ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.3

Placement Predict: A Review of Engineering Graduate Placement Statistics in India

Tanvi Gupta[1], Aparna Singh[2], Ajay Kaushik[3]
B. Tech Student[1, 2], Assistant Professor[3]
Department of Information Technology
Maharaja Agrasen Institute of Technology, GGSIPU, Delhi, India

Abstract:
Prediction of graduate jobs is a key problem in the analysis of employability which can be addressed with data-driven strategies. Aspiring Minds dataset describes aptly the attributes of a freshly graduate engineering student, making it one of the most suitable datasets for prediction of graduate placements as well as for building machine learning models for analysis of employability. By implementing various machine learning repressors, we show that Ridge and Lasso regression performed marginally better than Random forest, followed closely by Support Vector Machines. The objective of this study was to predict the target salary and to analyse the overall employability of an engineering graduate. This methodology of data-driven approach can also serve as a foundation for future studies towards prediction of salary and placements, and identification of key features linked to employability of candidates.

Keywords: graduate-employment, machine learning, engineering-jobs

I. INTRODUCTION

With increasing investment in the educational sector, rising unemployment among young graduates in India has become a critical issue. Concerns about employability are raised when the problem is discussed with respect to quantitative dimensions of education. Employers have raised concerns about the caliber of the graduates and their adeptness to meet the requirements of jobs they take up after graduation and expectation of employers [1-5]. Some companies have addressed the issue by introducing training courses in which the graduates are apprised of the specific skills required for their positions, for example, Wipro introduced an extensive training course for fresh graduates which runs over one year and covers a variety of topics. Such courses require time and resources and smaller employers may find it adversarial to their immediate interests, hence they desire an 'oven ready and self-basting' graduate [6]. The origins of employability skill scarcity lie in the education system. The rate of drop-out at primary and secondary school level is high and the admission levels in higher education are deprived. This leads to a thin educated workforce. Institutes and universities provide poor infrastructure which brings further challenges before employability. The standards in more than 60% of institutes and 90% of universities in India are detracted. Therefore, the quality of graduates produced is low, making them less employable. The lack of interaction between industry and institutes results in an ever widening gap in academia. The extant gap in skills, particularly listening, and teamwork and collaboration, and knowledge of the organization and process; product, solutions, and services; and consumer behaviour, is a matter of grave concern [7]. There is a major skill gap extant among Indian engineering students, which is a reason strong enough to compel engineering colleges and institutions to focus more on the quality of engineers produced each year and their employability. As per a survey, 64 percent of the employers were not completely satisfied with the quality of engineering graduates' skills. They identified integrity, reliability and teamwork as the top three most important general skills, while the top three most important specific skills were entrepreneurship, communication in English and use of modern tools and technologies. The employers were relatively satisfied with the graduates' communication skills in English, but not with their reliability (Federation of Indian Chambers of Commerce and Industry (FICCI) and the World Bank. July 2007). Aspiring minds released their data publicly for analytical purposes for the year 2016 which contains information about different aspects of a potential engineering graduate looking for employment. These factors include their logical abilities and numeracy, proficiency in English, technical knowledge in different branches of engineering and programming skills; and (qualities) such as conscientiousness, agreeableness and openness to new experiences. These data were collected for a range of engineering graduates hailing from different branches such as computer science, mechanical, electrical, civil etc. To address and showcase this problem of high skilled unemployment among engineering graduates from institutes all across the country, we present a statistical analysis involving various machine learning methods to predict the overall employment potential of a candidate using the aforementioned factors as a basis.

II. MATERIALS AND METHODS

1. Data from Aspiring Minds

Data of the scores on AMCAT or Aspiring Minds Computer Adaptive Test, was released by Aspiring Minds. This data comprises of the personal details and scores of 4000 candidates who took the test. For every potential candidate, the data is divided into sets of dependent and independent variables. The

salary offered is the target predictor, which is a dependent variable. It is determined through a range of independent variables such as gender, date of birth, college ID, degree, specialization etc. These data include the scores of the candidates in 10^{th} and 12^{th} standards, college and on the AMCAT. They have to be normalized due to the varying ranges of the values; for example, salary ranges from 35,000 to 40 lakhs in rupees, overall scores in 10^{th} standard (43 to 97.6), 12^{th} standard (40 to 98.7) and college examinations (6.45 to 99.93) have varying ranges, and scores on each section (logical, verbal and quantitative) of the AMCAT range from 120 to 900.

2. Dependent vs independent variables

Dependent variables are those whose outcome, or variation in outcome, is to be studied. The independent variables represent the arbitrary values of input, or the potential reasons for variation in the outcome. They are controlled by the experimenter. Models are deployed to study the effects independent variables have on dependent variables. The dependent variables in the data used are salary, date of joining, designation and the city in which the candidate is offered the job. Independent variables in the data include gender, date of birth, overall scores in 10^{th} and 12^{th} grade examinations, college ID, degree, specialization, overall score in college, scores in each section of the AMCAT etc. On an intuitive basis, the target predictor, salary (dependent variable), is dependent on sets of independent variables, majorly the degree, specialization and scores on each section of the AMCAT, whereas independent variables like college ID have no effect.

3. Preprocessing

Data preprocessing is a vital step in the data mining process. Analysis of data that has not been carefully screened for problems, such as out-of-range values, presence of redundant information or noisy data, can impede knowledge discovery during the training phase and produce misleading results. Thus, representation and quality of data before running an analysis has primacy, making data pre-processing the most important phase of a machine learning project. The product of data preprocessing is the final training set. Variables in the data used hold different types of values (numeric, textual, binary) and have to be transformed into a purely scalable numeric form so that they can be used to analyze and predict the values of the target, the salary. To perform such transformations, we wrote some descriptive functions to deal with specific nuances of these independent variables. For example - the scores of quantitative/verbal ability of a candidate were divided into 3 classes. More importantly, we realized (through exploratory data analysis) that the target predictor, salary, was extremely skewed *(+6.64)*. To correct this disparity, we normalized the salary using a logarithmic transformation *(log1p)*.

III. RESULTS AND DISCUSSIONS

1. Nature of AMCAT data

The quality of results obtained via machine learning are contingent upon the nature of data, selection of relevant features as well as evaluation metrics. Aspiring Minds provides one of the most well-curated data of the personal details and scores of the candidates in various examinations such as 10th and 12th standard examinations, college examinations and the AMCAT, facilitating their data-driven predictions. These meticulously curated data of the details of fresh graduates comprise of 4000 candidates and 17 independent variables (Supplementary Table T1). Associations between salary and the independent variables were inhomogeneous and were loaded with exceptionally large values of the predictor (salary) as well as the independent variables. When seen with the perspective of salary, the data were skewed (+6.44) with exceptionally large values of salary (Figure 1 and Figure 2). A correlation matrix of the 17 independent variables was computed (S1 Fig.).

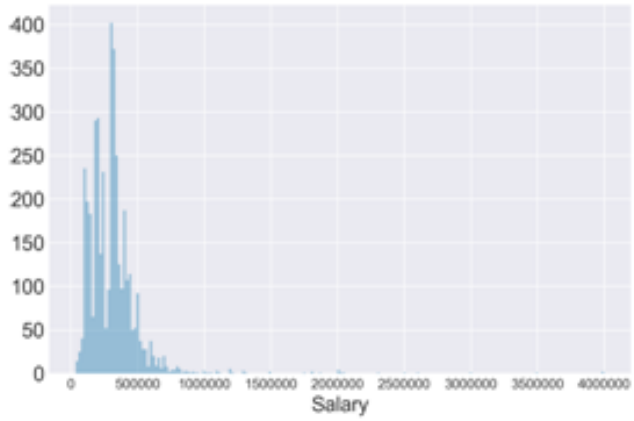

Figure.1. Skewed nature of our predictor variable i.e. Salary.

Figure.2. Normalized salary after applying log1p transformation.

After selection of features, we created scatter plots of the independent variables against our target i.e. salary to determine their individual effects on the same (S2-S5 Fig.). Moving along, after feature selection, we checked the finalized independent variables for skewness and corrected the same. Conclusively, at the end of the process, we had a log normalized data ready for predictive analysis through machine learning. Nature of the problem demands an accurate evaluation metric for this regression problem by which we can measure the correctness of our results. For this purpose, we have deployed the use of mean absolute error (MAE) and root mean square error (RMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}.$$

2. Assessment of regression performance

With this premise, we intended to design a strategy that could be meaningfully implemented for the prediction of salary. Along with mean absolute error and root mean square error, we have also made use of residual plots for a number of algorithm used in the analysis (Support vector machines, Random forests (n-estimator = 100), Lasso regression (a = 0.001), Ridge regression (a = 0.1)). Results (Table 1) indicate that Ridge and Lasso regression (Figure 3 & 4) fair out marginally better than Random forest (Figure 5), with Support Vector Machines following closely behind.

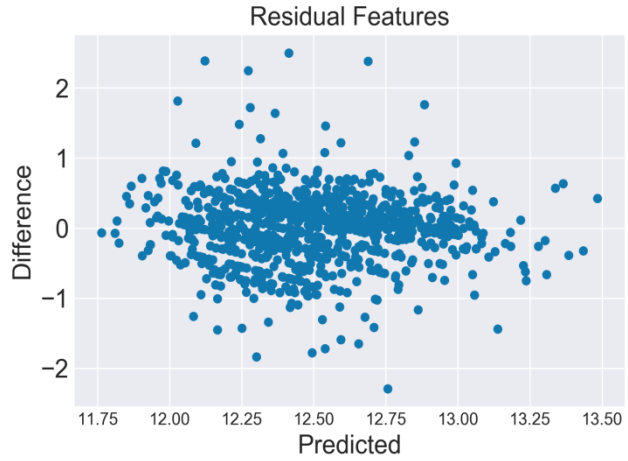

Figure.3. Residual plot for ridge regression.

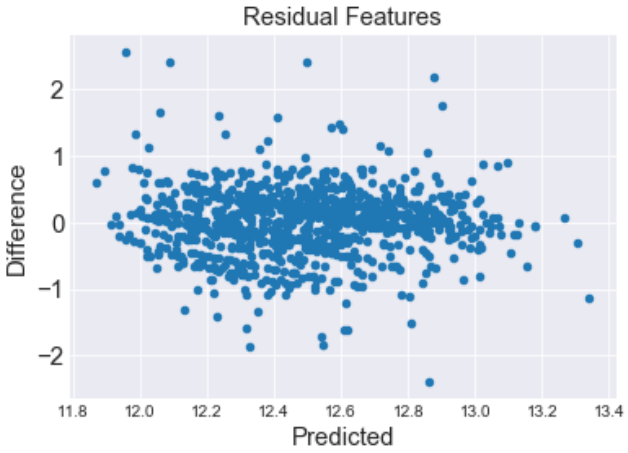

Figure.4. Residual plot for lasso regression

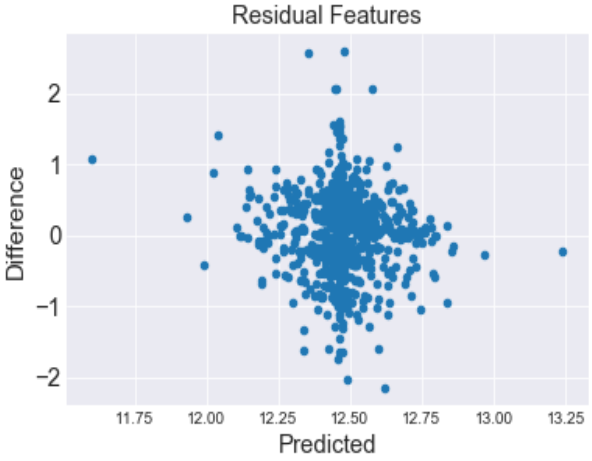

Figure.5. Residual plot for Support Vector Machine

Table.1. Comparison of various regressors based on their a) MAE and b) RMSE.

CLASSIFIER	MAE	RMSE
SVM	0.424635	0.424635
Random Forest	0.360008	0.493191
Ridge Regression	0.359375	0.488849
Lasso Regression	0.359208	0.489167

IV. SUMMARY AND CONCLUSIONS

Prediction of graduate placements is a key problem in the analysis of employability. Data-driven studies on empirical data have the potential to address this challenge. Aspiring Minds dataset is one of the most suitable dataset towards this objective as it appropriately describes the attributes of a freshly graduate engineering student. Thus, we have used it for prediction of graduate placements and salary, as well as for building machine learning models for analysis of employability. We highlight the inherent imbalance in these data that render studies solely relying on performance metrics such as degree, irrelevant. Such a methodology of data-driven approach can also serve as a basis for future studies towards prediction of salary and placements, and identification of key features linked to employability of candidates. The objective of this study was to predict the overall employability of a new engineering graduate. While we have used the inherent features in the Aspiring Minds dataset, one may use other features as well for the same data-driven strategy, such as any previous work experience etc.

V. ACKNOWLEDGEMENTS

T.G and A.S. thank the Department of Information Technology at MAIT, for providing the computational facilities required to complete the project and their continuous support.

VI. SUPPLEMENTARY DATA

S1 Fig: Correlation matrix of the independent variables
S2-S5 Fig: Scatter plots of the independent variables against the target (salary)
T1 Table: Training dataset

VII. REFERENCES

[1]. Hills, J.M., Robertson, G., Walker, R., Adey, M.A. and Nixon, T.D. (2003) Bridging the gap between degree programmed curricula & employability through implementation of work related learning. *Teaching in Higher Education*, 8, 311-231.

[2]. Knight, P. and York, M. (2002) Defining and addressing employability: A fresh approach. *Exchange*, 2, 15-18.

[3]. Lesslie, J. (2004) *the employer's perspective.* http:// www. bioscienc e.heacademy.ac.uk/events/reports/ (accessed 4 March 2005)

[4]. Little, B., Connor, H., Lebeau, Y., Pierce, D., Sinclair, E., Thomas, L., Yarrow, K. (2003) *Vocational HIGHER Education does it meet employers' needs?* London: Learning and Skills Development agency.

[5]. Miller Smith, C. (2002) A Business view of the graduate today. *Exchange*, 2, 8-11.

[6]. Atkins, M.J. (1999) Oven ready and self-basting: taking stock of employability. *Teaching in Higher Education*,4 (2), 267.

[7]. Survey for the Indian Banking, Financial Services, and Insurance Sector the Skills Gap Survey is an initiative of The Higher Education Forum supported by 1SOS & Westat 6 March 2010.