University Recommendation System

B R Pratheek
Computer Science and Engineering
PES University
Bangalore,Karnataka,India
pratheekraviprakash@gmail.com

Raghav S Tiruvallur
Computer Science and Engineering
PES University
Bangalore,Karnataka,India
rtiruvallur@gmail.com

Sai Manoj
Computer Science and Engineering
PES University
Bangalore,Karnataka,India
manojgh42@gmail.com

***Abstract*—The purpose of this paper is to find a method to recommend a list of universities to a student wanting to pursue higher education. The recommendation is based on certain scores obtained by the student on standardized tests, industry or research related experiences gained during their undergraduate coursework and also their publication history.**

Keywords—Recommendation System, Collaborative Filtering, Multivariate Linear Regression

Introduction

After completing their undergraduate coursework, a considerable number of students look forward to higher education in order to specialize in a particular domain of their interest which helps them get better career opportunities moving forward or they have deep rooted interest in research in a particular domain and are planning to obtain a doctorate degree in the same.

The students' journey begins when they start applying to various universities ,who then judge the students based on certain criteria including but not limited to their academic background which involves the students scores on standardized tests such as GRE,TOEFL and their undergraduate GPA. The university after a thorough process admits only those students who seem to have what it takes to complete the rigorous coursework.

The process of choosing the correct university can be considered the foundation of the entire process as it is a well known fact that the best universities provide the best opportunities to students.So in this paper we would like to discover answers to questions such as: Does GPA play a factor in a student being admitted into a university? Can a decent amount of experience in research/work give you higher chances of admission?

The research study's main motive is to develop a recommendation system of universities for students planning to pursue postgraduate study using data of past students and relevant information.

I. Literature review

A. The University Recommendation System for Higher Education

The authors [1] began their paper by explaining the concept of recommendation systems,the different phases of development of recommendation systems, the different types of systems, use case scenarios of these systems and their advantages and disadvantages.

Their approach was to test out various machine learning models which would provide a list of top 10 universities based on the probability of the student being accepted into those universities. The recommendation system would involve a SVM to classify universities that the applicant would find appealing and then the KNN algorithm would generate universities list which would meet the students qualifications. Some of the features that the system was built on included TOEFL, GRE scores,University rank,budget,weather,etc.

The accuracy of their model was 69%.

B. Recommendation System for Higher Studies using Machine Learning

The assessment [2] of the recommendation system is using various ML algorithms like Naive Bayes,K-means clustering.The author mentions the various factors that a student needs to keep in mind before applying to universities such as GRE,TOEFL,gpa,Cost of studying etc.

Recent papers have used Naive Bayes and Decision trees to implement this problem.They employed a profile based approach to solve their problem.

They used factors such as GRE quants score,GRE verbal score,TOEFLscore,CGPA to determine the list of recommended colleges.

However, they had to normalise the data using min-max normalisation.They normalised the GRE scores and TOEFL scores using this technique

In Collaborative filtering,the GRE quants score has a vital impact on the admission and universities prioritize this score over other scores.The steps to achieve this are to find Users similarity calculation(Pearson’s correlation,cosine similarity),N nearest neighbours,Prediction.

In this step,we get a list of universities based on a student's GRE and TOEFL score.The content based filtering also works in a similar manner.

The crux of the problem is solved using algorithms like K-means and Naive Bayes to achieve high accuracy to solve this problem.The author talks about using hybrid concept which implies using more than one algorithm to solve a problem.For example,using linear regression and SVM to attempt to solve this problem.

After testing with various algorithms like KNN,SVM,Linear regression,we see that KNN gives preference to GRE quants score,SVM gives preference to CGPA and Linear regression gives preference to TOEFL. On computing the accuracy,we see that KNN has 91% accuracy compared to SVM which has 81% and Linear regression which has 83%.

C.University Recommender System for Graduate Studies in USA

In the paper[5], the dataset consisted of 45000 samples after scraping the original dataset.It was cleaned through an extensive process of normalisation and imputation.

The goal of the author was to supply a list of 10 universities that would maximise the chances of a student getting an admit.

They initially used a baseline model which basically randomly chose 10 universities and had an understandably low accuracy of 22%.To improve this accuracy,techniques like Support Vector Machines,K-Nearest Neighbours,random forest were considered.The way they work is to choose one university that suits the student’s profile the most and then find 9 other universities that are similar to that university.

The author faced some issues like overfitting but stated it could be avoided by techniques like k fold cross-validation or PCA.

Hypertuning was done on the algorithms stated above to improve the accuracy of their model.

It was found that the KNN algorithm had an accuracy of 50.4% at 56 neighbours after trying various numbers of neighbours.

Random Forest was used, which is an ensemble of Decision Trees, and after experimenting with various numbers of trees an accuracy of 50.5% with 150 trees.

Support Vector Machine was then used, which is an advanced Machine learning algorithm and in this project, the Gaussian RBF kernel function was used which produced an accuracy of 53.4%.

From these algorithms, it was found that SVM had the highest accuracy of 53.4%.

It was also found that the features-Undergraduate university,GPA,GRE score,Research experience were found to explain the maximum variance in data and were used to build the final model.

The other features like journal publications,industry experience,internship experience did not provide any new information.

D. College Recommendation System[4]

The college recommendation system’s sole purpose is to generate a list of colleges that a student would be eligible to join based on his/her profile. The challenging task is to collect the database of all colleges and further generate a list of colleges from all the colleges which require elimination of those colleges in which the candidate is not eligible.

The goal of this system was to automate the college generation list. Several techniques such as Data Mining and Query Optimization Techniques were used.

The system would recommend colleges based on the student’s score and the rating of the college would be collected from students across various colleges. The resulting list would be filtered based on various criteria like caste information, region and minority.

SQL Database was used to store the reviews. The Naive Bayesian and Decision tree algorithms were used to implement the system. The Naive Bayesian algorithm was used since it was expected to have higher accuracy in recommending the best college. This technique was considered helpful to minimize the student’s time in searching for colleges.

II. Proposed Solution

A. Dataset

We have selected the university admission dataset [3] for recommending a list of universities based on a student's profile.

The dataset consisted of 26 columns and 53644 rows

As a whole our dataset had many NULL values which were treated accordingly.Many numerical values were represented as strings and numerous data points had unassociated links/strings which conveyed no information.

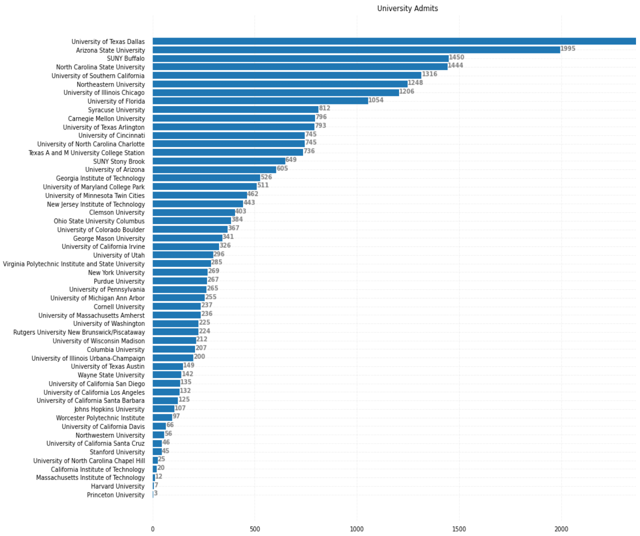
Rescaling was needed as certain features were very sparsely populated or were not updated to current trends.

Experiences were expressed in a number of weeks or days hence there was a need to rescale.

The scores columns(GRE,TOEFL) had a different scale(Eg:It was given out of 800 for quantitative portion of GRE but currently it is evaluated out of 170)
The cgpa column had invalid values which needed to be imputed , also each student's cgpa was on a different scale hence needed rescaling.This was done based on the topper's cgpa to get a rough idea on students performance compared to the rest of his/her batch.

B.Initial insights

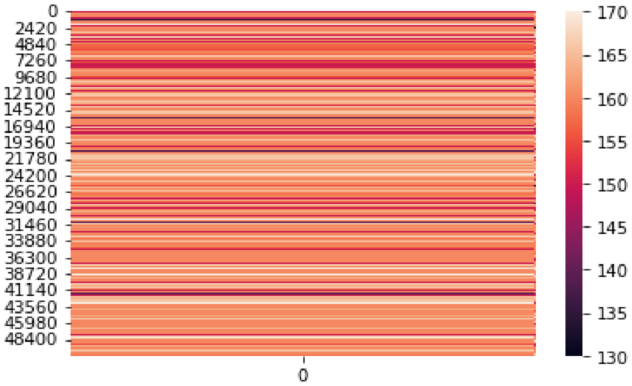
On our initial EDA, we find the following:



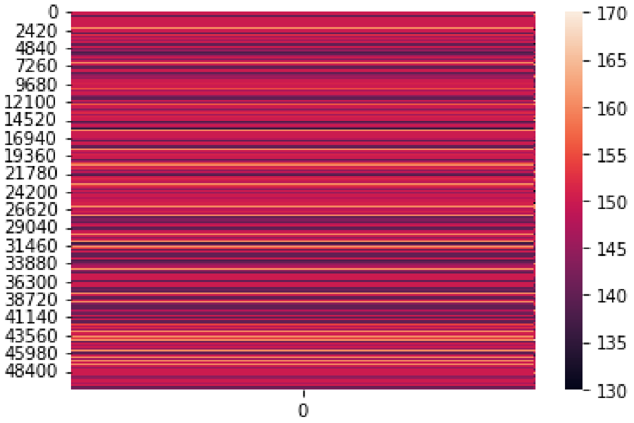
We can see from the above bar graph that University of Texas, Dallas had the highest number of admits from the assembled data.

We have plotted heat maps for gre quantitative and gre verbal scores to present the distribution of the scores.

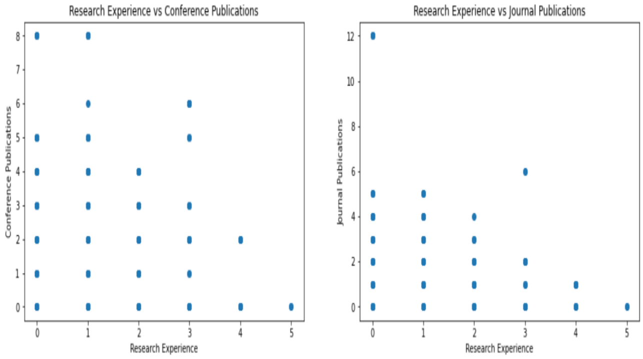
Heat map of gre quantitative scores:



Heat map of gre verbal scores:

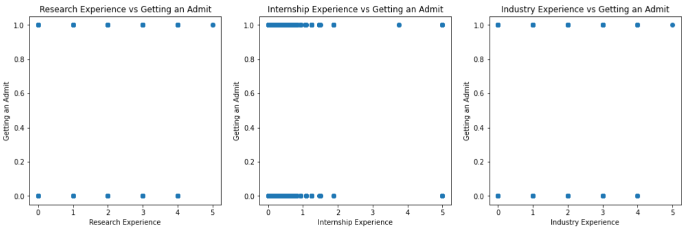


Scatter Plots of Research experience vs Publications:

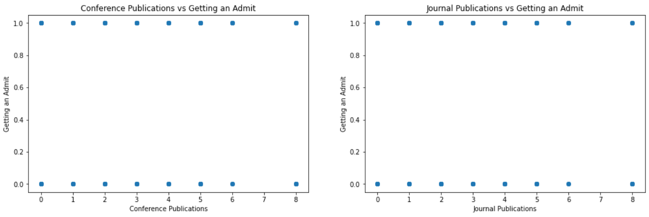


From the above plots, we can observe that research experience and publications are positively correlated.

On performing further EDA, We observed that there exists no linear relationship between the different types of experiences and getting an admit from the below scatter plots .Though not having any type of experience does not hurt the candidates chances, having it certainly increases the probability of getting an admit. Hence we can say it is not mandatory to have some sort of experience to get an admit.

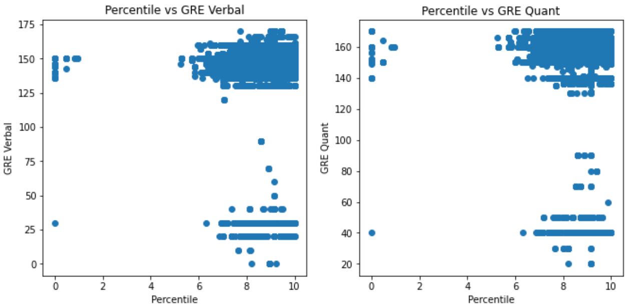


Correlation Scatter plots of Publications vs Getting an admit are plotted below:



The same inference that we learnt from the experiences can be applied to the above plots as well, i.e, it is not mandatory to have a publication to get an admit.

The plots of Percentile(relative gpa) vs GRE scores:



From the above plots, we observe that students with higher percentile usually score more in their Gre Verbal and GRE Quant compared to their peers who have a lower percentile.

C. Approach

After thoroughly understanding the data after completing preprocessing and visualization , we observed that the various types of experiences, the different scores obtained by the student and his percentile are indicators of him getting an admit.

Hence we plan to use collaborative filtering (model based techniques) on the above said attributes to help recommend a list of colleges that the user might get admitted to. We plan to test the accuracy of the recommendation using a self defined accuracy function.

The model that we intend to move forward with is Multi variable Linear Regression(MLR).

D. Fitting the model

We sought to use MLR to assign each university a cutoff based on it’s acceptance rate.This would imply a university with the least acceptance rate would be given the highest score.

The score assigned is a positive integer from the range [2,10], this assignment is based on the output of the MLR model, where each 0.1 increment is scored one higher and all universities that lie within that 0.1 bucket share the same score.

The regression model relates this score of the university to all the attributes like undergrad gpa,GRE,TOEFL etc.

For example, the equation would be something like this:

Target_Score=a1*(gpa) + a2*(gre) + a3*(toefl)+.......

We split our data into 80-20 sets for training and testing. x_training data consists of all the attributes and the y_training data consists of all the scores of those universities.This implies if a student has been admitted into a university, then he has scored at least the minimum cutoff of the university.

This model predicts the score of a student applying to universities, the score acts as their index which can be used to predict all of the universities they can target.A student having a certain score can target all of the universities below it as safe universities and some universities above it as moderate and ambitious universities.

We tested the accuracy of our model and found it out to be 65.72%.

The self defined accuracy function takes into account the colleges we did not predict and the user did not get into and the college we predicted and he got into.

This was the accuracy we computed after testing it on the split dataset.

III. Results and Inferences

For the MLR model we chose the accuracy achieved was 65.72%. Considering the size of the dataset and also the limited number of data points that actually helped in building the model we believe this score is commendable.

The effect that each attribute had on the final result can be tabulated as follows:

researchExp	0.1678
industryExp	-0.1413
toeflScore	0.0017
toeflEssay	0.0282
internExp	0.3436
greV	0.007
greQ	0.0171
journalPubs	-0.01329
greA	0.6271
confPubs	0.1207
Percentile	0.1054

As stated above publications and experience were not a must to get an admit but we see that having industryExp surely boosts the chances of getting an admit. We also observe that universities prefer those with good writing skills as toeflEssay and greA have higher coefficients compared to their counterparts.

IV. Conclusion

This report takes us through the full process of our project and journey into the dataset of graduate admissions.We started of with literature survey which included studying variety of research papers by fellow scholars and helped us to understand the various ways of approaching this problem and also gave us a better insight on how to model the dataset.

We started our project with a lot of data cleaning as our dataset was very noisy and had redundant information. EDA and data visualisation gave us a much better picture on how different attributes were related to one another and what were the necessary features we could use in our model. Then moved on to the crux of our project which was model building. We decided to use MLR(Multi variable Linear Regression) to build the recommendation system. The methodology of assigning rankings to colleges based on certain criteria and using MLR model to return scores based on which recommendations were made was something different we tried. After some further data transformation, we arrived at a reasonably good model which had an accuracy of 65.72%.

V. Future Prospects

We could use ensemble modelling techniques which include combining various models such as SVM, MLR, KNN,etc; to further learn underlying patterns that might exist in the dataset which in turn increases the accuracy of the model and also provides better predictions.

VI. Acknowledgment

We would like to express our gratitude to our course instructor Dr.Gowri Srinivasa and TA's for their continuous effort and guidance throughout this phase. We would also like to thank the Computer Science department of PES University for always giving us chances to gain hands-on experience in this field.

VII. References

[1] Aishwarya Nalawade, Bhavana Tiple,"The University Recommendation System for Higher Education",International Journal of Recent Technology and Engineering (IJRTE),ISSN: 2277-3878, Volume-8 Issue-6, March 2020

[2] Alcina Judy,Kesha D'cruz1, Janhavi Kathe1, Kirti Motwani, "Recommendation System for Higher Studies using Machine Learning",International Research Journal of Engineering and Technology(IRJET),e-ISSN:2395-0056,p-ISSN:2395-0072,Volume:07,Issue:04,April 2020

[3] https://www.kaggle.com/nitishabharathi/university-recommendation

[4] Vinit Jain, Mohak Gupta, Jenish Kevadia, Prof. Krishnanjali Shinde, "College Recommendation System", International Research Journal of Engineering and Technology(IRJET), ISSN: 2278-0181,Volume :5,Issue: 01, Published: 24th March, 2018

[5] https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/026

VIII. Appendix

A. Contributions

- B R Pratheek :Drafting the problem statement, Literature Survey, Part of data cleaning, helped in model building.
- Raghav S Tiruvallur :Literature Survey, Part of data transformation , helped in model building and came up with evaluation metric.
- Sai Manoj :Literature Survey, Part of data cleaning , Data visualization, Analysis of Results of visualization and model.