IBM Attrition

Kevin Carmody x15468652
School of Computing
National College of Ireland
Dublin, Ireland

https://youtu.be/B2CrCEGaVQ8

Youtube Link ^

Sirui Cheng x16107535
School of Computing
National College of Ireland
Dublin, Ireland

Abstract

Our reason behind our inspection was to explore and explain what reasons are behind employee attrition as attrition causes real problems within an organization no matter what size it may or may not be as they may need obtain new workers. This means they may need to provide training and education to new members of the group which could deplete valuable time and cash. To attempt and try and limit the danger of employee attrition and distinguish reasons behind attrition, historical data from IBM. Data Mining Techniques such as Naïve Bayes, Association Rule, Linear Regression and Decision Trees will play an important part in drawing out knowledge on employee attrition and explaining why they leave.

Keywords - Data Mining Techniques, Employee Attrition, IBM, Naïve Bayes, Association rule, Decision Trees

I. Introduction

Attrition has many definitions, but attrition can be known as “unpredictable and uncontrollable” and the reason behind reduction of work force due to “resignations, retirement, sickness or death”. Also known as turnover, which could have many structures, attrition can be manually intentional or can be down to automatic decisions as previous stated. These reasons can cause interference to firms and leave negative effects among current employees within the firm. Employee Attrition can cause firms / organizations a wide range of unnecessary costs such as Online Promotion for new positions, administration time expenses. Sources have demonstrated that 17%-35% of Employee Attrition are immediate expenses and between 73% to 88% Employee Attrition are put down to circumstantial differences. (Boles, Dudley et al., 2012).

There are many strategies within Data Mining Techniques such as Naïve Bayes, Linear Regression, Decision Trees. These techniques are utilized very regularly as they perform very well. Rules can be created and utilized for future outcomes. But when it comes to machine learning, certain courses of action can mistreat future attrition rates due to so much noise within data. Our paper is organized within various sections as follows:

Section 2 contains information about what we have researched in connection to employee attrition and what can be done about the employee attrition issue.

Section 3 contains information about how we have gathered our information and how we have prepared it for analysis and modelling. We also go through various aspects of the data and what most variables mean and explain them and how useful these variables will be to us when analyzing.

Section 4 contains information on how we approached our methodology within our project and how we utilized it.

Section 5 contains information on how the executed our chosen methodology within the project and how it was utilized.

Section 6 contains information about how we evaluated our data mining techniques on Employee Attrition and what we got as our results.

Finally, Section 7 contains information on our conclusion and future work if we had extra time.

II. RELATED WORK

A paper called “"Performance Analysis of Various Data Mining Algorithms" by K. Tamizharasi, K. Rajasekaram & Uma Rani work through various comparison of Precision and Accuracy values in between seventeen algorithms such as Decision Trees and Naïve Bayes.

It’s shown that there is a 51% accuracy rating with Decision Trees and Naïve Bayes with 54% accuracy, though Decision Tree had a Kappa of 0.21 and Naïve Bayes with 0.48, which shows you that there better argument for Naïve Bayes then Decision Trees. Overall Logistical Regression and PART were had the lowest accuracy rating with 47%.

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

Fig. 1 (Cohen’s Kappa)

In Dursen Delen’s journal on “Predicting Student Attrition with Data Mining Methods”, it mentions that using 8 years of institutional data, it’s shown that “*artificial neural networks performed the best, with an 81% overall prediction accuracy on the holdout sample*”.

While Amir Mohammad Esmaieeli Sikarodi Journal describes various reasons behind attrition using Data Mining methods and using a CRISP Methodology. It’s shown that Naïve Bayes can be used and is viewed to be the best technique to use with Attrition. Decision Trees are shown to perform well with PNN and MLP Techniques.

Various authors then mention a SMOTE Method can be used to fix an imbalance within data that can then become more satisfactory. Authors then go on to say that Naïve Bayes gave them a better prediction models than KNN and MLP. Naïve Bayes had a 88% rating while KNN and MLP were not too far behind with 75% and 82% respectfully. These figures give us a different perspective than the figure we found with K. Tamizharasi, K. Rajasekaram & Uma Rani (2014)

III. DATA EXPLORING

We download the data set from Kaggle named “Company”. There’re 1470 rows with 35 variables contains employees’ details, working condition, job details, attrition condition and so on. Each column with 1 being low and 5 being very high. In this data set, our goal is to understand “Attrition” to explore and solve this problem. This dataset was created by IBM to help find hidden understandings to what leads to Attrition within the company. Employees who have left IBM with have “yes” down under the “Attrition” Variable. Other variables such as Education will have how far studies have gone such as: 1= Below college, 2= College, 3=Bachelor, 4= Master, 5= Doctor.

Here, we will use the data set for simple analysis, the relationship between several major categories of factors and attrition. For example, gender, age, working hours, working years, salary levels, etc. From the data we can see that there are more men in the company than women, so the number of attritions is more, but not much, so there is no strong correlation with gender.

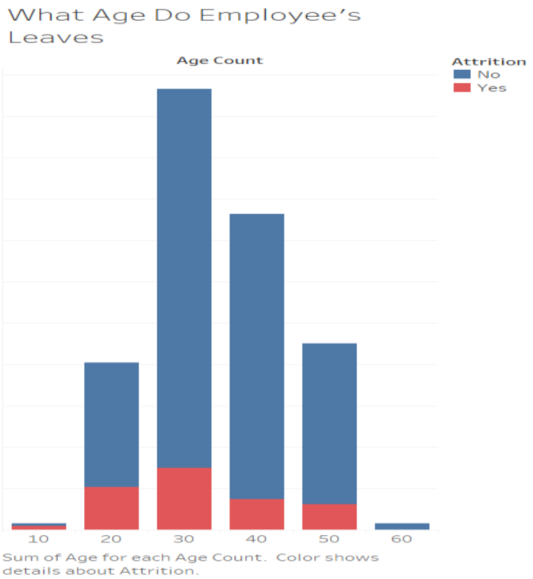


Fig.2 Age where Employee’s Leave

Then looking at the age, most people who leave the company are around 30 years old. People who are probably around 30 years old have more ideas and will choose their own lives. There isn’t much to going on in this plot as it’s impossible to say why they leave around there 30th birthday. This plot confirms a class that is imbalanced.

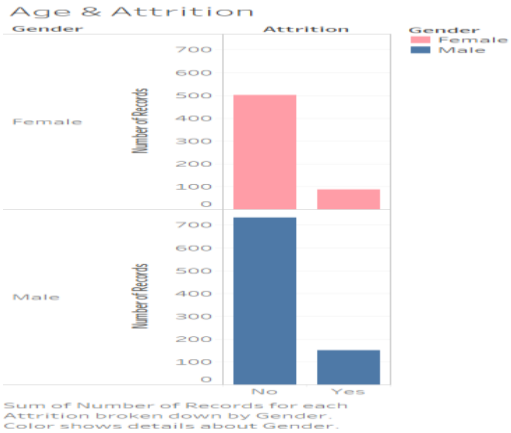


Fig.3 Gender & Attrition

Regarding Gender, More Males leave company's then their female counterparts. Here, in the figure above, we see that we have combined both attrition variables and gender variables. Shown in our plot, we see that more males leave the organization but it's not as outright as it seems with Males leaving being just a bit more than Females leaving. The concerning part is that the lack of employees leaving meaning that we now unfortunately once again show imbalanced classes. We can see that its variable called Attrition that is having classification problems where classes are not represented equally and so we must find a solution and create a strategy that could tackle classes that seems to be imbalanced.

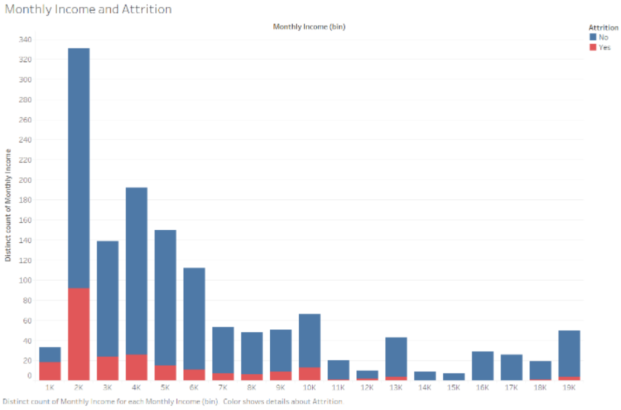
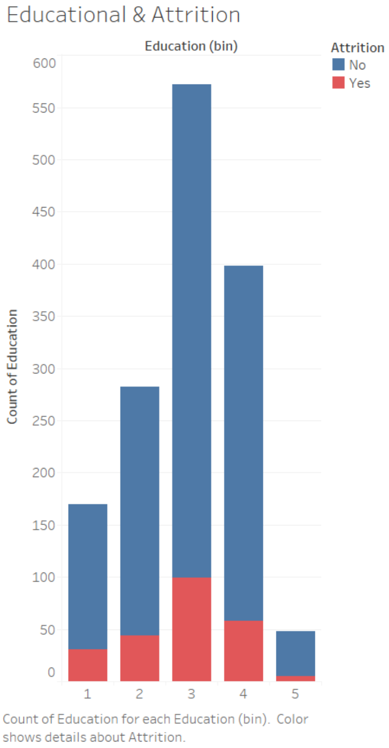


Fig.4 Monthly income and Attrition

For the salary issue, there is no doubt that the higher the salary, the lower the attrition rate. In the figure above, you can see that lower the wage (left side), Attrition (Red) is very high compared to the to the right side of the figure which is low or barely any attrition at all.

Fig.5 Educational Studies & Attrition



Regarding Educational Studies, our plot shows that people who have a degree or a tend to leave IBM which could be the reasoning behind many people leaving. In the plot above see that over 150 employees have left the company that have had a degree or better which could show that they have moved on in the industry.

The prediction of the relationship between other factors and attrition will be analyzed and explored in more detail below.

IV. Methodology

We selected three widely used prediction models to obtain predictions, including Naïve Bayes, association rules, and decision trees. These models will be trained and tested using the exact same data set and the accuracy will be insistent on.

Naïve Bayes

Naive Bayes is a straightforward, successful and well known, machine learning classifier. It is a probability classifier that makes classifications using a decision rules in a Bayesian setting known as Maximum A Posteriori. This technique is mainly used to classify the possibilities of each situation. Using this approach is useful for predicting what we want to do, because it could easily aid us to forecast employee attrition within that organization and can be used against other technologies to see differences. In the learning process, we found that the algorithm has the expertise to construct high precision, and the logic behind this is P(B|A) = P(A|B)P(B)/P(A). (A stand for feature and B stands for category).

Decision Trees

Decision trees are one of the most popular machine learning algorithms but also the most powerful. One of the bases is that they are so powerful is because they can be easily envisaged so that anyone can understand what's going on.

A decision tree is a map of the possible consequences of a string of associated choices. It shows organizations to compare possible factors against one another based on their value, expectations, and interests. They could be used to propel open discussions or to show an algorithm that forecasts the premier option mathematically.

Decision trees are a very ordinary classification technique. It is a kind of managerial learning. The so-called managerial learning is to give a bundle of trials. Each trial has a set of attributes and a factor. These factors are fixed in advance. Then grasps a classifier that can be used for new looks. The object gives the correct classification. Such machine learning is called supervised learning.

Association rule

Association rules are if-then statements that help to show the probability of relationships between data factors within a large data sets in different types of databases. Among them, the association rule XY has the support degree and the trust degree. Association rules are often used to find relationships between each attribute. It is primarily used to determine hidden rules in transaction data by applying purchases based on strong rule concepts and measurement of product frequency. By applying association rules to our data sets, we can clearly determine the extent of employee turnover and what factors are related to employee attrition.

V. Implementation

First steps within our project was to install certain packages for each data mining technique such as caret & e1071. Our CSV file can be read within RStudio and then when running, we can see what each column is named so then we can remove columns that we will not require within the project. Afterwards, we can then split the data into 2… called Train and Test.

Train will consist of 80% of the data which is 1176 rows and Test will consist of 20% of the data which will contain 294 rows of data. Code is then used to determine "yes" or "no" for Attrition between both splits and the original to show accuracy.

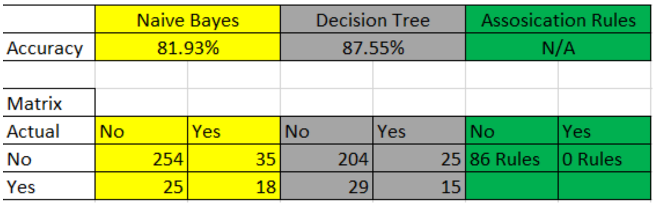
Code for Naïve Bayes is then used to show probability and then likelihood for all attributes available and in relation to Employee Attrition such as Salary and Age. We installed and loaded 3 packages, "e1071", "caret" & "gmodels". Columns were then removed that were unnecessary. We would also attempt to clean the data by looking for any na. Data was then split between 2 subsets, one would contain 80% of employee information and the 2nd subset would contain 20% of employee data. We then compared "yes" and "no" between the 2 subsets. We then conducted a naïve bayes function in R and used a predict () function and then used them within a confusionMatrix and CrossTable

Decision Trees required a few installed packages, most importantly, C50. Package that is to be used for plotting data are called Rpart, party and plot. We randomize first by setting a seed that would allow reproducibility and then create an 80/20 subset. We then gather their distribution within a prop table. We test them against each other and along with the original 1470 rows. We would train the model and then test the model and then evaluate with crosstables.

Association Rules was then used which would include a meaningful function known as inspect () which allows us to review variables. To try and avoid rules that are known to be redundant that could show up within our results, a remove redundant function was to be performed, then onwards, we would carry out how many various rules do we have with the "yes" or "no" for attrition within our IBM database. We then used visual plots for plotting out our redundant rules.

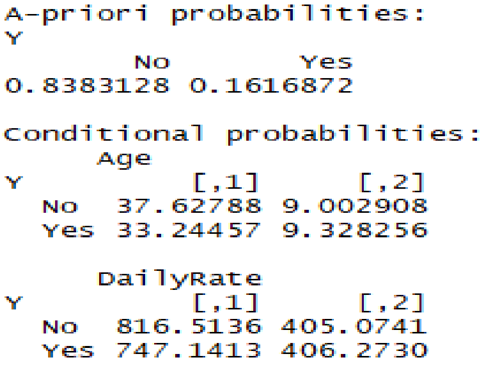
VI. Evaluation

The all-inclusive result that shows what we obtained within 3 different data mining techniques is shown below within our Table Result



	Naive Bayes		Decision Tree		Assosication Rules	
Accuracy	81.93%		87.55%		N/A	
Matrix						
Actual	No	Yes	No	Yes	No	Yes
No	254	35	204	25	86 Rules	0 Rules
Yes	25	18	29	15		

Fig 6. Our Results from our Used Techniques



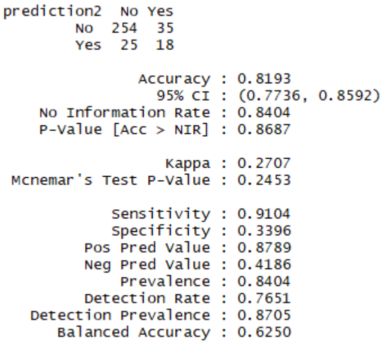
```
A-priori probabilities:
Y
       No       Yes 
0.8383128 0.1616872 

Conditional probabilities:
     Age
Y         [,1]     [,2]
  No  37.62788 9.002908
  Yes 33.24457 9.328256

     DailyRate
Y         [,1]     [,2]
  No  816.5136 405.0741
  Yes 747.1413 406.2730
```

Fig. 7 A-Priori

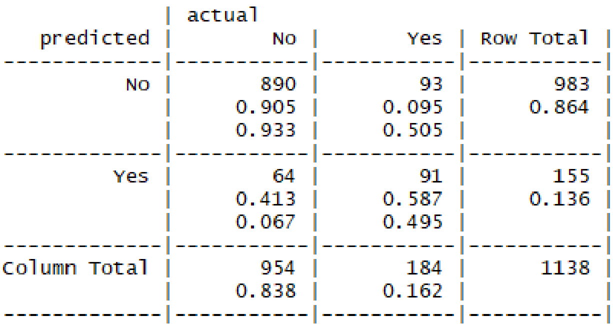
As we can see within our A-priori probability table above, the table shows us that 83.8% of employees are leaving IBM, where as the result to it, 16.1% are leaving the organization. Calculating the probability of age is not possible therefore standard deviation and the mean is calculated which is shown in provisional probabilities. This shows us that the age of 38 years old or even higher have a higher chance of not ever leaving the organization then it's counterpart, 34 years old and below has a better chance of leaving the company. Therefore, in simple English, younger employees are leaving the organization more than their older counterparts, as older employees tend to stay with there organization. Also included in the results is the variability, which in this instance are very close to each, "no" being at 9.00 and "yes" being at 9.32.



```
prediction2  No Yes
        No  254  35
        Yes  25  18
                                          
               Accuracy : 0.8193          
                 95% CI : (0.7736, 0.8592)
    No Information Rate : 0.8404          
    P-Value [Acc > NIR] : 0.8687          
                                          
                  Kappa : 0.2707          
 Mcnemar's Test P-Value : 0.2453          
                                          
            Sensitivity : 0.9104          
            Specificity : 0.3396          
         Pos Pred Value : 0.8789          
         Neg Pred Value : 0.4186          
             Prevalence : 0.8404          
         Detection Rate : 0.7651          
   Detection Prevalence : 0.8705          
      Balanced Accuracy : 0.6250          
```

Fig.8 Naïve Bayes Prediction

Next, we see our Naïve Bayes Model which shows us that our accuracy of this is **81.93%.** Our model was based on a portion of our dataset which was at 332 records. We also have a 95% confidence interval going into this model, so we are confident that the accuracy is going to be between **77.36%** and **85.92%.** We have a 5% chance that the accuracy could be wrong or that possibly it could be 5% up or down. The P-Value is above 0.05 therefore we can accept this. Also, within the results, we can see that our Kappa is 0.27. Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative items. With our Kappa being 0.27, we have a "fair agreement" that the accuracy is correct. We can also see that there is a huge difference between sensitivity and specificity, 91.04% and 33.96% respectfully, therefore shows that the model is skewed.



```
             | actual 
   predicted |        No |       Yes | Row Total | 
-------------|-----------|-----------|-----------|
          No |       890 |        93 |       983 | 
             |     0.905 |     0.095 |     0.864 | 
             |     0.933 |     0.505 |           | 
-------------|-----------|-----------|-----------|
         Yes |        64 |        91 |       155 | 
             |     0.413 |     0.587 |     0.136 | 
             |     0.067 |     0.495 |           | 
-------------|-----------|-----------|-----------|
Column Total |       954 |       184 |      1138 | 
             |     0.838 |     0.162 |           | 
-------------|-----------|-----------|-----------|
```

Fig. 9 Confusion Matrix

Next, we executed codes for Decision Trees. The following chart was the outcome of that code.

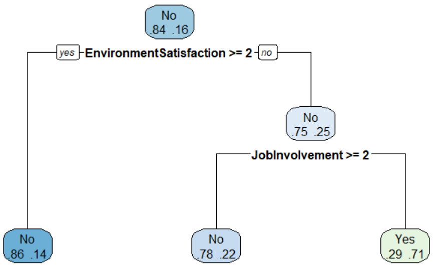
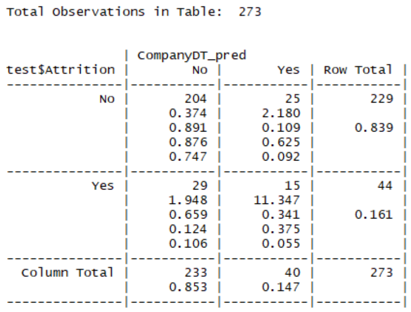


Fig. 10 Decision Tree

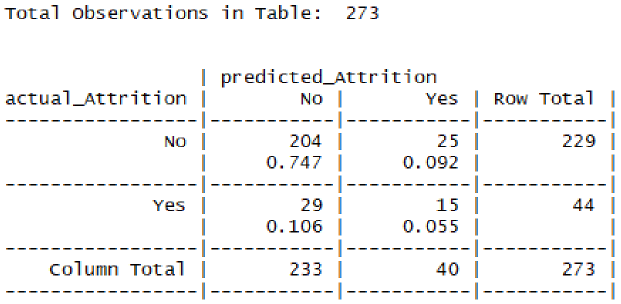


```
Total Observations in Table:  273

             | CompanyDT_pred
test$Attrition |        No |       Yes | Row Total |
---------------|-----------|-----------|-----------|
            No |       204 |        25 |       229 |
               |     0.374 |     2.180 |           |
               |     0.891 |     0.109 |     0.839 |
               |     0.876 |     0.625 |           |
               |     0.747 |     0.092 |           |
---------------|-----------|-----------|-----------|
           Yes |        29 |        15 |        44 |
               |     1.948 |    11.347 |           |
               |     0.659 |     0.341 |     0.161 |
               |     0.124 |     0.375 |           |
               |     0.106 |     0.055 |           |
---------------|-----------|-----------|-----------|
  Column Total |       233 |        40 |       273 |
               |     0.853 |     0.147 |           |
---------------|-----------|-----------|-----------|
```

Fig. 11 Confusion Matrix

From observing the table above, we can see that 29 from 233 NO's were incorrectly added and classified as "Yes" where as 25 from 40 YES where incorrectly added and classified as "No". This shows us that 204 employees are predicted to have NOT left the company and are classified as true negative employees and 15 employees fall into a "verifiable positive" category meaning the model has correctly predicted 15 members leaving the organization. 29 members have falling into "inaccurate negative" category which intends that we forecasted them not leaving but they left anyway. 25 members fall into a type I error category who predicted to depart from the country but didn't leave anyway.



```
Total Observations in Table:  273


                 | predicted_Attrition
actual_Attrition |        No |       Yes | Row Total |
-----------------|-----------|-----------|-----------|
              No |       204 |        25 |       229 |
                 |     0.747 |     0.092 |           |
-----------------|-----------|-----------|-----------|
             Yes |        29 |        15 |        44 |
                 |     0.106 |     0.055 |           |
-----------------|-----------|-----------|-----------|
    Column Total |       233 |        40 |       273 |
-----------------|-----------|-----------|-----------|
```

Fig. 12 Model Boosted

From the table above, after executed a boosted model to try and improve results, we have received a result of **87.55%.**

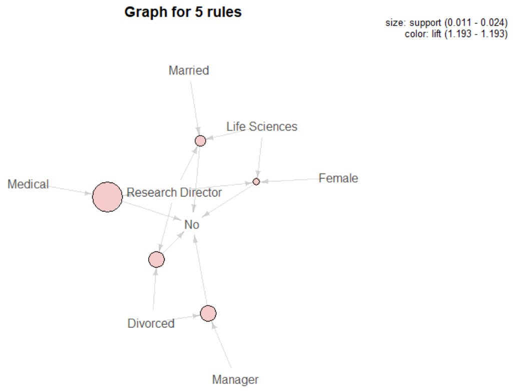


Fig. 13 Graph for 5 rules.

The arrow length within the graph above means support and intensity of color means confidence. The Graph above represents qualities between each other.

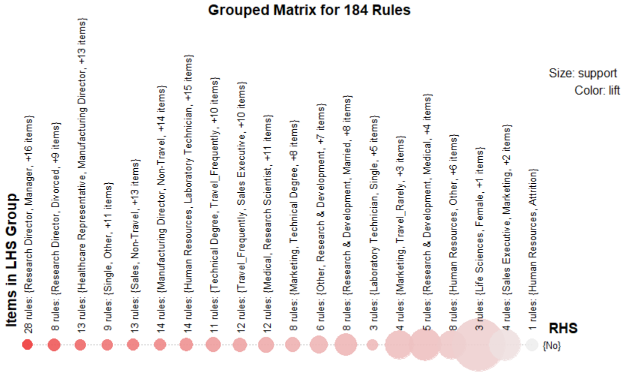


Fig. 14 Grouped Matrix

Our Diagram above explains our grouped matrix result by using lookalike balloon plots. The color of the plot explains the interest in the group and ratio of the balloon explains the aggregated support of the group.

VII. Results & Post Evaluation

This report explains aspects of various data mining methods that has the competence to assist huge organizations such as IBM explain and forecast attrition within the company. Our results from our research illustrates that our Decision Tree has the best algorithms in terms of accuracy then followed by Naïve Bayes. We also find areas within the file that contributes to Attrition such as Age, Daily Rate and Education.

There are lesser factors such as Gender that don't really have any influence on attrition within our project/data set. Our Study explains that younger employees tend to leave the company, around the age of 30, while employees over 40 tend to stay on with the organization.

The techniques used within our report can be of great aid to organizations such as IBM and want to gain a full insight into why attrition happens. This would be beneficial for companies and therefore help the company save money and time.

It is quite possible we could say right now that we have found some type of correlation between employees leaving there work place and not leaving there work place but we can not be certain as the data doesn’t seem to be legit as it’s shown to be fake data from Kaggle. Although the data may not be real, we can assume if this data was real, we would then be able to get similar outcomes.

If I was to do this in the future for a company, I would have likened to do a kNN or Random Forest to maybe have an expanded knowledge of this IBM situation and to find out what could produce a better result than the code and algorithms I ran. I would have also liked to use a SMOTE function that I could have tackled imbalanced classes that I had talked about in the report if I had more time for this project. Maybe having not one but multiple other companies and with real world data and then compare attrition to each other, maybe we could find the real reasons for attrition.

References

[1] Boles, J. S., Dudley, G. W., Onyemah, V., Rouziès, D., & Weeks, W. A. (2012). Sales force turnover and retention: A research agenda. Journal of Personal Selling & Sales Management, 32(1), 131-140

[2] K. Tamizharasi, D. Umarani and K. Rajasekaran (2014), "Performance Analysis of Various Data Mining Algorithms", Semanticscholar.org, 2019.[Online]Available: https://www.semanticscholar.org/paper/Performance-Analysis-of-Various-Data-Mining-Tamizharasi-Umarani/9334df883e232b6f223d44e8246c35bcb0ea714b. [Accessed: 04- Apr- 2019].

[3] Delen, D. (2011) ‘Predicting Student Attrition with Data Mining Methods’, Journal of College Student Retention: Research, Theory & Practice, 13(1), pp. 17–35.

[4] Amir Mohammad Esmaieeli Sikarodi (2015). [Online]. [4 April 2019]. Available from: http://www.jise.ir/article_10857_380ab2c2c84e1525e1f53647b46d6879.pdf

[5] M. Shinde, "COMPARATIVE STUDY OF DECISION TREE ALGORITHM AND NAIVE BAYES CLASSIFIER FOR SWINE FLU PREDICTION", Pdfs.semanticscholar.org, 2019. [Online]. Available: https://pdfs.semanticscholar.org/a7c5/5e09cf7e130ab8a88276c2f18fd3ef3c13a8.pdf. [Accessed: 04- Apr- 2019].

[6] "Introduction to Naive Bayes Classification", Towards Data Science, 2019.[Online].Available: https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54. [Accessed: 04- Apr- 2019].

[7] "What is association rules (in data mining)?", SearchBusinessAnalytics, 2019. [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining. [Accessed: 04- Apr- 2019].