Predictive Model on Employability of Applicants and Job Hopping using Machine Learning

Neeraj Khadilkar
Research Scholar
Department of CSE-IT
Vishwakarma Institute of Technology,
Pune, India

Deepali Joshi
Assistant Professor
Department of CSE-IT
Vishwakarma Institute of Technology,
Pune, India

ABSTRACT
The rate of rejection of various candidates is rising in spite of abundant job openings. Employability is a set of achievements, understandings and personal attributes that make individuals more likely to gain employment and to be successful in their chosen occupations. It is the ability of the candidate to check whether he/she is capable to gain employment or not. This is an automated effort to predict whether a person is employable or needs more training. This would help in the institutions to assess whether they are producing employable students or not, also this would provide a support for organizations in screening bundles of applications and finding the most suitable ones. Job hopping is another open problem in the industry wherein lots of efforts and resources are invested in the hiring process as well as in grooming the employee .It would be a great help to the employers if they are provided with tool which can predict the job hopping of the employees so that the managers can be prepared for the same. We are providing a solution to both the above stated problems in a novel way. In screening the resumes we use text mining and appropriate weighing in addition to personal attributes of the candidate which helps in improving accuracy of the system also our job hopping module is novice in terms of using text understanding from reviews about the company for hopping prediction. For employability prediction we got the highest accuracy for naïve based with 89% and for predicting whether the employee's going to quit the job or not we got the highest accuracy for decision tree with 85%.

Keywords
Employability prediction model, Machine learning, feature extraction, sentiment analysis, Classifiers.

1. INTRODUCTION
Current scenario states that around thousands of applicants are applying for one job. As per reports only a small proportion of Indian graduates are considered employable. This reflects in the fact that placement outcomes drop significantly as we move away from top tier institutions. Educational institutions train millions of youngsters but corporates often complain that they do not get the necessary skill and talent required for a job. According to Aspiring Minds National Employability Report, which is based on a study of more than 1, 50,000 engineering students who graduated in 2015 from over 650 colleges, 80% of them are unemployable and total 95% of the engineering graduates were unfit for programming jobs [15]

There is a process for hiring candidates such as Aptitude Test ,Interviews but when there are bulk of candidates then it becomes very time consuming and infeasible. The employability skills are necessary for jobs .there are various reasons we require Employability prediction model because it is time efficient, and for employer perspective it is very important thing to know whether the employee is going to quit the job or not and whether the employee is capable to gain the employment.

Existing system presents an empirical study that compares varied classification algorithms on two datasets of MCA (Masters in Computer Applications) students collected from various affiliated colleges of a reputed state university in India. One dataset includes only primary attributes, whereas other dataset is feeded with secondary psychometric attributes in it. The results showcase that solely primary academic attributes don't lead to smart prediction accuracy of students' employability, once they square measure within the initial year of their education. The study analyses and stresses the role of secondary psychometric attributes for better prediction accuracy and analysis of students 'performance. Timely prediction and analysis of students' performance can help Management, Teachers and Students to work on their gray areas for better results and employment opportunities.

In machine learning, the classifier is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. We are using several classifiers such as Decision tree, Random forest, Naïve Bayes and K-nearest neighbors approach.

2. LITERATURE REVIEW
Keno C. Piad, Menchita Dumlao et .al used the techniques called decision tree and naïve based to predict the employability of the IT graduates using various variables with maximum accuracy of 78.4 for logistic regression is implemented. The data was collected based on the five years of study of 515 students. In this the number of samples taken for data set is less. In the next approach**,** Qasem A. Al-Radaideh presente**d a** methodology where data mining techniques were utilized to build the classification model where CRISP-DM technology was adopted. Here decision tree algorithm is used which is very easy to understand for human but here only historical data is collected, it predicts the employability of graduates using nine variables.

Pooja thapar et al proposed methodology where several techniques such as multilayer perceptron-NN, J-48 and Random forest were used. This paper presents an empirical study that compares varied classification algorithms on two datasets of MCA (Masters in Computer Applications) students collected from various affiliated colleges of a reputed state university in India. One dataset includes only primary attributes, whereas other dataset is feeded with secondary psychometric attributes in it. The results showcase that solely primary academic attributes don't lead to smart prediction accuracy of students' employability, once they square

measure within the initial year of their education. The study analyses and stresses the role of secondary psychometric attributes for better prediction accuracy and analysis of students 'performance. Timely prediction and analysis of students' performance can help Management, Teachers and Students to work on their gray areas for better results and employment opportunities Data set can be improved by adding the significant attributes to enhance the model accuracy and other performance metrics.

Tripti Mishra et al proposed techniques such as J-48, C 4.5 Decision tree. the paper uses various classification techniques of data mining, like Bayesian methods, Multilayer Perceptron's and Sequential Minimal Optimization (SMO), Ensemble Methods and Decision Trees, to predict thee employability of Master of Computer Applications (MCA) students and find the algorithm which is best suited for this problem. For this purpose, a data set is developed with the traditional parameters like socioeconomic conditions, academic performance and some additional emotional skill parameters. A comparative analysis concludes that J48 (apruned C4. 5 decision tree) is most suitable for employability prediction with 70. 19% accuracy, easy interpretation and model building time(0. 02Sec) less than Random Forest, which has slightly better prediction accuracy (71. 30%),higher building time(0. 11) and difficult interpretation. Further, Empathy, Drive and Stress Management abilities are found to be the major emotional parameters that affect employability. As the paper discusses classifier couldn't attain high percentage of accuracy, and this research work has considered only MCA students.

3. METHODOLOGY

The model was divided in two modules first to predict the employability of the applicant and another is to predict the employee of the organization is going to quit the job or not.

1. Data collection phase: In employability prediction we select the sample resumes from that features were extracted such as prev_company, no_of_projects, certifications, achievements, publications, GPA (marks).the csv file was created and rank was calculated as algorithm described below. The rank played important role for resume ranking.

   For employability prediction we created CSV file of applicants where for each applicant, separately rank was calculated and data was labalised

   For predicting the employee's going to quit the job or not we required the various attributes such as employee reviews about their company, positive, negative, neutral reviews, salary, no_of_projects, year_of_experience, satisfaction level. The satisfaction level is calculated below. That played important role for making predictions. For that we created csv file of Employees.

2. Feature Vector selection: As we collected the data for employability prediction we extracted the features for data sets from resume using Regular Expressions and natural language processing.

   For predicting the employee's going to quit the job or not we used the employees review and we used the sentiment analysis

3. Implementing the classification Algorithms: we have used the Decision tree, Random forest, K-NN, Naïve Based approach for making predictions. And estimating the accuracy. The classification generally performed in two stages one is training and another is testing set. Let's consider the algorithms we used for classification.

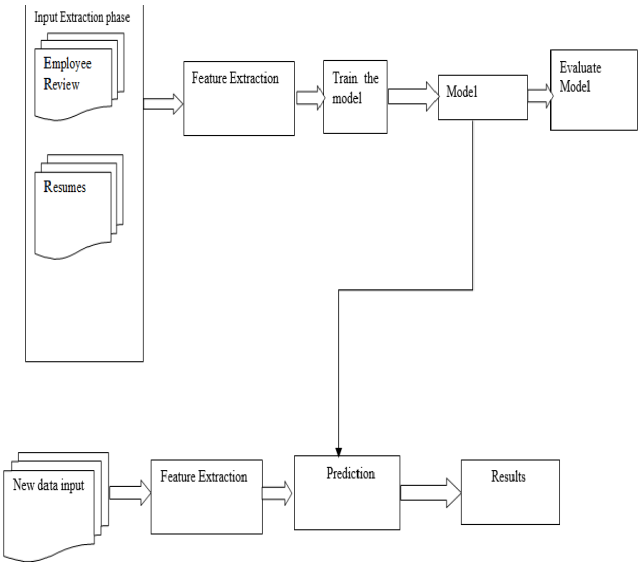


Fig1. System diagram for employability prediction and job hopping

3.1 K-Nearest Neighbors

It can be applied to the data from any distributions. The algorithm is very simple and it gives the good classification for large number of samples.

An object (new instance) is classified by the majority votes for its neighbor classes the object is assigned to the most common class

- Calculate the distance between new example (E) and all examples in the training set.
- Euclidean distance between two examples.
  - X = [x1,x2,x3,..,x_n]
  - Y = [y1,y2,y3,...,y_n]
  - The Euclidean distance between X and Y is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

- All the instances correspond to points in an n-dimensional feature space.
- Each instance is represented with a set of numerical attributes.
- Each of the training data consists of a set of vectors and a class label associated with each vector.
- Classification is done by comparing feature vectors of different K nearest points.
- Select the K-nearest examples to E in the training set.
- Assign E to the most common class among its K-nearest neighbors.

3.2 Random Forest algorithm:

1. A random seed is chosen which pulls out at random a collection of samples from the training dataset while maintaining the class distribution.

2. With this selected data set, a random set of attributes from the original data set is chosen based on user defined values. All the input variables are not considered because of enormous computation and high chances of over fitting.
3. In a dataset where M is the total number of input attributes in the dataset, only R attributes are chosen at random for each tree where R< M.
4. The attributes from this set creates the best possible split using the Gini index to develop a decision tree model. The process repeats for each of the branches until the termination condition stating that leaves are the nodes that are too small to split.

3.3 Decision Tree Classification:

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

1. **Categorical Variable Decision Tree:** Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example: - In given scenario of Employability prediction problem, where the target variable was "Applicant is employable or not i.e. YES or NO.
2. **Continuous Variable Decision Tree:** Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

   The various applications where the decision tree is used are medicine, production, Biomedical engineering, Pharmacology etc. [16]

3.4 Rank estimation using Resumes:

1. Select resumes
2. Define variable resume count
3. Feature Extraction for data set attributes
4. Estimate threshold value for each feature

   If feature value<= threshold

   Increment the range by lesser resume count

   Else

   Increment the range by greater resume count

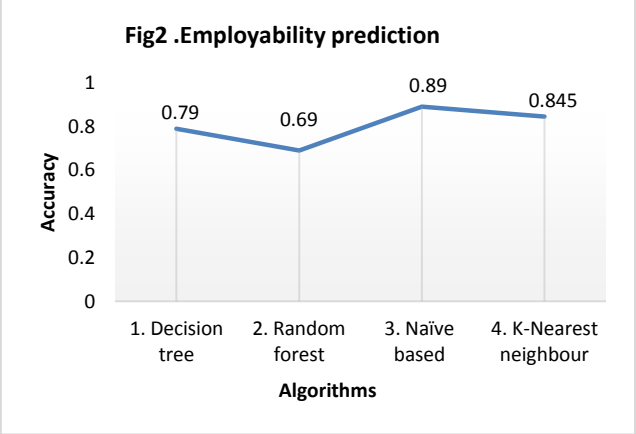
1. Add the total count which will be our rank for the resume
2. Add the rank in our data set for making predictions

3.5 Predicting whether the Employee's going to quit the job or not

1. Select the reviews of employee about their companies
2. Take the sentiment analysis and divide the positive, negative and neutral reviews for each review
3. Calculate the satisfaction level

   Satisfaction level = Max (values of positive, negative and neutral reviews) /total number of reviews

4. Add this attribute in our data set for making predictions



For Text mining through text understanding there is process

1. Part of speech (POS) tagging : the word in the text or in the sentence at tagged using POS-TAGGER so it assigns a label to each word
2. For each sentence the total positive and negative words are divided using the positive and negative words dictionary and the score is calculated such as +1 for positive, 0 for neutral and -1 for negative.

3.6 Mathematical Representation

$\{(X1, y1)\ldots (X_n, y_n)\}$== Data set D

Where X_n= feature and Y_i= class, X=new value, T=Threshold for feature

Let rank be the variable and W be the weightage for feature

Δ Be the features selected from resume

C=set of resumes: C= $\{c_1, c_2, c_3 \ldots c_n\}$

For i=1 …Cl where Cl: classifier

For i in C

if $\Delta_I <$T: Assign less value of W to Δ_I

Increment rank by lesser count

Else $\Delta_I >$= T: Assign greater value of W to Δ_I

Increment rank by greater count

Add Δ_I in data set D for classification

Do classification on X

4. EXPERIMENTATION

For Employability prediction the set of resumes are required. And total 5000 sample resumes were collected from [12], [13] and [14]. From set of resumes the features were extracted such as PrevCo, Year_of_exp, GPA, Degrees or Foreign Degrees, No_of_projects, Achievements, Publications, certifications from these csv file was generated.

For predicting job hopping or predicting whether Employee's going to quit the job or not, we collected 100 Employee reviews about their companies from indeed and glass door and their satisfactory_score was calculated from text mining of employee reviews we created csv file of attributes Employee review_id, sentiment values- pos, neg and neutral

values for review, Satisfactory level, Year_of_experience, no_of_projects, salary range.

For each model we have calculated

1. Accuracy
2. Misclassification rate
3. Confusion Matrix
4. TPR, FPR ,F1-score and precision

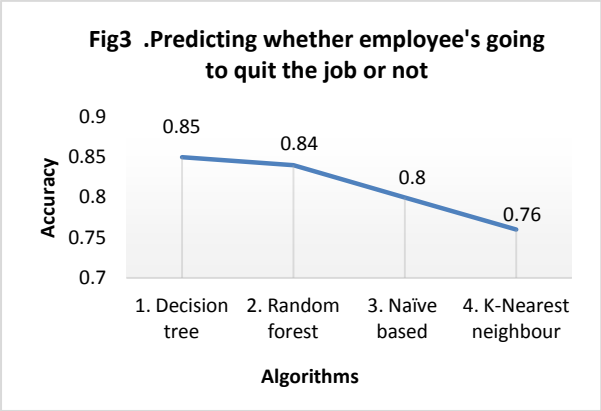
5. RESULTS AND DISCUSSION

There are results of the various classifiers in predicting the employability and to predict whether Employees going to quit the job or not. The following graph reveals the various accuracies of different classifiers. From the graph it reveals that for employability prediction naïve based approach appears to be the highest accuracy with 89% followed by the decision tree with 79%, K-nearest neighbors with 84.5%, random forest appears to be the lowest accuracy with 69%, and for Predicting whether the employee's going to quit the job or not Decision tree appears to be highest accuracy with 85%, Random Forest with 84%, Naïve based approach with 80% and KNN appears to be the lowest accuracy with 76%.

Once the model is formed the confusion matrix was calculated. The confusion matrix gives the better idea for the model we are developing it gives the four values such as TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative).

Condition Positive (p): total no of real positive cases in the data Condition Negative (n): total no of real negative cases in the data Below is the process for calculating a confusion Matrix.

We need a test dataset or a validation dataset with expected Outcome values.



1. Make a prediction for each row in your test dataset.
2. From the expected outcomes and predictions count
 1. The number of correct predictions for each class.
 2. The number of incorrect predictions for each class, organized by the class that was predicted. [10]

Precision is a fraction of retrieved instances that are relevant. Precision can be thought of as a measure of a classifiers exactness. A low precision can also indicate a large number of False Positives.

It is calculated as PRECISION= (TP)/ (TP+FP)

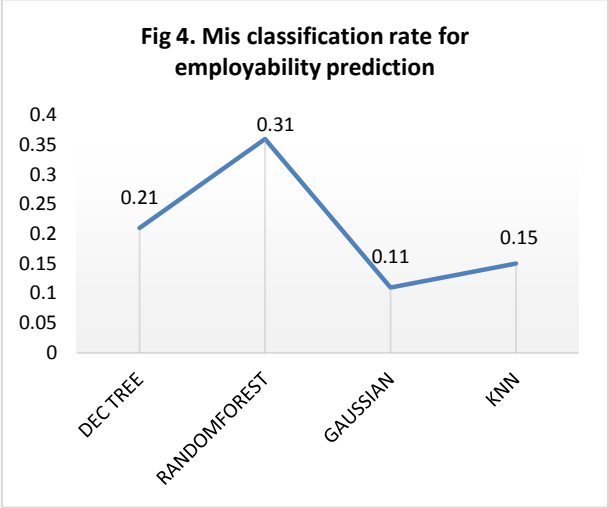
Recall is a fraction of relevant instances that are retrieved. Recall can be thought of as a measure of a classifiers completeness. A low recall indicates many False Negatives. It is usually expressed in percentages and is calculated as RECALL= (TP)/ (TP+FN)

The F1 score conveys the balance between the precision and the recall.

The TPR is calculated as TP/TP+FN and also called as sensitivity measures the proportion of positives that are correctly identified as such.

Where FPR is calculated as FP/FP+TN and also called as the specificity measures the proportion of negatives that are correctly identified as such.

The misclassification rate is calculated as the 1-Accuracy or FP+FN/P+N



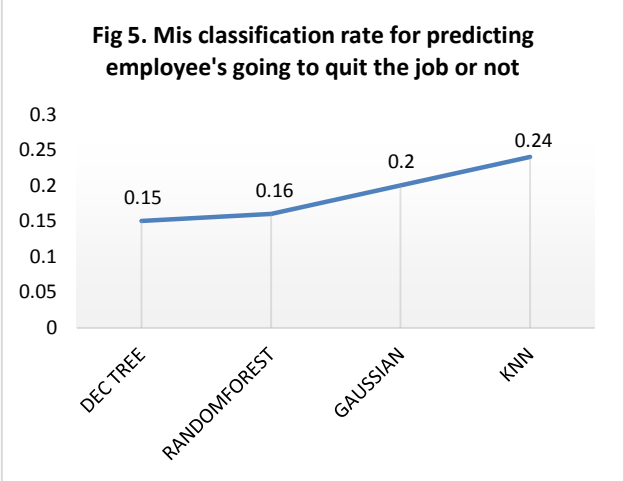


Table1. Employability prediction

Algorithms	TPR	FPR	F1-Score	Precision
1.Decision tree	0.96	0.833	0.87	0.78
2.Random forest	0.96	0.89	0.76	0.645
3.Naïve based	0.97	0.97	0.71	0.939
4.K-Nearest neighbour	0.96	0.81	0.91	0.84

Table 2. Predicting Job hopping

Algorithms	TPR	FPR	F1-Score	Precision
1. Decision tree	0.95	0.84	0.91	0.88
2. Random forest	0.94	0.68	0.9	0.87
3. Naïve based	0.93	0.75	0.88	0.85
4. K-Nearest neighbor	0.93	0.9	0.86	0.76

6. CONCLUSION AND FUTURE WORK

In proposed method we used several classifiers such as decision tree, K-NN, Naïve based approach, Random Forest whereas for employability prediction we got the highest accuracy for naïve based with 89% and lowest accuracy for Random forest with 69% and for predicting whether the employee's going to quit the job or not we got the highest accuracy for decision tree with 85% and lowest accuracy for KNN with 76%.we also calculated the accuracy, precision, F1-score etc.

In future, the data set can be improved by adding the significant attributes to the data sets and for employability prediction we used the resumes in the same format or template. The all resume were only in one format or template so for employability prediction we can use the resume in all formats. It is recommended to collect more proper data from several companies. Databases for current employees and even previous ones can be used, to have a correct performance rate for each one of them.

7. REFERENCES

[1] Tripti Mishra Department of Computer Science, Mewar University, Rajasthan, India. Dharminder Kumar Department of Computer Science, G. J. University, Hisar, Haryana, India. Sangeeta Gupta Department of Management, Guru Nanak Institute of Management, Delhi, India "Students' Employability Prediction Model through Data Mining"

[2] Keno C padd school of the computer studies" Predicting IT Employability Using Data Mining Techniques" http://ieeexplore.ieee.org/document/7529358/

[3] How to predict the employability of IT graduates using a classification algorithm? By Keno Piad, Bulacan State University.

[4] Bangsuk Jantawan, Cheng-Fa Tsai. "The Application of Data Mining to Build Classification Model for Predicting Graduate Employment". (IJCSIS) International Journal of Computer Science and Information Security, Vol. 11, No. 10, October 2013.

[5] Qasem A. Al-Radaideh, Eeman al Nagi "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012

[6] Estimating a prediction model for the early identification of low employability graduates in Malaysia http://repo.uum.edu.my/16583/

[7] In Hiring, Resume Info Could Help Employers Predict Who Will Quit http://www.psychologicalscience.org/index.php/news/minds-business/in-hiring-resume-info-could-help-employers-predict-who-will-quit.html

[8] Predicting Employability from User Personality using Ensemble Modelling http://www.gdeepak.com/thesisme/Sahil%20-%20801333023.pdf

[9] Introduction to Machine Learning https://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937

[10] Data Mining http://www.saedsayad.com/data_mining.htm

[11] What is a Confusion Matrix in Machine Learning http://machinelearningmastery.com/confusion-matrix-machine-learning/

[12] Applying the Decision Tree https://www.boundless.com/management/textbooks/boundless-management-textbook/decision making-10/considering-ethics-in-decision-making-79/applying-the-decision-tree-382-475/

[13] Sample Resumes http://www.jeffthecareercoach.com/sample-resumes/

[14] Circuit Gallery http://www.circuitgallery.com/resumes/

[15] Van Meter Williams Pollack LLP http://www.vmwp.com/resumes/

[16] Over 80% of engineering graduates in India unemployable: Study http://www.gadgetsnow.com/tech-news/Over-80-of-engineering-graduates-in-India-unemployable-Study/articleshow/50704157.cms

[17] A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python) https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/