Prediction of Employee Performance using Machine Learning Techniques

Anu Singh Lather
Delhi Technological University
Delhi India
anusinghlather@gmail.com

Ruchika Malhotra
Department of CSE
Delhi Technological University
Delhi India
ruchikamalhotra2004@yahoo.com

Priya Saloni
Department of CSE
Delhi Technological University
Delhi India
salonipriya.9@gmail.com

Prabhjot Singh
Department of CSE
Delhi Technological University
Delhi India
prabhjot.dua23@gmail.com

Sarthak Mittal
Department of CSE
Delhi Technological University
Delhi India
smittal.sa5@gmail.com

ABSTRACT

Any business's success depends on its employees. Businesses that realize this are concerned about employee output and productivity. Productivity has a compounding effect at the different levels in the workplace, meaning that high productivity at a lower level of organization paves way for higher productivity at the higher levels of the organization. Hence, analysis of performance of employees in any organization is the need of the hour. Performance of an employee cannot be attributed to any fixed quality. Different people have different skill sets and different behavioral characteristics. Thus, performance analysis requires data to be gathered from all walks of life.

The purpose of this paper is to analyze and predict the *performance* of employees in an organization on the basis of various factors, including, but not limited to, individual and domain specific characteristics, nature and level of schooling, socioeconomic status and different *psychological* factors.

This research paper uses Supervised learning techniques namely Support Vector Machines, Random Forest, Naive Bayes, Neural Networks and Logistic Regression which considers these factors and provides insights into the performance and commitment of employees. The employees are classified into 3 output classes indicating the level of their performance from low to high.

In this research paper, 10-fold validation technique is used to ensure the correctness of the prediction by the above-mentioned techniques. Support Vector Machines prove to be the most efficient in terms of accuracy. The result is accentuated by the high validation score obtained by the same.

Also, Employee's Creativity stands out as the most impactful feature.

AISS 2019, November 15–17, 2019, Singapore, Singapore

ACM ISBN 978-1-4503-7291-6/19/11…$15.00
https://doi.org/10.1145/3373477.3373696

CSS CONCEPTS

• **Computing methodologies~Supervised learning by classification** • Social and professional topics~Employment issues

KEYWORDS

Employee Performance, Supervised Learning, ATTA, TKi, FIRO-B, Motivation Analysis Test (MAT)

ACM Reference format:

Anu Singh Lather, Ruchika Malhotra, Priya Saloni, Prabhjot Singh and Sarthak Mittal. 2019. Prediction of Employee Performance using Machine Learning Techniques. In *Proceedings of 2019 International Conference on Advanced Information Science and System (AISS'19). Singapore, 6 pages.* https://doi.org/10.1145/3373477.3373696

1 Introduction

What is employee performance? How to measure employee performance? What are the factors which influence employee's performance and commitment the most? Can we ensure high performing employees?

Any company or organization is only as good as its employees. Hence, the performance of employees is directly related to the growth of the company. Although, the skillset of the employees can be tested at the time of recruitment, the level of performance can only be measured after his/her work in the company. This poses a great challenge to the human resources as to which person to hire and which team to place him/her in.

[1] There have been work predicting the human capability in the software industry on the basis of educational qualifications, technical expertise, personal characteristics etc. but the performance of personnel can depend upon various psychological factors as well as the level of creativity of a particular person. These factors are not taken into account in the numerous studies in this area. Also, the domain considered in the study does not span the various sectors of employment.

In this paper, we aim to understand the effect of psychological, socioeconomic and creativity factors on the performance and commitment of employees. The training data is

prepared with the help of standard tests which the employees took. These tests quantify various aspects of the employees' personalities. We use a number of supervised machine learning algorithms with performance and commitment ratings given by their respective managers as target variables for the training data. This training data contains responses of employees working in different sectors of the industry at different levels of management. This helps us build a robust system capable of predicting beforehand how an employee is most likely to perform in a new environment.
[1] The estimated performance and commitment indices thus obtained can serve as the basis for the decision of whom to hire and which type of project to place him/her in.
The rest of the paper contains four more sections: Section 2 encompasses the experimental setup. Section 3 gives an overview of the various analysis techniques used. Section 4 presents the results obtained as well as the analysis of results. Section 5 concludes the research paper and analyses the result.

2 Experimental Design

2.1 Hypothesis Formation

- The performance of an employee is affected only by the aforementioned variables.
- The performance and commitment ratings of the personnel given by their respective managers are unbiased.
- The candidates have responded to the questionnaires truthfully and correctly.
- The results of all the tests are accurate.

2.2 Empirical Data Collection

Data is collected from varied sectors like Automobile, Information Technology, Fast Moving Consumer Goods and Petroleum from employees working at different levels of management.
The individual, domain specific and socio-economic details are filled by the respondents. The variables are:

1. Individual Characteristics
    a. Gender
    b. Age Group
    c. Level of Education
    d. Marital Status
2. Domain Specific Characteristics
    a. Industry of employment
    b. Sector of employment
    c. Years of experience
    d. Level of management
3. Social and Economic Factors
    a. Socio - Economic Status
    b. Area of Location
    c. Income Group
    d. Family Type
    e. Type of Schooling
    f. Nature of School Studied at

For the collection of personality related factors, we have conducted various tests which are summarized below.

1. **Abbreviated Torrance Test for Adults (ATTA):** [3] ATTA is a shortened version of the Torrance Test of Creative Thinking. The former measures the creative thinking capabilities of adults on the basis of three activities (one verbal and two figural). These activities quantify the creativity of an individual in 18 creative thinking skills contained in the Torrance Incubation Model. They are divided into and
    a. Norm-referenced measures
        i. Fluency
        ii. Originality
        iii. Elaboration
        iv. Flexibility
    b. Criterion-referenced measures
        i. Verbal Responses
        ii. Figural Responses
2. **Thomas - Kilmann Conflict Mode Instrument (TKI):** [4] The (TKI) evaluates an employee's behavior in conflict situations. In the case of a conflict, a person's behavior can be described by two attributes
    a. Assertiveness (A), the degree to which the person endeavors to satiate his or her own interests, and
    b. Cooperativeness (C), the degree to which the person endeavors to satiate the other person's interests.

    5 conflict modes are possible depending on the value of these attributes, namely
    a. Collaborating Style: High A and High C
    b. Competing Style: High A and Low C
    c. Compromising Style: Medium A and Medium C
    d. Accommodating Style: Low A and High C
    e. Avoiding Style: Low A and Low C

    The raw score on any particular conflict-handling mode is equal to the frequency with which one chooses a TKI statement for that mode. The percentile scores are then scaled from 1 to 10.
3. **Fundamental Interpersonal Relations Orientation - Behavior (FIRO-B):** FIRO - B measures three needs namely
    a. Inclusion: This refers to the need to be included in activities
    b. Control: This refers to the need to maintain a balance of power and influence in work relationships
    c. Affection: Refers to the need to form close personal relationships with others.

    Each of these attributes is measured on two dimensions:
    a. Expressed: What one does with regard to others
    b. Wanted: What one expects others to do
4. **Motivational Analysis Test (MAT):** [5] MAT measures the motivation of employees on ten dimensions, in which five are ergs and five are sentiments.
    a. Ergs
        i. Mating Erg
        ii. Assertiveness Erg

iii. Fear (Escape) Erg
iv. Narcissism Comfort Erg
v. Pugnacity - sadism Erg

b. Sentiments
i. Self - Concept Sentiment
ii. Superego Sentiment
iii. Career Sentiment
iv. Sweetheart - Spouse Sentiment
v. Home - parental Sentiment

These tests fetched creative thinking abilities, conflict resolution style, decision making capabilities and Interpersonal Relation orientation and Motivation of employees.

The target variables, i.e., performance and commitment for the dataset are provided by their immediate supervisors on a scale of 0-5.

3 Research Methodology

3.1 Classification Procedure

Figure 1 shows the working of the classification engine.

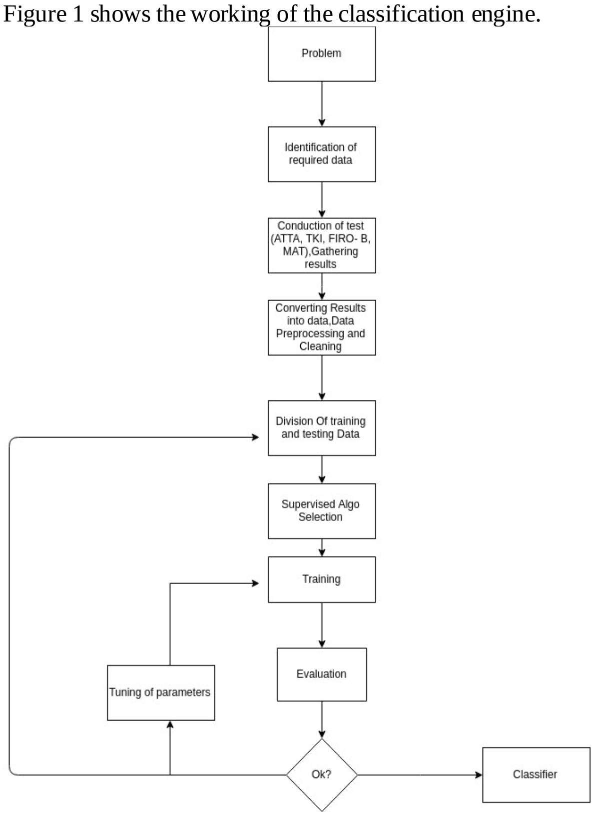


Figure 1: Performance analysis flow diagram

3.2 Data Analysis Techniques

Machine learning develops computer programs that access data and use it to train for themselves.

This research paper deals with supervised learning techniques:

3.2.1 Naive Bayes. According to Bayes' Theorem:
For a given data (t), best hypothesis(s) can be selected in the following manner:

P(t|s) = (P(t|s) * P(s)) / P(t)

Where

- **P(s|t)** - probability of hypothesis s given the data t. This is popularly known as the posterior probability.
- **P(t|s)**-given that the hypothesis s was true, probability of data t.
- **P(s)**- regardless of the data, the probability of the hypothesis s being true. This is popularly known as prior probability of s.
- **P(t)**-regardless of the hypothesis, the probability of the data.

We select the hypothesis with the highest probability and it is called the maximum a posteriori (MAP) hypothesis

MAP(s) = max(P(s|e))

or

MAP(s) = max((P(e|s) * P(s)) / P(e))

or

MAP(s) = max(P(e|s) * P(e))

Naive Bayes works on the assumption that unlike most of the real data, features are independent of each other. Hence final probability is calculated as P(e1|t) * P(e2|T) and so on.

3.2.2 Random Forest. Firstly, many decisions trees are created in Random Forest.

A new class is classified from an input data by putting the input data down each of the decision trees. A classification is given by each decision tree called "votes" for that class. Random Forest picks the class with maximum votes.

When the training data for the current decision tree is extracted by sampling with replacement, almost one-third of data are left out. This oob (out-of-bag) data gives a current impartial calculated value of the classification error as more decision trees are created in the forest. It is also giving estimates of feature importance.

After creation of each decision tree, data is fed down this tree, and we calculate proximities for every pair of cases. If two cases tend to occupy the exact terminal node, their proximity is added by one. At the end, these proximities are divided by the number of total decision trees to get normalized value.

3.2.3 Support Vector Machine. Support Vector Machine (SVM) is a machine learning algorithm which comes under supervised techniques. It deals with both classification or regression problems. In this research paper it is used to work on a classification problem. A hyperplane is created to differentiate between two classes, each coordinate denotes a feature. N-dimensional space created is a region for plotting each data point from dataset where n stands for the number of features in the dataset.

3.2.4 Neural Network. Multi-layer Perceptron (MLP) is a machine learning algorithm which comes under supervised learning techniques. Its basis is feedforward artificial neural network that learns a nonlinear function model by training on a dataset $f(\cdot): R^m \rightarrow R^o$. Here, M is the total number of features in input, O is number of features in output. For a given set of output Z and features $X = x_1, x_2, ..., x_m$, it can adapt a nonlinear function and approximator for either classification or regression. Neural Network has one or more non-linear layers, known as hidden layers between input and output unlike Logistic Regression.

3.2.5 Logistic Regression. This is a statistical method which analyses a dataset which has one or more independent variable. Logistic Regression predicts the outcome with a binary variable. In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.). The goal of logistic regression is to create a numerical equation by assigning weights to each feature in the dataset.

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

4 Results and Discussions

In this section, results of various classifiers are tabulated. In this research paper, hold-out validation method (85-15 ratio) and 10-fold validations is used. Table 1 encapsulates results of the two validation methods for different classifiers.

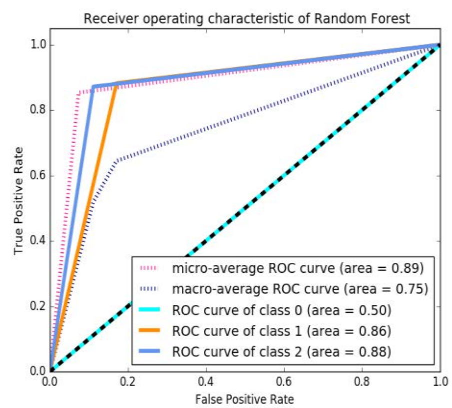
Area under Receiver Operating Characteristics (ROC) curve is obtained by plotting TPR against FPR. A good measure of the accuracy of the used classifier can be obtained through the area under these plotted graphs. The area can attain maximum value of 1 Sq. unit (maximum accuracy) and minimum value of 0 (minimum accuracy). Since Area-Under-ROC-curve considers, both FPR and TPR values, this metric is adopted over other parameters to compare accuracies between models. Following are the meaning of variables used:

- True positive = accurately predicted
- False positive = inaccurately predicted
- True negative = accurately discarded
- False negative = inaccurately discarded

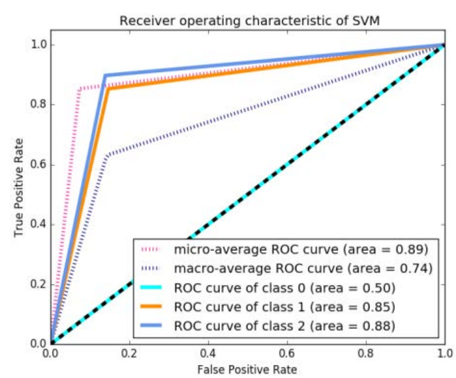
Table 1: Accuracy Scores of respective classification algorithm

Model	**85-15 split**	**10-fold cross validation**
Random Forest	0.853	0.799
Support Vector Machine	0.853	0.826
Neural Network	0.840	0.838
Logistic Regression	0.800	0.740
Naive Bayes	0.680	0.683

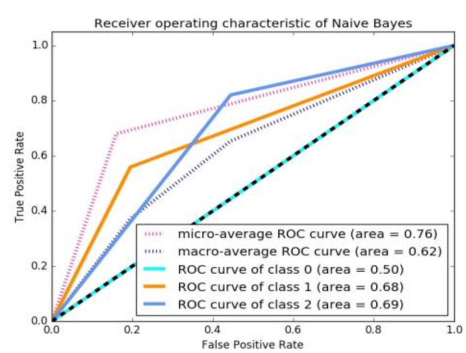
85-15 split is done to ensure we have all output classes in all folds else precision recall and k fold gives bad results due to less data points. From table 1, the Support Vector Machine outperforms the other four classifiers.



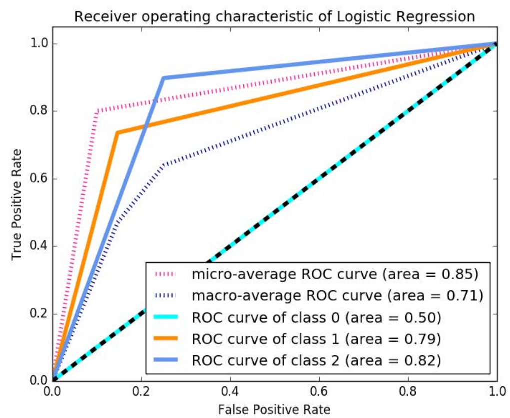
(i)



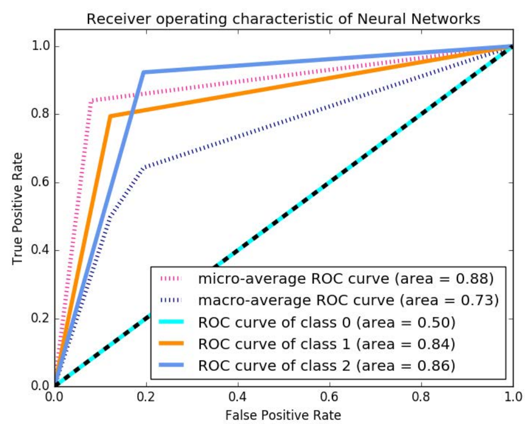
(ii)



(iii)



(iv)



(v)

Figure 2: Receiver Operating Characteristic curves for (i) Random Forest (ii) SVM (iii) Naive Bayes (iv) Logistic regression (v) Neural Network

Table 2: Class Label and the respective classes they Represent

Class Label	Subpopulation
0	Low
1	Medium
2	High

The Receiver Operating Characteristic curve is a metric used to assess the efficiency of each class of a classifier. Figure 2 shows the ROC characteristics of the 5 different classifiers.The advantage of using this metric is that it is unaffected by the imbalanced dataset.

Table 2 represents the class labels along with their respective classes. The 3 classes denote the level of performance of the employees. The Ranges of the mentioned labels were obtained by K-Means Clustering Algorithm, wherein the optimal number of classes was estimated using an elbow plot.

The quality of prediction of data points in a class is determined by the ROC curve. The farther the graph from the line at 45 degrees, the more effective the prediction. Hence, the higher the Area Under the Curve (ranging from 0 to1), the better the results.

Table 3 enlists the micro-average and macro average AUC values for the different models used. It must be noted that the area of 0.5 for class 0 (which does not provide any insights into the data) stems from the fact that the original dataset consists of very less data points and hence this can be ignored.

Table 3. AUC for applied supervised models

Model	Micro-average AUC	Macro-average AUC
Naive Bayes	0.76	0.62
Random Forest	0.89	0.75
Support Vector Machine	0.89	0.74
Neural Network	0.88	0.73
Logistic Regression	0.85	0.71

Based on AUC, it is observed that Random Forest model gives better performance than the other four classifiers (Figure 2 (i)).

Precision is the percentage of positive predictions that are correctly predicted. Mathematically,

$$P = \frac{T_p}{T_p+F_p}$$

Recall is the percentage of positive cases retrieved out of the total positive cases.

$$R = \frac{T_p}{T_p+F_n}$$

In a multi-class classification, precision and recall are calculated for each class and the results are averaged to get the metrics as shown in Table 4.

Table 4: Average Precision and Recall for applied supervised models

Model	Recall	Precision
Random Forest	0.85	0.87
Support Vector Machine	0.85	0.87
Neural Network	0.84	0.86

Logistic Regression	0.80	0.83
Naive Bayes	0.68	0.72

Feature Importance represents what effect each variable has on the training model as shown in Table 5.

Table 5: Feature Importance Table for Random Forest

Model	**Feature Importance**
Total Creativity	20.45%
Collaborating Score	9.06%
Total Need Score	8.83%
Commitment	8.49%

5 Interpretation

We observe that total creativity, collaborating score, total need score and commitment play a crucial role in predicting the performance of employees. Creativity, Collaborating Score and Need Score of an employee are measured by ATTA, TKI and FIRO-B tests respectively as explained in section 2.1. Commitment on the other hand is evaluated using a questionnaire which might differ from organization to organization.

From further analysis, we observe that the variables hold a direct relationship with the performance of an employee. Creativity, being the most prominent feature signifies the importance of thinking out of the box at the workplace. Collaborating mode of conflict handling is the mode wherein the employee is concerned about the interests of both, self and others. The results suggest that a high collaborating score leads to high performance. Other modes of conflict handling have a relatively lower effect on performance thus highlighting the efficacy of the collaborating style of behavior during conflicts. The need score brings to light the importance of the fulfilment of social needs of an employee vis-à-vis performance.

6 Conclusion

Performance analysis of employees is important for the growth of the organization as well as the individual. Since manual analysis of performance is a dull and tiresome work for the Human Resources. Apart from that, the process is highly disorganized and hence would benefit from automation. This can also be used in fair evaluation of employees. The performance score obtained from this model can be used to verify the ratings given by the managers and can thus prevent any biases in the workplace.

We have conducted various tests to gather insights into features that can impact employees’ efficiency at workplace. Further, data is preprocessed, and machine learning techniques are applied. Receiver Operating Characteristic (ROC) curves and scatter plots are used to analyze and visualize the result. Support Vector Machines prove to be the most efficient in terms of accuracy.

We aim to lay the foundation in the work of human capability prediction which can act as a stepping ground for more optimized automated systems.

7 REFERENCES

[1] G. S. Thakur, A. Gupta, and S. Gupta, “Data Mining for Prediction of Human Performance Capability in the Software Industry,” International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 53–64, 2015.

[2] Gupta, Sangita & V, Suma. (2014). Empirical Study on Selection of Team Members for Software Projects - Data Mining Approach.

[3] Trimadona B. Wiratrisna, “Using Abbreviated Torrance Test For Adults (Atta) To Measure Creativity Training Effectiveness in The Indonesian Context,” Credo Foundation-UNESCO Journals, August 2016.

[4] Thomas, K. and Kilmann, R. (2008). *Thomas-Kilmann Conflict Mode Instrument*. [online] Xicom, Incorporated. Available at: http://kilmanndiagnostics.com/wp-content/uploads/2018/03/TKI_Sample_Report.pdf [Accessed 10 Dec. 2019].

[5] Bernard, L., Walsh, R. and Mills, M. (2005). The Motivation Analysis Test: An Historical and Contemporary Evaluation. Psychological Reports, 96(2), pp.464-492.