Post-Graduate College Admission Recommender Using Data Analytics

S. Aarthi, M Sarvathanayan, B. Prithvi Kumar, Rakesh G S

Abstract: ***In this paper, we present an applied research on developing a post graduate college recommender system. This system can help students who pursue higher studies in choosing the college in which they can get an admission. This recommendation of college will be done using Data Analysis based upon various variables such as test scores like GRE, TOEFL and will also be done using university rating, undergraduate GPA, research experience, Statement Of Purpose and Letter Of Recommendation strength. The system uses an algorithm called Multiple Linear Regression to achieve this. Once all the values are entered, the proposed recommendation system will predict the rank of the college in which the student can get admission using a dataset which containing 500 different observations of students in the past.***

Index Terms**:** ***Data Analysis, Multiple Linear Regression, Recommender System.***

I. INTRODUCTION

Most students pursue post-graduation studies at universities after getting done with their undergraduate studies. To get an admission in a university has therefore become very important and competitive. Admission of students at Universities depend on various factors like GRE score, TOEFL score, IELTS score, Undergraduate CGPA, Research experience, etc. The cut-off marks of each university constantly changes every year as the ranking of universities keep changing. Hence, it is difficult to predict the university in which students can get admission with their score. Therefore, a prediction system which predicts the right university for students is required. A University admission recommender system for post-graduation can make this process easy.

We propose a system which will help students who are applying for higher education to have a clear understanding about the choices students have to make before applying for higher education Universities. Our proposed system will recommend universities to the users for post-graduate admission. This system uses Data Analytics to predict the chances of admission of students at universities. We use a dataset which has 500 observations and 9 variables. Among these 9 features, 1 is a dependent variable and 8 are independent variables. The weighted scores are computed from previous information of successful applicants such as UG CGPA, TOEFL, GRE Scores etc.

A. GRE and TOEFL

The Graduate Record Examinations is a graduate school entrance exam while the TOEFL is a test for English language skills. The colleges want to see GRE scores to make sure that we can handle the graduate-level coursework, and they want to see TOEFL scores to make sure that our English skills are good enough to do well at an English-speaking school.

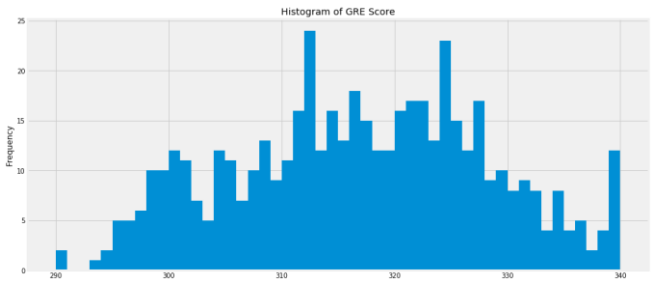

Fig. 1. Histogram of GRE Score

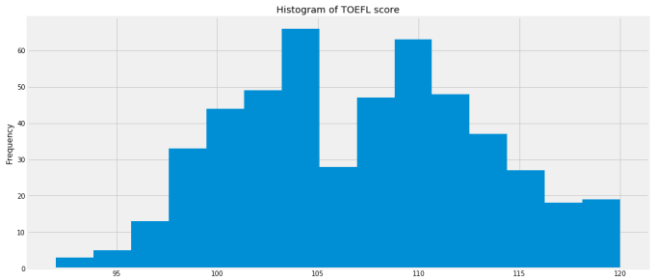

Fig. 2. Histogram of TOEFL Score

B. LOR

The Letter of Recommendation is required explicitly by an academic program and has to be sent directly to the university by the professor or employer. The document should be of 300-400 words and should present your character, accomplishments and abilities from an objective perspective.

Revised Manuscript Received on April 07, 2019.

Mrs. S.Aarthi, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

M Sarvathanayan, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

B.Prithvi Kumar, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

Rakesh G S, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

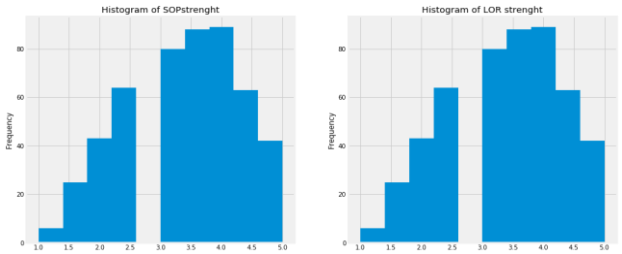

Fig. 3. Histogram of LOR and SOP

C. SOP

The Statement of Purpose is a long essay that is often asked by universities abroad. Usually about 1000 words, this essay seeks to understand the candidate's life, the motivations for the chosen career path and their goals.

The algorithm that this system uses is called Linear Regression. Regression with 1 independent variable or explanatory variable is called Simple Linear Regression. Linear Regression with more than 1 independent variable or explanatory variable is called Multiple Linear Regression. In the prediction, usually the variables that affect the system are more than one and hence we use Multiple Linear Regression.

II. METHODS

In this section, methods and measures which are applied are explained, for example, Multiple Linear regression as the method.

The dataset for graduate degrees admission (Masters), was originally found on "Kaggle" .The set includes the following Columns: Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, and Chance of Admit. Multiple Linear Regression is applied on this dataset.

Multiple Linear Regression

This type of regression tries to model the relationship between multiple explanatory variables along with response variable by fitting the linear equation to the observed data. All the values of independent variable '*x'* are linked with the value of dependent variable '*y'*. The population regression line is said to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ for '*p'* explanatory variables $x_1, x_2, \ldots, x_p$. This regression line tells by what means the mean response μ_y changes along with explanatory variables. The observed values vary about their means μ_y for '*y'* and are assumed to have standard deviation σ. Fitted values $b_0, b_1, \ldots, b_p$ are used to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the population regression line.

Since the observed values for '*y'* vary with their means μ_y, this model of multiple regression involves a term for this variation. In other words, it is conveyed as "DATA = FIT + RESIDUAL", where "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$ and "RESIDUAL" term represents deviations of the observed values for '*y'* from their mean μ_y. The notation for model deviations is ε.

Generally, multiple linear regression model with *n* observations, **is**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1, 2, \ldots n.$$

In the least-squares model, the best-fitting line is found by reducing the sum of squares of the vertical deviations from every single data point to the line. Therefore there are no cancellations between positive and negative values. The least squares estimates are usually computed by statistical software.

The values for the equation $b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$ are denoted $\hat{y}_i$, and e_i is the residuals which is equal to $y_i - \hat{y}_i$. The sum of the residuals is zero.

III. RESULT AND DISCUSSION

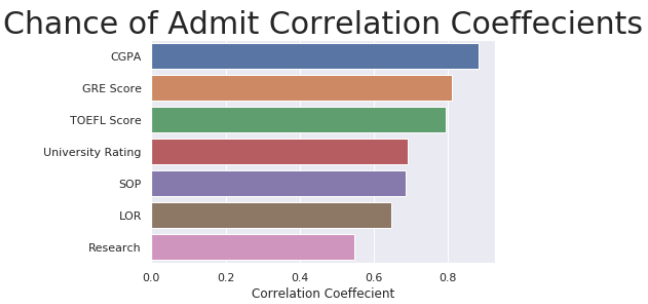

Fig. 4. Chance of Admit

It is observed that CGPA plays more important role than GRE score and TOEFL score. CGPA can tell how well a student has performed in studies in the previous semesters and how well will the student perform in future. That's the reason people with CGPA more than 8 tend to score good in GRE ad TOEFL. A student preparing to study MS in abroad hast to first study hard to get a good CGPA, then prepare for GRE to get better score and then finally prepare for TOEFL. This is the most common procedure students usually follow. There is a simple logic behind this small but very important and useful observation. CGPA is the cumulative GPA of every semester. To get a CGPA above 8 or 8 a student has to perform well in every semester consistent. If a student performs bad in any of the semesters it can directly affect the CGPA. Hence, we can tell how important it is to maintain the GPA score.

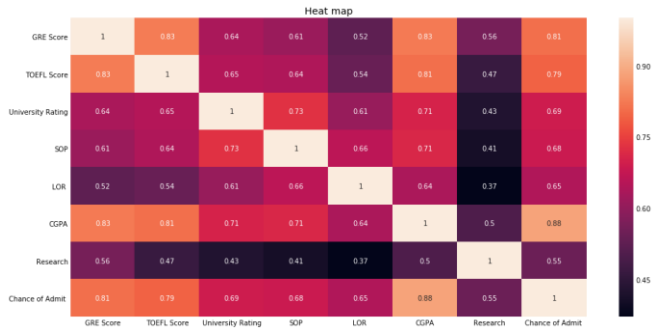

Fig. 5. Heat Map

IV. CONCLUSION

Data Analytics plays a major role in the current market and is also growing at good rate. In this case of post graduate college recommender system, multiple linear regression model is used. The conclusion is that the multiple linear regression has a high accuracy.In future, better techniques can be identified which can be used in place of multiple linear regression.

REFERENCES

1. Hui Li, "Integrative Method based on Linear Regression for the prediction of Zinc-binding Sites in Proteins", IEEE Access, 2017
2. Mahamudul Hasan, "Graduate School Recommender System: Assisting Admission Seekers to apply for Graduate Studies in appropriate Graduate Schools", 5th International Conference on Informatics, 2016
3. L. Duan, "Healthcare Information System : Data Mining methods in the clinical recommender system", Enterprise Information Systems, 2011
4. Thorsten Sommer, "A Web-Based Recommendation System for Engineering Education E-Learning Solutions", Research Gate, 2014
5. Michael D. Ekstrand, "Collaborative Filtering Recommender Systems", Foundation and Trends in Human-Computer Interaction, 2011
6. Cesar Vialardi, "Recommendation in Higher Education using Data Mining Techniques", Educational Data Mining, 2009
7. J. Ben Schafer, "Collaboration Filteration Recommender Systems", Research Gate Publication, 2007
8. Ben G. Weber, "A Data Mining approach to Strategy prediction", PCWP, 2017
9. Badrul Sarwar, "Analysis of Recommendation Algorithms for E-Commerce", GroupLens Research Group, 2007
10. Kavitha S, Varuna S, & Ramya R. (2016). A comparative analysis on linear regression and support vector regression. 2016 Online International Conference on Green Engineering and Technologies (IC-GET).
11. DC Montgomery, EA Peck, GG Vining, "Introduction to linear regression analysis", Wiley Series in Probability and Statistics, 2015.
12. H. Bydovska and L. Popelinsky, "Predicting student performance in higher education," in Proceedings - International Workshop on Database and Expert Systems Applications, DEXA, 2013.

AUTHORS PROFILE

S.Aarthi, Assistant Professor (O.G.) at Department of Computer Science Engineering, SRM Institute of Science and Technology, Chennai. Areas of interest include Software Engineering, Object Oriented Analysis and Design.

M Sarvathanayan, 3rd year B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai. Previously published research papers in the field of medicine and healthcare technology and developing security systems.

B.Prithvi Kumar, 3rd year B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai. Areas of interest are Data science, banking and finance and Python programming.

Rakesh G S, 3rd year B.Tech Computer Science Student at SRM Institute of Science and Technology, Chennai. Has previously worked in development teams for web apps and security system research.