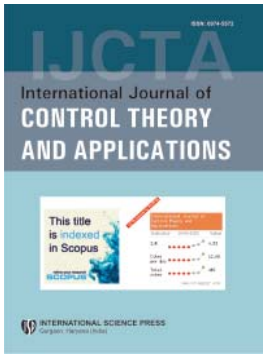

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 40 • 2016

Recruiting Parameters Recommended for students to be Considered Based on Data Analytics of Placement Trends in a University

Sasi Rekha Sankar[a] Anindya Ojha[b] and Upasana Pattnaik[c]

[a]*Asst. Professor, Department of Software Engineering, SRM University,Chennai, India.*

E-mail: sasirekha.s@ktr.srmuniv.ac.in

[b]*Student, Department of Software Engineering, SRM University, Chennai, India.*

E-mail: anindya.ojha.95@gmail.com, upasanapattnaik1995@gmail.com

Abstract: The placement trends recognition is a key role of all the placement officers in universities and colleges. The placement team identifies the appropriate student to face the interview of a software company based on the general parameters of CGPA and Arrears. The level of short listing and rejections can be reduced if they can map appropriate candidates to the software companies' requirements. The data analytics will be performed on the history of placement data and current year's placement trends to predict and map the students to the companies based on the criteria's like skill sets, interest, passion, requirements and may more. This method will reduce the rejection rate and increase the placement opportunities and also suggest the students to acquire the required skills for the forth-coming companies.

In this study, we have also employed a survey of alumni feedback to understand how college curriculum and individual academic prowess affect the employability of a student. Additionally, we have analyzed historical placement data from 2014 to 2016 to infer relevance of placement techniques and trends in recruitment over the years.

Keywords: *Data analysis, placement trends, college graduates, job.*

1. INTRODUCTION

In a country like India that boasts of a daunting population of a billion people, employment becomes a concern by sheer dint of the number of people it has to accommodate. More lucrative fields like engineering have to bear the brunt of the bulk as young students clamber to secure limited seats. With over a seven thousand options to choose from, the placement percentage of individual colleges is what ends up becoming the biggest discerning factor for students looking to turn their degrees into productive work. It is therefore, essential to analyze the placement process of colleges to see what makes one better than the other.

To narrow the analysis down, we concentrate our work on the placement data of SRM University. The main source of data being analyzed here is the historical placement data of the software engineering department from the year 2014-2016. Data analytics techniques like linear regression, logistic regression etc. with their various

key features have been employed in working on the data. With this project, we hope to assess how various factors namely GPA, history of arrears, round-wise strengths, curriculum, and placement training techniques etc. correlate and how the resultant correlation can be used to yield useful predictions and recommendations to improve the placement process in place.

2. DATA ANALYSIS AND RESULTS

The aim of the analysis in this paper is to predict the likelihood of students getting placed in a company. The prediction can be used by the department and especially the training & placement division of the software department to channel the students to the right company or train the students to make them attain their dream jobs if they consider them a potential candidate for a particular core company.

The placement division of the department gets the grades, arrears, history of arrears, 10^{th} & 12^{th} percentage from the individual faculty advisor; also they have a one on one session with individual students to collect information regarding their choice of company, the dream job profiles they prefer, company location, work culture and expected salary. The data gathered from the students can be mapped to a skill matrix.

The skill matrix will highlight the key potential a student possesses and point out the deficiencies they need to work on to attain their dream job.

We used linear regression to explore the relation between a dependent variable and one or more independent variables, through which we can predict the response of the dependent variable with respect to change in independent variable. Linear regression is represented with the equation y= c + bx, where y is the dependent variable, c is constant, b is coefficient of regression and x is an independent variable. This model allowed us to predict the number of students likely to be placed in a company.

In this analysis, the dependent variable is Cleared, which implies the number of students who have been placed in a company by qualifying eligibility criteria, aptitude tests, technical aptitude and personal interviews including both Technical and HR round. The independent variable Eligible represents candidates who have met the minimum requirements for individual companies. The next independent variable is Aptitude represents candidates who have cleared the aptitude round for individual companies. The Technical attribute of the independent variable represents the candidates who have cleared the technical round in the job interview. The dependent variable Cleared relies on the values of these independent variables.

We have analyzed the placement data of the Software Engineering department of a university from the year 2014 to 2015. The data received in excel sheet format.

2.1. Visualization

We visualized the data in 'Figure 1' to get a graphical representation of the companies' student intake from the year 2014 to 2016. The companies have been given pseudonyms to maintain privacy. For example, Company O only visited in the year 2014 during which less than 10 candidates were eligible and a miniscule fraction got placed (Cleared). In a second example Company G, an increase in eligibility and placement of candidates over the years can be observed. The inference that can be drawn from the data mentioned in the figure is that we can see the how frequently a company has visited the university. The purpose behind analyzing this data is to look for companies that only visited once and have failed to return in the span of two years and to find out the reason for it. The reasons can be elicited from company HR through a feedback system initiated by the university. The most common feedback received concerns the non-conformance of skill set with job profile and the students' response during interviews. 'Figure 2' can add to the credibility of the above statement as it also depicts the number of companies that visit each year and their recurrence or lack thereof. The analysis gathered from Company G can be further employed to find out the reasons for steady increase, the job profiles they offer, the interview process conducted via the same feedback process from HR and alumni.

The alumni feedback of Companies like 'G' can give deep insights to the university's placement department which can be used to aid the steady growth in placement pattern. This data will yield valuable information regarding a variety of important aspects namely – the kind of students who get placed, students who get retained in the company, reasons some students deferred the company when presented with multiple job offers, the reason for leaving the company within 6 months, the disappointments they faced in the company and based on all these attributes can map the right kind of student to this company. This can increase the value of 'Cleared' dependent variable by making alterations to the independent variables of the students' skills. In addition, the individual departments can make alterations to the teaching methods, placement training techniques, and curriculum so that they can focus on the core companies and mass recruiters for increasing their placement percentage each year. Thus, we get an overview of the placement statistics from 2014 to 2016.

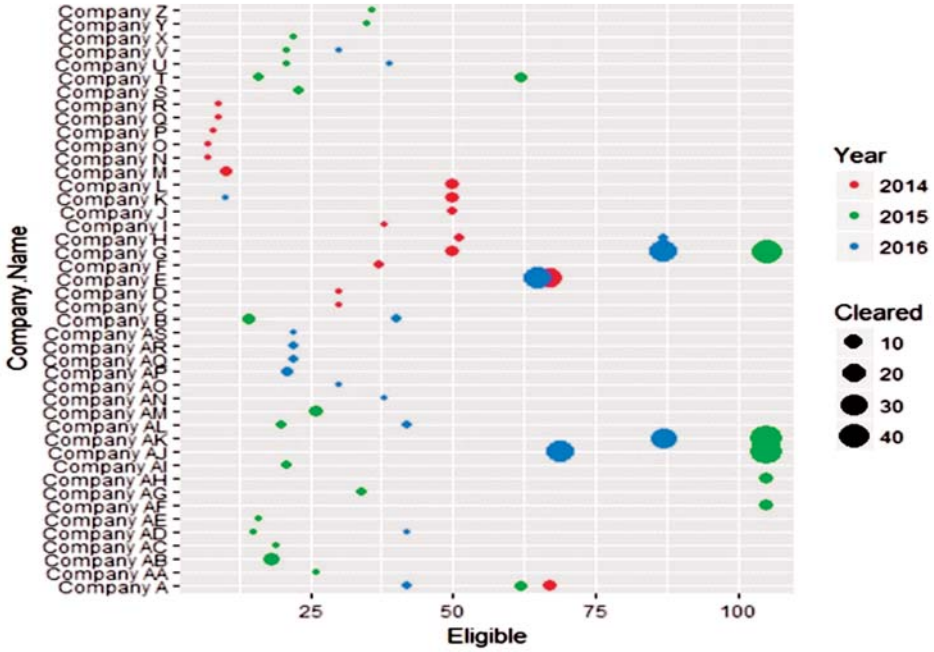

Figure 1: Eligibility vs. placement of candidates in companies from 2014 to 2016

As mentioned earlier, 'Figure 2' is a variation of the visualization provided by 'Figure 1'. The difference lies in how the company trends can be visualized where the number of companies and their reappearance can be gauged. Another main factor that can be obtained from this graph is the difference in placement pattern based on the company visit to the campus recruitment each year. The year wise demographics give the top management a view of the frequently visiting companies, impact in the proportion of the companies visited to the placed students. The 'Cleared' dependent variable can be compared with companies which have not recruited any students.

To analyze the data from a different perspective, we plot a graph of number of students who cleared various rounds to get placed against the count of companies. The graph is represented in 'Figure 3' and shows the intake of students in companies. We can see that single candidate placement is prevalent. This information can be valuable to the university's placement department in two ways. Firstly, these singular placed candidates and their profiles can be analyzed to gauge what helped them qualify companies that their peers could not. In other words, it will help them identify weaknesses in students as well as their training process. Secondly, by initiating a feedback system with the companies, the placement department can gain a better understanding on why these companies chose to take in a single candidate only. This answer, from the industry perspective will help them tweak their placement strategy and shift the graph towards a more favorable number. Predicting the candidate clearance, our next step is to train our model.

Figure 2: Number of companies and trend over the years

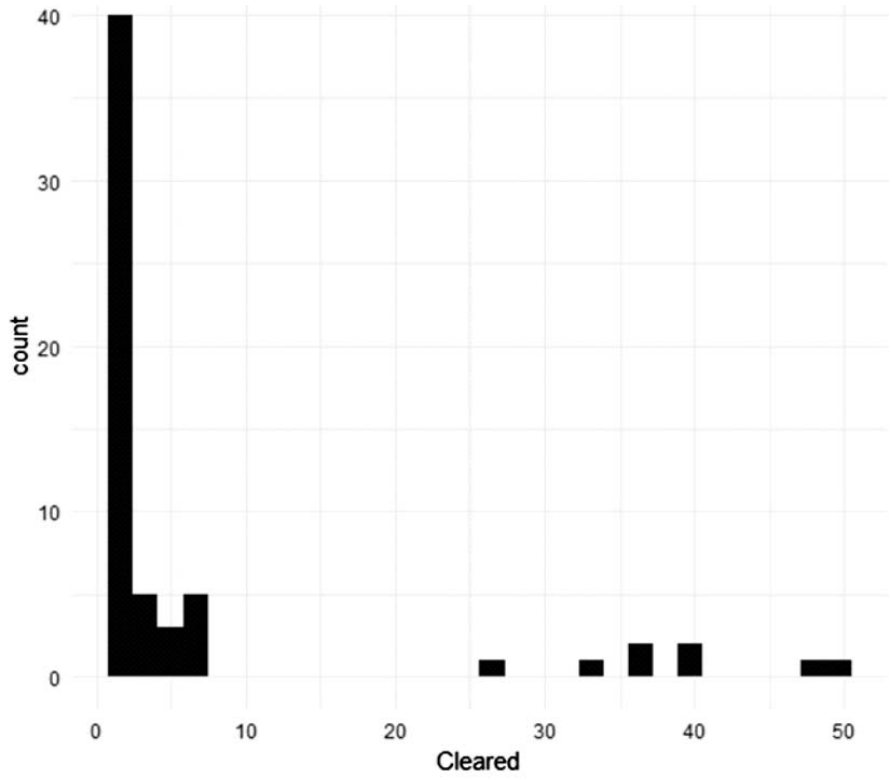

Figure 3: The count of candidates getting placed in companies

2.2. Analysis

The analysis was carried out by splitting the data into a training data and test data set. The training dataset is used to train the linear regression model. The data is fit into the linear regression model. What the model does is plot the response variable (Cleared) with respect to the explanatory variables and develop a pattern which can be predicted. The purpose of the test data set is to see how well the model fits and predict the values of the response variable.

Using R language to perform data analysis, our first action was to gauge the correlation between the fields.

Four major fields have been selected due to their relevance in predicting placed candidates. The 'Cleared' variable is the response variable which depends on the other three variables for prediction.

Eligible: Candidates meeting the company's eligibility criteria. They qualify to sit for the aptitude test.

Aptitude: Candidates who cleared the company's aptitude test and qualify for subsequent rounds.

Technical: Candidates who cleared the company's technical test and personal interview round.

Cleared: Candidates who got placed.

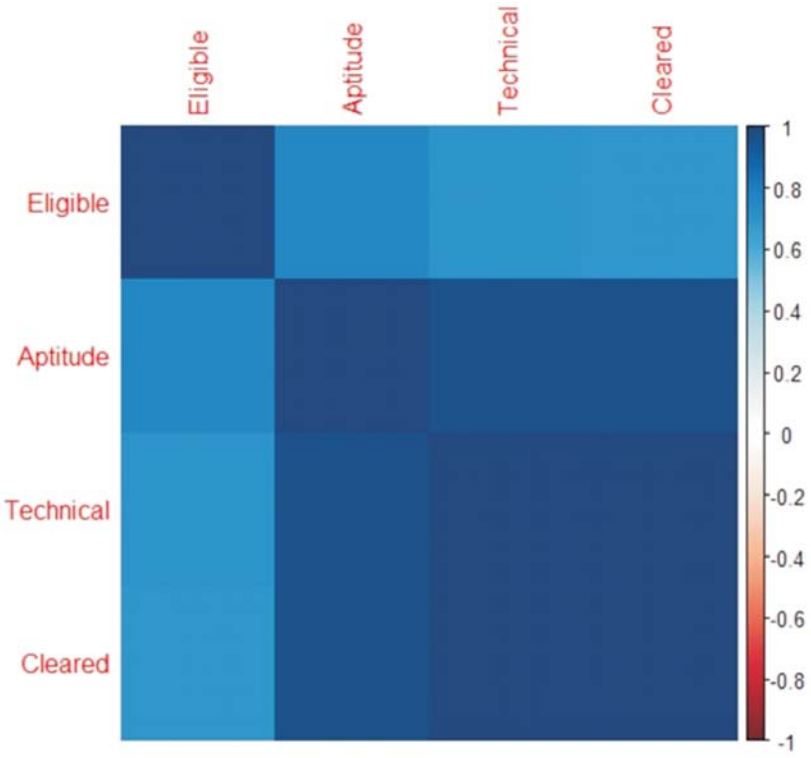

Figure 4: Correlation plot: Interdependence between fields

The data exploration results showed good correlation between the values of the Eligible, Aptitude and Technical fields with the Cleared field (Figure 4). The diagonal of the correlation plot (Figure 4) is ignored due to redundancy. It maps the same variables. We can observe that here is a very strong correlation between the Technical and Cleared variables. There is a strong relation between Aptitude and two fields- Technical and Cleared. There is a weak correlation between Aptitude and Eligible variables.

The inference gained from Figure 4 is the technical round results play a vital role in assuring the placement of a student. This information can be used by the placement department to increase focus on technical training of candidates. We can use these fields to train our model to predict the placement of the student.

We train our model with respect to the Cleared field and get the following model summary.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.443642   0.262009  -1.693    0.098 .
Eligible     0.004653   0.007190   0.647    0.521
Aptitude    -0.028659   0.031776  -0.902    0.372
Technical    1.037127   0.037297  27.808   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8938 on 41 degrees of freedom
Multiple R-squared:  0.9964,    Adjusted R-squared:  0.9962
F-statistic:  3818 on 3 and 41 DF,  p-value: < 2.2e-16
```

Figure 5: Model summary

'Figure 5' provides a summary of the entire model. Columns like Estimate indicate the slope value of the coefficients, Std. Error indicate the variability in the estimate for the coefficient, t value measures how meaningful the coefficient is to the model and P value which indicates the how irrelevant the coefficient is. We are concerned with the P value of the coefficient. Smaller the P value, greater is the significance of the coefficient. The most significant piece of information that can be observed from this summary is that the Technical Field has great statistical significance which can be inferred from the three stars in that row. It implies that this field is a good predictor and that students with stronger technical know-how would have a higher probability of getting placed as opposed to those slightly weaker. As mentioned above, this information can be utilized by the university's placement department to increase focus on improving technical skills of students during placement training.

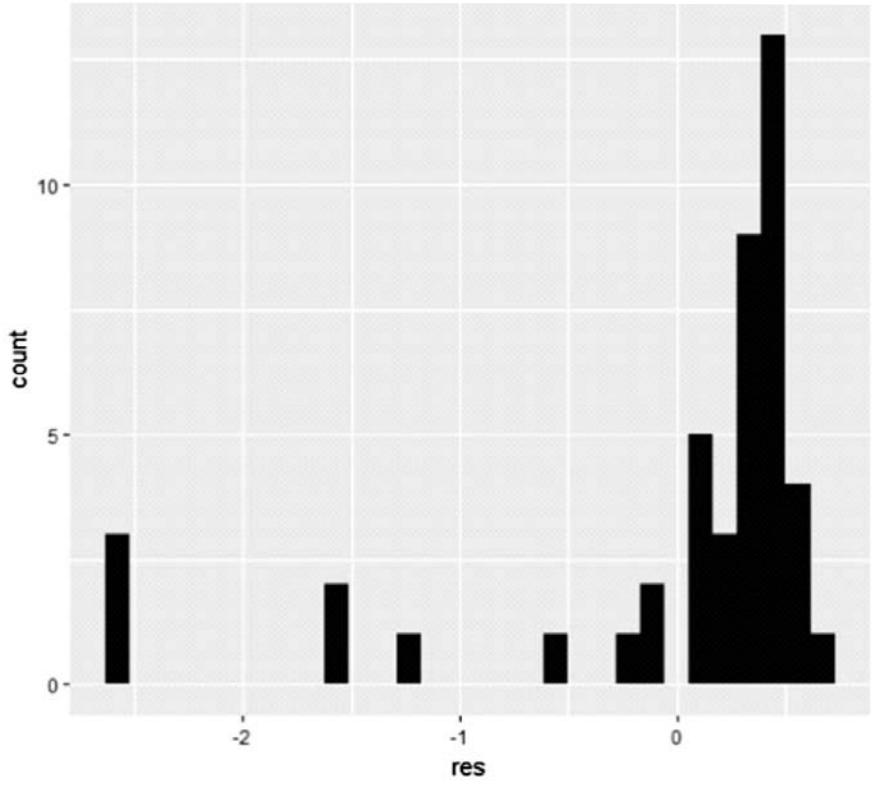

Figure 6: Residuals histogram

In 'Figure 6', we visualize the model by plotting the residuals. For our analysis the residuals is the difference between the actual 'Cleared' plots and predicted 'Cleared' plots. The residuals measure the offset between the actual plot and our model's plots. On reading the graph carefully, on can infer that a residual value of 0.4 occurs maximum number of times showing the 0.4 as the mean difference between actual values and those predicted by the model. It is also seen that a lot of the residues lie on the negative side of the graph. Negative values indicate that the values predicted by the model were smaller than the actual values. The reason for this occurrence is there is anomalous data and that the data set in itself in inconsistent which results from different companies having different kinds and number of rounds for selection.

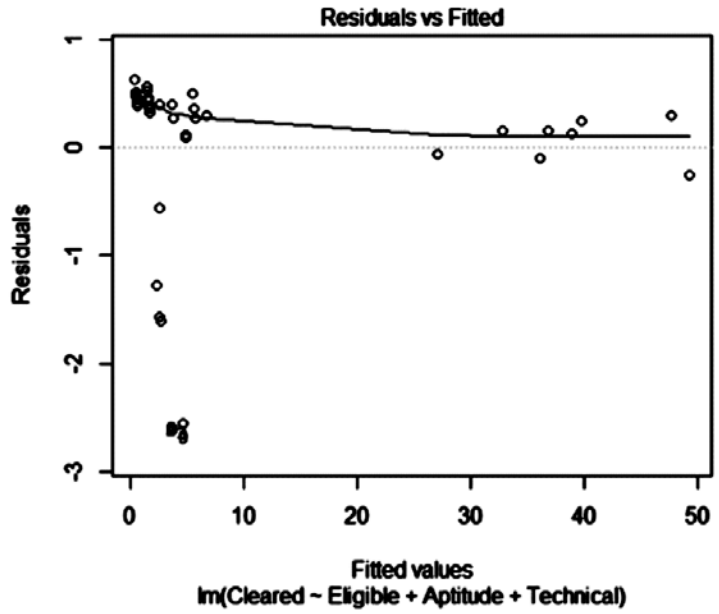

Figure 7.1: Regression Validation

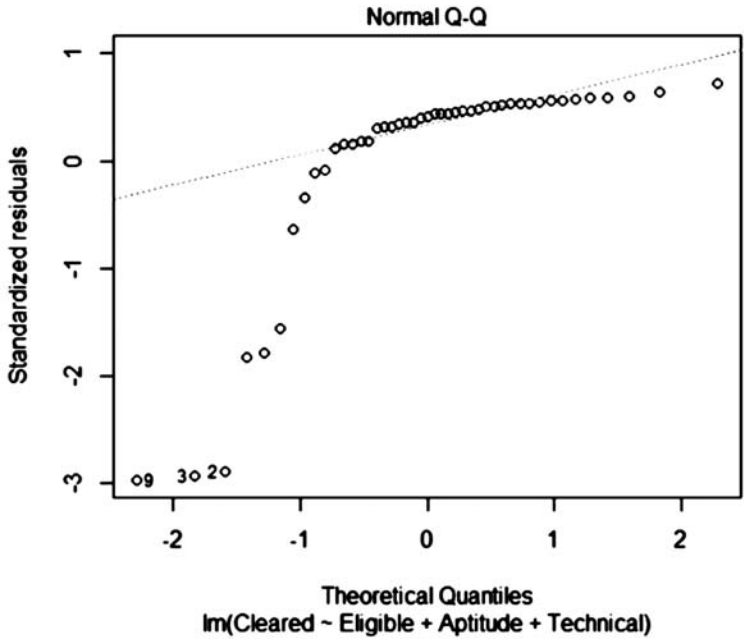

Figure 7.2: Regression Validation

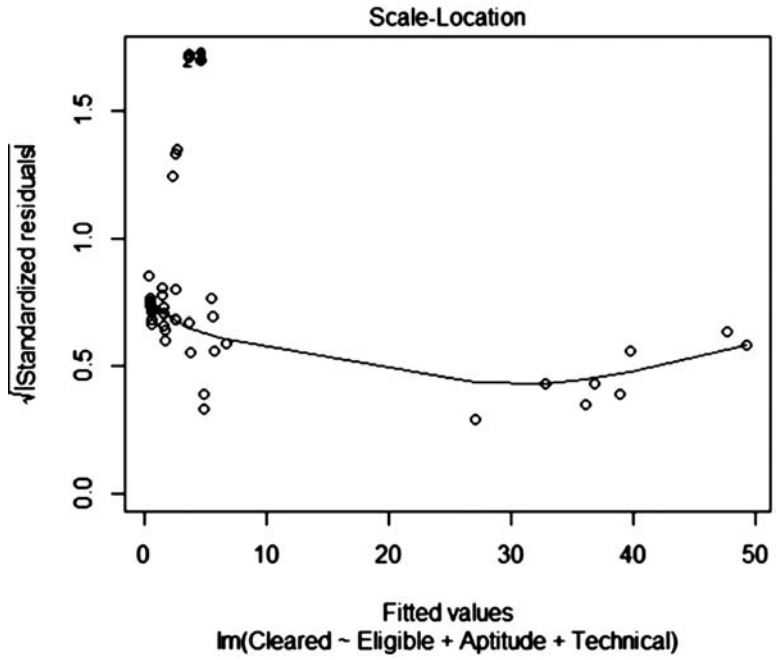

Figure 7.3: Regression Validation

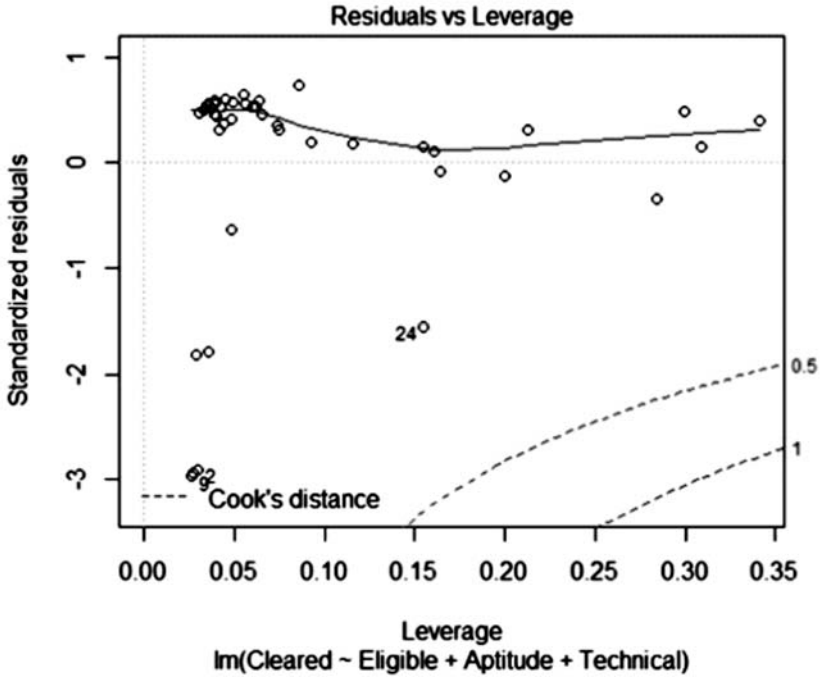

Figure 7.4: Regression Validation

The negative residuals observed in Figure 6 are analyzed and explored further in the subsequent diagrams. We plot the model and get the regression validation (Figure 7) of the residual, depicted through scatter plots. It displays the goodness of fit of the model and predictions. The goodness of fit implies the level of conformance between the actual values and predicted values. For example, 100% goodness of fit would mean that the predicted values are equal to the actual values at all points.

There are negative values in the model that are changed to zeros to improve the analysis. It removes anomalous data and provides better model results.

2.3. Prediction

The model is tested by predicting the test dataset. The model is trained with the training dataset and tested with the test dataset. By doing so we can evaluate how well our model works. Using the predict() function, we get a list of predicted values. This is then compared to the actual values and we evaluate the goodness of fit by evaluating these values.

It can be done via Mean Squared Error(MSE), Root Mean Squared Error(RMSE) and R- Squared Value for the model. The results are depicted in Figure 8. We get R-squared Value as 0.97. It is the goodness of fit of the model. It indicates the correlation strength. The higher the value, the greater is the correlation between the coefficients. This result implies that the predicted value of candidates placed to the actual value has 97% accuracy. This implies that there is a trend or a set of values for which placement prediction depends on. This visualization of placement prediction dependence on factors such as eligibility criteria, aptitude levels, technical know-how and student academic history can be further improved by obtaining more data which will happen as the project progresses.

```
> results$pred<- sapply(results$pred,to_zero)
> mse<- mean((results$real-results$pred)^2)
> print(mse)
[1] 0.5316147
>
> mse^0.5
[1] 0.7291191
>
> SSE= sum((results$pred-results$real)^2)
> SST= sum((mean(cdf$Cleared)-results$real)^2)
> R2= 1- SSE/SST
> R2
[1] 0.9792302
```

Figure 8: Evaluation of predictions

```
> print(results)
        pred real
6  3.6764137    2
11 4.7167113    5
15 0.5114203    1
19 4.8012054    3
20 1.5656767    2
22 3.5407373    4
28 0.5603784    1
32 0.4913259    1
33 0.5532967    1
35 1.6168590    2
42 5.6708273    6
47 1.5768747    2
48 0.5540381    1
49 0.5696843    1
50 0.6029962    1
59 1.6756596    2
```

Figure 9: Prediction vs. Real values for placement of a student from test dataset

The testing of the linear regression model is done by predicting the values of the test dataset. We can see the results of the prediction with respect to the actual results in Figure 9.

3. FUTURE OF THE PROJECT

While the existing data provides various useful insights into the correlations between different placement factors, it is still not sufficient to make accurate predictions and suggest reliable changes. To improve the accuracy of the analysis, in the future we would like to include an alumni survey to gather more data. This alumni survey would be sent to persons already placed. The survey would mainly aim to elicit three kinds of information. First, whether the participant is still at the company he/she got placed in and if not, why. Second, it would provide a scale to help participants rate their satisfaction with the job. And lastly, it would provide a scale to help participants rate subjects they learnt through the course of their degree on the basis of their relevance in the industry. This can help us identify curriculum relevance as well as student job compatibility to not only tweak the placement process to increase the placement percentage of the college but to also make changes that help students get jobs that match their skill sets. Furthermore, this can be turned into an application that can be used by the university exclusively to monitor and analyze all its placement statistics. By linking the analysis code with Tableau, we can help visualize the data better in the form of an interactive dashboard that will depict on-going placement statistics. The application can be made even more useful by creating user accounts for placement faculty and students to view and garner useful information from these statistics and to create a unified forum for the exchange of all placement related information.

4. CONCLUSION

In this project, data analysis techniques, primarily linear regression was used to analyze historical placement data using R-programming. Through various data manipulations and visualizations we received results that depicted high correlation values between the different contributing factors of the placement process. And the goodness of fit of the analysis was found to be 97%. To further improve this statistic, the data set will have to be expanded to include a survey which would then yield results that can be turned into productive recommendations.

5. ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to the Head of the Department of Software Engineering, placement coordinators and faculty of SRM University for providing the dataset and insight into the placement process. Their support and encouragement during the course of the research has been highly appreciated.

REFERENCES

[1] Nemy H. Chavez, Conrado I. Dotong, Nestor C. Camello and Jake M. Laguador, “Employability of Engineering Graduates of one Asian University as Basis for Curriculum Review”, EPH-International Journal Of Science And Engineering, Vol 1 Issue 6 June 2016, ISSN: 2454- 2016

[2] Yoram Bachrach, “Human Judgments In Hiring Decisions Based On Online Social Network Profiles”, Microsoft Research, Cambridge, United Kingdom, 2015 IEEE, 978-1-4673-8273-1/15

[3] Nemy H. Chavez, Evelyn L. De Castro, Nestor C. Camello, Joselito A. Dolot, Jake M. Laguador, “Relevance of School Related Factors to the Job Placement of Engineering Graduates”, EPH-International Journal Of Science And Engineering, Vol 1 Issue 6 June 2016 30, ISSN: 2454- 2016

[4] Grashiela M. Aguila, Evelyn L. De Castro, Conrado I. Dotong,”Employability of Computer Engineering Graduates from 2013 to 2015 in one Private Higher Education Institution in the Philippines”, Asia Pacific Journal of Education, Arts and Sciences, Vol. 3 No. 3, July 2016, P-ISSN 2362-8022, E-ISSN 2362-8030

[5] Julien Boulanger, “Job Analysis And Job Satisfaction”, Bachelor’s Thesis | Abstract, Turku University Of Applied Sciences, International Business Degree Programme| Tradenomi, Bachelor Of Business Administration, 2013.