UNIVERSITY OF CALIFORNIA,
IRVINE

Machine Learning with Keyword Analysis for Supporting Holistic Undergraduate Admissions in Computer Science

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Barbara Martinez Neda

Thesis Committee:
Assistant Professor of Teaching Sergio Gago-Masague, Chair
Assistant Professor Sangeetha Abdu Jyothi
Associate Professor Sameer Singh
Associate Professor of Teaching Jennifer Wong-Ma

2022

DEDICATION

To my mother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Dr. Sergio Gago-Masague for his guidance and support throughout my undergraduate and graduate academic career. I also thank my committee members, Dr. Sangeetha Abdu Jyothi, Dr. Sameer Singh, and Dr. Jennifer Wong-Ma for sharing their knowledge inside and outside the classroom. I would also like to thank the ICAPAI 2022 organizing committee and the anonymous reviewers for their review of *Feasibility of Machine Learning Support for Holistic Review of Undergraduate Applications*, on which this thesis is based[1]. Lastly, I would like to thank my mother, husband, family, and friends for their support and help along the way.

[1]Permission request not required for IEEE article reuse in a thesis - https://journals.ieeeauthorcenter.ieee.org/choose-a-publishing-agreement/avoid-infringement-upon-ieee-copyright/

ABSTRACT OF THE THESIS

Machine Learning with Keyword Analysis for Supporting Holistic Undergraduate Admissions in Computer Science

By

Barbara Martinez Neda

Master of Science in Computer Science

University of California, Irvine, 2022

Assistant Professor of Teaching Sergio Gago-Masague, Chair

College admissions processes have traditionally relied on academic characteristics like GPA and standardized testing, as well as supplementary application materials. In California, the introduction of Proposition 209 in 1996 prohibited the consideration of gender and ethnicity for admissions decisions. In an attempt to increase diversity, many universities adopted holistic review to fairly evaluate and consider applicants' abilities inside and outside the classroom. However, this increases subjective assessment which could have implications for human error and bias. As such, Machine Learning should be explored as a means of assistance while also reducing potential bias.

Minimal data regarding Machine Learning applications in undergraduate holistic review has been evaluated. In this thesis, we discuss performances of supervised classifiers that could provide verification of the scores that application reviewers assign. We utilize a dataset of applicants to the Computer Science department at the University of California, Irvine to train our models. Collected data includes demographics, academic history, high school information, and essay responses. The best-performing classifier was Logistic Regression trained

on a dataset that included all numerical and categorical variables along with extracted keyword bigrams from the text. This classifier obtained the highest accuracy of 0.789. With feature coefficient analysis, we observed the effects of academic achievement, extracurricular involvement and writing content on the model's predictions.

Chapter 1

Introduction

Undergraduate admissions have evolved over the last decades aiming to make selection processes as fair as possible for all applicants. In some universities, including those in the University of California (UC) system, modifications include the following: restrictions to eliminate consideration of race and gender, implementation of percentage plans to highlight top-performing students from each state high school, and removal of standardized testing. Additionally, the UC utilizes holistic review to ensure that reviewers “look at multiple factors beyond courses and grades to evaluate applicants’ academic achievements in light of the opportunities available to them and the capacity each student demonstrates to contribute to the intellectual life of the campus”[1].

At the University of California, Irvine (UCI), undergraduate applications that meet minimum requirements are reviewed and scored by two readers. If the two scores’ difference is over a certain threshold, it triggers a third review by a senior staff in the admissions office. Based on this information, we explored supervised classifiers’ performance with applicant data and scores to perform an additional screening. This screening could provide another layer of review by identifying any other applications that should be reviewed by an additional reader,

[1] https://admission.universityofcalifornia.edu/counselors/freshman/comprehensive-review/

potentially reducing the presence of mistakes or biases. The need for a third review would be triggered when the classifier's prediction significantly differs from the average reviewer score.

Currently, Machine Learning (ML) has been used by some universities to speed up admissions decisions [1; 2; 3], and in others to detect patterns [4]. We build on these findings to investigate their potential to be applied to holistic admissions. Since many universities are implementing holistic review and removing requirements for standardized tests, it is necessary to reduce or eliminate any bias that may arise. For instance, prior research has shown that letters of recommendation often reinforce stereotypes about applicants [5]. ML tools could assist by using classifiers to provide a supplementary review that triggers another human review whenever there is a difference over a certain threshold between the initial readers' scores and the model's prediction. Subconscious human bias could be present, and we aim to assist admissions selection committees to reduce its effects. Nonetheless, algorithmic bias could also be present in the proposed tools, but we argue that this could be controlled in all stages of processing by performing feature analysis and using supplementary tools, such as LIME [6], to explain and understand decision-making processes.

We worked with a dataset consisting of applicants to UCI's Computer Science (CS) department since it is one of the most selective undergraduate programs and presents a very low admission rate. In the last three application cycles, admission rate for CS and related fields, namely, Data Science (DS) and Computer Science and Engineering (CSE), has dropped to under 20 percent on average[2]. Therefore, it is of utmost importance to ensure that all applicants are given a fair chance at being admitted.

We used a total of 4,442 application records of California freshman applicants for the Fall 2021 cycle to the CS department to train our supervised models. The dataset contains demographics, academic history, high school information, and essay question responses. We

[2]https://datahub.oapir.uci.edu/Undergraduate-Admissions-Dashboard.php

extracted varying n-gram keywords from the essay question responses in the dataset to explore their usefulness in training and predicting with high accuracy. We trained Logistic Regression, Linear Support Vector Machines (SVMs), Naive Bayes, Decision Trees, Random Forest, AdaBoost, and Gradient Boost classifiers to analyze and compare their performances with our dataset.

Chapter 2

Background

2.1 Evolution of College Admissions

Proposition 209 prohibited the consideration of race and gender for college admissions and job applications in California [7]. After its introduction in 1996, minority representation in California campuses dropped [8]. In an effort to counteract the detrimental diversity effects of similar restrictions, universities in Texas, Florida and California implemented percent plans.

In 1997, Texas introduced the Top Ten Percent Plan after race could no longer be considered, admitting the top ten percent of students from each Texas high school [9]. The UC introduced Eligibility in the Local Context (ELC) in 2001, guaranteeing admission to the top 4% from each California high school, which eventually rose to 9%. However, "only 35% of the ELC... applicants for the 2012 entering class were underrepresented minorities compared to 37% of the overall pool of applicants" [10]. In other words, programs tailored to increase underrepresented minority representation in higher education were admitting students from these groups at lower rates than admissions without modifications. Florida's Talented 20

Program in 2000 was the most generous plan, guaranteeing admission to the top 20% of high school graduates to one state university [11]. Despite these efforts, Long argues that percentage plans do not achieve desired diversity effects, especially because admission is guaranteed for one of the universities in the system, not necessarily one of the most prestigious ones. After implementing percentage plans, institutions considered other modifications to selection processes, including comprehensive and holistic review.

Highly selective universities have implemented comprehensive and holistic review to consider applicants' socioeconomic status, life obstacles, and difficulty level of courses to analyze success inside and outside the classroom [9]. Additionally, standardized test requirements have recently been removed at some universities[1]. Since this reduces the number of quantitative variables left to analyze, admissions boards may then rely on more qualitative features. Currently, college admissions practices continue to be modified in an effort to improve processes and tackle new challenges.

Few research groups have explored the fitness of ML tools to assist with university admissions to tackle challenges admissions must handle. However, undergraduate admissions programs continue to search for improvements for selection practices. This thesis explores the feasibility of ML tools to assist in holistic review of undergraduate applications.

2.2 Related Work

Literature regarding applications of ML to the admissions process for university programs is limited. Additionally, a significant portion of these studies have focused largely on application review for graduate programs. This is insufficient to understanding their impact on undergraduate admissions as application components of interest vary. For instance, re-

[1] https://www.universityofcalifornia.edu/press-room/university-california-board-regents-unanimously-approved-changes-standardized-testing

search experience, undergraduate GPA, and vacancies at research groups are major data points influencing admission in graduate programs. Since these features are not relevant in an undergraduate setting, current studies' results cannot be generalized to undergraduate admissions. In addition, a large amount of studies focus almost exclusively on numerical features. While these features carry vital information, it is essential for new ML applications in undergraduate admissions to make use of entire applications, including text data that is often omitted in existing literature, to ensure that ML tools uphold holistic review goals.

The models described in the literature are closely related to college application review, and their goals fall into one of three main categories: (a) providing feedback for evolution of university practices, (b) tackling review of increased number of applications, and (c) implementing text analysis of admission essays. Their model performances, goals, and features used are introduced in the following sections as well as in Table A.1.

2.2.1 Contributions to the Evolution of University Practices

To begin, Applications Quest software makes recommendations for graduate admissions decisions that uphold diversity goals [12]. Holistic review is carried out by using clustering techniques, an unsupervised classification method, and selecting applicants from each cluster to reflect their diverse backgrounds. Numerical and categorical features are processed by the software, and text data is not. Instead, text is scored by readers, and that score is then fed to the clustering model. This combination of review by readers and software ensures that all application materials are analyzed before making a decision, reflecting holistic review. Applications Quest's purpose is closely aligned with universities' diversity goals, achieving satisfactory results in graduate admissions.

Likewise, [13] trained Logistic Regression, Random Forest, Gradient Boosting and ADABoost classifiers to predict graduate applicant success in Georgia Tech's Online Master's in Analytics Program while aiming to fulfill holistic review goals. In this instance, the authors extracted features that quantified essays and letters of recommendation instead of having readers score them. These were used in addition to more objective variables, like undergraduate GPA and GRE scores, to train models that could assist with holistic review in a graduate program. Other features included age, native language, and undergraduate major. The best-performing models received an Area Under the Receiver Operating Characteristic Curve (AUROC) score of 0.81. This group is motivated by similar holistic admissions goals, and their success suggests that similar approaches could be explored for undergraduate admissions.

At Roanoke College, [14] used SVMs and perceptrons to predict undergraduate admissions decisions as well as applicants' decision to enroll and achieved 94.57% accuracy. Their goal was to provide the university with valuable information to determine which features play the largest role in determining who should be admitted and who would accept an offer. The training dataset consisted of academic information such as GPA and SAT, as well as expected family contribution, citizenship, and calculated high school score. They also utilized applicants' number of campus visits and interviews to train the models, but they did not make use of any text data from their dataset.

In a different approach, [4] explored the ability for SVMs and linear discriminant analysis to detect hidden patterns in datasets of applicants to a specific undergraduate university. Authors determined that GPA and SAT requirements varied based on applicants' region of residence. Models achieved 95% accuracy in detecting admission decision variations based on region. No other features were explored, highly restricting performance to academic success and standardized testing, which has recently been removed from many state universities due to concerns of equity and access to preparation[2].

[2]https://www.universityofcalifornia.edu/press-room/university-california-board-regents-unanimously-approved-changes-standardized-testing

2.2.2 Speeding Up Review to Handle Increased Number of Applications

Alternatively, [3] trained Logistic Regression, Naive Bayes, Perceptron and Decision Tree classifiers to speed up graduate admissions using data such as test scores, GPA, research experience, and quality of admission essays. Traditional review would be necessary as a pre-training step to score text data. Logistic Regression and Naive Bayes classifiers predicted with 88.75% accuracy.

Similarly, [2] developed a system to speed up review and handle growing numbers of applicants into their Computer Science PhD program. They used Logistic Regression to detect students who were very likely to be either admitted or rejected, allowing reviewers to focus more time on those not detected by the model. The model achieved 87.1% accuracy, and was trained solely on academic achievement. The dataset contained undergraduate GPA, GRE scores and school reputation, and omitted often vital information provided in letters of recommendation and statements of purpose. While this model successfully reduced review time, it was discontinued in 2020 because training data was not updated, and its approach was counterproductive to increasing campus diversity [15].

In addition, [1] trained random forest classifiers to predict academic performance of applicants to UNC Eshelman's doctor of pharmacy program. Using application information like undergraduate GPA, university selectivity, PCAT score, and undergraduate major, the models achieved 77% accuracy. Authors suggested utilizing this tool as an initial filter for academic achievement, while still maintaining a formal analysis of materials that display soft skills, including letters of recommendation and personal essays.

2.2.3 Review of Text Application Materials

The last subsection of the literature describes NLP applications using admissions essay data, which could aid current approaches in incorporating a robust, complete review with models. To begin, studies have found that ML and NLP can accommodate for human bias in qualitative data and predict characteristics of applicants more precisely and efficiently. For example, analyzing writing styles of admission essays, a positive correlation between specific word usage, such as higher number of articles and prepositions, and college academic performance was detected [16]. The authors then argue that admissions essays can be analyzed to predict academic success before students enroll.

Stanford researchers have processed text in college application essays. In 2020, analogical tasks, a word vector evaluation method, were analyzed across data from students from different socioeconomic statuses. In general, models trained on writing from higher income students performed better than those trained on lower income student writing, highlighting the need to control for biases in NLP evaluation tasks [17]. Such limitations are especially important to consider in future developments in the field.

Lastly, [18] instead focused on applicant essay analysis to assess the ability to predict demographic characteristics solely from what applicants wrote. By using Naive Bayes, Logistic Regression, and a deep neural model, they predicted reported household income with as high as 68.94% accuracy, and gender with as high as 79.88% accuracy. With this tool, authors investigated the potential to detect implicit bias that could affect applicants during holistic review.

While these research groups have explored ML applications in admissions, none have considered their ability to assist with holistic undergraduate admissions by evaluating all submitted materials. The majority of the literature focuses on graduate level admissions, and the few applications in an undergraduate setting do not incorporate analysis of all submitted application materials. Significant differences between undergraduate and graduate application requirements highlight the need to further explore these approaches in an undergraduate setting. Given this gap in the literature, we explored the potential for ML applications to assist with holistic undergraduate admissions.

Chapter 3

Design and Implementation

In order to train our models and assess their performance, we obtained and preprocessed information about 4,442 in-state applicants to the CS department for the 2021 school year. We then trained and validated Logistic Regression, Linear SVMs, Naive Bayes, Decision Trees, Random Forest, AdaBoost, and Gradient Boost classifiers to compare their performances. We analyzed feature coefficients after training the classifiers to determine which variables had the largest impact in the top-performing classifier's decision-making processes.

3.1 Dataset Description

The independent variables in the collected dataset consisted of demographic information, academic history, data about high school attended, and responses to four chosen Personal Insight Questions (PIQs) out of eight available options. PIQs are questions that allow students to share information about themselves, their personalities, abilities and experiences[1]. This provides readers and trained models alike with valuable information.

[1] https://admission.universityofcalifornia.edu/how-to-apply/applying-as-a-freshman/personal-insight-questions.html

Demographic information variables were comprised of categories to describe a student's background. These included low-income status, first-generation status, military family relationship status, foster care status and residency status. While the dataset also included information regarding applicant gender and ethnicity, these were omitted throughout model training and testing to meet Proposition 209 criteria.

The dataset also included information regarding applicants' last high school of attendance. These variables included total number of Advanced Placement (AP) tests offered, percentage of students taking AP tests, and state where the student last attended high school. This information could be useful to assess the academic opportunities available to the student.

Multiple features regarding applicants' academic performance in the years leading up to applying to higher education were also provided. Three Grade Point Average (GPA) scores were provided: weighted, unweighted, and uncapped. Unweighted GPA follows the traditional 4.0 scale, while weighted and capped allow for advanced courses to follow a 5.0 scale, with capped limiting the number of courses that can follow the 5.0 scale. In-state applicant records also included a value for ELC, which recognizes the achievements of top-performing students in each high school by providing the class rankings of the top 9% of the student body from each school. Other features included total math courses, AP math courses, and AP tests taken, as well as AP scores for subjects like Computer Science and Statistics.

Lastly, provided information also included awards and involvement in extracurricular activities, allowing the student to highlight achievements inside and outside the classroom. Students could also share information about their experiences and abilities as a whole by responding to four Personal Insight Questions (PIQs) they preferred from the following eight available options:

- PIQ1 - Describe an example of your leadership experience in which you have positively influenced others, helped resolve disputes or contributed to group efforts over time.

- PIQ2 - Every person has a creative side, and it can be expressed in many ways: problem solving, original and innovative thinking, and artistically, to name a few. Describe how you express your creative side.
- PIQ3 - What would you say is your greatest talent or skill? How have you developed and demonstrated that talent over time?
- PIQ4 - Describe how you have taken advantage of a significant educational opportunity or worked to overcome an educational barrier you have faced.
- PIQ5 - Describe the most significant challenge you have faced and the steps you have taken to overcome this challenge. How has this challenge affected your academic achievement?
- PIQ6 - Think about an academic subject that inspires you. Describe how you have furthered this interest inside and/or outside of the classroom.
- PIQ7 - What have you done to make your school or your community a better place?
- PIQ8 - Beyond what has already been shared in your application, what do you believe makes you stand out as a strong candidate for admissions to the University of California?

Table 3.1 displays the amount of times each PIQ was chosen. PIQ 6, which specifically asks about an inspirational academic subject, is the most popular one. PIQ 8, on the other hand, is more open-ended and the least amount of students chose to answer it.

The last available feature is the final read score, which represents the review score awarded to each application. In admissions, two readers assign applications a score ranging from 1 to 3, and their scores are averaged together. In this instance, a score of 1 is the more favorable score, and 3 is the least favorable score that could be awarded.

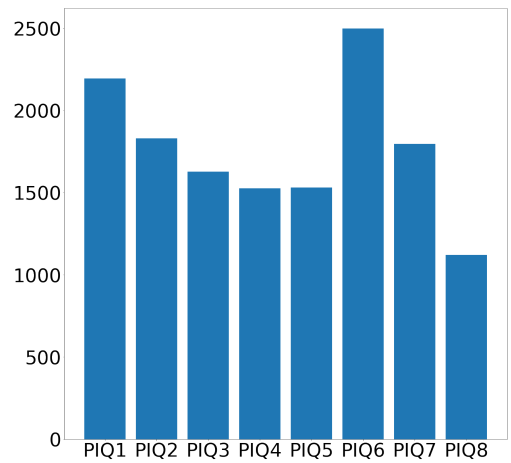


Figure 3.1: Number of times each PIQ was chosen

3.2 Data Cleaning and Preprocessing

For our binary classification task, a review score of 1 was mapped to 1, and less favorable scores were mapped to 0. This was done to determine which applications are expected to receive top scores. This step simultaneously balanced the dataset, with about 50% of records belonging to each class. Records without a read score were dropped because they would not be useful to train models or evaluate their performance.

In addition, the original dataset contained 6,658 applicant records, and it was reduced to 4,534 by dropping anyone whose last high school of attendance was outside of California. We decided to focus on California applicants because their records included a variety of potentially useful information that was not available for out-of-state students. Missing information included multiple variables regarding the high school of attendance as well as a percentage value for ELC program. As mentioned previously, ELC recognizes individual accomplishments of the top 9 percent of students from each high school in light of the opportunities offered by their particular school. It is of utmost importance to consider and explore records that include these features to ensure that the context surrounding a student's educational career are taken into account.

Records of applicants who did not have a GPA score were also dropped. Given that all related work that used numerical data to train classifiers included GPA in their datasets, we concluded that this could be one of the most important features for prediction and later verified our decision during feature coefficient analysis. Some features were dropped as well, notably gender and ethnicity to meet Proposition 209 criteria.

Features that consisted of different possible categorical values were one-hot encoded and transformed into binary columns that flagged the corresponding value of the original features. These features included primary major choice and residency status code. In addition, applicants could list extracurricular involvement in their application, and records included the following three datapoints: awards received, activities that the student participated in, and educational programs that the student was involved with. In order to retain some of the information that could be learned from these datapoints, these features were binarized. In this case, a value of 1 meant that a student had participated in activities or received awards, while a value of 0 meant that the applicant did not list any information regarding extracurricular involvement.

Other features that were dropped from the dataset would not provide any useful information for our current classification task. For example, TOEFL and IELTS exam scores are English language proficiency tests that international students have to take. Since we are focusing on in-state applicants for our approach, these features, along with mostly null columns, could be dropped.

3.3 Keyword Extraction

Students submitted responses to four out of eight PIQs where they could share more information about themselves that could not be reflected in other submitted application materials.

We analyzed key terms mentioned in the essays to gain insight into information discussed in these responses.

In order to obtain a list of topics of interest, we performed unsupervised keyword extraction utilizing YAKE!, which extracts keywords based on local and statistical features [19]. YAKE! follows an unsupervised and corpus-independent approach, making it a good candidate for keyword extraction from our dataset, which is not annotated with relevant keywords. YAKE! works by cleaning the input text, splitting, annotating, and identifying informative candidate keywords while dropping unimportant ones. It allows the user to set the desired n-gram size for extracted keywords, as well as the total number of keywords that should be extracted, and language to discard unimportant stop words in the given language.

For our keyword extraction task, we extracted unigrams, bigrams and trigrams to assess their usefulness in predicting applicants' awarded read scores. For each n-gram category, the top 50 most relevant keywords were extracted from each submitted PIQ response. Extracted candidate keywords were pre-processed by removing punctuation and numbers, tokenizing and stemming with the NLTK library [20]. For each PIQ, all extracted keyword stems were merged into a list along with their document frequencies. Keywords that appeared in under a specific number of responses were dropped, and each list of keywords was manually cleaned. In this step, relevant keywords to the question and to student experiences were kept while less relevant ones were dropped. Table 3.1 shows five sample keywords that were kept and dropped from each PIQ keyword set and for each n-gram category, as well as the general topic discussed in each PIQ. We can observe that YAKE! successfully extracts keywords related to the question topic, but manual cleaning of keywords is still necessary to remove any unnecessary keyword stems that may not add any useful information.

Following keyword extraction and processing, we merged all candidate keywords to generate a list of all possible terms of interest present in applicant writing. In its statistical analysis, YAKE! ensures that stop words are not included at the beginning or end of a keyword. Tri-

Table 3.1: Randomly sampled keyword tokens that were kept and dropped from each PIQ. Five samples were provided for each n-gram category.

	n-gram	Keep	Drop
PIQ 1 - Leadership	uni	respect, resolv, peer, mother, tenni	senior, month, fall, need, person
	bi	group work, band member, honor societi, patrol leader, older brother	past experi, start high, saturday morn, larg group, past year
	tri	pursu computer scienc, red cross club, water polo team, computer scienc class, school key club	decid to make, past two year, univers of california, want to make, year i join
PIQ 2 - Creativity	uni	game, code, algorithm, improvis, canva	past, add, top, sophomor, night
	bi	compos music, rubik cube, play chess, music video, adob illustr	high school, ten year, make thing, everyday life, spent hour
	tri	sheet of paper, leagu of legend, comput scienc principl, make video game, adobe after effect	deepli about press, escap from realiti, elementari and middl, spend countless hour, chocolat chip cooki
PIQ 3 - Talent	uni	design, teach, math, comput, tournament	high, month, now, think, entir
	bi	play guitar, water polo, computer scienc, program languag	start learn, sophomor year, grew older, junior year, start take
	tri	public speak skill, play the clarinet, math and scienc, began to develop, speech and debat	develop thi talent, year in high, open my eye, high school career, enter high school
PIQ 4 - Educational opportunity or barrier	uni	lab, tutor, calculu, lead, rigor	told, lack, watch, simpl, thing
	bi	honor class, colleg class, research paper, neural network, advanc class	school offer, school district, tenth grade, high school, past summer
	tri	comput scienc field, improv my english, html and css, scienc internship program, comput scienc major	start high school, oppportun to attend, middl college high, graduat high school, piqu my interest
PIQ 5 - Overcoming challenges	uni	requir, drop, weight, limit, sever	everyday, think, end, bad, hand
	bi	public speak, english class, social skill, lose weight, robot team	seventh grade, school year, young age, person life, grew older
	tri	immigr to america, time manag skill, diagnos with breast, help my parent, overcom my social	want to make, year in high, began to realiz, remember the day, high school student
PIQ 6 - Inspirational academic subject	uni	arduino, cosmo, function, librari, cultur	place, start, hour, push, chanc
	bi	circuit board, advanc comput, math problem, java program, artifici intellig	everyday life, take class, enjoy learn, call scratch, began learn
	tri	java program languag, python and java, synopsi scienc fair, creat video game, web design class	learn the basic, caught my attent, decid to pursu, subject that inspir, want to make
PIQ 7 - Improving community	uni	stem, sister, club, social, explor	talk, affect, decid, year, start
	bi	local librari, honor societi, eagl project, peer tutor, senior citizen	high schooler, unit state, enter high, school life, learn experi
	tri	boy and girl, link crew leader, scienc olympiad team, interest in stem, summer read program	elementari and middl, coupl of year, make the school, unifi school distri, high school career
PIQ 8 - Other strengths	uni	research, passion, teach, practic, work	stori, find, life, grow, look
	bi	computer scienc, video game, silicon valley, american cultur, social media	daili life, young age, spend time, entir life, junior year
	tri	girl who code, women in stem, leagu of legend, move to america, strong work ethic	univers of california, high school student, make me stand, opportun to learn, candid for admiss

grams could however include stop words in the middle word, so stop words were removed from trigrams prior to merging them with all other extracted unigram and bigram terms. This keyword list was later used as an input vocabulary for vectorization using TfidfVectorizer. In the next section, we introduce all feature extraction steps that were implemented.

3.4 Feature Extraction

As mentioned previously, students choose four PIQs to respond to from a list of eight questions. We used the TextBlob and textstat libraries to extract the following information from each chosen PIQ: character count, word count, sentence count, Flesch Reading Ease, Flesch-Kincaid grade level, polarity, subjectivity, and monosyllable vs polysyllable words ratio.

Flesch Reading Ease formula is based on the number of syllables per 100 words and the average sentence length to assess the readability and comprehension difficulty of a text [21]. Its scores range from 0 to 100, with most readable documents scoring closer to 100. This metric, as well as its predecessors, has been widely used over the last decades to objectively assess the difficulty level of text.

Flesch-Kincaid grade level is based on the same syllable and sentence length values multiplied by different weights [22]. This formula outputs a number that would correspond to the grade level, or total years of education, necessary to understand the text. This metric was initially calculated to determine text difficulty and then measure Navy personnel's grade level of reading comprehension. The same formula could be applied to submitted admission essays to explore relationships between this datapoint and awarded read score.

We used the TextBlob library to perform sentiment analysis and extract polarity and subjectivity. Polarity values range from -1 to 1 depending on how negative or positive the text is, and Subjectivity ranges from 0 to 1, with 0 representing more objective text and 1 more subjective text[2].

Additionally, we utilized the list of extracted keyword stems for a term frequency-inverse document frequency (TFIDF) representation of the text. This not only reduced dimensionality and training time, but it also allowed us to narrow down analysis to terms of interest. All writing components submitted by an applicant were combined, and we used TfidfVectorizer from the Scikit-Learn package for vectorization [23]. With TFIDF, word relevance is calculated by considering word frequency in the evaluated document as well as in the entire corpus, giving more insight into how useful each word can be for each document. After obtaining this information, text columns were dropped, and our dataset was ready to train the classifiers.

3.5 Classification Pipelines

We used Logistic Regression, Linear SVMs, Naive Bayes, Decision Trees, Random Forest, AdaBoost, and Gradient Boost to compare model performances utilizing different text features. We trained classifiers with data not including TFIDF representation of the text, and then including extracted keyword TFIDF representations with unigrams, bigrams, unigrams and bigrams, bigrams and trigrams, and unigrams, bigrams and trigrams together. Before implementing TFIDF vectorization of text, responses were pre-processed by removing stopwords, punctuation and numbers, tokenizing and stemming with the NLTK library [20].

We trained models using 5-fold cross-validation to ensure that the averaged results reflected more realistic performance metrics. For each data split, we implemented a Column Transformer that imputed null values of numerical features with the median. We then added an indicator variable to identify imputed values. Lastly, we scaled values with Robust Scaler.

[2]https://textblob.readthedocs.io/en/dev/

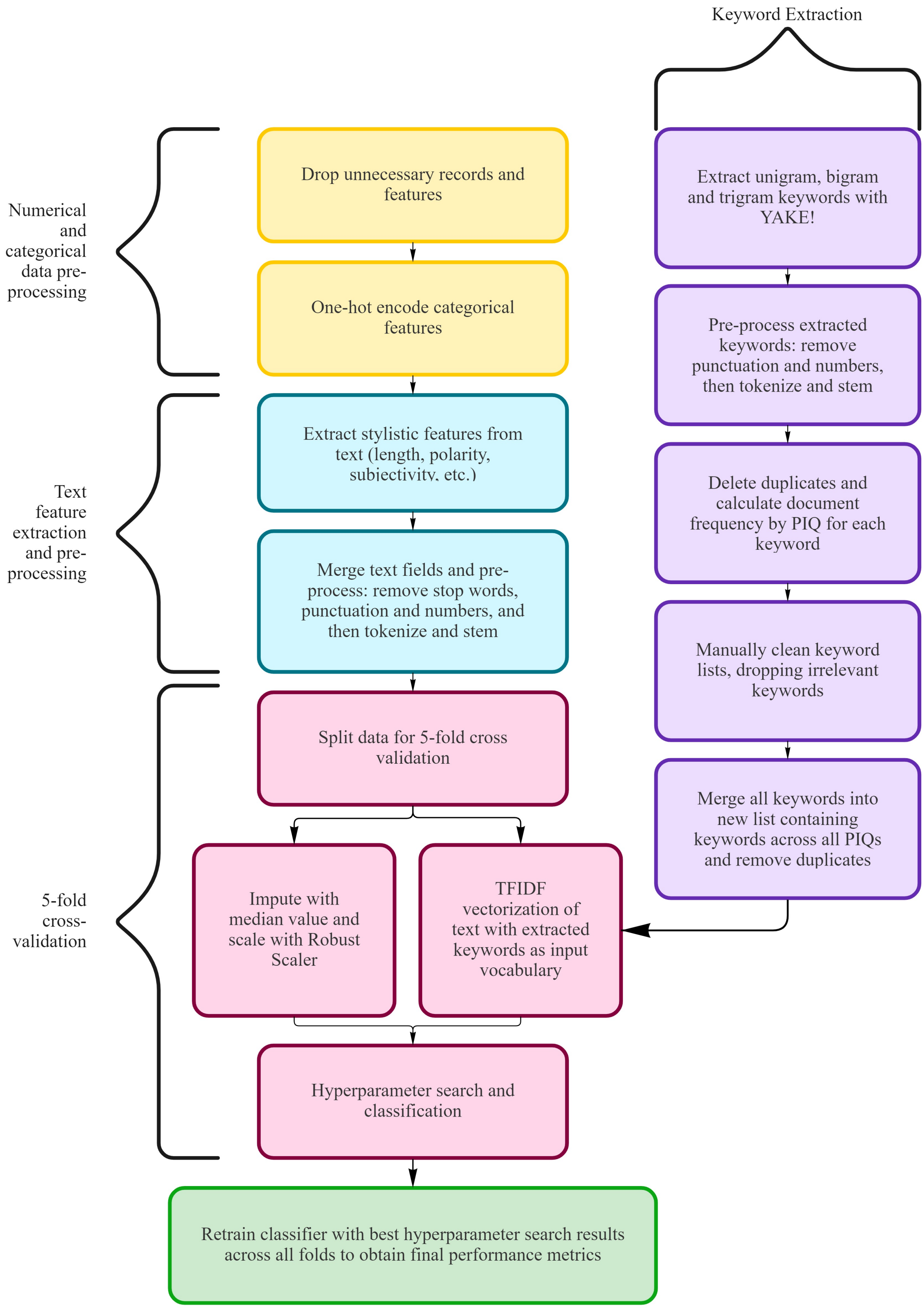


Figure 3.2: Data processing and classification pipeline

We transformed text features with TFIDF to obtain a matrix representation. Next, we performed hyperparameter tuning with GridSearchCV to find the best values for the parameters of the estimator. These steps, illustrated in Figure 3.2, were implemented for all seven classifiers we trained. We used the Scikit-Learn package to implement these steps [23].

3.6 Feature Coefficients

We evaluated the coefficients of the decision function to analyze feature relationships between the input variables and predictions for the best-performing classifier. As will be discussed in the next section, Logistic Regression using bigram keywords achieved the best accuracy. With Logistic Regression, positive coefficients increase the probability of the prediction being 1, and negative coefficients in turn reduce that probability. Feature coefficients will be examined in the discussion section to interpret the Logistic Regression classifier's predictions.

Chapter 4

Results

Performance metrics for all the trained classifiers are shown in Table 4.1, and the highest performance for each metric is bolded. For each classifier, we trained with numerical and categorical features along with text features utilizing the following keywords for vectorization: (i) no text features, (ii) unigrams, (iii) bigrams, (iv) unigrams and bigrams, (v) bigrams and trigrams, and (vi) unigrams, bigrams and trigrams. We calculated multiple relevant metrics and chose to evaluate model performance with accuracy as the primary score and AUROC as the secondary score if accuracies match across different classifiers. Accuracy gives us the percentage of correct predictions. Since the dataset's target class is balanced, it was not prone to give artificially high accuracies, making it a good metric to evaluate the model. Additionally, AUROC measures models' ability to differentiate between two classes, and it served as an effective metric for our binary classification model. The closer the AUROC value is to 1, the better the model is at differentiating between the two classes.

Based on model accuracy, the Logistic Regression classifier trained on numerical features and text features with keyword bigrams achieved the best performance for our classification task, achieving an accuracy of 0.789. Although Gradient Boost classification achieved a slightly lower accuracy of 0.787, it was the best in differentiating between classes across multiple n-gram sets used, achieving AUROC scores of up to 0.874. Additionally, Gradient Boosting classifiers saw the greatest improvements across multiple metrics with the addition of text features using bigram keywords, improving accuracy by 1.4%, precision by 1.4%, recall by 1.6%, F1-score by 1.5%, and ROC-AUC by 0.1%. This tells us that the classifier that makes use of binary keyword data is better at identifying a larger portion of true positive cases effectively.

Across all classifiers, Decision Trees and Random Forest were the only ones that saw a decrease in performance when incorporating keyword features. This could likely be a result of overfitting on the training set when increasing the dataset's dimensionality with the addition of keyword features.

Although precision, recall, and F1 scores were not the primary metrics we analyzed to select the best classifier, these values gave more insight to the model's ability to correctly predict application scores. In fact, the classifiers, especially Logistic Regression and Gradient Boosting, also achieved competitive scores with these metrics across multiple datasets.

Table 4.1: Performance metrics for trained classifiers with data including YAKE!-extracted n-gram keywords.

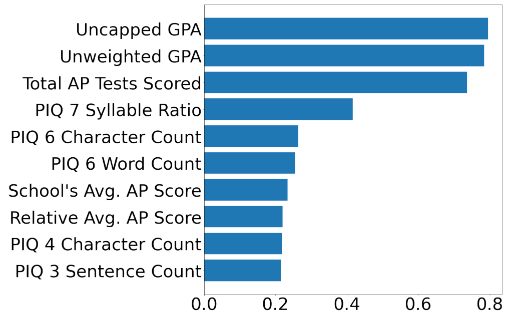
	Keyword n-grams Used	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	None	0.783	0.762	0.802	0.781	0.871
	Uni.	0.783	0.763	0.800	0.781	**0.872**
	Bi.	**0.789**	**0.769**	**0.805**	**0.786**	0.871
	Uni. and Bi.	0.783	0.763	0.800	0.781	**0.872**
	Bi. and Tri.	0.786	0.767	0.802	0.784	0.871
	Uni., Bi. and Tri.	0.784	0.763	0.802	0.782	**0.872**
Random Forest	None	0.778	**0.764**	0.784	0.774	0.859
	Uni.	0.774	0.743	**0.814**	0.777	**0.862**
	Bi.	0.774	0.751	0.798	0.773	**0.862**
	Uni. and Bi.	**0.782**	0.755	0.812	**0.783**	0.861
	Bi. and Tri.	0.757	0.739	0.770	0.754	0.855
	Uni., Bi. and Tri.	0.775	0.751	0.800	0.775	0.855
SVM	None	0.766	0.747	0.781	0.764	**0.864**
	Uni.	0.766	0.751	0.772	0.761	0.847
	Bi.	**0.773**	0.751	**0.793**	**0.771**	0.849
	Uni. and Bi.	0.768	0.753	0.774	0.764	0.849
	Bi. and Tri.	0.772	0.754	0.784	0.769	0.848
	Uni., Bi. and Tri.	0.769	**0.756**	0.772	0.764	0.849
Decision Tree	None	**0.736**	0.710	**0.767**	**0.737**	**0.773**
	Uni.	0.712	0.692	0.730	0.710	0.728
	Bi.	0.703	0.678	0.735	0.705	0.709
	Uni. and Bi.	0.713	0.690	0.740	0.714	0.731
	Bi. and Tri.	**0.736**	**0.711**	0.763	0.736	0.752
	Uni., Bi. and Tri.	0.721	0.710	0.716	0.713	0.730
Naive Bayes	None	0.685	0.652	0.747	0.696	0.773
	Uni.	0.691	0.658	0.751	0.701	0.774
	Bi.	0.695	0.672	0.723	0.697	**0.776**
	Uni. and Bi.	0.691	0.657	0.756	**0.703**	0.775
	Bi. and Tri.	**0.697**	**0.673**	0.728	0.699	0.775
	Uni., Bi. and Tri.	0.69	0.655	**0.758**	**0.703**	0.775
AdaBoost	None	0.767	0.735	0.812	0.771	**0.870**
	Uni.	**0.772**	**0.739**	**0.816**	**0.776**	0.869
	Bi.	0.767	0.735	0.812	0.771	**0.870**
	Uni. and Bi.	0.768	0.736	0.812	0.772	0.869
	Bi. and Tri.	0.767	0.735	0.812	0.771	**0.870**
	Uni., Bi. and Tri.	0.768	0.736	0.812	0.772	0.869
Gradient Boosting	None	0.773	0.754	0.786	0.770	0.871
	Uni.	0.774	0.752	0.795	0.773	0.873
	Bi.	**0.787**	**0.768**	**0.802**	**0.785**	0.873
	Uni. and Bi.	0.784	0.763	**0.802**	0.782	0.872
	Bi. and Tri.	0.771	0.751	0.786	0.768	0.870
	Uni., Bi. and Tri.	0.782	0.761	0.800	0.780	**0.874**

Chapter 5

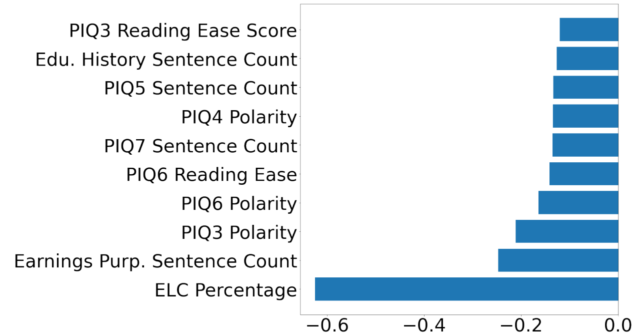
Discussion

We analyzed the features' coefficients with Logistic Regression, which was the most accurate classifier, and used the dataset with TFIDF vectorization using bigram keywords. We did this to explore which features had the highest impact in the decision-making process. Figures 5.1, 5.2, and 5.3 display the 10 features with the largest positive and negative coefficients for numerical, categorical and text features. Feature coefficient values range from -1 to 1, with positive coefficients increasing the log odds of the prediction being 1, and negative ones reducing that probability.

Features related to students' academic performance obtained some of the largest coefficients influencing the probability of predicting one class over another. In Figure 5.1, we can observe that unweighted GPA and uncapped GPA prove to have strong positive coefficients, so higher GPAs can make the model more likely to predict 1. These two features represent applicants' academic performance in high school, and uncapped GPA allows for advanced courses to award one extra grade point while unweighted GPA does not. Similarly, the number of Advanced Placement (AP) exams taken could be related to uncapped GPA, and it also has one of the largest positive coefficients, increasing the probability for the model to predict 1.



(a) Positive coefficients



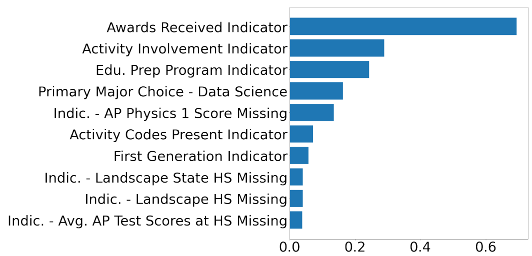
(b) Negative coefficients

Figure 5.1: Top 10 positive and negative largest coefficients for numerical features with logistic regression model and bigram keywords.

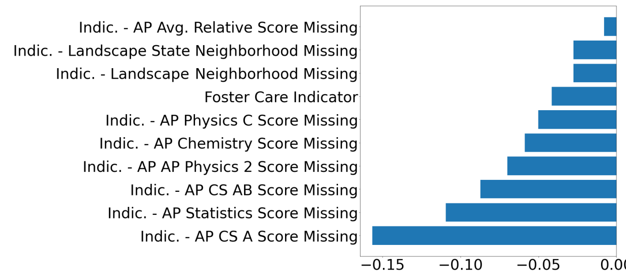
While academic features received large coefficients, it can also be observed that features related to the writing style of some PIQs were also assigned positive coefficients. Specifically, features regarding length of PIQs 3, 4 and 6 were in the top 10 largest coefficients for numerical features suggesting that the classifier preferred applicant records with longer responses for these questions. PIQ6 was not only the most popular response that students selected, but responses were also closely related to academic skills, so the classifier may prefer applications with more in-depth information regarding academic interests.

Interestingly, the strong negative coefficient for ELC reflects the model's ability to detect the significance of individual accomplishments by being in the top 9 percent of each high school's students. For this feature, smaller values indicate better academic performance, so records that have larger ELC values signal being further away from the top-performing group of students, and are more likely to be classified as a 0. Once again features regarding PIQs 3, 4, and 6 had larger coefficients when compared to all negative coefficients. In this case their polarity values had negative coefficients, suggesting that the classifier prefers writing with negative polarity scores.

Looking at the coefficients for categorical features in Figure 5.2, extracurricular involvement also has a large impact on the classifier's predictions as observed with the high coefficients for



(a) Positive coefficients



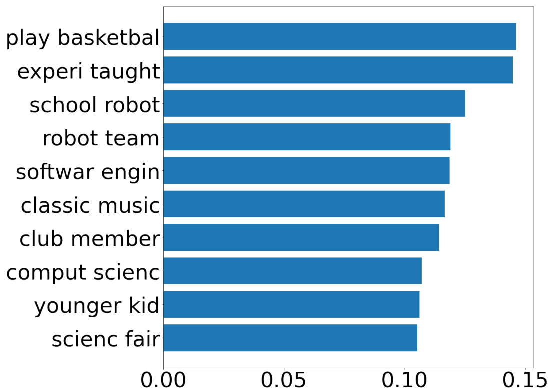
(b) Negative coefficients

Figure 5.2: Top 10 positive and negative largest coefficients for categorical features with logistic regression model and bigram keywords.

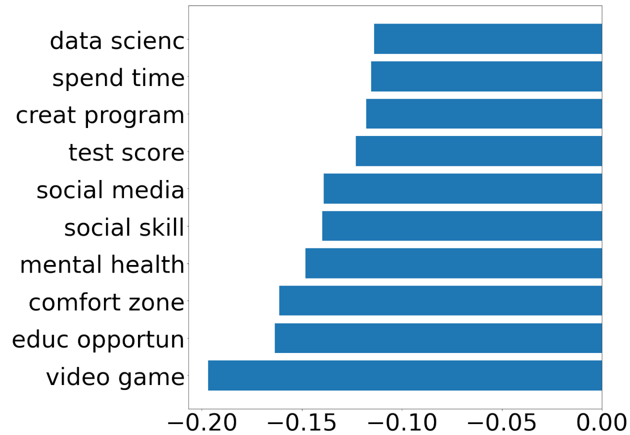
awards, activity involvement and educational program features. It is also interesting to note that the feature signaling that a student's primary major choice is Data Science obtained a positive coefficient, suggesting that students who apply to the Data Science program instead of Computer Science, which is much more popular and selective, are slightly more likely to be classified as a 1. The rest of the categorical features indicate whether certain values were available or not, and their coefficients were close to 0 so they had little impact on the final prediction.

Categorical features with negative coefficients were largely related to AP exam indicator variables added during the imputing step, which signaled that a student did not submit an AP exam score for that specific subject. AP CS and AP Statistics exam indicator variables obtained the largest negative coefficients. This suggests that the classifier picks up on the significance of students having enrolled in advanced courses for related fields during their high school years.

Figure 5.3 displays the bigrams with the largest coefficients, and we can see that the majority of these terms relate to STEM topics and extracurricular involvement. For example, features regarding learning from experiences, playing a sport, school robotics and computer-related topics were associated with positive coefficients. On the other hand, bigrams regarding less academic topics and more related to everyday life were associated with negative coefficients.



(a) Positive coefficients



(b) Negative coefficients

Figure 5.3: Top 10 positive and negative largest coefficients for bigram keyword stems with logistic regression model.

Responses talking about videogames, social media, social skills, and even mental health could slightly reduce the classifier's likelihood of predicting 1. Overall, feature coeffients of bigram keywords were much smaller than those awarded to numerical features. These results might suggest the need of a deeper analysis of text to extract a wider range of potentially more useful features.

Overall, the level of objectivity or subjectivity had little to no effect on the classifier's prediction as was reflected by small feature coefficients assigned to these features. Features that were used to explore whether any significant information could be extracted from individual word length also failed to provide useful information. These features obtained widely varying coefficient values across different essay responses, showing that this may not be a very useful feature to extract from text.

Chapter 6

Conclusion

The present work explores the possibility for ML classifiers to be successfully be developed and assist with holistic admission review for undergraduate programs. The dataset utilized for the classification tasks in this thesis includes features extracted from text material in addition to numerical features, which most previous work was limited to. We achieved the best accuracy in our experiments with Logistic Regression classifiers trained on data including TFIDF vectorization utilizing extracted bigram keywords. We found that academic achievement, extracurricular involvement, ELC recognition, and some writing components play a significant role in this model's predictions. On the contrary, subjectivity and Flesch-Kincaid features had little impact on the predicted values.

Our approach is one of the first to utilize a diverse dataset to train classifiers with the goal of assisting in undergraduate admissions. Our results show the potential for these classifiers to be further developed and improved to serve as an assistance mechanism during review to minimize human error and bias by ensuring that all applicants are reviewed and screened equally with fairness.

Tools based on these trained classifiers could serve as an additional review that screens all applications. In cases where the difference between classifier prediction and readers' awarded score is over a certain threshold, the application could be flagged to undergo another round of review, minimizing any potential errors or biases in applicant scoring.

While the majority of past approaches omitted text data or relied on human review to assign a numerical score to these features, we successfully extracted some information from the text to then train our models. Our models work towards upholding the policies of holistic review, which targets to analyze student performance and the context surrounding their education with all provided materials.

Chapter 7

Future Work

Future goals include further fine-tuning keyword extraction and text processing for PIQ responses, as well as awards and activities' self-reported application entries. These steps could provide even more data to analyze for ML applications. Since YAKE! extracts keywords based on statistical features, performance of other methods of keyword extraction could also be explored. For instance, TextRank and TopicRank perform graph-based keyword extraction, and KeyBERT leverages BERT embeddings to find keywords that are most similar to the text. Overall, this analysis with keyword extraction could serve in exploring new methods of assistance for undergraduate admissions boards that do not depend solely on a classification task.

Additionally, increasing the size of our dataset could aid with expanding analysis to out-of state and international students. Records may contain different data, and model performance should be analyzed. New tools could also be developed to provide campus-wide insights and assistance.

Lastly, we could pursue a multi-class classification approach to give the classifiers the ability to predict a wider range of scores rather than a binary score as with our presented approach.

With this, predictions could potentially give more resolution as to what score a new applicant would be expected to receive. In all, future steps could continue exploring approaches to assist with holistic review of applications to maximize fairness for all applicants.

Bibliography

[1] E. Muratov, M. Lewis, D. Fourches, A. Tropsha, and W. C. Cox, “Computer-assisted decision support for student admissions based on their predicted academic performance,” *American Journal of Pharmaceutical Education*, vol. 81, no. 3, 2017.

[2] A. Waters and R. Miikkulainen, “Grade: Machine-learning support for graduate admissions,” in *AI Magazine*, vol. 35, pp. 64–75, 2014.

[3] “A comparison of classification models in predicting graduate admission decision,” *Journal of Higher Education Theory and Practice*, vol. 21, pp. 219–230, Aug. 2021.

[4] C. Andris, D. Cowen, and J. Wittenbach, “Support Vector Machine for Spatial Variation,” *Transactions in GIS*, vol. 17, pp. 41–61, Feb. 2013.

[5] C. Houser and K. Lemmons, “Implicit bias in letters of recommendation for an undergraduate research internship,” *Journal of Further and Higher Education*, vol. 42, pp. 585–595, Jul. 2018.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “”Why should i trust you?” Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[7] R. C. Johnston, “Black, Hispanic Admissions Rebound at University of California,” *Education Week*, vol. 18, no. 31, 1999.

[8] J. L. Santos, N. L. Cabrera, and K. J. Fosnacht, “Is ”race-neutral” really race-neutral?: Disparate impact towards underrepresented minorities in post-209 UC system admissions,” *Journal of Higher Education*, vol. 81, pp. 675–701, Nov. 2010.

[9] P. Sacks, “The Geography of Privilege.,” vol. 17, no. 1, pp. 6–9, 2004.

[10] G. Orfield, “Alternative Paths to Diversity: Exploring and Implementing Effective College Admissions Policies,” *ETS Research Report Series*, vol. 2017, pp. 1–78, Dec. 2017.

[11] M. C. Long, “RACE AND COLLEGE ADMISSIONS: AN ALTERNATIVE TO AFFIRMATIVE ACTION?,” 2004.

[12] J. E. Gilbert, “Applications Quest: Computing Diversity,” tech. rep., Auburn University, 2006.

[13] S. Staudaher, J. Lee, and F. Soleimani, "Predicting Applicant Admission Status for Georgia Tech's Online Master's in Analytics Program," *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale*, pp. 309–312, Aug. 2020.

[14] T. Lux, R. Pittman, M. Shende, and A. Shende, "Applications of supervised learning techniques on undergraduate admissions data," *2016 ACM International Conference on Computing Frontiers - Proceedings*, pp. 412–417, May 2016.

[15] L. Burke, "The death and life of an admissions algorithm."

[16] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver, "When Small Words Foretell Academic Success: The Case of College Admissions Essays," *PLoS ONE*, vol. 9, Dec. 2014.

[17] N. Arthurs and A. J. Alvero, "Whose Truth is the "Ground Truth"? College Admissions Essays and Bias in Word Vector Evaluation Methods," Jul. 2020.

[18] A. J. Alvero, N. Arthurs, A. L. Antonio, B. W. Domingue, B. Gebre-Medhin, S. Giebel, and M. L. Stevens, "AI and holistic review: Informing human reading in college admissions," in *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 200–206, 2020.

[19] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, Jan. 2020.

[20] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[21] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[22] J. Kincaid, R. Fishburne Jr., R. Rogers, and B. Chissom *Naval Technical Training Command Millington TN Research Branch*, 1975.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] B. Martinez Neda and S. Gago-Masague, "Feasibility of machine learning support for holistic review of undergraduate applications," in *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*, 2022.

Appendix A

Relevant Literature Highlights

Table A.1 introduces relevant literature regarding ML applications in admissions, which fall into one of three categories: (a) providing feedback for evolution of university practices, (b) tackling review of increased number of applications, and (c) implementing text analysis of admission essays. The goals, approach, relevant datapoints, and model performance are presented. Relevant techniques and datapoints could be especially helpful in giving insight into what is relevant to consider when exploring this subject.

Table A.1: Summary of research conducted in applications of ML in college admissions.

	Category/Goals of the ML Tool	**Techniques Used**	**Dataset Size & Description**	**Best Performance**
[12]	Providing Feedback: Assist with holistic review. Recommends applicants for admission consideration	Hierarchical Clustering	N/A; Academic history, demographics, standardized test scores, essay	N/A
[4]	Providing Feedback: Search for implicit patterns	SVM	>10,000; GPA, SAT, geolocation	95%
[13]	Providing Feedback: Predict admission status. Long term goal: predict success in program	Logistic Regression, Random Forest, Gradient Boosting, and ADABoost	9044; Demographics, academic history, standardized test scores, undergraduate university prestige, letter of recommendation tone and complexity	0.81 ROC-AUC
[14]	Providing Feedback: Predict admissions decisions and students' decision to attend	Multi-Layer Perceptron and SVM	About 10,000; Demographics, academic history, standardized test scores, Expected Family Contribution, and number of visits, interviews, and institutional contacts	94.57%, MCC score of 0.89
[2]	Speed up admissions: speed up review of computer science applications	L1 regularized Logistic Regression	1467; GRE scores, GPA, highest degree attained, residency status, research area of interest, previous institutions and reputation, preferred advisor, and letters of recommendation	87.1%
[3]	Speed up admissions: Speed up graduate admissions and analyze past trends	Naïve Bayes, Logistic Regression, Multilayer Perceptron, Random Forest and Decision Tree	400; GRE and TOEFL scores, GPA, Statement of Purpose score, Letter of Recommendation score, research experience flag, and rating of university receiving application.	88.75%
[1]	Speed up admissions: Improve admissions. Consider utilizing as a preliminary filter	Random Forest	1389; GPA, PCAT score, undergraduate major, degree flag, university ranking, and performance in specific classes	77%
[16]	Text analysis: Explore correlations between college application essays and academic success	LIWC and statistical analysis	50,000 application essays; Demographics, academic history, SAT score, rate of use of personal pronouns, impersonal pronouns, auxiliary verbs, articles, prepositions, conjunctions, negations, and common adverbs.	N/A
[17]	Text analysis: Challenge fairness of current vector quality evaluation methods	Intrinsic evaluation tasks to evaluate word vectors	826,624 application essays; College admissions essays, reported household income (RHI)	N/A
[18]	Text analysis: Explore prediction of applicant demographic characteristics	Multinomial Naïve Bayes, Logistic Regression, and Deep Neural Model	283,676 application essays; College admissions essays, gender, and RHI	66% for RHI, and 77% for gender