Predicting NBA Salaries to Help Players and Agents

DSI 11 Capstone – Markell Jones-Francis

Problem Statement

Negotiating NBA contracts is complicated. While outside factors such as interpersonal relationships and star power play a role, the statistics that a player records on the court and their ability to help a team are the most significant factors at the negotiating table. To streamline this process for players and agents, we will be building a machine learning model using each player's basic, advanced, and team statistics in order to accurately predict their salary. To do this we will be using Linear Regression, Lasso, Ridge, KNN, Decision Tree, Random Forest and SVM models and, given the scale of NBA contracts, will be looking to make predictions that are accurate within $2,000,000 per year, using RMSE as our metric.

We will be using data from the 2016-17 to the 2018-19 seasons for these projections, because the 2016-17 season saw the NBA salary cap increase to $94M, from $70M the season before, as a result of the NBA's then newly signed television deal which brought in additional revenue. This cap increase resulted in salary increases across the board for NBA players, so much so that any salaries 2016-17 bear little comparative value when evaluating statistical performance. We are also ending the analysis at the 2018-19 season as that is the most recent completed NBA season.

With this tool we hope to assist agents find the fairest salaries for the players they represent, and determine which teams are most likely pay their clients the salary that they deserve.

Data Collection and Cleaning

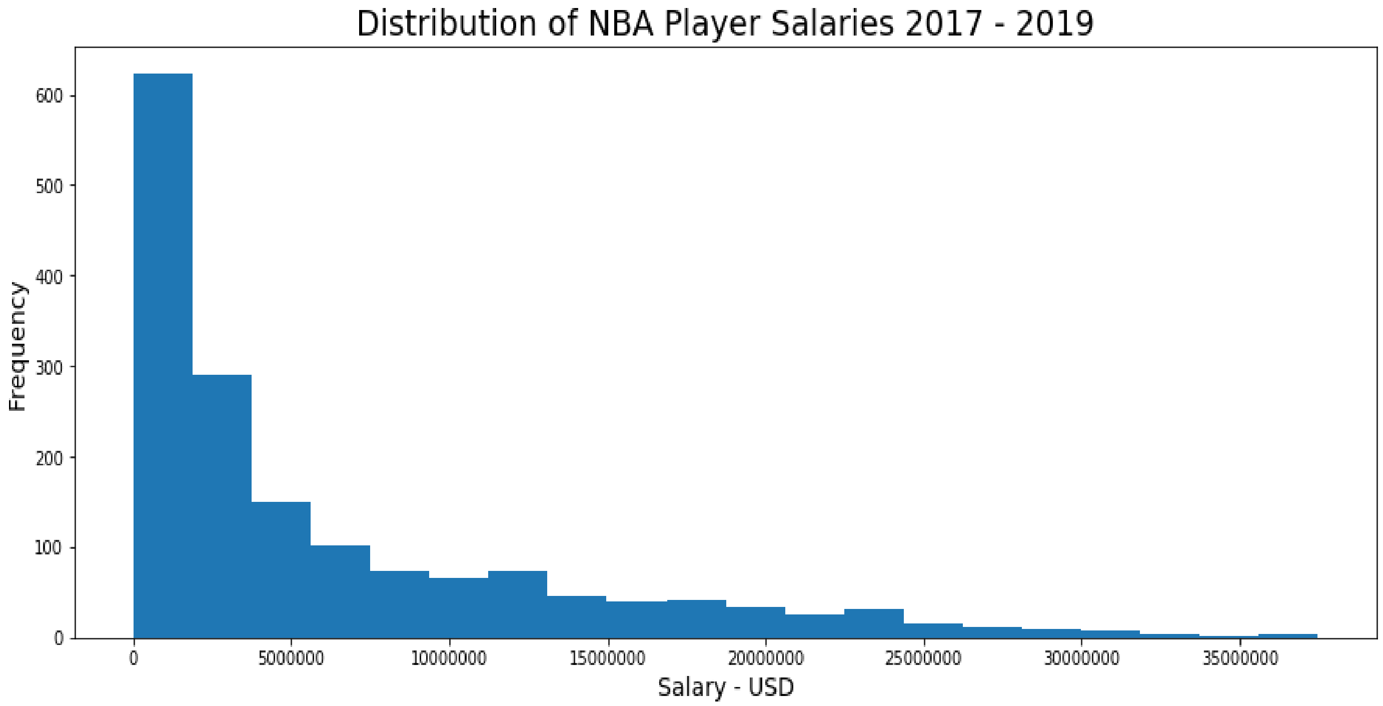
Data Collection

- Examining 1,651 player seasons from the 2016-17 Season through the 2018-19 Season
- Used BeautifulSoup4 to scrape NBA statistical data from multiple webpages:
    - Basic and Advanced individual stats from Basketball Reference
    - Team Stats from ESPN
    - Salary Data from Hoops Hype
- Ignored aggregate data for players that played for more than one team in a given season
- 61 Features excluding Salary and Player Name

Data Cleaning

- Standardized player alphabet by removing accents from player names
- Converted all data to the correct data types
- Nulls were replaced with 0s – the only null values were shooting percentage stats for players who did not qualify because they had made 0 shots in that category

Not Sharing The Wealth



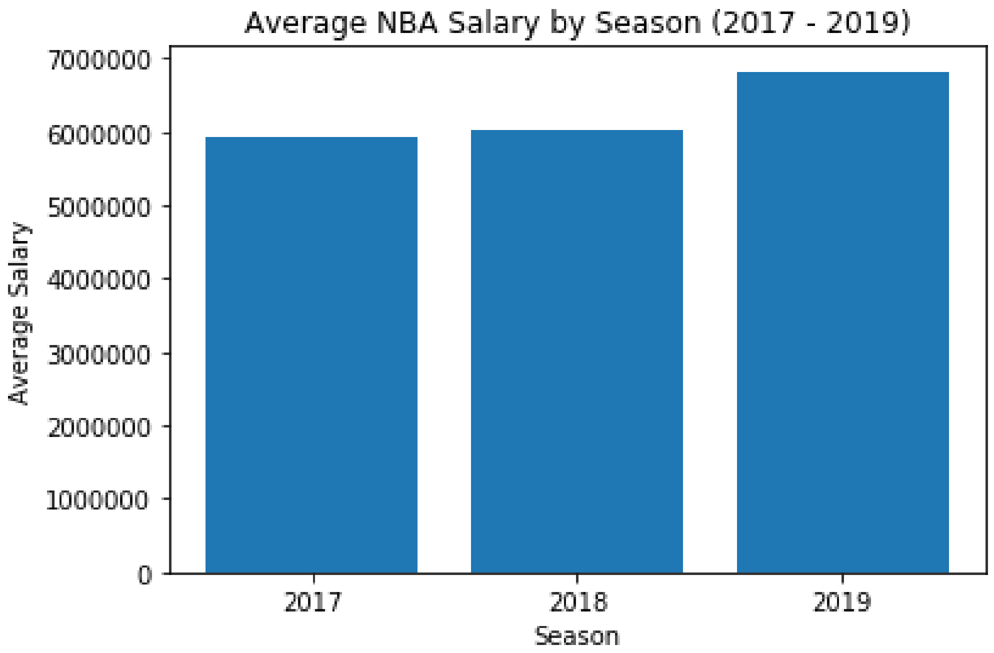
- ❖ Average Salary - $6,263,335.33
- ❖ Median Salary - $3,000,000
- ❖ Lowest Salary - $24,022.00
- ❖ Highest Salary - $37,457,154
  - ❖ Steph Curry – 2018-19

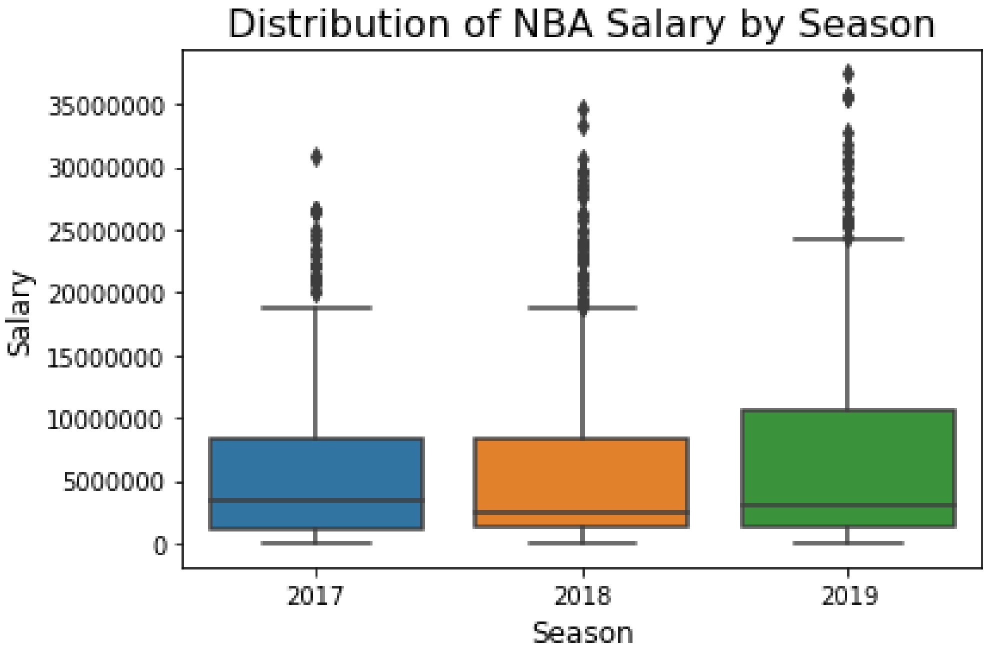
- The NBA has what is known as a soft salary cap, meaning that each team has a maximum amount of money that they can spend on salaries for their players and each team has 12 active players at any given time

- Most teams have at least one player signed to a “Max” Contract that can be 25-30% of the salary cap depending on certain criteria that the player meets

- A few select players are also eligible for the “SuperMax” Contract which allows them to paid 35% of the salary cap

Seasonal Changes



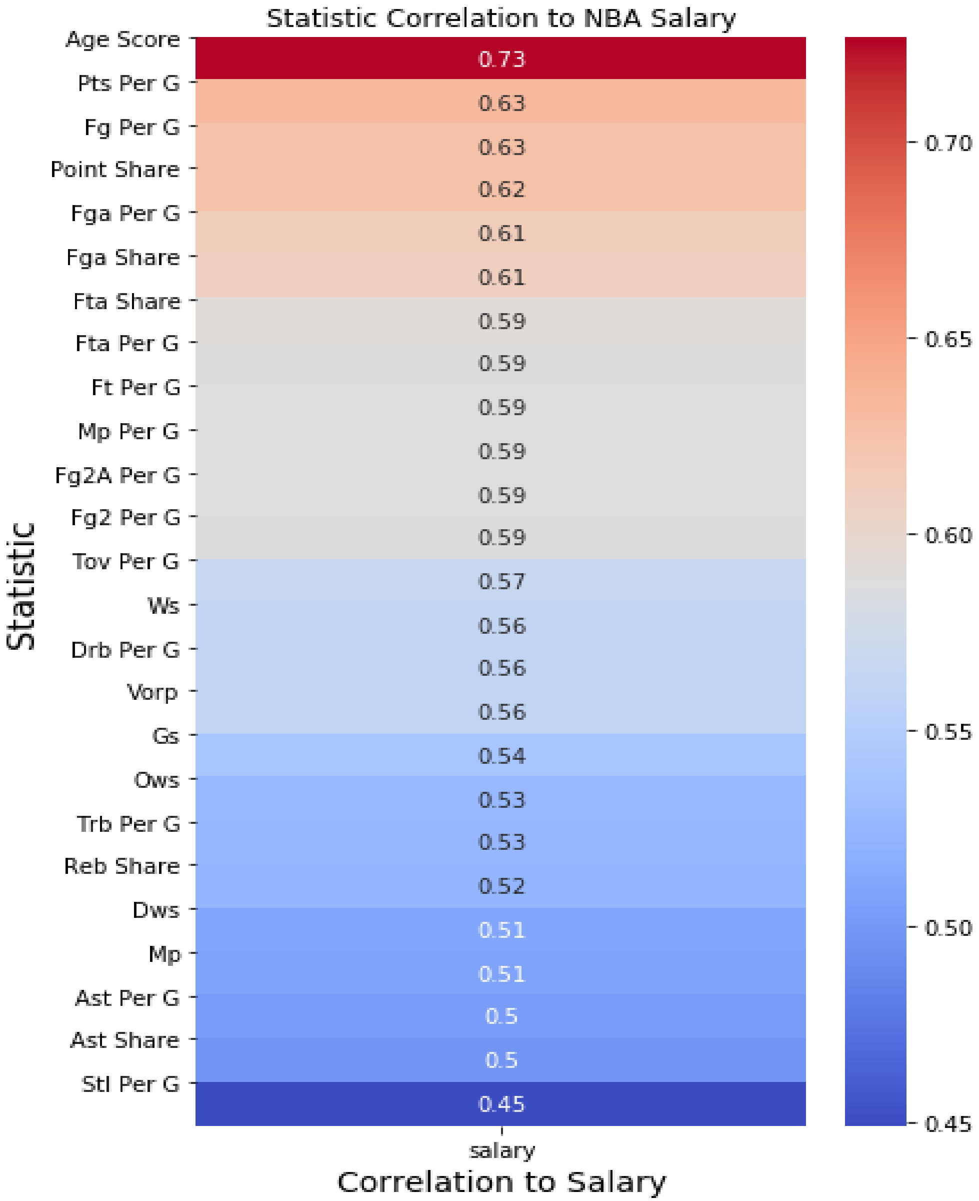


- The salary cap increased from $94M in 2016-17 to $99M in 2017-18 to $101M in 2018-19. Because of this players signing contracts in subsequent years will have higher salaries on average than similar players that signed in the years before.

- The average NBA Salary increased by $912,256.39 between 2016-17 and 2018-19.

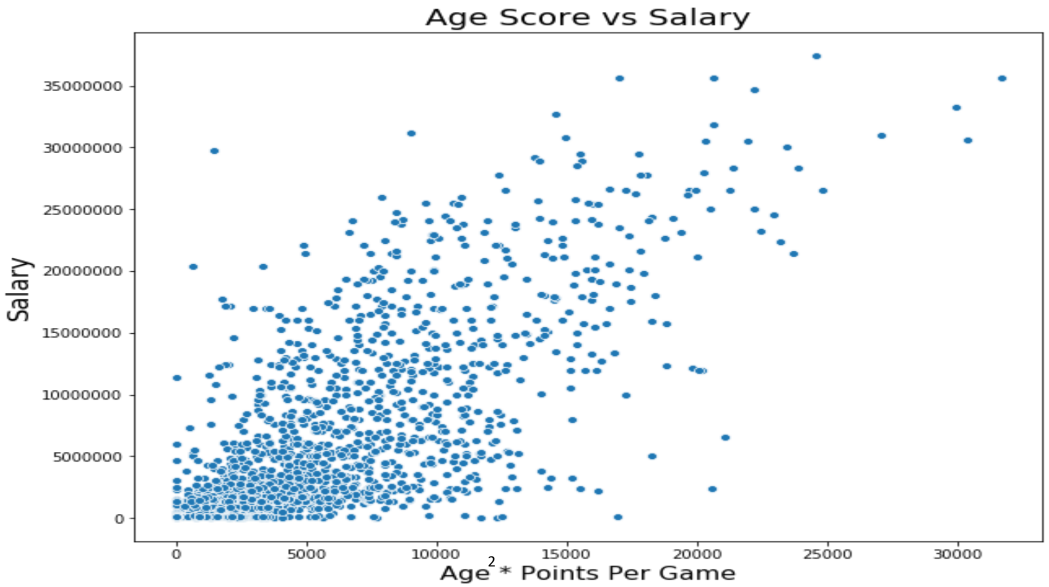
- The highest salary increased by $6,493,704.00 between 2018-19

Important Features

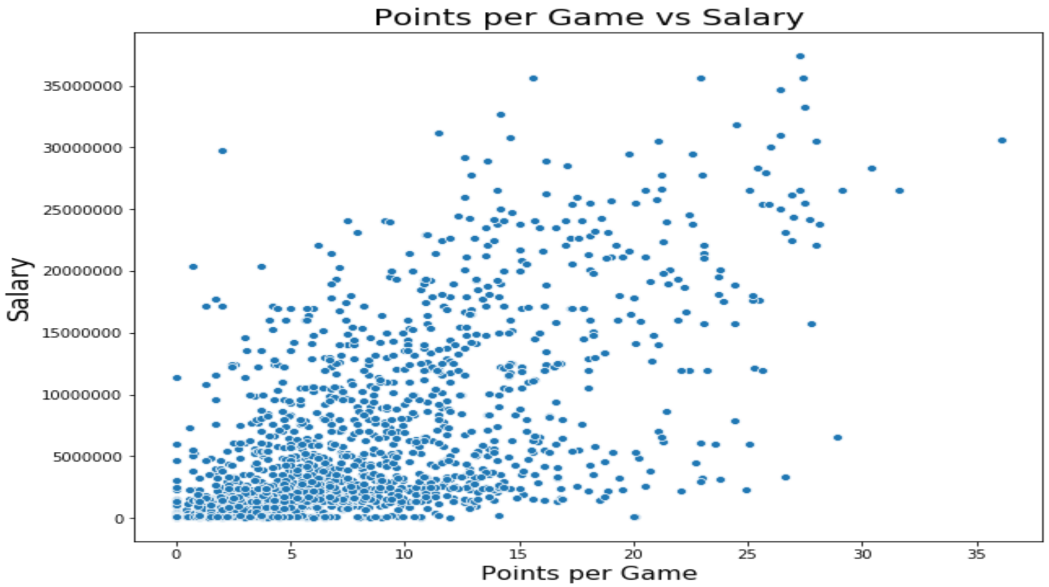


- The statistic most correlated to Salary is our created Age_Score feature (Age^2 * Points Per Game)
- Only 4 of the top 25 are Advanced Statistics
  - Win Shares (WS), Offensive Win Shares(OWS), Defensive Win Shares(DWS), and Value Over Replacement Player(VORP)
- The statistics most highly correlated to Salary are overwhelmingly related to scoring
- The only statistic with a negative correlation to Salary is 3-Point Field Goal Attempts per Field Goal Attempts (-.091)
- Team Statistics all have a correlation to Salary that is .08 or lower

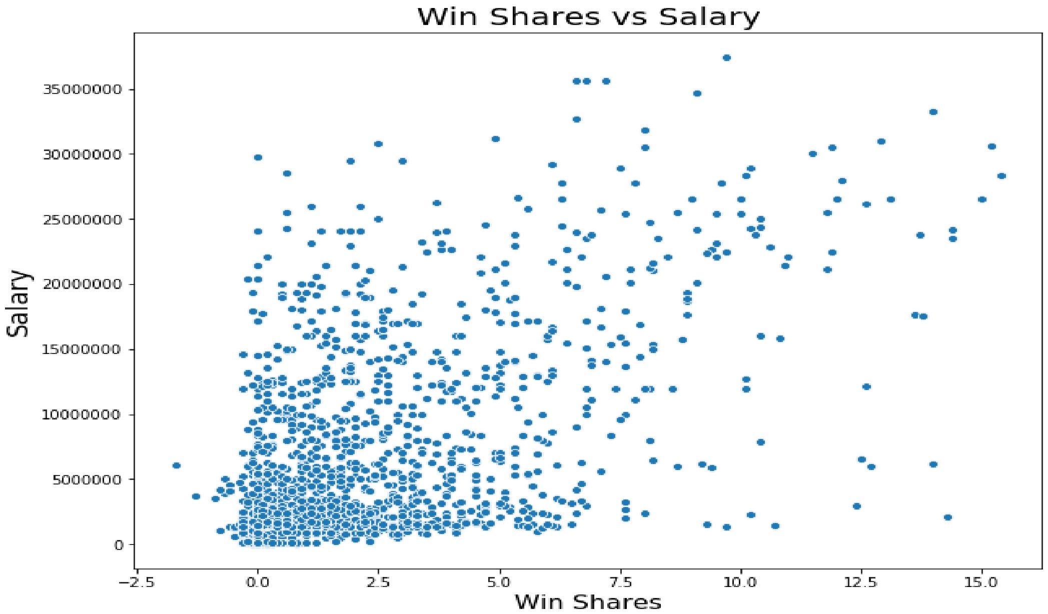
Important Features



Age Score vs Salary

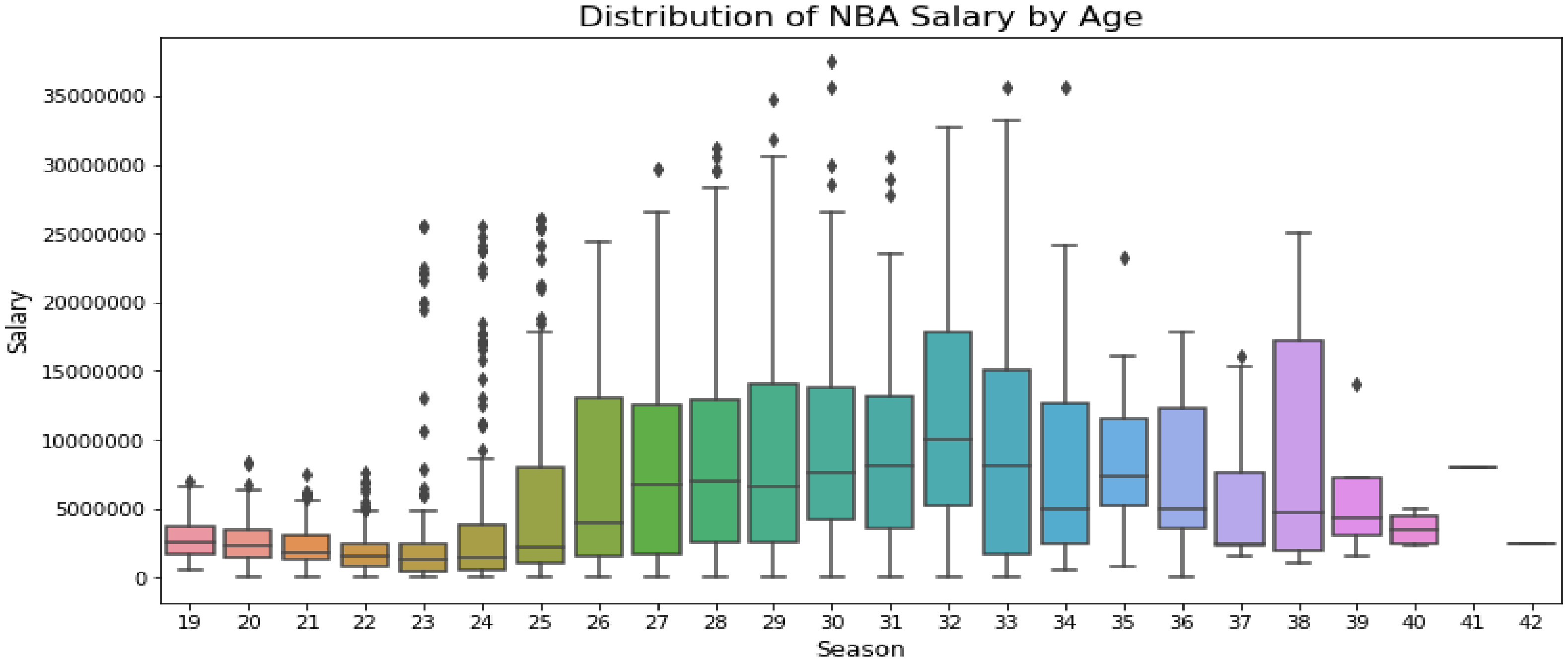


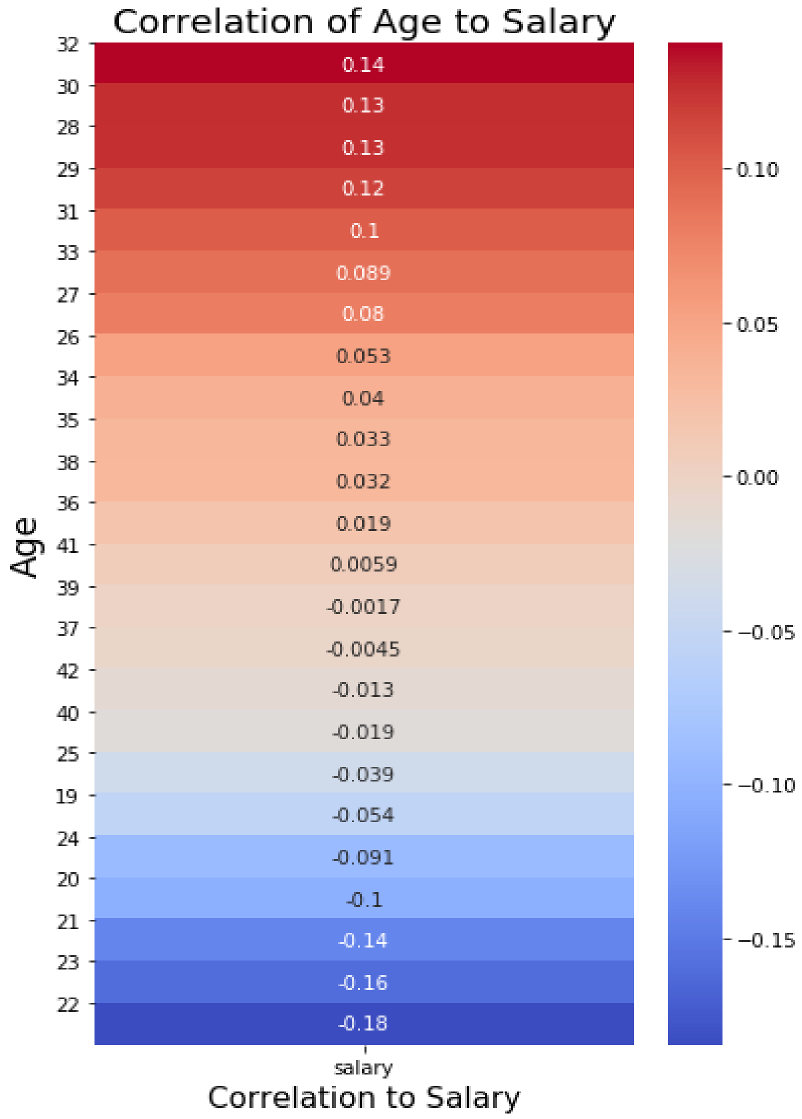
Points per Game vs Salary



Win Shares vs Salary

The Impact of Age


Distribution of NBA Salary by Age


Correlation of Age to Salary

- Average NBA Player Age is 26

- Players drafted in the first round are subject to a fixed Rookie Salary Scale in which their salary is determined by their draft position

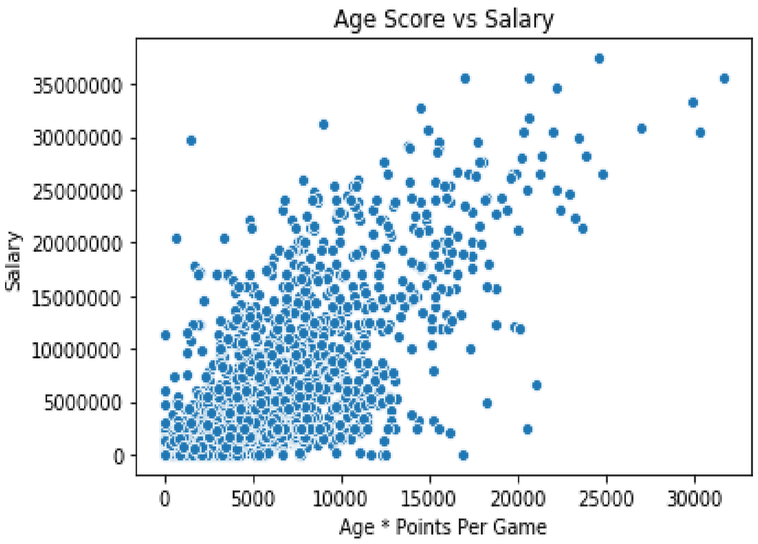
- The highest average salaries are when players are between 27 and 32 years old – this coincides with many players are eligible for large extensions

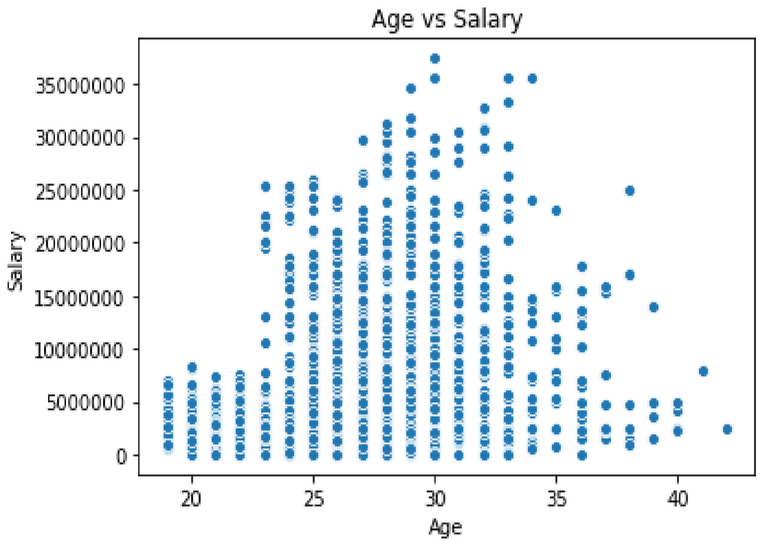
K-MEANS Clustering

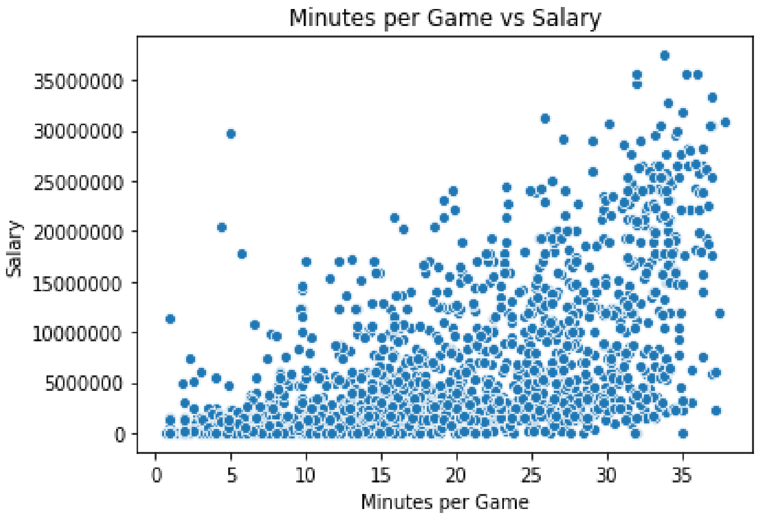
- Created 12 clusters of players in order to create distinctions between players that may share some similar stats but are likely to have different salaries

- Clustered players on the created Age_Score, Age, and Minutes Played Per Game

- Cluster criteria chosen in order to account for scoring, differentiate between age, and account for importance

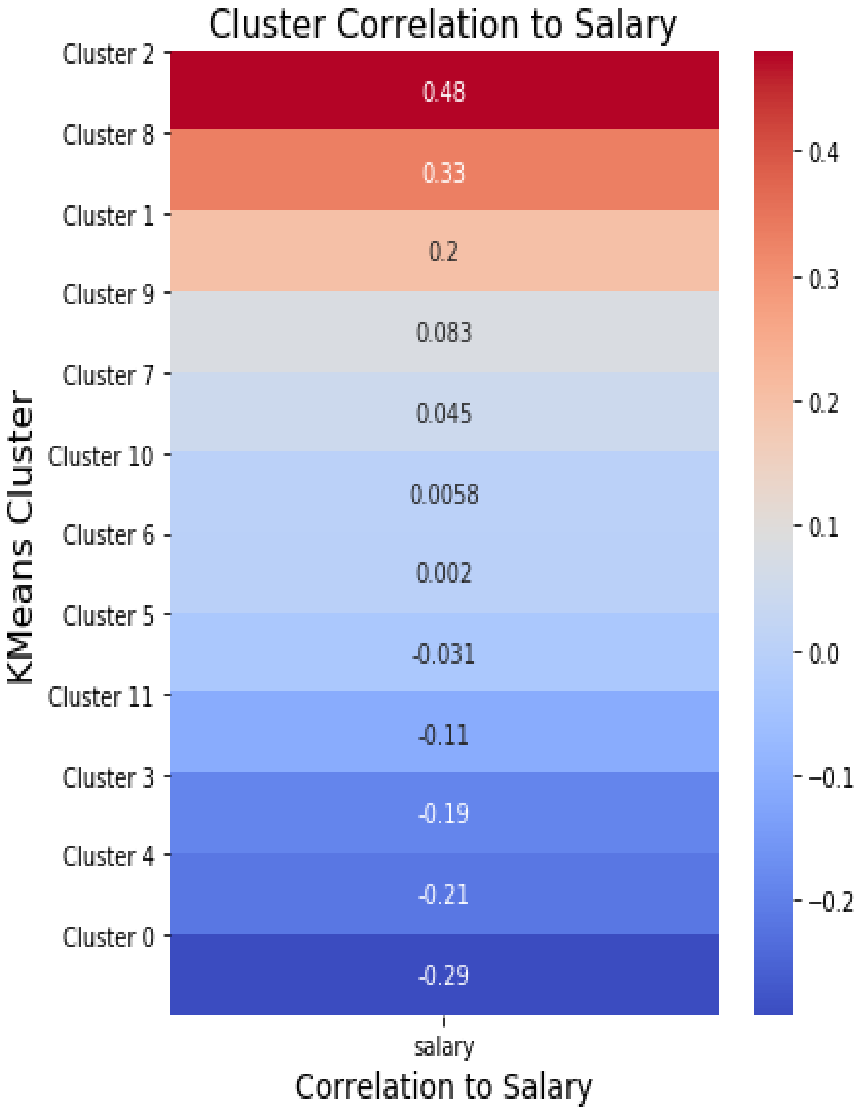

Age Score vs Salary


Age vs Salary


Minutes per Game vs Salary

Clustering Results

Cluster	Age	Games Played	Games Started	Minutes Per Game	Points Per Game	Salary
2	29.83	64.45	63.67	33.91	22.4	$23,533,836.09
8	26.52	60.66	55.39	32.07	18.34	$14,802,999.57
1	30.35	56.13	38.2	26.55	10.77	$11,062,603.30
9	36.09	54.64	17.89	22.22	9.46	$9,763,421.40
7	25.98	56.75	29.3	25.28	9.93	$7,234,204.70
10	22.33	64.35	54.02	30.52	15.04	$6,411,301.87
6	28.69	43.24	9.16	17.1	6.04	$6,307,986.41
5	33.9	35.17	5.18	12.68	4.01	$5,353,815.78
11	28.2	18.17	1.03	7.17	2.3	$3,447,058.55
3	21.56	53.69	18.12	19.99	7.41	$2,341,780.59
4	24.15	38.98	4.24	14.45	4.98	$2,238,899.18
0	22.33	16.14	0.69	6.57	2.1	$957,660.34

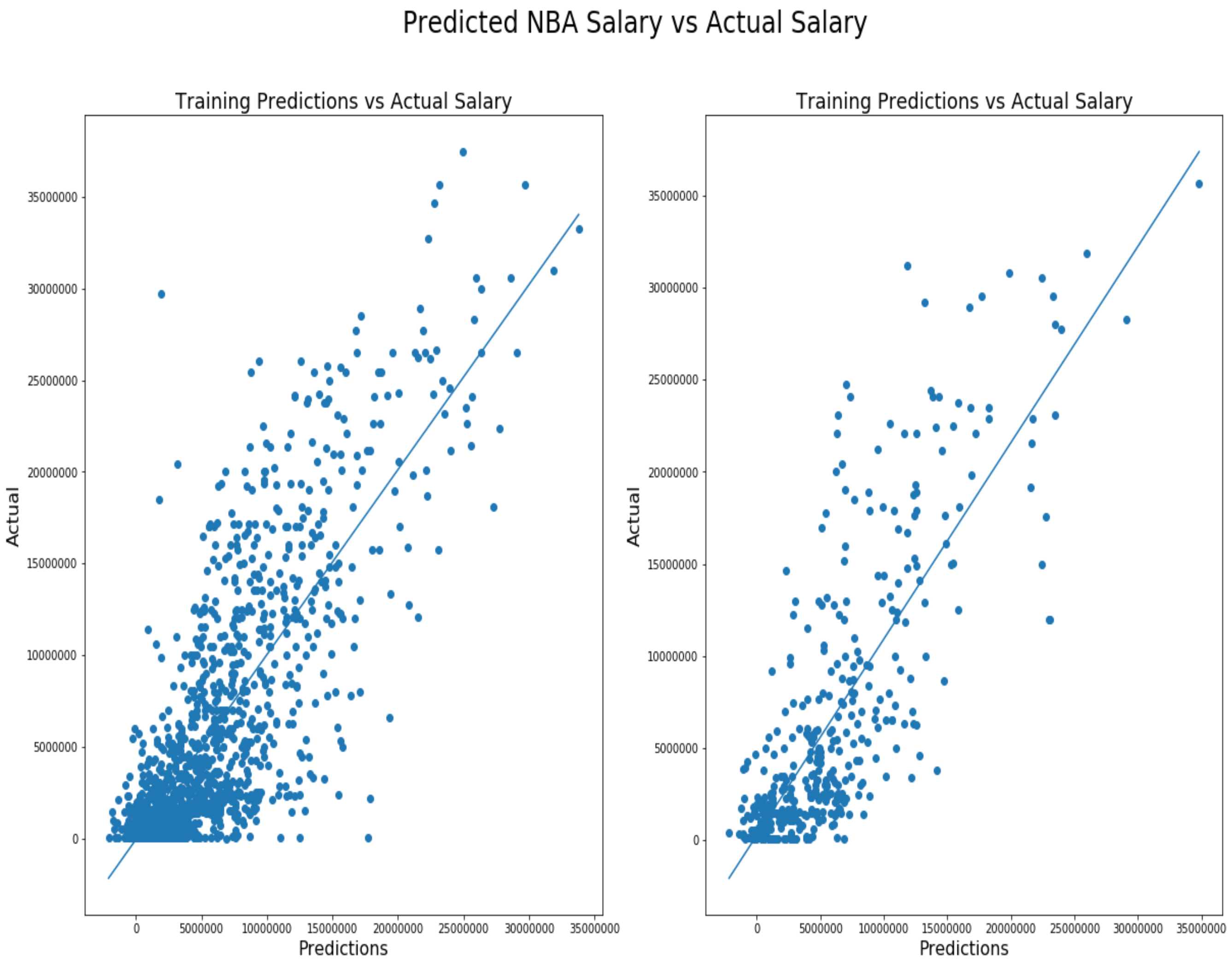


Model Selection

Model	Train RMSE	Test RMSE
Ridge	$4,395,017.13	$4,662,334.20
Linear Regression	$4,415,305.36	$4,738,569.10
Lasso	$4,336,575.82	$4,806,486.83
Random Forest	$2,551,553.13	$4,859,316.60
K Nearest Neighbors	$0.00	$5,147,463.72
Decision Tree	$4,353,970.10	$5,242,285.06
Baseline	$7,116,919.02	$7,579,804.27
Support Vector Machine	$7,809,224.83	$8,422,557.34

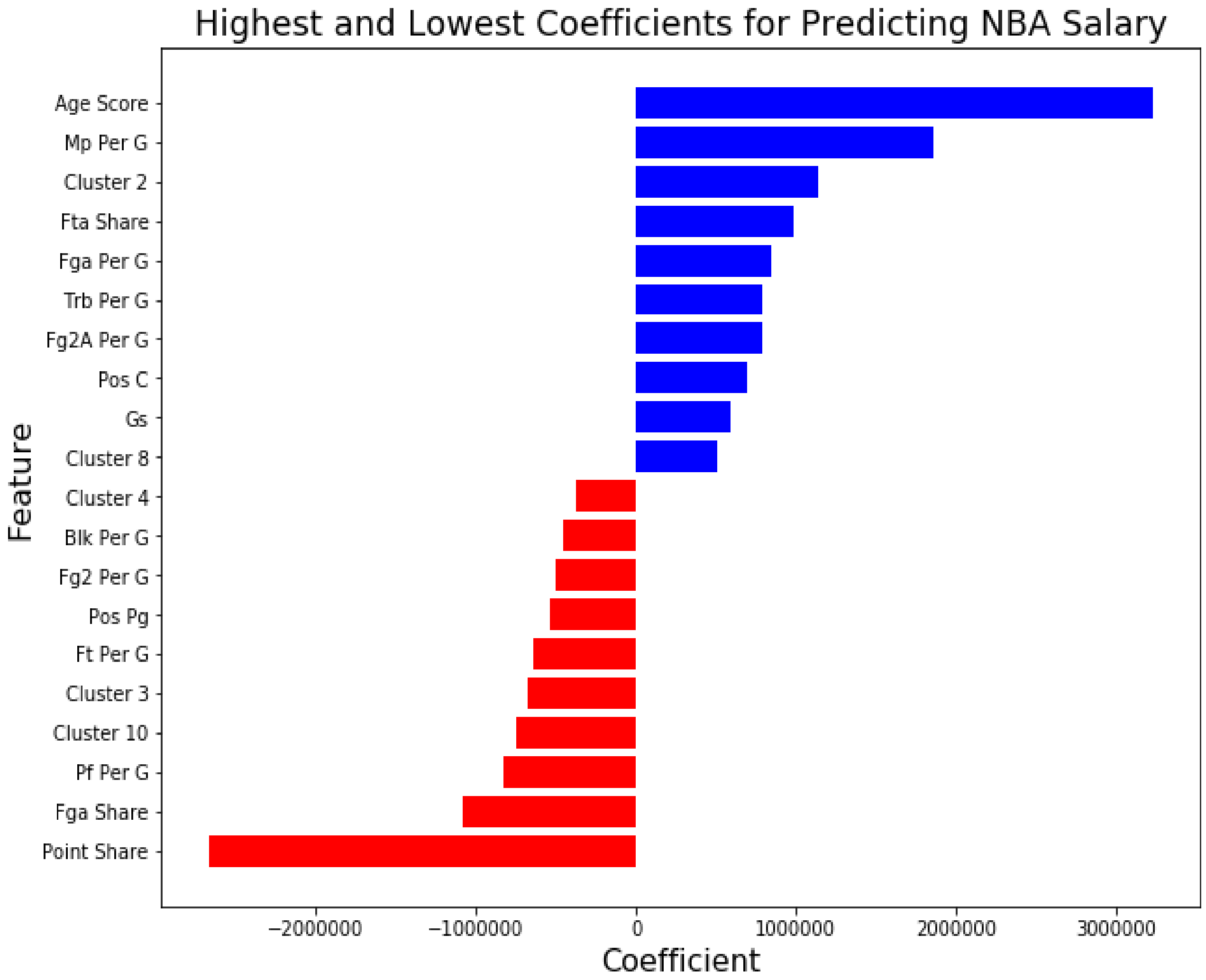
- Ridge was our highest performing model
- The feature set used to train the ridge model included the clusters, basic statistics, created features, seasons, and positions
- Ridge has a high degree of error but offers a significant improvement over the baseline
- Fit on the correct number of features as the train and test score are very close

Model Predictions



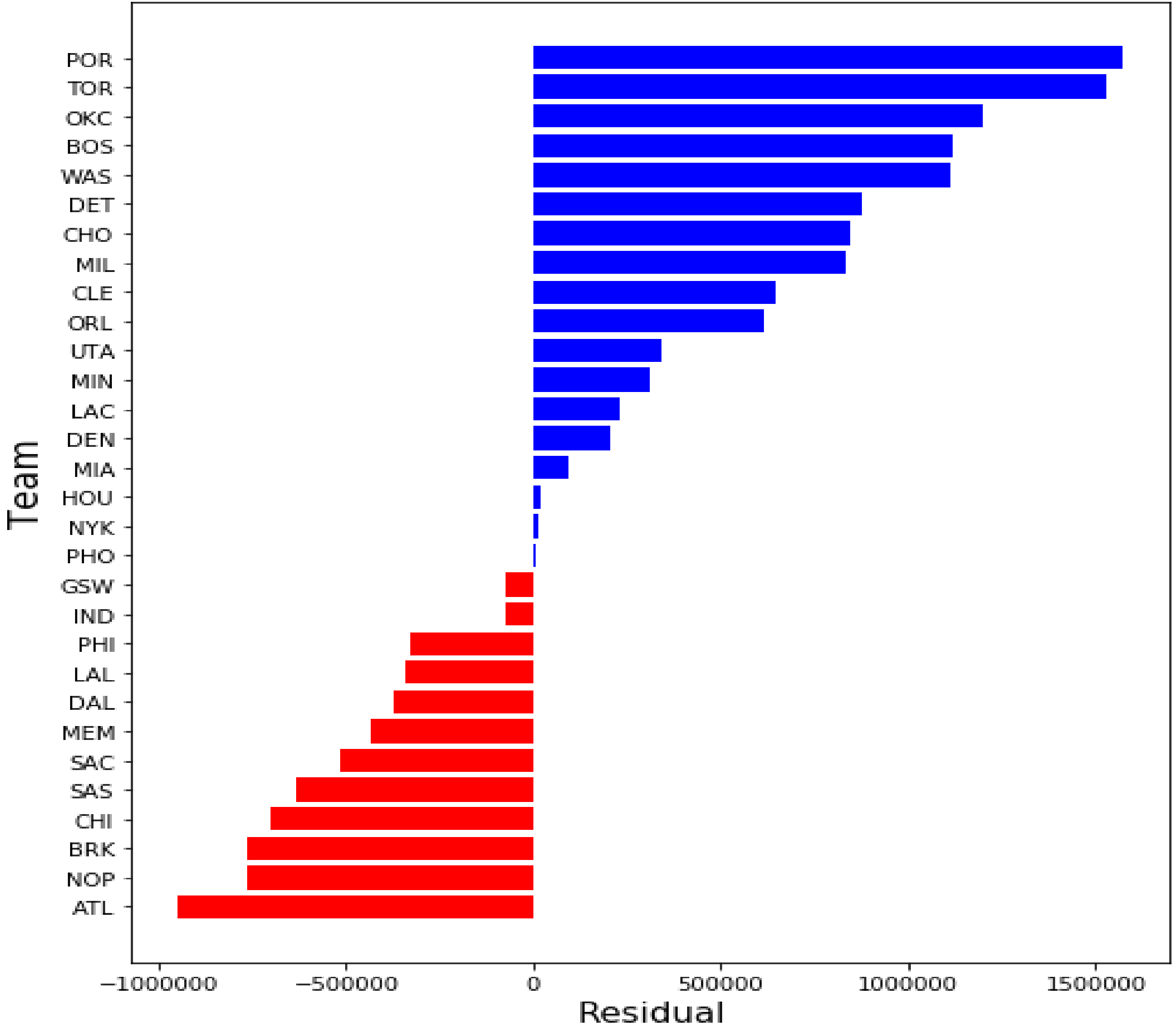
- Our model has an average error of $4,662,334.20
- Our Model is better at predicting the highest salaries
- Struggles with in the mid-to high range
- Often underestimates when predicting
- Provides negative predictions – the least productive players in the NBA
- Only 61 – 62% of the variance in Salary is explained by the features in our model

Most Impactful Features



- Our model was trained on a total of 52 features
- Age_Score had the most impact on the prediction of Salary
    - For each unit that Age_Score increased, Salary increased by $3,232,871.86 (with all other factors being held equal)
- Many of the features in the feature set are related to one another
- 29 of the features had a positive coefficient
- 23 of the features had a negative coefficient

Team Trends



Average Over-Pay and Under-Pay by Team

- The Portland Trail Blazers are the most likely to overpay:
  - Portland overpays players by $1,572,797.87 on average
  - 67% of the players signed by Portland have contracts that exceed their statistical value

- The Atlanta Hawks are the most likely to underpay:
  - Atlanta players are underpaid by $947,813.93 on average
  - 79% of players signed by the Hawks are underpaid

- Overpaying and Underpaying are not always a good barometer to judge how well a team is managed

Exceptions To The Rule

There are many factors that influence NBA player salaries that cannot be accounted for using basic and advanced statistics. Many players' contracts are affected by both team and individual circumstances leading to our model can having trouble with predictions in which these external factors play a major role.

- Injury:

Player	Team	Age	Points Per Game	Minutes Played per Game	Salary	Prediction	Residual
Gordon Hayward	BOS	27	2	5	$29,727,900.00	$1,884,748.95	$27,843,151.05

- Impact Beyond Statistics:

Player	Team	Age	Points Per Game	Minutes Played per Game	Salary	Prediction	Residual
Draymond Green	GSW	26	10.2	32.5	$15,330,435.00	$12,415,347.83	$2,915,087.17
Draymond Green	GSW	27	11	32.7	$16,400,000.00	$13,637,601.69	$2,762,398.31
Draymond Green	GSW	28	7.4	31.3	$17,469,565.00	$12,715,912.46	$4,753,652.54

- Interpersonal:

Player	Team	Age	Points Per Game	Minutes Played per Game	Salary	Prediction	Residual
DeMarcus Cousins	SAC	26	27.8	34.4	$15,756,438.00	$23,086,878.09	-$7,330,440.09
DeMarcus Cousins	NOP	26	24.4	33.8	$15,756,438.00	$18,042,911.79	-$2,286,473.79
DeMarcus Cousins	NOP	27	25.2	36.2	$18,063,850.00	$27,312,045.20	-$9,248,195.20
DeMarcus Cousins	GSW	28	16.3	25.7	$5,337,000.00	$15,604,517.81	-$10,267,517.81

Conclusion and Recommendations

What We Found

- There is far more than just raw statistics that goes into negotiating NBA contracts
- It is easier to predict the salaries of NBA players in the open market than those that have policy-driven restrictions
- Players looking to extract more value out of their contract should sign with Portland, Toronto, or Oklahoma City

How to Use this Tool

- This model should be used to establish a starting point for salary negotiations
- Agents and players can use this tool to determine likely value and find the seek out the most lucrative contract
- General Managers can use this model evaluate their own signings, on a player by player basis

Next Steps

- Further improve our model's predictive capability, by adding more features to our dataset that go beyond the scope of on-court statistics

- Added information such as injury history, draft position, playoff success, and contract eligibilities would provide greater context for each player so that we can make more accurate predictions

- We would also like to add data from more seasons with appropriate salary cap context, so that our model will be able to distinguish between eras