

A survey on Predictive Analysis in Employment Trends

Nita Radhakrishnan¹, Mehul Awasthi², P. Mahalakshmi³

^{1,2}Student, ³Assistant Professor (O.G)
Department of Computer Science and Engineering,
SRM Institute of Science and Technology

Abstract

This paper addresses the theories of using predictive analysis and Data Mining in arriving at suitable patterns and predicting paths and trends in the current Employment Scenario more specifically to the Engineering sector. India produces 1.5 million engineers every year, and yet there is a significant gap between their skills and the jobs and corresponding salaries they are offered. Recognizing the factors that influence this gap can help us bridge it. The survey shows that the ideal route to doing so, is by employing various Predictive analysis and Data Mining techniques on appropriate data sets, which help in addressing these issues. As per the survey, appropriate visualization techniques have also been used to extract the meaning from the prediction and analysis.

Keywords: Predictive Analysis, Employability, Data Mining.

1. Introduction

To begin with defining predictive analytics, it is the use of data along with statistical and machine learning algorithms to identify patterns and other outcomes that are likely to occur at instances in the future based on historical data. Organizations use predictive analytics to solve complex problems and discover new opportunities. Therefore, applying predictive analytics to solve the problems faced in employment only seems like a logical solution.

As the Indian economy burgeons, the demand for skilled labor is also proportionately rising. According to the National Skill Development Corporation, the growing skills gap in India is estimated to be more than 250 million workers across various sectors by 2022 [1].

It therefore makes sense that the skill gap in labor market is analyzed carefully and appropriate remedial measures are provided where necessary. An approach to analyzing this could involve the evaluation of content in employee profiles and their employability standard. Thus, the concepts of prognostic analysis and data mining can help various users in planning the trajectory of their careers.

1.1 Data Mining

Data mining is a practice where we discover patterns and extract knowledge from an enormous amount of data that often does not make any sense until it is properly cleaned and analyzed appropriately. Data mining techniques have been expansively employed in various popular technological features offered by firms like Facebook, Google and so on. Data mining is based in data handling technology, statistics and probability, information extraction and visualization and artificial intelligence. Data Mining encompasses the concepts of extraction, classification, prediction, analysis and evaluation and so on. [2]

1.2 Predictive Modelling

In Predictive modelling, a model is developed to determine a future outcome of an event. If the outcome is definite and textual then it is called classification but if the outcome is empirical then it is called regression. The two prominent tasks in data mining are Prediction and Classification, and the category of data that is nominal, ordinal, ratio or interval often dictates the algorithm used for the two. [2]

2. Literature Survey

According to the survey, for studies on employment, some of the classification and prediction concepts and algorithms used are, Neural Networks, Tree-Based Classifiers (J48), Random Forest, Bayesian Methods, Logistic regression and advance concepts of Ensemble Learning methods. The subsequent paragraphs also provide enough evidence to the fact that WEKA is used as the primary data mining tool to build predictive models.

The authors of [3] defined a framework to predict the unemployment rate using neural network (NN) and support vector regression (SVR). Through various stages of querying, optimization and validation, it was concluded that a data mining framework that uses NN is a viable method to forecast trends.

To predict the employability of MCA students, Mishra who is the author of [4], used five classification methods. These methods included Bayesian algorithms, Ensemble Learning methods and Decision Trees. Academic stand, emotional quotient and socio-economic conditions were some of the parameters used to build predictive models in the dataset.

A graduate employability model was built by the authors of [5]. Bayesian algorithms and Decision Tree were compared in this paper. It was concluded from this comparison that J48 is a better choice as an algorithm with high level of accuracy, for classification and prediction, as compared to Bayes algorithms.

The authors of [6] proposed a new method called Monte-Carlo Tree Search (MCTS). This method was presented by them in order to reduce computational workload for a system when it came to job and candidate recommendation in the employment network. The authors also claimed that the two obstacles of an enormous amount of data and personalization of different users was made

possible to work through. Their goal is to help create a fast and accurate employment network.

The authors of [7] developed a Grey-Markov forecasting model to predict the unemployment rate of graduates in China. Owing to the fluctuations in data collected for the analysis, the simple Grey model was found to be less efficient, while the Markov model proved to be a better model for forecasting purposes. Therefore, a combination of the two models, called the Grey-Markov model was built and implemented by the authors and was seen to be an efficient and simple forecasting model. It took the employment trends and the random fluctuations in the data into account to make the prediction.

To predict the performance of students, and more specifically their GPA, the authors of [8] carried out a study on first year undergraduate students in the computer science course. Their study involved the use of WEKA tool to build predictive models. The Naive Bayes classifier was found to be the most successful algorithm used for this prediction owing to simple parameters that were considered for this task as well as the simple classification of the class label called "Grade Point Average".

In [9] the authors work on a model in order to enhance the performance of a model that predicts the academic performance of students in a secondary school. The authors do this using data mining techniques. The authors note that education at the lower level is essential to the advancement of any country. Furthermore,

predicting a student's academic performance therefore becomes important in order to ensure their success. This can be done using data mining techniques. The authors realise that the best result was obtained by Naïve Bayes Classification. They further go on to propose that a classifier using Support Vector Machine (SVM).

To study the work in his paper [10], Gao constructed a data mining model using WEKA. Apart from using Standard Decision Tree classifiers for analysis and comparisons, Gao used the information from employment statistics as the attributes.

A Graduates Employment Model was built by the authors of [11] which could predict whether graduates were unemployed, employed or in an undetermined state. The Planning Office Maejo University in Thailand was from where the study used its sample. The various processes and stages involved in KDD and CRISP-DM were employed to make the classifications. The highest accuracy of 99.77% was found by the study, which belonged to the WAODE algorithm belonging to Bayesian methods.

The authors of [12] presented a study that uses a neural network approach to use the unemployment trends to predict future unemployment trends. This study has also implemented concepts of the Box-Jenkins method in order to reduce the dependence on the parameters. This paper has concluded that the semiparametric approach developed by its authors can accurately predict the time-series process as long as its characteristics have been captured correspondingly.

3. Overview of Algorithms Used

ALGORITHM	ABOUT	VERDICT
Neural Networks (NN)	A computational model based on biological neural networks. It consists of a connected group of artificial neurons and processes information using a networked approach. [13]	Successful with high level of accuracy in classification and prediction.
J48	It is based on the concept of C4.5 algorithm which itself is based on the ID3 algorithm. This algorithm is used for statistical classification. Information Gain forms the foundation of this algorithm. By building a decision tree based on various conditions and values, J48 has become one of the most popular algorithms used in WEKA	Due to its precision and speed of calculating the information gain and generating the decision tree, it competes with Neural Networks in being employed for prediction.
Averaged One-Dependence Estimators (AODE)	This is a semi-Naïve Bayesian Technique used for classification. This aggregates the predictions for multiple one-dependence classifiers where all the attributes depend on same single parent attribute as well as class.	Since it is a probabilistic classification learning method, it is more accurate than Naïve Bayes. It also reduces the dependence on attributes. It can be considered for slightly complicated classification.
Logistic Regression	This is a statistical method to analyze a data set in which there are one or more independent variables that determine an outcome, usually one of two possible outcomes.	This is a highly accurate and efficient algorithm which is used to classify objects into binary outcomes. For instance, to predict if an event will work or not, Logistic Regression is used.
Ensemble Learning	It is a machine learning model where many learning modules are trained to solve a single module. This employs a combination of popular algorithms to improve the efficiency of the task. It is a form of supervised learning	Ensemble Learning constructs multiple hypothesis and then learns from that, as opposed to ordinary ML techniques. While its implementation is complex and time consuming, it is very efficient in solving complicated tasks involving large data set with many attributes.

4. Existing Systems Employing Similar Concepts

Over time, there have been multiple attempts at resolving the gap that has been persistent in the employment sector. Some of these attempts have successfully employed one or more of these algorithms in addition to developing a very interactive user interface that makes it easy for users to access the information they need. A few of the most noted tools are as follows.

4.1 Glassdoor

Glassdoor is used by jobseekers and employees alike in order to get an idea of the work culture, salaries, job descriptions etc. in

other companies. The data on Glassdoor is generated via user posts on the website. Users can post reviews of the company they work for anonymously, thus giving scope for honest reviews about a particular company. Information on Glassdoor is thought of as trustworthy by its users and it does indeed give a fair idea about what it would be like to work for a particular company.

4.2 PayScale

PayScale is used by job seekers in order to figure out what kind of compensation they are entitled to given their set of skills and job profile. The way PayScale works is fairly simple. Users of PayScale are required to enlist their professional information such as their current job profile, compensation they receive currently etc. This then enables users to look at other users profiles with

similar skill sets and positions. Therefore, the users get a rough idea of how much they are worth in the current job market.

International Conference on 26-27 Feb. 2015, Publisher: IEEE, DOI:10.1109/ECS.2015.7124833.

4.3 Salary.com

Salary.com has a variety of free salary calculators and tools available for job seekers, employees and employers. It provides details on a jobseeker's worth, benefits package, and so on. Salary.com also provides services wherein an employer can get information that will help them with hiring and setting pay. There are a variety of business tools available for companies to get salary data and information.

5. Conclusion

From the survey conducted it is evident that there exists a pool of Data Mining and Machine Learning Algorithms from which choosing either one or more to make the career path predictions and to provide a futuristic view of how to bridge the gap between skills and the appropriate role assignment based on the accuracy of the result and the parameters considered, would enable us to implement a comprehensive and efficient tool to address the current pressing issues.

References

- [1] "The skills gap and what it means for your business," www.financialexpress.com/article/industry/jobs/the-skills-gap-and-what-it-means-for-your-business/138700, accessed: 2016-01-29.
- [2] "Predictive Analysis and Data Mining among the Employment of Fresh Graduate Students in HEP"
- [3] Nor Azziaty Abdul Rahman, Kian Lam Tan, Chen Kim Lim, <https://doi.org/10.1063/1.5005340>
- [4] Xu, W., Li, Z., Cheng, C., & Zheng, T. (2012). Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications*, 7(1), 33–42. <https://doi.org/10.1007/s11761-012-0122-2>
- [5] Mishra, T. (2016). Students' Employability Prediction Model through Data Mining, 11(4), 2275–2282.
- [6] Sapaat, M. A., Mustapha, A., Ahmad, J., & Chamili, K. (2011). A Data Mining Approach to Construct Graduates Employability Model in Malaysia, 1(4), 1086–1098
- [7] Job and Candidate Recommendation with Big Data Support: A Contextual Online Learning Approach., GLOBECOM 2017 - 2017 IEEE Global Communications Conference
- [8] Xianyin Li, Wanming Chen (2009). A Grey –Markov Prediction for unemployment rate of graduates in China
- [9] Aziz, A. A., Ismail, N. H., Ahmad, F., & Hassan, H. (2015). A framework for students' academic performance analysis using naive bayes classifier. *JurnalTeknologi*, 75(3), 13–19. <https://doi.org/10.11113/jt.v75.5037>
- [10] Student Academic Performance and Social Behavior Predictor using Data Mining Techniques, Computing, Communication and Automation (ICCCA), 2017 International Conference on
- [11] Gao, L. (2015). Analysis of Employment Data Mining for University Student based on Weka Platform, 2(4), 130–133
- [12] Jantawan, B., & Tsai, C. (2013). The Application of Data Mining to Build Classification Model for Predicting Graduate Employment. *International Journal of Computer Science and Information Security*, 11(10), 1–8. <https://doi.org/10.1016/j.bdr.2015.01.001>
- [13] Chih-Chou Chiu' and Chao-Ton Su2. (2014). A Novel Neural Network Model Using Box-Jenkins Technique and Response Surface Methodology to Predict Unemployment Rate
- [14] Dr. Yashpal Singh, Alok Singh Chauhan, Neural Networks in Data Mining.
- [15] T. Padmapriya and V. Saminadan, "Inter-cell Load Balancing Technique for Multi- class Traffic in MIMO - LTE - A Networks", International Conference on Advanced Computer Science and Information Technology , Singapore, vol.3, no.8, July 2015.
- [16] S.V.Manikanthan and K.srividhya "An Android based secure access control using ARM and cloud computing", Published in: Electronics and Communication Systems (ICECS), 2015 2nd