Elite College Admissions

Renee Aziz, Sarah Chon, Ebube Obi-Okoye

Introduction

The college admissions process is undoubtedly one of the most crucial points of a student's academic journey. Despite the significance and care this process requires, it is often difficult for students to wholly understand what the college admissions process looks like. College admissions is complex, with multiple factors influencing a student's decision to apply. Factors include academic rigor, extracurricular activities, financial considerations, courses, and personal preferences. The ambiguity surrounding college admissions sparked our interest in diving deeper into the topic that once felt like the most pivotal moment of our lives.

While some have universities in mind starting freshman year of college, for most, the research process does not begin till junior year. So much time and energy is put into researching the many college options available before the list is finalized. This issue resonated with us personally, and this was a common sentiment amongst other students who had gone through the process. Thus, for this analysis, we attempted to predict whether someone will apply to a university. We believed this would be an interesting question to explore because students, parents, and educators are constantly seeking to gain insights into the process to make well-informed decisions.

For this project, we began by performing some exploratory analysis on our dataset. We analyzed trends in the data and features that we believed would be important before we set out to focus on predicting rel_apply. After this initial analysis, we cleaned the dataset to ensure it was ready for modeling, removing anything that could affect model performance. Once our data was in the adequate format, we built a linear regression model. This model allowed us to see the most significant features in the dataset. Next, we fit a random forest model, and then an XGBoost model to further refine our predictions. Using the XGBoost model, we visualized feature importance to understand which variables contributed most to the prediction of rel_apply. We also analyzed feature importance and SHAP values from the XGBoost model. Finally, we evaluated the accuracy of our model using root mean square error (RMSE) as our performance metric.

Related Work

A lot of research has been done on this area; however, most seek to answer whether or not a student will be accepted into a specific college rather than whether or not the student will apply. One of the most recent college admissions cases was Chetty et. al,[1] the lawsuit against Harvard for unfair discrimination against Asians. In this case, expert testimonies were given for and against this case in an economical and statistical perspective. In the cited case, the expert witness argues that the statistical analysis, although proving that Asian Americans generally had higher test rates, failed to consider that Asian Americans, on average, score lower on the extracurricular scale.

Also, in a recent study by Parkhi et al. (2023), a machine learning model was developed to predict the likelihood of a student being admitted to specific colleges based on JEE exam scores in India.[2] This approach used decision trees and K-Means clustering to give students a list of potential college options based on historical admission data. Additionally, many popular algorithms work to predict whether a high school student will be accepted into a university. These include websites like Collegevine,[3] Campus Reel,[4] and even the US News and World Report.[5] We think it would be interesting to see what machine-learning models are behind these predictions. While these are similar models, our project aimed to explore and understand application behavior, an area of the college application process that we believe has not been thoroughly studied.

By shifting from predicting admission outcomes to predicting application behavior, our model adds a new layer of insight. They will give universities an understanding of the factors that attract certain types of applicants. This can help them tailor their admissions strategies as well as figure out brackets of people to pour outreach efforts into. The results of this project could also be useful to students, providing them with data-driven predictions that could help guide their college application decisions and narrow down their options.

Data Description

This data set, collected by Opportunity Insights, provides a comprehensive look at the components of US college admissions. It focuses on 139 colleges and universities in the United States, both public and private. The data set looks at student and college data from the entering

[1] D. Card, “United States District Court for the State of Massachusetts,” projects.iq.harvard, https://projects.iq.harvard.edu/files/diverse-education/files/expert_report_-_2017-12-15_dr._david_card_expert_report_updated_confid_desigs_redacted.pdf (accessed Oct. 14, 2024).

[2] P. N. Parkhi, A. Patel, D. Solanki, H. Ganwani and M. Anandani, "Machine Learning Based Prediction Model for College Admission," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/ICETET-SIP58143.2023.10151595.
keywords: {Machine learning algorithms;Databases;Education;Web pages;Prototypes;Machine learning;Information processing;Machine Learning;Decision Tree;K-Means},

[3] “College Admissions Calculator.” CollegeVine, www.collegevine.com/admissions-calculator. Accessed 14 Oct. 2024.

[4] “2024 College Chances Calculator - as Seen in Forbes.” CampusReel, www.campusreel.org/college-acceptance-calculator. Accessed 14 Oct. 2024.

[5] College Admissions Calculator - U.S. News, www.usnews.com/best-colleges/admissions-calculator. Accessed 14 Oct. 2024.

classes of 2010-2015 and aggregates data from around 2.4 million domestic students. It consists of 1946 observations with 81 variables.

The data set is aggregated by university name and parental income level. Each row presents a wide range of observations including test scores, application rates, and attendance outcomes for each unique combination of income percentile and university.

Pre-processing

In order to prepare the dataset for the different models we would later build, we began by cleaning up the data. First, we removed all unweighted columns and redundant tier names such as “tier_name,” “test_band_tier,” and “flagship.” The decision to exclude the unweighted columns was made to concentrate solely on the weighted values, which were more relevant to our analysis.

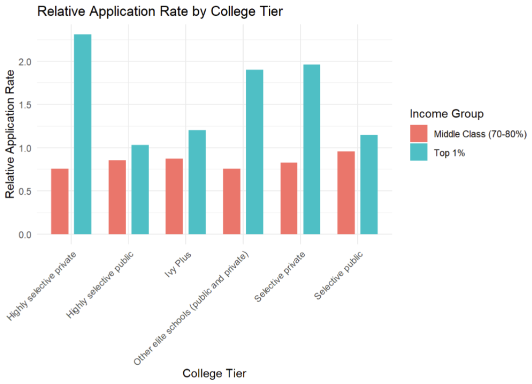
Next, we split the dataset into two subsets: one for public schools and another for private schools. This separation allowed us to address differences in the data structures between the two types of colleges. After splitting the data, we removed rows with more than four missing values. For the remaining missing values, we use MICE to impute them. Something we noticed early was that certain variables specific to public schools were not available for private schools. By splitting the data before handling missing values, we ensured that no crucial information was lost. Lastly, for both the public and private datasets, we deleted the “public” column and changed the “tier” and “name” columns to factor.

As an important step for our models, we split both our datasets into test and train data using an 80/20 ratio. This left the private dataset with 844 rows of training data and 212 rows of test data. On the other hand, public had 489 rows of training data and 123 rows of test data.

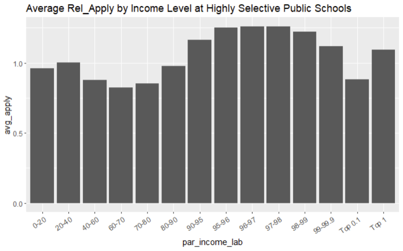
XGBoost Model

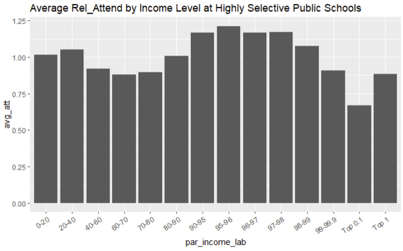
In order to build the XGBoost model, further data cleaning and transformation were necessary. We removed the variable “name” as this is an identifying variable and also removed par_income_lab as it was redundant with par_income_bin. We also made the par_income_bin variable to be of numeric type. Next, we converted categorical variables representing school tiers into numeric factors for both private and public schools in the training and test datasets. This was an essential transformation of our dataset as the XGBoost model required numeric input rather than categorical variables.

Summary Graphs

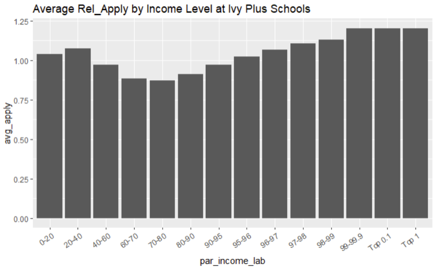

Relative Application Rate by College Tier

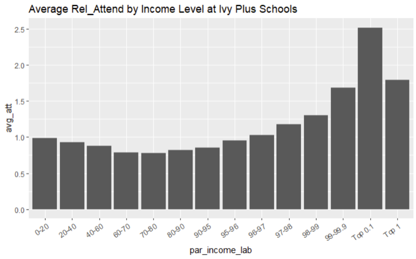
- Students in the top 1% income bracket apply the most of all types of colleges. However, their application rate relative to other income brackets is much higher for highly selective private schools, selective private, and other elite schools.


Average Rel_Apply by Income Level at Highly Selective Public Schools

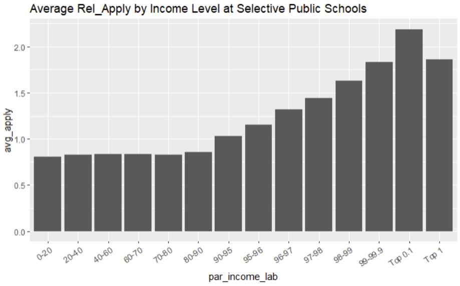

Average Rel_Attend by Income Level at Highly Selective Public Schools

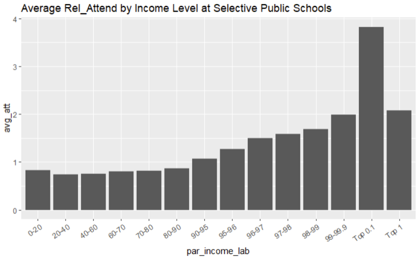
- The application rate for different groups in this tier is very close to the attendance rate


Average Rel_Apply by Income Level at Ivy Plus Schools


Average Rel_Attend by Income Level at Ivy Plus Schools

- Pretty significant differences for Ivy Plus colleges. People from all income brackets seem to apply at a high rate, however, most people in the Top 1% end up attending.





- No much people from the lower income brackets applying, most who attend are in the top 1%

Methods

Linear Regression Model

To begin our analysis, we started by applying a linear regression model to both the public and private school data frames. This was done to get a preliminary understanding of which variables would have the most significant impact on the rel_apply variable. We did not include the variables *name* and *super_opeid*, as these variables were just identifying variables. Although the linear regression model is simple to interpret and apply, due to its assumption of linearity of predictors and the response variable, it can oversimplify the predictive relationship between variables. Furthermore, due to the nature of our dataset, and how interrelated a lot of our variables are, we are very prone to being impacted by multicollinearity. Thus, we looked to other machine learning models to get higher predictive power.

Bagging Model

Next, we decided to apply the Bootstrap Aggregation (bagging) model to our data frame. We made the number of trees to be 200, and mtry as 53 for public schools and 23 for private schools (for the number of variables). After the model was created, we printed RMSE (root mean squared error) values for each bagging model. Bagging can assist in reducing the impact of outliers and overall variance due to its ability to rerun various models through the number of trees. However, we still sought a more accurate and comprehensive ML method.

XGBoost Model

After preparing both training and test matrices for the XGboost, we attempted to apply an XGBoost model to our dataframes. We started by training the models for private schools. We tested the models with different learning rates, also known as *eta*. We tested eta to be 0.005, 0.01, 0.05, 0.1, and 0.3

and their respective RMSEs. We plotted the error rates against the number of trees. As the number of trees increases, the error will decrease. However, with this graph, we could select the learning rate with the lowest error. Afterward, we calculated the SHAP scores to get a better sense of the importance of each feature. After calculating the SHAP values, we decided to rerun the XGBoost on a smaller data set, to see the impact of predictors without the overwhelmingly significant variables to see a more detailed analysis on features that were pushed out our initial analysis due to their strength. To finish our XGBoost model, we calculated the RMSE values for both public and private data sets and their respective XGBoost models. The XGBoost model considers the specificities of each input of the model. Thus, it can produce a relatively accurate predictive model. With the use of SHAP values, we can also create another model that explores the features past the two main dominating features.

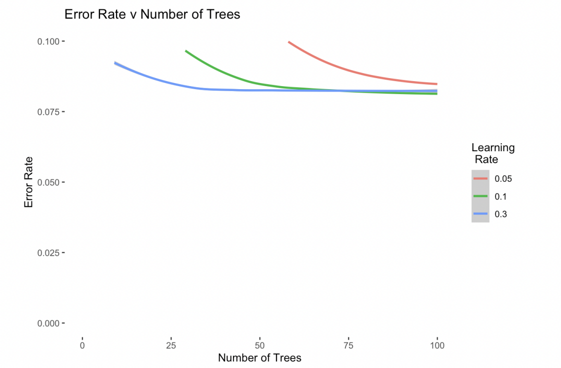
Results

In our linear regression model, both public and private schools had a large number of significant results, but it was clear that private schools had more significant variables in predicting rel_apply than the group of public schools did. To evaluate the predictive power of our linear regression models, we utilized the adjusted R-squared values. The adjusted R-squared value presents the percentage of variance that can be explained by the explanatory and also adjusts for the number of variables included in the model. For public schools, we saw an adjusted R-squared value of 0.9934, while the linear regression model for private schools saw an adjusted R-squared value of 0.9831. Both adjusted R-squared values are high, indicating that the models are able to predict a large percentage of the variance in rel_apply. However, we saw that the public schools' data frame had a slightly higher adjusted R-square value.

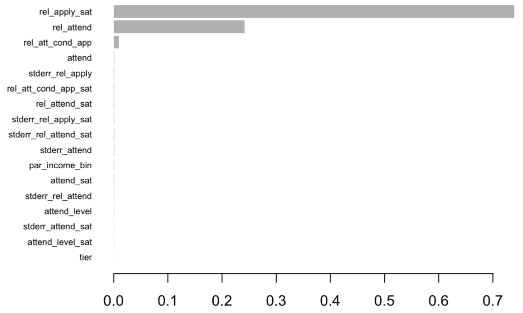
With our bagging model, the percentage of variance explained was 95.91% for private schools and 92.77% for public schools. We utilized RMSE values to evaluate predictive power, and it is often utilized to measure how accurate the predicted values of a model are compared to the test values. RMSE calculates residuals' standard deviation, showing the mean difference between the predicted values and actual values. In terms of printed RMSE values, the private school bagging model printed 0.0999, while the RMSE for the model for public schools was 0.1098. Thus, as seen by the higher percentage of variance value and the lower RMSE, the private school dataframe's bagging model had higher predictive power for the rel_apply variable. We wanted to note the higher predictive power of the private school bagging model despite the public school bagging model including 20 more variables.

Commented [1]: what??

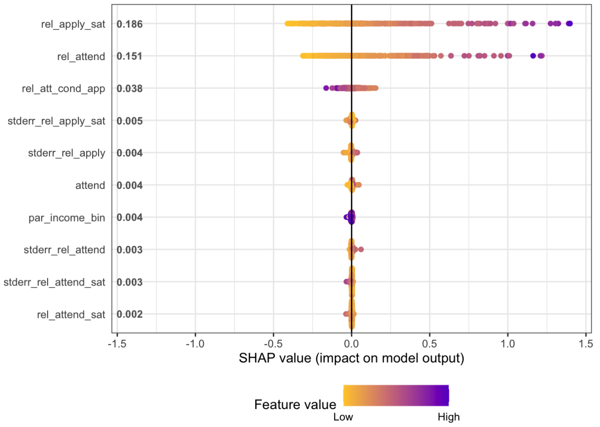
Next, we attempted to build an XGBoost model for the Private School dataframe. We needed to determine an essential hyperparameter for our model: eta. Our testing of various learning rates, *eta*, showed that a learning rate of 0.1 would yield the lowest error. Below are our findings; we dive more into these in-depth explanations within the *Discussion* section.



Thus, with our finalized XGBoost model, we used a cross-validation fold of 5, the number of rounds as 100, and our learning rate as 0.1. We also plotted feature importance for private schools at this point. This graph showed us that rel_apply_sat and rel_attend were the most important predictors of rel_apply.

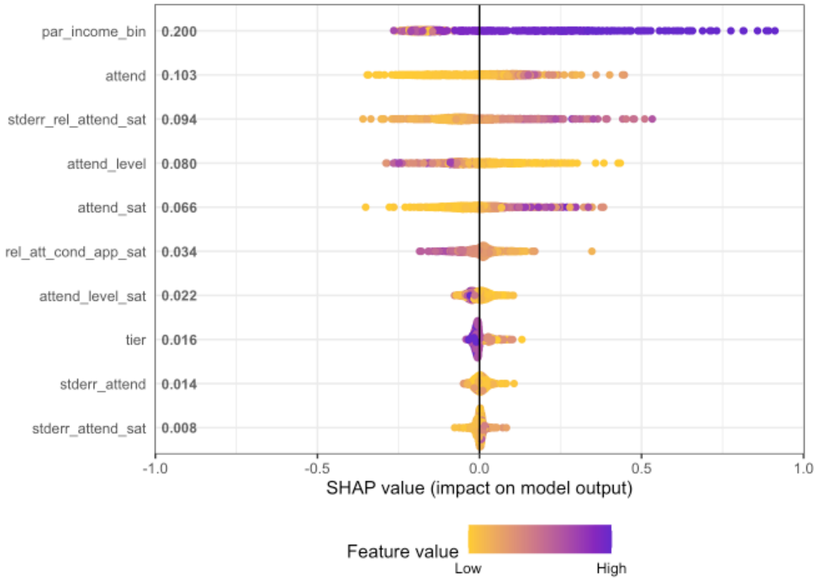


Our SHAP values analysis also allowed us to see what other predictors are impactful in predicting *rel_apply*.



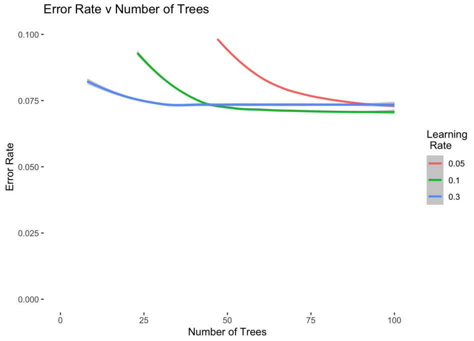
This SHAP value analysis showed us the other predictors that had an influence on *rel_apply*. According to rel_apply_sat, rel_apply still matters even when measured within the test score band, which is most common for that particular college. According to rel_attend, it seems that relative application also had an impact on application. The greater the relative historical attendance, the more likely to apply.

This smaller subsetted analysis allows us to delve deeper into feature effects that were not *rel_apply_sat* and *rel_attend*; because of the high significance of *rel_apply_sat* and *rel_attend,* we were unable to see the impact of variables *par_income_bin, attend, stderr_rel_attend_sat, attend_level*, and several other variables beneath.

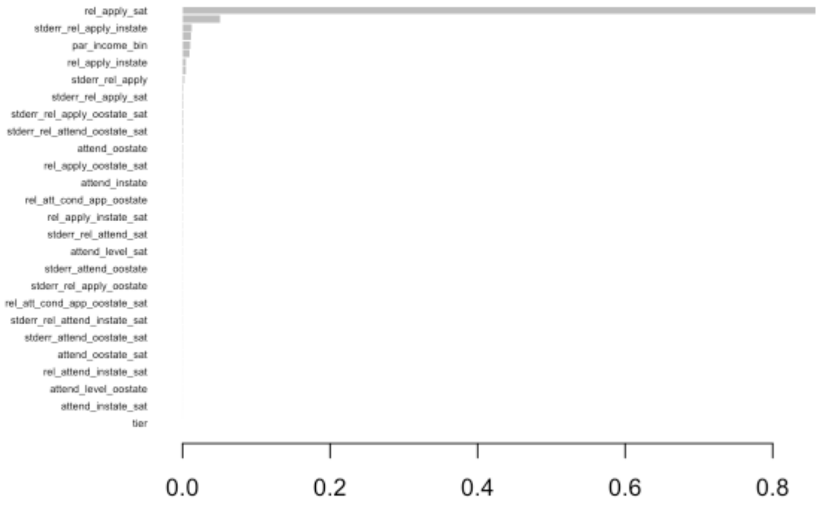


Looking at the detailed feature analysis, we can see that par_income_bin has high values scattered throughout. However, only low incomes are pictured as a negative SHAP value, meaning that because of one's income, they would be less likely to apply to a particular college. Under income is attendance, where we can see that high historical attendance means one is more likely to apply. Still, an income bin with low historical attendance does not necessarily mean one is more likely to apply than others.

Next, we moved on to complete the same analysis for public schools. Our ideal learning rate was also 0.1 for this model.

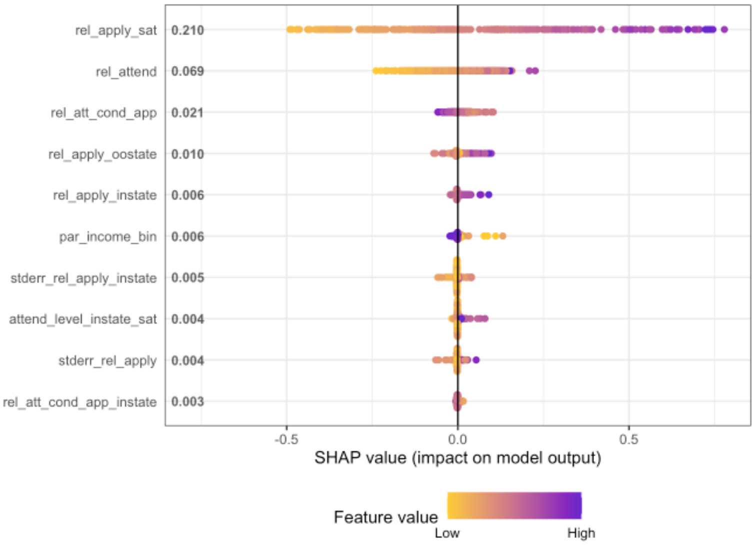


Here is our feature importance:

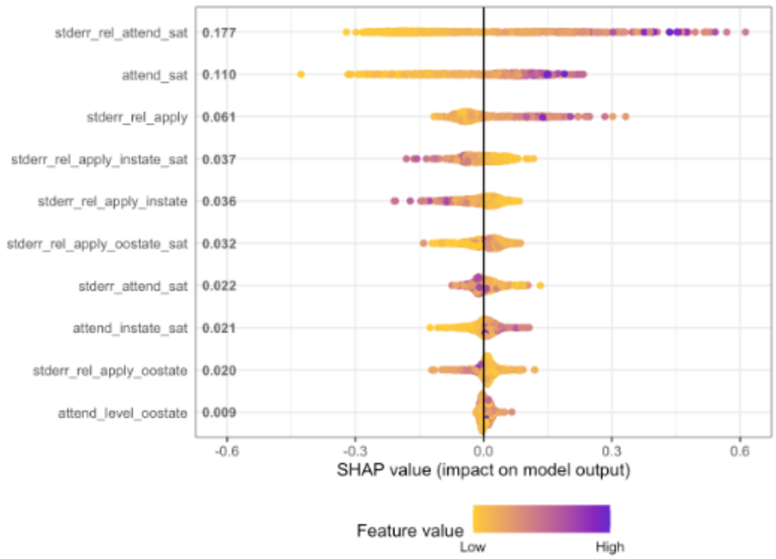


We re-did this larger model vs. smaller model SHAP variable analysis for public schools as well. The initial larger model variable analysis showed that variables *rel_apply_sat*, *rel_attend, rel_att_cond_app*, and *rel_apply_oostate* had the most significant impact on the rel_apply response variable.

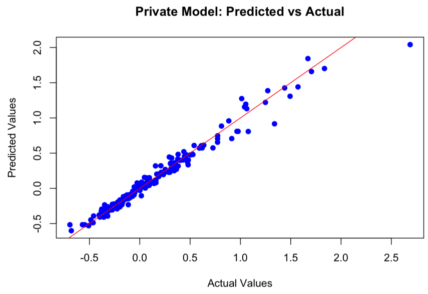
Essentially, rel_apply_sat conveyed that rel_apply still matters even when measured within the test score band most common for that particular college. However, this feature is more spread out for public than private schools. We can also see that a high historical relative attendance (rel_attend) means someone is more likely to apply. Interestingly, a high ratio between relative attendance and application means that someone is less likely to apply. This means that students from certain income bands are likely to attend once they apply. This might lower application because only students who are confident about attending would apply to these schools.

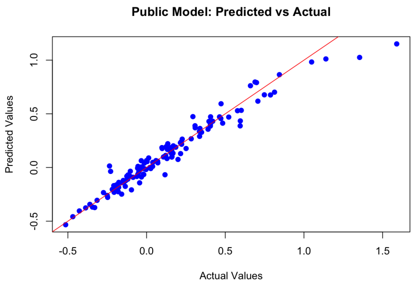


After adjusting for the smaller model, we saw the introduction of variables *stderr_rel_attend_sat, attend_sat*, and *stderr_rel_apply* having the biggest impact on the rel_apply variable. Essentially, attendance and application matter once again, and they matter more than application and attendance from instate v. out of state.



The RMSE of our private school model was 0.08376, whilst the RMSE of our public school model was 0.08307. Both of these RMSE values are small, indicating that our model had high predictive accuracy. These values are also very similar, showing that both have relatively similar predictive power.





Challenges

Some challenges that we encountered when analyzing our XGBoost model were:

1. Making conclusions from complex, scaled variables.

We mitigated this issue by doing an in-depth reading on variable definitions and integrating these definitions in the context of our problem and our explanations.

2. Thinking about the intercorrelation between features.

Most of our variables relate to rel_apply and rel_attend, so it only make sense that many rel_apply-adjacent features have purported importance. For example, rel_apply_sat might appear much more important than just being a pure reflector of SAT scores.

3. Doing a SHAP analysis with dominating features that would crowd out other features.

We mitigated this issue by performing a more detailed analysis by eliminating super significant features that overshadowed the others. However, one issue with this is that SHAP values measure individual importance as well as the importance of the feature within the model. In this manner, we inherently change the model by eliminating features; this might change how the SHAP summary perceives each variable.

Discussion

Many insights from the model are discussed through the methodology of our modeling process. However, there are many insights that can be derived through a more inquisitive approach. Our main findings are that XGBoost had the best model performance, SAT scores and historical attendance have a positive significant effect on both private and public schools, and income has a positive significant effect on private school application.

I. XGBoost had the best model performance.

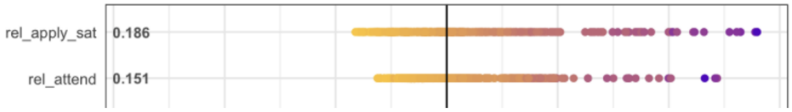
This is the most straightforward section of the three main insights. Ultimately, XGBoost had the best RSME, meaning it was the most accurate to our actual values and the best predictor. We expected this from the beginning, as XGBoost's "boosting" rather than "bagging" properties allow for accurate prediction. Although a "boosting" method poses the risk of overfitting, it doesn't matter in this context; the college admissions and applications process is quite standardized, but perhaps things like cultural fit affecting application could be an outlier that was overfitted into the data.

It appears as if our public model is slightly more accurate than our private model. Our private model reflects a slightly larger RMSE (0.0838) versus our public model (0.0807). This is likely due to more variables, instate and out-of-state information, available in our public model but absent in our private model.

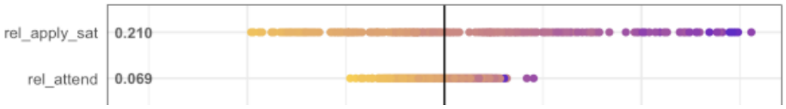
II. Private Schools: Income and SAT scores have positive, significant effects on application, but SAT scores affect public schools more. Historical attendance is also important, but nuanced.

Out of the overall feature importance and SHAP value analysis, we concluded that rel_apply_sat, which reflects relative application rates dependent on the middle 50 points test score band, is significant, followed by relative attendance. Interestingly, these two top variables' high points are more spread out and sparse than the ones from public schools, implying that some income bins (likely high ones) high attendance rate and high SAT scores makes them even more likely to apply to certain colleges.

Top values for private schools



Top values for public schools

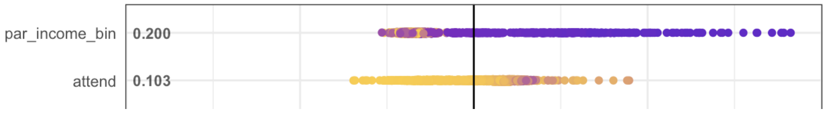


Here, we can also see that the differences between the two strongest SHAP values reflect that SAT scores affect public schools more than private schools. This is because the difference between rel_apply_sat and rel_attend is much greater for public schools than for private schools. Another interesting observation is that high rel_apply_sat values tend to have a near-equal impact on application whereas, in public schools low values for rel_apply_sat affect application much more than in private schools. This means that application might be more sensitive to low SAT scores in public schools rather than private schools.

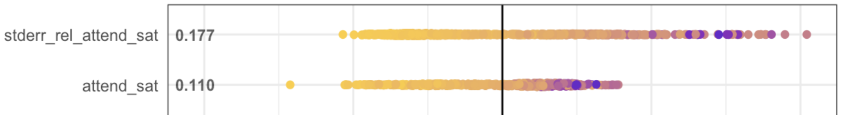
Historical attendance patterns also affect private school application, as it seems to have a greater effect on private schools than public schools by looking at the pure SHAP value. However, the private school model does not include any information on instate and out-of-state application and attendance, meaning that this state information might be with private schools' rel_attend feature instead of broken out like it is in public schools. In this manner, it seems like historical attendance patterns do, in fact, have a positive significant effect on private application, but it is difficult to say how this compares to attendance's impact on public application.

One interesting conclusion from our more detailed analysis resurfaced our notion that income matters more for private schools than for public schools.

Top detailed analysis features for private schools



Top detailed analysis features for public schools

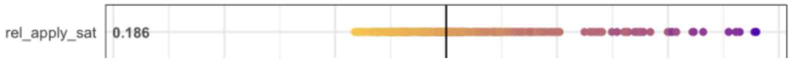


As you can see, higher income bin values have a positive effect on application, while lower income bins always have a negative effect on application. However, income bins are much lower on the detailed SHAP value list for public schools, meaning they don't have quite as great an effect on application as they do for private schools.

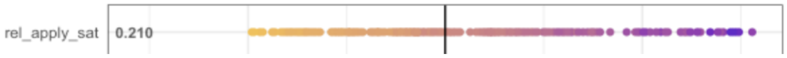
III. Public Schools: Historical attendance patterns and SAT scores have positive, significant effects on application.

As for public schools, we can see that rel_apply_sat also has a great effect on application, and to a greater extent. It appears as if historically low application based on the middle range of SAT scores for a college can penalize public school applicants more than private applicants since there is a lower range for the SHAP values in public schools

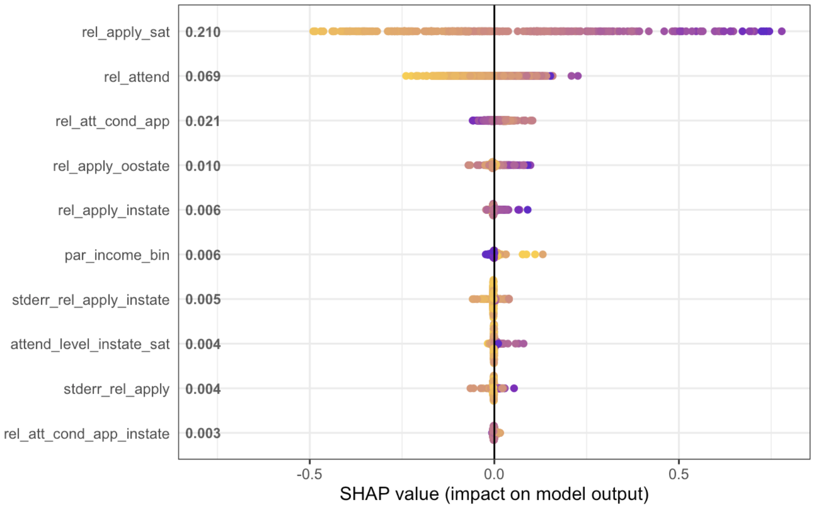
Private rel_apply_sat



Public rel_apply_sat



As discussed under public schools, we can also see that instate vs. out-of-state application has a great effect on the overall application rate as well from the All Features SHAP value summary:



As you can see, application instate and out-of-state have an effect on rel_apply, but to a lesser degree than rel_attend, or overall historical attendance. It also becomes apparent that rel_apply is more sensitive to out-of-state applications than they are to instate applications, given the SHAP value for rel_apply_oostate is almost 2x the value of rel_apply_instate.

Future Actions

Ultimately, it appears that income matters more and SAT matters less for private schools. This poses the issue of inequity; many of the T20 universities are private schools. This isn't to say that rankings are necessarily fair or an accurate measure of a better education, but their name and prestige undoubtedly has an impact on job and grad school admissions.

Here are some high-level actions that can be taken to reduce inequity in the application process:

- More need-based scholarships for private schools
- More reasonable financial aid for middle-class families applying to private schools
- Programs that encourage high schools in lower-income residential areas to apply to T20 schools

Conclusion

The American college admissions process is a byproduct of the college application process. Many high school students spend hours looking at statistics regarding admission, but not many consider who applies in the first place.

Our dataset explores applications for 139 colleges submitted by around 2.4 million domestic students. We ran linear regressions, random forests, and XGBoosts to predict what makes students more likely to apply. Our dataset presented many challenges due to its grouping by income bin and not by individual students. The fact that there were many similar features also presented some confusion. However, our thorough analysis including feature importance, full analysis of SHAP values, and detailed analysis of SHAP values also helped us reach the main conclusions.

Ultimately, our XGBoost models revealed the most information to us and had the best RSME. Our main findings are that XGBoost had the best model performance, SAT scores and historical attendance have a positive significant effect on both private and public schools, and income has a positive significant effect on private school application. A more thorough analysis of detailed SHAP values revealed that income bins are very significant for private schools and that instate and out-of-state metrics are essential when looking at public school applications.

Future Work

Given more time and resources, we would have liked to perform a more thorough analysis of SHAP values. There are many rel_apply-related features in our analysis that we did not eliminate so that we could draw further conclusions (e.g., rel_apply_sat revealed the importance of SAT scores; however, it would not be accurate to state that SAT scores are thus more important than rel_attend since this measure is, in some manner, a combination of both

SAT scores and rel_apply). Eliminating these features and adding them back in would provide a more thorough analysis with less error in our extrapolation. We would also like to find a way to mitigate the fact that there is no instate and out-of-state data for private schools, since this skews the comparison between private and public schools' rel_attend and rel_apply. It would also be interesting to examine other factors that influence application, such as culture fit.

In conclusion, our analysis was more of an exploratory analysis than a definite one but reveals promising results into suggesting what causes domestic students from a specific income group to apply.

Contributions

Ebube - Intro, Related Work, Data Desc.
Sarah - Methods, Results
Renee - Discussion, Conclusion and Future Work

Bibliography

1. D. Card, "United States District Court for the State of Massachusetts," projects.iq.harvard, Dec. 2017. [Online]. Available: https://projects.iq.harvard.edu/files/diverse-education/files/expert_report_-_2017-12-15_dr._david_card_expert_report_updated_confid_designs_redacted.pdf. [Accessed: Oct. 14, 2024].
2. P. N. Parchhi, A. Patel, D. Solanki, H. Ganwani, and M. Anandani, "Machine Learning Based Prediction Model for College Admission," in 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/ICETET-SIP58143.2023.10155195.
3. "College Admissions Calculator," CollegeVine. [Online]. Available: https://www.collegevine.com/admissions-calculator. [Accessed: Oct. 14, 2024].
4. "2024 College Chances Calculator - as Seen in Forbes," CampusReel. [Online]. Available: https://www.campusreel.org/college-acceptance-calculator. [Accessed: Oct. 14, 2024].
5. "College Admissions Calculator," U.S. News. [Online]. Available: https://www.usnews.com/best-colleges/admissions-calculator. [Accessed: Oct. 14, 2024].