Data Science College Admissions Data Review
For College Admissions Professionals

Ruchitha Rajaghatta (rsr262)
Julia VanPutte (jhv42)
Emmett Breen (ejb284)

Table of Contents

Abstract

In this report, we explore how standardized testing, demographics, and applicant characteristics shape the college admissions process, utilizing data from the National Center for Education Statistics. First, we analyzed how the pandemic affected standardized testing. It was found that because a large number of universities have switched to test optional, standardized testing is on the decrease, and this decrease is forecasted to continue into the future. Second, we researched the role that demographics will continue to play in college admissions. We found that although overall college enrollment is decreasing for both males and females, enrollment of minority groups is on the rise and is forecasted to continue. Lastly, after looking specifically at Cornell University's admissions over the last three years, machine learning models found that having a unique trait in your application was the most influential factor in admissions, rather than SAT, GPA, volunteering, leadership, or athletics. Our findings can be used by college admissions professionals to understand the decrease of standardized testing, the increase in minority enrollment, and the impact of having a unique trait in a college application in order to follow the trends of other universities across the country.

Introduction

We have analyzed the dataset provided by the National Center for Education Statistics on Postsecondary Education data (NCES, 2020). This database contains university statistics involving standardized testing scores, demographics, and acceptance rates. In addition, we have created a smaller database of scraped data from college confidential containing more specific characteristics of students. We have decided to ask three main questions:

- **Descriptive**: How did the pandemic affect standardized testing in college admissions?
- **Predictive**: What role will demographics continue to play in college admissions and enrollment?
- **Prescriptive**: What specific features of college applications are most influential in the admissions process at Cornell University?

These questions are specifically interesting to college admissions professionals, and through our data analysis we intend to give these professionals significant data-driven insights to guide decision making in the future of admissions. Particularly, recent events such as the COVID-19 pandemic have caused a shift in standardized testing's role in college admissions. It is important for admissions professionals to see how universities across the country have responded to fewer administered standardized tests, in addition to the institutions beginning to question how important these tests really are. Also in recent years, universities have been continually striving to make their admissions process more holistic, and our exploration of both demographics and application features provides professionals with insight into what types of applicants have a better chance at acceptance. This analysis can reveal the makeup of enrolling classes, which are important for admissions professionals to understand in order to avoid potential biases. We address how the choices made by college admissions professionals in crafting their admissions process affects prospective applicants based on their individual circumstances (income, demographic, access to academic resources) to all universities in the United States. In addition, we analyze how different application traits (volunteering, athletics, leadership) affect admissions decisions at Cornell University. Due to limited information regarding individual applications to colleges, we narrowed down our analysis of specific applicants to only Cornell University since all data needed to be manually scraped and cleaned.

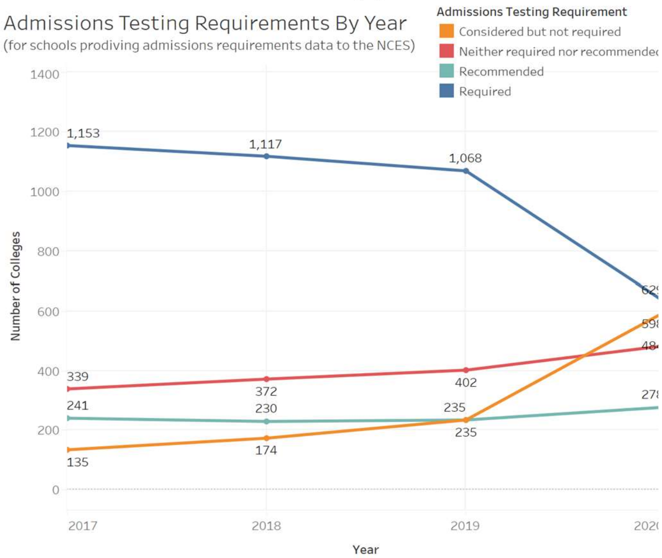
Question 1: How did the Pandemic Affect Standardized Testing in College Admissions?

It is well known that the widespread national COVID-19 pandemic has significantly affected the operations in the U.S. for the past two years. Because of factors such as social distancing, protecting the health of Americans, and the risk factors of COVID-19, this question was proposed to explore how the COVID-19 pandemic affected the admissions processes of postsecondary degree-granting institutions in the US, and how these national changes to college admissions process will affect admissions requirements in the future.

Analysis Concerning Admissions and Standardized Testing Requirements in 2020

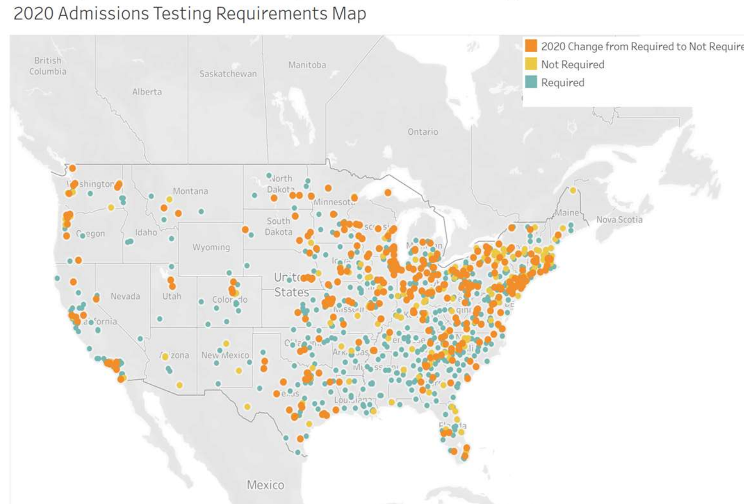
To answer this question, we conducted an analysis of the postsecondary admissions process data for 2019, before the COVID-19 pandemic, and for 2020, after the COVID-19 pandemic. *Figure 1* was constructed through filtering for the schools that provided admissions testing requirements data to the NCES and performing aggregation and data cleaning techniques in SQL. Then, the data was plotted to easily see the trends in schools' testing requirements by year. A drastic change is visible: in 2019, 44.95% of schools did not require test scores and 55.05% did require them. In 2020, this changed to 68.38% not requiring test scores while 31.62% of schools still required them. This can further be explored geographically in *Figure 2*.

Figure 1: Admissions Testing Requirements for US institutions by year



A linear plot of the aggregate number of degree-granting U.S. institutions grouped in color by the institutions' application requirements for standardized test submission plotted against year of application cycle. (NCES, 2020)

Figure 2: 2020 Testing Requirements Map



A tableau map of the aggregate number of degree-granting U.S. institutions grouped in color by the institutions' application requirements for standardized test submission plotted against year of application cycle. (NCES, 2020)

This drastic change in testing requirements can be further explored in *Figure 2*, a map of the US degree granting institutions' change in admissions requirements in 2020. Universities were mapped and grouped by their admissions requirements as a result of the COVID-19 pandemic using Tableau. This data was initially aggregated using SQL as well to control for outliers and determine the change in requirements. *Figure 2* reveals that many schools did not require standardized test scores in 2020.

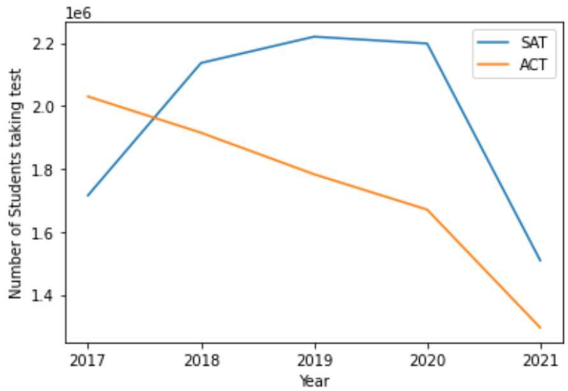
The majority of schools that switched from requiring scores in 2019 to to not requiring scores in 2020 (orange dots) are located in the northeast of the US. This is likely due to the difference in approach that each state took towards COVID-19. Institutions in the northeast are very progressive and took an aggressive approach to controlling the spread of COVID-19. Institutions in the south and the west did not.

Forecasting Standardized Testing Popularity

With the change in schools' requirements of standardized tests in 2020 due to the pandemic, the question arises: are these scores necessary for students to submit in their applications? It is well known that standardized tests have been criticized as biased and unreliable indicators of student success in the long term. Because of this widely known criticism, the pandemic's change in testing requirements and testing popularity is interesting for analysis. *Figure 3* shows the number of students taking standardized tests annually has decreased in both 2020 and 2021, which is significantly determined by the change in testing requirements for schools. This plot was generated through data cleansing and filtering in SQL and then using python notebooks for further analysis.

Figure 3: Students taking Standardized Tests (SAT and ACT) per Year

(NCES, 2020)



A tableau-generated plot of the aggregate number of degree-granting U.S. institutions grouped in color by the institutions' application requirements for standardized test submission plotted against year of application cycle.

In *Figure 4* and *Figure 5*, various forecasting models were run on the SAT and ACT dataset in python to attempt to forecast the future numbers of students taking the SAT and ACT annually, taking into account the decrease in number of test takers for the past 2 years.

For the SAT forecasting in *Figure 4*, the Holt model and the Holt Winters model, with no trend and a multiplicative trend, were fitted to the data. Simple Exponential Smoothing was also run but due to an inaccurate prediction, this model has been discarded from the visualization. For the ACT forecasting in *Figure 5*, 4 forecasting models were fit to the data: Simple Exponential Smoothing, the Holt Model, and the Holt Winters model, with no trend and with multiplicative trend. These forecasting models and plots reveal that the number of students taking these tests has decreased significantly in 2019 and 2020 and most of the fitted models predict a continually decreasing or constant number of students taking these tests in the future.

Figure 4: SAT Forecasting

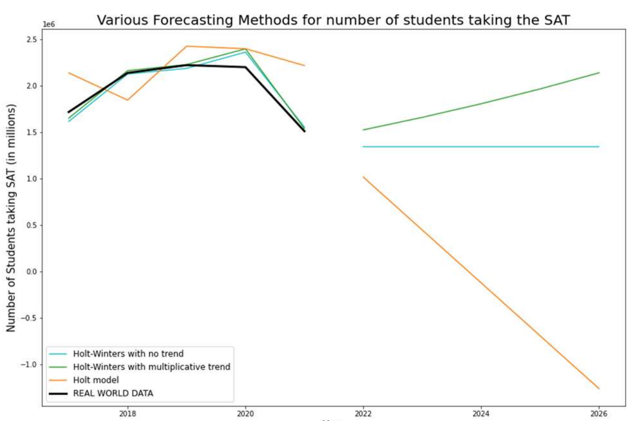
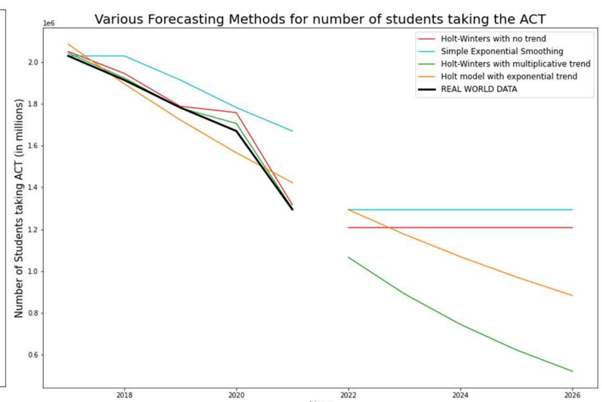


Figure 5: ACT Forecasting



Two plots of fitted and forecasted data, for the number of students taking the SAT (left) and ACT (right) annually. The plot was generated from 2017-2021 data and forecasts 2022-2026 using various models. (NCES, 2020)

Due to the constraints from the NCES database, we were only able to access annual data for the number of students taking the SAT. Therefore, the forecast is limited in its accuracy. However, the general trend in the decreasing forecasts is shown in the above plots. Specifically to the SAT plot, the forecasts from different models disagree on whether the future will see an increase or decrease in the number of test-takers. This leads to uncertainty that should be taken into account when interpreting results. Contrarily, the ACT forecast results are highly accurate, especially the Holt-Winters model with a multiplicative trend (shown in *Figure 5* in green).

Findings and Conclusions

Because of these limitations, these forecasts are not 100% accurate, but it can be seen that due to the pandemic limiting the number of testing centers and changing many university's admissions requirements, fewer students are taking standardized tests each year. Standardized tests have been used to determine postsecondary admissions in the United States since 1900 (National Education Association, 2020). It would be unlikely that the COVID-19 pandemic has a lasting effect on the testing process in the US, but the pandemic may have paved the way for other factors that will continue to decrease standardized testing, such as the general public questioning standardized testing's importance. Future analysis is recommended as future data becomes available, but it has been shown and forecasted that standardized testing will continue to play a smaller role in college admissions, since fewer universities are requiring standardized testing and fewer students are taking it.

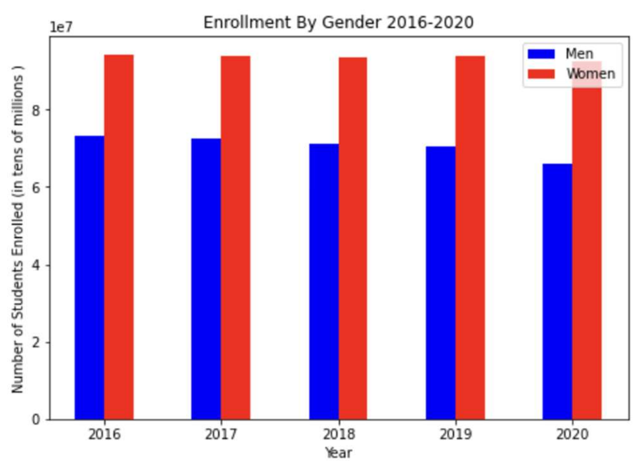
Question 2: What Role Will Demographics Continue to Play in College Enrollment?

As college admissions evolve and admissions offices boast "holistic" admissions decisions, demographic factors like race and gender become key components. Especially as professionals try to craft a well-rounded class of admitted students, the question arises: how do race and gender affect enrollment numbers?

Examining the Relationship Between Gender, Race and College Enrollment

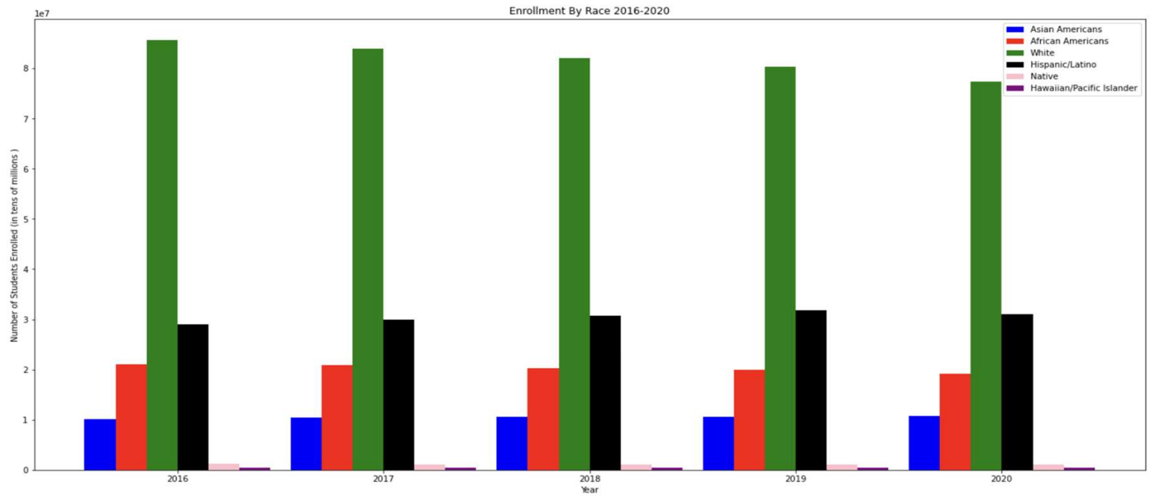
In order to examine this relationship, we accessed the NCES database, pulling a dataset containing post-secondary enrollment numbers by sex and race from the years 2011 to 2020. From this set, we extracted the data for the most recent 5 years (2016-2020) with the following categories: Male and Female for Gender and Asian, African American, White, Hawaiian/Pacific Islander, Hispanic/Latino, and Native American for races. After obtaining this data, we created multiple bar graphs to examine trends in enrollment based on these categorical variables, first by gender, shown in *Figure 6*, and then by race, shown in *Figure 7*.

Figure 6: 2016-2020 Enrollment, Male and Female



Comparative bar graph of male and female enrollment from 2016-2020.

Figure 7: 2016-2020 Enrollment, Divided by Race



Comparative bar graph of Asian, African American, White, Hawaiian/Pacific Islander, Hispanic/Latino, and Native American enrollment numbers between 2016-2020

Analysis of Relationship

Examining *Figure 6*, one clear observation can be made: female enrollment historically exceeds male enrollment by an approximate difference of 15 million students, and this difference stays relatively constant. Furthermore, between 2016-2020, male enrollment seems to slightly, but steadily decline. From *Figure 7*, it's clear that white enrollment far exceeds the enrollment of other minorities, exceeding double the next largest enrollment by race, which is Hispanic/Latino, depicted as the black bar in *Figure 7*. The slight, steady decline of the green bars indicates that white enrollment has been experiencing a decreasing trend, inverse to the minorities, all of which have been steadily increasing during the 5 year period. Given these trends, we elected to forecast the trends for enrollment by gender and see the predicted numbers for the next 5 years.

Forecasting Enrollment By Gender

We used a Holt Linear Model since the data shows a relatively linear trend, which can be seen in the blue lines in *Figure 8* and *Figure 9*.

Figure 8: Male Enrollment Forecasting

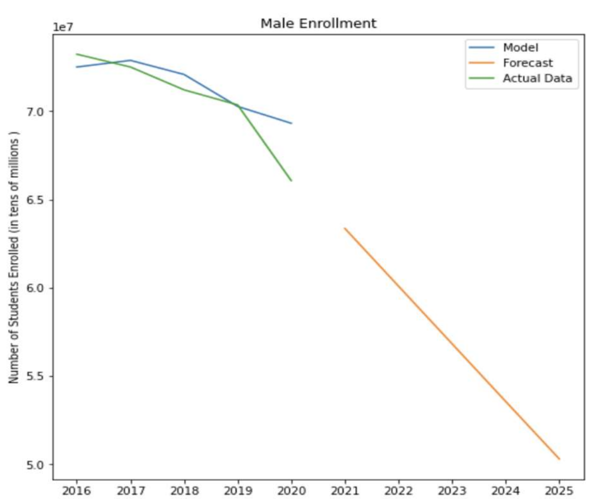
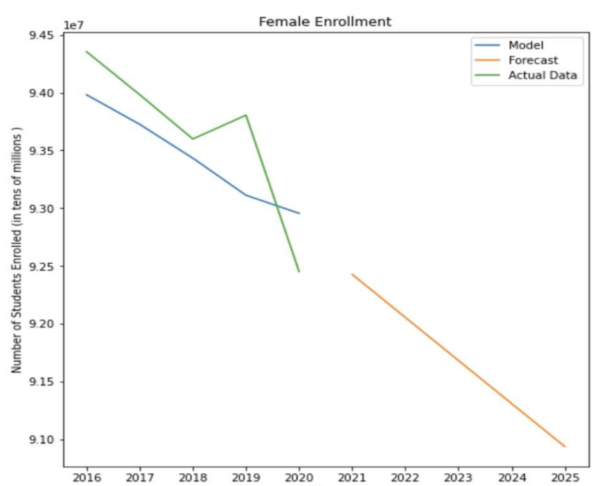


Figure 9: Female Enrollment Forecasting



Two line plots of enrollment data by gender, showing a steady decline forecasted using the Holt model (NCES, 2020)

Findings and Further Examination

We can draw a few basic conclusions from the above data visualizations. The first bar graph demonstrates that female enrollment is higher than male enrollment, and both are declining, with a more visible decline in male enrollment. Examination by race determined that white enrollment is decreasing as minority enrollment increases. Forecasting the next 5 years on enrollment by gender reveals a predicted decline for both males and females, with the male decline having a steeper declining slope than the female prediction. Overall, we also assessed the data for limitations. We do understand that the enrollment slightly differs from admissions as it isn't the accepted class, but rather, those who choose to enroll after acceptance. Therefore, the data doesn't show trends in admissions decisions, but the trend in those actually attending, which is still relevant to admissions professionals who want to ensure a high yield for those they admit. Representatives trying to create a well-rounded class reflective of the student pool must take these trends into account. Especially by gender, female enrollment is higher than males. Consequently, women should make up at least 50% of an admitted class if not more to reflect appropriate proportions. This is especially important to ensure gender equity in all such decisions. Furthermore, as white enrollment decreases and minority enrollment increases, the proportions of these races change and representatives should account for this change as well.

Question 3: What Specific Features of College Applications are Most Influential in the Admissions Process at Cornell University?

The difficult nature of finding a dataset with specific applicant features forced us to narrow our exploration to one institution to decide what features of an application have the largest effect on admissions. As an admissions officer, it is important to understand what features other schools look for in their applicants.

Deciphering what Features Most Effect Admissions at Cornell University:

To begin, we created a dataset composed of 75 data points manually scraped from college confidential's admissions data for Cornell in the last 3 years. The posts on the online forum were compiled into a dataset with the following features: SAT, unweighted GPA, whether they play a sport, volunteer on a monthly basis, are the head of a club, and have a unique trait in their application. The unique trait is partially subjective, although it was narrowed down to the following types of applicants: national champions / first generation college students / founders of large organizations, startups, or freelance work / particularly outstanding essays.

After collecting this dataset, we then wanted to understand which feature affected college admissions the most. As a first step, we trained a logistic regression model. This helped us understand the overall influence of each feature, as the statsmodels python package returns a summary of the data, including the p-values of each feature, as seen in *Figure 10*:

Figure 10: Logistic Regression Model Summary of Features and Respective P-Values

	coef	std err	z	P>|z|	[0.025	0.975]
const	-13.9568	5.743	-2.430	0.015	-25.213	-2.701
x1	0.0051	0.003	1.683	0.092	-0.001	0.011
x2	1.0714	1.253	0.855	0.393	-1.384	3.527
x3	-0.3217	0.725	-0.444	0.657	-1.742	1.099
x4	0.5878	1.076	0.546	0.585	-1.521	2.696
x5	1.5010	1.237	1.213	0.225	-0.924	3.926
x6	1.8008	0.657	2.740	0.006	0.513	3.089

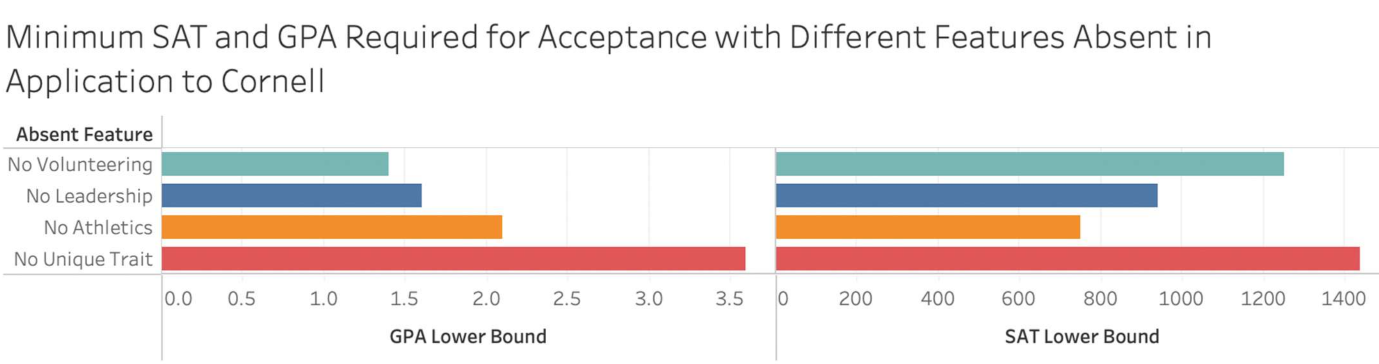
The model summary ordering is: GPA, SAT, playing a sport, volunteering, leadership, and having a unique trait. The column P >|z| corresponds to the respective p-value. The constant row is not relevant to our analysis.

The logistic regression output shows that the only p-value that is below .05 (for the standard 95% confidence level) is x6, which corresponds to having a unique trait. This means that the logistic regression model indicates that this is the only feature that is significant in prediction admission to Cornell.

Due to the linear limitations of logistic regression, we decided to further explore our data using the K Nearest Neighbors (KNN) classifier, a more complex model that can make more accurate predictions about admissions through sophisticated nonlinear boundaries. This model creates a distance metric between points and returns the most common result of the k closest training points to the test point. A train / test split was used to randomly split the data into 63 training points and 12 test points. After training, we input the test points (but not their corresponding admissions decisions) into the model. It was found that with k=5, the model obtained 100% test accuracy, predicting every admission / rejection correctly. Therefore, the more complex classifier we constructed does indeed generalize our dataset and can be used for Cornell admissions predictions and analysis.

Our trained and tested KNN model was then used to find the lower bound of both SAT and GPA for perfect applicants by slowly decreasing the respective statistic until the classifier predicted that they would be rejected. This process was then repeated with almost perfect applicants, who were missing one of the following features: athletics, volunteering, leadership, and unique traits. The lower bounds of both SAT and unweighted GPA for these almost-perfect applicants to be accepted are seen in *Figure 11*:

Figure 11: Lower Bounds for Acceptance to Cornell



The 5-NN boundaries between accepted / rejected applicants for SAT and GPA for "almost perfect" potential Cornell applicants. (Student-constructed K-NN classifier, Manually scraped dataset)

Analysis of Application Characteristics on Lower Bound of Acceptance:

First, draw attention to the volunteering, athletics, and leadership bars. Out of these three, volunteering increases the minimum GPA the least but increases the minimum SAT the most.

Inversely, athletics increases the minimum GPA the most but the SAT the least. Lacking leadership increases the lower bounds in between those for lacking volunteering and athletics. Therefore it cannot be concluded that not having one of these particular traits makes it much more difficult to get into Cornell than not having one of the others. However, the unique trait GPA and SAT minimum *significantly exceeds* the leadership, athletics, and volunteer minimum. Therefore, it is clear that not having a unique trait makes acceptance into Cornell the most difficult out of the applicant features sampled. Our model shows us that the most important applicant feature considered is a unique trait, because of the higher lower bounds when it's absent.

It is important to also notice the unrealistically low values obtained for the SAT and GPA thresholds. This is due to the fact that there are no applicants in the dataset with all perfect stats and a 2.0 GPA or 900 SAT, because realistically applicants with all perfect statistics except for one extremely low one do not exist. It is for this reason that the *numerical* value obtained for the lower bound is not meaningful. Rather, it is the *comparison* of these lower bounds with one another generated by an unbiased, generalizable model that has great significance.

Cornell University's Admissions Approach:

Cornell University is one of many selective colleges known to have strict numerical cutoffs for SAT and GPA. However, our data exploration found that even universities like Cornell take a more holistic admissions approach. This information was clearly backed by two machine learning models that were used to analyze how having certain features in your application will impact your chances of admissions. The logistic regression showed that the only statistically significant feature affecting admissions was the unique trait. Then, the KNN model generated lower bounds for SAT and GPA for applicants with and without leadership, volunteering, athletics, and a unique trait. The absence of a unique trait caused the highest lower bounds for SAT and GPA, meaning that it is the hardest to get into Cornell when the absent feature is a unique trait. As an admissions professional, it is important to notice that even at institutions like Cornell University there are not strict numerical cutoffs in the admissions process. Although limited access to data restricted our model, we claim that the inferences gained from our classifier can be extended as insight into other highly-selective universities.

Recommendations and Conclusions

Overall, the college admissions process is extremely complicated. Despite this, our data analysis has provided admissions professionals with insight into how standardized testing, demographics, and application characteristics are considered, and how recent trends can help us make predictions about the future. First, standardized testing is on the decrease, and the pandemic has accelerated this decline. A large number of universities have switched to test-optional, and it is likely that many more will follow suit, and fewer and fewer students are taking standardized tests each year. This leads to the conclusion that standardized tests are becoming less important in the admissions process. Second, demographics is definitely considered in admissions and our forecasts show that both male and female enrollment is overall decreasing while minority groups' enrollment is on the rise. This is important to consider as diversity efforts and gender equity are essential to create a well-rounded class of admitted students. Lastly, after zooming into Cornell University's admissions process, our machine learning algorithms found that the most influential factor in admissions is having a unique trait. This reveals that even top universities in fact use a holistic approach to admissions and don't have strict numerical cutoffs in their decision making. This same idea can be seen as a model for other selective universities.

It is important for admissions professionals to understand the decrease of standardized testing, the increase in minority enrollment, and the impact of having a unique trait in a college application in order to follow the trends of other universities across the country.

References

Cornell Ed class of 2022 results. College Confidential Forums. Retrieved March 29, 2022, from https://talk.collegeconfidential.com/t/cornell-ed-class-of-2022-results/1956793.

Cornell ED class of 2024 results. College Confidential Forums. Retrieved March 29, 2022, from https://talk.collegeconfidential.com/t/cornell-ed-class-of-2024-results/2075303

Cornell RD class of 2023 results only. College Confidential Forums. Retrieved March 29, 2022, from https://talk.collegeconfidential.com/t/cornell-rd-class-of-2023-results-only/2045435

National Center for Education Statistics. (2020). Use the data. The Integrated Postsecondary Education Data System. Retrieved March 31, 2022, from https://nces.ed.gov/ipeds/use-the-data

National Education Association. (2020, June 25). *History of standardized testing in the United States*. National Education Association. Retrieved March 29, 2022, from https://www.nea.org/professional-excellence/student-engagement/tools-tips/history-standardized-testing-united-states

Enrollment trends by race/ethnicity and gender. National Center for Education Statistics. Retrieved May 8, 2022, from nces.ed.gov website: https://nces.ed.gov/ipeds/SummaryTables/report/270

Appendix A: Logistic Regression and K-Nearest Neighbors Models

Source code:
https://colab.research.google.com/drive/1C0p4OJArit1dkFDFTScAd9RIdcA6xjOu?usp=sharing

```
Welcome to the Cornell Admissions Simulator!


This simulator uses a k-nearest neighbors binary classification machine learning algorithm

How it works: The algorithm takes in a series of features and finds the k most similar training points
 to it and returns the most common admissions decision of these points

Now that you know how it works, how many neighbors do you want the classifier to use to make its decision?

3

Now, time to fill out your application!

SAT: 1450

Unweighted GPA (max 4.0): 3.85

For the following questions, 1 = yes, 0 = no

Do you play a sport: 0

Do you volunteer on a monthy basis: 1

Are you the head of any clubs or organizations: 1

Do you have a unique trait that very few applicants have?: 1


Your application was most similar to the following training points:
[55 54 19]


Congratulations!!! Your application resembles many accepted Cornell students
```

A working example of the KNN classifier interface on a Jupyter Notebook. After being trained with data from College Confidential, it takes in a user input test point. The classifier returns its prediction and the indices of the points in the training data most similar to the test point.

For further information, we have attached the source code for the knn classifier in the form of screenshots below:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# KNN FUNCTIONS
def l2distance(X,Z):

    n,d1=X.shape
    m,d2=Z.shape

    S = np.repeat(np.reshape(np.diag(X @ X.T), (n,1)), m, axis = 1)
    R = np.repeat([np.diag(Z @ Z.T)], n, axis = 0)
    G = X @ Z.T

    D = np.sqrt(S - (2 * G) + R)
    return D

def findknn(xTr,xTe,k, inp):
    D = l2distance(xTr, xTe) # n x m

    indices = np.argsort(D, axis = 0)
    dists = np.sort(D, axis = 0)
    if inp == "user":
      print("\n\nYour application was most similar to the following training points:")
      print(indices[:k,0])

    return indices[:k,:], dists[:k,:]

def knnclassifier(xTr,yTr,xTe,k, inp):
    yTr = yTr.flatten()
    I,D = findknn(xTr,xTe,k, inp) # k x m
    m,d = xTe.shape
    preds = []

    for i in range(m):
        nn_xTe = I[:,i]
        labels, counts = np.unique(yTr[nn_xTe], return_counts=True)
        preds.append(labels[np.argmax(counts)])

    return preds

def classify(use,k):

  # Test point processing
  if use == "user":
    print("Welcome to the Cornell Admissions Simulator! \n\n")
    print("This simulator uses a k-nearest neighbors binary classification machine learning algorithm\n")
    print("\nNow, time to fill out your application!")
    inp = [''] * 6
    inp[0] = input("\nSAT: ")
    inp[1] = input("\nUnweighted GPA (max 4.0): ")
    print("\nFor the following questions, 1 = yes, 0 = no")
    inp[2] = input("\nDo you play a sport: ")
    inp[3] = input("\nDo you volunteer on a monthy basis: ")
    inp[4] = input("\nAre you the head of any clubs or organizations: ")
    inp[5] = input("\nDo you have a unique trait that very few applicants have?: ")
    xTe = [0.0] * 6
    for i in range(len(xTe)):
      xTe[i] = 500.0 * float(inp[i])
      if i == 0:
        xTe[i] /= 500.0
    xTe = np.array(xTe)
    xTe = xTe.reshape((1,6))

  if use == "test":
    test_data = pd.read_csv('test_data.csv')
    xTe = test_data[['SAT', 'GPA', 'Sports', 'Volunteering', 'Leadership', 'Unique_Trait']]
    xTe['GPA'] = xTe['GPA'] * 500.0
    xTe['Sports'] = xTe['Sports'] * 500.0
    xTe['Volunteering'] = xTe['Volunteering'] * 500.0
    xTe['Leadership'] = xTe['Leadership'] * 500.0
    xTe['Unique_Trait'] = xTe['Unique_Trait'] * 500.0
    xTe = xTe.values
    yTe = test_data['Accepted'].values
```

```python
# Training data processing
df = pd.read_csv("training_data.csv")
xTr = df[['SAT', 'GPA', 'Sports', 'Volunteering', 'Leadership', 'Unique_Trait']].values
yTr = df['Accepted'].values

for i in range(len(xTr)):
  for j in range(len(xTr[0])):
    xTr[i][j] = 500.0 * float(xTr[i][j])
    if j == 0:
      xTr[i][j] /= 500.0


# Run KNN
preds = knnclassifier(xTr,yTr,xTe,k,use)


# Descision
if use == "user":
  if preds[0] == 1:
      print("\n\nCongratulations!!! Your application resembles many accepted Cornell students\n")
      print("The KNN classifier predicts your application will be: Accepted\n\n\n")
  else:
      print("\n\nUnfortunately, your application matches that of rejected applicants :(\n")
      print("The KNN classifier predicts your application will be: Rejected\n\n\n")
if use == "test":
  classified = 0
  for i in range(len(preds)):
    if preds[i] == yTe[i]:
      classified += 1

  print("The KNN algorithm has: " + str(100.0 * classified / len(preds)) + "% accuracy on the test data\n\n")
```

```python
classify("test",6)
```

```
The KNN algorithm has: 100.0% accuracy on the test data
```

Note that the KNN algorithm had a 100.0% accuracy on the test data