IEOR 142 Final Project Outline

Title
Predicting the 2023 NBA Champions

Background & Motivation:
Each NBA season culminates in the NBA Playoffs, which is an end of season tournament featuring the top 8 teams from each conference (Eastern and Western). During the regular season, all 30 teams each play 82 games over a period lasting approximately 6 months. Then, 16 teams qualify for the playoffs by being in the top 8 teams by win percentage in their respective conference at the end of the regular season (with the 7th and 8th seeds being selected via a Play-In tournament, which we won't get into here). There are a maximum of 4 rounds that a team can go through. Each round of the playoffs tournament involves winning a best-of-7 series (i.e first to 4 wins) against their opponent. The NBA finals involve the winner of the Western Conference facing off against the winner of the Eastern Conference in a best-of-7, and the winner of this series is crowned that year's NBA champion (having won 16 games total).

For sports fans and gamblers, the NBA playoffs is one of the most exciting times of the year, often riddled with an array of upsets, thrilling moments, and intense matchups. One of the biggest debates that ubiquitously arises amongst anyone who follows the sport is "who will be this year's NBA champion?" Using the power of past data and advanced statistics, we aim to provide a statistically-backed accurate prediction.

Our project will make use of common machine learning techniques such as multiple linear regression as well as random forests to predict the 2023 NBA champion (and those of prior years!) based on statistics collected during the preceding regular season.

Data
For the data, we will be scraping the last 25 years of regular season data from basketball-reference.com. This table is known as `NBA_playoff_contenders`. The variable our models will try to predict is `Champion Share Score`, which is the ratio of playoff games won by that team that season to the number of possible playoff wins that can occur (16 in recent seasons). For example, for the 2022 NBA champions, the Golden State Warriors, their `Champion Share Score` would equal 1.0, as they had to win 16 games in the playoffs to secure the championship.

Observe that `Champion Share Score` values can only be between 0 and 1, with 0 representing a sweep in the first round (no wins) and 1 representing a championship victory (16 wins).

The following is a table of some (but not all) important features that will be used to build our models:

Feature Name	Description
Seed	The playoff seeding of the team
Champion Share Score	# of Playoff games won / 16
Playoff Experience	Sum of roster's prior playoff games.
MOV	The team's margin of victory in the regular season
SRS	The team's simple rating system in the regular season
eFG%	The team's effective field goal percentage in the regular season
TOV%	The team's turnover percentage in the regular season
ORB%	The team's offensive rebound percentage in the regular season
FT/FGA	The team's free throws made per field goal attempt in the regular season
Opp eFG%	The opponent's effective field goal percentage in the regular season
Opp TOV%	The opponent's turnover percentage in the regular season
DRB%	The team's defensive rebound percentage in the regular season
Opp FT/FGA	The opponent's free throws made per field goal attempt in the regular season

A great deal of data engineering effort has been expended during the ETL process. Further detail on this can be found in the "ETL Process" section. This project does not make use of any pre-made datasets. All data was manually scrapped, processed, and aggregated.

Approach 1: Multiple Linear Regression
- Split into 70% train, 30% test
- Feature selection via VIF and p-vals
- Training: optimize for in-sample R2
- Testing: optimize for OSR2

Approach 2: Random Forest Regression
- Pro: feature selection is already “built-in” to regression trees
- Discuss pros surrounding bootstrap aggregation (bagging)
- Use in-sample R2 as well as OSR2 to evaluate model
- Could also use more “general” accuracy metrics like rate of correct champion predicted or rate that top 2 highest predictions are the champion.

Conclusion:
Short comparison of the 2 models, discuss the pros and cons of each and come to a conclusion as to which is better for our use case, citing metrics. Discuss the limitation of each model and what could be done to improve them.

Impact:
- Argue that k-folds cross validation could be used to create a more accurate model with less risk of overfitting, since there will be less overall discrepancy between training and test data.

“How might you expand the scope of your analysis to improve its impact
even more?” Answer: Maybe talking about using our optimal model to bet on Championship odds given by a popular sports betting casino, and see if it can be profitable? Maybe create a new model with betting odds as a feature?