Predicting NBA Player Salary using Linear Regression and Web Scraping

Prathap Rajaraman

Background

- **Motivation**
  - Disparities between contract size and player value can be the difference between championship and disappointment
- **Objective**
  - **A data driven approach for negotiating new contracts, trading for undervalued players, and trading away overvalued players**
  - Use NBA box score statistics to predict player salary via linear regression
  - Explore relationship and correlation between box score performance and salary

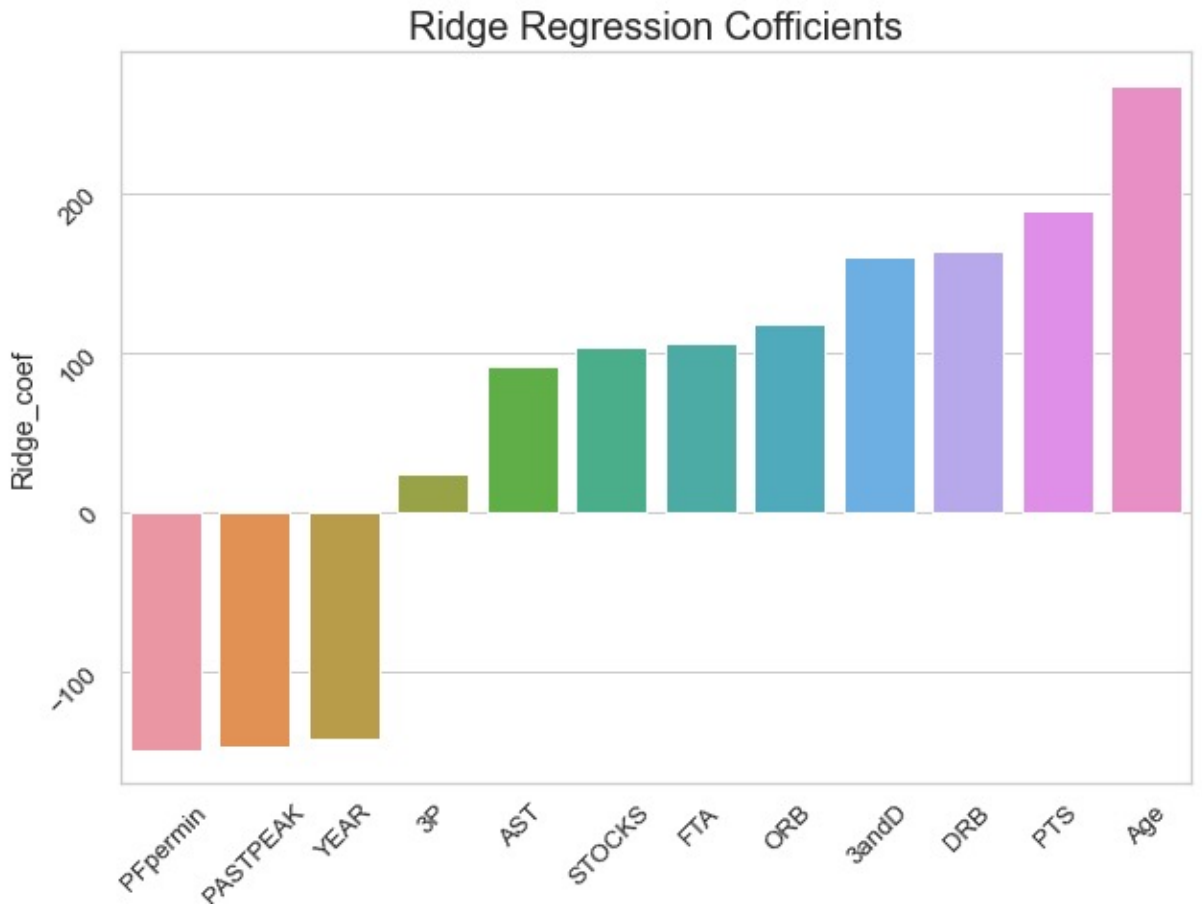
Methodology

- **Metrics**
  - Over 25 box score statistics analyzed, modeled, and evaluated for fit
  - Y variable is square root transformation of salary to reduce explosiveness and heteroskedasticity
- **Modeling**
  - Compared OLS, Ridge, and Lasso Regression models
  - Evaluation based on r squared, MAE, and intuitive fit

Feature Engineering and Variable Selection

- Variance inflation factor analysis was used to remove variables that have multicollinearity
- The following variables were also considered for analysis:
  - Stocks
  - 3 and D
  - Past Peak
  - Fouls Per Minute

Results



R Squared: 0.396
MAE: 816.89
MAE Squared: $667k

Anomalies

Player	Year	Points	Rebounds	Assists	Predicted Salary	Actual Salary
Pascal Siakam	2020	22.9	7.4	3.5	$15.5m	$2.4m
Andrew Wiggins	2019	18.1	4.8	2.5	$10.3m	$27.3m
Draymond Green	2019	7.4	7.3	6.9	$11.1m	$18.7m

- Large disparities can be attributed to:
  - Undervalued players
  - Bloated contracts
  - Players known for intangibles

Conclusion

- Ridge Regression offers the best combination of intuitive fit and accuracy
- Scoring, rebounding, versatility, and avoiding fouls seem to be the strongest variables linked to player salary
- Anomalies can be used to identify over or undervalued players
- This model is limited by the value that intangibles can bring to a team

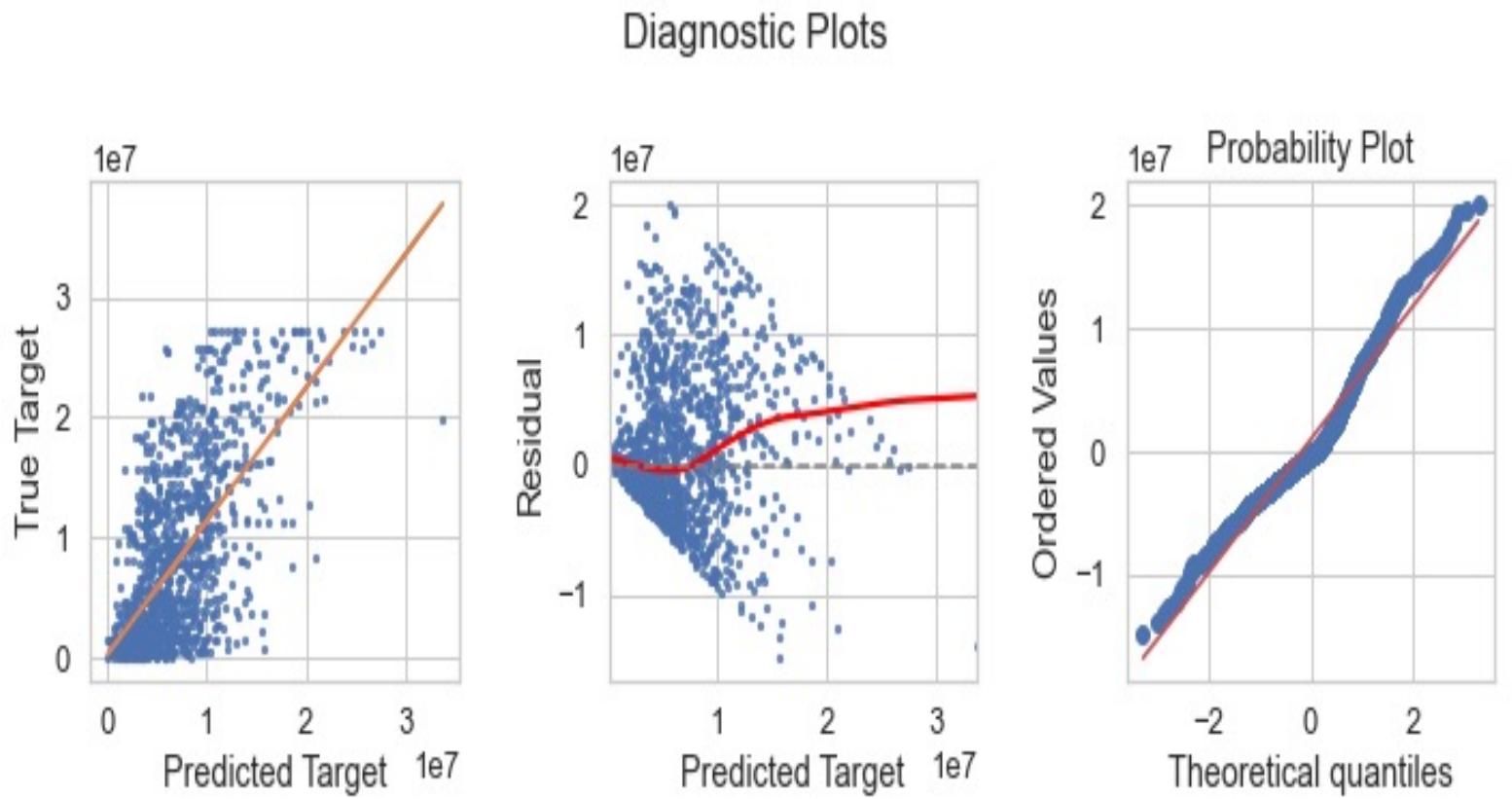
Appendix

Data and Tools

- **Data Filtering and Sources**
  - Stats from 2018 – 2021 via basketball-reference
  - Salary from same time frame via ESPN
  - No rookies or superstars
  - Salaries scaled to 2021 dollars

- **Tools Used**
  - Pandas and NumPy for data analysis and manipulation
  - Scikit-Learn and Statsmodels for statistical analysis
  - Matplotlib and Seaborn for visualization
  - Beautiful Soup for web scraping

Diagnostic Plots for Ridge Regression Model



Coefficients

Field	OLS Coefficient	Lasso Coefficient	Ridge Coefficient
YEAR	-138.16	-144.15	-141.08
Age	87.08	287.87	267.03
PASTPEAK	-1132.72	-160.26	-146.54
3P	-108.41	0	24.11
FTA	38.95	74.18	105.35
ORB	147.85	109.76	117.69
DRB	104.41	172.44	163.54
AST	54.9	81.71	91.82
STOCKS	80.71	91.21	103.36
PFpermin	-5313.62	-149.24	-148.51
PTS	54.12	241.89	188.22
3andD	181.42	168.35	160.02

Model Evaluation

Model	r^2 Train	r^2 Test	MAE	MAE squared
OLS	0.3804	0.398	809.7721	655,730.816
Lasso	0.3801	0.3998	813.6764	662,069.258
Ridge	0.3801	0.3966	816.8943	667,316.258