Human Resource Analytics

CSCI B 565 - Data Mining Project
Nilima Sahoo, Vinita Chakilam

April 27, 2017

Abstract

The main goal of this project is to analyze the past and present employee data of an organization to see why an employee has left the company and also to predict if the employee is going to leave the company in future. This kind of a classification would help companies predict who of their valuable employees might leave their organization in future.

Key words: Decision Tree, Random Forest Model, Support Vector Machine, Classification, Analytics, Visualization, Data mining

1 Introduction

This project focuses on classifying the past and present employee data based on a few important factors such as - Satisfaction Level, Last Evaluation, Number of projects worked on, Average monthly hours worked, Time spent in the company, Whether they have had a work accident, Salary and also if the employee has had a promotion in the last five years, to see if the employee is going to leave the organization or stay with them. This project also analyses the reasons behind employees who have already left the organization using multiple visualization techniques.

2 Data Set Description

The source of this data set is Kaggle.[2] The data set has 10 variables and 14,999 employee records. The different variables are: Satisfaction level, Last Evaluation, Number of Projects, Average monthly hours, Time spent at the company, Whether they have had a work accident, Whether they have had a promotion in the last 5 years, Department, Salary and Whether the employee has left.[4]

2.1 Correlation among variables

From the given data set, we ran the correlation plot in R to see if there is any correlation among the variables resulting in near multicollinearity. Below is the correlation plot:

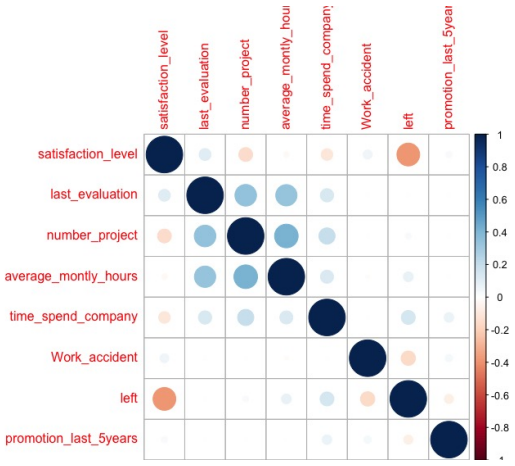

Figure 1: Correlation among variables

We can see that there is no strong correlation between any of the factors considered. Therefore, multicollinearity among variables is not an issue.

3 Analyzing Data

This data set has been analyzed from different perspectives to see which factor or factors have caused most number of resignations. We have also worked

to see the department from which most people have left, the distribution of employee satisfaction levels, salaries and various other factors to see if any of them are major contributors to the employee's quitting the company. Below are the different visualizations and insights that we have come up with to reach major conclusions.

3.1 Visualizations and Insights

We applied a few exploratory data analysis techniques to see how various factors have affected the attrition among the company. We visualized each feature against the employee leaving data to see which feature have a great impact and which features actually do not. We made use of the ggplot2 package in R to have visually appealing graphs.

To begin with, in the following graph, we tried to see the percentage of employees that have left the organization based on their satisfaction levels, and the department in which they are working to see if department has any affect on the employee quitting. [6]

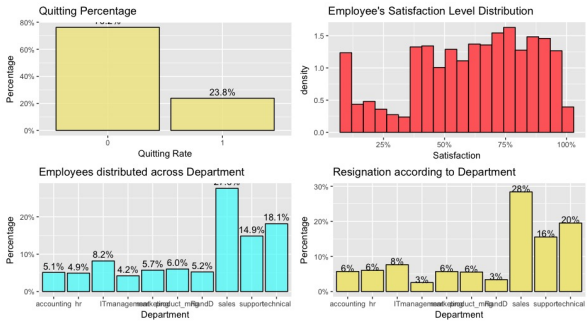

Figure 2: Effect of attributes on Employee Resignation

From the above graphs, we could see that, approximately 76 percent of the employees are with the company at present and 24 percent of the employees have left. Furthermore, on trying to see the department from which there are high number of employees quitting, we could see that sales department has the highest number of employees and so also is the resignation from the Sales department high. Distribution of Satisfaction level across employee leaving shows that, there are more number of employees with higher satisfaction than the number of employees with less levels of satisfaction.

Next, we tried to observe the resignation numbers within a department. The following graph interprets it. We use ggplot2 pacakge in R to use it. [7]

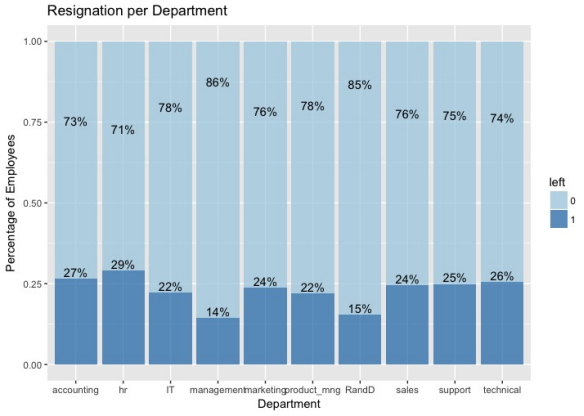

Figure 3: Attrition based on Department

From the above graph, we observed that Sales department is not the main culprit like we predicted in the previous graph, but the HR department is. The HR department has 29 percent of resignations when compared to other departments. Following the HR department is the Accounting department with 27 percent of resignations.

Furthermore, we tried to see the impact of each variable on the Employee's resignation and following is the list of variables and how we predicted it may affect the classification variable.

- Satisfaction Level: Less satisfaction level among employees, makes the employee more likely to quit
- Number of Projects: If the employee is working on higher number of projects, that might lead to more pressure and this might lead to the employee leaving the organization
- Last Evaluation: If the previous evaluation did not go as expected, this might lead to the employee getting frustrated and thus this means that there would be higher quitting rate
- Average Monthly Hours: If the average monthly hours of the employee is high, it might lead to lesser work-life balance which might aggregate the employee's decision to leave the company

- Promotion last 5 Years: If the person has not been promoted in the last five years, this might lead to employee feeling to need a change and ultimately leave the organization

- Work Accident: Higher rate of work accidents implies the employee is more likely to quit

- Time Spend in the Company: If the time spent in the company is high, the employee is more attached to the company and is less likely to quit. It could also mean, that the employee would want a change and is likely to leave the organization

But, the observed situation is different from the predicted one. Below are the graphs showing the real situation of how the count of resignations is based on each of the above mentioned features:

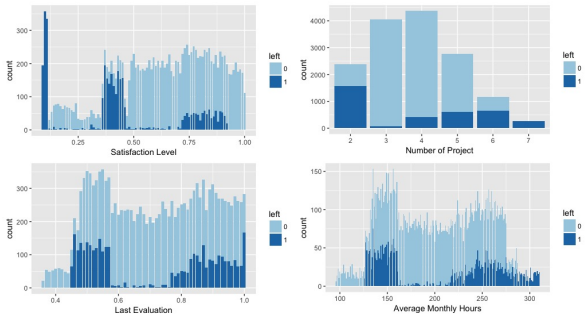

Figure 4: Effect of different variables on Attrition

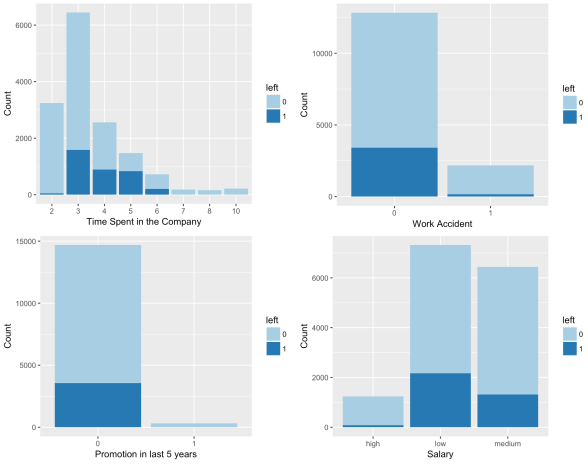

Figure 5: Effect of different factors on Attrition

From the above graphs, we can observe that, the lower the satisfaction level, the higher is the number of employees quitting which is as predicted. But there is also an anomaly that there are a few employee's who do not quit in spite of low satisfaction levels. We can also see that the Number of Projects has the reverse effect of our prediction- low number of projects is leading to employee resignation. The effect of Last evaluation is also different from our prediction. Even if employees have high evaluation ratings, they are still leaving the company. Based on the average monthly working hours, it is seen that employees who are working for 150 hours per month or 250 hours per month are equally willing to leave the job, therefore coming to conclusions from this variable is cumbersome. Based on the time spent in the company graph, we can see that the employees spending around 3 to 6 years in the company are mostly leaving. The longer they stay with the company, the lesser is their will to leave. From the salary graph, it is obvious that the lower the salary, the more the employees willing to leave the company. So, lower time spent in the company with low salary and employees whose promotion has not happened in last 5 years are more likely to quit.

Next, we tried to visualize the effect of salary, number of projects and last evaluation on Employees quitting in one graph.

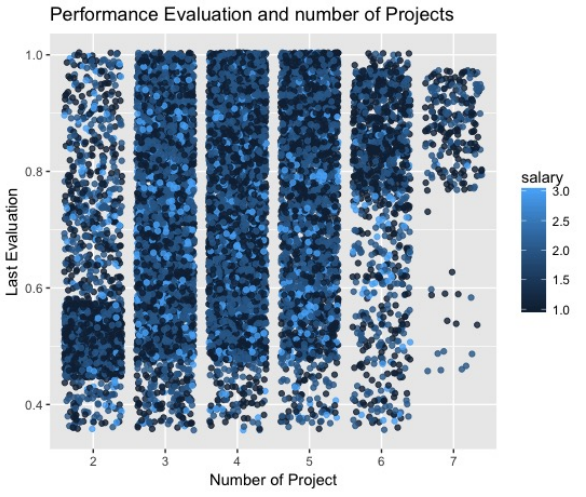

Figure 6: Combined effect of Salary, No of Projects, Last Evaluation on Employees Left

From the above graph, we noticed that even if

the employees are working on 7 projects and their evaluation rating is high, they are getting paid really less. This could also be one of the reasons for the employees to leave.

Hence, from all the above visualizations we tried to reach the factors which are strong predictors of employee leaving an organization.

4 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification[3] is to accurately predict the target class for each case in the data. In this data set the target variable is the variable 'left' which indicates if the employee had left the organization or not. Classification is mainly performed to help predict which employee would leave the firm in the future, based on the past and present employee data.

Each technique employed a learning algorithm to identify a model that best fits the relationship between the attribute set and class label 'left' of the input data. We used a number of techniques to classify the employee data set. We implemented Decision tree algorithm, Random Forest model and the Support Vector Machine to see which model has the highest classification percentage to be the best model for the employee data set. Taking the help of ROC curve and the classification errors, we reached a conclusion that random forest model best classified the data set.

4.1 Decision Tree Implementation

Decision tree is a graph that divides the data into nodes by using a branching method to illustrate every possible outcome of a decision. Decision Tree is a non-parametric supervised learning method used for classification. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. According to the decision tree algorithm, the best split of the data is the one which divides it into purer partitions. This would mean that the best split is the one whose difference between the degree of impurity between parent and child nodes is maximum. As the degree of impurity decreases with each split, the data becomes much classified. The larger their difference(gain), the better the test condition would be. The gain can be calculated by the equation:

$$Delta = \bar{I}(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

where I(.) is the impurity measure of a given node, and 'N' is the total number of records at the parent node, k is the number of attribute values, and $N(v_j)$ is the number of records associated with the child node, v_j. Decision tree induction algorithms often choose a test condition that maximizes the gain.

Before the decision tree algorithm has been implemented to this particular data set, the sales and salary variables are converted into numeric factors. The data set has been initially divided into training and testing data sets using the sample function. The training data accounts to 60 percent of the data set and the remaining 40 percent being the test data set. Decision tree algorithm has been implemented on the training data using the rpart library functions. Before the classification error has been calculated on test data, the decision tree is pruned to avoid overfitting. The complex parameter value for corresponding to the minimum cross validation error is 0.01. The classification error is calculated to be equal to 0.03059368. This means that the accuracy percentage for this algorithm is close to 96.9 percent. The decision tree has been plotted using the fancy r part plots from rpart.plot library.

The decision tree for the data is as below:

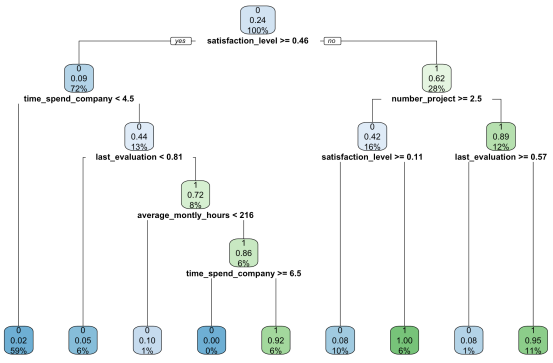

Figure 7: Decision Tree

4.2 Random Forest Model

Random forest is the class of ensemble methods specifically designed for decision tree classifiers. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors. Bagging using decision trees is a special case of random forests, where randomness is injected into the model-building process by randomly choosing 'N' samples, with replacement, from the original training set. Randomization helps to reduce the correlation among decision trees so that the generalization error of the ensemble can be improved.

The random forest model has been implemented on the test data using the random forest library in R packages.[5] The algorithm has been implemented with 1000 trees. Evaluation of the performance of this classification model is based on the counts of test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix. On calculating the confusion matrix, we see that the true positive and true negative values are pretty high. Although a confusion matrix provides the information needed to determine how well a classification model performs, summarizing this information with a single number would make it more convenient to compare the performance of different models. This can be done using a performance metric such as accuracy, which is defined as the ratio of number of correct predictions to the total number of predictions. The accuracy of this random forest model is 98.6 percent.

4.3 Support Vector Machine Model

Support vector machine[1] classification is based on a statistical learning algorithm. SVM generally works well with high dimensional data and avoids the curse of dimensionality problem. The basic idea of SVM involves constructing a hyper plane which divides the data into two planes maximizing the classification. This plane is called the Maximum margin hyperplane. It represents the decision boundary using a subset of the training data know as support vectors.

An improved version of the Support Vector Machine model gives an accuracy close to 96 percent. The model has been implemented on test data using the e1071 and quadprog libraries in R. The initial SVM model has given a lower accuracy percentage of 95 percent. The improved model has been built by changing the gamma and cost values. Intuitively, the cost parameter trades off misclassification of training examples against simplicity of the decision surface. Low value of cost tends to make decision surface smooth, while a high cost value tries all training examples correctly by giving the model freedom to select more samples as support vectors. Gamma parameter defines how far the influence of a single training example reaches, with low values connote far and high values connote the neighborhood. Therefore, our improved model is obtained by inputting the gamma value to 0.25 and the cost to 10. The plot of the training data is as below:

4.4 Which model works best and Why

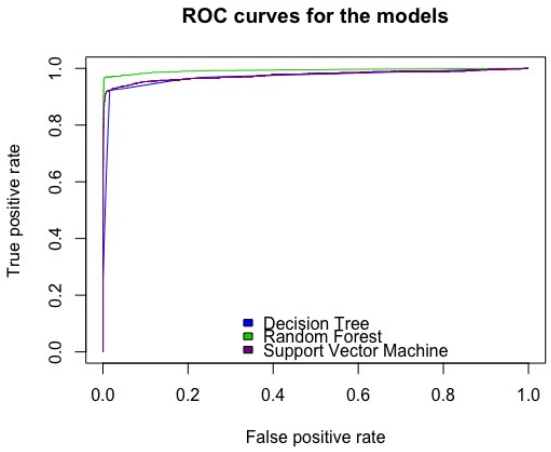

Figure 8: Receiver Optimist Characteristic curve for the three models

We know that the key objective of the learning algorithm is to build models with good generalization capability; i.e., models that accurately predict the class labels of previously unknown records. Most classification algorithms seek models that attain the highest accuracy, or equivalently, the lowest error rate when applied to the test set. Therefore, to decide which model best works for the data, ROC curve can help find the best performer. By plotting

the Receiver Optimistic Characteristic curve we could see that Random Forest model best worked for our data with the highest efficiency. The plot of the ROC curve for all the three models is as above.

5 Conclusion

The visualizations have unraveled multiple facets of employee resignation and the knowledge across its varied reasons. It has provided insights over the multiple reasons leading to resignation of employees through the various scatter plots and histograms. Our present work creates a basis for a classification of employee data. From the above ROC plots, we chose the final model to the employee data set classification to be the random forest classification model with 1000 trees Giving an accuracy of 98.7 percent.

6 Acknowledgements

We would like to thank our Professor Christopher Raphael for his support and advice all throughout. We would also like to thank our Assistant Instructors for their technical guidance and feedback.

References

[1] Jean-Philippe Vert. *Introduction to SVM in R.* NA, unknown.

[2] Ludovic Beninsant. *Human Resource Analytics for Employees.* Kaggle for Data Scientists, California, USA, 2015.

[3] Pang-Ning Tan; Michael Steinbach; Vipin Kumar. *Introduction to Data Mining.* NA, unknown.

[4] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015.

[5] Scott Scoltzman. *Plotting Trees for Randm Forest Models.* R Bloggers, California, USA, 2017.

[6] Hadley Wickham. Getting started with ggplot2. In *ggplot2*, pages 11–31. Springer, 2016.

[7] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer, 2016.