Important Variables in Employee Attrition

Submitted by

Kyle Anthony Gable

Management Information Systems (Specialization in Business Analytics)

To

The Honors College

Oakland University

In partial fulfillment of the

requirement to graduate from

The Honors College

Mentor: Dr. Yazan F. Roumani, Associate Professor of Quantitative Methods

School of Business Administration

Oakland University

April 7th, 2023

**Abstract**

This paper will investigate the important variables that contribute to the loss of employees (employee attrition) and seeks to answer the question of “What variables play an important role in employee attrition rates regardless of industry?” This paper aims to provide new knowledge by answering this question and to increase public knowledge and understanding of these important variables. It is hoped that by reading this paper employers and management personnel will stand to gain within their respective domains by gaining the knowledge necessary to improve employee retention rates and employees will stand to gain by having a more amiable workplace as a result of changes implemented by management. Discovering which variables are important in employee attrition will benefit employers, management, and employees.

**Introduction**

Employee attrition and employee retention are two sides of the same coin. The latter is the measure of how an organization retains employees, the former is a measure of how an organization loses employees. Employee attrition is, by definition, “the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death or retirement” (Garter). A high employee attrition rate means that more employees are leaving an organization than staying. By studying employee attrition, one can learn how to better retain employees.

The current study of employee attrition has been quite limited in scope. Existing research papers on the subject often only pertain to what machine learning (ML) methods are best at predicting employee attrition rather than the variables; for instance, (El-Rayes et al., 2020) focuses on tree-based models and the viability of tree-based models to predict employee attrition. Others such as (Guerranti, F., & Dimitri, G. M., 2023) attempt to compare multiple different methods such as neural networks, random forest, logistic regression, and other methods using the same dataset and comparing common metrics in order to determine which methods are superior. Less commonplace are papers such as (Haldorai et al., 2019) that have focused on significant variables contributing to employee attrition in specific industries, such as the hotel industry. While it is ascertainable that some variables that are significant in some industries are likely not significant in other industries, it has thus far been largely unexplored as to whether any variables persist to be important in contributing to employee attrition across multiple industries.

By utilizing various ML methods, one can hope to obtain a better understanding of the important variables that contribute to employee attrition. The objective of this paper is to identify what variables are important across multiple industries, if any. This paper will utilize the ML methods of classification and regression tree (hereafter referred to as CART), logistic regression based on forward selection, backward selection, and stepwise selection, discriminant analysis, neural networks, and random forest on a real HR dataset collected by Edward Babushkin and made available on the Kaggle website by Davin Wijaya. These ML methods will be performed through SAS Studio. Several recent papers have been conducted with fictitious HR datasets, such as Jain et al. (2020) and Chung et al. (2023), and while these are valuable papers with insightful results, ultimately fictitious datasets are fictitious datasets; no matter who made them, they are no replacement for real data. It is hoped that by reading this paper a manager or employer seeking to improve retention rates could learn what variables may be contributing to their organization’s employee attrition rates and may thus be better suited to implement changes that will contribute to a decrease in employee attrition rates.

**Aims and Objectives**

**Aims**

1. To determine if any of the independent variables in this dataset are important in the prediction of employee attrition (Attrition\_Status) across multiple industries.

2. To identify which modeling method is best at predicting employee attrition (Attrition\_Status) across multiple industries and to identify which modeling method is the best at prediction while also notably parsimonious given different sets of independent variables in this dataset.

**Objectives**

1. In cases where a numerical importance is provided from the modeling method, utilizing the provided numerical importance from such modeling methods to determine which variables are the most important in those methods.

2. In cases where a numerical importance is not provided from the modeling method, utilizing comparison between multiple variances of the same modeling method in order to determine which variables are the most important in those methods.

**Data Collection**

Analysis was performed on a dataset collected by Edward Babushkin and made available on the Kaggle website by Davin Wijaya. This dataset includes 1,129 observations and 16 variables. Comparison between the variables through exploratory data analysis (EDA) demonstrated no significant correlation between any of the variables, so no variables were removed for that reason. Additionally, there were no variables that existed solely as a primary key, such as an “EmployeeID” variable, so no variables were removed for this reason either. That stated, 9 observations had the oddly specific value of “30.40033257” for the variable “Employee\_Age” whereas all other values for this variable were integers, so these 9 observations were removed from analysis. Due to this dataset being originally in the Russian language, many of the variables’ names as well as the observation categories themselves had their names modified with the purpose of making clearer what the variable/observation means as compared to the original translation. This analysis used the translated dataset, so the variables listed below have been freshly translated.

**The Dataset**

**Dependent Variable**

Attrition\_Status is whether the employee is leaving the organization (0 means that they are staying, 1 means that they are leaving).

**Independent Variables**

Experience\_in\_Months is a measure of how much experience the employee has measured in months.

Gender is whether the employee is male or female (m means male, f means female).

Employee\_Age is how old the employee is in years.

Employee\_Industry is the industry that the employee is involved in (Agriculture, Banking, Building, Consulting, HoReCa, IT, Manufacturing, Mining, Pharma, PowerGeneration, RealEstate, Retail, State, Telecom, Transportation, and Etc; HoReCa is shorthand for Hotel, Restaurant, Catering; Etc means that the industry was not already mentioned).

Employee\_Profession is the profession that the employee is employed in (Accounting, BusinessDevelopment, Commercial, Consulting, Engineer, Finance, HR, IT, Law, Management, Marketing, PR, Sales, Teaching, and Etc; Etc means that the industry was not already mentioned).

Source\_of\_Hire is where the employee was sourced from (Advertising, EmployeeRecommendation, FriendRecommendation, FriendsWithEmployer, JobSite, JobSiteVacantPosition, Recommendation, and RecruitingAgency).

Coach is whether the employee had a coach; and, if the employee had a coach, whether that coach was also the employee’s supervisor (“no” means there was no coach, “yes” means there was a coach, “supervisor” means that the employee had a coach, and that coach was also the employee’s supervisor).

Head\_Supervisor\_Gender is whether the head supervisor for that employee is male or female (“m” means male, “f” means female).

Wage\_Color is whether the employee has a white wage or a gray wage (“white” means that the employee has a white wage, this means that the employee is earning no more than what that employee states on their tax paperwork; “grey” means that the employee has a gray wage, this means that the employee is earning more than what that employee states on their tax paperwork).

Way\_to\_Work is how the employee gets to work (bus, car, or foot).

Extraversion is a measure of extraversion in the employee going by the OCEAN big 5 personality traits, this variable ranges from 1-10 in the dataset (1 means a person is more introverted, 10 means a person is more extroverted).

Agreeableness is a measure of agreeableness in the employee going by the OCEAN big 5 personality traits, this variable ranges from 1-10 in the dataset (1 means a person is less agreeable, 10 means a person is more agreeable).

Conscientiousness is a measure of conscientiousness in the employee going by the OCEAN big 5 personality traits, this variable ranges from 1-10 in the dataset (1 means a person is less conscientious, 10 means a person is more conscientious).

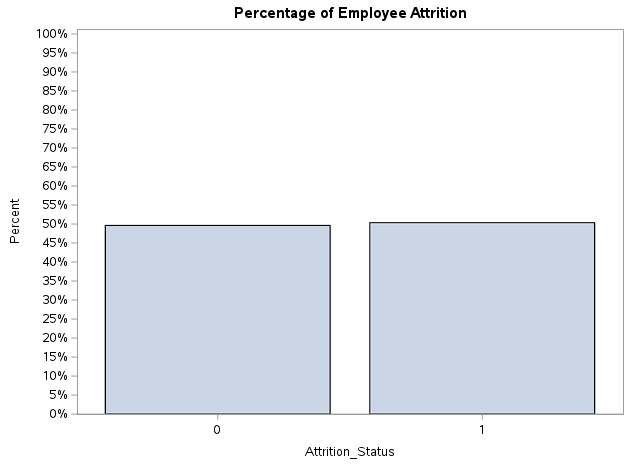
Neuroticism is a measure of neuroticism in the employee going by the OCEAN big 5 personality traits, this variable ranges from 1-10 in the dataset (1 means a person is less neurotic, 10 means a person is more neurotic).

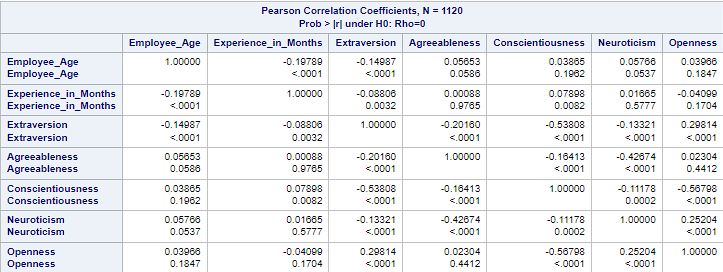
Openness is a measure of openness in the employee going by the OCEAN big 5 personality traits, this variable ranges from 1-10 in the dataset (1 means a person is less open, 10 means a person is more open).

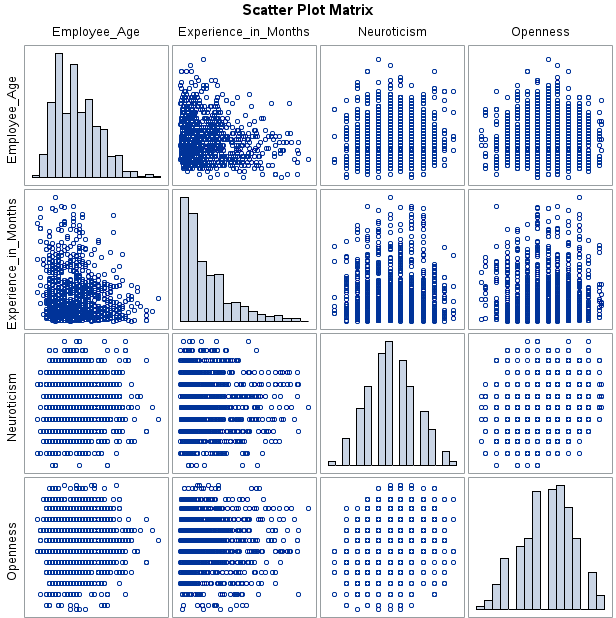
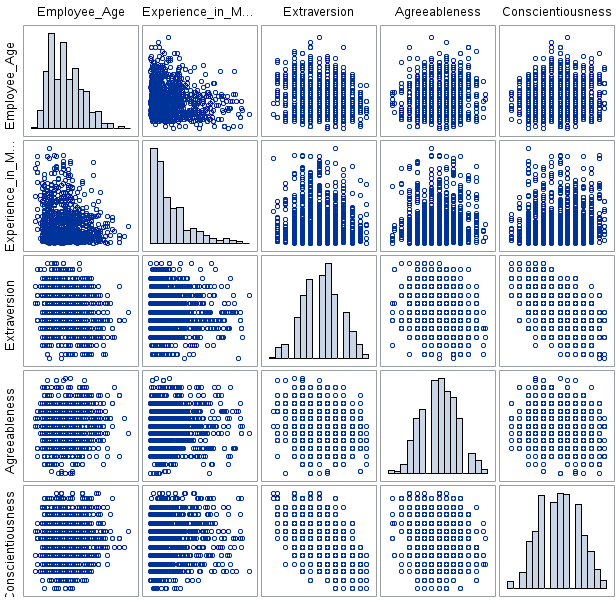
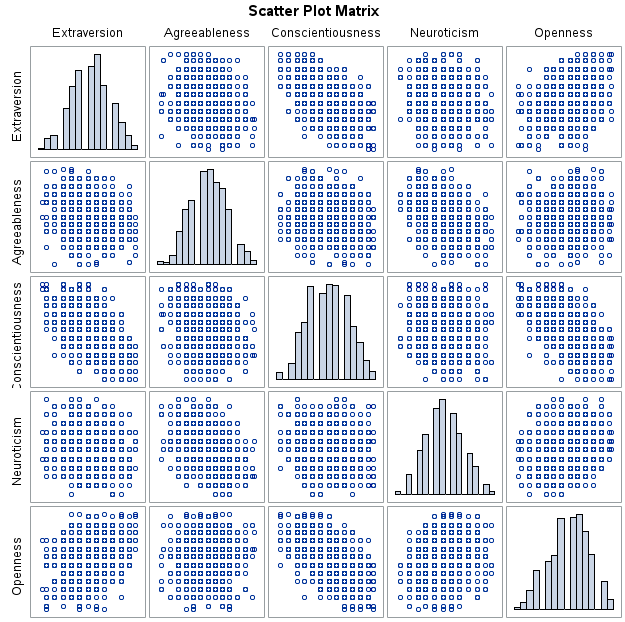
**EDA**

The purpose of EDA is to better understand the variables in a dataset. This is important for various reasons, one of which is the case in which multiple variables in the dataset are highly correlated, for instance, if there was a variable for weekly income and a variable for monthly income, these would be highly correlated. This can be problematic as it can throw off the importance of all the variables in the predictive model if these two variables were to be left in the analysis; therefore, one of these variables should be removed in this scenario. Such obvious variables do not exist in this dataset; however, sometimes variables can be not so obviously correlated, so it is still important to perform EDA to check for high correlation, as well as to help increase overall understanding of the dataset.

The histogram below (*see image below*) shows the overall distribution of employees who are leaving their organizations (Attrition\_Status is 1) and those who are staying at their organizations (Attrition\_Status is 0). As can be observed, the dataset consists of an almost perfect 50/50 distribution, so there will be no issue with trying to have the model predict what employees belong to a minority group, as there is no minority group.



 The images below (*see images below*) show the correlation between all the continuous variables in the dataset, including Employee\_Age, Experience\_in\_Months, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Notable are the correlation coefficients between Conscientiousness and Extraversion, Neuroticism and Agreeableness, and Conscientiousness and Openness. While these correlation coefficients are notable, ultimately, they are not large enough to warrant removal from the analysis, so all of these variables will remain in the analysis.



**Analysis**

Multiple ML modeling methods were chosen for analysis as the nature of the relationship between Attrition\_Status and each of the other variables is largely unknown, that is, whether each variable has a linear relationship, logarithmic relationship, another kind of relationship, or no relationship with Attrition\_Status. This ties into multiple ML methods being used in that each method utilizes different ways of looking for a relationship between the variables; logistic regression, for instance, works best with variables that can be converted into having a linear relationship as the result of a logarithm being applied; other ML methods, such as CART, are not so picky. All the models utilized a 70/30 split between the training set and validation set. The dataset was split so that there is data to predict on. If a dataset is not split, then a model is not predicting data; rather, it is fitting data, which is what is performed on the training set. The validation set tests the model that was built using the training set data and assesses how well it performs at predicting, as the data in the validation set is entirely new to the model, so it was not trained or fitted on that data.

The CART models output a numerical importance value for each of the independent variables, so this was the metric that was utilized in judging whether a variable was important in the CART models; the other models did not output such a value; therefore, the metric that was utilized in judging whether a variable was important in the model was the predictive accuracy (judging by sensitivity, specificity, AIC, BIC, and AUC when applicable) of the different models compared to models of the same method (ex. neural network compared to neural network, discriminant analysis compared to discriminant analysis, etc.).

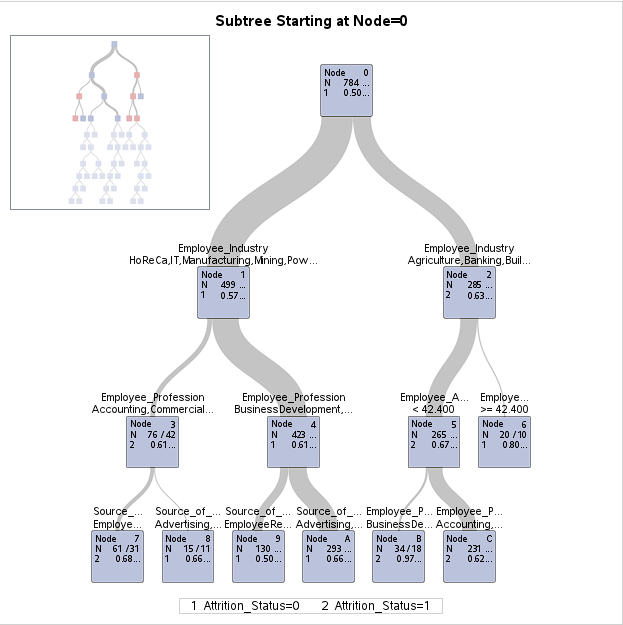
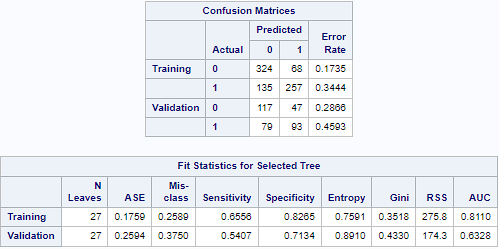
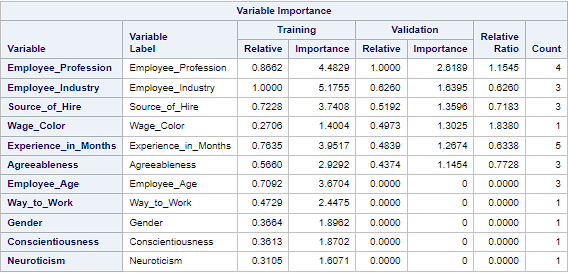
It is also worth mentioning that the CART models were the only models that did not have dummy variables run through them. The variables that did not have dummy variables created for them are Experience\_in\_Months, Employee\_Age, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, as they were deemed to be continuous variables as opposed to categorical variables.

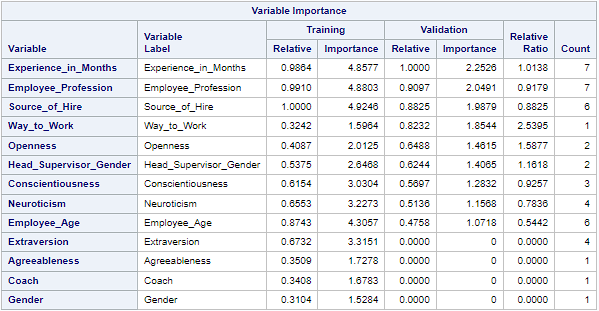
**CART**

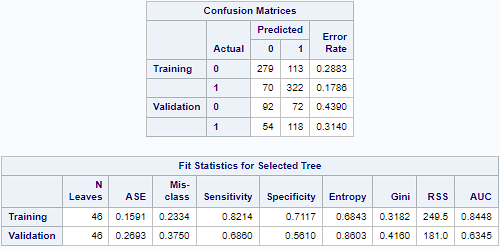
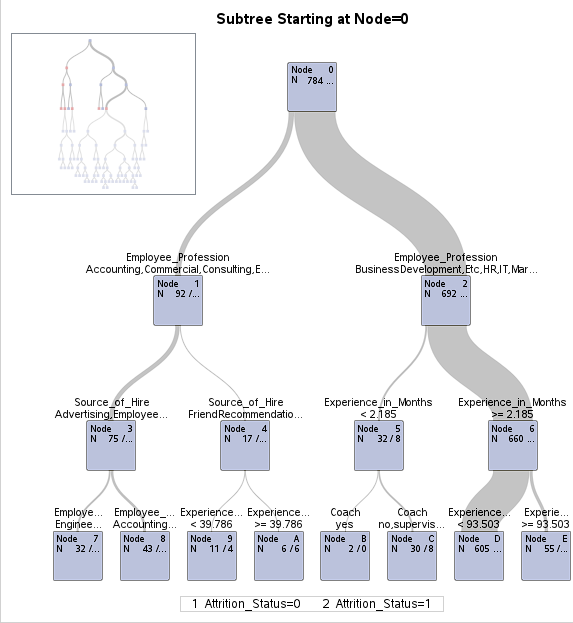
***Gini***

Two analyses were conducted for CART with Gini, one analysis included the Employee\_Industry variable, the other did not. This was the only difference between the two analyses. The results of the analyses are shown in the images below.

**With Employee\_Industry.**

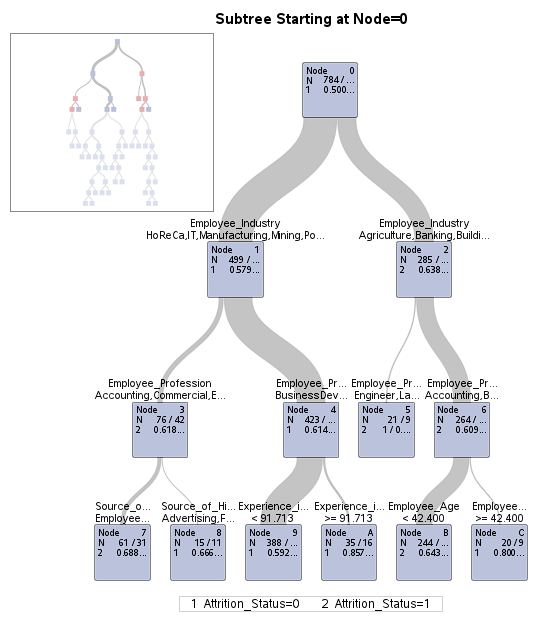
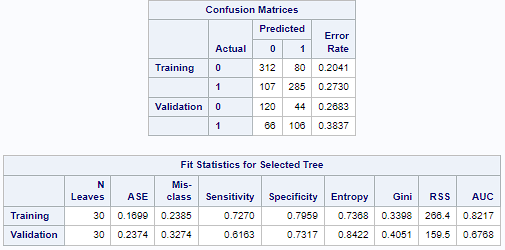


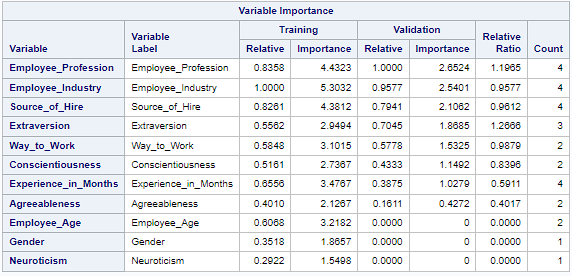
**Without Employee\_Industry.**

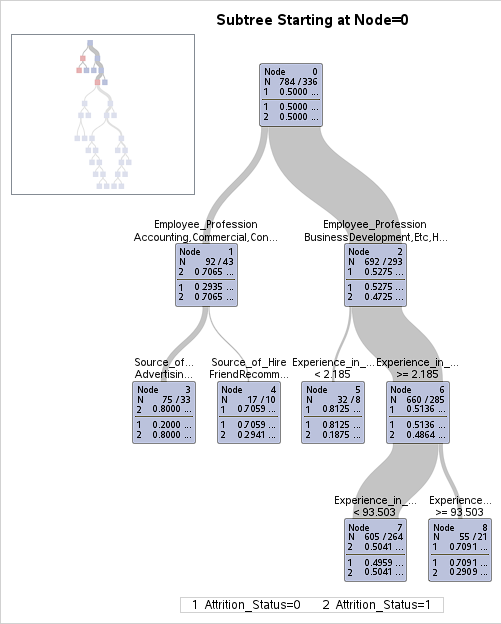
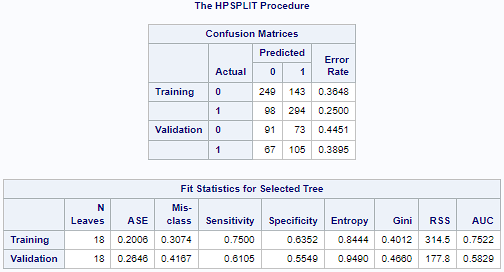


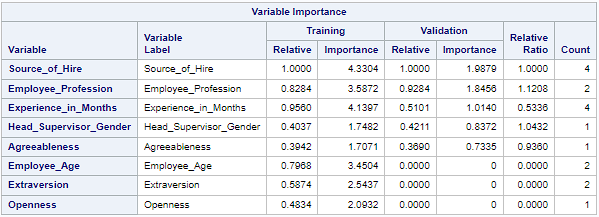
***Entropy***

Two analyses were conducted for CART with Entropy, one analysis included the Employee\_Industry variable, the other did not. This was the only difference between the two analyses. The results of the analyses are shown in the images below.

**With Employee\_Industry.**



**Without Employee\_Industry.**

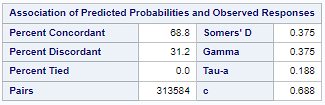
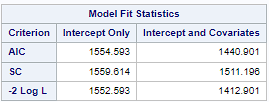


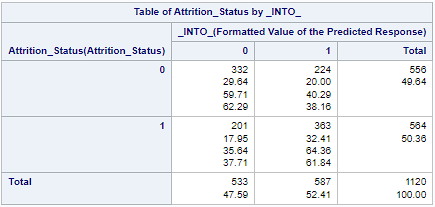
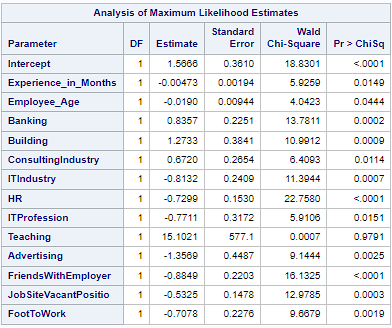
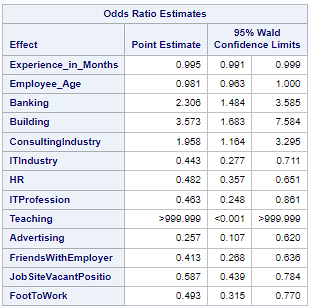
Both of the Gini analyses performed arguably worse than their Entropy counterparts in prediction; that is, CART with Gini including Employee\_Industry performed worse than CART with Entropy including Employee\_Industry and CART with Gini excluding Employee\_Industry performed worse than CART with Entropy excluding Employee\_Industry. Seeing as this is the case, Entropy is the clear winner between the two CART methods. That said, which Entropy analysis performed better is disputable. The most parsimonious model was based on Entropy without the Employee\_Industry variable; however, specificity dropped about 18% in the validation set without Employee\_Industry, meaning that the best predictive model was based on Entropy with the Employee\_Industry variable.

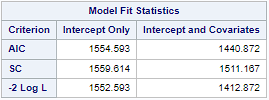
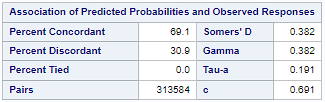
**Logistic Regression**

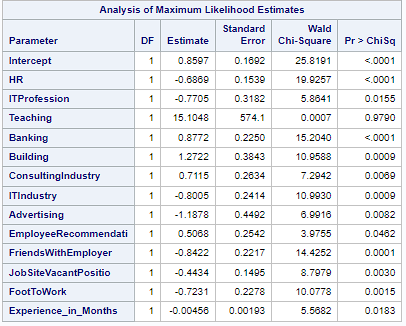
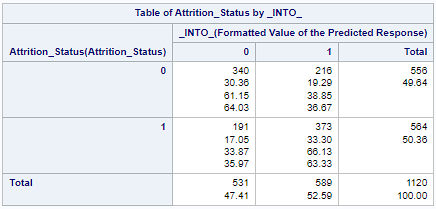
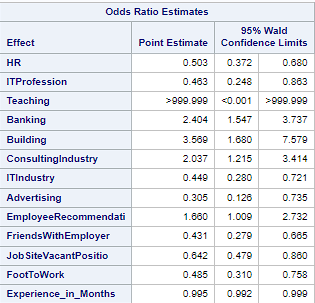
Nine analyses were conducted with logistic regression, three of the analyses were with forward selection, three of the analyses were with stepwise selection, and three of the analyses were with backward selection. None of the stepwise or backward selection analyses were viable due to there being a risk of a quasi-complete separation, so six of the analyses were dropped. This left three, all with forward selection. The first model included all variables, the second model included the variables deemed important by CART with Entropy including Employee\_Industry, and the third model included the variables deemed important by CART with Entropy excluding Employee\_Industry. The results of the analyses are shown in the images below.

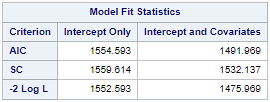
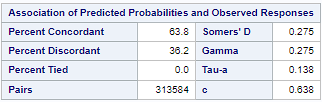
***All Variables***

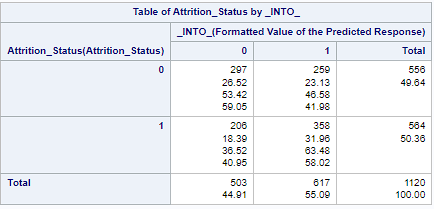
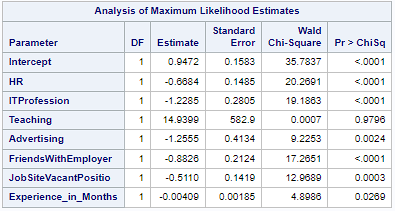
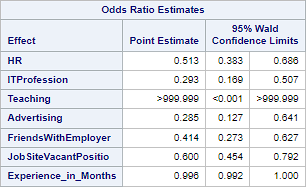




***With Employee\_Industry***



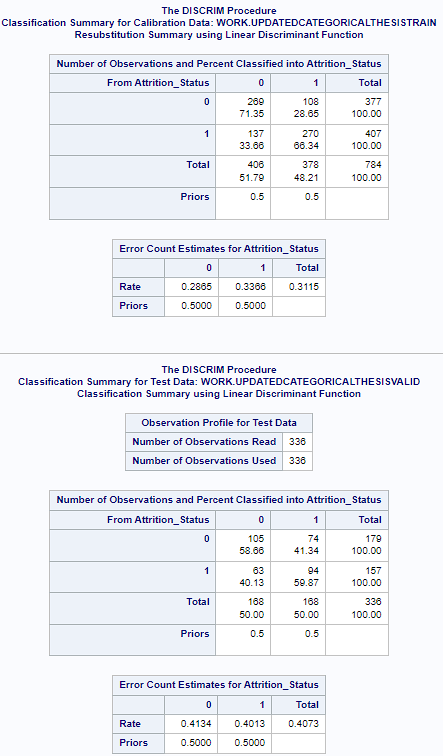
***Without Employee\_Industry***

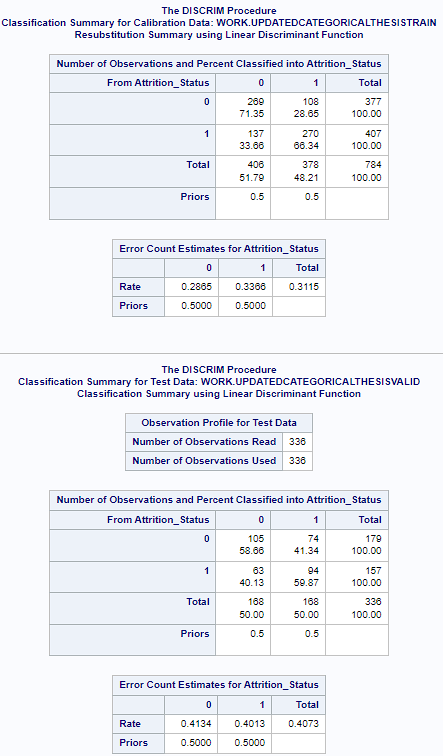


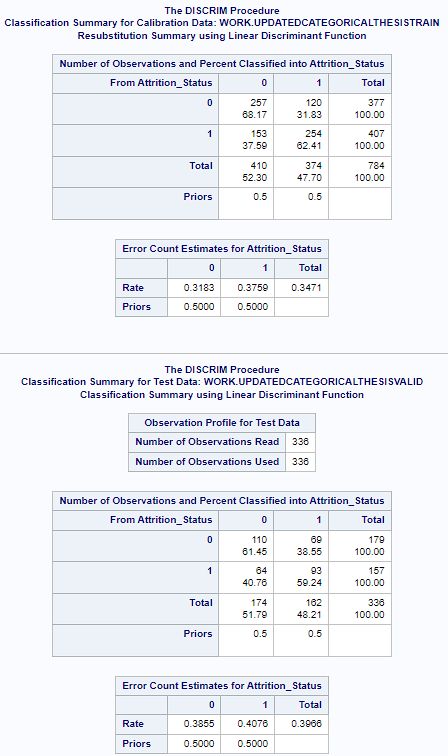
Overall, the best predictor was the second model, which included the variables deemed important by CART with Entropy including Employee\_Industry, the runner-up was the first model, which included all the variables, the worst model was the third model, which included the variables deemed important by CART with Entropy excluding Employee\_Industry. While the third model predicted the worst, it did not predict worse by that large a margin; so while the third model predicted the worst, it did not necessarily perform the worst, as it was the most parsimonious.

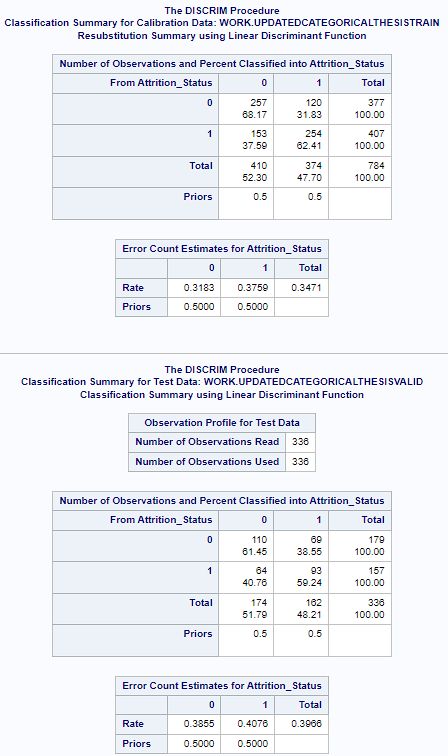
**Discriminant Analysis**

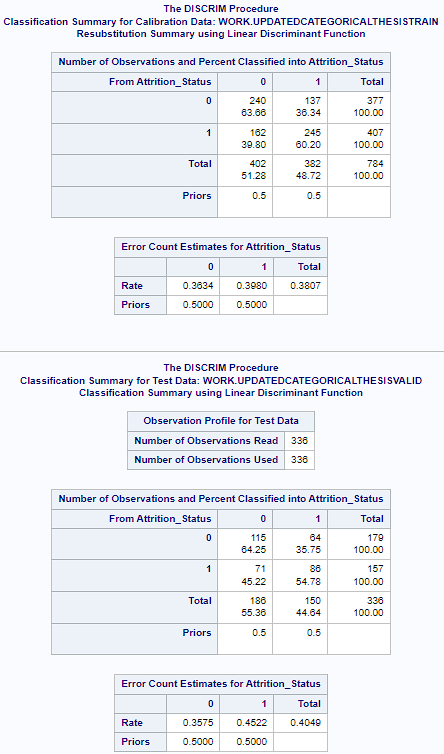
Similarly to what was performed for logistic regression, three models were created. The first model included all variables, the second model included the variables deemed important by CART with Entropy including Employee\_Industry, and the third model included the variables deemed important by CART with Entropy excluding Employee\_Industry. The priors were not adjusted for any of the models as the default 50/50 priors seemed well fit for this dataset. The results of the analyses are shown in the images below.

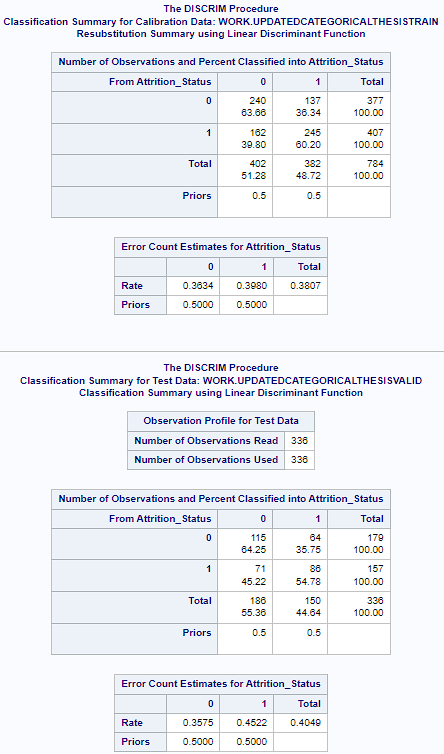
***All Variables***



***With Employee\_Industry***



***Without Employee\_Industry***

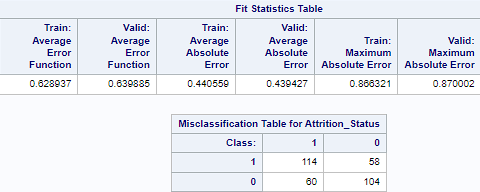


The best predictor was the second model, which included the variables deemed important by CART with Entropy including Employee\_Industry, the runner-up was the third model, which the variables deemed important by CART with Entropy excluding Employee\_Industry, the worst model was the first model, which included all the variables. This differs from the results of logistic regression, as the third model was the worst in that method; however, this method is mimicking logistic regression in that while the third model was not the best predictor, it did not necessarily perform the worst, as it was the most parsimonious.

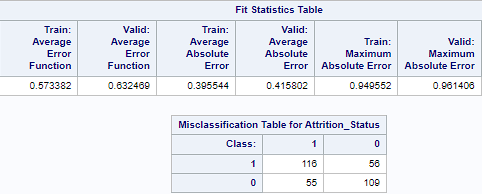
**Neural Networks**

Similarly to what was performed for logistic regression and discriminant analysis, three models were created. The first model included all variables, the second model included the variables deemed important by CART with Entropy including Employee\_Industry, and the third model included the variables deemed important by CART with Entropy excluding Employee\_Industry. The neural network with all the variables included was provided one hidden layer with 59 nodes and 10 tries, the neural network with the industry variable was provided one hidden layer with 40 nodes and 10 tries, the neural network without the industry variable was provided one hidden layer with 20 nodes and 10 tries. The number of nodes differs for all of these models as, through moderate experimentation, these are the combinations that yielded the best predictive results for each variation of the model respectively.

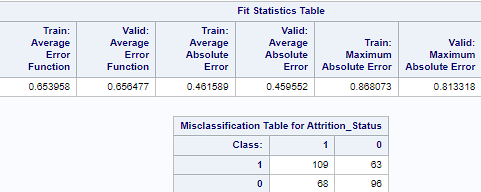
***All Variables***



***With Employee\_Industry***



***Without Employee\_Industry***

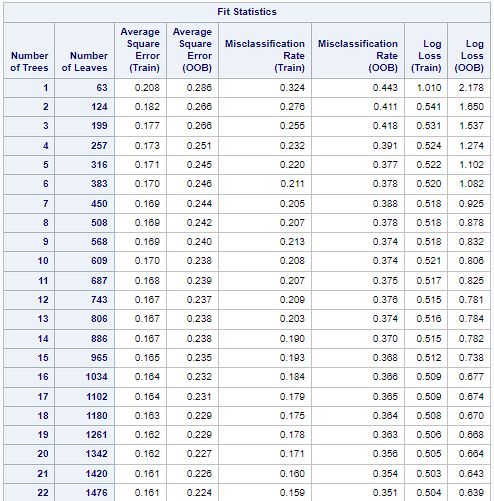


The best predictor was the second model, which included the variables deemed important by CART with Entropy including Employee\_Industry, the runner-up was the first model, which included all the variables, the worst model was the third model, which included the variables deemed important by CART with Entropy excluding Employee\_Industry. This mimics the results of logistic regression. Mimicking both logistic regression and discriminant analysis is that while the third model predicted the worst, it did not predict worse by that large a margin; moreover, it had at most half of the nodes of the two models, so while the third model predicted the worst, it did not necessarily perform the worst, as it was the most parsimonious in both variable count and node count.

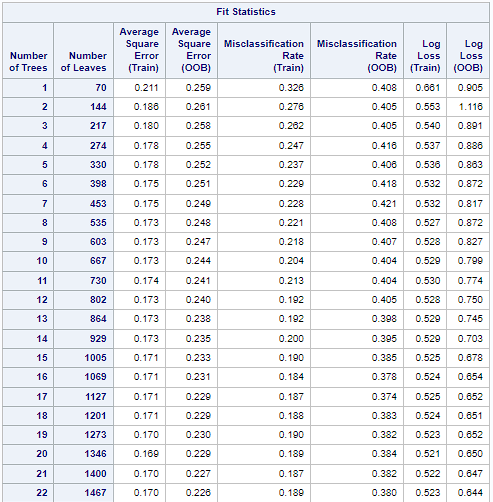
**Random Forest**

Similarly to what was performed for logistic regression, discriminant analysis, and neural networks, three models were created. The first model included all variables, the second model included the variables deemed important by CART with Entropy including Employee\_Industry, and the third model included the variables deemed important by CART with Entropy excluding Employee\_Industry. All of the random forest models all had a minimum leaf size set to 6. The maximum number of trees was set to 100 for all models; however, for consistency’s sake, only up to the 22nd tree is shown for all models, as performance began to largely stagnate past the 22nd tree for all models.

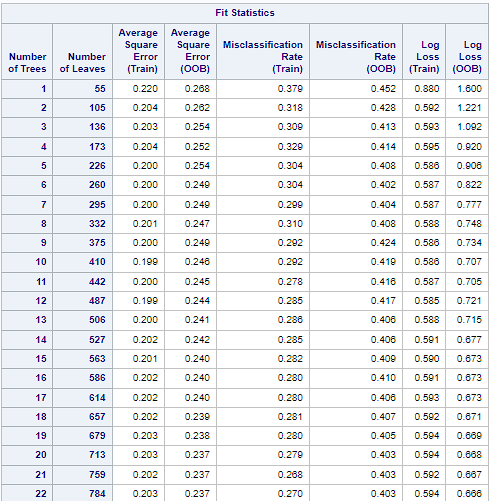
***All Variables***



***With Employee\_Industry***



***Without Employee\_Industry***

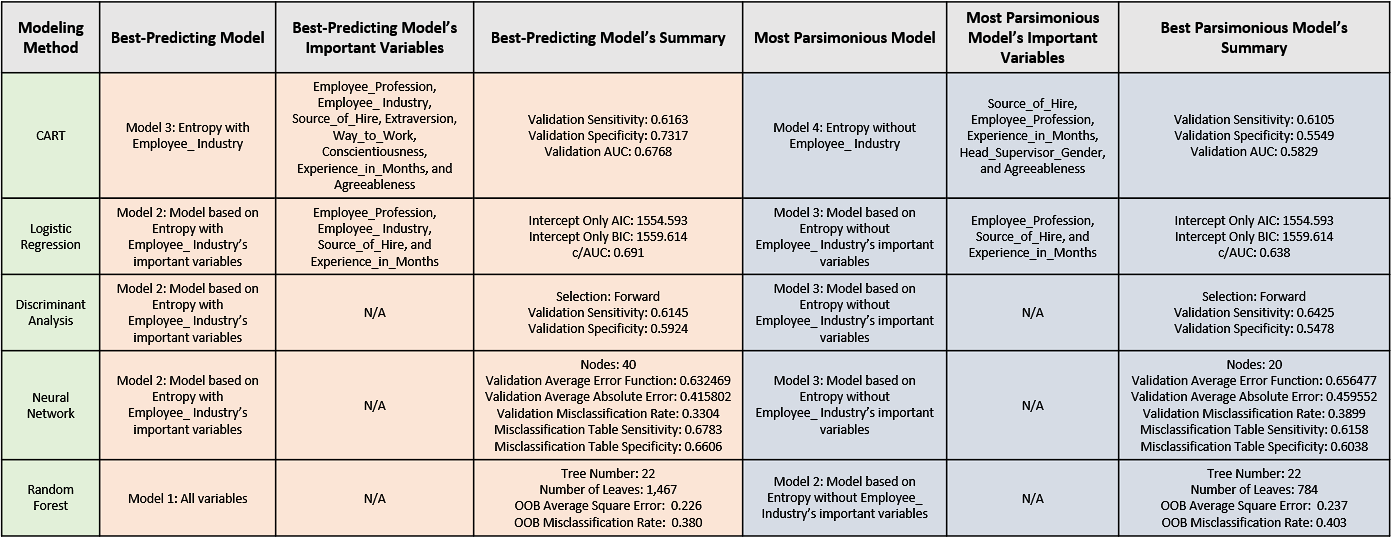


The best predictor was the first model, which included all the variables, the runner-up was the second model, which included the variables deemed important by CART with Entropy including Employee\_Industry, the worst model was the third model, which included the variables deemed important by CART with Entropy excluding Employee\_Industry. This differs from the results of logistic regression, discriminant analysis, and neural network. It is again worth noting that the OOB misclassification rate does not vary greatly between any of the models, so while the first model was the best predictor, it was not necessarily the best performer, as the third model was the most parsimonious.

**Results**

**The Models**

The best-predicting model overall, based on the comparison of metrics (*see image below*) seemed to be the neural network model with the variables deemed important by CART with Entropy including Employee\_Industry. The most parsimonious model seemed to be logistic regression with the variables deemed important by CART with Entropy excluding Employee\_Industry. Generally, the model variants with the variables deemed important by CART with Entropy excluding Employee\_Industry tended to perform the worst; however, the model variants with the variables deemed important by CART with Entropy excluding Employee\_Industry also did not tend to lag very far behind the best-predicting model variants.



This indicates that the hypothesis of there existing variables that are important to predicting employee attrition (Attrition\_Status) regardless of industry has weight, as if it was the case that Employee\_Industry was the only important variable, the models with the variables deemed important by CART with Entropy excluding Employee\_Industry would have lagged extremely far behind in prediction ability, which they did not. Taken as a whole, the models indicate that the industry that the employee is employed in (Employee\_Industry) is important in predicting attrition, but not the only important variable in this dataset.

**Important Variables**

Comparing all the models, the variables in this dataset that are important in predicting employee attrition (Attrition\_Status) are Employee\_Profession, Employee\_Industry, Source\_of\_Hire, and Experience\_in\_Months. Employee\_Profession, Source\_of\_Hire, Experience\_in\_Months, and Agreeableness all overlap as important in both of the CART models with Entropy, Employee\_Industry was not input into the second model; so, naturally, it was not important. Utilizing the results of the modeling method that was most parsimonious, logistic regression, provides additional confirmation of the importance of these four variables, as the two variant models of logistic regression that were based on the variables deemed important by CART with Entropy confirm the importance of Employee\_Profession, Employee\_Industry, Source\_of\_Hire, and Experience\_in\_Months. Notably, both logistic regression models do not incorporate Agreeableness into the models, unlike both of the CART models with Entropy. Seeing as both logistic regression models predicted comparably well to their CART counterparts, it is reasonable to conclude that Agreeableness as well as all other variables were not important in the prediction of employee attrition (Attrition\_Status) in this dataset.

**Conclusion**

All-in-all, these results indicate that there are, in fact, variables that are important in the prediction of employee attrition (Attrition\_Status) across multiple industries. These variables include Employee\_Profession, Source\_of\_Hire, and Experience\_in\_Months. What this means is that, in the effort to increase retention rates (and decrease attrition rates), organizations and managers should pay particular attention to the profession the employee is employed in, the avenue in which the employee was acquired, and how much experience the employee has.

The specific Employee\_Profession, Source\_of\_Hire, and Experience\_in\_Months will likely vary from organization to organization; but, based on this dataset, organizations should pay particular attention to employees who do not work in the HR or IT professions, employees who were sourced through avenues other than advertising, employee recommendation, employer connections, or a job website, and employees who have lower experience, as these employees are more likely to leave the organization. Specifically, going by the third logistic regression model’s odds ratio output, employees working in the HR and IT professions and employees who were sourced through advertising, friendship with the employer, or a job website are significantly less likely to be leaving the organization, and for each month an employee has gained experience (Experience\_in\_Months), an employee is 0.4% less likely to leave the organization.

**Application**

These results can be utilized by any organization with reasonable caution and care to adjust hiring and management practices in the hopes of increasing employee retention rates and decreasing employee attrition rates. To reiterate, the application of the findings of this paper should be conducted first with great thought, consideration, caution, and care, as every organization is different, and these differences could make the findings of this paper extremely relevant for one organization and extremely irrelevant to another. That said, once this caution has been taken, it is the advice of this paper that employers should, in the effort to lower employee attrition rates, focus on sourcing employees through personal connections, advertising, and job sites, as these employees are less likely to leave the organization. Employers should also focus retention efforts primarily on employees who are not in the IT or HR professions and focus those same retention efforts on employees with less experience, as this dataset indicates that these employees are more likely to leave the organization.

**Limitations**

As is the case with all datasets, there are limitations. Most notably for this dataset is the omission of a continuous variable regarding employee wage. This dataset includes the variable Wage\_Color, but this variable is categorical and much more limiting as a result. It could very well be the case that if there had been a continuous wage variable, it would’ve been important. Another limitation is with regards to the personality variables of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, being deemed as continuous; an argument could be made that they should have been treated as categorical, but it was ultimately decided that treating them as continuous made more sense. On the same topic of the personality variables, an additional criticism could be made that the personality variables and gender were both included, as there is research that indicates that some of these traits tend to be, on average, higher in one of the genders as opposed to the other (Weisberg et al., 2011); still, it was ultimately decided that keeping both the personality variables and the gender variable in would be preferable to the alternative of removing either the gender variable or all of the personality variables. Lastly, this dataset was gathered in the Russian Federation, meaning that caution should be taken when generalizing the results to nations besides the Russian Federation.

**Biographical Note**

The author of this paper, Kyle Anthony Gable, is graduating from Oakland University in April of 2023 with a Bachelor of Science in Management Information Systems with a Business Analytics specialization. He first took an interest in analytics when he had his first statistics class in high school. Having worked on multiple projects and for different employers since that class, he found himself disliking how inefficient and outdated many processes within organizations are and felt that it was his calling to make a difference where he could. He seeks to become a data analyst or a business analyst and make his name known in the realm of analytics. It is his goal to improve profitability, efficiency, and employee welfare in every organization he is employed by or works with. It is his belief that happier employees make for more profitable corporations, and more profitable corporations make for happier management, so he seeks to ultimately improve the lives of management and employees with his work. This paper is his first attempt at what he hopes will be many attempts to make his name known in the industry by providing to the public domain data that he sees as useful for every industry and every organization, to varying degrees.

**Citations and References**

Alsheref, F.K., Fattoh, I. E., & Ead, W.M. (2022). Automated prediction of employee attrition using ensemble model based on machine learning algorithms. *Computational Intelligence and Neuroscience*, 1–9. https://doi.org/10.1155/2022/7728668

Chung, D., Yun, J., Lee, J., & Jeon, Y. (2023). Predictive model of employee attrition based on stacking ensemble learning. *Expert Systems with Applications*, *215*. https://doi.org/10.1016/j.eswa.2022.119364

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. *International Journal of Organizational Analysis, 28*(6), 1273-1291. https://doi.org/10.1108/IJOA-10-2019-1903

Garter. (n.d.). *Attrition*. Retrieved April 5, 2023, from https://www.gartner.com/en/human-resources/glossary/attrition

Guerranti, F., & Dimitri, G. M. (2023). A comparison of machine learning approaches for predicting employee attrition. *Applied Sciences, 13*(1), 267. https://doi.org/10.3390/app13010267

Haldorai, K., Kim, W. G., Pillai, S. G., Park, T. (Eliot), & Balasubramanian, K. (2019). Factors affecting hotel employees’ attrition and turnover: Application of pull-push-mooring framework. *International Journal of Hospitality Management*, *83*, 46-55. https://doi.org//10.1016/j.ijhm.2019.04.003

Jain, P.K., Jain, M., & Pamula, R. (2020). Explaining and predicting employees’ attrition: a machine learning approach. *SN Applied Sciences*, *2*. https://doi.org/10.1007/s42452-020-2519-4

Krishna, S., & Sidharth, S. (2022). HR Analytics: Employee attrition analysis using random forest. *International Journal of Performability Engineering, 18*(4), 275–281. https://doi.org/10.23940/ijpe.22.04.p5.275281

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

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting employee attrition using machine learning approaches. *Applied Sciences*, *12*(13), 6424–6442. https://doi.org/10.3390/app12136424

Srivastava, P. R. & Eachempati, P. (2021). Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach. *Journal of Global Information Management (JGIM), 29*(6), 1-29. http://doi.org/10.4018/JGIM.20211101.oa23

Weisberg, Y. J., Deyoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, *2*, 178-178. https://doi.org/10.3389/fpsyg.2011.00178