Predicting the Characteristics of Basketball Winners

• • •

Minh Bui, Bryant Hicks, Cassandra Cohen

Project Motivation

- The motivation behind this project was to make novel discoveries within NBA statistics.
- Answers to the following questions were sought:
  - What characteristics are predictors for winning games in the NBA?
  - Which players recorded the best statistical performances?
  - What stats correlate the highest with win rate?
  - Does defense win championships?
  - Can "Jump Shooting Teams" win championships?

Dataset

- Based on NBA games played since 2004
- Dataset was from was broken up into multiple csv tables including:
  - Individual player statistics organized by game.
  - Team statistics and score for each game

Data Preparation

- Organized the multiple tables into an SQL database so they could be integrated together
- Removed redundant/irrelevant attributes
- Removed inconsequential players that could skew results:
  - Players that spent most of their time on the bench
 - Players who participated in less than a minimum threshold of games

Tools

- Github
  - Team collaboration
- Pandas and numpy
  - Data analysis
- Sqlite
  - Data Integration
- Jupyter Notebook
  - Code editor/IDE
- Discord

Exploratory Analysis

- Found total points vs net point differential based on teams and players
- Teams are sorted in order of highest winning percentage
- Discovery:
  - Plus-minus seemed like a good candidate attribute for predicting success
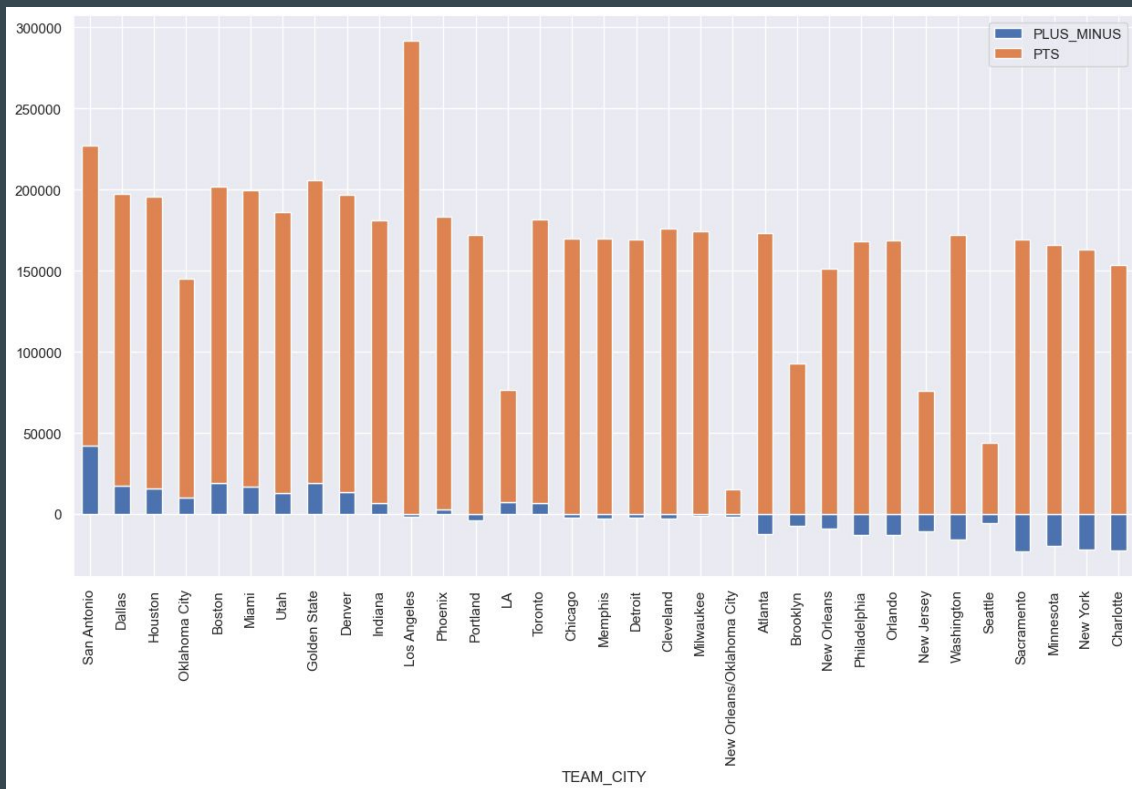
Market Basket Analysis

- Used apriori algorithm to find the most frequent groups of players that shared the court
- Discovery:
  - Most frequent groupings corresponded to the Spurs which had the longest running dynasty

| antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(Manu Ginobili)	(Tony Parker)	0.048114	0.052400	0.042426	0.881781	16.827929	0.039905	8.015659
(Tony Parker)	(Manu Ginobili)	0.052400	0.048114	0.042426	0.809665	16.827929	0.039905	5.001118
(Tim Duncan)	(Tony Parker)	0.045582	0.052400	0.041530	0.911111	17.387658	0.039142	10.660501
(Tony Parker)	(Tim Duncan)	0.052400	0.045582	0.041530	0.792565	17.387658	0.039142	4.601047
(Manu Ginobili)	(Tim Duncan)	0.048114	0.045582	0.039037	0.811336	17.799464	0.036844	5.058825
(Tim Duncan)	(Manu Ginobili)	0.045582	0.048114	0.039037	0.856410	17.799464	0.036844	6.629203
(Manu Ginobili, Tim Duncan)	(Tony Parker)	0.039037	0.052400	0.035686	0.914172	17.446065	0.033641	11.040643
(Manu Ginobili, Tony Parker)	(Tim Duncan)	0.042426	0.045582	0.035686	0.841139	18.453288	0.033753	6.007868
(Tim Duncan, Tony Parker)	(Manu Ginobili)	0.041530	0.048114	0.035686	0.859287	17.859255	0.033688	6.764734
(Manu Ginobili)	(Tim Duncan, Tony Parker)	0.048114	0.041530	0.035686	0.741700	17.859255	0.033688	3.710690
(Tim Duncan)	(Manu Ginobili, Tony Parker)	0.045582	0.042426	0.035686	0.782906	18.453288	0.033753	4.410871
(Tony Parker)	(Manu Ginobili, Tim Duncan)	0.052400	0.039037	0.035686	0.681041	17.446065	0.033641	3.012810
(Udonis Haslem)	(Dwyane Wade)	0.041803	0.050608	0.033933	0.811743	16.039887	0.031818	5.043059
(Dwyane Wade)	(Udonis Haslem)	0.050608	0.041803	0.033933	0.670516	16.039887	0.031818	2.908173
(Marc Gasol)	(Mike Conley)	0.041024	0.042309	0.030076	0.733143	17.328108	0.028341	3.588783
(Mike Conley)	(Marc Gasol)	0.042309	0.041024	0.030076	0.710866	17.328108	0.028341	3.316714

Decision Tree Classification

- Attempt to classify whether a team will win or lose a game
- Calculated information gain on attributes to see which were the most relevant
    - Highest gain from plus-minus followed by field goal and free throw percentages

Attribute	Information	Gain
PLUS_MINUS	0.799271	0.200727
FG_PCT	0.836033	0.163965
FT_PCT	0.836308	0.163689
FG3_PCT	0.836314	0.163684
PTS	0.836409	0.163589
FGA	0.836475	0.163523
FGM	0.836476	0.163522
FG3A	0.836476	0.163522
FG3M	0.836477	0.163521
REB	0.836478	0.163519
FTA	0.836480	0.163517
FTM	0.836483	0.163515
DREB	0.836485	0.163513
AST	0.836490	0.163508
BLK	0.836493	0.163505
OREB	0.836499	0.163499
TO	0.836501	0.163497
STL	0.836506	0.163492
PF	0.836507	0.163490

Linear Regression Model

- Model uses plus/minus to predict winning percentage
- Blue line represents model's prediction
- Red dots represent actual data points
- Graph shows high accuracy of linear regression model using plus/minus as predictor
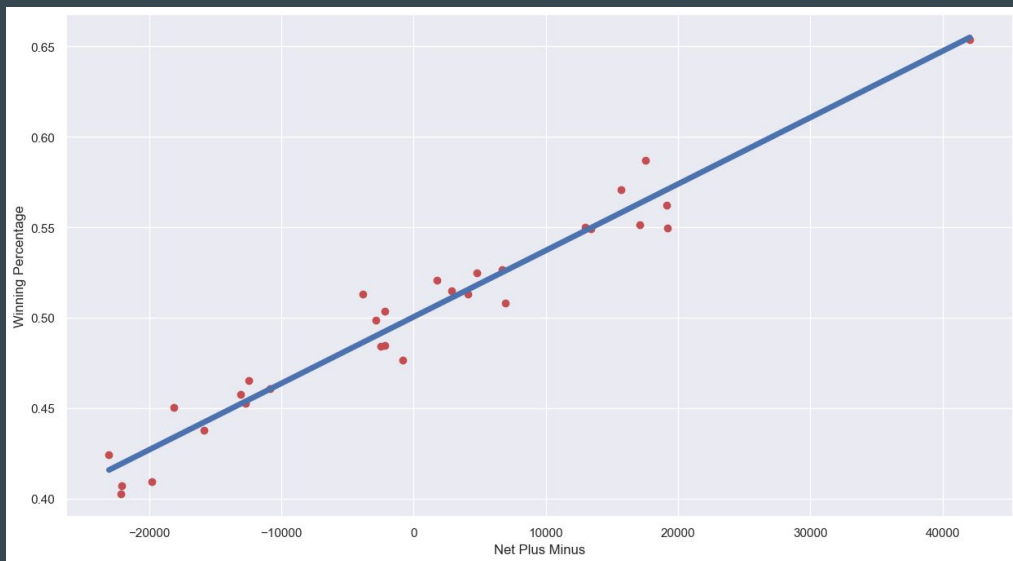
Application of Knowledge Gained

- Teams should highly consider plus-minus when signing individual players
  - Currently more emphasis is placed on offense only based stats like 3 point percentage
- Player impact on team performance should be valued above individual athlete statistics
- Draymond Green is a prime example of this phenomenon because his personal numbers do not appear very impressive but his high plus/minus reflects his value to the team