The Value of Data Analytics as it is Applied to the GE Synchrony Credit Model

Carrie McDowell

Southern New Hampshire University

Author Note

Carrie McDowell, Student at the Southern New Hampshire University

This research was to complete an assignment for DAT 690, Dr. Ricky J. Sethi

Correspondence concerning this article should be addressed to Carrie McDowell, Student, Southern New Hampshire University, New Hampshire, 03106

Contact: 364mc@att.net

Abstract

This paper will draft an enterprise strategy for Synchrony that creates business value from existing data sources. Included is an executive presentation of the summary of the analytic plan, the results of the initial pilot and final analysis, and an explanation of the value this plan will add to the organization. CRISP-DM methodology was used to articulate the business problem and then RStudio to analyze the data and create a predictor model based on attributes identified as bad credit risk applicants.

*Keywords:* data analytics, CRISP-DM, credit score, RStudio, Rattle, prediction model

Contents

[**Introduction** 6](#_Toc21866885)

[**Organizational Background** 6](#_Toc21866886)

[**Problem Statement** 7](#_Toc21866887)

[**Project Purpose** 7](#_Toc21866888)

[**Stakeholders Needs** 7](#_Toc21866889)

[**The Need for Data Analytics** 8](#_Toc21866890)

[**Project Type** 9](#_Toc21866891)

[**Project Pilot Evaluation** 9](#_Toc21866892)

[**Analytic Tools Used** 10](#_Toc21866893)

[**Data Understanding** 10](#_Toc21866894)

[**Data Preparation** 11](#_Toc21866895)

[**Pilot Success** 12](#_Toc21866896)

[**Pilot Failure** 13](#_Toc21866897)

[**Return on Investment** 13](#_Toc21866898)

[**Pilot Modification** 14](#_Toc21866899)

[**Project Plan Implementation** 14](#_Toc21866900)

[**Exploring Data** 14](#_Toc21866901)

[**Summarizing the Data** 15](#_Toc21866902)

[**Visual Distributions** 17](#_Toc21866903)

[**Correlation / Regression Analysis** 18](#_Toc21866904)

[**Transforming Data** 20](#_Toc21866905)

[**Building the Model – Project Plan Implementation** 29](#_Toc21866906)

[**Model Framework** 29](#_Toc21866907)

[**Data Mining** 29](#_Toc21866908)

[**Descriptive Analysis** 29](#_Toc21866909)

[**Predictive Analytics** 30](#_Toc21866910)

[**Variable Analysis** 30](#_Toc21866911)

[**Cluster and Association Analysis** 30](#_Toc21866912)

[**Model Selection and Creation** 32](#_Toc21866913)

[**Model Selection** 33](#_Toc21866914)

[**Model Performance** 36](#_Toc21866915)

[**Measures of Performance** 36](#_Toc21866916)

[**ROC chart.** 38](#_Toc21866917)

[**Scoring** 38](#_Toc21866918)

[**Project Plan Conclusion** 41](#_Toc21866919)

[**Recommendations** 42](#_Toc21866920)

[**Appendix A** 46](#_Toc21866921)

[**Appendix B** 47](#_Toc21866922)

[**Appendix C** 48](#_Toc21866923)

[**Appendix D** 49](#_Toc21866924)

[**Appendix E** 50](#_Toc21866925)

[**Appendix F** 51](#_Toc21866926)

[**Discussion A** 52](#_Toc21866927)

[**Discussion B** 57](#_Toc21866928)

[**Discussion C** 59](#_Toc21866929)

[**Capstone Component Two: Personal and Professional Reflection** 62](#_Toc21866930)

# **Introduction**

According to Amadeo, the financial crises of 2008 were the worst since the great depression of 1929. The stock market fell 777 pts in one day, banks and large businesses declared bankruptcy, unemployment rose to more than 9%, housing prices feel almost 32%, and many defaulted on loans which only made matters worse. The nation was slow to recover (McDowell, 2019).

The financial crises spawned this project which questioned whether a predictor model can be created to identify a customer as a good or bad credit risk at the application stage. This project began in August of 2019. This report is a precursor and includes visual and text-based charts and reports based on findings from analyzing heuristic credit data (Appendix A). Included in this report is a proposed timeline and cost associated with building and implement a predictor model (see Appendix C and D), as well as discussions on security and ethical issues surrounding the collection, storage, and use of customer data (see Discussions A, B, and C). In addition, I have included a stakeholder’s registry and the proposed project team (see Appendix E and F).

# **Organizational Background**

General Electric Company (GE), incorporated on April 15, 1892. Currently GE is a global company with products and services ranging from aircraft engines, power generation, and oil and gas production equipment to medical imaging, financing and industrial products and has served customers in 180 countries (General Electric Co (GE.N) Company Profile).

For the purpose of this project, data was extracted from one thousand Synchrony’s retail financial customers and was analyzed. Synchrony customizes financing programs across retail, health, auto, travel and home, and consumer banking. Synchrony has more than 80 million accounts which equated to $140 billion in sales financed ("Synchrony").

# **Problem Statement**

In response to the financial crisis of 2008–2009, I was tasked to see if it is profitable to build a model that will accurately predict if a new customer presents a bad credit risk. The model must predict a credit risk without bias. Using the predictor model will significantly reduce the rate of defaults, reduce costs, and increase profit.

While loan default has decreased since 2009, losses from default totaled $4,692,000 in 2018; up $626,000 from the previous year. Creating a predictor model to identify a high-risk customer before extending credit will significantly reduce the rate of defaults, increase profit, and cut down on human error which will also remove the company from claims of discrimination. While there will still be gray areas, most applicants can be vetted accurately using a predictor model without human intervention.

# **Project Purpose**

## **Stakeholders Needs**

Stakeholders of Synchrony will be of high-level company executives, stock holders, all employees, any company Synchrony does business with, and customers. Knowing each stakeholder’s need allows the analyst to tailor real-time reports to them.

Speaking solely of employees of Synchrony, it’s important to understand each Stakeholders interest and influence early in the project and to share the data to be used in the model to get buy-in. They must agree on the dataset, a data dictionary, and how the data will be applied to the predictor model. As a data analyst, it is important to understand each stakeholder’s expectation of not only the outcome of the model but also the project process. Additionally, project crepe can become a problem if data issues are not resolved before the beginning of the project. While not popular, blockers must either be removed from the project or made to comply (McDowell, 2019). Appendix D outlines the project timeline and major milestones. At each of these milestones, issues will be address and stakeholders will determine if the project should proceed or be scrapped and sign off accordingly.

## **The Need for Data Analytics**

Data analytics will compile and analyze data and then return results which will give our company leaders an edge to better manage bad credit risks customers and the loss of resources caused by them. Until now, without data analytics, the agents have not been able to identify who will be a good or bad credit risk. Therefore, the answer to the problem is evident; some process involving vetting customers must begin.

The purpose of first researching a problem is two-fold. It allows the analyst the opportunity to test theories using the data available, identify missing or invalid data, and then further defining the original hypothesis or change it altogether (Shuttleworth, 2008). In addition, it is possible that research has already been conducted and can be duplicated. There is no reason to spend time and money where it is not necessary.

In the case of Synchrony’s high default, it is very possible to use data to research and identify which offices and even agents have a higher rate of default. The company many wish to invest in training, or make other decisions based on sound data. More often than not, research will dictate that the environment changes and therefore business process must be modified to fit the new risks facing Synchrony.

Data analytic tools such as RStudio, uncover patterns in large datasets quickly that cannot be detected by the human eye or found in a timely manner. RStudio’s suite of tools in Rattle include cluster, association, and linear analysis not to mention the modeling tools such as a basic decision tree, Random Forests, and more. RStudio and Rattle are open software and free to use and share. You can access and download your version here: <http://cran.rstudio.com>.

# **Project Type**

For this project I used the descriptive data provided (see Appendix A) to create the predictive model. This data taken from 1,000 existing and past customers is the base on which the prediction model is made as the data reveals patterns that will accurately identify a bad credit risk customer. According to Shearer, using variables such as income, home ownership, and others against behavioral variables such as payment and credit history will give the best result (McDowell, 2019).

# **Project Pilot Evaluation**

During the pilot phase, I took a closer look at the data captured. I noticed data missing that would be invaluable moving forward. I also consulted with stakeholders and managers to discuss the ongoing security of customer data as well as ethical standards in the use of customer data. It was decided at that time that variables that could identify a person’s sex, age, or race will not be used in the final model. Additional data taken from new customers will be utilized in a modified model in the future. The cost to modify the model will be minimal as the research is already done. The new variables that will be captured are: ADDRESS, PHONE, AREA CODE, LAST\_NAME, ZIP\_CODE, and IP\_ADDRESS.

Following the CRISP-DM methodology and utilizing RStudio and RStudio GUI Rattle, I began to analyze the data. The CRISP-DM methodology has six important parts. The process includes: 1) business understanding, 2) data understanding, 3) data preparation, 4) modeling, 5) evaluation, and 6) deployment. Data preparation and data understanding are the most time consuming of all the steps and they are revisited often throughout the project.

## **Analytic Tools Used**

Based on the business objective, RStudio is the analytic tool I used to manipulate the data and build the predictor model. I chose RStudio because it has a graphical interface called Rattle which makes it easy for non-technical users to analyze data. Not only does RStudio provide robust data mining tools, it also provides detailed reports. The ease and power of the R packages make this tool a good choice for GE as it will require the least training and provide real reports faster (McDowell, 2019). Lastly, the tool is free to use and there is a plethora of information and training material found on the web.

## **Data Understanding**

Data understanding has four phases which include collecting then describing the data; and exploring and verifying data quality. This includes cleaning the data and then analyzing for any initial insights (McDowell, 2019). I identified a few variables with naming convention issues that prevented the data from being used in the model. Initially, my insights were based on what I understood on a personal level relating to commercial credit such as a person’s credit score, income, job stability, etc. As it turned out, through further data understanding and preparation, I found that I was wrong to “guess” on what I thought was the right algorithm to model.

## **Data Preparation**

During this phase selecting the right data was key to the success of the model. It was necessary to transform some variables and construct others based on how important a variable was to those identified with customer default (McDowell, 2019). Also, as agreed, I omitted the following variables from the dataset provided: AGE, FOREIGN, MALE\_DIV, MALE\_SINGLE, and MALE\_MAR\_WID, and then began to work with the rest of the dataset.

Using Rattle, I exposed the levels of each variable, and then with the help of a Random Forest model, this final dataset was identified as the most important variables to use to target a bad risk customer. A Random Forest model operates by constructing multiple of decision trees at training time and outputting the results as a mean prediction of the individual trees

|  |  |
| --- | --- |
| Variable Name | Description |
| CHK\_ACCT\_.0.0. | Checking account balance = $0 |
| CHK\_ACCT\_.0.1. | Checking account balance less than $200 |
| CHK\_ACCT\_.2.3. | Checking account balance greater than $200 |
| HISTORY\_.0.0. (Credit) | No credit history |
| HISTORY\_.0.1. | Good credit history |
| HISTORY\_.3.4. | Poor credit history |
| NEW\_CAR\_.0.0. | Loan default = False |
| NEW\_CAR\_.0.1. | Loan default = True |
| GUARATOR\_0.0. | Loan default with a guarantor = False |
| GUARANTOR\_0.1. | Loan default with a guarantor = True |
| DURATION | 4 to 72 months |
| AMOUNT | $250.00 to $18,424.00 |
| REAL\_ESTATE\_0.0. | No real estate owned |
| SAV\_ACCT\_.3.4. | No savings account |

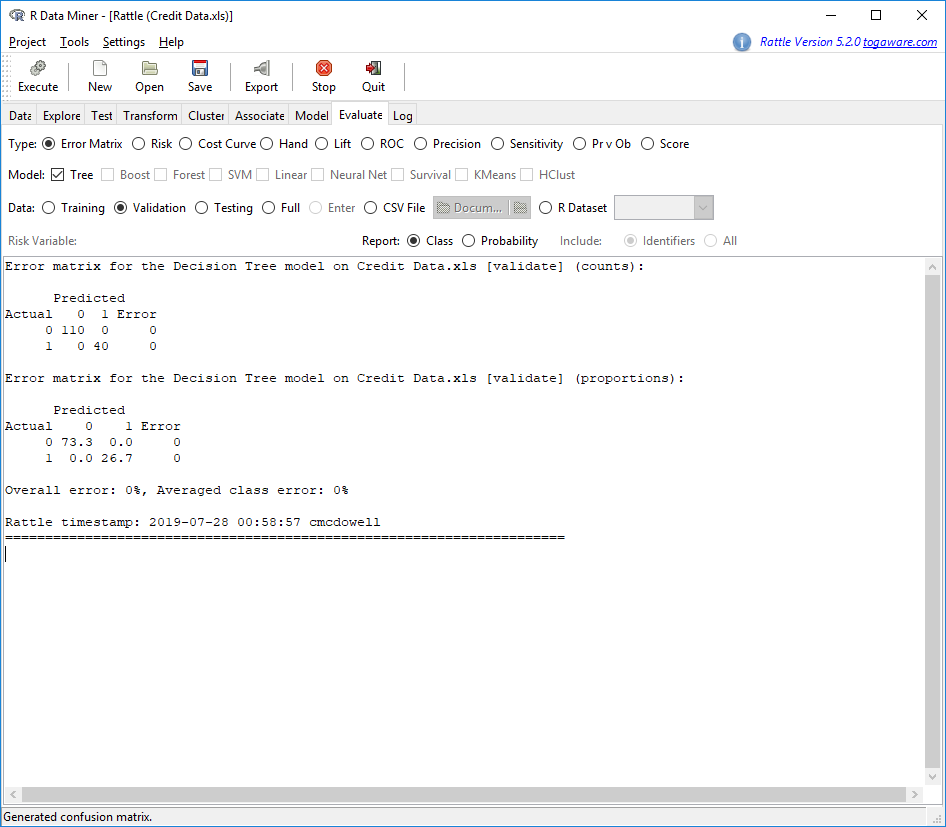
## **Pilot Success**

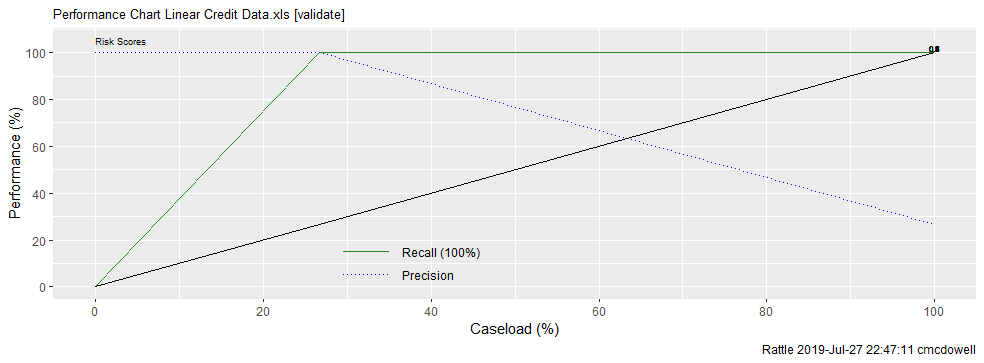
Overall the model performed well and will meet the needs of the organization which was to build a predictor model that will determine if an application presents a bad credit risk.

The initial pilot run successfully identified variables associated to customers with high-risk tendencies. The predictor model did show positive results when variable levels were exposed and only the top fifteen variables were used including the target variable DEFAULT\_0.1. which identifies as DEFAULT = Yes.

I evaluated the Random Forest model to discover the following:

1. The Error Matrix chart showed an overall error of 0% and class error of 0%



1. The performance chart also shows a 100% match.

What does this mean? This means that the predictor model I built will be accurate 100% of the time.

## **Pilot Failure**

It is clear that the pilot needed modification. 100% accuracy tells me that I over fitted the data to the model. The challenge was finding that happy medium. I should have taken more time to analyze the data using RStudio and the tools in Rattle to find an actual good model fit. By honing in on only the few variables identified as important, I removed other variables that were apparently significant. Typically, before scrubbing and analyzing data I would define each variable and create a data dictionary. Taking variables at face value may prove to be an issue. It is important that the model return accurate results so taking my time was imperative.

In order to create an accurate model, fit for production, I modified the plot by starting again at the beginning and taking my time and to be sure to go through the steps outlined in the CRISP-DM structure. I concluded that 60% accuracy was ideal.

# **Return on Investment**

Before continuing on, I wanted to be sure that the cost of this project was worth the investment. The net charge-off for 2018 totaled $4,692,000.00 (Bishop, 2019). While there are always extenuating circumstances, identifying 60% of high-risk customers may have saved the company $2,815,200. The cost to implement this project is $632,570 and will take six months to complete (see Appendix C). All things being equal, implementation of this predictor model will save the company $2,032,630 in the first year. Considering ongoing efforts to maintain the model and secure customer data will cost approximately $150,000 a year. A five-year return on investment is $12,693,430.

Additionally, I believe that other models working in conjunction with this first predictor model will increase savings even more. These additional models will look at agent behavior which may result in additional training and track in-depth details about applicants.

# **Pilot Modification**

The original pilot plan was accurate at 100%. I felt that it would cause too many false positives, lower profit, and cause agents extra work. I started back from scratch, this time not getting so granular with the data. In the last pilot, I committed to digging deep into the data by inspecting variable significance, combining variables and renaming some to better explain the variable’s use. To find the actual important variables, I ran a linear regression on the original dataset using the variable DEFAULT\_0.1 as the target. RStudio did all of the heavy lifting by considering all the ways that variables are dependent of each other and chose only those making a difference on the outcome of the model. That way, when a customer or agent enters data into the graphical user interface (aka, online application), the system will look at the algorithm of most important variables first and in the end return an approval or credit denied message.

# **Project Plan Implementation**

## **Exploring Data**

Working with the dataset Credit\_Data\_v3 I began to explore the dataset. I deleted the data that could be construed as discriminatory and renamed other variables so that Rattle will recognize and use the data contained within the variables moving forward. For instance, I changed the variable name RADIO/TV to RADIO\_TV, CO-APPLICANT to CO\_APPLICANT and OBS# to OBS. This standardized the names to match the naming convention of the other variables.

## **Summarizing the Data**

Using RStudio, I opened the tool set Rattle and then the dataset Credit\_Data\_v3. Explanation for the code below begins with a hashtag (#)

# Loading the toolset Rattle into RStudio

> library(Rattle)

> Rattle()

# Loading the dataset Credit\_Data\_v3 into RStudio

> > library(readxl)

> Credit\_Data\_v3 <- read\_excel("Capstone/Credit\_Data\_v3.xls")

> View(Credit\_Data\_v3)

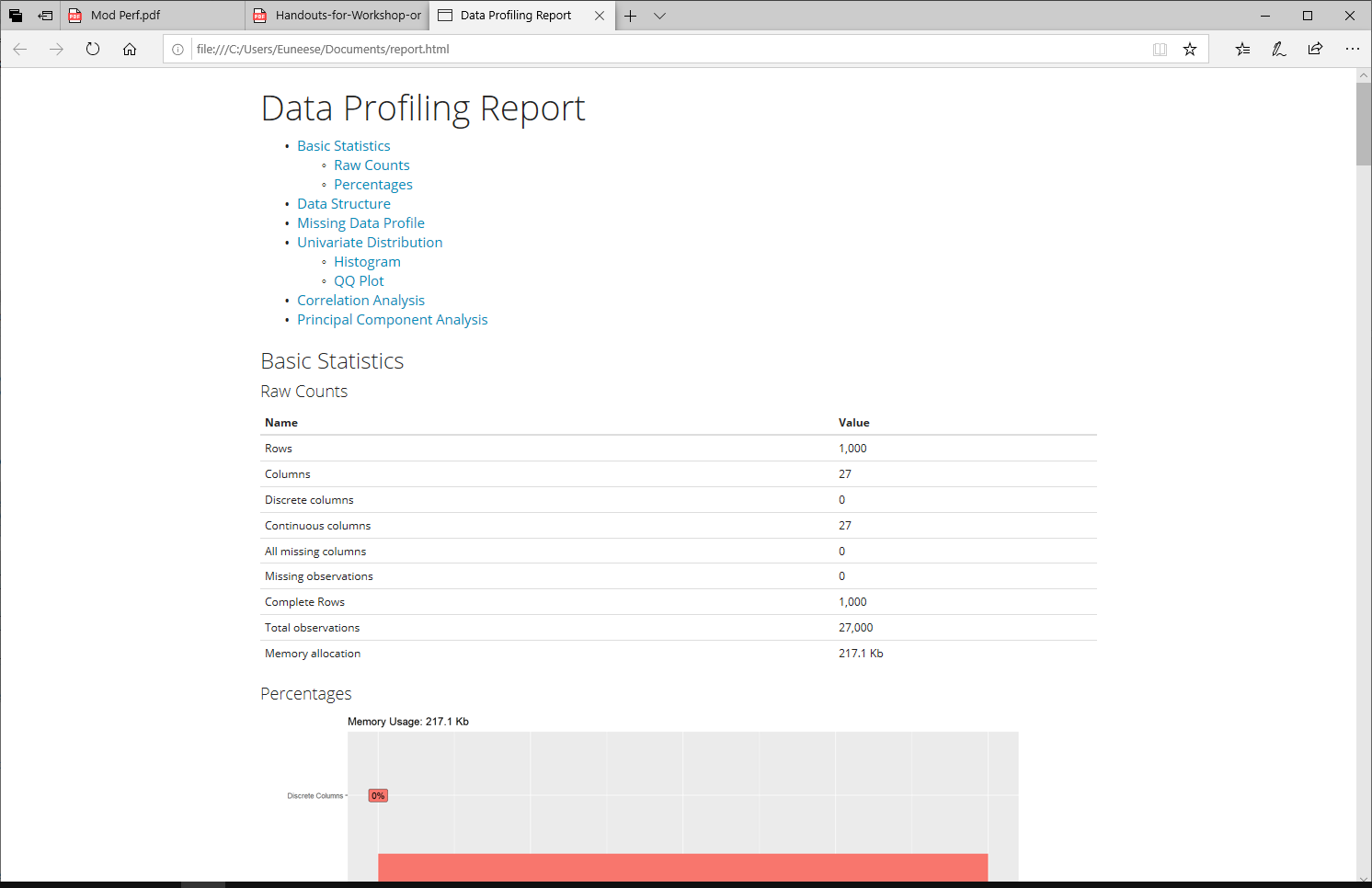
In RStudio I ran a simple report that explores the data and returns results.

# Run a simple report to explore the data in the dataset Credit\_Data\_v3. I first needed to download the library DataExplorer.

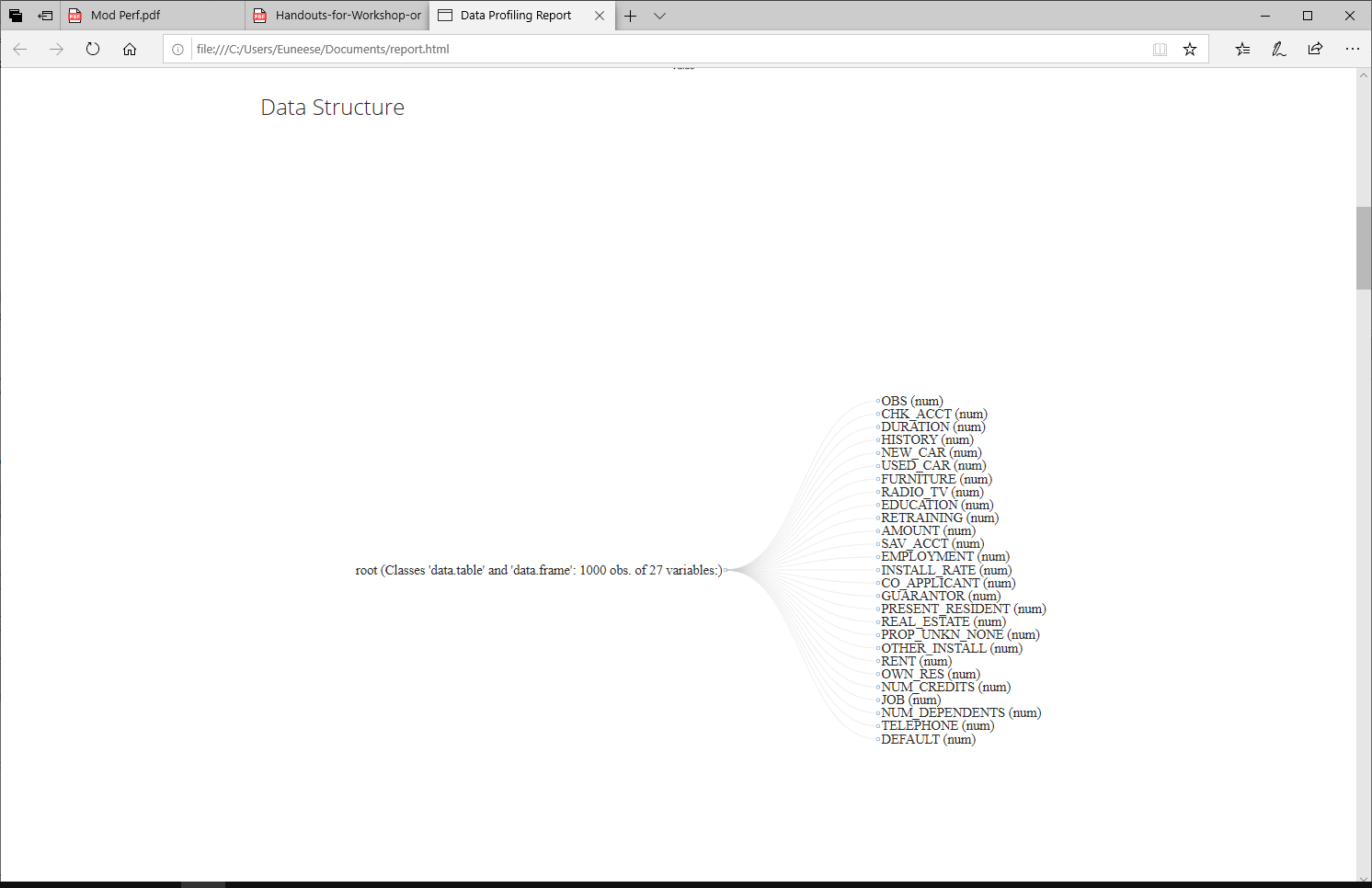
> library(DataExplorer)

> DataExplorer::create\_report(Credit\_Data\_v3)

The basic statistics from the report shows that the data set contains 27,000 observations and 27 columns (variables), with no missing data.



The data structure of the dataset Credit\_Data\_v3 is as follows:



Data Structure

Dataset Credit\_Data\_v3

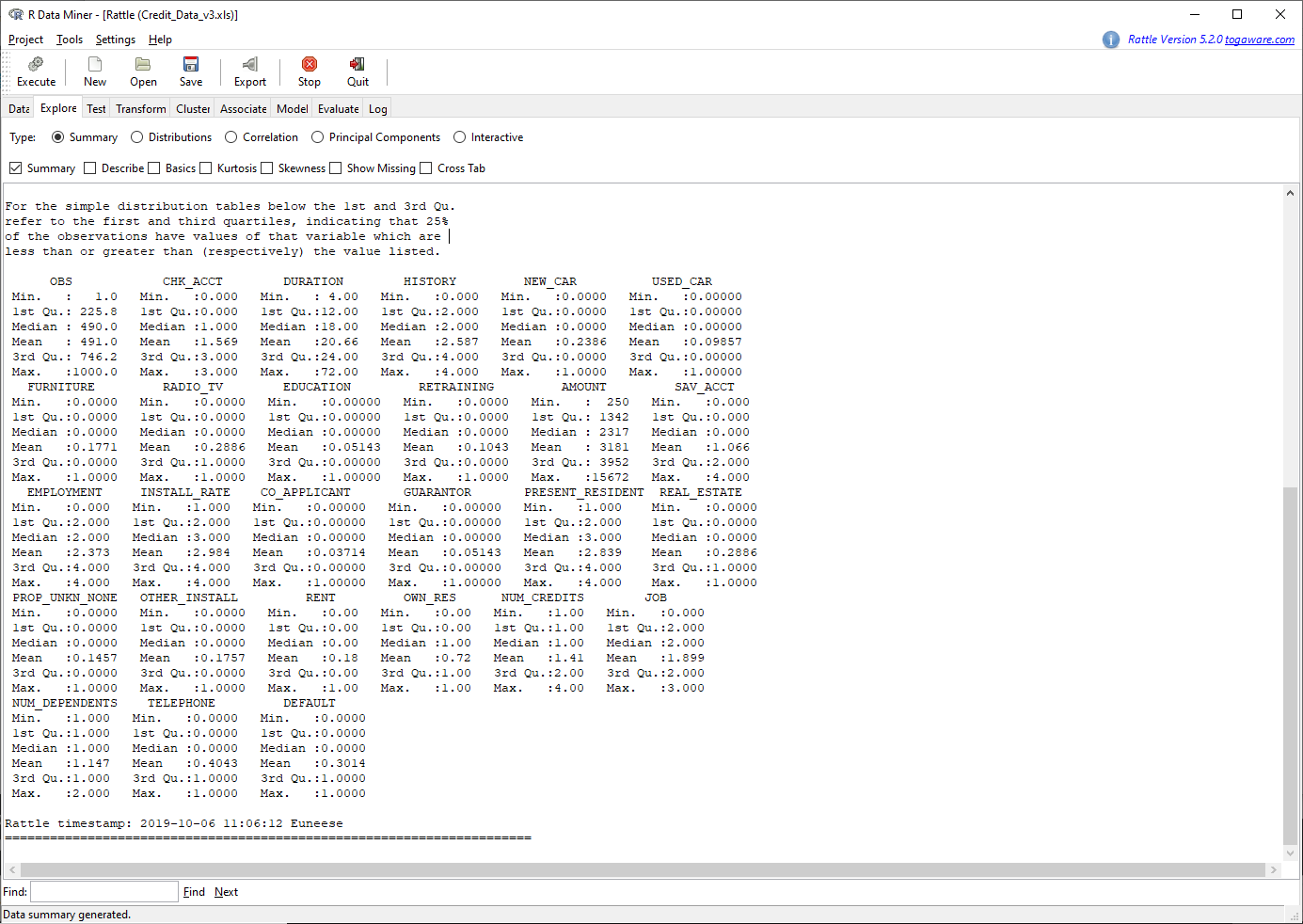
Next from the same report, I looked at the correlation between all variables with no target variable assigned. The greater the correlation, the brighter red the square.



Looking at the correlation analysis above, I saw that there was a strong association between the loan AMOUNT and DURATION, as well as NUM\_CREDITS AND HISTORY. Two variables with a strong association was not enough to build a model on, and I did not assign a target variable so I decided to use RStudio to run a linear regression model to identify the proper variables to use in the model but first I want to look closer at the data.

## **Visual Distributions**

Rattle makes it easy to analyze the data. Under the explore tab in Rattle, I ran a summary report to see the distributions of the data in the 1st, and 3rd quartile for each variable. As you see, the report also outlines the minimum, median, mean and maximum of each variable as well.



## **Correlation / Regression Analysis**

To better identify which variables to use to catch a bad credit risk customer, I needed to target observations (customers) from the data set that have defaulted and then look at what variables are associated with default. Meaning what attributes do these customers have in common, (i.e. checking account balance, the number of dependents, if they own a home, etc).

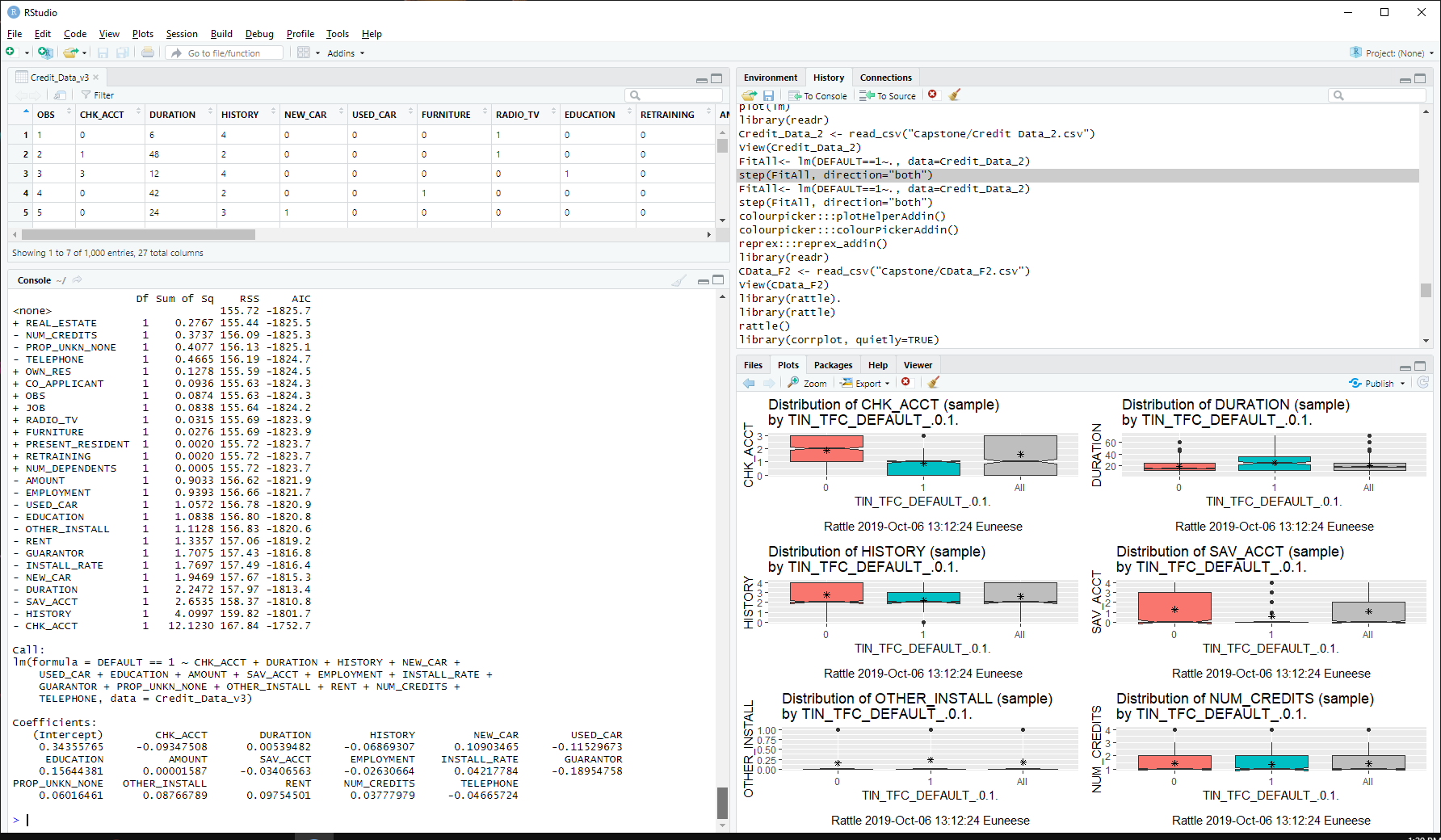
I ran a regression model on the Credit\_Data\_v3 dataset using DEFAULT==1, which is DEAFAULT.0.1. as the target. Remember DEFAULT.0.1 means that the customer did indeed default. The regression model looks at all variables and how they are associated to one another and how these associations work against the target variable. In the end, the model told me the best variables to use to create the prediction model.

#Fit the data to use in the regression model and set DEFAULT==1 as the target

>FitAll<- lm(DEFAULT==1~., data=Credit\_Data\_v3)

# Told the model to check forward and backward for a good fit to the model.

> step(FitAll, direction="both")

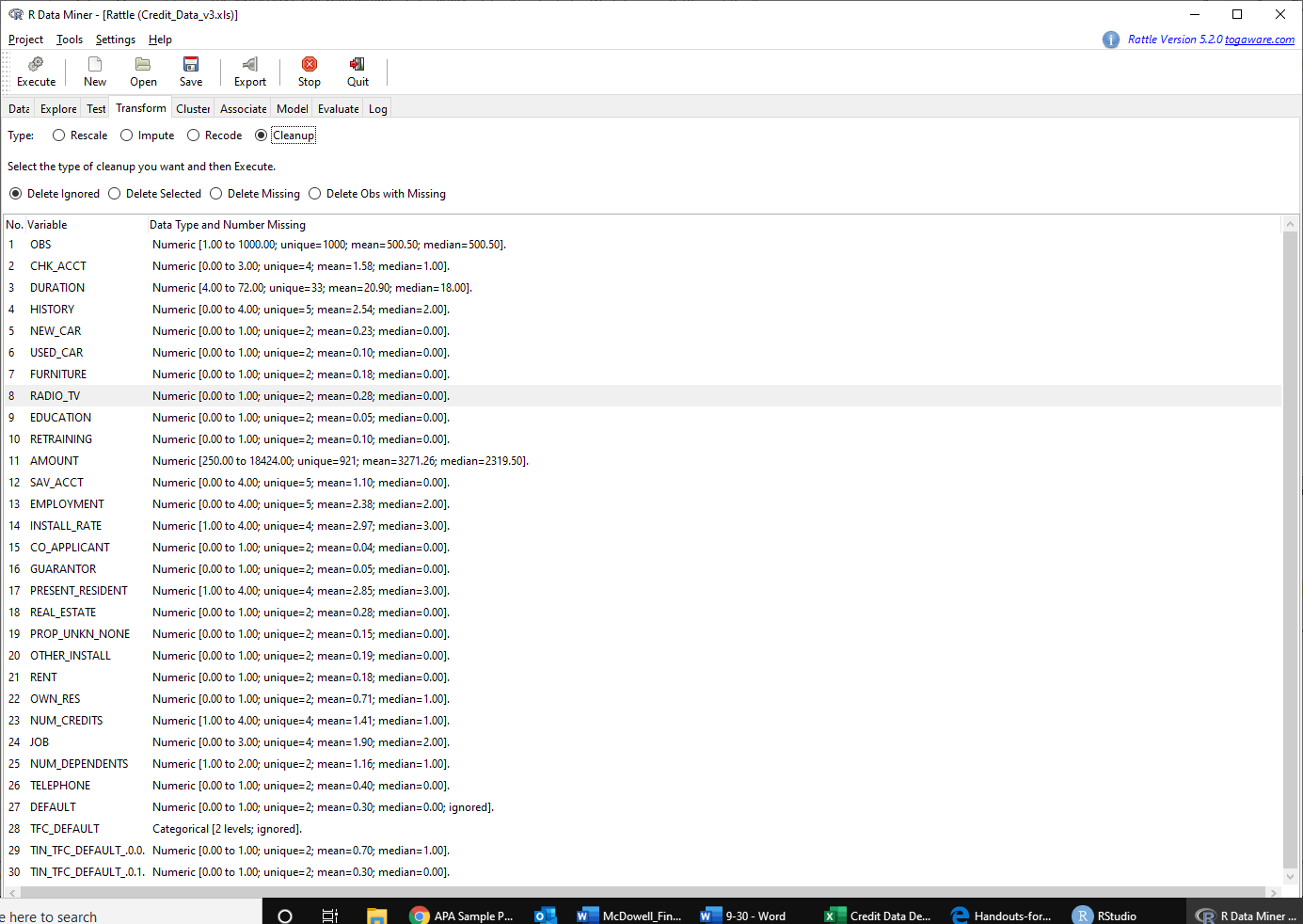
 The regression model proposed the following variables for the prediction model:

Due to this result, and the decision to omit other variables, these variables that have little association to default have been omitted:

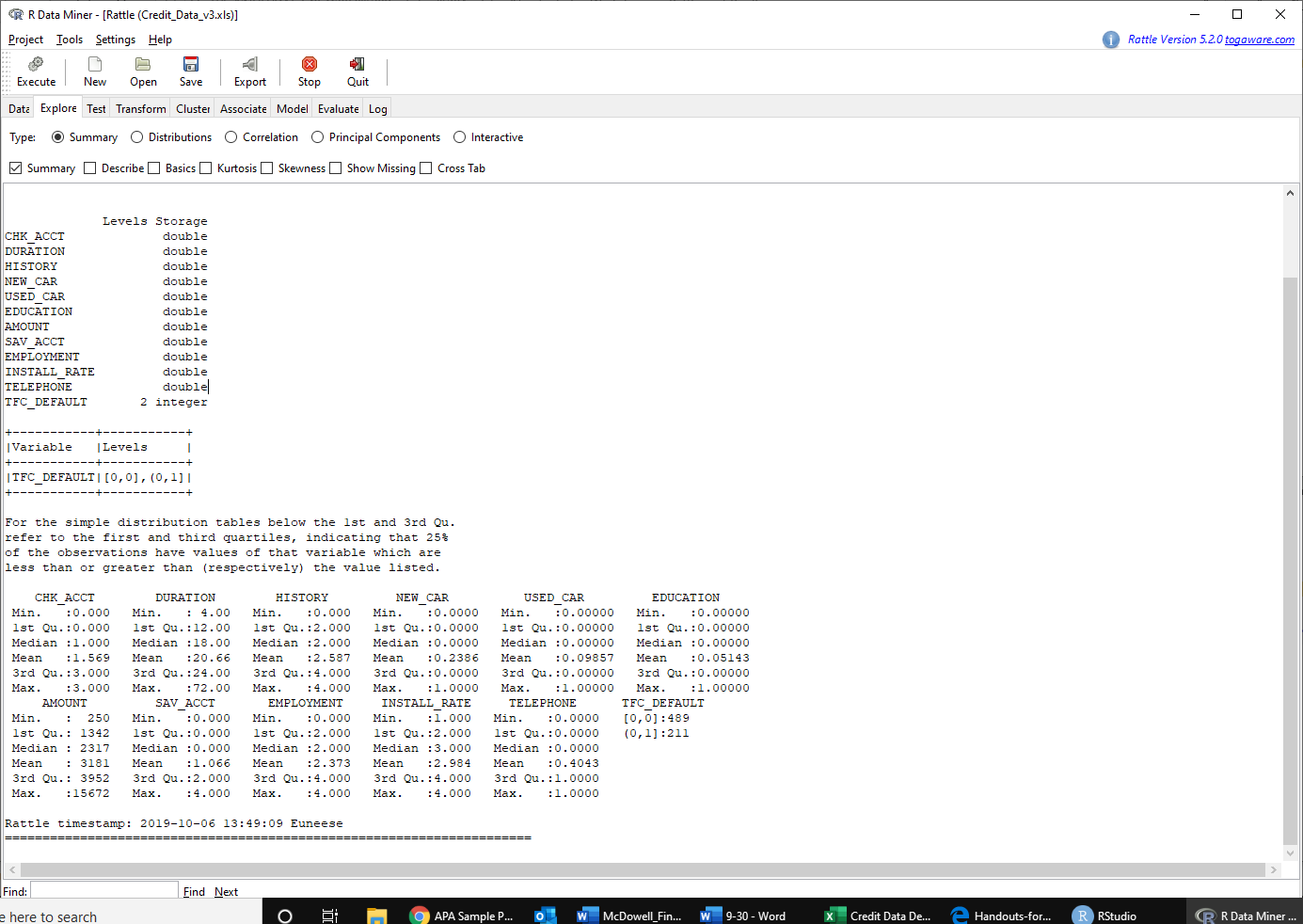
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Name** | **Variable Name** | **Variable Name** | **Variable Name** |
| FURNITURE | PRESENT\_RESIDENT | OWN\_RES | CO-APPLICANT | NUM\_DEPENDENTS |
| RADIO/TV | REAL\_ESTATE | NUM\_CREDITS | GUARANTOR |  |
| RETRAINING | PROP\_UNKN\_NONE | JOB | OTHER\_INSTALL |  |

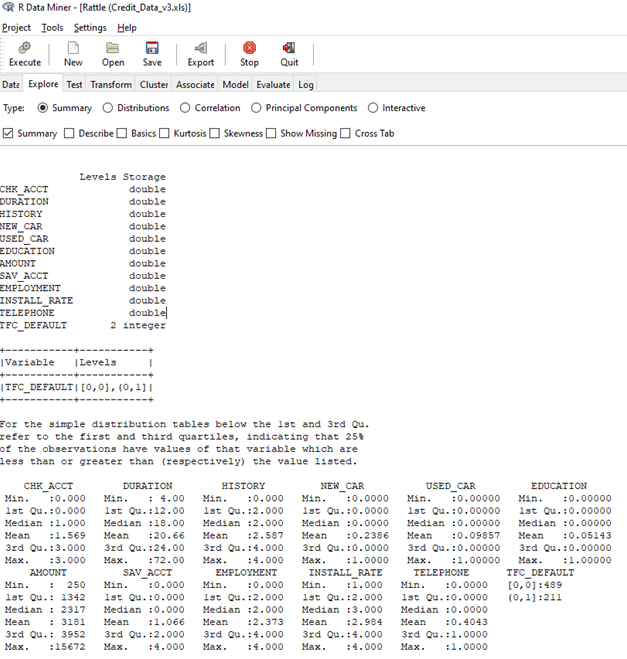
Back to Rattle to delete the variables that I will not use.

1. On the data tab, I set the variables above to ignore then selected execute.
2. Under the transform tab, I selected cleanup and then Delete Ignored



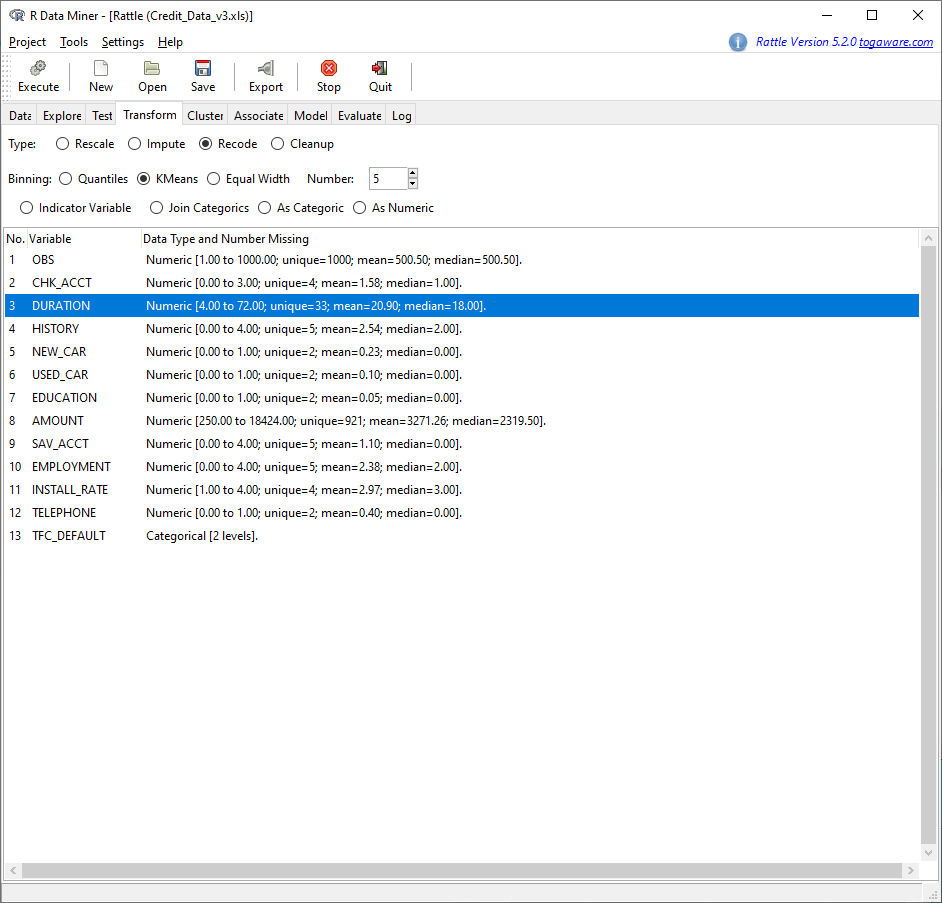
The steps above deleted the variables, leaving only 13 variables left in the dataset.



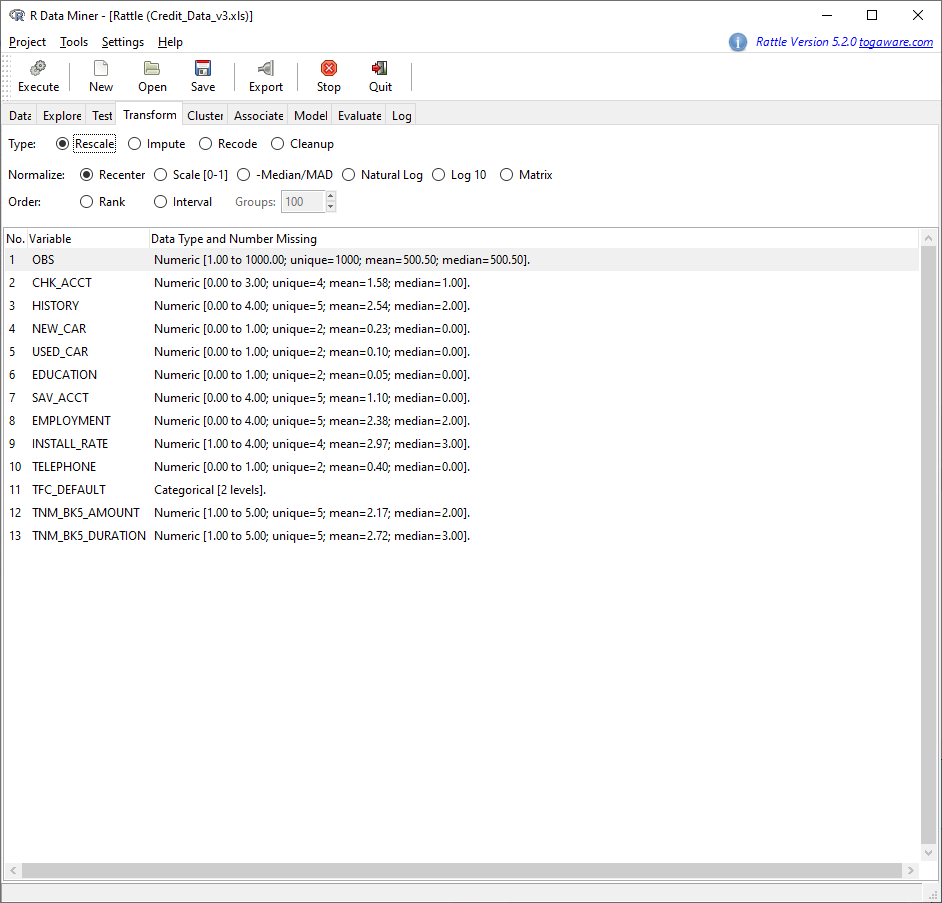


# **Transforming Data**

Using the explore tool in Rattle, I looked deeper into the correlation of variable levels using default yes and no as the target. First, I rescaled the sizes of the variables DURATION and AMOUNT to 5 variable levels using the Recode option under the Transform tab in Rattle. I used the KMeans to place the variables into 5 bins.



After omitting old variables of DURATION and AMOUNT, and transforming the newly created ones to As.Numeric, I will be working with this dataset:



Using the Distributions type under the Explore tab, I opened each variable to reveal the observations of Default yes (0,1) vs Default no (0,0).

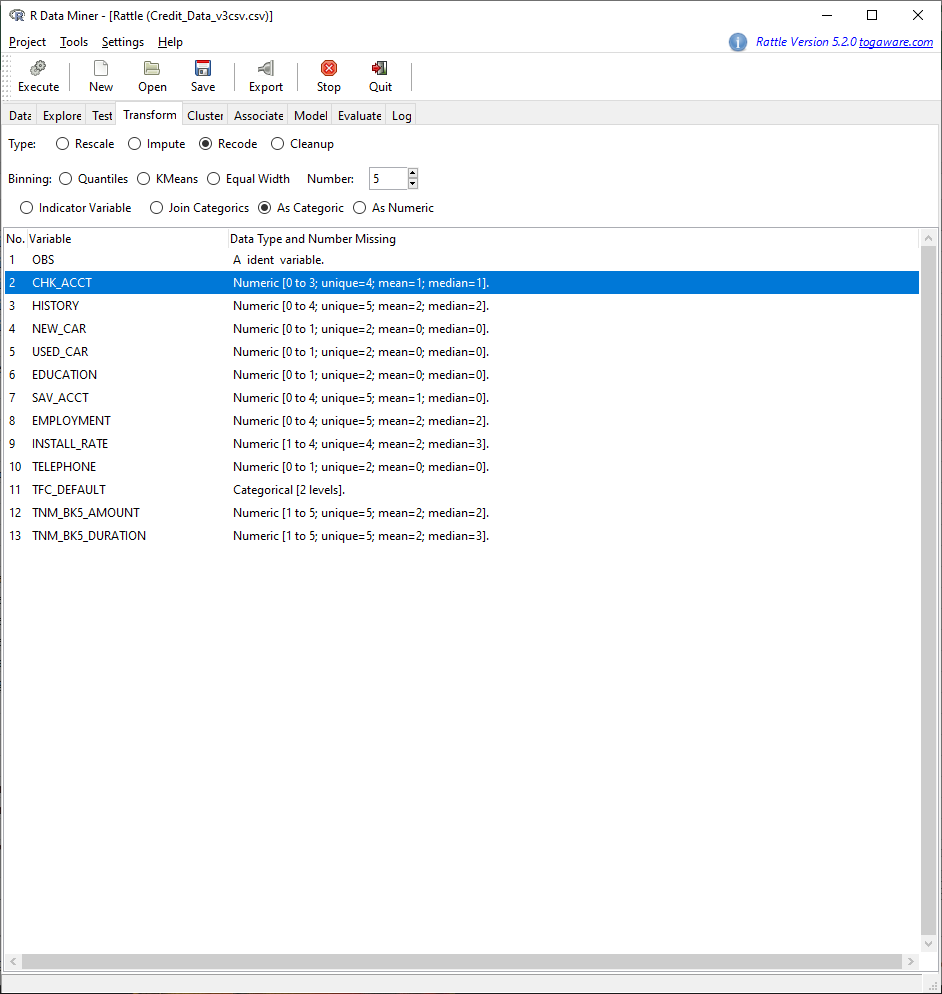
|  |  |  |
| --- | --- | --- |
|  | | Levels of the variable CHK\_ACCT:  0 = Account empty  1 = Less than $200  2 = Greater than $200  3 = No checking account  It appears that those having less than 200 are susceptible to default. |
|  | Levels of the variable SAV\_ACCT:  0 = Less than $100  1 = $100 - $500  2 = $500 - $1000  3 = Greater than $1000  4 = No savings account  It appears that those having less than $100 are susceptible to default, however there are many outliers in the data. | | |

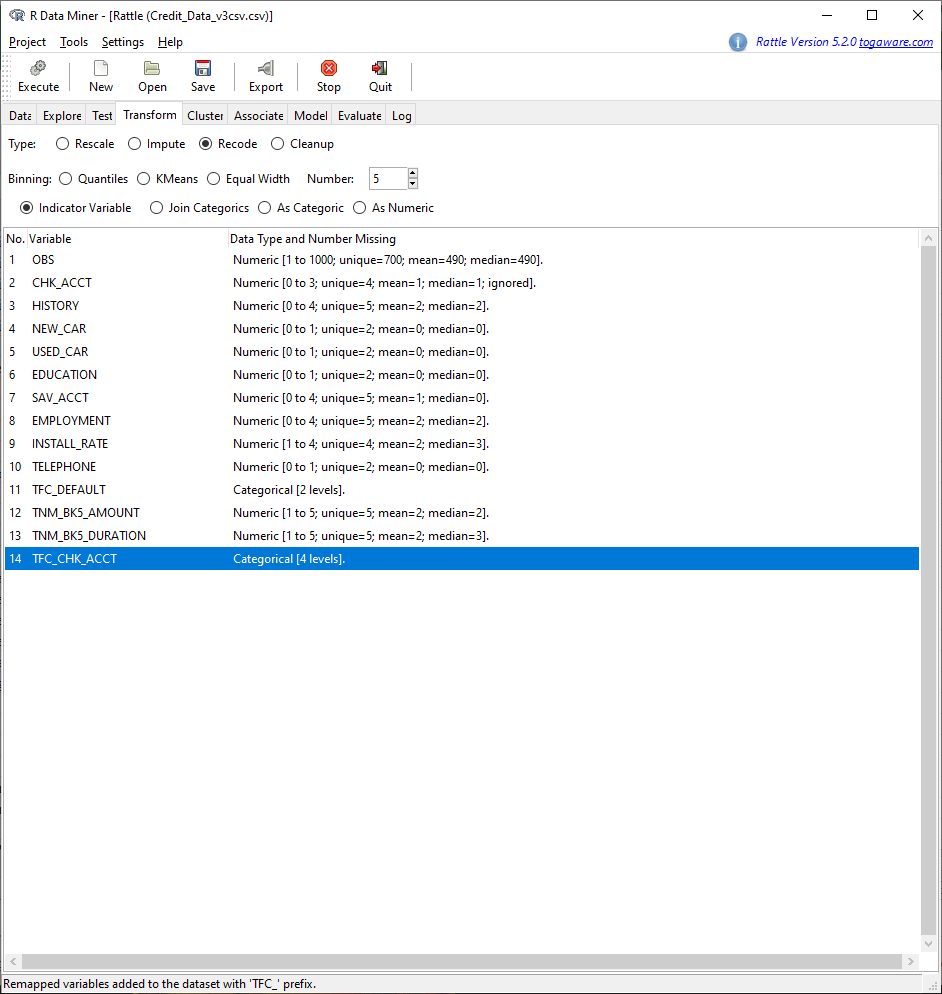
|  |  |  |
| --- | --- | --- |
|  | | Levels of the variable HISTORY:  0 = No credit history  1 = Good credit  2 = Good credit  3 = Poor credit  4 = Deny service  It appears that most of the credit risk customers have both good and poor credit. |
|  | Levels of the variable EMPLOYMENT:  0 = Unemployed  1 = Employed less than 1 year  2 = Employed 1 to 4 years  3 = Employed 4 – 7 years  4 = Employed more than 7 years.  It seems that most customers who defaulted were employed between 1 and 7 years. | |

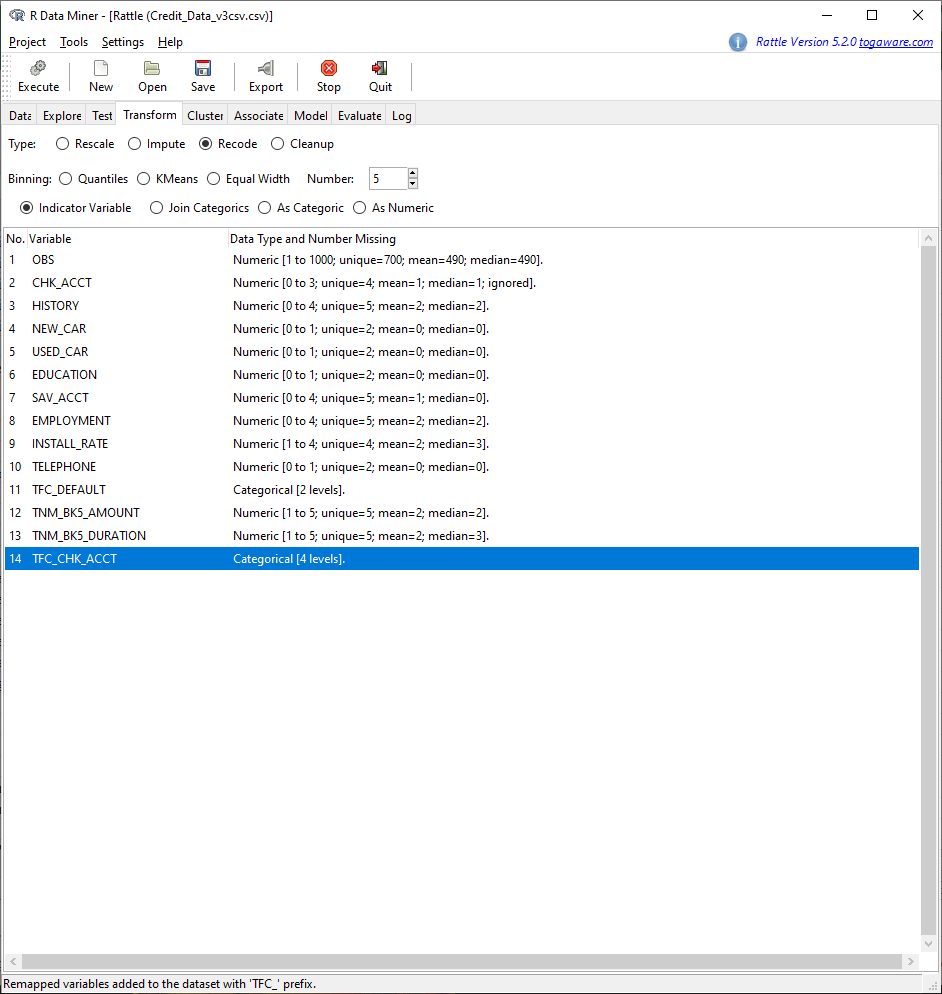
|  |  |  |
| --- | --- | --- |
|  | | Levels of the variable AMOUNT:  It does appear that those borrowing smaller amounts have a higher rate of default. |
|  | Levels of the variable DURATION:  It does appear that the loan duration for those susceptible to default is mid-range. | |

|  |  |
| --- | --- |
|  | Levels of the variable INSTALL\_RATE:  1 = 1% of disposable income  2 = % of disposable income  3 = 3% of disposable income  4 = 4% of disposable income  Most defaulted at 3%, there are some at 2% and 4% as well. |

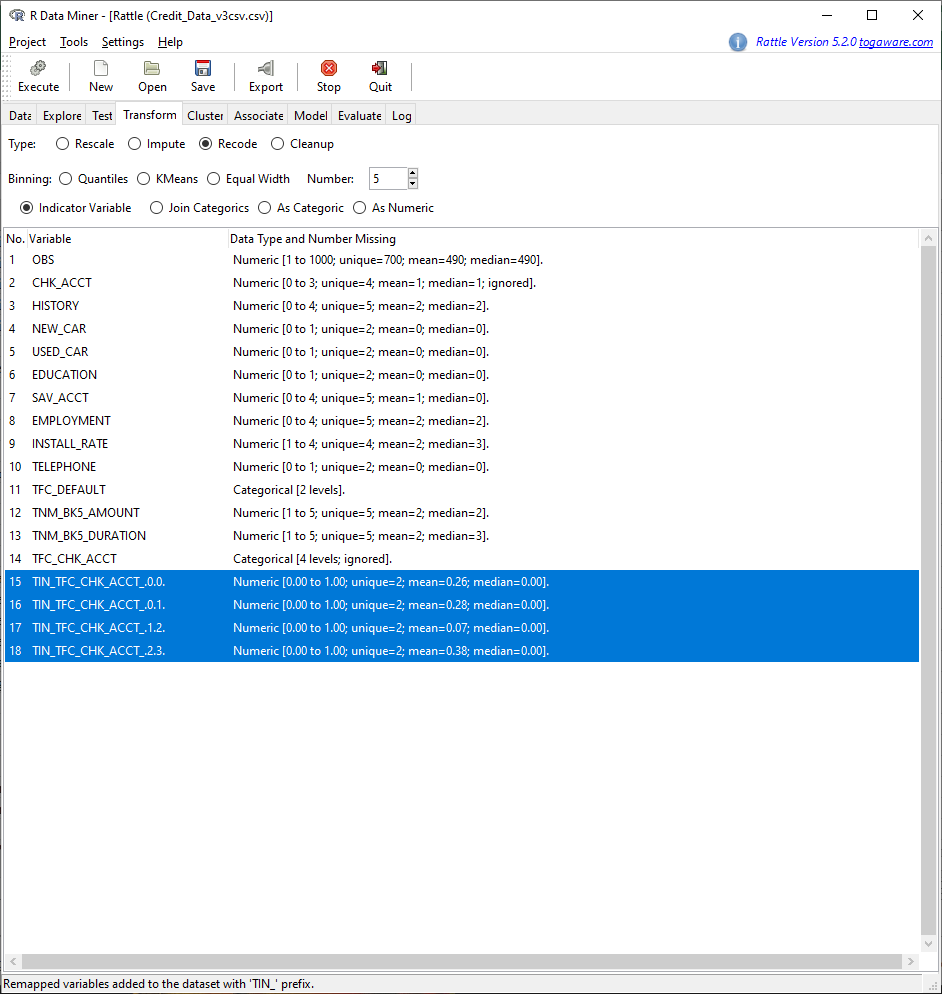
Once I opened the levels of the variables above, I could perceive data that can be used to build thresholds into the algorithms that can be used to score data read from the user interface. It is possible to weed out bad credit risk customers at the onset, at the point of application.

Since the last time, I overfit the data to the model, I didn’t want to do that again but I did open the levels of each variable to be read into the model. To do this, I first had to transform all of the variables from numerical to categorical using Rattle.

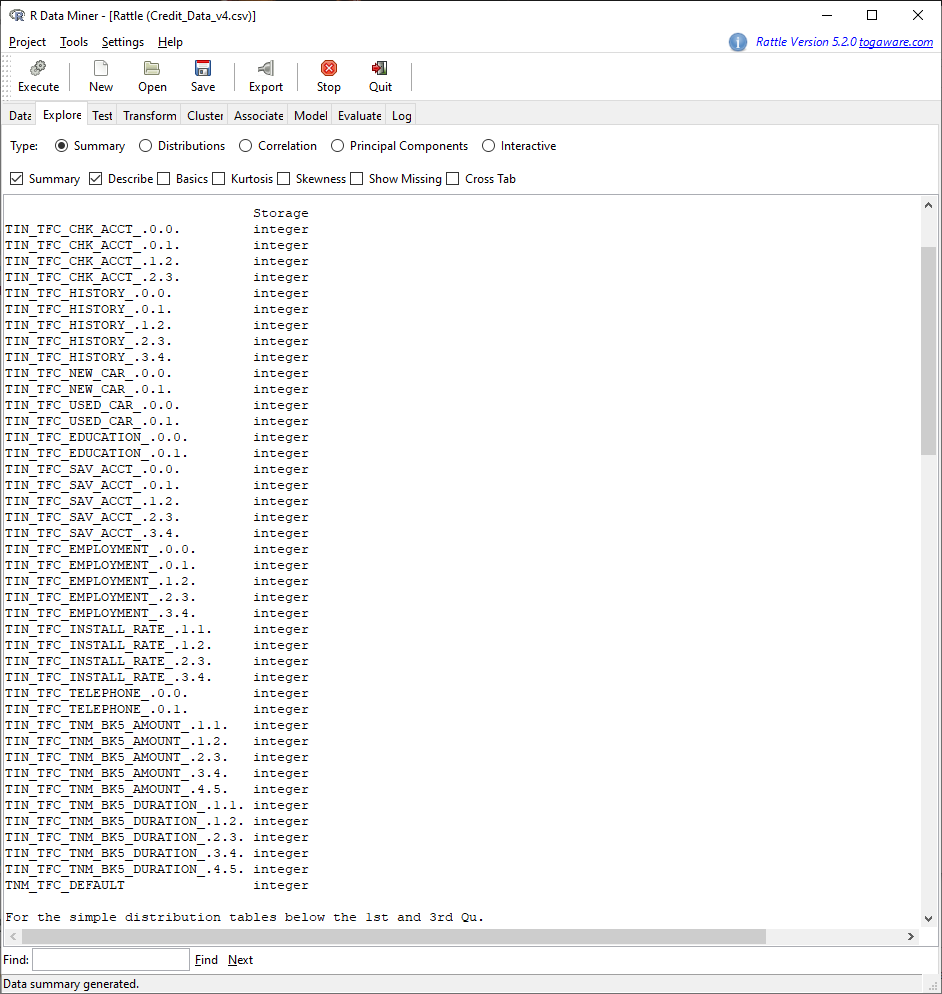
Next, I changed the new variables to an indicator variable.



You will see the 4 levels of CHK\_ACCT now open. Rattle will atomically set the old CHK\_ACCT variables to ignore so that they can easily be deleted using the Cleanup tool as we’ve done before.



Once I created the new variable levels, I saved the data with a new file name of Credit\_Data\_v4 and then continued on. This final data set can be used to compare/contrast other variable types to answer other business questions such as; of those with good credit how many have been employed more than 7 years. Here is the final dataset I used to create the model:

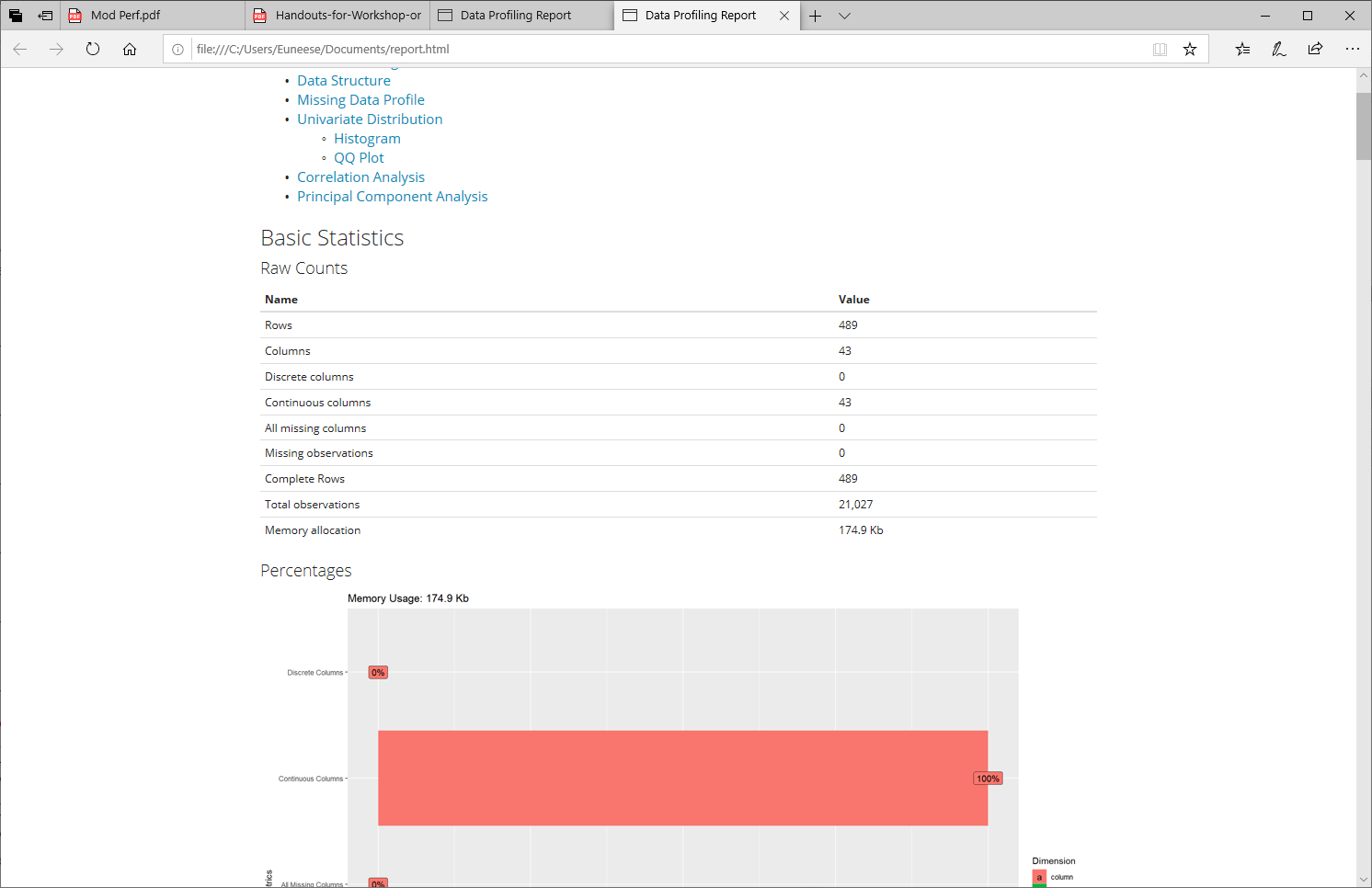


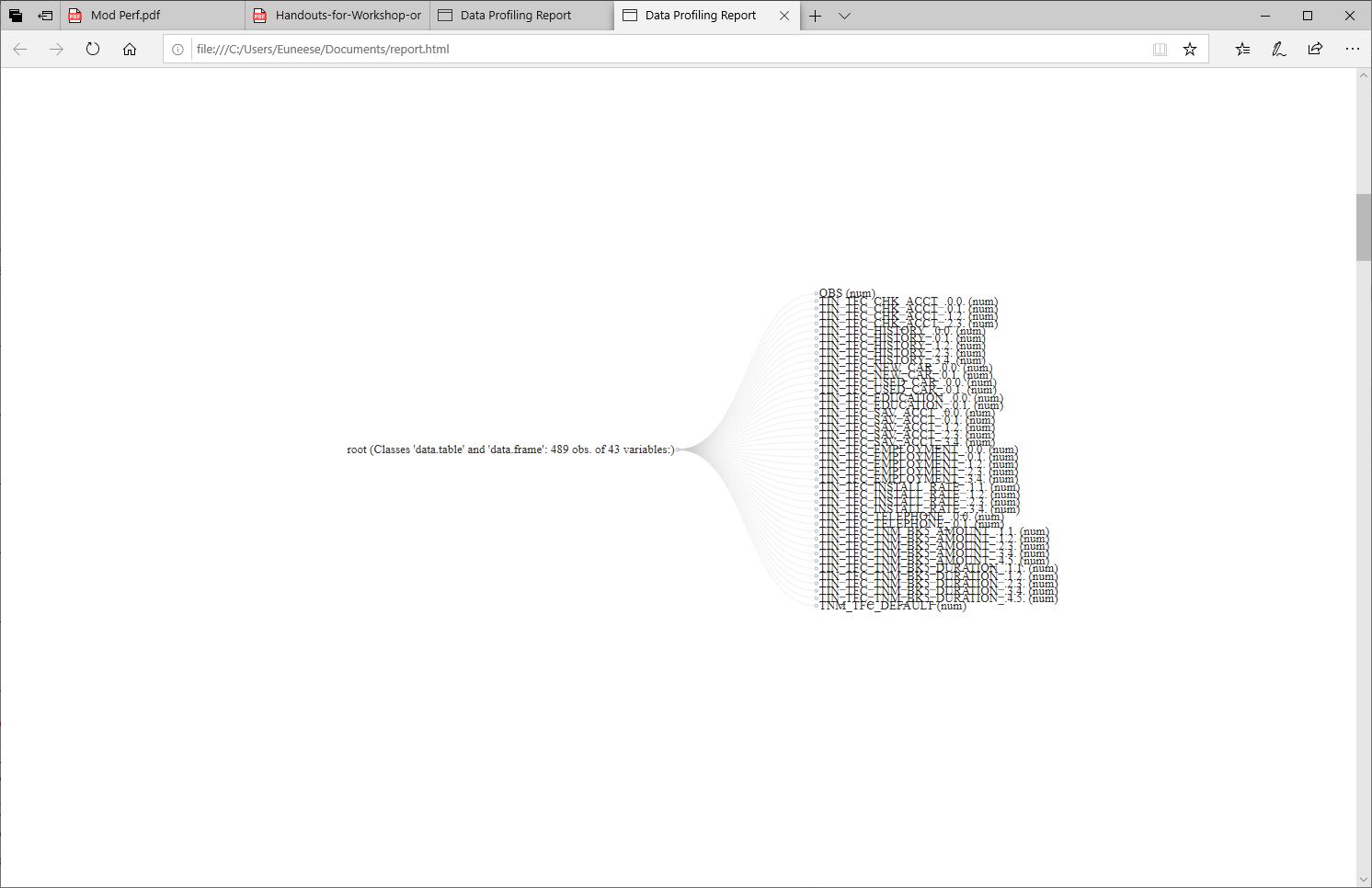
Again, in RStudio I ran a simple report that explores the data and returns results.

# Run a simple report to explore the data in the dataset Credit\_Data\_v3

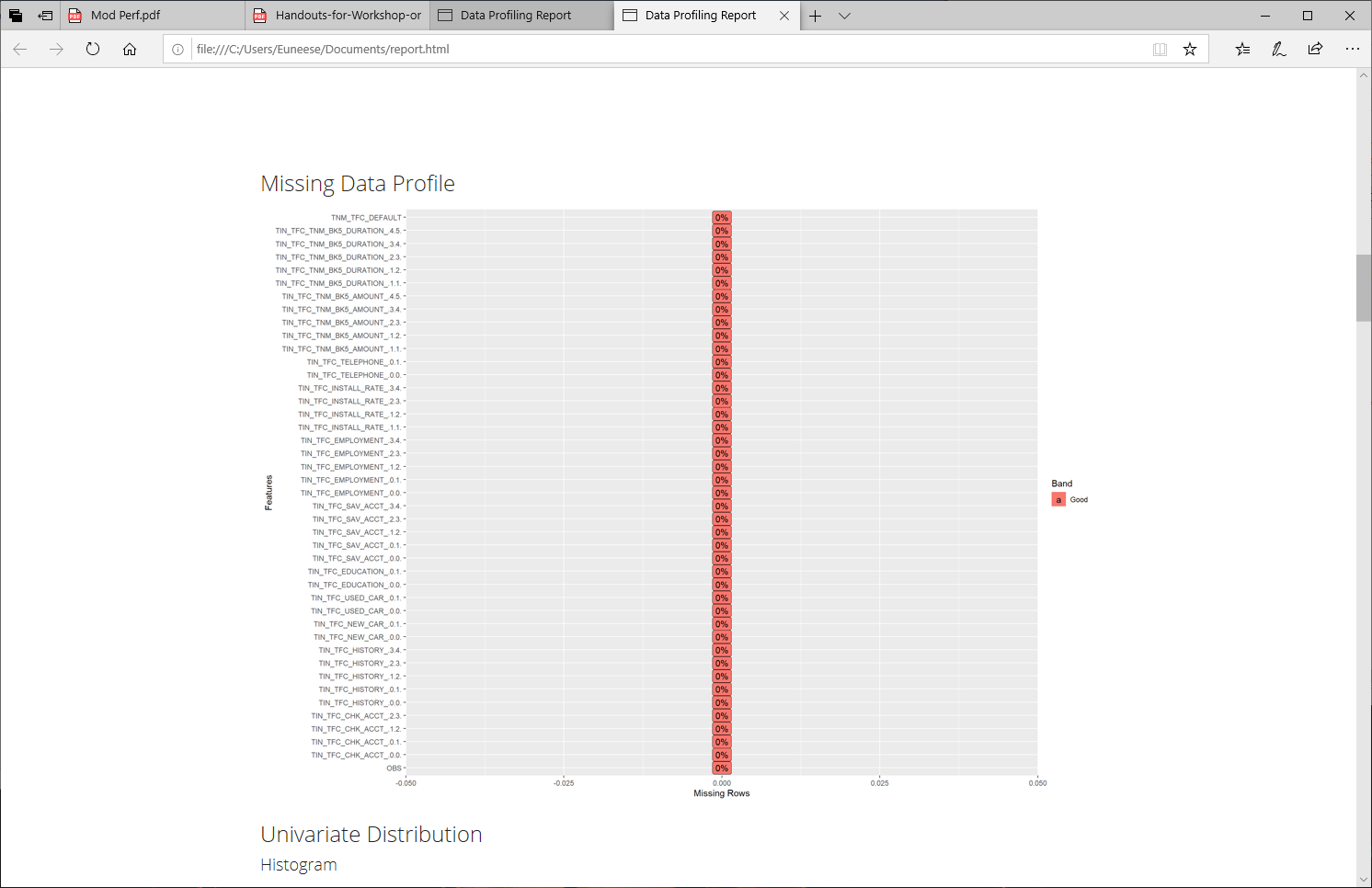
> library(DataExplorer)

> DataExplorer::create\_report(Credit\_Data\_v3)

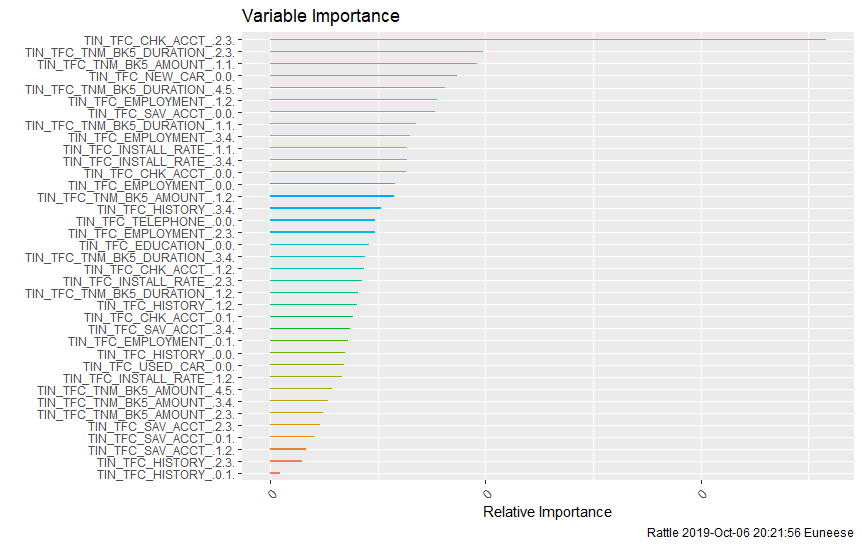
The basic statistics from the report shows that the data set contains 21,027 observations 43 columns (variables), and 489 customers with no missing data.



The report also identifies any missing data, as you see here, there is no missing data:



Using Rattle, I retrieved this chart of variables and their level of importance to the model:



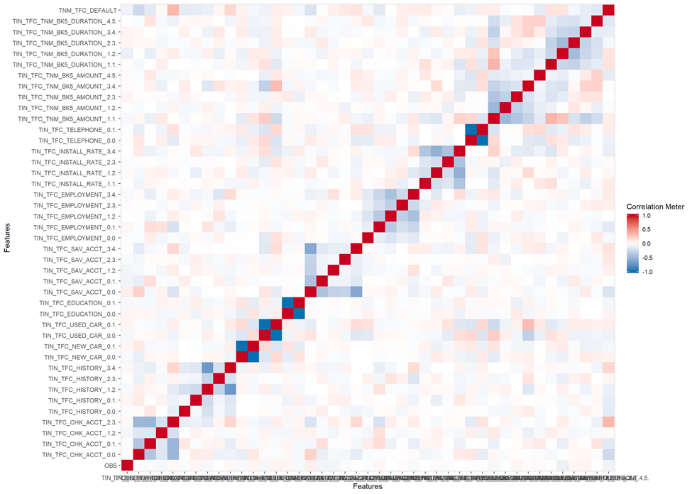
# **Building the Model – Project Plan Implementation**

## **Model Framework**

Building a model to correctly classify a customer as a bad credit risk takes time. The reason I chose to expose the variable levels is because in doing so, the model was able to reduce the overall error by 6% (see Table 1). I chose to use the default settings on Rattle to build and test the model. The default setting automatically partitions the data set into 3 datasets by selecting a random number of observations to equally represent the whole population. 70% of the original dataset is used to train the model. 15% is used to validate the model and then the last 15% of the original dataset is used to test the model. The validation and test datasets were not included in the data used to train the model so this data is assumed to be “new customer” data when applied. I also used a random number of data, 42 units, which will allow anyone to reproduce my work.

# **Data Mining**

## **Descriptive Analysis**

 Descriptive analysis is used to identify any pattern not seen by the naked eye. This type of modeling is unsupervised with no target variable applied. We are simply attempting to discover any association or pattern that in the data that we should look at closer. Association or patterns found here may be the foundation of predictive analytics.

## **Predictive Analytics**

Using RStudio and Rattle I was able to build a model that will predict an event: the event of a customer posing as a bad credit risk. I used a classification model which put new customers into one of two possible classifications. Good or bad. This is also displayed as 0.0., and 0.1. The target variable I used in the model is simply “DEFAULT” and that includes both levels of this variable. In the pilot run, I used .0.1. or yes as the target variable and received 100% accuracy. We know how that turned out, so I would prefer 60% accuracy with some human intervention included. This differs from descriptive analytics as predictive analytics is considered supervised machine learning since we already know the association of all of the variables to the target “DEFAULT” from the heuristic dataset provided.

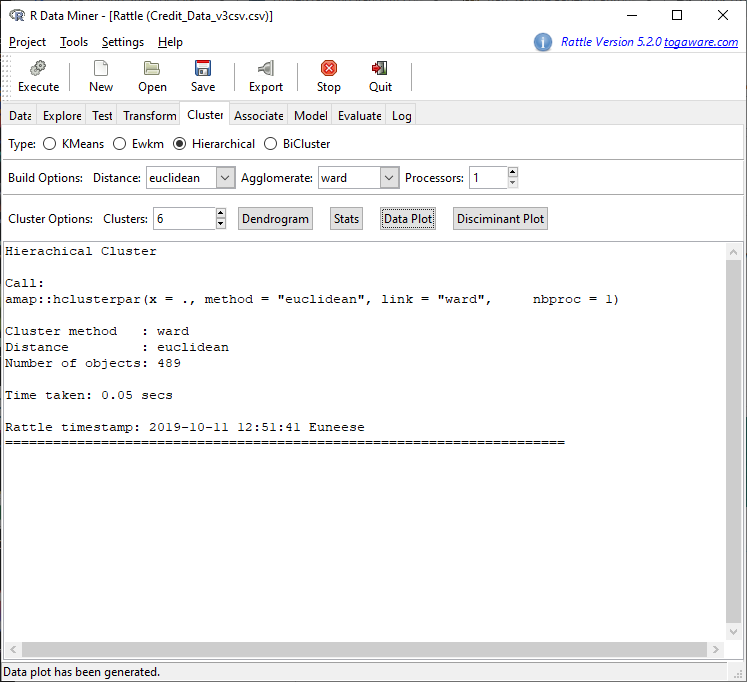
## **Variable Analysis**

## **Cluster and Association Analysis**

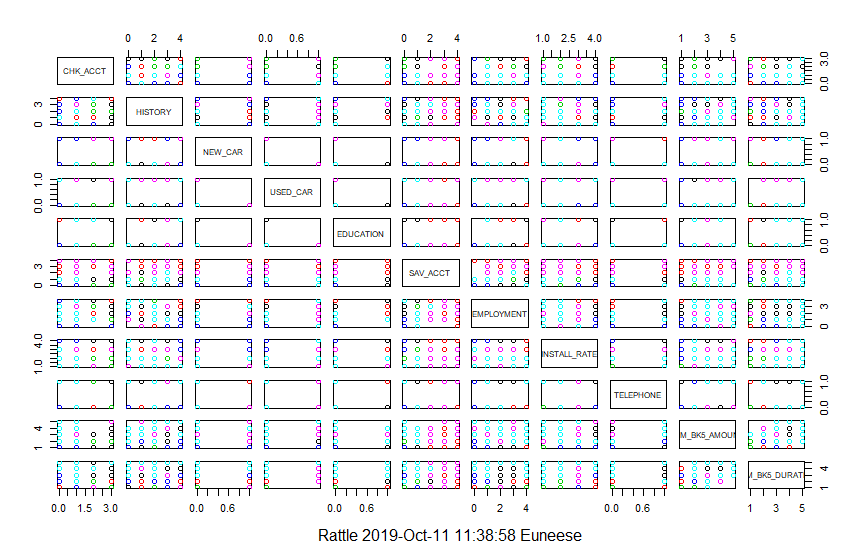
Clustering groups data into sets with similar attributes is a great example of descriptive data analytics. Clustering gives the data analyst the ability further identify data within the whole dataset that are more alike making the clustering of data via attributes more meaningful (Williams, 2013).

Without setting a target variable, Using Rattle I created this cluster of data. Under the Cluster tab, I selected KMeans, changed the number of clusters to 6 and left everything else at the default setting.

Using the Hierarchical cluster type and 6 cluster, the Data Plot clearly shows which variable have attributes in common. For the sake of clarity and time, I used the dataset that is compartmentalize, Credit\_Data\_v3. It is the same data, just all tied up into a variable bundle.



As seen below using data analytics tools to visualize data is an easy way to uncover patterns in large datasets quickly that cannot be detected by the naked eye. For instance, the three variables for loan type do not have many data points that are common with other variables. However, the others are very much associated.



## **Model Selection and Creation**

Using Rattle, it is easy to create a predictor model and there are many classification model types to choose from. I chose to create my model using a Random Forest. Below I experimented with two data sets and a simple decision tree to compare the results. Once created, I used the testing dataset to evaluate the accuracy of the model. The data contained within each data set is the same data. The first Credit\_Data\_v3 is compartmentalized, whereas v4 is the result of opening the variable levels.

In this error matrix; Table 1, you see that the class error for each model was the same however using the dataset v4 performed better when it came to the overall error. This tells me that a model built with Credit\_Data\_v4 will perform better in the long run.

Table 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| % | Credit\_Data\_v3 | |  |  | % | Credit\_Data\_v4 | |  |
|  | Predicted | | Error |  |  | Predicted | | Error |
| Actual | 01 | 00 |  |  | Actual | 01 | 00 |  |
| 01 | 12.3 | 25.5 | 67.5 |  | 01 | 8.1 | 23.0 | 73.9 |
| 00 | 6.6 | 55.7 | 10.6 |  | 00 | 2.7 | 66.2 | 3.9 |
| Overall error:  Class error: | | | 32%  39% |  | Overall error:  Class error: | | | 26%  39% |

The Receiver Operating Characteristic (ROC) curves as seen in Table 2 tell the story of thresholds of the data points above. It signifies the performance of each variable as a point on the curve. They do seem to be a little different. Using data that is still compartmentalized appears to give us a better true positive result at the beginning of the model and overall than the model using data points from all variable levels. Incidentally, the closer to 1.0 the better.

Even after this result, I continued on with the Credit\_Data\_v4 version of data with the variable levels exposed.

Table 2

|  |  |
| --- | --- |
| Credit\_Data\_v3 Area under the curve = 71% | Credit\_Data\_v4 Area under the curve = 68% |
|  |  |

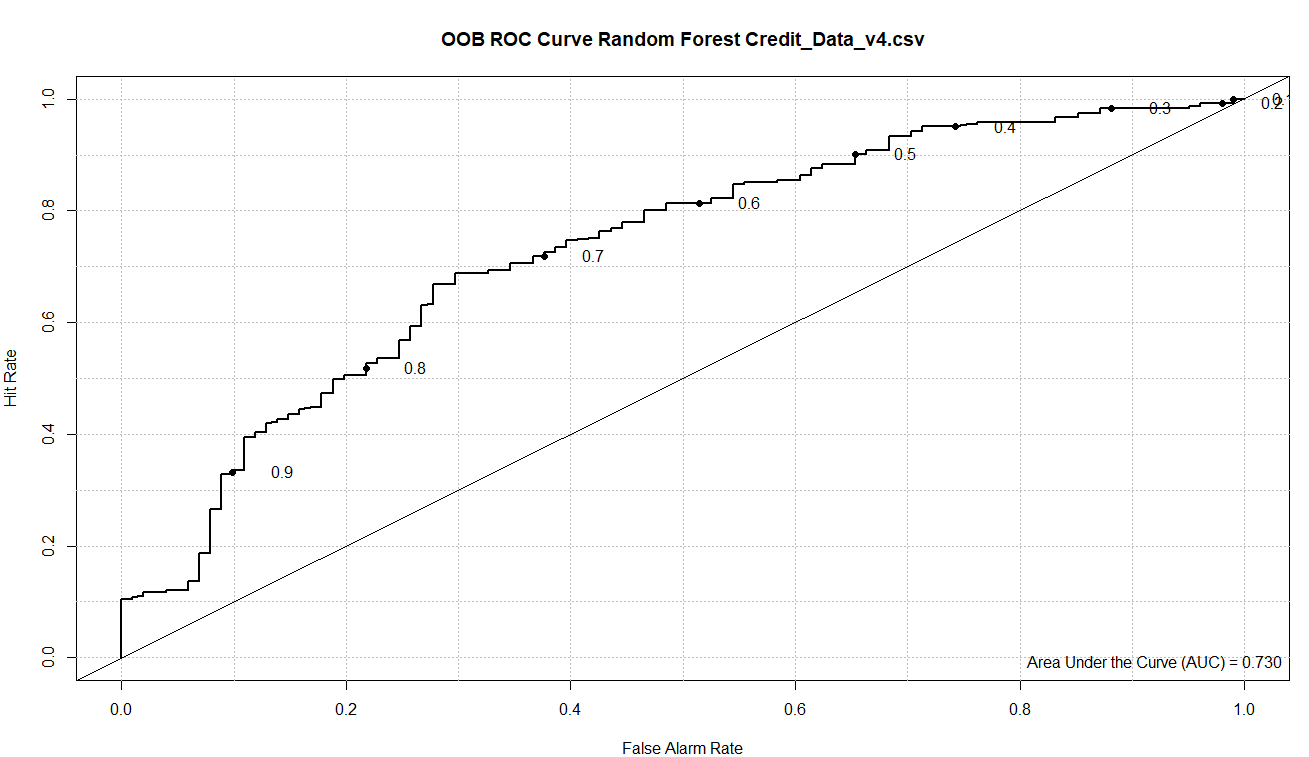
## **Model Selection**

Using the v4 data I looked at the results received from all models. As seen below, the Random Forest model seems to be the most accurate.



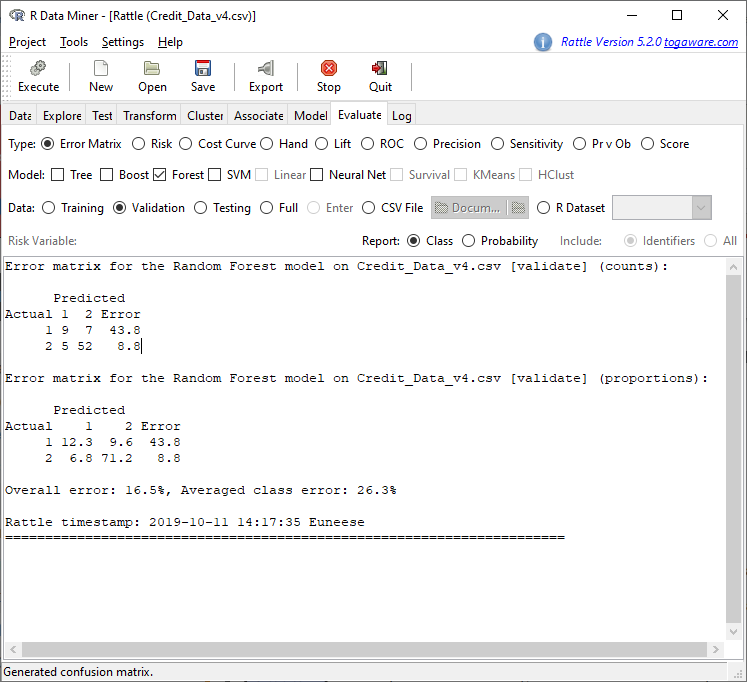
 Using about two thirds of the training dataset, the Random Forest algorithm builds several classification decision trees using a method called bagging. Each bag contains the same number of random samples which is used to build classification decision trees independent from each other. Each tree has different performance behaviors and on each node of the tree a different variable selected. When deploying each decision tree individually, it has equal weight to all other trees in the outcome of the final result. In the case of my Random Forest, I produced 500 individual decision trees 6 variables. Each one with a slightly different outcome. The overall Out of The Box (OOB) estimate of error rate is 25.15%. Looking at the error chart below, it appears that the trees line up around the 350th tree.

The area under the ROC curve is 73% which is a little lower than I like, however again, I do not want to overfit this model. As mentioned before other data collected will be useful to bring this accuracy percentage up.



# **Model Performance**

## **Measures of Performance**

 Rattle has a great tool which allowed me to evaluate the results of my Random Forest Model.

Using the validation and testing datasets I received the following result.

**Error matrix.**  While the training data would prove to be 100% accurate, the validation and training datasets below are accurate more than 60% of the time.



******ROC chart.** The ROC charts the true positive against the false positive rate at various thresholds. A false positive rate is a false alarm rate and true positive scores the rate of a positive hits. The area under the curve is the probability that the model will accurately predict new customer data. In this case, the model will accurately predict new customer data 68% of the time. The accuracy using the validation dataset is very good. If I tweak the data further, I fear I will overfit the model. This signifies that the model will accurate predict an event, the event of a bad-credit risk.

# **Scoring**

I chose to score my model using a scorecard method. I ran a new regression model on v4 and used most of the important data points to score the model. I scored the following variables:

Loan Amount, Loan Duration, History, Checking Acct, and Saving Acct. Here is the code used in RStudio.

#Filter variables

> {

> dt\_f = var\_filter(Credit\_Data\_v4csv, y="DEFAULT")

[INFO] filtering variables ...

#Breaking into training and test datasets. I set the threshold at 60% accuracy and the seed the same as I used when creating the model, 42.

> dt\_list = split\_df(dt\_f, y="DEFAULT", ratio = 0.6, seed = 42)

> label\_list = lapply(dt\_list, function(x) x$DEFAULT)

The weight of evidence (WOE) binning is a simple calculation of the importance of an independent variable to the dependent variable. In our case the code will bin variables-based frequencies of the dependent variable distribution. Then the code will combine the variables with similar WOE values (Bhalla). These new values will be used to score the model. In refence to this model. The calculation is as follows:

#woe binning

> bins = woebin(dt\_f, y="DEFAULT")

[INFO] creating woe binning ...

# converting train and test into woe values

> breaks\_adj = list(

+ LOAN\_AMOUNT.or.LOAN\_DURATION=c(1,2,3,4),

+ other.HISTORY.or.CHK\_ACCT.or.SAV\_ACCT=c(0,1,2))

> bins\_adj = woebin(dt\_f, y="DEFAULT", breaks\_list=breaks\_adj)

[INFO] creating woe binning ...

> dt\_woe\_list = lapply(dt\_list, function(x) woebin\_ply(x, bins\_adj))

[INFO] converting into woe values ...

#Ran a regression linear model using DEFAULT as a target on the newly created WOE values.

> m1 = glm(DEFAULT ~ ., family = binomial(), data = dt\_woe\_list$train)

> m\_step = step(m1, direction="both", trace = FALSE)

> m2 = eval(m\_step$call)

> pred\_list = lapply(dt\_woe\_list, function(x) predict(m2, x, type='response'))

> perf = perf\_eva(pred = pred\_list, label = label\_list)

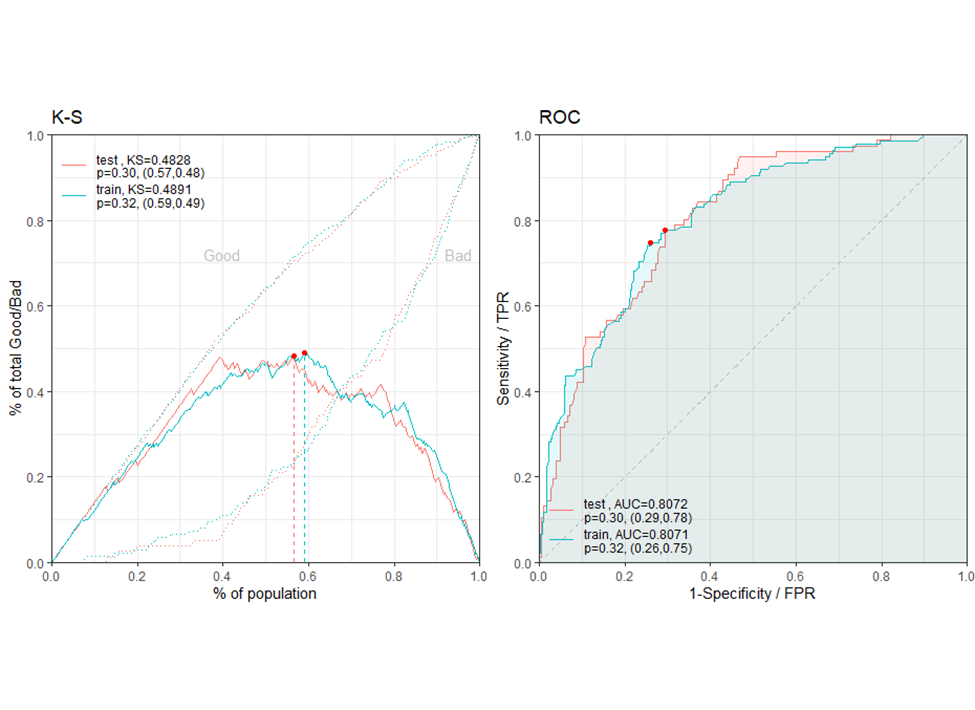
[INFO] The threshold of confusion matrix is 0.3163.

> card = scorecard(bins\_adj, m2)

> score\_list = lapply(dt\_list, function(x) scorecard\_ply(x, card))

> perf\_psi(score = score\_list, label = label\_list)

}

 The charts below show how well the model performed. It appears that the test and trainin­g data are very similar. Both the training and test data proved that the model can discern from good-risk customer data and bad-risk customer data 80% of the time.

The performance evaluation ( perf\_psi ) calculates the stability for the score and

variables. It is used to run ongoing tests on the model to ensure the data received from the population does not change (Xie, 2019). This is important because overtime the data will change and the model should be tested and calibrated often. The formula is as follows:

When the PSI is less than 0.1 there is no real change in new data. When the PSI reaches 0.1 - 0.25 there is some minor change, and the model should be evaluated. Anything greater than 0.25 signifies a major change and model will need to be changed (Xie, 2019.

{

$pic

$pic$score

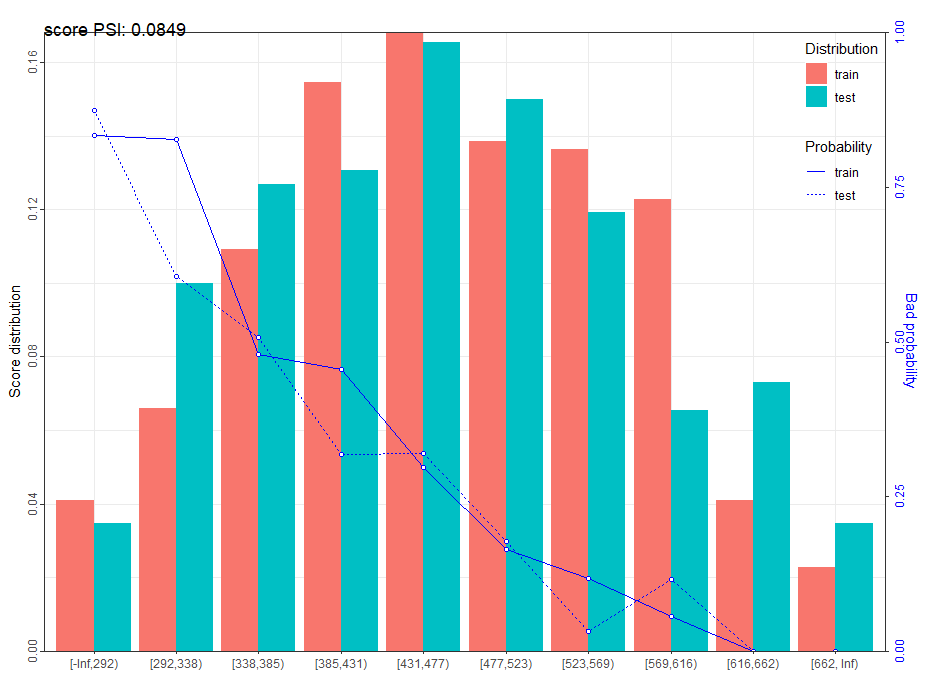
$psi

}

variable dataset psi

1: score train\_test 0.08494946

The PSI score for our model is fine at this time, however as previously stated, this should be tested often.



# **Project Plan Conclusion**

Overall, the model scored well. The model created in R Studio, and Rattle successfully predicted if a customer will be a good or bad credit risk on the onset. To recap, I used heuristic data provided from 1000 Synchrony customers. I scubbed the data and chose a set of variables to use to create a predictor model. I provided a positive return on investment with a projected five-year return of $12,693,430, and proved that a predictor model will indeed help to solve the business problem which is to significantly reduce the rate of defaults by identifying a bad-risk customer during the application process, to reduce costs and to increase profit. Furthermore, as discussed, with the addition of other models and new data captured, this model will increase in accuracy and identify considerably more bad-risk customers.

# **Recommendations**

I do recommend that a data dictionary be created to incorporate future variables as well as known variables. In addition, stakeholders will want to define what they will consider success for this project. Personally, I feel that 70% accuracy is successful, and we should strive to reach that goal. In the end, we will never be able to stop all fraud or loan default. Synchrony’s senior managers and officers should settle on an amount that they can live with or without in this case, in most cases this is 5% of company profits. Anything above this amount is excessive in my opinion and we should be able to control it with continual data processes.

I also recommend data be analyzed on each office and agent to look for outliers in approving bad-risk customers. It is possible to look at demographics and new training at locations identified. Perhaps the company may wish to offer a new low balance credit card for new customers who are on the threshold.

Lastly, I would also recommend that we capture new data from new customers and if we have it available on existing customers, I’d like to get my hands on it. The new data will help to identify bad-risk customers posing as other people. The new variables that should be captured are: ADDRESS, PHONE, AREA\_CODE, LAST\_NAME, ZIP\_CODE, and IP\_ADDRESS. This new data should be analyzed and mapped to existing customer data if we have it.

All of these recommendations will cut down on loan default and fraud, and will increase profits. Most of the recommended solutions are not costly and can be implemented right away.

References

Amadeo, K. (2018, November 06). When and Why Did the Stock Market Crash in 2008? Retrieved July 17, 2019, from <https://www.thebalance.com/stock-market-crash-of-2008-3305535>

Bhalla, D. (n.d.). Weight of Evidence (WOE) and Information Value (IV) Explained. Retrieved October 11, 2019, from https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

Bishop, S., & Ketron, G. (2019, January 23). PDF. Draper. Retrieved from <https://investors.synchronyfinancial.com/~/media/Files/S/Synchrony-Financial-IR-V3/press-release/q4-2018-earnings-press-release.pdf>

Code of Federal Regulations. (n.d.). Retrieved June 14, 2019, from <https://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&sid=1e9a81d52a0904d70a046d0675d613b0&rgn=div5&view=text&node=16:1.0.1.3.38&idno=16>

Consumer Credit Protection Act of 1968. (2018, November 29). Retrieved June 27, 2019, from <https://en.wikipedia.org/wiki/Consumer_Credit_Protection_Act_of_1968>

GE Overview. (n.d.). Retrieved from <http://snhu-media.snhu.edu/files/course_repository/graduate/dat/dat650/ge_overview.pdf>

General Electric Co (GE.N) Company Profile. (n.d.). Retrieved June 6, 2019, from <https://www.reuters.com/finance/stocks/company-profile/GE.N>

McDowell, C. (2019). Project Proposal for Synchrony. Unpublished manuscript, Southern New Hampshire University.

Privacy Policy. (2018, December 19). Retrieved June 11, 2019, from <https://www.ge.com/privacy>

Safeguards Rule. (2016, November 16). Retrieved June 14, 2019, from <https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/safeguards-rule>

Shuttleworth, M. (2008, August 2). Purpose of Research - Why Conduct Scientific Research? Retrieved from <https://explorable.com/purpose-of-research>

Synchrony. (n.d.). Retrieved June 6, 2019, from <https://www.synchrony.com/about-us.html>

Williams, G. J. (2013). Data mining with Rattle and R: the art of excavating data for knowledge discovery. New York: Springer.

Xie, S. (2019, September 1). PDF.

# **Appendix A**

The GE Synchrony Project Heuristic Data



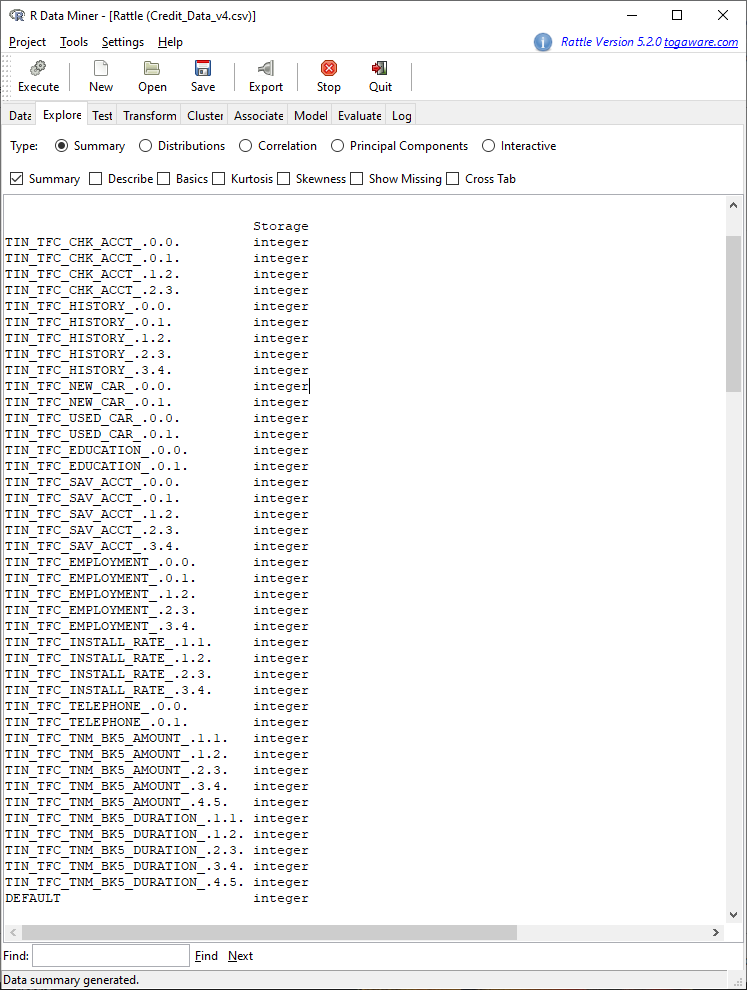
# **Appendix B**

The GE Synchrony Project Data

The GE Synchrony project involved analyzing heuristic data from GE’s Synchrony credit customers, and then build a model to predict default on future customers.

After analyzing and then transforming the original dataset seen in Table A1, the data in Table B1 was actually used to create the model. Some data variables from the original dataset were omitted.

Table B1

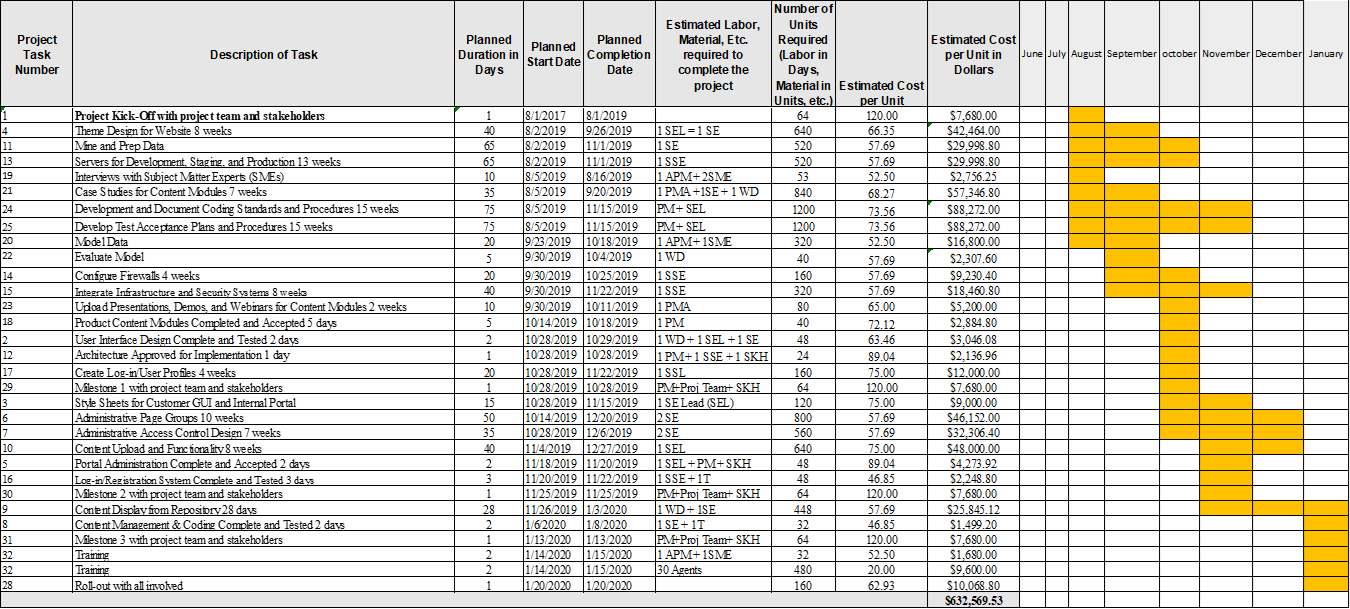


# **Appendix C**

The GE Synchrony Project Cost

The project scope as seen in Table C1 involves many project team members for approximately six months and is expected to cost just under $633,000.00.

Table C1



# **Appendix D**

The GE Synchrony Project Milestones

The GE Synchrony project is expected to take all of six months. Upper level management must sign off at each milestone as shown in Table D1.

Table D1

# **Appendix E**

The GE Synchrony Project

This stakeholder registry identifies project stakeholder’s and their level of influence and or support.



# **Appendix F**

The GE Synchrony Project Team

|  |  |  |
| --- | --- | --- |
| **Role** |  | **Name** |
| Project Sponsor | PS | Brian Doubles |
| PM - Project Manager | PM | James Waters |
| APM - Assistant Project Manager | APM | Sam Massoni |
| SEL - Software Engineer Lead | SEL | Steve Quan |
| Data Engineer | DE | Carrie McDowell |
| Business Analyst | BA | Susan Xiao |
| Software Engineer Lead | SEL | Cheryl Bolden |
| Software Engineer | SE | Richard Quest |
| Software Engineer | SE | Shannon Valley |
| Software Engineer | SE | Dennis Callahan |
| IT Systems Security | SSE | Beth Stroller |
| IT Systems Security | SSE | Herold Williams |
| Interface / Web Development | WD | Henry Stephenson |
| Training Specialist Lead | SME | Nicholas Bergh |
| Training Specialist | SME | Tyler Burns |
| UA Tester | SME | Kimberly Contos |
| Subject Matter Expert | SME | Monica Ianucci |
| Subject Matter Expert | SME | Jonathan Brant |

# **Discussion A**

The GE Synchrony Project Security Discussion

**This section is excerpted from Page 19 of the final project submitted for DAT-650**

**Security Goal Summary**

GE collects private data on each loan applicant. This data is invaluable to the company and should be treated as any other asset held by GE. The security goal for GE, it’s employees, and any third-party company that GE employs is to secure the data and to only use it in an ethical manner (McDowell, 2019).

**Data Source Security Requirements**

The Federal Trade Commission (FTC) has implemented a Safeguards Rule 16 CFR Part 314 which demands that all companies have measures in place to safeguard customer data. In addition, the company must ensure that any third-party company they do business with have policies in place to secure your customer data (Safeguards Rule, 2016).

Sections 501 and 505(b)(2) of the Gramm-Leach-Bliley act sets standards for developing and maintaining a security architecture to safeguard customer data (Code of Federal Regulations). This represents rules for all financial institutions governed by the FTC and extends to all service providers in which GE shares customer information with (McDowell, 2019).

Standards for safeguarding customer information include the development of a security program which will:

* Ensure the security of customer data
* Protect against threats to the program and customer data
* Protect against unauthorized access to customer data

The elements of the security program should include designating security personnel and conducting an internal and external risk assessment to include:

* GE personal training
* Risks associated to the company network and software design, and information system which include data processing, transmission, storage and disposal.
* Policies regarding detecting and responding to attacks and system failure
* The design of controls and procedures to regularly test the elements of the security program
* How GE oversees its service providers

Overseeing service providers will involve:

* Doing business only with service providers proven to safeguard GE’s customer data
* Requiring the service provider by contract to maintain a security policy

While there are laws on the book such as the Gramm-Leach-Bliley act and the Computer Fraud and Abuse Act of 1986 (CFAA), codified in 18 U.S.C. Sec. 1030 which protects individuals from others obtaining their private data via the internet unlawfully, the problem is that criminals that access private data are not in the US and cannot be prosecuted and in most circumstances the breach is due to the carelessness of a company employee or service provider. For this reason, pressure is on the company to take steps to ensure data privacy. GE must be abreast of all laws at a federal and state level in each country that they do business in (McDowell, 2019).

Regarding financial customers the following personal information is collected on the applicant, co-applicant, and co-signer:

* Government-issued ID numbers and Tax ID numbers
* US Social Security numbers
* Name, address, telephone number, email address, and fax number
* Biographical data
* Demographical data
* Purchase/credit history
* Financial information
* Location data

**Current Architecture for Data Privacy**

**Strengths**

GE has a robust data privacy policy. GE employs Chief Privacy Leaders for each business unit. These leaders can be found on the Privacy site at Support Central (McDowell, 2019).

Employees may access GE’s privacy portal here: integrity.ge.com, and within the GE Code of Conduct which can be accessed here: <https://www.sec.gov/Archives/edgar/data/1262449/000119312508061906/dex142.htm>, section 26 speaks on data privacy. It reads that GE is committed to handling data responsibly and in compliance with applicable privacy laws as governed in each country they do business with. For purposes of security GE follows these processes:

* Continual training
* Uses anonymous data, data that is encrypted
* Limits access to private data to individuals who need it for a legitimate business purpose
* Uses care to prevent unauthorized access to private data

GE states that everyone’s responsibility to look for and report:

* Unauthorized access to private data
* Violation of security protocols
* The sharing of private data
* The transferring of private data

Security breaches can come from physical locations as well as online. GE implements a Security and Crisis management (SCM) plan to secure employees, facilities, customer and business information, IT, and service providers from the threat of terrorist threats (McDowell, 2019). This policy includes:

* Business’s emergency planning and emergency drills.
* Adhering to the entry and exit rules at GE facilities, including wearing the appropriate badge.
* Maintaining a safe working environment
* Conducting appropriate background checks on new hires and contractors, wherever allowed by law.
* Screening all customers, suppliers, agents and dealers against appropriate terrorist Watchlists.

Additionally, everyone must report security lapses to business leaders, Crisis Management Leader or GE Ombudsperson (McDowell, 2019).

GE discloses their privacy policy to all to read here: <https://www.ge.com/privacy>. In it they explain the Privacy Policy and the measures they take to safeguard the personal information collected, the types of personal information they collect, how it is used, how it is shared, and how anyone may control the processing of the information (McDowell, 2019).

Additionally, GE complies with the APEC Cross Boarder Privacy Rules (CBPR) system. This ensures data protection of data privacy when transferred among participating APEC economies (McDowell, 2019).

GE has a policy to retain records only if they are relevant and to destroy records according to law. GE also provides their customers an option to choose how GE uses their private data and each year they send an email containing their privacy statement (McDowell, 2019).

# **Discussion B**

The GE Synchrony Project Ethical Standards Discussion

**This section is excerpted from Page 19 of the final project submitted for DAT-650**

On paper GE appears to have a robust policy regarding the ethical handling of private data.

GE has a very extensive Code of Conduct which can be accessed here: <https://www.sec.gov/Archives/edgar/data/1262449/000119312508061906/dex142.htm>

The author of the GE Code of Conduct asks that each person in the GE community which include service providers, commit to the Code of Conduct and to uphold a high ethical standard. Additionally, it explains further that GE leaders are responsible to foster this culture (McDowell, 2019).

The GE Code of Conduct includes:

* Obey the applicable laws and regulations governing our business conduct worldwide.
* Be honest, fair and trustworthy in all your GE activities and relationships.
* Avoid all conflicts of interest between work and personal affairs.
* Foster an atmosphere in which fair employment practices extend to every member of the diverse GE community.
* Strive to create a safe workplace and to protect the environment.
* Through leadership at all levels, sustain a culture where ethical conduct is recognized, valued and exemplified by all employees.

The statement also provides the rules of engagement and discloses the penalties if any party does not comply.

GE is committed to high ethical standards. They expect their employees to obey all laws governing data privacy and to conduct themselves becoming of the company. This includes third party doing business with or representing GE.

Section 39 explains that GE will monitor the use of company products including files on a computer, email transmissions, and phone calls to maintain privacy. While some may find

GE expects that anyone with access to private date follow this protocol:

* Collect, process, and use private data for business purposes only
* Do not share private data outside the use it was collected for

# **Discussion C**

The GE Synchrony Project Recommendations to Improve Ethical and Security Standards

**This section is excerpted from Page 19 of the final project submitted for DAT-650**

The goal of the finance department is to extend service to those identified as a good risk. To place an applicant into a “good risk” category it is important that GE does not appear to be conducting unethical data analysis. To address issues of ethical behavior I recommend using software to analyses private data using algorithms and untouched by human hands (McDowell, 2019).

As the extent to which the applicant’s data can be used fraudulently, it is very important to ensure the data is kept private.

**Recommended Software**

To analyze customer data and identify an applicant as a good or bad risk, I recommend using R with Rattle. To improve security standards, for data storage I recommend the use of AWS, a cloud storage and data transfer service tool offered by Amazon. The AWS WireWheel Enterprise Package along with their Barracuda Cloud Security Guardian offers data privacy and protection as well as compliance to all data privacy laws (McDowell, 2019).

AWS’ Ionic Enrollment and Key Management Services allow users to authenticate using stored and managed data encryption keys which serve as a customer-controlled gate. Finally, the CrowdStrike solution offers real-time detection and protection to threats.

To improve security at the UI level, the UI must be back end secured. The industry standard is through the Transport Layer Security (TSL) which encrypts the data between a server and client to secure data during online transmission. This is required by the Payment Card Industry (PCI)

  In addition to TSL, good strong security software like Kaspersky should be purchased. Kaspersky offers robust anti-malware products. The security team should continue training and always stay abreast of the latest security software (McDowell, 2019).

**Securing shared data or data touched by employees**

There should be a two-step authentication to access company servers or the database.

Employees who have access to the data should go through an extensive background check. Their emails and files should be accessed by the company security team and no customer data file should be accessible for electronic transfer. GE employees should go through yearly compliance training and be held accountable for their actions. There should be strict rules with severe consequences when data is mishandled (McDowell, 2019).

GE should allow only employees with certain permissions to access files. In addition, employees should be restricted where or how they access company servers and log in using a strict password policy that must be changes quarterly (McDowell, 2019).

If off site, employees should access the database only from a VPN, and they should not be allowed to access personal email or social media from company equipment or while logged into company servers (McDowell, 2019).

Virus protection should be installed on all company computers. Employees should not have permissions to download software or modify downloaded software. It is important that the IT group routinely access each computer to gauge its health and update all software (McDowell, 2019).

**ECOA and the CCPA**

In the interest of compliance, while it may be easy for long time agents to predict a credit risk on an applicant using their own intuition, it is best to calculate the risk leaving human intervention out of the equation. Using data analytics guarantees that GE is in complaint with ECOA and CCPA laws (McDowell, 2019).

The GE institution follows the Equal Credit Opportunity Act (ECOA) and therefore they will create a model using data to predict this for future applicants. The ECOA is a united states law (codified at [15 U.S.C.](https://en.wikipedia.org/wiki/Title_15_of_the_United_States_Code) [§ 1691](https://www.law.cornell.edu/uscode/text/15/1691) et seq.), enacted in 1974 that makes it unlawful to discriminate based on race, color, religion, national origin, sex, marital status, or age or the fact that any or all of the applicants income derives from public assistance (Equal Credit Opportunity Act, 2019)

In addition to ECOA laws, GE also follows the Consumer Credit Protection Act (CCPA) United States law [Pub.L.](https://en.wikipedia.org/wiki/Act_of_Congress" \o "Act of Congress) [90–321](http://legislink.org/us/pl-90-321), 82 [Stat.](https://en.wikipedia.org/wiki/United_States_Statutes_at_Large) [146](http://legislink.org/us/stat-82-146), enacted in 1968. Among other things, this act ensures that GE will disclose its terms and cost with borrowing and how those costs are calculated (Consumer Credit Protection Act of 1968, 2018).

# **Capstone Component Two: Personal and Professional Reflection**

Carrie McDowell

Author Note

Carrie McDowell, Student at the Southern New Hampshire University

This research was to complete an assignment for DAT 690, Dr. Ricky J. Sethi

Correspondence concerning this article should be addressed to Carrie McDowell, Student, Southern New Hampshire University, New Hampshire, 03106

Contact: [carrie.mcdowell@snhu.edu](mailto:carrie.mcdowell@snhu.edu)

**The Capstone Project**

The project I chose to tackle was the creation of a model that will accurately classify a credit applicant as a good or bad credit risk at the application stage. I feel that the model I created is a great representation to stakeholders which will convince them to invest in a data analytics process for continued success. The tools that I used contained in RStudio were very easy and basic; once I got the hang of it. I’m not too tech savvy when it comes to coding.

The challenges I faced creating this model was learning RStudio and learning how to code. The book, Data Mining with Rattle and R was very useful however basic. I wanted to do more, I wanted pizzazz and sparkle. I found a plethora of learning material and tutorials on line which expanded my world and blew my mind because I became overwhelmed at that point; in a good way. The greatest challenge that I encountered was to learn to get out of my own head. I had a hard time reeling it in and settling on one algorithm and sticking to it when I saw a new shinny anything out of the corner of my eye.

One great thing that worked well was reaching out to my instructor for help when I was stuck. I realized early on that he was my greatest tool and I was not too proud to beg. During the process, I had a running spreadsheet that I used to compare how different algorithms worked with each model. That worked well in deciding on the final algorithm and model type to use.

**Strength and Weakness**

As a senior manager, my strengths for this capstone experience come from knowledge driving continual improvements and experience with every aspect of a functional project. I have acted as a senior advisor/stakeholder, product manager, project manager, and SME. I have been on both sides of the business analysis table.

Having been a manager for many years did not prepare me for my next role, one in data analytics. I had always known that I could use data to solves business problems. I worked hard at solving my quest to prevent fraud before starting this course but failed. My weakness came from the lack of experience in coding and the knowledge to understand the reports generated from data analytics. After completing this course, I now feel more comfortable in presenting and understanding the outcome of my work. I must admit though, I struggle with color so I try to stick to basic boring colors for my charts.

Leading up to this capstone project, the courses taken previously prepared me for the final project. In particular learning project management, legal implications, visual aids were most useful in this final project. Every course provided the ability to write sound and specific reports/essays. As an adult student this has been most useful. This capstone experience summed up the program as a whole.

**Implementation**

As a data analyst, if I were to implement my analytic project/solution as, in a real company or organization today, I would begin with the project manager to outline the project, and prepare all documents, including a project scope and estimation of cost. I would depend on the project manager to get the buy in from the business owner, and develop the team.

Once buy-in occurred, I would create a simple report to present to all stakeholders outlining the need for and anticipated value that the project will return. I would develop a data dictionary and rules for data integrity before meeting with managers to create a gap analysis and record user case stories while creating the data dictionary. I would also work with the web developer to create a standard application outlining the data format for the back-end.

**Ethical Consideration**

I am of the mindset that people, companies, and the government have no business in personal affairs…until they do. Coming from a background of fraud detection, I would do anything to catch the bad guys, this included associating others with them and keeping our products safe from them. There is such a fine line between using a person’s personal data for good or evil. In the minds of the public it is all evil. In today’s society where people and companies are condemned on social media it is so important the companies keep a good reputation. Death by social media is quick and permanent. Most companies do not indulge what they really do with a person’s data in the background. Yes, they may and should disclose what they use the data for, that they will not sell it, and allow the user to edit and unsubscribe to emails, but that really is a small part of it when it comes to keeping the company safe. The company has an obligation to keep customer data secure. To keep a company, secure from fraud, they do need to use the private data of their customers to keep the bad guys out.

My capstone project did pose ethical issues. First off, it could be construed that if the algorithm to deny credit included the data point of sex or age, this could pose a problem. If word got out, the company could suffer the wrath of social media and pay millions in lawsuits. The executive decision was made to omit any data regarding sex, age, and race from the algorithm. While company experts would have used this data to create a better model, others may not understand the appropriateness within the algorithm.

Companies should have the utmost integrity when it comes to using and securing the data. The company should have strict guidelines that are fully enforced, and create annual learning modules. Externally, the company should let the applicant know what data they collect, how it is stored, how long it is stored and how it is used. They should give the customer access to it and allow them to unsubscribe to any marketing emails. They should also provide the privacy statement to them on a yearly basis.

It is important to secure the data since the company is liable for any intrusions. The company should hold customer and employee data on a server separate from their POS or email systems. The security team should have strict rules and assign permissions where appropriate.

**Personal Life**

The courses I have taken have prepared me for my future endeavor. They have increased my comprehension of data analytics, writing, and understanding what I read.

During the time spent in the final course I lost my job. I will use this project as an example of my skill and critical thinking abilities. I plan to upload this project into GitHub and include it in a portfolio of my work.

What’s next you ask? I plan to attend a six-month boot camp at Vanderbilt university to learn scripting languages such as Python and SQL. I think it would be a great addition to this major. Here is a link to the curriculum: <https://bootcamps.vanderbilt.edu/data/>

I have been in management since I was seventeen years old. I am at a point in my life were I no longer want to manage others. I have a passion for data analytics and want to land a job as a data architect or engineer. This program has set my foot on the right path. My experience with collaborating in cross functional and diverse teams will benefit me in this career move as well.

**Fin**

I cannot say enough about my experience with SNHU, the professors, and staff. My counselor diligently reached out to me frequently as a sponsor to an addict would with constant encouragement. She was wonderful. All of the professors were very helpful any time I needed assistance. My experience with SNHU was unlike anything I encountered before. It was very much agile and in the terms of data, it was semi unsupervised vs supervised learning. The classes were structured in a way that after instruction, it is expected that the student research and work hard to for solutions. Loved it, and thank you!