SmartAdmit - Admission Auto-Decision-Maker

Siyi Liu
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
sliu0228@vt.edu

Wangzhi Zhan
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
wzhan24@vt.edu

Hun Liang
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
hunliang@vt.edu

Divya Polavarapu
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
DivyaP24@vt.edu

Abstract

Admission to higher education institutions is an important milestone for students and poses a complex challenge for both applicants and admissions committees. For applicants, it involves navigating various requirements, presenting their qualifications effectively, and standing out among a competitive pool. For admissions committees, it requires evaluating a large volume of applications fairly and consistently while identifying candidates who best align with institutional goals and values. This project seeks to address these challenges by analyzing the application process to determine the most impactful factors that influence admissions decisions. By identifying these key components, the project explores the development of an automated decision-making system to streamline the admissions process. By leveraging data-driven techniques and machine learning algorithms, the system can evaluate applications efficiently and objectively, reducing the workload of admissions committees and ensuring consistent decision-making.

Keywords

College Admissions, College, Auto-decision Grader, MLP, Clustering, Word clouds

ACM Reference Format:
Siyi Liu, Wangzhi Zhan, Hun Liang, and Divya Polavarapu. 2018. SmartAdmit - Admission Auto-Decision-Maker. In . ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 1 Introduction

1.1 1.1 Problem Description

Introduction to Problem and Data. Admission is the first step towards successful higher education, which is a challenge for both the applicants and the admissions committee. One task is to find the most important parts of the application process and to provide guidance for applicants. Another task is to build an automatic decision-maker, to make the decision process easier. Our dataset is the college admission data of a large, public U.S. university. The input comprises 3 types of features, numeric, categorical, and textual features. The labels include raters' scores and the final decision of admission, rejection, or wait-listed.

Admission Auto-Decision-Maker. Our first task is to build an automatic decision-maker (ADM). The most challenging aspect of this task is that the input contains several different modalities, so we should consider how to combine them effectively. The ADM should be able to output both raters' scores and the final decision.

Machine Learning-based Data Analysis. Besides the ADM, we also want to use machine learning techniques to analyze the data, so that we can provide insights of the admission process. Specifically, we will conduct a feature importance study to find important features from the input; finally, we will study the clustering of applicants, to find whether there are favored clusters and their corresponding attributes.

1.2 Research Gap

Previous works for university admissions decision-making can be roughly grouped into two categories: statistical analysis and numeric models. Works of the first category, like [3] and [5], designed various metrics to analyze the input data. The advantage of such methods is that there is a high probability that highly accurate conclusions can be reached. However, such methods need a large amount of human engineering and therefore rely on the expertise of statistics. Works of the second category, like [13], utilize simple machine learning techniques like linear regression to analyze the data. The advantage of such methods is generalization, which means the same method can be used to process various datasets, without the need to redesign a new model, although the hyperparameters may need to be tuned again. However, such methods can only make use of simple numeric features, and cannot process other types of features like textual features or categorical features.

Our dataset, however, includes numeric features, textual features, and categorical features. To make full use of all the provided features, a more feature-compatible model is needed.

1.3 Intuition behind the Current Technique

As previously mentioned, the primary obstacle to a comprehensive understanding of the university application decision making data is that the feature types are complicated. Therefore, we need to design different ways to pre-process different types of features. This operation will lead to clean and dense embeddings that can be input

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

into a downstream machine learning model which will output the final predicted decision.

2 Current Technique

2.1 Related Works

There have been previous works on projects that are relevant to our own. The first is a comparative look at the accuracy of different regression models on graduate admission data sourced from Kaggle. The dataset covered parameters such as test scores like the GRE and TOEFL, statements of purpose, letters of recommendation, undergrad GPA, etc. This is similar to some of the parameters covered by our dataset, namely, the statement of purpose, test scores like SAT, and high school GPA. A noticeable difference is that our dataset is covering undergraduate admission data not graduate admission data like this paper. They covered four regression algorithms; Multiple regression, Polynomial regression, decision trees, and random forest. The data was divided into an 85:15 split of training and testing data and it was found that multiple regression had the highest accuracy of an 80.48% chance of a student's admission into a university [15]. While interesting we ultimately did not end up using any of the covered regression methods in this project but it could be used for future works to see what insights are gained from utilizing regression.

The next paper covers grading text-based questions in courses with a large number of students using AI and NLP. While highly promising for automating such tasks, further refinement is required to match the reliability and accuracy of human graders. Especially when it comes to nuanced student responses [4]. This paper was particularly relevant to our work since we were trying to figure out how an auto grader can work with admission data and data such as open-ended statements of purpose.

For feature importance and identification, we explored two primary sensitivity-based methods: LIME [16] and SHAP [12]. Among these, SHAP stands out due to its ability to provide both global and local explanations, whereas LIME is restricted to local explanations. Additionally, SHAP has the potential to detect nonlinear associations depending on the underlying model, an area where LIME falls short because it relies on a local linear approximation [17]. Despite the high-quality explanations offered by SHAP values, their precise computation is efficiently feasible only for decision tree-based models through the Tree SHAP algorithm[11]. However, the computational expense associated with using conditional expectations in the algorithm remains a challenge[20]. To address this limitation, several studies [14] and [20] have combined the Tree SHAP algorithm with surrogate models, achieving a balance of high accuracy and interpretability. Given the high dimensionality and complexity of our data, Using either method would have resulted in substantial computational demands and could have generated explanations that were less interpretable or insufficiently reflective of the model's true behavior.

Handling mixed data types, such as essay responses represented as vectorized textual data, rater ratings, test scores, and extracurricular activities, which have both numerical and categorical variables, poses a notable challenge in data analysis. Two widely adopted techniques for clustering mixed data are variations of the K-means clustering algorithm, such as K-prototypes [7], and FAMD (Factor Analysis of Mixed Data) [1]. While K-prototypes are adept at managing mixed data by integrating numerical and categorical variables, their performance can degrade in high-dimensional datasets with numerous categories, resulting in elevated computational complexity[9]. We ended up using PCA and K-means clustering instead of K-prototypes and FAMD.

2.2 Dataset

Our project uses a real-world undergraduate admission dataset from a large, public American university, with the permission of data usage for research purposes. The final decision labels being classified include being admitted, being waitlisted, and being denied. In order to rule out the class imbalance issue, we randomly chose 4000 samples from each class and built a balanced dataset (N=12,000) that covers these three classes. Specifically, there are three types of features in the dataset, described in detail below:

The first type of features are numerical features, including:

- `sat-v`: SAT Verbal (or Evidence-Based Reading and Writing) score used for admissions.
- `sat-m`: SAT Math score used for admissions.
- `sat-comp`: SAT Composite score.
- `act-eng`: ACT English score.
- `act-math`: ACT Math score.
- `act-read`: ACT Reading score.
- `act-sci`: ACT Science score.
- `act-comp`: ACT Composite score.
- `nc1`: Additional criterion 1.
- `nc2`: Additional criterion 2.
- `nc3`: Additional criterion 3.
- `ncav`: Additional criterion averaged.
- `hs-gpa`: High school GPA.
- `college-gpa`: College GPA before the student transfers to the university.

The second type of features are text features, including:

- `eq1`: Answers to the first essay question about group/community experiences.
- `eq2`: Answers to the second essay question about discrimination and equity.
- `eq3`: Answers to the third essay question about leadership.
- `eq4`: Answers to the fourth essay question about goal-setting.

The third type of feature is a categorical feature, including:

- `uni-coll`: The specific college in the university that the students apply for.

2.3 Auto-Decision-Maker

Feature Pre-processing. The dataset provides three different types of features: numerical features, textual features, and categorical features. For the numerical features, we linearly normalize them on each dimension. For the textual features (i.e., answers to each essay question), we use the large language model designed in [19] to extract a dense embedding out of each textual paragraph. For the categorical features, we first counted how many categories are

there for each feature, and then transferred each feature into one-hot embeddings. After separately pre-processing all three types of features, we concatenate them together to obtain a comprehensive initial representation of each data entry. For labels, since the task for the current work is multi-class classification, we transferred the labels into a one-hot format. Figure 1 illustrates the aforementioned pre-processing pipeline.

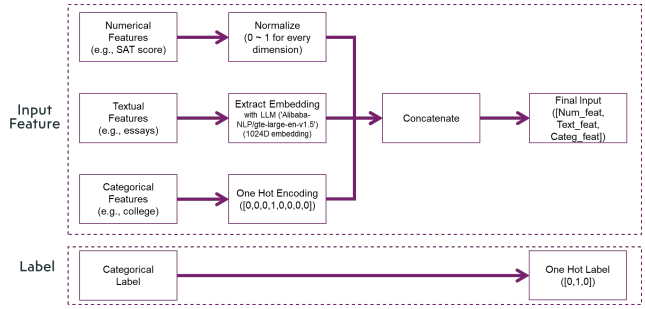

Figure 1: Feature Pre-processing Pipeline

Backbone Model. After the preprocessing step, each data entry is reformatted into a numerical vector, and all the input data is composed of a tensor. We then use a multi-layer perceptron to process the input tensor. The overall architecture of the backbone model is illustrated in Figure 2. For the first two layers, the input of each layer undergoes a linear projection ensued by a ReLU non-linear activation. For the final layer, the input undergoes a linear projection and a softmax activation. Finally, after an argmax operation, a prediction for the class label can be obtained. Moreover, the output of the first two layers is saved as representations, which can be used in the later data clustering stage.

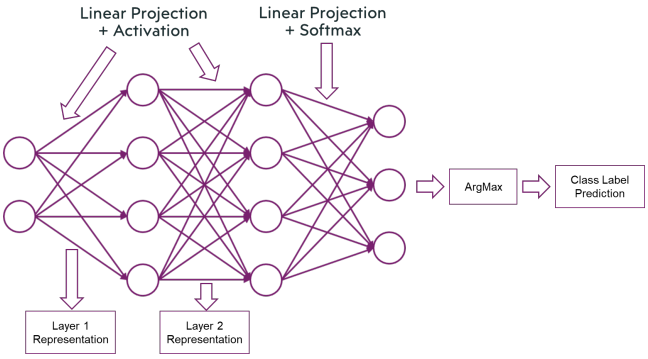

Figure 2: Backbone Model

2.4 Clustering

To investigate whether the latent space representations learned by the Multilayer Perception model contain any meaningful structure, we applied unsupervised clustering to analyze the grouping patterns of applicants based on their input features. The goal was to identify distinct clusters that align with performance-based groups (e.g., admitted, rejected, and waitlisted applicants) and to validate the effectiveness of MLP in transforming input features into more structured representations.

Our clustering methodology consists of the following steps. First, we passed the input features, including standardized test scores (SAT/ACT), GPA (high school, or college if the applicant was transferring), and essay scores, through the MLP model to obtain intermediate latent representations. Specifically, we extracted the Layer 1 representation (output of the first hidden layer) and the Layer 2 representation (output of the second hidden layer) for further analysis. These presentations reflect the progressively abstracted feature interactions learned by the network. Second, we applied K-means clustering to group the data points in the latent spaces. K-Means minimizes intra-cluster variance by iteratively assigning points to clusters based on their proximity to cluster centroids. For consistency, we set the number of clusters $k = 3$, corresponding to the three primary admission outcomes: admitted, rejected, and waitlisted. Third, to facilitate visualization and interpretability of the clustering results, we performed Principal Component Analysis (PCA) to reduce the dimensionality of the latent space representations. PCA projects the data onto two principal components while retaining the maximum variance, allowing inspection of the cluster structure in a two-dimensional space. This enabled us to visually compare the clustering behavior of the layer 1 and layer 2 representations.

2.5 Explainability

We decided to focus on the feature importance of essay answers, given that our input features covering three different types are very complex. To interpret which words/phrases in each essay contribute the most to the applicant's being admitted in the final decision, we output a word cloud for answers to each essay question. For each essay question, the methods we used are as follows.

First, we only input embeddings representing each essay into the MLP model, and output the classification result (i.e., prediction on the final decisions) for each essay. We only considered the text features without any other numerical and categorical features, since we wanted to rule out the influences of these variables on the classification. Second, we used a Spacy model `en_core_web_trf` [6] to lemmatize each essay, using an nltk package [2] to remove the stop words, deleted non-alphabetic characters, and converted the lemmatized essay into lowercase. Third, we used a TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer [18] to generate the unigram and bigram features for each essay. We set `ngram_range` as (1,2) since we only focused on individual words and two-word phrases. We also set `min_df` as 15 and `max_df` as 0.8, given that we aimed to focus on those unigram and bigram features included in more than 15 essays and discard those features included in more than eighty percent of essays. Fourth, we standardized all the unigram and bigram features since the vectorizer represented each of them as vectors consisting of raw frequency counts of single words used in the unigram or bigram. We input all these standardized feature vectors into a Ridge classifier, with the classification results output in the first step as the labels. We then retrieved the model's learned coefficients for each feature and

class and calculated the absolute coefficient values for the final interpretability. Last, we selected the top 30 features with the highest absolute coefficients and used them to generate a word cloud for each essay question.

2.6 Comparison to the existing techniques

As mentioned in section 1.2, previous methods on university admission decision-making generally rely on human-engineered statistical metrics or use simple machine learning models to process numerical features. Our method, however, uses the feature preprocessing pipeline depicted in section 2.3 to pre-process the three different types of input features. Therefore, with a downstream generic purpose model (MLP), our method can make use of all the input features and provide the final prediction of admission decisions.

In regard to clustering, traditional clustering methods such as K-Means [10] or hierarchical clustering [8] are often applied directly to raw features like GPA or test scores. While these techniques can group data points, they fail to fully capture non-linear interactions or relationships among high-dimensional features which are critical for datasets like the one we used. Our approach addresses these limitations by leveraging the MLP-based representation learning pipeline, which aligns with recent advances in representation learning methods where neural networks enhance feature transformation for downstream tasks [**?**]. Additionally, our use of PCA visualization allows for intuitive validation of clustering quality against admission outcomes or intended college labels offering a more robust and interpretable solution for analyzing admission data.

The existing feature importance techniques that we will compare in the next section are SHAP[12] and LIME[16]. SHAP[12] and LIME[16] are two popular model-agnostic methods for explaining feature importance in machine learning models. Specifically, SHAP[12] assigns contributions to features by considering all possible combinations of feature subsets, ensuring fair attribution based on Shapley values. SHAP[12] offers both local explanations for individual predictions and global importance summaries, making it suitable for complex models.

LIME[16] generates local explanations by fitting simple surrogate models, such as linear regressions, around specific predictions. This method works by sampling data points near the instance of interest and learning a simplified model to explain the prediction. However, it focuses solely on local interpretability, can be unstable due to randomness in sampling, and its explanations depend heavily on the choice of the surrogate model.

3 Evaluation

3.1 Results of Auto-Decision-Maker

Experiment Setting. We used an MLP as illustrated in Figure 2, with two hidden layers. The latent dimension is 64. We trained the model for 20 epochs, with the learning rate being 1E-3.

Classification Results. The final classification results are listed in Table 1. The overall classification accuracy is 0.58, while the F1 score is 0.57. The first two classes, admission and denial have much higher accuracy than the third class, waitlist. This is because the wait-listed samples are located somewhere near the border between admission and denial. Therefore, the decision boundary of waitlist class is much more blurred than the first two classes, which explains why the third class is hard to distinguish.

Table 1: Classification Results

Metric	Result
Class 1 (Admission) Accuracy	0.72
Class 2 (Denial) Accuracy	0.65
Class 3 (Waitlist) Accuracy	0.36
Overall Accuracy	0.58
F1 Score	0.57

3.2 Results of Clustering

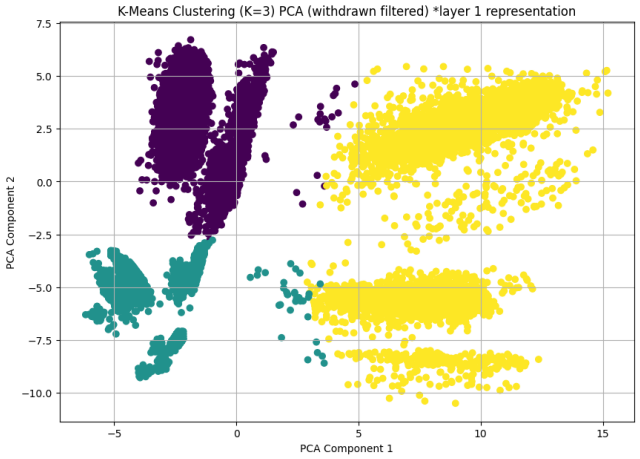

Figure 3: K-Means + PCA with Admission Results (Layer 1 Representation)

Figure 3 and 4 analyze the structure of the latent space representations generated by the MLP model where we applied K-Means clustering and PCA visualization. In these figures, data points are color-coded on the basis of their K-Means cluster assignments. In the layer 1 presentation, the K-Means algorithm partitions the data into three primary clusters where it displays noticeable overlap and inconsistency in their boundaries. While some regions show compact groupings, a significant number of points from different clusters appear intermixed, indicating that the features at this stage do not yet fully separate performance-based patterns. This observation suggests that the layer 1 representation captures only basic feature interactions, leading to partially formed clusters with limited structure. In contrast, the layer 2 representation shows a marked improvement in cluster separation and definition. The K-Means clusters are more cohesive, with clearer boundaries and less intermixing compared to layer 1. The points within each cluster are more tightly grouped, reflecting a higher degree of homogeneity in the learned features. This improvement shows the MLP's ability

to transform input features into a more structured latent space as data progresses through deeper layers.

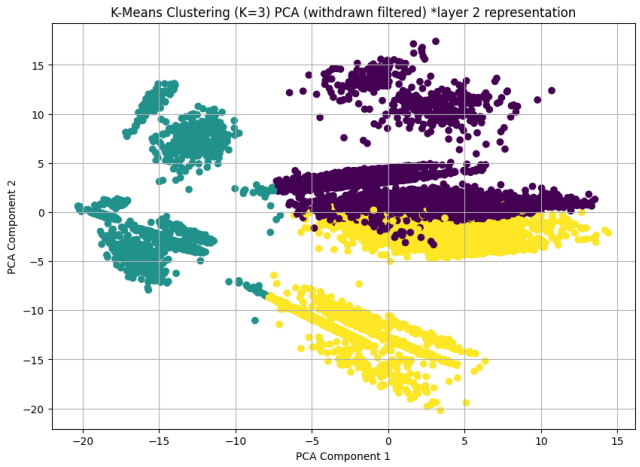

Figure 4: K-Means + PCA with Admission Results (Layer 2 Representation)

Further analysis was conducted to explore the clustering of admitted applicants based on their intended colleges as shown in Figure 4. By visualizing college-specific groups in the layer 1 latent space, distinct patterns emerged for different colleges such as Engineering, Liberal Arts, and Agriculture. Applicants from the College of Engineering, for example, formed more compact and well-defined clusters, reflecting the homogeneity in their performance profiles. In contrast, candidates for liberal arts exhibited greater dispersion, indicating higher variability in their scores and features. It underscores the model's capacity to capture nuanced differences between college-specific applicant groups.

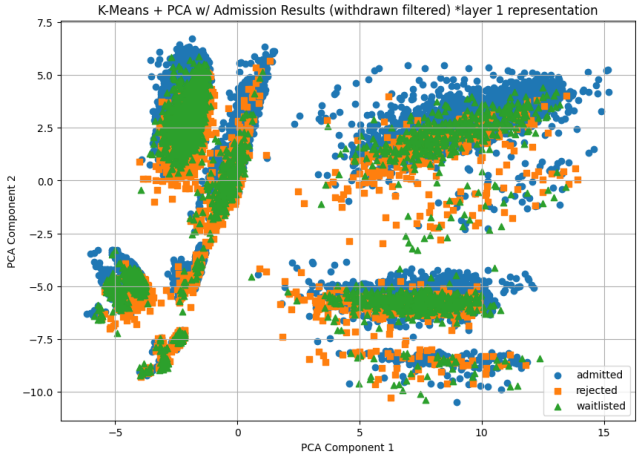

Figure 5: K-Means + PCA with College Labels (Layer 1 Representation)

3.3 Results of Explainability

Figure 5,6,7 and 8 display the feature importance of the answers to each essay question, respectively. The existing feature importance methods such as SHAP[12] and LIME[16] did not work for our textual features. The corresponding reasons are below. First, we used a transformer model to convert all essay texts into embeddings, while both SHAP[12] and LIME[16] cannot directly analyze how embeddings influence the classification results through the MLP model. Secondly, LIME[16] can only generate local explainability, while our goal was to output a global summary of a word or a phrase's importance. Hence, we chose to use our methods described in 2.5 for feature importance analyses, rather than use SHAP[12] or LIME[16].

In figure 1, some keywords include honors society, special Olympics, national charity, charity league, and farmers. Writing about involvement in these organizations makes for strong undergraduate admission essays about group and community experiences because these experiences demonstrate meaningful engagement, leadership, and service. These groups often involve collaboration toward a shared purpose, showcasing a student's ability to work with others while making a tangible impact. Through such participation, students can reflect on personal growth, the development of key skills such as leadership and empathy, and lessons learned from supporting diverse communities. Additionally, these experiences often align with the values universities seek, such as civic responsibility and a commitment to service.

Figure 6: The top 30 textual features in responses to the first essay question that contribute the most to admission decisions.

Similarly, in figure 2, some important topics include white supremacist, self esteem, sexual profiling, police brutality, speak louder, United

States. These are compelling subjects for undergraduate admission essays about discrimination and equity because they reflect a deep awareness of critical societal issues. Writing about these topics allows students to demonstrate intellectual engagement, social consciousness, and a commitment to fostering justice and equality. Universities value applicants who are not only aware of societal challenges but also willing to engage in meaningful dialogue and action. By connecting personal stories or perspectives to broader societal contexts, students can show their potential as thoughtful, change-oriented members of a diverse academic community.

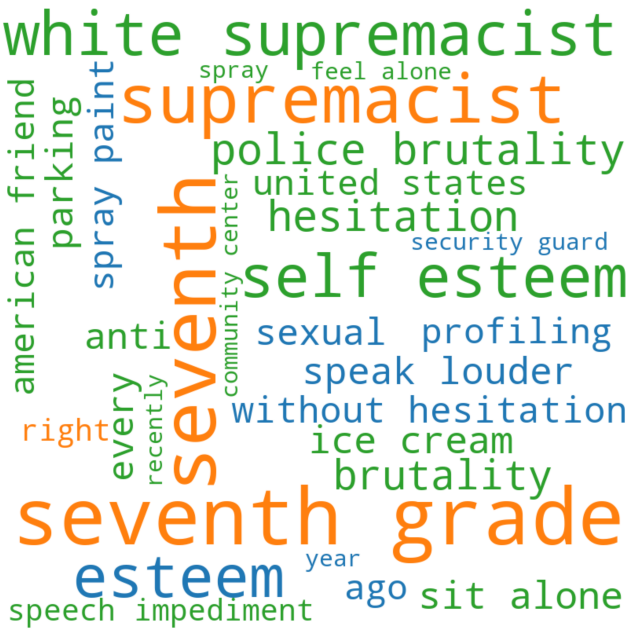

Figure 7: The top 30 textual features in responses to the second essay question that contribute the most to admission decisions.

In figure 3, some important topics are scout ranch, pep rally, philmont scout, sporting event, business leaders, non-profit, self esteem, and work tirelessly. These topics are a great fit for undergraduate admission essays about leadership because they illustrate diverse experiences where leadership qualities can be developed and demonstrated. These settings provide real-world examples of taking initiative, motivating others, and overcoming challenging aspects of leadership. Whether organizing a pep rally, leading a team at a sporting event, or guiding peers during a challenging Philmont Scout trek, students can showcase resilience, teamwork, and decision-making under pressure. Experiences with business leaders or non-profit organizations highlight strategic thinking and community impact. Building self-esteem or learning to work tirelessly toward a goal, underscores perseverance and dedication. Altogether, these stories reveal a well-rounded leader ready to contribute meaningfully to a university environment.

Lastly, comfort zones, cadets, marine corps, marching, trial error, and study abroad are some topics in the essay about goal-setting that have the highest contribution to undergraduate admission.

Figure 8: The top 30 textual features in responses to the third essay question that contribute the most to admission decisions.

These topics involve setting ambitious objectives, pursuing personal growth, and overcoming challenges. Stepping out of one's comfort zone reflects a willingness to embrace new experiences and push beyond limits, as an essential trait for achieving goals. Participation in cadet or Marine Corps programs highlights discipline, commitment, and resilience developed through structured training and leadership roles. Marching in a band or military unit requires teamwork, precision, and practice toward a shared goal. Trial and error illustrate perseverance, adaptability, and the learning process essential for success. Studying abroad involves setting academic and cultural goals while navigating unfamiliar environments, and fostering independence. These experiences demonstrate how students set, pursue, and achieve meaningful goals through hard work and adaptability.

4 Conclusion

We were able to develop and implement several methods that provided valuable insights into the decision-making process for college admissions. By employing clustering techniques, we grouped applicants based on their diverse features, including academic performance, essay responses, and test scores. This approach not only allowed us to identify patterns and similarities between applicants but also helped us uncover hidden relationships that may influence admission decisions. The clustering results served as a foundation for understanding how different attributes interact and impact admission outcomes, thus offering a more nuanced view of the decision-making criteria.

This work lays a solid groundwork for future endeavors in automating and improving the college admission process, with the

Figure 9: The top 30 textual features in responses to the fourth essay question that contribute the most to admission decisions.

potential for developing more equitable, transparent, and efficient systems for evaluating applicants.

4.1 Limitations and Challenges

Some limitations we ran into comprised the fact that our dataset was confidential so we were unable to work with the dataset. This wasn't a significant issue since we were able to conduct all our work on subsets of fabricated data and then apply it to our actual dataset to glean results.

Additionally, it's important to note that the dataset currently being used is limited to a single year's worth of admission data. This means it may not fully capture the variability and trends that could emerge across multiple years or represent broader historical patterns in college admissions. Additionally, the criteria for college admissions are dynamic and can change over time, as seen with the shift to test-optional policies following the COVID-19 pandemic. As such, it would be critical for future work to consider these historical shifts and adjust the model accordingly. This could involve continuously updating the model to reflect new trends in admissions criteria, ensuring that predictions remain accurate and relevant.

One of the key challenges was analyzing global feature importance across all variables, given their inherent complexity. The diversity of variable types and their intricate interactions made this task particularly difficult. As a result, we chose to narrow our focus to textual variables, which allowed for a more manageable and targeted analysis within the scope of the project.

4.2 Future Works

There are several avenues for further development that could significantly enhance the work we've done thus far. One key direction would be to create a more robust and accurate model by integrating additional features that are typically part of the college admissions process but were not included in the current analysis. For instance, incorporating recommendation letters, which provide insights into an applicant's character and achievements, could improve the model's ability to evaluate applicants holistically. Other potential features might include interviews, portfolios, or other non-traditional elements of the application that could offer deeper insights into an applicant's fit for a particular institution.

Additionally, a thorough analysis of feature importance across the three primary types of features; numerical, categorical, and textual would further strengthen the model. By investigating how each feature category contributes to the admissions decision-making process, we could gain an understanding of which factors play the most significant roles in predicting successful outcomes. This could lead to the identification of previously overlooked features that could offer valuable predictive power.

Another important improvement would be the development of a user-friendly interface. This tool could allow prospective applicants or admissions counselors to enter their personal statistics—such as GPA, test scores, extracurricular activities, and essays and see how their profile compares to those of past applicants. By visualizing their likelihood of admission based on these factors, users could receive valuable feedback on how to strengthen their applications. This would be particularly useful for applicants seeking to understand which areas of their profile need improvement.

5 Statement of Work

Each team member's contributions to the project are listed as follows: Siyi Liu, idealization, dataset, literature review, feature importance; Wangzhi Zhan, building the auto-decision-maker; Hun Liang, literature review, clustering; Divya Polavarapu, literature review, clustering. Each person contributed the same amount of work when writing the final report. Wangzhi Zhan made the project website.

Acknowledgments

To Professor Eldardiry, James Weichert, and Vasanth Reddy: Thank you for such an informative and fun class.

References

[1] Amir Ahmad and Shehroz S. Khan. 2019. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access* 7 (2019), 31883–31902. https://doi.org/10.1109/ACCESS.2019.2903568

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc. https://www.nltk.org/book/

[3] Stephen L DesJardins, Halil Dundar, and Darwin D Hendel. 1999. Modeling the college application decision process in a land-grant university. *Economics of Education Review* 18, 1 (1999), 117–132.

[4] Rujun Gao, Hillary E. Merzdorf, Saira Anwar, M. Cynthia Hipwell, and Arun R. Srinivasa. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence* 6 (2024), 100206. https://doi.org/10.1016/j.caeai.2024.100206

[5] William Ho, Prasanta K Dey, and Helen E Higson. 2006. Multiple criteria decision-making techniques in higher education. *International journal of educational management* 20, 5 (2006), 319–337.

[6] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://doi.org/10.5281/zenodo.1212303

[7] Xin Jing and Hao Gao. 2024. An Improved K-PROTOTYPE Clustering Algorithm and Its Application. In *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing* (Sanya, China) *(MLNLP '23)*. Association for Computing Machinery, New York, NY, USA, 182–186. https://doi.org/10.1145/3639479.3639517

[8] S C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32 (1967), 241–254. https://api.semanticscholar.org/CorpusID:930698

[9] Hiba Jridi, Mohamed Aymen Ben Hajkacem, and Essoussi Nadia. 2020. *Parallel K-Prototypes Clustering with High Efficiency and Accuracy*. Springer, Cham, 380–395. https://doi.org/10.1007/978-3-030-59065-9_29

[10] S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[11] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. *CoRR* abs/1802.03888 (2018). arXiv:1802.03888 http://arxiv.org/abs/1802.03888

[12] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[13] Hanan Abdullah Mengash. 2020. Using data mining techniques to predict student performance to support decision making in university admission systems. *Ieee Access* 8 (2020), 55462–55470.

[14] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. 2019. Model-Agnostic Interpretability with Shapley Values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. 1–7. https://doi.org/10.1109/IISA.2019.8900669

[15] Ch. V. Raghavendran, Ch. Pavan Venkata Vamsi, T. Veerraju, and Ravi Kishore Veluri. 2021. Predicting Student Admissions Rate into University Using Machine Learning Models. In *Machine Intelligence and Soft Computing*, Debnath Bhattacharyya and N. Thirupathi Rao (Eds.). Springer Singapore, Singapore, 151–162.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[17] Ahmed M. Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz. [n. d.]. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems* n/a, n/a ([n. d.]), 2400304. https://doi.org/10.1002/aisy.202400304 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202400304

[18] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21.

[19] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669* (2024).

[20] Zhipu Zhou, Jie Chen, and Linwei Hu. 2022. Shapley Computations Using Surrogate Model-Based Trees. arXiv:2207.05214 [stat.ML] https://arxiv.org/abs/2207.05214