Predicting NBA All-NBA Players

By Aidan Chandrasekaran, Casey Ng, Brandon Tat, Eric Tran

Background on the NBA

- The National Basketball Association
- 30 Teams
  - Approximately 15 players per team
- What are All-NBA teams?
- 3 teams, 1st, 2nd and 3rd team
  - Comprised of the 5 best players each year (2 guards, 2 forwards, 1 center)
  - In total, the 15 best players, as voted by sportswriters and broadcasters after the end of each NBA season

Questions of Interest

- Can we use certain NBA statistics to predict whether or not a player makes the All-NBA team?
- What are the statistical differences between All-NBA players and non All-NBA players?
- Is KNN or Neural Networks better at modeling the data?
- Is there a difference in our model's prediction for balanced and unbalanced data?

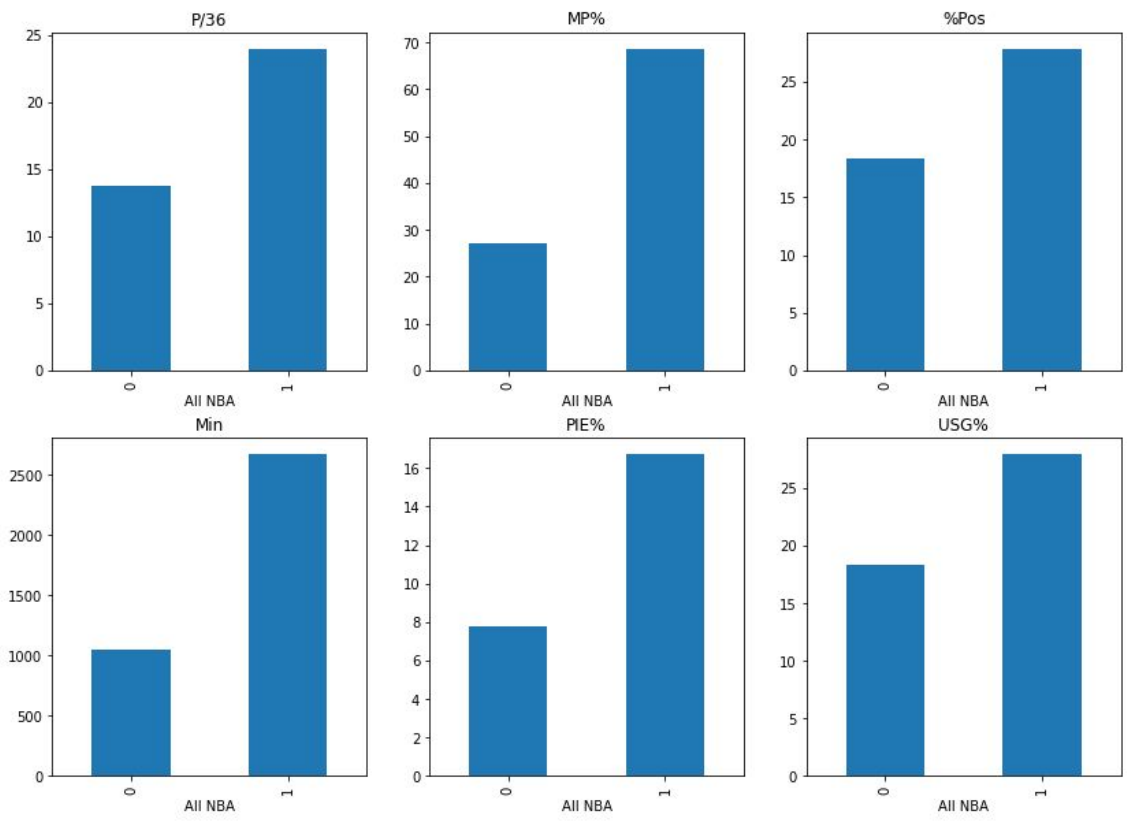
The Data

- NBA Data
  - https://github.com/fivethirtyeight/nba-player-advanced-metrics
  - read_csv()
- All NBA Teams
  - https://www.basketball-reference.com/awards/all_league.html
  - Using XML, Xpaths
- Only used 2000-2020 because old data might not be reliable
- Merged the data
  - Found highest correlations
  - Normalized the data

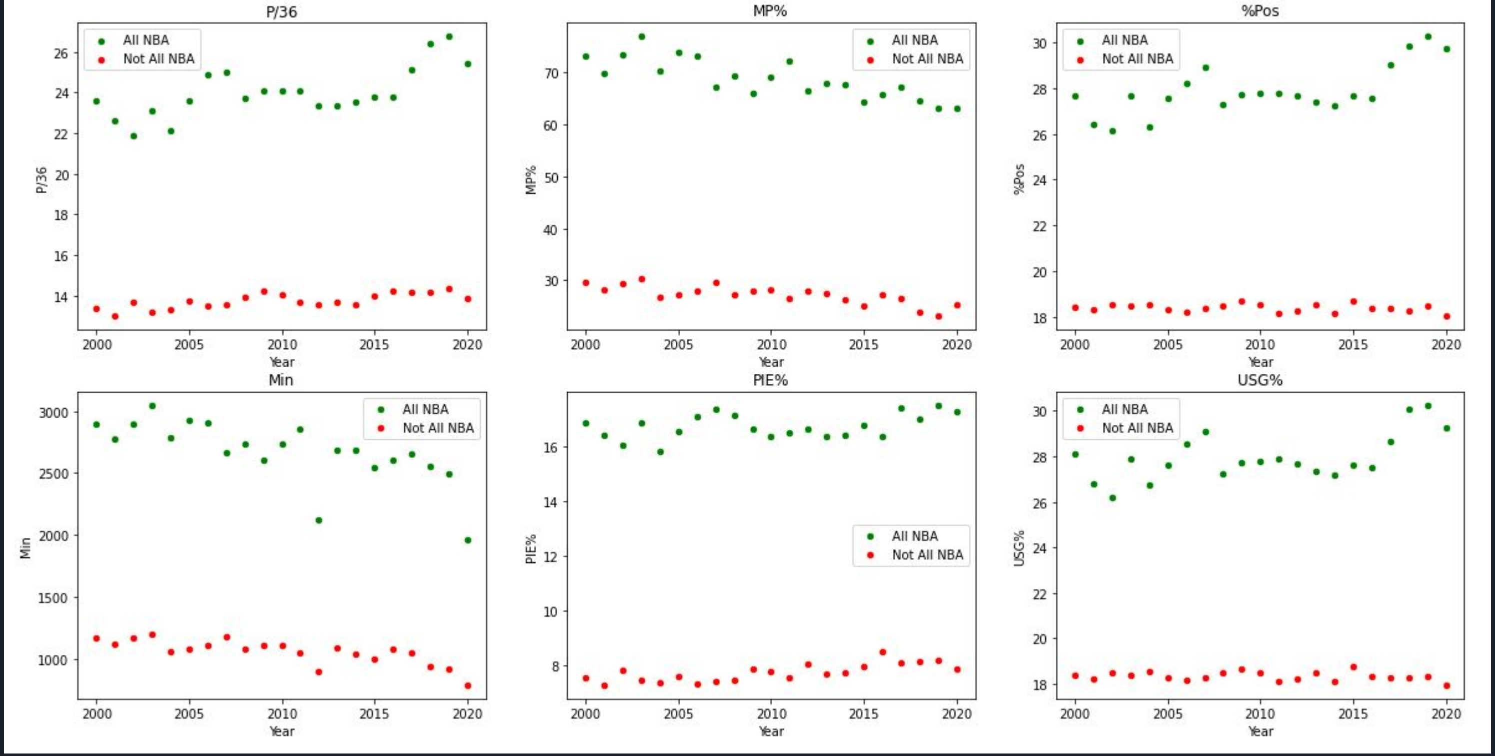
The Data

- Highest Correlations:
    - P/36 (Points per 36 min)
    - MP% (Percentage of available minutes played)
    - %Pos (Share of team possessions used on court)
    - Min (Total minutes played)
    - PIE% (Player impact estimate)
    - USG% (Usage rate)

Data Visualization



Data Visualization



Building our models and methodologies

- We decided to use K Nearest Neighbor and Neural Network
- We did not use K-Means Clustering because we felt like the clusters would be too uneven for it to work effectively
- We wanted to test how our models would perform against each other and see which one is best

Building our models and methodologies (cont.)

- We trained both types of models using 2 different methods of data:
  - Balanced means 1:1 ratio of classes
  - Unbalanced means any other ratio, in our case ~36:1
  - Balanced (~300 all-NBA and 300 non all-NBA players)
  - Unbalanced (~300 all-NBA and ~20000 non all-NBA players)
- Z-score to normalize all data

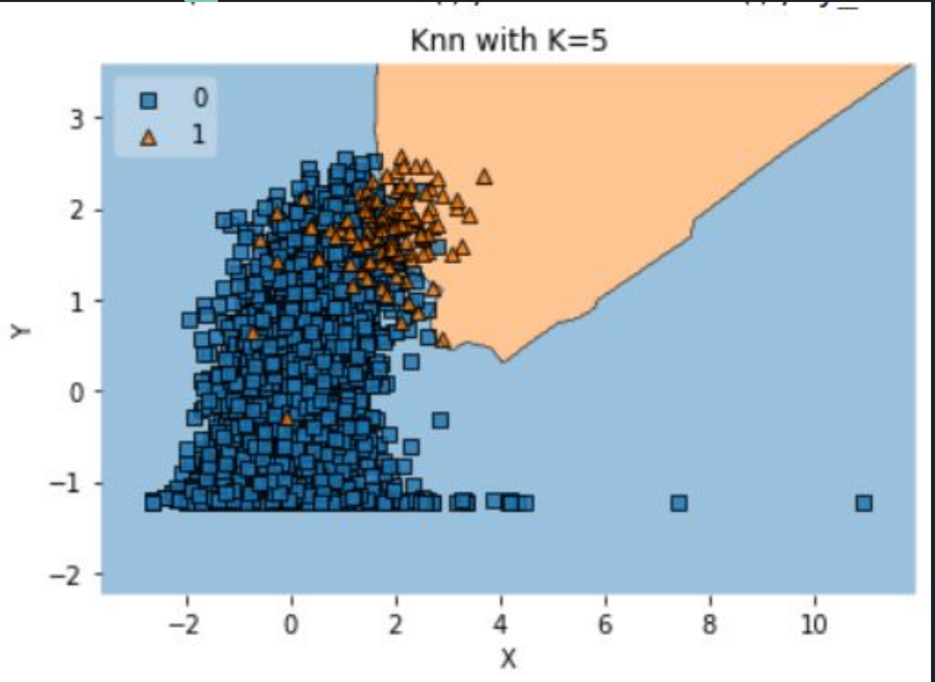
KNN

- 5 neighbors worked best for us
- Anything less caused too much variance
- Anything more had no noticeable improvements
- Results were:
    - ~94% for balanced on balanced
      ~98% for unbalanced on unbalanced
      ~55% for unbalanced on balanced data

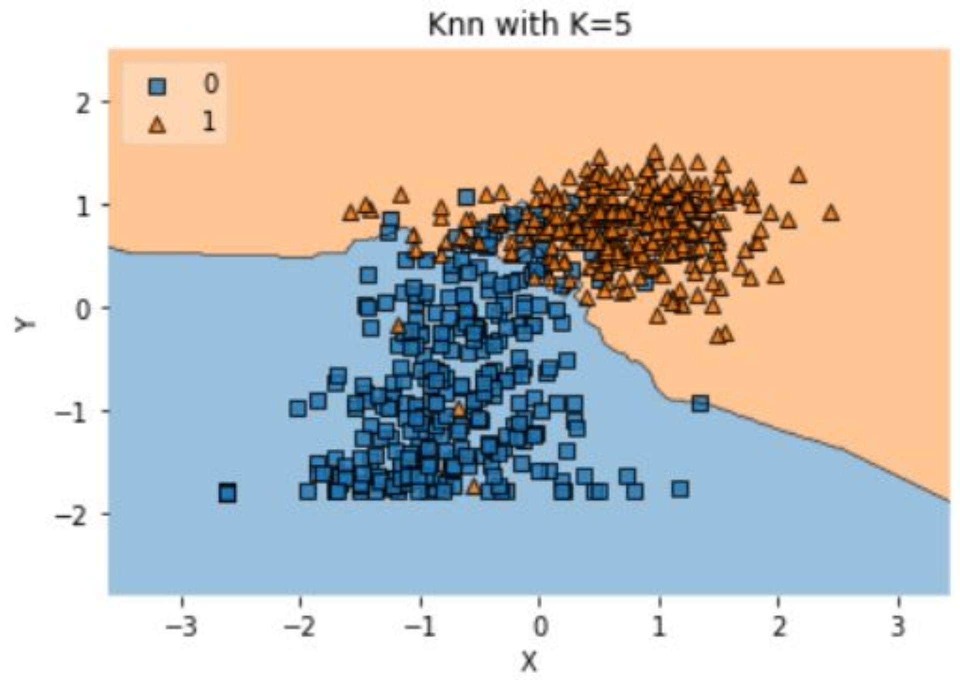
- Reasons?
    - The unbalanced model was probably guessing a lot of false for when it was being tested on the unbalanced dataset so when it was being tested on balanced it marked many false even if they were actually all nba

KNN Results Graphs

Unbalanced



Balanced



Credit to: https://towardsdatascience.com/knn-visualization-in-just-13-lines-of-code-32820d72c6b6

Neural Networks (NN)

- Attempting to do binary classification, All-NBA vs non All-NBA
- Used class weights for the unbalanced NN
- Trained on randomize balanced datasets for balanced NN
- 3 hidden layers for each

NN Results

- ~ 95% accuracy on unbalanced data
- ~ 96% accuracy on balanced data
- ~ 54% accuracy for unbalanced model on balanced data
- Why?
  - Unbalanced NN got trained on a dataset that skewed more towards non All-NBA so it was more likely to mark players as such
  - Balanced NN was the best for accurately determining if a player was all nba or not

Predicting Non All-NBA vs. All-NBA with NN

- With our balanced and unbalanced data models, can we try some hypothetical player statistic inputs and predict if that player was voted into an All-NBA team that year?
- Input players with standardized (z-score) stats: Points/36, MP%, %Pos, Min, PIE%, USG%
- keras model.predict_classes

- Unbalanced Model:

Star: 25, 75, 20, 3000, 15, 30 -> Yes

Regular: 10, 5, 10, 200, 10, 10 -> No

Good Sub: 20, 5, 20, 300, 10, 25 -> No

- Balanced Model:

Star: 20, 50, 25, 1500, 10, 25 -> Yes

Regular: 15, 50, 20, 1500, 10, 25 -> Yes

Good Sub: 10, 40, 10, 1500, 5, 10 -> No

In conclusion, what did we learn?

- Both models worked well on our balanced dataset (important for training and testing), ~95% accuracy
- A model trained on unbalanced dataset will perform well on unbalanced dataset; however, it has a hard time distinguishing between the two classes when given a balanced dataset
- Across the six variables most highly correlated with whether or not a player was All-NBA, we found major differences in the value of players who were All-NBA vs. who weren’t

Major obstacles faced along the way

- Getting good data is difficult
- Data cleaning (even with good data to start with)
  - Removing the diacritics/accent marks so the datasets would align
- Normalizing and balancing your data is vital to getting an accurate model or else it will get trained poorly
- Visualizing data is important to show our results in a concise way