# Machine Learning on Vegas NBA Analysis

The main question to the capstone project is centered on the accuracy of Vegas’ lines in comparison with the actual results and how gamblers can take advantage of gaps. This question can be framed as a machine-learning problem in separate smaller questions: what is the average point spread differential for each team? What are the average points scored per game per team? What is the average home court advantage in point spread per game? By breaking down these three questions I can begin to create a model to predict the point spread between two teams as well as the over/under in a specific game. To take this a step further, I can analyze the actual results vs. Vegas’ prediction vs. my predicted model to find the information gap between a solely data driven prediction. This is a supervised, regression problem to discover relationships between variables that impact the end result of each game.

The primary features to be used are the individual stats for each team, which stats are most impactful on the ending point spread per team, and the Vegas predicted and final results of the NBA games. After using these features and not quite getting the desired results for accuracy, I researched externally for common stat predictors used to predict scores for teams. This resulted in creating multiple new features as more individual stats for each team.

Initially, the goal is to create a foundation prediction model using linear regression. By discovering where there is noise in the data and where the most important features are, I can create the most accurate model possible. From this state, I will look into creating decision tree models using the stats of each team to predict what their point spread per game will be. The final learning technique will be a Random Forest Model. This technique entails using the dataset to fine tune the model on which features are most impactful rather than arbitrarily choosing the details of the model myself.

Evaluating success will be based upon the accuracy of the predicted model for point spread per team in comparison to the actual results. This way I can gauge my model against Vegas’ as well as compare it to the actual results to identify gaps. The metric used for the linear model will be R-squared values: multiple R-squared and adjusted R-squared values. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The key limitation: R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots. In general, the higher the R-squared, the better the model fits your data.