DAT-690 Capstone Component Three: Complete Data Analytic Presentation with Personal and Professional Reflection

Beverly Gagnon

Southern New Hampshire University

Prepared for: Dr. Mary McDonald

November 15, 2020

Table of Contents

Data Sources and Business Value 4

CRISP-DM Application 4

Purpose of Research 5

Stakeholder Needs: Identification 5

Stakeholder Needs: Assessment 5

Stakeholder Needs: Application of Analytics Plan 6

Purpose of Analytic Structure 6

Articulation of Tool Selection 7

Additional Data Fields 8

Ethical Implication 9

Ethical Recommendations 9

Model Creation: Applicability 11

Model Creation: Value 12

Model Creation: Pilot Plan 13

Model Creation: Pilot Test 14

Pilot: Report Creation 15

Pilot Successes and Challenges 17

Pilot: Feedback 18

Pilot Lessons Learned 18

Modification Strategies 19

Actionable Steps 19

Potential Issues.. 20

Data Quality 20

Data Structure 20

Model Evaluation... 21

Areas of Concern 21

Fit of Model 22

Model Results 22

Model Deployment 22

Presentation Summary 22

Complete Personal and Professional Reflection 24

References.. 27

Appendices... 30

Data Sources and Business Value

"General Electric Company (GE) is a worldwide digital engineering company. GE's products and services extend from power generation, aircraft engines, and oil/gas production equipment to industrial, medical imaging, and financing products (Reuters, n.d.). GE is so large and diverse; it demands the greatest employees and attention to detail to run efficiently. Middle managers within The Company have perceived a trend of top talent leaving GE in the last year. This theory is confirmed by the rise in requests for new job openings. Not wanting to fall behind, GE has decided to act quickly and defeat the turn-over rate by using the data actively available. The Human Resources department has a current dataset containing the current and past employee attrition, this dataset will be a great beginning. This project will use the Cross-Industry Standard Process for Data Mining (CRISP-DM) to process the data and create models to estimate *what is the likelihood of attrition* within GE? Knowing this GE can focus on improvement in the area(s) needed for resolution. This project will describe the purpose of analytics, evaluate ethical implications, apply model creation, develop a pilot plan, report the results and include a presentation summarizing the steps in this report.

CRISP-DM Application

The CRISP-DM life cycle is composed of six phases – business understanding, data understanding, data preparation, modeling, evaluation, deployment. Using this methodology a data mining team can understand the data mining process and can use CRISP-DM as a recipe to adhere to while working on a data mining project (see Appendix A). During the business understanding phase analysis of the project outputs or goals is created (Shearer, 2000).

Purpose of Research

To remain a leader, GE has decided to act quickly and defeat the turn-over rate by using the data actively available through their Human Resources department. GE is eager to target the employees more likely to leave by discovering the variables present in the departing employees. Being able to predict the common variables GE can lessen their employee attrition rate, which will save time and money searching, hiring and training new employees.

According to studies, the cost of losing a salaried employee could cost six to nine months' salary or even twice the annual salary for an upper-level salaried executive. These expenses are from for recruitment and training of the leaving employee and their replacement. According to the Work Institute, 41.4 million US employees left their positions voluntarily. The Work Institute also predicts that by 2023 there will be a 35% turnover rate (Merhar, 2020).

Stakeholder Needs: Identification

It is vital to set expectations and deliver the results expected, so all stakeholders involved can understand the value of the project. If the project shows a positive perceptive and the stakeholders still have unanswered questions, the work put forth into a project will go unrecognized (Hasan, 2020). The stakeholders' needs will benefit The Company through cost and time spent on hiring. Employees will be less likely to leave when the variables are identified and addressed on a local level.

Stakeholder Needs: Assessment

Once the employee attrition variables are identified, through predictive analytics, GE will benefit from knowing by addressing the situation(s) before the employee gives notice or preventing job hunting. Being able to retain their employees will give GE time to focus on success and growth. Taking care and keeping employees will give GE more confidence in current and future project assignments and management.

Stakeholder Needs: Application for Analytics Plan

Valid predictive modeling will help all departments of GE by revealing likely variables that have been causing employee attrition. This project will use the Cross-Industry Standard Process for Data Mining (CRISP-DM) to process the data and create models. The CRISP-DM methodology will describe the purpose of analytics, evaluate ethical implications, apply model creation, develop a pilot plan, report the results and include a presentation summarizing the steps in this report (Shearer, 2000).

The available dataset will be analyzed using Rattle from the RStudio analytics tool. RStudio's platform can perform both descriptive and predictive data analysis (R-Project.org, n.d.).

Purpose of Analytic Structure

The analytic structure is a way of describing the methods being used to analyze the data. The purpose of a descriptive-analytic structure would be to search historical data to find a meaning or pattern (Rouse, n.d.) and describing the data being used for the modeling phase. Analyzing datasets to predict a future or unknown is the purpose of a predictive analytic structure (Imanuel, n.d.). This project is planning to identify what is the likelihood of attrition, so business users can preserve their talent (predictive analytics) through analysis of the Human Resources Employee Dataset.

Both the Naïve Bayesian and the decision trees algorithms are classification algorithms. A Naïve Bayesian predictive model performs as a great standard for contrast to other models, while the decision trees algorithm is the most instinctive and commonly used. Naïve Bayes models are used primarily for text-based information. They can be used to filter spam and other text by putting the data into categories. Naïve Bayes models believe if Event X is present then Event Y is probably happening. For example: If you put the word *money* in an email the model would automatically think the email is spam (HackerEarth, 2017). A Decision Tree is a classification and regression model that is used to predict the outcome of a target variable by using learned decision rules applied to the (training) data (Chauhan, 2019).

A Random Forest is another type of classification algorithm. It contains a huge number of different decision trees that works as a collaborative. Each tree in the random forest gives a predicted variable and the variable with the most outcomes is our model's prediction (Yiu, 2019).

Articulation of Tool Selection

Choosing the right data analytics tool could be something that can make the project results work the best fit. Josh Levy, manager of analytics at Aspirent, in an interview, said that choosing an analytic tool can be the largest challenge in data analytics. Levy created a five-step guide to help organizations find the right tools to use. The first is to consider the current state of analytic capabilities. GE, being a vast company, has a healthy analytic infrastructure in place, and information should be easy to acquire with simple interviews with the people who use the data every day. Second, the company should look at their current landscape of analytic tools and their application. Third, the analytics team needs to compare to other industry leaders on which tools are being used and how they are being used. Not all tools are created equal. The fourth step is looking at The Company's needs, what do we need this tool to do, and prioritize the needs to align with the use of the available tools suggested by the industry. Lastly, a decision can be made and followed through use by The Company (Bayern, 2018).

The HR Employee Attrition Dataset is kept in a Microsoft Excel spreadsheet. For the initial look and scrub of the dataset, the data scientists on our team will be using Excel to make modifications and become familiar with the variables and the information within the dataset.

The data analytics team feels uncomfortable to rely on a single tool. With the ease of use, compatibility and ability to show statistical graphics Rattle and R will be the tool of choice to use with this project. The data analyst team is most familiar with this tool. R is used to create great statistical techniques like linear/nonlinear, time series analysis, clustering, and more. R is shared cost-effective for the organization. It can be integrated with various other formats and can be used with Windows, macOS, and Linux. Rattle works as a point-and-click interface which makes it easier to control without scripting for results (R-Project.org, n.d.). All members of the data analytics team are familiar with Rattle and R so this tool will be used company-wide for this project.

Additional Data Fields

The thirty-five data variables available from GE's Human Recourse's Employee Database show a very thorough snapshot of each current and past employee. Some added data fields or variables should be included. They mostly show variables that could lead to being unsatisfied at work or signs leading up to a job seeking and leaving a current position even if it is well-paid. The additional fields that would aid in this project are the number of time off request approval/denial, expense amount approval/denied from business travel, team building event attendance, weekly attendance to staff meetings, and attendance to project/summary meetings.

Also, GE needs to look at the “shocks” within The Company to see if these workload changing events could be the reason for top and middle-level executives are searching for the same position in a competitor’s company (University of Missouri-Columbia, 2019).

Ethical Implication

Privacy and data security depend on trust. If privacy is violated this is a risk and constitutes a threat to security. Ethics can be the framework of the law when data is being referenced. Personal data is everywhere. Having full data privacy would include the ability to stop unauthorized use and access, ensure accuracy and correctness when being collected, having the right to inspect, update or correct your data, and have the feeling of ownership (Lee, 2016). Unfortunately, no one’s data is safe from everything.

If the Human Resources Employee Database was used on daily basis by everyone in the department for various reasons, what could stop an HR staffer to look further into an employee’s private information to learn about other subjects like performance ratings, relations satisfaction or work/life balance variable when all the HR staffer needed to do was update the current cell phone number of an employee? There is not an accountability stop in place that could prevent the rest of the data to be manipulated or read. Does the employee have to permit the use of data to take part in this project? Some of the variables that would cause some concerns of ethics are gender, race, pay rate, marital status, relationship satisfaction, stock option and other personal variables that are not for public knowledge.

Ethical Recommendations

It is recommended that GE should adopt the six protection principles from the personal data privacy ordinance in Hong Kong. These recommendations can be used in the gender, race, pay rate, marital status, relationship satisfaction, stock option and other personal variables not for public knowledge mentioned previously. They can be summarized as:

1. Data collection and Purpose Principle – Data is collected legally; data subjects are notified of purpose and only collect necessary data.
2. Accuracy and Retention Principle – Data must be accurate and only kept for the period needed to complete its purpose.
3. Data Use Principle – Only use data for the purpose it was collected.
4. Data Security Principle – Data personal need to takes steps to keep the data safe and to prevent a breach.
5. Openness Principle – Data team will be transparent with how the data will be used.
6. Data Access and Correction Principle – Data subjects should be able to correct and check for the accuracy of their data (Lee, 2016).

These principles will be able to keep GE's HR Employee Dataset up-to-date. As for the former employees, the recommendation would be to keep them archived only to be used in a demonstration purpose to help with the history or employees, not individuals. A proposed way to execute this would be to erase or hide the names of the data profiles of former employees of The Company.

Employee names and other identifying information is not included in the intended dataset. Employees remain anonymous.

Model Creation: Applicability

Before GE decides on which data analytic strategies to be used in their employee attrition project, The Company needs to explore which strategies will work best. Alex Bekker recommends a company should ask these questions about the project.

1) What is the state of data analytics at GE?

2) How involved do we need to get in the data, is the answer obvious?

3) How far are the current data insights from the insights needed for this project (Bekker, 2019)?

GE needs to invest in resources to recognize forecasters of attrition and develop advanced techniques to retain their sought-after employees. The first is to identify retention methods to examine data analytics strategies and create a model that will support The Company's increasing employee retention rate and turnover rate.

Descriptive analytics uses data aggregation and mining and will offer an understanding of the past and assists in finding what happened that made employees leave GE (Bachar, n.d.). Studying the data from the human resources department will give a great overall look at the reasons employees left The Company. The findings will make an incredible baseline to the turnover rate at GE. Researching the different employee variables of marital status, job satisfaction, environmental satisfaction, hour/monthly rate, training tune and others could uncover a statistic that was overlooked.

Predictive analytics will be used to create models and forecasting methods for understanding what will happen in the future (Bachar, n.d.). Using the employee variables as targets in the created models can give insight on which variables in the future could be improved to decrease the turnover rate. Some models that can be used in this project are correlation, clustering, logical regression and decision trees.

Model Creation: Value

The analytic models created for this model will show the variables that are present in the voluntary employee attrition rate. GE could use their gut instincts on what are the popular factors but the data will reveal the actual reasons. The correlation analysis will likely identify the variables that are influencing employee turnover. A clustering model will show the variables that have the most influence by grouping the data in groups within the model visualization. Logical regression is a classification algorithm that will determine the probability of the variables that are responsible for voluntary employee attrition. Decision trees and random forests can be used to research the variables shown in the clustering analysis to find the culprit and likelihood of the variable(s) that cause the employee attrition.

Knowing what is making the employees search for other opportunities could help GE hold on to their employees which they need to be successful in their industries and head into the future with employees that want to be part of GE's success.

During model creation, the variables which should not be considered are Over 18, Performance Rating, Percent Salary Hike, Standard Hours, Employee Count and Employee Number. These variables do not give any value to the project, for example, the number of standard hours worked is 80 hours for everyone and the employee number does not represent a value.

When decision trees and random forest models are being fashioned the depth, splits and number of trees for the random forests will be altered to seek a true outcome variable through the evaluation phase. With evaluation, the model's percentage of showing true positives during their production will be noted and considered when altering model construction.

Model Creation: Pilot Plan

For this analysis, a pilot plan is being developed to gain insight into the GE attrition rate. Employees that leave GE are given a survey and the results have been recorded, this dataset will be used to train on for this project. This analysis is to identify the main variables/reasons for the employee turnover rate. GE wants to retain its employees to remain to have a competitive edge and the cost of time training its employees. The pilot plan will be following the stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000).

The first step of the pilot plan is data preparation. The GE Employee Attrition dataset selected is collected by the HR Department will be used. It includes the data that is needed to answer the question of *What are the variables behind the employee attrition at The Company?* The dataset will go through a cleaning process to ensure that all values make sense and there are no missing values or data. If the data needs to be constructed, integrated or formatted it will be included in this step. The main objective is to have the data ready to be fed into modeling software (Brown, n.d.).

Once the data is prepared the dataset will be loaded in R/Rattle to be explored and modeled. The dataset will be loaded in Rattle using a 70/30/0 percent partition. The target variable is Attrition. The variables that are ignored are Employee Count, Employee Number, Over 18 and Standard Hours (same for all employees) these variables will have no impact on the predictability of the models being created.

This plan will examine the relationships between the variables. The modeling that is proposed to use is a correlation, clustering, decision trees and random forests. After the models are run, they will be evaluated for their margin of error. If the models perform poorly, another modeling strategy will be selected to show the prediction power of the data.

The evaluation technique being used is the receiving operating characteristic (ROC) curve. It will plot a visualization that will show the error or success of the model being used in comparison to the regression line representing the data (Williams, 2011). The outcome of the pilot plan will be recorded and plan modifications will be done tailored when needed to discover further employee attrition analysis.

Model Creation: Pilot Test

A pilot test was created and deployed. The data was loaded into R and analyzed by Rattle. A Summary was run to find out the means and median of the variables being reviewed for this employee attrition project. Along with the summary, a "missing data" report was run, in which it is found that the dataset is complete and no data is missing.

Out of curiosity, a Correlation Plot was made and with a reduced data filter the plot did show the datapoints mostly around the -0.2 – 1 values on the plot. There was an outlier of the variable for Performance Rating that landed in the 0.2 value but after a closer look, it seems that was, just that, an outlier.

Two decision trees were fashioned. They both were scattered and involved too many variables to study. When the variable was brought down to 5 for the split and the bucket with a maximum depth of 7 a more, readable version of the decision tree was shown. It had put the Overtime variable to the top and filters out to Monthly Income, Years with Current Manager and so forth. According to the decision tree, it seems that Overtime was a factor along with age since it showed up on both branches of the tree (see Appendix B). After running a ROC curve on the decision tree, it gave a 66% chance of being accurate in predicting employee attrition (see Appendix C).

To further manipulate the dataset two random forests were executed. One with 500 trees and 10 variables. The results can out as Overtime, Age, Monthly Income and Total Working Years were the top variables for employee attrition. The second had 500 trees and 5 variables, Overtime, Age, Total Working Hours and Job Role were the top variables for this model. A ROC curve was run for each random forest model and the results were 83% accuracy of predicting the variables for employee attrition (see Appendix D).

In the conclusion of the portion of the project, it would be a reasonable plan to run a variety of predictive models to be able to observe the data and see what the data's story is. If the various modeling steps were skipped there would be no other models to compare the evaluated results too. If the data analysis terminated after the decision tree models, the predictability would only be 66% correct. When studying this data with predictive modeling it shows the top variable involving employee attrition at GE is Overtime.

Pilot: Report Creation

The following is a report and visuals created from the proposed pilot plan involving the GE employee dataset supplied by the HR department. The steps are as follows: two decision tree models and two random forest models.

The first decision tree model was run as a traditional model with the variables brought down to 5 for the split and the bucket with a maximum depth of 3 for a more, readable version of the decision tree was shown. The variable of overtime is shown at the top (see Appendix E). The evaluation technique being used is the receiving operating characteristic (ROC) curve. It will plot a visualization that will show the error or success of the model being used in comparison to the regression line representing the data (Williams, 2011). The ROC curve (see Appendix F) reveals that even though there are not any false positives to report on, the accuracy/error rate is only at 65%. This may be too low to bank on for this project.

The second attempted decision tree was run as a condition model with the variables at 5 split and bucket and a maximum depth of 3, showing overtime at the top as well (see Appendix G). The ROC curve from the evaluation of the second decision tree also shows there are no false positives but the error rate is just 5% more than the first decision tree created with a 70% error rate (see Appendix H).

The decision trees show that overtime could be a factor in the employee attrition rate at The Company. The error rates of both are not great percentages and the team was looking for a better rate percentage to be more certain of the factoring attrition variable.

The modeling technique of random forests will be used to see if the team can receive a smaller error rate over the decision trees. To further manipulate the dataset two random forests were executed. The first random forest has 500 trees and 10 variables. The results came out as Overtime, Monthly Income and Job Role were the top variables for employee attrition. The out of the bag (OOB) error rate shows 14% and the ROC curve puts the model at an 83% percent accuracy rate (see Appendix I).

The second random forest created had 500 trees and 5 variables, Overtime, Monthly

Income and Age were the top variables for this model. A ROC curve was run and the results were 85% accuracy of predicting the variables for employee attrition (see Appendix J).

Pilot: Successes and Challenges

When the user case dataset was analyzed with a correlation plot, the variable of Performance Rating came up as an outlier. At this point the team was not confident that this was a true reason for employee attrition, it was only an outlier. At this point, this variable was ruled out as not being a cause of employee attrition and taken out the possibilities.

The decision trees that were created showed Overtime as the deciding variable followed by Monthly Income and Years with Current Manager. The ROC curve showed a 66% chance as being the correct predication. The team was not satisfied with this mundane percentage.

Next, two random forest models were done to see if the results would differ and have a higher correct predictive percentage. The results for the first model shown Overtime, Age, Monthly Income and Total Working Years were the top variables for employee attrition. The second predicted Overtime, Age, Total Working Hours and Job Role were the top variables for this model. A ROC curve was run for each random forest model and the results were 83% accuracy of predicting the variables for employee attrition. These models presented a percentage rating that was more acceptable for this project, see the table below.

Type of model	Top Variables	ROC Curve Percentage
Decision Tree #1 & 2	Overtime Monthly Income Years With Current Manager	66%
Random Forest #1	Overtime Age Monthly Income Total Working Years	83%
Random Forest #2	Overtime Age Total Working Hours Job Role	83%

Summary of the Modeling Results

Pilot: Feedback

The predictive modeling project feedback that was received was to take full notes on each model being created. If the models are made without know which variables are being analyzed, how can modifications be done on the continuing models? Having a record, of the variables for each model, the team will not repeat their efforts inadvertently and be distracted by the work that has been already performed and assessed.

Pilot: Lessons Learned

While building the models for this project in DAT-650, there was not a strategy on which variables were to be selected for the models. There were attempts to seek out the “potential” culprits to focus on, but a method was not discovered. There must be a more direct technique to narrow the variables used for the model building.

The other lesson learned was documenting all the models being build. It was a lesson taught in DAT-640 but the practice was not carried over into the next semester’s project. Once realized, the remaining models were well documented in the form learned in the Predictive Analytics course.

Modification Strategies

After refocusing on the business problem, the purpose of this project is not trying to solve which factors cause employee attrition, it is what is the *likelihood* of employee attrition with the variable is present. Once the variables are found, the modeling can be built predicting the *likelihood* of employee attrition when x, y, and/or z is present.

During modeling building, the number of models built should be higher than the four decision tree models and four random forest models that were used. Adjusting the number of models built could generate more legit findings.

Another approach could include dividing the departments into 3 different sections. The separation of the data could show very different results since Job Role is a top variable in the random forest model.

Actionable Steps

This phase of the project could use a scrubbed down version of the employee attrition dataset. Some of the variables were not used or needed instead of <Ignore> them in the modeling tool they could be removed from the dataset. Also, the Naïve Bayes and regression models should be created and studied to see if they perform better than the decision tree and random forest models.

Lastly, the team is using Rattle to do the modeling and some of the members are suggesting using RapidMiner as an alternative tool to help with the predicting. We will make attempts to carry out modeling with RapidMiner as well.

Potential Issues

Issues with scrubbing down the employee attrition dataset could pose a potential issue. Take out the data needed for the modeling could occur. If more data were available through GE's Human Resources department other variables could be studied and the likelihood of attrition may tell a different story from the sample of data presently being applied.

Data Quality

The dataset being used has been obtained from the HR department of GE. It includes 1,470 entries and 35 variables (columns). A modified version of the dataset will be used which includes 1,470 entries with 27 variables. Repetitive and non-related variables were dismissed. This was done to simplify the modeling process. The variables removed were employee count, employee number, monthly income, monthly rate, over 18, performance rating, standard hours and stock options level. From this point forward the term *dataset* will be referring to the modified dataset of GE attrition. The dataset was loaded into Rattle and shows NO MISSING DATA (see Appendix K).

Data Structure

The data structure is a 70, 15, 15 partition. 70% of the dataset is used as a "training data" dataset, while the other two partitions are used for validation and testing of the data, 15% for each. Even though the variables are numerical and categorial the values were kept "as is" to keep the integrity of the data (Williams, 2011). A variable correlation graph was built to see if there were any relationships (see Appendix L).

Model Evaluation

In my Milestone 2: Pilot Modification, it was suggested a separation of job roles should be studied. I ran a test under the <Explore> tab (see Appendix M). It shows that sales executives as the role most likely leave.

The **decision tree** below is made from the GE attrition dataset (see Appendix N). At the top, the model elects OVERTIME as being the top variable. According to the ROC Curve, the decision that was created only has a 66% correct prediction rate (Williams, 2011) (see Appendix O).

A **random forest** model was created because the decision tree model has a low percentage rate of correct predictability. The random forest was produced with 500 trees and using 5 variables. The ROC Curve shows that that model has an 81% area under the curve (AUC) (see Appendix P). The test dataset came out with an 80% AUC (see Appendix Q). Both prediction rates are higher than the decision tree model and seem like the model to use in this project.

Areas of Concern

The area of concern is whether the random forest model will be reproducible. The model from DAT 650 performs similarly. Notes were produced this semester so the doubt to produce the results again is diminishing. Variables were taken out and OVERTIME still tops the list for the likelihood of attrition.

Fit of Model

Looking at the sample data is showed little to no difference from the likelihood of attrition. It shows the variables of age, education, hourly rate and job satisfaction. The model picked out the variables that needed attention regarding attrition (see Appendix R).

Model Results

After looking at the results in Appendix R, the obvious variables that could cause attrition were just lukewarm with its results. After running the decision tree and the random forest models, both models used over time as the reason for the likelihood of attrition in GE employees. The error matrix shows a 79.5% error rate (see Appendix S). It comes close to the AUC of 80-81% in random forest ROC Curves (see Appendix P and Appendix Q).

The performance chart shows the truth about the predictability of the random forest model (see Appendix T). The risk score is low at 20%; the recall line is above the regression line giving it a strong validation of being a good model showing the attrition rate (Williams, 2011).

Model Deployment

To keep the currently created model in a virgin state a series of flowcharts and reduced code was produced to make this project reproducible. The flowcharts show the steps that should be taken for data preparation and model creation (see Appendix U). The scaled-down version of R code includes code to load the data into R/Rattle and the code for the random forest that was used for this project (see Appendix V).

Presentation Summary

The presentation that accompanied this report has been submitted under another cover. It covers a summary of the report, with an introduction to General Electric and their problem statement, the steps are taken through CRISP-DM elements, test pilot with results and modifications, plus the new model with data preparation, data exploration, model building, results and recommendations for the use of the random forest predictable model.

Complete Personal and Professional Reflection

During the past year, I have had high anxiety about this course. I wondered what it will involve. Is it an internship? Will I survive it? During this semester I was relieved to know that we were continuing the work that was started last semester during Advanced Data Analytics. When I completed the final project for last semester, it seemed *more* work should be done on the project. I enjoyed following through and finishing the work, this semester. It was challenging.

This Capstone's workload could be done as part of a real-world scenario. The assignments are set up in the order the tasks should be done when working on a similar project, outside of class. The assignments made the final component presentation easier to assemble by having all the work done before the deliverable had to be created.

This semester handed me several firsts. This was the first time I had to write a reflective journal about my assignments, thoughts and work leading up to the completion of them. Also, I had never created a flowchart. Putting a pen to paper to organize the information that needed to be represented in the visuals was helpful. Processes had to be kept in mind when forming the flowcharts. And lastly, this was the first time I had to record my voice in a Microsoft PowerPoint presentation. I am a former radio personality and I found this hard. The timing, the tongue twisters and the slides with the robust scripts were tough. I kept my focus and was able to push through and get it done. I needed to do some research on how to insert my voice into the presentation (Gunnell, 2019).

Being an adult student was an uplifting experience for me at Southern New Hampshire University. It took a bit of discipline to find the time to do the work for the assignments but after about a year I was able to get into a routine. My first semester was the hardest, Foundations in

Statistics. I was worried about my math ability and I applied myself when I realized that data analytics is all about statistics. Understand and live statistics.

As the semesters progressed, I felt I was a natural. The hardest part of the assignments, in my mind, was trying to figure out what was needed from the loosely explained rubrics. I started to believe they were constructed that way leaving students to interpret it in their way. Google is my best friend now. Help was available when I got lost in the Milestones and tried to remember to Keep It Simple.

The past few weeks not only have I been working on the assignments for this semester I have also been working on my professional experience resume. I learned that adding projects conducted in school is the only *experience* I can use on my resume. I wish our program included an internship, for credit or as an option for experience (MySNHU, n.d). As of this week, I got the “thumbs up” from the representative at our Career Center and about 90% done with my Handshake profile. I am meeting up with her soon to make my LinkedIn profile ready to act as an awesome promotional tool for me as a data analytics professional.

Time management was a huge part of being able to finish my assignments on time (Kurzawka, 2018). In the first couple of semesters, I was struggling until I structured my schedule and kept with it every semester. Unfortunately, two years ago I changed jobs and was drained every day from it and found it hard to work on my homework every day. I saved the bulk of the work for the weekends. Waiting for the weekend to finish my work made it hard to ask questions to the professor and get an answer in a reasonable time to incorporate their feedback and I also disappeared from my friends and family because of it. I learned to do scan the module on Monday or Tuesday and get my questions to the professor shortly after that. I ended up taking a 32-hour workweek a year ago and this made it easier for me to complete my assignments.

Continuing my education was an amazing experience. This major did not even exist when I initially attended college. I have thanked my friend who suggests this school and major to me. I found it inspiring and I am excited to start my new mid-life career.

References

Bachar, D. (n.d.) *Descriptive, Predictive and Prescriptive Analytics Explained*. Retrieved from: https://www.logility.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/

Bayern, M (2018, November 2). *How to choose the right data analytics tools: 5 steps*. Retrieved from: https://www.techrepublic.com/article/how-to-choose-the-right-data-analytics-tools-5-steps/

Bekker, A. (2019, May 14). *4 Types of Data Analytics to Improve Decision-Making*. Retrieved from: https://www.scnsoft.com/blog/4-types-of-data-analytics

Brown, M. (n.d.). *Phase 3 of CRISP-DM Process Model: Data Preparation*. Retrieved from: https://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-preparation/

Chauhan, N. (2019, December 24). *Decision Tree Algorithm – Explained*. Retrieved from: https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4

Gunnell, H. (2019, December 19). *How to Record Voiceover Narration in PowerPoint*. Retrieved from: https://www.howtogeek.com/449836/how-to-record-voiceover-narration-in-powerpoint/

HackerEarth. (2017, April 19). *Naïve Bayes Theorem | Introduction to Naïve Bayes Theorem | Machine Learning Classification*. Retrieved from: https://www.youtube.com/watch?time_continue=1&v=sjUDlJfdnKM&feature=emb_logo

Hasan, E. (2020, February 27). *3 Principles of Stakeholder Management for Data Scientist*. Retrieved from: https://retina.ai/blog/3-principles-of-stakeholder-management/

Imanuel, (n.d.). *What is Predictive Analytics?* Retrieved from: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/

Kurzawka. K. (2018, July 12). *15 Most Effective and Proven Time Management Techniques*. Retrieved from: https://www.timecamp.com/blog/2018/07/15-effective-proven-time-management-techniques-2020/

Lee, W. (2016, December 24). *An Ethical Approach to Data Privacy Protection*. Retrieved from: https://www.isaca.org/resources/isaca-journal/issues/2016/volume-6/an-ethical-approach-to-data-privacy-protection

Merhar, C. (2020, June 2). *Employee Retention – The Real Cost of Losing an Employee*. Retrieved from: https://www.peoplekeep.com/blog/employee-retention-the-real-cost-of-losing-an-employee

MySNHU. (n.d.). *SNHU Career*. Retrieved from: https://my.snhu.edu/Offices/COCE/SNHUCareer/Pages/createyourresume.aspx

Reuters. (n.d.). *About General Electric Company*. Retrieved from: https://www.reuters.com/companies/GE.N

R-Project.org. (n.d.) *What is R?* Retrieved from: https://www.r-project.org/about.html

Rouse, M. (n.d.) *Descriptive Analytics*. Retrieved from: https://whatis.techtarget.com/definition/descriptive-analytics

SAS. (n.d.). *Predictive Analytics What it is and why it matters*. Retrieved from: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, 5*(4), Fall 2000, 13-22. Retrieved June 22, 2020, from https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf.

University of Missouri-Columbia. (2019, May 29). *Why do top executives leave companies? It's personal, study finds*. Retrieved from: https://phys.org/news/2019-05-companies-personal.html

Williams, G. (2011) *Data mining with rattle and r – the art of excavating data for knowledge discovery*. Spring-verlag.

Yiu, C. (2019, June 12). *Understanding Random Forest*. Retrieved from: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Appendix A

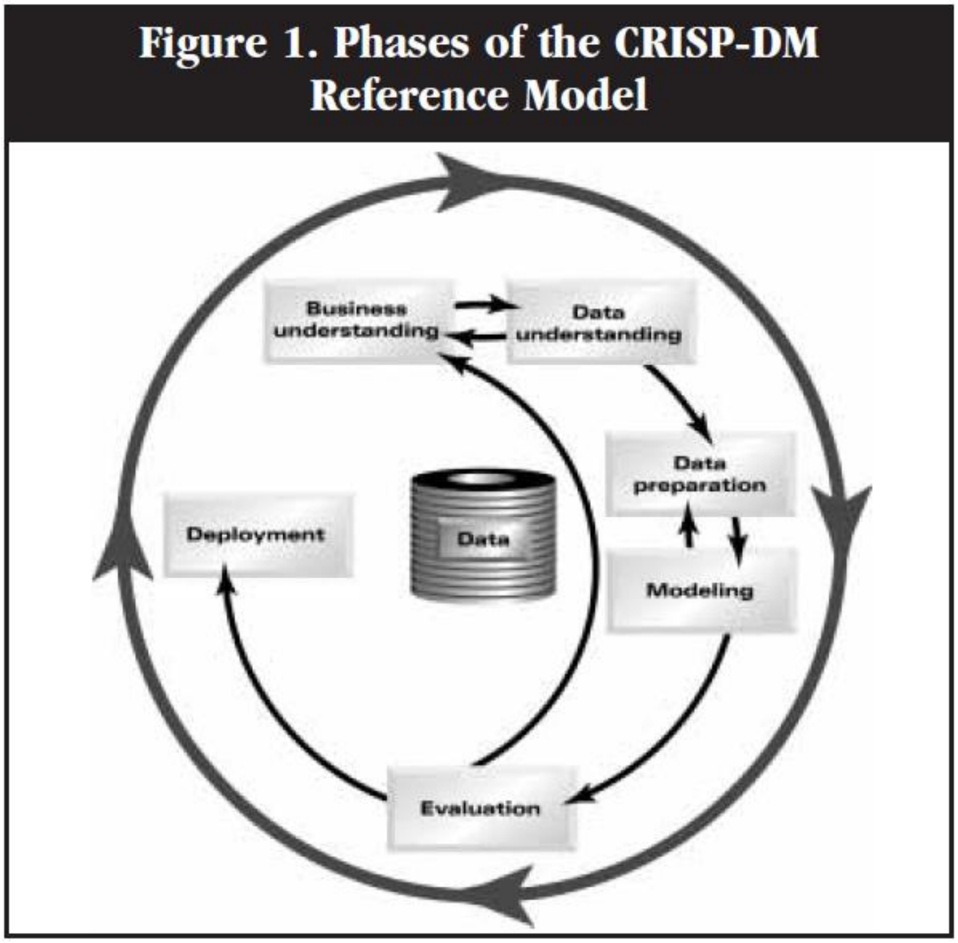
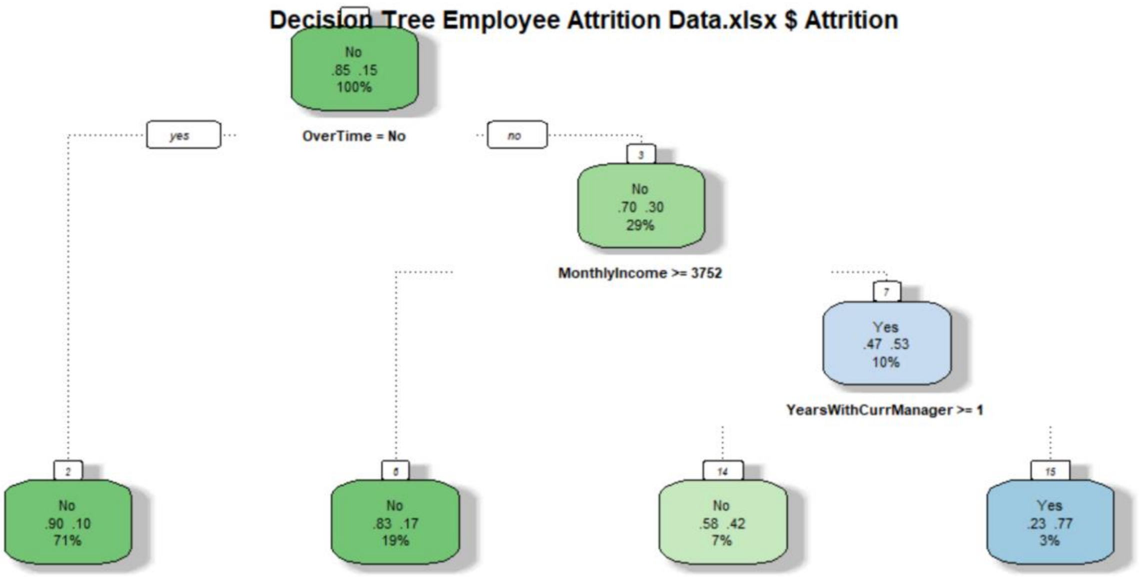


Figure 1. Phases of the CRISP-DM Reference Model

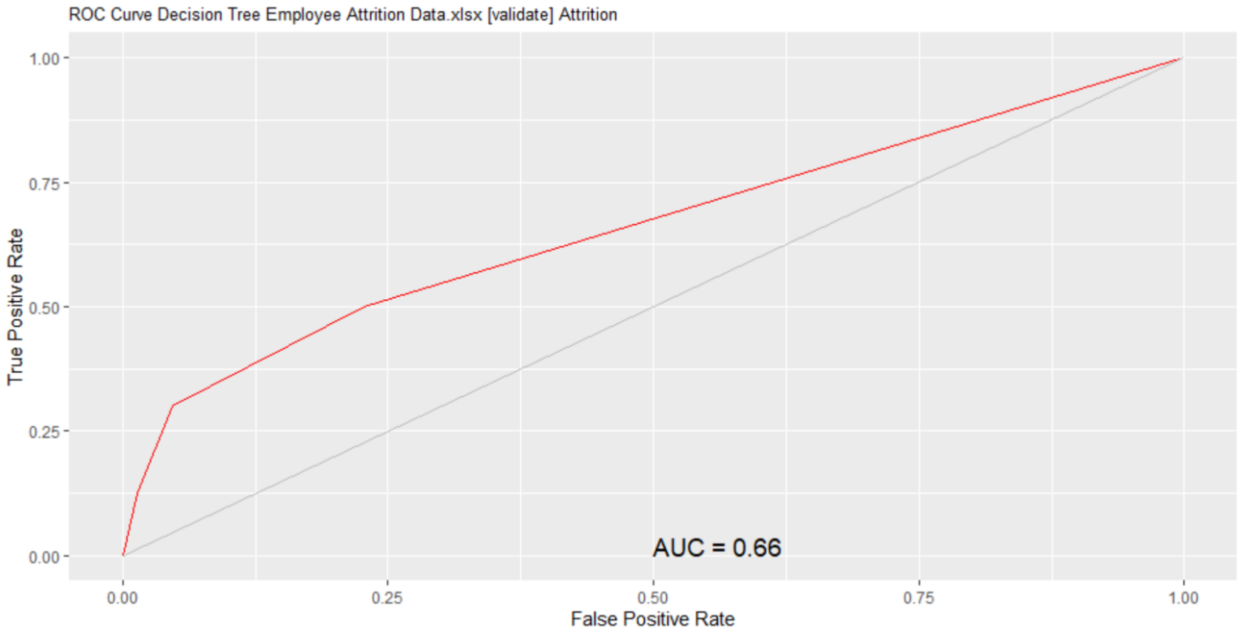
Phases of CRISP-DM Reference Model

Appendix B



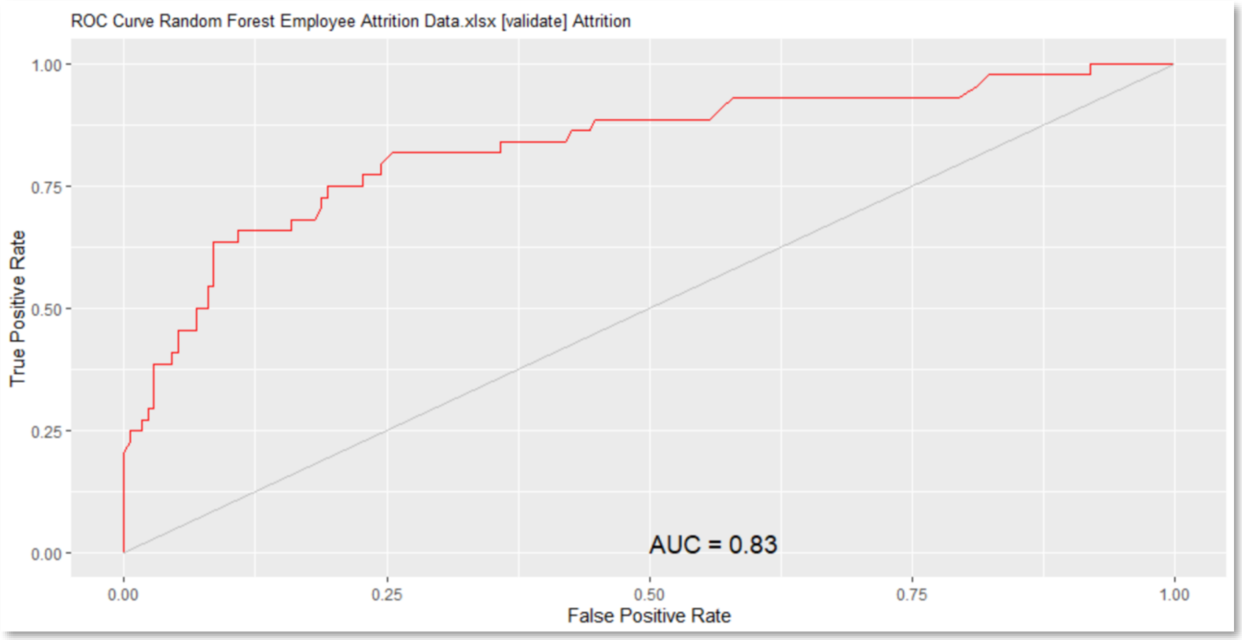
Decision Tree made from Attrition Dataset

Appendix C



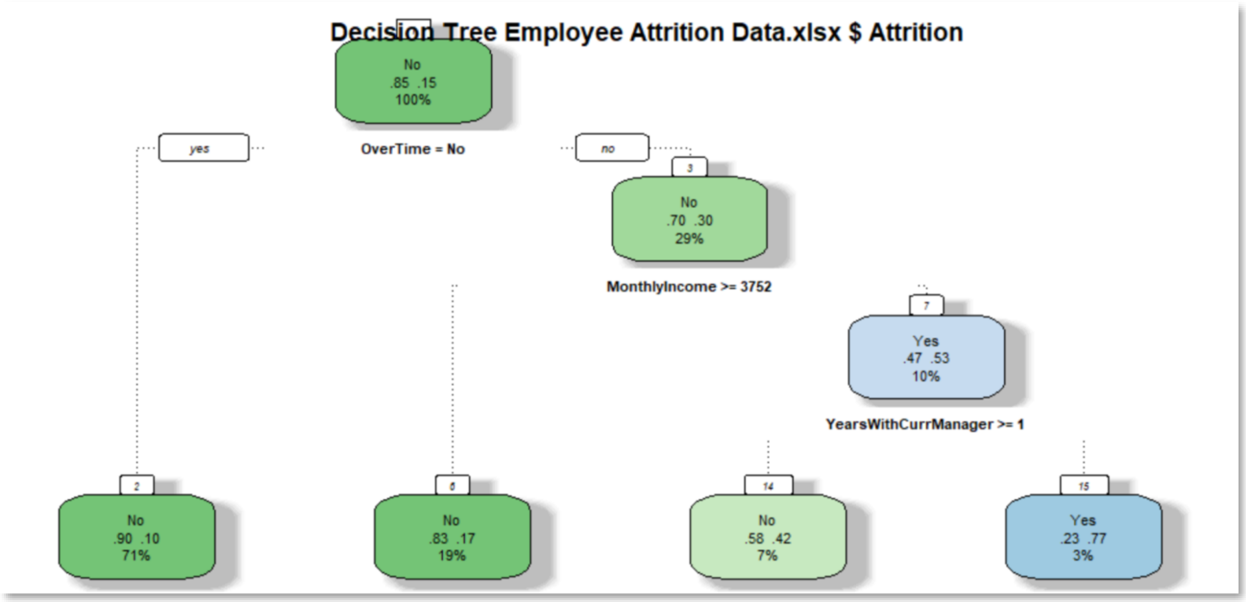
ROC Curve showing the AUC of the Decision Tree Model

Appendix D



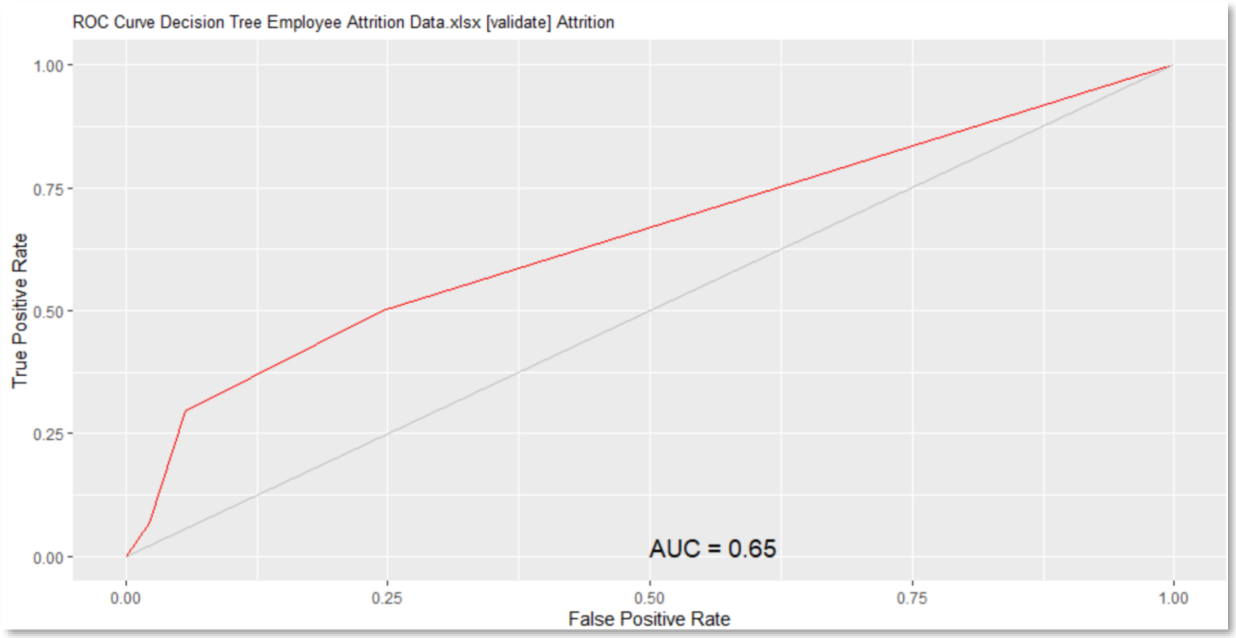
ROC Curve showing the AUC of the Random Forest Model

Appendix E



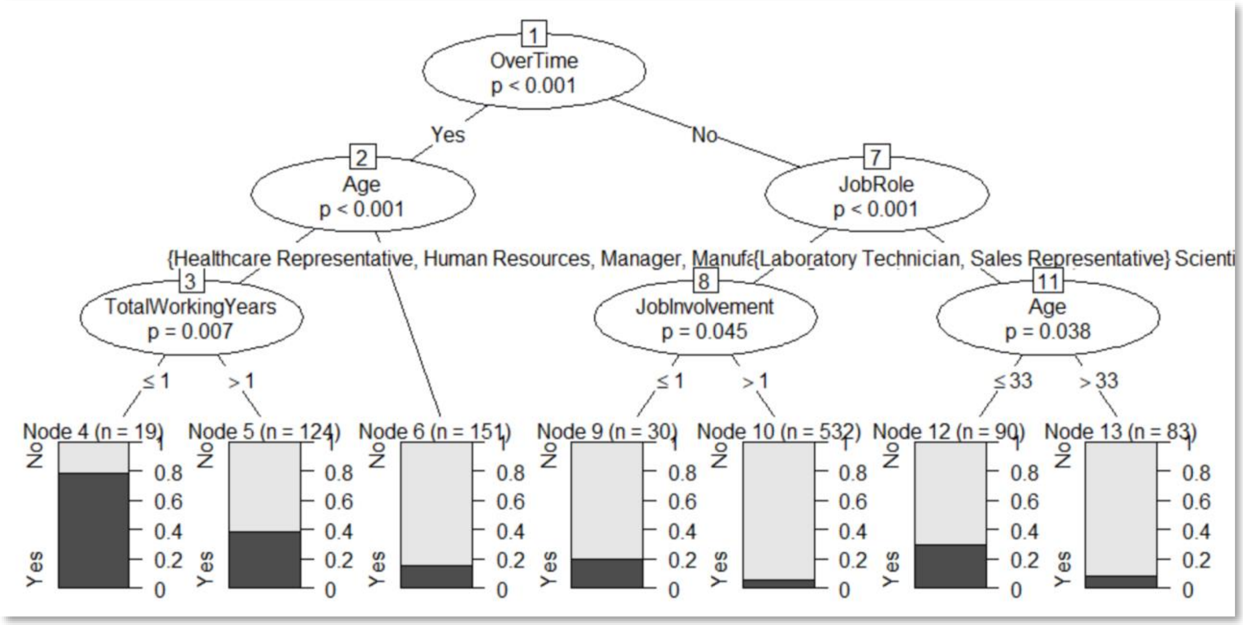
First traditional decision tree with 5 for split and bucket and 3 as maximum depth

Appendix F



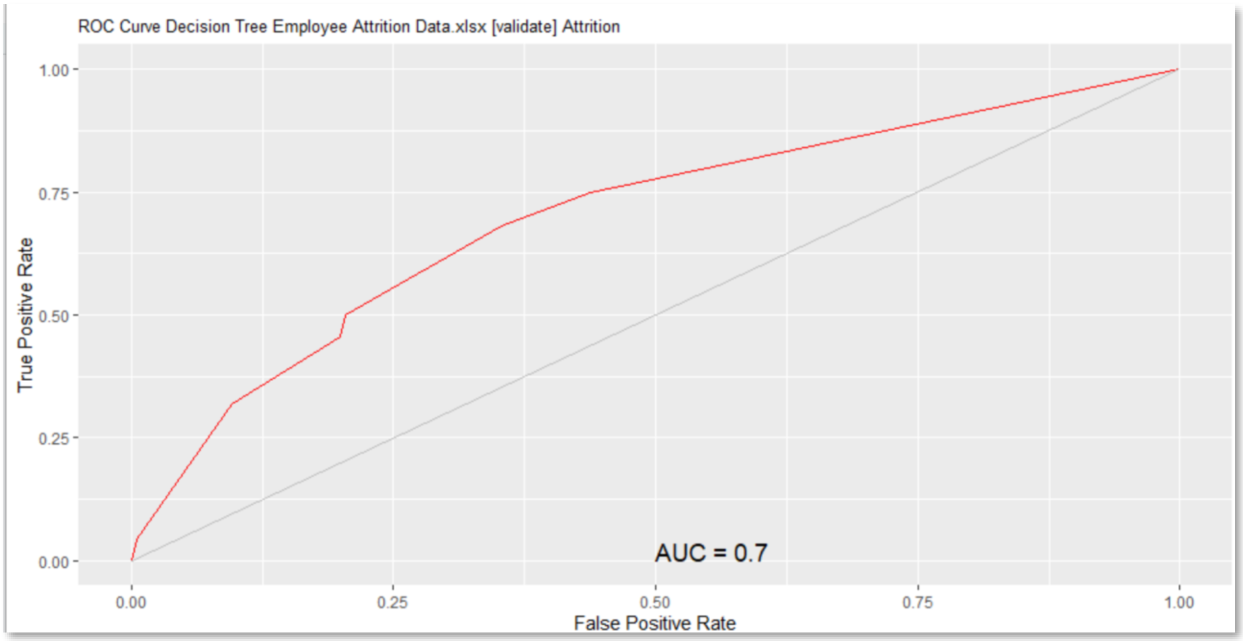
ROC curve from the first decision tree created

Appendix G



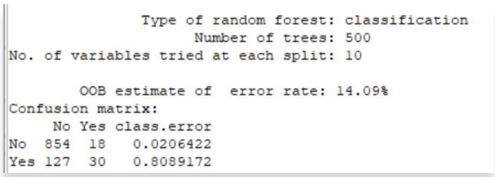
Second conditional decision tree with 5 for split and bucket and 3 as maximum depth

Appendix H

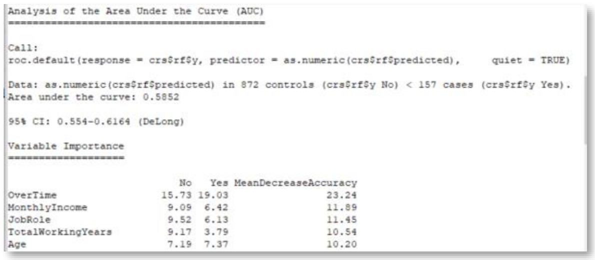


ROC curve from the second decision tree created

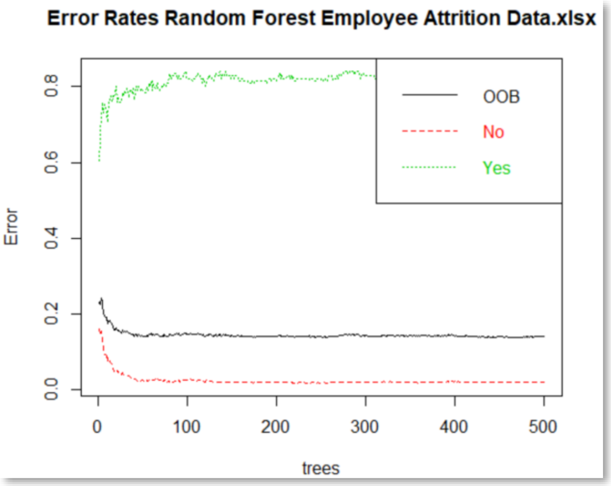
Appendix I

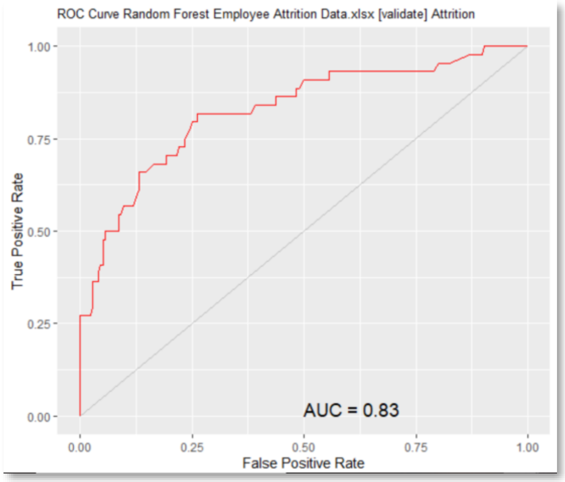


```
                Type of random forest: classification
                      Number of trees: 500
No. of variables tried at each split: 10

        OOB estimate of  error rate: 14.09%
Confusion matrix:
     No Yes class.error
No  854  18   0.0206422
Yes 127  30   0.8089172
```



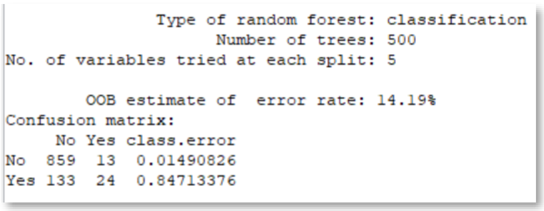
```
Analysis of the Area Under the Curve (AUC)
==========================================

Call:
roc.default(response = crs$rf$y, predictor = as.numeric(crs$rf$predicted),     quiet = TRUE)

Data: as.numeric(crs$rf$predicted) in 872 controls (crs$rf$y No) < 157 cases (crs$rf$y Yes).
Area under the curve: 0.5852

95% CI: 0.554-0.6164 (DeLong)

Variable Importance
===================

                     No   Yes MeanDecreaseAccuracy
OverTime          15.73 19.03                23.24
MonthlyIncome      9.09  6.42                11.89
JobRole            9.52  6.13                11.45
TotalWorkingYears  9.17  3.79                10.54
Age                7.19  7.37                10.20
```



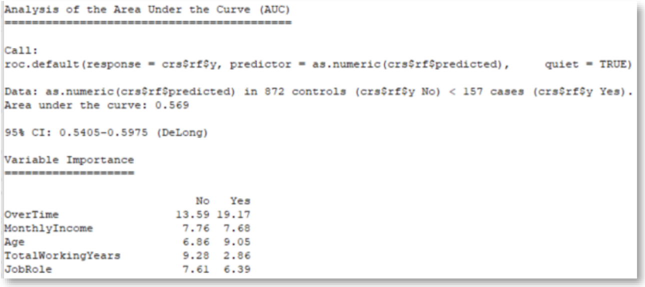


Reports, error rates and ROC chart from the first random forest created

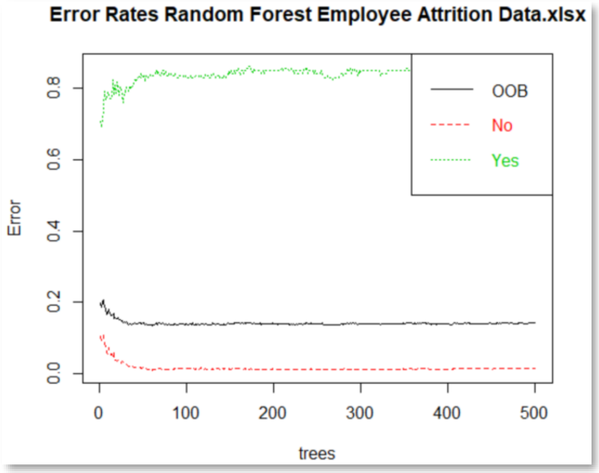
Appendix J

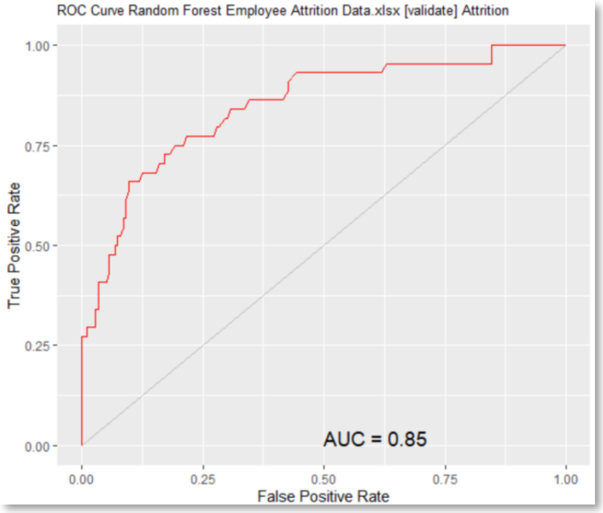


```
                   Type of random forest: classification
                         Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 14.19%
Confusion matrix:
     No Yes class.error
No  859  13  0.01490826
Yes 133  24  0.84713376
```



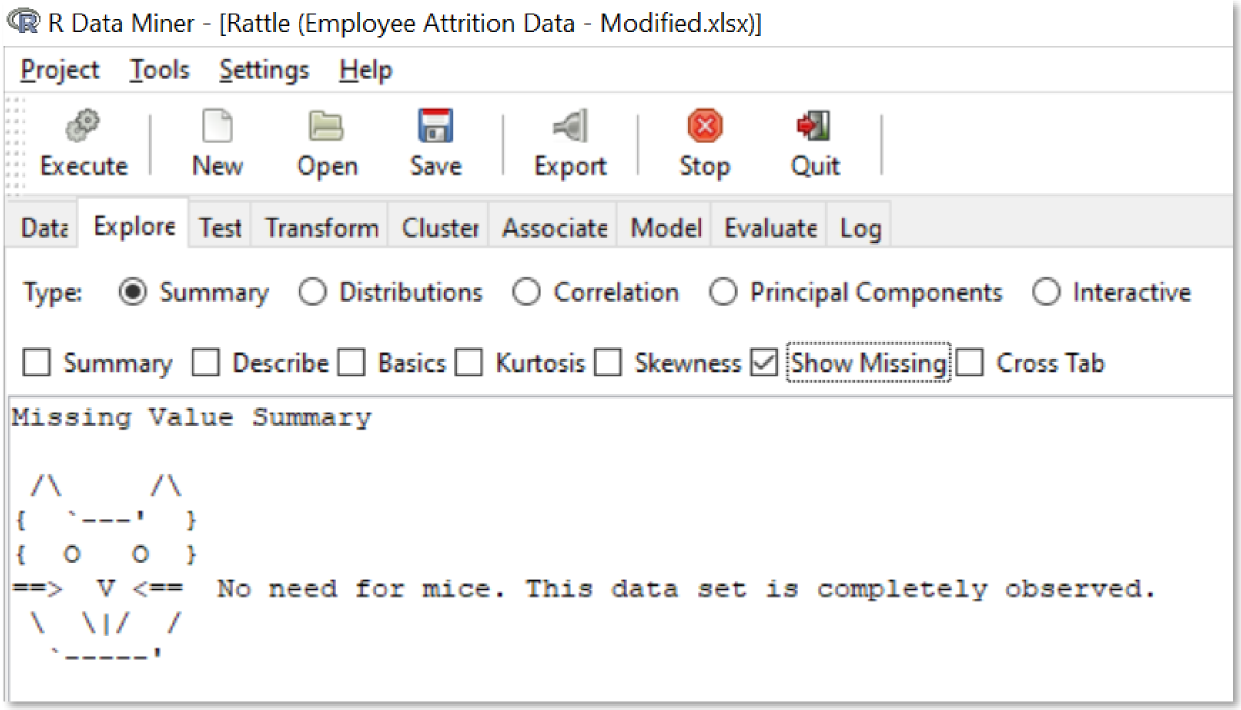
```
Analysis of the Area Under the Curve (AUC)
==========================================

Call:
roc.default(response = crs$rf$y, predictor = as.numeric(crs$rf$predicted),     quiet = TRUE)

Data: as.numeric(crs$rf$predicted) in 872 controls (crs$rf$y No) < 157 cases (crs$rf$y Yes).
Area under the curve: 0.569

95% CI: 0.5405-0.5975 (DeLong)

Variable Importance
===================

                         No   Yes
OverTime              13.59 19.17
MonthlyIncome          7.76  7.68
Age                    6.86  9.05
TotalWorkingYears      9.28  2.86
JobRole                7.61  6.39
```





Reports, error rates and ROC chart for second random forest model

Appendix K



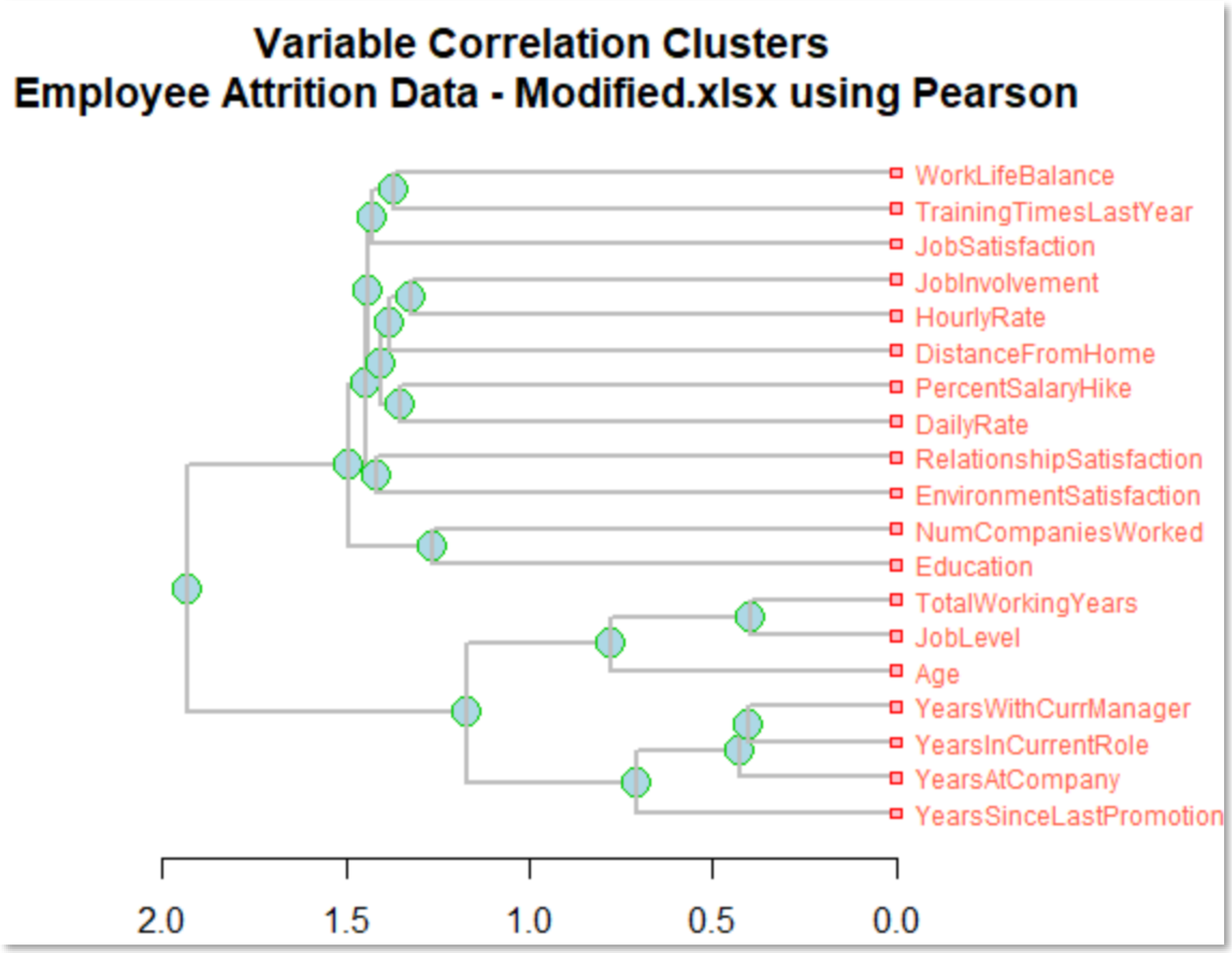
Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

□ Summary □ Describe □ Basics □ Kurtosis □ Skewness ☑ Show Missing □ Cross Tab

```
Missing Value Summary

 /\      /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'
```

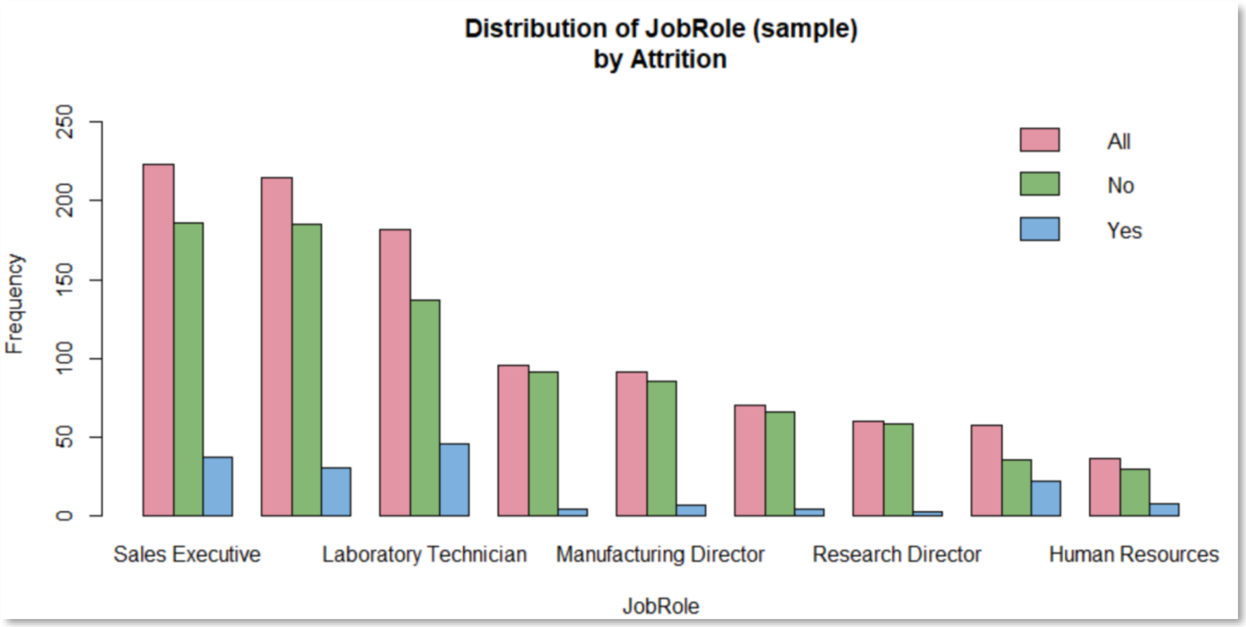
Data Quality, shows no missing values in the dataset

Appendix L



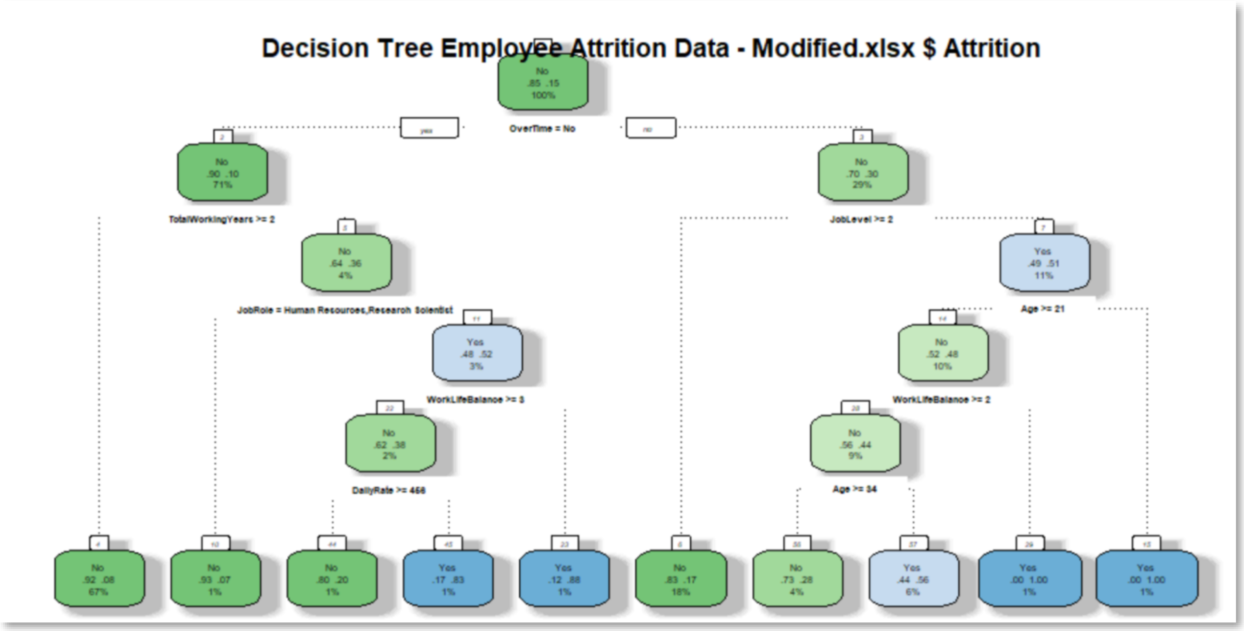
Correlation using the Pearson Hierarchical Method

Appendix M



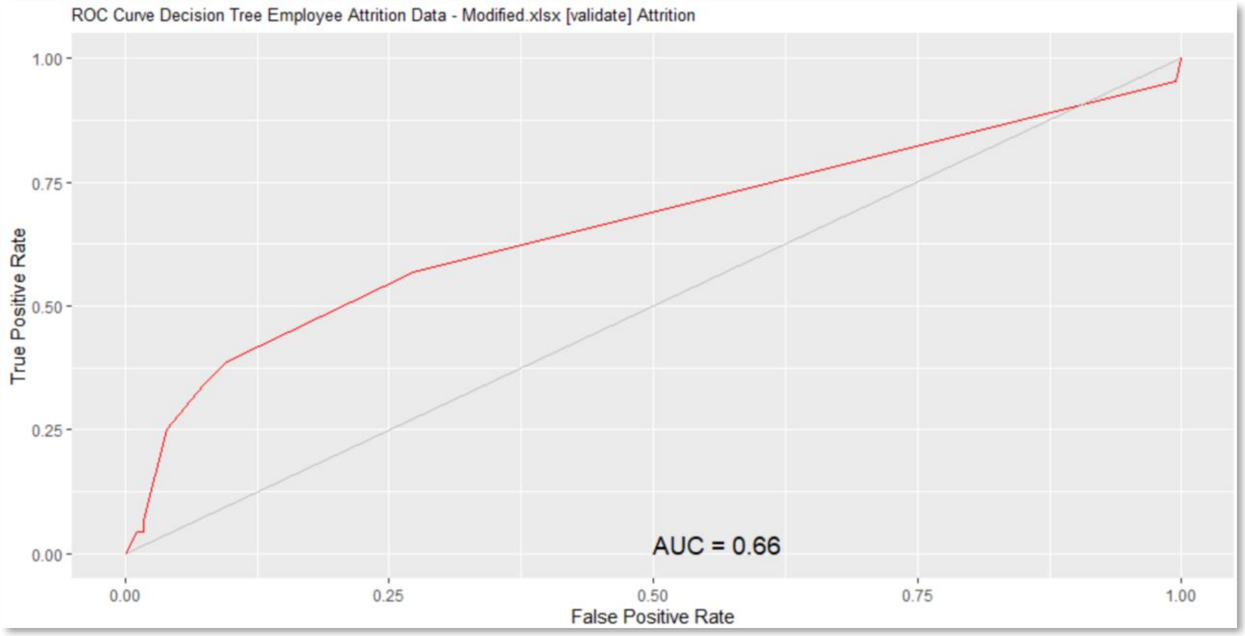
Attrition dataset is broken down into job role

Appendix N



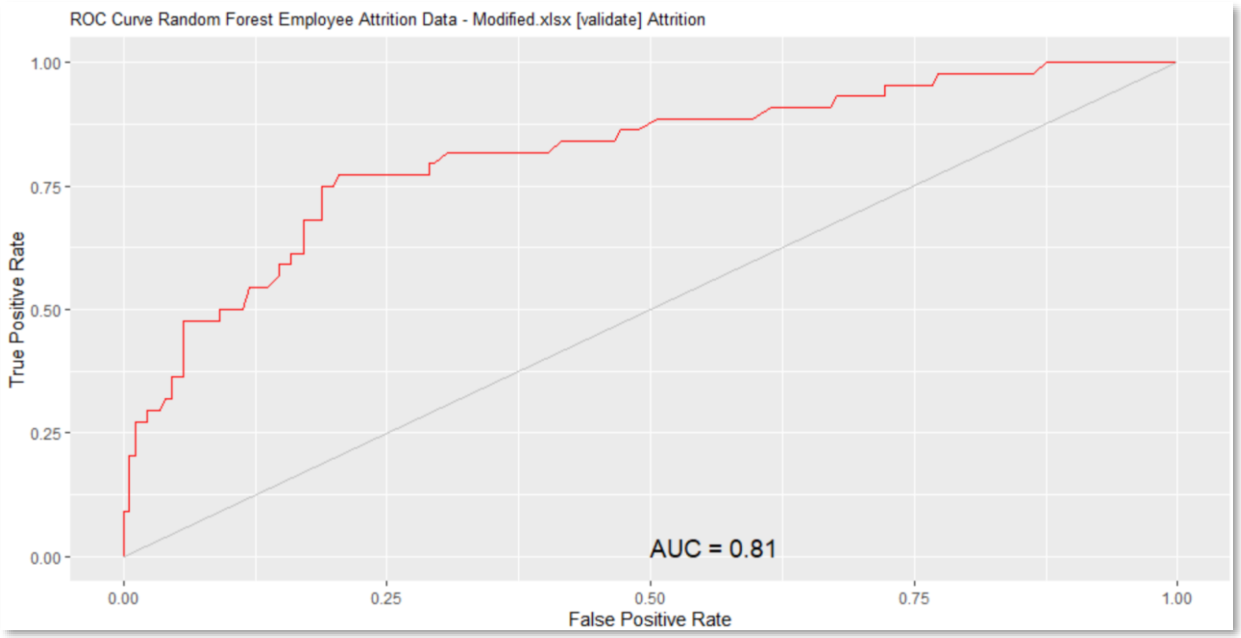
Decision Tree made from attrition dataset

Appendix O



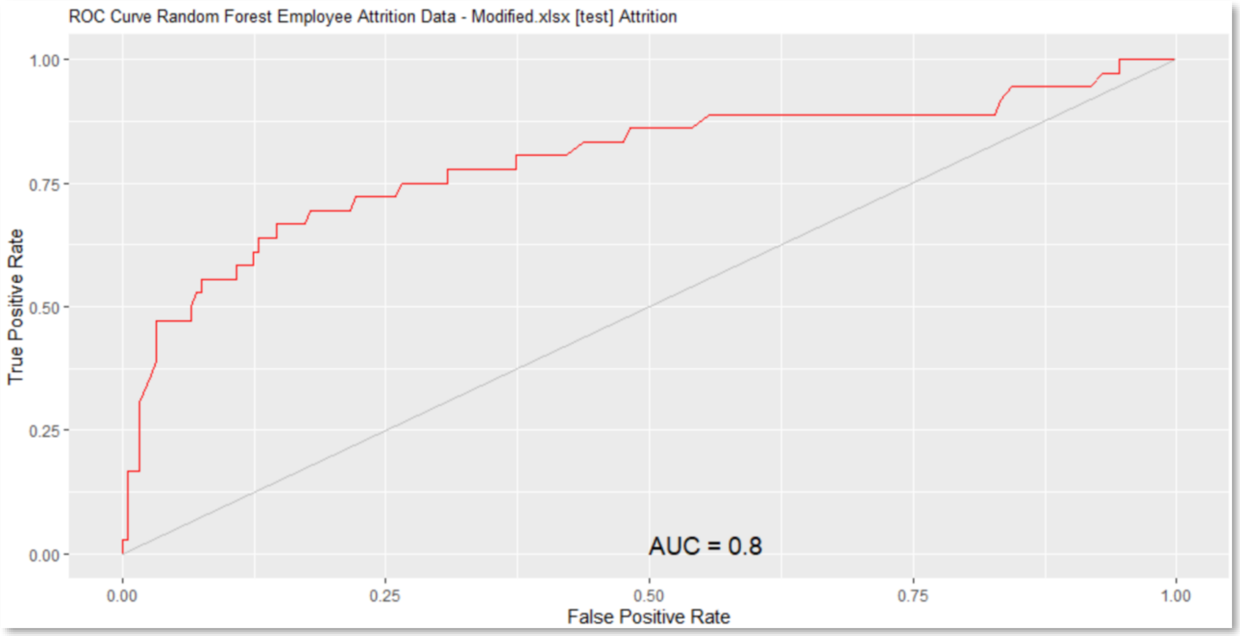
ROC Curve for the decision tree model

Appendix P



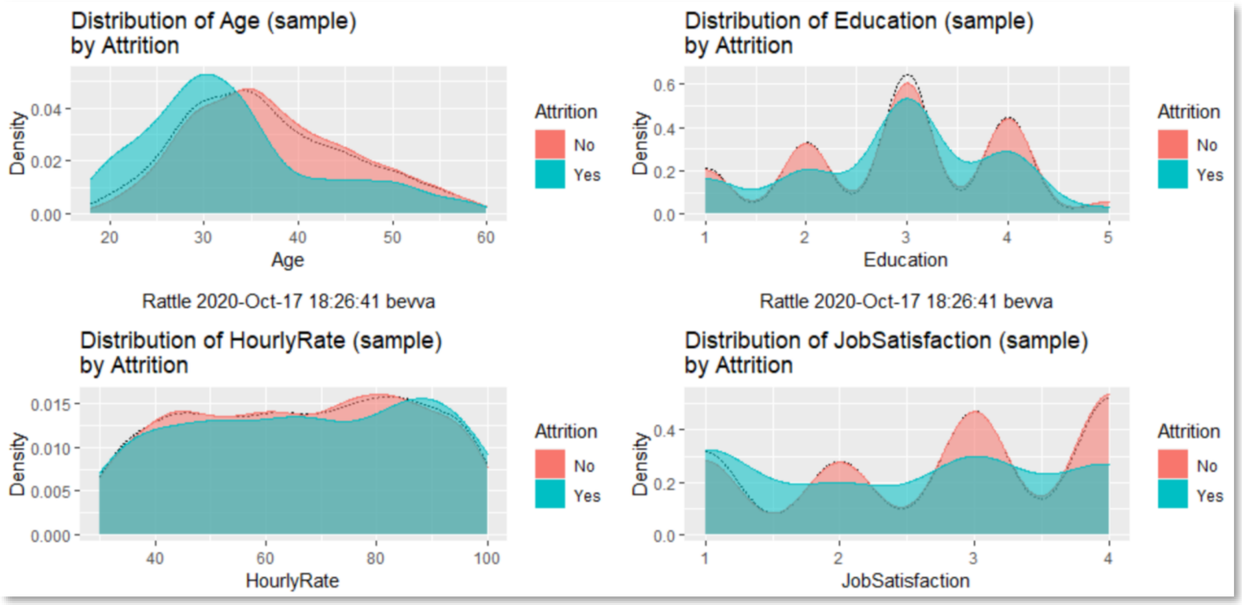
ROC Curve for the random forest model

Appendix Q



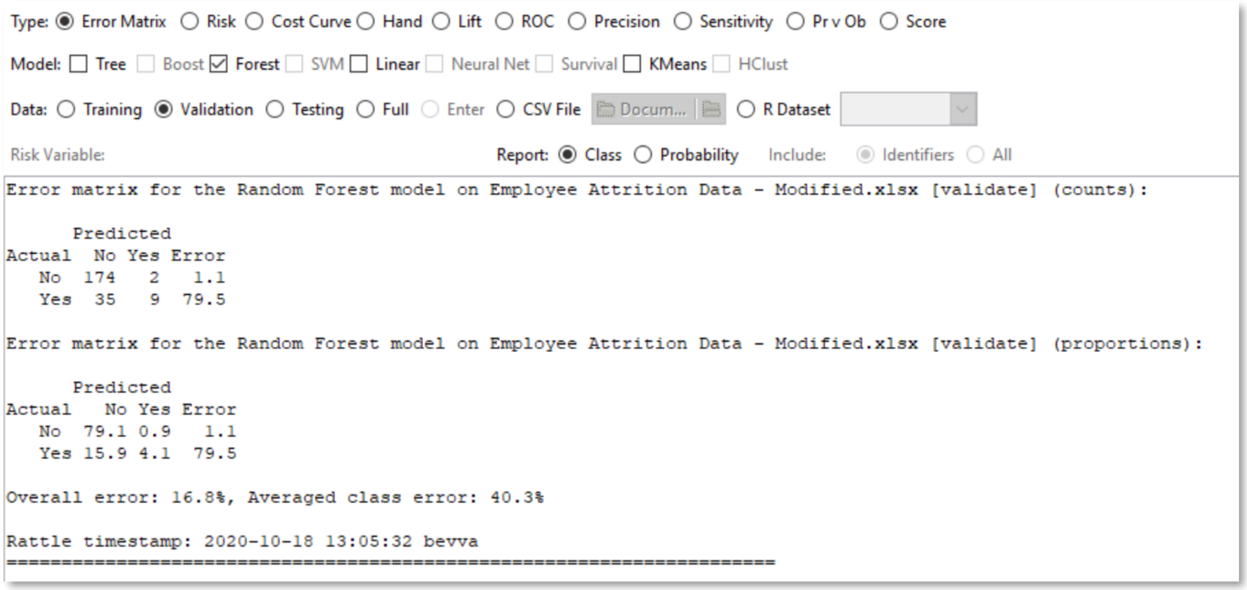
ROC Curve for the random forest model test dataset

Appendix R



Exploring the dataset

Appendix S

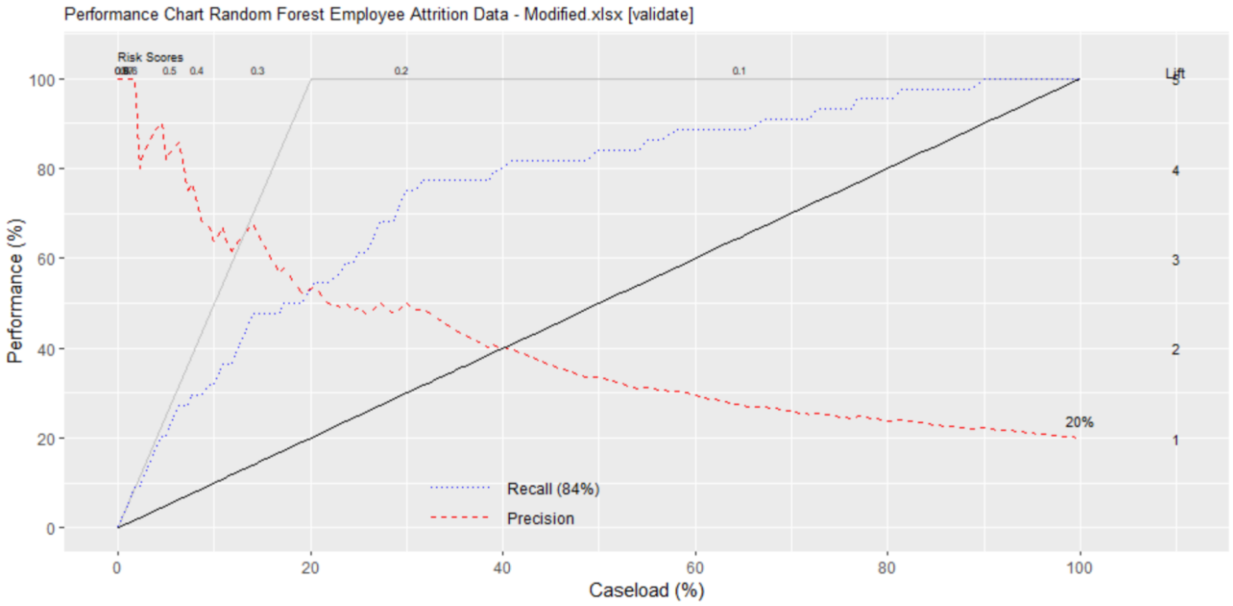


Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score
Model: □ Tree □ Boost ☑ Forest □ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust
Data: ○ Training ● Validation ○ Testing ○ Full ○ Enter ○ CSV File Docum... ○ R Dataset
Risk Variable:
Report: ● Class ○ Probability Include: ● Identifiers ○ All

```
Error matrix for the Random Forest model on Employee Attrition Data - Modified.xlsx [validate] (counts):

      Predicted
Actual  No Yes Error
   No  174   2   1.1
   Yes  35   9  79.5

Error matrix for the Random Forest model on Employee Attrition Data - Modified.xlsx [validate] (proportions):

      Predicted
Actual   No Yes Error
   No  79.1 0.9   1.1
   Yes 15.9 4.1  79.5

Overall error: 16.8%, Averaged class error: 40.3%

Rattle timestamp: 2020-10-18 13:05:32 bevva
======================================================================
```

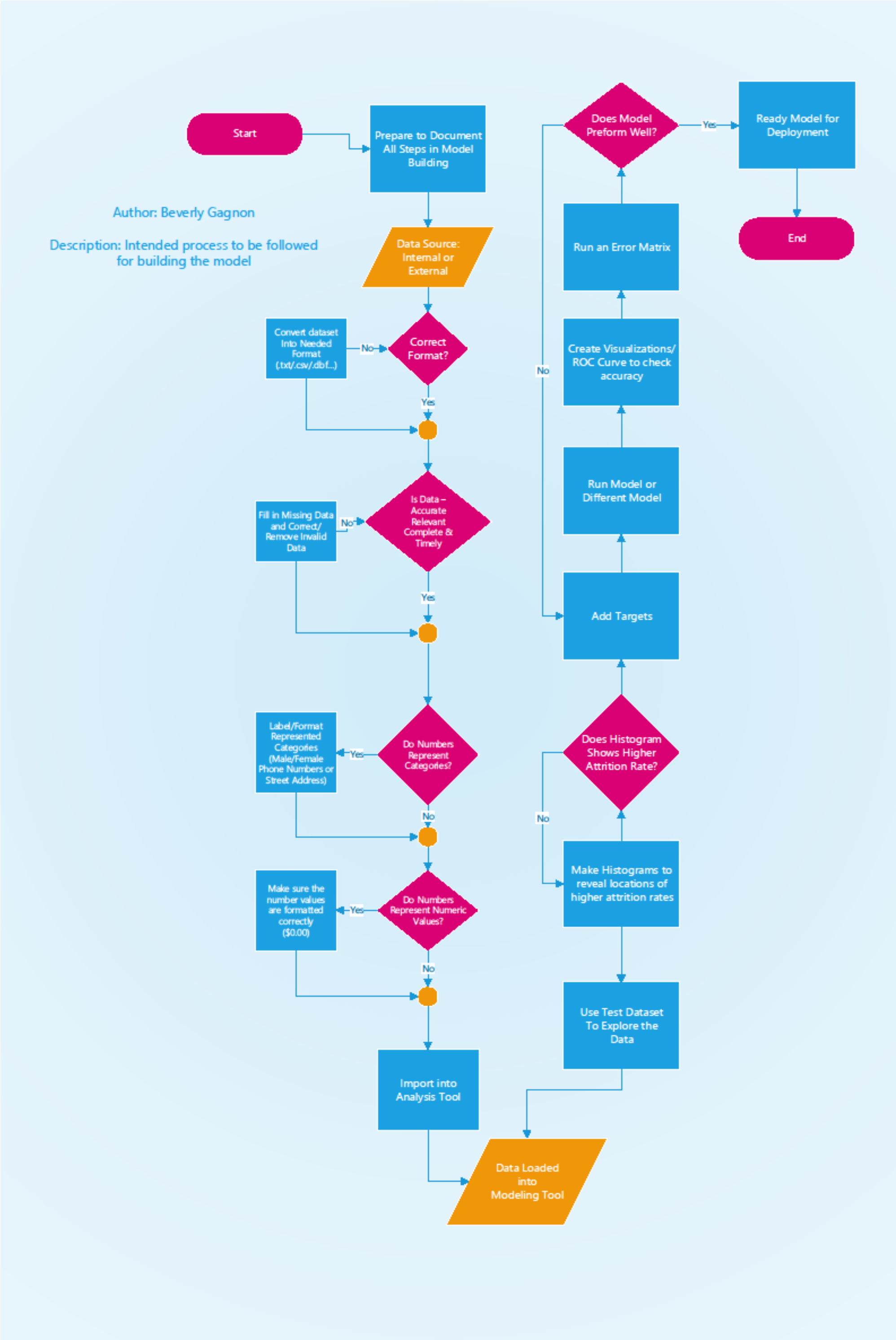
Error Matrix of the random forest model

Appendix T



Performance chart from the random forest model

Appendix U


Start
Prepare to Document All Steps in Model Building
Author: Beverly Gagnon
Description: Intended process to be followed for building the model
Data Source: Internal or External
Convert dataset Into Needed Format (.txt/.csv/.dbf...)
No
Correct Format?
Yes
Fill in Missing Data and Corect/ Remove Invalid Data
No
Is Data – Accurate Relevant Complete & Timely
Yes
Label/Format Represented Categories (Male/Female Phone Numbers or Street Address)
Yes
Do Numbers Represent Categories?
No
Make sure the number values are formatted correctly ($0.00)
Yes
Do Numbers Represent Numeric Values?
No
Import into Analysis Tool
Data Loaded into Modeling Tool
Use Test Dataset To Explore the Data
Make Histograms to reveal locations of higher attrition rates
No
Does Histogram Shows Higher Attrition Rate?
Add Targets
Run Model or Different Model
Create Visualizations/ ROC Curve to check accuracy
Run an Error Matrix
No
Does Model Preform Well?
Yes
Ready Model for Deployment
End

Flowchart show steps to be taken for data preparation and model creating

Appendix V

```
# Load a dataset from file.

library(readxl, quietly=TRUE)

crs$dataset <- read_excel("C:/Users/bevva/Desktop/School File/Employee
Attrition Data - Modified.xlsx", guess_max=1e4)

crs$dataset

#=======================================================================

# Action the user selections from the Data tab.

# Build the train/validate/test datasets.

# nobs=1470 train=1029 validate=220 test=221

set.seed(crv$seed)

crs$nobs <- nrow(crs$dataset)

crs$train <- sample(crs$nobs, 0.7*crs$nobs)

crs$nobs %>%
seq_len() %>%
setdiff(crs$train) %>%
sample(0.15*crs$nobs) ->
crs$validate

crs$nobs %>%
seq_len() %>%
setdiff(crs$train) %>%
setdiff(crs$validate) ->
crs$test

# The following variable selections have been noted.

crs$input     <- c("Age", "BusinessTravel", "DailyRate",
"Department", "DistanceFromHome", "Education",
"EducationField", "EnvironmentSatisfaction",
"Gender", "HourlyRate", "JobInvolvement",
"JobLevel", "JobRole", "JobSatisfaction",
"MaritalStatus", "NumCompaniesWorked", "OverTime",
"PercentSalaryHike", "RelationshipSatisfaction",
"TotalWorkingYears", "TrainingTimesLastYear",
"WorkLifeBalance", "YearsAtCompany",
"YearsInCurrentRole", "YearsSinceLastPromotion",
"YearsWithCurrManager")
```

```
crs$numeric   <- c("Age", "DailyRate", "DistanceFromHome",
"Education", "EnvironmentSatisfaction",
"HourlyRate", "JobInvolvement", "JobLevel",
"JobSatisfaction", "NumCompaniesWorked",
"PercentSalaryHike", "RelationshipSatisfaction",
"TotalWorkingYears", "TrainingTimesLastYear",
"WorkLifeBalance", "YearsAtCompany",
"YearsInCurrentRole", "YearsSinceLastPromotion",
"YearsWithCurrManager")

crs$categoric <- c("BusinessTravel", "Department",
"EducationField", "Gender", "JobRole",
"MaritalStatus", "OverTime")

crs$target    <- "Attrition"
crs$risk      <- NULL
crs$ident     <- NULL
crs$ignore    <- NULL
crs$weights   <- NULL

#=======================================================================

# Build a Random Forest model using the traditional approach.

set.seed(crv$seed)

crs$rf <- randomForest::randomForest(Attrition ~ .,
data=crs$dataset[crs$train, c(crs$input, crs$target)],
ntree=500,
mtry=5,
importance=TRUE,
na.action=randomForest::na.roughfix,
replace=FALSE)

# Generate textual output of the 'Random Forest' model.

crs$rf

# The `pROC' package implements various AUC functions.

# Calculate the Area Under the Curve (AUC).

pROC::roc(crs$rf$y, as.numeric(crs$rf$predicted), quiet=TRUE)

# Calculate the AUC Confidence Interval.

pROC::ci.auc(crs$rf$y, as.numeric(crs$rf$predicted), quiet=TRUE)FALSE

# List the importance of the variables.

rn <- round(randomForest::importance(crs$rf), 2)
rn[order(rn[,3], decreasing=TRUE),]
```