Predicting Medical School Admission Outcomes with Personal Statements and Research Aptitude Scores

Erica Chio
NYU Sackler Institute of Graduate Biomedical Sciences

Marina Marin
Institute for Innovations in Medical Education

June 30, 2020

A thesis in fulfillment of the Masters in Biomedical Informatics Degree

1 Summary

As medical school applications increase, the burden on application screeners also increases. With an increase in workload, errors and biases arise. Thus, Artificial Intelligence is meant to be the next step in reviewing applications, to counter those issues. In an application, there is the structured data, containing grades, hours, and numerical data. The other aspect is the unstructured data, the essays and text blurbs written by the applicant. This data is harder to process, but could potentially aid in determining an applicant's admission decision. Focusing on the personal statement and research experience, the text data is analyzed to determine whether it has any predictive power in an applicant's final admission decision. After adding a number of features, derived from the text data, and normalized for the human bias in faculty members, the research experience was able to perform decently well, achieving the best F1 score of 0.69. On the other hand, the personal statements models did not perform well, and the null hypothesis was accepted - there was no predictive power in personal statements.

2 Introduction

Medical school applications are increasing. New York University School of Medicine's applications in 2019 went up 47% to 8,932 applicants, up from 6,069 applicants from the year before (Weinstock, 2019). This increase in applications has led to a huge increase in workload for those responsible for screening these applicants. These admission officers have to go through primary applications, secondary applications, and interviews, to try and understand their applicant as deeply as possible. Medical school admission officers also have a responsibility to admit students that excel academically and have the personal qualities required for them to matriculate successfully (Witzburg and Sondheimer, 2013). In an application, there is the applicant's grade point average (GPA) and the Medical College Admission Test (MCAT) scores, research experiences, personal statements and letters of recommendation. All of these aspects come together to try and describe who an applicant is holistically, but do some aspects have more predictive validity to an admission outcome?

Applicant's GPA and MCAT scores are often studied to see if there is any predictive validity in the applicant's performance in medical school and clinical performance. Generally, the GPA and MCAT scores are ranked most important to an admission officer when looking at an application (Monroe et al., 2013). An applicant's GPAs are deemed a good predictor in medical school performance (Murden et al., 1978; Salvatori, 2001). On the other hand, the MCAT has differing results on the quality of its predictive validity. Some say that it has little to no predictive value in school or clinical performance, and instead its value lies in being used as a filter to screen out those with low test scores (Murden et al., 1978). There are other studies that state that the MCAT is good at predicting performance in school, and the problems with medical school applications lie within the unstandardized selection

tools such as essays and interviews (Salvatori, 2001). With much debate on the validity of the MCAT, it is hard to say definitively that there is any value in the MCAT scores.

Looking at an applicant's numerical (structured) data cannot detail everything about the applicant. If an applicant has strong extracurriculars, such as leadership positions or "extraordinary" stories, admission officers may be more lenient if their GPA is lower. Admission of these applicants lowers correlation of GPA to medical school performance, especially if these "outstanding" students go on to do well in medical school (Murden et al., 1978). Furthermore, there is the question of the reliability of GPAs from different schools, as each institution has different standards (Blue et al., 2000). From this, it can be gathered that, while there is merit to the GPA and MCAT scores when looking at an application for medical school, it cannot always provide the full picture as to who the applicant is.

These standardized aspects cannot represent the applicant's personal qualities. There is a need for personality traits such as compassion, integrity, and professionalism in future physicians (Price et al., 1971). The interview is meant to be a way for the school to determine personal traits of applicants and is often considered of the same importance as GPA and MCAT in an application (Price et al., 1971). However, there are studies that question the validity and reliability of an interview. Smith et al. studied two acceptance classes' grades, one accepted with interviews, and one without (Smith et al., 1986). Their grades in medical school showed no significant difference. In another study, Murden et al. found that interviews could have predictive success, such that those with good personal characteristics were more likely to receive outstanding reviews in internships (Murden et al., 1978).

The interview is meant to determine personal qualities that admission officers cannot discern from paper applications. However, there is an issue in determining which personal qualities are the most compelling and how to measure these qualities in a standardized and objective way (Albanese et al., 2003). Price et al. describes 87 personal qualities of a good physician, but realistically, is impossible to measure in each applicant (Price et al., 1971). Another issue in the interview is that the qualities demonstrated in the applicants are not necessarily stable (Albanese et al., 2003). In a literature review, it has been concluded that medical school environments actually decreased the compassion and humanism in students (Rezler, 1974). As such, interviews cannot be the only way to determine an applicant's personal qualities. Several solutions include using personal statements and letters of recommendation, which can also be used to deepen the admission officer's understanding of the applicant (Albanese et al., 2003).

A recommender's endorsement for an applicant is strengthened and becomes more reliable with longer term mentorship and partnership. Thus, another crucial aspect of an application is the research experience. Research experience demonstrates personal initiative, dedication to learning and problem solving (Murphy, 2020). However, not all research experience is considered significant, depending on duration, content, and resulting peer-reviewed papers (Kowarski, 2019). Research experience is something valued in current medical students, who are pushed

to pursue and dedicate time to research, especially in the earlier years of medical school (Honer and Linseman, 2004; Ommering et al., 2018). Earlier exposure to research experience was among the top three factors to be associated with career achievement in academic medicine (Brancati et al., 1992). Furthermore, early exposure was found to lead to a positive view on research, encouraging students to be more informed and confident when choosing a career in research (Boyle et al., 2017).

Personal statements are another unstructured aspect of the application, a place for the applicant to inform the admission officer know why they are interested in going to medical school. It is stressed how a bland personal statement is the reason why one applicant can be picked over another, especially when grades are similar. However, there aren't any studies or sources stressing the importance of the personal statement (Shemmassian, 2018). Furthermore, Albanese et al. conducted a literature review and concluded there hasn't been any studies on using the personal statement for assessing personal qualities, and that this could be an untapped resource (Albanese et al., 2003). Salvatori conducted a literature review on studies on the personal essay. She was able to conclude that subjectivity of content and rating are too varied to determine reliable predictive validity. Instead, she found personal essays limiting in being a key tool in selecting applicants, especially because there is no guarantee on how accurate the personal essay represents an applicant (Salvatori, 2001). Albanese et al. was able to conduct a survey on first year medical students to determine how well their personal statements represented themselves. A majority of the students surveyed over the three years, 53% - 84% felt that their essay "adequately represented some element of their personal characteristics." Their survey also included that 41% - 44% had input from others, and 2-6% had professional help (Albanese et al., 2003). Salvatori also found investigators that had positive results, with both validity and reliability, but felt that there needed to be more results before personal statements could be a reliable selection tool (Salvatori, 2001).

A lot of the issues that come from personal statements is the subjectivity from reader to reader, applicant to applicant. Thus, with structured and standardized changes to how an essay is graded and written, it could potentially have more predictive validity. To combat subjectivity, there is an effort to create a system of grading personal statement essays. Dong et al asked a journalist to create a rubric based on journalism literature. From there, they graded each essay based on elements such as coherence, persuasiveness, passion for medicine, confidence, grammar, and others.Then, they examined the essays scores to the students' school and internship performance to determine if the scores had any relationships. There were only weak correlations among the application essay scores and their future performance, but admit to limitations in their study such as small sample size and only one journalist grader (Dong et al., 2013). However, this was the only study done on standardizing the grading of essays in medical school applications.

Automated essay grading systems would potentially be beneficial to medical schools. It would offer efficiency and lower workloads for admission officers. However, through a literature review, there are no studies or reviews

on using automated essay scoring systems on medical applications. There are, however, studies on using this method on other standardized essays, such as Common Core Tests and Test of English as a Foreign Language Test (TOEFL) (Haberman, 2011; Wilson, 2018). Automating grading essays uses computing methods such as Natural Language Processing (NLP), latent semantic analysis, and machine learning (Shermis and Burstein, 2003, 2013). These automated essay scoring systems have scored essays consistent with human graders, and could potentially decrease human bias and error (Shermis, 2014). However, critics argue that if essays are to be measured by machines, writers will eventually write for machines, instead of humans - the intended reader (Deane, 2013). But, with the use of automated essay grading in standardized testing, it has the potential to expand into application screening processes.

Automated essay grading systems utilize natural language processing (NLP), a field dedicated to turning qualitative data to quantitative. In other words, it is meant to take text data and parse it into something that computers can understand the way humans can. It is not a single technique, but rather multiple techniques to create this understanding. NLP works by breaking down text into something quantitative that can be analyzed with machine learning techniques (Chary et al., 2018). When combined with machine learning techniques, NLP has the potential to help lowering burdens and responsibilities within the medical world. For example, NLP has already been successful in structuring clinical notes, which are often hard to interpret and filled with abbreviations, and predicting potential medical complications from physician notes (Kreimeyer et al., 2017; Murff et al., 2011). Thus, NLP could potentially help screen medical school applications, as there is a lot of unstructured data - such as personal statements and letters of recommendation - that admission officers have to go through for each applicant.

This literature review found that automated essay scoring used semantic and syntactic analysis (Wilson, 2018). Syntactic analysis is essentially the grammar and structure of a sentence. This involves identifying the subject, predicate, nouns, verbs and pronouns. However, it is hard to distinguish these different parts of a sentence, especially when a word can be both a noun and a verb. Semantics on the other hand, is the study of the meaning behind the sentence. It focuses on the way the sentence was written, and what it is trying to convey (Redd and M, 2014).

Syntactic analysis of the personal statement could be in the form of scoring the readability, complexity and grade level in which the personal statement was written. The Flesch Readability Index was introduced by Rudolph Flesch to determine how easy it is to read a text. The score is calculated by average sentence length and average syllable count, percentage of “personal words”, and average percentage of personal sentences. A score closer to 100 indicates the text is very easily read and understood, while a lower score indicates a difficult text (Flesch, 1948).

In an attempt to represent the personal statement’s semantics, word embedding is an option to represent that information. Word embedding encodes each word as a vector, allowing words that are commonly found in

similar contexts to be near each other in the vector space. This allows simple algebraic operations to accurately represent a word: “wv(“Madrid”) - wv(“Spain”) + wv(“France”) is expected to be close (Zhang et al., 2016).” A popular software utilizing these word embeddings is word2vec (Mikolov, 2013; Mikolov et al., 2013). Word2vec has two variants proposed - a continuous bag of words (CBOW) model and a continuous skip-gram model. CBOW looks at the context and the words around to determine the meaning of a word, while skip-gram looks at a word to determine the context of the surrounding words (Zhang et al., 2016).

These techniques will make it possible to try and model medical school applications, to classify which applications could potentially be admitted or rejected.

3 Methods

3.1 Data

The data was obtained from NYU Medical School 2014-2019 applicants. Application data can be split into two types of data, structured and unstructured. The structured data includes grades, hours and experiences, number of peer-reviewed publications, MCAT grades and percentiles, etc. The other aspect of the application can be categorized as unstructured data - text data. This data includes text blurbs where applicants detailed their research experience and personal statements. The unstructured data is to analyzed to determine if it could aid pre-screening of medical school applications. Thus, the research experience text blurb and personal statements were pulled from the applications to see if they could be classified. The research experience was classified by faculty members, who scored the applicant’s research experiences. However, since the personal statements did not get individually scored, it would be added into the structured data model as another feature, and to be classified by outcome of the application.

3.1.1 Research Experience Data

The research experience was scored by two faculty members, using a rubric. The scores ranged from 1-5, where 1 indicated no research experience, and 5 indicated significant research experience. Scores 2 and 3 indicated summer experience with no significant research progress, while scores 4 and 5 demonstrated long term (spanning more than one year) experience. Scores 4 and 5 acknowledged the presence of significant research progress through contribution to a peer-reviewed publication. Thus, the scores were binarized into two classes with scores 4 and 5 being positive, and anything lower, negative.

Key factors that differed the positive and negative classes are length of duration, being an author on a peer-reviewed publication, and whether or not the applicant received a letter of recommendation from their research

mentor. The applicant gave the dates in which they worked at each organization, which was converted into months. From there, whether the experience lasted three, six, or twelve months was calculated. This, along with the number of publications an applicant had was also added into the model as features.

3.1.2 Personal Statement Data

Instead of scoring each personal statement, they were classified by outcome of the applicant - whether the applicant was invited, held, or rejected. Besides the actual text data, readability indexes, and other features were added to give it numerical data.

Features such as Word and sentence count of the personal statement was added. Lexical density, which is the number of lexical (context) words divided by total number of words, was calculated. The number of times ”NYU” or ”New York University” was mentioned within the essay was counted. The Flesch Reading Ease and Standard Score were also added in as features of the personal statement to further categorize the text. The Flesch Reading Ease measures the readability of the text by grade level. It is scored from 0-100, where scores 90-100 is considered very easy to read and scores 0-29 is considered very difficult. The Standard Score is a composite score created by textstat (a python package), which is calculated from multiple grade level tests such as Dale Chall, Gunning FOG, and Automated Readability Index. The resulting score would correlate to the grade level, so a score of 4 would indicate a fourth grader level and a score of 12 would indicate a twelfth grader. A score of 13 indicates a college student, and any score higher indicates a college graduate to professor level.

3.2 Word Embedding

The text data, both research experience and personal statements, have to be converted in a way that the model and computer will be able to understand it. Word embedding was chosen to vectorize the text because it maps the words to vectors, allowing the computer to understand it. In word embedding, word2vec is a common way to process the words into vectors in a way where the vectors represent the meaning of the words.

Word2vec can be hypertuned to get the best possible representation of the data. There is the option of using a pre-trained word2vec model, such as the Google News model, or training the vectors on the text data (corpus) itself. The Google News model was trained on 3 billion words and has vectors for 3 million words. Each vector length is 300 features long. Training the word2vec model from scratch allows more flexibility and customization, but is trained on a much smaller dataset. When training from scratch, an embedding size of 300 was chosen to maintain uniformity between the Google News vector model and the self-trained model. There are slight differences in the way each model maps the same words (**Figure 1**).

Other hypertuning options for word2vec is whether to use CBOW or skip-gram. To further enhance the vectors

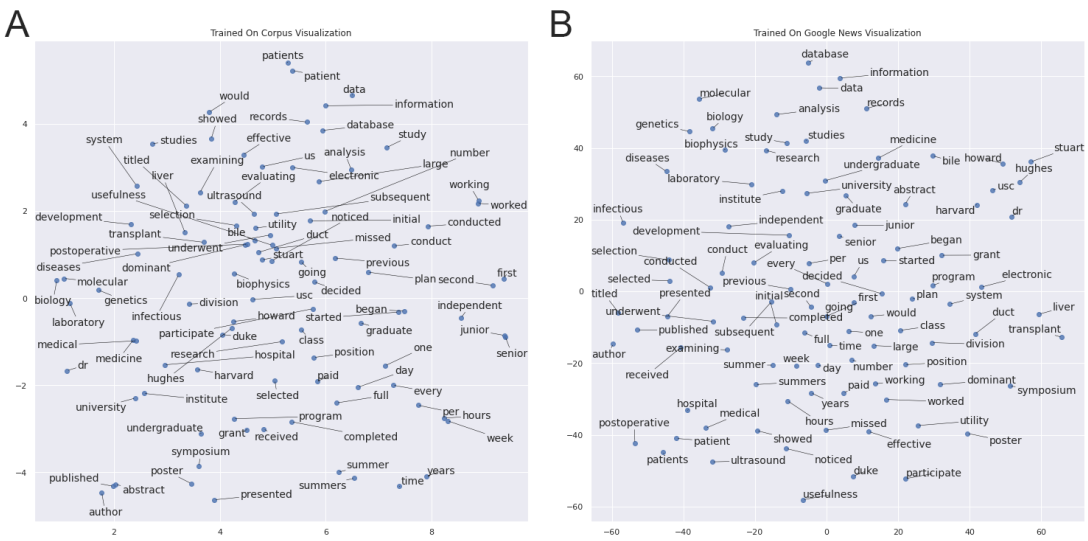

Figure 1: word2vec Visualization (A) A visualization of 100 words trained on the corpus itself. **(B)** A visualization of 100 words trained on Google News model (previously trained on 3 billion words.

representing the text data, Term Frequency — Inverse Document Frequency (TFIDF) was introduced into the model. TFIDF is a weight used to evaluate how important a word is based on the frequency it appears in each text blurb, but is offset by how often it appears in the entire corpus.

3.3 Models

3.3.1 Research Aptitude Model - Faculty ID Normalization

To account for potential faculty member bias, there was an attempt to normalize against harsher or more lenient scorers. One method was to only trust the scores in which both faculty members agreed that an applicant's research experience was positive (a score 4 or above) or negative (a score 3 or below). Thus, those applicants who had one faculty member who scored their experience as negative and one who scored their experience as positive would be removed from the dataset, as the scores would be contradictory.

Another way to normalize is to input a feature into the model to signify how harsh or lenient a faculty member is at scoring research experience. First, each faculty member had all their scores averaged. Then, this individual faculty member average score was subtracted from the average of the entire dataset's scores. If a faculty member's score was lower than the overall average, their difference would be negative, indicating they were a potentially harsher grader. If a faculty member's difference was positive, it would suggest that they are more lenient (**Figure 2**). This difference would be added into the model as a weight for each faculty member.

Aside of normalizing the faculty ID with an input, another option is to normalize the output / predicting value. The normalized score would be the z-score (**Equation 1**). The raw score would be the applicant's aptitude score.

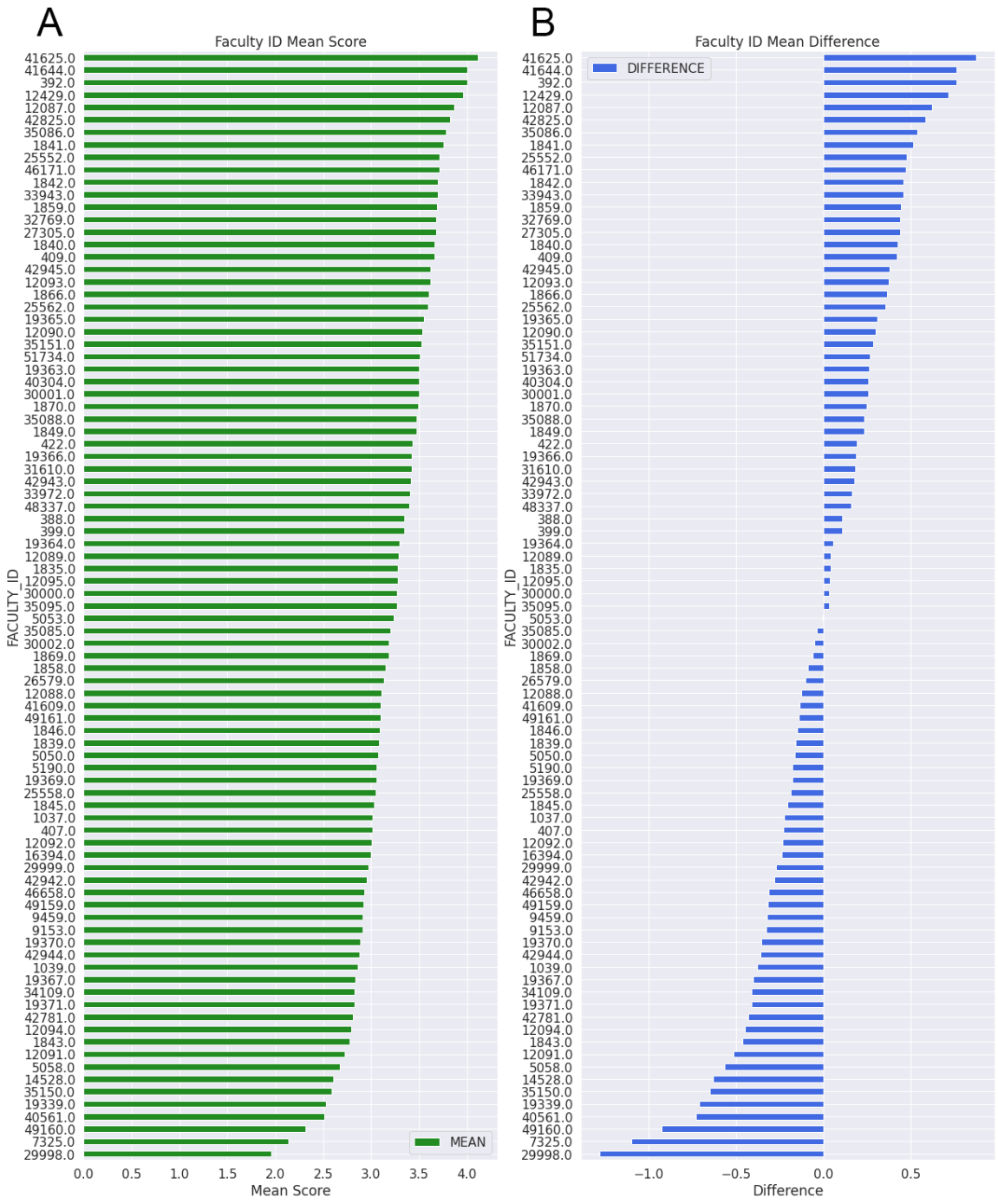

Figure 2: Faculty Mean Scores and Difference (A) Each faculty member's mean score. There is a clear range, with some faculty members averaging a score of 4.0, while some average closer to 2.0. **(B)** If a faculty member's average score was lower than that of the dataset mean score, that would indicate that that faculty member was a harsher grader. On the flip side, a positive difference would indicate a more lenient faculty member

The population would be that faculty member's scores. Then, these new scores would be the new predicting value. To maintain the same class distribution, the cutoff of positive was kept the same as the original aptitude scores. Thus, since the scores 4 and 5 represented 13% of the original scores, the cutoff for 13% in z scores was 0.25 (**Figure 3**). This new score would represent what the score would be if the faculty member was not bias.

$$\text{z-score} = \frac{(\text{raw score} - \text{mean of population})}{\text{standard deviation of population}} \tag{1}$$

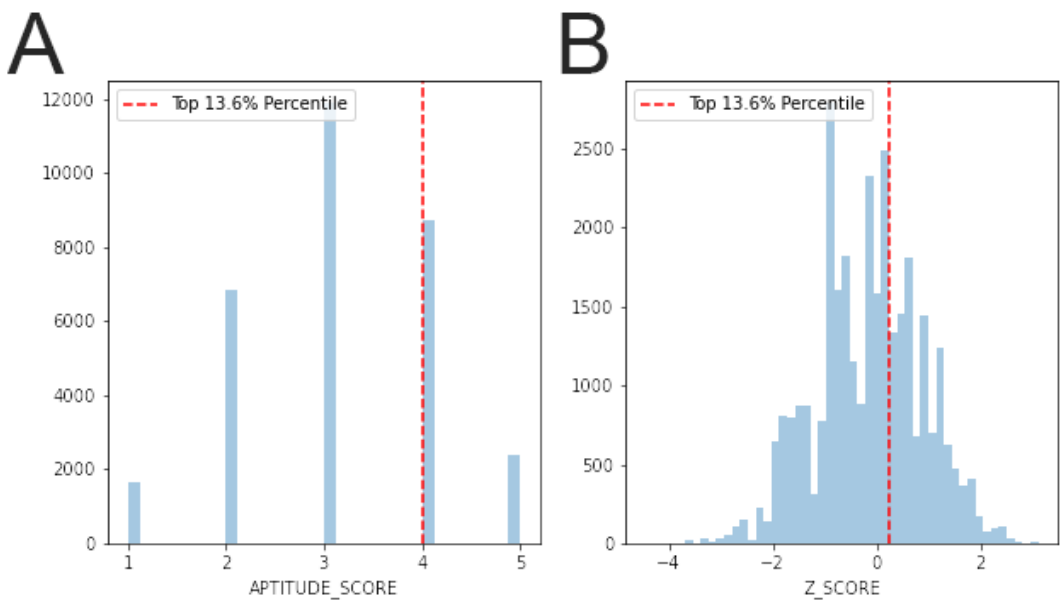

Figure 3: Distribution of Original and Normalized Scores (A) The distribution of the original aptitude scores. **(B)** The distribution of the newly calculated z-score output. The red line indicates the cutoff point of what is considered "positive" class (significant and good research experience) and "negative" class (insignificant research experience).

Thus, four comparative models were created from this normalization. The first is where the data was input as is, and the predicting aptitude score was the average of the faculty members' scores. The next model would be only concordant scores between the faculty members, signifying confidence in the applicant's aptitude score. Then, there would be one model where the faculty member bias is normalized with an added weight in the model, and one where the output was normalized.

3.3.2 Personal Statement Model

The personal statements resulted in three different models. One in which just the personal statement word embeddings and features were modeled. Then, one where just the structured data was modeled, and finally, one with structured data and the personal statement features. This is to determine whether or not the personal statement features added had any classifying power, by itself and when added to structured data.

3.4 Classifiers

Five different classifiers were tested to see which would perform the best. The different classifiers were stochastic gradient descent (SGD), Logistic regression, Gaussian Naive Bayes, Random Forest Classifier and XGBoost Classifier.

3.5 Evaluation Metrics

The model was evaluated through several metrics. Since the model is meant to pre-screen, False Positives are more preferable than False Negatives. This means that those that are classified as positive would continue to be evaluated. Thus, precision and recall are important metrics to be considered for these models. Thus, the F-1 score, which is calculated from precision and recall is looked at. Finally, the AUC-ROC curve is plotted to determine the True Positive Rate versus the False Positive Rate.

4 Results

4.1 Research Aptitude Model

The models were compared against each other with the evaluation metrics such as F-1 score and AUC-ROC curves (**Table 1**).

Within the added feature, the dataset that detailed the applicant's letter of recommendation author had little overlap, resulting in a much smaller dataset (8627 applicants to 2337 applicants). This feature was checked for feature importance, and eventually removed from all other models.

4.1.1 Data As Is

In this model, the data was plugged into the model without further manipulation to the scores or normalization from faculty member biases. It contained all of the added features, such as number of publications, duration of experiences, number of unique experiences, and total hours. The best classifier was Gaussian Naive Bayes with an AUC-ROC score of 0.71, precision of 0.46, recall of 0.44, and a F1 score of 0.45.

4.1.2 Concordant Scores

This data was manipulated so that the only applicants that had both faculty members agree on class was kept, resulting in a better performing model. The best classifier was XGBoost with an AUC-ROC score of 0.85, precision of 0.72, recall of 0.58, and a F1 score of 0.65.

Model	Classifier	AUC-ROC	Precision	Recall	F1 Score
Data (as is)	GNB	0.71	0.46	0.44	0.45
Concordant Ratings between Faculty	XGB	0.85	0.72	0.58	0.65
Concordant Ratings, Faculty ID - normalized with an added feature	XGB	0.89	0.75	0.64	0.69
Concordant Ratings, Faculty ID - normalized with an added feature + Letter of Recommendation Feature	XGB	0.86	0.73	0.61	0.66
Concordant Ratings, Faculty ID - normalized output	XGB	0.85	0.70	0.55	0.62

Table 1: Research Aptitude Model Performance The best classifier of each model and each model's performance in AUC-ROC score, Precision, Recall, and F1 Score.

4.1.3 Concordant Scores, Faculty ID Normalized with a Input / Feature

This model performed the best of the research aptitude models. It had an added weight for each faculty member to represent how harsh or lenient of a grader they were. The best classifier was XGBoost with an AUC-ROC score of 0.89, precision of 0.75, recall of 0.64, and a F1 score of 0.69. The feature importance graph of this model indicated that the added weights of the faculty member did have some significance. But still, the most important feature is how many peer-reviewed publications an applicant had (**Figure 4**)

Since this model performed the best, the letter of recommendation author feature was added to check its significance. After running this model, the best classifier with this added feature was XGBoost with an AUC-ROC score of 0.86, precision of 0.73, recall of 0.61, and a F1 score of 0.66. Both models performed equally, indicating that this new feature did not have a huge impact in performance. All of the added features' importance was also plotted (**Figure 5**). It is evident that whether or not the letter of recommendation came from where the applicant conducted research has very little importance. Thus, it was subsequently removed as a feature from other models to have a larger sample size.

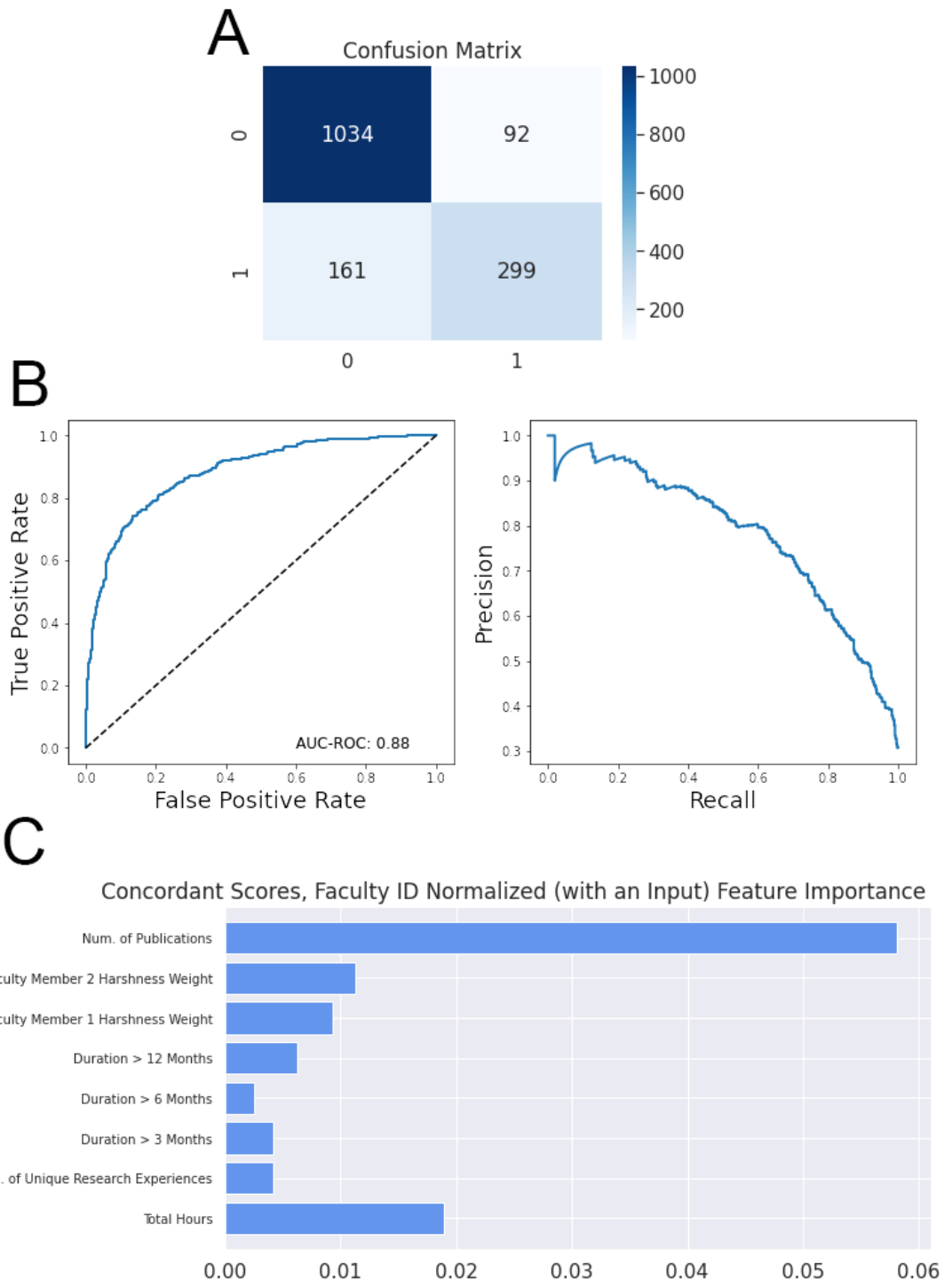

Figure 4: Visualization of Concordant Scores, Faculty ID normalized with an Input Model The best performing classifier was XGBoost. **(A)** The confusion matrix of the model. 1 represents the positive class, while 0 represents the negative class. The class split was close to 30:70 ratio. **(B)** AUC-ROC and Precision Recall Curve of the model. **(C)** Feature Importance Plot of the Model. Number of Publications was the most important feature, with total hours being second. The two weights added to account for faculty member bias was the next two important features.

4.1.4 Concordant Scores, Faculty ID Normalized as an Output / Predicting Variable

This last model of the research aptitude models had a different predicting variable than the other models. This new score was calculated in an attempt to normalize and reduce the faculty member bias present. The best classifier

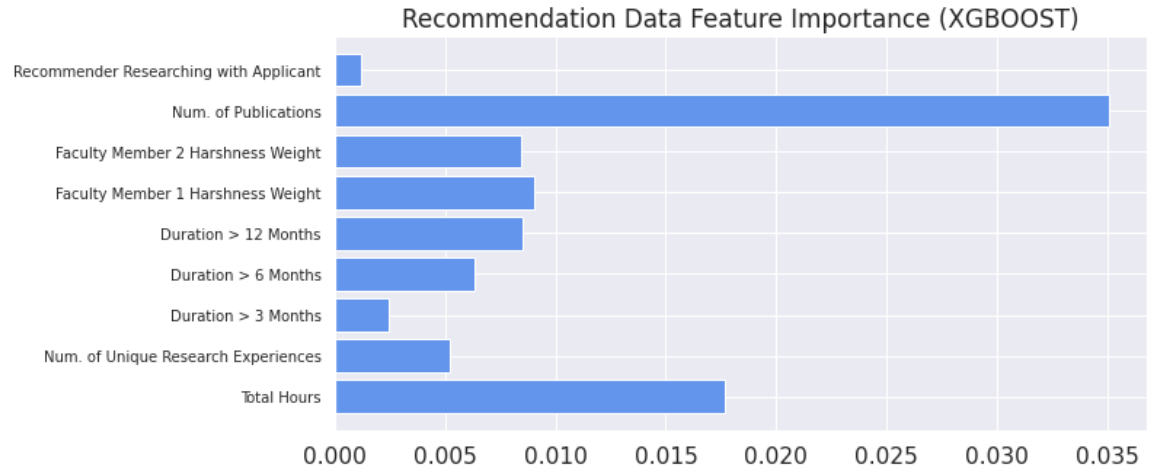

Figure 5: Recommendation Feature This is the feature importance plot of the model containing letter of recommendation author data. Whether the applicant had a letter of recommendation by their research mentor was the least important of the added features.

was XGBoost with an AUC-ROC score of 0.85, precision of 0.70, recall of 0.55, and a F1 score of 0.62.

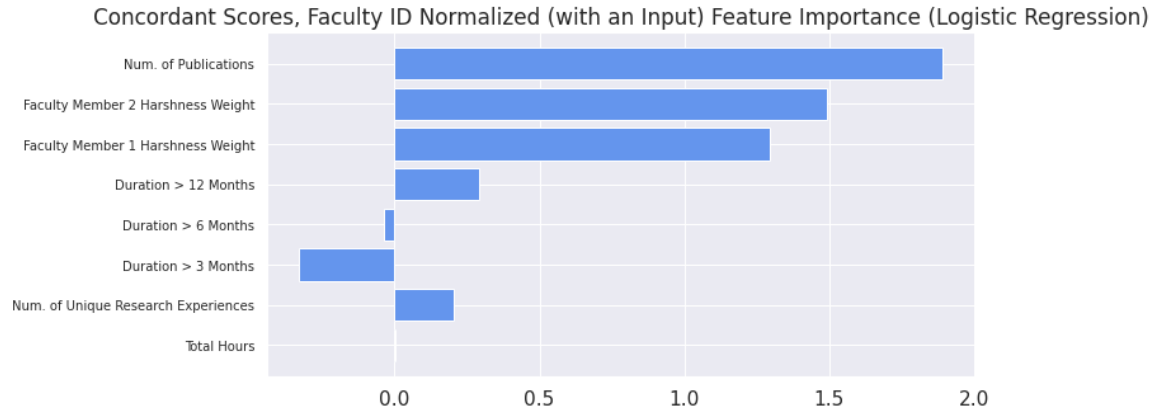

Figure 6: Logistic Regression Feature Importance This is the feature importance plot of the logistic regression classifier for the concordant, faculty member normalized with an added weight model. The values that are negative, are more likely to predict the negative class. The positive valued features are more likely to predict the positive class. Most features predicted the positive class. However, there are two features predicting the negative class - how many experiences lasted more than 3 months, and how many experiences lasted more than 6 months.

The XGBoost classifier performed consistently well for all these models, but the Logistic Regression classifier performed second best. When looking at the Logistic Regression model's feature importance, it reports which class the feature is more likely to predict. Interestingly, two features were related to the negative class: research experience lasting around 3 months and research experience lasting around 6 months (**Figure 6**).

4.2 Personal Statement Model

The features calculated from personal statements, such as readability index, standard score, NYU mentions, lexical density, were plotted to tentatively determine if there was any relevance in these scores to the outcome of the application. From these plots, it is evident that there is no difference in distribution of these features whether or not an applicant was admitted (**Figure 7**).

The models were compared against each other with the evaluation metrics such as F-1 score and AUC-ROC curves (**Table 2**).

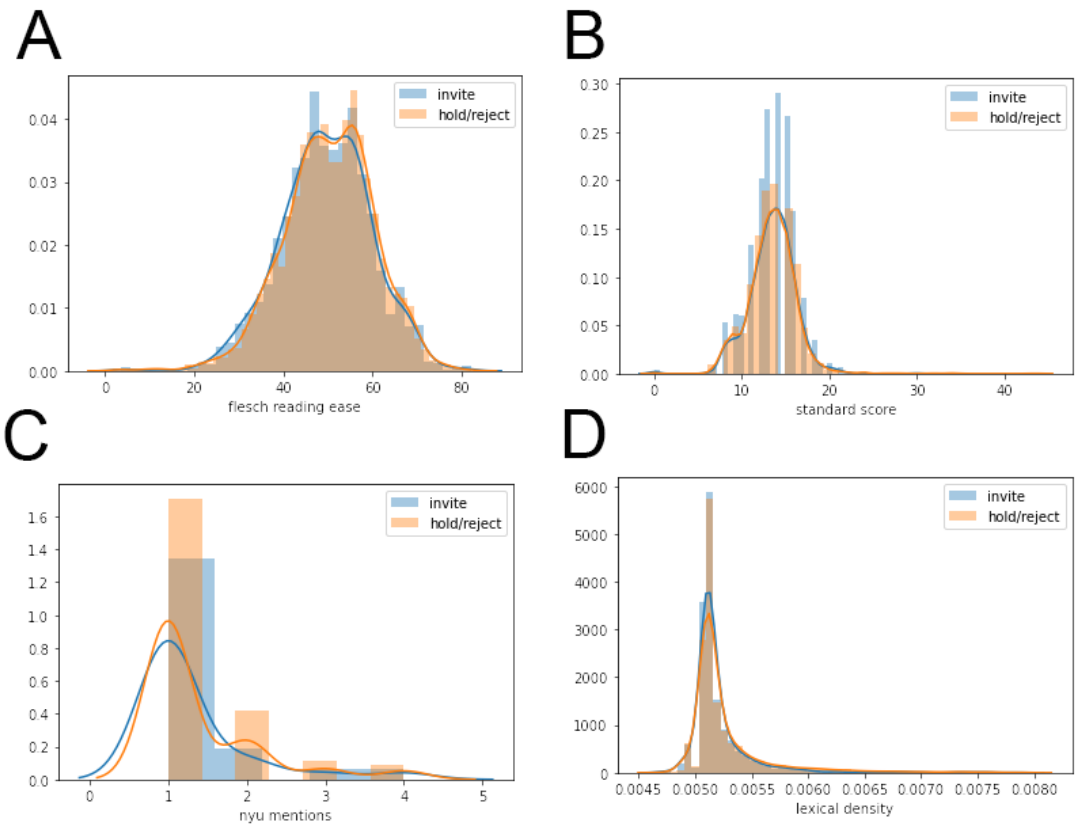

Figure 7: Personal Statement Features Relevance The distribution of those who were invited and those who were rejected overlap for all four features. This indicates there is no significance relevance of the feature to the outcome of the applicant. **(A)** The distribution of Flesch Reading Score between those applicants who were invited and those who were held or rejected. **(B)** The distribution of standard score. **(C)** The distribution of NYU mentions in the personal statement. **(D)** The distribution of lexical density.

4.2.1 Personal Statement Features

The personal statement and the added features itself had very little predictive validity, as the classifiers were not able to correctly predict the outcome of the application solely based on the personal statement. The best classifier was Guassian Naive Bayes with an AUC-ROC score of 0.60, precision of 0.32, recall of 0.57, and a F1 score of

Model	Classifier	AUC-ROC	Precision	Recall	F1 Score
Personal Statement Features	GNB	0.60	0.32	0.57	0.41
Structured Data	XGB	0.82	0.62	0.49	0.55
Structured Data + Personal Statement Features	XGB	0.81	0.64	0.46	0.54

Table 2: Personal Statement Model Performance The best classifier and corresponding AUC-ROC, Precision, Recall, and F1 Score are documented.

0.41 (**Figure 8**).

4.2.2 Structured Data

This model's performance is a baseline to determine if the personal statements features would add anything. The best classifier was XGBoost with an AUC-ROC score of 0.82, precision of 0.62, recall of 0.49, and a F1 score of 0.55 (**Figure 8**).

4.2.3 Structured Data and Personal Statement Features

The best classifier was XGBoost with an AUC-ROC score of 0.81, precision of 0.64, recall of 0.46, and a F1 score of 0.54 (**Figure 8**).

The personal statement feature importance was plotted to see if there was any feature of importance. Although some features were more important than others, it was of 0.0030 importance, essentially insignificant (**Figure 9**).

5 Discussion

5.1 Research Aptitude Model

Interestingly, whether or not the letter of recommendation came from a research mentor was also shown to have low feature importance (**Figure 5**). Originally, this feature was thought to have significance because in the rubric, those applicants with a letter of recommendation from a research mentor were supposedly bumped up one level. However, in practice, almost 60% of the applicants received a letter of recommendation from their research mentor. This would render the feature insignificant as a majority of the applicants had a letter of recommendation from their research mentor.

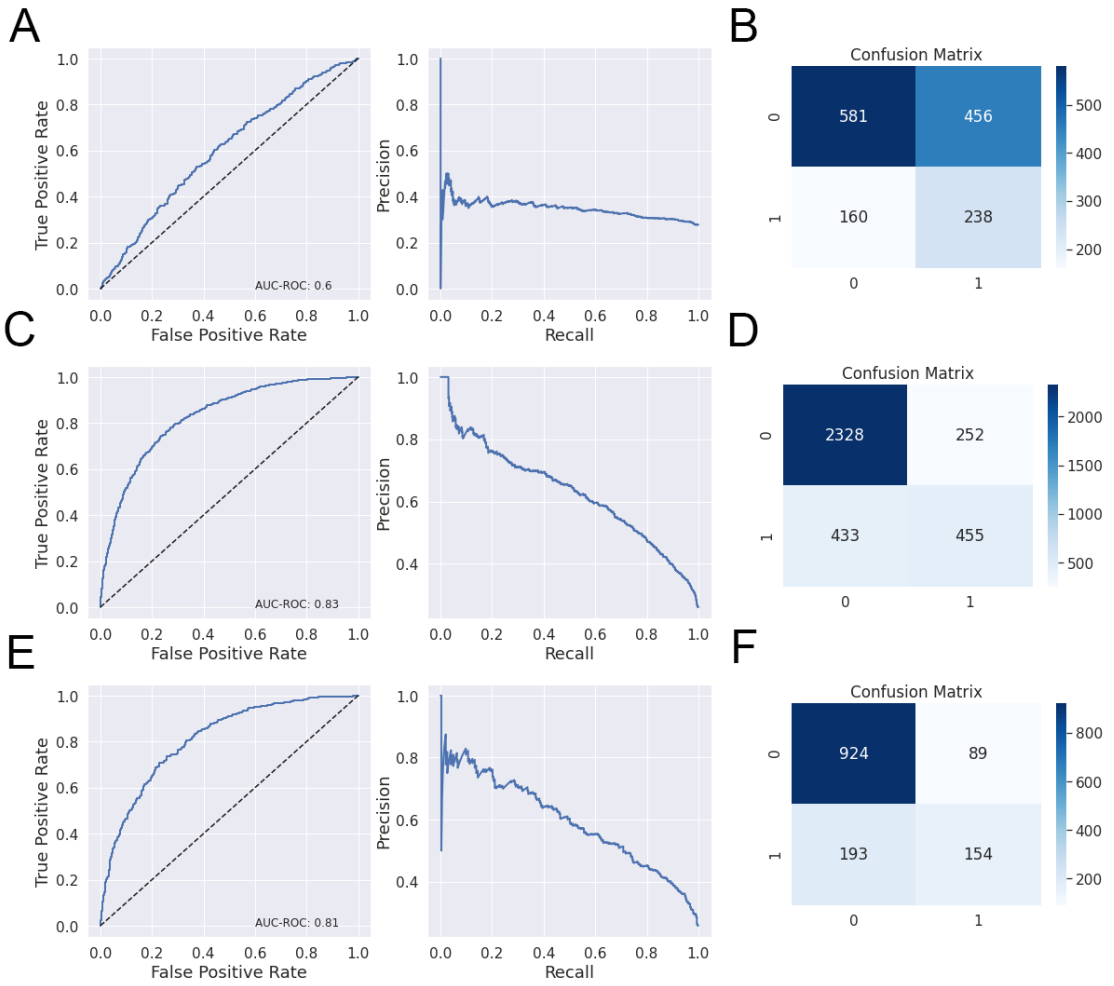

Figure 8: Personal Statement Model Visualizations (A) The AUC-ROC and Precision Recall curves of the model that only contained personal statement features. **(B)** Confusion matrix of the personal statement features model. The class split was around 25:75, positive to negative. **(C)** This is the AUC-ROC and Precision Recall Curve of the structured data model. **(D)** The confusion matrix of the structured data model. **(E)** The AUC-ROC and Precision Recall curve of the combined structured data and personal statement features model. **(F)**The confusion matrix of the model.

The logistic regression model had two features which predicted the negative class (**Figure 6**). According to the rubric, scores 2 and 3 are associated with summer research, generally 3-6 months long. Thus, the features should be associated with the negative class. The third duration feature, whether the experience lasted over 12 months predicted positive class. Scores 4 and 5 expected long term, potentially year long experiences. Thus, duration of experience was a significant feature in the model.

The original model with no normalization or data manipulation did the worst. There are a lot of factors that could be attributed as to why this model did the worst. Most importantly, this model did not account for human bias by the faculty members, especially as it had been screened by multiple faculty members - meaning there was no uniformity in the bias.

With concordant scores between faculty members, the model was able to predict more accurately. This can be

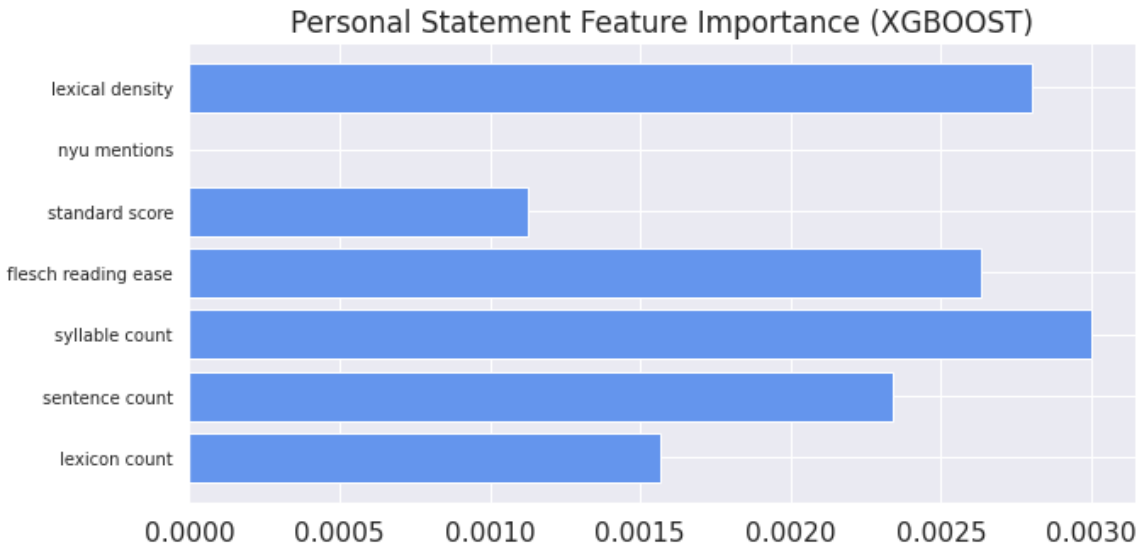

Figure 9: Personal Statement Feature Importance Some of these features were more important than others. ”NYU” mentions had no significance in the model, but syllable count was found to be of more importance than other features.

attributed to the fact that these scores were more ”confident,” as both graders agreed on its score. This would remove all the middle, more subjective applicants, allowing the model to more easily distinguish between the positive and negative classes. In 30% of the applications, the two faculty members did not grade the research experience concordantly, which were removed. When this discordance is removed, the performance of the model improved, meaning that those cases were potentially throwing the model off since they were not confidently considered one class or the other.

The research aptitude model did best when faculty member harshness and leniency were factored in. It becomes evident that faculty member bias plays a huge part in the scoring of research aptitude. The weight was the difference of the faculty member’s mean score and the entire applicant pool’s average score. This feature was within the top 5 important features. With more relevant features, it would make sense that this model performed the best (**Figure 4**).

Another attempt was to try and normalize the output scores so that it would be what the faculty member would have potentially graded the experience, had there been no bias. This model did worse than the other faculty member normalization model. This could be because it lacked two extra features, so it would have been more comparable to the second model, the model with just concordant scores. However, this model performed worse than that model as well. This could indicate that the new outcome variable was not normalized accurately. One factor is that aside from human bias, a faculty member could only have worse applicants in their screening pile. This would mean their average grade would be lower, but through this normalization, these applicant’s scores would be raised, even if they are not deserving of it.

In conclusion, this model would still need to be worked on before it could be implemented in production to help

pre-screen medical school applicants. In this experiment, best model was determined through F1 score, which is calculated with precision and recall. In production, instead of placing more value on the F1 score, recall could be more heavily considered. In a pre-screening model, false negatives would be more detrimental than false positives. False negatives would result in losing potentially good applicants, but false positives would just require further screening to remove.

This model is more analytical than ready to be in production because it relies a lot of post review data, eg: faculty member biases. This model could potentially be integrated with the structured data model to create a singular pre-screening model. However, since this model isn't completely accurate, several tests would still have to be done. One would be to test whether the model was a more accurate predictor of research aptitude or faculty members, testing which set of scores would more accurately predict admission outcome. Another route that could be taken with this model is to implement it between interview invitations and final admission outcome to determine if research aptitude played a larger role in that portion of the application process.

5.2 Personal Statement Model

The personal statement features model did very poorly by itself. The AUC-ROC was 0.60, which is only a little better than when a model randomly predicting the classes as 50/50. Furthermore, the model predicted almost half of the observations as one class, and half as the other even though the class split was supposed to be 25:75. This meant the model was not able to learn at all.

However, the personal statement was meant to be combined with the structured data model to see if it could help the structured data perform better. The two performances between structured data and structured data and personal statement model are essentially the same. Thus, it can be said that the null hypothesis is true, that personal statements do not relate to admission outcomes. None of the personal statement readability indexes, or features were able to predict an application outcome which was expected due to earlier analysis of the distributions (**Figure 7, 9**). Instead of focusing on how well an essay was written, future analysis should focus more on semantics.

References

Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., and Farrell, P. M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine*, 78(3):313–321.

Blue, A. V., Gilbert, G. E., Elam, C. L., and Basco, W. T. (2000). Does institutional selectivity aid in the prediction of medical school performance? *Academic Medicine*, 75.

Boyle, S. E., Cotton, S. C., Myint, P. K., and Hold, G. L. (2017). The influence of early research experience in medical school on the decision to intercalate and future career in clinical academia: a questionnaire study. *BMC medical education*, 17(1):245.

Brancati, F., Mead, L., Levine, D., Martin, D., Margolis, S., and Klag, M. (1992). Early predictors of career achievement in academic medicine. *JAMA*, 267(10):1372–1376.

Chary, M., Parikh, S., Manini, A., Boyer, E., and Radeous, M. (2018). A review of natural language processing in medical education. *Western Journal of Emergency Medicine*, 20(1):78–68.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24.

Dong, T., Kay, A., Artino Jr., A. R., Gilliland, W. R., Waechter, D. M., Cruess, D., DeZee, K. J., and Durning, S. J. (2013). Application essays and future performance in medical school: Are they related? *Teaching and Learning in Medicine*, 25(1):55–58.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology,*, 32(3):221–233.

Haberman, S. J. (2011). Use of e-rater®in scoring of the toefl ibt®writing test. *ETS Research Report Series*, 2011(2):i–13.

Honer, W. G. and Linseman, M. A. (2004). The physician-scientist in canadian psychiatry. *Journal of psychiatry neuroscience : JPN*, 29(1):49–56.

Kowarski, I. (2019). What type of research impresses med schools?

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., and Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of biomedical informatics*, 73:14–29.

Mikolov, T. (2013). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. pages 3111–3119.

Monroe, A., Quinn, E., Samuelson, W., Dunleavy, D. M., and Dowd, K. W. (2013). An overview of the medical school admission process and use of applicant data in decision making. *Academic Medicine*, 88(5):672–681.

Murden, R., Galloway, G. M., Reid, J. C., and Colwill, J. M. (1978). Academic and personal predictors of clinical success in medical school. *Academic Medicine*, 53(9):711–9.

Murff, H. J., Fitzhenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., and Speroff, T. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *Jama*, 306(8).

Murphy, B. (2020). How research experience can strengthen your medical school application.

Ommering, B., van Blankenstein, F. M., Waaijer, C., and Dekker, F. W. (2018). Future physician-scientists: could we catch them young? factors influencing intrinsic and extrinsic motivation for research among first-year medical students. *Perspectives on medical education*, 7(4):248–255.

Price, P. B., Lewis, E. G., Loughmiller, G. C., Nelson, D. E., Murray, S. L., and Taylor, C. W. (1971). Attributes of a good practicing physician. *Academic Medicine*, 46(3):229–37.

Redd, M. V. and M, H. (2014). Semantical and syntactical analysis of nlp. *International Journal of Computer Science and Information Technologies*, 5(3):3236–3238.

Rezler, A. G. (1974). Attitude changes during medical school. *Academic Medicine*, 49(11):1023–30.

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*, 6:159–175.

Shemmassian, S. (2018). Medical school personal statement: The ultimate guide (examples included).

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.

Shermis, M. D. and Burstein, J. (2003). *Automated Essay Scoring: a Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates.

Shermis, M. D. and Burstein, J. (2013). *Handbook of automated essay evaluation: current applications and new directions.* Routledge Taylor Francis Group.

Smith, S. R., Vivier, P. M., and Blain, A. L. (1986). A comparison of the first-year medical school performances of students admitted with and without interviews. *Academic Medicine*, 61(5):404–6.

Weinstock, M. (2019). Qa: 'the debt burden of these students is unsustainable in the 21st century'.

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology*, 68:19–37.

Witzburg, R. A. and Sondheimer, H. M. (2013). Holistic review — shaping the medical profession one applicant at a time. *New England Journal of Medicine*, 368(17):1565–1567.

Zhang, Y., Rahman, M. M., Braylan, A., Dang, B., Chang, H.-L., Kim, H., McNamara, Q., Angert, A., Banner, E., Khetan, V., McDonnell, T., Nguyen, A. T., Xu, D., Wallace, B. C., and Lease, M. (2016). Neural information retrieval: A literature review.