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 employees. These strategies include large increases in compensation, liberal perks, frequent job rotations, as well as travel and stay abroad. However, the 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 analyze employee attrition using appropriate models.

I. INTRODUCTION AND BACKGROUND: -

"Analysis of Employee Attrition Rate and Performance"

1. Predicting how various factors affect employee attrition and satisfaction.

2. Identification of significant sources affecting 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 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 that 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.

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 those factors which are more responsible for employee attrition. This study aimed to analyze employee attrition using appropriate models.

II. PREVIOUS WORK: -

[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 the 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 the 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 that 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.

III. PROPOSED SOLUTION: -

Using Python, the following observations were made on the dataset as a part of the initial Exploratory Data Analytics and Visualization part.

The target attribute in the dataset is ‘Attrition’ which is a binary attribute with the 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 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 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 include: 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.

Two different models i.e. Decision tree Classifier and Logistic Regression Models are built to compare the results of them and find which would be a better classifier to solve the given problem statement based on the parameters defined for both of the models.

A] Decision Tree Model: -

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.

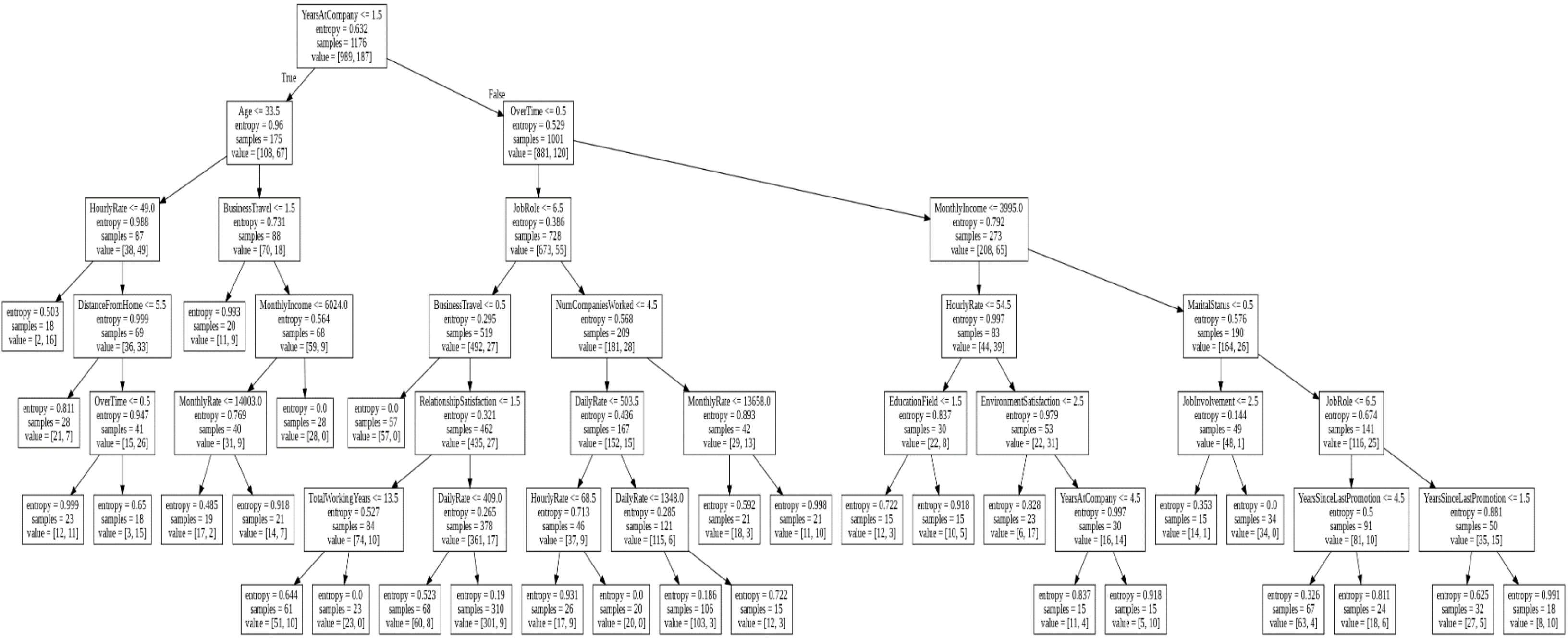
The method used in the construction of the decision tree for the dataset is as follows:

1. Transforming the categorical attributes in the dataset into dummies.
2. Diving the dataset into training and testing dataset so as to employ machine learning algorithms to build a model and make the model learn.
3. Construction of the decision tree using the 'DecisionTreeClassifier' of the 'sklearn.tree' library of python.
4. Fitting of the training dataset into the constructed decision tree.
5. Depiction of the constructed Decision Tree Classifier.

EXPERIMENTAL RESULTS: -

6. From the obtained decision tree, the attributes such as Job level, Overtime, Environment satisfaction, YearsAtCompany, Age, and Monthly income were found to have the most impact on employee attrition as compared to the others. And also among these Job level was found to have the highest contribution.
7. The constructed Decision tree provided the following results:
a. With a test dataset size of 20%, the ACCURACY of the TEST RESULT was found to be 82.31%.
b. With a test dataset size of 30%, the ACCURACY of the TEST RESULT was found to be 83.44%.

The decision tree construction involves careful selection of the parameters like max_depth, min_samples_split, min_samples_leaf, the number of features to be considered for



the best split etc as it might have an impact on the accuracy and precision of the model constructed as well as the time taken for the model construction.

(A clear picture of the decision tree can be seen in the execution of the Python notebook submitted)

B] Logistic Regression Model:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

To predict an outcome variable that is categorical from predictor variables that are continuous and/or categorical Used because having a categorical outcome variable violates the assumption of linearity in normal regression The only “real” limitation for logistic regression is that the outcome variable must be discrete Logistic regression deals with this problem by using a logarithmic transformation on the outcome variable which allow us to model a nonlinear association in a linear way It expresses the linear regression equation in logarithmic terms (called the logit)

The method used in the construction of the logistic regression for the dataset is as follows:

1. Import the necessary libraries and the data set. The data was taken from Kaggle and describes information about employee attrition and the factors affecting them. We will be predicting the value of attrition(yes/no)

2. Converting the categorical objects by data encoding

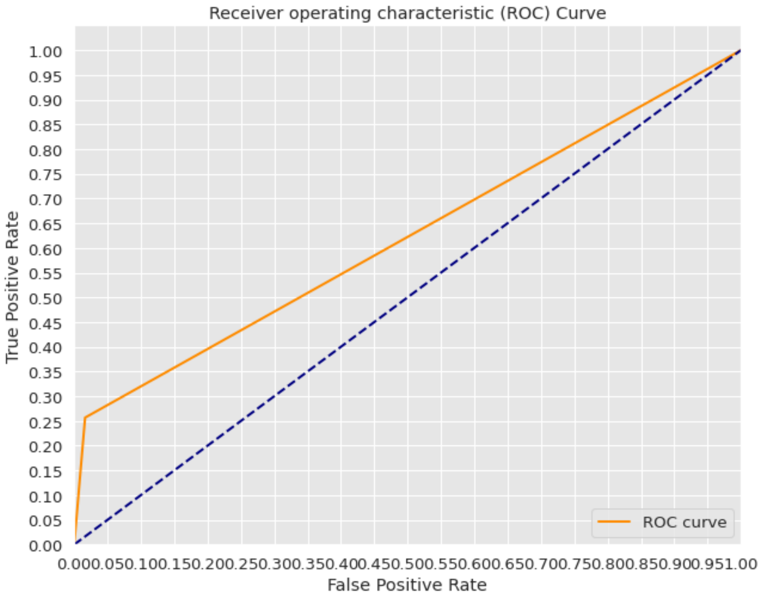
3. Dividing the dataset first into training and test data using sklearn library.

Checking the dimension of train and test data.

4. After diving the dataset into train and test data with the help of sklearn library we create our logistic regression model.

5. Next, we pass the training dataset to the model and train our model.

6. Now the model is ready to predict.



EXPERIMENTAL RESULTS: -

7. The constructed logistic Regression provided the following results:

a. With a test dataset size of 20%, the ACCURACY of the TEST RESULT was found to be 88.775 % and the AUC value was found to be 0.6202.
b. With a test dataset size of 30%, the ACCURACY of the TEST RESULT was found to be 86.16 %.

8. The accuracy and precision of the logistic regression model would however vary according to the values of the various parameters like choosing the right predictor variables, we should avoid the use of the highly correlated variable, handling continuous input variables, the assumption regarding relationship between input and output variables, Even wisely choosing the ratio of training and test data is important.

IV. CONCLUSIONS: -

The problem that we intended to solve through this analytical process was to reduce the rate of Attrition of Employees in a Company. Employee Attrition which is a process of the workforce dwindle at a company, following a period in which a number of people retire or resign, and are not replaced, has been an increasing concern in all the companies in the present years. Studies show that several attributes contribute to the unhappiness and dissatisfaction of the employees which causes an increase in attrition. These attributes include the number of years worked, the salary drawn per month, the work environment, job level, number of foreign travels, etc to name a few.

Through our analytical process, we wish to build machine learning models where the models are given a set of previous instances of an employee who left the job and the model hence learns and predicts the factors which have a larger influence on the attrition over the others.

The Decision tree gives a tree representation by ranking the factors contributing to attrition and hence predicts the same for a given new instance of an employee. The Logistic Regression model also performs on a similar basis. The results of the two models have been obtained as mentioned above.

These models can be used by the company frequently of course by making constant changes as well, in order to predict and control the attrition rate of the employees by improving on the various factors when found to have an impact on the same.

Thus, this analytical process was to identify the factors which have a great impact on Employee Attrition and hence take necessary steps to control the same and strive for Employee Retention and improvement of the performance which is extremely vital for any company.

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