Emani Jones Portfolio

Table of Contents

Introduction

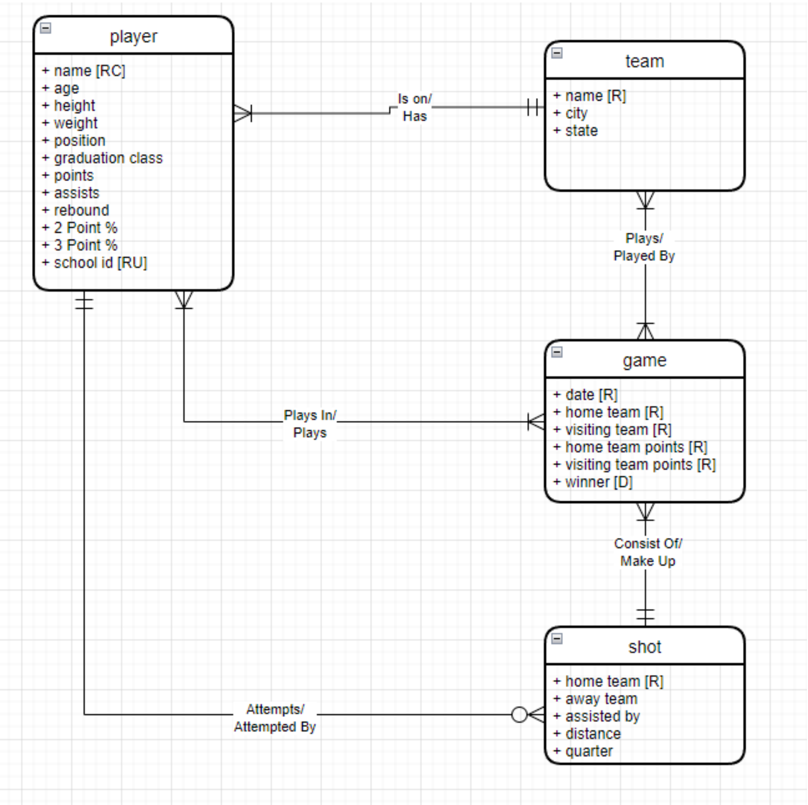
During my undergraduate studies I minored in sports analytics, those courses were the introduction to the world of data analytics and data science. Falling in love with it, I decided that that was going to be my career path, so when I needed to choose an internship for my capstone, I knew that I wanted it to be in sports analytics. I interned at Wasserman Management Group and while it was a great experience for me, I realized that my data analytics skills were lacking and that is when I decided that I wanted to pursue my Masters in Applied Data Science.

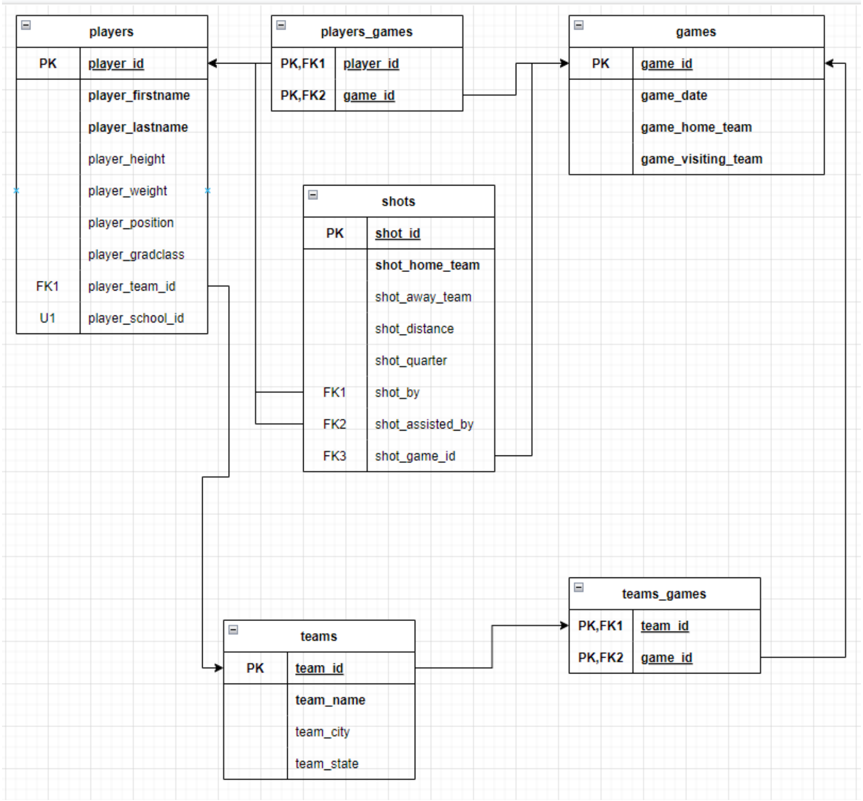
Throughout my 2 years in the program, my knowledge of the field of data science has expanded so far beyond what I thought I would be doing. Not only was I able to sharpen my skills in data collection, analytics, visualization, and database management, but I was also able to learn new skills such as machine learning, natural language processing, and deep learning.

IST 659

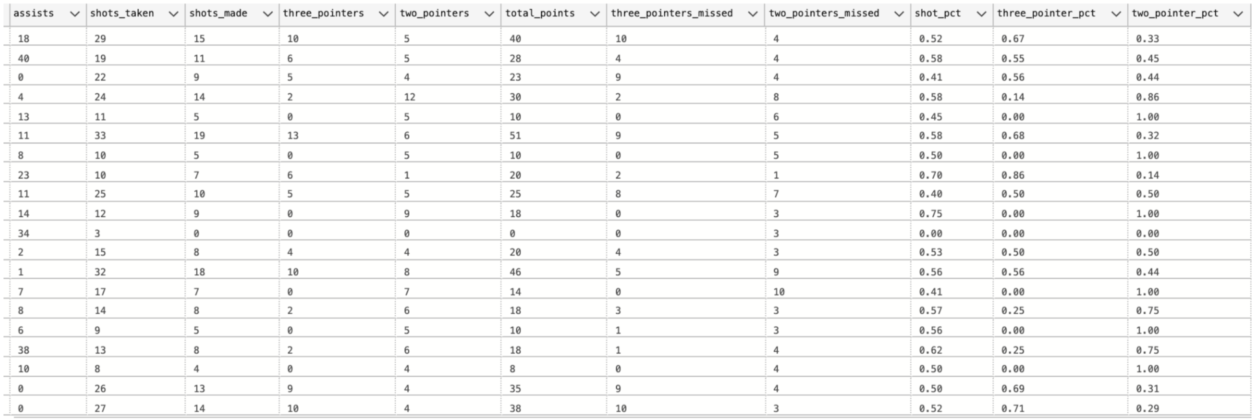
In IST 659, I acted as if I was a sports analyst for the Syracuse basketball team and I created a database of games from a basketball invitational that included the best high school players in the world. In undergrad, I minored in sports analytics and those courses were my introduction to data analytics and data science. I still have a major interest in sports analytics and would like to work in that field eventually.

The internal model of the database consisted of a table for player information, game information, each shot taken, and team information. We created stored procedures that will allow users to easily add player, team, and game data, without having to interfere with the internal model.


player
+ name [RC]
+ age
+ height
+ weight
+ position
+ graduation class
+ points
+ assists
+ rebound
+ 2 Point %
+ 3 Point %
+ school id [RU]
Is on/
Has
team
+ name [R]
+ city
+ state
Plays/
Played By
game
+ date [R]
+ home team [R]
+ visiting team [R]
+ home team points [R]
+ visiting team points [R]
+ winner [D]
Plays In/
Plays
Consist Of/
Make Up
shot
+ home team [R]
+ away team
+ assisted by
+ distance
+ quarter
Attempts/
Attempted By



In addition, through aggregation, I created views that will show statistics such as points, assist, and shooting percentage.



assists	shots_taken	shots_made	three_pointers	two_pointers	total_points	three_pointers_missed	two_pointers_missed	shot_pct	three_pointer_pct	two_pointer_pct
18	29	15	10	5	40	10	4	0.52	0.67	0.33
40	19	11	6	5	28	4	4	0.58	0.55	0.45
0	22	9	5	4	23	9	4	0.41	0.56	0.44
4	24	14	2	12	30	2	8	0.58	0.14	0.86
13	11	5	0	5	10	0	6	0.45	0.00	1.00
11	33	19	13	6	51	9	5	0.58	0.68	0.32
8	10	5	0	5	10	0	5	0.50	0.00	1.00
23	10	7	6	1	20	2	1	0.70	0.86	0.14
11	25	10	5	5	25	8	7	0.40	0.50	0.50
14	12	9	0	9	18	0	3	0.75	0.00	1.00
34	3	0	0	0	0	0	3	0.00	0.00	0.00
2	15	8	4	4	20	4	3	0.53	0.50	0.50
1	32	18	10	8	46	5	9	0.56	0.56	0.44
7	17	7	0	7	14	0	10	0.41	0.00	1.00
8	14	8	2	6	18	3	3	0.57	0.25	0.75
6	9	5	0	5	10	1	3	0.56	0.00	1.00
38	13	8	2	6	18	1	4	0.62	0.25	0.75
10	8	4	0	4	8	0	4	0.50	0.00	1.00
0	26	13	9	4	35	9	4	0.50	0.69	0.31
0	27	14	10	4	38	10	3	0.52	0.71	0.29

Database administration is one of the most fundamental skills for a data scientist. Databases are the backbone of data management. They allow for organized storage, retrieval, modification, and deletion of data. Good database skills ensure efficient data organization and accessibility, leading to better decision-making and improved operational efficiency.

IST 707 Machine Learning

In the project, I predicted the outcome of games played in the National Basketball Association (NBA). I used many different machine learning algorithms to predict the outcomes. The dataset used had an array of team statistics for both the home and away teams for each corresponding matchup, scraped from the NBA Stats API. I also feature-engineered additional features. I used five different machine learning models: Multiple Logistic Regression, Naive Bayes, K Nearest Neighbors, Random Forest, and Neural Network. After conducting our initial models, I created a k-means cluster analysis on NBA players and added the clusters to our features in the models.

This project allowed me to use a variety of data science techniques. I used web APIs on Python to scrape the NBA stats. However, I used initially R to scrape the NBA players from basketball-reference to cluster them. After discovering that Python also has the rvest package, I migrated over to Python, so that all of my code could be on one file. The fact that I can use either software that is best for the task at hand is a useful skill as it is important to be able to use multiple software.

In addition, since I was the sports analyst in the group, I took the lead in this project. As a team member, I am flexible and can adapt to the group environment, Since I had the best understanding of sports analytics I took the lead, however, if it was a topic I was less familiar with I would be okay with taking the backseat.

IST 709 Natural Language Processing

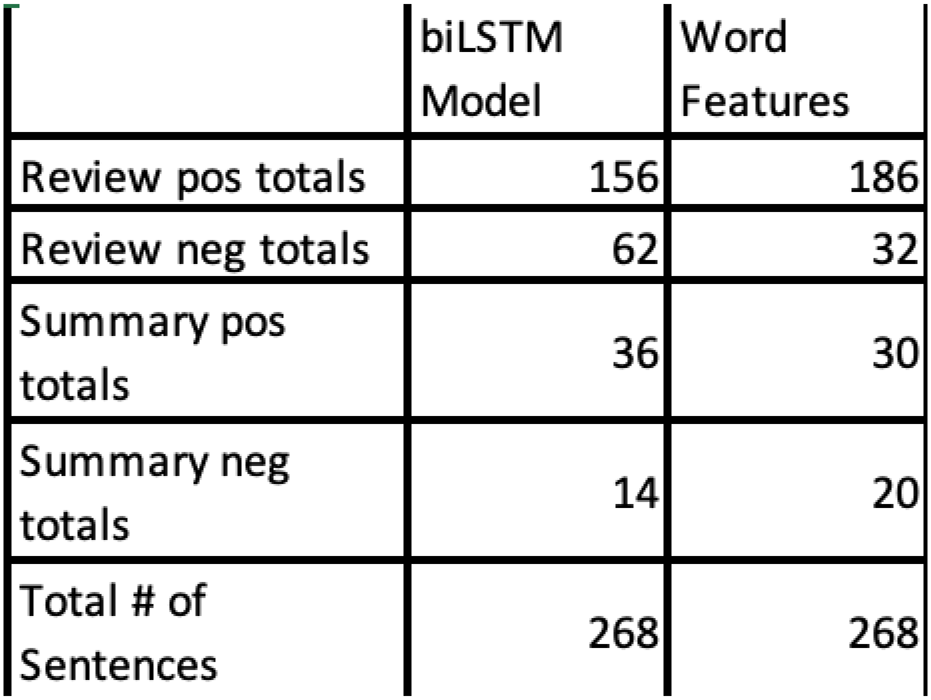
This course was the most challenging in the Applied Data Science program. It challenged me from both a technical and conceptual standpoint. Even with a solid understanding of coding and machine learning, the course was difficult for me. In the course, we touched on several topics such as Part-of-speech (POS) tagging, named entity recognition, and neural networks. It was a lot of information to learn over only one semester.

There were 3 assignments in the course, each one building upon the other. All of the assignments were based on the review contents from Amazon Product Data provided by Julian McAuley at http://jmcauley.ucsd.edu/data/amazon/. The dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 to July 2014. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

In the first assignment, I was tasked with extracting data from these fields: “reviewText”, “overall”, and “summary” and pre-processing tasks to the text data in “reviewText” and “summary”. This required me to create a for loop to extract the columns. I then lowercased all of the text and tokenized them. After that, I removed stop words and lemmatized the remaining words. After preprocessing the text, I had to find the most frequent 50 words, as well as the most frequent 50 bigrams. I then found the top 50 content words in the context of the word “video” or “videos”

In the second assignment, I first extracted the sentences that contained adjective words, capital word(s), exclamation marks, or any combination of them. This required me to do POS tagging on all of the sentences in the reviews and extract the sentences with adjectives, exclamations, and capital words. After that, I was tasked with performing sentiment analysis on the reviews. Since the reviews are not already labeled, I created a bag of words using a labeled training corpus of movie reviews from the NLTK package. I then included negation to further improve the model. I then used a naive Bayes classifier to classify the reviews.

Our last assignment was to once again, perform sentiment analysis on the reviews, but this time using deep learning to do so. I applied a BiLSTM classifier BERT word embeddings of the reviews. I then compared the models to my previous sentiment analysis. Below are the results:



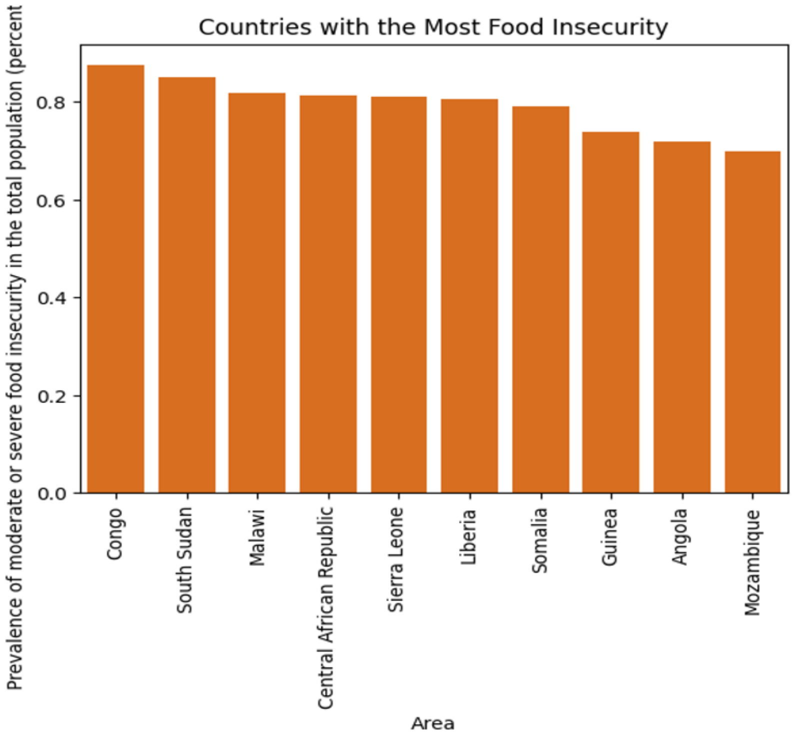
	biLSTM Model	Word Features
Review pos totals	156	186
Review neg totals	62	32
Summary pos totals	36	30
Summary neg totals	14	20
Total # of Sentences	268	268

These assignments were different in that I didn’t have labeled data. There was no way to test how good my models were because I couldn’t test for things like accuracy, precision, and recall. That is, however, the reality of working as a data scientist, a large portion of the data you receive will be unlabeled data. If I were to receive a dataset like this in the real world, however, I would probably use AMT to provide labeling services, because I would like to ensure that I am providing good models for prediction

IST 652 Scripting for Data Analysis

We investigated what type of factors will predict whether or not certain individuals and countries are food insecure. The datasets we used are from online resources from kaggle.com, as well as the World Bank data catalog. We used multiple datasets such as food insecurity by country as well as a dataset containing economic indicators and demographic data as well.

In the project, we found what countries were mostly affected by food insecurity (Figure). We found that all of the countries with the most food insecurity are in Africa (mainly Central Africa).



In addition, we found what economic and demographic factors influence food insecurity. We Prevalence of food insecurity We found that food insecurity negatively correlates with GDP per capita. Which means that the higher the food insecurity, the lower the GDP, on average. Prevalence of food insecurity positively correlates with inflation (Figure .

For this project I focused more on the technical aspect of the project, given that my coding skills were above that of my partner. At the time of the project, I was more comfortable with R, but since this was a python class, I needed to use the technical skills I learned from this course. That was actually one of the reasons I enrolled in the course, as my Python skills were on a very basic level. Python is the number one programming language and it was imperative that I got more comfortable with the programming language.

This project, more so than my other projects, addressed ethical concerns. Food insecurity is a global issue that affects millions of people and should be analyzed in the most ethical way possible. When dealing with human demographic data, you must be aware of the potential ethical issues. “The collection and use of demographic data in psychological sciences has the potential to aid in transforming inequities brought about by unjust social conditions toward equity. However, many current methods surrounding demographic data do not achieve this goal. Some methods function to reduce, but not eliminate, inequities, whereas others may perpetuate harmful stereotypes, invalidate minoritized identities, and exclude key groups from research participation or access to disseminated findings (Call, Eckstrand,2022).

IST 718: Big Data

The goal of this project was to predict the popularity level of new music that is released on the platform. To accomplish this, the variable measuring the overall popularity of the track will be converted to a factor variable with three levels: Low, Medium, and High based on the the 1/3rd and 2/3rd percentiles of the distribution of the track popularity scores. In addition, I Used K means Clustering, to cluster songs to provide listeners with song recommendations, based on the song they are currently playing.

There were several challenges that we encountered in the project. The first problem was that popularity, which was the target variable, was a numerical category from 0-100, so in order to do classification, we needed to convert it to a categorical variable. We decided on using a multiclass classification, with 3 levels. By doing so, we limited the number of classification models we could use, so we only were able to use multiclass classifiers.

After tuning and running our models, our most accurate classifier (Naive Bayes) only achieved an accuracy of 51.83%. We were disappointed in the lack of accuracy of our models, however, it was a good representation of real-world data, not always easy to classify. There are a lot of factors such as playlisting, label investment, and internet virality that can affect how popular a song becomes. In addition, the data dates back to the 1950s, what made a song popular then may not be what makes a song popular in the 2020s. Factoring release decade may or may not improve the model

Model	**Accuracy**
Decision Tree	41.83%
Random Forest	43%
Naïve Bayes	51.83%

Conclusion

When joining this program, I only had a background in data analytics, with a specialty in sport analytics. Now I am a data scientist, with skills in machine learning, neural networks and database administration. Throughout my two years in the program, I was able to develop those skillsets through assessments, assignments and projects, with the help of my professors and my peers.

Just as important as gaining the technical skills, I learned how to convey my findings to a range of different audiences. I can explain the technical details to my fellow data scientists, but I can also verbally and visually communicate findings to individuals that don’t know anything about data science.

With my technical and communication skills that I’ve built, I am confident that I can bring this assets to the work force and be great addition to the company I will be working for.