General Electric (GE) Credit/Financing Risk Predictive Analytic Solution

Charles Adkins

JUNE 5, 2020

General Electric (GE) Credit/Financing Risk Predictive Analytics Solution

Data Mining and Business Value

General Electric (GE) has expressed its interest in implementing analytic strategies and methods to identify potential business value and reduce organizational risk exposure in their credit/financing risk departments. In response to the financial crisis of 2008-2009, the credit branch has been asked to reassess the method used to determine if an application presents a bad credit risk (DAT 650, 2020). By implementing analytic strategies and associated methods, GE will be able to utilize predictive modeling to quickly identify factors that make an applicant at higher risk of default, subsequently, reducing their risk exposure and protecting organizational financial stability.

The project problem is to determine the likelihood of default given the variables provided in the dataset.

GE has provided a dataset that consists of 1,000 past credit applicants, of which are described by 31 variables that include aspects of the applicants' financial circumstances and history, as well as specific personal demographic information. By analyzing such data, GE can better characterize potential customers and by such characteristics, make data-driven decisions regarding associated financial risk and liability.

This project will be accomplished by following the CRISP-DM (CRoss Industry Standard Process for Data Mining) process, of which outlines an iterative framework for gaining an in-depth understanding of the business domain, preparing relevant data, data modeling, model evaluation and deployment (Shearer, 2000). The business understanding stage, which is the first stage in the CRISP-DM process, provides context to the project scope and emphasizes areas of concern that will need to be reiterated throughout the lifecycle of the project. This stage includes objectives that address criteria for business success, risk and contingencies, terminology and data mining goals and success criteria (Shearer). The second and third stages involve data understanding and data preparation. These stages involve describing and exploring the dimensions and quality of the data, as well as cleaning, constructing and formatting the data for subsequent requirements (Shearer, 2000). These stages are crucial to the success of the project, as they directly impact the accuracy and actionable insights gained from the project results themselves.

Later, the modeling, evaluation and deployment stages occur. The modeling stage involves selecting a modeling technique, designing the model, building and tuning the results (Shearer, 2000). Upon obtaining the desired features and function of the model, its results are evaluated in the evaluation phase. This phase includes referencing the model results to the business success criteria, of which was laid out in the first stage, as well as determining the next action steps that either grant or halt proceeding into the final stage (Shearer, 2000). The final stage is the deployment stage. During this stage, the project plan and associated model are deployed, monitored and maintained for effectiveness and efficacy, presented, reviewed and documented for future business initiatives (Shearer, 2000). By following the CRISP-DM process, the integrity of the project scope will be maintained and its associated success criteria will be achieved in an iterative and comprehensive manner.

Analytic Structure and Organizational Relevance

For this particular project, both descriptive statistics and a predictive model technique will be used. Descriptive statistics describe the basic features of the data in both numerical and graphical forms and can be presented by measures of frequency, distribution, variation, histograms and box-plots. These measures summarize the features from a collection of data, which can be used to better understand and analyze the behavior and patterns contained within (Lund Research, Ltd., 2018). Predictive modeling is defined as a modeling process that uses data mining and probability factors to predict outcomes (PAT Research, 2020). Each variable within the dataset will be given a probability value as it relates to the target outcome, default in the case, and variables with higher probabilities will have greater association with the target. By implementing such an analytic structure, GE will be able to determine the respective probability factors which may presage specific customers that are higher risks for default. This valuable insight will minimize application processing time, lower organizational risk exposure, safeguard financial assets and prevent customers from pursuing detrimental financial decisions.

Analytic Tools

The statistical programming software, RStudio, will be utilized for the purpose of this data mining and analytics project, as well as MS Excel, which houses the dataset currently. RStudio provides a wide range of applications and packages, of which includes the Rattle package. Rattle is a graphical user interface package that presents several different data mining options, including machine learning modeling and data visualizations (Williams, 2011). The features of both RStudio and Rattle make the requirements of a data mining project simple and intuitive. Users can explore data and its specific characteristics, examine modeling, parameter tuning features and visualizations, with a few short lines of code and by electing the desired program options. When datasets are complex in variety, volume, variability and veracity, the ability to efficiently and accurately analyze by way of statistical computing and model development is paramount in extracting and capitalizing on its intrinsic value. RStudio and Rattle both provide that ability and assurance.

Security, Privacy and Ethical Needs

Current Security Requirements and Ethical Strategies

General Electric (GE) has obtained, analyzed and capitalized on several aspects of customer credit/financing data, but due to historical loss, has determined the need for a reassessment of application processing criteria. The nature of such data must be protected by the data management strategies outlined in regulatory literature, otherwise, customer privacy and confidentiality may be violated and subsequent damaging financial and reputational risk events may implicate GE's future. Ethical concerns must also be addressed to protect applicants from discriminatory and prejudice conclusions, but also GE from negligent determinations based upon false-pretenses. Ensuring that an applicants data is utilized in an appropriate setting and application will bolster these efforts and produce a more credible and respected outcome. Specific security frameworks and ethical strategies will help GE manage such risks and concerns.

Currently, GE has the applicant data stored in a vulnerable state, more specifically, an editable spreadsheet. This is a breeding ground for violations in security and data integrity, as anyone with access to the file can purposefully or accidentally manipulate the data. The current data lacks preventative measures and frameworks that deter such violations.

In terms of ethical concerns, data evaluated by the application processing criteria, must be directly related to the target outcome. For instance, in terms of credit approval, an applicants marital status should not hold similar weight to that of their credit history, nor should their age. These particular demographics are more descriptive in nature, specifically to accurately identify an applicant, more so than determine their creditability as it relates to financial and credit liability.

Current Strengths and Weaknesses

The current architecture carries some security and privacy strengths that can be praised, but still contains areas that need more attention in order to fulfill the demand for a highly functional data risk management framework. For instance, the data is internal, meaning that it has been collected and maintain within the confines of the enterprise, which limits the amount of Personal Identifiable Information thats contained within this particular dataset (Wilson, 2014). The data does not contain names, addresses, contact information or account numbers, but due to the data being internal, such data can be linked to particular customers when cross-referenced to other sources. This provides security in that if the data were to be accessed without permission, the scope of damage one might be able to inflict on specific targets, theoretically, would be minimal.

Without a dynamic system of security measures in place, unauthorized persons with intent may be able to access not only this data, but other assets within the organization. With this in mind, the organization and this particular data would greatly benefit from the implementation of a layered security framework. This approach establishes protocols from the outside in, such that, it begins with perimeter security, network security, platform security, application security, data security and user security (Berson & Dubov, 2011). Although the organization may already have several of these security zones in place, it is worth revisiting the purpose and duties of relevant domains to ensure they are up to adequate levels for the needs of this particular dataset.

Improving Security and Privacy Strategies

Prior to implementing any additional security and privacy strategies, it's important to review and analyze the aspects of the current GE framework. More than likely, GE has relied upon several security features on the outer layers of their threat deterrence system, of which include fire walls, encryption, vulnerability assessments, virus protection, input validation and access controls, within the perimeter, network, platform and application security domains (Berson & Dubov, 2011).

In addition to these features and areas, there must be ample attention placed around the data itself, specifically upon encryption and access control. The value contained within the data can be protected by enforcing authentication, authorization and administration protocols, which can be defined as the ability to verify ones identity correctly, enforcement of access permissions to specified users and protecting, establishing, auditing and maintaining data integrity and availability, respectively (Berson & Dubov, 2011). By incorporating these strategies, in their fullest scope of applicability and functionality, the data will be protected from both internal and external threats and maintain a high-level of trusted integrity.

Ethical Strategy

The proposed architecture and strategy not only protects the value and integrity of the data from threats, but it also protects the data from unethical use. Such security and privacy measures are based upon the existence and capability of threats, while at the same time, upon specific data governance standards. These standards include the cultural awareness of data stewardship within the organization and how the security and the information lifecycle of the data is managed and exploited for particular purposes. By acknowledging, understanding and carrying out such standards and policies, the data will be protected and utilized with the integrity that it deserves and for the intended purpose and scope that it was initially obtained.

Data Ethics

In terms of data ethics, the privacy and confidentiality of the data must be protected, as the nature of the data contains sensitive Personally Identifiable Information (PII). Protecting the privacy of the data involves not exposing personal details to outside of uninvolved entities or individuals, while confidentiality rests upon the extent to which the data can be shared and with whom it is shared (Uria-Recio, 2018). Customers are trusting organizations and their respective data stewards to be aware of these protective concepts and grant them the power to utilize their data for specified reasons, while also protecting the value contained within from exposure to threats. In addition to these measures, a transparent view of how the data is being leveraged should be disclosed and well-understood. Both the carrier and owner should present each other with comprehensive outlines of concerns, understandings and procedures that involve such leveraging efforts. Finally, the data should not institutionalize unfair bias, of which models can absorb unconscious biases introduced to them, which ultimately amplifies these skewed perspectives (Uria-Recio, 2018). It will be the responsibility of the organization and the data stewards to reflect upon the data they use, how they use it and how they might join data elements to provide a more comprehensive view of the data and its intrinsic value and meaning.

Security Applications and Tools

For the needs of GE and their credit/financing data, a robust security framework application may be of some use. Such an application may be included in IBMs Security Products catalog. IBM has a broad portfolio of security products which include Advanced Fraud Protection, Data Security, Identity Access Management, Intelligence Analysis and Investigations, Mainframe Security, and more (IBM, 2020), all of which play a role in imitating and upholding the aforementioned requirements for GEs' credit/financing data. In addition to these security applications, an IBM data governance product, such as IBM Data Governance will be of value, as this tool helps establish an organizational environment that is conducive for effective security operations.

Model Creation

Overview

Prior to designing and initiating a data model, it must first be understood how that model and its respective results will be utilized within the organizational landscape. Understanding such utility rests upon the details of how the model is structured, how it incorporates the data and how the model results can be leveraged to generate the desired value and assist in data-driven business decisions. In the case of the General Electric (GE) credit/financing risk data, it is understood that the organization desires to manage their financing risk more effectively, by analyzing historical applicant data, to apply the findings towards future financial investments.

Data Analytic Strategies

Both descriptive and predictive analytic methods will need to be utilized to developed a full picture of this use case data. The key difference between the two types of learning techniques is that Unsupervised learning has no determined outcome, but rather the goal of uncovering clusters or pattern segments within data, whereas Supervised learning has an outcome of classes, which is determined by the inputs (Brownlee, 2019).

Descriptive statistics present the distribution (clusters and patterns) of the data across important variables, while predictive analytics apply appropriate algorithms to address the issue at hand to draw conclusions about the future. Descriptive analytics uncover intrinsic groups formed within the data, provide summary statistics and as such, bring to light how the variables are related. However, due to the nature of this strategy, descriptive analytics would not sufficiently address the issue at hand; There needs to be a conclusive understanding of which variables directly predict the target variable, not simply display how the observations are clustered together. Given this, it is appropriate to select from the various types of Supervised Learning predictive model algorithms to help address the use case scenario outcome, while in tandem, using descriptive statistics to present the nature of the data itself, to be used as supportive evidence for the subsequent use case conclusion.

Data and Analytic Structures

The dataset is composed of 31 variables and a target variable, ‘DEFAULT,’ which characterizes this structure as a binary categorical target variable dataset, where we understand that the 31 preceding variables, in some manner, influence and ultimately, produce the result of the target variable. As previously mentioned, the target variable is ‘DEFAULT’ with binary values of ‘0’ and ‘1’ - ‘No’ and ‘Yes’ respectively, and as such, this requires a predictive classification algorithm.

Logistic regression and Random Forests are two types of classification algorithms we will explore to assess their appropriateness for this particular dataset. These classification algorithms will effectively answers questions regarding the Use Case by establishing decision boundaries based on each variable and predicting the outcome probability.

Logistic Regression

Logistic regression uses the binary target variable and the datasets remaining variables, to determine the probability by which the variables predict the log of the odds of the binary categorical target variable (StatQuest with Josh Starmer, 2018). As such, we will be able to predict, based upon the values of the 31 variables, whether or not an applicant would be a high risk applicant (DEFAULT = 1 or ‘Yes’) or the applicant is of low risk (DEFAULT = 0 or ‘No’). The result of this algorithm will present the variables that are found to be statistically significant in influencing such an outcome. Limitations with this model include its’ lack of flexibility. The model is only able to determine the relationship between significant variables and the outcome, as it places no emphasis on thresholds for such influencing values.

Random Forests

Random Forests are a collection of decision trees that generate decision criteria for each variable based upon the best value available, that leads to a specified outcome (Augmented Startups, 2017). Random Forests alleviate the limitations of the logistic regression algorithm by generating a model that displays specific decision criteria for the best variables that predict the outcome. This algorithm handles bias well and displays variable value ranges that can be used to assist in decisions regarding “boundary” applicants. For instance, although an applicants’ variable values lead it towards producing a high or low risk determination, those specific values may be within close range of the variable decision criteria, which can be analyzed by a credit branch associate, to further consider the individual nature of the applicant. This feature mitigates instances where an applicant might be on the cusp of being labeled a high-risk applicant, but places emphasis on the acceptable ranges of values, which may appropriately deem the applicant low-risk, with optional contingencies.

Business Value

The combination of descriptive statistics, and both logistic regression and random forest modeling algorithms, will effectively determine which variables are most impactful towards a specified outcome, high-risk or low-risk default, while also providing the value thresholds that an applicant may possess, of which places them in one group or the other. The power of such models will accurately place applications in one outcome group, while also granting credit branch professionals the freedom to assess applications that may reside on the cusp of either outcome group. The combination of both methods of analysis and determinism improves accuracy and efficacy. The built-in precision of these models also provides the organization with metrics of sensitivity and specificity. Sensitivity is the proportion of positives that are correctly identified, whereas specificity reflects the proportion of negatives that are correctly identified (Patel, n.d.), both of which display the likelihood of default and reduce error in the decision making process. By understanding the sensitivity and specificity of a model, the organization will be able to more confidently make decisions with supporting statistical evidence. This combination of systematic determinism and professional data assessment, ultimately increases application processing efficacy, effectively categorizes lending risk, increases the probability of collecting ROI value, alleviates instances of misaligned application judgments and provides flexibility in making decisions that may have otherwise been thwarted by lack of support from data.

Model Pilot Plan

Advanced Analytics Overview

Advanced analytics address complex and time-sensitive issues by providing calculated solutions, many of which consist of deriving valuable data-driven insights used for improving business processes, managing risk and generating innovative ways of providing services and products to customers alike. By implementing advanced analytic techniques, GE will be able to leverage their credit/financing risk data to better understand their organizational environment, their clients and how the two entities can collaborate in creating a more cohesive and beneficial partnership.

Advanced analytics refer to a wide range of analytic tools and techniques which includes data mining, machine learning, forecasting/predicting, and pattern matching functions and metrics (Wagner, 2018). One of the most valuable elements of an analytics project, is the quality of data that is being used to derive insight. If the quality and referential scope of the data is lacking in accuracy or completeness, the results of the project will be useless to the organization. In addition to this, the data must contain the necessary dimension of complexity, to arrive at results that reflect real-world conditions and that are not hindered by sources or actions of bias.

Internal and External Data and Sources

The necessary measures required for the GE credit/financing risk Use Case will follow the stages included in the CRISP-DM process. Given that we have discussed the business understanding (stage one), as well as the data understanding (stage two), it is time to initiate the data preparation (stage three) and modeling (stage four) stages. The characteristics of the dataset provided by GE have been discussed in detail in prior sections of the report, of which will be included in the final presentation, but the dataset may benefit from additional sources. These additional sources may come from both internal and external sources, ultimately providing us with a more complete scope of understanding and thus, generating more reliable and detailed results.

Internal Data

Internal sources may include historical data that details instances of business that the applicant has conducted with GE in the past. For example, the GE credit/financing dataset is a dataset that includes 1000 observations, or applicants. This translates to 1000 applications, of which does not specify whether it is the applicants' first application submitted to GE, or whether they have submitted several applicants in the past, some being approved, disapproved or some mixture of the two outcomes. Such data may already be specified in the '**OTHER_INSTALL**' variable and the '**NUM_CREDITS**' variable, as they represent whether the applicant has other installment plan credits and the number of existing credits at the bank. The nature and value of such variables may, and should, play a key role in determining what type of behavior an applicant may exhibit in the future.

Another internal factor that will be of value to the project would be the applicants' geographical location. This data would be used for incorporating cost-of-living ratios. For example, the length of a fixed, monthly income in Seattle, Washington, would not go as far as it might in Topeka, Kansas, and as such, the details of the applicants' data, should include the geographical data to drill down into the risk and plausibility of the application. This data may be considered internal or external data, as it could be obtained outside of the organization, but also from within, as the organization undoubtedly knows the location of their customers, as it would be part of the rudimentary process of collecting personal contact information, typically obtained at the start of such business discussions. Nevertheless, geographical data will add dimension to the dataset, with will bolster the results in respect to real-world applicability.

External Data

Given the nature of this project and its' intended purpose, extending effort towards incorporating external data sources will be an avenue worth exploring as well, as it generates a more complete landscape of an applicants financial circumstance. The current dataset includes several financial and demographic factors, but these do not exhaust the list of possible key factors that may be of great assistance in arriving at an objective conclusion regarding application approval/disapproval. Such external factors might include the applicants' credit score, income, current debts and any potential property collateral. The latter factors are elements that are typically utilized for calculating ones credit score, as some call these characteristics the five C's of credit. The five C's include the persons character (reflected by their credit history), their capacity (their debt-to-income ratio), their capital (the amount of money they have), any collateral (an asset that is used as security for the loan), and the credit conditions (the purpose of the loan, the amount and the interest rate) (Segal, 2020). To gauge how these factors can be included into our current dataset, we will discuss each one further.

The applicants' credit history is a variable that is already included in the dataset and is established as a categorical type variable. The '**HISTORY**' values are:

'no credits taken' (0),
'all credits at this bank paid back duly' (1),
'existing credits paid back duly until now' (2),
'delay in paying off in the past' (3), and
'critical account' (4).

This communicates the status of any cases of credit that the applicant had in the past or currently has. Their capacity is their debt-to-income, which as mentioned before, income and debts would both be external data points that would need to be obtained to calculate that ratio. The income data would also be referentially impacted by the inclusion of geographical location, used to account for cost of living ratios. Their capital, as it is the amount of money they have, can be determined to be a combination of several factors. Their income being one, their savings being another, which is included in the dataset already, and the amount in their checking account. The dataset does include a variable that pertains to their checking account status (‘**CHK_ACCT**’), but it is also categorical, with values of:

‘< 0’ (0),
‘0 <…< 200’ (1),
‘=> 200’ (2), and
‘no checking account’ (3).

Given this, due to both the savings account and checking account status variables being categorical, it would be difficult to calculate capital, as they would need to be specified in a numerical sense. Furthermore, the savings account variable is an averaged, categorical value, which does not specify the timeframe, by which is was calculated to be the average. For these reasons, capital may want to be included in the dataset, but given respect, as it will not be an exact figure. The collateral aspect could be attributed to the variables already included in the dataset. The dataset specifies whether the applicant owns real estate, property or their residence in the variables ‘**REAL_ESTATE,**’ ‘**PROP_UNKN_NONE,**’ ‘**OWN_RES,**’ respectively. These values may be utilized to determine whether or not collateral is necessary. Finally, the conditions of the credit, as they are characterized to be the purpose of the loan, the amount and the interest rate, are of great importance to the purpose of the project. The purpose is credit has been included in the dataset, with limited examples of such, while the amount is also included. The interest rate is included as well, but is being presented as the ‘installment rate as % of disposable income'. This particular variable, '**INSTALL_RATE,**' needs further investigation to determine the context of such value.

Utilizing the power and detail of both additional internal and external sources of data will strengthen the results produced from the analytic efforts. The importance of such efforts remains focused on the quality and the relevance of such data, as the inclusion of diverse data sources does increase the likelihood of bias and unreliable results, if not given the proper attention during the data preparation and modeling stages. The overall scope of such attention is to minimize risk for both the applicants and the credit branch, as beneficial outcomes resemble credit being paid in-full, on time and producing a trustworthy partnership that can be favorably reflected upon in the future, which may lay the foundation for new business opportunities.

Data Preparation and Model Creation

Data Preparation

We can see from the results in **Appendix A**, the dimension function returns that the data frame contains 1000 observations and 32 variables and contains no values of 'N/A'. It's important to mention that R names 32 variables, because it has determined the 'OBS' or number of observations, to be a distinct variable. Thus, we will remove that variable column and let the data frame rows, represent the observations. As previously mentioned, other preparation efforts are required, but it is out of the scope of this particular section.

Then, we can explore the summary of the data frame by instituting the appropriate commands displayed in **Appendix B**. The summary function displays statistical data for each 8 variable, such as the minimum, 1st quartile value, median, mean, 3rd quartile and maximum. This outputs helped us understand the distribution of such variable values, providing us with intel regarding how the data values are spread out across observations.

Model Creation

Once the data preparation is complete and verified to be accurate, the dataset is then ready to be imported into the modeling package Rattle. The Rattle package displays our dataset, after being imported, and allows for the dataset to be partitioned into training, validation and testing sets. This action mitigates bias and helps ensure we gain reliable results. We verify that the variables are appropriately selected for their type (Identifier, Input and Target), then select 'Execute'.

Moving to the 'Model tab,' we have several radio selectors to choose from, by which will generate a specific model type using our dataset. In the case for a Decision Tree or Random Forest model, we would select 'Tree,' or 'Forest,' respectively. We then select 'Execute' again and the Rattle package generates the details and results of the model. Upon inspecting these results, we will then determine how effectively that model portrays the dataset and determine whether or not we can make any preliminary conclusions based upon said results.

For the Random Forest model, one can tune the parameters to prune the output to fit the accuracy of the model. Such pruning might include the number of trees plotted and the number of variables evaluated at each split. Evaluating the OOB error rate and confusion matrix will help in the analysis of accuracy and the needs for tuning. Preliminary results shown in **Appendix D**.

Model Pilot Plan Conclusion

Cross referencing the results of the models will help establish the reliability of each model and will help conclusively determine which variables are statistically significant in predicting factors an applicant may have, that ultimately lead to defaulting on a loan. More in-depth analysis of the results and effort placed in parameter tuning will create more accurate models that are more easily interpretable. As the Random Forest model has been found to be the most appropriate model, more time and effort will be placed into developing and tuning this model to produce conclusive results, which will be presented in the next section, Pilot Plan Results. With this information, GE credit branch will be able to make more proficient determinations in the credit/financing application process, with respect to maximum likelihood of such applicants defaulting on their credit obligations.

Model Pilot Results and Solution Application

Model pilot results and solution application are included within the presentation aspect of the proposal. The presentation grants a comprehensive look at the model results, the interpretation of the results, analysis of error and accuracy of the model, loss value and the return on investment, and day-to-day operational utility.

References

Augmented Startups. (2017). Random Forest - Fun and Easy Machine Learning. Retrieved May 10, 2020 from https://www.youtube.com/watch?v=D_2LkhMJcfY&t=343s

Berson, A. & Dubov, L. (2011). Master Data Management and Data Governance. (2nd ed.). McGraw-Hill. Retrieved April 20, 2020 from https://mbsdirect.vitalsource.com/#/books/1260121240/cfi/6/2!/4@0:0

Bekker, A. (2017). Big Data: Examples, Sources and Technologies explained. Retrieved May 15, 2020 from https://www.scnsoft.com/blog/what-is-big-data

Brownlee, J. (2019). Supervised and Unsupervised Machine Learning Algorithms. Retrieved May 7, 2020 from https://machinelearningmastery.com/supervised-and-unsupervised-machinelearning-algorithms/

DAT 650. Use Case Description Document. Credit/Financing Risk. Retrieved April 14, 2020.

IBM. (2020). IBM Security Products. Retrieved April 26, 2020 from https://www.ibm.com/security/products

Lund Research, Ltd. (2018). Descriptive and Inferential Statistics. Retrieved April 15, 2020 from https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php

Patel, K. (n.d.). Machine Learning & Business Value. Retrieved May 11, 2020 from https://www.datascienceassn.org/sites/default/files/Machine%20Learning%20%26%20Business%20Value%20by%20Kush%20Patel.pdf

PAT Research. (2020). What is Predictive Modeling? Retrieved April 16, 2020 from https://www.predictiveanalyticstoday.com/predictive-modeling/

Segal, T. (2020). Five Cs of Credit. Retrieved May 14, 2020 from https://www.investopedia.com/terms/f/five-c-credit.asp

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing. Vol. 5, Number 4. Retrieved April 14, 2020 from https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

StatQuest with Josh Starmer. (2018). StatQuest: Logistic Regression. Retrieved May 9, 2020 from https://www.youtube.com/watch?v=yIYKR4sgzI8

Uria-Recio, P. (2018). 5 Principles for Big Data Ethics. Retrieved April 24, 2020 from https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d105cd3

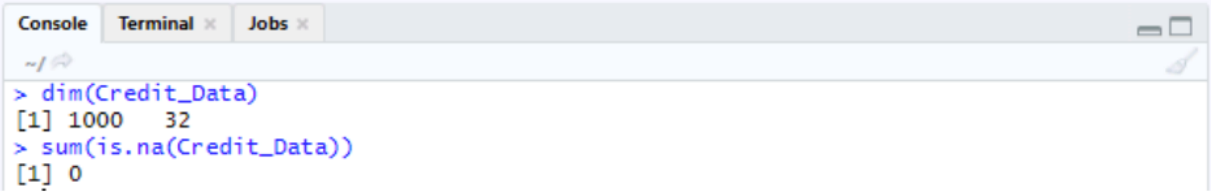
Wagner, J. (2018). Advanced analytics vs. artificial intelligence: How are they different? Retrieved May 13, 2020 from https://www.zylotech.com/blog/advanced-analytics-vs.-artificial-intelligence-how-are-they-different

Williams, G. (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer. Retrieved April 18, 2020.

Wilson, S. (2014). The strengths and weaknesses of Data Privacy in the Age of Big Data. Retrieved April 23, 2020 from https://www.constellationr.com/blog-news/strengths-and-weaknesses-data-privacy-age-big-data

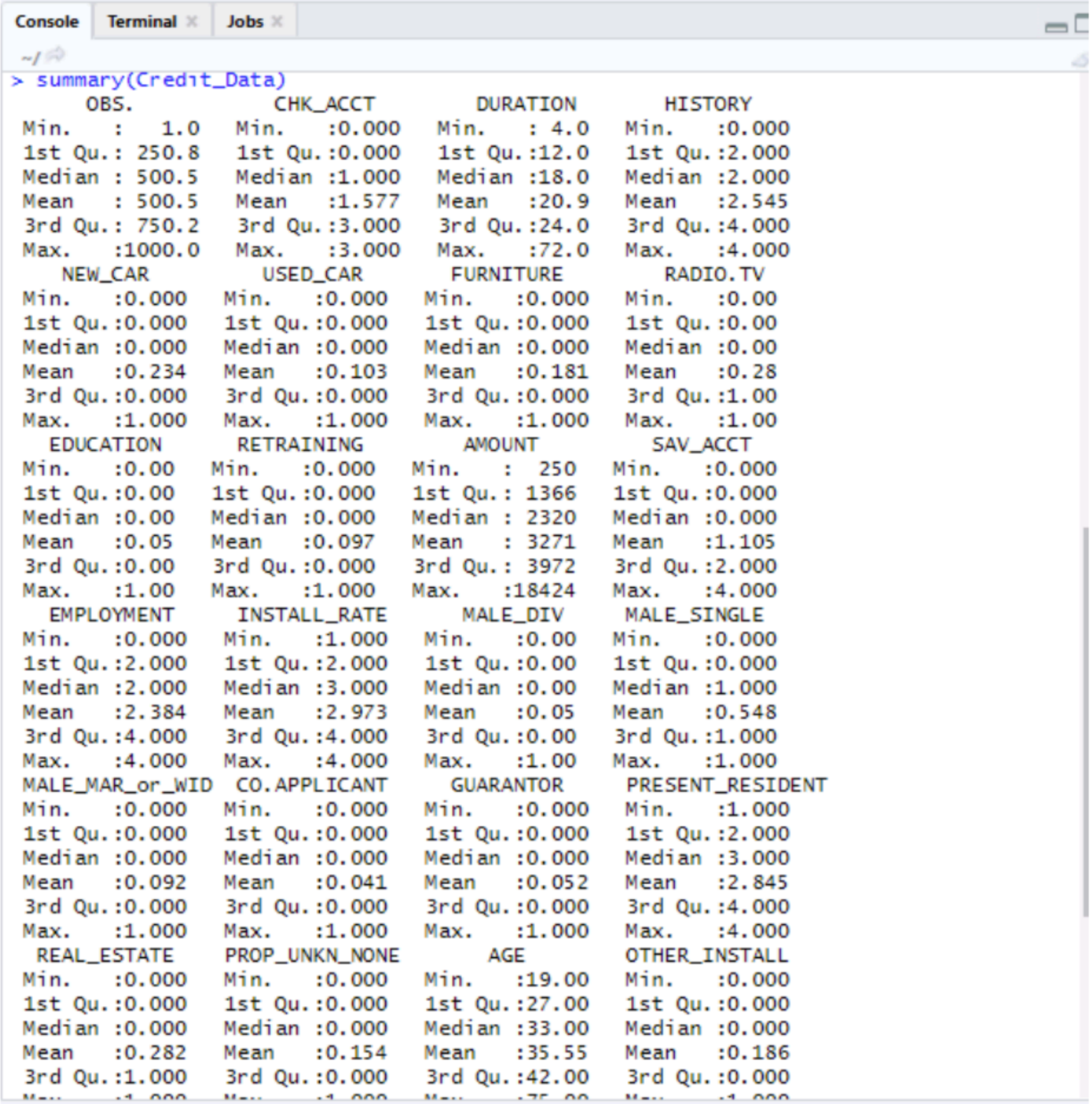
Appendices

Appendix A: Dimension Function and N/a Results



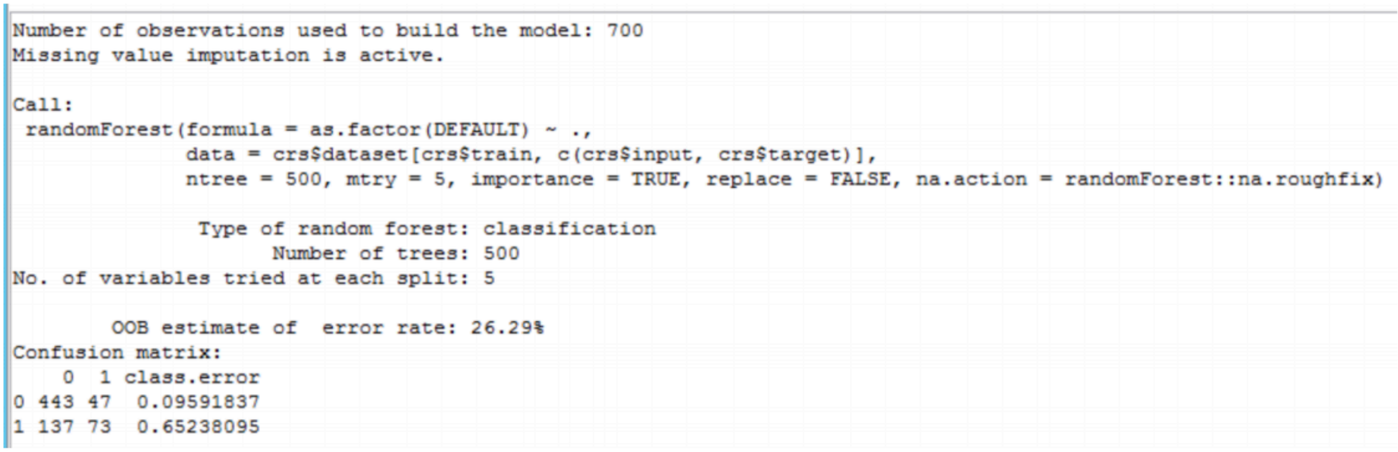
```
> dim(Credit_Data)
[1] 1000   32
> sum(is.na(Credit_Data))
[1] 0
```

Appendix B: Data Summary



```
> summary(Credit_Data)
      OBS.           CHK_ACCT        DURATION        HISTORY     
 Min.   :   1.0   Min.   :0.000   Min.   : 4.0   Min.   :0.000  
 1st Qu.: 250.8   1st Qu.:0.000   1st Qu.:12.0   1st Qu.:2.000  
 Median : 500.5   Median :1.000   Median :18.0   Median :2.000  
 Mean   : 500.5   Mean   :1.577   Mean   :20.9   Mean   :2.545  
 3rd Qu.: 750.2   3rd Qu.:3.000   3rd Qu.:24.0   3rd Qu.:4.000  
 Max.   :1000.0   Max.   :3.000   Max.   :72.0   Max.   :4.000  
    NEW_CAR         USED_CAR       FURNITURE        RADIO.TV   
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00  
 Median :0.000   Median :0.000   Median :0.000   Median :0.00  
 Mean   :0.234   Mean   :0.103   Mean   :0.181   Mean   :0.28  
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.00  
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.00  
   EDUCATION     RETRAINING        AMOUNT         SAV_ACCT    
 Min.   :0.00   Min.   :0.000   Min.   :  250   Min.   :0.000  
 1st Qu.:0.00   1st Qu.:0.000   1st Qu.: 1366   1st Qu.:0.000  
 Median :0.00   Median :0.000   Median : 2320   Median :0.000  
 Mean   :0.05   Mean   :0.097   Mean   : 3271   Mean   :1.105  
 3rd Qu.:0.00   3rd Qu.:0.000   3rd Qu.: 3972   3rd Qu.:2.000  
 Max.   :1.00   Max.   :1.000   Max.   :18424   Max.   :4.000  
   EMPLOYMENT     INSTALL_RATE      MALE_DIV      MALE_SINGLE   
 Min.   :0.000   Min.   :1.000   Min.   :0.00   Min.   :0.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.00   1st Qu.:0.000  
 Median :2.000   Median :3.000   Median :0.00   Median :1.000  
 Mean   :2.384   Mean   :2.973   Mean   :0.05   Mean   :0.548  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.00   3rd Qu.:1.000  
 Max.   :4.000   Max.   :4.000   Max.   :1.00   Max.   :1.000  
 MALE_MAR_or_WID  CO.APPLICANT     GUARANTOR     PRESENT_RESIDENT
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :1.000  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:2.000  
 Median :0.000   Median :0.000   Median :0.000   Median :3.000  
 Mean   :0.092   Mean   :0.041   Mean   :0.052   Mean   :2.845  
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:4.000  
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :4.000  
  REAL_ESTATE    PROP_UNKN_NONE       AGE        OTHER_INSTALL  
 Min.   :0.000   Min.   :0.000   Min.   :19.00   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:27.00   1st Qu.:0.000  
 Median :0.000   Median :0.000   Median :33.00   Median :0.000  
 Mean   :0.282   Mean   :0.154   Mean   :35.55   Mean   :0.186  
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:42.00   3rd Qu.:0.000  
```

Appendix C: Random Forest Preliminary Results (Excerpt)



```
Number of observations used to build the model: 700
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(DEFAULT) ~ .,
              data = crs$dataset[crs$train, c(crs$input, crs$target)],
              ntree = 500, mtry = 5, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 26.29%
Confusion matrix:
    0  1 class.error
0 443 47  0.09591837
1 137 73  0.65238095
```