Expediting International Student Admission Process Using Data Analytics

Mounicasri Valavala\textsuperscript{1}, Wasim Alhamdani\textsuperscript{2}, Hemalatha Indukuri\textsuperscript{3}

\textsuperscript{1}University of the Cumberlands, Williamsburg KY 40769, USA
mvalavala1456@ucumberlands.edu

\textsuperscript{2}University of the Cumberlands, Williamsburg KY 40769, USA
wasim.alhamdani@ucumberlands.edu

\textsuperscript{3}S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh 534204, India
indukurihemalatha@gmail.com

Abstract. The student admission process is a critical step from both student and university perspectives. The students will benefit if the universities take admission decisions sooner rather than later, leaving them time to opt for another university in case of rejection. However, with the level of manual work involved in the student admission process, universities often take long durations for the admission decisions. Expediting the admission process using Data analytics starts new research across different tasks in educational institutes. The research presents a model to automate each step of the international student admission process in universities across the USA, using Data analytics. The model's distinguishing capability is that it performs both predictive and descriptive analytics to analyze different student's application sections and adapt to different admissions types with minor changes.

Keywords Student Admissions. Data Analytics. Word Cloud. Machine Learning. Natural Language Processing.

1 Introduction

Student admission is a tedious process for both students and the universities. Nevertheless, the technology helped eliminate the paperwork, making it far better than a decade ago. The admission process depends on the degree, student accomplishments in the selected area, course grades, recommendation letters, Statement of Purpose (SOP), and other factors like nationality. With the increase in globalization, there is tremendous interest to study in prestigious universities across the globe for better opportunities.

The 2020 report on International Educational Exchange discloses that over a million international students admitted to USA universities during 2019/2020, and the United States continue to enjoy international students' top destination position \cite{1}. Hegarty (2014) points out that international student education is a significant industry contributing to the US economy \cite{2}. The admissions continue to grow despite laborious visa processes and stringent immigration rules \cite{2}. Considering the criteria mentioned above, we chose to expedite the international student admission process in the USA using Data analytics, which is a promising solution across different industries.

The student admission process contains a student application that takes a student's personal information, education history, test scores, activities participation, and Statement of Purpose (SOP). Also, most of the universities require recommendation letters that support the applicant's capabilities. Some university website asked the applicants to supply no more than two recommendation letters, as it will take more time for the staff to read and evaluate the recommendation letters \cite{3}. The other documents accepted are transcripts, general or subject test (GRE), English proficiency test (IELTS, TOEFL) score reports.

This research will categorize the application requirements into two categories and provide a holistic analysis of finalizing student admission. The first category is to perform predictive analytics using supervised classification algorithms on education history and score reports. The second category is to apply natural language processing techniques to speed up the analysis of recommendation letters and SOP. We will present the model theoretically and present the implementation results in the upcoming articles.

2 Related Work

The research of using Data Analytics in the education sector is scattered around different academic levels and across different activities within an educational institution. There is extensive research in this area to predict student
performance, perform association rule mining from both teacher and student perspectives, and sentiment analysis to
decide students' interest levels to admit into a particular university. According to our knowledge, there is limited to no
research specific to international student admissions automation using Data Analytics. There is an urgent need for
research to implement a data analytics framework in the education domain [4]. The model presented in [4] addressed
using data science models that provide information to the instructor about the need for intervention to improve student
performance.

A case study presented in [5] presents a predictive model to foresee students' performance based on the given
information during Nigeria's admission process. The limitation of this research is that it performs only predictive
analytics. Hence, this model is not adaptable to admissions across the globe where the applicant often needs to submit
textual information such as recommendation letters and SOP. Robin & Brian (2010) presented research on students' technology usage in undergraduate admissions [6]. However, it covers only the communications means to connect students with the universities.

The authors in [7] provided a review of existing research using predictive analytics to evaluate student
performance. According to this survey, most of the research relies mainly on the cumulative grading point average
(CGPA) and other internal data assessments [7]. The models presented in [8] predict poor student performance in
academics using academic data and socio-economic details. International Data Mining society coined a term called
Education Data Mining (EDM) to explore the educational domain's data and draw insights to understand students and
the learning environment [8, 9].

The conceptual model discussed in [10] explains how the universities can use data analytics in different stages of
student-college association. The authors mentioned that data analytics helps in admission decision making using
predictive analytics to predict a student’s performance in the future and natural language processing techniques to
analyze qualitative content [10]. We consider the author’s inputs on data analytics in the admission process to create a
data analytics model for the student admission process. The research in [10] extended the conceptual model to
implement data analytics during the student and post-student stage.

The prediction model presented in [11] predicts the student enrollment using the admission data, such as the
selected academic program, current course plan, intended course plan, age group, residency status etc. The test scores,
and school grades, which influence the admission decision, are not considered and the authors mentioned that this
should be part of future work [11]. The current research addresses this research gap by including the test scores,
school grades, and community activities.

3 Research Contribution

The existing research in this area is confined to predicting student grades within the coursework, technology usage in
the education sector, analyzing university and student statistics, and sentiment analysis of social media content. The
authors in [10] provided a high-level view of the model for prospective students in universities. We will delve into
further details and create a theoretical model for international student admissions in the USA. The proposed model is
adaptable to different admission types with minor changes.

The current research will automate the student admission process using predictive and descriptive analytics. The
proposed model ensures that it analyzes both the qualitative and quantitative content of a student application and
provides the student performance prediction and word clouds to represent the textual information. As a result, the
time spent on manual analysis of recommendation letters, essays, and SOPs can be reduced drastically. Also, the
predicted performance measure helps the faculty to make an admission decision. We also perform qualitative analysis
to get the feedback of the proposed model from the admission departments across different universities.

4 Methodology

The proposed model expedites the student admission process using Machine Learning (ML) and Natural Language
Processing (NLP) techniques. We analyze application data using ML and textual content, such as recommendation
letters and SOP using NLP. The model contains two main modules, namely the Application data analyzer and
qualitative data analyzer, and the model's outcome is input to the admission office to make an admission decision.

The model reduces the extent of manual effort required by the admissions department to make the admission
decision. As a result, the turn-around time for an application will be considerably less. The below schematic design
gives an overview of the admission model.
4.1 Application Data Analyzer

The Application Data Analyzer (ADA) contains a feature extractor, data preprocessor, ML admission classifier, and ML performance predictor. ADA takes application data as input and outputs the report with the predicted student future performance and a Boolean value that indicates whether a student with similar academic credentials was admitted to the school. We use regression and classification ML models to generate the ADA outcome. Using ADA eliminates the need for manual review of application data to determine the student eligibility.

4.1.1 Feature Extractor: The application contains different sections, such as personal information, family information, education history, test scores, extracurricular activities, and essays. The ML models in the ADA module only need education history, community activities and test scores. Hence, the feature extractor retrieves the required information from the application data and sends it to the Data preprocessor.

The feature of the Dataset for ADA are Career Interest, Intended Highest Degree, Graduation Class Size, Class Rank, GPA Scale, Rank Weighting, CGPA, Schedule Type, ProgressionCategory, Honors Exists, Honors Title1, Honors Grad Level1, Honors Recognition1, Honors Title2, Honors Grad Level2, Honors Recognition2, Honors Title3, Honors Grad Level3, Honors Recognition3, Honors Title4, Honors Grad Level4, Honors Recognition4, Community Based Participation, and Community Based Participation Count. We create the dataset structure based on the student application in the common app [12]. The Common App provides a common student application for different universities across the USA.

4.1.2 Data Preprocessor: Data preprocessing is an essential step before applying the model, as data quality impacts the model accuracy. The Data preprocessor cleans the data in four steps, such as cleaning the missing values, converting the categorical variables, standardizing the data, and handling multicollinearity. We will follow the strategies, such as deleting the record, imputing the missing value, and deleting the column based on the percentage of missing values to clean the missing data. We will use one-hot encoding to handle the categorical variables and standardize the data to fall on a fixed scale to ensure that the impact of values in one column does not dominate the model result. Finally, we will check each column's correlation with other columns in the Dataset using a Heat map and eliminate the highly correlated column to avoid the dual impact of a feature on the model.
4.1.3 Machine Learning Admission Classifier: The ML classifier follows a supervised learning approach, where the admission history of students in the school constitutes the training dataset. The classifier's outcome is "Yes" or "No," indicating whether a student can be admitted or not. We will use Random Forest to perform the classification, based on the data size and data sparsity.

Random Forest (RF) is an ensemble model that works based on the aggregation of Decision Trees' outcome and is robust enough to handle variance. RF is emerged as a standardized classification algorithm in innovative environments [13]. RF, as a classification model, chooses the label using a voting mechanism, where the output of the highest number of trees is considered.

The advantages of RF are the robustness to overfitting, model effectiveness with variable data sizes and unbalanced datasets, and support to fault tolerance [14]. For our research, we will generate the data to test the model, and the data will be replaced with the real data from the universities when the model is implemented in real-time. RF will ensure that the model fits both scenarios.

4.1.4 Machine Learning Performance Predictor: The ML predictor uses a linear regression model to predict the CGPA of a student based on his/her educational history and test scores. The different variants of linear regression are Bivariate linear regression, where the model can fit two independent variables to predict a dependent variable, and Multivariate linear regression, where the model can handle more than two independent variables to predict a dependent variable. A study to predict the CGPA of engineering students proved that usage of Multivariate linear regression is a promising solution to predict the performance [15-17]. The below equation represents Multivariate linear regression, where \((x_1,x_2,...,x_n)\) are independent variables, and \(y\) is the dependent variable:

\[
y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n
\]

The Multivariate regression model's advantages are the support for complex data, personalized nature, and each independent variable's contribution to the result [16]. There is extensive research to predict student performance across different courses, and multivariate regression is proven as an efficient model. The performance predictor uses the Multivariate regression to predict the CGPA of a prospective student in the future. We will perform hyper-parameter tuning to achieve maximum accuracy.

4.2 Qualitative Data Analyzer

The qualitative data analyzer analyzes the textual content, such as Recommendation letters and SOP. The model will drastically bring down the time required to read the content by creating content summaries and word clouds. We use natural language processing techniques to perform these activities. This module's components are a data preprocessor, a word cloud generator, and a text summarization model. The module takes recommendation letters and SOP as input and sends the summary and word cloud generated from this content to the admission office. The below sections will explain the component of this module in detail.
4.2.1 **Data Preprocessor**: Data preprocessing is a mandatory step for NLP to focus on the required content. The different preprocessing steps are special characters removal, tokenization, stop words removal, and lemmatization. We use all the above-mentioned preprocessing steps for the word cloud generator, whereas we only first three preprocessing techniques for the text summarization model. The reason to skip the lemmatization in the text summarization model is to maintain the word's state in the sentence, which will impact the summary extraction.

![Data Preprocessing Steps for Qualitative Data Analysis](image)

In special characters removal, the preprocessor will remove the symbols, numbers, and any other character that is not an alphabet. In tokenization, the sentence will be broken into words, which will be used as a bag of words for word cloud and text summary generation. Stop word removal eliminates the connecting words in the sentence, such as "is", "a", "the", "an", "to" etc. Finally, lemmatization will take a word to it's base form. For example, the word "studies" is used in a sentence; the lemmatization step converts this word to “study”.

4.2.2 **Word Cloud Generator**: The word cloud generator generates a word cloud with the key terms in the content. Word Cloud is a graphical representation of words where color, size, and weight represent the features like word frequency. Word clouds are considered as an effective means to draw insight into the content, construct the relationship between different keywords, and quickly retrieve the knowledge from information [18]. There are different word cloud applications, such as student response assessment, content comparison, feedback analysis, etc.

The word clouds can be generated using the reporting tools like Tableau, programming languages like Python, and other open access websites. We use Python to generate the word clouds for quantitative content. We chose to use word cloud over Named Entity Recognition (NER), as NER functionality is restricted to named entities. In the case of student content, identifying named entities alone does not provide the complete picture. We use the WordCloud package of Python to generate the word cloud and Matplotlib to create the word cloud's graphical representation. The below word cloud shows the key-terms in a recommendation letter and gives a high-level view of the student's achievements and skill.
4.2.3 **Text Summarization Model:** The text summarization model summarizes the content in a recommendation letter or SOP. The two types are summarizers are extraction summarizer and abstraction summarizer. The extraction summarizer extracts the summary based on word frequency count, and the summary is a combination of the sentences with high-frequency words. The abstraction summarizer is a semantic-based summarizer that works based on the content's semantics.

Extractive summarizers often result in better performance than abstract summarizers due to the practical issues with semantic representation and the need for complex analysis techniques. There are no pure abstractive summarizers available today. Extractive summarizer is used before abstractive summarizer [19]. We use an extractive summarizer to summarize the content in recommendation letters and SOP, and the implementation follows the below steps.

- Calculate the word frequency.
- Tokenize the content into sentences.
- Calculate the score for each sentence using the word frequency.
- Identify the sentences with a score greater than the average score of all the sentences in the content.

The model will create a report with the content summary for recommendation letters and SOP. This report will help the admission officer to get a view of the textual content presented by the prospective student. Word Cloud that gives highly used terms and a summary that gives an abstract view of the content will provide a complete overview of the submitted documents, resulting in faster application screening. We will review the outcome of the textual data analysis with the admission officers to get their feedback. Further improvisations will be implemented based on the qualitative analysis of the admission officer's interview outcome.

5 **Conclusion**

The proposed model speeds up the application process without losing the integrity of the application screening process. It also provides a complete design to cover both qualitative and quantitative content. There is scope to improvise this model based on the feedback from the admissions department. The model helps to manage resources efficiently due to the prediction and text processing reports. We will present the implementation details and results of this model in the next articles of this series.

We can incorporate the design to analyze the extracurricular activities section in future work, contributing to some universities’ admission rate. The model can be extended to decide funding to a student and predict student
performance in non-academic activities. Though the model is directed towards international student admissions, it can be extended to other admission types.

References

[1] United States Hosts Over 1 Million International Students for the Fifth Consecutive Year. (2020, November 16). Institute of International Education. https://www.iie.org/en/Why-IIE/Announcements/2020/11/2020-Open-Doors-Report
[2] Hegarty, N. (2014). Where We Are Now –The Presence and Importance of International Students to Universities in the United States. Journal of International Students, 4(3), 223-235. Retrieved from https://www.jjis.org/index.php/jjis/article/view/462
[3] How to Apply: International. (2020, October 26). Undergraduate Admissions. https://admissions.unh.edu/apply/international-students
[4] Liu, M.-C., & Huang, Y.-M. (2017). The use of data science for education: The case of social-emotional learning. Smart Learning Environments, 4(1), 112. https://doi.org/10.1186/s40561-016-0040-4
[5] Oladokun, V., Adebanjo, A., & Charles-Owaba, O. (2008). Predicting Students’ Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course.
[6] Robin, L., & Brian, F. (2010). Using Technology in Undergraduate Admission: A Student Perspective. Journal of College Admission, 10–17. https://eric.ed.gov/?id=EJ893890
[7] Shahiri, A. M., Husain, W., & Rashid, N’. A. (2015). A Review on Predicting Student’s Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157
[8] Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje, 10(3), 119–125. https://doi.org/10.1109/rita.2015.2452632
[9] educationaldatamining.org. (n.d.). Educational Data Mining Society. https://educationaldatamining.org/
[10] Attaran, M., Stark, J., & Stotler, D. (2018). Opportunities and challenges for big data analytics in US higher education: A conceptual model for implementation. Industry and Higher Education, 32(3), 169–182. https://doi.org/10.1177/095042221879937
[11] Cirelli, J., Konkol, A. M., Aqlan, F., & Nwokoji, J. C. (2018). Predictive analytics models for student admission and enrollment. Proceedings of the International Conference on Industrial Engineering and Operations Management, 2018(SEP), 1395-1403.
[12] Common App. (n.d.). Common Application. https://apply.commonapp.org/common/6/33
[13] Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19(1), 567. https://doi.org/10.1186/s12859-018-2264-5
[14] Liaw, A. & Wiener, M. Classification, and regression by randomForest (2002). R News 2, 18–22
[15] Rajalaxmi, R. R. Natesan, P. Krishnamoorthly, N. Ponn, S. (2019). Regression Model for Predicting Engineering Students Academic Performance, International Journal of Recent Technology and Engineering, 7(6S3), 71-75.
[16] Elbadrawy, A., Studham, R. S., & Karypis, G. (2015). Collaborative multi-regression models for predicting students' performance in course activities. Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15, 103–107. https://doi.org/10.1145/2723576.2723590
[17] Khan, W. Z., & Zubaidy, S. A. (2017). Prediction of Student Performance in Academic and Military Learning Environment: Use of Multiple Linear Regression Predictive Model and Hypothesis Testing. International Journal of Higher Education, 6(4), 152. https://doi.org/10.5430/jhe.v6n4p152
[18] DePaolo, C. A., & Wilkinson, K. (2014). Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. TechTrends, 58(3), 38–44. https://doi.org/10.1007/s11528-014-0750-9
[19] Allahyari, M., Pouriyeh, S.A., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. ArXiv, abs/1707.02268.