Predicting Momentum Shifts in NBA Games

Whitney LaRow, Brian Mittl, and Vijay Singh

Motivation and Task Definition:

- NBA games are often decided by a few moments of rapid scoring: can these streaks be determined via machine learning techniques and used to better predict the outcome of a game?
- We define a streak as a point where a team scores at least 8 points while limiting their opponent to 0 points

Sample Data:

Source data from basketball-reference.com:

4:00.0	J. Hill makes 2-pt shot at rim (assist by E. Davis)	+2	22-18	
3:46.0			22-20	+2 S. Curry makes 2-pt shot from 19 ft (assist by L. Barbosa)
3:27.0	Turnover by K. Bryant (lost ball; steal by A. Iguodala)		22-20	
3:19.0			22-22	+2 S. Curry makes 2-pt shot from 11 ft

Scraped JSON data from basketball-reference.com:


```
"awayTeam": "Los Angeles Lakers",
"homeTeam": "Golden State Warriors",
"date": "November 1, 2014",
"plays": [{'awayScore': "22",
        'homeScore': "20",
        'team': "away",
        'time': "3:27.0",
        'playType': "turnover"},
    {'awayScore': "22",
        'homeScore': "22",
        'team': "home",
        'time': "3:19.0",
        'points': "2",
        'playType': "madeShot"},
... ]
```


Pre-Processing:

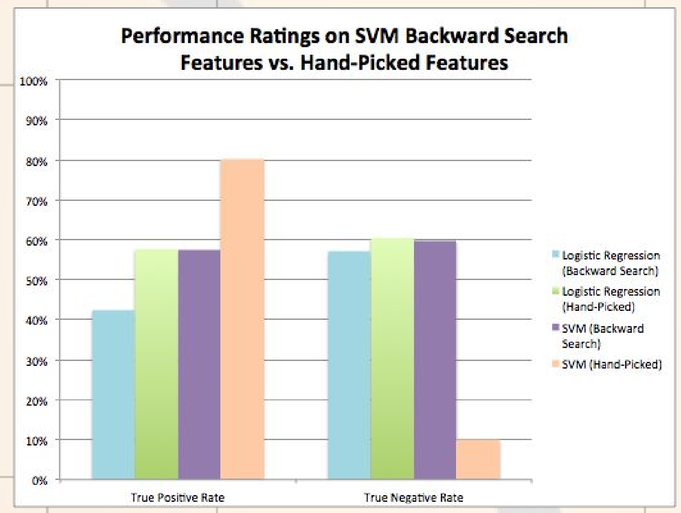
- Converted JSON file into CSV file
- Full data set consists of over 1,000 NBA games and over 500,000 in-game events where a streak prediction occurs

Methods:

- Backward search for feature selection
- K-fold cross validation for parameter selection
- Logistic regression
- Support vector machine

Results and Analysis:

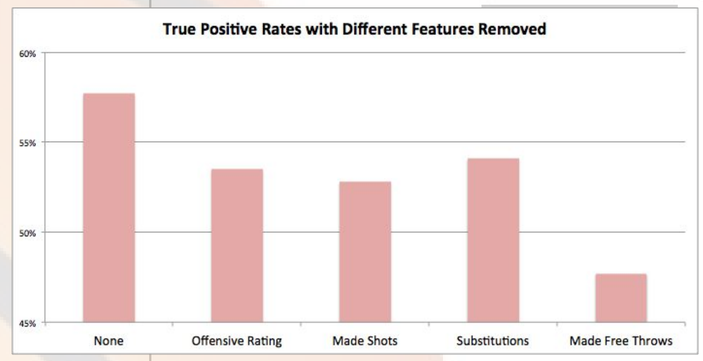
- Optimal features include consecutive points, turnovers, field goal percentage, timeouts, free throw percentage, rebounds, and substitutions
- Momentum streaks are difficult to predict due to high variance between games, but they are not entirely arbitrary

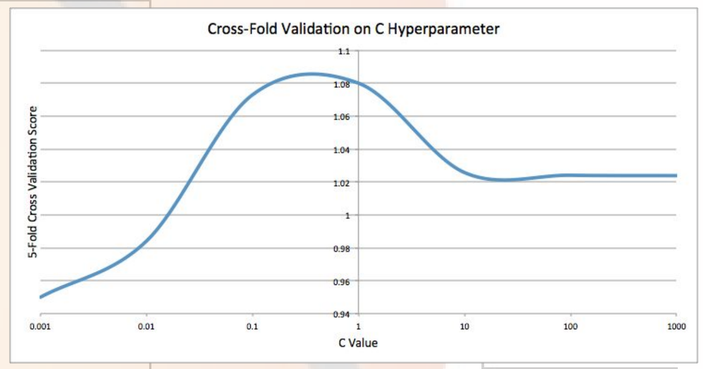


	True Positive Rate	True Negative Rate
Logistic Regression Train	55.00%	60.55%
Logistic Regression Test	57.69%	60.48%
SVM Train	56.55%	60.46%
SVM Test	57.34%	59.82%

Challenges:

- Data collection
  - Using Python library “Scrapy” to collect data from basketball-reference.com
- Recency of events
  - Currently using “window” approach that only considers last 10 possessions in the feature vector





Future Work:

- Continue fine tuning parameters with crossfold validation
- Explore more features that use individual players and time as parameters