Problem

Professional basketball in the nba is an entertaining yet mindful sport. Over the course of a season, the best teams typically have the most wins. In this project, the objective is to predict a team's win percentage in a season.

Using a linear least squares regression model, total wins can be predicted. This model will use common nba team statistics which are published every game. A few of the numeric variables are 3 pointers success percentage, 3 pointers attempted, two-pointer success percentage, two-pointers attempted, free throws success percentage, free throws attempted, offensive rebounds, defensive rebounds, assists, steals, blocks, turnovers, personal fouls, and points. A categorical variable included will be the season played.

Stakeholders might be the average nba enthusiast, sports analysts, and possibly gambling bookies. A sports analytics company might use this project if any new insights or methods are discovered. Many more advanced statistics exist on the official nba website, but they have a strict policy on use of their data.

A few obstacles exist regarding true independence between all statistical variables. Points scored is composed of two-point shots, three-point shots, and free throws but there are only a finite number of scoring opportunities in each game. For example, more 3 point attempts on average will lead to less two-point attempts. This can be considered a dependence between all 3 variables since an inverse relationship exists. Also, a league-wide shift lead to an increase in three pointers made during the 2010s. So each of the statistical categories may have a partial dependence on the season.

A worse problem is that, good teams tend to stick together while bad teams will trade away players. This phenomenon leads to conflating or overrepresentation of winning teams and players. One workaround is a stratified sampling of the dataset. Instead of selecting each team every season, a single team could be sampled randomly every 4 years. The 4-year period is an arbitrary choice to avoid clustering of superior performing teams.

This problem is novel because wins are a powerful prediction. Including the season played may uncover which statistics are useful in different time eras. Also, there is no concept of points allowed or defense which is an important facet of the sport.

The Plan

The data used to train the model will consist of the following dataset on Kaggle. https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18?select=nba_team_stats_00_to_23.csv

A few issues exist concerning missing data.

Due to differing number of games played during lockout seasons or covid, each statistic will need its value divided by the total number of games played in that season. These are called per game basketball statistics.

A worse problem in the dataset is that several teams have moved cities or changed names over the seasons. This will pose a difficulty in the stratified sampling technique used in this project.

For the teams without any relocations, we'll use stratified sampling. We'll randomly assign each of these franchises a number 1 through 4. This is an offset and the 4 is arbitrary and represents a period of 4 years. In order to reduce the bias of having identical teams in a season, we'll sample each team one time every 4 seasons. Each season will be labeled sequentially 1-4 with a period 4 frequency. The relocated teams will be considered on a case-by-case basis and use stratified random sampling when appropriate.

This project will use a multi-variable linear regression model to predict win percentage. We will use 20% of our data points as a validation or test set. Once we have trained the model, the mean squared error and correlation score can be determined from the test data. This determines the accuracy of the model and whether the model is a good predictor of team wins %. If the predictive model is not accurate, it is likely a fault of the experiment design. Further analysis would be needed such as correlation analysis between different variables. One variable that can be adjusted is the 4-year period which separates each team's sample.