Understanding Labor Markets

Saurabh Banerjee
Sapient Global Markets
(+91) 9980072561
sbanerjee3@sapient.com

Paviya Chemparathy
Sapient Global Markets
(+91) 9845151902
pchemparathy@sapient.com

Yaasna Dua
Sapient Global Markets
(+91) 9971652910
ydua2@sapient.com

Kanishk Agarwal
Sapient Global Markets
(+91) 9901117618
kagarwal4@sapient.com

ABSTRACT

Employability has been a concern for Indian college graduates. The IKDD CODS 2016 data challenge is a great opportunity to get an empirical understanding of factors that impact employment outcome for Indian engineering graduates.

In this paper, we have elaborated our approach for data understanding, data preparation, predictive modelling and model evaluation for predicting salary of new graduates entering the labor market. We have also included some insights from our analysis.

The Salary range is wide, varies in a non-linear manner with most of the input features. Thus, the best algorithm to predict salary should consider the non-linearity and the outliers.

We have detailed our approach and observations in the following sections.

Keywords

Regression; Predictive Modeling; MARS; AMCAT; Salary; Prediction; IKDD; Data Science

IKDD CODS ' March, 2016, Pune, Maharashtra, India.

1. INTRODUCTION

The dataset provided by Aspiring Minds has 5 dependent variables and 32 independent variables. The training dataset consisted of 3,998 records and test dataset has 1,500 records. The dependent variables DOJ, DOL, Designation and Job City are used only for data exploration since these are not available in the test set. The dataset contains some free form self-reported text information containing inconsistent or incorrect data (e.g. 10th and 12th board names, Job city, designation, specialization). The data had to be cleansed for our analysis.

We have used R for our analysis and followed the CRISP-DM methodology for data mining.

2. DATA UNDERSTANDING & PREPARAITON

In this section we have detailed statistical analysis performed to understand the dataset and also mentioned the data preparation steps on the some of the attributes.

2.1 Salary

Salary is spread across a range of ₹35k - ₹40lac and is right skewed with the median at₹3lac and the interquartile range from ₹1.8lac to ₹3.7lac.

2.2 10^{th} Percentage & 12^{th} Percentage

The 10^{th} and 12^{th} percentage has no missing values and outliers. 10^{th} percentage is slightly left skewed and 12^{th} percentage follows normal distribution. 10th and 12^{th} percentages have higher correlation with salary compared to college GPA.

2.3 Collage GPA

College GPA has max value of 99.93 and minimum value of 6.45. This indicated that the values are not in the same scale. Some of the college GPA could be in percentile and some in absolute numbers. There was no indicator to distinguish those. All the values which were below 10 were converted to 100 point scale.

2.4 Quant, Logical and English

Quant, Logical and English scores don't have any missing values. The data follows normal distribution with no skews. These AMCAT scores show a high correlation with the salary and Quant has the highest correlation.

2.5 Domain

Domain score has 246 missing values and was imputed using the Quant, logical and English scores. We found a good correlation between the domain and these attributes hence created a linear regression to generate the missing domain scores. The data follows a bimodal distribution.

2.6 College Tier

90% of the students in this dataset are from Tier 2 colleges. In the data given that the median salary of the Tier 1 college graduates is higher than the Tier2 college graduates.

2.7 Graduation Year

The dataset contains graduates with different years of experience. As expected, the salary range is less for graduates with less years of experience.

2.8 Gender

Approximately 75% of the graduates in this dataset are males. The data indicates there is no salary variation based on gender.

2.9 Degree

Approximately 92% of the graduates in this dataset has B.Tech/BE degree. There are only 2 MSc. (Tech) degree holders which we merged with M. Tech. The data shows that the median salary of B.Tech / BE graduates is higher than MCA post graduates.

2.10 Specialization

This categorical field had 50 unique values which were grouped into few specializations like Computer, Electronics, Mechanical, Civil and Chemical. All others were grouped into one separate category as "Others". 55% of the students in this dataset are from computer branches and 35% from electronics.

2.11 10^{th} Board and 12^{th} Board

10^{th} board and 12^{th} board had more than 200 unique categorical values. We grouped them into 3 categories, namely CBSE, ICSE and state board.

2.12 Feature Creation

2.12.1 Aggregate Percentile:

Attributes like ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg had sparse data. To consider these attributes in our model, we decided to create a new attribute named aggregatePercentile by merging the scores in these 7 subject test. These individual scores were converted to percentile scores and then we took the mean to get the aggregate percentile. This new feature showed a higher correlation with Salary than Domain.

2.12.2 ImprovementIn12 and ImprovementInCollege:

Added two new features by finding the difference between scores in 10th and 12th and College GPA

2.12.3 Age

Added age as a feature by subtracting DOB from 2015

3. Modeling and Evaluation

After performing the tasks of exploratory data analysis, data cleaning and new feature creation, we explored different regression models, starting with simpler models and then more complex ones. In the following sections we have documented our observations using linear regression, SVM and MARS. Models were developed with the transformed dataset after data preparation.

We split the dataset into training and validation sets. For evaluating the model we have looked at the R-squared and the RMSE values. We also examined the residuals.

3.1 Model 1: Multiple linear regression

We used the R caret package to train the lm model which shows the importance of different features. We also validated the feature importance using the step function and selected the best features to refine the model. We also removed salary outliers.

Attributes used:

Salary ~ Quant + GraduationYear + X10percentage + AggregatePercentile + collegeGPA + English + X12percentage + CollegeTier + CollegeCityTier + conscientiousness + extraversion + openess_to_experience

Model output:

R-Squared from the model was 0.1022
RMSE calculated over the validation set was 191575.2

The residuals are well distributed for most of the salary range but had uneven distribution for salaries greater than ₹6 lac.

3.2 Model 2: SVM

Linear Regression only works for linear relationships. To explore non-linearity we used SVM with radial kernel. We used R package e1071 to train the SVM model.

Attributes used:

Salary ~ Quant + GraduationYear + X10percentage + AggregatePercentile + collegeGPA + English + X12percentage + CollegeTier + CollegeCityTier + conscientiousness + extraversion + openess_to_experience

RMSE calculated over the validation set was: 187550.6

RMSE improved slightly. However, based on our analysis of residual plot, SVM model was unable to predict salaries greater than ₹5lac properly.

3.3 Model 3: MARS

In the dataset we had 24 students with an annual salary greater than 12lakh. The models we have explored so far were not able to predict the higher salary ranges. We were able to achieve better results with Multivariate adaptive regression splines (MARS). MARS is a non-parametric regression technique that automatically models nonlinearities and interactions between variables. We used the implementation in R package "earth".

Attributes used:

Salary ~ X10percentage + X12graduation + X12percentage + CollegeID + CollegeTier + collegeGPA + CollegeCityID +

CollegeCityTier + GraduationYear + English + Logical + Quant + Domain + conscientiousness + agreeableness + extraversion + nueroticism + openess_to_experience + AggregatePercentile + X10CBSE + X10ICSE + X10Stateboard + X12CBSE + X12ICSE + X12Stateboard + Btech + Mtech + MCA + numericalGender + computer + biotechnology + electronics + mechanical + chemical + aeronautical + other + civil + Andhra.Pradesh + Assam + Bihar + Chhattisgarh + Delhi + Goa + Gujarat + Haryana + Himachal.Pradesh + Jammu.and.Kashmir + Jharkhand + Karnataka + Kerala + Madhya.Pradesh + Maharashtra + Meghalaya + Orissa + Punjab + Rajasthan + Tamil.Nadu + Telangana + Union.Territory + Uttar.Pradesh + Uttarakhand + West.Bengal + YOB + improvementInCol + improvementIn12

Final Parameters after grid search:

Degree =1, nk =19, trace = .5, penalty = -1, thresh =0.001,nfold=0

	coefficients
(Intercept)	528892.6
CollegeTier	-81000.3
h(74.5-X10percentage)	-632.9
h(X10percentage-74.5)	2479.1
h(62.9-collegeGPA)	-7484.3
h(collegeGPA-62.9)	1948.2
h(2010-GraduationYear)	-123.6
h(GraduationYear-2010)	-60151.6
h(GraduationYear-2013)	60309.5
h(265-English)	1687.9
h(English-265)	184.4
h(640-Logical)	-80.1
h(Logical-640)	1898.2
h(390-Quant)	234.2
h(Quant-390)	220.9
h(0.994051-Domain)	-55070.4
h(Domain-0.994051)	-14474562.2

Figure 1. MARS hinge functions

Model Output(figure 1 and figure 2):

65 predictors were provided as input for the model. Out of the 65 predictors, the model used 8 predictors based on the importance.

The important predictors identified were Quant, Graduation Year, X10percentage, CollegeTier, Domain, English, CollegeGPA, and Logical. These can be seen in the hinge functions and the corresponding coefficients are shown below.

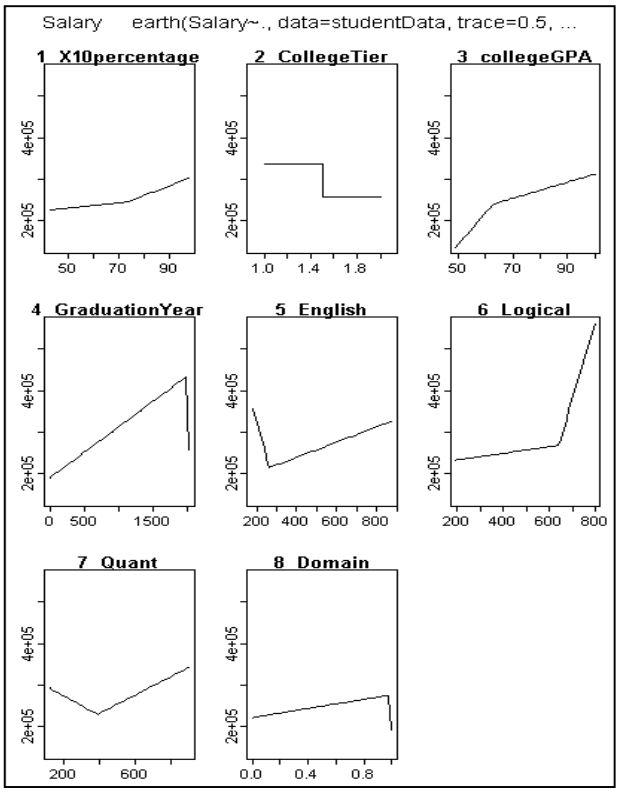

Figure 2. MARS Regression lines

R-Squared from the model was 0.1698
RMSE calculated over a cross validation test set was: **181600.4**

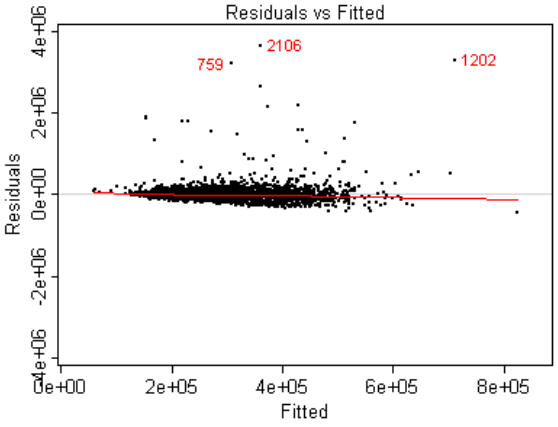

Figure 3. Residuals vs. Fitted for MARS

Interpretation from residuals vs. fitted (figure 3):

There is lot of variation in the salary (above ₹12 lacs) which the model is still not able to predict correctly. The model has ignored outliers trying to fit the mostly occurring values in the data set better. Hence the residual plot shows the value of the residuals as high as ₹35 lacs. It is able to predict better for salaries up to ₹8 lacs with minimal residuals.

Modifications tried on the feature set-

- There were a few outliers (Salaries above 6.5 lakhs) which we were not able to predict. We had 128 such outliers. We tried finding why these graduates got such high salaries, so that we could predict the high salaries better for the test data. Thus, we applied Decision Trees, Random forest to this data to see if there was a pattern, but we remained unsuccessful in finding any pattern.
- We thought that there would be some correlation between the work experience and the salaries, but unfortunately that data was not available for testing. Hence, we took the date of graduation and subtracted it from the year 2015 to get an approximate measure of the work experience which we included as an input to the model. However, this did not have any impact.

4. Conclusions

The best performing models from each category were compared against each other. MARS model with tuning parameters like degree, nk, penalty, thresh etc. resulted in better outcome. However, our MARS model is unable to predict salaries greater than ₹8 lacs. This can be because of fewer examples in that range and missing features which influence higher salaries.

Our analysis leads us to the conclusion that merit does count in determining the salary of a student. AMCAT quant scores and high school results are very good predictors for salary compared to the other inputs. College GPA is next best predictor of salary. The students pursuing specializations related to IT tend to get more salary than their counterparts in other branches.

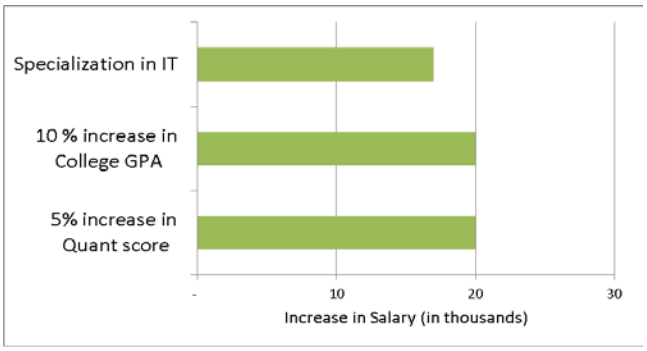

Figure 4. Salary Variations

Also, students from tier 1 colleges get better salary on average. This could be because companies who offer higher salary don't visit the college campus of lower tier colleges for placements. Thus students belonging to lower tier colleges don't get the opportunity to compete for high paying jobs. Also some companies may offer differential salary to candidates based on their college tier. CBSE students earn slightly more compared to state boards. State in which the college is located also has an impact on salary.

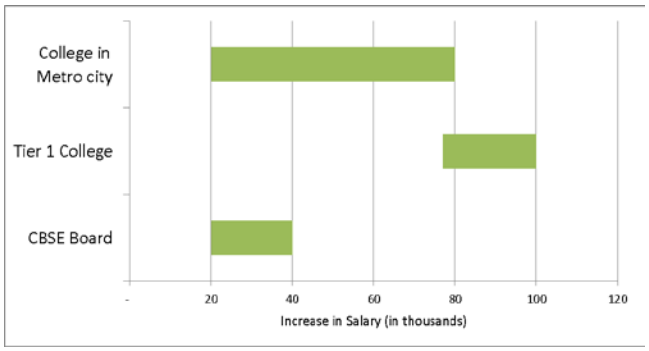

Figure 5. Range of Salary Variations

We did not find any impact of gender; personality test scores like extraversion, neuroticism and agreeableness on salary.

The salaries in tier 1 cities are higher than tier 2 cities for SW Development jobs. The difference in salary can vary between 45k and 60k annually. Meanwhile, the differences in average salary between tier 1 and tier 2 job cities are insignificant for Non-IT and Support jobs.

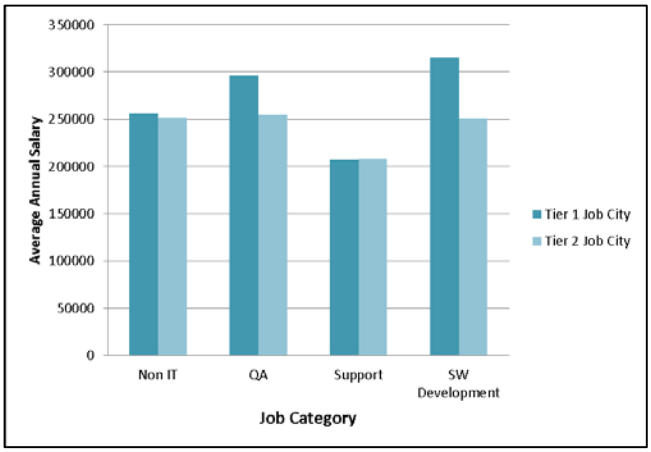

Figure 6. Average Salary by Job Category & Job City Tier

We did not find any impact of gender; personality test scores like extraversion, neuroticism and agreeableness on salary.

4.1 Salary Predictor (Model Deployment)

We developed an interactive user interface using the deployed model. Users can select different values for key predictors using easy to use sliding controls. Predicted salary is displayed in the salary dial by invoking the deployed model. The tool helps us gain better insight on how different inputs impact the predicted salary.

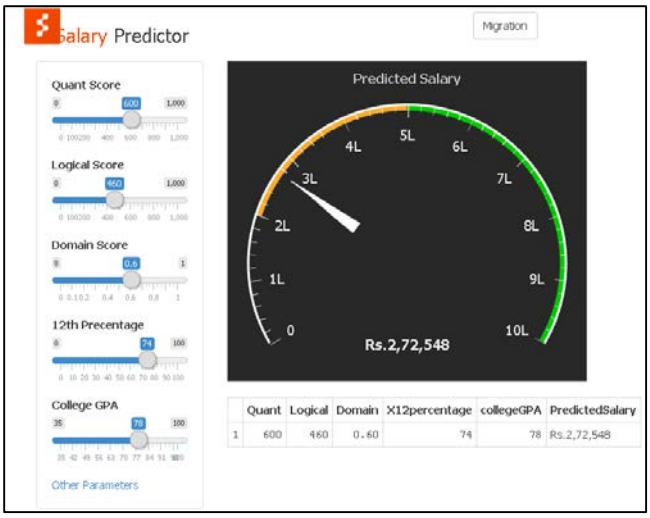

Figure 7. Salary Predictor

4.2 Data Visualization - Migration of Job seekers

We developed an interactive visualization depicting migration of students from their college location to their respective Job cities in a map of India. Users can select a "To" or "From" location to observe migration of job seekers. We can observe heavy migration to popular job destinations like Bengaluru and NCR.

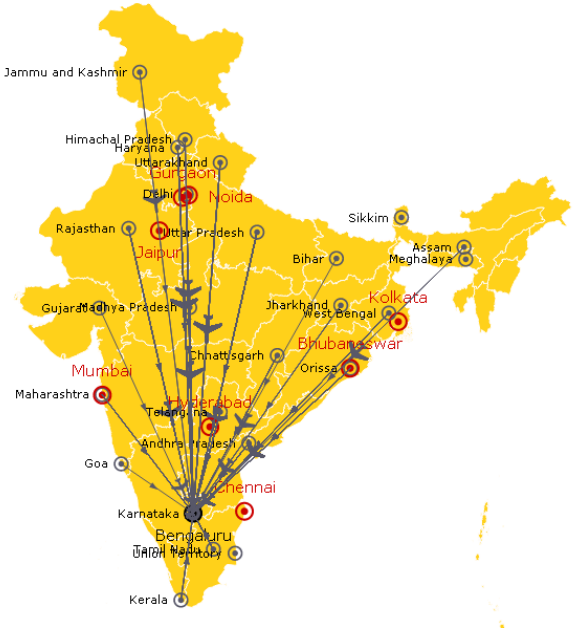

Figure 8. Map depicting job student migration to Bangalore

4.3 Areas of Applications

4.3.1 Recruitment

Profile scoring model can be used to customize salary at the individual level.

4.3.2 Career Counselling

Provide better guidance to students by matching student profile using applications leveraging quantitative models.

4.3.3 Education Reform

AMCAT quant and logical scores are the best predictors for salary. Quant scores are perceived to reflect employability. Education boards can emphasize higher order thinking skills from early schooling to develop quantitative and logical aptitude.

5. REFERENCES

- Earth: Multivariate Adaptive Regression Splines DOI= https://cran.r-project.org/web/packages/earth/index.html
- Caret: Classification and Regression training DOI = https://cran.r-project.org/web/packages/caret/index.html
- Which engineer gets a job? by Aspiring minds DOI= http://www.aspiringminds.com/research-articles/which-engineer-gets-a-job
- Five Technical Aspects of Compensation Data By Jonas Johnson DOI= http://www.erieri.com/PDF/Comp-JJ.pdf