Predicting Default Status on Credit Cards

Anthony Iarussi

16 December, 2019

1. Introduction

1.1 Problem Statement

This data set allows exploration of how different variables impact defaulting on credit card payments. We will look at individual interactions of these variables, as well as using machine learning to see which variables are most useful for predicting an individual defaulting on their credit card payments. This is an important issue today, that companies have turned to data science to help solve. From this data set we will model high risk user characteristics and credit card payment history. As this model includes longitudinal data, it can be used as a tool to assess risk in real time. Unlike credit scores the model will also take into account important personal characteristics such as age, gender, marital status and education level. The problem of assessing credit risk is decades old. Due to the influx of credit card users in many countries, it has become necessary to use all means to assess risk in individuals.

1.2 Background

In recent years, there has been the arrival of data, in all industries. Data accumulated in recent years has been used to evaluate possible decisions and reinforce decisions being made. Industries that deal with risk have benefited greatly from various data mining methods. For companies that deal with credit cards, assessing risk can make or break the company. In a paper published by Shenghui Yang and Haomin Zhang, they compare several data mining methods to predict default rate in the same Twain data set. The three methods used are logistic regression, neural networks, and SVM pattern recognition (Yang). Another paper published by I-Cheng Yeh and Che-hui Lien compares six data mining methods to predict default rate in the same Twain data set. The

six methods used are K-nearest neighbor, logistic regression, neural networks, Naive Bayesian, discriminant analysis, and classification trees. In determining the conclusion I-Cheng Yeh and Che-hui Lien used simple linear regression to produce a line of best fit for the predicted v.s. the actual probability of default rate for each method in order to determine the most accurate probability (Yeh).

2. Data Acquisition and Preprocessing

2.1 Data Sources

The main data source for this project will be 'The Default of Credit Card Clients Dataset", which can be found on Kaggle. The original dataset comes from the University of California's School of Information and Computer Science, from the online machine learning repository. The dataset has information on 30,000 individuals from Taiwan, over the 6 month time period of April 2005 to September 2005. The data has 25 variables, including attributes like gender, education, marital status, and age. The data also contains variables that follow the billing and statuses for each credit card holder monthly. Included among these variables is a response variable, 'default_pay', telling whether or not the individual had defaulted on the credit card. Lastly, the education variable is the only column that contains unknown responses. The unknowns in the education variable make up approximately 1.56% of the responses.

2.2 Data Cleaning

The first step in data cleaning was to change the names of various columns to better reflect our data. The original data has numbers reflecting months in the various columns for bill amount and payment amount, I inserted months to make the data easier to understand. Next, I used the describe() function to get a basic feel for the data I am working with. Some initial observations after describing the data by column is that education and marriage have a minimum value of zero, which does not have an assigned value for both education and marriage. The education column originally held 6 values which represents levels of education. The data set considers values of 5 and 6 'unknown', and values of 4 'other', for the education column. So, I grouped

together the following values in the education column to make a collective unknown group: 0, 4, 5, 6. I did a similar process for the marriage column, I grouped all 0 values into the preexisting 'other' group. For all categorical groups provided in the data set: sex, education, marriage, and default pay; a numerical value is in place to represent the groupings. For an example, in the sex column, a 1 represents 'male' and a 2 represents 'female'. I went through all categorical groupings and replaced the numerical value with the actual name of the group the value represented. Changing the values for the groupings allowed for a straightforward data exploration.

2.3 Feature Engineering

I opted to use all of the features provided by the data set in my initial exploration. I disregarded the ID column in all exploration and machine learning. I created a new column to use in my initial data exploration called 'AGE_GROUPED'. The age grouped column blocks the numerical column 'AGE' into 5 blocks. The numerical age group did not have any individuals below 21 and older than 79 years of age. Therefore, I blocked the age group into five groups: [21-29], [30-39], [40-49], [50-51], [60+]. I used feature engineering to create a new data frame named 'default_trend' based on the six PAY_Month variables. The default trend data frame tells us whether or not an individual paid in a given month.

Table 1. 5 of the 30,000 rows in the 'default_trend' data frame.

	April	May	June	July	August	September	Months_Paid
0	1	1	1	1	0	0	4
1	0	1	1	1	0	1	4
2	1	1	1	1	1	1	6
3	1	1	1	1	1	1	6
4	1	1	1	1	1	1	6

Each row in the default trend data frame represents an individual. I used an if else statement nested in a for loop to append a binary output to each column labeled as a month. An output of '1' means that the individual paid at least the minimum due for the month of interest. The outputs of -2, -1, or 0 in the PAY_Month variable from the original data frame make up all 1's for the corresponding month in the default trend data frame. An output of '0' means the

individual did not pay the minimum due for the month of interest. The outputs of 1 through 8 in the PAY_Month variable from the original data frame make up all 0's for the corresponding month in the default trend data frame. The meaning of the values in the PAY_Month variable, -2 through 8, are explained in Table 10. The 'Months_Paid' variable is the sum of all columns across each row. Therefore, the months paid column holds a count for the number of months each individual paid at least the minimum amount due for the 6 month period the data was collected over. I then merged the columns 'April' and 'Months_Paid' from the default trend data frame to the original data frame. I merged 'April' as 'PaidInApril' and changed the binary output to 'Yes' and 'No' and I merged 'Months_Paid' as 'MonthsPaid. The two new variables will be valuable for basic data exploration and machine learning.

3. Exploratory Data Analysis

3.1 Categorical Data Analysis

The categorical data analysis will provide value counts for each group and break down the default rates by different groupings.

3.1.1 Default Pay

The response variable for the data set is the 'default_pay' column. Each individual in the data set either defaults or does not, majority do not.

Table 2. Value count and percentages for default pay group.

NODefault	23,364	77.88%
YESDefault	6,636	22.12%

3.1.2 Sex

Initially when exploring the gender variable I expected to find that males would have a higher default rate than females. I hypothesised this because males are typically a more risky demographic. I explored this hypothesis by creating a table (Table 3).

Table 3. Default rates and percentages by gender.

default_pay	NODefault	YESDefault	% defaulted
SEX			
Female	14349	3763	20.78%
Male	9015	2873	24.17%

I conducted a 2 proportions Z-test to see whether or not there is a statistically significant difference in the proportions of default rates of males and females. After performing the test I got a z of 6.921 and a p-value of 2.25e-12. The p-value of 2.25e-12 is far below .05. Therefore, I can reject the null hypothesis that the proportion of males that defaulted is equal to the proportion of females that defaulted. I can conclude that the proportion of males that defaulted is greater than the proportion of females that defaulted. Therefore, males are more likely to default.

3.1.3 Education

The initial hypothesis for the education variable was that the higher education, the lower the default rates. This is a reasonable hypothesis because a higher education implies a higher salary, and a higher salary means the ability to pay credit card payments. I explored this hypothesis by creating a table (Table 4).

Table 4. Default rates and percentages by education.

default_pay	NODefault	YESDefault	% defaulted
EDUCATION			
Graduate School	8549	2036	19.23%
High School	3680	1237	25.16%
University	10700	3330	23.73%
Unknown	435	33	7.05%

I conducted a chi-squared test for independence to see whether or not there is a statistically significant relationship between default and level of education. After performing the test I got a chi-squared statistic of 160.41 and a p-value of 1.495e-34. The p-value is 1.495e-34 which is far less than .05 Therefore, we can reject the hypothesis that there is no statistically significant

relationship between default and level of education. There is definitely some sort of relationship between whether or not an individual defaults and the level of education the individual has.

3.1.4 Marriage

I did not know what to expect when exploring the marriage variable. It seems reasonable that either single or married individuals could have higher default rates.

Table 5. Default rates and percentages by marital status.

default_pay	NODefault	YESDefault	% defaulted
MARRIAGE			
Married	10453	3206	23.47%
Other	288	89	23.61%
Single	12623	3341	20.93%

I conducted a 2 proportions Z-test to see whether or not there is a statistically significant difference in the proportions of default rates of individuals who are single and married. I did not include the 'other' group in this analysis. After performing the test I got a z of -5.259 and a p-value of 7.257e-08. The p-value of 7.257e-08 is far below .05. Therefore, I can reject the null hypothesis that the proportion of individuals that are single that did default is equal to the proportion of individuals that are married that did default. I can conclude that the proportion of individuals that are single that did default is less than the proportion of individuals that are married that did default. Therefore, individuals who are married are more likely to default.

3.1.5 Age Grouped

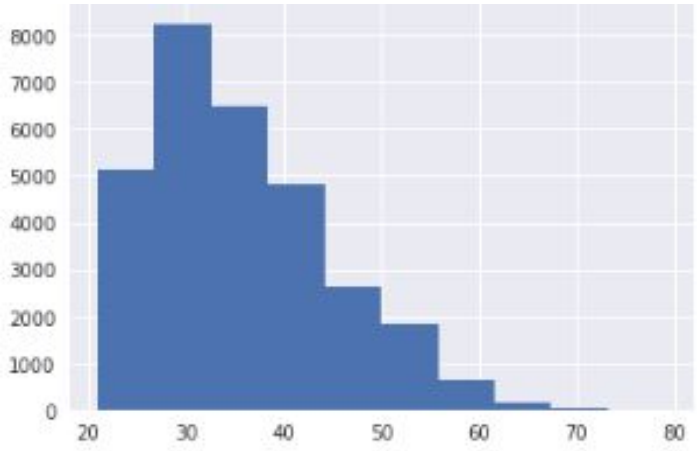
The initial hypothesis for the age grouped variable was that individuals who are younger would have higher default rates. This is a reasonable hypothesis because younger demographics tend to be more risky. I explored this hypothesis by creating a table (Table 6).

Table 6. Default rates and percentages by age grouped.

default_pay	NODefault	YESDefault	% defaulted
AGE_GROUPED			
[21-29]	7421	2197	22.84%
[30-39]	7841	2002	20.34%
[40-49]	4296	1298	23.2%
[50-59]	1449	481	24.92%
[60+]	243	96	28.32%

I conducted a chi-squared test for independence to see whether or not there is a statistically significant relationship between default and age grouped. After performing the test I got a chi-squared statistic of 40.87 and a p-value of 2.863e-8. The p-value is 2.863e-8 which is less than .05 Therefore, we can reject the hypothesis that there is no statistically significant relationship between default and age grouped. There is definitely some sort of relationship between whether or not an individual defaults and age group of the individual. Along with breaking down age into groups categorically. The data set provides us with the specific age of each cardholder, all of which are older than 20 and younger than 80.

Figure 1. The Distribution of ages in the data set.



3.1.6 Months Paid

The initial hypothesis for the months paid variable was that individuals who paid the minimum amount due for more months will have a lower default rate and vice versa. This is a reasonable hypothesis because the more months you pay the less likely you are to default.

Table 7. Default rates and percentages by number of months paid.

default_pay	NODefault	YESDefault	% defaulted
MonthsPaid			
0	398	943	70.32%
1	127	171	57.38%
2	406	545	57.31%
3	567	587	50.87%
4	1163	736	38.76%
5	3106	1320	29.82%
6	17597	2334	11.71%

I conducted a chi-squared test for independence to see whether or not there is a statistically significant relationship between default and months paid. After performing the test I got a chi-squared statistic of 4971.99 and a p-value of 0. The p-value is 0 which is far less than .05 Therefore, I can reject the hypothesis that there is no statistically significant relationship between default and the number of months paid. There is definitely some sort of relationship between whether or not an individual defaults and the number of months the individual paid.

3.1.7 Paid in April

The initial hypothesis for the paid in April variable was that there would be a higher default rate among individuals who did not pay in the first month of the data set. As we can see in Table 8, I hypothesized correctly, as 52% of the individuals who did not pay in the first month, ended up defaulting.

Table 8. Default rates and percentages by paid in April.

default_pay	NODefault	YESDefault	% defaulted
PaidInApril			
No	1468	1611	52.32%
Yes	21896	5025	18.67%

I conducted a 2 proportions Z-test to see whether or not there is a statistically significant difference in the proportions of default rates of individuals who did not pay the minimum amount due in April. After performing the test I got a z of -42.624 and a p-value of 0. The p-value of 0.0 is far below .05. Therefore, I can reject the null hypothesis that the proportion of individuals that paid in April that did default is equal to the proportion of individuals that did not pay in April and did default. I can conclude that the proportion of individuals that paid in April that did default is less than the proportion of individuals that did not pay in April but did default. Therefore individuals who did not pay April are more likely to default.

3.2 Numerical Data Analysis

The numerical data analysis will provide basic data exploration and insight into the numerical variables that make up the data set.

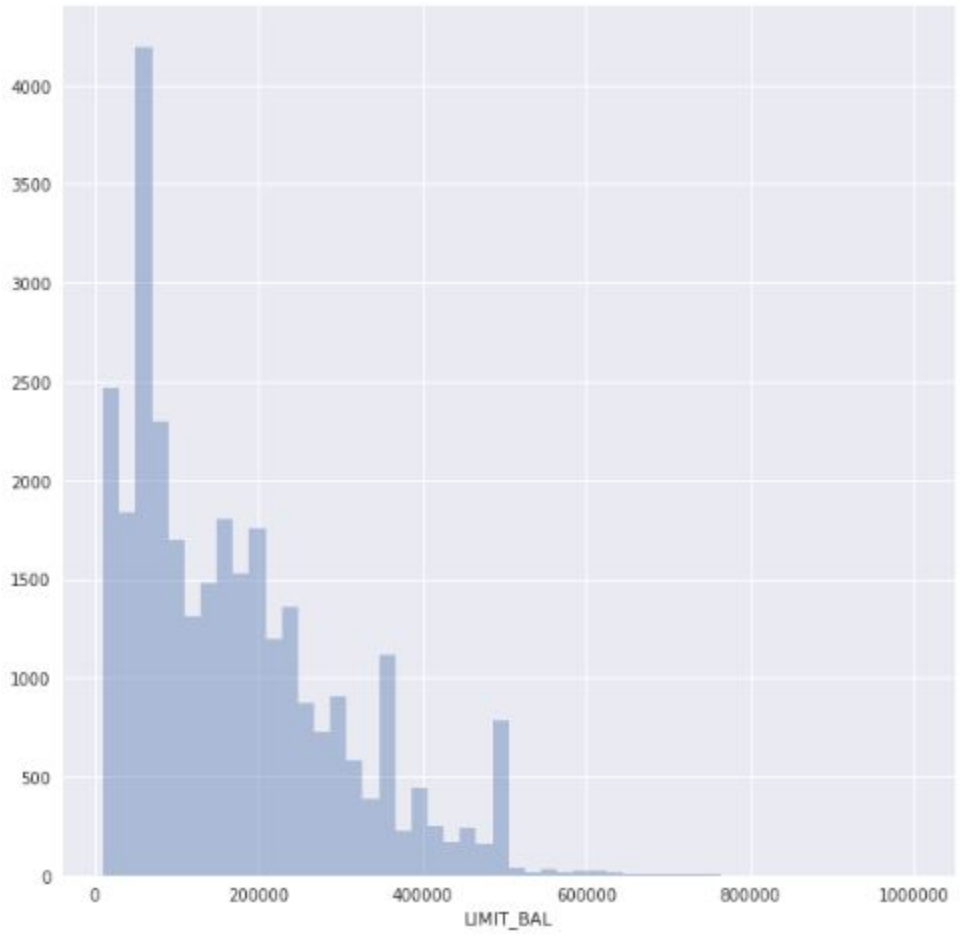
3.2.1 Limit Balance

Limit Balance is the spending limit granted to each individual in the data set. I visualized the distribution of limit balances granted to all individuals in the data set (Figure 2). The distribution of limit balance appears to be within a reasonable range. Although it appears a large fraction of clients are receiving a limit balance that exceeds 200,000 NT dollars. I looked at credit card holders that paid a balance of more than 275,000 NT dollars in September, and all numbers seem reasonable and relative for the outliers (Table 9).

Table 9. A look at 3 of the 13 individuals with a limit balance greater than 275,000 NT dollars.

	LIMIT_BAL	PAY_Sept	PAY_August	BILL_AMT_August	PAY_AMT_Sept	BILL_AMT_Sept	default_pay
2687	500000.0	-2	-1	367979.0	368199.0	71921.0	NODefault
3220	310000.0	-2	-2	298887.0	298887.0	0.0	NODefault
5687	480000.0	0	0	400000.0	302000.0	106660.0	NODefault
6773	470000.0	0	0	488642.0	300000.0	491428.0	YESDefault

Figure 2. The distribution of limit balance.



3.2.2 Repayment Status

The PAY_April variable is the individual's payment status in April, or at the end of the 1st month. PAY_May variable is the individual's payment status in May, and so on through to the PAY_September variable. Based on the summary it is clear that the values in the PAY_Month variables range from -2 to 8. Each of the 11 possible values hold meaning as explained in Table 10.

Table 10. Values in the six PAY_Month variables meaning (Kagan).

[-2]	Balance paid in full and no transactions this period, credit card inactive for a period. This is seen with individual #3220 in Table 9.

[-1]	Balance paid in full, but account has a positive balance at end of period due to recent transactions for which payment has not yet come due.
[0]	Customer paid the minimum due amount, but not the entire balance. The customer paid enough for their account to remain in good standing, but did revolve a balance.
[1-8]	Number of months payment has been delayed for the individual.

3.2.3 Bill and Pay Amount

This data set contains 6 bill amount variables, one for each month. The bill amount variable shows the current amount of a cardholders bill in the specified month, in NT dollars. This data set contains 6 payment amount variables, one for each month. The PAY_AMT_month variable is the amount of a previous payment in the month specified.

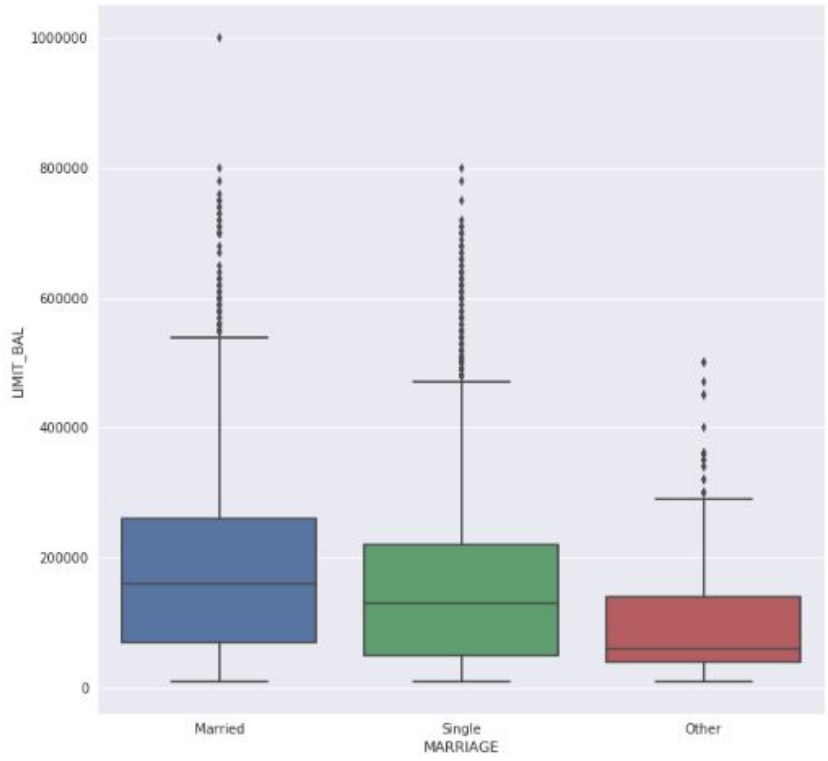
3.3 Exploratory Surprises

After exploring default rates by specific attributes and establishing the rates according to the categorical variables, it is beneficial to look at the mean limit balance according to the same categorical variables. This will gauge whether or not the company follows any risk assessment when dealing with cardholders.

3.3.1 Limit Balance by Marital Status

In Table 5, I found that married individuals have a default rate of 23.47%, as single individuals have a default rate of 20.92%. This trend in default rate is not reflected in how the company deals its limit balance (Figure 3). As married individuals on average are granted a higher limit balance than single individuals.

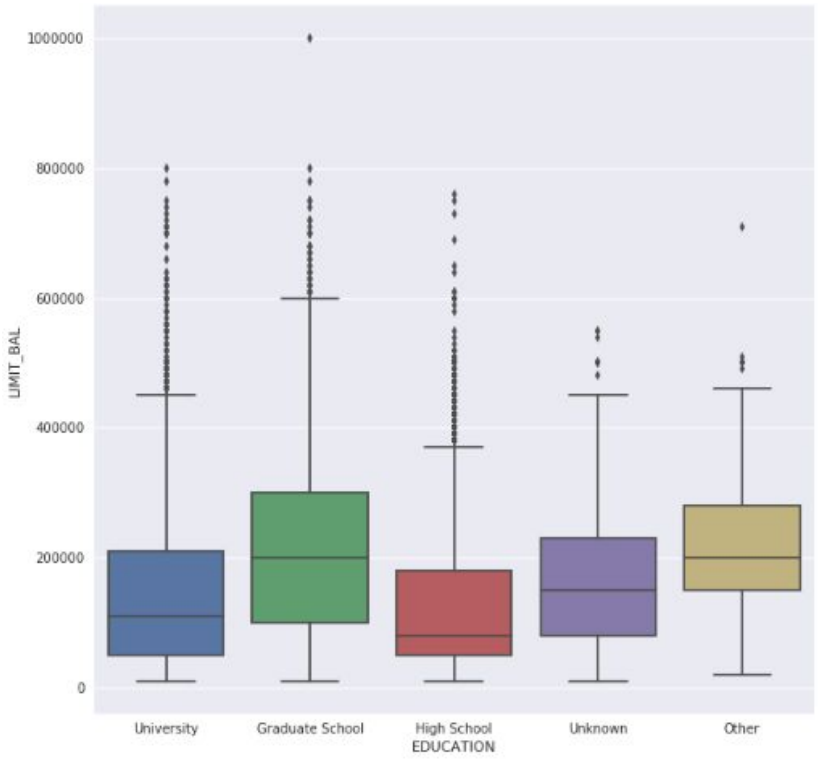
Figure 3. Box plots comparing mean limit balance by marital status.



3.3.2 Limit Balance by Education

This did not come as a surprise, as the order of mean limit balance follows the exact order of default rate percentage for the 3 main groups: high school, university, and graduate school. The lowest default rate and highest mean limit balance is the graduate school group, then university, and lastly high school with the lowest mean limit balance and highest default rate. The patterns following those of the default rates proves that the company dealing to the cardholders has at least a loose system in check to assess risk when dealing credit. It is clearly seen in education and gender. But not for marriage status, as married individuals have a higher default rate and a higher mean limit balance.

Figure 4. Box plots comparing mean limit balance by education.



4. Predictive Modeling

I used classification models to predict whether or not an individual defaulted on the credit card. The classification models will give me an idea of what features are important in predicting whether or not an individual defaulted. In the data exploration, I proved each variable has a statistically significant relationship to default rates. Therefore, the classification models will reveal which of these variables default rates depend most on.

4.1 Data Cleaning

I cleaned the original data set to prepare for the predictive modeling. I used the original data set because it kept the outputs for all variables as integers. Like the cleaning I did for the exploratory analysis, I created a collective unknown column for the education variable containing the values: 0, 4, 5, 6. The unknown group under the education variable amounted to 468 individuals. I combined the ‘unknown’ group with the ‘university’ group because the university group has the

largest population size under the education variable. I then assigned a sequential order that demonstrates value to all groups left within the education variable. I assigned the graduate school group to 3, the university group to 2, and the high school group to 1. Next, I did a similar process for the marriage variable. I assigned all 377 ‘unknown’ and ‘other’ individuals to the ‘single’ group because it has the largest population under the marriage variable. This action turned the marriage variable into a binary variable with only two groups, ‘married’ and ‘single’, making it applicable to machine learning. Lastly, I added the ‘Paid_April’ variable to the data frame which is the ‘April’ variable from the default trend data frame (Table 1).

4.2 Feature Selection

I have decided to only use data given in the first month to predict whether or not an individual defaulted. A company would not take on a client for six months to analyze risk. Therefore, only spending data from the month of April will be used. The categorical features I selected were the marriage, education, paid in April, and sex variables. The numerical features I selected were age, limit balance, pay amount in April, and bill amount in April variables.

4.3 Performance of Models

The five main metrics I evaluated each model with was log loss, accuracy, number of true positives, false positives, false negatives and true negatives. The lower the log loss the better, and the higher the accuracy the better. The number of true positives can be interpreted as the model predicts an individual is going to default and they actually did default when compared to the actual values. The higher the number of true positives the better. A dumb model could predict all individuals defaulted and still be 77.88% accurate according to Table 2. Therefore, the more a model can predict actual defaults, the better. The number of false positives can be interpreted as the model predicts that an individual defaults and they do not default. The lower the number of false positives the better. The number of false negatives can be interpreted as the model predicts that the individual does not default and they do default. The lower the number of false negatives the better. Between false negatives and false positives, false negatives are considered worse due to the model clearing at risk individuals. Lastly, the number of true negatives can be interpreted

as the model predicting that an individual does not default and they do not default. The higher the number of true negatives the better.

Table 11. Performance of classification models. Best performance labeled in red.

	Decision Tree	Random Forest	Logistic Regression	SVM
Log Loss	.501	.588	.490	.523
Accuracy	.791	.778	.789	.773
No. of True Positives	307	329	285	63
No. of False Positives	244	364	237	137
No. of False Negatives	1320	1298	1342	1564
No. of True Negatives	5629	5509	5636	5736

The best overall model here is the decision tree model, as it has the best accuracy at 79.1%. The logistic regression model was very close to the decision tree model overall. But, it is important to note that the accuracy of all models is below or barely above a dumb model that would be 77.88% accurate by predicting no individuals default. Compared to this metric, all models do not provide a sufficient prediction of default.

4.4 Feature Importance

Although the models did not provide a sufficient prediction of default. It is still useful to view the importance of all features used in the model. The decision tree model has the highest accuracy. Therefore, I will analyze the importance of all features used in the decision tree model.

Table 12. Feature importance of decision tree model.

	Feature	Importance
5	Paid_April	0.561160
3	LIMIT_BAL	0.233959
6	PAY_AMT6	0.095485
7	BILL_AMT6	0.081493
2	AGE	0.017290
4	MARRIAGE	0.010613
0	SEX	0.000000
1	EDUCATION	0.000000

5. Conclusions

In this study, I analyzed the relationship between variables describing an individual's attributes and spending, and default rates. I was able to prove that all attributes describing an individual including education, sex, age group, and marrital status have a statistical relationship to default rate. But, as I proved in the feature importances of the decision tree model, these attributes are not significant in predicting default. The feature I engineered 'Paid_April' is the most important in the decision tree model. This proves that whether or not an individual paid the minimum amount due in one month is very significant in predicting default. But, the models I produced using only spending information from April were not exceptional. I can conclude that more than one month of data is needed for an improved model using the following classification models: decision trees, random forest, logistic regression, and SVM.

5.1 Future Directions

Although the results I produced explain the importance of key variables to a certain degree, they do not provide exceptional results when using decision trees, random forest, logistic regression, and SVM to predict default. Predicting default to an exceptional accuracy may not be possible with only one month of information. Using multiple months and possibly other classification methods may be necessary to produce a model that can predict default with exceptional accuracy.

References

Kagan, Julia. “What Is a Minimum Monthly Payment?” *Investopedia*, Investopedia, 26 Feb. 2018, www.investopedia.com/terms/m/minimum-monthly-payment.asp.

Yang, Shenghui, and Haomin Zhang. “Comparison of Several Data Mining Methods in Credit Card Default Prediction.” Scientific Research, Scientific Research Publishing Inc., 2018, www.scirp.org/html/1-8701465_87507.htm.

Yeh, I-Cheng, and Che-hui Lien. “The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients.” Sciencedirect.com, Elsevier, 2009, bradzzz.gitbooks.io/ga-seattle-dsi/content/dsi/dsi_05_classification_databases/2.1-lesson/assets/datasets/DefaultCreditCardClients_yeh_2009.pdf.