

**Problem 2 Report**

Abdillah Aldeghaither

Abdullah Alhejailan

David Deans

Santosh Gurung

Huy Pham

# 8 December 2019

**Mission**

“Using descriptive analysis and data mining models, create a report for Stark Enterprises to evaluate high-risk borrowers in the data provided.”

Understanding the lending market is no small feat, especially when half of the market was dusted away until just recently, but taking the time to do this is an excellent investment that will allow Stark Enterprises to overcome the blip and operate with confidence. Knowing the differences between low-risk and high-risk borrowers and establishing a prediction engine on those premises will provide an edge in an already highly competitive industry, regardless of the timeline. There are many providers of this service, and almost all of these competitors are applying similar data analytics practices in their own firms. To stay involved with the industry, one must put resources into market research and understanding. Going further, investing in a prediction model will allow Stark Enterprises to avoid wasting resources on loaning to those who should not be receiving them, and increase the rate at which they can loan to low-risk consumers while still providing the due diligence of making sure they are worthy (maybe even worthy enough to lift Mjolnir).

**To that end, our formal objectives for Stark Enterprises are as follows:**

Use Business Analytics methods and data mining techniques to establish an understanding of the historical data in the lending market. We will interact with this information in an exploratory data analysis to create a set of descriptive statistics. This information will be applied to a data mining method to predict the impacts of the variables given on the risk of defaulting on a loan.

**What this objective will specifically accomplish for Stark Enterprises:**

· Establish confidence that the historical information available has been fully and correctly interpreted and cleaned.

· Responsible business analytics methods of cleaning the available data, providing the most accurate matrix of information possible. To this end we will be removing variables that do not support our goals.

· Application of this matrix to a predictive algorithm that supports Stark Enterprises’ main goal of gathering information on high-risk consumers to support the financial segment’s understanding of the lending market.

· Provide an interpretation of models and an analysis of what information should be added to the model in future iterations to create a more robust prediction method.

· Let the sun rise on a grateful Mortgage Market

**Descriptive Analysis**

**Table 1:** Descriptive Statistics of Numerical Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Max** | **Min** | **Str** |
| **Credit Score** | 751.10 | 759.00 | 834.00 | 601.00 | 44.55 |
| **Number of Units** | 1.00 | 1.00 | 4.00 | 1.00 | 0.24 |
| **DTI** | 34.06 | 35.00 | 50.00 | 1.00 | 9.41 |
| **Original UPB** | $241,750 | $223,000 | $1,039,000 | $17,000 | $119,301 |
| **LTV** | 73.06% | 77.00% | 104.00% | 6.00% | 16.50% |
| **Original Interest Rate** | 3.756 | 3.750 | 5.875 | 2.250 | 0.466 |
| **Original Loan Term** | 318.60 | 96.00 | 360.00 | 96.00 | 74.02 |

After looking at the data, we’ve found a piece of interesting information that may lead to valuable findings. The table above shows a brief descriptive statistics summary of numerical variables. First, the credit score, which summarizes the borrower’s creditworthiness, and it falls between 301 and 850. The mean of credit score is 751, where the median is 759. Significantly, the minimum score is 601, and the maximum score is 834. In terms of the number of units in the mortgage, the average number is 1, where the maximum is four units. The average original debt-to-income (DTI) ratio is 34.06%, the median is 35%, the max is 50%, and the minimum is 1%. The average if the UPB of the mortgage on the note date is $241,750, the median is $223,000, the maximum is $1,039,000, and the minimum is $17,000. The average original loan-to-value (LTV) is 73.06, where is median is 77, the peak is 104, and the minimum is 6. The original interest rate has an average rate of 3.76%, which is higher than the median by 1%. The maximum original interest rate is 5.88%, and the minimum is 2.25%. Finally, the average original loan term is 319 monthly payments, the median is 360, which is the same as the maximum, and the minimum is 96 monthly payments.

Figure 1 illustrates the correlation between 9 numerical variables; the average original loan-to-value (LTV), mortgage insurance percentage (MI%), risk level, DTI ratio, original interest rate, original loan term, credit score, number of units, and original UPB. It can be clearly seen that there are significant positive relationships between LTV, MI, and the risk level. The correlation between LTV and risk level is 0.42, and between MI and risk level is 0.68. Also, obviously, there is a notable positive relationship between the original interest rate and the original loan term, which is 0.64. As the number of scheduled monthly payments of the mortgage spreads, the original interest rate also rises. Significantly, the correlation between risk level and credit score is -0.06, which indicates that there is no significant relationship between these two. Moreover, there is a negative relationship between credit score and the initial interest rate. As the credit score decreases, the interest rate increases.

****

Figure 2 shows the average risk level in terms of channel and property type. Figure 3 indicates whether it is a retail, broker, or correspondent, and how each differs in terms of the average risk level. We can see that correspondent, an entity that typically sells the mortgages it originates to other lenders, has the highest average risk level, which is about 0.120. In contrast, broker, a person or entity that specializes in loan originations receiving a commission to match borrowers and lenders, has the lowest average risk level high than 0.075. Besides, the graph in the right shows how the type of property, which states whether the property type secured by the mortgage is a condominium, leasehold, planned unit development, cooperative share, manufactured home, or single-family home differs in terms of the risk level. The graph proves that planned unit development has the highest average risk level, which is higher than 0.10, whereas cooperative share significantly has the lowest average risk, which is lower than 0.01. For manufactured housing, single-family home, and condominium, they all have comparable average risk levels that fall between 0.08 and 0.10.



Prepayment penalty mortgage is a mortgage concerning which the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal. Figure 4 shows how the prepayment penalty mortgage affects the risk level. Clearly, the graph shows that the average risk level is higher if there is a prepayment penalty. Significantly, the graph shows that the average risk level with no prepayment penalty is not remarkably lower than with the penalty. The average risk level with a penalty is about 0.125, and the average risk level, if there is no penalty, is about 0.095.

**Model Analysis**

The Data mining tool we used to determine which customers are high-risk is a classification tree. We chose the classification tree technique since it performs well across a wide range of situations, and it is easily interpreted. Another reason for choosing a classification tree is because we could classify or predict an outcome based on numerical and/or categorical variables.

The goal of tree model is to classify or predict an outcome based on a set of predictors(numerical or categorical). There are two key ideas underlying classification trees. The first is the idea of recursive partitioning. Recursive partitioning is repeatedly splitting the records into two parts to achieve maximum homogeneity of the outcome within each new section. The algorithm inputs different values to maximize purity in the initial split. Pure here means containing records that belong to just one class. After the algorithm gets a split where maximum purity is achieved, the algorithm repeats the process for a second split, and so on. The second, on the other hand, is pruning using validation data. Pruning the tree is simplifying the tree by removing sections that have little meaning, and; therefore, avoid overfitting. The idea behind the pruning process is to find the point at which the validation error is at a minimum. One of the advantages of a classification tree is that it does not require the assumptions of statistical models.

In building our classification tree, we used the training set. Also, before creating our final model, we balanced the data. Data is unbalanced if there is an overwhelming proportion of 1 class, which in this case was median/low risk. An unbalanced data could lead to inaccurate results. For our tree, we included only variables that had a significant relationship with the risk level. All information that led to the final model is found in the appendix.



As shown in Figure 5, the first branch of the tree indicates the first rule, which is whether the customer has a Loan-To-Value(LTV) higher or equal to 91. If the customer had an LTV lower than 91, the tree shows that the customer has a median or low risk. However, if the customer had an LTV of 91 or higher, the tree takes another split. The second branch of the tree shows the second rule, which is whether the customer has a Debt-To-Income ratio higher or equal to 91. If the customer had a DTI ratio above or equal to 29, then that customer has high risk. Oppositely, If the customer had a DTI ratio below 29, then that customer has a median or low risk. In determining the accuracy of the tree, we applied the tree we created from our training set to our validation data. The accuracy of the tree in predicting the risk level (high or low/median) was 0.9924. The tree predicted 100% of the consumers with high risk while predicting 99.167% of the consumers with low/median risk.

The results shown in the tree are essential since it would help you identify which customers are most likely to default on their loans, and; therefore, save you a lot of money.



One way to show the accuracy of our model visually is through Figure 6, which shows the cumulative gains graph. In the graph, the grey dotted lines show the predictive power if we predicted randomly(no predictive model), while the dotted red line shows the optimal gains (the perfect model). As shown in the graph, our model, the red line, is exactly on top of the dotted red line, which means that our model has a perfect prediction for predicting risk level.



Another way to show the accuracy of our model visually is through Figure 7, which shows a type of graph called the cumulative lift graph. The lift line (the red line) is a measure of the effectiveness of our predictive model calculated as the ratio between the results obtained with and without the predictive model. Consequently, the greater the area between the lift curve and the baseline, the better the model. Therefore, our predictive model, according to the lift curve graph, does an excellent job.

## 

## Recommendations

The implementation of our Classification Tree isolated 2 variables as being the most important to determining a borrower’s default risk: LTV (Loan-To-Value) as the most important variable and DTI (Debt-to-income) as the second most important. As such, it is recommended that these ratios that should be the most looked when deciding whether or not a borrower is high-risk, medium risk, or low risk. Our recommendations and a further breakdown of the implications of these two variables can be further justified as following:

The loan-to-value ratio describes the size of the loan you take out versus the value of the property you are getting the loan for. The higher the LTV, the riskier the loan is, since the loan would be shouldering most of the cost for the property.[[1]](#footnote-0) This is why it makes sense for LTV to be the first rule in our classification tree (Figure 5), as it is the most indicative of how risky the loan is.

The debt-to-income ratio describes your monthly debt payments divided by your gross monthly income. The higher the DTI ratio, the more monthly debt the borrower has, ergo the less likely they are to stay on top of loan payments. As a benchmark, a 0.43 DTI ratio is the max ratio a borrower can have while still being eligible for a qualified mortgage.[[2]](#footnote-1) Because it paints a better picture of how reliable borrowers have been in paying past debt on a monthly basis, it is a good metric as the second rule in our classification tree.

## Sustainability

Looking forward, the next questions to be asked are the practicality of collecting data long term and how often the model should be updated. The answer to the former is fairly simple: if a potential borrower is confirmed as reliable (low or medium risk) and Stark Industries accepts them as a client, their behavior in paying loans will already be monitored. Stark Industries can then look to see if there are any irregularities in their behavior.

This leads to the next question: how often should the model be updated? Looking at the two variables used to measure customer dependability, the loan-to-value ratio is more of a ‘historical’ variable as it checks how big a customer’s loans have been relative to what they are using it for, so it will not be as valuable of a metric to look at since it only updates every time a borrower takes out a new loan. The debt-to-income ratio on the other hand, is measured each month since it looks at monthly income versus monthly debt payments[[3]](#footnote-2), so it would be a better metric to observe overall. In other words, our recommendation is to update the model each month, as a borrower’s DTI changes at that frequency.

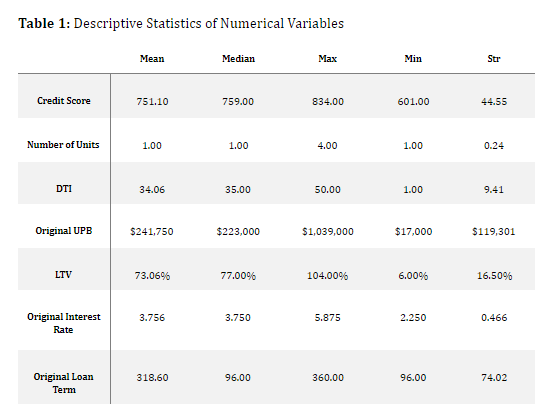
**Future Considerations**

For future modeling , there are a few improvements that could be made to the information available to create stronger models. As we are measuring consumer risk, having a stronger understanding of the history of an individual would be very helpful. The first-time homebuyer flag is beneficial, but understanding the reasons why one may be on their second home might make them and even larger risk than a first time home buyer.

There are also several variables pertaining to application of applicant income and loans that would prove beneficial in future modeling. For a loanee’s income, understanding how their budgeting is split would prove useful. This can be in the form of a percentage of one’s income already dedicated to financial obligations. Would an unforeseen financial event, like paying hospital fees, put our loanee at risk of defaulting on their loan?

When providing a loan to someone in the future, we should know what percentage of the applicant’s purchase will be covered by the loan. This will help inform how much of the loan is required towards the purchase of this item. Is the applicant looking to speed their purchase towards something they were already saving for? Would the purchase being made otherwise be impossible without a loan? These are very important questions we cannot answer without access to these details.

**Appendix**



A figure presenting the descriptive statistics for each of our important variables. We are observing the Mean, Median, Maximum, Minimum, and Str for our numerical values.



Figure one is a plot of variables and their corresponding correlation to one another. Those with a red background are associated with a negative relationship and those with a blue background are associated with a positive relationship. By doing this in our Descriptive Statistics, we can begin to work out which variables may have the largest impacts when constructing models. Additionally, we may be able to spot out variables that aren’t worth using.



Bar chart color coding average risk level. This is done to get a better idea of the distribution of values in our dataset. For this plot, we are using ggplot2, a graphing function found in the tidyverse package. We used channel and property type as our filters to measure the columns by.



Figure 3 is another bar chart from the ggplot2 function. Color coded, we have Mean\_Risk on the x axis and Property Type on the y. We are measuring the average risk associated with properties of specific types in this figure. We have also flipped its coordinates.



Figure 4 is another color-coded bar chart on mean risk. This time we are seeing if the mean risk is affected by where a prepayment penalty mortgage is detected in the data. What we see is that if there is, the mean risk associated will be higher.

**Prediction**

After working on the messy data containing numbers of different variables we came to the conclusion that classification model is the best way to analyze it rather than the KNN model. It is because of the following reasons:

**KNN MODEL**

**Strengths**

1. K-NN algorithm is very simple to understand and easy to implement. To classify the new data point K-NN algorithm reads through whole dataset to find out K nearest neighbors.
2. K-NN does not explicitly build any model, it simply tags the new data entry based learning from historical data.

**Limitations**

1. KNN cannot perform well with a large data set of numbers of variables. It struggles to predict the output of new data points.
2. K-NN algorithm is very sensitive to outliers as it simply chose the neighbors based on distance criteria.
3. K-NN inherently has no capability of dealing with missing value problem.
4. K-NN slow algorithm. As the size of the dataset grows efficiency or speed of algorithm declines very fas

**Classification model**

**Strengths**

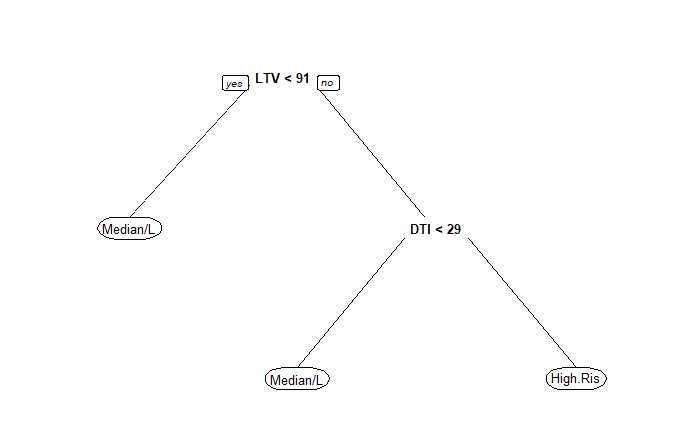
1. Classification model performs well across a wide range of situations, and it is easily interpreted.
2. we can classify or predict an outcome based on both numerical and categorical variables.
3. Classification predictions can be evaluated using accuracy, whereas regression predictions cannot.

**Limitations**

1. Decision tree often involves higher time to train the model.
2. A small change in the data can cause a large change in the structure of the decision tree causing instability.

Despite having own strengths and limitations of KNN model and the classification model we decided to go with the classification model mainly because of the versatility of classification model. Since our data has both categorical as well as continuous variable classification model would be the best fit.

While working on the descriptive analysis, we figured out that there are three main variables on the data i.e. Loan to value, mortgage and insurance percentage, and credit score. Also, while studying the data, we found that the data was unbalanced, and; therefore, we balanced the data. After balancing the data through the rose package, we made two decision trees with different branch levels. The first decision tree’s branch limit was set at one, and the accuracy of the tree very high when used on the validation set. However, we felt that the tree provided limited information since there was only one rule( whether LTV is higher or lower than 91). Therefore, we chose the second decision tree, which had a limit of ten, to use in our report. Even with a high branch limit, the tree produced only two rules, which we think was due to the data being too ‘clean.’

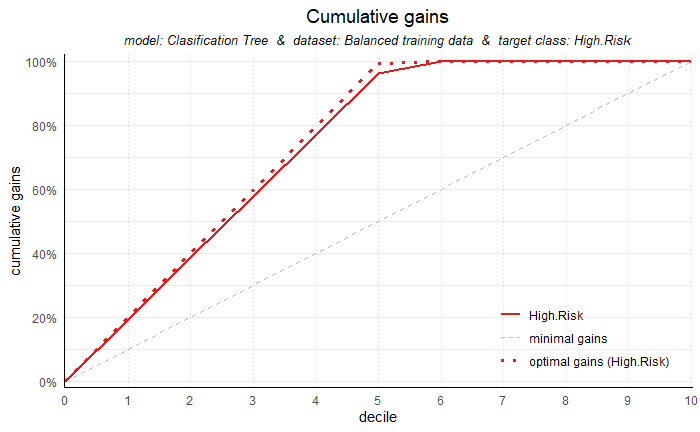


The figure shown above can be explained as follows:

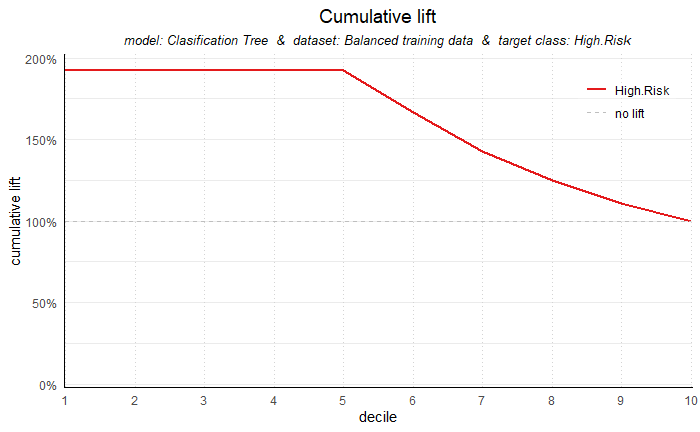
There are two ways to distinguish customers with a different loan to value number. If a customer LTV lower than 91, the customer is considered as a lower or medium risk.. However the customer has LTV of 91 or higher, there are two possibilities, either medium risk or high risk and this is measured by the Debt-To-Income ratio. Customer with a DTI of lower than 29 is considered as a medium or lower risk. Whereas customer with DTI higher than 29 is considered as a high risk for business.

After the classification model and the outcome from it we decided to check the accuracy. “How accuracy is our model and our output?”

The model we use is nearly perfect for the given data set. i.e our accuracy of prediction and the optimal accuracy level are around the same boundary. The dotted line shows the maximum accuracy that can happen to the given data set. Whereas the solid red line shows the accuracy of our model. It turns out that the optimal gain occurs at the 5th percentile as shown in the figure below.



We also checked the accuracy of our model using different method than just a cumulative gain. And it is none other than the cumulative lift. This model helps to measure the effectiveness of our model calculated as the ratio between the results obtained with and without the predictive model. In this model we considered 100% as a base. the 50th percentile was twice as much on a cumulative lift. However, there is a drastic change in the cumulative lift after the 50th percentile. It decreases dramatically to the 100 (base) again as it reaches the 100th percentile.



1. Hayes, Adam. “How the Loan-to-Value (LTV) Ratio Works.” Investopedia. Investopedia, December 3, 2019. https://www.investopedia.com/terms/l/loantovalue.asp. [↑](#footnote-ref-0)
2. “What Is a Debt-to-Income Ratio? Why Is the 43% Debt-to-Income Ratio Important?” Consumer Financial Protection Bureau. Accessed December 8, 2019. https://www.consumerfinance.gov/ask-cfpb/what-is-a-debt-to-income-ratio-why-is-the-43-debt-to-income-ratio-important-en-1791/. [↑](#footnote-ref-1)
3. “What Is a Debt-to-Income Ratio? Why Is the 43% Debt-to-Income Ratio Important?” Consumer Financial Protection Bureau. Accessed December 8, 2019. https://www.consumerfinance.gov/ask-cfpb/what-is-a-debt-to-income-ratio-why-is-the-43-debt-to-income-ratio-important-en-1791/. [↑](#footnote-ref-2)