POMS.6220 DECISION ANALYTICS – UNIVERSITY OF MASSACHUSETTS, LOWELL

SPRING 2019

TERM PROJECT

**Human Resources Predictive Analysis: Understanding Factors of Employee Attrition**

BY

SRIDHAR RANGAN

JASON TRAN

**Abstract**

One of the biggest challenges for Human Resources in any company is the rate at which employees leave the company (attrition rate). The attrition rate is impacted by both employees voluntarily ending employment or being terminated by human resources. The goal for human resources is to lower the rate at which employees leave. This is due to the high cost attached to finding, hiring and training new employees. This paper examines using predictive analytics to determine the leading causes of attrition. Several classification models were employed such as Random Trees, C&R Trees, C5.0 Trees, CHAID and Neural Net. An Information Fusion-Based Sensitivity analysis was performed to identify the overall strongest predictors of attrition. This paper aims to demonstrate the advantages of using predictive analytics to supplement decision making to lower attrition rate.

Keywords: Human Resources, HR Analytics, Classification, Attrition, Neural Net

1. **Introduction**

One of the most key departments in any major firm is Human Resources (HR).  Some of the responsibilities for HR include managing employee records, finding good talent and ensuring other departments are operating well.  One of the biggest challenges faced by managers in human resources is attrition. The attrition or attrition rate refers to whether or not an employee will leave the company.  This measures the rate of leaving based on either an employee willingly leaving or being terminated by HR.

For any decision maker in human resources, the attrition rate should always be closely looked at.  The major reason that the rate should be as low as possible is due to cost. For finding new talent, HR must spend the time and effort to finding job seekers.  Additionally, once a new employee is hired, they have to go through training and orientation which is all at expense of the company. If the attrition rate is high, HR constantly needs to find new talent which will cost the firm more in the long run.

The managers in human resources have access to all employee data.  They can use this data to conduct not only to look at the descriptive statistics but to do an in-depth analysis to figure out the driving factors of the attrition rate.  Based on several models, managers can make hiring decisions accordingly to lower the rate at which employees leave. The use of HR analytics is a growing trend in any firm and “enables people managers and teams to comprehend more about the complexities of the people factor in their organization” (Kumar, 2016).

This report will demonstrate multiple classification models to determine the key factors which impact the attrition rate.  Using analysis decision makers within the firm can take proper steps to ensure employee satisfaction to reduce the likelihood of leaving.

1. **Motivation**

Human Resources uses a variety of ways in order to find the best talent for their organization.  The recruiters typically implement soft skills to determine a candidate's potential to fit with the organization and the culture environment.  While soft skills have merit, analytics has become more prevalent in isolating the best possible workers. Human resources additionally is responsible for ensuring employee satisfaction and keeping the attrition rate low.  In the past, HR has used elements such as company outings, bonuses, raises and promotions as means of keeping employees satisfied however, data can be used to determine leading causes of employee attrition.

This project aims to look at employee data and determine not only the causes of employee attrition but to develop possible solutions for lowering the employee attrition rate.  While in the past, proactive approaches might have been successful, the use of HR analytics can provide valuable insight and can be used as a basis to reduce company expenses.

1. **Literature Review**

This section demonstrates the critical findings from certain articles pertaining to Human Resource analytic in relation to employee attrition. Based on a published article from An International Journal, Business schools in India is having difficulty in recruiting adequate faculty members that are in position to be assign research activities, teach, and also support administrative responsibilities. In order to uphold the standards of excellence that they seek, it is necessary that the business schools retain faculty members as well as keep the attrition levels low which would help. This study is extremely helpful for business school, where email analytics could be utilized to predict the faculty members intention in leaving their business school. Innovative/creative even though previous studies have studied attrition; this research utilizes predictive analytics and draws the conclusion on the faculty member’s reason to quit. This research supports business schools to forecast the probability of the faculty members leaving the business school which is of great use, as appropriate action can be taken to maintain and control attrition. (Raman, Bhattacharya, & Pramod, 2019)

According to a different research, an organization performed an analysis in order to help better understand the root cause of employee turnover rate. This was executed by means of a series of deep data mining methods such association rules and decision trees (Girmanova & Gašparová, 2018).

Another research, in accordance with employee retention suggested that employee attrition poses concern to their organizations. This is due to the sheer fact that when employee turns away from an organization carry along with them as stated “invaluable tacit knowledge” which is considered to be leverage for the organizations (Nagadevara, Srinivasan, & Valk, 2008). This causes substantial risk and challenges for industry such as the Indian software organization that has seen rapid growth within its sector.

“This research explores the relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee turnover. The unique aspect of this research has been the use of five predictive data mining techniques on a sample data of 150 employees in a large software organization. The results of the study clearly show a relationship between withdrawal behaviors and employee turnover. Age and marital status emerged as key demographic variables. The findings of this study have implications for both research and practice. There is a need to expand the scope of this research to include multiple organizations and a large sample, which will allow for more robust predictions. For practitioners, it emphasizes the need for greater use of models and analytical tools in engaging with human resource strategies and plans, and in particular that HR professionals will need to understand, appreciate and apply such models in future to be able to perform their roles as strategic business partners.” (Nagadevara, Srinivasan, & Valk, 2008).

One organization in particular took a technological approach to the matter at hand. Innovative technology allowed organizations to collect exceptional knowledge of employee conduct allowing managers to exhibit informed decisions.

The consulting firm Deloitte utilized a smart badge system to track employee movement. The data collected allowed for Deloitte to make a visualized heat map of all employee office activity (Bosanac, 2015). This innovative technology played a huge role in the retention of employees. It enabled organizations to exploit people analytics.

“Credit Suisse Group looks at a variety of data points, including performance reviews and a manager's team size, to determine which workers are likely to jump ship. That allows upper management to intervene before someone quits, instead of having to conduct an exit interview to find out what the problems are” (Bosanac, 2015).

These research reviews that were conducted did not perform multiple classification algorithms. The most that were performed were two models at best. Majority of the research did not demonstrate the crucial predictors.

Additionally, besides utilizing human resource analytics to gauge in employee attrition, it is also utilized for recruiting, compensation and benefit, and skills inventory. The objective for performing this research is to determine the type of analytics company apply to improve their decision-making process and the type of analytic making tool to be utilized (Sousa, 2018).

These research reviews that were conducted did not perform multiple classification algorithms. The most that were performed were two models at best. Majority of the research did not demonstrate the crucial predictors that affect attrition.

1. **Methodology**

Prior to modeling, the Cross Industry Standard Process for Data Mining was used in conjunction with the dataset.  The methodology was chosen based on its well-known status in data mining and research (Shearer, 2000). Figure 1 outlines the CRISP-DM process.

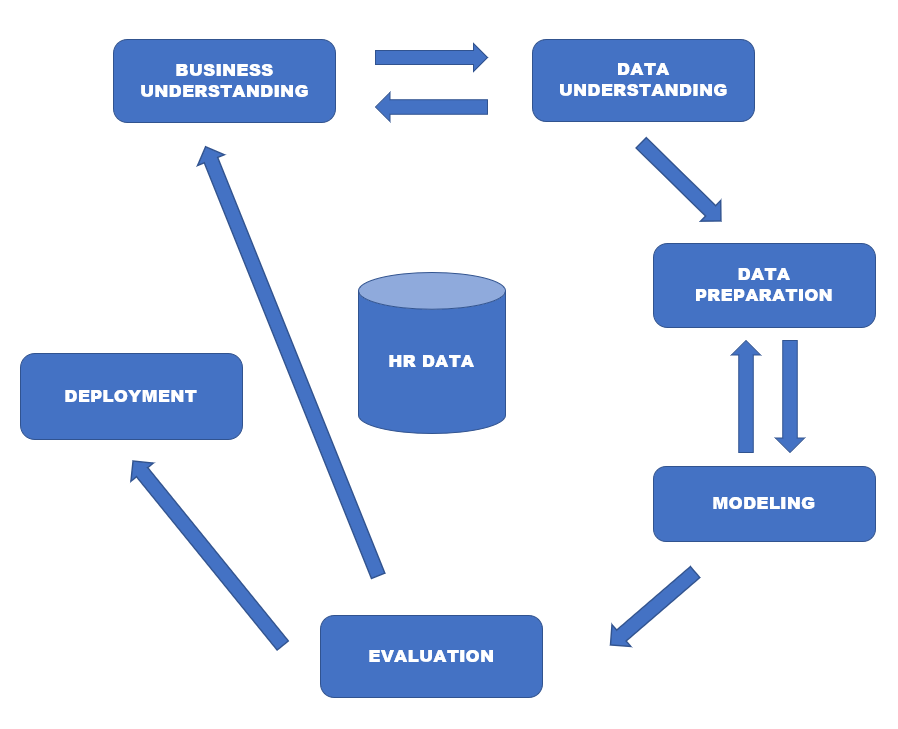


Figure 1: CRISP-DM Diagram

**Step 1: Business Understanding**

The first step in the CRISP-DM methodology serves to find the business objective, assess the situation and determine data mining goals.  For any human resources department, the main business objectives include hiring correct personnel, ensuring employee satisfaction and maintaining a low percentage attrition rate.  The data mining goal for this project is to predict whether or not an employee will leave the company. Using predictive modeling, insight can be gained about primary causes of employee attrition.

**Step 2: Data Understanding**

The second step in the CRISP-DM methodology is to collect, describe and explore the data as well as verify the data quality.  The dataset used in this paper is from Kaggle.com and is publicly available (Kaggle, 2017). While the data set is fictional, the attributes and fields accurately represent data that HR managers would analyze. The data has 35 attributes and 1471 rows of data for each employee. The target attribute or variable selected was *Attrition* to measure whether or not an employee will leave the company.

**Step 3: Data Preparation**

The third step for CRISP-DM is data preparation which involves making sure the dataset is ready for running the different data mining models. Since the dataset is fictional, there were no missing values for any of the attributes. To ensure no missing values, the dataset was run through python using the read\_excel command from the Pandas library. The original dataset was provided in a csv format. However, there were issues in importing the dataset into the IBM SPSS Modeler.  Therefore, we converted the csv file into a standard excel file in order to import and run the appropriate models. One final piece of cleaning performed was the removal of the *EmployeeNumber* column. When experimenting with running different models using IBM SPSS Modeler, an issue was discovered in the predictors. Several of the models indicated that the *EmployeeNumber* field was a significant predictor of employee attrition. However, based on the dataset given, this column simply keeps record of each employee number and cannot have any bearing on employee attrition. Using the CRISP-DM methodology, the original dataset was revisited. The column *EmployeeNumber* was removed entirely from the dataset to ensure that the models would not believe that the employee ids were significant predictors of employee attrition.

Some of the variables in the dataset contain rating scales based on each employee. In Figure 2, the rating scale is displayed showing either range of 1 to 5 or 1 to 4 depending on the variable.



Figure 2: Rating Scales for Variables

Figure 3 outlines all of the different attributes given in the dataset along with a brief explanation about the field as well as variable type. The variable highlighted in red indicates the *EmployeeNumber* variable which was removed prior to running the model. The variables which are highlighted in blue are the ones corresponding Figure 2 with rating scales. The target variable *Attrition* is highlighted in green.



Figure 3: Input and Target Variables for Modeling

1. **Modeling**

The objective of this project was to correctly predict whether an employee will leave the company or not. Since the target variable, *Attrition* was identified as either a binary or flag variable, several classification models were used in determining the best method of accurately predicting the employee attrition rate. This paper utilizes several different models in IBM SPSS: Random Trees, C&R Trees, C5.0 Trees, CHAID and Neural Net. For construction of each model, the data was split into either training or testing data. This paper utilizes a 70% partition for the training data and a 30% partition for the testing data. Figure 4 displays each of the models as constructed by IBM SPSS Modeler.

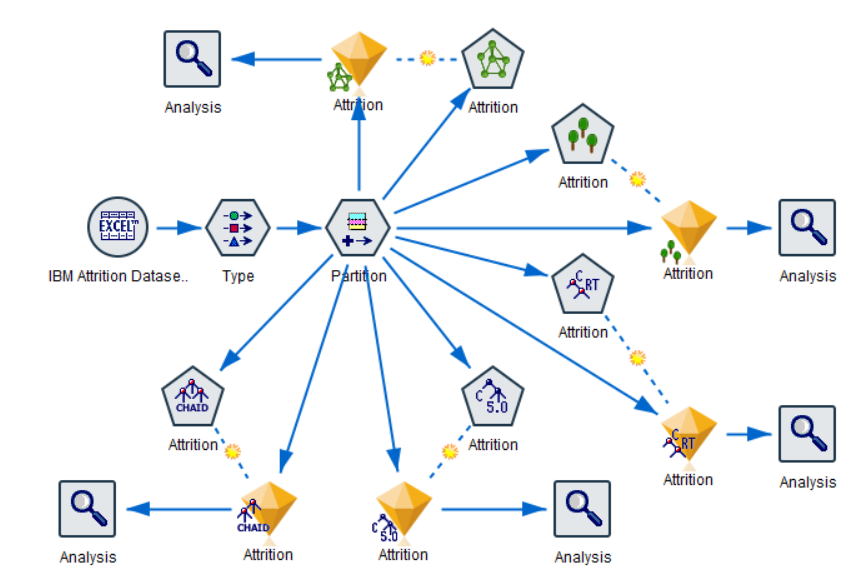


Figure 4: Classification Models Visualized

**5.1 Random Trees Method**

The first model implemented in this paper is Random Trees. According to developers at IBM, Random Trees “are robust when you are dealing with large data sets and numbers of fields. Due to the use of bagging and field sampling, they are much less prone to overfitting and thus the results that are seen in testing are more likely to be repeated when you use new data” (Ruiz, 2015). Since the HR dataset was relatively large, the Random Trees approached appeared to be a logical first step in examining the data. Figure 5 highlights the accuracy of the model and Figure 6 shows a screenshot of some of the decision rules found for the target variable.

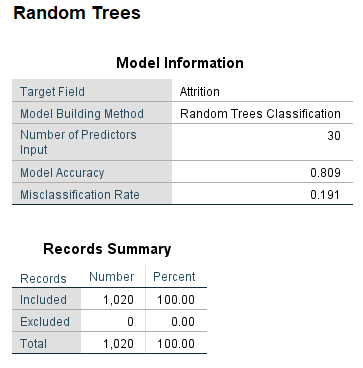


Figure 5: Random Trees Accuracy

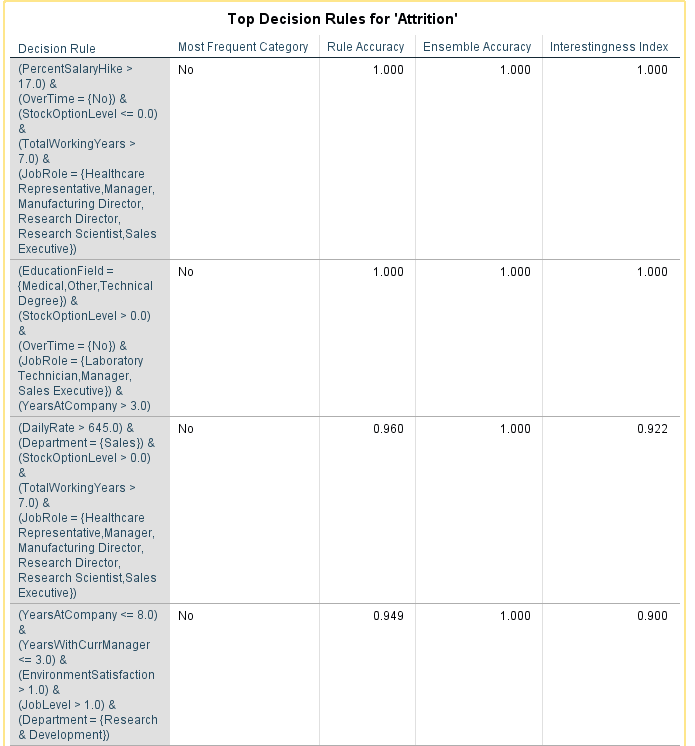


Figure 6: Key Decision Rules – Random Trees Model

**5.2 C&R Trees**

The second model used by this paper is C&R Trees which is one of the decision trees methods. This specific approach was used in order to learn more about the predictors of attrition. The Knowledge Center at IBM specifies the significance of decision trees and how they can be used:

Decision tree models allow you to develop classification systems that predict or classify future observations based on a set of decision rules. If you have data divided into classes that interest you (for example, high- versus low-risk loans, subscribers versus nonsubscribers, voters versus nonvoters, or types of bacteria), you can use your data to build rules that you can use to classify old or new cases with maximum accuracy. For example, you might build a tree that classifies credit risk or purchase intent based on age and other factors. (IBM Knowledge)

The C&R Trees model was able to deliver more insight regarding the different predictors of attrition. Additionally, the model gave a set of rules in determining leading causes for employees leaving the company. Figure 7 is a screenshot of the output of the rules for the C&R model.

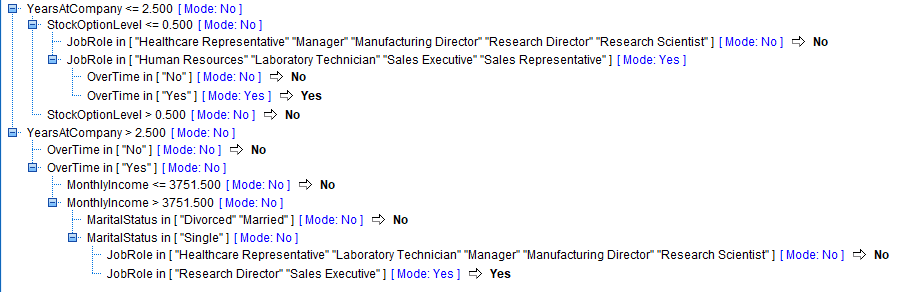


Figure 7: C&R Trees – Decision Rules

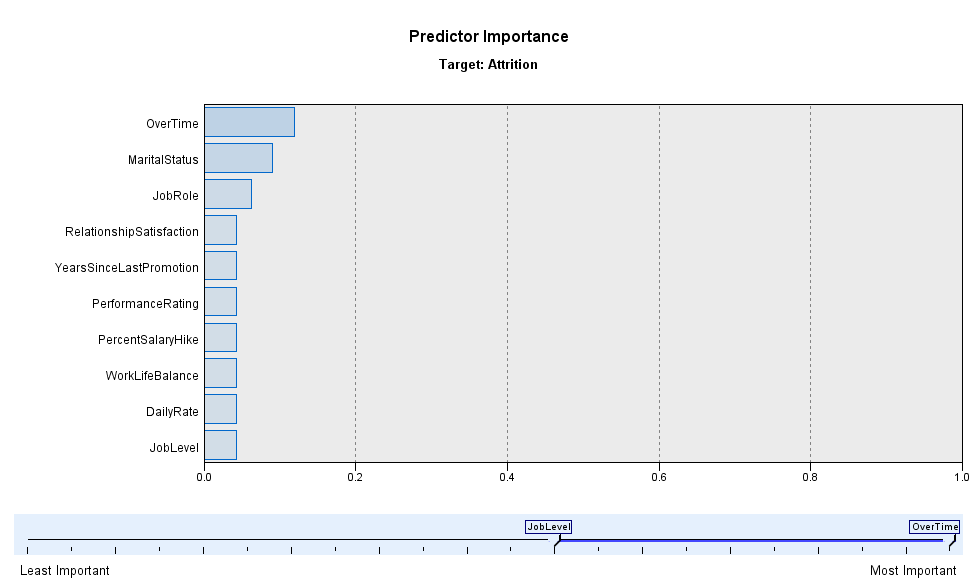


Figure 8: C&R Trees – Significant Predictors

**5.3 C5.0 Trees**

Another model used to predict employee attrition was the C5.0 model. Similarly, to C&R Trees, the C5.0 is another decision tree model. The model was able to deliver a different set of decision rules as well as different significant predictors from the C&R Trees model. Figures 9 and 10 are screenshots of the output for the rules and predictor importance.

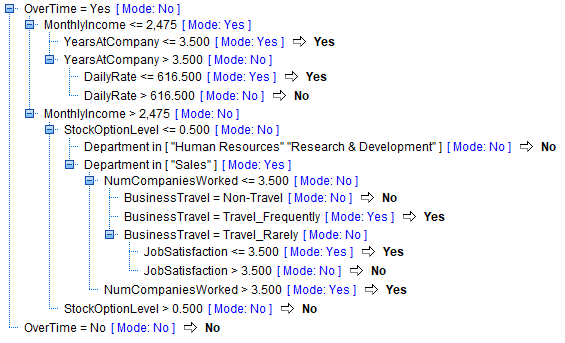


Figure 9: C5.0 Model – Generated Rules

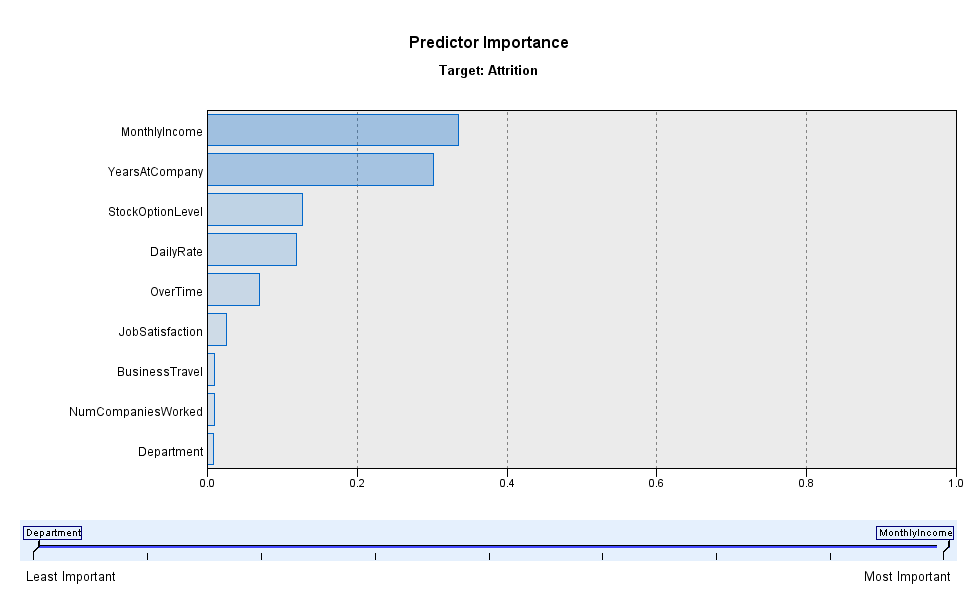


Figure 10: C5.0 – Significant Predictors

**5.4 Chi-Square Automatic Interaction Detector (CHAID)**

Similarly, to both C5.0 and C&R Trees, the CHAID model was used to determine key predictors of attrition. The CHAID model can be used for “prediction as well as classification, and for isolating the most persuasive variables against target goals, such as website visitation or supplier recommendation.” (Beacon Tech, 2010). Figure 11 shows the output of predictors as per the CHAID model.

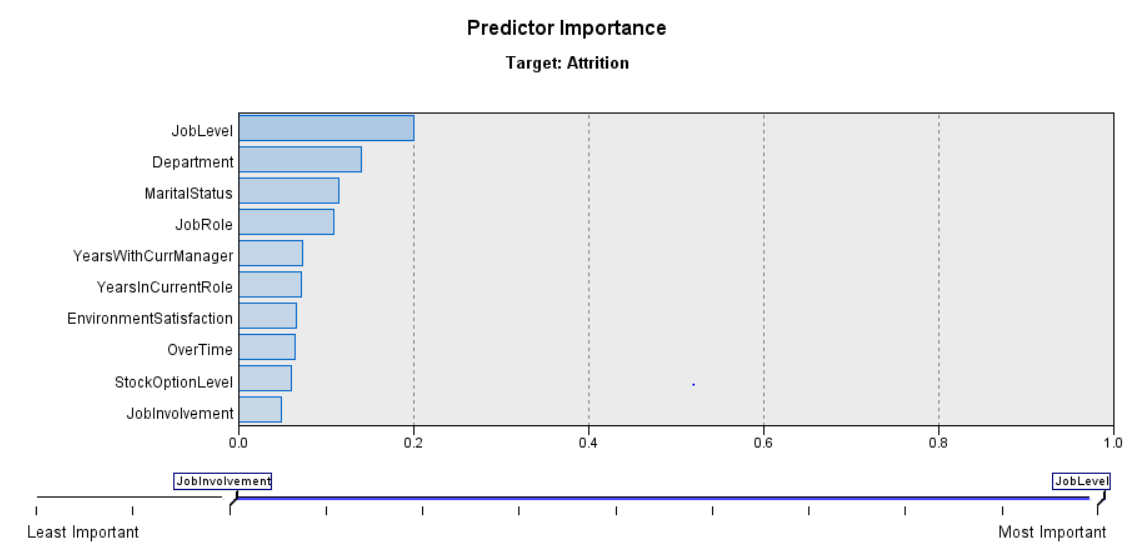


Figure 11: CHAID - Significant Predictors

**5.5 Neural Net**

The final model which was used in determining key predictors was the Neural Net Model. This model approaches the classification problem differently. Unlike the previous models, the neural net operates similar to the human brain and identifies key predictors through the use of patterns (Skymind, 2019). Figure 12 shows the output of the neural net and its identification of key predictors. Figure 13 is the neural network visualized as given by IBM SPSS modeler.

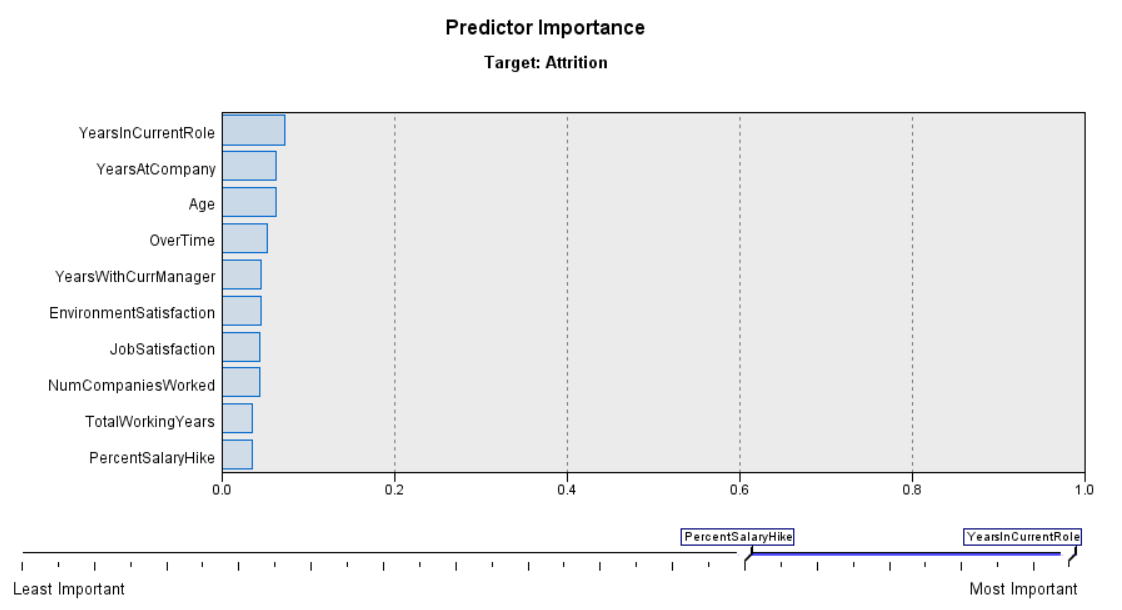


Figure 12: Neural Net - Significant Predictors

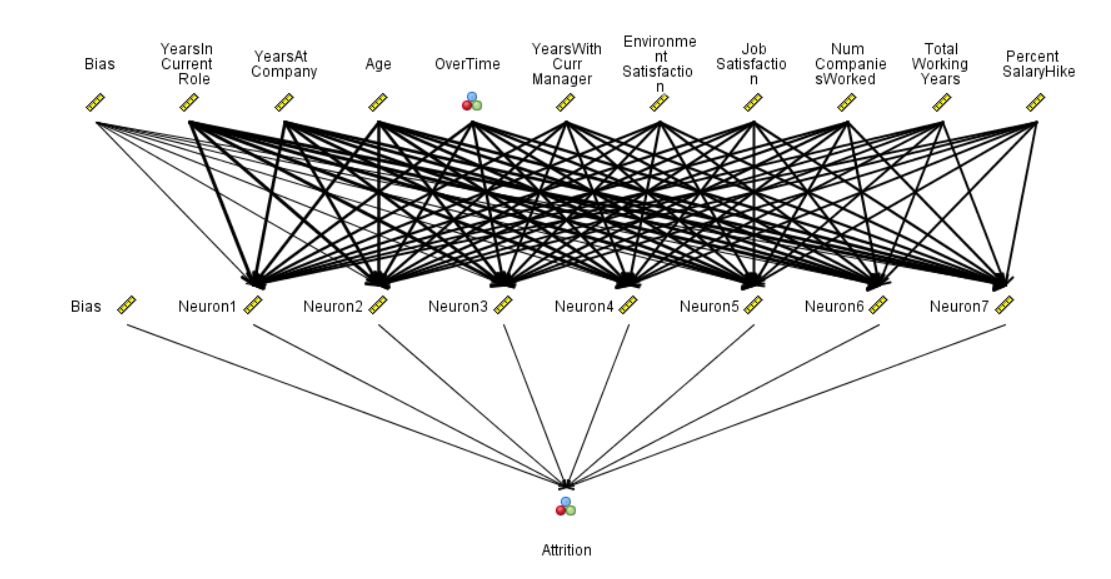


Figure 13: Neural Network

1. **Results and Discussion – Performance and Predictors**

After creating and running the models, the accuracy, sensitivity, specificity and precision were all calculated and compared to find the best model. Figure 14 is a comprehensive chart which shows the values for each model. The values in the figure were either calculated using IBM SPSS output or by use of formulas.



Figure 14: Calculated Results

Figure 15 (Provided by Professor Asil Oztekin) shows the equations used in the calculation of the various metrics. The True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values were provided by the analysis output from the individual models. The IBM SPSS Modeler provided two confusion matrices for each model (Training and Testing). For the calculations, the testing matrix for each model was used via equations in Figure 15.

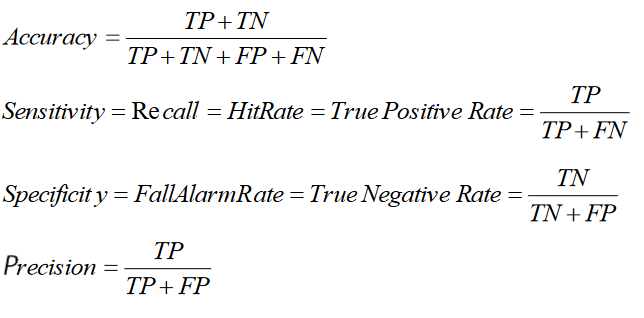


Figure 15: Equations for Calculations

Based on Figure 14, the model which had the highest accuracy was the Neural Net model with an accuracy of 84.6%. The accuracy is a determinant of how often the model is correct overall. In comparison with other values, Neural Net performed the best. The sensitivity score for this dataset represent how often the model predicts an employee to leave the company when the data shows that the employee actually left the company. The values of the sensitivity score were all low. The Neural Net model had the highest sensitivity with 30.6% which means that only in 30% of cases the model predicts an employee to leave when they actually left. The specificity score indicates how often the model predicts an employee to stay at the company when the data shows the employee actually stayed at the company. For all models, this score was the highest with all scores being above 95%. The Neural Net model has the highest score (97.7%) meaning that it was the best at determining the likelihood of an employee to remain at the firm.

**6.1 Key Predictors for Each Model**

For each model, the top predictors were identified by using the IBM SPSS Modeler software. Figures 16 and 17 show the predictor importance for the different models given the target variable as *Attrition*.

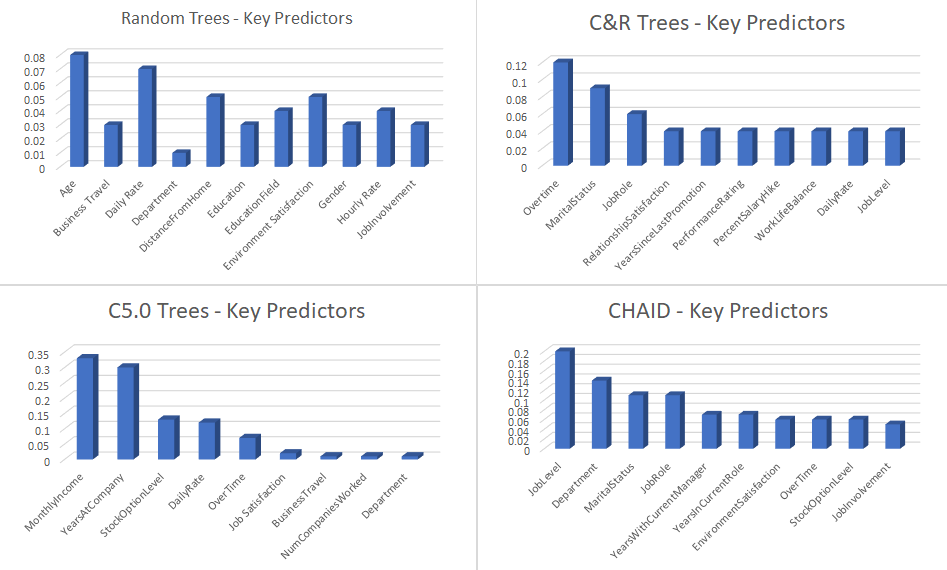


Figure 16: Important Predictor Variables Comparison

Since the Neural Net model had the best overall accuracy, it has been separated from the other models. The key predictors for the Neural Net are displayed in Figure 17.

Figure 17: Neural Net – Key Predictors

While all the models provide different insight into the factors that can impact *Attrition*, it is important to understand the overall top factors based on all models. The best approach for determining overall strongest predictors is through use of the information fusion-based sensitivity analysis.

**6.2 Information Fusion-Based Sensitivity Analysis**

This type of sensitivity analysis allows for weighing of all the models based on accuracy and determining the top key predictors overall. Figure 18 shows the formula which is used for the analysis.

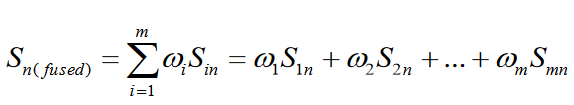


Figure 18: Mathematical Formula – Fusion Analysis

The equation in Figure 18 relies upon two variables. The S variable is the sensitivity for each model. The ω variable represents the weight of the model. This paper utilizes weight based on model accuracy. Depending on the accuracy from Figure 14, the weight of the model was calculated by taking the model accuracy and dividing by the total accuracy of all of the models used. Figure 19 displays the calculated weights for all of the models used.



Figure 19: Weight Calculation



Figure 20: Calculation of Strongest Predictors

Figure 20 shows the information fusion-based sensitivity analysis which was performed using Microsoft excel. Using the weights calculated in Figure 19, the overall importance of the key predictors was calculated. The top 5 overall key predictors have been highlighted in green in Figure 20. Figure 21 is a visualization of the top 5 key predictors from the fusion-based sensitivity analysis. Based on all models, the best predictors of *Attrition* are Age, Job Level, Overtime, Monthly Income and Years at Company.

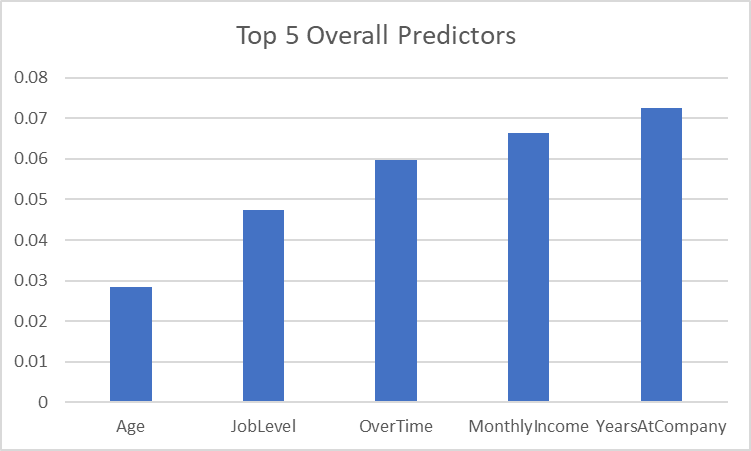


Figure 21: Graph of Top 5 Overall Key Predictors

1. **Managerial Implications**

From the information fused-based sensitivity analysis, the top 5 predictors of employee attrition are Age, Job Level, Overtime, Monthly Income and Years at Company. For Human Resources managers, it is recommended to evaluate these factors for each employee to understand the likelihood of attrition. The specialists within HR can utilize this information to make informed decisions in hiring and maintaining a strong business culture.

The key predictors identified serve to indicate which factors for employees need to be examined. Since the number of years at the company is the most significant predictor, HR managers may look at the descriptive statistics of the *YearsAtCompany* field to determine which types of employees are at risk of leaving. Additionally, since monthly income is another significant predictor, compensation or incentives for employees can be evaluated to ensure that long term costs of training new employees can be mitigated. These changes can help ensure a more satisfactory workplace environment with great benefit for the company. Since age and job level were identified as a key factor of attrition, HR specialists should consider reexamining potential new employees and their desire to rapidly rise through the ranks. While it is important to have driven employees, it is also important to temper expectations of being promoted quickly.

The purpose of this paper is to demonstrate the benefits of predictive analytics as a supporting instrument for human resources decisions. The managers of HR may want to consider a balanced approach between core human resources knowledge and predictive analytics. HR analytics has the capability to improve other areas of an organization such as understanding key performance indicators of employees and the value added to a company (Gelbard et al, 2018). Through the use of analytics to support decisions, HR can improve both the environment and the satisfaction of employees in the workplace.

1. **Conclusion**

This paper utilizes different classification modeling approaches to determine the likelihood of employee attrition. The models used for this paper include: Random Trees, C&R Trees, C5.0 Trees, CHAID and Neural Net. The dataset used was partitioned by splitting the data in 70% training and 30% testing.

After running all models, various evaluation metrics were used such as accuracy, sensitivity, specificity and precision. The Neural Net had the highest accuracy therefore performing the best out of all models. Each of the models identified key predictors that impacted attrition. To determine the overall strongest predictors of attrition from all models, an information fusion-based sensitivity analysis was performed. The strongest predictors were Age, Job Level, Overtime, Monthly Income and Years at Company. From the predictors, HR personnel can utilize this type of information in decision making.

The objective of this research paper is to demonstrate the value of applying predictive analytics to human resources challenges. A common issue for HR is lowering the rate at which employees leave. It is important to understand the factors which contribute to employee turnover. By using a combination of HR knowledge and analytics, managers can make better decisions to improve the overall satisfaction of their employees to reduce the employee attrition rate.

**References**

Beacon Tech. (n.d.). CHAID Analysis. Retrieved from <http://www.beacontech.com/market-research-tools-we-use/quantitative-analysis/chaid-analysis.asp>

Bosnac, A. (2015). Mining Your Employees for Data. *Canadian Business*, *88*(12), 10. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=110094177&site=ehost-live>

Gelbard, R., Ramon, G. R., Carmeli, A., Bittmann, R. M., & Talyansky, R. (2018). Sentiment analysis in organizational work: Towards an ontology of people analytics. *Expert Systems*, *35*(5), 1. <https://doi.org/10.1111/exsy.12289>

Girmanová, L., & Gašparová, Z. (2018). Analysis of Data on Staff Turnover Using Association Rules and Predictive Techniques. *Quality Innovation Prosperity / Kvalita Inovácia Prosperita*, *22*(2), 82–99. <https://doi.org/10.12776/QIP.V22I2.1122>

IBM Knowledge Center. (2012, Spring). Decision Tree Models. Retrieved from <https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm>

Kaggle. (2017, March 31). IBM HR Analytics Employee Attrition & Performance. Retrieved from <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

Kumar, R. (2016). HR Analytics: Where Do We Stand? (cover story). *Human Capital*, *20*(5), 12–15. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=119115236&site=ehost-live>

Nagadevara, V., Srinivasan, V., & Valk, R. (2008). Establishing a Link between Employee Turnover and Withdrawal Behaviours: Application of Data Mining Techniques. *Research & Practice in Human Resource Management*, *16*(2), 81–99. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=36853817&site=ehost-live>

Raman, R., Bhattacharya, S., & Pramod, D. (2019). Predict employee attrition by using predictive analytics. *Benchmarking: An International Journal*, *26*(1), 2–18. <https://doi.org/10.1108/BIJ-03-2018-0083>

Ruiz, A., & Spss. (2015, October 12). Random Trees algorithm in SPSS Modeler 17.1. Retrieved from <https://developer.ibm.com/predictiveanalytics/2015/10/11/random-trees-algorithm-in-spss-modeler-17-1/>

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*.

Skymind. (n.d.). A Beginner's Guide to Neural Networks and Deep Learning. Retrieved from <https://skymind.ai/wiki/neural-network>

Sousa, M. J. (2018). HR Analytics Models for Effective Decision-Making. *Proceedings of the European Conference on Management, Leadership & Governance*, 256–263. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=133413027&site=ehost-live>

**Biography**

Sridhar Rangan is a graduate student at the University of Massachusetts, Lowell. He is currently pursuing a Master of Science in Business Analytics degree and concentrating on marketing analytics. Prior to graduate studies, he received two undergraduate degrees at the University of Massachusetts Lowell in Computer Science and Business Administration with concentrations in Finance and Marketing. He has also worked at Brigham and Women’s Hospital in Boston as a research assistant and has contributed to several medical papers in cardiovascular science. After completing his graduate studies in December 2019, he hopes to pursue a career in consulting or business intelligence.

Jason Tran was born and raised in Lowell, Massachusetts. He is currently enrolled in the Manning School of Business graduate certificate program in Business Analytics. Jason is currently a full-time student. He worked for a family business throughout his teen years and adulthood, apart from that he has also work for a startup Biotechnology Company. He graduated from the University of Massachusetts at Lowell in with a Bachelor of Science in Business Administration, dual concentration in Finance and Management degree. He decided to pursue this certificate program to acquire additional academic credentials to obtain a competitive edge in the current job market.