River Wilkinson

Southern New Hampshire University

DAT-650-Q1231 Advanced Data Analytics 21TW1

Professor Huang

Final Project

12 December 2021

**USE CASE SCENARIO**

Many middle managers have raised awareness in the human resources department that many high-potential employees have left the company to pursue others and are evident in the recent increase in job postings. There is a need to identify how to retain talent within General Electric (GE), given the need to remain competitive and the total cost and time required to train new employees. In the current environment, the average cost of attrition for an individual is 80% of their annual salary.

The current data environment that GE utilizes is an HR web-based desktop system that contains information about all employees, past and current, and metadata about each employee. HR maintains the environment via a web-based Java client-server application, and the data is stored in an Oracle database in transactional form. HR uses the data warehouse created by the IT department, which is updated every evening with the current day data. The data warehouse data is the source of their reports. It can have an ad hoc extract to select data into Excel for unique research purposes. Using the extract tool, GE has compiled a file of employee data relevant to analyzing the attrition problem. The HR department would like to determine if this data can be used to identify whether or . Understanding attrition drivers is essential for metadata like high performers, role types, and other pertinent groupings from the analysis.

**ANALYTIC STRUCTURE**

Descriptive analytics interprets historical and current data better to understand business changes (Frankenfield, 2020). Descriptive analytics answers the question, "What happened?" Descriptive analytics describes a range of historical and current data to draw comparisons (Pierson, 2019). These measures all describe what has occurred in a business during a set period. GE can implement descriptive analytics to describe the types of employees that have left the business during a set period. Data elements applicable to this analysis are age, education, gender, job involvement, job level, job role, monthly income, work-life balance, relations satisfaction, and job satisfaction. Using descriptive analytics, GE can use large volumes of data they have collected by breaking it down to give important areas more focus. This type of analytics has become a vital part of business operations because it will help the HR department understand their current employee attrition problem and compare it to the past.

Predictive analytics uses statistics and modeling techniques to determine future performance based on current and historical data. This type of analytics answers the question, "What will (or might) happen?" Although this type of analytics focuses on current and historical data, predictive analytics involves complex model building and analysis to predict a future event or trend (Pierson, 2019). GE will benefit from this type of analysis by generally identifying employees that may leave. This information will provide GE with a general employee profile with the characteristics of an employee who might leave the business. GE will then develop a strategy to retain more employees and save money and time when training new employees.

**ANALYTIC TOOL PROPOSAL**

Data mining allows an organization to discover hidden patterns in data by use of machine learning. Sophisticated algorithms are the mining tools used in predictive analytics to extract hidden values from newly discovered patterns. Organizations can examine and process their data to better understand their customer base, improve their operations, outperform their competitors, and better position themselves in the marketplace (Bari, A., 2016). Predictive analytic models are categorized in various ways. These categories are sorted by the business problems the model solves and the primary business functions they serve (such as sales, advertising, human resources, or risk management), and the mathematical implementation used in the model (such as statistics, data mining, and machine learning). A clustering and classification model, decision model, or association model may be used based on an organization's business objective.

**CRISP-DM FRAMEWORK**

The CRISP-DM model (Cross Industry Standard Process for Data Mining) builds data mining projects. CRISP-DM provides a framework for planning and managing a data mining project across any domain. This model includes six steps: 1. Business understanding, 2. Data understanding, 3. Data preparation, 4. Modeling, 5. Evaluation, and 6. Deployment (Manna, 2014). Implementing the CRISP-DM methodology will form a pilot plan for GE to describe and generally identify employees that may leave the business.

**ETHICAL IMPLICATIONS IDENTIFIED IN DATA**

Ethical issues may arise related to data consent. The data collected by the HR department includes demographic information that some employees may not be comfortable with within the employee attrition analysis. This demographic data includes but is not limited to age, marital status, gender, and monthly income. GE can implement a strategy to provide transparency for employees on how and why their data is of value. Providing transparency will build trust within the business and possibly improve data literacy by informing how data can improve business decisions.

**DATA ANALYTIC STRATEGY**

Employee Attrition Analytics is a type of predictive analytics focused on identifying why employees voluntarily leave, what might have prevented them from leaving, and how to use this data to predict attrition risk (Killham, 2020). Businesses are able to understand and design effective intervention strategies that will reduce unwanted attrition. The capability of predictive analytics enables the design of an employee retention model to keep high-performing employees on board. The cost of recruiting, hiring, onboarding, and training new employees can cost a business billions of dollars. In GE’s current environment, the average cost of attrition for an individual is 80% of their annual salary. The goal of predictive analytics for employee attrition is to be proactive and get ahead of attrition before it becomes a problem.

The purpose of the predictive analytic model is determining who is leaving the organization, when they are leaving, and why they are leaving to predict future patterns. The answers to these questions can be found in the environment satisfaction, job satisfaction, and relations satisfaction data collected in the Employee Attrition dataset and by creating a demographic of the employees who left the company voluntarily. Analyzing this demographic will reveal information about attrition in different job roles, job levels, job involvement, and departments to reveal who is leaving and when. Answering the who, what, and why questions and combining the data to see other similarities and differences between employees who stayed versus those who left is the foundation of employee attrition analytics (Killham, 2020).

**ANALYTIC VALUE**

An effective employee attrition model is unique to an organization and focuses on its biggest challenges. Many middle managers have raised awareness in the human resources department that many high-potential employees have left the company to pursue other positions and are evident in the recent increase in job postings. There is a need to identify how to retain talent within General Electric (GE), given the need to remain competitive and the total cost and time required to train new employees. Because employees have different reasons for leaving, GE needs to be flexible and creative in the steps it takes to retain high-performing talent. Data analysis can be used to establish turnover benchmarks for employees. The tracking of these benchmarks over time reveals how the employee experience is changing for better or for worse, if the reasons employees are leaving have changed, or if the attrition pattern is different. These benchmarks show GE what actions are needed to reduce attrition effectively and alert middle management to adjust if needed.

**PILOT PLAN**

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a process model that provides an overview of the data mining life cycle. The CRISP-DM methodology includes descriptions of the typical phases of a project, the tasks involved with each stage, and an explanation of the relationships between these tasks (IBM, 2021). A pilot plan is used in data analytics to ensure that the CRISP-DM methodology is used correctly and that the data can be deployed to solve the business problem.

Diagram

Description automatically generated

**CRISP-DM Process**

**BUSINESS UNDERSTANDING**

Gaining as much insight as possible into the business goals for data mining is the first stage of the CRISP-DM process. The CRISP-DM methodology provides a structured way to accomplish this stage. The tasks involved include:

1. Gathering background information about the current business situation
2. Documenting specific business objectives decided upon by key decision-makers
3. Agreeing upon criteria used to determine data mining success from a business perspective

Understanding an organization’s business situation helps identify what resources are available, what problems need to be solved, and what goals are to be achieved. A concrete primary objective should be agreed upon by the project sponsors and other business units affected by the results.

**DATA UNDERSTANDING**

The data understanding phase of CRISP-DM involves looking at the data available for mining. This step is essential to avoid unexpected problems during the next stage, data preparation. Data understanding involves collecting initial data, exploring data, and verifying data quality (IBM, 2021).

The current data environment that GE utilizes is an HR web-based desktop system that contains information about all employees, past and present, and metadata about each employee. The HR department would like to determine if this data can be used to identify attrition of employees that may leave. Understanding attrition drivers is essential for metadata like high performers, role types, and other pertinent groupings from the analysis.

The statistical programming languages R and RStudio were used to understand the employee attrition dataset better. The skim() function is an alternative to using the summary() function and provides a quick and broad overview of the dataset (RDocumentation, 2021). One of skim() benefits is that it lists the number of missing values per variable. The dataset contains 35 variables (columns), nine character variables, and 26 numeric variables. There are also 1470 rows. The skim() has provided the mean for variables such as age (36.9), monthly income (6503), numbers of companies worked (2.69), hourly rate (65.9), etc. Using the ggplot package, I was able to visualize a couple of variables to understand the data distribution.

Calendar

Description automatically generated with low confidence

Calendar

Description automatically generated with low confidence

**Skim() function results in RStudio**

Chart, histogram

Description automatically generated

**Visualization of the age distribution of employees**

**Chart, bar chart

Description automatically generated**

**Visualization of the job roles of employees**

**DATA PREPARATION**

Data preparation is the process of manipulating raw data into a form that can be readily and accurately analyzed for business purposes. Data preparation involves the following tasks:

1. Merging datasets and/or records
2. Selecting a sample subset of data
3. Aggregating records
4. Deriving new attributes
5. Sorting the data for modeling
6. Removing or replacing blank or missing values
7. Splitting into training and test datasets

The employee attrition dataset was loaded into Rattle. The dataset containing no missing values was split into a training dataset containing 1029 observations and 32 variables. The variables included in the training dataset are BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, and OverTimeAttrition.

**MODELING**

The modeling stage aims to create a model to determine if the business problem is solvable. Modeling is usually conducted in multiple iterations; typically, several models are run because it is rare for a data mining question to be answered satisfactorily with a single model and execution.

Decision trees are usually represented upside down with the root at the top and the leaves at the bottom. Three splits occur from the single trunk into two or more branches starting from the root. This continues until a leaf is reached, a node that is not further split. The root and the leaves are also referred to as nodes. Associated with each non-leaf node will be a test or question that determines which branch to follow. The leaf nodes contain the “decisions.”

Table

Description automatically generated

**Decision Tree Model created in Rattle**

Random forest algorithms produce more accurate models because the ensemble reduces the observed instability when building single decision trees. The random forest algorithm tends to be much more robust to changes in the dataset, and small changes in the dataset will have little, if any, impact on the final decisions made by the resulting model   
(Williams, 2011). Random forest models are generally more competitive with nonlinear classifiers. Random forests also perform well with the input variables that have little bearing on the target variable and are also suitable when there are many input variables and not so many observations. Graphical user interface, text

Description automatically generated

**Random Forest Model created in Rattle**

**EVALUATION**

Model evaluation is the phase that decides whether one model performs better than the other. It is essential to consider the model outcomes according to every evaluation method and applying different methods can provide different perspectives.

Decision trees, also called classification and regression trees, are the traditional building blocks of data mining and the classic machine learning algorithm. Decision trees are easy to view, understand, and explain. Decision trees do not consistently deliver the best performance and represent a tradeoff between performance and simplicity of explanation (Williams, 2011). The decision tree had a root node error of 0.15% (157/1029 observations). The variables used in the tree construction are shown in the appendix below.

Random forests combine multiple decision trees into a single ensemble of models (to build a forest of trees). The goal of a random forest model is to create a high-performance model with less need to interpret the results. A random forest model can use every variable in the dataset without transposing any categorical data into numerical data, saving time and resources for the pilot run. The random forest model created in Rattle has five variables tried at each split, and the OOB estimate has an error rate of 14%. As for the analysis of the Area Under the Curve (AUC) is 0.57%. The variables are listed by level of importance. The three most important variables are OverTime, Age, and MonthlyIncome.

**MODEL EVALUATION**

Model evaluation is the process of identifying the best amongst different models. This process allows us to understand what is to be expected from different models when scoring new observations. Evaluation of model performance also helps us to identify whether any mistakes were made in the choice of input variables (Williams, 2011). The different measures for model evaluation are as follows:

**Error Matrix**

            The error matrix, also known as a confusion matrix, shows the true outcomes against the predicted outcomes. By performing the error matrix in Rattle, two tables are presented: the count of observations and the proportions. The cells of the matrix are referred to as True Negatives, False Positives, True Positives, and False Negatives. In Rattle, the Confusion Matrix is the default on the Evaluate tab. Clicking the Execute button will run the selected model(s) against the chosen dataset to predict the outcomes for each of the observations in that dataset. The predictions are compared with the actual observations (Williams, 2011).

Graphical user interface, text, application, email

Description automatically generated

**Error Matrix created in Rattle**

Graphical user interface, text, application, email

Description automatically generated

**Precision and Recall Plot created in Rattle**

The precision of a model is the ratio of the number of true positives to the total number of predicted positives (the sum of the true positives and the false positives). It is a measure of how accurate the positive predictions are, or how precise the model is in predicting.

The recall of a model is just another name for the true positive rate. It is a measure of how many of the actual positives the model can identify, or how much the model can recall. The recall is also known as the sensitivity of the model. Another measure that often arises in the context of sensitivity is specificity. This is simply another name for the true negative rate (Williams, 2011).

Chart, line chart

Description automatically generated

**ROC Curve Plot created in Rattle**

            The Receiver Operator Characteristic (ROC) curve compares the false positive rate to the true positive rate. There is a trade-off between the number of observations that are incorrectly classified as positives against the number of observations that are correctly classified as positives. The ROC has a form and interpretation like the risk chart, though it plots different measures on the axes. ROCR is used by Rattle to generate these charts (Williams, 2011). The ROC figure below shows that the random forest model is correct 77% of the time.

Chart, line chart

Description automatically generated

**Sensitivity/Specificity Plot created in Rattle**

Sensitivity is the metric that evaluates a model’s ability to predict true positives of each available category. Specificity is the metric that evaluates a model’s ability to predict true negatives of each available category. Specificity determines a model’s ability to predict if an observation does not belong to a specific category. It requires knowledge of the model’s performance when the observation actually belongs to every other category than the one being considered. These metrics apply to any categorical model (Mitrani, 2019).Chart, line chart

Description automatically generated

**PLOT IMPLEMENTATION**

Implementing an advanced data analytic solution will help GE derive insight into their employee attrition problem.There is a need to identify how to retain talent within General Electric (GE), given the need to remain competitive and the total cost and time required to train new employees. In the current environment, the average cost of attrition for an individual is 80% of their annual salary. Implementation of an advanced data analytic solution will combat the average cost of attrition and to identify what needs to be done in order to save more related annual salary. These costs are attributed to onboarding, lost productivity, poor engagement, and hiring costs. By investing in this solution, a model can be used to predict which combination of employee attributes are more likely to remain at the company based on the dataset provided by HR. The random forest model results indicate that G.E can predict with 77% accuracy the employees who will leave the company. Although the accuracy of the model can be improved, full implementation of the pilot model will lend itself o saving money in the long run with employee incentive programs, increased employee engagement and increased morale for every employee at G.E.

**References**

[Bari, Anasse, Mohamed Chaouchi, and Tommy Jung. Predictive analytics for dummies. John Wiley & Sons, 2016.](http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119267005.html)

Halton, Clay. 30 June 2021. Behavioral Economics: Predictive Analytics. Retrieved on 24 October 2021 from <https://www.investopedia.com/terms/p/predictive-analytics.asp>.

IBM. 2021. SPSS Modeler CRISP-DM Guide. Introduction to CRISP-DM. Retrieved on 21 November 2021 from <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-introduction-crisp-dm>.

Killham, Emily. 06 May 2020. Perceptyx. Employee Attrition Analytics: The Who, When & Why Of Employee Turnover. Retrieved on 05 November 2021 from https://blog.perceptyx.com/employee-attrition-analytics

Manna, M. (2014, December 18). *The data science project lifecycle*. Data Science Central. <https://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle>.

Mitrani, Alex. 05 December 2019. Evaluating Categorical Models II: Sensitivity and Specificity. Retrieved on 15 December 2021 from https://towardsdatascience.com/evaluating-categorical-models-ii-sensitivity-and-specificity-e181e573cff8.

Pierson, P.E., Lillian. Brown, Meta S., McFredries, Paul. John Wiley & Sons, Inc. 2019. Back to Basics: Data Analysis.

RDocumentation. 2021. skim: Skim a data, getting useful summary statistics. Retrieved on 11 November 2021 from <https://www.rdocumentation.org/packages/skimr/versions/2.1.3/topics/skim>.

Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!, Springer.

Agrawal, Samartha. 17 May 2021. Towards Data Science. How to split data into three sets (train, validation, and test) And why? Retrieved on 06 February 2022 from <https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c#:~:text=Train%2DValid%2DTest%20split%20is,of%20these%20datasets%20is%20below>.

Chan, Carmen. N.d. Displayr. What is a ROC Curve and How to Interpret It. Retrieved on 06 February 2022 from <https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/>.

Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!, Springer.