Module 4 Data Science Project: Predicting All-NBA Team Selections

David Goldstein

dgoldstein24@gmail.com
dg2996@columbia.edu

Hypotheses: Prompt

- The All-NBA First ,Second, and Third teams contain the best players in the NBA for a given season.
- Sports writers select the recipients of these honors
- Questions that follow:

    - Do the selections reflect the best players statistically or are there other factors/biases?

        > If not, then the NBA Collective Bargaining Agreement may be inherently flawed

            *NBA players are eligible for ”SuperMax” contracts only if they are selected to All-NBA teams

    - Can future All-NBA players be predicted from their stats from the previous season?

        > If so, then such a predictive model could be used to identify players poised for a breakout season, which would be very helpful to General Mangers of NBA teams

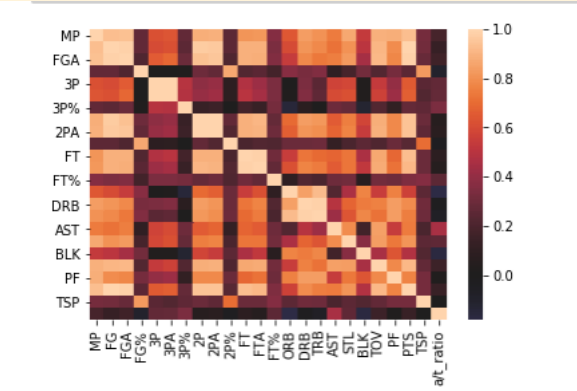
Obtain and Scrub: Compile data and clean from Basketball Reference and NBA

Out[3]:

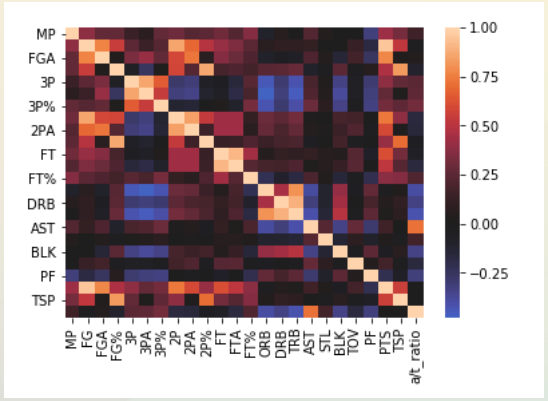
	Year	Player	concat	count	Pos	Age	Tm	G	GS	MP	...	ORB	DRB	TRB	AST	ST
0	2019	Álex Abrines	2019Álex Abrines	1	SG	25	OKC	31	2	588	...	0.3	2.6	2.9	1.2	1.
1	2019	Quincy Acy	2019Quincy Acy	1	PF	28	PHO	10	0	123	...	0.9	6.4	7.3	2.3	0.
2	2019	Jaylen Adams	2019Jaylen Adams	1	PG	22	ATL	34	1	428	...	0.9	4.1	5.0	5.5	1.
3	2019	Steven Adams	2019Steven Adams	1	C	25	OKC	80	80	2669	...	5.3	5.0	10.3	1.7	1.
4	2019	Bam Adebayo	2019Bam Adebayo	1	C	21	MIA	82	28	1913	...	3.1	8.1	11.2	3.5	1.

5 rows × 32 columns

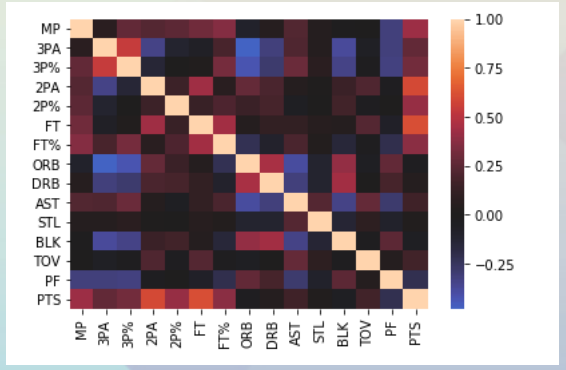
Explore: Multicollinearity graphs for overlapping features



Adjustment of stats to per 36 minutes – totals correlated to strongly with amount of minutes played



Removal of overlapping features



Model Pt. 1: Calculate prediction accuracy of a year's All-NBA selection based on statistics from that NBA season

Many combinations of models and features tried to increase model accuracy and increase the interpretive meaning of the results:

- With/without minutes played as a feature
- With/without manually generated advanced statistics
  - True Shooting Percentage – accounts for player's shooting efficiency based on free throws, 2-point field goals, and 3-point field goals
  - Assist to turnover ratio – Measurement of ability to create successful offensive plays while maintaining ball security
- Simple machine learning models vs. grid search models for best-tuned ML models
- SVM vs. Adaboost vs. Random Forest
- Machine Learning vs. Connected Neural Networks
- Connected Neural Networks with varying layers, regularizations, loss functions, and activation functions

Validation and Next Steps

- Some interesting findings
  - Discrepancy between which All-NBA team a player should have made vs. the honor they received
    - > i.e. the model says Joel Embiid should have been 1st team All NBA in 2019
    - > I and many NBA thought leaders agree

- The most accurate ML and CNN models both had an accuracy of about 97.8%
  - Inference: All-NBA selections are primarily determined from current season statistics
  - Inference: findings may have been skewed by the ratio of selected players to all players in the database being too small: the model could be accurate by predicting every player to not be All-NBA

- A more interesting exploration would be predicting players to make All-NBA next year based on this year's statistics
  - More actionable information for GM's
  - Likely less direct correlation (and therefore more for the model to have to do) as compared to relationship between current season stats and current season All-NBA

  - Additional thoughts – Filter players by the number of minutes played to increase the ratio of selected players to total players evaluated by the model & remove 1st, 2nd, 3rd team differentiation

Obtain and Scrub: Shift player selections back a year, simplify labels to selected vs. non-selected, and filter by min played

In [241]:
```
#Recategorize any non "No_Honors" category as "Selected"

new_h = []
for item in df_4['Honors_Next_Year']:
    if item != 'No_Honors':
        new_h.append('Selected')
    else:
        new_h.append('No_Honors')

df_4['Honors_Next_Year'] = new_h
df_4
```

In [58]:
```
#iterate throught the dataframe
#find every entry with an honor
#store the honor in the player's entry for the previous year by savin in 'new_honors'

indices_to_change = []
honors_to_change = []


for i in range(len(df_2)):
    if df_2.iloc[i]['Honors'] != 'No_Honors':
        last_year = df_2.iloc[i]['Year'] - 1
        name = df_2.iloc[i]['Player']
        honor = df_2.iloc[i]['Honors']

        for index in range(len(df_2)):


            if (df_2.iloc[index]['Player'] == name) & (df_2.iloc[index]['Year'] == last_ye
                indices_to_change.append(index)
                honors_to_change.append(honor)
                break
```

Out[241]:

	ORB	DRB	STL	BLK	TSP	a/t_ratio	Honors_Next_Year
303	1.2	7.3	1.4	0.8	0.621176	2.170732	Selected
1193	3.5	8.4	0.7	1.2	0.617315	1.040000	Selected
1226	1.2	2.7	1.0	0.4	0.534810	1.000000	No_Honors
1464	0.7	5.0	1.6	0.6	0.600245	1.651163	Selected
2661	0.6	3.5	1.0	0.3	0.532338	1.761905	No_Honors
...	...	...	...	...	...	...	...
10292	0.6	3.6	1.3	0.3	0.498532	1.230769	No_Honors
10293	1.1	2.4	1.1	0.6	0.521403	1.243243	No_Honors
10319	0.5	2.6	1.2	0.1	0.532454	1.678571	No_Honors
10337	1.6	6.0	1.5	0.5	0.506859	1.468750	No_Honors
10346	0.7	1.9	1.5	0.2	0.508811	2.100000	No_Honors

Model Pt. 2: Calculate prediction accuracy of next year's All-NBA selection based on statistics from current NBA season

- Run through combinations of models and tuning parameters similar to part 1 in order to find the most accurate prediction model

- Best model: Machine learning model: Grid Search-tuned Random Forest Classifier
    - 79.48% accuracy
    - Not every guess was No Honors – improvement over issue with part one

Validation and Results

• In this case, the evidence of true positives and no glaring false positives is more important than false negatives - it is more important to find diamonds in the rough than to scrutinize over every option you might have missed

• The model did not have any glaring misses in terms of predictions – no players who were False Positives were far away from All-NBA consideration

• The below players were correctly predicted to be All-NBA players based on their stats the previous year, even though they had not made All-NBA in that previous year

- By continuing to build on the model, this analysis could be a useful tool to identify players poised for breakout seasons

Karl-Anthony Towns
2017 -> 2018

James Harden
2016 -> 2017

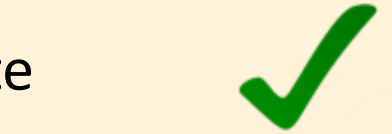
Blake Griffin
2011 -> 2012

Dwight Howard
2006 -> 2007

Paul Pierce
2001 -> 2002

Interpret: Recommendations

1. Rec 1: Trust that All-NBA teams are an accurate reflection of statistical excellence



2. Rec 2: Do not try to predict future All-NBA appearances of players who have played very few minutes

3. Rec 3: Utilize predictive modeling to find breakout candidates, and continue to develop the model to increase usage in this capacity

Interpret: Project Takeaways

1. ML and CNN are both very useful, CNN probably better for big, differentiated datasets. ML worked better for this project

2. Tuning for both ML and CNN is essential

3. Feature engineering based on domain expertise can drastically improve results

4. Predictive modeling has meaningful real-world applications in professional sports

Thank You

David Goldstein