David Owen
Puja Subramaniam
Mohammed Saqib Asghar
Shelby Watson

What Motivates People to Leave their Jobs?

Description of Project Goals

People Analytics has increased in importance over the years. As employers collect more data about their employees, they've begun to use these data to their advantage to better understand what motivates their employees, and how they can retain top talent. Understanding the reasons behind why employees leave a company, both voluntary and involuntary, can help employers understand what motivates individuals. Knowing which factors most significantly affect attrition allows companies to foster the best work environment and conditions for employees to stay.

The broad applicability of this topic led us to choose our dataset. Having insight on this data can help an employer target key changes to make in the business in order to improve job satisfaction amongst employees, and reduce attrition rates. These analytics can also help to pinpoint what resources management might need and can even spark new change in the company culture. While employee attrition is unavoidable and there are many factors out of a company's control, there are ways for companies to avoid unnecessary attrition. Excessive attrition is ultimately costs the company money, and can affect the general environment and morale of the company (Gray 2016). Each time an employee leaves and a new employee is hired, the company must utilize time and resources to interview, hire, and train said new employee, which results in long periods of lowered productivity (Markovich 2019). Thus, the value of people analytics is essential for the Human Resources team at a company to continue to engage and retain top talent.

The goal of this project is to better understand what factors play a role in employee attrition, and how we can leverage these to reduce attrition rate. We do our best to determine what factors, specifically those in the employer's control, can be used to foresee whether an employee is likely to leave the company soon. We looked for a variety of patterns in the data to see what specific factors related to their job may affect their decision to stay at the company. We broke this down by age group to see specifically what motivates each generation of the workforce. We also looked for trends in educational backgrounds, to better understand which job functions best suit which individuals. We then see if we can take a sample of employees and determine their likelihood of leaving the company.

Exploratory Data Analysis

We start by exploring some basic facts about our data set. Graphs visualizing some of these observations can be found in the appendix at the end of the report. We have information on a total of 1470 employees, and there is a ~16% attrition (Figure E). The average age of our employers is 37, with 60% of them male and 40% female. We also do a breakdown of marital status, and observe that 32% are single, 46% are married, and 22% are divorced (Figure A).

Next, we identify some patterns. One of the first things that we noticed was that attrition is highest amongst the job roles of Sales Executives, Lab Technicians, and Research Scientists (Figure C). Looking at different job departments, it seems like attrition is higher in Sales and R&D (Figure D). In addition, attrition is similar in gender across various departments. Looking at more personal aspects for individual employees, it looks like attrition is higher among employees who are single (Figure B). When employees live over 8 kilometers away from work, they are more likely to quit (particularly married people). In regards to work-life balance, a level between a 2.5 and a 3.0 on a 4.0 scale has the least attrition. Additionally, attrition seems to be higher in the lower earning level, and among employees who have been working at the company for a lesser amount of time. Also, attrition seems to be higher among employees who do not have stock options. This makes sense, as they do not have any direct stakes in the company other than their job.

Next, we note a few things that seem somewhat abnormal. For example, at Job Level 5, employees are more likely to quit if their distance from home is between 2.5 and 5.0 kilometers, but are less likely to leave if the distance from home gets closer to 7 kilometers (Figure F). Another interesting thing to note is that Percent Salary Hike (the rate at which an employee's salary increases) does not appear to have a major impact on the attrition rate (Figure G). One last abnormality to note is the attrition is high even for employees promoted in the last three years, indicating that this might not be a significant factor in assessing the attrition rate.

Solutions and Insights

In the case that we are presenting our model as a judgment of what qualities of the workplace can be changed to improve employee retention, there are many features that must be dropped as an employer cannot affect them. For this model, we performed the first round of dimensionality reduction by hand. They were the following:

- Age
- Distance from Home
- Education / Education Field
- Marital Status

- Number of Companies Worked
- Over 18
- Performance Rating
- Relationship Satisfaction
- Total Working Years
- Years at Company

All dummy variables for the fields were removed as well, reducing dimensionality by 20.

Decision Tree:

For predictive purposes, we built two classification trees and utilized principal component analysis to decide upon their complexity. After one-hot encoding our categorical features, there were 55 total columns to consider.

The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. Given the distribution of the data, each value in the dataset will have the sample mean value subtracted, and then divided by the standard deviation of the whole dataset.

As is required by PCA, we standardized the data to a mean of 0 and variance of 1, saving the transformation in order to use it on future data. We further reduced the dimensionality as much as possible while retaining 95% of the initial dataset's variance, which was achievable with 15 features. The following had a significance level above 0.02:

- **Overtime: Y/N**
- **Monthly Income**
- Job Level
- Employee Number
- Stock Option Level
- Job Role = Sales Executive
- Daily Rate
- Job Satisfaction
- Years with Current Manager
- Environment Satisfaction
- Job Involvement

The first two features (highlighted in bold) maintained an importance of about 0.14, at least three times greater than any of the others. This quality was consistent across multiple iterations of the model. The rest were minimally important, but a visual inspection of the model showed that they did increase node purity.

The model itself achieved an out-of-sample accuracy of 64.9%. This performance metric takes into account a heavy imbalance of the labels in our data, where Attrition = No is 6 times more common than Attrition = Yes. Without accounting for this, the overall prediction accuracy was 86.6%, but such a metric is meaningless. For comparison, a model which did not take this imbalance into account and simply guessed 'No' for an entire test set would be correct just over 80% of the time.

We planned to present a second tree, where all variables were considered and it could be used as a prediction algorithm for a management team to asses the attrition risk of their employees, but we found that including the other 20 variables, even when most were removed again by PCA, significantly reduced the model's prediction accuracy. For that reason, we chose to stand by our first model.

If this model were to be presented to a hypothetical management team, we would present the feature importance first. While the actual scores changed, their rankings relative to each other were very consistent across every model tested, and we are confident in saying that they are significantly correlated and should be considered when assessing attrition risk.

Logistic Regression:

Since our problem is one of classification, we also created a model using logistic regression. Like with the decision tree model, we eliminate variables that were out of the employer's control and dummy coded any necessary predictors. After accounting for our imbalanced data in the model, we are able to achieve an out of sample accuracy of 60.9%. Similar to our decision tree, we find that the same predictors have the most weight in our logistic regression model.

Things we can see from the data, in both models:

- Individuals who work overtime are more likely to leave a company.
- The higher the stock option, the less likely the individual is to leave.
- The less job involvement, the more likely the employee is to leave.
- Those who travel frequently are also more likely to leave a company.

If we dive deeper into the features, we see some underlying correlation between these variables. Individuals in sales have to travel frequently to visit their clients; their travel often requires long hours, causing them to work overtime. While these three predictors are highly related, we can also look at them individually to see what changes we as an employer can push. Sales reps specifically undergo extensive training, and losing those with experience can be costly for an employer. Instead, our analysis brings to light the idea that perhaps we need to find ways for sales reps to travel less frequently; for example, if we have locations across the country, maybe we create more satellite offices for sales reps to work out of so that their travelling distance is minimized.

Thus, we can see that more than using logistic regression as a way to predict whether an employee will leave or not, it is helpful to look at the weights of each of the predictors. These values help drive changes that can be made to various roles.

Appendix

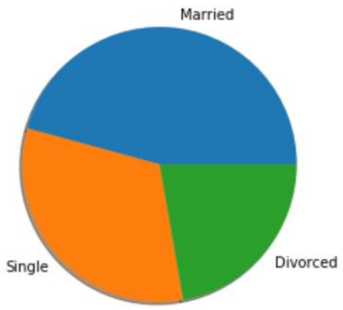


Figure A

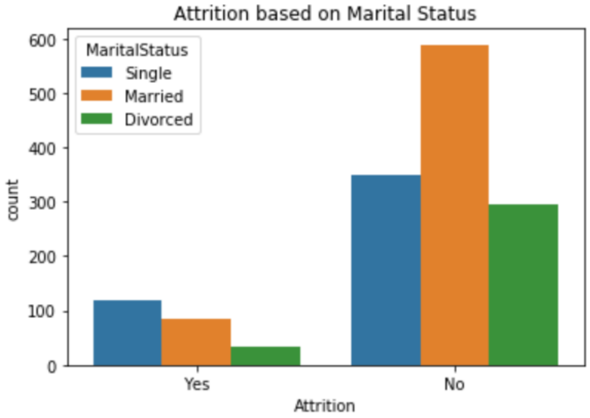


Figure B

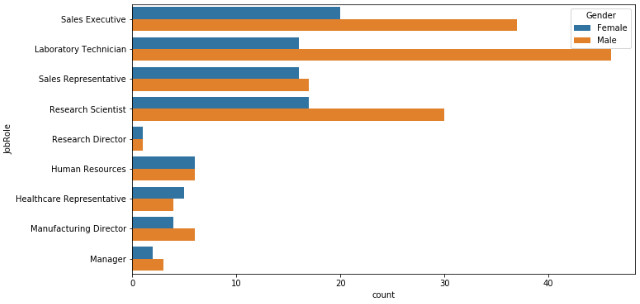


Figure C

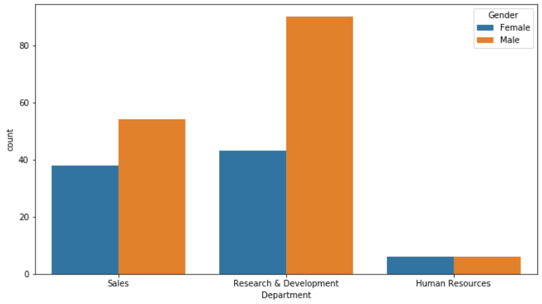


Figure D

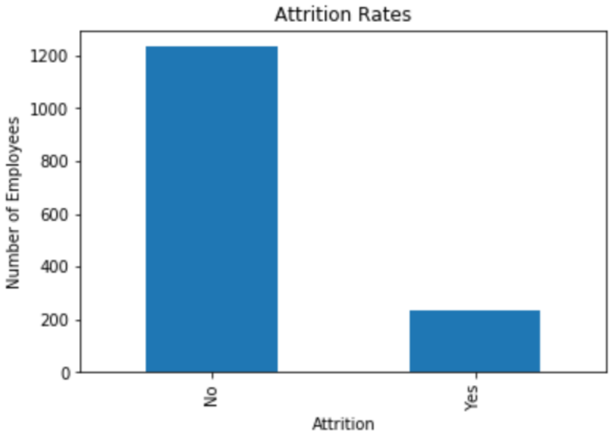


Figure E

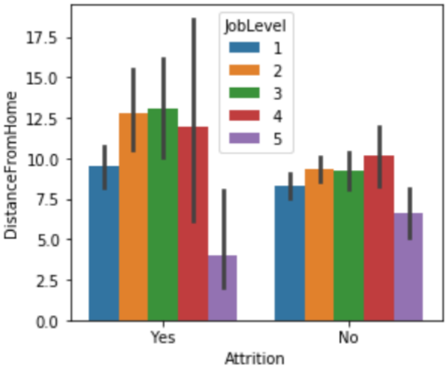


Figure F

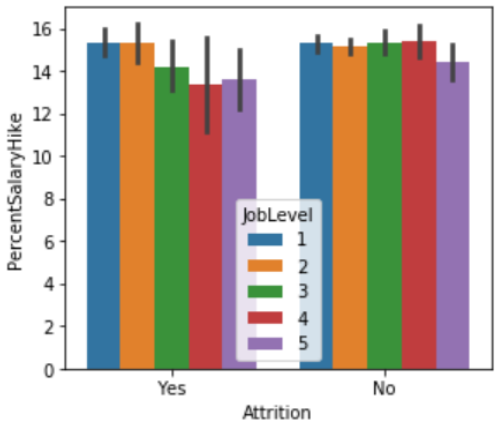


Figure G

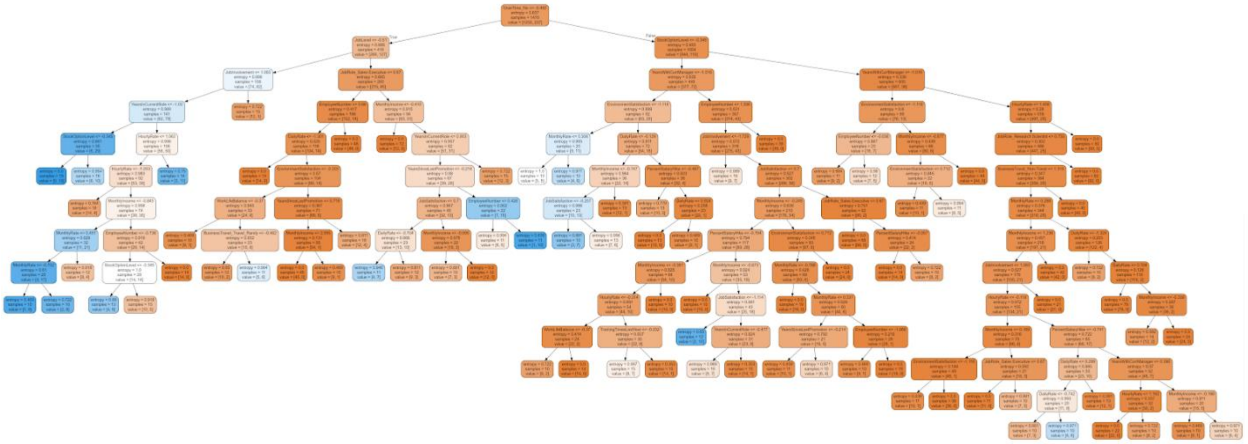
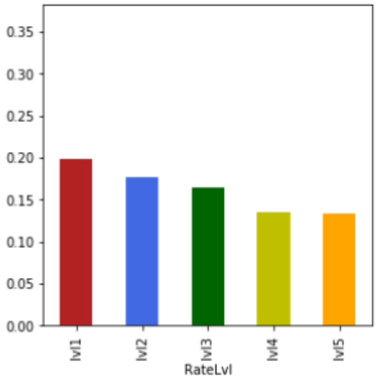
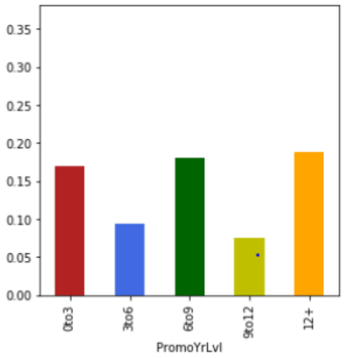


Figure H





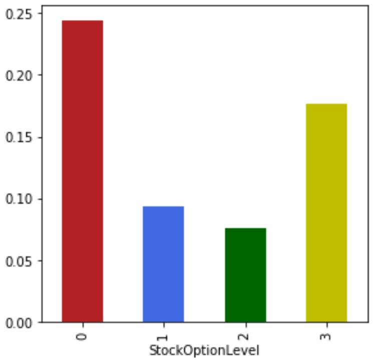


Figure I

Figure J

Figure K

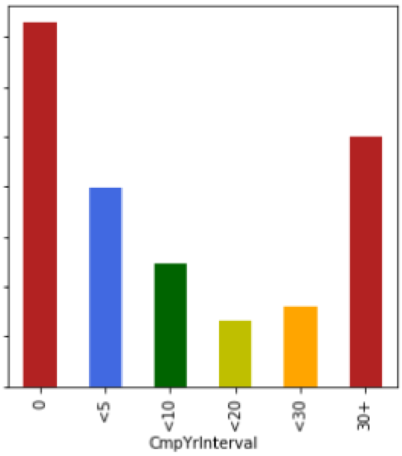


Figure L

References

Gray, C. (2016, October 26). The Impact of Attrition on a Business. Retrieved August 7, 2019, from https://smallbusiness.chron.com/impact-attrition-business-21348.html

Markovich, M. (2019, February 04). The Negative Impacts of a High Turnover Rate. Retrieved August 7, 2019, from https://smallbusiness.chron.com/negative-impacts-high-turnover-rate-20269.html

Pavansubhash. (2017, March 31). IBM HR Analytics Employee Attrition & Performance. Retrieved August 1, 2019, from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset