Vagmi Bhagavathula, Jake Friend, Jordan Wecler

Data Science 2010

8 February 2023

Project Proposal

Our project proposal centers around assessing the National Basketball Association’s (NBA) offensive performance as the season progresses. By analyzing each team from the NBA, we can predict how much they are expected to score as the season progresses, perhaps for the postseason. The dataset we would like to be using is a Kaggle database that contains about 64000+ games and the box score statistics of each. This way, we can analyze the offensive performance through points scored, the rate of their scoring, the time during the season, etc.

For our project proposal, we would be studying regression to explore whether a team is shooting lights out or in a slump, thus eventually regressing or progressing to the mean. We plan to implement methods such as taking a certain number of games as the training data and retrieving the other games as the testing data to predict how many points they will score. We want to develop visualizations that describe the offensive performance across the league throughout the season and create models that optimize the number of training parameters but limit the testing error.

As per the project proposal, we seem to have no concerns, other than how we’re expected to develop the model and what methods you would prefer us to use. While we understand regression fits best for the current problem and the dataset, we’re unsure how to implement it. Currently, we would like to use this Kaggle database (link below). However, we’re also unsure if the data is too overwhelming or not compatible with the modeling we are expected to do. To that end, we could consider controlling the dataset by year or by era. Furthermore, consider team data as opposed to player data, or even by conference.

Link to Kaggle Database: <https://www.kaggle.com/datasets/wyattowalsh/basketball>