Beating Vegas: NBA Game Outcomes

Will Swindell

Motivation

- 2018: Supreme court strikes down federal ban on commercial sports betting
- Ensuing Sports Betting industry has grown to become a $150 billion a year industry
- Sports Books, or “Vegas” in general, have advanced statistical models to determine favorites, point spreads, individual performances, and much more
- Best question is: what *can’t* you bet on these days?

Motivation (Continued)

- Why basketball?
  - Personally experience playing and close follower of the NBA
  - With 82 games a year plus playoffs, a wealth of data is available
- The goal:
  - From historical data, the projected winning team by Vegas's model wins 67.9 percent of the time
  - Can we do better than this using ML models?

The Dataset(s)

- Combines 4 datasets found on [kaggle.com]
  a. Data about each individual game played since 2004 [link]
  b. Player performances in each of the above games [link]
  c. Team standings at the beginning of each of the above games [link]
  d. Seasonal Team statistics from each year for the above games [link]
- After combining, each row of the final dataset contains data from exactly one matchup between two teams, home and away.

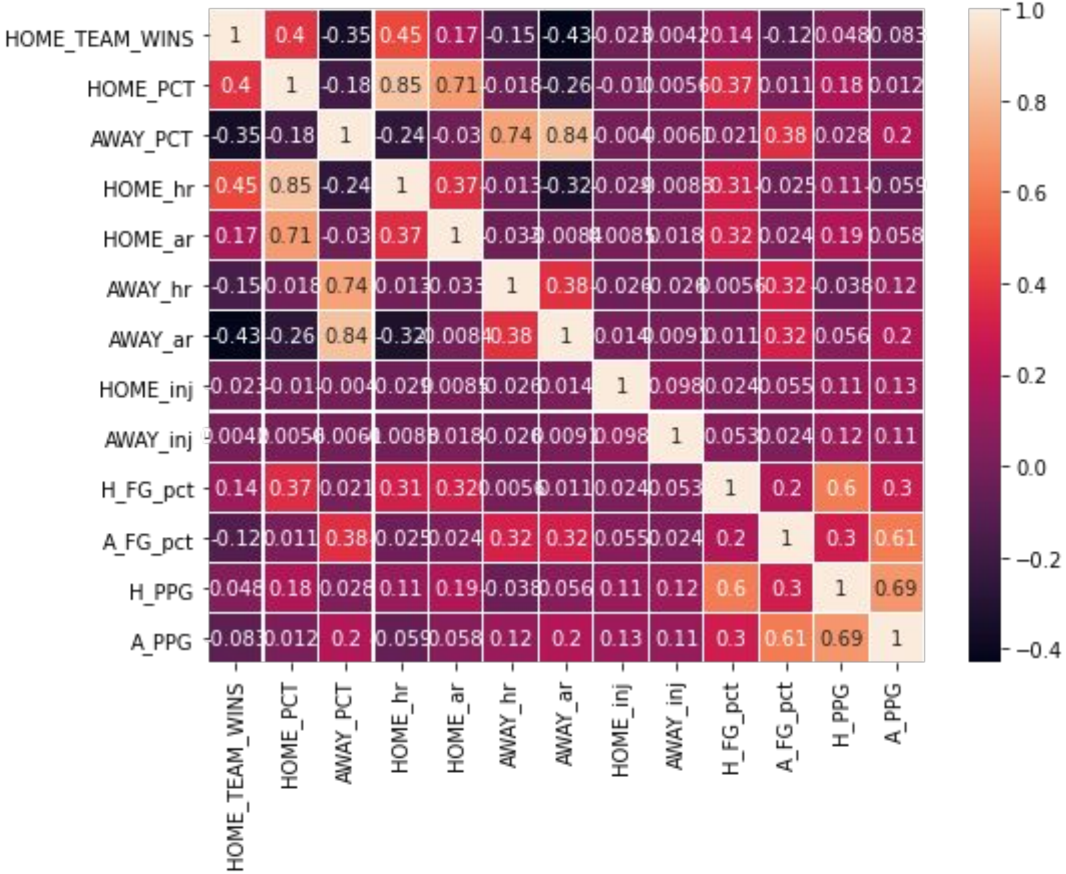
Variables of Interest

From the Datasets, the following data points were extracted for both the Home team and the Away team for each individual game:

- Record (Percentage)
- Home Record (Percentage)
- Away Record (Percentage)
- Number of Injured Players
- Season Average Shooting Percentage
- Season Average Points per Game

Each row also has an indicator variable for whether or not the home team won the matchup in question

Correlation with Home team Winning



Machine Learning Task

- Classification - who will be the winning team?
- Tried 4 different classification approaches
  - Decision Tree
  - K-Nearest-Neighbors (KNN)
  - Logistic Regression
  - Support Vector Machine (SVM)
- With all models, data points were normalized and biased using their correlation with class (home team winning) as weight

Decision Tree

```
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier

num_tree = tree.DecisionTreeClassifier(criterion='entropy', max_depth = 7)
num_tree.fit(X_train,Y_train)

print("Baseline train score:", num_tree.score(X_train, Y_train))
print("Baseline test score:", num_tree.score(X_test, Y_test))
```

```
Baseline train score: 0.7565419654971893
Baseline test score: 0.7393410852713178
```

- Findings: most important variables
    - Unsurprisingly Home record at home and away record away
    - Win percentage for each team
    - Points per game and field goal percentage are moderately important
- Unimportant: Injuries
    - Unsurprising - this value was pretty much always very close to zero and did not specify which players were injured

KNN

```
[0.72170543 0.72015504 0.69767442 0.75968992 0.74418605 0.70930233
 0.76356589 0.75426357 0.7620155  0.77054264 0.71782946 0.7248062
 0.74496124 0.71395349 0.70775194 0.75426357 0.74321179 0.73545384
 0.75640031 0.75640031]
cv_scores mean:0.7379066459787949
```

- Best Model: Used all variables
- Could not get a Cross-Validation score higher than the above
  - No matter which variables were in use
  - Worse before normalization and adding bias

Logistic Regression

```
[0.74031008 0.73488372 0.69689922 0.7744186  0.75426357 0.73953488
 0.76899225 0.76511628 0.76434109 0.7620155  0.72790698 0.73023256
 0.74496124 0.71705426 0.71937984 0.76046512 0.74631497 0.74864236
 0.7742436  0.75950349]
cv_scores mean:0.7464739807915516
```

- Noticeably better than KNN
- Removing Injuries resulted in around a 0.3 percent higher accuracy with cross validation
- Any other removals and removing bias made the model perform worse

Support Vector Machine

```
[0.74031008 0.74108527 0.70465116 0.77829457 0.75271318 0.73953488
 0.76744186 0.76511628 0.76124031 0.76124031 0.73100775 0.72945736
 0.73953488 0.7124031  0.72015504 0.75581395 0.7416602  0.742436
 0.77269201 0.76105508]
cv_scores mean:0.745892164468 5803
```

- Similar in performance to the Logit model
- Any modifications were unsuccessful in improving the model
  - Removing injuries
  - Using unbiased data

Let’s Give it a Try

FINAL	1	2	3	4	T
Warriors (53-29, 22-19 Away)	28	22	17	28	95
Grizzlies (56-26, 30-11 Home)	38	39	42	15	**134**

7:00 PM - ESPN
Heat (53-29, 24-17 Away)
76ers (51-31, 24-17 Home)

Use the model here to test two games from this year’s NBA playoffs

- One from last night
- One that will be played tonight

Possible Improvements

- More Data!
  - Add player performances to the classifier - might add more importance to the injury statistic if we know which player is hurt
  - Indicator for playoffs vs. regular season
- Realistically, it is difficult to expect a model to perform much better than this as it already correctly predicts winner 75 percent of the time given historical data
- Favorite has won only 67.9 percent of the time historically