ECON 8320 Project Writeup

In early 2021, the NCAA began allowing college athletes to earn money based on their Name, Image, and Likeness (NIL). Many college athletes have taken this opportunity; however, to the author's knowledge, there isn't a well-defined list of athletes with NIL deals. Consequently, for my final project for ECON 8320, I worked on compiling information regarding college athletes' NIL deals. By using the tools taught in Tools for Data Analysis, I collected information on 8,390 NIL-sponsored college athletes. This write-up will explain the methods, difficulties, and results of this research.

Methods

The website I used for compiling the NIL data was https://nilcollegeathletes.com/. I then used many Python packages to collect data regarding NIL deals. I utilized the requests package to gather the website's HTML code, and I also used the BeautifulSoup method (imported from the bs4 package) to parse the HTML code into a better format to work with. In most cases, the text within the website's HTML code isn't uniform; therefore, I also used Python's regular expression module (re) to extract text information. I used the Pandas library to store the college athletes' names, universities, sports, sponsors, Instagram, and Twitter accounts. I also used Pandas to create new columns—these include the number of sponsors each student has and the sports conference they play in. Once I collected this data, I created some plots using the Plotly.Express (px) package.

The bulk of this project was the creation of the web scraper. Firstly, once I parsed the webpage using the BeautifulSoup method, I searched for ways to parse the information for each student. I found that the "tr" tag included the name, university, sport, and sponsors for each student. There were about 20 students on each page, so I iterated through the "tr" tag to get the

information for all these students. However, finding the social media accounts was more of an ordeal. While the “tr” tag had a variety of information, it did not include the students’ Instagram and Twitter accounts. The website had a subpage for each student, so I used the requests and BeautifulSoup libraries to collect and parse the social media information in each of these subpages. Then, I merged (using a left join where each student’s name was the key) the social media information into the main DataFrame to create the final dataset.

Once I gathered all this NIL data, I inputted it into a Pandas DataFrame, and I began the data cleaning process. The web scraper function inadvertently created some extra columns, which I believe resulted from merging the social media data with the original dataset. Therefore, I dropped these columns and renamed the resulting set. The social media information was also somewhat incorrect; if a student did not have an Instagram/Twitter account, the function was supposed to return “N/A” as their account. However, this didn’t encompass all the use cases—I believe that some students’ subpages did not have a tag with the “Instagram” or “Twitter” text, causing this error. Consequently, I created a noneCheck function to make the output uniform, and I applied it to the DataFrame. Additionally, I created columns based on the schools’ sports conferences, and I created some plots for exploratory data analysis.

I also tried to fit a clustering model to the data; I used the sklearn.clustering package to create this. I ended up creating dummy variables for each conference (the columns got a value of 1 if they were a part of that conference). Additionally, I created a dummy variable for Twitter and Instagram accounts. Once I created all these dummy variables, I inputted these columns into a K-means clustering algorithm to try and find patterns that I wouldn’t be able to extrapolate.

Difficulties

The most challenging part of this project was the collection of NIL-sponsored athletes. The general process of creating a web scraper wasn't extremely challenging, but figuring out the best tags to search through was difficult for me. I'm not too well versed in HTML code, so I didn't understand what an "a" or "li" tag encompasses, to put it lightly. Consequently, I used a lot of trial and error to create the final code, which took a while to synthesize. I believe there are more efficient ways to get the same information, but I'm not sure how to find this. However, the most challenging part was collecting students' social media information. I didn't necessarily want to crawl through each student's subpage of data (due to efficiency concerns), but since this information wasn't present on the main page, I had no choice. Then, I couldn't figure out how to get each student's subpage, but I eventually realized to look in the "a" tags in the student list I generated for each page.

In the future, I hope to have a greater grasp of HTML code and also understand the best ways to deconstruct parsed information. More specifically, I hope to understand the main tags to search for and also understand how to create a scraper that catches all use cases. This mastery only comes with practice, but I strongly believe that I've improved greatly over this semester in this facet. Additionally, I would also research the structure behind the website before I start scraping it—even though the website I used was structured nicely, I perhaps needed to look at other websites' structures.

Creating the clustering model was also difficult. This was mainly because I didn't understand how to fit everything together. There were many ways to create a clustering model, and it was difficult for me to find the "best" model for my purposes. Consequently, I spent a lot of time consulting the sklearn documentation, but in the end, the model turned out to be

insightful. In the future, I hope to understand better ways to choose a model, and hopefully, we'll learn more about this process in Business Forecasting.

Results

As stated earlier, I utilized a web scraper to collect information on 8,390 NIL-sponsored students and conducted an exploratory data analysis using this data. For me, the most surprising result was the differences between sports conferences. I thought that the Power Five conferences would have a majority of NIL-sponsored students; however, a majority of students in the data were not from these conferences. I also thought that the Southeastern Conference (SEC) would've had the most NIL deals in the Power Five, but they were third (behind the Big 10 and the ACC). This behavior was still present when I looked at "NIL deals per capita" (dividing the total deals in each conference over the number of schools in that conference). I believed that the SEC was king of NIL deals—since they dominate college football—but football isn't the only college sport.

Another interesting observation was the variety of sports that NIL-sponsored students were participating in; I didn't realize that students were getting NIL deals for participating in water skiing, equestrian, sailing, and squash, to name a few. I also thought it was interesting to see how people from different conferences use social media. Using the clustering model, I found that students in Power Five conferences use Twitter and Instagram at higher rates than those not in Power Five conferences. Additionally, I found that both groups used Twitter at a higher rate than Instagram. Consequently, this project allowed me to gain a deeper understanding of the NIL market and the myriad of opportunities students have access to.