Forecasting NBA Game Attendance using Historical Data Analysis

Elijah Sartin & Dr. Robert Kelley

Bellarmine University Data Science Program

Abstract

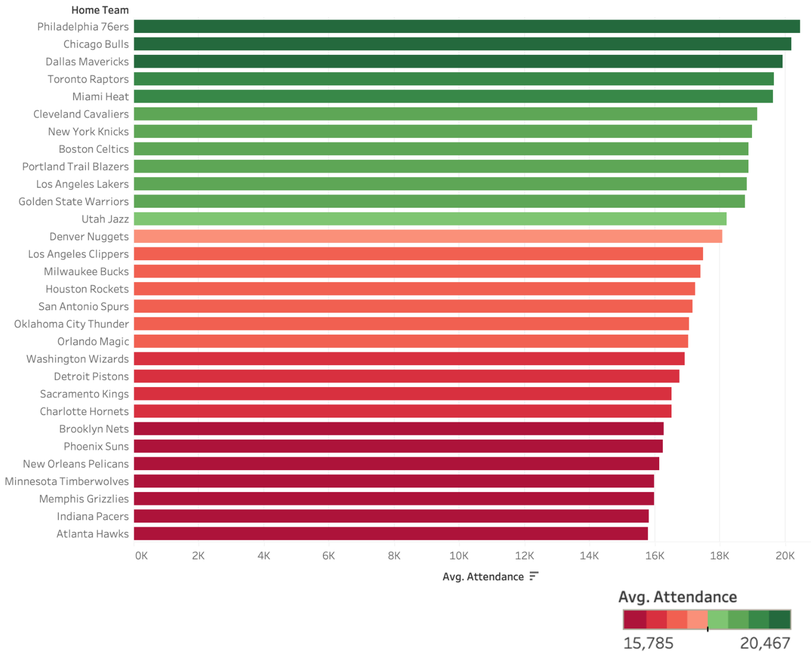
This project aims to develop a predictive model that accurately forecasts attendance at NBA games based on pre-determined factors, such as start time, home and away teams, and day of the week. To achieve this, data will be collected from various online sources and analyzed using Python and MySQL. The results will be displayed using Tableau. The project's end goal is to create an adaptable model that can be extended beyond the NBA to predict attendance for other sports. Through the use of advanced data analysis techniques, this project will provide valuable insights for sports organizations to improve attendance and increase revenue.

Objectives

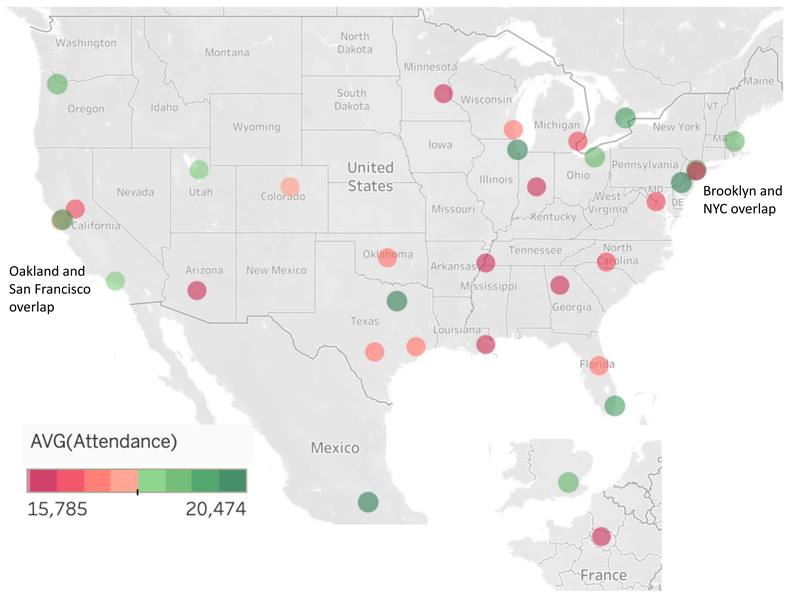
- Build a robust data collection system that can efficiently gather current and historical data on NBA games from https://www.basketball-reference.com. This system will be designed to capture a wide range of data, including venue, attendance, and team records, and will be optimized for scalability and reliability.
- Create an interactive data dashboard that enables users to gain deep insights into the collected data. This dashboard will feature a variety of visualizations, allowing users to explore trends and patterns in the data, and gain a deeper understanding of the factors that drive NBA game attendance.
- Develop a predictive model using a random forest machine learning method that can accurately forecast attendance for upcoming NBA games. This model will leverage historical attendance data, as well as other relevant features such as team records, venue, and game schedules, to make accurate predictions about future game attendance.

Materials & Methods

- To gather data for analysis, the project uses web scraping techniques implemented in Python to extract data from https://www.basketball-reference.com. Specifically, the BeautifulSoup and requests packages are employed to automate data collection. The data will be stored in a local MySQL database for subsequent analysis.
- Data exploration is conducted using Python to derive insights and identify patterns in the collected data. Statistical techniques are employed as necessary to perform an exploratory data analysis.
- To enable easy visualization and interpretation of the data, a dashboard is created using Tableau. The MySQL database is connected to Tableau to facilitate real-time data updates and enable dynamic dashboard generation. The resulting dashboard will be intuitive and user-friendly, presenting the data in an accessible and actionable format.

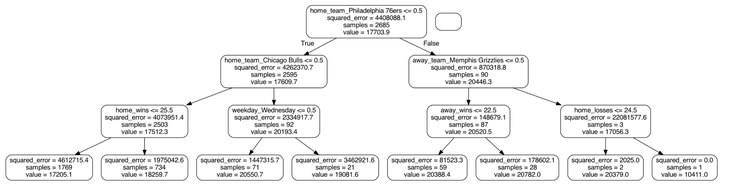


Results



Mean Attendance: 17,738

Min: 6,960, 1st Quartile: 16,449, Median: 18,071, 3rd Quartile: 19,432, Max: 22,983



The data was split into train/test and ran through a Random Forest Regression as well as a Linear Regression. After running each model with various splits, it was determined RF was more accurate on average. Below are the results for the two models using the same train/test split.

	MAE	MAPE	MSE	R^2
Random Forest	808.04	4.98%	1,458,290	0.6611
Linear Regression	968.32	5.93%	1,718,082	0.6008

References

King, Barry. (2017). Predicting National Basketball Association Game Attendance Using Random Forests. Journal of Computer Science and Information Technology. 10.15640/jcsit.v5n1a1.

https://www.basketball-reference.com/

https://www.python.org/

https://beautiful-soup-4.readthedocs.io/

https://www.mysql.com/

https://www.tableau.com/

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com