Analysis of IBM HR data

Ria Pan, James Barbour, Fahmi Khalid

Abstract

The purpose of this project is to analyze data provided from IBM's Human resource department in order to understand the common characteristics in employee attrition, or possible employee discrimination. In order to weight the importance of each variable presented, we filtered the data based on variables that were integer type and performed a correlation analysis. By doing so we were able to develop a basic understanding of which variables to include in our predictive model. The predictive model that we have developed has proven to predict employee attrition with an 84% accuracy rate. We also wanted to consider relationships in non-integer type variables in hopes of observing possible discrimination trends that can be supported by the dataset. We developed a number of visuals in order to justify our conclusions. Lastly, we strived to develop a user-friendly Shiny app for possible managers to use in order to predict employee attrition. All code and details shall be presented in the appropriate Rmd files as part of the final project.

Overview and Motivation:

The issue of keeping one's employees happy and satisfied is a perennial and age-old challenge. If an employee you have invested so much time and money leaves for "greener pastures", then this would mean that you would have to spend even more time and money to hire somebody else. Alongside these reason maintaining happy employees have resulted in increase profit for companies based on various studies done in industry. By understanding the main reasons why employees leave and possible discrimination, the increase probability your company will succeed.

Related Work

In class we had discussed effective techniques of communication. As a team we took that to heart and wanted to develop models that were not only easy to understand but visually pleasing. We have looked into various psychology journals, which study employee habits and utility in order to develop a healthy reasoning of what possible variables are important based on published studies.

Initial Questions

- Does Marital Status have a significant effect on wages?
- How does performance rating contribute to monthly income?
- How do different departments contribute to employee attrition?
- How do different job roles contribute to employee attrition?

- Can we give a general idea of working atmosphere based on department or job role attrition?
- Can we predict employee attrition?
- Does a higher education level contribute to lower attrition levels?
- How can we tell if an employee is happy?
- Does work life balance concur with years in position or years with company? How about with monthly income?

Some of these questions were answered by using visuals such as scatter plots, bar graphs, and others. While with other questions such as predicting employee attrition, we developed a healthy predictive model based on the random forest classifier, which enabled an 84% accuracy rate. Though some of the questions were straightforward and were answered with yes or no, a few of them were impossible to determine. One impossible question to answer was "How can we tell if an employee is happy." Essentially this question boils down to looking into background studies, which observe employee utility from psychology journals. Yet with what these studies propose we did not have the data for thus forcing us to abandon a few questions. One of the main issues with the data set was not the organization but the type of the variables and transforming different types to allow for smooth analysis. For example; on two of our team members laptop the "Age" variable operated as a normal numeric yet with the last team mate the "Age" variable only produced errors. Running into small technical issues which were resolved using Google while also performing other yak shaving activities did slow us down but in the end allowed for worthwhile analysis.

Exploratory Data Analysis

Considering our two goals of predicting employee attrition and possible discrimination, we aimed at creating visuals that would tell us a story on either subject.

Exploratory plot 1:
Represents the average monthly incomes based on gender which we then faceted against Job roles to observe the possible pay gap between genders across various Job roles. From first looking at the graph you may see slight gaps between the genders in certain roles such as Research Director and Manager which one might conclude as insignificant. But if we take the sum for the year of monthly income for Research Director then males will be making a total of $199,896 versus females making a total of $181,728. A $18,168 pay gap!

Exploratory plot 2:
We wanted to know based on Job roles where was the attrition coming from? With the bar graph we found that the top three job roles with the highest attrition by percentage were Healthcare Representative, Manufacturing Director, and Laboratory Technician. So why does this graph matter? We suggest that IBM looks into the underlying reasons why these roles have the highest loss. From the definition we understand that any loss will result in cost for the company. Are employees unhappy in these roles, are they treated

fairly, and are there opportunities for growth?

Exploratory plot 3:
After grouping by relationship status, we averaged the monthly income in order to observe any possible trends that occur based on marital status. We concluded that employees who were single observed a deficit of $897 in average monthly income relative to the divorced and married categories.

Exploratory plot 4:
Similar to the first plot we grouped our x-axis by “Relationship Satisfaction” and observed average monthly income of the 4 groups. We observed that as the level of satisfaction an employee perceives in their relationship the higher average monthly income they earn. Some psychology studies have concluded that a happy and successful relationship outside of work results in being more productive and thus earning more at work. Although we have the background theory, we do not have further data to support these claims, but to simply make observations and state our hypothesis.

Exploratory plot 5:
We observed the average monthly income based on Job satisfaction and found an interesting trend. As the level of satisfaction for one’s job decreases, the average monthly income that one earns increases. Our reasoning begins with the idea that with increased pay comes increased amounts of responsibility, which in turn causes more stress to perform well and ensure that one’s team is performing well.

Exploratory plot 6:
This is a summary of a linear regression of Work life balance against Years since last promotion in an attempt to distinguish the overall atmosphere and attitude an employee who stays longer with the company develops. The results were inconclusive.

Exploratory plot 7:
We wondered whether Performance rating would have an effect on Monthly income thus we used a box plot to compare the difference performance ratings beside each other. The conclusion was that there was not enough significance between the rating groups to emphasize any trend.

Exploratory plot 8 + Exploratory plot 9:

Performing a correlation against Age, daily rate, distance from home, education, and employee number we did not find any significant correlation. Thus later on we decided to perform a correlation on all numeric variables against each other.

Exploratory plot 10:
As a general plot we wanted to understand, regardless of job roles, education, or other factors how does the average monthly income differ between males and females. The

surprising conclusion from the graph was that females overall made $306 more money than males. It would be then be helpful to look at our first exploratory plot to understand where this trend may be coming from.

Exploratory plot 11:
Understanding how many females and males account for job roles is also an important question to ask. We may move away from monthly salary and ask based on the amount of each gender is there a tendency to hire or attract a certain gender into certain positions. Questions to ask for instance would be; in the Laboratory technician positions is the gap there due to a shortage of qualified female laboratory technicians or are the hiring managers discriminating against females. Now statistics with inadequate data will not give you the immediate answer but it will provide a compass of where to look.

Exploratory plot 12:
By grouping according to gender and Job role we then summarized the mean monthly income in order to create a scatterplot that in theory would emphasize the difference between genders across all job roles. Though a scatterplot in this situation we found not to be the most influential of graphs.

Exploratory plot 13:
We could have coded a correlation for each variable we may have thought was important individually but this would have taken too much labor. Thus what we ended up doing was segmenting the original data set and by taking only the integer based variables and performing a correlation analysis. The deeper the blue circle, the higher the correlation between the two variables is. This gave us a more specific set of variables to test against one another. The variables we found to be over .70 correlated are:

1. Job Level VS Monthly income
2. Total working years VS Job level
3. Total working years VS Monthly income
4. Total working years VS Age
5. Performance salary hike VS Performance Rating
6. Years at Company VS Years with Current Manager
7. Years at Company VS Years in Current Role
8. Years in Current Role VS Years with Current Manager

Exploratory plot 14:
By filtering the employees who have quit, we then graphed monthly average income by amount traveled by employee and observed that the largest group who quit were those who “Rarely traveled.” Now what does this mean? Inconclusive.

Exploratory plot 15:
We placed working years as the x-axis and monthly income as the y-axis to develop a visual of where certain positions lie in terms of working year experience. As expected the more working years an employee may have the higher the monthly income. Also with

increased working years results in higher positions such as Manager or Director.

Exploratory plot 16:
Using the entire dataset we set Performance rating as the independent variable and Performance salary hike as the dependent variable. What we observed was that Performance rating only resulted in a rating of 3 or 4, which made an interesting find as a trend of the company but at the same time did not provide much to conclude on.

Final Analysis:

We were able to accomplish our two main objectives. Develop an accurate model for predicting employee attrition and identify possible areas of discrimination based on various characteristics. By utilizing various types of graph we were able to emphasize the trends by possible gender discrimination, and employee attrition. Though as we answered our objective questions, we began developing questions that could not be answered by the dataset such as how happy are IBM employees, was this data comparable to other technology companies, and also what incentive systems are in place to ensure accurate responses from employees and incentive to perform well.

The most prominent income gap between genders was presented in higher-level job roles such as Manager and Research director. One of these gaps represented a $18,168 difference in yearly salary, which should be alarming to the Human resource department. Other characteristics of attrition we found were that based on job roles Healthcare representatives, manufacturing directors, and laboratory technicians, which represented the top three jobs employees leave. What we would recommend is reviewing the hiring process and criteria for these positions, understanding the responsibilities they hold, and whether their working environment emphasizes a long-term relationship with the company.