12th International Conference on Recent Development in Engineering Science, Humanities and Management
The International Centre Goa, Panjim, Goa (India)
14th April 2019 www.conferenceworld.in ISBN: 978-93-87793-85-9

PROFILE EVALUATION FOR HIGHER STUDIES USING EDUCATIONAL DATA MINING MODEL

Ketan Dalvi[1], Harsh Chaludia[2], Milan Desai[3], Arnab Chakraborty[4], S.P.Khachane[5]

[1,2,3,4] *B.E., Department of Computer Engineering, MCT Rajiv Gandhi Institute of Technology (India)*

[5] *Asst. Professor, Department of Computer Engineering, MCT Rajiv Gandhi Institute of Technology (India)*

ABSTRACT

This paper describes a methodology for evaluating and recommending students planning for pursuing their higher education in abroad countries on basis of their academic, GRE, TOEFL/IELTS and various other profile scores, also by grading their extracurricular activities. Starting from collection of databases, which was a result of scraping data from various sources, Educational Data Mining (EDM) techniques will be used to discover a set of students having similar academic profiles who got admits from the various universities, thus gaining information about the prospective universities that the students can apply for getting the admits. Hence the students can get a better picture of where they stand and can make an intelligent well-formed decision. A thorough analysis of the profile is done which not only suggests the universities but also gives the required steps necessary for profile improvement in order to get admit into the Dream University.

Keywords - admits, Educational Data Mining, evaluation, rejects, scraping

I. INTRODUCTION

Educational Data Mining (EDM) is a sub-domain of data mining which deals with data from academic student databases which is used to develop various techniques and to recognize patterns that are unique.

A university education has become a basic part of an aspirant willing to gain expertise in their desired domain. Admission to university is therefore a topic of importance. How a student chooses a university, and conversely how a university chooses a student, determines the success of both sides in carrying through the education. Every year, hundred of thousands of these students, and many others who don't get admission, apply to American post-graduate programs and, in the process, discover that there is a dearth of reliable sources to aid them in making an informed decision. There are several sources that provide admission related statistics, but do not cater to individual needs, thus, leaving the applicant with the only option of guessing and hoping for the best[1].

A typical university application packet comprises of transcripts, standardized test scores, letters of recommendation, a statement of purpose that expresses students aims, ambitions and research interests, and descriptive answers to a few additional questions. Test scores include GRE, language test scores - such as TOEFL or IELTS etc. Universities, then, evaluate application packets based on rules or heuristics which are unknown to students, and release

decisions[1]. Since requirements, deadlines and the specific process to meet them is university specific, the applicant needs to first choose the universities he would apply to. Given the uncertainty, a naive solution is to apply to a large number of universities. But, the more the number of applications, the higher the investment of time and energy. This also implies a large monetary investment, which is a major concern for applicants from developing countries[2]. One of the strategies to circumvent this is to categorize the universities into buckets so that one only applies to a few representatives from each category.

1.1. Evaluation

The evaluation model involves a user friendly web interface in which engineering undergraduate students can evaluate their scores for getting to know their dream universities and chances of getting an admit. Apart from normal entrance scores, the extracurricular activities are also taken to understand their domain of interest. With the help of accurate outcomes we eliminate the margin of errors by comparing various machine learning algorithms like LR, LDA, RF, NB-C, and SVM.

1.2. Abbreviations and Acronyms

The terminology which are used throughout the paper are shown below in the following Table I :

Table I : Abbreviations and Acronyms

GRE	Graduate Record Examination
TOEFL	Test Of English as a Foreign Language
IELTS	International English Language Testing System
LR	Logistic Regression
SVM	Support Vector Machine
NBC	Naive Bayes Classifier
LDA	Linear Discriminant Analysis
RF	Random Forests
CART	Classification and Regression Trees
LOR	Letter Of Recommendation
SOP	Statement Of Purpose
CV	Curriculum Vitae

1.3. Existing Problem

In the past, a lot of work on employing data mining techniques in the field of education were undertaken to predict outcomes but most of them included their customized version of a solution to a problem which is not a comprehensive approach to deal with it.

In current scenario the students who are looking to pursue higher education in abroad have to depend on study abroad consultancies, human decisions based on past experiences and web forums for taking suggestions regarding the university selection from current students and alumni .This is one of the major barrier for students who are planning to pursue higher education abroad.

There is a large amount of data available on internet present in disparate form, and the major barrier to obtain the required meaningful data is the growing size of database and the versatility of the domains[3]. To overcome these barriers and to find the useful patterns of the data, various regression and classification techniques can be used to find useful information from data. Regression analysis entails looking at dependent variables (outcomes) and an

independent variable (the action) while also assessing the strength in the association between them.

1.3.1. Current Scenario

Admission candidates share parts of their application information on the online website portals such as The Grad Cafe, Yocket, etc. Information collected from such websites is an approximation of students' perspective. All of the metadata from application can be used to extract features and then can be fed into a classification algorithm[1]. The benefit of this modelling is that it uses the same independence assumption which is practically valid, and is flexible enough that it allows experimentation with multiple classification algorithms. The goal of this modelling is to identify as many correct decisions as possible so that the student has a fair estimate of his chances of getting into a school beforehand.

Yocket is one of the platform for students aspiring to study abroad. It has a community-based approach that helps an individual seeking admissions abroad to build a rapport and approach students for advice and testimonials. The main USP of this community platform is that it enables students to seek help from peers, seniors, and even professional experts. Another interesting feature is that Yocket has implemented a machine learning tool that helps users know their probability of getting into a particular university.

1.3.2. Drawbacks Of Existing Systems

All the existing systems present online working in this field are all concerned with providing an instant decision with clustering algorithms[3]. We analysed admission predictors of various external agencies like GyanDhan and Yocket and we found out that these models are inconsistent because when practical value was compared with the original decision then it turns out to be false with huge margin of difference.

The inconsistency was because of the following parameters:

i. **Accuracy** – Just choosing a random algorithm to predict an outcome does not gurantee in providing an accurate outcome even if the accuracy in that algorithm is more than 90%[4].

ii. **Dynamicity** – Predictions based on sample data is not efficient to provide a decision every now and then. It is because every year the admits and rejects data keeps on changing and database needs to be updated regularly.

II. PROPOSED SYSTEM

The figure below (Fig. I) depicts the block diagram of the proposed system which will be implemented as web-based application.

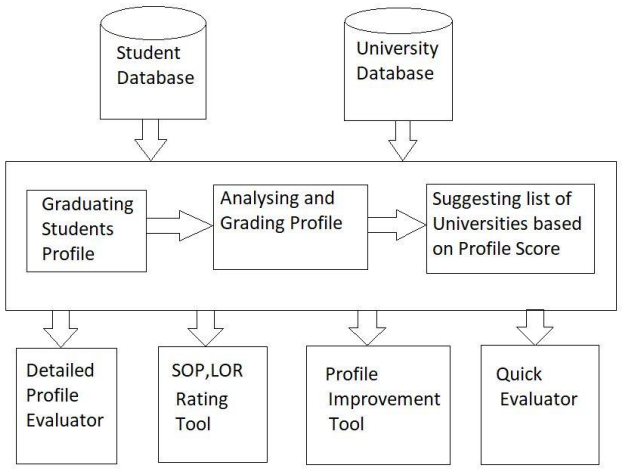

Figure I : Block Diagram

The university database is obtained from the official university websites wherein they mention the previous year cutoffs and details of the admitted students. Most of the M.S. aspiring students use the freely available online services Yocket, Edulix, GradCafe, etc[5]. which also helped us to create a University Database wherein the details of the admitted, rejected students from different universities are available. The student database contains the profiles of the M.S. aspiring students willing to evaluate their respective profiles. It will be the live input where one enters the details required for profile evaluation.

The system has following four evaluation features :

i. **Complete Profile Evaluation :** It is the primary feature of the system which can only be used by the registered users. Student will have to enter all details like GRE, TOEFL/IELTS scores, GPA, No. of international/local publications, no. of internships, work experience, etc. A thorough analysis of the profile will be done and thus an invoice report will be generated which will contain details like probablity of getting into a particular university along with recommended courses. The recommended universities will be classified as Dream, Safe and moderate based on the strength of the profile and the university ranking.

ii. **Get Rating :** This feature is used for rating the documents like SOP, LOR and CV based on their strength. The rating would be done on the scale of 1 – 5 where 5 will be for outstanding document and 1 will be for below average document. One would get the glimpse of the strength of document and thus it can be improved if its rating is not satisfactory. Even this feature is accessible only to registered user.

iii. **Profile Enhancement :** Every student dreams of a university where one would like to study but are unable to get into it due to lower test scores or even weak profiles. Many students think of giving multiple attempts of GRE, TOEFL examinations and are also desperately figuring out ways for improving profile in order to get admit into dream university. Profile Enhancement tool is thus used for finding out the loopholes in the profile and thus suggesting measures for improving profile required for getting into desired university.The result may be anything like the more GRE, TOEFL/IELTS marks, GPA needed to get into Dream University, or the lack of work experience, less no of research publications as per university criteria.

iv. **Quick Evaluator :** Quick Evaluator is a common feature which can be used by any person visiting the website. Upon entering the minimal parameters like GRE, TOEFL/IELTS scores, GPA, SOP, LOR and no of Research Papers published, it gives the cluster of 12 Best universities based on the overall strength of profile. The main purpose of this feature is to give the glimpse of the system to the user.

III. METHODOLOGY

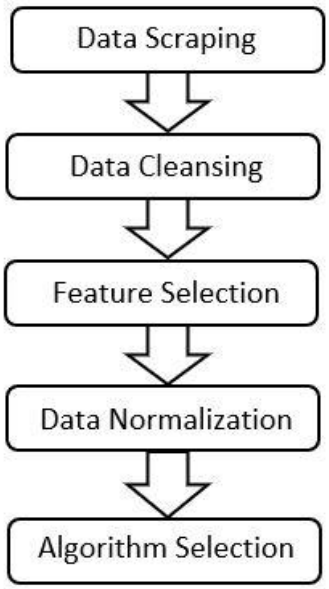

Figure II : Methodology Flow Diagram

1) **Data Scraping :** Identifying correct dataset is the crucial step in training any automated system and thus gathering of correct dataset was of utmost important[2]. Thus, we collected previous 10 years of data of M.S. pursuing students by scraping from the popular websites like Yocket, Edulix, etc. Data was scraped by using custom made python scripts and also by using python libraries like Beautiful Soap and Scrapy.

2) **Data Cleansing :** It is necessary to fix the erroneous records as it may decline the efficiency of prediction. If the record did not have mandatory fields such as GRE Verbal & Quantitative, or GPA, then such a record was completely deleted. Missing year and missing term were replaced with the mean of existing values. Records which had invalid values such as negative GPA, or out of range value for GRE score components were removed in a rule-based fashion. Records having incompatible data-types such as string for GPA, or numerical value for degree were removed. Sometimes students refer to the same degree using various acronyms such as BE or BEng for Bachelor of Engineering. Similar is the case for departmental majors such as CS or Comp Science for Computer Science. Also, sometimes there are spelling mistakes and typos in these fields. Each of these fields is, hence, normalized to manually created codes e.g. {ba, bs, be, btech etc.} instead of original strings. The code assignment was done using regular expressions or a manual mapping of string value to code. There were no such labels present in dataset as Waitlisted, or Conditionally Accepted. We excluded any application that was classified as Result Not Available because it simply represents either a missing value or a pending decision[1].

3) **Feature Selection :** As the dataset we have might contain different parameters like username, user-id which is usually not meant to be used in the algorithm, needs to be ignored and important features should be selected. The following are the main predictive features in the model:

i. **Admits and rejects of universities** – This is one of the past applicant's decision from their shortlisted universities which is one of the model's highest weighted features.

ii. **Undergraduate CGPA** – Most of the universities emphasizes on selecting students with higher CGPA even when you have an

ISBN: 978-93-87793-85-9

average GRE & TOEFL score. This is because the university assumes the student to be a hard working student which might make an excellent graduate and a good fit for the admission.

iii. **GRE Quant score –** Every university has their average scores listed on their website from past year records. Also high quant score is expected from students because every university knows that the student might come from rural areas and they might not be well versed with communication in english.

iv. **GRE Verbal score** – An average verbal score is expected from students. Research study says that the universities usually accept low verbal score in comparison with quant score.

v. **Research papers** – Students applying to specialization in branches are expected to show their talent in terms of research work or achievments in those fields. This is because many ivy leagues prefers students with some research in a field that they want to work for.

vi. **Letters of recommendation & SOP rating** – It is one of the basic factors that constitutes for an evaluation of a profile. A self made documents are expected which is free from plagarism.

vii. **TOEFL score** – After GRE score, this is what they look for, which is the english test. The sub-section from toefl exam are reading, listening, speaking and writing. A minimum of 20 is expected in each section and it differs from university to university.

After a thorough theoretical evaluation, the importance of these features is calculated practically with the dataset whose comparison is shown in Fig. III. It is evident from the below table that CGPA plays an important role in the admission process followed by GRE, TOEFL, LOR, University rating and research. Thus, we get to know what the university is looking in an excellent graduate that they can provide an admission for.

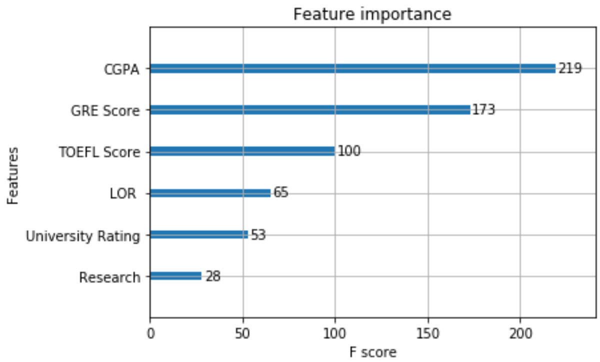

Figure III : Feature Comparison

4) **Data Normalization :**

As we are using the optimization algorithm for finding the best-fit parameters, it is necessary to normalize the data. For that purpose, we are using the sigmoid activation function. We'd like to have an equation we can give all of our features and it will predict the class. In the two class case, the function will spit out a 0 or a 1. The sigmoid is given by the following equation:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \ldots(1)$$

At 0 the value of the sigmoid is 0.5. For increasing values of x, the sigmoid will

approach 1, and for decreasing values of x, the sigmoid will approach 0. On a large enough scale, the sigmoid looks like a step function. For the logistic regression classifier we'll take our features and multiply each one by a weight and then add them up. This result will be put into the sigmoid, and we'll get a number between 0 and 1 wh. Anything above 0.5 we'll classify as a 1, and anything below 0.5 we'll classify as a 0, which is shown in following graph (Fig. IV).

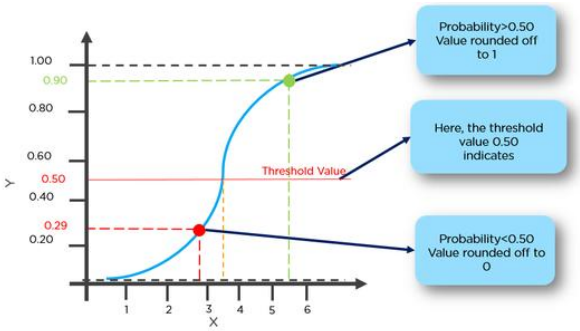

Figure IV : Sigmoid function variation

The input parameters like GRE, TOEFL, University rating, LOR, CGPA and research are normalized on a scale of 1 to 5 as shown in Table. V. The idea behind normalization of these scores is to eliminate the huge difference values between parameters. This not only helps in eliminating erroneous data but also helps to predict with different strategies like point system prediction or point based prediction.

Table II : Sample data of Normalized Parameters

GRE Score	TOEFL Score	University Rating	LOR	CGPA	Research
4.7	4.642857	3.75	4.375	4.567308	5.0
3.4	2.678571	3.75	4.375	3.317308	5.0
2.6	2.142857	2.50	3.125	1.923077	5.0
3.2	3.214286	2.50	1.875	2.996795	5.0
2.4	1.964286	1.25	2.500	2.259615	0.0

5) Algorithm Selection : The algorithm selection factor is very much important because it decides the overall accuracy of the model. By testing sample data in various algorithms like LDA, LR, KNN, CART, RD, NB, & SVM as shown in Fig. VI, we decided to choose logistic regression as the best fit for our version of model which was giving 87.5% accuracy followed by RF,NB giving 86% accuracy and LDA giving 85.5% accuracy. We neglect the remaining algorithms because they are not a fit for our model being evident from comparison. But these accuracies does not constitute in predicting a value which is true. This is basically tested with the error rate provided by the above mentioned algorithms. Lesser the error rate, less are chances that the outcomes might be false when compared with the sample dataset.

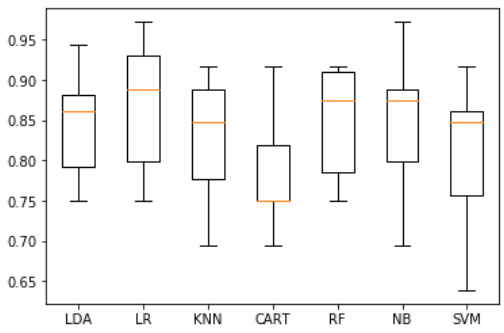

Figure V : Machine Learning algorithm comparisons

It was only with logistic regression we were able to achieve maximum average accuracy while testing various dataset. We following the method of predictive analysis using logistic regression implementation.

IV. ALGORITHM IMPLEMENTATION

i. The first step is to implement the sigmoid function.

It converts a continuous input into admit value between zero and one. This value can be interpreted as the class probability theta(x):

$$\varnothing(x) = \frac{1}{1+e^{-x}}$$

……(2)

ii. The Gradiant Descent is implemented as a second step followed by the cost function 'J' parameterized by theta. In each iteration the values of theta changes based on the gradiant descent equation given below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update θ_j for all j).

…(3)

iii. Now we calculate the cost function which is expected to reduce in each iteration which is given by:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

…..(4)

Now that we have the optimal model parameters as shown in Fig. VII for our data set. Next we need to write a function that will output predictions for a dataset X using our learned parameters theta. We can then use this function to score the training accuracy of our classifier.

```
Model --> Logistic Regression
Overall Accuracy: 90.0
              precision    recall  f1-score   support

         0.0       0.92      0.92      0.92        26
         1.0       0.86      0.86      0.86        14

   micro avg       0.90      0.90      0.90        40
   macro avg       0.89      0.89      0.89        40
weighted avg       0.90      0.90      0.90        40
```

Figure VII : LR Parameters

V. CONCLUSION

The idea behind this research was to develop a model that can be used by the aspiring students who wants to pursue their education in the USA. Various classification and regression algorithms were tested, developed and used for this research. Logistic Regression proved to be a best fit for our model when compared with various algorithms. Our system model can be used by the students for complete profile evaluation, getting document ratings, enhancing their profile and a quick evaluation their chances of getting shortlisted in a particular university with an average accuracy of 70% being tested on various datasets. A web application is made to provide user an interactive interface which was easy to use for the users from various non-technical backgrounds. PHP framework – Codeigniter v3.3 was used to create the user interface. The overall objective of our model to eliminate the

external dependencies was achieved successfully as the system allowed the students to get to know their profile and where it stands without investing time on multiple websites. Also, it will help them save money that they would be spending on external agencies, and consultants where they would be still confused as earlier. As our ultimate objective was to automate human decisions followed by providing faster evaluation procedure at once, we not only are working to make a system that helps students but also to make university decisions simplified in their perspective.

VI. FUTURE SCOPE

Creating the model with additional parameters for various exams like GMAT/ SAT etc. which will also help undergraduate and MBA aspirants who wants to pursue their education abroad. Generalizing a universal script that makes admission prediction process easy by taking all desired criteria or parameters into consideration which satisfies for any given entrance exam.

To make a common portal for storing admits and rejects which will provide a dynamic database. This will help the model to constantly change according to university requirements. Also, we can implement a system in university perspective that will help them automate decision by selecting aspiring student who are a best fit for their requirements. If such a system is implemented then it will fasten university decisions and will help students save application fees which is normally wasted.

Python scripts are normally inoperable in windows hosting server and normally used in Linux/ubuntu hosting servers. Since these scripts are integrated with php framework, the time taken to run the script usually -takes 7-8 seconds. If more processing power is provided then multiple scripts can run parallelly and making the user experience a faster experience of functionalities in the system.

REFERENCES

Theses :

1. Thesis by Narendra Gupta, "*AMERICAN GRADUATE ADMISSIONS: BOTH SIDES OF THE TABLE*", Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2016

Journal Papers :

2. Joe Manley Gnanasekaran, Ramkishore Swaminathan, Swetha Krishnakumar, "University Recommender System for Graduate Studies in USA"., Proceedings of the *28th Annual ACM Symposium on Applied Computing*, ACM New York, NY, USA , *pp. 1852-1858* March 2013.
3. Tirtha Chavan, Amruta Barretto, Tanmay Chavan – "Data Evaluation and Knowledge Extraction for Students Using Clustering and Classification", 2013 *(IJCSIT) International Journal of Computer Science and Information Technologies,, pp.187-193 ,2012*.
4. Simon Fong and Robert P. Biuk-Aghai, "An Automated University Admission Recommender System for Secondary School Students", *The 6th International Conference on Information Technology and Applications (ICITA 2009)*
5. [Kassegne, 2016] Kassegne, S. (2016). Edulix - premier site for scholars – "education crowdsourced". Web.