

Predictive Model for loan defaults

Data Analytics Capstone



July 1, 2018

Robert cordrey

SNHU DAT-690

**Table of Contents**

page

Modeling Scenario 2

Data Security 8

Construction of the Predictive Model 13

Additional Considerations 22

Pilot Plan 25

Predictive Modeling Results 26

References 30

Addendum attached – Data Set for modeling and testing

Addendum attached – R Script used for the Predictive Model

Addendum attached – Data Processing Flow Chart

Addendum attached – Exploratory Analysis Flow Chart

**Modeling Scenario**

**Introduction**

The modeler will be using the raw data set as provided by SNHU to determine the likelihood of a customer to default on a loan will assist in minimizing the risk posed to the lending institution. Understanding this likelihood will allow for business rules to be constructed to dictate the action to be taken, by the institution, based upon a scoring model. The lending institution is primarily interested the possibility of default while other business rules and/or actions may be created. The modeler will use the framework of Cross Industry Processing for Data Mining (CRISP-DM) throughout the life cycle of this project. CRISP-DM is a well-proven methodology that contains six major steps, Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The use of a model that follows a proven and standard practice is required to maintain industry standards for data mining which allows for the proper execution of a data solution.

**Business Understanding**

The goal of this business problem is to develop a model that will minimize the financial risk posed by a lending institution from a customer based upon preselect measured criteria. These criteria will use standards and scores that have been preapproved by the lending institution as having met the model requirements. The model will be deemed successful with a prediction accuracy of that determined by the lending institution for identifying a customer that is likely to default on a consumer loan. The data set does not use any criteria that is directly related to a specific individual or class and/or grouping of individuals. Each observation is represented by a unique identifying value. The received data set is assumed to be true and accurate, without any misrepresentations.

**Data Understanding**

The data to be used has been received as a cvs file from Southern New Hampshire University (SNHU) titled as CREDIT\_DATA. The data set contains 1000 observations consisting of 32 variables, each with an appropriate, identifying, header. The formatting and contents of the file appear to meet the requirements to build a successful model. The specific target for this model is the variable named “default”. Many variables used contain a value equally a yes or no response or contain a value based upon a predetermined scale correlated to the response received. For those observations outside of this window, a summary of the data is shown below.

AMOUNT AGE

Min. : 250 Min. : 19.00

1st Qu. : 1361 1st Qu. : 26.00

Median : 2218 Median : 32.00

Mean : 3192 Mean : 35.16

3rd Qu. : 3833 3rd Qu. : 41.00

Max. :18424 Max : 75.00

The AMOUNT shown above is that of the current credit limit of each observation. The AGE is based upon the response submitted by each observation.

**Data Preparation**

Before the model construction begins, all data quality problems and preparation activities must be completed. The structure of the data set is that each value is represented by the Numeric data type. Most inputs are the result of answering a specific question resulting in an integer, yes or no, or selecting a specific, pre-determined, choice from available options. Based upon the question posed, each answer was converted into an integer value. Within the various data types, the Observation Identifier is unique. It is critical to have a unique identifier for each observation. Without this unique identifier, the resulting outcome would not be accurate. There are not any missing values within the dataset for each observation. In the case of a missing value, one would need to create a rule as to the method of handling this issue. Examples of potential rules for handling a missing value is that the row containing the missing value is removed, the mean or median of the column containing the missing value would be substituted in the empty cell. While there are limited choices as to how this should be handled, it is not necessary to give any consideration for this data set. There does not appear to be any issues with the quality of the data set. Based upon the quality of the data set, there are no reason for any data to be excluded for use in the model. The cleanliness of the data appears to be very high and is sufficient in its current state. It will not be necessary to merge this data with another data set due to the completeness of the received data. Additional aggregation of the data is not necessary for this model.

**Modeling**

This model will use logistic regression as the basis of construction as its tool. The use of logistic regression will result in receiving a probability between 0 and 1, as a non-continuous outcome. Logistic Regression is used when predicting an outcome that is obtained from the use of categorical data. The use of logistic regression will allow for the model to interpret the data in a linear fashion. Because the question to be answered deals with the relative importance of each variable as it is related to the target, logistic regression can answer this question with a high degree of accuracy. An additional benefit of using logistic regression is that the response is used to predict the likely hood of an event happening based upon the value of multiple values from a sufficient population. The value that is targeted is “default” and the lending institution desires to know the likelihood of a customer default occurring. To test the validity of the model, it will be necessary to divide the data set into three sets, using a 70/15/15 split among the data. The numbers represent how the data will be split for training, validation and testing, respectively. The seed will be set as a consistent number to have the ability for reproducing the results to allow for further review and validation by the modeler and others. An additional point for possible further study is that of correlation of variables. While it is possible that an independent variable may on cause a slight variation on the target, when pairing this variable with one or more other variables, it is possible that the importance of this may be increased. This question is not directly posed but can present cause for an additional study if deemed necessary.

**Evaluation**

Within this step, it will be necessary to test the accuracy and general results of the model. This testing is necessary to ensure that model answers the business question posed and that the answer is accurate within the desired scope. Other results uncovered during the data mining process may be studied during this phase, as time permits. Any additional results uncovered may also reveal additional challenges and possible future questions that need an answer. During the evaluation phase, it is necessary to decide what steps are required to proceed and how complete these steps. Within each step, it may be necessary to plot an action and reaction to ensure that the process remains within the scope of the business question posed.

**Deployment**

Within this phase, it is necessary to review the evaluation results and formulate a plan for its deployment. This deployment is critical, as all other phases, to the success of the project. For this project, a written report and/or a PowerPoint presentation with the necessary analytical charts and/or graphs as needed to support the facts presented. If it is necessary create a model using the same variables from different observations, this can be easily and quickly created. The model will have been tested and proven to be accurate. If the variables are changed when needing refreshed report, it will be necessary to create a new model. The accuracy of the model is dependent on the quality of the data and the variables represented. Any changes to this will result in a model that is no longer considered accurate. Within the conclusion of the deployment phase, the modeler can provide additional documentation summarizing any important experiences gained during the project.

**Analytic Structure**

The chosen model will rely on the use of predictive analytics. Predictive analytics relies on the technique used to identify the likelihood of a target based upon the historical data associated with known observations. As a proven method to reduce risk, knowing the likelihood of a customer to default on a consumer loan lends to the institution being more fiscally responsible. As a steward of being fiscally responsible, positive rewards will be reaped. Predictive analytics will also allow the lending institution to generate additional sales based from programs designed to meet a specific campaign. A predictive model does require care, which in turn, generates a stronger model. To accurate determine which step to take next, one must first understand how the business reached it current standing. Predictive analytics will adequately and sufficiently show what needs to be done next based upon the historical actions completed. The primary focus of predictive analytics is the of a relation between a target and other known variables.

**Modeling Tool**

The chosen tool for this model is that of using logistic regression within R. R is an open-source language that is free to install, use, update, or modify. Initially, R is challenging to learn. From a data science prospective, there are two primary choices, R and Python. Each language has its strengths and weaknesses thus R is able adequately handle the needs of this model. The use of logistic regression is the choice of analysis when there is a need to explain the relationship of a target to one or more variables. Increasing the number of variables will lead to an improved statistical validity. Conversely, fewer variables will lend itself to a lower value of statistical validity. The modeler must be sure that an excessive number of variables are not added to the model that would cause the model to be inefficient and allow for over-fitting to occur.

The modeler will be using the raw data set as provided by SNHU to determine the likelihood of a customer to default on a loan will assist in minimizing the risk posed to the lending institution. Understanding this likelihood will allow for business rules to be constructed to dictate the action to be taken, by the institution, based upon a scoring model. The lending institution is primarily interested the possibility of default while other business rules and/or actions may be created

**Data Security**

**Model Success**

From a security prospective, the model will be successful with the return of a value that does not use or reveal any data that is personally identifiable while maintaining the all legal requirements and the community standards for its ethical use.

**Data Security and Privacy**

Any information stored by a business or individual must be properly protected. Information by itself is a word used to cover many facets of our life and in our surroundings. This information may be of varying importance from keeping a diary of your eating habits, to maintaining health information, to recording of various financial records, just to give a few examples. Various laws and regulations have been enacted to require specific types of data are protected in a specific manner. These laws and regulations are brought about by various agencies within the federal government, state governments, federal or state based regulating agencies, along with various local municipalities. The specific concern of each law or regulation is to set forth requirements for the safekeeping and accuracy of all lawfully obtained data. That data which contains personally identifiable or financial information is held to a much higher degree of security, due to the nature of its contents. The discussion on the importance of security and privacy is easily a lengthy topic. The purpose of this content is to summarize its overall importance as more detailed information is available from numerous online sources.

**Security concerns or risks for data storage**

To determine if there are any security concerns or risks, one must first understand the meaning of a security concern and/or risk. The storage of data is easily put at risk by one of four major groups, as to cover most issues. These four groups are theft, loss, insecure practices, or neglect. Theft can occur by means of individual acts or deliberate attacks on a system. The number one cause, in my opinion, is caused by insiders of a business. The loss of data can occur when media or papers are inadvertently left behind. Examples of places where lost media can be found is in a taxi, airport, restaurant, or any other place where one may choose to conduct business. Insecure practices cover many different aspects from poor password managements systems to the storing of unencrypted data. Anytime that data is collected, regardless of its purpose, any security concerns for this data should be addressed immediately. Neglect results in the careless handling of data. An example of this is that when a computer is sold and the hard-drive is not properly erased. There are tools available, for free, online that would allow for one to recover data that the owner had intended to remove from the hard drive. For those hard drives or servers containing very sensitive data, the best prevention is to physically destroy the hard drive and/or server.

**Data Set Security**

The data set, as received, does not appear to have any security concerns. Within the data set, there is not any personally identifiable data.

**Data Set Ethical Concerns**

While the collection of data may be permissible, this does not mean that it is necessary to use all data. Collecting of data occurs at every turn in our life and our actions. For example, when a user clicks on a web site, this is collected about that user. The tracking of customer purchases raises not only privacy concerns but ethical concerns. Many people are familiar with the story of Target determining that a teenage girl was likely to be pregnant due to her purchasing habits. While this practice was not illegal, questions were raised as to the ethical practices used and the resulting outcome of this data mining event. The damaging part of Targets action is due to the age of the female, not because the algorithm pointed to her as a likely candidate to be pregnant. It is important to determine exactly what information is needed and then decide if the obtained information is both legal for use and ethically correct. Everyone and each business is accountable for upholding their legal responsibilities and acting in a manner that is ethically responsible and correct. There does not appear to be any data that would raise the concern for being ethically challenged in the data set. While it is legal to collect data, this does not mean that the data collected is ethically right. For the responders to intentionally answer a question with a wrong answer does not mean that an illegal act occurred but it would raise the question about the answers being ethically correct. Within the received data set, it is assumed that all responses were gathered in a manner that is both legal and ethically correct. Another example of an action that is not illegal but is ethically wrong is to collect data and not being truthful about its intended usage. The act of lying, is not illegal but does open the door to other legal actions such as fraud or misrepresentation, to name a few possibilities.

**Data Set Security Strategy**

Using the data set provided, there does not appear to be any privacy or ethical issues. The lack of privacy or ethical issues does not require any extraordinary security standards. While each variable name is self-explanatory, the value assigned to most variables can hide its true meaning. Reviewing the data, there are many variables that a user may try to guess its actual meaning such as does the applicant own a car. Showing a pattern with responses being a 0 or 1 could lead to the conclusion of a yes or no response since this is a simple question to answer. There are other variable responses represented with an integer response that are using more than a 0 or 1 since these variables are not yielding a simple yes or no response.

In looking at the various tools available for data security, there is not a “one-size-fits-all” solution. As vulnerabilities evolve, so must the enterprise security. For those companies working in an Agile environment, this will be an almost daily task to be reviewed. As a released is published, security standards and protocols must remain sufficient to maintain the required level of protection from internal and external threats. Having a thoroughly trained security specialist within the IT department will minimize the potential of a threat becoming a reality. Without this security specialist, there are many available online but do their qualifications meet your minimum requirements?

The use of encryption is to secure any sensitive information from unauthorized access. It is possible to enforce encryption requirements throughout the enterprise or only in a local setting. For any machine accessing sensitive data, it should also be mandatory that either the external media is disabled or encryption is required. There have been many cases where unauthorized individuals have accessed sensitive data or an external media was misplaced and found by an unauthorized user. In addition to the encrypting of the data, hard-drive encryption of the local machine is not to be overlooked. Regardless of the situation, steps must be taken to prevent any access by an unauthorized user.

Passwords for accessing sensitive data are not to be taken lightly. Sadly, there appear to be individuals that will use a bot to repeatedly try to gain access your data. This takes very little effort on the part of the programmer and the results, if successful, can be very rewarding. More times than not, companies do not enforce rigorous passwords for the typical enterprise user. This policy can only be made with the agreement of those responsible stockholders. Leaving the final decision to an IT Security Specialist would probably result in requiring a very lengthy and difficult password. Imagine how the users would revolt when their selected password of “password123” would no longer work!

Any script that is used to access the data must be protected from Cross-site Scripting. Malicious JavaScript can access the session tokens stored within a cookie. The use of this session token could allow the unauthorized user to impersonate an authorized user. Cross-site scripting was once thought of a problem only for the user. If the problem can affect the user, it can affect the security of your data.

Most programmers seem to reuse scripts from previous applications for use in a new application. In general, this practice does not bring with it any unusual security concerns. When reusing a script, it must be secured that would prevent an unauthorized user from exploiting a bug that would result in a malicious action. It is also important to consider whether the script is accessing a legacy application that may present unique vulnerabilities not previously considered.

**Summary**

The security of the model will be deemed successful when all legal requirements have been fulfilled and the data is only used in a manner in accordance with the ethical uses as deemed appropriate by commonly accepted community standards. The security of data is of great importance for many reasons. To create a security policy that is complex and requires a lot of work from the user will create an environment that is easy to ignore and will likely fail. To improve the chances of success, an enterprise environment should employ many techniques that require little input from a user prospective.

**Construction of the Predictive Model**

**Model Creation**

The purpose of this segment is to introduce a plan for the analyzing of data in a method and manner that is completely reproducible. Any conclusions drawn from research that is not reproducible should not be trusted. It is paramount that the analyst allows for the research to be reproduced by sharing of all data and code. The doubt cast by controversies regarding the replication of results will quickly cast doubt with the validity of conclusions that are observed. Without regard to the social or profession standing of the analyst, complete transparency must occur during the entire analyzation.

**Predictive Algorithm Method**

The data being used is provided by the lending institution without any input or suggestions from a known third-party. The data appears to be complete and without any missing values. The lending institution desires to whether it can expect a customer to default on a loan based upon the data from 32 variables all of which are numeric in nature. There are 32000 observations in the data set provided. The initial input of the variables where originally received as either a binary or classification response, except for assigning a unique value for each observation. Considering the simplicity of the question, the completeness of the data provided, and the yes or no response desired, a logistic regression is best suited for this research. The targeted outcome of the analysis is to determine if a customer is likely to default on a loan using quantitative, qualitative, or a combination of methods to analyze the variables in predicting the targeted outcome.

**Predictive Model Recommendation**

A logistic regression is best suited for this research for a variety of reasons. Most importantly, logistic regression is well-suited when needing to understand a binary response of an outcome based upon various decisions and/or choices that were made earlier. The goal of using a logistic regression is to determine the relationship between the targeted value and that of independent variables. The use of logistic regression is to produce a formula that will predict the probability of it characteristics in relation to the targeted variable. The time needed to prepare the data for use is minimal in the presence of missing values. Since this does not appear to be an issue with the data provided by the lending institution, no further explanation is needed. In the event of outliers within the data, a logistic regression is not effected. Considering this to be a non-influencing data point, the analyst does not need to investigate the positioning of potential outliers. A logistic regression offers a format that is easy to explain and understand. As mentioned earlier, the transparency of the results is an important factor in all research projects. The clarity and transparency of the summary is important in the decision-making process. The random forest is capable of easily handling a combination of values with the associated features with ease. In the result of a logistic regression, a summary representation will be presented in a simple and easy to understand format that generally requires no further explanation. Unfortunately, a logistic regression on a small data set is prone to over-fitting of the data. Without being properly fitted, the results may not prove to be as accurate as expected. As a measure of overfitting, the area under the Receiver Operating Characteristic (ROC) curve should not be greater than 95%. Any value greater than 95% would dictate a closer examination of the results. The ROC curve provides a summary for the predictive powers of the associated variables within the data set. Using the Area Under the Curve (AUC), as it is shown in the ROC curve, is an index for the accuracy of the model. A simplistic interpretation of a AUC of 90% is that the model would prove to be true 90% of the time.

**Reproducible Research**

When considering the case of reproducible research, the analyst must consider that degree of validity that will be assigned to the outcome. Many times, a project may be started without concern of needing to reproduce the work later. As a project moves forward, it may be within the scope to create a “what-if” or sample set using randomly selected observations. The outcome may yield positive results and you wish to share these results with a fellow analyst for review. How would the reviewer mimic your results other than to simply read your script and assume that the results are accurate? Whether one has the intention of producing results to be reproduced, an emphasis should always be placed on if the results will need to be reproduced. It is irrelevant who will be reproducing the work, with setting the seed for reproduction, the exact results will vary. The workflow of the research should be in a simplified manner without confusion or redundant processes. In a simplistic manner, the workflow begins with data management, followed by the analysis, the conclusion of results, and finally the written report of the analysis. A written record of each process within the various steps are required for later verification and reproducibility. This documented record of events will allow for the reproduced analysis in determining the relationship between the given variables as provided by the lending institution For the analysis of the data set from the lending institution will complete all work within the framework of a R repository, with the complete script being available for review and verification. For the purposes of reproducing the analysis, the set.seed function will be used. The set.seed function in R allows another analyst or researcher to mimic the results as originally obtained. Any time that random numbers need to generate, such as to choose random rows in a dataset, the identical results will be returned by the user due to using the functionality of set.seed. When a dataset needs to be divided into x different sets by random choice, the set.seed function will produce identical results each time the identical value is used as in the original analysis. This function is paramount for being able to reproduce the results of an analysis.

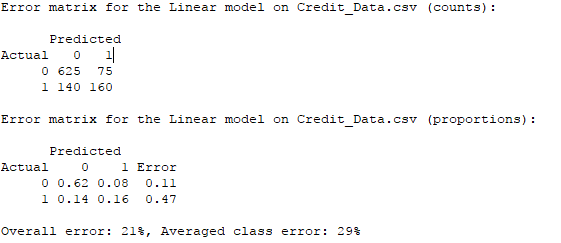
**Evaluation**

For this exercise, I will discuss the structure of the model using the dataset as provided by the lending institution. As the model is constructed, each step will be evaluated and discussed explaining the reasoning for each process. The steps for implementing a continuous scoring engine will be discussed and explained as the model is being built. There are various features available that will be used to determine the validity and where improvements, if any, can be made to the model. And finally, documentation will be provided for the results of the summary used for the predictive model.

The data used for this model consists of 1000 observations of 32 input variables, in a csv formatted file. The target question is to determine whether a customer will default on a loan. The variables will not be assigned a weight of importance, only a yes or no response is desired as the targeted response. A logistic regression method will be used within R. The default settings will be used. If any value other than a default setting is chosen, this will be specifically addressed.

**Implementation Steps**

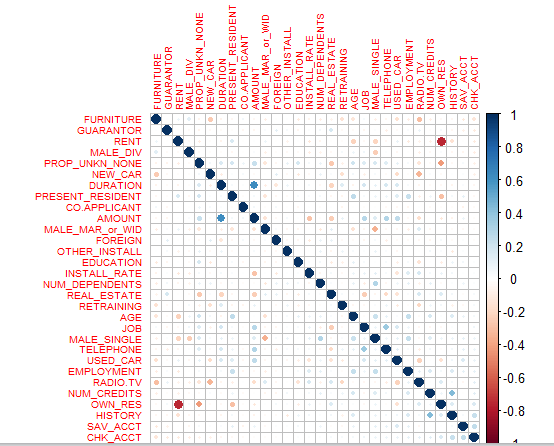
To continue building the model, the evaluation of a Confusion Matrix returns the below chart for review. The purpose of the first step is to review the results within the Confusion (Error) Matrix. When reviewing the results of the Confusion Matrix, it is more important to focus on the value returned at the Actual/Predicted intersection, the upper left quadrant. Within the Count portion of the Confusion Matrix, the numerical value is expected to be higher than the other three values shown. In reviewing the Proportions section, the value at the intersection of the Actual/Prediction is expected to be high. The closer to 100%, the better the model.



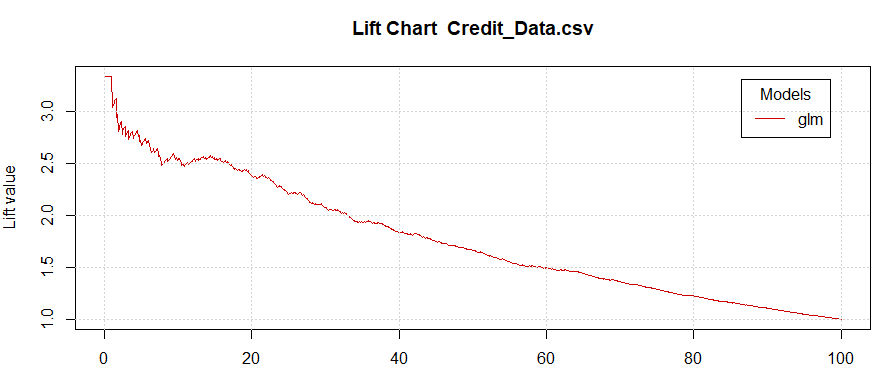
**Model Evaluation**

It is necessary to evaluate the various models available based upon the data set in use. It may be necessary to create numerous model selections so that each can be compared against one another.

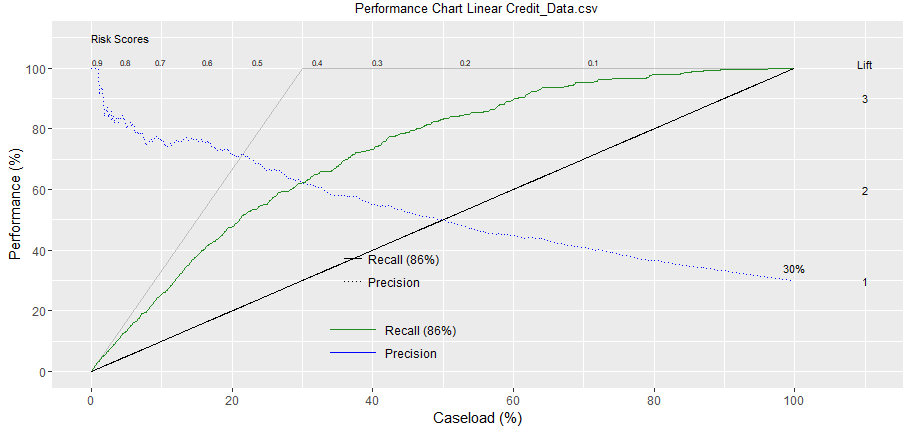
To determine the relation between the target variable and each of the independent variables, a correlation matrix is plotted. Using the legend on the right, it is observed that owning a residence shows the strongest degree of correlating to the target variable of defaulting on a loan.



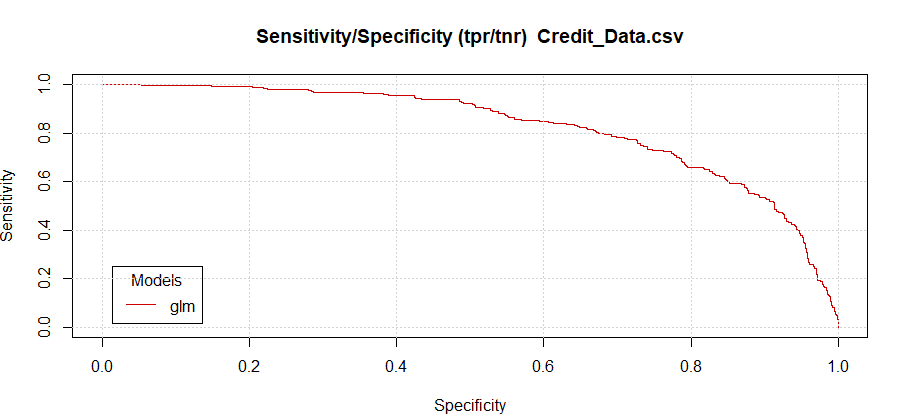
For reaching the top 80% of customers, a Lift of approximately 1.25 is observed. The purpose of understanding the Lift value is to measure the ratio of the results that will be obtained with and without a predictive model. It is also important to consider the consistently of results as plotted in each respective Lift chart. At a quick glance for a Lift chart, the greater the distance between the plotted line and the baseline, the better the model.



The next point to consider is that of Risk. Often it is suggested that these results should be more appropriately titled as Gain instead of Risk. The gain returned is associated with each dataset variable. Any adjustment that is needed to the variable is adjusted as a measure of its importance correlated to its specific gain, or risk. The optimum risk, or gain is where the Area under the Recall curve is maximized. The logistic regression model returned an area under the curve at 86% with a consistent plotted curve. The results for the logistic regression model is shown below.



Moving to the sensitivity of each model, generally a visualization determination can be made for choosing the optimum model. A line will be plotted that considers the variable sensitivity to any change within the model. What we are looking for is the plotted line that maintains the highest value on the sensitivity scale before dropping down to meet the specificity axis at the bottom. The optimum model will maintain a higher value along the sensitivity scale for a longer duration, or length. Below, the Sensitivity chart for the logistic regression model is shown.



Finally, we will consider the Scoring results for each model. These results are attached to this paper for individual review, if desired. The purpose of scoring of variables is to allow for further exploration using other tools and/or for the actual model deployment. The ordering of the score is done in the same order as shown in the original data set. The selection was made before executing the score selection to return the values as a probability instead of selecting a class. The selection of class will return a 0 or 1 value. This may not prove to be useful when plotting the data since only a 0 or 1 is shown. The csv file named Credit\_Data\_score\_idents\_class\_identifiers.csv returns the binary response for each customer regarding the likelihood of defaulting on a loan. Plotting the probability will allow for further and more detailed examination of the data. The probability plotting may consist of using box plots, histograms, or many other visual representations, as deemed as appropriate. The csv file named Credit\_Data\_score\_idents.csv returns a score assigned to each customer, using the given variables, as a percentage of the likelihood of defaulting on a loan. A customer with a higher probability is more likely to default in comparison to a customer with a lower percentage score. These files are shown in the addendum for individual review. The below table shows the example of the expected results with and without using the predictive model. The improvement to the profitability when using the predictive model will lend itself to increased fiscal responsibility.



**Expected Results**

As various tests and charts have revealed the results of the logistic regression model, this model has consistently proven to be the best for use with this dataset. The logistic regression algorithm, using the traditional settings, has consistently returned the most favorable results. This conclusion is valid on for this dataset and is expected to show a high degree of reliability in determining whether a customer is likely to default on a loan using the variables presented.

**Additional Considerations**

There are additional data sources that may be considered which will lend to an improved predictability response for the likelihood an applicant to default on a loan. There are many sources of external data available. It is the responsibility of the lending institution to determine those sources and measures that meet the lending criteria considerations. Below, I have shown six examples of additional data to be used and the suggested method for recording each within the data set. Following the six examples, I have submitted a scenario showing the value of the improved predictability indicators and the effect on the profitability of the lending institution.

**Current debt to income ratio**

The current debt to income ratio (DTI) is a good indicator of the potential risk from an applicant based on the current situation. While various industries may have different requirements, as a rule, those with a DTI of less than 28% have proven to be a lesser risk than an applicant with a higher DTI. The DBI does not paint an accurate picture of the total risk based upon the applicant but does present an accurate overall assumption of the potential risk. It is recommended to use categorical values as determined by the lending institution for the required parameter limits.

**Post-loan debt to income ratio**

It is important to consider the effects on the applicant if approved for the loan. The lending institution must be aware of the changed financial position the applicant will be in if approved for the applied loan. The criteria from the lending institution should specify various requirements to ensure that the applicant will be in a financial position for loan repayment. Obviously, offering a loan to an applicant that is not financially able to repay a loan is not conducive to the lending institutions financial practices and profitability. It is recommended to use categorical values as determined by the lending institution for the required parameter limits.

**Regional unemployment rate movement**

While being unemployed does not directly affect a credit rating, it is important to understand the employment rate in a regional market. Looking at this from a macroeconomic point, if other similar or connected industries are moving towards a more conservative business practice or reducing labor costs, this will most likely affect the applicant’s employer. It is equally important to understand the business climate as a consideration of the potential for a future event. If the applicant is employed by a company that is not broadly diverse and operated in a global economy, it would be prudent to understand the economic conditions of those countries in which the applicant’s employer conducts business. It is recommended to use categorical values as determined by the lending institution for the required parameter limits.

**National unemployment rate movement**

The effect on the national unemployment rate is more of a generalized indicator of those actively seeking employment that are without employment as reported by various state agencies. Government agencies report only on those individuals actively seeking employment while collecting monies available through unemployment benefits. Once these benefits have been exhausted, an accurate count of those being unemployed may become skewed. An individual receiving partial unemployment benefits due to receiving pay from a less-than full-time job is counted as being unemployed. During this time, it is thought that this person would accept a full-time position if offered. While receiving unemployment benefits, these individuals are under-employed. It is recommended to use categorical values as determined by the lending institution for the required parameter limits.

**Applicant owns a small business**

As a business is created, typically it is operating on a shoestring budget with limited funds. These funds are generally provided by the owner since the actual business does not have a credit history. The applicant seeking a loan for a small business is not a new concept and this event happens frequently. It may be necessary for the applicant to seek a personal loan using their creditworthiness as collateral for loan approval. The review of a submitted business plan would lend additional creditability to reaching a loan decision. It is recommended to use a binary value as determined by the lending institution for the required parameter limits.

**Level of education**

The level of one’s education is not reported by any financial instruction as part of a credit history. There is a direct correlation between the type of college degree and the income earned. There appears to be a distinctive line between that of a professional degree and undergraduate degree. Those with a graduate degree or higher tend to achieve a higher credit score than those without a degree of this type. It is recommended to use categorical values as determined by the lending institution for the required parameter limits.

**Improved predictability scenario**

The primary benefit to using additional data sources for the loan making decision process is to improve the rate of repayment by mitigating the default rate. The lower default rate will have a direct impact on the profitability of the financial institution. For example, it the lending institution has a default rate of 4.5% equaling 1,500,000 dollars, cutting this rate to 3% would improve the profitability by 500,000 dollars. If the default rate is reduced an additional .5%, this would amount to an expected gain of more than 666,666 dollars. As the purpose of the lending institution is to make a profit from customers by means of monetary loans, it is prudent that these offers are made only those applicants that meet the necessary requirements while minimizing the exposure the any negative impacts or results.

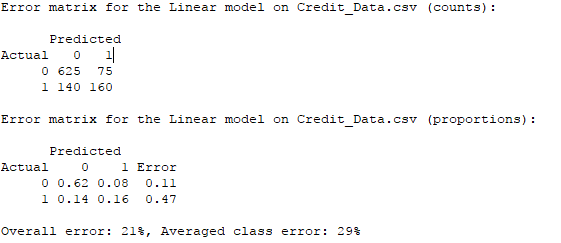
**Pilot Plan**

The score file created is in the form of a csv formatted file, which will be exported to the directory or server as specified by the lending institution. This file type is needed to load the data into a custom-built cloud-application using another tool named Shiny. A custom-built interface will be built using Shiny that will generate the predictive results in a table-like format for all loan applicants. Shiny can create a friendly user interface in a simple to read and easy to understand format. The scoring model created within Shiny will show the predictive score for each applicant, along with other data as determined by the lending institution. The final presentation of the user-interface for the predictive results are at the sole discretion of the lending institution. It is recommended that a button is available within the user-interface that will allow for the predictive model to be updated as the demand is deemed necessary. As data values within the dataset are updated, the user will be able to select a button to generate the new predictive score. The predictive model will be exported in the Predictive Model Markup Language(PMML) as an automated process. PMML is readily recognized and accepted by the current computer technology used within the statistical computing community.

**Predictive Modeling Results**

Before the implementation of the credit risk model, there are important questions that must still be answered. It is important to keep in mind that the lending institution desires to maximize profits from a loan while minimizing risks. The risks to be considered are that of the potential borrow and the associated interest rate to be charged to the potential borrower. Answering these two questions appear to be straight-forward since the lending institution has policies in place to provide the answers. It is important that the results of the model are directly related to the business plan and goals of the lending institution. A scoring model can be extracted to provide more depth to the insight of individual applicants but this does not directly answer the posed question of whether an applicant is likely to default on a loan.

The Confusion (Error) matrix for the model is shown below. The model returns a 78.5% of accuracy of a correct classification. The error rate for this model is 21.5%. The precision for this model or when it predicts yes, is it correct, is currently at 68%. To increase the accuracy and precision of the model, additional variables may be needed.



The sole determination of whether an applicant is likely to default on a loan does not lend itself to the ability to maximize profits. The stated sole purpose of the lending institution is to only make loans to the applicants that are not likely to default on a loan is to reject those that were determined as likely to default. That is, any applicant with a “Yes” for the likelihood of defaulting would be automatically rejected. It is those customers that are likely to repay the loan that present not only the smallest risk but also the lowest profitability. The customers that are not likely to default are also aware of their payment history and understand the importance of receiving a low interest rate for a loan.

In addition to assigning a Yes or No determination to the likelihood of defaulting, a level of credit risk should be devised based upon the applicant’s profile of the model results. Those applicants with a low likelihood of defaulting should be receiving an appropriate interest rate for loan repayment. The next classification would be for those that are likely to repay the loan but have scored higher using the score as represented by the logistic model. The model will select those customers that are still likely to repay the loan but with a slightly weaker profile than the premier loan applicants. Those applicants in this classification will represent only a slightly higher risk but should generate considerably more profit in the form of charging a higher interest rate. The assignment of an interest rate has a direct impact on the monthly loan payment. It is this monthly loan payment that applicants have the most sensitivity. It is necessary to find an equal balance between that of the lowest interest rate for the customer and the ability to maximize the potential profit.

The final classification of applicants should consist of those in the lower end of having the possibility of defaulting on a loan. It is within this classification that the lending institution has the potential to accept the highest degree of risk and the potential for the highest profitability. The determining factors for this classification are made by the lending institution in accordance with established policies and regulations. The financial condition of the lending institution can be improved dramatically by simply managing the potential risk, not necessarily eliminating all risks. The below chart reflects potential outcome for the suggested three tiers of applicant classifications.

It is observed that the applicant’s receiving the lowest interest rate are also the least likely to default represent the minimal amount of financial return, with an estimated gain of $603.20. Those applicants with a slightly higher risk but are still likely to repay the loan represent a slightly better return, approximately $1032.32. For the third classification of customers that may default but present on a slightly higher risk, the financial return is considerably more improved at approximately $2151.20. All loans were calculated for borrowing $10,000 for a 48-month term, with a payment due on the last day of each month. The interest rates charged for this example .029, .049, and .099, respectively.

The implementation of this model, using a tier calculation will lend itself to an improved financial position by the gain of additional revenue while mitigating risks. This improved financial standing will have a direct impact on the profitability of the lending institution and an expected approval rating from the institutions shareholders. It is the responsibility of the institutions executive team to optimize a strategy that will optimize the financial return of investments while maintaining adherence to the policies and regulations set forth. Having the ability to use various profiles or strategies will also the institution to remain nimble in any changes to the processes involved while remaining in a manageable condition. The possession of a single strategy does not allow for the lending institution to reach the determined business goals as this model is three strategies for the focus of minimizing risk and maximizing profit.

The below table shows the example of the expected results with and without using the predictive model. The table was derived from using values directly received in the data set provided by the lending institution. Any procedures that were applied were done with uniformity and consistency. These results yielded a significant difference between that of using no model and using the suggested predictive model. The improvement to the profitability when using the predictive model will lend itself to increased fiscal responsibility.



**References**

Story, M., & Congalton, R. G. (1986). Accuracy assessment: a user’s perspective. *Photogrammetric Engineering and remote sensing*, *52*(3), 397-399.

Tranmer, M., & Elliot, M. (2008). Binary logistic regression. *Cathie Marsh for census and survey research, paper*, *20*.

Lemeshow, S., & Hosmer Jr, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology*, *115*(1), 92-106.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 705-724.

Hilbe, J. M. (2009). *Logistic regression models*. CRC press.

Christensen, R. (2006). *Log-linear models and logistic regression*. Springer Science & Business Media.

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. In *Data Mining and Knowledge Discovery in Real Life Applications*. InTech.

Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*.

Spooner, D. L. (1988, October). The Impact of Inheritance on Security in Object-Oriented Database Systems. In *DBSec* (pp. 141-150).

Hinke, T. H., Delugach, H. S., & Wolf, R. (1997). A framework for inference-directed data mining. In *Database Security* (pp. 229-239). Springer US.

Shulman, A., & Co-founder, C. T. O. (2006). Top ten database security threats. *How to Mitigate the Most Significant Database Vulnerabilities*.

Aissa, A. B., Abercrombie, R. K., Sheldon, F. T., & Mili, A. (2010). Quantifying security threats and their potential impacts: a case study. *Innovations in Systems and Software Engineering*, *6*(4), 269-281.

Chandrashekhar, A. M., Ahmed, S. T., & Rahul, N. (2015). Analysis of Security Threats to Database Storage Systems. *International Journal of Advanced Research in data mining and Cloud computing (IJARDC)*, *3*(5).

Brauch, H. G. (2011). Concepts of security threats, challenges, vulnerabilities and risks. *Brauch, Hans Günter; Oswald Spring, Úrsula; Mesjasz, Czeslaw*, 61-106.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965-980.

Harrell, F. Logistic Regression Model.

Steyerberg, E. W., & Eijkemans, M. J. (2000). Prognostic modeling with logistic regression analysis. *network*, *10*, 11.

Chan, Y. H. (2004). Biostatistics 202: logistic regression analysis. *Singapore medical journal*, *45*(4), 149-153.

Condous, G., Okaro, E., Khalid, A., Timmerman, D., Lu, C., Zhou, Y., ... & Bourne, T. (2004). The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. *Human Reproduction*, *19*(8), 1900-1910.

Chamberlain, G. (1979). Analysis of covariance with qualitative data.

Mihail, N., Cetina, I., & Orzan, G. (2007). Credit risk evaluation. *Theoretical and Applied Economics*, *4*(4), 47-52.

Palmer, F. P. (2011). *U.S. Patent No. 8,086,523*. Washington, DC: U.S. Patent and Trademark Office.

Tom, M. D. (1997). *U.S. Patent No. 5,696,907*. Washington, DC: U.S. Patent and Trademark Office.

Palmer, F. P. (2013). *U.S. Patent No. 8,407,139*. Washington, DC: U.S. Patent and Trademark Office.

Frey, R., & McNeil, A. J. (2003). Dependent defaults in models of portfolio credit risk. *Journal of Risk*, *6*, 59-92.