University Recommender System for Graduate Studies in USA

Ramkishore Swaminathan
A53089745
rswamina@eng.ucsd.edu

Joe Manley Gnanasekaran
A53096254
joemanley@eng.ucsd.edu

Swetha Krishnakumar
A53087417
swk032@eng.ucsd.edu

Aditya Suresh kumar
A53092425
asureshk@eng.ucsd.edu

ABSTRACT

For an aspiring graduate student, choosing which universities to apply to is a conundrum. Often, the students wonder if their profile is good enough for a certain university. In this paper, this problem has been addressed by modeling a recommender system based on various classification algorithms. The required data was scraped from www.edulix.com, and data-set containing profiles of students with admits/rejects to 45 different universities in USA was built. Based on this data set, various models were trained and a list of 10 best universities were suggested such that it maximizes the chance of a student getting an admit from that university list.

Keywords

Recommender System,University Admission

1. INTRODUCTION

Every year the number of students seeking admission for the graduate studies is constantly increasing. As a result the competition gets tougher and the chances of admission becomes unpredictable. Given the growth of new programs and number of admissions, a student is often unaware of the existence of such programs. In this paper, a justifying attempt using K Nearest neighbours, Random Forest and Support Vector Machines was made to provide a solution to these issues by considering the target university's perspective to evaluate whether a student's profile is competitive enough to be admitted into their university. Hence the students could get a better picture of where they stand and can make an intelligent well-formed decision.

As a first step, information regarding 45 universities were collected along with the details about students' profile and their admission results. Section 2 explains about the problem that has been addressed and the approach to solve it.Section 3 includes a brief description about the existing literature on similar topics outlining their approach, techniques adopted, pros and cons analysis, results and conclusions. Section 4 describes the data set, populated by scraping and data cleansing and transformation. It also gives detailed explanation about the selection of relevant features and their impact on the model. Section 5 gives information about various models and a comparison between each other in terms of accuracy. Finally, the performance of the chosen models are analyzed using the results obtained and a summary are included under Section 6 and Section 7.

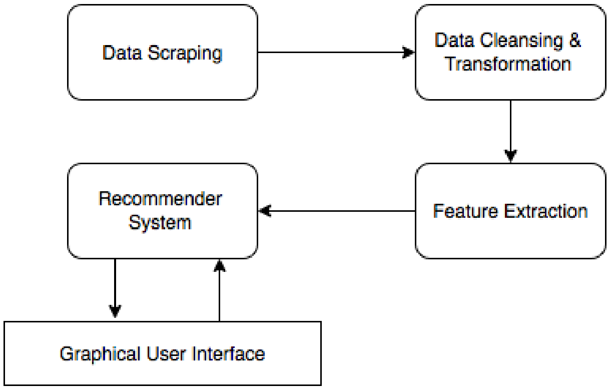

Figure 1: Flow Diagram of complete process

2. THE RECOMMENDATION PROBLEM

In today's fast-paced world, every technological innovation influences the importance of higher education, especially the ones which serve as hubs to the latest researches and trends. Given that, the United States will be one of the top destinations for any student across the world. For international students who wish to pursue graduate studies in the United States of America, choosing a suitable college and earning an admit is a challenge. Although, many internet resources and forums are available, they do not offer satisfactory suggestions, as most of them are based on assumptions from college rankings and not the actual statistical relations. From a student's point of view, the cost of the application and the amount of dedication to the process is also high. Thus, to guide the students in an efficient manner, the university recommender system has been developed, based on the input of the students' academic data. Since the problem is extensive, for the sake of simplicity, a select list of 45 universities were considered.

3. LITERATURE REVIEW

In the past, a lot of work on employing data mining techniques in the field of education were undertaken. Few recommender systems to suggest course and university based on a student's academic record were developed. Those systems employed decision tree classifier and fuzzy c-means clustering techniques using WEKA tool kit and it was aimed to help the students choose a stream which will suit their skill sets[4]. Another different recommender system was built to help the students with their academic itineraries. They help in making decisions about what course to select based on a student's schedule, stream and professors. Here, the model was trained based on past 7 years data for a particular university and classifiers for every subject was modeled based on cumulative GPA[8]. On the other hand, some recommender systems were modeled to help the university to know about their students by keeping track of their time, extracurricular activities and achievements, in addition to their academic potential. This helps them to identify and categorize the students depending on the need using two-step algorithm and K-means[5]. However, there was no access to any of the data-set used in the above mentioned works. Although similarities exist with the topic considered in this paper, it is not appropriate to compare results directly with any previous work because the data set used in this paper is completely different.

4. DATA SET IDENTIFICATION

The first step in building any recommendation system is the identification of the data set. For this particular problem, academic details and background information which are provided during the application process, forms the core data. In order to build the classification model for the recommender system, this data has to be organized with appropriate labels. This core data for the application process is not readily available on the internet for direct consumption. Though there were few forums which had some vital information regarding the same in terms of scores, the distinguishing information regarding the students' research interest and knowledge in a particular topic remains unknown. However, this whole approach is based on making maximum use of the available information.

Variation of the number of admissions to any graduate program based on undergraduate universities is represented by Figure 2. It was found that Mumbai University(1587), National Institute of Technology(1467), Visvesvaraya Technological University(1426) and Anna University(1032) were some of the undergraduate universities with highest number of admits.

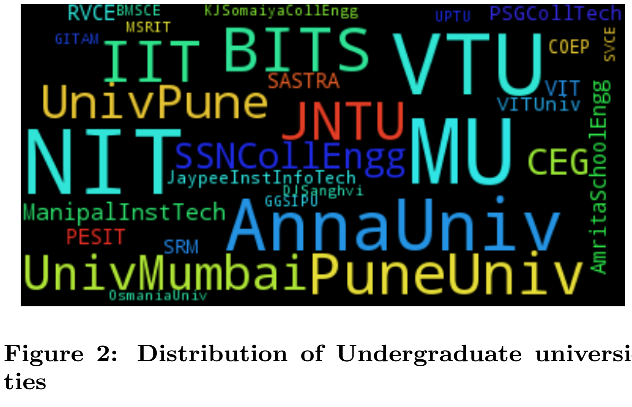

Figure 2: Distribution of Undergraduate universities

The 'Edulix' forum is one of the most popular forums for students planing to pursue graduate studies. This is the hub for students who wish to take part in discussions and queries regarding any information about graduate studies. This forum basically collects the academic details of its users to evaluate their profile against past experiences. Out of all these data, some data like the candidate's undergraduate university, CGPA, GRE and TOEFL scores, number of research publications, work experience etc.were identified as prospective features. By writing a web crawler script, relevant data necessary for this model was scraped off from their website, cleaned and then transformed into appropriate forms to be used as input data for the models

4.1 Data Scraping

Initially the list of 45 universities was narrowed down, which had enough data to be scraped. Universities with skewed data were dropped down. Then a crawler was built to get the list of students and the links to their profiles on Edulix. Once the unique set of students was identified, the data was scraped from each profile and then the required data was extracted from the HTML by using the python library 'BeautifulSoup'. The tabular structure of Edulix's web page, helped to identify the required data labels and points. The usual way of accessing the required elements by using the XPath did not work out for this case, because the HTML was malformed in many cases.

4.2 Data Cleansing and Transformation

About 45000 samples of raw data was obtained by as a result of scraping. Each sample corresponds to the profile of a student. The data points extracted included GPA, undergraduate university, GRE verbal score, GRE quantitative score, GRE analytical writing score, number of journal publications, number of conference publications, industry experience, research experience, internship experience and pursuing major. Cleansing the data of undergraduate universities had to be done, since this field was just a text box and not a select field. So input from different students created anomalies and this was corrected by trimming the string and removing spaces found in them. The GRE scores(Verbal, Quantitative and AWA) were also cleansed since they contained scores of both old and new versions of the examination. Similarly the GPA scores available were based on different point systems, so all the GPA scores were uniformly scaled to 4 point scale. Also, certain categorical features like the student's undergraduate university and department to which they apply were considered as separate features. A total of 1435 distinct undergraduate universities and 53 distinct majors were obtained after filtering and each of these were used as binary features.

4.3 Feature Extraction

The most important property of a feature is its correlation with the predicted output. Exploratory analysis was done by plotting the feature values for two different universities and observing their variation. Variation of features CGPA and GRE for two different universities(Purdue and NJIT), has been shown in Figure 3 and Figure 4 respectively.

Initially, when all the features in the data set were considered the accuracy was comparatively low(40%). The forward selection algorithm[7] was used to select the best set of features for the model. In the first iteration of the algorithm, the single best feature was identified that best describes the variance in the data. In the second iteration, the best fea-

Table 1: Statistics of the features

	Research Exp.	Industry Exp.	Intern Exp.	GRE Verbal	GRE AWE	Journal Publications	Conference Publications	CGPA
Mean	0.29	3.46	0.39	148.31	5.29	0.03	0.04	0.75
Std. Deviation	2.42	11.11	2.26	15.39	1.48	0.25	0.32	0.36
Min	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00
Max	53.00	132.00	96.00	170.00	6.00	12.00	8.00	0.98
25%	0.00	0.00	0.00	145.00	3.00	0.00	0.00	0.70
50%	0.00	0.00	0.00	150.00	3.50	0.00	0.00	0.77
75%	0.00	0.00	0.00	154.00	4.00	0.00	0.00	0.84

ture was fixed and the the next best feature was found. This process was repeated till the accuracy no longer improved. Based on this method, undergraduate university, research experience, GRE and GPA were found to be the most effective features. After using forward selection algorithm, the accuracy improved.

During this process, a situation arose, when the accuracy did not show any improvement, even though the best features were chosen. This was because, the numerical features like CGPA and GRE score were based on different scales, and so had an an adverse implication on the model. However when scaled from 0 to 1, there was a significant improvement in the accuracy. Hence, all the numerical variables were then normalized to a scale of 0 to 1 by using the following formula,

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where X is value of any feature.

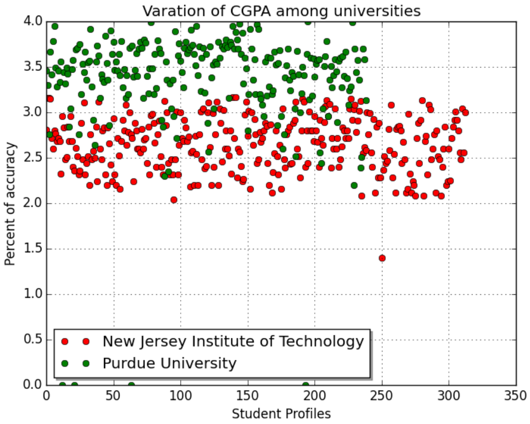

Figure 3: Variation of CGPA among two universities

5. RECOMMENDATION MODELS

The baseline model is one in which it randomly predict 10 universities out of a total of 45 universities for each user. The accuracy of this model was found to be 22%.

Three different models, Support Vector Machine, K-Nearest Neighbors and Random Forest, were built using a combination of all the features mentioned above, to classify a student profile to the best university that they must apply to, among the available 45 universities. Once the best university was found for the student, the 9 most similar universities in terms of the selected features was found by computing euclidean distances to give a total of 10 universities, that the student must apply to.

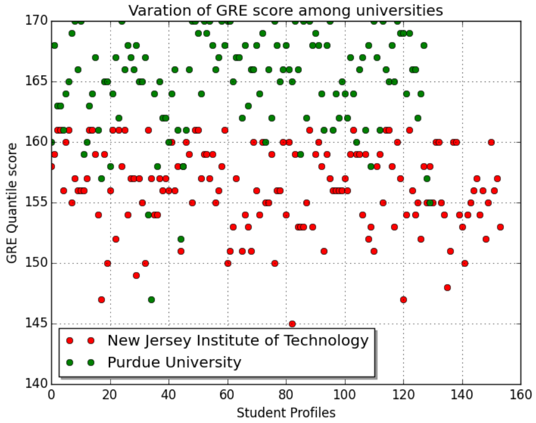

Figure 4: Variation of GRE among two universities

The data was split as 80:20 for training and testing. The model classified the training data with good accuracy but had a high error rate for test data. This problem was due to over-fitting and can be avoided by techniques like Cross validation to test the model on more datasets or by techniques like Principal Component Analysis to reduce the dimension(number of features used) of the model or more datasets can be used [3],[1]. The first technique has been employed in this project. k-fold cross validation mainly prevents overfitting as it reduces the variance by averaging over k different partitions, so the performance estimate is less sensitive to the partitioning of the data[6]. The entire data set is divided into 5 sets and each time 4 sets are used as the training data and the model is tested on the remaining 1 set which is used as the test data. The accuracy of the model is determined and this process is repeated 5 times. Each time a different set is used as the test data. The error rates obtained are all averaged to obtain the final error rate.

The following subsections describe the models that we tried.

5.1 K-Nearest Neighbours

K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. In K-NN classification, the output is a class membership. An object is classified by majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.The K-NN model was run by varying the number of neighbours that were used and it was found that the best accuracy of 50.6% was obtained when the number of neighbours was equal to 56. The variation of accuracy with the number of trees constructed is shown in Fig 5.

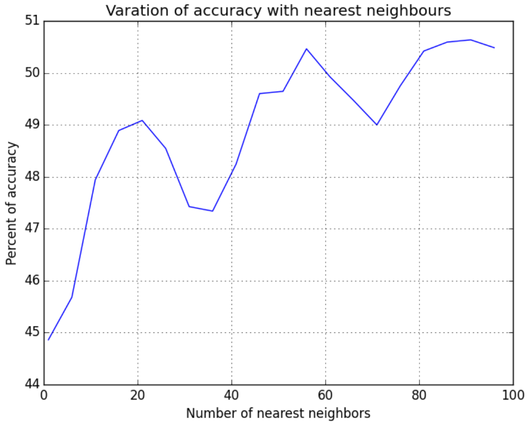

Figure 5: K-Nearest Neighbours Model

5.2 Random Forest

Random Forest is an ensemble of decision trees. Unlike single decision trees which are likely to suffer from high Variance or high bias (depending on how they are tuned). Random Forests use averaging to find a natural balance between the two extremes. Since they have very few parameters to tune and can be used quite efficiently with default parameter settings (i.e. they are effectively non-parametric). Random Forests are good to use as a first cut when you don't know the underlying model, or when you need to produce a decent model in a short time. The Random forest model was run by varying the number of trees that were used and it was found that the best accuracy of 50.5% was obtained when the number of trees was equal to 150. The variation of accuracy with the number of trees constructed is shown in Fig 6.

5.3 Support Vector Machine

Support Vector Machine is an advanced machine learning technique used for classification of both linear and non linear problems. When the training patterns are linearly separable,a linear kernel is used. The linear SVM can be extended to a nonlinear classifier by first using a kernel function to map the input pattern into a higher dimensional space. The nonlinear SVM classifier so obtained is linear in terms of the transformed data but nonlinear in terms of the original data[2]. In this project, the Gaussian RBF kernel function has been used as shown in the equation,

$$K(x,y) = exp(-\frac{||x-y||^2}{2 * \sigma^2})$$

The best accuracy of 53.4% was obtained with SVM.

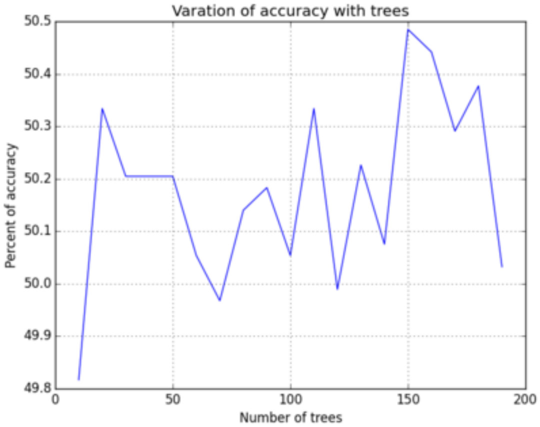

Figure 6: Random Forest Model

6. RESULTS

In this work, K-Nearest Neigbour, Random Forest and Support Vector Machine were considered for recommending the 10 best universities for aspiring graduate students and their performances are summarized below :

Table 2: Accuracy of the models

Baseline	K Nearest Neighbour	Random Forest	SVM
22.2%	50.6%	50.5%	53.4%

From table 2, it is seen that Support Vector Machine performs better when compared to the Random Forest and K-Nearest Neighbor for recommending the 10 best universities.

Support Vector Machine model had a regularization parameter of 1, and an 'RBF' kernel was used and the degree of the polynomial kernel function was found to be 3. Since Support Vector Machine includes a regularization parameter in addition to the k-fold cross validation technique, the accuracy improved well for the test data when compared to the other models.

The K-Nearest Neighbour method is a lazy learner and so, the algorithm did not learn anything from the training data, thereby not generalizing well for the test data. Also not being robust to noisy data, the K-Nearest Neighbour was not successful as a good recommender.

For the Random Forest model, a total of 150 trees were found to constitute the best model, thereby making it very slow for real time predictions.

The overall accuracy turned out to be more than twice the accuracy the baseline. Since this is a multi-class classification problem with 45 classes, the accuracy could not be improved further.

The features - Undegraduate university, GPA, GRE Score and Research experience were found to explain the maximum variance in the data and were used to build the final model. Table 3 shows the importance of the different features used to build the models.

The features - number of journal publications, number of conference publications, industry experience, internship experience and pursuing major did not provide any new information about the data and hence did not contribute to the model.

Table 3: Feature importance

Feature	Importance
Undergraduate University	32.94%
GPA	24.09%
GRE Score	23.83%
Research Experience	19.13%

7. CONCLUSION AND FUTURE WORK

Random Forest, K-Nearest Neighbor and SVM models have been successfully used for the building the university recommendation system. The Support Vector Machine model is found to be comparatively more accurate.

New features like Statement of Purpose, Letter of Recommendation etc. can be analyzed using text mining techniques and could be incorporated if found to improve accuracy. Also, as an extension to this work, recommendation of university with respect to research interest can be made with further study.

8. ACKNOWLEDGMENTS

We would like to thank Prof. McAuley for his guidance and suggestions.

9. REFERENCES

[1] G. C. Cawley and N. L. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107, 2010.

[2] M. H. P. Himani Bhavsar. A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering and Technology*, 1:185–189, 2010.

[3] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.*, 11:1957–2000, 2010.

[4] S. K. Kumar and S.Padmapriya. An efficient recommender system for predicting study track to students using data mining techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 3:7996–7998, September 2014.

[5] J. Luan. *Data Mining and Its Applications in Higher Education*. New Directions for Institutional Research, 2002.

[6] R. B. Rao and G. Fung. On the dangers of cross-validation. an experimental evaluation. pages 588–596. SIAM, 2008.

[7] T. Ruckstieb, C. Osendorfer, and P. van der Smagt. Sequential feature selection for classification. volume 7106 of *Lecture Notes in Computer Science*, pages 132–141. 2011.

[8] C. V. Sacin, J. B. Agapito, L. Shafti, and A. Ortigosa. Recommendation in higher education using data mining techniques. pages 191–199. www.educationaldatamining.org, 2009.