BU Athletics: Team-B

Report and Code Submission

For each deliverable, you must submit both a report and the associated code. Your report should include the following sections for Deliverable 1:

- A brief introduction to your problem statement.

The intention behind this deliverable was to clean up the data we were given and then our team made it a goal to answer all the baseline questions we were given by BU Athletics.

- Details of the data collection or cleaning steps you've undertaken.

We processed the initial Excel table we were given as a CSV then cleaned up the table in pandas. For example in the CSV we deleted the column 'Student' which was a unique student ID number along with a few other columns that were all null values including the 12th-22nd semester GPA values for students which were 0.0 in the dataframe for all students. We also renamed some columns to easier names such as 'English Language Primary Language Of' to 'Language'. We lastly filled in some null values with more valuable assumed values such in the 'Country' column where all foreign-born students had their home country listed. We filled in the other remaining values with 'USA'.

- Exploratory Data Analysis (EDA).

Our EDA is provided in the Jupyter Notebook files that we answered our baseline questions in. We have also included graphs to see if we could visualize any interesting analytics results. See Google Slides presentation for graphs.

- If your analysis has led to answers for any of the questions or if you've formulated hypotheses, especially for at least questions.

How do these students with high SAT/ACT scores or high GPA perform academically at BU compared to student athletes with low SAT/ACT scores or low GPA? - Rob

- **High school stats are not great indicators of how students perform academically at BU. We tried to make a model that would predict how the student would perform in college. The Decision Tree classifier does not end up being a great model for predicting if athletes will fall into the A, B, C, or lower range of average GPAs. This, along with our scatter plots of HS GPA v. College GPA and SAT vs. College GPA might be an indication that high school stats are not good indicators of how a student will perform in college. These scatter plots did not have any well defined trends between these measurements.**

Does the average student athlete improve their GPA throughout their college year? - Rob

- **The average student does not improve their GPA much during their college year. After computing clusters using KMeans based on HS GPA, College GPA, and High school SAT, we did not find any surprising results. As HS GPA and SAT scores go up, the student usually performs better in college. However as we saw in our plots students that start off with lower GPAs usually improve at a greater rate than their peers who start off with higher GPAs as reflected by our graphs. This is not surprising as these students have more room to improve to begin with.**

What is the range of accepted SAT/ACT scores and highschool GPA for student athletes? - Bargav/Mounika

- **Range of SAT (ACCEPTED STUDENTS) - 950 - 1600 (950 be the lowest and 1600 be the highest)**

- **Range of high school GPA (ACCEPTED STUDENTS) - 2.2 - 4.3 (2.2 be the lowest and 4.3 be the highest)**

What percentage of these student athletes are domestic students? What about international students? - Bargav/Mounika

- **The percentage of International students is 15 percent and the percentage of domestic students is 84 percent**
- **Clearly there is more percentage for domestic students then the international students**

Does English being the primary language of the country impact the students' performance? - Bargav/Mounika

- **If you ignore the outliers, the distribution is pretty much the same, suggesting to us that English as a primary language or not, is not making any necessary difference for athletes in grades. In Fact the lower quartile range is slightly better for Non-English students.**

How does the academic performance of students with similar HS GPAs compare? (those with SAT/ACT vs. those without SAT/ACT score submitted)? Ex: A student with a 3.0 HS GPA compared academic performance in college and AI (have submitted SAT/ACTscore) - Bargav/Mounika

What is the range of BU GPA for student athletes? - Bargav/Mounika

- **The range of BU GPA is lying in between 0.0 to 4.0**

Do students from certain geographic areas (Northeast) perform better academically than another geographic area (Southwest)? - Jeff

- **No, in fact students from the Northeast perform marginally worse than the other regions. The big difference lies within the standard deviation; athletes from the Northeast are more alike to each other academically, as evidenced by the smaller discrepancy and the box chart. My hypothesis is that access to quality education is more equitable in the Northeast which might explain the difference with the larger gaps in other regions. In my normal distribution graph, the Northeast group is marginally thinner, which supports my hypothesis. It is taller than the other three groups because most students are from the Northeast, which is trivial since our university is located in the Northeast.**

Are there any significant differences in the academic performance of student-athletes based on their sport? - Jeff

- **Yes, there are considerable differences in average GPA between sports. Furthermore, some sports do have a wider range of GPAs. For example, Women's Golf has a significantly higher average GPA, and the difference is marginal as evidenced in the gradient of the heat map. With this, we can hypothesize the conduciveness of each athletic department to athletes' academic performance by its color gradient. Lastly, I performed Kruskal-Wallis because we are not using ANOVA. ANOVA assumes normality and equal variances. Since we see that there are jumps in GPA from the heat map, and the rate of change of color shade, we can't assume that and Kruskal-Wallis allows us to negate that margin of error. After calculating, there are differences in academic performance between sports, but we don't have evidence to reject the claim that they are equal.**

What are the most common majors for student athletes? Do student athletes tend to switch majors? - Shangzhou

- **We plotted the distribution of FIRSTMAJOR and the distribution of LASTMAJOR over the number of students. We found that initially Business Admin & Mgt has the most students and later more students transferred to this major.**

What percentage of student athletes stay in the same major? - Shangzhou

- **The percentage of student athletes stay in the same major was calculated by the number of students who had the same FIRSTMAJOR and LASTMAJOR over total students. 232 students changed their majors during studies. Based on the calculation, we can find that 76.6% of students stay in the same major.**

- Individual contributions of each team member. We recommend that each team member writes 3-4 lines about their contributions, which can then be compiled into the report.

Of the roughly ~8 base questions that we were assigned by BU athletics and spoke about in the previous section of this deliverable we tried to roughly divide up the questions evenly to individually deliver insights about. As you can see after each baseline question the person assigned to answering that question is named.

Rob - I did my two questions and the data cleaning/pre-processing. If not noted before I have taken on the role as team lead and have been trying to head meetings and do some of the non-technical work (reports, this deliverable). I have also been corresponding with Matthew at team meetings.

Bargav - Worked with Mounika on numerous questions and produced a variety of graphs to deliver new insights into core questions. Starting to look more into what new questions can be asked based on these insights provided.

Mounika - Worked with Bargav on numerous questions and produced a variety of graphs to deliver new insights into core questions. Starting to look more into what new questions can be asked based on these insights provided.

Shangzhou- I did two questions assigned to me and plotted figures to support my calculations. I submitted a pull request but it has not been approved. I need to find the correlation between majors and other features, and try to find out the possible reasons for student athletes changing majors.

Jeff - I did two questions assigned to me. I made visual graphs to be presented on Friday and support some of my hypotheses. I also used statistical methods (i.e. Kruskal-Wallis) to provide further clarity to my findings. Otherwise, all the base questions are answered. We just need to do further exploration to solve ambiguities that came from Kruskal-Wallis and visual discrepancies.

All of our code for this deliverable has either been merged into the team-b branch of the BU Athletics Spark Github repo by our TPM Matthew Batacan or is currently up as a PR awaiting review. The code can be found at ds-bu-athletics-performance/fa23-team-b/code. There are also certain code chunks that are currently up for PR review at the time of this submission and can be found in the GitHub repository.