Anna Eliason

Zakaria El Jbari

Tong Fu

You Zhou

Data Mining Group Project:

Bank Loan Default Prediction

University of South Florida

ISM 6136.020

2019-10-13

1. **Introduction**

One of the most important business sectors that drives the economy is the financial sector. It is practically impossible for an individual to go through life without resorting to the services provided by the financial industry. Financial intermediaries are used for all kinds of services including, but not limited to: credit cards, savings and checking accounts, investment accounts, vehicle loans, personal loans, mortgages, home equity loans, and certificates of deposit. The department that ultimately drives financial institutions’ profits is lending. The system to evaluate borrowers has changed dramatically throughout the years as technological advances enable financial institutions to develop more relevant metrics and models. We will be investigating some of the factors that lead borrowers to default on their home loans.

The dataset needed to start this project would be one that includes a significant amount of information about loans issued by financial institutions and has to at least include if the loan was paid in full or defaulted. The dataset used for this project was a generic bank dataset from 2018 that included information on lending. Each data point represents a loan and each column represents a variable that information was collected on. The variables included our minimum requirement because it had a loan amount column and total amount of principal paid column. The raw data was represented over 30 variables(columns) and 887,380 observations(rows).

1. **Key Business Questions**

Descriptive statistics on our dataset showed that the proportion of home loans that went into default was as high as 64.36%. This clearly poses a significant issue for the bank and its shareholders, one that would need to be immediately addressed by the C-suite executives. In an effort to tackle the matter at hand, we will attempt to develop a better underwriting process. The two important questions that this or any financial institution’s CEO would be interested in, in order to solve the delinquency rate would be:

1. Is it possible to build a data mining model good enough to fully automate the underwriting process while lowering the delinquency rate?
2. Is it possible to come up with improved benchmark metrics that loan officers can base loan decisions on?
3. **Relevance to contemporary problems in business**

The two questions aforementioned directly address several significant contemporary problems in business, specifically in the consumer lending industry. Defaulted loans have negative effects on the consumer, the bank, and the market as a whole. Consumers’ credit scores will drop if their loan defaults, which in turn affects their ability to take out loans or enter other credit contracts in the future. If a single financial institution holds too many delinquent loans, it can eventually lead to bank failure. The bank will then be forced to sell its existing mortgages to other lenders. There are many occurrences of this, including when Washington Mutual and IndyMac went bankrupt in 2007. The former was one of the biggest bank failures in US history (Kagan, 2019).

Default mortgages were one of the main reasons behind the 2007 banking crisis, the 2008 financial crisis, and the Great Recession. In the early 2000s, mortgage interest rates were low, regulations on lending were lenient, and house prices were increasing rapidly. People were able to get large loans with low monthly payments. However, in 2007, home prices abruptly stopped increasing, and under-qualified borrowers stopped making payments on their mortgages. This resulted in people defaulting on their loans in record numbers, and banks started losing money (Pritchard, 2019). When financial institutions repossessed more houses, real estate value crashed creating a domino effect that ultimately plunged the entire economy into a recession, and the rest is history. The 2008 financial crisis caused by the housing bubble had devastating effects on both the US and the global economy.

It becomes evident that it is in the bank and the borrower’s best interest that the underwriting processes for mortgage applications are strict and highly reliable. As previously mentioned, descriptive statistics on our dataset showed that the proportion of home loans that went into default was as high as 64.36%, which places this bank in the risk zone for bank failure. Developing an underwriting process that will improve its default rate is thus very relevant to contemporary problems in the fields of finance and economics.  

1. **Data exploration and modifications**

Since this is a generic dataset, we had to clean the data first. We started by exploring our data and understanding each variable and the information it presents. The dataset was very complete and did not have missing data for any variables. Some variables could very intuitively be eliminated from our analysis such as the loan ID, and the year and date issued. The dataset originally included many types of loans, such as credit card, car, and medical loans. Since we were only interested in investigating mortgages, we filtered the loan type variable to show only house loans and eliminated all other observations. This narrowed down our data to 3,707 observations. Once our key business questions were defined, we added a few variables in order to help us investigate potential solutions.

Initially, the data set included the variables “Loan Amount” and “Principal Recovered,” but no binary target variable. In order to answer the key business questions, we had to create a new variable with a binary target based on the “Loan Amount” and “Principal Recovered” variables. We named this variable “Default” and classified “1” for default loans and “0” for non-default loans. Each observation where principal recovered was lower than the initial loan amount was classified as “1,” and each observation where principal recovered was equal to the initial loan amount was therefore classified as “0.”

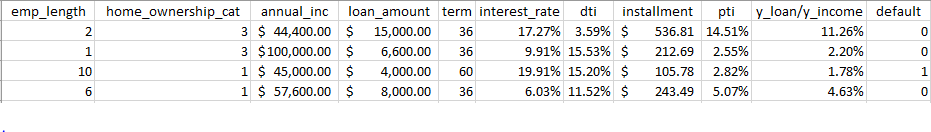
**DTI** is the **debt-to-income ratio** and is an important financial health indicator that many banks use in order to determine whether or not to approve a loan. It divides a person’s monthly debt obligations obtained from their credit report by gross monthly income, including everything from vehicle loans, house loans, and credit cards. The **DTI** ratio was already present in the original dataset. However, another important metric that is commonly used along with the **DTI** ratio is the **PTI** ratio.



**PTI** stands for **payment-to-income**, and is calculated by taking a borrower’s estimated mortgage payment divided by individual's gross monthly income. This ratio was not present in the data set, however, loan installment and yearly income were. First, yearly income was divided by 12 to obtain a monthly income. Then, PTI was obtained by dividing the observation’s loan installment by monthly income. We used these variables to add an additional column that calculated PTI for each observation, and utilized this ratio to investigate if we could come up with any suggested cutoff values that this bank’s loan officers can base loan decisions on.

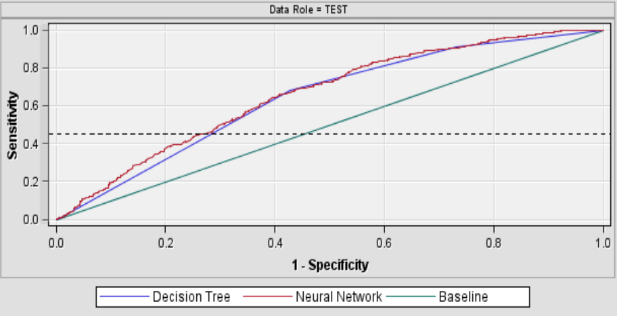
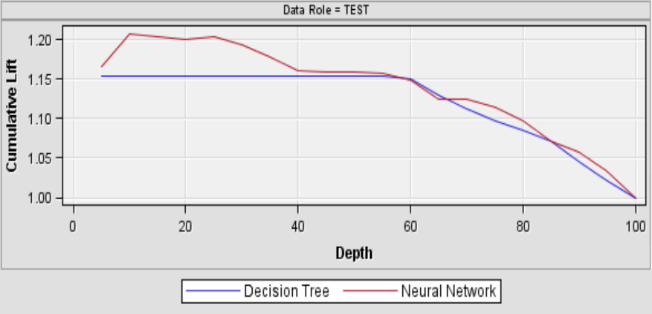


We also created a new variable named **y\_loan/y\_income.** The idea behind this variable is to potentially develop another metric that is useful in predicting default risk borrowers. This variable was constructed by converting the total loan amount to an annual loan principal by dividing the initial balance by the term in months and then multiplying it by 12. The reason for constructing this variable was to try and capture the borrower’s annual income compared directly to the principal amount for the same period. The **DTI** metric can be misleading sometimes as borrowers have the ability to consolidate their loans and lower their monthly obligations giving the false illusion that they can afford the debt. However, what this metric fails to capture is the total remaining debt that the borrower still needs to repay.

After adding these three variables, we decided to drop the variables in our dataset that would be irrelevant in answering our key business questions. These variables included home\_ownership and term\_category, which were redundant since they were represented by the variables home\_owenership\_cat and term. A variable called “Income Category” was also eliminated since the dataset already recorded total income. For the same reason, we dropped an “interest payment” variable as well since interest rate in percentage was also recorded. Leaving these variables would result in high multicollinearity because they will probably predict our dependent variable in the same way as the variables kept. We also dropped application type since all observations had the same input for that variable, namely “individual.” The image below provides a snapshot of sample data from our final dataset:

1. **Analysis**

***Question 1)***    *Is it possible to build a data mining model good enough to fully automate the underwriting process while lowering the delinquency rate?*

In order to answer this question, we will address what type of model best predicts default risk and whether the local optimum from our resulting models is good enough to be used in practice. We used SAS Enterprise Miner to build and compare different predictive models, including a decision tree and a neural network. The decision to compare these two models was made because they are both able to capture non-linear relationships between variables, and they both take into account interaction between variables should there still be any after the data modifications made. By altering the properties for the decision tree and the neural network in order to get the best possible models (more info about the alterations can be found under the “Trade-offs and Model Property Settings” section), two models of similar performance were developed. However, as the images below depict, the neural network had slightly higher lift and ROC values, making it the better model of the two.

After identifying the better model of the two, the question of whether or not the best model is good enough to recommend the bank to use can be investigated. This is done by looking at the lift curve, ROC curve, and misclassification rate.

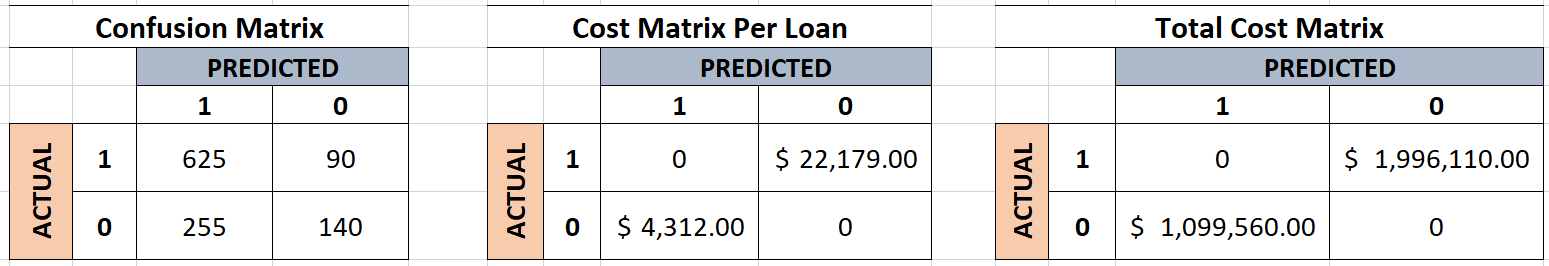
The goal of the model is to identify loan applications that have a high default risk so that the bank can avoid approving these loans. The cumulative lift chart shows the advantage of using the predictive model to choose what loans to deny. It does this by depicting how much more likely the bank is to identify at-risk loans using the model than it would be if it simply denied a random sample of applications. SAS E-miner “risk-scored” the validation dataset by attaching to each case an estimated probability that the loan would default, i.e. that the case’s default variable would equal 1. The cumulative lift chart above shows that, by using the neural network model, the bank should be able to capture 1.2 times as many high default-risk applications if they investigate the applications with risk scores in the top 20 percentile. If the bank instead investigated applications in the top 10th percentile, it would be able to identify 1.21 times as many high default-risk loans. This is not a great lift, but it provides a small increase in accuracy, which might result in significant savings for the bank. To further investigate the model’s performance, the ROC and misclassification values would also have to be taken into consideration.

The receiver operating characteristic curve (ROC), plots the True Positives on the Y-axis against  the False Positives on the X-axis. It depicts how good the model is at distinguishing between classes, i.e. how good the model is at predicting defaults as defaults and non-defaults as non-defaults. The goal is to maximize the AUC, the area under the curve. The neural network model for predicting bank defaults has an ROC value of 0.66, meaning that there is a 66% chance that the model will be able to distinguish between defaults and non-defaults. An ROC value of 0.5 means that the model is completely useless in distinguishing between positive and negative classes, meaning that our model’s ROC rate of 66% is fairly poor.

The neural network model has a misclassification rate of 0.32, which means that the model will predict the wrong outcome 32% of the time, or in other words, the model has an accuracy rate of 68%. In this particular scenario, the misclassification rate means that the model may predict that a loan applicant’s loan will default, when in fact it would not, or vice versa, 32% of the time. The problem with an inaccurate prediction is that it will cost the bank a lot of money. For example, if the bank chooses not to approve an applicant's loan request because the model predicted that it would default, the bank would lose money on potential interest earned if the model’s prediction was inaccurate. If the bank would approve an applicant’s loan which later defaulted, the bank would lose the defaulted loan amount plus the interest that would have been paid on that amount.

The best model for the loan default dataset had relatively low lift, ROC, and misclassification rates. The question is, are they too low for this model to be used to fully automate the bank’s underwriting process? Keeping in mind that the bank currently approves as many as 64.36% default loans, deeper investigation is needed in order to respond to this question. Perhaps the model is not great, but it is better than using no model at all? In order to determine whether or not these performance values were acceptable, a cost matrix can be developed to calculate average monetary losses the predictive model could result in, and compare it to average monetary losses currently suffered.

After identifying the average loss of **false negatives** and **false positive**s, the following cost matrices could be constructed:



The matrices reveal that the model results in a lower number of false negatives than false positives, but that false negatives are the most expensive mistakes. A false positive means that the model predicts that the loan will default, when in fact it will not. This means that the bank will lose money on lost interest payments, which would cost the bank on average $4,312 dollars per loan. This cost was computed by first calculating the amount of cumulative interest the bank lost on each loan that was falsely predicted and summing those values to get a total dollar value from interest that would otherwise have been gained. Then, we would divide the sum by the number of loans that fall into this category to get an average loss per loan value, which is essentially the opportunity cost.

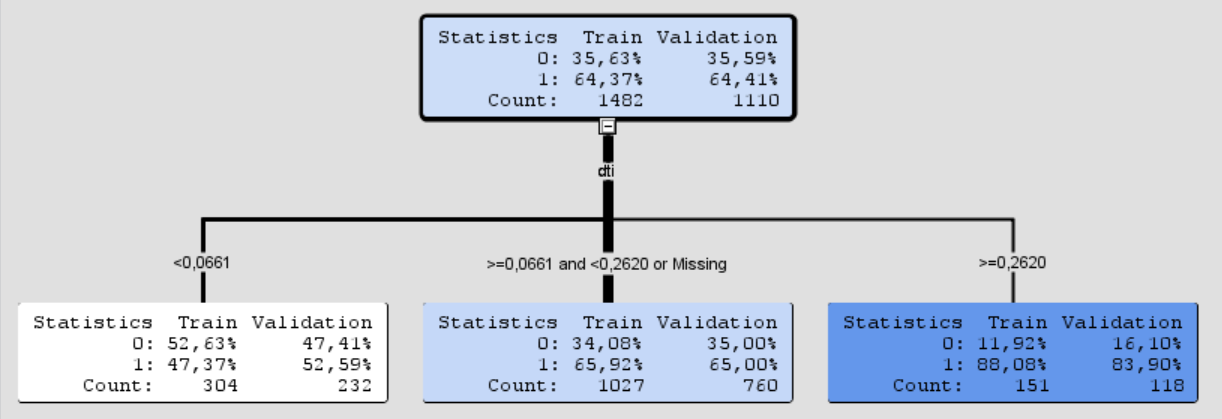
 A false negative means that the model predicts that the loan will not default, when in fact it will. This is a more costly mistake since it will cost the bank both the loaned principal and the interest payments on that loan, resulting in an average loss of $22,179 per loan. This value was computed by first summing the principal initial balances, and cumulative interest lost of loans that did indeed default, adding those two values together, and then dividing by the amount of loans to get the final average loss per loan.

Even if the model only classified 90 false negatives in the validation dataset, the average total cost of those misclassifications would be $1,996,110 dollars. In addition, the bank would also lose an average of $1,099,560 on the 255 false positive loans. Thus, using this model would result in a total loss of $3,095,670 if applied to our sample data.

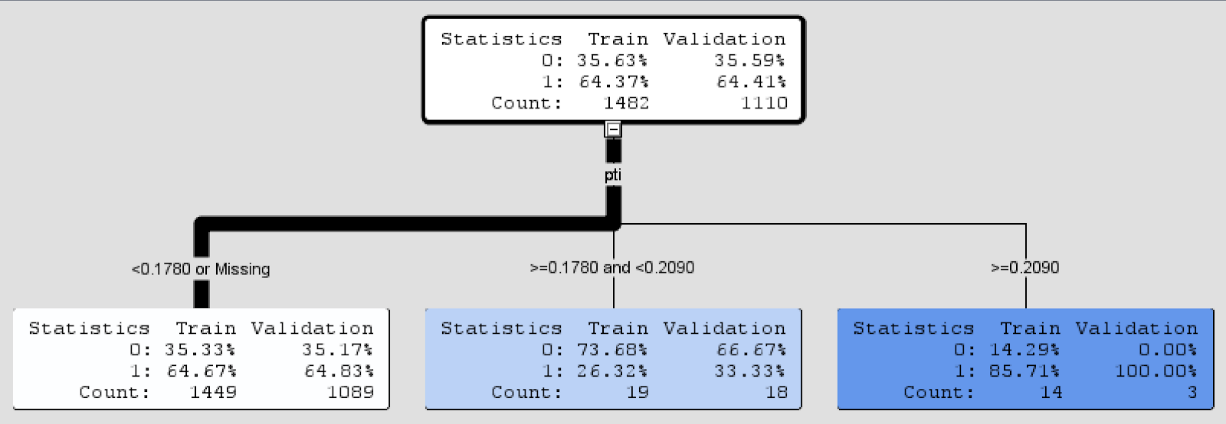
The total loss resulted from using the predictive model may seem very high. However, keeping in mind that the bank currently defaults 64.36% of their loans, not using the model cost the bank $52,920,086. Thus, the relatively low lift, ROC, and accuracy rates would, in fact, be acceptable in this case. Using the predictive model will save the bank $49,824,416. Therefore, it is possible to build a data mining model good enough to fully automate the underwriting process while lowering the delinquency rate, at least for this specific bank.

***Question 2)*** *Is it possible to come up with new metrics and/or improve current benchmark metrics that loan officers can base loan decisions on?*

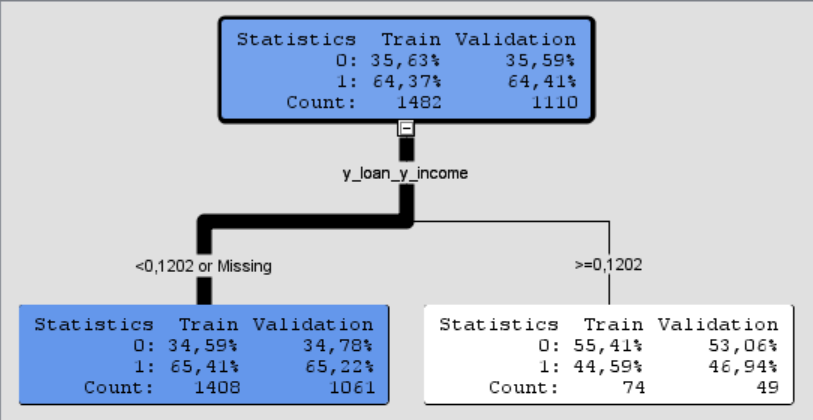
In order to come up with recommendations concerning metric benchmarks, we had to isolate each “metric” variable and observe how well it can split our data by the binary default category. The three metrics we chose to isolate and provide benchmarks on were **DTI**, **PTI**, and the variable we constructed named **y\_loan/y\_income**. We decided to use an interactive decision tree, which contains an algorithm that splits nodes based on the best value that isolates the two levels of our dependent variable. Rather than allowing the decision tree to grow as large as it wants and prune back automatically, we limited the number of branches to three and set the leaf node size to minimum ten. We wanted three branches in order to get a representation of possible low, medium, and high risk taking decisions. For example, the three branches made it possible for us to determine that approving loans to applicants with **DTI** values lower than 6.61% would be considered low risk, approving applicants with **DTI** values between 6.61% and 26.2% would be considered medium risk, and approving applicants with **DTI** values of 26.2 and above would be considered high risk. The image below describes our SAS enterprise output findings for **DTI**:



We applied the same methodology for the PTI to have cutoff values that would divide into the three aforementioned categories: low, medium, and high risk. For PTI, the cutoff values from the model were under 17.80% for low risk, between 17.80% and 20.89% for medium risk, and 20.90% or above for high risk category as shown below:



For **y\_loan/y\_income**, the model was only able to split into two leaf nodes because further splitting would violate the algorithm rules, probably leading to a node that has less observations than the minimum categorical size. Thus, this variable will be placed into the two categories only. The values were under 12.02% for low risk and 12.02% or above for high risk as shown on the SAS Miner image below:



These values will be utilized to provide recommended courses of actions that will be discussed in further detail in the Recommendations section.

1. **Trade-offs and model property settings**

We encountered several trade-offs during the model properties modification stage. We noticed a pattern when adjusting the leaf node categorical minimum property for example. When we set the minimum to a low number, the model outputs better splits with higher accuracy but nodes were very small, which could expose our model to the risk of overfitting. Inversely, when the minimum was set to a high number, the model outputs less accurate splits, but the resulting nodes contain a significant amount of observations in each. These results would probably be more accurate in showing how the general population of loans would be split.

For example, we tried setting the leaf node to minimum 100 because we wanted to ensure the model is practically useful. However, with the default limit of 5 minimum observations per node, we were actually able to get a better performing tree. The tree revealed that if we investigated applicants who had **DTI** scores below 6.61% and **PTI** scores above 17.90%, we were able to capture 83.33% non-defaults and 16.67% default in one leaf node, which is a relatively good split. However, this node only included 6 observations, so the model would not be practically useful since it only captures about 0.16% of the banks loan takers. After setting the leaf node to 100, the model revealed that a **DTI** cutoff higher than 26.2% would capture 16.10% non-defaults and 83.90% defaults in a node that included 118 observations. Thus, it gave a very similar split but it captured more of the bank’s loan takers. These tradeoffs gave us an overall worse performing decisions tree, but they made the model more useful in practice.

To optimize the performance of the neural network, we experimented using backpropagation to update each of the weights in the network so that they caused the actual output to be closer to the target output. We also experimented with different numbers of iterations, other training techniques, and different seeds. However, the neural network performed best when set to use the default settings, so we decided to leave the properties unaltered.

1. **Limitations**

***Causality Limitation:***

One of the limitations of our analysis is that our results do not take into consideration all the factors that may lead a borrower to default on their loan. Some factors that will influence the outcome, but are not captured in our analysis include: unexpected life events such as serious illness, sudden loss of employment, or economic recession period. Even excellent borrowers are sometimes pushed into default due to reasons that are beyond their control. It would be a good idea to eliminate such cases for the dataset to obtain more accurate and relevant results.

***Regulation & Law Limitation****:*

There are strict regulations put in place to prevent financial institutions to make credit decisions based on specific demographics. While it might be considered unethical to base decisions on demographics such as age and education, these demographics might have a high correlation with the rate of default from a purely financial and analytical point of view.

***Data Limitation:***

The dataset was obtained from a specific bank, which might provide specific solutions for said bank, but might not generalize accurately to all financial institutions. Moreover, the result recommendations will highly depend on the specific financial institution’s risk adversity.

1. **Results and Recommendations**

***Question 1)***

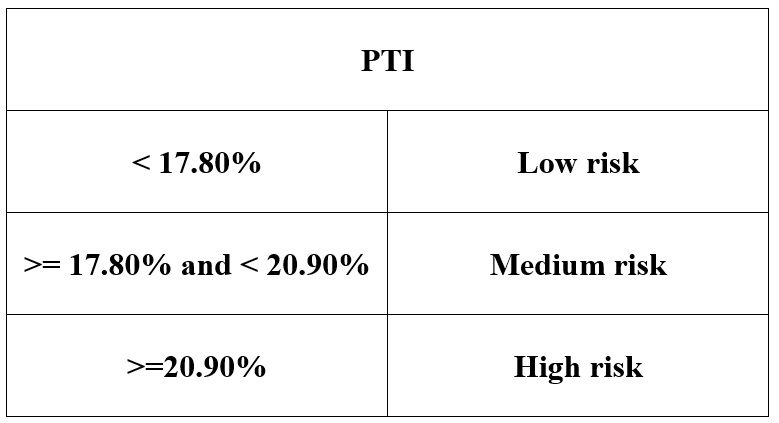
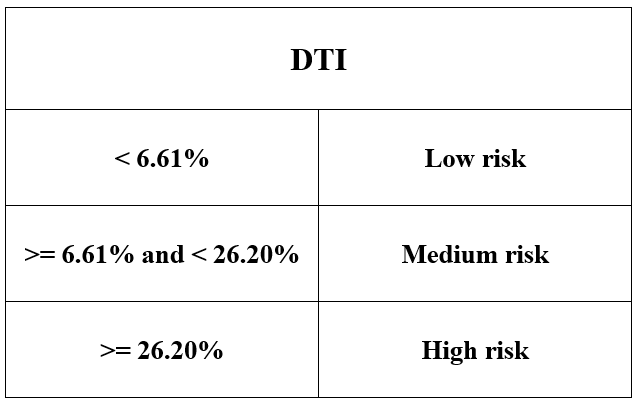
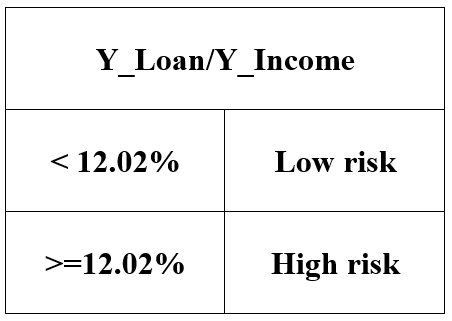
As stated in the analysis section, the best model we were able to come up with in order to automate the underwriting process had a relatively low lift, ROC, and accuracy rate. The models performance is thus relatively poor. Therefore, we would not recommend banks that do not have a significant delinquency rate issue to implement this model as a tool for automating the underwriting process. However, considering that this specific bank currently defaults 64.36% of their loans, the relatively poor predictive model would still save the bank $49,824,416.

That being said, we would recommend the CEO of this specific bank to integrate the model to automate the underwriting process, and use it as a first step in the decision. Getting classified as a default borrower would automatically result in rejection. If the borrower was classified as a non-default, the loan officer would then need to check the metrics requirement decision (detailed under Question 2) and accept only if the borrower satisfies the metrics as well.

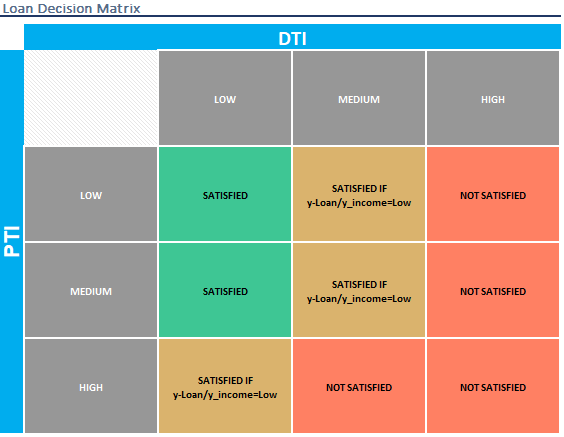
Even if using this neural network is better than not using any model at all, we still believe there is room for great improvement in the model. Therefore, we also recommend that the CEO gather more information about their loan applicants that could potentially improve the model. Potential ways to do this will be discussed in the Potential Follow-up Analysis section.

***Question 2)***

For this bank, we propose that the loan officers use the following cutoff values for **DTI**, **PTI**, and **y\_loan/y\_income** to benchmark and categorize each borrower by their risk level, if the borrower passed the predictive model’s test:

After categorizing the borrower in each of the metrics, we recommend loan officers use the following matrix to determine whether the metrics requirement has been satisfied:



A combination of all the three metrics and the neural net model is recommended for this bank to make a loan decision. Prospective borrowers would have to first get classified by the model as a non-default, and then satisfy the metric requirements to be approved.

Before any other bank uses this model, it would need to utilize the cost matrix to first calculate the potential savings they can earn by using their borrower data as input. After comparing their current losses to potential losses after model implementation, they would then make a decision on moving forward accordingly.

We recommend small financial institutions and institutions with high delinquency rates to use the same acceptance criteria we developed for our bank. We recommend this because these institutions are usually more risk averse and often cannot sustain significant losses without failure. Credit Unions and other financial institutions with stricter regulations would benefit from utilizing this acceptance criteria as well.

Larger financial institutions and banks that do not typically struggle with delinquency can approve borrowers that either satisfy the proposed metrics or are classified as non-default borrowers by the model, rather than both. The reason being that they can afford to take on more risk and possibly generate more benefits by doing so. Thus, we recommend accepting a certain amount of high risk loans, but implement contingencies and regulations to cap the proportion of high risk loans in the institution’s overall mortgage portfolio to a certain value. This value should be defined by each institution’s risk department.

1. **Additional data for potential follow-up analysis**

There are several possible follow-up phases that could be pursued by other researchers. One useful follow-up analysis would be to add a variable to the model that would eradicate one of the Limitations previously mentioned. A good example would be a binary variable named job\_loss defined as a 1 if borrower ever experienced job loss and 0 if they did not. The accuracy of our metric benchmarks could be improved if we eliminate borrowers who defaulted due to a sudden job loss. This would result in cleaning some of the noise in our data that is causing our false predictions, and improving the overall model accuracy.

Another very useful follow-up analysis would be conducting the same analysis, but with an unbiased random sample of borrowers from various institutions. The results from this model would more likely be more accurate. Moreover, the results would be more meaningful because it a random sample is representative of the true population of borrowers, so the model could directly be implemented at any institution.

**References**

Kagan, J. (2019, September 21). What is Bank Failure? *Investopedia*. Retrieved October 10, 2019, from <https://www.investopedia.com/terms/b/bank-failure.asp>

Pritchard, J. (2019, February 4). What Caused the Mortgage Crisis? *The Balance.* Retrieved October 10, 2019, from <https://www.thebalance.com/mortgage-crisis-overview-315684>.