NBA Analysis

Team:
Danny Abouchakra
Danielle DiGioacchino
Keshav Gupta
Shantanu Vaidya

Project Topic

NBA Analysis

Using NBA data sets, our team will run supervised machine learning to answer whether or not the outcome of a game can be predicted with high accuracy.

Why we selected this topic?

Large dataset available and it is good when it comes to analysis.
We share interest in the sport.
It is a famous sport across North America.
Lastly, If we predict game outcomes with high accuracy we could profit from gambling.

Our Data source?

https://www.basketball-reference.com

What questions do we hope to answer?

1. Will we be able to predict the outcome of a game based on a team's roster and player statistics?
2. As we add more features to our data (for example: age or total points) will we be more successful at predicting the winner?
3. Will certain player statistics be less or more important towards the accuracy of our prediction?
4. Does including more years worth of player statistics help to better predict our outcome, versus just looking at the last season?

Data Exploration phase of the project

As per plan, we reduced the data to the columns that are important to our analysis. As we build and create supervised model we may plan to add other features to our data and use other machine learning models that may have an impact on increasing accuracy at predicting a winner

Analysis phase of the project

Data has been scraped from the website, cleaned and processed using pandas. The data has been uploaded to SQL database using SQLAlchemy and saved in CSV. After using input data, we used multiple linear model to predict the outcome of game based on roster, and player statistical history. We used Scikit learn to train test and split data. We also trained and tested with random forest regression to try and find what gives us higher accuracy with this type of data.

We were looking for linear correlation based on data and what we are trying to achieve, we used supervised learning and multiple linear regression model and achieved 65% accuracy.

Technologies, tools, algorithms, languages used

- Conda 4.10.3
- Python 3.9.6
- Jupyter-Notebook 6.3.0
- Pandas 1.2.4
- Leaflet 1.7.1, D3.js v5
- PostgreSQL 14.0
- pgAdmin 4 5.7,
- Imbalanced-learn 0.8.1
- Scikit-learn 1.0.1
- Tensorflow 2.9.0, Keras-tuner 1.1.0
- JavaScript
- HTML