NBA Salary Prediction

A Regression Project

By: Justin Zhang

Introduction

- Goal: using advanced and basic game statistics from players' on court performance to predict player salary.

- Salary data is from:

  (https://data.world/datadavis/nba-salaries)
- Advanced and basic statistics are from:

  (https://www.basketball-reference.com/)

- Application: provide additional measure for the NBA and NBA players during salary and contract negotiation.

Data Wrangling

- Basic stats data shape: 3589 rows X 31 columns.
- Advanced stats data shape: 3575 rows X 28 columns.
- Merging Statistics Data
  - Six years of advanced/basic statistics divided into 12 files.
  - Missing values must be filled in before data merge.
    - Missing values are caused by 0 in denominator.
    - NaNs are filled in with 0.
  - Advanced and basic statistics are merged through inner join.
    - Merged on columns:

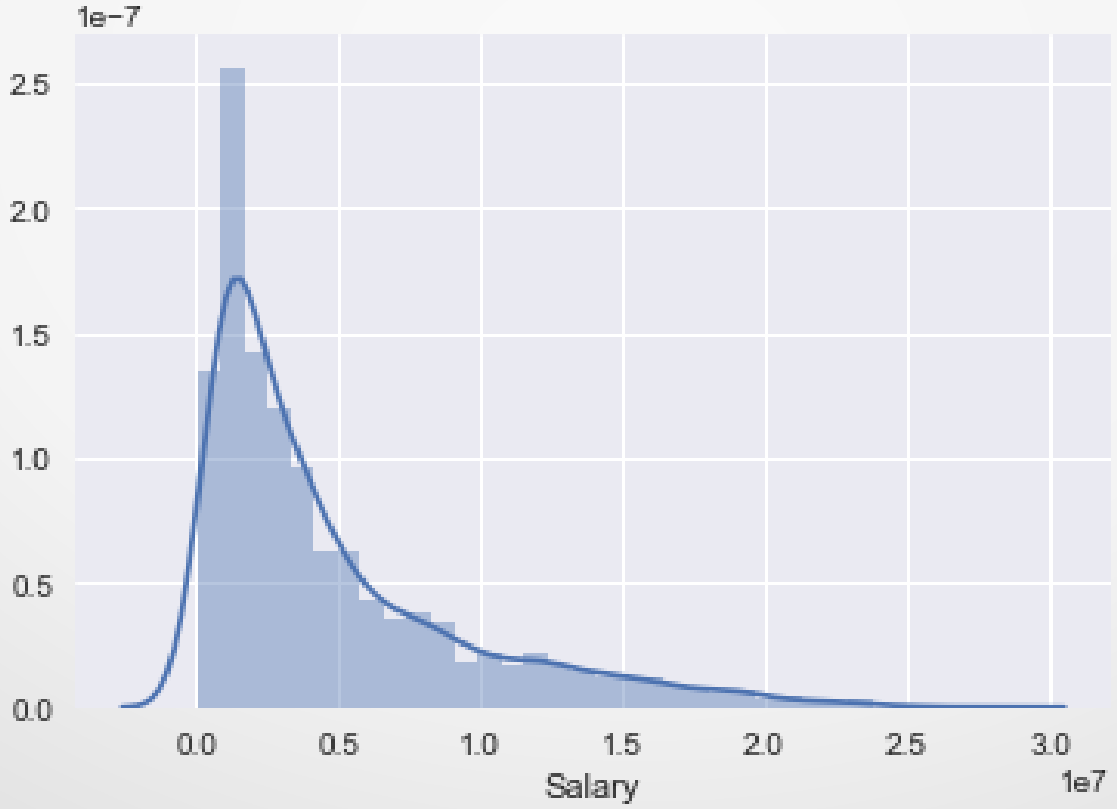
      Year, Player, Rk, Pos, Age, Tm, G, MP

Data Wrangling

- Merged stats data shape: 3575 rows X 51 columns
- Salary data shape: 2788 rows X 9 columns
- Merge stats data with salary data
  - Create uniformity between columns to be merged on.
    - ‘Player’ column in stats consist of player name and player ID.
      - Separating via regular expression then separating the two variables.
    - ‘Team’ column in stats comprised of team abbreviation.
      - Replace abbreviated team names with full team names.
- Salary dataset had no missing values.
- Merging salary and stats dataframe
  - Left (salary) join on ‘Player, ‘Team’, and ‘Year’ column
- Result dataframe shape: 2788 rows X 58 columns

Exploratory Data Analysis

- Salary statistics:
  - Count: 2378Mean: 4,416,256
  - STD: 4,706,159 Min: 19,914
  - Median: 2,677,320 Max: 30,453,800

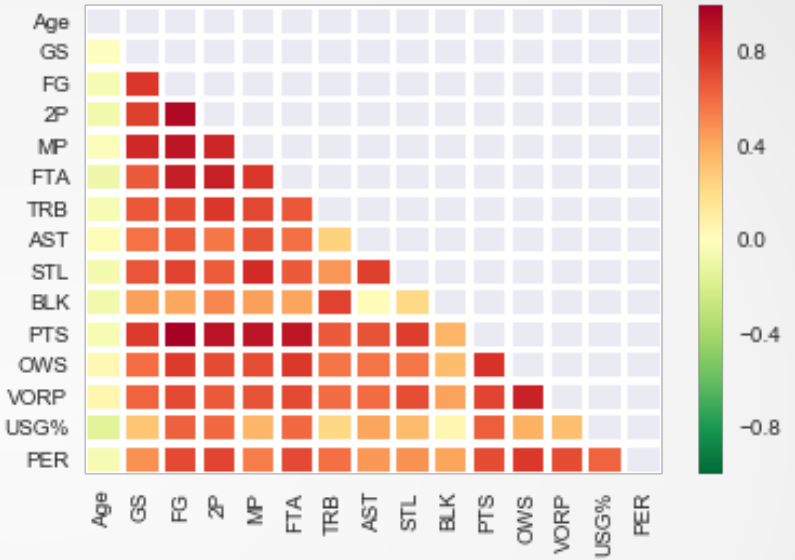


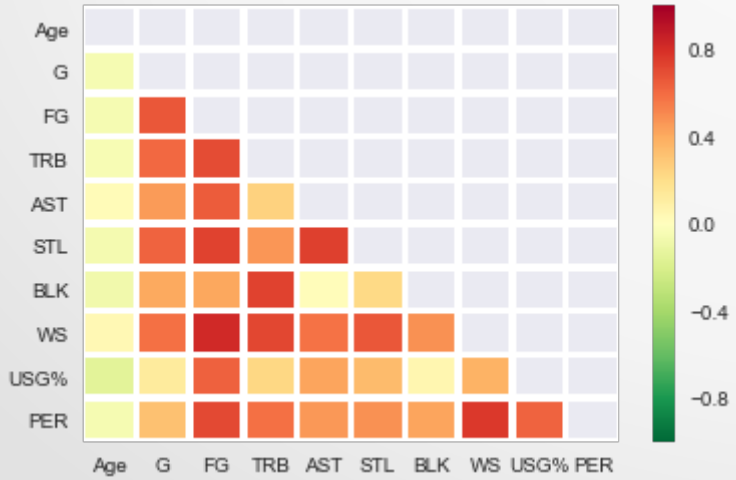
Exploratory Data Analysis

- To pre-screen the predictor variables, it is graphed against the response variable.
  - Good relationship is observed when there is a wide spread on x-axis and the best fit line has a strong positive or negative relationship.
  - Variables are visually picked out:
    - Age, GS, FG, 2P, MP, FTA, TRB, AST, STL, BLK, PTS, OWS, VORP, USG%, PER

Exploratory Data Analysis

- Correlation matrix is constructed to visualize multi-collinearity.

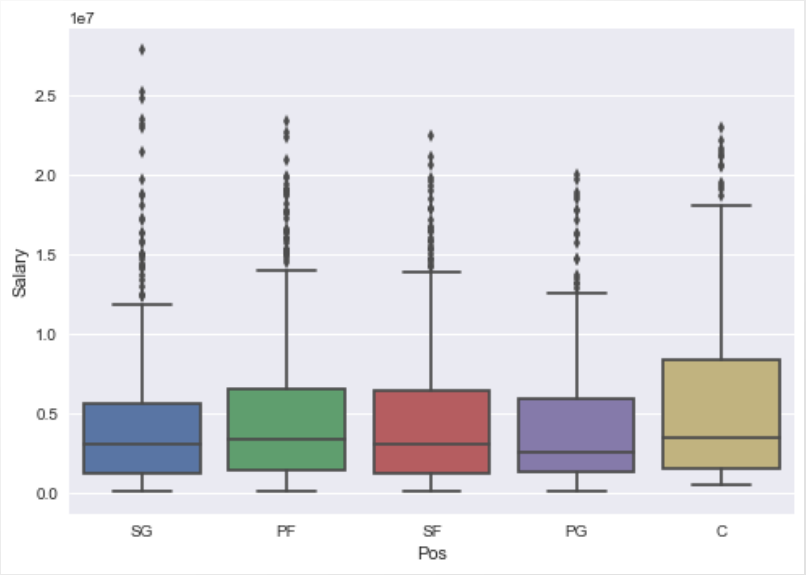




- Dark red relationships are removed.
- Reducing multi-collinearity should help with prediction accuracy

Exploratory Data Analysis

- Using hypothesis testing can help decide whether numbers are caused by chance.
  - Null hypothesis: salary difference does not exist between PG and SF
  - Alternative hypothesis: difference exist between PG and SF salary.
    - The player position graph is similar in value.
    - To find out if the differences are meaningful, a T-test can be done with alpha=0.05 which is 95% confidence interval.
- PG and SF are chosen due to its similar in value.
- T-test yields p-value of 0.1578 which is too high, therefore alternative hypothesis is rejected.

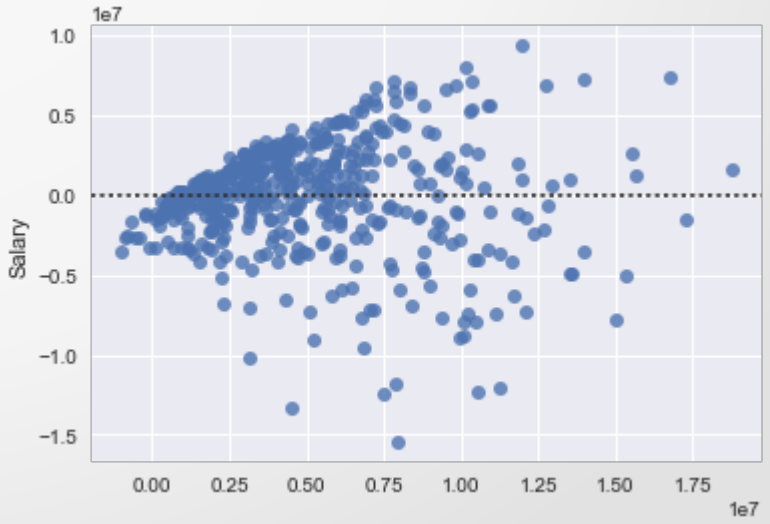


Model Building

- The machine learning algorithms applied:
  - Linear Regression
  - Ridge Regression
  - Random Forest
  - Gradient Boosting Regressor
- Train-test-split splits training and testing samples into 75% and 25% respectively
- Scores are calculated with coefficient of determination which is the proportion of the variance in the dependent variable that is predictable from the independent variable.

Model Building

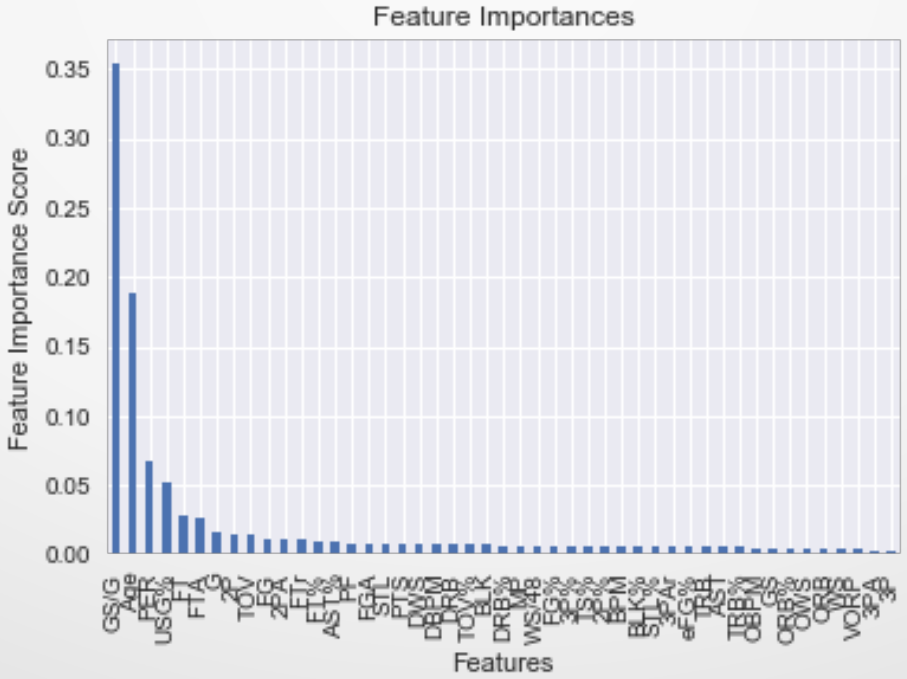
- Linear Regression
  - Fast and easily interpretable.
  - Low predictive power
- Works best for variables with a normal distribution
- Variables can be transformed using log.
- Hetoroscedasticity caused by predictor variables error terms not having the same variance.



Model Building

- Ridge Regression
  - Fast and alleviates multi-collinearity amongst regression predictor variables.
  - Alpha is the only parameter necessary for tuning

- Random Forest
  - Slow algorithm, but with high predictive power. Deals with over-fitting quite well.
  - Have a method which lists top features of importance.



Model Building

- Pick out the top 15 most important features from random forest method.
- Gradient Boosting yields the best result
  - GBM combines the weak to improve prediction accuracy.
  - Slow algorithm with high prediction power.
  - Low interpret-ability, often considered as a black-box algorithm.

Results

	No parameter: training	No parameter: testing	Tuned parameter: training	Tune parameter: testing
Linear (pre-selected X)	0.5323	0.5177		
Linear (all X)	0.5736	0.5686		
Ridge (pre-selected X)T	0.5323	0.5177	0.5323	0.5177
Ridge (all X)	0.5720	0.5678	0.5661	0.5629
Random forest (all X)	0.9200	0.5424	0.8875	0.6577
Random forest (top 15 X)	0.9251	0.6236	0.8971	0.6630
GBM (all X)			0.7841	0.6566
GBM(top 15 X)	0.7930	0.6652	0.7476	0.6753

Conclusion

It was unnecessary to pre-screen the predictor variables, since some methods have properties that can rank feature importance.

The accuracy increased from 53.23% to 67.53% meaning 67.53% of the testing points can be explained by the model.

Testing score is a generalization of how the model would perform on out of sample data or real world data.

This project can be easily converted to a classification problem for better interpret-ability.

- For example: if it was a binary classification problem, a socre of 50% or higher indicates a working model, since it is better than blind guessing.