Predicting Player Performance with NBA Data

Johnny McGregor
General Assembly

Project Goal

Find out how well a machine learning model can predict a player's daily fantasy output given certain statistics from the player and his opponent

- Traditional and Advanced Statistics
- Draft Kings fantasy points formula
- Root mean squared error to evaluate model

Draft Kings Fantasy Points Formula

- 1 point for each point scored
- 1.25 points for each rebound
- 1.5 points for each assist
- 2 points for each steal
- 2 points for each blocked shot
- 1.5 points for a double double
- 3 points for a triple double
- Lose .5 points for every turnover

NBA Advanced Statistics

- True Shooting Percentage
- Effective Field Goal Percentage
- Total Rebound Percentage
- Total Assist Percentage
- Total Block Percentage
- Total Turnover Percentage
- Usage Percentage
- Offensive Rating
- Defensive Rating

Exploratory Data Analysis / Preprocessing

- Web scraping to collect data
  - Current season and five previous seasons back to 2014-2015
  - 117,000+ rows of data using 879 unique players
- Convert all data to appropriate data types
  - Most of the values need to be floats
- Rename some columns
  - Multiple columns had same name
- Create features out of the original data
  - Up to date averages
  - Five game rolling averages

More Preprocessing

Offensive Stats

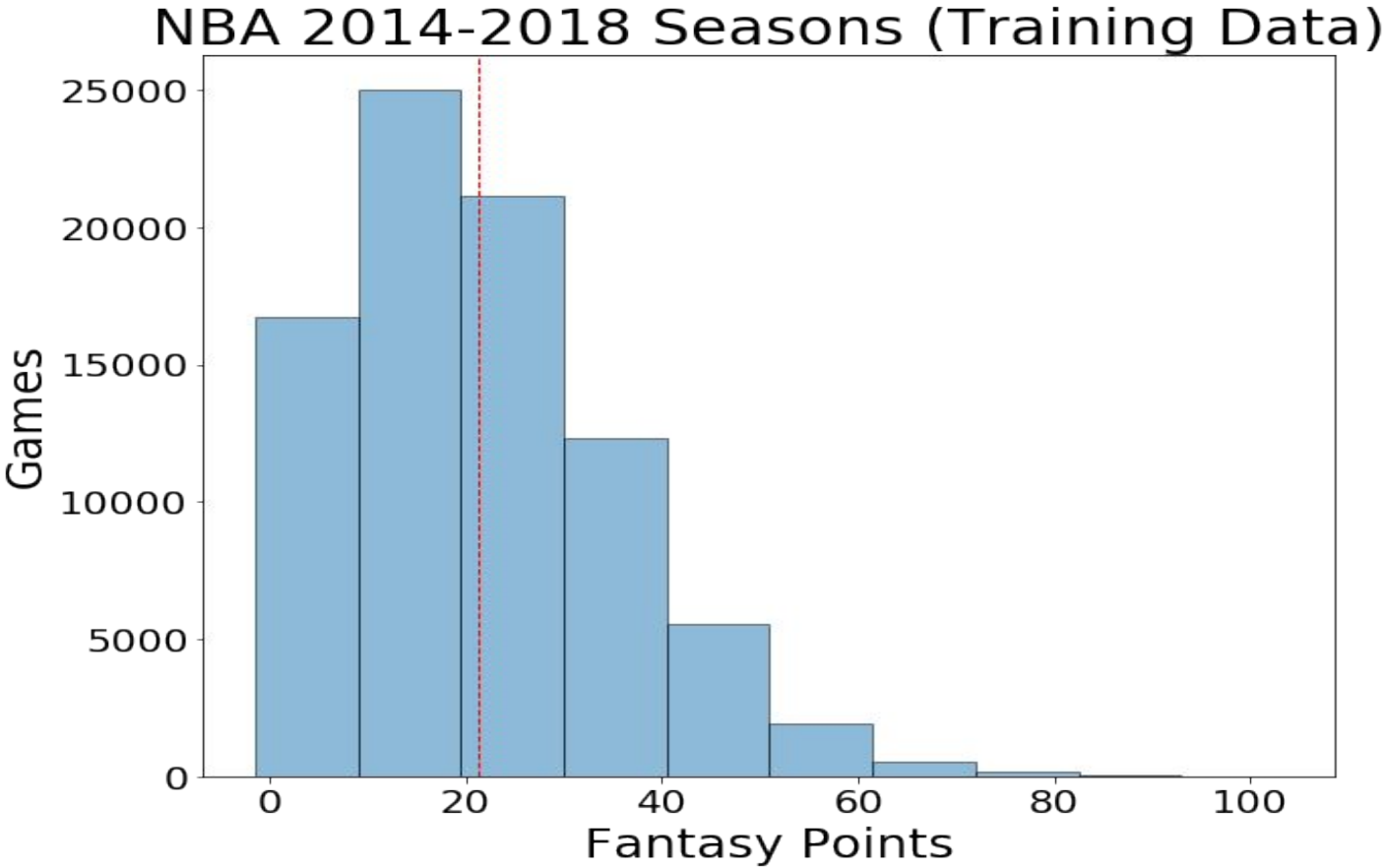
date	player	gs	mp	trb	ast	stl	blk	tov	pts	dub_dub	trip_dub	fantasy_points
2017-12-06	kevin_durant	1	35	11	10	0	2	3	35	1	1	72.25
2017-12-08	kevin_durant	1	38	10	7	1	5	2	36	1	0	73.00
2017-12-11	kevin_durant	1	36	9	5	0	3	4	28	0	0	51.75
2017-12-14	kevin_durant	1	34	11	7	0	2	5	36	1	0	65.25

More Preprocessing

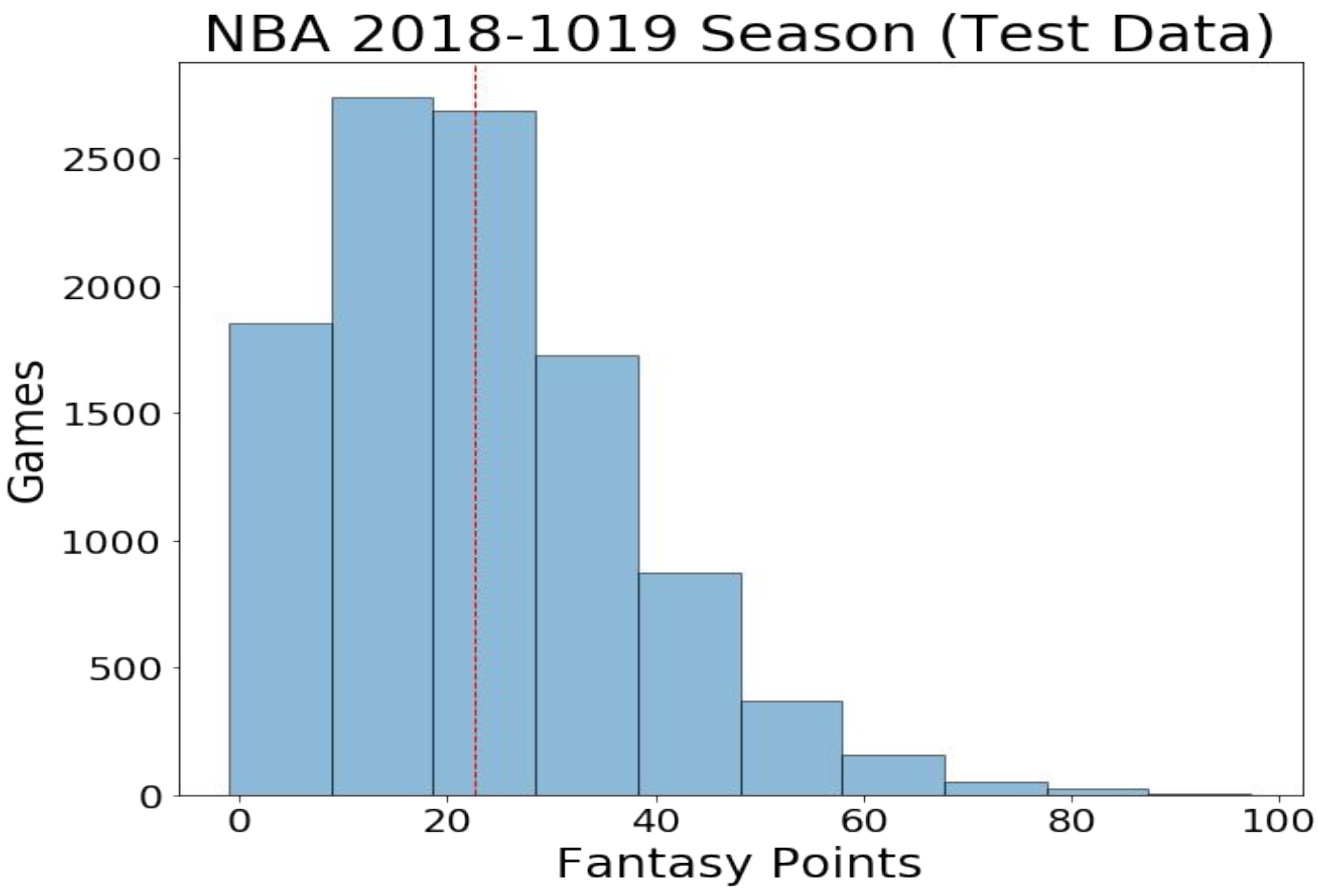
Opponent Stats

date	player	opp_ortg	opp_drtg	opp_pace	opp_trb_pct	opp_score	opp_pts_allowed
2017-12-06	kevin_durant	90.0	104.4	96.7	51.0	87	101
2017-12-08	kevin_durant	101.6	105.8	96.4	52.1	98	102
2017-12-11	kevin_durant	111.1	118.6	93.6	48.1	104	111
2017-12-14	kevin_durant	105.5	121.8	92.0	43.1	97	112

Visualizing the Target Variable



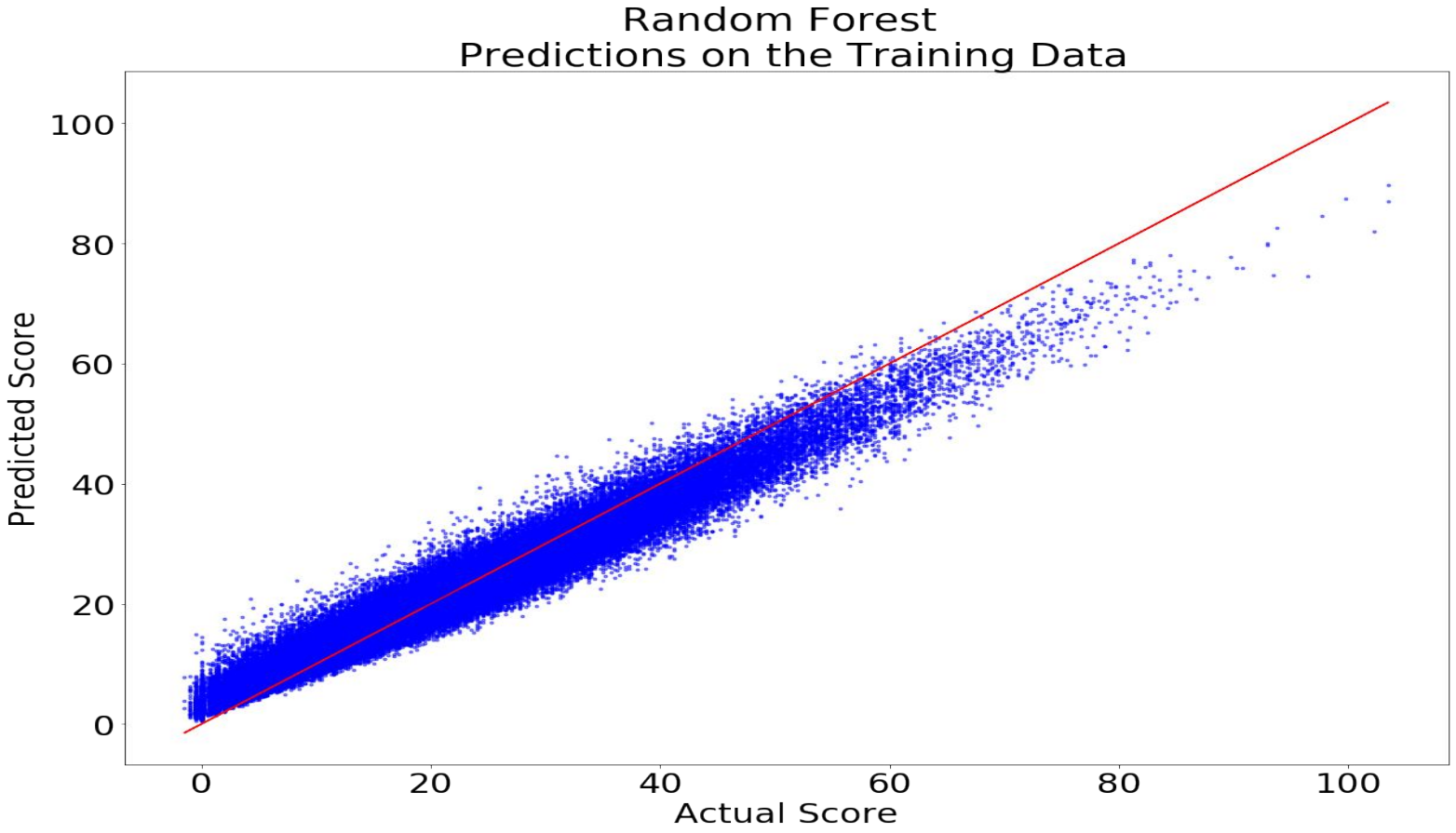
Visualizing the Target Variable



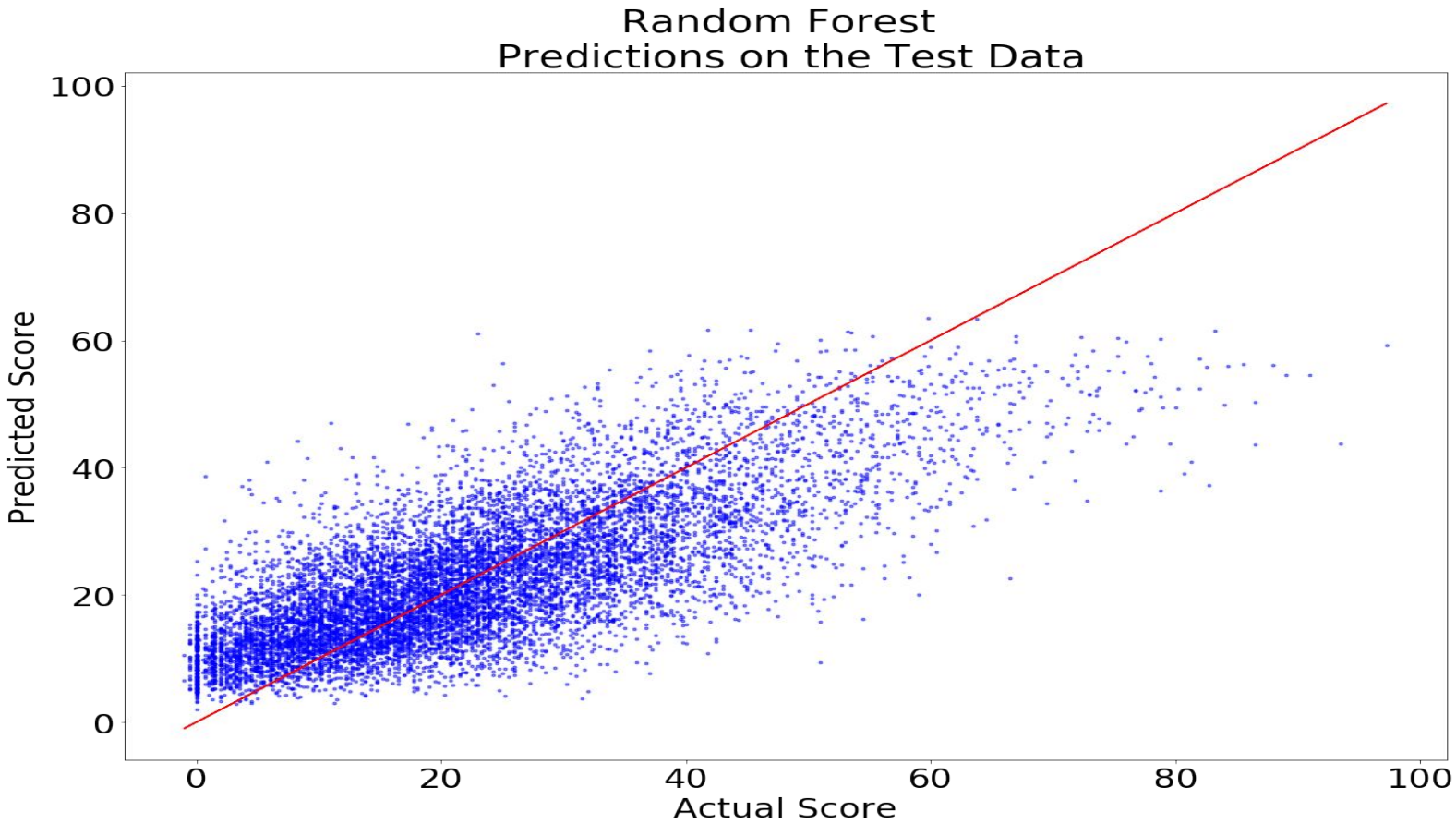
Regression Models

- R^2 Score
  - Percentage of Explained Variance in Data
- Root Mean Squared
  - How “off” on average each prediction is
  - Baseline RMSE is 14.62
- Random Forest
  - R^2 scores - .939 on the training data and .561 on test
  - RMSE - 9.68
- Gradient Boosted
  - R^2 scores - .57 on the training data and .568 on test
  - RMSE - 9.61
- Linear Regression
  - R^2 scores - .559 on the training data and .57 on test
  - RMSE - 9.58

Regression Models



Regression Models



Classification Models

- Multiclass
  - Tier 1: 50 or more fantasy points
  - Tier 2: 30 to 49 fantasy points
  - Tier 3: 20 to 29
  - Tier 4: Below 20
- Confusion Matrix
  - Table of predictions vs actual labels
- Accuracy Score
  - Random Forest - 100% on training / 60.7% on test
  - Gradient Boosted - 64.8% on training / 60.9 % on test
  - Logistic Regression - 62.8% on training / 59.9% on test

Random Forest Confusion Matrix

	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Actual 1	145	342	15	15
Actual 2	75	1403	401	495
Actual 3	7	695	489	1426
Actual 4	0	273	383	4333

Gradient Boosted Confusion Matrix

	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Actual 1	145	342	14	16
Actual 2	79	1401	441	453
Actual 3	6	695	549	1367
Actual 4	1	254	433	4301

Logistic Regression Confusion Matrix

	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Actual 1	35	450	9	23
Actual 2	7	1378	197	792
Actual 3	1	596	240	1780
Actual 4	0	197	161	4631

Conclusions / Recommendations

- Spread and randomness make it difficult to accurately predict target
- Incorporate average Draft Kings salary history into data
- Incorporate natural language processing to spot players who may be starting due to an injured teammate.
- Look for players with low season averages that have had sporadic high fantasy point totals