Graduate School Application Data Mining: A Systematic Approach

1st Tuan Vinh
Emory University
tvinh@emory.edu

2nd Shuyang (Simon) Bian
Emory University
simon.bian@emory.edu

3rd Nazif Azizi
Emory University
mazizi4@emory.edu

***Abstract*—Recent trends in graduate program enrollment have sparked debates over increased competitiveness, a perceived preference for domestic students, and the declining importance of standardized test scores and academic publications. The economic uncertainty post-COVID has intensified these discussions, leading some to question the value of higher education investments. This study leverages nearly one million data points (853,142 results) from The GradCafe to examine these claims critically. We have found that overall competitiveness remains consistent over the years, with limited diacrhonic impact on admission decision. As decision tree was modified, we found that fine mapping the data to limited categories improved accuracy on decision outcome decision. An interpretable approach on the admissions was attempted, and the nationality and major have been the two factors that exerted the greatest effects on students' application results. We used geographic data mining on school and admission rate, and has found a strong impact on eastern coast on the competitiveness in admission decisions. Our next steps involved using large language model to further process and probe prospective students' decision-making processes, and explore developing a school recommendation system based on students' profiles using established and innovative models. By providing empirical evidence on these fronts, this research has illuminated the realities of graduate program applications in the current socio-economic context and contribute to developing tools that support informed decision-making for prospective students.**

***Index Terms*—Data mining, Big Data, Graduate Course**

I. Introduction

Over the years, graduate program enrollment has changed significantly, marked by intensified competition (Council of Graduate Schools, 2020), an observable preference for domestic candidates (Inside Higher Ed, 2020b). Under such new circumstances, the traditionally valued metrics, including absolute academic performance level, exemplified by standardized test scores (Council of Graduate Schools, 2020), shift their significance. Post-COVID, further uncertainties have initiated students' reassessment of investments in terms of times and financial requirements towards their pursuit of a higher degree (McKinsey & Company, 2020).

Since its foundation in the early 2000s, the GradCafe (https://www.thegradcafe.com) has been an essential resource for prospective graduate students. While there exists a fundamental use of introduction of students to each other through its "Forum" feature, its more enriched - while seldomly explored resources - have been its accumulation of over 740,000 submissions over its collection of nearly all graduate schools in the United States. By sharing admissions experiences, students have provided each other with valuable insights, fostering a sense of community among those navigating the complicated and arduous graduate school application process. Several challenges lay ahead for using the potential GradCafe data to help extrapolate for an underlying trend within the PhD applications. First, it has been noticed that these statistics are voluntarily provided. Therefore, many would submit non-cleaned data points, potentially challenging data pre-processing. Second, while the given data shall be quite considerable from a longitudinal standpoint, the number of attributes is limited for a single data row, and researchers need to closely monitor these expanded attributes to allow for data mining on exquisite features. These limited attributes, correlated with a natural-language-based description of "non-score" experience, are subjective and can be prone to manipulation by the individuals who submit. Therefore, more significant efforts should be put into improving the data quality, and careful clustering must be done to avoid over-generalization or over-simplification of the given statistics.

II. Related Work

From a data-mining aspect, Liu, Su et al. (2017) investigated determinants of admission into Computer Science (CS) graduate programs in the United States. Using Generalized Linear Models (GLM), Generalized Additive Models (GAM), Random Forests and Gradient Boosting, and Bayesian Networks, they found heightened competitiveness for international students, particularly in Ph.D. programs. Key predictors of admission success identified include the number of publications and faculty size. Nevertheless, additional factors directly processed through natural language parsing, such as publication records, have yet to be supplemented in the analysis to enrich insights for prospective applicants. Makkinje (2015) similarly used statistics for the Physics Program but focused only on GPA and GRE as necessary quantifiable attributes. Other works, such as Gupta (2016) and Ferreira (2022), have similarly focused only partially on quantifiable attributes, with data tracing back to pre-COVID times that need to be more diachronic for the shifting dynamics in the current US higher education constructs.

Meanwhile, nowadays, the field of Natural Language Processing (NLP) has evolved from analyzing smaller text corpora to, with the bolster of higher memory and better central pro-

cessing units, allowing users to process vast document collections. Recent NLP advancements have introduced capabilities such as co-reference resolution, sentiment analysis, and topic modeling, essential for extracting meaningful patterns from large-scale datasets. Cutting-edge NLP technologies, including Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), spaCy (by Honnibal et al., 2013), Stanford CoreNLP (by Manning et al., 2014), Stanza (by Qi et al., 2020), and Gensim (by Rehurek and Sojka, 2011), offer potent tools for data analysis in educational research, as demonstrated by our application to the GradCafe dataset.

III. Results

I. Proposed Approaches

The data was crawled by Wget, a free GNU project that retrieves content from web servers alongside GNU Parallel to multithreaded crawl for up-to-date data.

Subsequently, the following GradCafe dataset was obtained: one that spans 19 years, from 2006 to 2024, encompassing a total of 847,725 entries. Each year's contribution to the dataset fluctuates significantly, beginning with 8,018 entries in 2006 and reaching a peak of 73,147 in 2021, before a decrease to 6,911 in the preliminary data for 2024. This variance highlights the evolving volume of data collected over time, reflecting potentially increased participation or changes in data collection methodologies. A figure of the submission by date is visualized below. We observe consistent submission since 2010 - with year 2022 and beyond possibly limited by the timing of data collection.

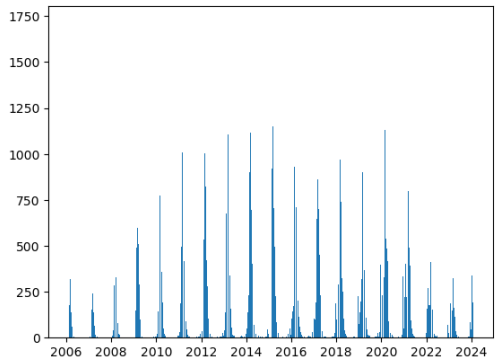

Fig. 1: A Description of the Data Collected of GradCafe User Submission

Within this dataset, the attributes reveal a mix of numeric and categorical data, pointing to a rich source of potential insights. Specifically, 'GRE_Stats' and 'GPA' are notable numeric entries, though they exhibit a high percentage of missing values—76.22% and 69.83%, respectively. This substantial absence of data suggests the need for careful consideration in data handling and preprocessing steps. Conversely, the dataset is predominantly categorical, with attributes such as 'Major_School', 'Degree', and 'Citizenship', among others. These categorical attributes provide a multifaceted view of the applicants' backgrounds, aspirations, and academic journeys.

II. Proposed Approaches

Data is merged using regular Python pandas. Data intermediates will be cleaned using KNIME. Data Transformations are conducted as follows for the categorical attributes.

1. School.

During transformation of the variations in school names, we employed the fuzzywuzzy library for fuzzy string matching, improving consistency across entries. We referenced a predefined list of schools to identify the most similar match for each entry, systematically replacing common abbreviations with their full forms before matching. A caching mechanism optimized this process, storing previously matched names to reduce computational redundancy.

2. Major

For the 'Major' field, standardization was essential due to the diverse ways fields of study can be referred to. We derived a reference list of STEM majors from the official 2023 STEM Designated Degree Program List published by the Department of Homeland Security. Utilizing fuzzywuzzy, we matched each entry to the closest corresponding major in the reference list, ensuring accurate and consistent representation of fields of study. This also allowed analyzing potential outcome differences between STEM and non-STEM majors.

3. Other non-quantifiable attributes

Other non-quantifiable text data were extracted based on the results of the frequent pattern mining as indicated below in the analysis part.

III. Experimental Results

Our group used Python, with standard scientific packages, including SciPy, Pandas, Numpy, and Matplotlib as background, and all other system-defined packages as basic foundations.

1. Natural Language Processing Analysis: A Sentiment and Frequent Itemset Mining

The dataset was initially analyzed using frequent pattern mining with the a-priori algorithm. To optimize memory usage, the 'low Memory' option in mlxtrend was activated, resulting in efficient computational processing. Afterward, a filtration process was applied to the 1-itemset nouns, resulting in a table of support values and frequent patterns consisting of single items. The analysis of itemsets containing two or more elements revealed a predominance of stopwords, which were deemed irrelevant for further analysis. Therefore, the investigation was confined to single itemset frequent pattern mining. The pattern was evidenced by the significant support values for terms such as 'research' (28,768), 'work' (9,175), and 'paper' (4,288), underscoring their relevance to academic research. Meanwhile, the dataset also shows considerable support for abstract terms such as 'luck' (19,394), 'hope' (11,224), and 'thanks' (8,702), highlighting the emotional dimensions that underpin students' perceptions of admission decisions. This warrants a closer examination of variations in sentiment later.

This dichotomy between concrete terms that are closely associated with the mechanics of academic research and

abstract terms reflecting the emotional landscape of student experiences highlights the importance of exploring how these focal points shift over time, which requires a temporal analysis. Using word clouds as a visual tool for this investigation provides an intuitive means to discern patterns and trends in the data, as shown in the following conclusions.

interview	61264	mail	8664
program	49001	post	8438
offer	41875	information	7892
application	32198	international	7648
letter	30136	review	6810
research	28768	news	6740
decision	24092	person	6700
school	23173	professor	6618
experience	22111	committee	6387
status	20564	director	5915
luck	19394	degree	5743
department	16625	weekend	5555
week	15873	response	5336
official	14761	fall	5333
university	13514	word	5308
march	12169	engineering	4914
call	11254	group	4903
hope	11224	process	4783
phone	11207	contact	4650
time	10974	author	4540
choice	10945	science	4531
people	10927	dream	4480
work	9175	number	4359
student	8781	track	4328
thanks	8702	paper	4288

Lexical Changes Over Time The analysis of the word clouds over time shows a shift in the vocabulary used in the context of graduate school applications. For instance, in the earlier years, words like ”PhD,” ”fellowship,” and ”status” were more prominent, while in later years, words like ”waitlist,” ”portal,” and ”GPA” became more common. This reflects changes in the concerns and priorities of applicants and admissions offices over time.

Trends and Patterns Certain words maintain a strong presence across multiple years, indicating persistent themes in graduate school applications. Words like ”know,” ”applied,” ”email,” and ”poster” appear consistently, suggesting they are key components of the application discourse.

Co-occurrence and Associations By analyzing which words frequently appear together, we can infer potential associations. We observe that ”email” and ”status” often co-occur, which might suggest that email is a common medium for sharing application status updates.

Sentiment Analysis Over Time As the sentiment-bearing words change over time (e.g., a shift from ”excited” to ”anxious”), it could indicate a change in the emotional tone of the applications. This thus paves way for future analysis on the sentiment change in the report.

2. Decision Tree Construction and Layout

In our study, we aimed to analyze the predictive importance of various academic and demographic features on the admission status of applicants. We employed a Decision Tree Classifier within a pipeline that included preprocessing steps for both numeric and categorical data.

Fig. 2: A Word Cloud of GradCafe data, 2008-2023

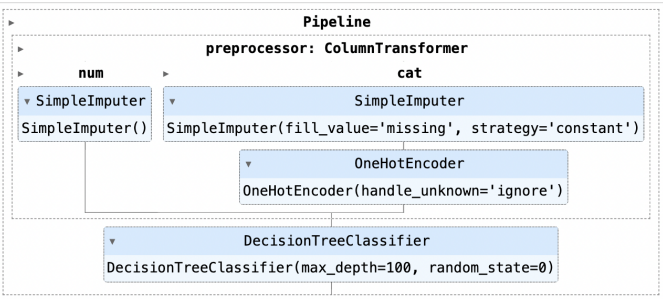

Fig. 3: Overview of decision tree

Initially, our raw dataset incorporated a blend of numeric scores (GRE Quantitative - `GRE_Q`, GRE Verbal - `GRE_V`, and GRE Analytical Writing - `GRE_AW`), GPA, as well as categorical attributes such as Degree, Citizenship, School, and Major.

To enhance the uniformity of the school names and to mitigate the discrepancies arising from different representations of the same institution, we applied a fuzzy matching normalization process (`School_Sim`). Additionally, a binary attribute, `Is_STEM_OPT_Eligible`, was included to indicate eligibility for STEM Optional Practical Training, a critical consideration for international applicants in STEM fields. Stepwise incorporation significantly ($p < 0.0001$, ANOVA test) improved precision upon integrating these features, as depicted in the result comparison. This suggests the predictive strength lies in a multifaceted evaluation of candidates, rather than a narrow focus on traditional scores such as GPA and standardized tests alone.

In a strategic modification to our feature set, we omitted the `Citizenship` and `GPA` attributes to assess their impact on decision-making. The accuracy of our model was then adversely affected, indicating that these factors may be very critical for admission decisions as previously presumed. This result thus prompts for further investigation on diachronic changes in the results of admission outcomes, which belies the so-called holisticness within the graduate school application.

Posteriorly, we were able to deploy the decision tree model on pythonanywhere.com, a free website that could host a

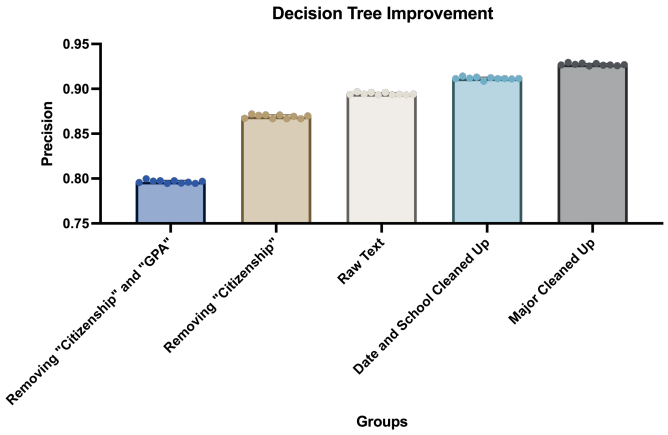

Fig. 4: A Decision Tree Precision measurement, varied attributes selected for, $p < 0.001$ for all groups compared using ANOVA

Python based web application, allowing for deepened functionality. We used Flask to create the web app, and the sklearn model was dumped as a re-usable model. A link to access the application is at: https://shuyangbian.pythonanywhere.com/. Users could enter their respective details, and upload Non-null attributes to the system for an immediate decision prediction.

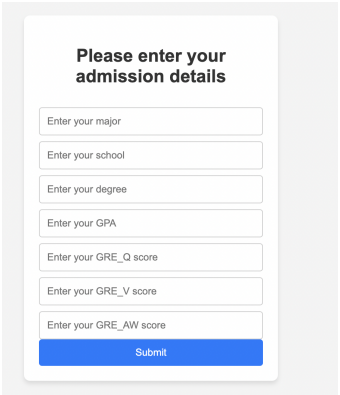

Fig. 5: A screenshot of the interface designed, deployed at pythonanywhere

3. Statistical Analysis: Diachronic View of GPA and international student applications

STEM OPT Eligible Fields

- **American Students**: There seems to be a stable or slight increase in the percentage of American students applying, being admitted, and enrolling over the years in STEM OPT eligible fields.
- **International Students**: The trends for International students in STEM OPT eligible fields show more volatility. There are significant fluctuations in applications, admissions, and enrollments, with a notable peak and subsequent decline in the ”Applied” category.

STEM OPT Non-eligible Fields

- **American Students**: The application and admission rates for American students in non-STEM fields appear relatively stable, but enrollment rates have a decreasing trend.
- **International Students**: For International students, there is a notable decline in the percentage of applications, admissions, and enrollments in non-STEM fields over the years.

Comparing American and International Students

- In STEM OPT eligible fields, International students have a higher percentage of ”Applied” but not necessarily a corresponding higher percentage in ”Admitted” and ”Enrolled” categories, especially in later years.
- In non-STEM fields, American students maintain a more stable presence across all three categories compared to International students, who show a decreasing trend.

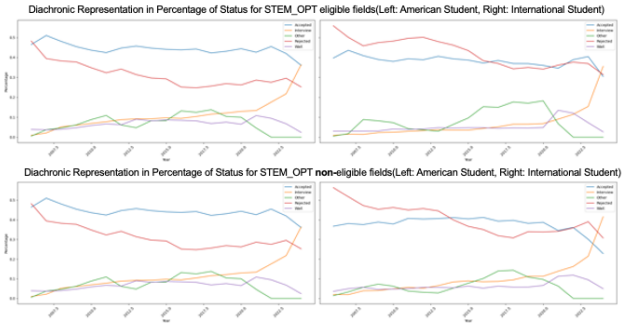

Fig. 6: STEM OPT Yearly Trends Graph.

4. A Visualization of Application Clustering

We further aim to explore the potential for spontaneous clustering of graduate school applicants based solely on their application data, without any supervised labeling or knowledge of admission outcomes. The goal was to investigate whether the actual admission process could be considered haphazard, with applicants' outcomes attributed largely to chance circumstances beyond the data captured in their applications.

Again, similar to section 3, numeric features were imputed with mean values and scaled, while categorical features were imputed with the most frequent values and one-hot encoded. To determine the optimal number of clusters, we employed the elbow method using K-means clustering on dimensionality-reduced data. The elbow plot suggested a flat drop, indicating no clear optimal number of clusters. Based on this result, it appears necessary that we designate a required number of cluster for further analysis. Since the admission decisions are mostly ”Accept”, ”Reject”, ”Waitlist” or ”Interview”, we proceeded by forcing the clustering into four groups for further analysis. Based on the clustering designated above, we construct a 3-dimension PCA component and visualize the cluster representation. Extracting information from each of the principal component suggests that the top ranking components

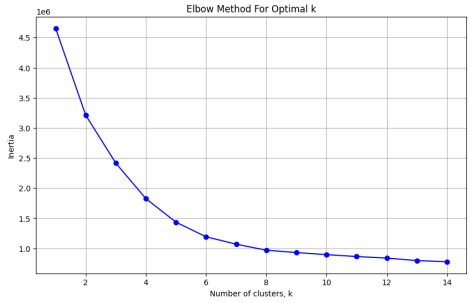

Fig. 7: Elbow plot over number of clusters.

would be as follows that drives the clustering:
1. GRE Verbal and Quantitative scores
2. Degree type (PhD vs. Masters) and citizenship status
3. GPA, GRE Analytical Writing score, with some influence from major (Computer Science)
Notably, while numeric attributes like GRE scores and GPA played a significant role, the applicant's major field of study also emerged as an important factor in the clustering. Since geographically, different schools have their respective strength program, it therefore prepares the foundation for further geographic-based investigation on the applicants' decision status over the years. These results would be briefly discussed in Section 5. We observe that the "purple" dots in the

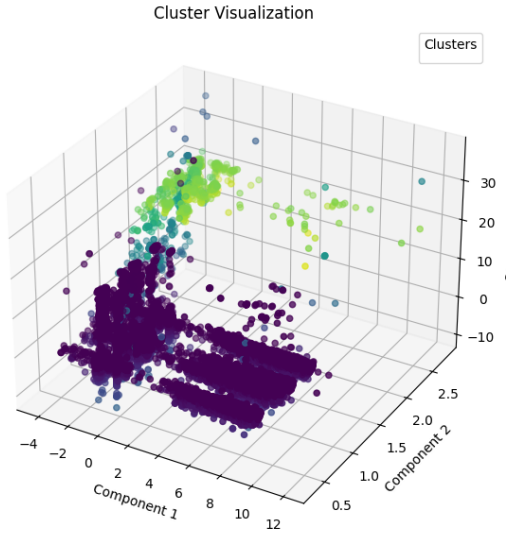

Fig. 8: Clustering of the $k = 4$ using the primary 3 Principle Component for Visualization

cluster visualization were occupying a significant proportion, but "blue" and "green" are intermingled together. In reality, however, people who are interviewed should have had at least cross overlap with people who are accepted. This intuition leads us to evaluate the clustering performance using a confusion matrix. The results indicated that most applicants were incorrectly classified, supporting the hypothesis that the actual admission process may be quite haphazard, with outcomes largely attributed to chance circumstances beyond the captured data. In summary, our unsupervised clustering analysis on

Fig. 9: Confusion Matrix of the $k = 4$

graduate school application data suggests that while certain numeric attributes and academic backgrounds influence the natural clustering of applicants, the admission process itself appears to be highly unpredictable based solely on the available data. This finding reinforces the notion that numerous unmeasured factors may contribute to the seemingly arbitrary nature of admission decisions. Therefore, a more descriptive analysis on the outcomes of post-application statuses of these posted data entries were investigated.

5. Mining Spatial Data: A Geographic-based approach

Recent studies have shown that geographic mining would be effective if segmentation is used upon data. Segmentation of fine-labeled data to more coarse regions could sometimes yield better results for data visualization purposes, allowing for the production of insights and therefore, engagement of readers with data. To this end, the varied institutions reported by the GradCafe users were mapped to a most likely, normalized institution using the data cleaning pipeline as introduced in the methodology, thereby converting reported institutions to one of the two hundred possible ones as obtained from US news report, since GradCafe is mostly limited to U.S. higher education institutions.
To achieve the effects of segmentation of the varied regions, Tableau was used such that different institutions' latitude, longitude, city and state could be better visualized. An overall plot for the acceptance rate and rejection rate was subsequently visualized, with each circle size representing the relevant degree of rejection rate for each city, and the degree of competitiveness represented by the hue of each state's filling. As represented in the following two figures below (Fig 8, Fig 9), when the overall rates of acceptance rate are compared, there exists no palpable difference across the regions. Nevertheless, within the nation, there exists a significant skew towards the east coast when rejection rate is compared, suggesting a potential high interest in the region, and a uniform standard of academics across the nation. Our subsequent step is to visually compare the overall acceptance and rejection rate by

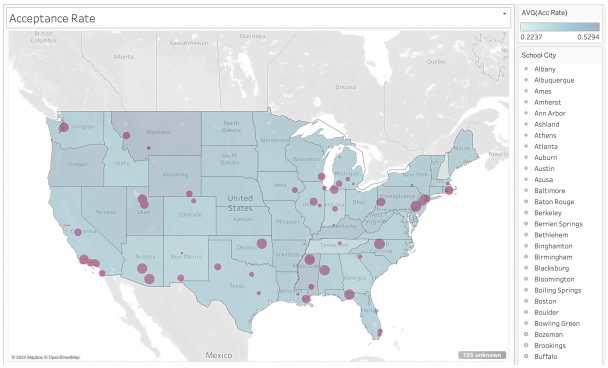

Fig. 10: Acceptance Rate of GradCafe Results by Geographic Visualization

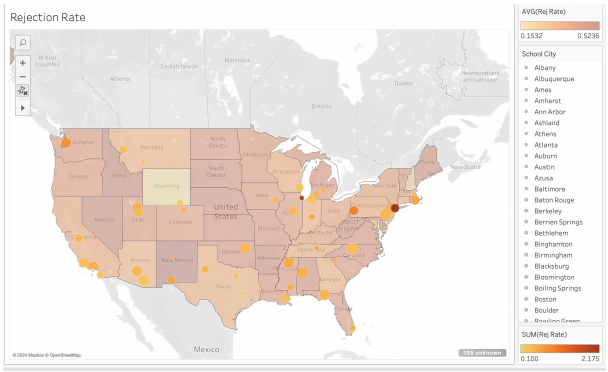

Fig. 11: Rejection Rate of GradCafe Results by Geographic Visualization

using the mapping method as described in section 3, to further visualize the effects of nationality and STEM majors' impact.

Region	Status Citizenship	Accepted	Interview	Other	Rejected	WaitList
Midwest	American	36.90	11.35	5.90	40.54	5.30
	International	32.28	8.69	7.93	43.85	7.25
Northeast	American	37.69	10.47	7.08	37.57	7.19
	International	32.27	6.66	6.96	47.64	6.47
West	American	100.00	0.00	0.00	0.00	0.00

Division	Status Citizenship	Accepted	Interview	Other	Rejected	WaitList
East North Central	American	36.90	11.35	5.90	40.54	5.30
	International	32.28	8.69	7.93	43.85	7.25
Middle Atlantic	American	24.93	12.02	2.23	55.19	5.64
	International	23.16	10.99	2.34	56.37	7.13
Mountain	American	100.00	0.00	0.00	0.00	0.00
New England	American	38.10	10.42	7.23	37.01	7.23
	International	32.69	6.46	7.18	47.22	6.44

In the above section, chi-square contingency tests were conducted, and a less than 0.001 outcome was observed.

6. Discussion

Educational data mining, the use of data exploration on student outcomes, has been widely studied. For example, Essa and Abdullah (2023) analyzed more than 270000 student records for graduation outcomes. Using feature extraction and dimensionality vector features on the tSNE algorithm, student achievement was then predicted using a machine learning approach. The dataset size of our study was comparable to that of this previous study, and several salient features were also identified. While our study suffers from a lack of precise data points for several metrics used in Essa's study (including gender and specific school the students come from), our study achieves similar results in predicting using prior scores and using a sophisticated model to estimate the score.

Another feature of the current work involves the comparison of the use of the decision tree algorithm to analyze the salient features of the dataset. Using a stepwise improvement in refining the categories of the number of features, our experiments demonstrate how a refined mapping of the categories could improve the outcome of decision making in the graduation process, thereby facilitating the school choice decision. While our data had very different characteristics from that of the literature review, a similar use of methodology (from decision tree to clustering) has allowed our work to be comparable to other work available at large, facilitating future benchmarking efforts.

Furthermore, the study has been innovative in the data mining approach in constructing a human-computer interaction (HCI) interface, allowing users and readers to more effectively interact with the decision process. This incorporation of an interactive component adds to the functionality of the program, allowing for the potential use for prospective students in making informed choices for their graduate school applications. Nevertheless, contrasting to popular school decision tools for US undergraduate admission process available such as niche.com, the current application may need further improvement in its GUI design and interface outlook.

7. Conclusion and Future Directions

This study has made significant advancements in the data mining approaches for analyzing graduate school applications. By fine mapping categories to more coarse-level data, we significantly improved the accuracy of predicting admission outcomes, underscoring the utility of refined data categories in educational data analysis. Additionally, the use of single-factor analysis emerged as a crucial tool in extrapolating deep insights, even when faced with data challenges such as null values and self-reporting biases.

However, several limitations need to be addressed in future studies:

- **Data Quality and Completeness:** The data set, primarily user-generated, contains null values and lacks sufficient numeric attributes, potentially hindering comprehensive analysis. Efforts to improve data collection and preprocessing will be crucial for enhancing the robustness of future analyses.
- **Comparative Analysis:** The nascent stage of related research and a lack of benchmarks for state-of-the-art comparisons underscore the need for continued research that could contribute to a more established benchmark for the field.

Moving forward, we propose:

- **Advanced Predictive Models:** Implementing an explain-

able RNN or GPT model to refine our predictive capabilities regarding student acceptance likelihood. These models can leverage large language capabilities to understand and predict complex patterns, offering a nuanced approach to decision support systems in education.

- **Democratizing Information:** There is an urgent need to democratize knowledge about graduate school admissions to level the playing field for all applicants. By developing tools that provide transparent and accessible insights into the admissions process, we can help demystify the elements that influence admission decisions.

As we look towards the future of graduate admissions, and the potential to leverage data mining, there is significant promise in the application of large language models (LLMs) to enhance decision making processes and support systems for prospective students. LLMs such as OpenAI's GPT have shown exceptional capability in generating human-like text and understanding complex patterns in data (Tu, Zou, Su, & Zhang, 2024). By integrating LLMs as such into the college admissions landscape, we can develop sophisticated tools that assist students in crafting applications, preparing for interviews, and make decisions informed by data about their education paths based on predictive analytics.

Furthermore, the market for such tools is expansive and growing. As the landscape of higher education becomes increasingly competitive and complex, prospective students approach educational consultants to receive a competitive edge (2019). The educational consultant industry's industry wide revenue is $2.9 billion in 2023 with a projected growth of 0.2% (2024). There is a substantial opportunity to develop such tools to help educational consultants, university admissions departments, and directly to students and their families.

Moreover, the deployment of LLMs in this context addresses an urgent need to democratize access to information and resources related to college admissions. Historically, knowledge about the admission process and the factors that significantly impact acceptance likelihood has been tacit and unevenly distributed, often benefiting those with greater resources or insider knowledge (Kromydas, 2017). By creating an data-driven platform that analyzes extensive admissions data and offers personalized recommendations, we aim to level the playing field, allowing students from diverse backgrounds to access high-quality guidance and insights. The development of AI tools for graduate admissions is poised not only to enhance the process's efficiency and effectiveness but also to transform the market by democratizing access to essential information and resources.

REFERENCES

[1] Ferreira, Orvell. "College Recommendation System: Comparison Of Different Machine Learning Models Orvell Ferreira, Clint Ferreira, Sloan Dcunha And Prof. Parshvi Shah." Stochastic Modeling.

[2] Liu, Chen, Zehao Su, and Jiayao Zhang. "Understanding Admission Results of CS Graduate Programs in US Universities."

[3] Gupta, Narender. "American graduate admissions: both sides of the table." (2016).

[4] Sylvan, Lesley, Andrea Perkins, and Carly Truglio. "Student experience applying to graduate school for speech-language pathology." Perspectives of the ASHA Special Interest Groups 5.1 (2020): 192-205.

[5] McKinsey & Company. (2020). COVID-19 and US higher education enrollment: Preparing leaders for the fall. Retrieved from https://www.mckinsey.com/industries/education/our-insights/covid-19-and-us-higher-education-enrollment-preparing-leaders-for-fall

[6] Yong, L., Sachau, D., & Lassiter, A. L. (2011). Developing a measure of virtual community citizenship behavior. Knowledge management & e-learning: an international journal, 3(4), 682.

[7] Yong, Luman. Exploring the Antecedents of Organizational Citizenship Behavior in Knowledge-based Virtual Communities. Minnesota State University, Mankato, 2011.

[8] Inside Higher Ed. (2020). Graduate enrollment grew in 2020 despite pandemic. Retrieved from https://www.insidehighered.com/news/2020/11/12/graduate-enrollment-grew-2020-despite-pandemic

[9] Inside Higher Ed. (2020). Final fall enrollment numbers show pandemic's full impact. Retrieved from https://www.insidehighered.com/news/2020/12/17/final-fall-enrollment-numbers-show-pandemics-full-impact

[10] Council of Graduate Schools (CGS). (2020). The Impact of COVID-19 on Graduate Education. Retrieved from https://cgsnet.org/impact-covid-19-graduate-education

[11] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018, arXiv:1810.04805

[12] Honnibal, M., Goldberg, Y, and Johnson M. (2013). A non-monotonic arc-eager transition system for dependency parsing. In Proceedings of CoNLL.

[13] Manning C, Surdeanu M, Bauer J, et al. (2014) The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Stroudsburg, PA, USA, 2014, pp. 55–60. Association for Computational Linguistics. DOI: 10.3115/v1/P14-5010.

[14] Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." arXiv:2003.07082v2 [cs.CL] 23 Apr 2020.

[15] Rehurek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

[16] Tu, Xinming; Zou, James; Su, Weijie; Zhang, Linjun. "What Should Data Science Education Do With Large Language Models?" Harvard Data Science Review, vol. 6, no. 1, Jan. 19, 2024. Available: https://hdsr.mitpress.mit.edu/pub/pqiufdew

[17] "Education consultants in the US - market size, industry analysis, trends and forecasts (2024-2029): IBISWorld," IBISWorld Industry Reports. [Online]. Available: https://www.ibisworld.com/united-states/market-research-reports/education-consultants-industry/.[Accessed: Apr.21,2024].

[18] https://www.tandfonline.com/doi/pdf/10.1080/12460125.2022.2151071

Appendix: Task assignment for each member and Time schedule

Sprint #	Date	Item	Author
1	2024-02-10	Gathering Data	Simon, Nazif
2	2024-02-15	GPA, GRE, program	Tuan
3	2024-02-20	Program Popularity	Nazif, Simon
4	2024-02-25	Time Analysis	Tuan
5	2024-03-01	Predictive Model	Tuan
6	2024-03-06	Sentiment Analysis	Simon
7	2024-03-11	Topic Modeling	Simon
8	2024-03-16	Scholarship	Nazif
9	2024-03-21	Geographical	Simon
10	2024-03-26	Recommendation Sys	Tuan
11	2024-04-07	Clustering	Simon, Nazif