Fantasy Basketball Predictor

By Robert Cheer

Overview:

My project used data from basketball-reference.com to predict what a players statistics would be in the next NBA season. Players were ranked according to value in each of 9 traditional statistical categories used in standard Fantasy Basketball leagues. Players received a value for each category based on the number of standard deviations above the mean they were in that category.

Exploratory Data Analysis and Key Features:

The features that I ended up using included previous performance, including last years traditional and advanced stats as well as the usage rank, which included whether or not the player was a primary option on his team. Then I included features on the players career, age (multi-dimensional polynomial), and the amount of years they've been playing in the league.

Finally, I included the teammates as a feature, which would account for a recent trend with star players leaving your team. The features I used included maximum teammate minutes*usage rate, points, field goal attempts and turnovers, all indicators of a ball dominant player.

What I found was that having a ball dominant player join or leave your team was the best indicator that a player would deviate from their career trajectory. A great example of that is Russell Westbrook elevating to MVP after Kevin Durant left. It also happens the other way though as Kevin Love and Dwyane Wade saw big drops when they joined Lebron James on the Heat and Cavs respectively.

Tech Used:

In order to create my models I used Python classes, functions as well as packages SQLAlchemy, Flask, psycopg2, pandas, numpy and sklearn.

In addition I used a lot of PostgreSQL in order to set up the data, breaking up my original data into many subtables that could store aggregated values (career stats, teammate stats, players that switched teams, etc)

Model Selection:

For each category, I tested Gradient Boosting and Random Forest Regressors against Rectified Linear Unit based Neural Networks and Linear Regression. I ended up selecting a mixture of models based on which model performed better on the test set.

Evaluation:

I used data from 2007-08 season through 2016-17 to create the model, choosing the model based on k-fold cross validation mean-squared error, then applied the model to 2017-18 data. After ranking the players and applying Rank-Bias Overlap to compare my model to Yahoo's preseason rankings, my model outperformed at every level (top 10, 50, 100, 150, 200).

Next Steps:

To go even further into teammate data, usage rate doesn't capture how difficult it is to play with someone. If I could find data on how many dribbles each player takes and the total movement data from each team I think there would be a lot more signal. In addition, finding a way to rank the rookies who I did not have data for currently.

https://github.com/RCheer3/Fantasy-Basketball-Capstone-Project

Robert L Cheer

r.cheer3@gmail.com. 530-219-1969

Education

University of California, Berkeley, Berkeley, CA **January 2009 - August 2013**
Bachelor of Arts, Double degrees in Statistics and Applied Mathematics

Data Science Coursera Certificate, John Hopkins **September 2016-March 2017**

Objective

Obtain a role on a data science or analytics team at a company that is focused not only on the present, but on the future and is using data to help make better decisions.

Work History

Galvanize, San Francisco, CA **September 2018-December 2018**
Student
- Three month 700+ hour data science immersive training program.
- Covered topics in core data science areas including Bayesian statistics, A/B testing, Random Forests, Natural Language Processing, data mining, mongoDB, Alternating Least Squares, matrix decomposition, Linear Regression and Logistic Regression. Primarily coded in Python libraries numpy, pandas, and sklearn.
- Four full-day Case studies cleaning data in Python and creating predictive models.

Verisk Insurance Solutions, San Francisco, CA **February 2016-August 2018**
Senior Data Analyst
- Independently formulated hypothesis to improve effectiveness and efficiency of our products and created prediction models to improve existing logic.
- Used machine learning techniques to predict the probability of an auto policy having uninsured drivers.
- Created SQL functions and Stored Procedures to automate and upgrade operations.
- Created and marketed new products by providing presentations and ROI predictions for prospective customers.
- Gathered positive metrics and predicted potential results helping with contract negotiation and marketing material.
- Performed A/B testing on changes to language on mailer and script updates to improve contact rate and customer experience.
- Created projects and exercises to train entry level members of the team best practices in SQL and data analyses.

Verisk Insurance Solutions, San Francisco, CA **December 2013-January 2016**
Project Analyst
- Performed various analyses on large datasets in SQL server to improve as well as promote Verisk products.
- Worked with developers on daily and monthly processes to ensure maximum client satisfaction.
- Performed technical projects for product enhancements as well as technological migrations.
- Performed generalized linear model regression analytics in R
- Created visualizations in Tableau for client presentations.

Skills

- Proficient in database querying with SQL. Experience writing stored procedures and functions.
- Machine learning techniques including random forests, boosting, gradient descent and time series modeling
- Proficient with probability theory, A/B testing and random sampling
- Experience creating multi-variable predictive models using linear and logistic regression modeling on large datasets.
- Experience with revenue and net income projections.
- Programming in Python, including pandas, numpy, spicy.stats and matplotlib.
- Data visualizations in Tableau, Excel and Python
- Exposure to AWS Elastic Compute Cloud, NoSQL and MongoDB.
- Exposure to Natural Language Processing
- Experience creating data models with statistical computing language R
- Exposure to Spark, creating recommender systems using Alternating Least Squares
- Experience with Non-negative Matrix Factorization, Singular Value Decomposition and Neural Networks.