Forecasting NBA Game Attendance Using Historical Data Analysis

Elijah Sartin

[esartin@bellarmine.edu](mailto:esartin@bellarmine.edu)

22 April 2023

**Executive Summary**

This project aimed to predict attendance at any given NBA game by analyzing various factors, such as date, home and away team, and day of the week. The project collected data using Python and BeautifulSoup, which extracted information from basketball-reference.com, and stored it in a MySQL database. The data was cleaned and analyzed using exploratory data analysis and visualizations created in Tableau. Two models were used to predict attendance: a random forest regression and a linear regression. The results showed that the project was a success, and the web scraper was the most satisfying part of the project. However, the project could be improved by including additional variables, such as arena capacity and median income, and running the data through other models. The project demonstrates the importance of data preparation and the potential of machine learning models in improving attendance and increasing revenue for sports organizations.

**Introduction**

The goal of this project was to accurately predict attendance at any given NBA game. Various factors that are determined prior to the game starting were used to make these predictions. The variables included statistics such as the date of the game, the day of the week, the home team, and the away team. It was intended for the data set to be frequently updated with current data, changing as more NBA games occur. To achieve this, data was collected from online sources, stored in MySQL, and analyzed using Python. Tableau was used to create a dashboard that delivered broad insights into the data set. Ideally, the model would be adaptable and able to be used with other major sports leagues to make similar predictions. Through the use of a machine learning model, the overall goal would be to provide valuable insights about attendance so that sports organizations are able to improve the number of people at games as well as increase revenue.

**Project Details**

To begin creating any model data must first be collected, which was the first sub-project needed. I decided to use Python and the BeautifulSoup package to gather data from basketball-reference.com. Other packages such as HTTPlib2 and requests were used to extract the JavaScript from the backend of the website. BeautifulSoup allows the JavaScript to be maneuvered through intuitively, with the ability to find information with tags and various modifiers such as strongtext, childtext, and others. I located the information, as well as its context, using the inspect function in Google Chrome and then used BeautifulSoup to hone in on it in Python. Some of the data was not easily read or clean when it was discovered with BeautifulSoup, I used the regular expression package to split the data into a usable format. I then put all of the information for a given game into a tuple that contained all of the statistics for that game. Once I was able to accurately extract each of the desired variables from a specific game URL, I then began to create a loop that collected URLs in a given date range. Due to limitations from the website, I was only able to hit it less than 20 times a minute which limited the speed of the script. I collected a single month of URLs that were then put into the script that extracts the data and then appended to a list. This list was then taken and uploaded to a local MySQL database for storage. I collected the data for all regular season games since 2017 except the 2020-21 COVID bubble season as these games were not attended by fans.

Following the collection of all the data it then needed to be cleaned and analyzed. Due to the nature of data collection, it was largely cleaned by the time it arrived in the MySQL database. I removed outliers caused by COVID restrictions that lingered and caused decreased attendance at a few games, I wanted to ensure that these outliers did not interfere with the results. Aside from the removal of those few games, the data was already prepared for analysis. I began an exploratory data analysis to take a deep dive into the data. I analyzed the frequencies of all the categorical data and the distributions of all the numerical data. I also created bar plots to compare numerical statistics with the teams. As an extension of this exploratory data analysis, I connected MySQL to Tableau in order to create more visualizations. I created a dashboard displaying the attendance by the team, and day of the week, as well as a map that showed attendance based on location both across the US and other countries.

After completing the exploratory data analysis I put the data into a couple of different models. I had to tweak variables - converting the date into the day of the week and having the month, day, and year as separate variables. The first model that I ran it through was a random forest regression. Next, I ran the data through a linear regression model to use as a baseline to compare the random forest regression to. After completing all of the code for the project I then created a poster that discussed the goals of the projects, the methods used, and the results that were garnered.

**Reflection**

Overall the project was a success, data was collected, updated as time passed, analyzed, and then put into a model to predict attendance at various games. The creation of the web scraping tool was the most satisfying part of the project. The tool was able to collect any given period of games' worth of data with the only restriction being the amount of time the script would take to run. Aside from the possibility of collecting data from other websites, it is hard to see any room for improvement in the success of the web scraper. The other data could include various factors such as the popularity of each team, arena capacity, and median income in each city. These new data would provide more context and give the models more quantifiable information to predict better estimates. Exploring the data would most likely have provided much more interesting results if more statistics surrounding the teams were present, using statistics like winning percentage as opposed to wins and losses being variables could have given more useful insights than them being together. If given more time I also would have liked to include attendance at the home team’s last home game and attendance last time the two teams faced each other. Including this past data for reference would give the model much more context for predicting attendance at a game. In addition to including more variables, I would also be interested in running the data through a few more models than just the two that I used. Things such as a neural network and investigating the use of XGBoost would be first on the list of models to apply.

I learned quite a lot while doing this project. I had previously heard the sentiment thrown around about the preparation of the data taking up the bulk of a data scientist’s time and this held true for me throughout this project. Much more time was built creating the web scraper and diving into the data than the time that was spent creating my models. My recommendation to anyone who is starting a similar project is to gather all of the data that you could potentially want to include from the very start. The web scraper that I built did what it was intended to rather well, but I wish that I had gathered more data from other sites at the same time. My models ended up having decently strong results but I think they would have been much stronger if more information was provided. It is rather possible to go back and make those adjustments now, but it would have been less time intensive to gather all of the data at the same time. Of course, it is impossible to know exactly how everything is going to work out by the time you get to make your models, but having a clear roadmap and gathering everything you need from the start should greatly help the overall success of your project. Personally, by the time I got to make my models, I felt as if I was in a pretty good position. My only major issue throughout the project was getting linear regression to run properly. For most train/test splits it provided results that were decent yet not quite as good as my results from random forest regression. Unfortunately for other train/test splits it would return results that made little to no sense and were overall inconclusive. Below are the results comparing the two models in the same random state, one that worked well with the linear regression.

|  | **MAE** | **MAPE** | **MSE** | **R2** |
| --- | --- | --- | --- | --- |
| Random Forest | 808.04 | 4.98% | 1,458,290 | 0.6611 |
| Linear Regression | 968.32 | 5.93% | 1,718,082 | 0.6008 |

As part of learning how time intensive the data collection and preparation process is I learned how to use BeautifulSoup to effectively create a web scraper. I also built my confidence in the use of Python to interact with MySQL databases and upload data to them as well as execute calls into Python. It was both upsetting and informative to have the difficulties that I did with the modeling. I think that encountering roadblocks along the way can often be beneficial to the process of learning. As I proceeded through the project there were multiple times when I continuously received error codes and began to get frustrated, but every time figuring it out was very rewarding.

**Conclusion**

Overall this project was a great experience to have under my belt. It is the first time that I have had to undergo a project like this from start to finish. I have always previously worked with data that is already acquired and it was informative to collect and organize that data myself. I was able to create a web scraper in Python, organize the data, store it in MySQL, complete an exploratory data analysis, and run the data through a couple of different models. Along the way multiple roadblocks were encountered and overcome, helping me to know how I could better approach a project like this in the future. If I were to continue working on this going forward I would want to incorporate more contextual data prior to running the data through more models to see just how accurate the predictions could get. Once the model was heavily refined for NBA games I would then seek to apply it to other major sports leagues.