DAT 690- Capstone Project

Data analytics Solution

Duncan Thompson

2021

# Project Summary & Analytic Plan

## CRISP-DM Business Understanding Phase: Business Problem

General Electric Company (GE) is one of the most diversified corporations in the world. They produce products that include electronic and electrical equipment, aircraft engines, and financial services (“General Electric | History, Acquisitions, Products, & Facts | Britannica,” 2020).

GE’s Human Resources department has noticed a spike in job postings, costing the company most due to the top-talented individuals' attrition. The main issue that we would like to address is the attrition rate of employees and the cost of attrition—the issue is retaining top-tier and top-performing employees in all company business lines.

## CRISP-DM Business Understanding Phase: Research Question

The primary research objective is to understand and identify the attrition of employees that may leave. The predictive model is going to guide by answering and solving the following questions:

#### What is the likelihood of attrition?

#### What employees are leaving within the first year of employment, and what are the major drivers?

* What policies or strategies can be adopted based on the results to improve employee attrition?

## CRISP-DM Business Understanding Phase: Business Solution

Building the predictive model to address the research questions would allow GE to retain top-tier talent and reduce the cost of attrition. By retaining these employees, the company can remain competitive in all industries that they produce products in and attract other top-performing individuals that would allow the company to have a more diverse group of intellect to move into more industries. The cost of attrition reduction brings value to the company on a monetary scale, which would allow the salaries to be increased and better employee engagement efforts to stay ahead of voluntary attrition.

## CRISP-DM Data Understanding

The data provided by General Electric Company (GE), the plan is to look at the data at a high level. We are identifying the data types and how many rows of data that we have access to. The data examination will uncover whether we can use the data to meet all of the requirements. After describing the data, the exploration task is performed by identifying the target variable/attribute and using statistical analyses to address the data mining goals. Transformation of the data may stem from this step. Verifying the quality of the data is next. Data quality verification will use the exploration data analysis report to ensure the information does not contain errors or missing values.

## CRISP-DM Data Preparation

Select the columns that relevant to the data mining goals and would add value to the model. Next, the data transformation from numerical to categorical data types is performed. The construction of new columns is unnecessary for this project and integrating them into aggregated columns is not required.

## CRISP-DM Modeling

We have come to the point where we select the best modeling technique during the project phase, which we have two to choose from (Classification or Regression). A detailed description of the modeling technique(s) documents the modeling technique's assumptions (s). A test design is then built based on the modeling technique, and quality is tested a run of the model on the prepared dataset to create one or more models. The results are described and interpreted. Lastly, for this part of the project, there is an evaluation of the model to check the performance, and based on the performance, we will know if we have met the data mining goals.

## CRISP-DM Model Evaluation

The model evaluation plan evaluates the model's performance and accuracy to ensure that we have achieved the business goals. We are generating multiple models using different methods to evaluate each model's effectiveness—the application of different business rules for each model using various measures to determine each model's sustainability. We will also evaluate whether the model has to be more sensitive than specific or more specific than sensitive. Suppose the results of the models are efficient and satisfy the business needs. In that case, a document is developed to summarize the process and highlight activities missed, which should be repeated. When determining the next steps in the project, the project team will review the process and decide if it moves on to deployment. We are going to list the possible actions along with reasons for or against each option.

## CRISP-DM Deployment

The final model will work in production by writing a deployment plan that summarizes the necessary steps and performs them. The model will need to have to be monitored and may be retrained. These steps would have to be documented in detail. After the deployment and monitoring plan is written, a final report is written and presented to the customers. At the end of the project, there will be a retrospective period to identify what went wrong, what went well, and what could have been improved.

# ****Data Understanding and Data Preparation****

## Data Understanding

The data is taken from the data warehouse and maintained by HR staff. The data was extracted into an Excel spreadsheet with pre-approved columns to be extracted.

There are 13 numerical variables and 22 categorical variables. The next step is to convert the categorical variables to numerical variables because the target variable is the Attrition column, a categorical variable. The data does not contain any missing values, but variables must be changed, so data cleansing is required.



Figure :Sample Data

In a review of the data, EmployeeCount, Over18, and StandardHours have only one distinct value, and EmployeeNumber has all distinct values because it is a unique identifier. These features are removed because they are redundant and do not bring any value to the model results. The Performance rating metrics have two values because we want to predict the top-performing employees to identify why they are attrition. Based on a linear correlation coefficient measure, the most positively correlated features are MonthlyRate, HourlyRate, NumCompaniesWorked, distance from home, and Target (Attrition). The ex-employees' average ages are 33.8, and the current employees are 37.5 based on the target attribute. The percentage of the employees left was a majority on the Human Resources and Technical Degrees, 25% of the 204 ex-employees. The single employees have the highest attrition than Married or Divorced. Loyal employees have a higher salary, and people who live further away from work have a higher attrition rate than those within proximity to work.

## Data Preparation

The first step is to clean the data by label encoding the columns with 2 or fewer unique values. Then converting the categorical variables into dummy variables these tasks creates more attributes. The next stage is to scale the features using MinMaxScaler to shrink the range between 0 and 5 because machine learning algorithms perform better when numerical input variables fall within a scale.

Recursive feature elimination using cross-validation is used for feature selection. The determination is to use 31 of 42 features to run the model with the 10 stratified K-fold accuracy.

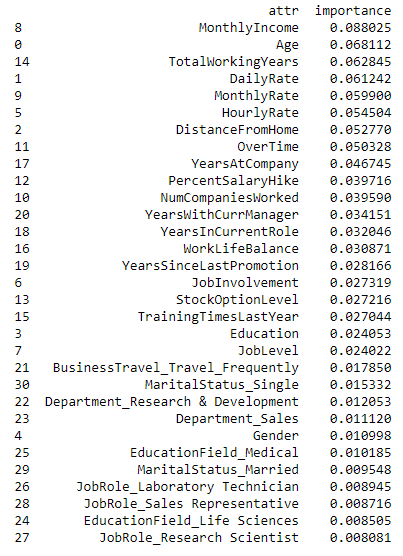


Figure : List of Features for Model

Figure

# Modeling & Evaluation

**CRISP-DM Modeling Phase: Data Quality**

The data quality produced by the analytic plan's implementation follows all of the data rules established during the business understanding phase. We manipulated the categorical data type to numerical. The data does not have any missing values.

**CRISP-DM Modeling Phase: Data Structure**

The data is structured and pulled directly from a database. Logistic regression models use quantitative data to predict employee attrition. After running the model, the data structure did not change. We are also classifying which employees will leave the company. The data was not scaled to fit the model.

**CRISP-DM Evaluation Phase: Model Evaluation**

The model that was run to identify employee attrition areas exceeded the accuracy score that we established for the project. Using the model, a company can determine which areas to look after to make workplace decisions and feel confident. Also, the logistic regression model exceeds the goal of having a model with an accuracy range of 70%. That was the success criteria set at the beginning of the project. The success rate was in the 80% range.

**CRISP-DM Evaluation Phase: Area of Concern**

After running the model and perform a stepwise regression to identify the features/variables that significantly impacted the model's prediction, I found that salary was not a contributing factor for employee attrition. It was strange because the median hourly income is $4,946, with an employee's average years at the company being five years. This finding or concern was more apparent when I saw that one of the driving factors for employee attrition was the job role of a sales representative or job role in general.

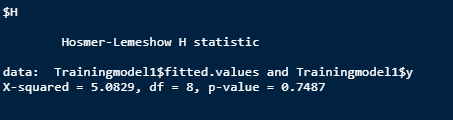
**CRISP-DM Evaluation Phase: Fit of Model**

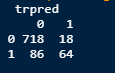
The hypothesis testing for the fit of the model was performed using the AUC score and the Hosmer-Lemeshow test, a function in RStudio. The H0 is "The model is a good fit," and H1 is "The model is not a good fit" if the p-value>0.05, we will accept H0 and reject H1. The Hosmer-Lemeshow test results showed that the p-value = 0.7487, which means that the model was a good fit. Also, to ensure that the model is a good fit, an AUC score was used, and a ROC curve was plotted (See Artifacts section for ROC curve). The AUC score for the training data set was 0.8684, and for the testing, data set the score was 0.8178. Since the AUC score is close to 1, we can feel confident that the model's data is a good fit. We can improve the model by changing the independent variables that we indicated by the stepwise function and feature selection.

**CRISP-DM Evaluation Phase: Model Results**

The model performed well because the training data model had an accuracy of 88.26% and when the testing data was input into the model the accuracy was 89.58%. The model fitted with the testing data had few false positives. The model predicted correctly for the testing data of true positives 314 and true negatives 30. The Precision rate of the test model is 95% and the Recall of the model 93%. So with these statistics we know that when the model predicts attrition that 95% of the time it is correct.

**CRISP-DM Modeling Phase: Artifacts**





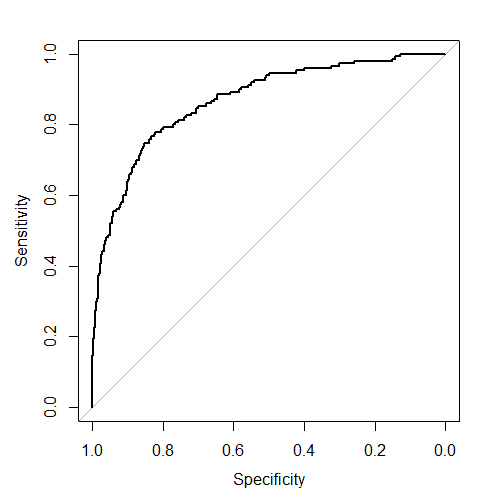


Figure : Training Data ROC Curve

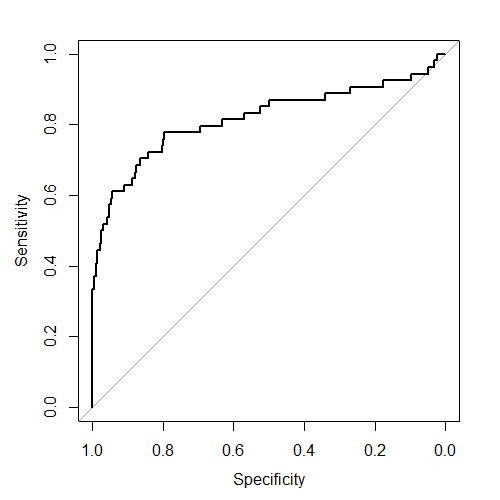


Figure : Testing Data ROC Curve

Appendix



Figure : Data Analytic Solution flowchart

1. This General Linear Model will predict the employees that will leave the company and the what causes the employees to leave. Our measure of a successful model is that the accuracy of the model is within the 70% range and to have enough information to develop an employee retention plan.
3. ## Step 1:Import the data
4. ```{r}
5. library(readxl)
6. library(caret)
7. library(MKmisc)
8. library(pROC)
9. library(Metrics)
10. library(ggplot2)
11. library(ROCR)
12. library(InformationValue)
13. library(cvms)
14. EmployeeData<-read\_excel("C:/Users/Hi/Desktop/Southern New Hampshire University/DAT 690 - Capstone Course/EmployeeData.xlsx")
15. head(EmployeeData)
16. summary(EmployeeData)
17. ```
18. ### Step 1a: Check for missing values
19. ```{r eval=FALSE, include=FALSE}
20. anyNA(EmployeeData)
21. prop.table(table(EmployeeData$Attrition))
22. ```
23. ## Step 2: Feature Engineering
24. The next step is to change the data types from character to factors and changing the Target Variable(Attrition) from Yes - employee left the company/ No - employee did not leave the company to (1-Yes/0-No).
25. ```{r}
26. EmployeeData$Attrition[EmployeeData$Attrition == "Yes"] <- 1
27. EmployeeData$Attrition[EmployeeData$Attrition == "No"] <- 0
28. EmployeeData$Attrition <- as.numeric(EmployeeData$Attrition)
29. EmployeeData[,c(2,4,6,7,11,15,17,22)]=lapply(EmployeeData[,c(2,4,6,7,11,15,17,22)],as.factor)
30. EmployeeData$Over18[EmployeeData$Over18 == "Y"] <- 1
31. EmployeeData$Over18 <- as.numeric(EmployeeData$Over18)
32. ```
33. ## Step 3: Splitting the data into "training" and "testing" datasets
34. -With the training data set I will build up the model and test its accuracy using the Test Data set.
35. ```{r}
36. set.seed(1000)
37. ranuni=sample(x=c("Training","Testing"),size=nrow(EmployeeData),replace=T,prob=c(0.7,0.3))
38. TrainingData=EmployeeData[ranuni=="Training",]
39. TestingData=EmployeeData[ranuni=="Testing",]
40. nrow(TrainingData)
41. nrow(TestingData)
42. ```
43. \* The above code shows that we have successfully split the entire data set into two parts. Now we have 886 Training data and 384 Testing data.
45. ## Step 4: Building the Model
46. + 4a. Identify the independent variables or the predictors
47. + 4b. Incorporate the dependent variables or target "Attrition" in the model
48. + 4c. Transform the data type of the model from character to formula
49. + 4d. Incorporate Training data into the formula and build the model
50. ```{r}
51. independentvariables=colnames(EmployeeData[,2:35])
52. independentvariables
53. Model=paste(independentvariables,collapse="+")
54. Model
55. Model\_1=paste("Attrition~",Model)
56. Model\_1
57. class(Model\_1)
58. formula=as.formula(Model\_1)
59. formula
60. ```
61. \* Now I am going to put the training data in the formula using glm() and build the logistic regression model
62. ```{r}
63. GLMModel=glm(formula=formula,data=TrainingData,family="binomial")
64. ```
65. \* The model will be designed using the "Stepwise selection" method to get the significant variables of the model. This will allow for a better fitting model and have the best variables to make an accurate prediction
66. ```{r}
67. GLMModel=step(object = GLMModel,direction = "both")
68. summary(GLMModel)
69. ```
70. \* Based on the results we can see, Business travel (Frequently), Distance from home, Environment Satisfaction, Job Involvement, Job Satisfaction, Number of Companies Worked, Job Role ( Sales Executive, Sales Representatives, & Laboratory Technicians ), Overtime, Total Working years, Years since last promotion, Relationship Satisfaction. All these are most significant variables in determining employee attrition. These can be the areas that the company can focus on to reduce attrition.
72. \* Need to perform a goodness fit test on the data set, to determine the accuracy of the predicted probability of the model. I am going to use the Hoshmer-Lemeshow test.
73. \* The hypothesis is
74. + H0: The model is a good fit if the p-value > 0.05
75. + H1: The model is not a good fit if the p-value < 0.05
76. ```{r}
77. library(MKmisc)
78. HLgof.test(fit=GLMModel$fitted.values,obs=GLMModel$y)
79. ```
80. \* Based on the results of the test the model is a good fit because the p-value = 0.7487.
82. ## Save and Read Model to and from RDS file
83. ```{r}
84. #saveRDS(GLMModel, "GLMModel.rds")
85. #read.GLMModel <- readRDS("GLMModel.rds")
86. ```
87. ## Model Evaluation
88. \* Applying the model for prediction
89. ```{r}
90. GLMModel.training <- predict(object = GLMModel, newdata = TrainingData, type = "response")
91. #GLMModel.testing <- predict.glm(object = GLMModel, newdata = TestingData, type = "response")
92. ```
93. \* Model performance (Displays ROC curve and AUC score)
94. ```{r}
95. library(caret)
96. library(pROC)
97. ```
98. \* ROC and AUC score
99. ```{r}
100. #GLMModel.training.roc <- roc(TrainingData$Attrition, GLMModel.training, plot = T)
101. #GLMModel.training.roc$auc
102. GLMModel.train.roc <- plotROC(TrainingData$Attrition,GLMModel.training,returnSensitivityMat = TRUE)
103. ```
105. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
107. \* Model performance (Displays confusion matrix)
108. ```{r}
109. GLMModel.training <- ifelse(test = GLMModel.training > 0.5, yes = 1, no = 0)
110. table(TrainingData$Attrition, GLMModel.training) # Displays the classification table
111. GLMModel.train.CM <- confusionMatrix(TrainingData$Attrition, GLMModel.training)
112. print(GLMModel.train.CM)
113. #GLMModel.testing <- ifelse(test = GLMModel.testing > 0.5, yes = 1, no = 0)
114. #table(TestingData$Attrition, GLMModel.testing) # Displays the classification table
115. #GLMModel.test.CM <- confusionMatrix(TestingData$Attrition, GLMModel.testing, threshold = optimal.GLMModel.training)
116. #print(GLMModel.test.CM)
117. ```
118. ```{r}
119. library(Metrics)
120. ```
121. ### Model Performance Statistics
122. ```{r}
123. #calculate accuracy
124. GLMModel.accuracy <- accuracy(TrainingData$Attrition, GLMModel.training)
125. GLMModel.accuracy
126. #calculate sensitivity
127. GLMModel.sensitivity <- sensitivity(TrainingData$Attrition, GLMModel.training)
128. GLMModel.sensitivity
129. #calculate specificity
130. GLMModel.specificity <- specificity(TrainingData$Attrition, GLMModel.training)
131. GLMModel.specificity
132. # calculate precision
133. GLMModel.precision <- precision(TrainingData$Attrition, GLMModel.training)
134. GLMModel.precision
135. # calculate recall
136. GLMModel.recall <- recall(TrainingData$Attrition, GLMModel.training)
137. GLMModel.recall
138. #calculate total misclassification error rate
139. GLMModel.misClassError <- misClassError(TrainingData$Attrition, GLMModel.training)
140. GLMModel.misClassError
141. ```

Figure 7: RStudio code for the Logistic Regression model to predict employee attrition

## Production Turnover

**Date:** 2021-01-31

**Business Department:** Human Resources

**Project Name:** Talent Retention

**Project Description:** The intent of the project is to provide General Electric Company (GE) the capability to proactively identify employees that are prone to voluntarily leaving the organization. A predictive model will be developed to identify these key drivers of the why employees leave the company by using historical data so the human resources department can develop an initiative or focus on the areas to reduce attrition, which will reduce the monetary loss for hiring new talent and retaining the high performing talent.

**Model Baselines:** The mode baseline accuracy is between 88% and 93%. The ROC should be around 0.8619. We can say that the model is running successfully if the model accuracy is 70% or above, if it falls below the threshold of 70%, then the data must be checked for null or missing values.

**Model Performance:** The model’s runtime should be around 1 to 5 minutes depending on the amount of data that is being ran through it. If the model goes beyond 5 minutes, please contact the Analytics team to troubleshoot the issue and perform data quality checks.

**Comments:** If additional columns are added to the data when inputting into the model advise the Analytics team. The Over18 field may have some issues if it is not converted to a numerical value like 1-Yes over 18 or 0-Not over 18 because the model has for transform the non-numerical values in order for it to run efficiently through the model.