Understanding the Indian Labour Market: A Data Centric Approach

Shabana K M, Tony Gracious, Hrishikesh Subramonian
R&D Department
Flytxt
Trivandrum-695581, India
shabana.meethian,tony.gracious,hrishikesh.subramonian@flytxt.com

ABSTRACT
India produces 1.5 million engineers every year. Identifying the significant factors that influence the salary and the jobs these engineers are offered can help us understand the inefficiencies or skill gaps in the labor market, which will be extremely useful for policy making and constructive interventions. Predictive modelling using different machine learning techniques was performed on a data set that included both employee profiles and their employment outcomes. Feature analysis, correlation analysis, t-test and decision tree analysis were performed to identify the significant factors that influence the annual salary offered to a candidate. Visualizations generated based on employee salary, designation and job city revealed interesting insights.

Keywords
Data Mining, Predictive Modelling, Feature Analysis, Decision Tree Analysis, Visualization

1. INTRODUCTION
As the Indian economy grows at a decent pace, the demand for skilled labour is also on a significant rise. Even though a large number of students graduate from colleges in India every year, only a small fraction of them are considered employable. According to National Skill Development Corporation (NSDC), the growing skills gap in India is estimated to be more than 250 million workers across various sectors by 2022 [1].

Hence it is very important to analyze and understand the skill gap in the labour market and make suitable corrective measures wherever required. A data centric approach to this analysis involves the study of data involving both the student profiles and their employment outcomes. Identifying the significant factors affecting the annual salary offered to a candidate could give important pointers to the various skills expected by the industry and would also provide insights on the factors which influence the return in the labor market. A study in this line has been conducted using the AMEO 2015 [2] dataset. This dataset on Aspiring Minds' Employability Outcomes captures the academic and demographic information of engineering undergraduates appearing for AMCAT, Aspiring Minds's test of job skills, along with the employment outcomes (annual salaries of students' first jobs).

Predictive modelling using machine learning algorithms was performed to predict the salary a particular undergraduate would get on graduating based on various features such as AMCAT scores, personal information, pre-university information, university information and demographic information. The significant factors that influenced the annual salary offered to a candidate were studied using a variety of techniques such as feature analysis, correlation analysis, t-test and decision tree analysis. Visualizations generated based on job locations, designations, salary and other independent variables revealed interesting insights.

2. PREDICTIVE MODELLING
Developing a good machine learning model for salary prediction involved performing outlier removal, feature engineering and application of a suitable supervised learning technique.

2.1 Outlier Removal
Outlier removal was performed by removing observations with salary greater than 10 lakh (99.05th percentile). The chances of a candidate being offered a salary greater than 10 lakh on his/her first job is quite less and in many cases, on examining the designations it was felt that the candidate must have mistakenly added an extra zero while mentioning their annual salary. Hence only the observations with salary less than 10 lakh were used for training.

2.2 Feature Engineering
As a part of feature engineering, exploratory data analysis was performed to identify features that had a significant influence on salary. A detailed description of the experiments performed as a part of this exercise can be found in the next section. Feature extraction was performed based on the results of this study. Some of the meta features added after experimenting with various polynomial features include:

- acadperf : Quant * Logical * English * Class 10 percent * Class 12 percent * College GPA
- DC : Domain * Computer Programming

- QLE : Quant * Logical * English
- GpaSqrt: square root of GPA

For categorical features, the values with a minimum frequency of occurrence were retained as such and the others were converted to a single value. These features were then represented using one-hot encoding.

2.3 Salary Prediction

Supervised machine learning techniques such as ridge regression, random forests, etc. were used to learn a model for salary prediction using the new feature set. The models were implemented using scikit-learn [3]. It was observed that a random forest regressor with 300 estimators gave the best result on the test set.

3. DATA INSIGHTS AND RECOMMENDATION

3.1 Decision Tree Analysis

For the decision tree analysis, samples were divided into three salary classes.

- Salary <3 lakhs : Class A
- Salary >=3 and <=4 lakhs : Class B
- Salary >4 lakhs : Class C

The features CollegeID, CollegeCityID, CollegeState, Board10, Board12 and degrees Mtech, MSc(Tech) were excluded from the analysis as the dataset did not capture their distribution well. All the observations with less frequent specializations were also excluded. Feature engineering was done for categorical data with one-hot encoding.

A decision tree of depth 4 with minimum samples 100 for split node and minimum samples of 20 per leaf node was learned on the dataset using scikit-learn [3] with Gini impurity as the criteria for split .

At each node a split that reduces the Gini impurity score the most is found. The decision tree learned from the data has been visualized in Figure 1. Each box contains the condition of split, Gini impurity measure, number of samples at that node, number of samples per class at that node and the name of the majority class at that node. The color and its intensity at each node is determined by the majority class and its percentage at that node.

The important insights gained from this analysis are the following:

1. The most important features that determine the salary offered to a candidate are his/her quantitative skills, class 10 percentage, graduation year, college GPA, computer programming skills and English skills
2. Quantitative skill is very important as 61% of samples with quantitative score less than 524.5 (53rd percentile of quant score) were offered annual salary less than 3 lakh

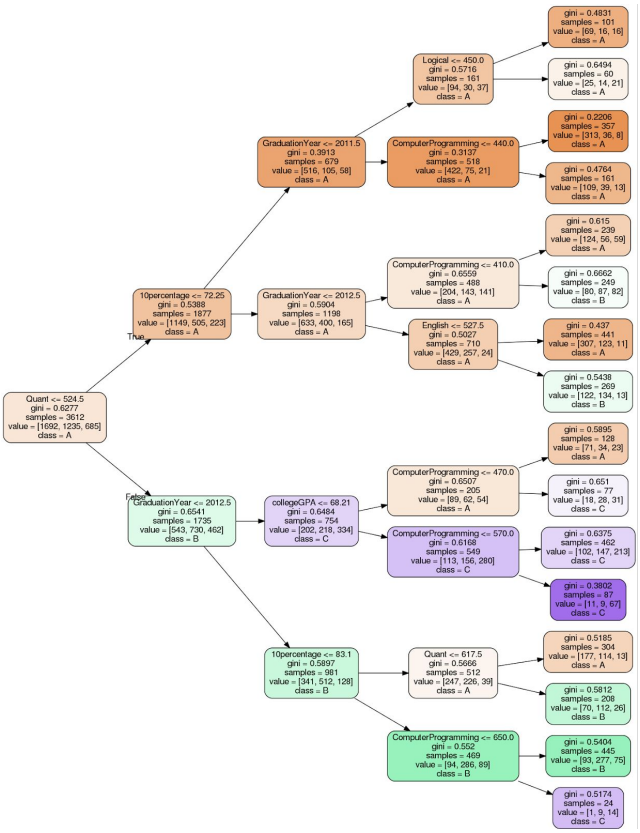

Figure 1: *Decision Tree Constructed based on Salary Classes*

3. Computer programming skills help in improving salary
4. Candidates with annual salary greater than 4 lakh have good quantitative score, college GPA and computer programming skills

3.2 Feature Analysis

The most important features as identified by the random forest regressor are given in Table 1.

The feature with the highest score is acadperf which implies that continuous good performance in school and college along with good quantitative, logical and English skills contributes to a higher salary. Graduation Year has a negative influence on salary as the median salary has been found to decrease over the years 2010-'15.

Pearson Correlation test and Welch Two Sample t-test in R [4] were also used to identify the important factors that influenced the salary of a candidate. The plots were generated using the R package ggplot2 [5] The main inferences drawn from the analysis have been listed below:

1. It was observed that the annual CTC of a candidate has significant positive correlation with his/her Quant, Logical and English scores. This was verified using Pearson Correlation test. It was also observed that

Table 1: Top Features and Importance Scores

Feature	Percentage Score
acadperf	42.51
GraduationYear	18.21
DC	4.15
QLE	2.24
openess_to_experience	2.20
CollegeCityID	2.07
X12graduation	2.00
CollegeTier	1.99
X12percentage	1.98
CollegeID	1.85

candidates with higher Quant, Logical and English scores tend to have higher salaries

2. The annual salary offered to a candidate was found to have a significant positive correlation with the product of the candidate's class X percentage, Class XII percentage, college GPA, quant, logical and English scores (Figure 2), as verified using Pearson Correlation Test. This observation implies that candidates with consistent academic records across school and college with good quantitative, logical and English skills tend to land up in a job with a good salary

3. Candidates studying in tier 1 colleges were observed to have significantly higher salaries as compared to candidates from tier 2 colleges, based on Welch two sample t-test with unequal variance

4. Candidates from Jharkhand and Karnataka were observed to have significantly higher salaries as compared to students from other states. This was also verified using Welch two sample t-test with unequal variance

5. It was interesting to note that candidates with civil engineering specialization had higher salaries as compared to students with other specializations as evident from the box plot in Figure 3.The same was also verified using Welch two sample t-test

6. Candidates with good domain knowledge and computer programming skills tend to earn a higher salary, as indicated by a Pearson Correlation Test

7. Candidates who passed class 12 after 2004 tend to have lower salaries with each passing year, as verified by a Pearson Correlation test. In similar lines, the candidates who graduated after 2010 were observed to have lower salaries with each passing year, as could be seen from Figure 4

8. Candidates who studied CBSE syllabus in their 10^{th} or 12^{th} tend to have higher salaries as compared to the other students. This was verified using a Welch two sample t-test and could be observed from the density plot in Figure 5. The CBSE students were also observed to have significantly higher quant, logical and English scores as compared to the non CBSE students.

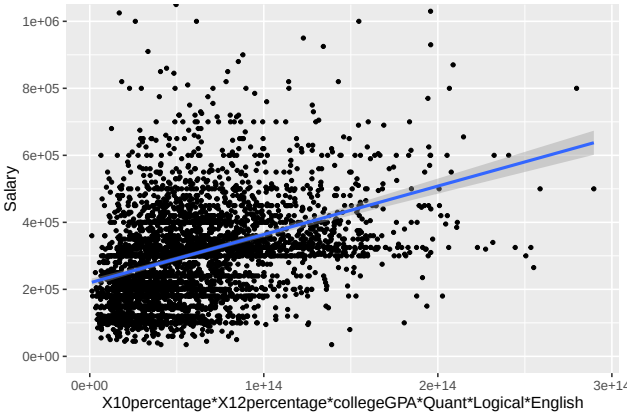

Figure 2: *Scatter plot of Class X percentage * Class XII percentage * College GPA * Quant * Logical * English against Salary*

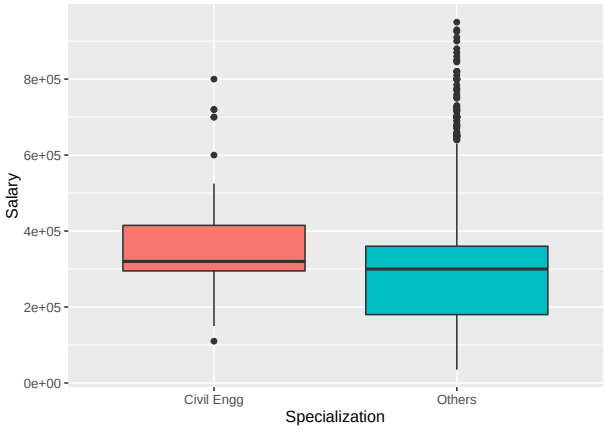

Figure 3: *Box plot of Salary of Civil Engineering graduates*

4. VISUALIZATIONS

It was observed that the job locations provided in the data set was noisy. Hence the data was cleaned by computing Levenshtein distance (Edit Distance) between the given job locations and the names of major cities in India and replacing the job locations with the city names having the minimum distance. After cleaning, only the job locations with a minimum frequency of 10 were considered for further analysis.

The median salary of employees per location was computed and plotted on a map of India as circles with radius and color proportional to the median salary at each location, as shown in Figure 6. The median salary was chosen instead of mean salary as the median was less susceptible to the presence of outliers. From the figure, it could be noted that the highest median salaries were offered at Mumbai, Pune and Bengaluru. Similarly the number of employees working

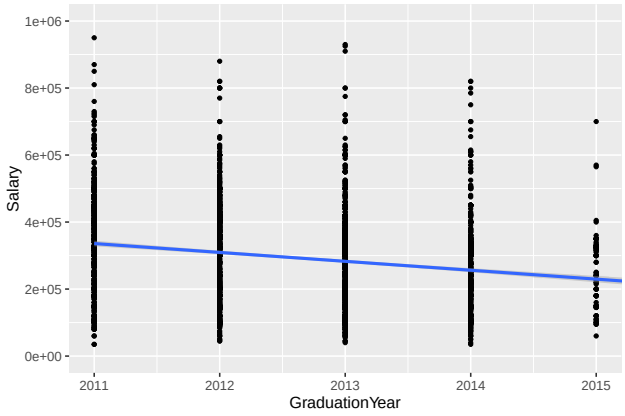

Figure 4: *Scatter plot of Graduation Year against Salary*

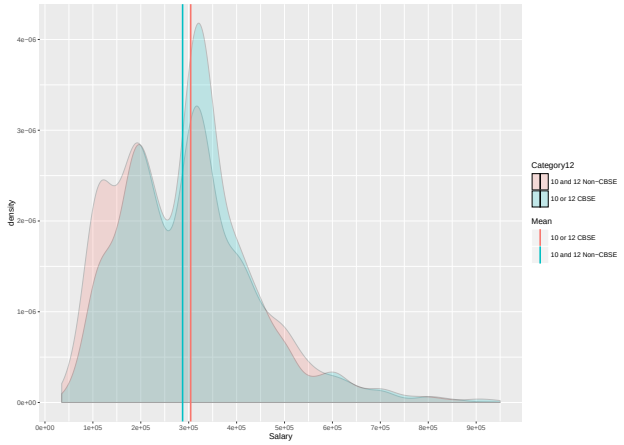

Figure 5: *Density plot of Salary for CBSE and Non-CBSE students*

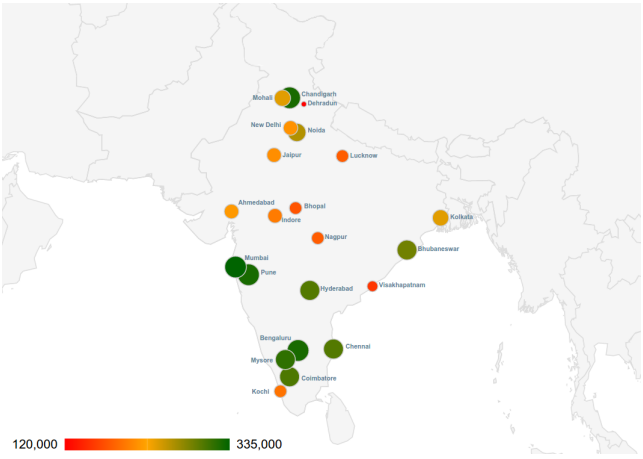

Figure 6: *Median Salary of Employees at each Location*

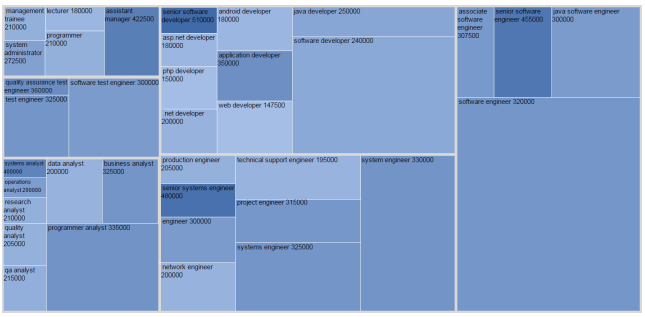

Figure 7: *Tree-map of Designations*

	Ahmedabad	Bengaluru	Bhopal	Bhubaneswar	Chandigarh	Chennai	Dehradun	Gurgaon	Hyderabad	Indore	Jaipur	Kochi	Kolkata	Mumbai	Mysore	New Delhi	Noida	Pune
Delhi	0.75	6.72	0	0	2.99	4.48	0	22.39	3.73	0.75	0.75	0	0	1.49	0.75	21.64	26.12	7.46
Punjab	0	12.32	0	0	7.97	5.07	1.45	19.57	2.17	0	0	0	2.17	2.9	0.72	4.35	23.91	9.42
Haryana	0	5.8	0	0.72	0.72	0.72	0	26.09	2.17	0.72	0.72	0	0.72	4.35	0.72	21.74	20.29	6.52
Uttar Pradesh	0.27	12.67	0.13	0.13	0.81	3.23	0.27	10.24	3.23	0.13	0.54	0.13	0.67	3.77	0.54	16.17	33.02	5.93
Uttarakhand	0	15.22	0	0	2.17	3.26	7.61	11.96	4.35	0	2.17	0	1.09	2.17	2.17	10.87	22.83	7.61
Himachal Pradesh	0	30.77	0	0	15.38	0	0	7.69	0	0	0	0	0	7.69	0	15.38	15.38	7.69
Rajasthan	1.49	10.45	0	0	0.75	3.73	0	8.21	1.49	0	29.85	0	1.49	7.46	0	5.97	14.93	13.43
Madhya Pradesh	1.92	14.1	11.54	0	0.64	1.92	0	5.77	5.13	13.46	0.64	0	0.64	11.54	1.92	3.21	5.13	22.44
Maharashtra	0.47	5.16	0.47	0.47	0	2.35	0	0.47	0.47	0	0.47	0	0.94	21.13	0.94	0	0.47	61.5
Bihar	0	16.67	0	0	0	16.67	0	0	16.67	0	0	0	0	0	0	16.67	33.33	0
Telangana	0	12.09	0	0	0	6.96	0	0	74.73	0	0	0.73	0.37	1.47	0.73	0	0	2.2
Andhra Pradesh	0.52	28.35	0	1.03	0	14.43	0	0	43.3	0	0	0.52	0	2.06	1.55	0	0.52	3.61
Orissa	0.72	23.19	0	24.64	0	4.35	0	5.07	9.42	0	1.45	0.72	9.42	3.62	2.9	0.72	2.9	10.87
Tamil Nadu	0	19.41	0	0	0	62.83	0	0.66	1.32	0.33	0.33	0.66	0.66	1.97	0.99	0.99	1.32	1.97
Union Territory	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
Gujarat	61.54	7.69	0	0	0	0	0	0	0	0	0	0	0	15.38	7.69	0	0	7.69
Assam	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Goa	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Karnataka	0	80.49	0	0	0	1.52	0	0.61	1.83	0	0.3	0	1.52	3.35	2.74	2.44	0.91	3.96
Jammu and Kashmir	0	40	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	20
Sikkim	0	0	0	0	0	0	0	66.67	0	0	0	0	0	0	0	33.33	0	0
Kerala	0	27.78	0	0	0	5.56	0	0	5.56	0	0	38.89	0	5.56	0	11.11	0	5.56
Chhattisgarh	0	41.18	0	0	0	17.65	0	11.76	5.88	0	0	0	0	11.76	0	0	0	11.76
Jharkhand	0	31.58	0	0	0	0	0	5.26	5.26	5.26	0	0	10.53	0	5.26	5.26	21.05	0
West Bengal	0	22.98	0	0.62	0	3.73	0	1.86	3.73	0	0	0.62	50.93	3.11	1.86	0	4.35	5.59
Meghalaya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0

Figure 8: *Heat map depicting the patterns in job location and college state*

at each location was also plotted and it was observed that a good fraction of employees worked at Bengaluru, Noida and Hyderabad.

For the analysis of job designations, k-means clustering was performed using a bag-of-words approach. The number of clusters was determined by using the elbow method. The most popular designations in each cluster and the median salary of each designation was visualized using a tree-map, as seen in Figure 7. The area of each rectangle is proportional to the number of employees with the designation and the color is representative of the median salary offered for the designation. It could be observed from the visualization that 'Software Engineer' is the most common designation with a median annual salary of 3.2 lakh INR. The highest median salary was offered for 'Senior Software Developer'.

The relation between the job location and college state was also explored and visualized using a heat map (Figure 8). It could be observed from the figure that most employees worked at the nearest industrial hub with respect to the college state. For instance, a good majority of the students in the northern states worked either at New Delhi, Noida or Gurgaon.

By using the join year and year of leaving of employees, the

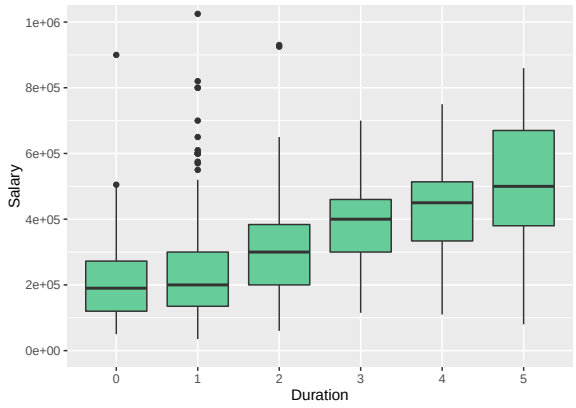

Figure 9: *Box-plot of First Job Salary across Duration at Work in Years*

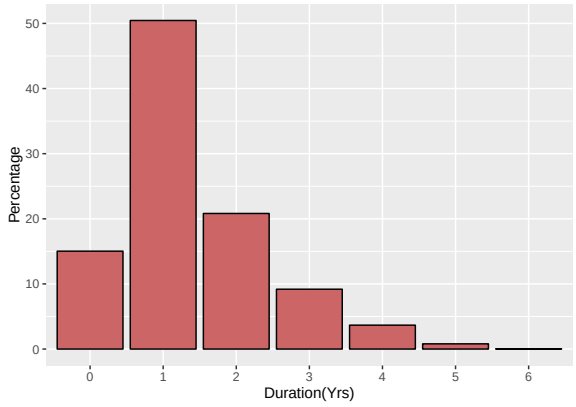

Figure 10: *Histogram showing the Number of Years employees spend at their first job*

duration (in years) spent at the first job was computed and analyzed. The duration at work was inversely related to the annual CTC offered, as could be seen from Figure 9. It could be also observed from Figure 10 that about 50% of the employees spend only about one year at their first job.

The median salaries offered for various designations over 2010 to 2015 could be seen in Figure 11. The size of points is proportional to the number of employees with that designation. It could be observed from the plot that the median salary of business analysts have remained almost stable from 2013-'15 whereas for most of the other designations there has been a negative trend.

The median salary offered to civil engineering graduates have remained almost stable from 2012-'15 whereas a clear negative trend could be observed for mechanical engineering, information science engineering and electronics and communication engineering graduates (Figure 12).

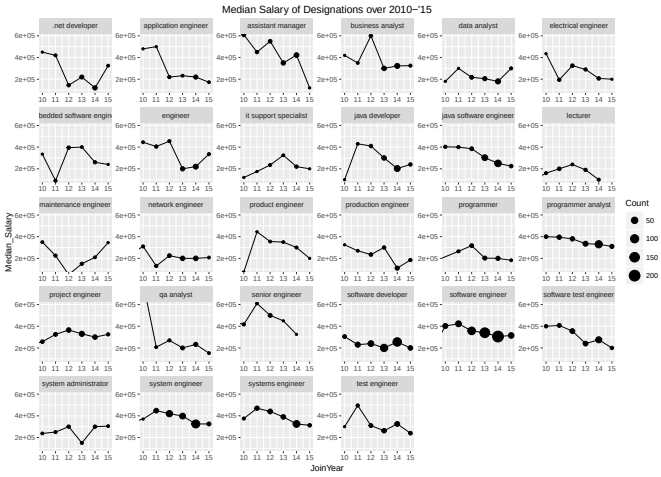

Figure 11: *Median Salaries offered for various Designations over 2010-'15*

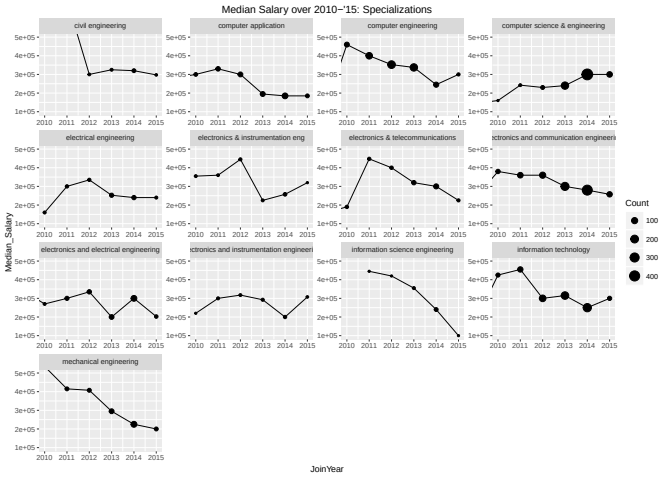

Figure 12: *Median Salaries offered for various Specializations over 2010-'15*

5. REFERENCES

[1] The skills gap and what it means for your business. `http://www.financialexpress.com/article/industry/jobs/the-skills-gap-and-what-it-means-for-your-business/138700`. Accessed: 2016-01-29.

[2] A. Minds. Aspiring minds employment outcomes 2015, 2015.

[3] F. Pedregosa and G. Varoquaux et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[5] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.