ANALYSIS OF EMPLOYEE ATTRITION AND PERFORMANCE

Team: Data Wizards

(Prathik B Jain, Nitish S,
Supreet Ronad, Sandeep Bhat)

ABSTRACT: -

Employee retention is one of the biggest challenges in IT companies all over the world. The cost of employee attrition would be the cost related to the human resources life cycle, lost knowledge, employee morale, and organizational culture. Different companies adopt different strategies to retain the employees. These strategies include large increases in compensation, liberal perks, frequent job rotations, as well as travel and stay abroad. However, literature on turnover indicates that a person's intention to quit is a function of demographic characteristics, job characteristics and organizational characteristics. Individuals who have an intention to quit are also likely to engage in other withdrawal behaviors like absenteeism and late-coming. This study aimed to analyse employee attrition using appropriate models.

I. LITERATURE SURVEY

[1] ***Title:*** HR ANALYTICS: EMPLOYEE ATTRITION ANALYSIS USING LOGISTIC REGRESSION
(By: I Setiawan*, S Suprihanto, A C Nugraha and J Hutahaean, Department of Computer Engineering and Informatics, Politeknik Negeri Bandung, Bandung, Indonesia.) - April 2020 **[1]**

Relation: The article relates to the data and problem definition of our project analysis i.e. probability of employee attrition. The dataset is also very similar to what we have chosen for our project analysis.

Claims: Some of the main claims made based on the results obtained are:

- An employee with a single marital status has a more significant number in attrition than those who divorced and married.
- Some employees, especially "junior" employees still want to have more experience.
- An employee with a small number of working years and companies worked has a more significant probability of attrition.
- To reduce employee attrition rate, the company needs to improve the human resource department by evaluating the working environment, job satisfaction, employee workload, and interaction between manager and employee.

Takeaway: This study aimed to analyze employee attrition using logistic regression. The result obtained can be used by the management to understand what modifications they should perform to the workplace to get most of their workers to stay.

[2] ***Title:*** ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS - Alao D. & Adeyemo A. B**.** Department of Computer Science, University of Ibadan, Ibadan, Nigeria - March 2013 **[2]**

Relation to our project, claims, assumptions, key-takeaways:

In the above referred paper, the method of decision tree learning and corresponding rule-set generation is used to develop a predictive model in order to predict new cases of employee attrition based on various factors and aid in improving those factors in order to reduce attrition and improve performance of employees.

One of the main or standard assumptions made in any decision tree learning method is that instances belonging to different classes have different values in at least one of their features.

Repeated implementation of the decision tree learning algorithm helped to reduce the error rate and the results from the study based on the collected dataset showed that employee salary, length of service, salary hike, employee ranking were the main factors which were interconnected and influenced the attrition rate and performance.

The results obtained from the developed classifiers were convincing to the organizations and as such did not see any limitations or lacuna in the evaluation performed.

Decision trees generally tend to perform better when working with discrete/categorical data. Since our dataset mainly contains both discrete and categorical data, the method proposed in the paper can be suitable for model building. As a result, the author of this paper used decision tree learning among the other data mining techniques due to its advantages like: comprehensibility, robust nature, good performance even for large data etc.

[3] ***Title:*** THE EFFECTS OF EMPLOYEE EMPOWERMENT ON ACHIEVEMENT MOTIVATION AND THE CONTEXTUAL PERFORMANCE OF EMPLOYEES **-** African journal of business management - August 2011 **[3]**

Relation: The article relates to the data and problem definition of our project analysis i.e. probability of employee attrition. The dataset is also very similar to what we have chosen for our project analysis.
Claims: Some of the main claims made based on the results obtained are:
- structural equation model (SEM) was used.
- SEM was preferred because of the fact that SEM is an analysis method which enables theoretical models to be tested as a whole.
- SEM has two basic elements. These are; measurement model and the structural model. The period which evaluates the structure between latent variables and observed variables is called the "measurement model"
- to measure the variables regarding validity and reliability of the factors. Structural model however, is used to model the relationship between latent variables by defining explained and unexplained variables

Takeaway: This study aimed to analyze employee attrition using structural equational model.
Empowerment perception of employees affects their achievement motivation positively and their contextual performance positively.

[4] Title: EARLY PREDICTION OF EMPLOYEE ATTRITION IN SOFTWARE COMPANIES-Application of Data Mining Techniques1 - Aug 8-11, 2007 **[4]**

Relation: This paper relates to our project analysis i.e Employee attrition. The data mining techniques used here can be implemented in our project.

Assumptions: Assuming that the employees in the sample are young, with fast growth in the company and continuing in the same team for a fairly long time. They are predominantly male and single.

Same dataset was used for training as well as for testing because of the dataset contained only a limited number of observations.

Claims: Here various data mining techniques are used. But we are focusing on mainly ANN (Artificial Neural Network)

Comparing these methods, we can see that:

The ANNs appear to have the lowest predictive accuracy percent where as the best prediction was possible with discriminant analysis.

The role of age as a variable in the Indian context is particularly significant.

It is also likely that many employees engage in a process of career exploration in their first few jobs. Therefore, the relationship between age and turnover has to be examined in the Indian context.

While these predictive accuracies are specific to the data used in the analysis and to the specific company in question, the study has shown that it is possible to predict the employee attrition, and identify those who are likely to leave the company even before they had made their final decision to leave.

Advantage of this model analysis:

It is proposed that the models developed and tested could be used with data on other employees in the company for cross validating the model.

II. PROBLEM STATEMENT

"Analysis of Employee Attrition Rate and Performance"

-> Predicting how various factors affect employee attrition and satisfaction.

-> Identification of significance source affecting the employee satisfaction and ranking them in the order of most to least important.

It is well understood that "human" assets are one of the most reliable sources of organizational performance, efficiency and effectiveness, organizations are expecting from their employees to demonstrate higher levels of efficiency, effectiveness, and performance. Work processes which are getting more complex and gradually challenging conditions of competition are the other factors which heighten the expectations of organizations from their human resources especially in the face of rapid developments in the areas of communications and information technologies, This requires human resources to have various additional competences.

To know the various reasons for employee Attrition, to know the Strategies for Employee Retention. In order for an organization to continually have a higher competitive advantage over its competitors, it should make it a duty to minimize employee attrition. Reasons for employee attrition are spread over various factors. Our goal is to find such factors and build a predictive model so as to help the organizations in improving on those factors which are more responsible for employee attrition.

III. UNIQUENESS IN OUR APPROACH TO SOLVE THE PROBLEM

Among all employee related problems, employee attrition is one of the key problems in today's scenario despite the changes in the external environment. Attrition is said to be a gradual reduction in the number of employees through resignation, death and retirement.
The main aim of the analytics is:

- To know the Various Reasons of Employee Attrition.
- To know the Strategies for Employee Retention.
- To examine the impact of three R's i.e. RESPECT, RECOGNITION and REWARDS on the retention of the employee and development of the organization.

Besides identifying the factors responsible for employee attrition, by targeting the same, an attempt is also made on suggesting ideas for employee retention which are easy to implement and cost efficient.
In order to reach the above aim, we intend to adopt two different methods. They are :

- > *DECISION TREE APPROACH:* -A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

-> *LOGISTIC REGRESSION APPROACH:* - Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

The few advantages of these models:

A. DECISION TREE: A significant advantage of a decision tree is that it forces the consideration of all possible outcomes of a decision and traces each path to a **conclusion**. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis.

B. LOGISTIC REGRESSION: Logistic regression is easier to implement, interpret, and very efficient to train. It makes no assumptions about distributions of classes in feature space. It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions. Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.

One of the main or standard assumptions made in any decision tree learning method is that instances belonging to different classes have different values in at least one of their features. On the other hand, the requirements to use a logistic regression model are absence of multicollinearity and no strong influence of outliers.

The methods used previously in solving the problem using the above approaches mentioned overlook the presence of multicollinearity and outliers which have a significant negative impact on the model constructed. However, the analytical process used by us ensures that all the assumptions and requirements are met.

The main reasons for adopting the above two approaches in our analytics process is because the dataset chosen mainly contains categorical(nominal) attributes and involves a binary classification problem (ie yes/no).
Using two methods, we can obtain better results by comparing both the models. This way we can also achieve maximum accuracy.

IV. EDA AND VISUALIZATION REPORT:-

Using Python, the following observations were made on the dataset as a part of initial Exploratory Data Analytics and Visualization part.
The target attribute in the dataset is 'Attrition' which is a binary attribute with values 'Yes/No'.
As a part of the pre-processing and exploratory analysis of data the following observations were made:
1) NUMBER OF COLUMNS: **35**
2) NUMBER OF SAMPLES: **1470**
3) NUMBER OF QUANTITATIVE VARIABLES: **16** (Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager)

4) NUMBER OF QUALITATIVE VARIABLES: **17** (Attrition, BusinessTravel, Department, Education, EducationField, EnvironmentSatisfcation, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, Over18, OverTime, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance)

5) The entire data was described w.r.t to the five-point summary, mean and count of each attribute using the python library.
6) It was also found using a python function that there are no missing values or NaN values in the entire dataset (i.e in

any of the rows of the dataset). As a result , there was no need for any kind of missing value treatment to the dataset.

7) Missing values were checked for using a plot and an inbuilt missing value count function of python.

8) Boxplots were plotted for a few attributes in order to find the presence of outliers which may affect the further analysis and model building.

9) A heatmap was plotted to find the correlation coefficients of each attribute against every other attribute present in the dataset.

10) Since the dataset is mainly constituted of categorical/nominal attributes, reassigning of the target variable to a numerical value was performed in order to ease the process of analytics and model building.

11) Also, a few of the attributes seemed to have no significant impact on the target attribute as well as the model building using plots and heatmap. As a result, those attributes were dropped as a part of the pre-processing of data.

12) Since the target attribute is ‘Attrition’, boxplots were plotted against each attribute values and the target attribute, in order to find the count and the impact of each value on the target attribute. This is one of the major visualization done which would aid the process of model building.

13) Histograms were plotted to find the value distribution of a few attributes which may also help in further analytical processes.

14) Swarmplots were constructed for a few attributes to show the distribution and count of values in relation to the target attribute.

15) Diagnostic plots which includes:
histograms, probability plots and boxplots were constructed for a few of the numerical attributes in order to find the presence of outliers and distribution of values across the dataset.

NOTE: All the results obtained and the plots that have been drawn are done using python libraries and the proof for the same is present in the Google Colab Notebook attached along with the submission in the Google Drive.

V.REFERENCES

[1] IOP Conference Series: Materials Science and Engineering HR analytics: Employee attrition analysis using logistic Regression
(https://iopscience.iop.org/article/10.1088/1757-899X/830/3/032001)

[2] Analyzing employee attrition using decision tree algorithms
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1012.2947&rep=rep1&type=pdf

[3] The effects of employee empowerment on achievement motivation and the contextual performance of employees
https://academicjournals.org/journal/AJBM/article-abstract/253107414671

[4] Early Prediction of Employee Attrition in Software Companies
https://www.researchgate.net/profile/Vishnuprasad_Nagadevara/publication/237223699_Early_Prediction_of_Employee_Attrition_in_Software_Companies-Application_of_Data_Mining_Techniques/links/0f3175386602a0613d000000/Early-Prediction-of-Employee-Attrition-in-Software-Companies-Application-of-Data-Mining-Techniques