Final Project Report - A Data Mining Approach on Attrition Rate Analysis of Employees

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

Introduction: Problem and Motivation

Attrition rate is a measure of the number of individuals or items moving out of a collective group over a specific period. It is one of the two primary factors that determines the steady-state level of customers a business will support. It is often used by Human Resources professionals to determine a company's ability to retain employees. Attrition rate is increasingly used in the marketing world as a figure to project the number of new sales necessary to maintain the status quo. The development of an evaluation model to determine the attrition rate of employees for Human Resource Management is highly necessary because employees of an organization play a vital role in determining the organization's performance and picture the complete profit scale. Every strategy of an organization is directly or indirectly associated with the employee's talent. Hence, it is vital to retain the employees who serve to increase the organization's profit. Employee attrition occurs when employees break ties with an organization. In fact, attrition statistics show that positive relationships with employees lead to employee loyalty and retention, while negative experiences result in higher employee attrition rates and lower profits.

Attrition rate is also one of the key business metrics, as the cost of retaining an existing employee is far less than acquiring a new one. As employee retention becomes a valuable metric, it is in companies' best interest to reduce employee attrition rate. Thus, in our project, we have tried to provide a distinction between prevailing employees and employees who left the organization by examining the correlation among employee details from the dataset we use. By mining 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. Determining the cause of employee attrition helps an organization make positive changes and deliver better experiences to meet employee's expectations and ultimately reduce attrition rate.

Key Idea

The key idea behind experimenting with HR Data Analytics dataset is to use interesting techniques such as **Principle Component Analysis (PCA)**, **Clustering**, **Apriori algorithm for Association Analysis** and determine the attrition rate of employees.

The problem of finding the cause of employee attrition, as mentioned, can be reduced to finding association between the important factors contributing to attrition. Dimensionality

reduction is one of the key ideas that helps us to eliminate the attributes which give the least variance in the dataset. We use the PCA technique for dimensionality reduction so that the existing set of variables is reduced to a smaller set and the reduced set still contains the most of the important information to determine the attrition rate.

We further perform **Clustering** on the dataset with reduced set of attributes that we obtained from PCA. Clustering a large human resource analytics data for attrition rate is a challenging task in itself given the volume and size of data which we are dealing with and exploring them could lead to several interesting results. We implemented Assignment Based Clustering on the dataset and got some interesting results on why employees left the organization.

Finally, we apply **Apriori** algorithm for associative analysis on the pre-processed dataset to find the correlation between various attributes in the dataset. This will help us uncover the dependencies between attributes on which assignment based clustering could not be performed (Since, assignment based clustering is not directly applicable to categorical attribute values in the dataset). All these approaches put together, will help us determine the most important factors that contribute to an increasing attrition rate.

The Data

Overview

Human resource analytics dataset of employees working in a particular organization is publicly available[1]. For our project, we have used a dataset from Kaggle which consists of around 15,000 records and 10 parameters with 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**.

Data Preprocessing

Data pre-processing is a crucial step to carefully screen out attributes which do not yield results and add attributes which will help us in making a good judgement towards our analysis. The following are the data pre-processing and feature engineering steps done over the HR data to yield logical and well-reasoned results.

1. **Improper Evaluation**: 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.

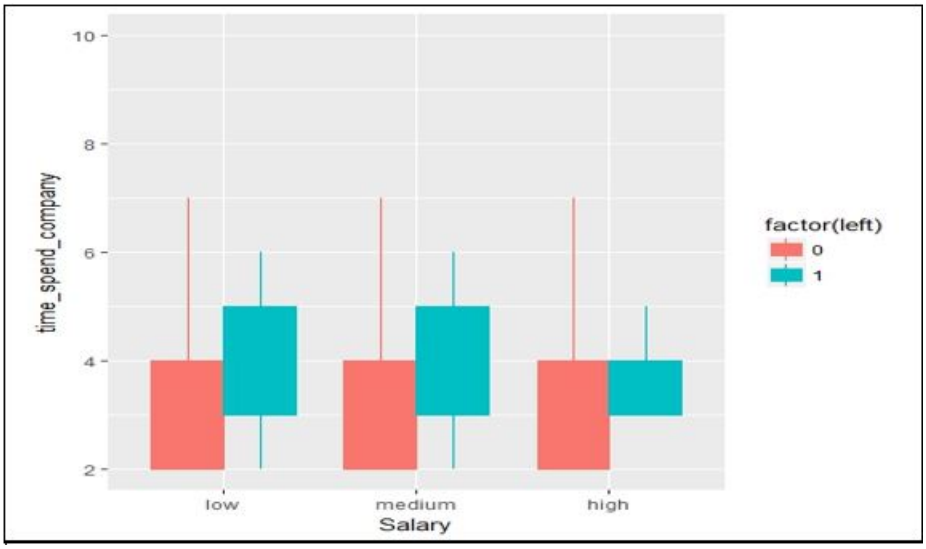
2. **Overrated**: For all employees who have the satisfaction level, last evaluation and the number of projects less than the median value are technically low performing employees. Albeit if they are promoted, then they are considered as overrated employees. Hence this column holds a Boolean value as **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. It is the process of summarizing the major characteristics of the attributes present in the data. Data visualization acts as an important technique in laying out important attributes and helps in exploring the data in an effective manner. Following are a few observations that were made based on explorations done with seaborn plots in python and gg plots in R. Only few of the plots have been shown here in the document. The rest of them can be accessed using the links attached for the corresponding observations.

1. Employees leaving the company have spent more years with *salary* levels low and medium. The plot is shown below,



2. Employees who are under performing are over rated. The plot is shown in the *link* here.

3. Only a small number of employees have a high salary. The plot is shown in the *figure* here.

4. Management department has the least attrition rate as it has a higher proportion of highly paid employees.

5. There is almost no attrition in High Salary Paid Employees.

6. We can see that the employees with salary level low have not been promoted much in last 5 years.

7. There are over rated employees who are under performing but promoted.

8. We can note that as the experience of the employee increases, they tend to leave the company.

9. Employees with number of projects as 2 have mostly left the company. Secondly, employees with 6 and 7 projects have left the company.

10. We can voice saying employees with few projects and the ones burdened with projects have left the company.

11. Employees with higher satisfaction level are evaluated incorrectly. The plot is shown in the *link* here.

Observations from all the above interactions

1. Employees from low and medium *salary* grade have mostly left the company.

2. Employees with *satisfaction level* lower than the median value have left the company.

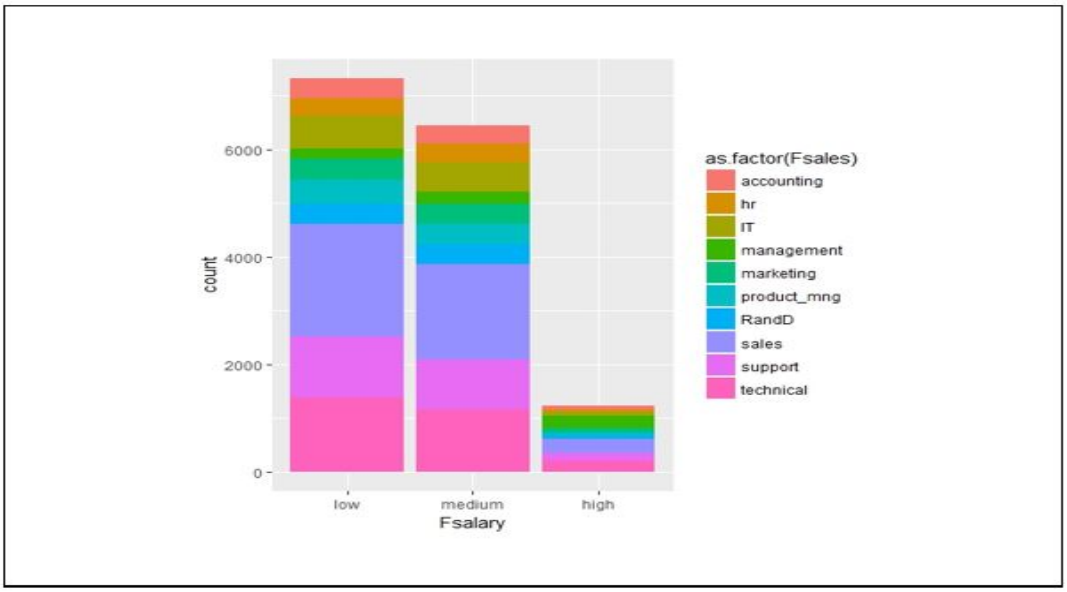
3. Employees with *high average daily hours* have left the company.

4. Employees who fall under the *Improper evaluation* category have left the company.

5. Employees with *experience* greater than 4, and still no promotion have left the company. The plot is shown in the *link* here.

6. Employees with higher *satisfaction level* and higher *average daily hours*, lower *satisfaction level* and lower *average daily hours*, lower *satisfaction level* and higher *average daily hours* have left the organization. The plot is shown in the *diagram* here.

7. We could see that *sales* department has majority of the employees falling under *salary* level low and medium. We could also identify that *support* and *technical* departments have *salary* level low and medium. The plot is shown below,



What We Did

Principal Component Analysis (2 Components)

1. **Basic Approaches Tried**

   (a) **Description**
   Principal Component Analysis (PCA) is a dimensionality reduction technique that can be used to reduce the attribute space such that the existing set of variables is reduced to a smaller set with the most important information about the dataset. The first principal component accounts for as much of the variability in the data as possible. In the similar way, each succeeding principal component accounts for as much of the remaining variability too.

   (b) **PCA on HR Analytics Dataset**
   Since HR analysis dataset is around 13 dimensions and a few of the attributes were causing some noise in the data, we performed PCA to reduce the dimensions of the dataset. As it would be impossible to get a good visualization of all the 13 attributes (dimensions) put together, we used PCA to reduce it into 2 components. This is done by performing linear combination of attributes and the attributes with maximum variance and second highest variance are further extracted.

   While performing 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 $left$ feature (ex-employee or current employee) is represented in binary (0 or 1), we consider it as output label and slice the remaining dataset such that there are only 7 factors considered for PCA implementation. They are *satisfaction level*, *last evaluation*, *number of projects*, *average monthly hours*, *time spent in company*, *work accident*, *promotion in last 5 years.*

   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. We can further find eigenvalues and eigenvectors from the covariance matrix. The covariance matrix is calculated as,

   $$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

   (c) **Eigenvectors and Eigenvalues**
   The **eigenvalue** is calculated as follows,

   $$|(A - \lambda I| = 0$$

   Here A is the matrix, λ is the eigenvalue and I is the identity matrix. Eigenvectors are uncorrelated linear combinations of the original set of random variables.

The **eigenvectors** for the features are calculated as,

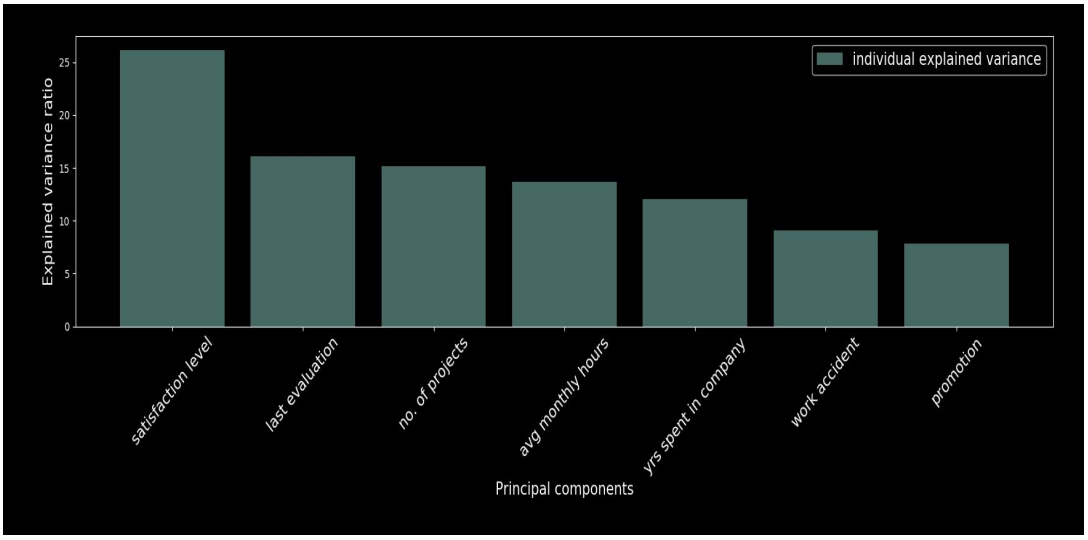
$$AV_i = \lambda V_i$$

Here V_i denotes the eigenvectors.

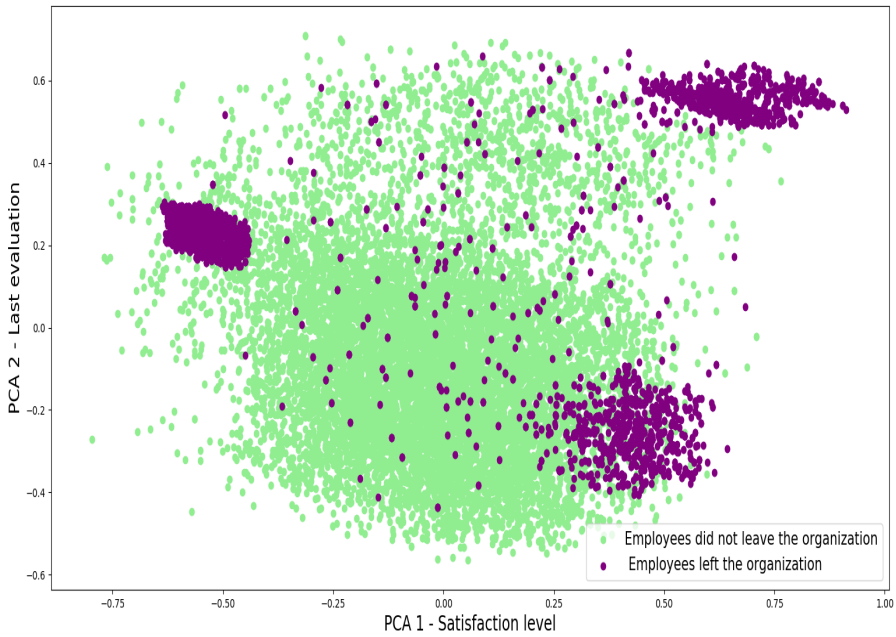
When we find the 7 eigenvectors/values of the attributes data set, 2 of the eigenvectors will have large eigenvalues and other eigenvectors will have relatively lower eigenvalues. Eigen value for a given factor measures the variance in all variables which is accounted for by that factor.

The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data and the corresponding attribute is ignored for analysis. The explained variance gives us an idea on how much information (variance) can be attributed to each of the principal components.

2. **Results and Learning**
We noticed that maximum variance is given by the first principal component i.e., *satisfaction level*, with 26%. The second principal component which provides the next maximum variance is *last evaluation* feature with a variance of about 17%. The plot for explained variance is given as below.



From the above graph of explained variance, we get that the *satisfaction level* and *last evaluation* are the best components that can be represented in 2 dimensions using PCA. The clusters below (based on *satisfaction level* and *last evaluation*) indicate an analysis of employees who left the organization.



Clustering

Basic Approaches Tried

Given that there are several clustering formulations out there in the wild, if only we had a fairy or an oracle throwing out the clustering formulation that would be best suited for our given problem, whenever we needed, things would have been great. But sadly since there seems to be no such fairy existing, we thought why don't we learn and get to see if Assignment Based Clustering works on our dataset. Since we were interested in exploring the different clustering formulations on the HR Data Analytics, we decided to take up different combinations of attributes in the dataset and find some interesting observations on why employees left the organization. The details of our clustering technique is given below,

Assignment based clustering (K-means)

1. **Description**
   The clustering technique that we tried to explore for our given problem is K-means clustering. It is an assignment based clustering technique and given the number of clusters to be formed it gives us information about which clusters will our data points belong to. The notion of clusters in our data set would represent the combination of factors that led to employees leaving the organization. The programming platform that we used to perform K-Means is **Python**.

   Since clustering can be performed on continuous set of values, we have considered attributes such as **satisfaction level**, **last evaluation** and **average daily hours** to perform clustering on and evaluate the basis on which employees left the organization.

The **algorithm** that we followed to perform **K-Means Clustering** is,

(a) Initialize cluster centers by randomly picking k nodes from data set.

(b) Next we compute the closest center for each node which takes each data point in our data set and computes the center that is closest to it.

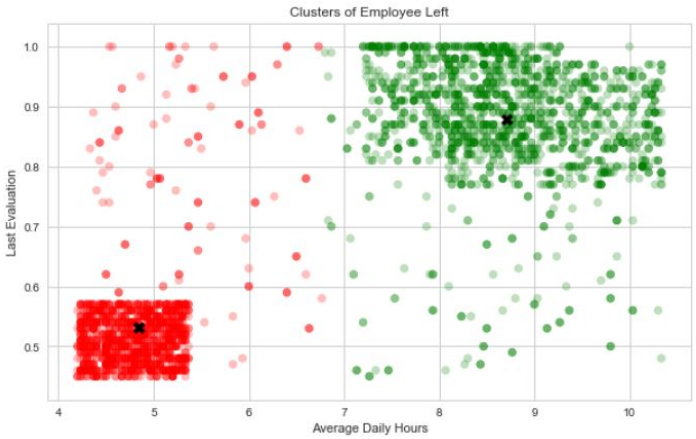
(c) Once we have found the centers and the nodes close to these centers, we now find the new center by computing the mean of the nodes that are in a single cluster. This is done to each node in a cluster.

(d) Now we again group the nodes based on the cluster center that is closest to them.

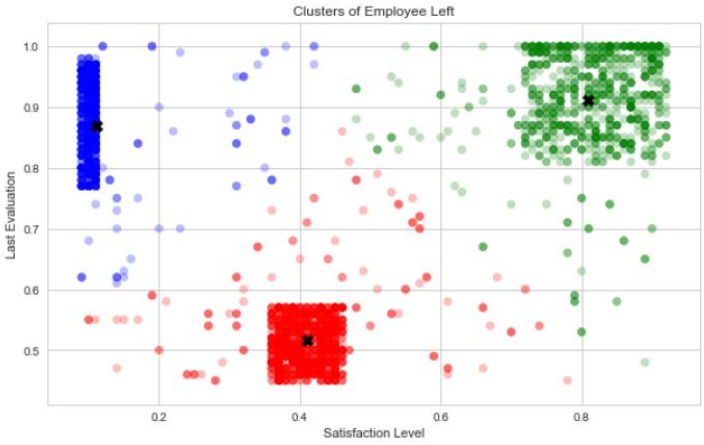
(e) We continue the above steps until we have the required k number of clusters and cluster centers which happens when the distance between the nodes and their present closest center is lesser than the previous center.

2. **Results and Learning**
We wanted to visualize the clusters obtained at the end of the algorithm. We set a fixed value of k for each combination of attributes (satisfaction level, average daily hours, and last evaluation) and tried plotting the data set values for employees who left the organization based on these attributes. Following are the results we could obtain,

(a) We set k=2 for the first analysis. The first set of analysis based on columns - **average daily hours** and **last evaluation** shows that employees who had maintained a very less average daily hours and had a bad evaluation, tend to leave the organization. Similarly, employees who tend to work for more hours and have a very high evaluation also tend to leave the organization (may be in search of better opportunities). The graph of the analysis is shown below,



(b) We set k=3 for the second analysis. The second set of analysis based on columns - **satisfaction level** and **last evaluation** shows that employees with very low satisfaction level and very high last evaluation tend to leave the organization. Similarly, employees with average satisfaction level and average last evaluation tend to leave the organization. The third cluster suggests that employees with very high satisfaction level and very high last evaluation also tend to leave the organization. The plot of the analysis is shown below,



(c) We set k=2 for the third analysis. The third set of analysis based on columns - **average daily hours** and **satisfaction level** shows that employees with medium satisfaction level and very low average daily hours tend to leave the organization. Similarly, employees with very low satisfaction level and very low average daily hours or employees with an average average-daily-hours and high satisfaction level also tend to leave the organization. The plot of the analysis is shown *here*.

Apriori Algorithm for Association Analysis

1. **Basic Approaches Tried**

   (a) **Association Analysis Description**
   Association Analysis is a technique for uncovering the interesting relations between the variables that are hidden in larger datasets. It is used to find association between attributes with categorical values and discrete values (which could not be performed in assignment based clustering). The association rules indicate a strong relationship between records in the dataset. It is denoted by $X \rightarrow Y$. The certainty or the strength of the rule is determined by Support and Confidence.

   All items in a rule indicate an itemset. **Support** indicates how popular an itemset is, that is measured by the proportion of transactions in which an itemset appears. **Confidence** measures how often each item in Y appears in transactions that contain X. In order to consider the popularity of both X and Y, another measure **lift** tells how likely an item Y is to occur when item X occurs, while controlling the popularity of Y.

   (b) **Apriori Algorithm on HR Analytics Dataset**
   The **Apriori Algorithm** is an influential algorithm for mining frequent item sets for boolean association rules. The **Apriori principle** is as follows: If an item set is frequent, then all its subsets must also be frequent.

   $Support(X \rightarrow Y) \Rightarrow$ Number of transactions which contains an item set X / (Total number of transactions)

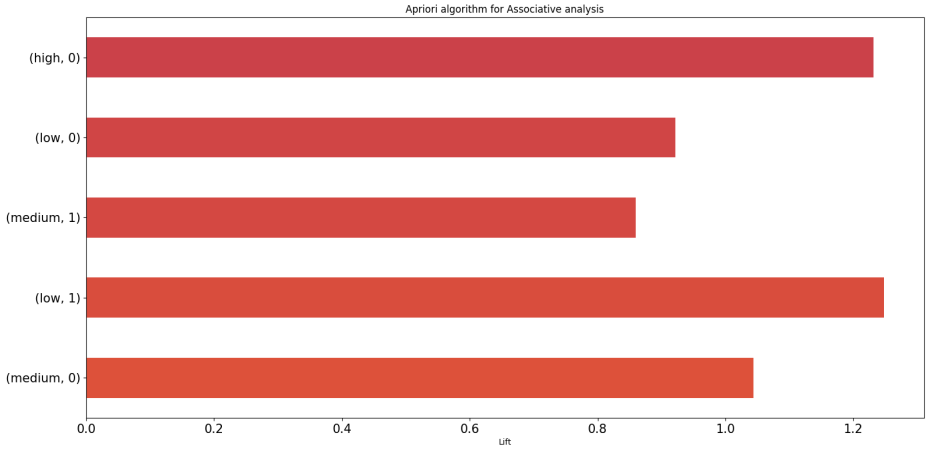
$$Confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

$$Lift(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X) * support(Y)}$$

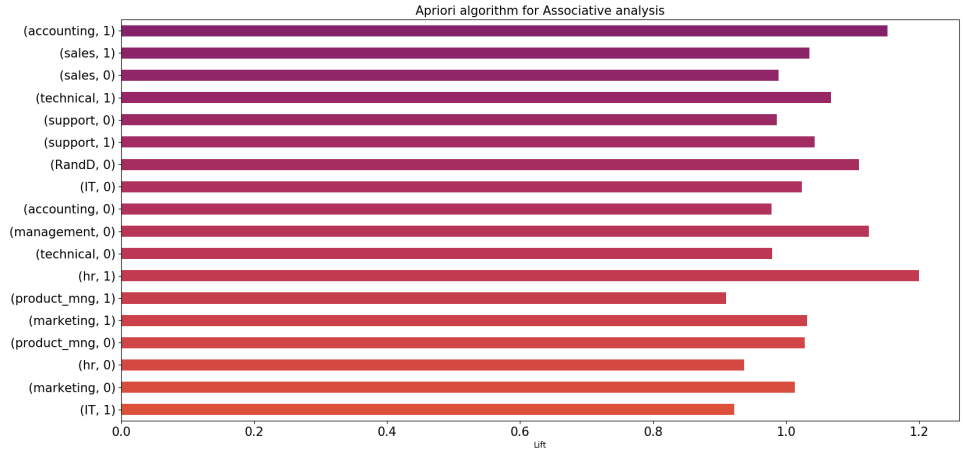
First, we considered the LHS (Antecedent) as **department** and RHS (consequent) as **employees who left the organization** and have kept RHS fixed through out the process in order to make an analysis on the attrition rate of the employees.

After pre-processing the data, we generate association rules on this data using minimum support = 0.01 and minimum confidence = 0.03.

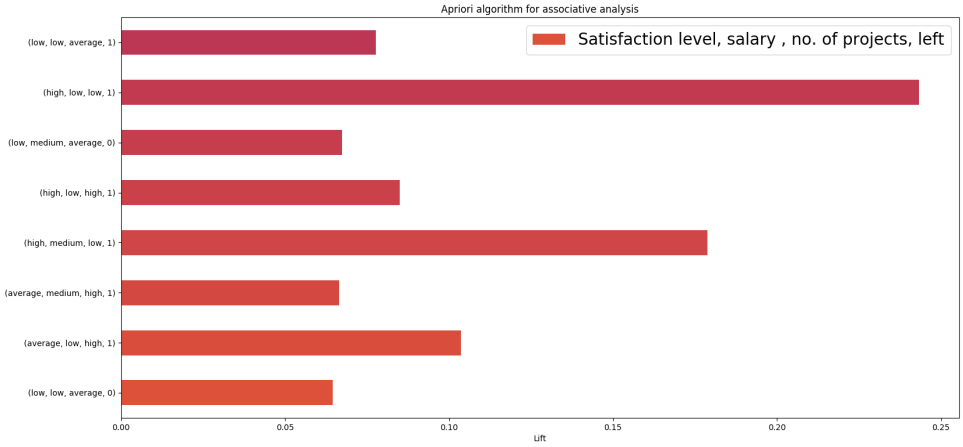
2. **Results and Learning**
The plot below depicts the associative analysis between *salary* (high, low, medium) and $left$ (current employee or ex-employee) feature with apriori algorithm.



The plot below depicts the associative analysis between *department* and *left* (current employee or ex-employee) feature with apriori algorithm. From the plot below, we can see that that employees from HR department are highly likely to leave the company. In the plot, (HR,1) has a lift value of 1.2, which is the highest. Next we can see that employees from accounting department are more likely to leave the organization with lift value 1.1. RandD and management department are the least likely to leave the organization i.e., those two departments have higher number of employees surviving in the organization.



The plot that provides association using apriori algorithm between 3 parameters such as *satisfaction level*, *salary*, *number of projects* is shown below. Also, association between *satisfaction level*, *last evaluation*, *average daily hours* is depicted in *plot2*.



Based on these two plots, we can summarize that the characteristics of an employee who has left the company are,

(a) An employee with *high satisfaction level*, *low salary* and *low* number of projects.

(b) An employee with *high satisfaction level*, *medium* salary and *low* number of projects.

(c) An employee with *average satisfaction level*, *low evaluation* in the last term and *high average daily hours*.

(d) An employee whose *average daily working hours* were *low*.

(e) An employee whose *salary* was *low* or *medium*.

Overall Results and Learning

All the data mining approaches used above and the corresponding analyses tend to state the the major factors responsible to evaluate the employees attrition rate. They are,

1. **satisfaction level**
2. **last evaluation**
3. **number of projects**
4. **average daily hours**
5. **salary**

From the factors stated above, employee *satisfaction level* is the most important parameter to determine the attrition rate of the employees. It could also be inferred (from the various plots and graphs) that employees who under-worked or over overworked generally left the organization. Also, employees with very less *number of projects* tend to leave the organization. The *evaluation* of employee in the previous term plays a significant role in deciding if the employee should stay or leave the organization. Employees with *low* or *average* salaries are more likely to leave the organization.

Also, it is to be noted that an employee with any one or two of the above factors having a *high* value and other attributes having a very *low* value will tend to leave the organization. An equal balance between these factors strictly needs to be maintained for the employee to be happy with the organization and not leave it.

Distribution of Work

All team members contributed to each part of the project equally. Coding of every module and every write-up was discussed and worked-upon together.

References

[1] https://www.kaggle.com/rohandx1996/human-resource-analytics/data

[2] J. Yang, D. Zhang, A. Frangi, and J. Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, (1):131137, 2004*

[3] Y. Jia, G. Xia, H. Fan, Q. Zhang, X. Li, An Improved Apriori Algorithm based on Association Analysis, *J. Bacteriol. 15 (15) (2012) 208211*

[4] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, An efficient K-means clustering algorithm: Analysis and implementation, *IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881892, Jul. 2000.*