CS 6140: Intermediate Report - A Data Mining Approach on Attrition Rate Analysis of Employees

Group 4: Sanjana Aravindan, Vinu Sreenivasan, Harshini Keerthi Vasan

Dataset

The dataset that we are working on is HR analytics dataset collected from Kaggle (publicly available). It consists of around 15,000 records and 10 parameters that are the factors related to the employee details. The factors that describe each employee's record are employee satisfaction level, last evaluation, number of projects, average monthly hours, time invested for the company, work accident, promotion in the last 5 years, department, salary, current employee or ex-employee.

Progress towards the proposed goal

Our proposed goal is to analyze attrition rate of the employees. Attrition rate is the measure of the number of individual's moving out of the organization. By examining the correlation among employee details from the dataset, we will be able to provide a distinction between prevailing employees and employees who left the organization. By studying this data, we were able to interpret what kind of employees are currently employed in the organization and what kind of employees left the organization . Going further we would design strategies to decrease the attrition rate.The sections further would give a detailed view of the basic approaches tried , what worked and what did not work.

Basic Approaches Tried

Data Preprocessing

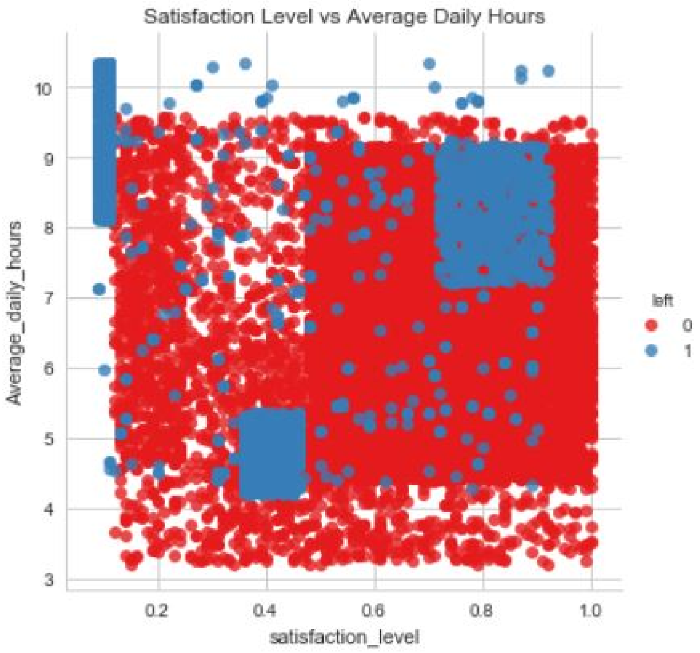
The following are the data pre-processing and feature engineering steps done over the HR data to yield logical and well-reasoned results.

1. ImproperEvaluation : For all employees who have a good evaluation (0.87, as its the 3rd quartile value), and the salary grade as low = Assigned Boolean value as yes for the new column.
2. Overrated : For all employees who have the satisfaction level, last evaluation and the number of projects less than the median value are technically low performing employees. Albeit if they are promoted, then they are considered as overrated employees. Hence this column holds a Boolean value as yes for those who satisfy the criteria.
3. Average daily hours : Column which helps us evaluate how much everyone has worked on average daily basis.The column "average_daily_hours" should be rounded off to two decimal places as the process of cleaning the data set.

Data Exploration

Data exploration is an important task when working with complex data set. We have used Data visualization as a technique in laying out important attributes and helps in exploring the data in an effective manner. Following are the observations that were made based on explorations done with seaborn plots in python and gg plots in R :

1. Employees from low and medium salary grade have left the company.
2. Employees with satisfaction level lower than the median value have left the company.
3. Employees with high average daily hours have left the company.
4. Employees who fall under the Improper evaluation category have left the company.
5. Majority of the employees left fall under sales, support and technical departments.
6. Employees with experience greater than 4, and still no promotion have left the company.The plot is shown in the *link* here.
7. Employees with higher satisfaction level and higher average daily hours, lower satisfaction level and lower average daily hours, lower satisfaction level and higher average hours have left the organization. Shown below is the plot for the same.



Association Analysis

Association Analysis is a technique for uncovering the interesting relations between the variables that are hidden in larger datasets. It is used to indicate the likely occurrence of an item based on the occurrences of other items in the dataset. It can be represented in the form of Association rules or sets of frequent items. The associative model is denoted by X → Y. The certainty or the strength of the rule is determined by Support, Confidence and Lift.

$$\text{Support is } S(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$
$$\text{Confidence is } C(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$
$$\text{Lift is } L(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X) \times \sigma(Y)}$$

By using Apriori algorithm and generating associative rules between certain parameters in the dataset, we could figure out that the attributes department,number of projects strongly influence the "left" attribute.To implement this, we considered the LHS(Antecedent) as "department"(one of the attributes) and RHS(consequent) as "left" . The RHS is fixed in order to make an analysis on the attrition rate of the employees.
Based on the *plot*, we can see that the people from HR department are highly likely to leave the company. From the plot, (HR,1) has lift value as 1.2, which is the highest. Further, we can see that accounting department is highly likely to leave the organization with lift value 1.1. RandD and management are the departments whose employees are least likely to leave the organization i.e., those two departments have higher number of employees surviving in the organization.

Principal Component Analysis

It is a dimensionality reduction or data compression method that can be used to reduce the attribute space such that the existing set of variables is reduced to a smaller set that still contains most of the information in the large dataset. While performing the PCA for the HR dataset, we drop off the factors whose values are alphanumerical. We consider only those factors whose values are continuous or discrete. Since left features is represented in binary(0 or 1), we may consider that as output label and slice it such that there are 7 factors for performing the PCA.
The dataset needs to be standardized for better performance results. We standardize the variables by shifting the distribution of each variable with mean zero and standard deviation of one .On implementing PCA to our dataset, we noticed that maximum variance is given by the first principal component i.e., satisfaction level, with 26%. It can be noticed from the *graph* that the last principal component (promoted_last5years) gave the least variance of about 10% which needs to be eliminated. Thus, principal component analysis has helped us in improving the dimension of dataset to 6 and data has been simplified with uncorrelated factors.

Anomaly Detection

Anomaly detection method is used in identifying data points that don't conform to expected behavior. Unexpected data points are also known as outliers, exceptions etc. Anomaly detection provides critical and actionable information. We performed machine learning based anomaly detection algorithm to detect and eliminate outliers. It did not work for our dataset and it produced all points as outliers.

Thoughts for Improvement

It would be interesting to see how creating an associative analysis with others factors such as salary, duration of working hours, performance evaluation would affect the attrition rate of employees. We are planning to try the clustering based anomaly detection technique to remove outliers.

Experiments run to demonstrate the effectiveness

1. We adopted the approach of testing our algorithms on relatively small, medium and large datasets. We found that our associative analysis algorithm works well for small datasets too which was expected.

2. We tried data exploration with small, medium and large data set and could see that results did not change for strong parameters like satisfaction level, last evaluation, department and average daily hours whereas the results for parameters like promotion in the last 5 years fluctuated.