Springboard Capstone

Marshall Pagano

July 2018

Overview

- The objective of this project was to estimate an NBA player's first non-rookie contract based on his NBA body of work.
- Phases to this project were:
  - Data wrangling: locating and pulling down relevant NBA data to perform the above analysis
  - Data exploration: interrogating the data set to look for insights to help guide and shape the eventual analysis
  - Data analysis: building two models, one generic, one ensemble, to compare and contract their merits and analyze the value of the final model.

Objective

- The goal of this capstone project is to be able to predict an NBA player's first non-rookie salary.
- The hope is that the model would provide value to a potential client. A couple scenarios in which a model of this type would be useful:
  - NBA team could use this when determining how much to pay their players (that are just coming off Rookie contracts). Allows determination of what is a "fair price" for a qualifying player.
  - NBA agents could use this as a negotiation tool to ensure that they are not being undersold by an NBA team.
  - This model could serve as a proxy for overall player quality. Quality and salary are different things, but correlate extremely closely.

Data Wrangling

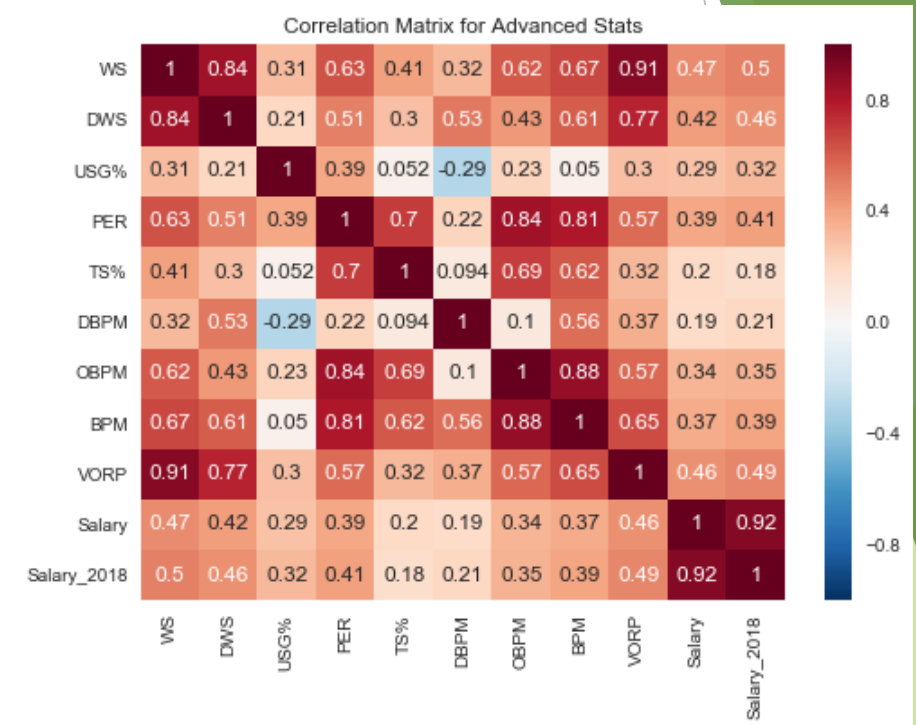
- Data used for this capstone project was primarily from BasketballReference.com.
- First, a list of players must be generated.
  - Base list: all players drafted from 1990-2013.
  - Final list: all players from base list that played long enough in NBA to receive a second NBA contract.
- Data from each qualifying player's individual webpage included:
  - Traditional Per Game statistics
  - Advanced statistics
  - Contract information

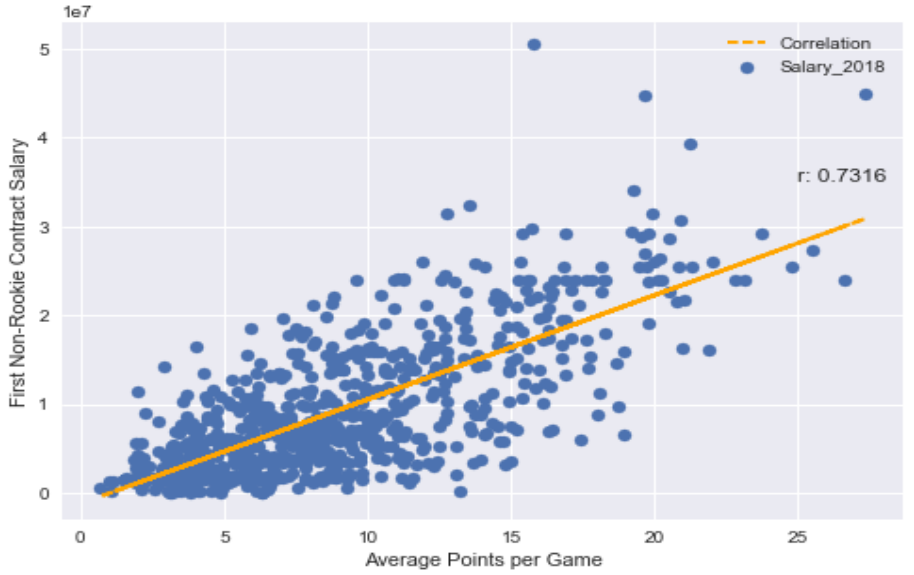
Data Wrangling continued

- Traditional Per Game Statistics:
  - Age, Pos, G, GS, MP, FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, eFG%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS
- Advanced Statistics:
  - Age, Pos, G, MP, PER, TS%, 3PAr, FTr, ORB%, DRB%, TRB%, AST%, STL%, BLK%, TOV%, USG%, OWS, DWS, WS, WS/48, OBPM, DBPM, BPM, VORP
- See Appendix A for explanation of these statistics
- Salary cap information was gathered to normalize salaries based on the league wide salary cap which varies from year to year. Finally all salaries were converted into 2018 dollars.

Data Exploration

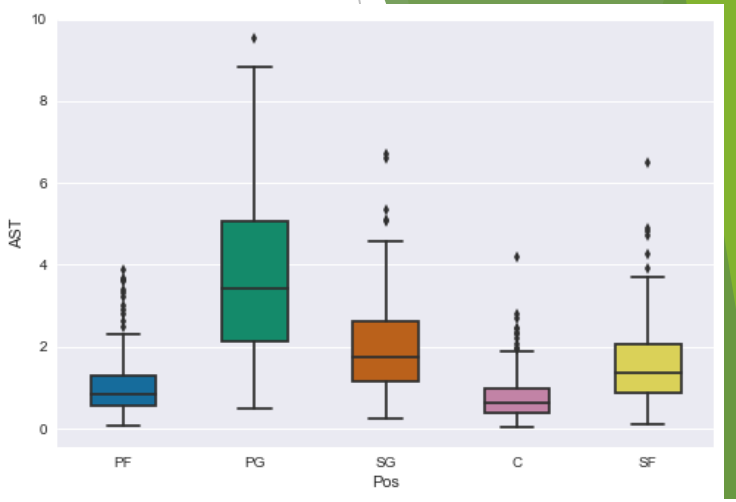
- Correlation matrices were used to find strong correlations with target variable (desired) and amongst each other (not desired)
- Variables were individually analyzed to confirm correlation levels and look for trends in relationships to 2018 adjusted salary.

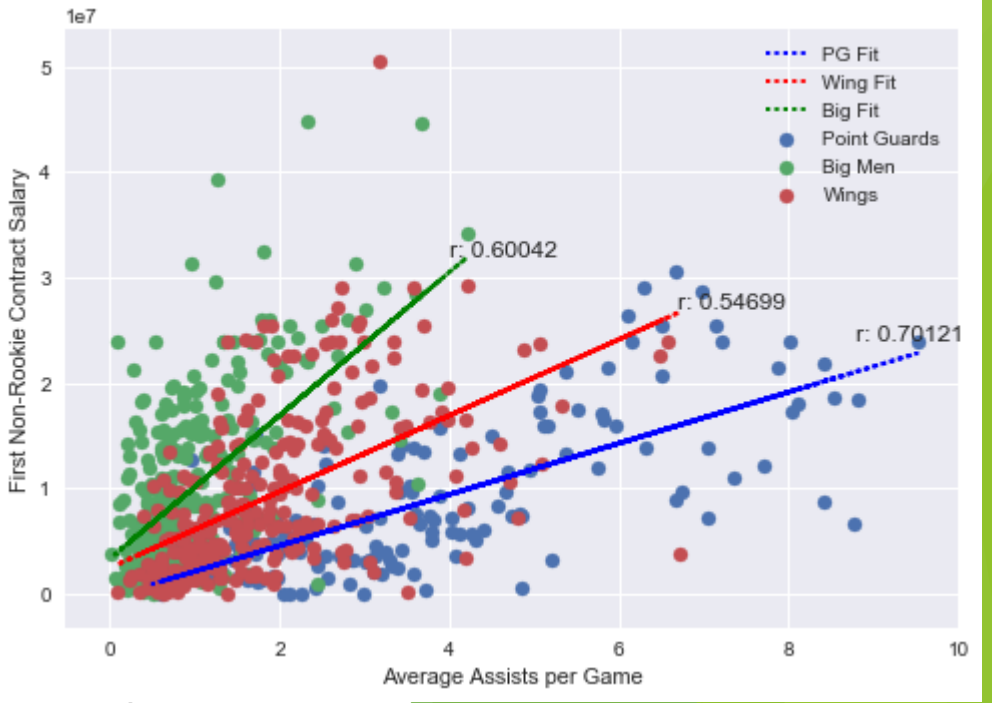




Data Exploration continued

- Variables were analyzed and grouped by other variables, like position, to help determine structure of eventual ensemble model
- Variables with statistically significant differences in correlation levels across position were used in the position specific portion of the ensemble model (e.g. Assists)
- Variables without statistically significant differences correlation levels across position were used in the general portion of the ensemble model





Data Analysis

- Two models were created:
  - General model – using a traditional OLS approach
  - Ensemble model – using position-specific OLS sub-models
- Test-train split analysis was performed to ensure model was not overfitting to the data
- Weights of general vs. position-specific sub-models were optimized to minimize overall root mean squared error
- Ensemble model performed slightly better than General model based on root mean squared error
  - General model: RMSE = $**4,796,258**
  - Ensemble model: RMSE = $**4,562,856**

Data Analysis continued

- General model output explanation:
  - Each incremental Point per game increases a player's predicted salary by $718,000
  - Each incremental Block per game increases a player's predicted salary by $1.51M
  - Overall model explains nearly 67% of the variance of the players' first non-rookie contract salary

R Squared: 0.6725

Variable	Coefficient
PTS	718,183
AST	323,0491
TRB	345,619
BLK	1,512,113
DWS	1,831,869
Age	-389,763

Root Mean Squared Error: **4,796,257**

Conclusion

- Ensemble model proved to show modest increase in overall quality of model.
  - Points per game are the main determinant of a player's future salary
- Both models could be improved through inclusion of new data (both more variables, and more data points).
- Existing model is has limited use given its relatively high RMSE values.
  - Models could be used as a guideline for salary determination, but further improvements would need to be made for model to have real use

Appendix A

- Age - Age
- Pos - Position
- G – Games played
- GS – Games started
- MP – Minutes played
- FG – Field goals made
- FGA – Field goal attempts
- FG% - Field goal percentage
- 3P – Three pointers made
- 3PA – Three point attempts
- 3P% - Three point percentage
- 2P – Two pointers made
- 2PA – Two pointers attempted
- 2P% - Two point percentage
- eFG% - Effective field goal percentage
- FT – Free throws made
- FTA – Free throw attempts
- FT% - Free throw percentage
- ORB – Offensive rebounds
- DRB – Defensive rebounds
- TRB – Total rebounds
- AST - Assists
- STL - Steals
- BLK - Blocks
- TOV - Turnovers
- PF – Personal fouls
- PTS - Points

Appendix A continued

- PER – Player efficiency rating
- TS% - True shooting percentage
- 3PAr – Three point attempt rate
- FTr – Free throws attempt rate
- ORB% - Offensive rebound percentage
- DRB% - Defensive rebound percentage
- TRB% - Total rebound percentage
- AST% - Assist percentage
- STL% - Steal percentage
- BLK% - Block percentage
- TOV% - Turnover percentage
- USG% - Usage percentage
- OWS – Offensive win shares
- DWS – Defensive win shares
- WS – Win shares
- WS/48 – Win shares per 48 minutes
- OBPM – Offensive box plus-minus
- DBPM – Defensive box plus-minus
- BPM – Box plus- minus
- VORP – Value over replacement player