Assignment 2: EDA Using Watson Analytics

Daanish Ahmed
DATA 610 9080
Fall 2016
dsahmed2334@yahoo.com
Professor Vahe Heboyan

October 30, 2016

Introduction

In recent years, data analytics has become increasingly accessible to the public due to new technologies such as IBM's Watson Analytics. This software allows users to import and shape their data at any time while also providing them access to powerful visualization tools that can generate models based off of existing data trends. Using these tools, a user can conduct an exploratory data analysis, or EDA, which can lead to a deeper understanding of the trends and relationships occurring within a dataset (Kane, 2015). My goal in this assignment is to use Watson Analytics to perform an EDA on an imported dataset focusing on employee attrition. The results of my analysis should provide some insight into the different factors which influence employee attrition, as well as some other trends which may occur in the data.

The dataset that I am analyzing was retrieved from IBM's website of sample datasets, and its main purpose is to determine the leading factors influencing employee attrition (Stacker, 2015). It is quite robust, as it features 1470 cases and multiple inputs which could shape the outcome of the analysis (see Table 1 in Appendix A). Some of the inputs include educational factors such as field of study and education level. Additionally, there are important career-related factors such as the employee's job level, years in the workforce, and monthly income. Finally, there are also some useful personal inputs such as gender, age, marital status, and distance traveled from home to work. In my analysis, I found that most of these inputs generally influenced the outcomes of certain target variables. Some variables for which I developed predictive models include employee attrition, job role, and the total number of companies an employee worked for. However, one interesting observation is that not all of the target variables were strongly influenced by the inputs. In fact, a few of the questions answered by Watson Analytics yielded little connection between the data at all. But before even importing the dataset into Watson Analytics, I needed to thoroughly examine and cleanse the data in Microsoft Excel to ensure that it was free of missing values, errors, and outliers. To deal with outliers in particular, I chose the method of truncation as outlined by Osborne and Overbay (2004) in which the outliers would be truncated to a value three standard deviations away from the mean. According to the authors, this method preserves the order and size of the data while limiting the effect that the outliers would have caused.

Exploration of Initial Questions

With the data cleansing process complete, the next step is to explore the initial starting points provided by Watson Analytics. Here, I found that several weak connections between the variables were yielded in the analysis. The first suggestion on the list, a spiral graph showing the factors driving environmental satisfaction, found no key drivers at all. The only valuable insight it provided was that environmental satisfaction is independent from the other inputs in the dataset. Another starting point involved comparing environmental satisfaction and attrition to the percentage of an employee's salary hike. The bar graphs generated in Watson Analytics (see Figure 1 in Appendix A) suggested that neither variable had a significant influence over salary hike. Experimenting with different filters produced variations in the results, but no visible trends were generated even after doing so. Though neither graph is useless, they have limited analytical value. This demonstrates that although Watson Analytics is incredibly powerful, the technology is still developing and has not reached a point where it can replace human judgment.

Yet regardless of these shortcomings, Watson Analytics was still able to generate some insightful starting points. One of these discoveries is a line graph relating monthly income, age, and attrition (see Figure 1 in Appendix A). The graph clearly indicates that workers' earnings increase significantly with age. However, it also suggests that attrition has a greater negative impact on the salaries of older employees. This is an interesting finding that can strongly benefit from additional research. Another insightful diagram is a treemap showing the breakdown of the average number of companies that employees work at based on their education level and attrition (see Figure 1 in Appendix A). This model shows that attrition severely hurts the number of companies in which an employee can work at. Also, workers who are generally well-educated and do not suffer from attrition tend to work for the highest number of companies. But most interestingly, those who have attained the very highest levels of education generally work for significantly fewer companies. Altogether, these starting points may not always produce the most specific results, but they still provide valuable research questions to explore in the future.

Further Exploration

Overall, I found that some of the most useful insights came from the more specific questions that I entered into Watson Analytics. One of these questions was to determine the influence that an employee's field of education would have on their job role within the company. The resulting visualizations (see Figure 2 in Appendix A) feature three bar graphs, two of which show the popularities of different job roles and educational fields among employees. The third graph shows a clear breakdown of each job role and the fields of study for each worker within those roles. At first glance, most of these results seem logical. For instance, the majority of healthcare representatives, laboratory technicians, and research scientists studied either life sciences or medicine. However, a clear percentage of employees have job roles that differ from their fields of study. For example, there is a large population of sales executives that studied life sciences, and there is also a relative lack of manufacturing directors with technical degrees. One way to make these results more relevant is to utilize filters based on education level. For this model, I utilized a global filter to see if changing educational level will alter the results of the graph. By setting the filter to include only those of the highest educational level, I found that each of the roles became more specialized, and there were far fewer individuals whose field of study did not match their job role. This process can greatly benefit my organization because it can allow the company to assign better candidates towards important job roles by looking at their fields of study and education levels.

The last set of questions in my analysis revolves around this dataset's main purpose—determining the factors that influence attrition. By using a spiral graph, I found that some of the most relevant inputs include job level, years in the workforce, monthly income, age, gender, marital status, and distance between home and work. The graphs produced by Watson Analytics (see Figures 3 and 4 in Appendix A) all show that each of these inputs has a direct impact on attrition. After experimenting with different visualization types, I found that stacked bar graphs, area curves, and treemaps were the most effective formats for visually conveying the existing relationships in this dataset. The results indicate that younger, lower-paid employees tend to have higher attrition rates compared to older, experienced and higher-earning workers. These results are consistent with findings by Roberts (2015), who claims that employees are most likely to leave their jobs during their first few years of working. In addition, lower-ranking employees

who are made to work overtime suffer particularly high attrition rates. The results also suggest that attrition rates are high for single workers and those who live farther away from work, and that attrition is generally higher among males. These findings are extremely useful and can help my organization to pinpoint the causes of attrition so that it can work towards solving them. To add more value to this analysis, I included job satisfaction as a local filter to see its impact on attrition and monthly income. The results indicate that job satisfaction is crucial to the analysis as well, since employees with both the lowest incomes and job satisfaction rates have attrition rates near 100%. And finally, I introduced a calculation of annual salary by multiplying each employee's monthly income by 12. This calculation can be useful for analyzing long-term data, such as comparing attrition based off of an employee's salary and total years at the company.

Data Refinement

The final process that I will discuss in this paper is the refining of the dataset to improve its quality score. After importing the dataset into Watson Analytics, I found that its data quality score was 70%. I began the refinement process by filtering out the variables with quality scores of 0%. These variables, which include employee count and standard hours, contribute little to the dataset since all of their entries have the same values. Hiding these columns thus increased the quality score to 76%. From here, I also filtered out columns that were of significantly lower quality than the dataset as a whole. However, some of these columns—such as attrition and education field—are utilized by the discoveries in this project. Hiding such columns will still increase the data quality, but it will also remove those variables from all of the discoveries in which they are present. Therefore, I chose to hide only unused columns such as department, business travel, and performance rating. Doing so increased the data quality by one percent. Furthermore, I can include some of the data groups I created for the analysis, allowing me to hide the columns from which these data groups were derived. For example, I could filter out monthly income and total working years because these columns were only used via their data groups. This process once again raised the data quality by one percent, bringing it to 78%. Finally, I attempted to create a hierarchy that consisted of monthly salary rate, daily rate, and hourly rate (and therefore hide those three columns). However, doing this reduced the dataset's quality by two percent. Since these variables had by far the highest data quality (98-100%), utilizing a hierarchy in this case may not necessarily benefit the data refinement process.

Conclusion

The results of my EDA show that Watson Analytics is an indispensable tool for data analysis. By quickly providing accurate solutions that are easy to understand, this software can help many organizations towards solving issues such as employee attrition. Of course, Watson Analytics is not perfect and there is a long way to go until this type of technology can fully replace the human decision-making process. In organizations like my own, there may even be resistance towards the adoption of this technology due to its issues as well as its complexity. But based on the insights from my analysis, I believe that the benefits of using Watson Analytics far outweigh its shortcomings. If a program can easily answer questions about employees that may otherwise be difficult for a human to answer, then that program can transform any company.

References

Kane, D. (Performer) (2015). *EDA and Model Selection* [Web]. Retrieved from https://www.youtube.com/watch?v=St-xm35a_nk

Osborne, J., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation,* 9(6). Retrieved October 24, 2016 from http://pareonline.net/getvn.asp?v=9&n=6

Roberts, P. (2015, January). The CFO and CHRO Guide to Employee Attrition. *Workforce Solutions Review,* 6(1), 8-10. Retrieved October 28, 2016, from UMUC Library.

Stacker, M., IV. (2015, April 2). Guide to Sample Data Sets. Retrieved October 25, 2016, from https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets//

Appendix A

EducationField	EnvironmentS...	JobSatisfaction	Attrition	EducationLevel	JobRole	MaritalStatus	OverTime	Gender	Age	TotalWorkingY...	YearsAtComp...	YearsInCurrent...	TimeWithCurr...	NumCompani...	JobLe
Medical	1	3	No	4	Research Dir...	Married	No	Female	47	26	20	15	5 to < 7	3	
Medical	4	3	No	1	Sales Execut...	Married	Yes	Female	24	5	5	4	2 to < 5	1	
Marketing	4	4	No	5	Sales Execut...	Married	No	Male	32	7	4	3	2 to < 5	4	
Life Sciences	4	2	No	3	Laboratory T...	Married	No	Female	34	7	5	4	2 to < 5	5	
Medical	2	3	No	5	Research Sci...	Married	Yes	Male	41	7	4	2	2 to < 5	2	
Other	3	3	No	4	Laboratory T...	Divorced	No	Male	40	11	8	7	7 to < 9	3	
Life Sciences	1	4	No	2	Sales Execut...	Divorced	No	Male	31	13	13	7	9 to < 12	1	
Marketing	1	4	Yes	3	Sales Execut...	Divorced	No	Male	46	28	7	7	2 to < 5	4	
Life Sciences	1	1	Yes	3	Laboratory T...	Single	Yes	Female	39	11	1	0	less than 2	2	
Life Sciences	3	2	Yes	5	Manufacturin...	Single	No	Female	31	10	10	8	7 to < 9	1	
Medical	2	1	No	3	Healthcare R...	Divorced	Yes	Male	45	24	7	7	7 to < 9	2	
Medical	4	2	No	2	Human Reso...	Single	No	Female	31	8	3	2	2 to < 5	9	
Life Sciences	2	4	Yes	3	Laboratory T...	Married	Yes	Male	31	7	2	2	2 to < 5	6	
Technical De...	2	4	No	3	Manufacturin...	Married	Yes	Male	45	10	3	1	2 to < 5	4	
Marketing	3	1	No	3	Sales Execut...	Divorced	No	Male	48	15	2	2	2 to < 5	4	
Technical De...	1	3	Yes	4	Human Reso...	Married	No	Female	34	2	2	2	2 to < 5	1	
Medical	4	2	No	1	Research Dir...	Divorced	No	Male	40	16	9	8	7 to < 9	2	
Medical	4	1	No	3	Sales Execut...	Single	No	Male	28	6	5	4	2 to < 5	0	
Life Sciences	3	3	No	3	Laboratory T...	Single	No	Male	44	7	5	2	2 to < 5	3	
Medical	4	2	No	3	Research Dir...	Single	No	Male	53	34	9	8	7 to < 9	4	
Technical De...	1	3	No	4	Healthcare R...	Married	No	Male	49	20	3	2	2 to < 5	2	
Medical	3	3	No	3	Research Sci...	Divorced	No	Male	40	8	3	1	2 to < 5	2	
Life Sciences	4	2	No	3	Research Sci...	Single	No	Male	44	6	5	2	2 to < 5	1	
Medical	4	1	No	3	Sales Execut...	Married	No	Male	33	5	4	3	2 to < 5	0	
Other	4	1	No	3	Sales Execut...	Single	No	Male	34	15	13	9	12 and above	3	
Life Sciences	4	3	No	1	Sales Execut...	Married	No	Male	30	4	2	1	2 to < 5	8	
Medical	1	4	No	2	Laboratory T...	Single	No	Female	42	12	12	9	7 to < 9	1	
Marketing	1	3	No	5	Sales Execut...	Married	No	Female	44	11	1	0	less than 2	7	
Technical De...	3	3	No	3	Research Sci...	Divorced	No	Male	30	1	1	0	less than 2	1	
Life Sciences	2	3	No	2	Research Sci...	Married	No	Male	57	13	12	9	7 to < 9	0	
Life Sciences	3	2	No	4	Healthcare R...	Divorced	No	Male	49	29	8	7	7 to < 9	3	
Medical	2	1	No	3	Research Dir...	Divorced	No	Male	34	16	14	8	9 to < 12	7	
Technical De...	1	3	Yes	3	Sales Repres...	Married	No	Male	28	5	3	2	2 to < 5	3	

Table 1. Employee Attrition Dataset.

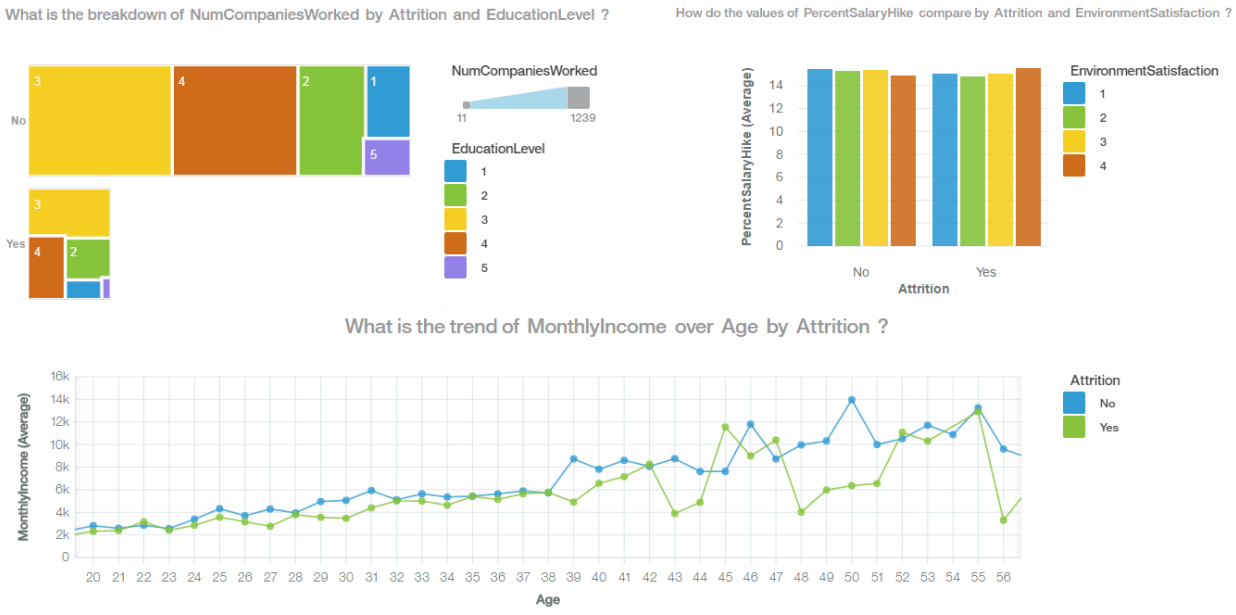

Figure 1. Starting Points by Watson Analytics.

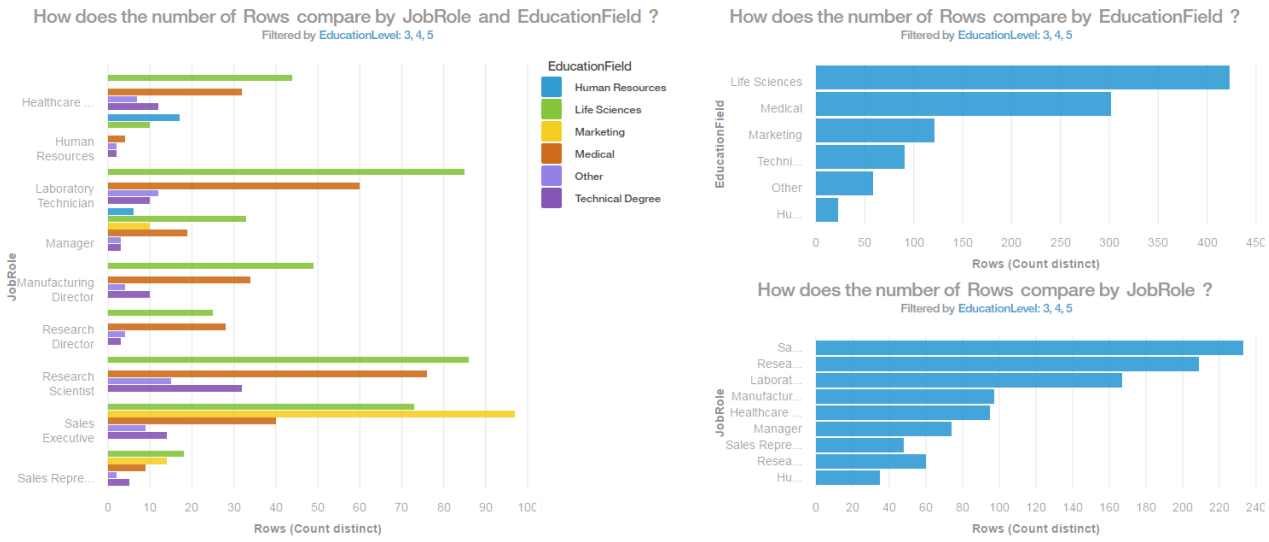

Figure 2. Comparison of Employee Education Field and Job Role.

Figure 3. Career Factors Influencing Attrition.

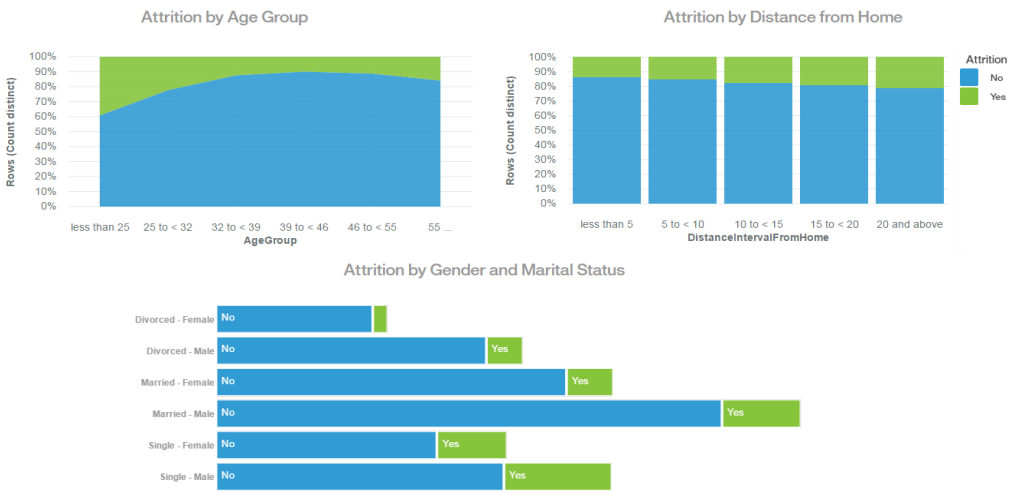

Figure 4. Other Factors Influencing Attrition.