NBA Greatest of All Time Campaign Recommendations

Terry A. Meyer, Student, USF Data Bootcamp
March 26, 2024

Problem Statement

Provide data visualizations that support NBS sports apparel line decisions for NBA Most Valuable Team, Most Valuable Player, Best Player by Position, and best Fantasy Dream Team of All Time using public data from 1950-present day.

Context: The CEO of Greatest Apparel of All Time (GAOT) sporting gear is interested in launch a new apparel line around the NBA’s Greatest of All Time (GOAT). She wants to use publicly available data to save cost and potentially stay closer to the fan base.

Success Criteria

- Prevent proposal using an executive level dashboard using Tableau
- Analysis uses a common set of metrics that level team and player statistics by accounting for changes to the game of basketball
- Four different GOAT player/team recommendations for merchandising

Scope:

- Only NBA teams and players from 1950-Present
- Publicly available data
- Team and player costs are not a criteria

The Dataset

NBA Stats (1947-present) (kaggle.com)

There are 3 leagues represented: the National Basketball Association (1950-present), the NBA's predecessor in the Basketball Association of America (1947-1949) and the NBA's past competitor in the American Basketball Association (1968-1976)

On the team side, there are 7 files:

- totals & opponent totals
- per game & opponent per game statistics
- per 100 possessions & opponent per 100 possessions statistics (starting from 1974)
- team summaries

On the player side, there are 10 files:

- player totals
- player per game stats
- player per 36 minute stats
- player per 100 possessions stats (starting from 1974)
- player advanced stats
- player play by play stats (starting from 1997) - percentage of time spent at different positions, fouls drawn & committed, etc
- player shooting stats (starting from 1997) - success rate and attempt rate from different shot distances
- end of season teams (All-Defense, All-Rookie, All-League)
- end of season team voting (All-League)
- all-star selections
- awards voting results (Rookie of the Year, Sixth Man of the Year, Most Valuable Player, Defensive Player of the Year, Most Improved Player)

Glossary | Basketball-Reference.com (reference dataset feature definitions)

Data Wrangling

- Selected the Player Totals.csv and Team Totals.csv from data package
- Dropped ‘birth year’. Not useful.
- Dropped the rows from ABA and BAA leagues; this also resolved missing ‘age’ values
- Filled missing ‘fg_percent’ values with 0. The players played very little.
- Filled missing ‘x3p_percent’, ‘x2p_percent’, ‘ft_percent’ values with the calculation or 0 if no shot attempts.
- Kept all of the features with NA values if they were statistics introduced in later years. I did not use most of these in later steps as they did not allow for fair comparison of players and teams.
- Dropped a ‘League Average’ value from the ‘team’ column.
- Dropped a ‘Baltimore Bullets’ single row. The team folded in 1954 and is not part of a current franchise.
- Dropped the ‘mp’ column for the Team Totals dataframe. Not useful and many missing values.
- Saved data as nba_player_totals.csv and nba_team_totals.csv for EDA use

Exploratory Data Analysis

- Found a ‘TOT’ value in ‘tm' that totaled player points that played for multiple teams in a season. This was making the sum of player career points in error. Dropped these roads.
- Verify split histogram plots were correct. There were multiple years with player lock outs that limited player and team stats based on fewer games played.
- Consolidate player positions into SG, PF, C, PG, and SF. The dataset tracked position combinations where a player filled multiple roles.
- Produced multiple interactive plots using the Bokeh library to examine multiple features and correlations.
- Consolidates all of the team names that are part of a present day team franchise under the current name to allow comparison of most successful franchises. Let the ‘abbreviation’ alone so the contributions of these teams could be recalled later.
- Tale of Two Eddies. Discovered there were two players named Eddie Johnson in the NBA. The dataset did not distinguish between them. Identified the correct rows and labeled them ‘Eddie A Johnson’ and ‘Eddie L Johnson’.

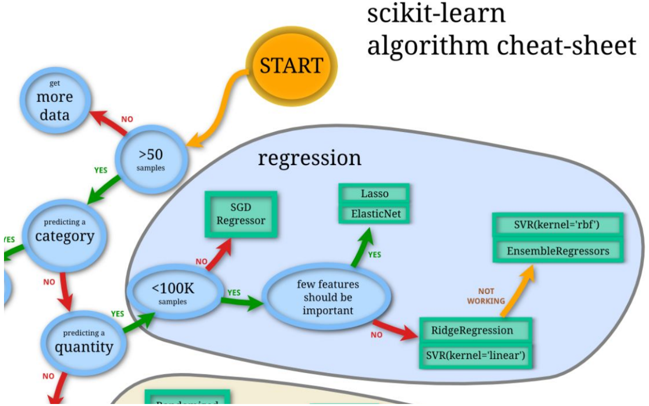
Exploratory Data Analysis

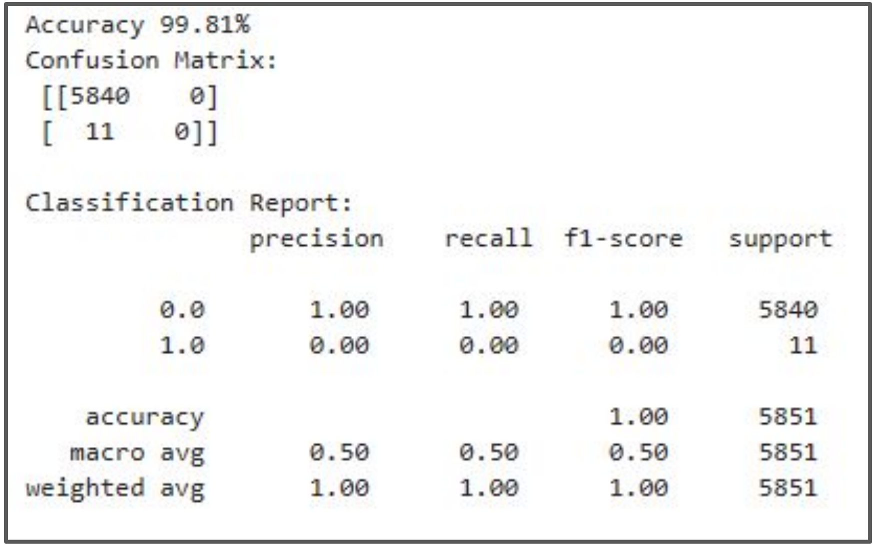
- Created a ‘changes’ feature that segmented seasons by rule changes to investigate the impact of these changes on the teams and players
- Created a ‘decades’ feature that segmented the data by 10 year periods for later analysis
- Created a boolean feature ‘allstar’ to identify if a player was voted to that years allstar team
- Created a boolean feature ‘mvp’ to identify if a player was declared the League MVP.   This is the target feature for modeling in the next step.
- Created a boolean feature ‘championship’ to identify if a team won the NBA Championship that season.
- Created box plots to identify outlier values. The outliers were not errors, but indications of stand out performances by teams and players.
- Create heatmaps for features selection; strong positive coefficients indicate influence on potential target variables for modeling. I did not select those features with greater than .80 as these represent multicollinearity issues for modeling.

Modeling (1 of 2)

Logistic Regression

- Fast
- Predicts probability of an instance
- Widely used for binary classification tasks



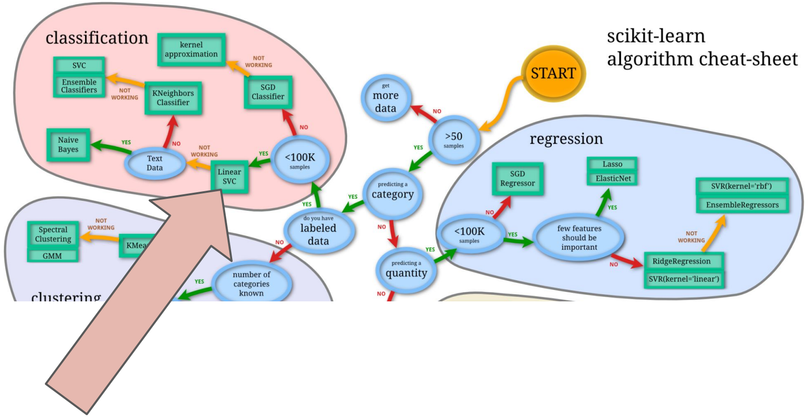


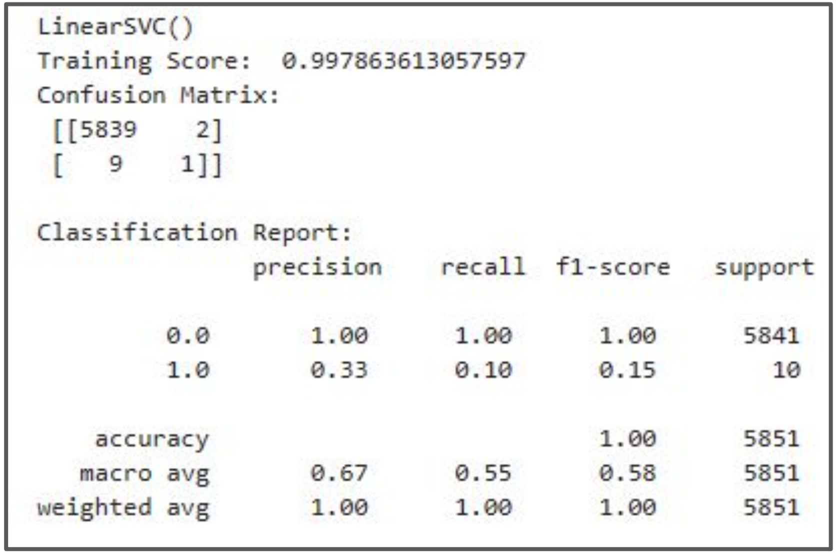
```
Accuracy 99.81%
Confusion Matrix:
 [[5840    0]
 [  11    0]]

Classification Report:
               precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      5840
         1.0       0.00      0.00      0.00        11

    accuracy                           1.00      5851
   macro avg       0.50      0.50      0.50      5851
weighted avg       1.00      1.00      1.00      5851
```

Modeling (2 of 2)

LinearSVC (Best)

- Fast support vector classifier
- Predicting a boolean, labeled category
- Less than 50,000 samples





```
LinearSVC()
Training Score:  0.997863613057597
Confusion Matrix:
 [[5839    2]
 [   9    1]]

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      5841
         1.0       0.33      0.10      0.15        10

    accuracy                           1.00      5851
   macro avg       0.67      0.55      0.58      5851
weighted avg       1.00      1.00      1.00      5851
```

Analysis and Recommendations

NBA GOAT Player and Team Franchise Analysis | Tableau Public

Practical Considerations & Suggestions

- Greater segmentation of the players by season for more direct comparison
- Use more in depth statistical methods to examine the standard deviations of top player performances and the probability of those efforts being repeated
- More in-depth study of the players that made up the teams in the franchises to get a more nuanced appreciation of each team over the NBA seasons