UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

Project 2 – Group 7

STAT 6021 ~ Fall 2022

12/14/2022

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

Agenda

01 **Questions of Interest**

Project Motivation

02 **The Data**

Data and Variable Descriptions

03 **Visualizations**

Addressing the Questions of Interest

04 **Models**

Answering Questions of Interest with Models

05 **Results**

High-level analysis of results

First Topic

Questions of Interest

Linear Regression 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?

Logistical Regression Question

What is the likelihood of credit card default considering variety of demographic factors and recent monthly payments?

Second Topic

The Data

The Data

Data Source

- Default of Credit Card Clients Dataset
- Retrieved from Kaggle.com
- Contains data related to credit card statements from clients in Taiwan during certain months in 2005
- Publicly available dataset includes payment and demographic data of credit card holders from an unspecified Taiwanese bank.

Data Structure

- The raw dataset included 30,000 observations on 24 variables
- Combined Numeric and Categoric Variables
- Credit card holder data, wide data

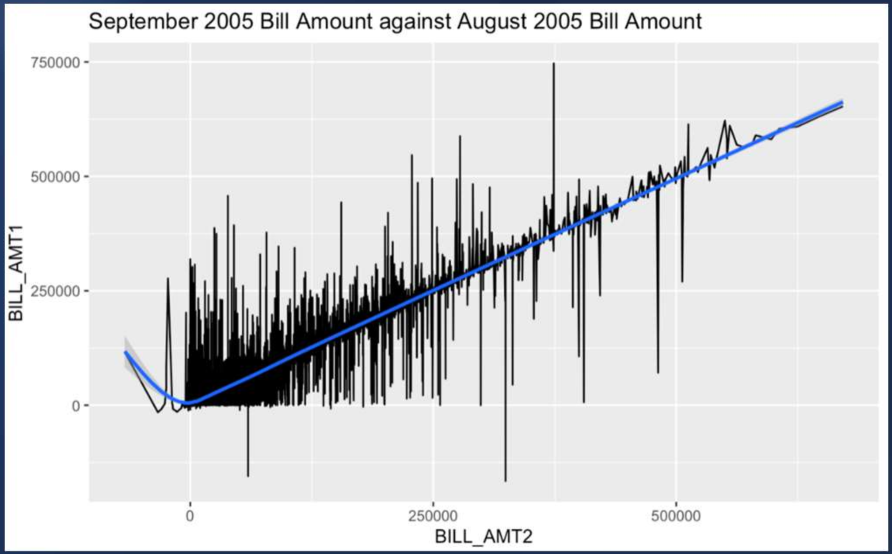
Variables

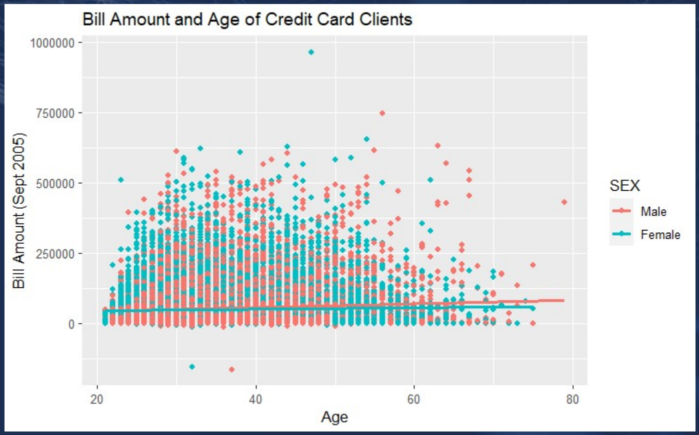
DEFAULT: binary factor representing defaulting on next payment
LIMIT_BAL: numeric representing maximum card limit
BILL_AMT1 … BILL_AMT6: numeric representing monthly bill amount
PAY_AMT1 … PAY_AMT6: numeric representing amount of monthly payments
PAY_2 … PAY_6: factor representing number of months behind payments
SEX: factor representing male or female
AGE: numeric representing age
MARRIAGE: factor representing marital status
EDUCATION: factor representing education level
AVG_UTILIZATION_RATE: numeric representing
OVERSPENDER: factor representing payments exceed limit balance
TOTAL_EXCESS_BALANCE: numeric representing running amount over the limit balance

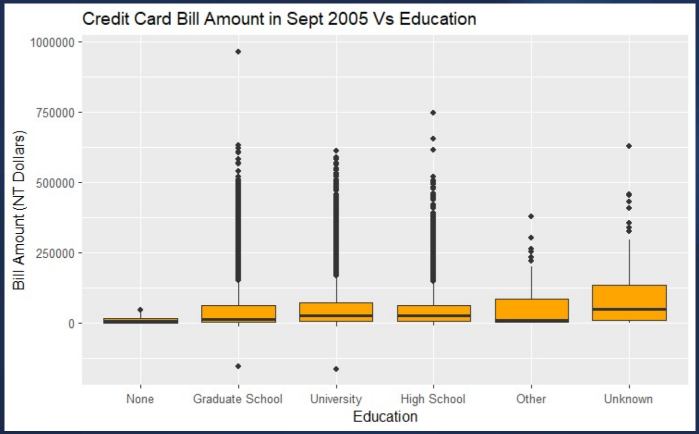
Third Topic

Visualizations

Linear Visualization

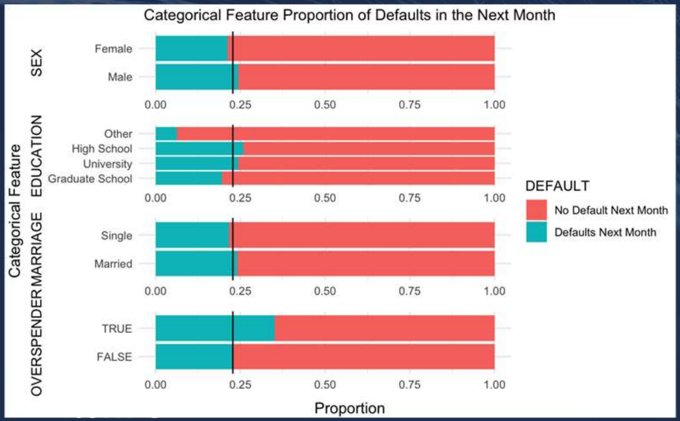




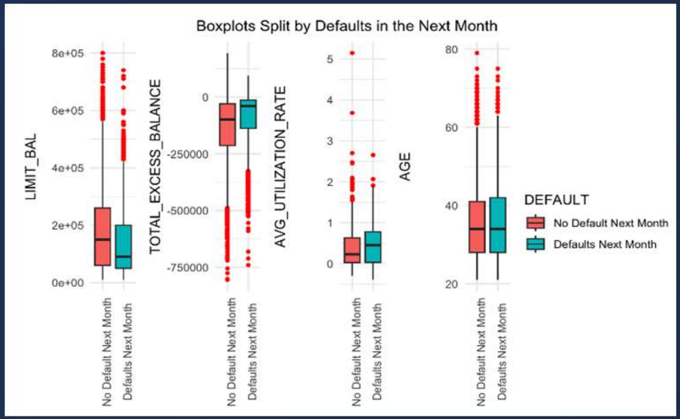


Logistic Visualizations

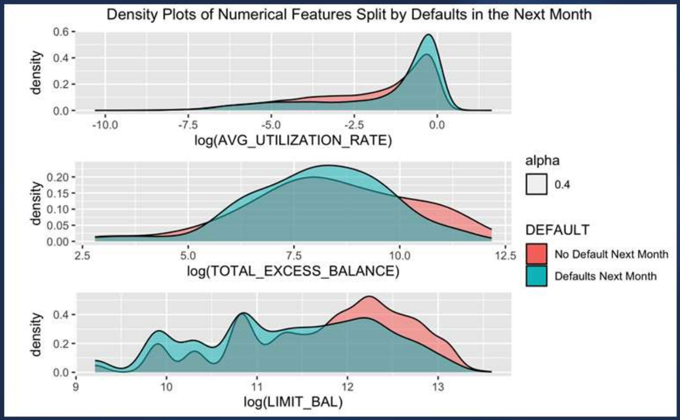
A)



B)



C)



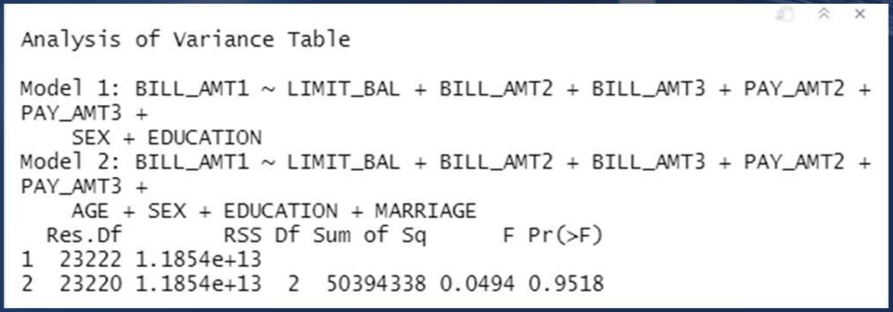
UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

Fourth Topic

Models

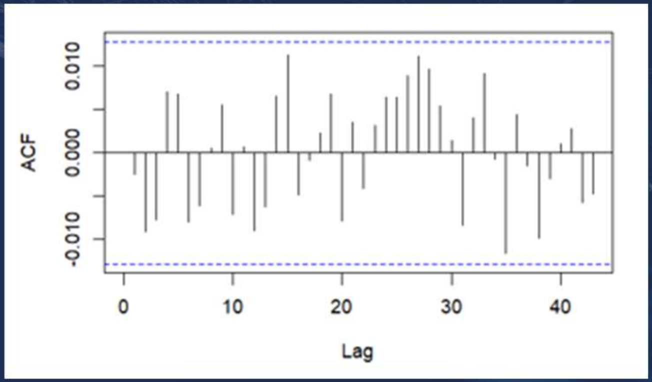
UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

Linear Model



```
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
```



$$
\begin{aligned}
\mathit{BILL\_AMT1} = {} & -869.60 + .0145\mathit{LIMIT\_BAL} + .8842\mathit{BILL\_AMT2} + .1059\mathit{BILL\_AMT3} \\
& - .0800\mathit{PAY\_AMT2} + .0693\mathit{PAY\_AMT3} + .0689\mathit{SEX\_Male} \\
& + 152.30\mathit{EDUCATIONHigh\ School} + .0068\mathit{EDUCATIONOther} \\
& + .0017\mathit{EDUCAITONUniversity}
\end{aligned}
$$

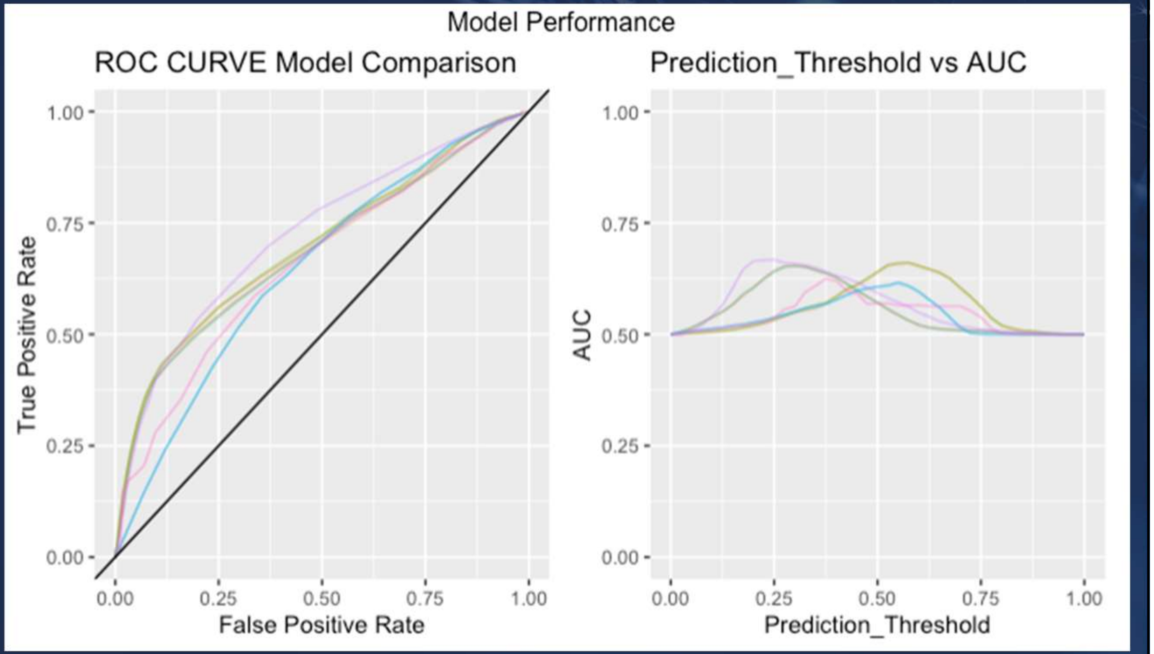
Logistic Model

Generic Model Form:

$$\log\left[\frac{P(DEFAULT = 1)}{1 - P(DEFAULT = 1)}\right] = \alpha + \beta_1(X_1)+\ldots+\beta_n(X_n)$$

Models

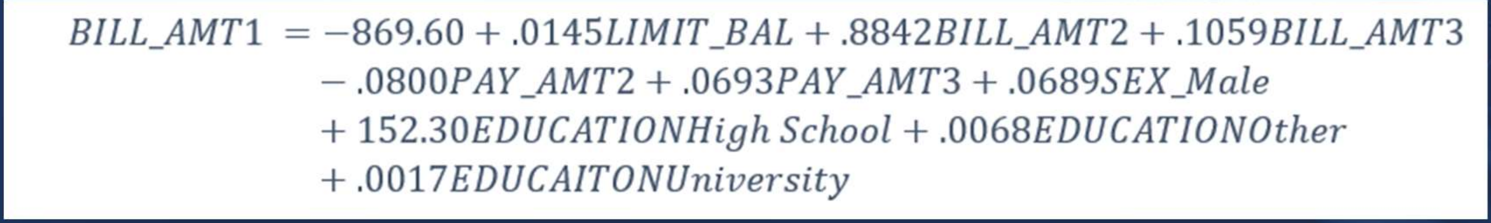
- ● Model 1: Everything No Scaling No Adjustment
- ● Model 1: Everything RobustScaler Over-Sample
- ● Model 2: Interactions MinMaxScaler Over-Sample
- ● Model 2: Interactions No Scaling No Adjustment
- ● Model 3: Two-Stage MinMaxScaler Over-Sample
- ● Model 3: Two-Stage No Scaling Over-Sample
- ● Model 4: Log Transformation No Scaling No Adjustment
- ● Model 4: Log Transformation RobustScaler No Adjustment



Fifth Topic

Results

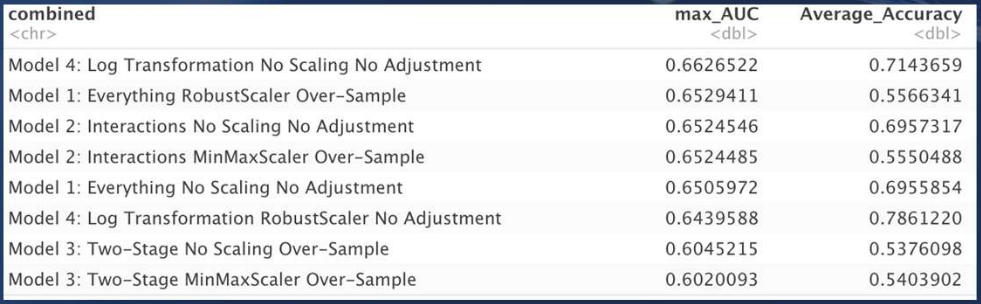
Linear Conclusions



$$\begin{aligned} BILL\_AMT1 &= -869.60 + .0145LIMIT\_BAL + .8842BILL\_AMT2 + .1059BILL\_AMT3 \\ &\quad - .0800PAY\_AMT2 + .0693PAY\_AMT3 + .0689SEX\_Male \\ &\quad + 152.30EDUCATIONHigh\ School + .0068EDUCATIONOther \\ &\quad + .0017EDUCAITONUniversity \end{aligned}$$

Logistic Conclusions

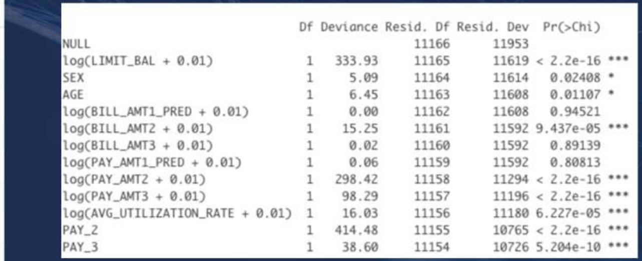
Model Comparison



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

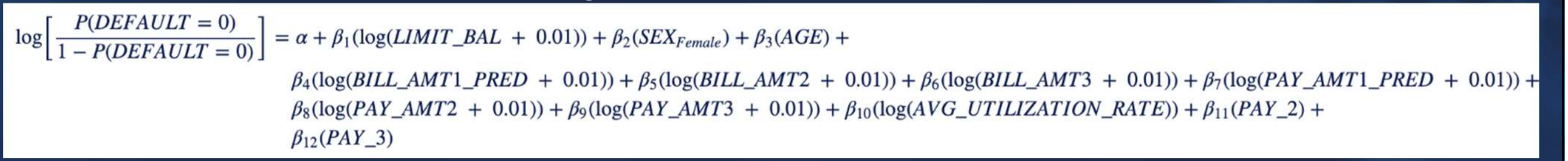
Selected Model Test Performance:

Anova(mod4, test = "chisq"):

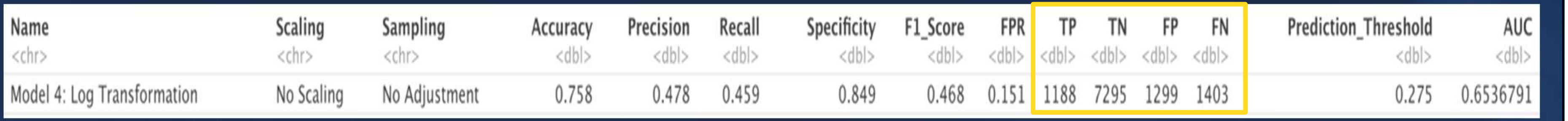


```
                                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                            11166      11953              
log(LIMIT_BAL + 0.01)             1   333.93    11165      11619 < 2.2e-16 ***
SEX                               1     5.09    11164      11614   0.02408 *  
AGE                               1     6.45    11163      11608   0.01107 *  
log(BILL_AMT1_PRED + 0.01)        1     0.00    11162      11608   0.94521    
log(BILL_AMT2 + 0.01)             1    15.25    11161      11592 9.437e-05 ***
log(BILL_AMT3 + 0.01)             1     0.02    11160      11592   0.89139    
log(PAY_AMT1_PRED + 0.01)         1     0.06    11159      11592   0.80813    
log(PAY_AMT2 + 0.01)              1   298.42    11158      11294 < 2.2e-16 ***
log(PAY_AMT3 + 0.01)              1    98.29    11157      11196 < 2.2e-16 ***
log(AVG_UTILIZATION_RATE + 0.01)  1    16.03    11156      11180 6.227e-05 ***
PAY_2                             1   414.48    11155      10765 < 2.2e-16 ***
PAY_3                             1    38.60    11154      10726 5.204e-10 ***
```

Recommended Model Selection: Model 4: Log Transformation...



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



Name <chr>	Scaling <chr>	Sampling <chr>	Accuracy <dbl>	Precision <dbl>	Recall <dbl>	Specificity <dbl>	F1_Score <dbl>	FPR <dbl>	TP <dbl>	TN <dbl>	FP <dbl>	FN <dbl>	Prediction_Threshold <dbl>	AUC <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