Predicting NBA Game Outcomes Using Machine Learning

By
Shawn Fang

The Problem:

Can we build a model as good or better than current models which predict outcome of NBA Games

The Solution:

Test and tune various models and identify the most predictive precise model

The Data

	Date	GameID	season	H_Team	A_Team	H_Team_Elo_Before	A_Team_Elo_Before	H_Team_Elo_After	A_Team_Elo_After	H_Last_10_gm_avg_PTS	.
0	2008-10-30	20081030PHO	2008-09	PHO	NOP	1508.416895	1508.416895	1493.762230	1523.071560	103.0	.
1	2008-10-31	20081031PHI	2008-09	PHI	NYK	1486.830104	1505.657856	1505.383439	1487.104521	84.0	.
2	2008-10-31	20081031TOR	2008-09	TOR	GSW	1513.169896	1491.583105	1517.801701	1486.951300	95.0	.
3	2008-10-31	20081031MIA	2008-09	MIA	SAC	1494.342144	1496.115314	1510.314181	1480.143277	115.0	.
4	2008-10-31	20081031BOS	2008-09	BOS	CHI	1505.657856	1509.8509	1517.162690	1498.346066	90.0	.
...	...	...	...	...	...	...	...	...	...	...	.
14134	2020-08-13	20200813BKN	2019-20	BKN	POR	1518.795548	1528.36943	1514.104890	1533.060088	117.9	.
14135	2020-08-14	20200814TOR	2019-20	TOR	DEN	1683.137229	1548.559957	1687.014564	1544.682622	111.0	.
14136	2020-08-14	20200814IND	2019-20	IND	MIA	1563.263075	1566.395438	1575.196348	1554.462166	110.4	.
14137	2020-08-14	20200814LAC	2019-20	LAC	OKC	1642.593683	1586.866995	1646.563430	1582.897248	118.8	.
14138	2020-08-14	20200814HOU	2019-20	HOU	PHI	1569.491973	1533.078029	1534.811696	1567.758306	114.8	.

14139 rows × 58 columns

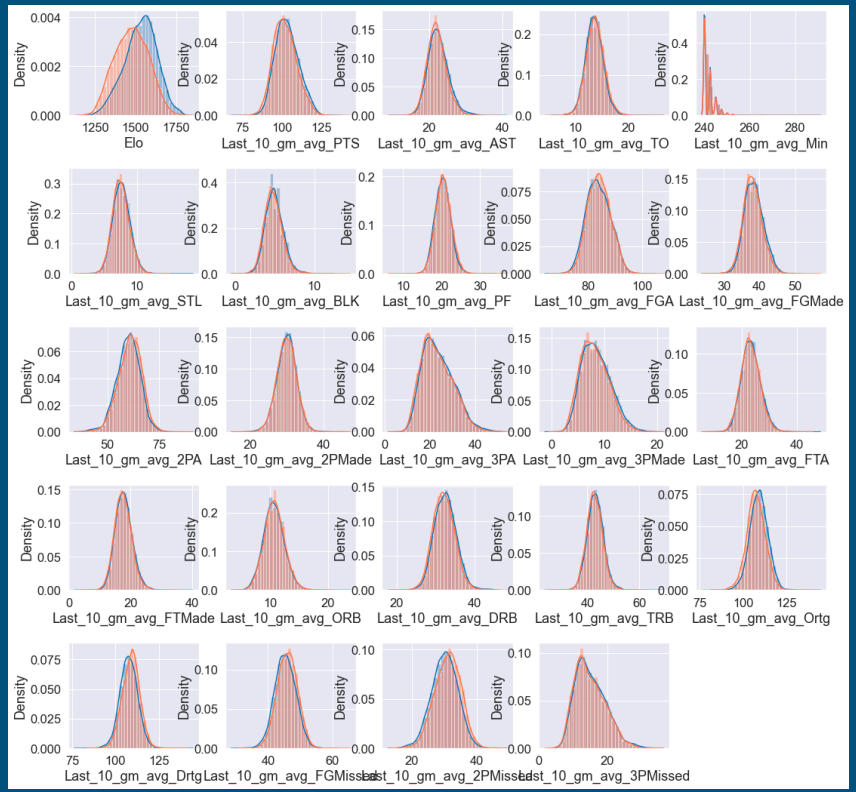
Data Source: https://www.kaggle.com/heeebsinc/nbaseasonstats201820, https://www.kaggle.com/rafaelgreca/nba-games-box-score-since-1949

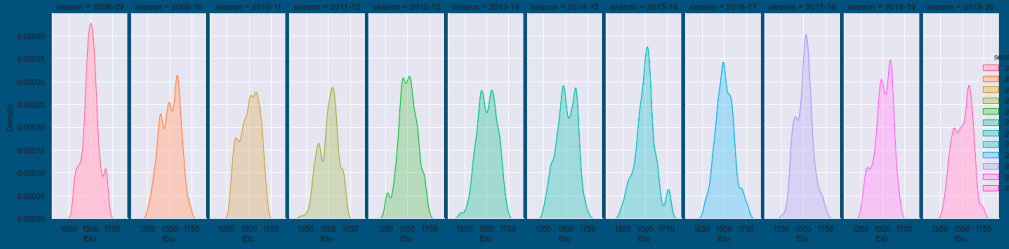
Data Wrangling

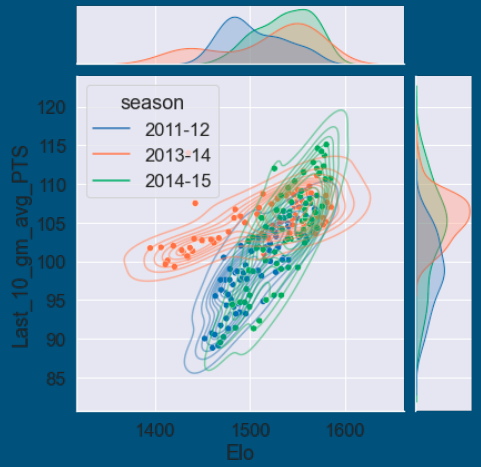
Original datasets:
Team stat (15348, 124)
Player stats (37369, 26)

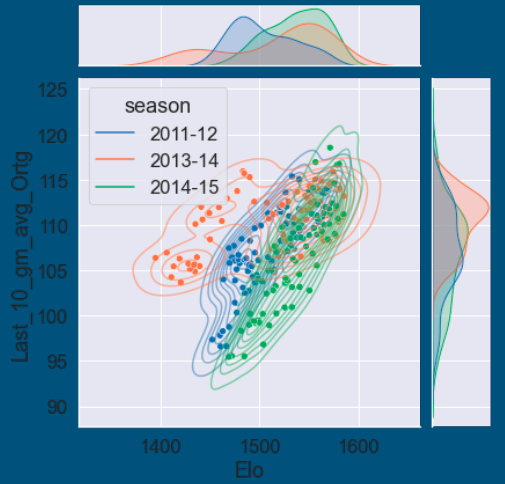
- Only consider data from regular season
- Drop irrelevant column, and % columns
- Drop rows 0 minute NaN's
- Created Game ID for both data frames that is joinable
- We ended data wrangling with cleaned player stats (291008, 24), and cleaned team stats (14348, 80)

Exploratory Data Analysis

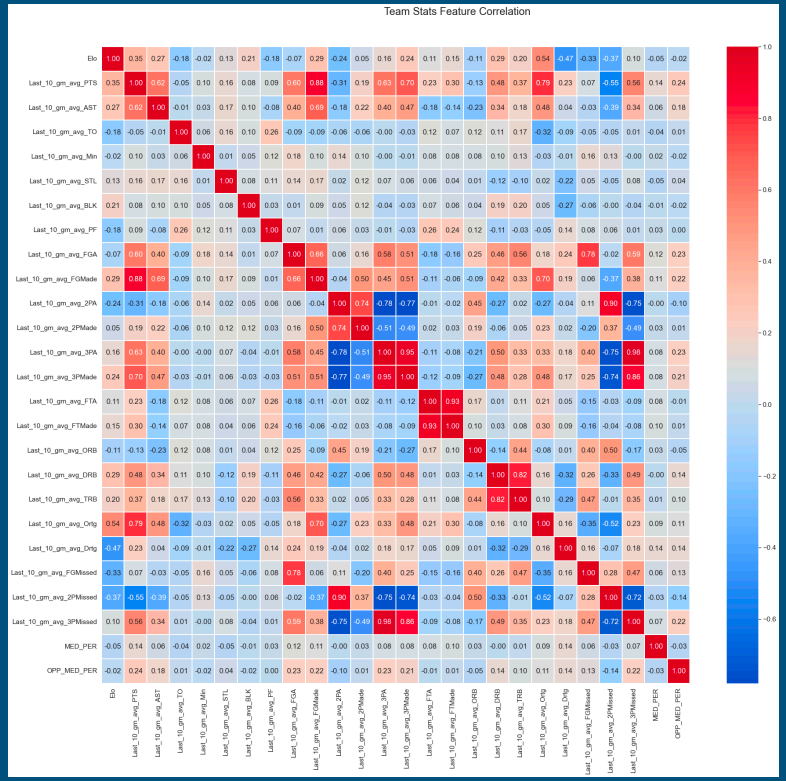


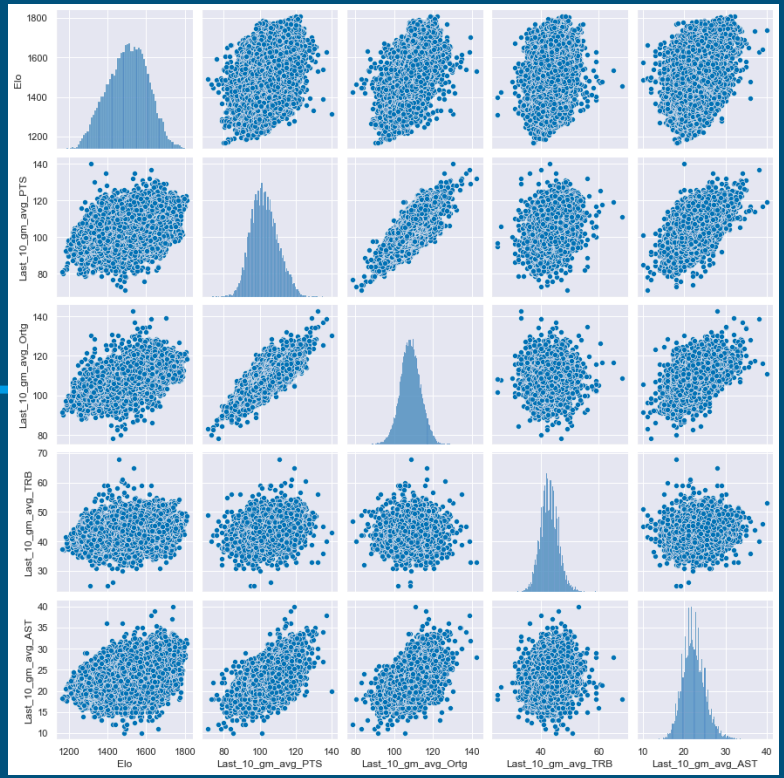






Exploratory Data Analysis Cont.

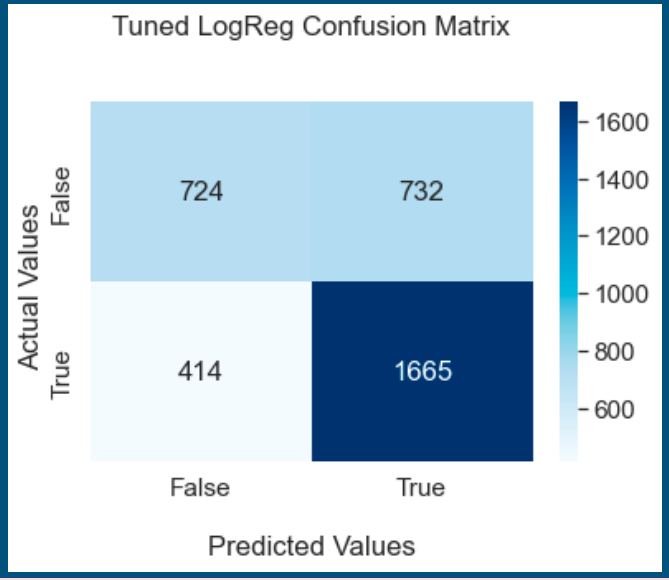




Model Selection

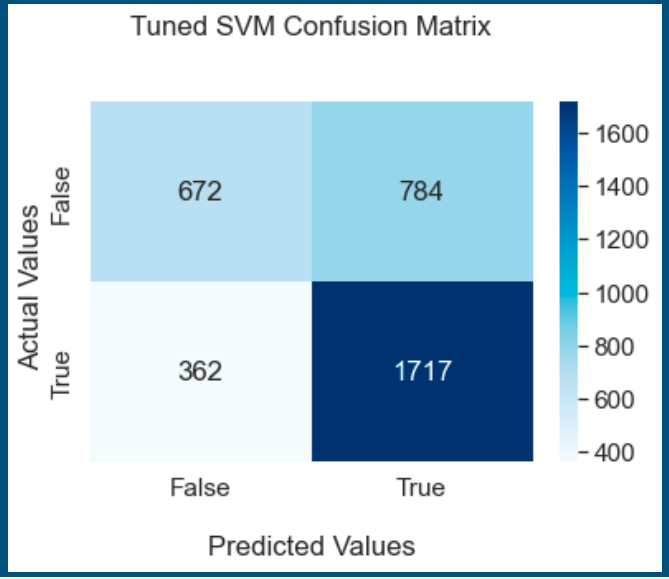
- Logistic Regression
- Random Forest Classifier
- Support Vector Machines
- K-Nearest Neighbor
- Gradient Boost

Model Comparisons: Precision and F1 Score



0.6758132956152758

	precision	recall	f1-score	support
0	0.64	0.50	0.56	1456
1	0.69	0.80	0.74	2079
accuracy			0.68	3535
macro avg	0.67	0.65	0.65	3535
weighted avg	0.67	0.68	0.67	3535



0.6758132956152758

	precision	recall	f1-score	support
0	0.65	0.46	0.54	1456
1	0.69	0.83	0.75	2079
accuracy			0.68	3535
macro avg	0.67	0.64	0.64	3535
weighted avg	0.67	0.68	0.66	3535

Hyperparameter Tuning

Grid Search Cross Validation with Logistic Regression Model

Precision Increase by Site

- Accuracy: 0.6752 to 0.6758
- Precision: 0.63 to 0.64
- F1: No Change 0.56 to 0.56

Takeaways

- Logistic Regression was consistently best model, but SVM is competitive
- Summarizing Stats are helpful, but only if focused on offense or defense, not together
- A lot of features are redundant

Future Research

- Increase collection of data especially the player defensive data for building defensive summarizing features
- Create features that utilize progress stats with seasonal averages to determine player/ team progression trend.
- Build a player career trajectory model to used with progression trend.

Thank You!

NBA enthusiast for collecting and organizing NBA Game Statistics Datasets, and Kaggle for hosting.

Mukesh Mithrakumar for guidance during project, and Blaine Bateman for coding support.

Questions?