Will Swindell

Professor Barsky

CS 370

5/20/2022

Beating Vegas: Classifying NBA Outcomes

Motivation:

In 2018, the supreme court of the United States lifted the federal ban on sports-related wagering and opened the door for the legalization of sports betting. Today, sports betting is a $1510 billion industry in the United States and it is almost impossible to find a professional sporting event that can't be bet on in some respect. My interest in the NBA and also my curiosity about betting lines prompted me to explore the classification of NBA outcomes.

Process:

Like all machine learning projects, the longest and most arduous process involved collecting and processing the datasets used in building a model. The datasets I used come from Kaggle.com and comprise historical outcomes from every single matchup from 2004 until now. Other datasets include team records on each date, and team averages over the course of the season. Combining these datasets accurately was the most time-consuming process and due to the large amount of data, took several

minutes to actually run. After the data was cleaned-up, the model exploration phase began. I tried four different models for this project with varying degrees of success. First, to visualize the important variables, I experimented with a decision tree classifier, which showed that the most important inputs were the home and away team winning percentages, as well as the home team's record at home and the away team's record on the road. Less important for predicting outcomes were the number of injured players on each team and the number of points per game scored by each team. After determining the important factors, I experimented with a logistic regression model, a k-nearest-neighbors classifier, and support vector machines.

Best Results:

The most effective model was the logistic regression model with all variables included except for the number of injuries on each team. This model was accurate on the test data with a 20 fold cross-validation about 75 percent of the time. Given that according to the Las Vegas lines, the favored team wins about 68 percent of the time, this was an excellent result.

Visualization:

After completing the model, I created a visualization webpage that can be used to predict game outcomes from user input. At the moment, this webpage can only be locally hosted.

Link to Datasets:

https://www.kaggle.com/datasets/nathanlauga/nba-games?resource=download&select=ranking.csv

https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats?select=Team+Stats+Per+Game.csv