Final Project - Exploring the Relationship Between Middlebury College's Admissions Rating and Predicting Freshman and Sophomore GPA

Kyle McCausland, Jacob Sherf , & Ciara Murphy

2024-12-13

Our Dataset

The dataset we analyzed focuses on 657 Middlebury College students from 2010 and includes key variables such as first-year GPA (fygpa), sophomore GPA (sophgpa), gender, athletic status, and average academic rating (AAR). The AAR is a critical metric used by the admissions office to evaluate applicants holistically on a scale of 1 to 7. It reflects the academic rigor and performance of a student in the context of their high school. Admissions officers consider various factors, including high school class rank, the difficulty and breadth of the student's course load, both weighted and unweighted GPAs, and writing ability as demonstrated in application essays. Importantly, these metrics are not assessed in isolation but are weighted against the resources and opportunities available at the applicant's high school. For example, admissions evaluates how many advanced courses the student chose to take relative to the number offered at their high school, and test scores are interpreted in the context of the school's average performance.

In addition to academic performance, athletic status is another critical component of the dataset. While Middlebury does not offer athletic scholarships, athletes benefit from an additional layer of support in the admissions process. Coaches advocate for prospective student-athletes by submitting evaluations that include transcripts, test scores, and an athletic rating. This rating, also on a scale of 1 to 7, reflects the athlete's expected contribution to their sport at Middlebury. A higher athletic rating often correlates with more flexibility in admissions criteria, provided the student meets Middlebury's baseline academic standards.

Our analysis aims to explore the relationship between AAR and academic outcomes, focusing on first-year and sophomore-year GPAs. To better understand these dynamics, we categorized the students into four distinct groups: male athletes, female athletes, male non-athletes, and female non-athletes. These categories allow us to investigate whether and how gender and athletic status influence the relationship between admissions metrics and subsequent academic performance.

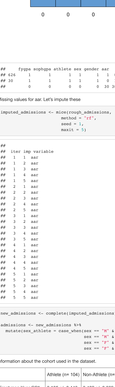
```
library(ggplot2)
library(naniar)
library(mice)
library(tidyverse)
library(rpart)
library(rpart.plot)
```

Let's load in our admissions dataset

```
rough_admissions <- read.csv('/Users/ciaramurphy/Downloads/admissions.csv')
```

Let's see if there is any missing values

```
md.pattern(rough_admissions)
```




```
##     fygpa sophgpa athlete sex gender aar
## 626     1       1       1   1      1   1 0
## 30      1       1       1   1      1   0 1
##         0       0       0   0      0  30 30
```

Missing values for aar. Let's impute these

```
imputed_admissions <- mice(rough_admissions,
                           method = "rf",
                           seed = 1,
                           maxit = 5)
```

```
##
##  iter imp variable
##   1   1  aar
##   1   2  aar
##   1   3  aar
##   1   4  aar
##   1   5  aar
##   2   1  aar
##   2   2  aar
##   2   3  aar
##   2   4  aar
##   2   5  aar
##   3   1  aar
##   3   2  aar
##   3   3  aar
##   3   4  aar
##   3   5  aar
##   4   1  aar
##   4   2  aar
##   4   3  aar
##   4   4  aar
##   4   5  aar
##   5   1  aar
##   5   2  aar
##   5   3  aar
##   5   4  aar
##   5   5  aar
```

```
new_admissions <- complete(imputed_admissions)

admissions <- new_admissions %>%
  mutate(sex_athlete = case_when(sex == "M" & athlete == 0 ~ "Male Non-Athlete",
                                 sex == "M" & athlete == 1 ~ "Male Athlete",
                                 sex == "F" & athlete == 0 ~ "Female Non-Athlete",
                                 sex == "F" & athlete == 1 ~ "Female Athlete"))
```

Here is information about the cohort used in the dataset.

	Athlete (n= 104)	Non-Athlete (n= 522)
Freshman Year GPA	3.135 +/- 0.446	3.402 +/- 0.369
Sophomore Year GPA	3.207 +/- 0.416	3.421+/-0.359
AAR	3.82 +/- 1.13	5.00 +/- 0.988
Sex		
Male	59	255
Female	45	297
Gender		
Male	59	255
Female	45	297

Using Decision Trees - A Supervised Learning Model

Exploring our dataset using a known supervised learning model.

Before using our newly learned SVM approach, we wanted to understand our data better using a decision tree, which we have used numerous times in our Statistical Learning class before. The results of this tree show us that the data was indeed split by sex, aes, and athlete status when predicting fygpa, indicating to us these variables have a potential relationship in being able to predict first year GPA. Using these variables that the decision tree has made splits on, we can explore this relationship further in our SVM. While this tree is just for fygpa, we made the assumption that there is a similar relationship for sophgpa too.

```
tree1 <- rpart(fygpa ~ athlete + aar + sex + gender, data = new_admissions,
               control = rpart.control(cp = 0.01))

rpart.plot(tree1)
```

Using Support Vector Machines - A Supervised Learning Model

We chose to use Support Vector Machines to address our research question. A Support Vector Machine (SVM) is a supervised machine learning algorithm designed to find the optimal separation between data points. While commonly used for classification, SVM can also be adapted for regression tasks, such as predicting continuous outcomes like GPA in our case. In regression, the algorithm fits a linear hyperplane through the data to minimize prediction error while maximizing the margin, or tolerance, around the predictions. This ensures the model is not overly sensitive to variations in the data.

For this analysis, we used a linear kernel, which assumes a straightforward relationship between the predictors (AAR, sex, and athletic status) and the target variable (GPA). A linear kernel is ideal when the relationship between predictors and the target is expected to be linear, which we can safely assume as true in this case. The cost parameter C was set to 1 to balance model accuracy and generalizability, ensuring the predictions aren't overly influenced by outliers in the training data. By using this approach, the SVM constructs a "widest street" through the data, making predictions with minimal error while maintaining simplicity.

One key advantage of SVMs is their flexibility in managing multi-dimensional predictor spaces while maintaining efficiency. This makes SVMs particularly well-suited to our dataset, where AAR, sex, and athlete status collectively represent distinct but intersecting dimensions of influence on GPA. By leveraging the linear kernel, we could explore these interactions without introducing unnecessary complexity, which is vital for maintaining interpretability in an academic context.

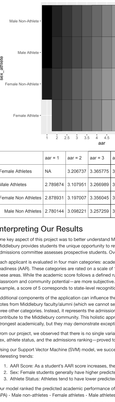
Train our Support Vector Machines (SVM) model on the entire dataset to predict First year GPA.

```
svm_model <- svm(fygpa ~ aar + sex + athlete, data = admissions, kernel = "linear", cost = 1, scale = TRUE)
```

Predict `fygpa` for each `aar` in the dataset

```
admissions$predicted_fygpa <- predict(svm_model, admissions)

admissions %>%
  ggplot() +
  labs(x = "aar", y = "sex_athlete", title = "How does aar predict first year GPA across different Sex/Athlete co
mbos?") +
  geom_tile(aes(x = factor(aar), y = sex_athlete, fill = predicted_fygpa),
            color = "black") +
  scale_fill_gradient(low = "black", high = "white")
```



	aar = 1	aar = 2	aar = 3	aar = 4	aar = 5	aar = 6	aar = 7
Female Athletes	NA	2.950462	3.103934	3.265265	3.427167	3.580080	NA
Male Athletes	2.730335	2.889236	3.048136	3.207040	3.365941	3.524843	NA
Female Non-Athletes	2.834985	2.993887	3.152788	3.311690	3.470592	3.629493	3.788395
Male Non-Athletes	2.773795	2.932661	3.091563	3.250464	3.409366	3.568267	3.727169

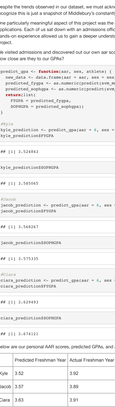
Now let's do it for sophomore GPA

```
svm_model <- svm(sophgpa ~ aar + sex + athlete, data = admissions, kernel = "linear", cost = 1, scale = TRUE)
```

Predict `fygpa` for each `aar` in the dataset

```
admissions$predicted_sophgpa <- predict(svm_model2, admissions)

admissions %>%
  ggplot() +
  labs(x = "aar", y = "sex_athlete", title = "How does aar predict sophomore GPA across different Sex/Athlete com
bos?") +
  geom_tile(aes(x = factor(aar), y = sex_athlete, fill = predicted_sophgpa),
            color = "black") +
  scale_fill_gradient(low = "black", high = "white")
```



	aar = 1	aar = 2	aar = 3	aar = 4	aar = 5	aar = 6	aar = 7
Female Athletes	NA	3.290737	3.365775	3.524613	3.427167	3.683851	NA
Male Athletes	2.789874	3.107951	3.266899	3.426027	3.365941	3.585080	NA
Female Non-Athletes	2.878931	3.197007	3.356045	3.515063	3.470592	3.674127	3.833160
Male Non-Athletes	2.780144	3.098221	3.257259	3.416297	3.409366	3.575335	3.734373

Interpreting Our Results

One key aspect of this project was to better understand Middlebury Admissions' approach to evaluating candidates and reviewing applications. Middlebury provides students the unique opportunity to read their application with an admissions officer and gain insight into how the admissions committee assesses prospective students. Overall, we were all impressed with Middlebury's holistic admissions process.

Each applicant is evaluated in four main categories: academics, the outside of the classroom, community potential, and overall academic readiness (AAR). These categories are rated on a scale of 1 to 7. The average admitted Middlebury student typically scores around a 4 across these areas. While the academic score follows a defined rubric that accounts for a student's school, the other categories—such as life outside the classroom and community potential—are more subjective. These scores reflect the applicant's impact during their high school years. For example, a score of 5 corresponds to state-level recognition, a 6 indicates national-level recognition, and a 7 reflects international recognition.

Additional components of the application can influence the AAR score, such as alumni interviews, athletic ratings provided by coaches, and notes from Middlebury faculty/alumni (which we cannot see any notes post application process). Notably, the AAR score is not an average of the three other categories. Instead, it represents the admissions officer's overall assessment of how well the applicant is likely to perform and contribute as the Middlebury community. This holistic approach acknowledges that a student with a high AAR score may not necessarily be the strongest academically, but they may demonstrate exceptional potential to positively impact the broader campus community.

From our project, we observed that there is no single variable that can directly predict a student's GPA. However, certain key factors—such as sex, athlete status, and the admissions rating—proved to be reliable indicators.

Using our Support Vector Machine (SVM) model, we successfully predicted first-year and second-year GPAs. The results revealed some interesting trends:

1. AAR Score: As a student's AAR score increases, their predicted GPA also tends to increase.
2. Sex: Female students generally have higher predicted GPAs compared to male students.
3. Athlete Status: Athletes tend to have lower predicted GPAs compared to non-athletes.

Our model ranked the predicted academic performance of students in their freshman year as follows: - Female non-athletes (highest predicted GPA) - Male non-athletes - Female athletes - Male athletes (lowest predicted GPA)

Despite the trends observed in our dataset, we must acknowledge the time commitment student athletes dedicate to their sports, and also recognize this is just a representation of Middlebury's constantly evolving community.

One particularly meaningful aspect of this project was the opportunity to visit Middlebury Admissions and review our own high school applications. Each of us sat down with an admissions officer (Paige) to discuss our scores across the four categories and receive our AAR score. This hands-on experience allowed us to gain a deeper understanding of the admissions process while also gathering personalized statistics for our project.

We visited admissions and discovered out our own aar scores. How well does our model do when predicting a GPA off an aar score in 2024? How close are they to our GPAs?

```
predict_gpa <- function(aar, sex, athlete) {
  new_data <- data.frame(aar = aar, sex = sex, athlete = athlete)
  predicted_fygpa <- as.numeric(predict(svm_model, new_data))
  predicted_sophgpa <- as.numeric(predict(svm_model2, new_data))
  cat("Predicted:\n",
      "FYGPA = predicted_fygpa,
      SOPHGPA = predicted_sophgpa)
}
```

```
#Kyle
kyle_prediction <- predict_gpa(aar = 6, sex = 'M', athlete = 1)
kyle_prediction$FYGPA
```

```
## [1] 3.524843
```

```
kyle_prediction$SOPHGPA
```

```
## [1] 3.585065
```

```
#Jacob
jacob_prediction <- predict_gpa(aar = 5, sex = 'M', athlete = 0)
jacob_prediction$FYGPA
```

```
## [1] 3.568267
```

```
jacob_prediction$SOPHGPA
```

```
## [1] 3.573135
```

```
#Ciara
ciara_prediction <- predict_gpa(aar = 5, sex = 'F', athlete = 0)
ciara_prediction$FYGPA
```

```
## [1] 3.629493
```

```
ciara_prediction$SOPHGPA
```

```
## [1] 3.674127
```

Below are our personal AAR scores, predicted GPAs, and actual GPAs:

	Predicted Freshman Year	Actual Freshman Year	Predicted Sophomore Year	Actual Sophomore Year
Kyle	3.52	3.92	3.59	3.88
Jacob	3.57	3.89	3.58	3.93
Ciara	3.63	3.91	3.67	3.94

Interestingly, while our predicted GPAs were slightly lower than our actual GPAs, this discrepancy can potentially be due to the age of the dataset. The data we used was originally collected in 2010. Over the past 15 years, Middlebury has evolved as an institution. Post COVID19 pandemic, students and faculty alike have observed grade inflation, with higher GPAs across various disciplines compared to 2010 levels. We believe that if we had access to a more current dataset, our model's predictions would more accurately align with our actual GPAs.

Acknowledgments

We would like to extend our heartfelt gratitude to Professor Alex Lytvyn, our Statistical Learning professor, for his guidance, patience, and inspiration throughout the semester and for his generous office hours! We also thank the Middlebury Admissions Office for their transparency, openness, and passion to build the best community they can when explaining the admissions process and allowing us to view our own applications. Most importantly, thank you for accepting us into Middlebury. We are proud to be Math Kids!