Journal of Data Science 16(2017), 553-574

AN EMPIRICAL STUDY OF STARTING SALARIES AND EMPLOYMENT TRENDS OF ENGINEERING STUDENTS IN INDIA

Shrihari Vasudevan, Ritwik Chaudhuri, Madhavan Pallan and Sudhanshu S. Singh[1]

[1] *IBM Research - India*

Abstract: This paper presents an empirical study of a recently compiled workforce analytics data-set modeling employment outcomes of Engineering students. The contributions reported in this paper won the data challenge of the ACM IKDD 2016 Conference on Data Science. Two problems are addressed - regression using heterogeneous information types and the extraction of insights/trends from data to make recommendations; these goals are supported by a range of visualizations. Whereas the data-set is specific to a nation, the underlying techniques and visualization methods are generally applicable. Gaussian processes are proposed to model and predict salary as a function of heterogeneous independent attributes. Key novelties the GP approach brings to the domain of understanding workforce analytics are (a) statistically sound notion of uncertainty of prediction that is data dependent, (b) automatic relevance determination of various independent attributes to the dependent variable (salary),(c) seamless incorporation of both numeric and string attributes within the same regression frame- work without dichotomization; specifically, string attributes include single-word or categorical (e.g. gender) or nominal attributes (e.g. college tier) or multi-word attributes (e.g. specialization) and (d) treatment of all data as being correlated towards making predictions. Insights from both predictive modeling approaches and data analysis were used to suggest factors, that if improved, might lead to better starting salaries for Engineering students. A range of visualization techniques were used to extract key employment patterns from the data.

Key words: Gaussian process, predictive modeling, association mining and workforce analytics.

1. Introduction

Workforce analytics is an active area of research, the broad mandate being the understanding of labour markets. This study focuses on the problem of data-driven enablement of the Engineering student population of India. Of particular interest, are the problems of predicting starting salary and identifying factors, which if improved, can lead to higher starting salaries.

Globally, several recent studies have addressed the problem of understanding student expectations of starting salary given the educational qualifications they pursue. The Dutch study in Webbink & Hartog (2004) suggests that students make education choices considering the

return (salary), a view also suggested in Schweri & Hartog (2015). The authors of Webbink & Hartog (2004) asked students their expected starting salary and reconciled this information with their actual salaries when they took up jobs. It concluded that there was no systematic difference between expectations and realizations, suggesting that students make realistic estimates of their starting salaries. A study from Switzerland (Wolter & Zbinden (2002)) reported similar findings but also reported that expected salary gains in the first ten years of employment exceeded actual gains. Students in the United Kingdom tended to overestimate their starting salaries, as was suggested in Jerrim (2008). The paper van der Merwe (2011) focused on salary expectations of South African students; it reported that first year students of Durban University of Technology were able to quite accurately predict their starting, medium and long term salaries; it also reported a significant relationship between the expected rate of return to the educational investment and to factors like field of study and educational accomplishment of parents. The authors of Ogunrinola & Adebayo (2010) developed ordinary least squares models for expected job-search period and starting salary for final year university students of select universities in South-Western Nigeria. They reported generally optimistic views despite high unemployment; private university students were generally more positive than public ones.

These studies are mostly aimed at reconciling student expectations with realizations and making conclusions thereof. Models that were developed (e.g. in Ogunrinola & Adebayo (2010)) were based on multiple linear regression. This work focuses on the problem of modeling starting salary as a function of candidate (an Engineering student) background; it also the focuses on the identification of key trends and factors that influence salary. The latter could guide prospective job seekers towards focused improvement of key metrics towards obtaining higher starting salaries. Lack of India-specific data-sets suited to addressing such problems has not helped with progress in this area.

This work is a summary of statistical analysis done on a recently compiled dataset, Aspiring Minds (2015), capturing employment outcomes. The analysis was done during a time-bound data science com- petition held within the framework of a major conference on data science - ACM IKDD CODS (2016). The dataset Aspiring Minds (2015) contains background information about engineering candidates and their employment outcomes. A number of parametric, non-parametric and machine learning methods were attempted towards predicting salary for job-candidates, given their background information. This paper reports the use of Gaussian processes to model salaries; it demonstrates the use of kernels as a means of integrating heterogeneous background information types. The technique is new to the domain of workforce analytics. The paper also presents several insightful trends observed from the data and makes recommendations for candidates seeking higher starting salaries. This paper thus views the workforce- analytics / labour market / student salary expectation problem from a data science perspective; it reports on the application of a state-of-the-art probabilistic modeling approach to the problem and trends that are intended to enable the Engineering student population of India.

2. Data Description

The data-set, Aspiring Minds (2015), contains background information of engineering candidates and their employment outcomes. For each candidate, the data contains details on gender, date of birth, date of joining (DOJ), date of leaving (DOL) (if available), class 10 grades, class 10 board, class 12 grade, class 12 board, GPA in college, college Major, college reputation (Tier 1 or Tier 2), college city, college city tier, college ID, permanent location for the candidate. The dataset also contains Aspiring Minds' AMCAT test scores taken optionally by candidates to assess their cognitive, domain and personality attributes. Employment outcomes captured in the data set are the annual salary (in Indian Rupees (INR), job-title (Designation) and job-location (JobCity) for the first job the candidate landed. The dataset was observed to be incomplete and noisy, with many of the entries in the string variables spelled wrongly. Also, the salary attribute contained a few high-value cases.

Table 1: Summary of variables with missing values

Variable name	Number of missing values	Percentage of missing values
Domain	246	6.15%
ComputerProgramming	868	21.71%
ElectronicsAndSemicon	2854	71.38%
ComputerScience	3096	77.44%
MechanicalEngg	3763	94.12%
ElectricalEngg	3837	95.97%
TelecomEngg	3624	90.64%
CivilEngg	3956	98.64%

The dataset was divided into two sections - training and test. Training data contained information on 3998 candidate profiles while the test data contained 1500 different candidate profiles. Within the test data, information on DOL, DOJ, designation and job-location were not provided. These variables along with the salary comprised five dependent attributes; the remainder were the independent attributes. For salary-prediction purposes, the above mentioned attributes were ignored from the training data. They were however considered for data visualization to draw inferences on job locations and designations.

The training data contained missing values. Missing entries have been denoted by -1. A summary of missing values for different variables within the training data can be obtained from Table 1. For other

variables there were no missing entries in the training data; there were however discrepancies among various entries within the variables 10board, 12board, CollegeSate, etc.

3. Gaussian Processes

3.1 Theory

Gaussian processes (GPs; see Rasmussen & Williams (2006)) are stochastic processes wherein any finite subset of random variables are jointly Gaussian distributed. They may be thought of as a Gaussian probability distribution in function space. They are characterized by a mean function m(x) and the covariance function k(x, x') that together specify a distribution over functions. In the context of the salary prediction problem, each x ≡ (attribute$_1$, attribute$_2$,..., attribute$_d$), independent attributes representing candidate profile (d dimensional) and f (x) ≡ z, the salary of the candidate. Although not necessary, the mean function m(x) may be assumed to be zero by scaling/shifting the data appropriately such that it has an empirical mean of zero.

The covariance function or kernel models the relationship between the random variables corresponding to the given data. It can take numerous forms as shown in Rasmussen & Williams (2006, chap. 4). The stationary squared exponential (or Gaussian) kernel (SQEXP) is given by

$$k_{SQEXP}(\mathbf{x}, \mathbf{x}', \Sigma) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \Sigma (\mathbf{x} - \mathbf{x}')\right), \quad (1)$$

where k is the covariance function or kernel; $\Sigma = \text{diag}[\, l_1 , l_2 , ..., l_d \,]^{-2}$ is a d x d diagonal length-scale matrix (d = dimensionality of input data), a measure of how quickly the modelled function changes with a change in each of the input attributes. The set $\{ l_1 , l_2 , ..., l_d \}$ constitute the kernel hyperparameters; they define a family of functions rather than a single function.

The non-stationary neural network (NN) kernel, shown in Neal (1996), Williams (1998a,b), takes the form

$$k_{NN}(\mathbf{x}, \mathbf{x}', \Sigma) = \frac{2}{\pi} \arcsin\left(\frac{2\bar{\mathbf{x}}^T \Sigma \bar{\mathbf{x}}'}{\sqrt{(1 + 2\bar{\mathbf{x}}^T \Sigma \bar{\mathbf{x}})(1 + 2\bar{\mathbf{x}}'^T \Sigma \bar{\mathbf{x}}')}}\right), \quad (2)$$

where $\bar{x}$ and $\bar{x}'$ are augmented input vectors (each point is augmented with a 1), Σ is a (d + 1) x (d + 1) diagonal length-scale matrix given by $\Sigma = \text{diag}[l_1 , l_2 , ..., l_d , \beta \,]^{-2}$, β being a bias factor and d being the dimensionality of the input data. The set $\{ [l_1 , l_2 , ..., l_d , \beta, \sigma_f \}$ constitute the kernel hyperparameters. The NN kernel represents the covariance function of a neural network with a single hidden layer between the input and output, infinitely many hidden nodes and using a Sigmoidal transfer function Williams (1998a) for the hidden nodes. Hornik, in Hornik (1993), showed that such neural networks are universal approximators and Neal, in Neal (1996), observed that the functions produced by such a network would tend to a Gaussian process. Prior work in Vasudevan, Ramos, Nettleton & Durrant-Whyte (2009a) found the NN kernel to be more effective than the SQEXP kernel at modeling complex or discontinuous data.

Regression using GPs uses the fact that any finite set of training (or evaluation) data and test data of a GP are jointly Gaussian distributed. Assuming noise free data, this idea is shown in Expression 3 (hereafter referred to as Equation 3). This leads to the standard GP regression equations yielding an estimate (the mean value, given by Equation 4) and its uncertainty (Equation 5).

$$\begin{bmatrix} \mathbf{z} \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{3}$$

$$\bar{f}_* = K(X_*,X)\,K(X,X)^{-1}\,\mathbf{z} \tag{4}$$

$$\text{cov}(f_*) = K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*) \tag{5}$$

For n training points (X, z) = $(x_i, z_i)i=1...n$ and n* test points (X_* , f_*), K(X, X_*) denotes the n × n_* matrix of covariances evaluated at all pairs of training and test points. The terms K(X, X), K(X_* , X_*) and K(X_* , X) are defined likewise. In the event that the data being modeled is noisy, a noise hyperparameter (σ) is also learnt with the other GP hyperparameters and the covariance matrix of the training data K(X, X) is replaced by [K(X, X) + σ^2I] in Equations 3, 4 and 5. GP hyperparameters may be learnt using various techniques such as cross validation based approaches Rasmussen & Williams (2006) and maximum-a-posteriori approaches using Markov Chain Monte Carlo techniques Rasmussen & Williams (2006), Williams (1998b) and maximizing the marginal likelihood of the observed training data Rasmussen & Williams (2006), Vasudevan, Ramos, Nettleton & Durrant-Whyte (2009a) - the approach used in this paper. The marginal likelihood to be maximized is described in Equation 6.

$$\log\ p(\mathbf{z}|X,\theta) = -\frac{1}{2}\mathbf{z}^T K(X,X)^{-1}\mathbf{z} - \frac{1}{2}\log|K(X,X)| - \frac{n}{2}\log(2\pi) \tag{6}$$

3.2 Why GPs

Gaussian processes (GPs) and specifically Bayesian GPs (referred to here as GPs) are a generic and flexible approach for addressing a range of analytics problems that occur across domains; they have numerous capabilities that make them a versatile methodology for salary prediction given candidate profiles and more broadly, workforce analytics. GPs provide data driven (multi-scale) modeling as they learn a continuous manifold from available data; they provide data driven uncertainty characterization in that every prediction has an associated uncertainty that depends on available support data (for prediction) and prior knowledge available (e.g. apriori known noise). The underlying Gaussian assumption and the properties of this distribution enable the derivation of closed form expressions for a variety of complex settings. The Bayesian modeling paradigm along with the nonparametric nature of the GP model allow for a convenient mechanism to incorporate prior knowledge about the problem at hand. The work Vasudevan, Ramos, Nettleton & Durrant-Whyte (2009b) compared GPs (which perform Kriging interpolation) with a variety of standard interpolation techniques. Results suggested that for simple data (e.g. smooth variation, dense data) GPs would be competitive with standard interpolation techniques; for complex data (e.g. sparse, discontinuities etc.) GPs would outperform existing interpolation techniques. GPs naturally provide Automatic Relevance Determination (ARD) Rasmussen & Williams (2006) in a multi- factor model by weighing

different factors based on their relevance to the desired outcome. This is similar to the variable selection employed in traditional regression settings. GPs also provide a mechanism to explicitly incorporate uncertainty in the factors Girard (2004). GPs can also be used to combine multiple data-sets (data fusion) as demonstrated in Vasudevan (2012) or to predict multiple outcomes simultaneously Vasudevan, Melkumyan & Scheding (2015). Whereas the Gaussian assumption is indeed a strong one for real world settings, non Gaussian data may be handled through transformations both explicit (e.g. the Box-Cox transformation) or implicit, as in Snelson, Ghahramani & Rasmussen (2004). GPs thus bring a lot of different traditional analytical capabilities together under one common flexible framework that can also be extended in different ways to meet various data challenges; they promote automation of analytics, reduce manual efforts and improve consistency of outcomes.

3.3 Predictive modeling of Salary data

The data-set, Aspiring Minds (2015), provides salary information for various candidates given their background information. Candidate background data includes three kinds of attributes numeric (e.g. scores), nominal (categorical or single-word e.g. gender) and phrasal (multi-word e.g. specialization). The normal approach to handling such data is to create a new high-dimensional data-set with the latter attributes dichotomized and apply any number of analytical approaches. Instead, this paper uses GPs to integrate all three kinds of attributes within the aforementioned regression framework.

The paper Girolami (2006) integrated heterogeneous feature types within a Gaussian process classification setting, in a protein fold recognition application domain. Each feature representation was represented by a separate GP. Fusion used the idea that individual feature representations were considered independent and hence a composite covariance function could be defined in terms of a linear sum of Gaussian process priors. The authors of Reece, Roberts, Nicholson & Lloyd (2011) integrated "hard" data obtained from sensors with "soft" information obtained from human sources within a Gaussian process classification framework. It used heterogeneous information-types as mutually independent sources of information that were transformed into the kernel representation (a kernel for each kind of information) and combined using a product rule. This paper follows a similar approach but proposes the use of a string matching kernel that reduces to a nominal kernel for single-word or categorical attributes and to an intersection kernel (squared form, normalized by length of strings and passed through an exponential function) for multi-word attributes i.e. multi-word attributes are treated as sets of words. The form of the string matching kernel proposed for a pair of string attributes is given as

$$k_{SM}(\mathbf{x}, \mathbf{x}', \Sigma) = \exp\left(\frac{1}{l^2}\frac{|\mathbf{x} \cap \mathbf{x}'|^2}{|\mathbf{x}||\mathbf{x}'|}\right) \tag{7}$$

where |x| is the measure/length of the set x, k is the covariance function or kernel; $\Sigma = \text{diag}[l_1, l_2, ..., l_d]^{-2}$ is a d x d diagonal length-scale matrix (d = dimensionality of non-numeric input data). The set of parameters $\{ l_1, l_2, ..., l_d \}$ are referred to as the kernel hyperparameters.

Given heterogeneous candidate attributes, the kernel function for such data is obtained as a product of kernel matrices obtained for numeric and string attributes considered individually. The resulting kernel matrix is obtained as

$$K(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \, . \, K_{SM}(\mathbf{x}_{str}, \mathbf{x}'_{str}, \Sigma_{str}) \, . \, K_{num}(\mathbf{x}_{num}, \mathbf{x}'_{num}, \Sigma_{num}) \quad (8)$$

where σf is the signal variance and Knum can be based on the SQEXP (Equation 1) or the NN (Equation 2; used in this paper) or any other kernel applicable to numeric data. The data is assumed to be noisy and consequently a noise hyperparameter is incorporated - the kernel matrix K(X, X) in eqns 3, 4 and 5 are replaced by [K(X, X) + σ2I]. The set of hyperparameters { l_1 , l_2 , ..., l_d , β , σ_f , σ} including length scales for each attribute, bias term (for NN kernel), signal variance and noise variance are learnt from the data; these taken together with the data, define the GP model.

The proposed GP approach brings several novel features to the problem of understanding labour markets (and more broadly, workforce analytics). Equation 5 produces an uncertainty estimate for a prediction. Thus, not only does the approach provide salary predictions but it also provides upper and lower bounds for these predictions. These uncertainties are dependent on the support data used to make the prediction. GPs provide automatic relevance determination of various independent attributes to the dependent variable (salary). GPs (specifically, kernel methods) provide a means for seamless incorporation of both numeric and phrasal (single or multi-word) attributes within the same regression framework without dichotomization - this relies on the fact that sums and products of kernels will be kernels. Finally, GPs consider all data points as being correlated; it tries to model and use this correlation to improve outcomes.

The data-set, Aspiring Minds (2015), was subject to a data-cleaning step wherein incorrect or in- correctly spelt or non-standardized attributes were rectified. For string attributes, missing values were replaced by 'NA'. Within the model, each 'NA' was assumed to be different from each other, i.e. missing values were regarded as being unequal. The salary data observed a log-normal behavior and was hence log-transformed prior to modeling. The model hyperparameters were then learnt from a randomly drawn sample (1000 points) of the data-set. The GP model is defined by both the hyperparameters and the training/evaluation data. A ten fold cross validation was conducted on the training data. This involved randomly splitting the data into ten subsets and repeatedly testing one subset against the model defined by the other nine subsets of data. The average MSE and its root (RMSE) were computed as were the predictions for test data, which were accompanied by uncertainty bounds. To compare the performance of GPs, a baseline approach of multiple linear regression was used; this choice was motivated by its use in prior/related works mentioned in Section 1. String attributes in the data set were recoded into separate dichotomous variables prior to creating the model; the recoding is popularly known as "dummy-coding", it significantly increases the dimensionality of the data set. A ten-fold cross validation process were performed for this approach as well. Performance outcomes for the two methods, using both numeric and string attributes, are depicted in Table 2. It shows that the Gaussian process approach compares favorably with the baseline approach while leveraging

properties of kernels to develop a clean way of integrating heterogenous attributes without any recoding of the data being required. The GP modeling code was developed in R.

Table 2: 10-fold cross validation outcomes

Approach	RMSE (salary prediction error in Indian Rupees)
Gaussian Process (product of SM and NN kernels)	187,492
Linear Regression (baseline approach)	202,422

The hyperparameters of the GP model define the (relative) importance of the factor towards salary prediction. Note that this is one possible combination of relevant parameters and that other combinations may be obtained from solutions to the non-convex optimization problem of maximizing the log-marginal likelihood. As per the GP model, the attributes most relevant to salary prediction were collegeGPA, 12graduation, GraduationYear and Quant, in that order. Various other approaches tested also suggested parameters which if increased could increase the salary of the candidate. Linear modeling approaches like linear regression, regularized (LASSO) linear regression and PCA (loadings) suggested that increasing Quant, 10percentage, English, 12percentage, Logical, Domain and collegeGPA leads to better salaries.

4. Data Insight and Visualization

This section draws on data tabulation and visualization techniques to extract interesting patterns observed in the data. It also provides recommendations based on these trends for a potential job-candidate aspiring for a higher starting salary in their first job.

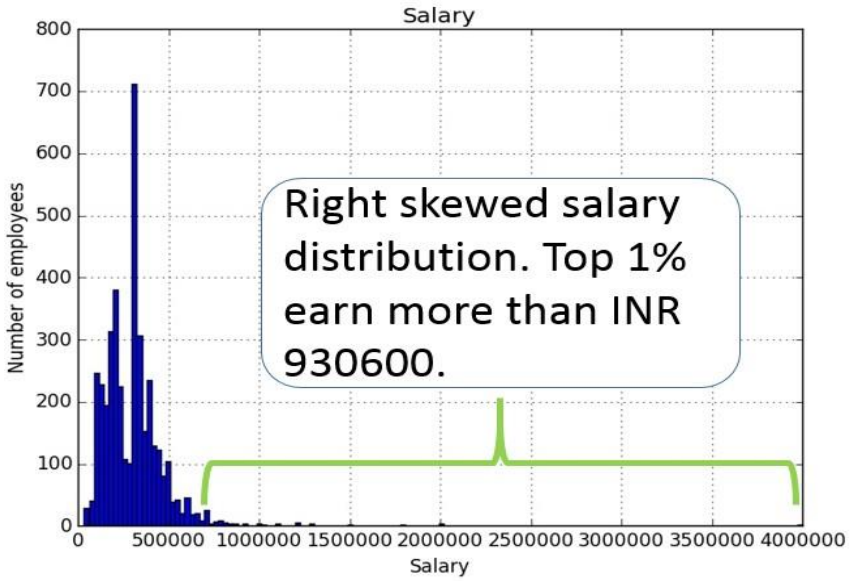


Figure 1: Salary distribution of employees

The data set captured salary of fresh college graduates; the distribution of salary was found to be right skewed with 99% of the employee earning less than INR 0.93 million, while the highest salary obtained was INR 4 million. The salary distribution is shown in Figure 1.

As part of exploratory data analysis, a correlogram plot was used to identify variables correlated with salary; this is shown in Figure 2. The figure shows that salary had a stronger correlation with attributes Quant, Logical, 12 percentage, and relatively weak correlation with English, Domain and GPA. This suggested that with relatively high scores in Quant, Logical and 12 percentage, one has a better chance of getting a higher-salaried job.

An objective of the analysis was to understand how scores in different subjects affected the salary of employees. This would provide recommendations on subjects fresh graduates needed to perform well, to bag a high salaried job. To capture this information in an interpretable manner, each of the attributes Quant, Logical and 12percentage were divided into four sections using the following rule:

- Low: less then or equal to 25 percentile score
- Medium: Above 25 percentile but less then or equal to 75 percentile score
- High: Above 75 percentile but less then or equal to 99 percentile score
- Very high: Above 99 percentile score

To understand how salary varied with different level of scores in different subjects a 3-D bubble chart was created with the levels of scores in X, Y and Z axis and the color of the bubble signifying the average salary of the employees. The size of the bubble was proportional to the number of employees in each group. In Figure 3, the color of the bubbles map with the color bar of average salary on the right. Also, the size of the bubble signifies how many people are there in each of the group. A large bubble represents a large number of people in that group while a smaller bubble represents the opposite. As evident from the figure, low scores in all three attributes, Quant, English and 12 percentage, resulted in a low salary while a high score in Quant, English and 12 percentage fetched a high-salary job. Also, it was observed that very few fresh graduates actually landed a high-salary job; most graduates obtained a medium-salary job with medium level scores in all three subjects. The figure also suggested that it is more important to do well in Quant and Logical, than 12 percentage. A candidate with a low 12-percentage score but with a high Quant and English score may end up bagging a higher salary job.

Since a major portion of the dataset was based on the candidates who were graduates in an Engineering domain, it was interesting to understand how English scores affected the salary of employees. A pie-chart visualization of the data was used to enable this objective; this is shown in Figure 4. The figure suggested that more employees were hired for medium salaried jobs when their English score was high. Also, employees with higher English scores were more likely to get a medium-salary job and less likely to get a low-salary job. But, going against this trend, it was found that no candidate with a very high English score landed a high-salary job. Taken together with previous findings, this suggested that candidates who obtained very high scores in English,

did not do as well in their Quant, Logical or 12-percentage scores. Often, it is of importance to observe whether people doing well in one subject are also doing well in other related subjects. As a case in point, it would be useful to investigate whether people with strong Quant scores also did well in Logical. A positive outcome would suggest that candidates with high Quant scores have a higher chance of having strong Logical abilities as well. To understand this, a scatterplot between Logical and Quant Scores was used. This is shown in Figure 5. It can be observed from the Figure 5 that the correlation between high Quant scores and high Logical scores was only 0.21. Hence, candidates who did well in Quant did not necessarily do as well in Logical. The correlation between high scores in Quant and high Domain scores was observed to be stronger.

From the scatterplot shown in Figure 6, it can be observed that correlation between high scores in English and high scores in Domain was 0.09 while the correlation between all English scores and all Domain scores is 0.22. This suggested that candidates who did well in English did not get strong Domain scores.

Fresh graduates tend to choose job cities which can afford them significant opportunities for career growth. It was thus important to determine the major cities where the employee concentration was maximum. Cities (the variable JobCity) were divided into three different tiers based on Wikipedia information. During exploratory analysis of the training data, it was found that for 461 employees the JobCity were missing. There were 11 instances where JobCity was outside India and for another 8 instances JobCity was unknown. These instances were removed before further analysis. Additionally, there were typographical errors in the names that required fixing prior to analysis.

A horizontal barplot was used to understand employee concentration in various cities; this is shown in Figure 7. The figure suggested that most employees were located in the cities of Bangalore, Noida, Hyderabad, Pune, Chennai, Gurgaon, New Delhi (NCR Region), Kolkata, Mumbai, Jaipur, Lucknow and Bhubaneswar in that order. In Bangalore there were 686 employees.

Understanding the employee percentage concentration in different cities was also important. With further investigation, this could enable an understanding of why some of the major cities lag behind the top job-cities in attracting employees. There were 197 job cities in the data. A pie-chart representation of employee percentage concentration was used to draw inferences on job cities. This is shown in Figure 8. The figure suggested that almost 19.4% of the employees were based in Bangalore while 22% of the employees were placed in 188 small cities, A new variable was created to represent job states depending on the job cities where employees are placed. This could enable state-level macro-economic trend analysis. The data set represented a sample of the Indian labour market, the employee concentration was thus calculated within different Indian states. To understand employee concentration of India, a heat map of India was created in the shades of red as shown in Figure 9. The deeper the shades of red, the more is the concentration of the employees in a state.

From Figure 9, it was observed that Karnataka, Maharshtra, Uttar Pradesh, Tamil Nadu and Te- lengana had a high concentrations of employees. Orissa, West Bengal, Madhya Pradesh, Rajasthan had moderately high concentrations of employees. Low concentrations of employees were observed in the states of Bihar, Punjab, Gujarat and Uttarakhand. The states of Himachal

Pradesh and Jharkhand had very low concentrations of employees while states in entire North East, Jammu and Kashmir had almost no employees. The representation of employee concentration depended on how representative the dataset was of the fresh graduate employee population across the states of the country. Assuming a uniform data collection process and/or accounting for known factors in data collection across states, this information could enable state-level comparisons and growth recommendations for individual States.

While on the subject of state-level analysis of employment outcomes, it was deemed interesting to understand if employees had to relocate from their home states. To understand this, the attributes of job-state and college-state were looked into. Employees who had completed their college education in one state and obtained a job in another were designated to have "Out-of-State" jobs whereas those that did not have to relocate were designated to have obtained "In-State" jobs. A pie-chart representation of this information is shown in Figure 10. The figure suggested that almost 45.9% of the employees obtained "In-State" jobs which is substantially large; the remainder had "Out-of-State" jobs. Taken together with the number of samples from each state, the information could shed light on development and employment/business climate in states.

Understanding typical job designations offered to fresh graduates could enable college students in course selection and career planning. To clarify this, a horizontal bar-plot of the top twelve designations was prepared. This is shown in Figure 11. The dataset comprised of information on fresh college graduates who had mostly graduated in the discipline of Engineering; hence, it was not unusual that maximum number of employees were hired for the position of "software engineers". The next most hired job roles were "system engineers" and "software developers".

Given (from Figure 11) that the top five designations by employee concentration were software engineers, system engineers, software developers, programmer analyst and java software engineer, it was deemed insightful to analyze the kinds of specializations in college that fetch jobs with such designations. College students may use this information in choosing their specializations, to target particular job roles that may interest them. This information is shown through a stacked bar plot in Figure 12. The figure suggested that most software engineers (sw.engg.) and software developers (sw.dvlpr.) had a Computer Science and Engineering (CSE) background. Maximum system engineers (sys.engg.) were from an Electronics and Communication Engineering (ECE) background. For the designation programmer analyst (pg.anlyst.), roughly equal proportions of employees had specializations in Computer Engineering (CE), Computer Science and engineering (CSE) and Information Technology (IT). Amongst java software engineers (java sw. engg.), most employees had a Computer Science and Engineering (CSE) or Information Technology (IT) background. Except for the job designation of software developer, for all of the other four top designations, the proportion of employees with college specialization Computer application (CA) was least.

The next objective was to understand candidate performance in various AMCAT tests in the top five salaried jobs. This also reflected on the nature of candidates that best fit those job roles. Based on the multiple bar plot in Figure 13, senior software engineers had the highest average Domain score and the lowest English score compared to other four job designations. Assistant

managers had the lowest average score in each of Quant, Logical, English and Domain. System engineers had the highest average Quant and Logical scores closely followed by application developers. Programmer analysts had relatively low average scores in Quant, Logical and English as compared to system engineers and application developers. For visualization, charts and plots were prepared using Python. The libraries used were numpy, pandas, matplotlib of the Scipy stack (see Jones, Oliphant, Peterson et al. (2001)). An editable map of India was accessed using the site http://www.mapsofindia.com/editable-maps/customise-your-map.html.

5. Conclusion

An empirical study of a recently compiled data-set modeling employment outcomes of Indian Engineering students was presented. It focused on data-driven enablement of the Engineering student population of India. To this end, a Gaussian process model for prediction of starting salary given candidate (Engineering student) background was developed. The approach enabled providing data driven characterization of prediction uncertainty, resulting in predictions accompanied with bounds. Key independent attributes relevant to salary prediction were automatically determined during modeling. The developed regression model seamlessly integrated both numeric and string attributes within the same regression framework without the need for dichotomization. Recommendations and insights from both predictive modeling approaches and from data analysis suggest that better Quant, 10-percentage, English, 12-percentage, Logical, Domain and collegeGPA attributes leads to better salaries. Visualization of data suggested that maximum number of employees worked in Bangalore. Tier-1 job-cities (Ahmedabad, Bangalore, Chennai, Delhi, Hyderabad, Kolkata, Mumbai, Pune) had roughly 63% of all employees. Roughly 45.9% employees had in-state (same college-state and job-state) employment. Top 5 designation categories based on employee concentration were software engineer, system engineer, software developer, programmer analyst, java software engineer. Top five salaried job designations having at least 50 employees were senior software engineers, assistant managers, application developers, programmer analysts and system engineers. For system engineers, it was important to have good scores in Quant, Logical or Domain as their job responsibilities require quantitative skills.

References

[1] ACM IKDD CODS (2016), 'Conference on Data Sciences (CODS)'. http://ikdd.acm.org/Site/CoDS2016/index.html..

[2] Aspiring Minds (2015), 'Aspiring Minds Employment Outcomes 2015'. Data-set available online.

[3] Girard, A. (2004), Approximate Methods for Propagation of Uncertainty with Gaussian Process Models, PhD thesis, Department of Computing, University of Glasgow.

[4] Girolami, M. (2006), Bayesian data fusion with gaussian process priors: An application to protein fold recognition, in 'Workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB)'

[5] Hornik, K. (1993), 'Some new results on neural network approximation', Neural Networks 6(8), 1069– 1072.

[6] Jerrim, J. (2008), Wage expectations of uk students: how do they vary and are they realistic?, Working paper.

[7] Jones, E., Oliphant, T., Peterson, P. et al. (2001), 'SciPy: Open source scientific tools for Python'.
URL: http://www.scipy.org/

[8] Neal, R. M. (1996), Bayesian Learning for Neural Networks, Lecture Notes in Statistics 118, Springer, New York.

[9] Ogunrinola, O. & Adebayo, A. (2010), 'Labour market expectations of final year university students in south-western nigeria', African Journal of Business and Economic Research 5(1), 69–89.

[10] Rasmussen, C. E. & Williams, C. K. I. (2006), Gaussian Processes for Machine Learning, MIT Press.

[11] Reece, S., Roberts, S., Nicholson, D. & Lloyd, C. (2011), Determining intent using hard/soft data and gaussian process classifiers, in 'Proceedings of the 14th International Conference on Information Fusion (FUSION)'

[12] Schweri, J. & Hartog, J. (2015), Do wage expectations influence the decision to enroll in nursing college?, Working paper.

[13] Snelson, E., Ghahramani, Z. & Rasmussen, C. E. (2004), Warped gaussian processes, in S. Thrun, L. Saul & B. Sch¨olkopf, eds, 'Advances in Neural Information Processing Systems 16', MIT Press, pp. 337–344.

[14] van der Merwe, A. (2011), 'Earnings expectations of typical south african university of technology first- year students', Education Economics 19(2), 181–198.

[15] Vasudevan, S. (2012), 'Data fusion using gaussian processes', Elsevier Journal of Robotics and Au- tonomous Systems . Available online 25 August 2012..

[16] Vasudevan, S., Melkumyan, A. & Scheding, S. (2015), 'Efficacy of data fusion using convolved multi- output gaussian processes', Journal of Data Science . online 2014, based on arXiv report 1210.1928.

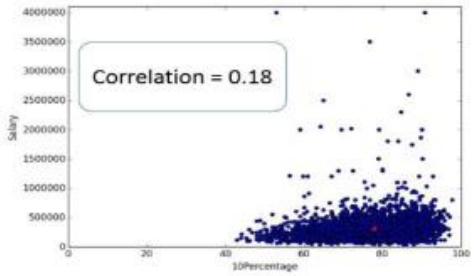
[17] Vasudevan, S., Ramos, F., Nettleton, E. & Durrant-Whyte, H. (2009a), 'Gaussian Process Modeling of Large Scale Terrain', Journal of Field Robotics 26(10), 812–840.

[18] Webbink, D. & Hartog, J. (2004), 'Can students predict starting salaries? yes!', Economics of Education Review 23.2, 103–113.

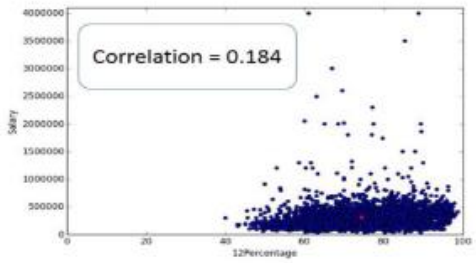
[19] Williams, C. K. I. (1998a), 'Computation with infinite neural networks', Neural Computation 10(5), 1203– 1216.

[20] Williams, C. K. I. (1998b), Prediction with Gaussian processes: From linear regression to linear prediction and beyond, in M. I. Jordan, ed., 'Learning in Graphical Models', Springer, pp. 599–622..

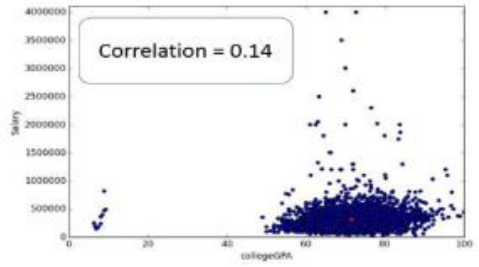
[21] Wolter, S. C. & Zbinden, A. (2002), 'Labour market expectations of swiss university students', Interna- tional Journal of Manpower 23(5), 458–470.



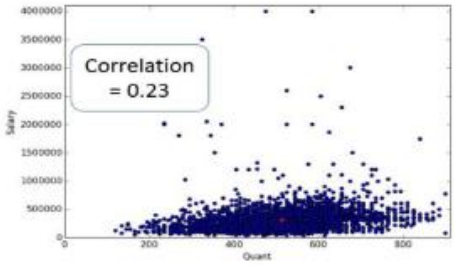
(a) Scatterplot of Salary and 10percentage



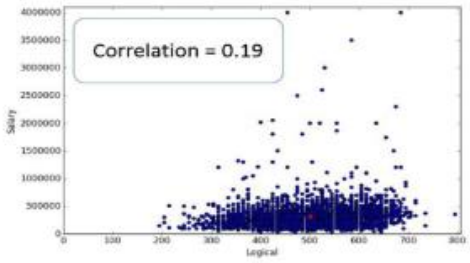
(b) Scatterplot of Salary and 12percentage



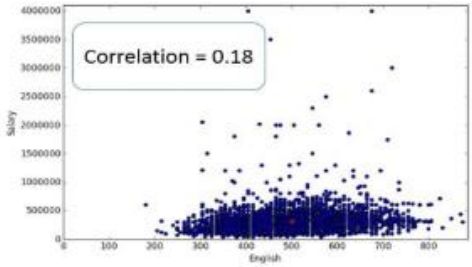
(c) Scatterplot of Salary and GPA



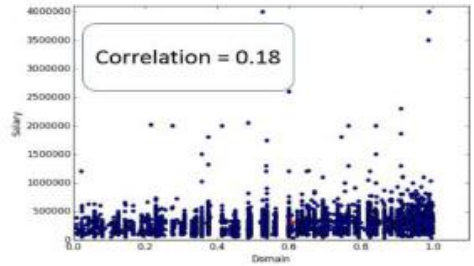
(d) Scatterplot of Salary and Quant



(e) Scatterplot of Salary and Logical



(f) Scatterplot of Salary and English



(g) Scatterplot of Salary and Domain

Figure 2: Correlation between salary and other variables

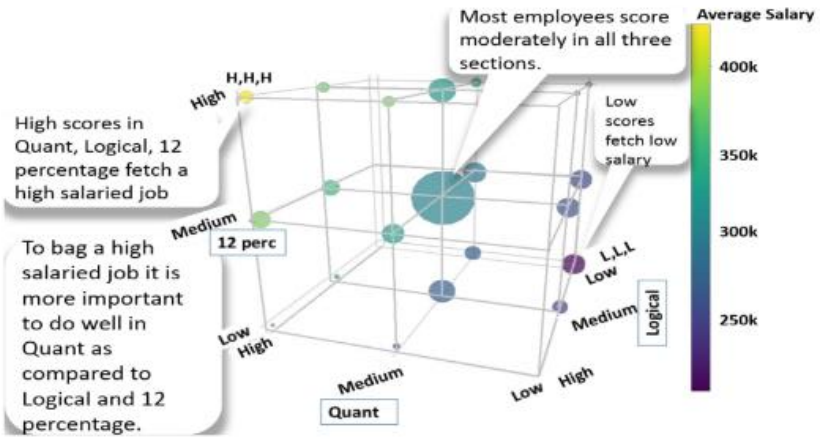


Figure 3: Average salary of employees against high, medium, low scores in Quant, Logical and 12 percentage

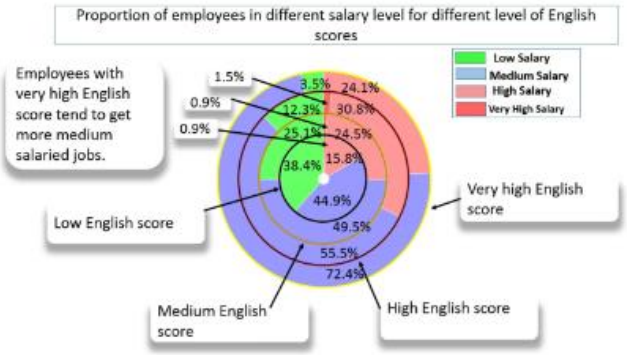


Figure 4: Proportion of employees in different salary for different level of English scores

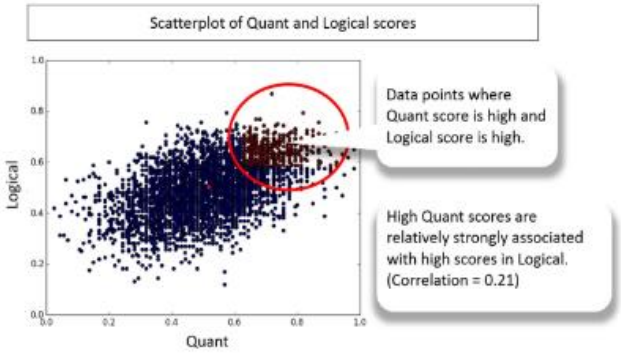


Figure 5: Scatter plot of Logical and Quant scores

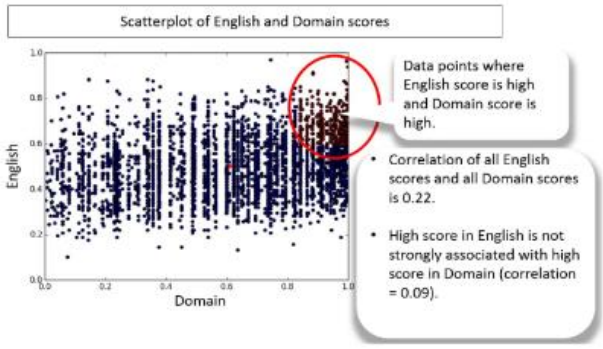


Figure 6: Scatter plot of Logical and Quant scores

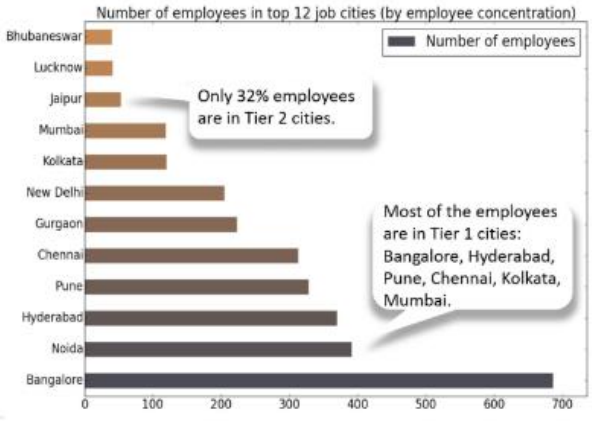


Figure 7: Number of employees in different job cities

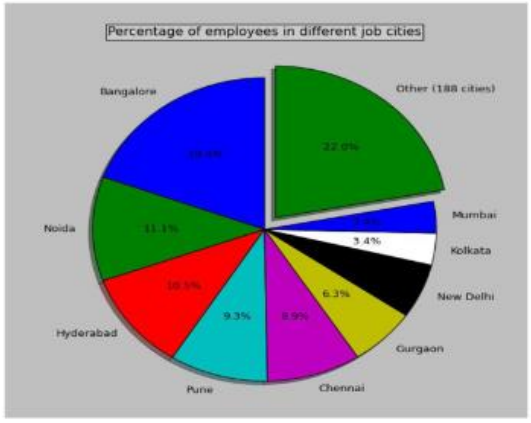


Figure 8: Proportion of employees in different job cities

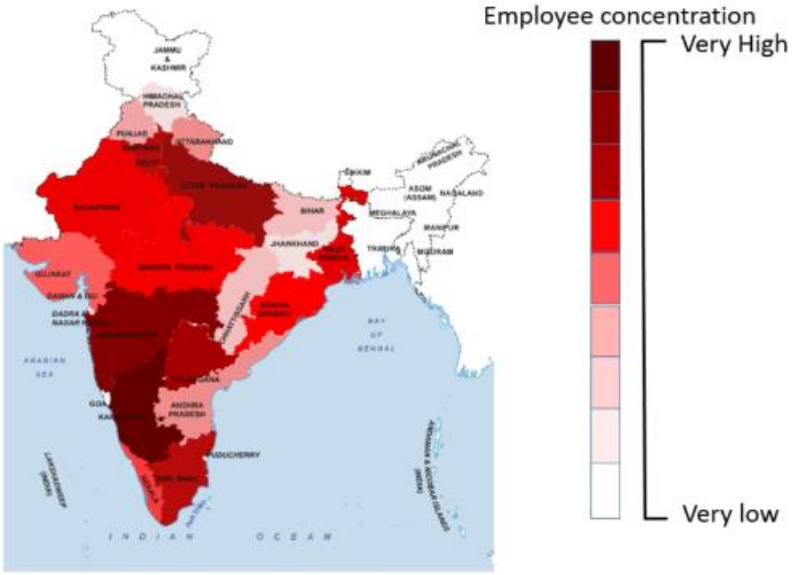


Figure 9: Concentration of employees in different job states(Shades of red signify the concentration of employees. Dark red is very high and white is very low)

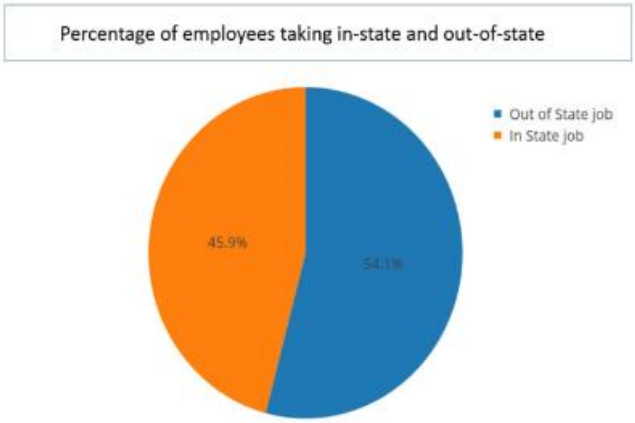


Figure 10: Proportion of employees with In-State Jobs and Out of-State Jobs

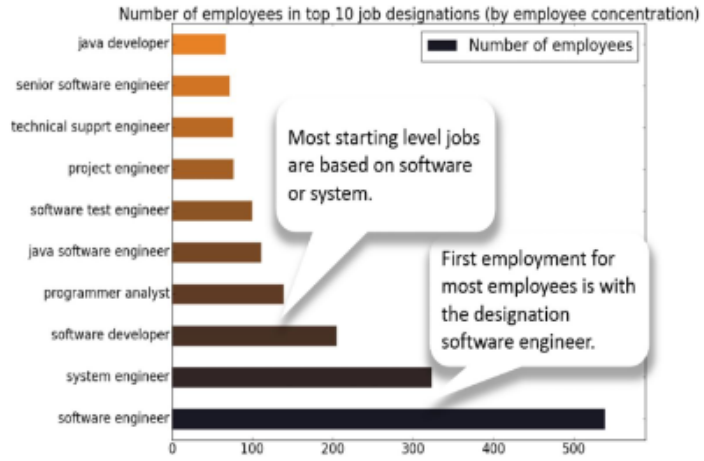


Figure 11: Number of employees in top 12 job designations

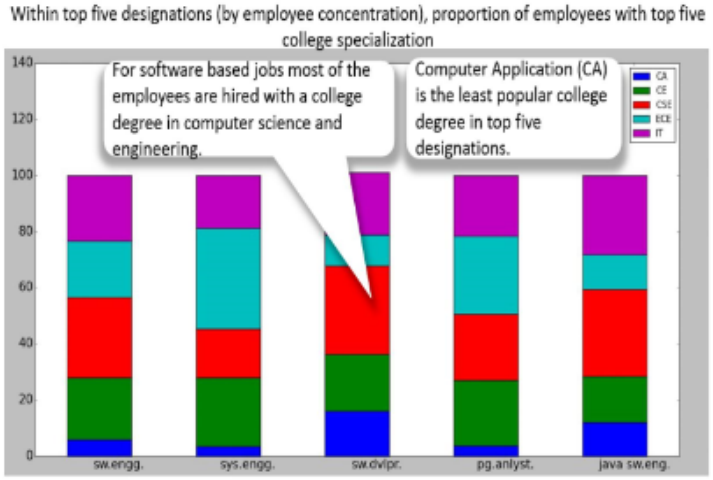


Figure 12: Within top five designations, proportion of employees with top five college specialization

Figure 13: Comparison of average scores in Quant, Logical, English, Domain in top five highest paid jobs