Line Fitting Project

S. Roiz
e-mail: samuel.roiz.123@my.csun.edu

Aug. 4th, 2020

Contents

Data: Data was obtained from people's accounts with random numbers.

`http://www.wisc.edu/writing/Handbook/ScienceReport.html`

You are highly encouraged to take a look at this web-site!

1 Introduction:

1.1 What is curve fitting

What is curve fitting: Curve fitting is one of the most impressive and most broadly utilized examination devices in programs such as Matlab, Origin, Excel, etc. It is used every day in the working world for banks, scientists, etc. Curve fitting looks at the connection between at least one indicator and a reaction variable, with the objective of characterizing a "best fit" model of the relationship. By looking at this model, we can predict and see what variables are better nor worse. For example, let's say I work for a bank. I have 10 customers who took loans and have credit cards. Only 4 of them paid the loan. The other 6 did not. The bank wants to know how can we avoid the people who do not pay the loan. We would use variables of the customers by looking at their accounts. To see like the number of credit cards they have, the number of times they been late to pay the credit card after 30 days. These variables will be used to create a curve-fitting model. The model will tell us the patterns of who to trust by using this model.

1.2 My project scenario:

My project scenario: A bank calls me and asks me to make a model. I have to make a model that they can use in order to trust someone when someone wants a loan. The bank gives me data on accounts. The data can be very messy, so it needs to be cleaned up. What to do to clean it up is to make many variables. The main variables that I used are the outcome, Age of Account, Credit Card Balance to Loan, auto balance ratio, number accounts, number accounts ever 30 days past due, number accounts ever 60 days past due, number accounts ever 90 days past due, number inquiries last 6 months, number inquiries last 12 months, and number mortgage. The outcome is the first step. The outcome is the number of people who paid and who did not pay. You will see below what a table of outcome looks like. However, I had to find data to play the scenario. I used the credit card model data examples to look at then I produced a code to generate random numbers.

2 Data:

2.1 Outcome

Here are two tables. The first table is the data I got sent from the bank. I only displayed 40 outcomes, and I was received with 7,500 outcomes. Since it was so messy, I had to use the frequency table to clean it up. Using the frequency table is very useful and it shows how much easier it is to read data. The two have a huge difference. Imagine 7,500 lines in a table. By applying the frequency code, it only needs two lines for the table.

2.1.1 Outcome and Age of Accounts

Here is a short summary table of when you put the two variables together. I applied it with the frequency code to count the ones that have the same variable output which is group count. For example, the outcome has a zero. The amount of age of accounts is 3 months old. There is only one that has an outcome of zero and has a 3-month-old account.

Table 1: Outcome

1
0
1
0
1
0
1
0
1
1
1
0
1
0
1
0
1
0
0
0
1
1
1
0
0
1
0
1
0
1
0
1
0
1
0
1
0
1
0
1

Table 2: Frequency Outcome

outcome	amount
0	2500
1	5000

Table 3: Outcome and Age of Accounts

outcome	amount	groupcount
0	1	3
0	3	1
0	4	1
0	5	3
0	6	3
1	0	1
1	1	4
1	2	4
1	3	3
1	4	2

2.1.2 Sample Weight

Since my data is so big, I have to reduce it. However, reducing the data can affect the numbers. To avoid messing up the data, you need to use Sample weight. The sample weight is a way to reduce the data by keeping the ratio together. For example, my outcomes are 5,000 ones and 2,500 zeros. There are more 1's than 0's. So if I reduce the data to 270 numbers total. I would have 85 ones and 85 zeros with a sample weight of 7. The sample weight only applies to the ones. When I get a table of outcomes of ones, I will multiply the variables for ones by 7's. So it goes back to its real data.

Table 4: Without Sample Weight

outcome	amount	groupcount
1	0	1
1	1	4
1	2	4
1	3	3
1	4	2

Table 5: With Sample Weight

outcome	amount	groupcount
1	0	7
1	1	28
1	2	28
1	3	21
1	4	14

2.2 Full Table

We have over twenty-eight variables. Here is a summary of the full table that we gathered from the bank data.

Table 6: Full Table

A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	O
1	72	-102	35	35	63	45	26	2	0	-101	255	-101			
1	100	-102	-101	-101	80	-101	-101	4	4	29	296	110			
0	67	-102	18	53	230	58	58	8	0	142	230	142			
1	48	1	-101	-101	15	-101	-101	10	-101	30	115	62			
1	58	7	-101	-101	38	-101	-101	7	2	7	122	65			
1	85	8	-101	-101	84	-101	-101	6	4	36	176	122			
0	84	99	-101	-101	5	-101	-101	5	1	20	173	145			
1	75	-104	59	69	122	-101	-101	15	14	-101	150	-101			
0	100	-104	-101	-101	98	-101	-101	44	-101	-101	151	-101			
1	212	-104	-101	-101	136	-101	-101	76	-101	-101	423	-101			

A = outcome B = avgAgeOfAccounts C = highestBalToLimitRatio D = timeSinceLast30DaysLate

E = timesinceLast60DaysLate F = timeSinceLastChargeOff G = mosSncMostRcntDtOPnd

H = TimeSinceLastInquiry I = timeSinceOldestCreditCard J = timeSinceOldestMortgage

K = autoBalanceToLoanRatio L = creditCardBalanceToLoanRatio M = mortgageBalance

N = numberAutoLoans P = numberCreditCardsOpenededLast2Yrs Q = numCreditCards90PctUtilized

Table 7: Full Table

R	S	T	U	V	W	X	Y	Z	ii	jj	dd	pp	zz
2	0	-102	5	3	1	1	0	7	26	13	13	13	7
2	0	-102	0	0	3	1	4	17	32	0	0	0	7
1	0	-102	1	0	1	1	1	1	14	5	2	1	1
6	2	0	0	0	0	0	2	9	19	0	0	0	7
7	0	0	0	0	3	1	4	10	25	0	0	0	7
4	0	0	0	0	3	1	4	9	23	0	0	0	7
1	1	1	0	0	2	2	3	12	17	0	0	0	1
5	0	-102	0	0	0	0	0	3	14	2	1	0	7
2	0	-102	0	0	0	0	0	3	9	0	0	0	1
1	0	0	0	0	0	0	0	2	3	0	0	0	7

3 Graphs

Since I have gathered all of my data and applied sample weight and the frequency distribution. I need to graph them to compare each other in order to see what variable is the best fit.

3.1 Age of Accounts

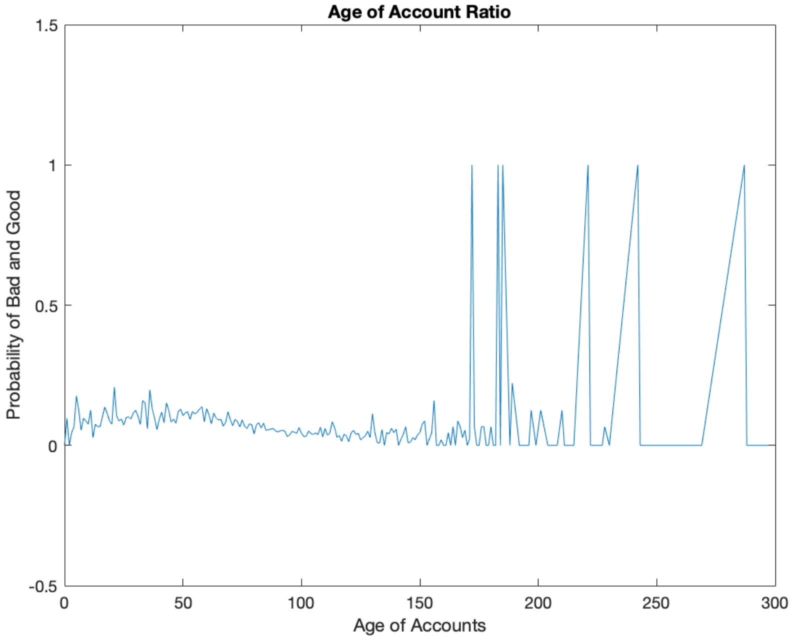

Figure 1: Noisy Age of Account in month

The graph is a little bit messy as you can see above. It has many spikes after 170 months of the account. The spikes on the graph are known for noise. I need to trim down the max range of age of account so it can be more accurate and clear.

After reducing the spikes, you can see how clear it is. What this graph tells us is that the month of account at fifty months has a higher probability of customers not paying back. However, it is not enough information if we need to trust someone based on their age of accounts. The graphs on the next pages will help determine who we should trust.

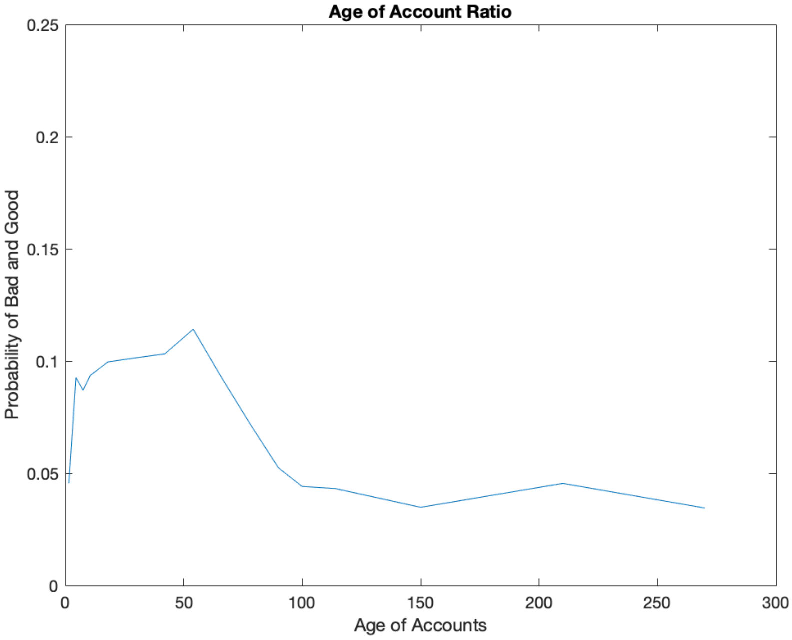

Figure 2: Age of Account in month

3.1.1 Summary of Age of Accounts

The summary of the Age of Accounts, the more months you had for the account. The less risk you give to the bank because it shows that the account has not been shut down and paying its dues correctly. The accounts that are young have a higher risk because they do not have much to lose. The bank has more to lose which is why the probabilities of risk are much higher.

3.2 Credit Card Balance to Loan

A credit card consolidation loan allows you to pay numerous credit cards and reduce credit card debt into a loan with a fixed rate and term. It can likewise assist you with setting aside cash by decreasing your loan cost or making it simpler to take care of your debt quicker.

In figure three, it shows progression after credit balance is having some spikes towards one hundred meaning that if the account has at least fifty credit card balance to loan, it will have a higher probability of the person not paying by at least ten percent. We need to reduce the noise again to have a closer look.

In figure four, when the noise is suppressed, it shows more accuracy and reliable data source on whether whose account has risk.

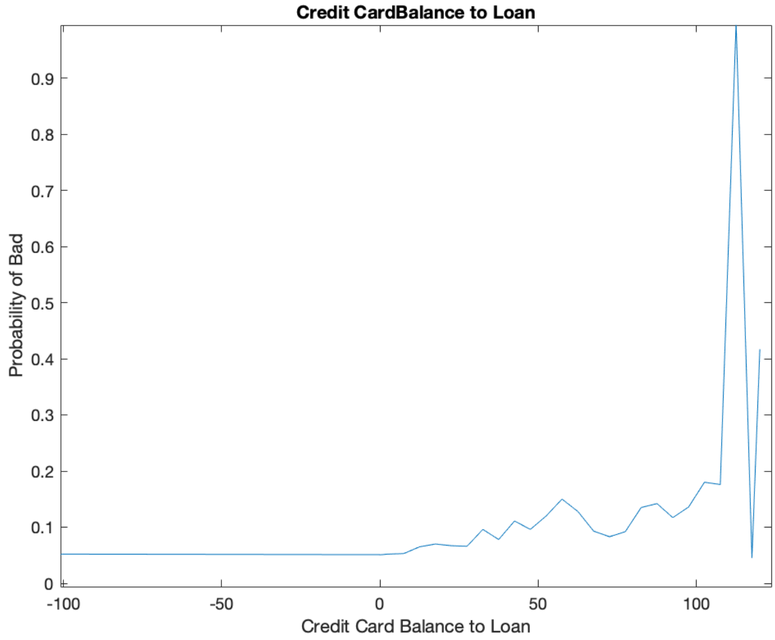

Figure 3: Noisy Credit Card Balance to Loan

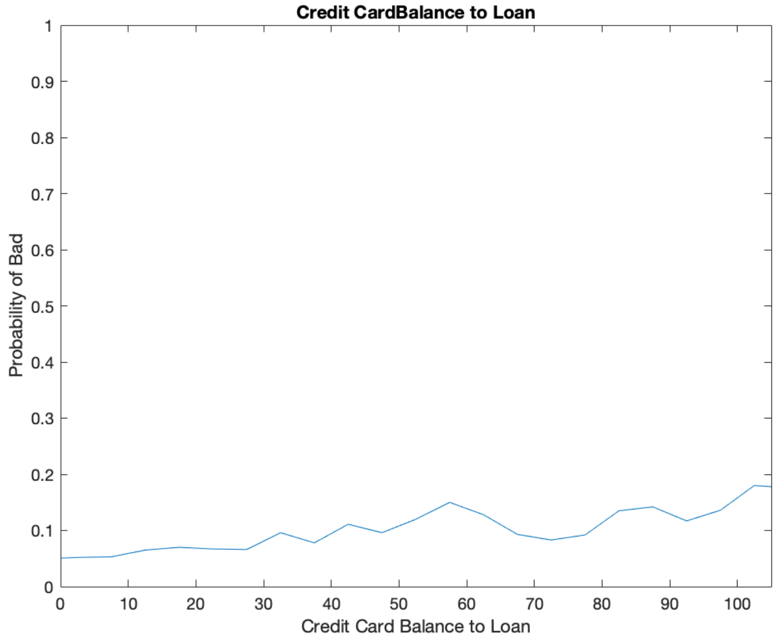

Figure 4: Credit Card Balance to Loan

3.2.1 Summary of Credit Card Balance to Loan

The lower the number is for credit card balance to the loan has lower risk even though they start at five percent risk. If it has a credit card balance to loan, it will increase risk slowly as the higher the number goes up to at least ten percent after thirty-two. Once it passes a hundred, it will be at least fourteen percent risk.

3.3 Auto Balance Ratio

This number is significant because it tells credit scoring organizations the amount of your accessible credit you are utilizing. Specialists propose utilizing close to 30 percent of your cutoff points, and ”less is better”. As you can see from the graph below, the accounts with a lower ratio have a higher probability of bad because the graph has negatives. It is not displayed because of its account has a negative for Auto Balance ratio, it means it owns no credit and has maximum credit. The accounts that are not negative are a higher risk.

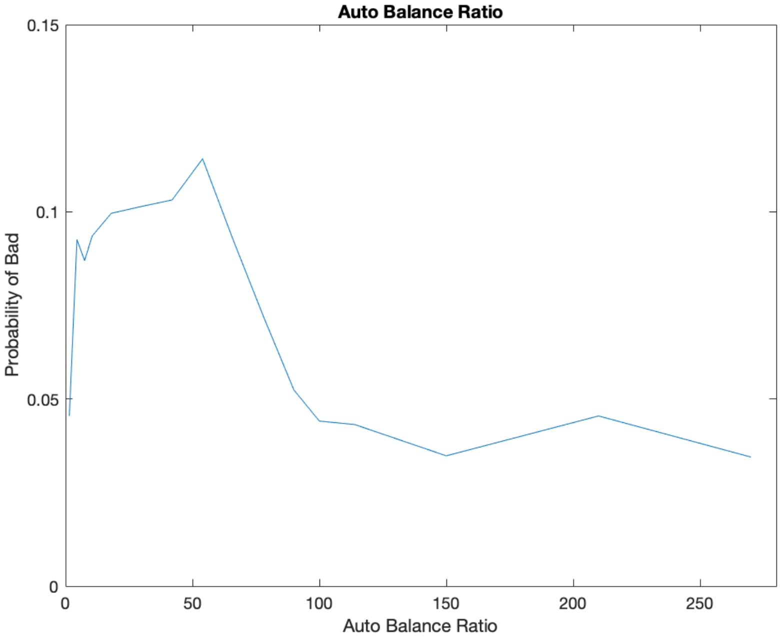

Figure 5: Auto Balance Ratio

3.3.1 Summary of Auto Balance Ratio

The summary of the Auto Balance Ratio, if the account has a positive number for the auto-balance ratio. It is going to have risks. If it is negative, it has no risk. The accounts that have an auto-balance ratio between one to fifty have are a huge risk of ten to twelve percent. However, the risk is not too high comparing to other variables. Auto Balance ratio variable is not too reliable unless if the account lands a ratio of one to fifty.

3.4 Number Accounts

Number of Accounts are just the amount of accounts does the client have in banks. We will see if the number of accounts will play a factor in the probability of the client will buy the loan back. The first graph (figure 6) is very noisy towards the end. It needs to suppress the noise. The second graph is much more clear. The number of accounts between forty and forty-five has a higher risk than others. Even though they are all at least five percent probability of bad, it still a huge difference as more accounts they have that it slowly increases by three percent for every ten accounts.

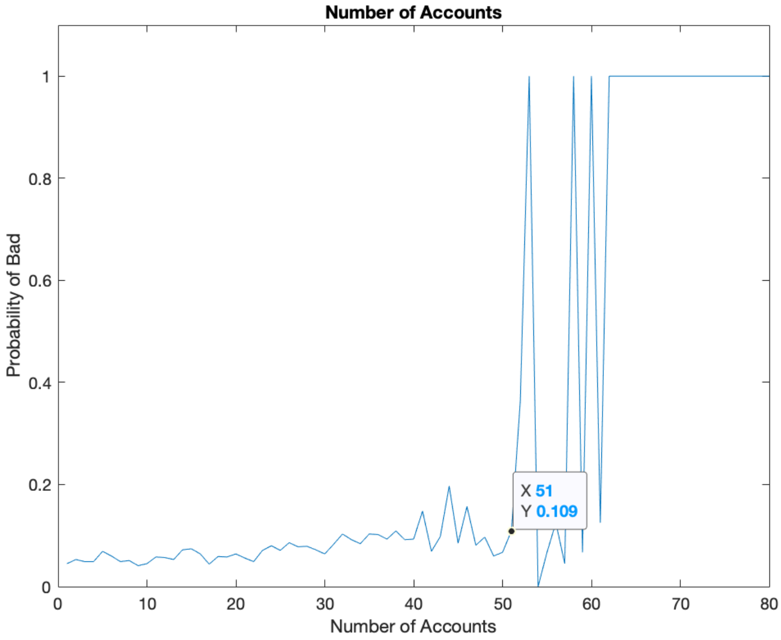

Figure 6: Noisy Number Accounts

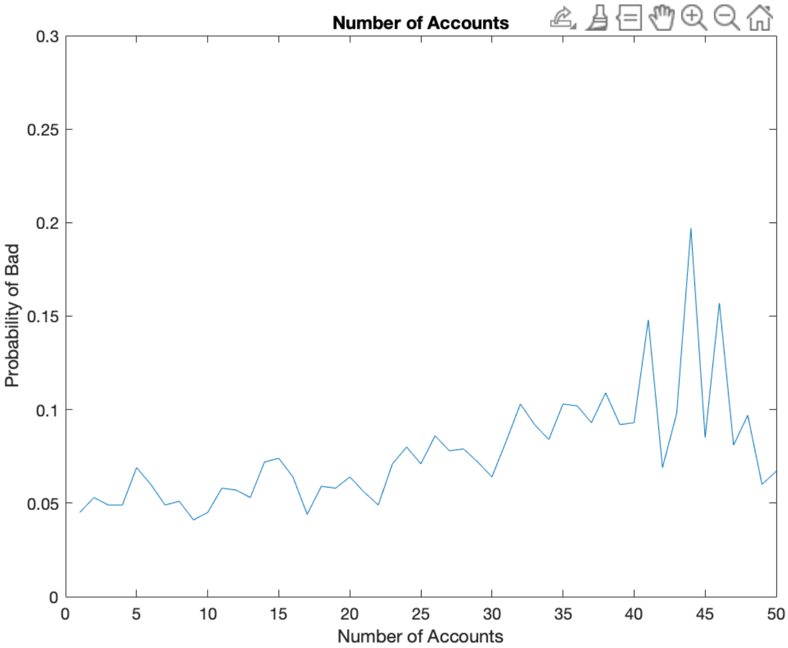

Figure 7: Number Accounts

3.4.1 Summary of Number of Accounts

The summary of the Number of Accounts, the more accounts you have, the risk increases slowly. It increases by the more accounts because if the client has more responsibility and we do not know if the accounts are stable. The fewer accounts, the bank can risk more because the client has no other account that can affect them financially. It went from one account to four accounts with a risk of five percent. Then at forty account to forty-five accounts is up to twenty percent risk. Forty to forty-eight are the high risk, but after fifty it is a very high risk due to not enough account data.

3.5 Number of Accounts Ever 30 Days Past Due

The Number of Accounts Ever 30 days past due shows the number of times the accounts have been late after 30 days of payment. Accounts that have zero times of been late after thirty days are only four percent of not paying back. As the more times, the account has been late, it increases its probability of not paying back. It went from never been late with a probability of four percent to nine times late with a probability of thirty-six percent. It is a huge dramatic difference.

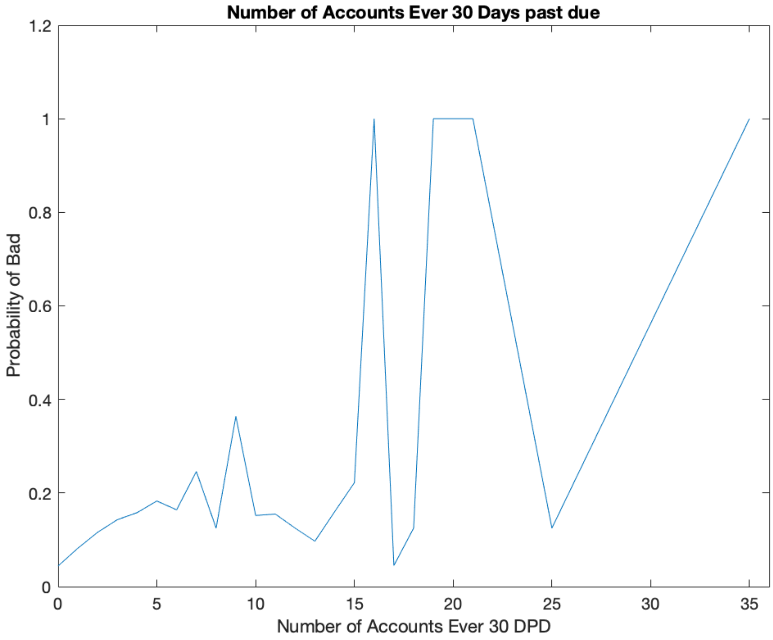

Figure 8: Number of Acc. 30 Days Past Due

3.6 Number of Accounts Ever 60 Days Past Due

Number of Accounts every sixty days past due is a very similar graph except the risk is significantly higher after each time its been late. The reason why it drops to zero risks is that no account had a number of times been late for thirteen times. However, we need to look at 90 days past due to see before we can say anything yet.

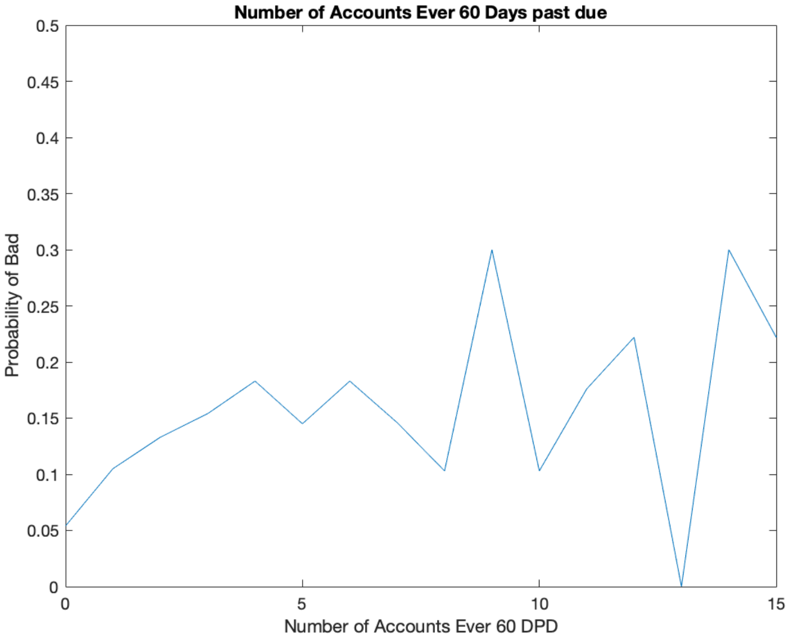

Figure 9: Number of Acc. 60 Days Past Due

3.7 Number of Accounts Ever 90 Days Past Due

The risk is similar to sixty days where if the account has nine accounts a past due day, it has at least twenty five percent of not paying back. For sixty and ninety days, the accounts that at least fourteen accounts late are a high probability of up to one hundred percent. It is clear that these variables play a big role in knowing if someone is stable.

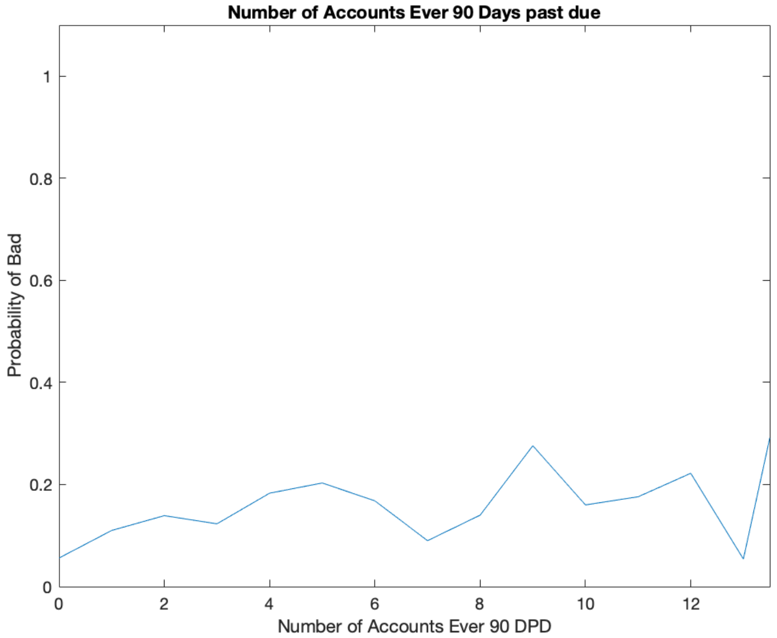

Figure 10: Number of Acc. 90 Days Past Due

3.7.1 Summary of Accounts Ever Days Past Due

The summary after these three variables of DPD, it is somewhat accurate. For 30 DPD, it tells us that we should not trust any client that is in the range of five or more. If want to take the risk, then we should not go beyond 15 because that is a whopping eighty percent difference. For sixty days, it is the same but less number of accounts. We should not trust anyone beyond zero because it starts at 5 percent. If the risk is needed, no more than three or we will have a risk at least seventeen percent of the risk.

3.8 Number Inquiries Last 6 months

Inquiries can have a more noteworthy effect on the off chance that you have scarcely any records or a short financial record. Huge quantities of requests likewise mean more serious hazards. Factually, individuals with six requests or more on their credit reports can be up to multiple times bound to bow out of all financial obligations than individuals without any requests on their reports. The graph has a very high risk. The more inquiries, the higher the probability will be as you can see on the graph. One number of inquiries is seven percent and it rises up drastically up to 30 percent with only 6 more inquiries total.

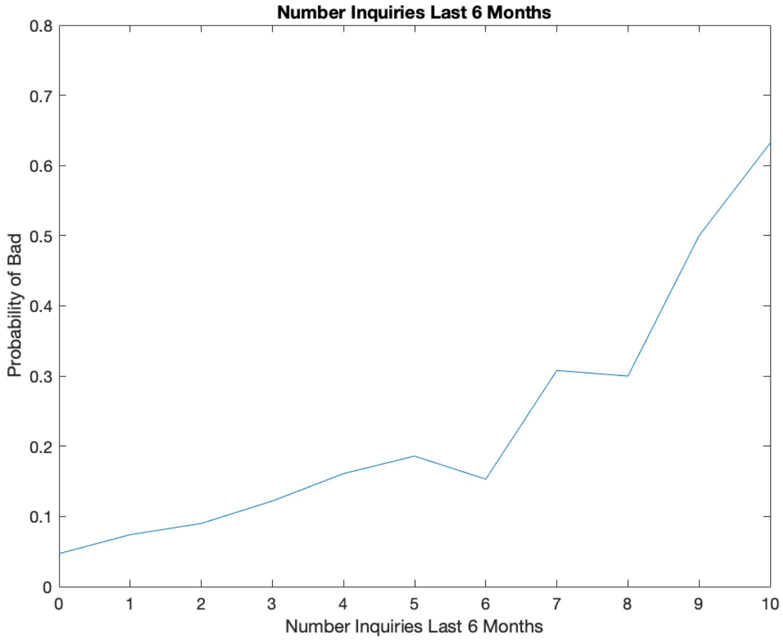

Figure 11: Number Inquiries Last 6 months

3.9 Number Inquiries Last 12 months

The graph is very similar to the last six months. The graph should have trimmed to twelve accounts so we can ignore the noisy section. However, after twelve times of been late. It is a very high risk at a whopping sixty percent.

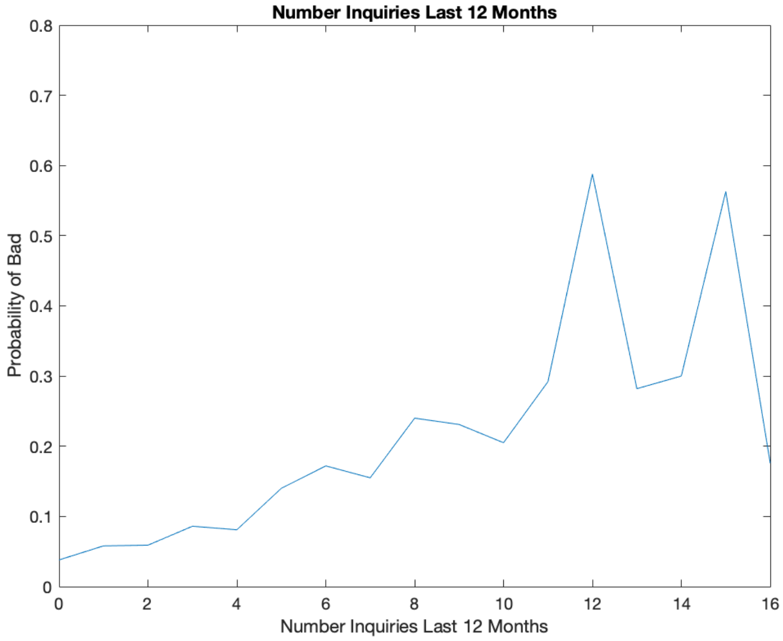

Figure 12: Number Inquiries Last 12 months

3.9.1 Summary of Inquiries

The summary after these two variables of Inquiries, it is very accurate and tells us what the client is a huge risk. It is very clear that we should be careful when it comes to a single number of inquiries where they start at least five percent risk. Then it increases drastically by three percent for six months. For twelve months, it increases by ten percent for six accounts until when it hits ten accounts. It increases by forty percent from ten accounts.

3.10 Number Mortgages

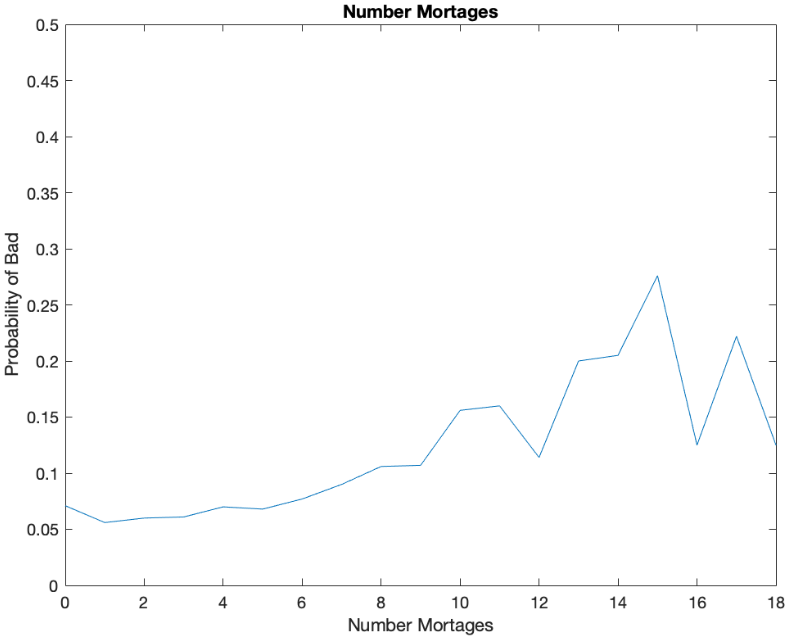

Figure 13: Number of Mortgages

3.10.1 Summary of Number of Mortgages

It all starts accounts at risk of five percent then if it has more number of mortgages. It increases slowly up to twenty-six percent of the risk. It is clear to say that the more number of mortgages, the more risk the account has to the bank. The reason why it has more risk is that they have to pay off that loan and if they take another loan, they have to pay more loans. The less number of mortgages, the less risk.

4 Results

In this section, I will be talking about what I thought about the project. The challenges. The fun. The mistakes.

4.1 Interesting

What is interesting about curve fitting is that you can do it in many ways. As long as you get the data and tables in. You can get it done by doing different ways, but it has to be the same graph. Curve Fitting is a very interesting method that it is so simple when yet its so difficult to do. Curve fitting is a long process for the way I did it. It is something you have to put time and it was very interesting to me because I am used to making graphs within minutes. What is also interesting is that you can apply curve fitting for anywhere or any situation. You can curve fit the number of people that eat eggs. You can do anything with it which makes it very fun and can be played around with.

4.2 Challenges

I had so many challenges. The first challenge for me was where to start. You cannot graph without data and data was the hardest part of the project because if you want to graph the data, you cannot just graph the data or it is not going to make sense when you look at it. It was very hard to put the data together because I had 7,500 accounts with a bunch of variables. So I had 7,500 accounts with 28 variables, which is 210,000 numbers in the data. There was no way for me to graph that from the start. Another challenge I had was applying the frequency table code to the data. It was supposed to be "tabulate()" but I could not figure out. So, I did the long route. I would say the biggest challenge for me was time. I was doing everything in the hard way when there way many ways to make it more efficient and easier. I even found out that there was a credit model program in the mat lab after I finished my data which I will use next time. The graphs were not hard. It just was the data that was a big challenge. Another challenge I had was latex. The latex format is very strange but once you get it, you get it. The biggest issue from latex was the figure order. It kept going to the bottom after my text. I had to put the "/float barrier" to prevent it from going anywhere.

4.3 Learned

I am proud of this project because I have learned so much from it. I may have learned how to fit curves, but I learned many important techniques that I can use outside of the mat lab. If someone asked me to do the fitting curve method on a different program, I would know what to do and start because I know the basics and order of making a curve. If you know how to use any program, but not know how to do a method. You are basically stuck. It is better to learn the method before you learn how to program because you need to understand it. That is what I love about the mat lab. I understand it.