International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Random Forest Analysis of Job Satisfaction

Jackulin Mahariba A., Tanvi Mahajan, Sparsh Heda

Abstract*: *Job satisfaction plays an important role in the productivity of an organization. Satisfaction of an employee cannot only be determined through variables like salary and location. There are lots of factors that affect satisfaction which further affects the performance of an organization. The main goal of this project is to obtain better knowledge of the parameters responsible for job satisfaction and based on it how various organizations differ from each other with respect to their working conditions. It presents the result of an empirical study of how factors like age, gender, department, education, marital status, hours per week, overtime, hike, native country etc. affects job satisfaction of the labor force of a particular country using machine learning. The study is based upon the results obtained through supervised algorithm, Random Forest. Through this research we attempt to discover how the different aspects of job satisfaction are related to job prevailing parameters. The model will help organizations increase productivity of its employees by ensuring a better working condition.

***Key Words: Job Satisfaction, Job Performance, Parameters, Classification, Hierarchical Clustering*.**

I. INTRODUCTION

The term 'job satisfaction' is furnished because the angle of content, associate worker possesses in his or her current position in a company. For example, an employee getting a decent salary working for a MNC might not be satisfied due to his stagnant position in the company regardless of the fact he is working hard. In the previous couple of decades, many studies have been conducted over the satisfaction of workers' as a central research variable. It's thought to be an obligatory attribute that is often measured by organizations so as to encourage a healthy and positive attitude of an employee towards the work that he engages in. Although the world support the idea of considering employees as an asset of the company, the profit-making mind-set of the organisations is hindering the monetary remuneration of the employees and goes against the welfare of the employees. Therefor there have been increasingly more instances of employees leaving and being replaced in organizations throughout the vertical hierarchies. This shows why job satisfaction levels are so important among the employees of any organization and how they affect the organizational effectiveness.

This paper is based upon the results obtained through machine learning algorithms. Machine learning is the ability of a machine to learn and predict accurate outcomes based upon the previous results while updating itself as new data becomes available without directly programming the machine. Algorithms of Machine Learning are further classified as supervised and unsupervised.

Supervised algorithms need data scientist to train the machine initially and once the training is complete, the algorithm will be applied to what was learned to new data while the latter need not to be trained with the desired output. Rather, it uses an iterative approach like clustering to review and classify data by identifying commonalities and react based on the presence or absence of such commonalities. This paper uses Classification under supervised form of learning which is implemented when the output variable is a category. The model included is decision tree that classifies the instances by sorting them depending upon values. Each node of the tree represents a feature that will be classified.

II. LITERATURE SURVEY

Hoppock in 1935 defined job satisfaction as combination of physiological, psychological and circumstances caused by environment that can leads a person to say, "I am satisfied with the kind of job I am doing". His study surveyed 450 flight attendants of an airline. In the result based upon the study, he concluded upon four factors – job environment, the nature of the job, characteristics of the organization, and social dimension. The results have shown that flight attendants' job satisfaction affects their positive productivity.

According to Vroom (1964), job satisfaction is more focused on the job role of an employee. He defined job satisfaction as an affective direction on the role of individuals in their work which they are presently occupying. The results were based on Simple random sampling and Purposive methods that concluded the general lack of motivation and low monthly salaries were the major parameters that lead to reduction of morale for high performance.

According to Spector satisfaction in job is more about the way how workers feel about the kind of job they are doing and its aspects. It is related to the point up-to which people can love or hate their job. The conclusion was based upon Multi-stage clustering sampling method and Regression Analysis stating Job satisfaction depends upon planning and implementation, time, goals and priority, need assessment, way of evaluation, welfare and services etc.(Spector 1997)

Armstrong refers feelings and attitude towards work as a reflection towards job satisfaction.

Manuscript published on 30 April 2019.
* Correspondence Author (s)
Jackulin A. Mahariba, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Tanvi Mahajan, B.Tech, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
Sparsh Heda, B.Tech, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

Positive and approving attitudes indicate job satisfaction. Negative and disapproving attitudes indicate dissatisfaction towards job. He conducted a survey based on some questions, designed to test his model .Testing for the model and hypothesis was done with the help of structural equation modeling analysis. Results indicated that positive involvement in the job significantly impacts organizational commitment and job satisfaction amongst workers.

In 2006, Soberman, Iyer and Christen provided a model for job satisfaction based upon job related parameters, job performance, firm performance and role perceptions.

George, J.M. and Jones, G.R. in 2008 in their paper stated, satisfaction of job depends upon various stances related to job like supervisors, subordinates, income workers and co-workers. Multiple Regression Analysis was implemented in order to test the hypothesis. According to the results based upon the T-test, gender does not have any impact on job satisfaction. ANOVA test found that there are significant differences on job satisfaction that could be related to educational level, age and personal income.

The study done by Kimball and Shapiro (2008) is based upon the income and substitution effects that cancel out in the overall labor market due to the fact that overall labor supply stays constant with respect to wage. The main variables that were used by this study were marital status, optimization over time, and fixed costs of going to work. Though it provided an in-depth look at flexibility in labor supply, but the difference in the income and substitution effects in various income brackets was left unexplored. The overall results show very high labor supply elasticity in response to a wage shock, suggests that the most of the people chose to reduce their working hours when given a hike and sudden increase in earnings.

Dr. Shatakshee Dhongd, Sina Mehdikarimi, Samuel Norris and Charles Stalzer in 2015 in their study analysed the usual number of hours worked per week and it's relationship with personal income. The regression analysis was largely in agreement with economic theory and partially in line with their hypothesis. The substitution effect appeared to be relevant because of the positive and significant effect of the log of adjusted income on usual hours worked per week. Shortcomings were the violations of the Gauss-Markov assumptions in multiple regressions.

As per a study done in 1996-2001 details, increase in job satisfaction is directly connected to a 6.6 percent increase in productivity per hour. Harvard Business Review in their study recently showed an analysis that depicts an average of 31 percent more productivity and 37 percent higher sales when employees are happy or satisfied.

III. IMPLEMENTATION

A. *Dataset*

The dataset is a census dataset which has nearly 50000 records with 16 attributes pertaining to the employees across the globe. There are attributes like age, work class, education, marital status, occupation, relationship, race, gender, income which directly or indirectly affect the job satisfaction level of an employee are considered for the prediction of the job satisfaction level of the employee.

B. *Data Pre-processing*

The dataset has many columns with string values, which are required to be in a float (numerical) form. We use pandas library to standardize the data into the format required by the algorithm used. The missing values of the dataset were evaluated and the result is as follows:

Column Name	Missing Values	Missing Values Percentage
workclass	2799	5.73%
occupation	2809	5.75%
native-country	857	1.75%

Figure 1 Description of the missing values in the dataset

Since the number of rows with missing values was very less and the prediction of the categorical columns would have distorted the data, we decided to replace the values with the most frequently occurring value of the respective column.

Correlation of the target value with the other data columns was considered for feature selection, where in the features with lesser correlation value were removed in order to attain a higher accuracy for the model. The same has been plotted on a heatmap in the following figure:

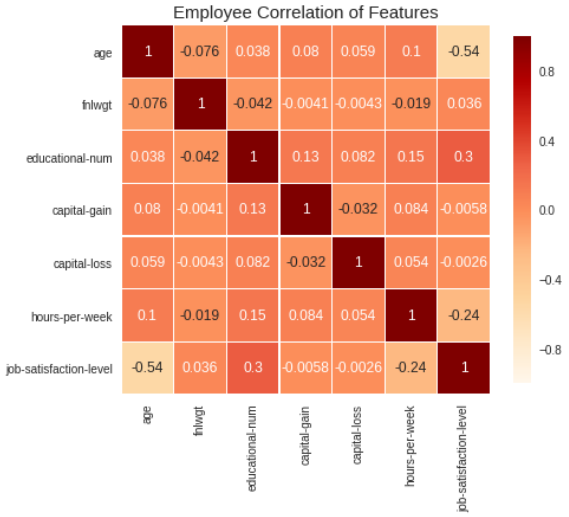

Figure 2 Heat map denoting correlation of dataset columns

The dataset was divided into training, testing and cross-validation subparts for the model training and testing purposes.

C. *Comparing Algorithms*

Linear Regression: This is a supervised method of learning that assumes two variables x as input and y as single out variable and their linear relationship. The linear equation combines a set of input value(x) and arrives at a solution, i.e. the predicted output(y) for x.

Predictions are simple once a linear equation has been generated.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Logistic Regression: This predictive method of analysis is used when the variable that is dependent is binary like yes vs. no, 0 or 1. The major assumptions in this method are that the dependent variable should be dichotomous, no outliers should be present in the data; avoid high correlations among predictors which could be assessed using correlation matrix. The main task is to estimate the log odds of a task.

Linear Discriminant Analysis: Logistic regression method of classification is restricted only to classification of two class problems. Also this method is unstable with well separate classes. LDA method overcomes these problems by considering statistical properties of the data which is calculated for every class. Multivariate Gaussian properties like mean and covariance matrix are estimated from the data and plugged into linear discriminant equation to get predictions.

With assumptions like Gaussian data and same variance for each attribute, model estimates the variance and mean for each class.

The mean (m) is calculated for each input (x) for each class (c) as

$$m=1/nc * sum(x) \quad (1)$$

and variance (s^2) across all input (x), with instances n and classes c

$$s^2=1/(n-c) * sum((x-m)^2) \quad (2)$$

The model makes predictions using Bayes' Theorem based upon the probability that every new input that arrives belongs to each class and class with the maximum probability is the predicted output.

K-Nearest Neighbor (KNN): This is a supervised method of learning that can solve both regression and classification problems. This method assumes similar kind of things exists in closer proximity means similar data points are closer to each other. Mathematically this algorithm calculates the distance between data points plotted on a graph like Euclidean distance. Algorithm requires initializing k to selected number of neighbors. Calculate distance and pick first k entries in sorted order. Get labels for the selected entries and calculate *mean* of k labels in case of regression and mode for classification. The major task is to choose the right value for k. We run the algorithm several times with variant values for k. The k that reduces errors while implementation and predicts accurately for the data seen before.

Classification and Regression Trees (CART): The method represents a binary tree generated using algorithms and data structures. Each root node represents input (x) and leaf nodes a new set of data is given to the machine is its uses tree to represents predicted output (y).The tree is partitioned into k dimensional spaces where each input variable is a dimension. When new data is tested, it is traversed and lands the data to an equivalent dimension and the output to that dimension is the predicted output.

Tree is constructed using Greedy Splitting by calculating Gini Impurity (G) for each dimension

$$G=1-(p \text{ of yes})^2 - (p \text{ of no})^2 \quad (1)$$

where p is the probability of occurrence of an event.

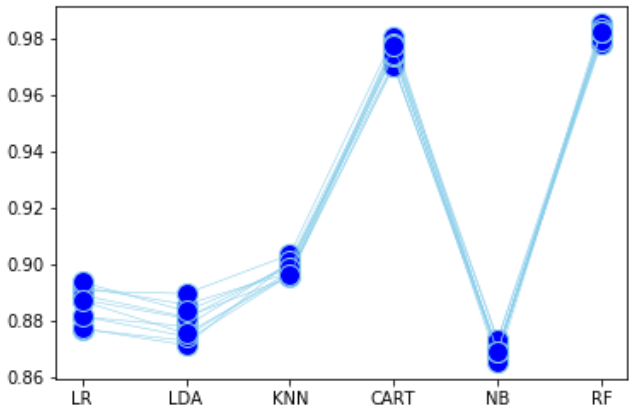

Figure 3 Chart showing results of accuracy of various algorithms used in percentage

D. *Final Algorithm*

Machine learning has many powerful algorithms to be used to predict data of the concerned type. After using the above mentioned algorithms we got the best result with Random Forest Classifiers. Random Forest is an algorithm based on decision trees.

A decision tree is a tree like structure wherein each node of a tree is a condition, and its branches are the possible outcomes. For a given set of values, the conditions are checked starting from the root node, and the tree is traversed till the leaf node is reached. The leaf node contains the final outcome which is considered for the final decision.

As the name suggests, Random Forest is a collection of many decision trees, randomly generated by a process called 'bootstrapping' from the given dataset. Bootstrapping chooses from the dataset a random set of rows, and processes it through the decision tree to obtain a result set. Many such combinations of random set of rows are taken, which result into different result sets. The average of these result sets is taken to get the final result.

The final result thus obtained is called as the result from the random forest algorithm.

IV. RESULT

After multiple trials with the data and feature selection, as well as the train-test data ratio, the model ultimately predicts the correct job-satisfaction level.

Decision tree classifiers follow closely behind in terms of the accuracy score, but the random forests have an added advantage over the decision trees that they are trained on various models of decision trees, taking into account different samples each time. This leads the random forest to have a better control over overfitting.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

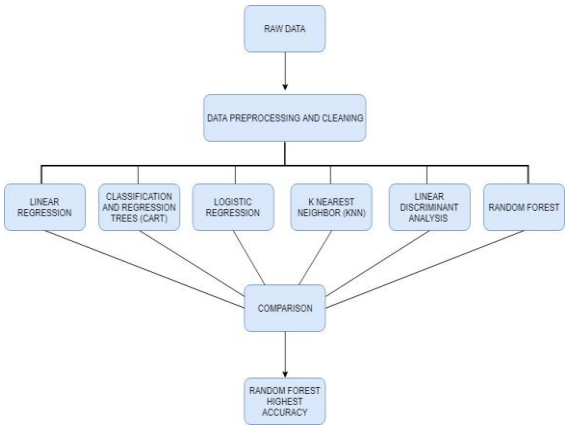

Figure 4 A flowchart describing the process of data analysis for job satisfaction level

V. CONCLUSION

According to the data that we used, age and income are the primary determinants to the level of job satisfaction in an individual working for any company. Large amount of research has been conducted in the area which proves the proportional relationship between the job satisfaction levels of an employee and the productivity he/she gives to the company. This kind of analysis can help the companies manage their human resources better by ensuring better output of the employees, which can be achieved by molding their work conditions according to the requirements. The best features according to our models were age, hours of work and income. The model hereby developed can be used for the companies' utilities, wherein the best features can be made mandatory as inputs of the data pertaining to their employees and other features, which also help determine the job satisfaction level, as optional. Such a model has the potential for a huge market, where the industries are becoming employee friendly, and the emotions and needs of the employees are increasing in ranks in the priorities of any successful organization.

REFERENCES

1. Armstrong, M. (2006) A Handbook of Human Resource Management Practice.
2. Christen, M., Iyer, G. and Soberman, D. (2006). Job Satisfaction, Job Performance, and Effort: A Reexamination Using Agency Theory, Journal of Marketing, Januaryr, Vol. 70, pp. 137-150
3. Christen, M., Iyer, G. and Soberman, D. (2006). Job Satisfaction, Job Performance, and Effort: A Reexamination Using Agency Theory, Vol. 70, pp. 137-150
4. Hoppock, R. (1935). Job Satisfaction, Harper and Brothers, New York, p. 47
5. Lawler, E.E. III and Porter, L.W. (1967). The Effect of Performance on Job Satisfaction, Industrial Relations, pp. 20-28
6. Locke, E.A. and Latham, G.P. (1990). A theory of goal setting and task performance, Prentice Hall, p.4
7. Spector, P.E. (1997). Job satisfaction: Application, assessment, causes and consequences, Thousand Oaks, CA, Sage Publications, Inc
8. Vroom, V.H. (1964). Work and motivation, John Wiley and Sons, New York, p.99

AUTHORS PROFILE

Jackulin Mahariba A. is an Assistant Professor (Sr. G) of Computer Science and Engineering at SRM Institute of Science and Technology. Her areas of interest include algorithm design and analysis, data structures and machine learning.

Sparsh Heda is an UG student of Computer Science and Engineering at SRM Institute of Science and Technology. His areas of interest include machine learning, and coding in Python.

Tanvi Mahajan is an UG student of Computer Science and Engineering at SRM Institute of Science and Technology. Her areas of interest include data analytics and web development.