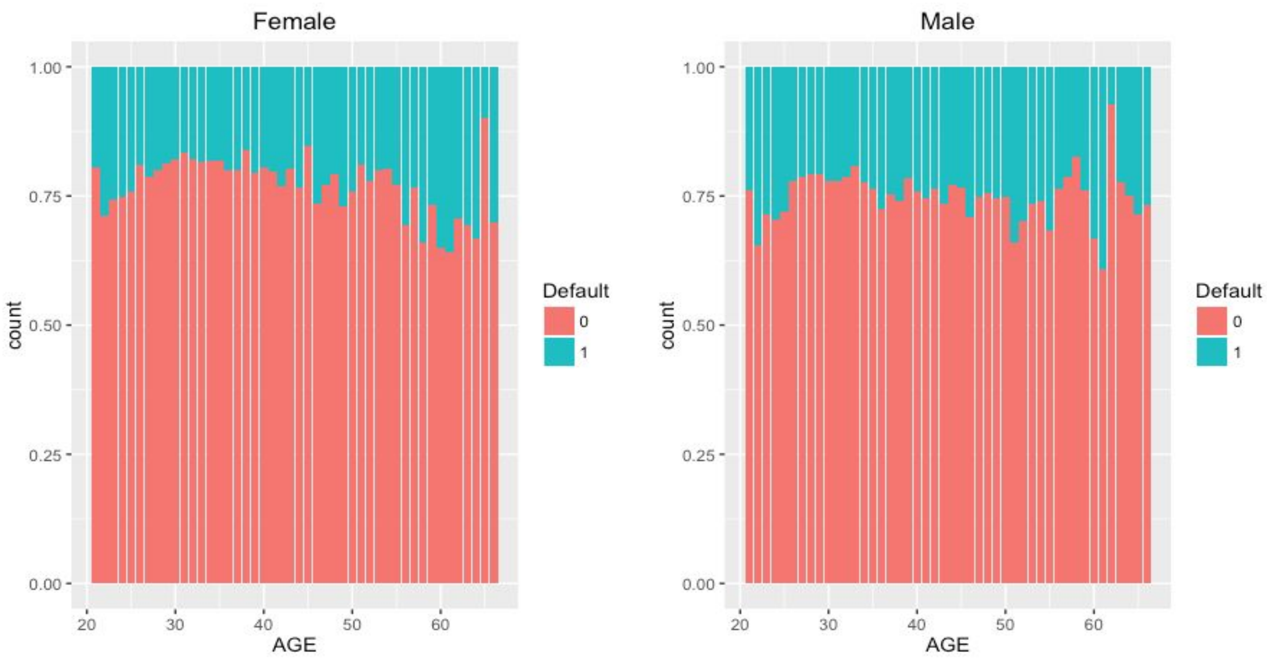


It is important to note that ages 67 and above have been omitted from the data due to a lack of samples, therefore making the data too sparse. From Figure 4, it appears that the data is spread relatively evenly across the ages. We then separated the ages by gender to see if that affect age's predictability of default rates.

Figure 5: Relationship between Age and Default Rates (separated by sex)

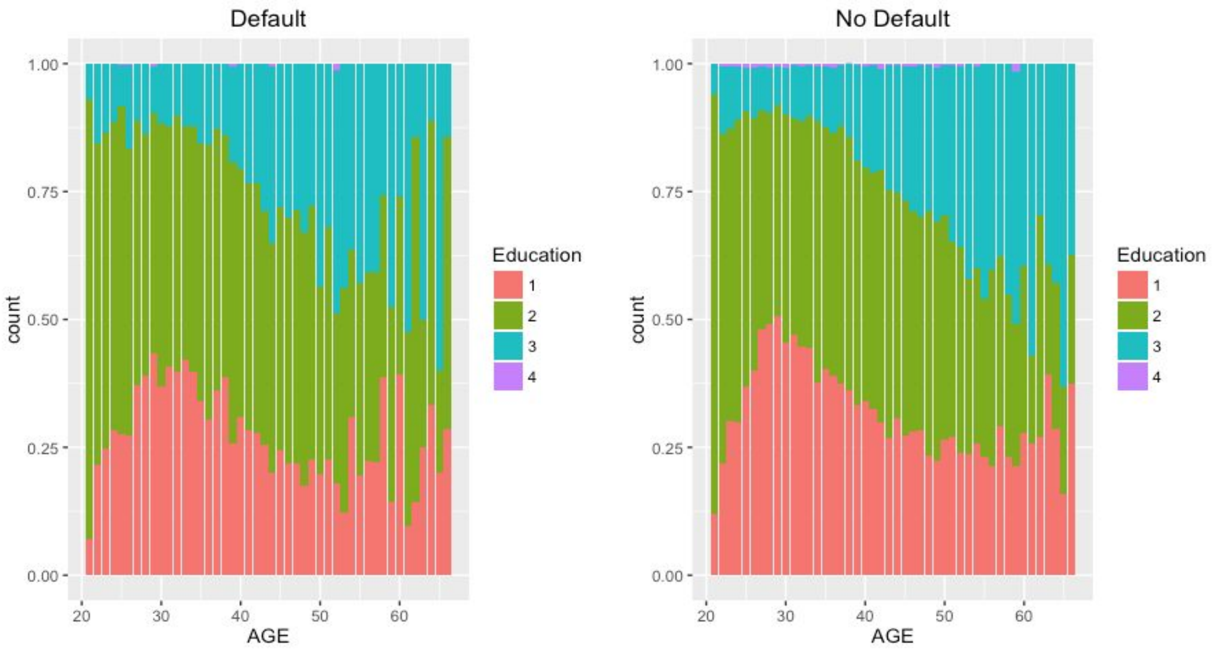
[4] https://www.insidehighered.com/news/2015/09/28/four-surprising-findings-debt-and-default-among-community-college-students

[5] http://www.bankrate.com/finance/debt/men-women-and-debt-does-gender-matter.aspx#ixzz4S8mfnGYb



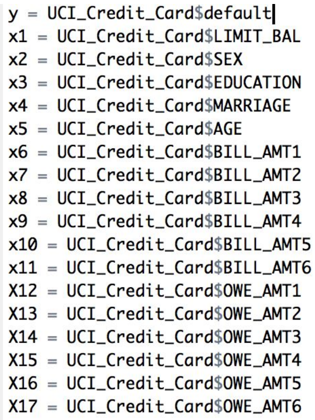
As seen in Figure 3, there is a trend in which the higher educated one is, the less likely one is to default one's loans. This trend is further deconstructed across ages. This trend holds, and is seen below in Figure 6.

Figure 6: Relationship between Age and Education (separated by default rate)



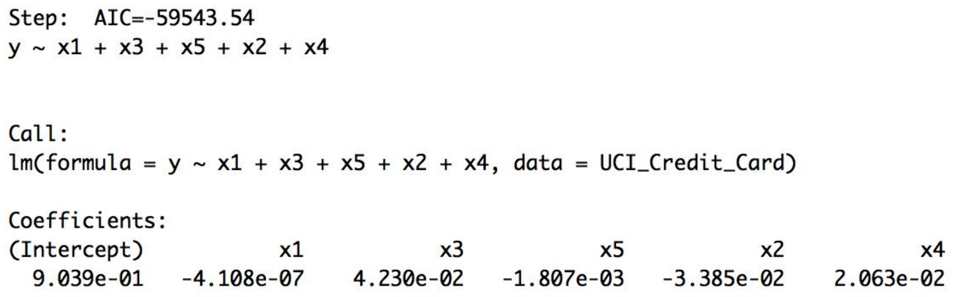
It is seen that the trend in Figure 3 holds true in that less educated people are more likely to default on their loans. Additionally, there is a decline in graduate and undergraduate educated people with increased age. This is probably due to the fact that Taiwan was a developing country. Consequently, the younger generation have had more exposure to education than the older generation.

Next, we checked whether some of the variables (sex, limit balance, sex, education, marriage, and age) can determine whether the person defaulted or not. To check whether there is a correlation between these variables we start with our response variable with no predictors and we add one predictor variable at a time. The variables are as follows:



```
y = UCI_Credit_Card$default
x1 = UCI_Credit_Card$LIMIT_BAL
x2 = UCI_Credit_Card$SEX
x3 = UCI_Credit_Card$EDUCATION
x4 = UCI_Credit_Card$MARRIAGE
x5 = UCI_Credit_Card$AGE
x6 = UCI_Credit_Card$BILL_AMT1
x7 = UCI_Credit_Card$BILL_AMT2
x8 = UCI_Credit_Card$BILL_AMT3
x9 = UCI_Credit_Card$BILL_AMT4
x10 = UCI_Credit_Card$BILL_AMT5
x11 = UCI_Credit_Card$BILL_AMT6
X12 = UCI_Credit_Card$OWE_AMT1
X13 = UCI_Credit_Card$OWE_AMT2
X14 = UCI_Credit_Card$OWE_AMT3
X15 = UCI_Credit_Card$OWE_AMT4
X16 = UCI_Credit_Card$OWE_AMT5
X17 = UCI_Credit_Card$OWE_AMT6
```

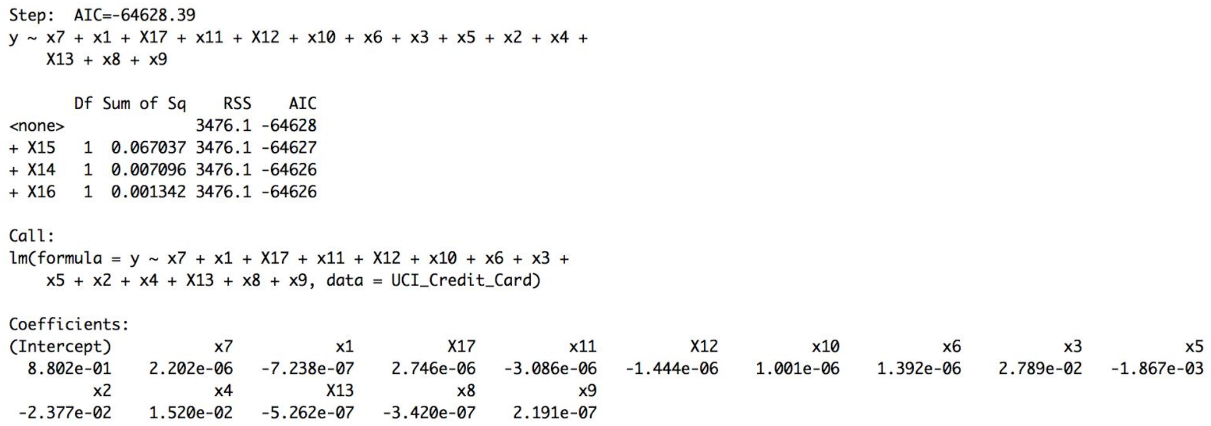
After iterating through, we get the final model pictured below:



```
Step:  AIC=-59543.54
y ~ x1 + x3 + x5 + x2 + x4


Call:
lm(formula = y ~ x1 + x3 + x5 + x2 + x4, data = UCI_Credit_Card)

Coefficients:
(Intercept)           x1           x3           x5           x2           x4  
  9.039e-01   -4.108e-07    4.230e-02   -1.807e-03   -3.385e-02    2.063e-02  
```

However, since the absolute value of AIC is quite large we can conclude that the five variables can not ultimately determine whether a person defaults or not.

The next test is whether or not the five predictors above with the bill and owe amounts over the next 6 months can determine whether one defaults or not. After iterating through, we get the final model pictured below:

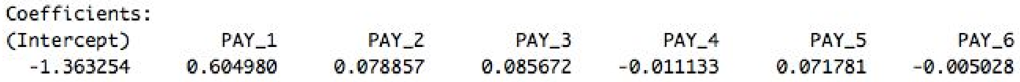


```
Step:  AIC=-64628.39
y ~ x7 + x1 + X17 + x11 + X12 + x10 + x6 + x3 + x5 + x2 + x4 +
    X13 + x8 + x9

       Df Sum of Sq    RSS    AIC
<none>              3476.1 -64628
+ X15   1  0.067037 3476.1 -64627
+ X14   1  0.007096 3476.1 -64626
+ X16   1  0.001342 3476.1 -64626

Call:
lm(formula = y ~ x7 + x1 + X17 + x11 + X12 + x10 + x6 + x3 +
    x5 + x2 + x4 + X13 + x8 + x9, data = UCI_Credit_Card)

Coefficients:
(Intercept)           x7           x1          X17          x11          X12          x10           x6           x3           x5
  8.802e-01    2.202e-06   -7.238e-07    2.746e-06   -3.086e-06   -1.444e-06    1.001e-06    1.392e-06    2.789e-02   -1.867e-03
         x2           x4          X13           x8           x9
 -2.377e-02    1.520e-02   -5.262e-07   -3.420e-07    2.191e-07
```

Here we see that owe amount 3, owe amount 4, and owe amount 5 were dropped. This means that they are insignificant values in determining whether or not the person defaulted. Adding more values only increased AIC, thus including owe amounts and bill amounts over the 6 month period only weakens the model in predicting default.

After looking at which categorical and continuous variables affect the default payment, we did logistic regression. We looked at how the history of payment record affected the default payment. For each logistic regression model, we split the data in 80% training and 20% testing. For each set of categorical variables, we found the model (coefficients), the confusion matrix, the accuracy, and the misclassification rate.

All variables:

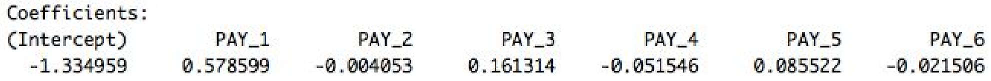


```
Coefficients:
(Intercept)       PAY_1       PAY_2       PAY_3       PAY_4       PAY_5       PAY_6
  -1.363254    0.604980    0.078857    0.085672   -0.011133    0.071781   -0.005028
```

	0	1
0	4641	93
1	994	272

Accuracy: 0.8188, Misclassification Rate: 0.1812

Males:

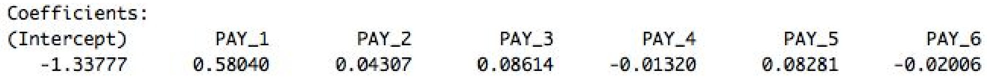


```
Coefficients:
(Intercept)       PAY_1       PAY_2       PAY_3       PAY_4       PAY_5       PAY_6
  -1.334959    0.578599   -0.004053    0.161314   -0.051546    0.085522   -0.021506
```

	0	1
0	8937	213
1	2155	583

Accuracy: 0.8008, Misclassification Rate: 0.1992

Females:

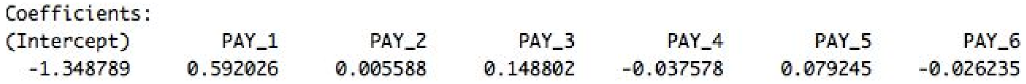


```
Coefficients:
(Intercept)       PAY_1       PAY_2       PAY_3       PAY_4       PAY_5       PAY_6
   -1.33777     0.58040     0.04307     0.08614    -0.01320     0.08281    -0.02006
```

	0	1
0	11895	243
1	2665	707

Accuracy: 0.8125, Misclassification Rate: 0.1875

Graduate School:

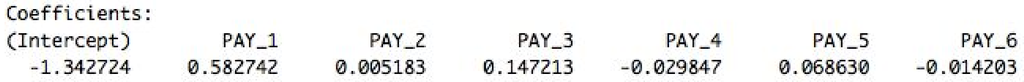


```
Coefficients:
(Intercept)       PAY_1       PAY_2       PAY_3       PAY_4       PAY_5       PAY_6
  -1.348789    0.592026    0.005588    0.148802   -0.037578    0.079245   -0.026235
```

	0	1
0	7903	200
1	1951	531

Accuracy: 0.7968, Misclassification Rate: 0.2032

University:

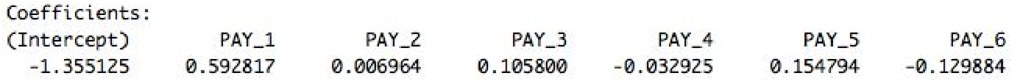


```
Coefficients:
(Intercept)       PAY_1       PAY_2       PAY_3       PAY_4       PAY_5       PAY_6
  -1.342724    0.582742    0.005183    0.147213   -0.029847    0.068630   -0.014203
```

	0	1
0	10683	235
1	2468	644

Accuracy: 0.8073, Misclassification Rate: 0.1927

High School:

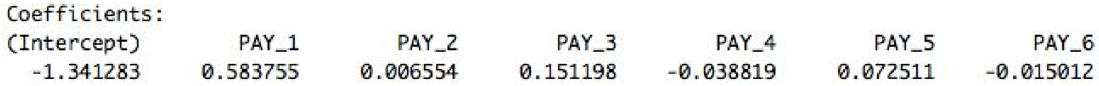


```
Coefficients:
(Intercept)       PAY_1       PAY_2       PAY_3       PAY_4       PAY_5       PAY_6
  -1.355125    0.592817    0.006964    0.105800   -0.032925    0.154794   -0.129884
```

	0	1
0	3731	81
1	900	205

Accuracy: 0.8004, Misclassification Rate: 0.1996

Married:

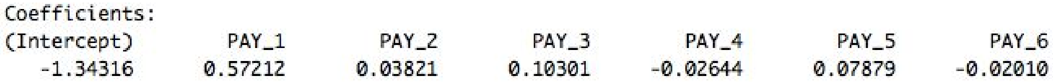


```
Coefficients:
(Intercept)      PAY_1      PAY_2      PAY_3      PAY_4      PAY_5      PAY_6
  -1.341283   0.583755   0.006554   0.151198  -0.038819   0.072511  -0.015012
```

	0	1
0	10398	230
1	2403	628

Accuracy: 0.8072, Misclassification Rate: 0.1928

Single:

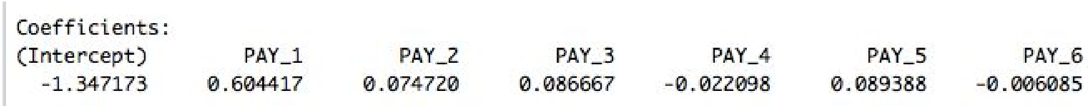


```
Coefficients:
(Intercept)     PAY_1     PAY_2     PAY_3     PAY_4     PAY_5     PAY_6
   -1.34316   0.57212   0.03821   0.10301  -0.02644   0.07879  -0.02010
```

	0	1
0	12231	248
1	2769	716

Accuracy: 0.811, Misclassification Rate: 0.189

Female, University, Single:



```
Coefficients:
(Intercept)      PAY_1      PAY_2      PAY_3      PAY_4      PAY_5      PAY_6
  -1.347173   0.604417   0.074720   0.086667  -0.022098   0.089388  -0.006085
```

	0	1
0	6990	140
1	1433	392

Accuracy: 0.8243, Misclassification Rate: 0.1757

After going through the categorical variables, we found that Single Female who had a University Education had the most accurate model in predicting default or no default for the next payment with an accuracy of 0.8243.

III. Conclusion

Overall we found that the variables that best predicted default rates is sex, marital status, and education. Specifically males, married people, and those who did not get an education are more likely to default. However, looking at the AIC values, we can not say that those three factors can always predict whether or not a person defaulted. There are many unknown, confounding variables that could be playing a role in this dataset. For instance, owning multiple credit cards, credit score, purchase history, etc.

Appendix

```
install.packages("leaps")
library(leaps)

summary(UCI_Credit_Card_csv$TOTAL_OWED)

attach(UCI_Credit_Card_csv)
corrgram(UCI_Credit_Card_csv)

ggplot(UCI_Credit_Card_csv,aes(x = AGE,y = TOTAL_OWED,color = factor(MARRIAGE))) +
geom_point()+geom_smooth()

plot(residuals(lm(TOTAL_OWED~AGE)))


## Figure 1: Age vs. Total Owed

ggplot(UCI_Credit_Card_csv,aes(AGE,TOTAL_OWED,col=factor(SEX))) +
geom_point(alpha=0.5) +
  labs(title='Age vs. Total Amount Owed',y='Total Owed in Dollars',x='Age') + geom_smooth()

UCI_Credit_Card_csv %>% filter(SEX == '1') %>% hexbin(AGE,TOTAL_OWED)

plot(hexbin(AGE,TOTAL_OWED,xbins=50))
summary(hexbin(AGE,TOTAL_OWED,xbins=50))

## default separated by sex

ggplot(UCI_Credit_Card_csv,aes(x = SEX,fill = factor(default.payment.next.month))) +
geom_bar(position = "fill")
chisq.test(SEX,default.payment.next.month) # statistically significant

# males tend to default more

## default separated by marriage

ggplot(UCI_Credit_Card_csv,aes(x = MARRIAGE,fill = factor(default.payment.next.month))) +
geom_bar(position = "fill")

##

Male = UCI_Credit_Card_csv %>% filter(SEX == '1', MARRIAGE != '0') %>%
  ggplot(aes(MARRIAGE,fill=factor(default.payment.next.month))) + geom_bar(position='fill') +
  labs(title='Male') + guides(fill=guide_legend(title='Default'))

Female = UCI_Credit_Card_csv %>% filter(SEX == '2', MARRIAGE != '0') %>%
  ggplot(aes(MARRIAGE,fill=factor(default.payment.next.month))) + geom_bar(position='fill') +
  labs(title='Female') + guides(fill=guide_legend(title='Default'))

grid.arrange(Female,Male,ncol=2)

# 1: SINGLE, 2: MARRIED, 3: OTHERS
# No interaction. Males still tend to default more. In general, singles tend to default more.

Male = UCI_Credit_Card_csv %>% filter(SEX == '1', EDUCATION != '0', EDUCATION != '5',
EDUCATION != '6') %>%
  ggplot(aes(EDUCATION,fill=factor(default.payment.next.month))) + geom_bar(position='fill') +
  labs(title='Male') + guides(fill=guide_legend(title='Default'))

Female = UCI_Credit_Card_csv %>% filter(SEX == '2', EDUCATION != '0', EDUCATION != '5',
EDUCATION != '6') %>%
  ggplot(aes(EDUCATION,fill=factor(default.payment.next.month))) + geom_bar(position='fill') +
  labs(title='Female') + guides(fill=guide_legend(title='Default'))

grid.arrange(Female,Male,ncol=2)

# The more uneducated, the less likely you are to default. Trend of male and female follow
general trend.

# Default, Age vs. Education

Default = UCI_Credit_Card_csv %>% filter(default.payment.next.month == '1', EDUCATION !=
'0', EDUCATION != '5', EDUCATION != '6', AGE < 67) %>%
  ggplot(aes(AGE,fill=factor(EDUCATION))) + geom_bar(position='fill') +
  labs(title='Default') + guides(fill=guide_legend(title='Education'))

No_Default = UCI_Credit_Card_csv %>% filter(default.payment.next.month == '0', EDUCATION
!= '0', EDUCATION != '5', EDUCATION != '6', AGE < 67) %>%
  ggplot(aes(AGE,fill=factor(EDUCATION))) + geom_bar(position='fill') +
  labs(title='No Default') + guides(fill=guide_legend(title='Education'))

grid.arrange(Default,No_Default,ncol=2)

# Omitted ages 67 and above because not enough data; too sparse.
# There is no interaction between age and education with regards to default rate. Trend is same
throughout.
```

```
# Age vs. Default

Age_Default = UCI_Credit_Card_csv %>% filter(AGE < 67) %>%
ggplot(aes(AGE,fill=factor(default.payment.next.month))) + geom_bar(position='fill') +
  labs(title='Relationship between Age and Default Rates') +
guides(fill=guide_legend(title='Default'))

##### Can we do this instead?
library(MASS)
stepAIC(glm(formula = default.payment.next.month ~ MARRIAGE + SEX + AGE + EDUCATION
+ OWE_AMT6 + LIMIT_BAL + BILL_AMT6, family = binomial, data = UCI_Credit_Card_csv))

hexbinplot(LIMIT_BAL~AGE, type = 'r')

summary(lm(LIMIT_BAL~AGE))


#if have positive value then they still owe money
UCI_Credit_Card$default = 0
for (i in 1:length(UCI_Credit_Card$TOTAL_OWED)){
  if (UCI_Credit_Card$TOTAL_OWED[i] > 0){
    UCI_Credit_Card$default[i] = 1
  }
  else{
    UCI_Credit_Card$default[i] = 0
  }
}

x1 = UCI_Credit_Card$LIMIT_BAL
x2 = UCI_Credit_Card$SEX
x3 = UCI_Credit_Card$EDUCATION
x4 = UCI_Credit_Card$MARRIAGE
x5 = UCI_Credit_Card$AGE
x6 = UCI_Credit_Card$BILL_AMT1
x7 = UCI_Credit_Card$BILL_AMT2
x8 = UCI_Credit_Card$BILL_AMT3
x9 = UCI_Credit_Card$BILL_AMT4
x10 = UCI_Credit_Card$BILL_AMT5
x11 = UCI_Credit_Card$BILL_AMT6
X12 = UCI_Credit_Card$OWE_AMT1
X13 = UCI_Credit_Card$OWE_AMT2
X14 = UCI_Credit_Card$OWE_AMT3
```

```
X15 = UCI_Credit_Card$OWE_AMT4
X16 = UCI_Credit_Card$OWE_AMT5
X17 = UCI_Credit_Card$OWE_AMT6



subs = regsubsets(y~x1+x2+x3+x4+x5, data=UCI_Credit_Card)
#parameters for each best p model
P = summary(subs)$which

#CP for each best model
CP = summary(subs)$cp

#rsqured for each best model
RP = summary(subs)$rsq

#adjusted r for each model
RAP = summary(subs)$adjr2

#aic and bic for each model
dat = lm(y~x1+x2+x3+x4+x5)
aic = AIC(dat)
bic = BIC(dat)
aic
bic

#forward step function model
cbind(p,CP,RP,RAP,AIC,BIC)
empty = lm(y~1, data=UCI_Credit_Card)
step(empty, scope=y~x1+x2+x3+x4+x5,direction="forward")
step(empty, scope =
y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+X12+X13+X14+X15+X16+X17, direction =
"forward")

UCI_Credit_Card$PAY_1 = UCI_Credit_Card$PAY_0

#all

train_num = ceiling(nrow(UCI_Credit_Card) * 0.8)
train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(UCI_Credit_Card),]
log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
            family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)


##depending on sex
male = UCI_Credit_Card[ which(UCI_Credit_Card$SEX == "1"),]
train_num = ceiling(nrow(male) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(male),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
            family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)


female = UCI_Credit_Card[ which(UCI_Credit_Card$SEX == "2"),]
train_num = ceiling(nrow(female) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(female),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
            family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)

##depending on education
gradschool = UCI_Credit_Card[ which(UCI_Credit_Card$EDUCATION == "1"),]
train_num = ceiling(nrow(gradschool) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(gradschool),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
        family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)

uni = UCI_Credit_Card[ which(UCI_Credit_Card$EDUCATION == "2"),]
train_num = ceiling(nrow(uni) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(uni),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
        family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)

high = UCI_Credit_Card[ which(UCI_Credit_Card$EDUCATION == "3"),]
train_num = ceiling(nrow(high) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(high),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
        family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)

##depending on marriage
married = UCI_Credit_Card[which(UCI_Credit_Card$MARRIAGE == "1"),]
train_num = ceiling(nrow(married) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(married),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
        family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
```

```
log_con = table(true = test$default.payment.next.month, model = log_pred)

single = UCI_Credit_Card[ which(UCI_Credit_Card$MARRIAGE == "2"),]
train_num = ceiling(nrow(single) * 0.8)

train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(single),]

log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
          family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)

#BEST 3
low = UCI_Credit_Card[ which(UCI_Credit_Card$SEX == "2" | UCI_Credit_Card$EDUCATION
== "2" | UCI_Credit_Card$MARRIAGE == "1"),]
train_num = ceiling(nrow(low) * 0.8)
train = UCI_Credit_Card[1:train_num,]
test = UCI_Credit_Card[train_num + 1:nrow(low),]
log_model = glm(default.payment.next.month ~ (PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
PAY_6), train,
          family = binomial)
log_pred = predict(log_model, test, type = "response")
log_pred = (log_pred > 0.5) + 1
log_con = table(true = test$default.payment.next.month, model = log_pred)
```