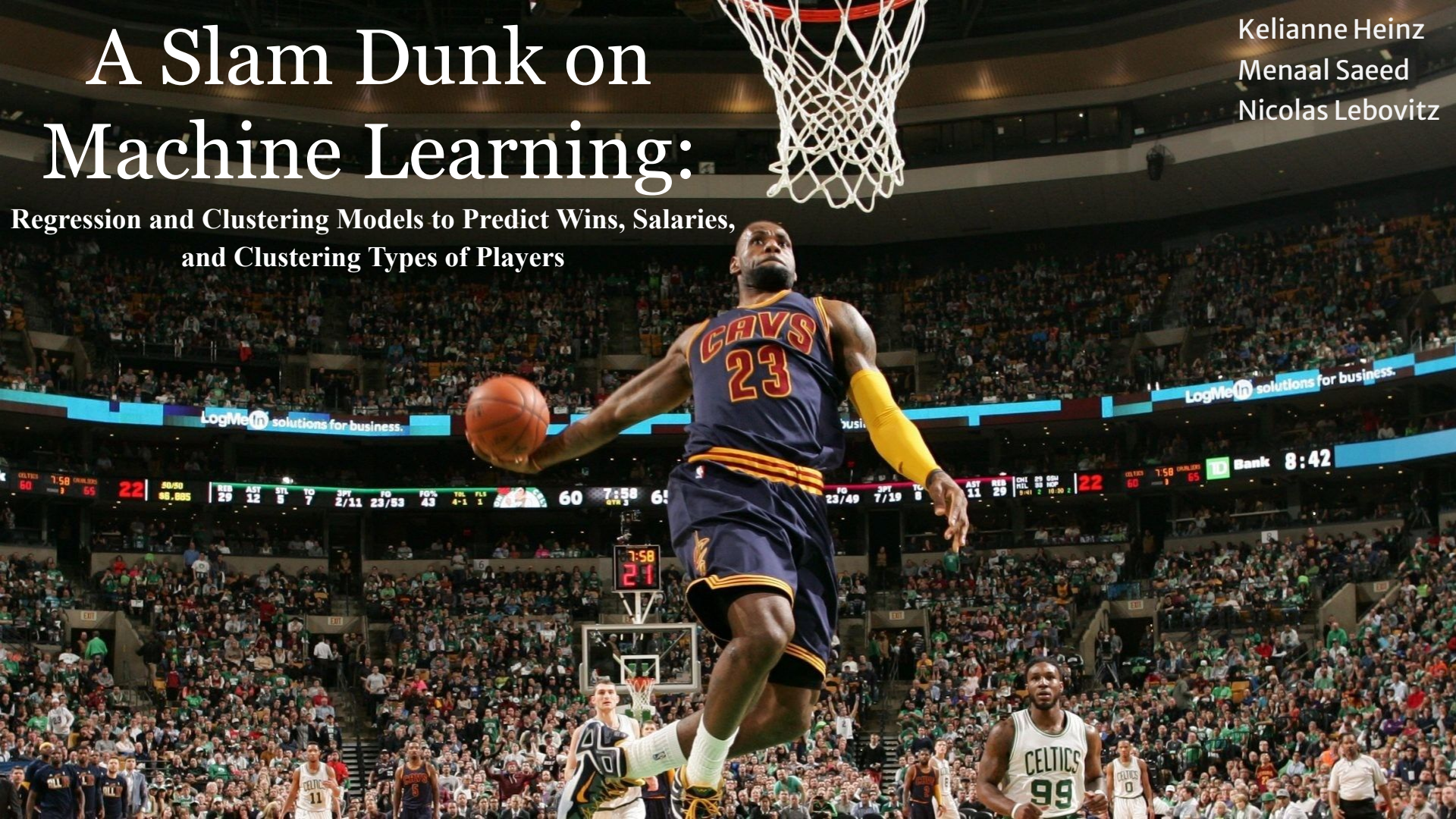
A Slam Dunk on Machine Learning:

Regression and Clustering Models to Predict Wins, Salaries, and Clustering Types of Players

Kelianne Heinz
Menaal Saeed
Nicolas Lebovitz



Agenda

1 Problem Statement and Objective

2 Approach and Methodology

3 Model Diagrams

4 Exploratory Data Analysis & Models

5 Evaluation (Measure of Success)

6 Comparison, Limitations, and Future Work

Create 3 Machine Learning Models on NBA Game and Player Data to determine likelihood of winning a game, predict salaries of players and group similar types of players.

Game Statistics

Home Team
Away Team
Season
Win/ Loss Records
Advanced team stats
Previous game boxscores

Salary Information

Player Name
Team Name
Season Start
Season End
Salary in USD

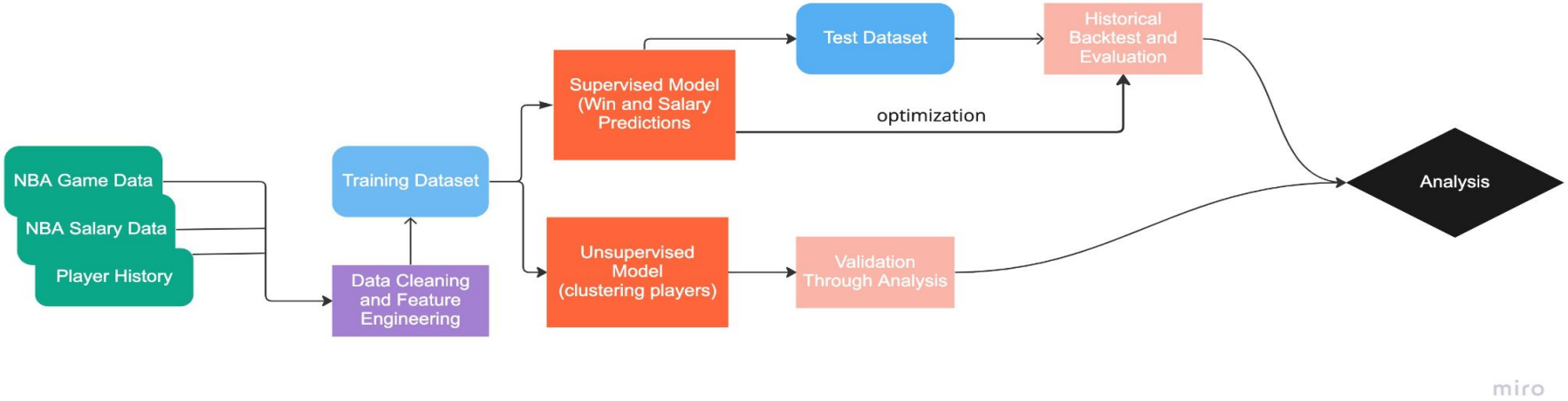
Advanced Player Statistics

Player Name
Team Name
Position
Age
Minutes Played
Games Started
Player Efficiency Rating
...

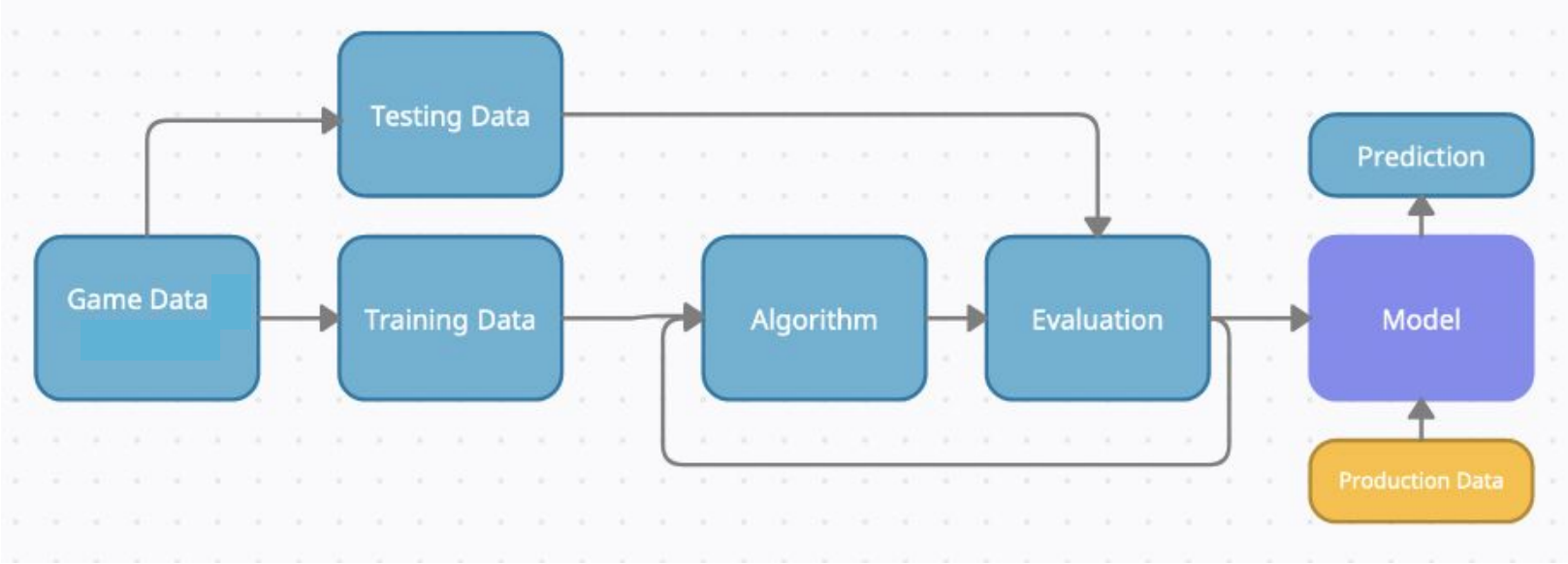
Types of Models to Build

	First Model	Second Model	Third Model
Method	Predict the winner of NBA matchups across each regular season, classifying each team's game outcome as a win or loss	Predict salary for a given player in the NBA given advanced statistics relating to games	Identify clusters of similar types of players in the NBA to determine trends in salaries and other advanced statistics for various types of players
Type	Supervised Logistic Regression	Supervised Linear Regression	Unsupervised Clustering
Models	Naive Bayes Decision Trees Stochastic Gradient Descent Random Forest	Linear Regression Stochastic Gradient Descent Regressor Gradient Boosting Regressor Random Forest Regressor	K Means Clustering Agglomerative Clustering Gaussian Mixture Method

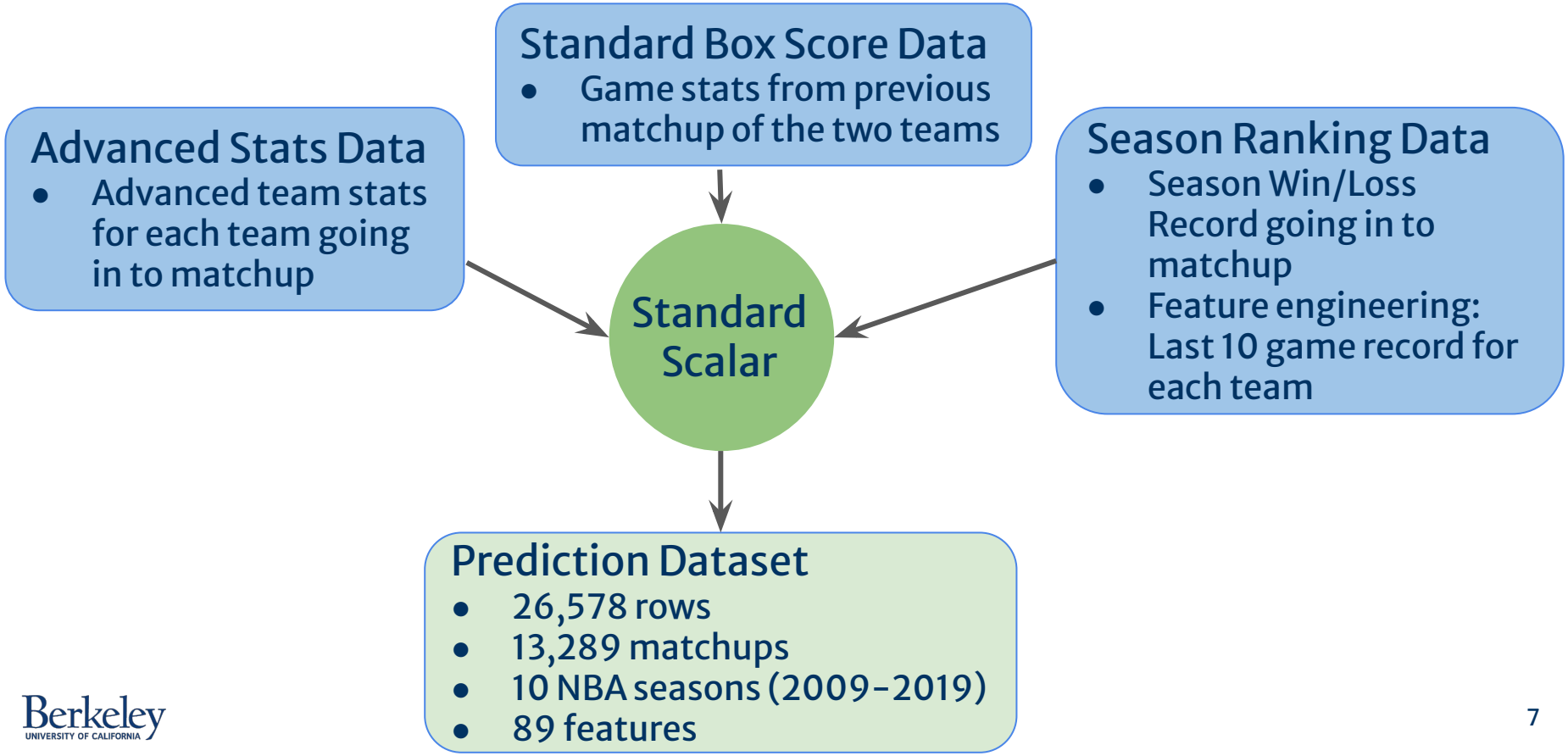
Block Diagram for ML Methods



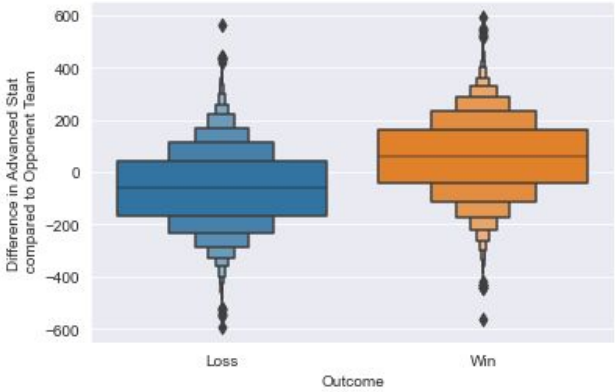
Block Diagram for Supervised ML Methods



Logistic Regression for Predicting Game Wins: Data

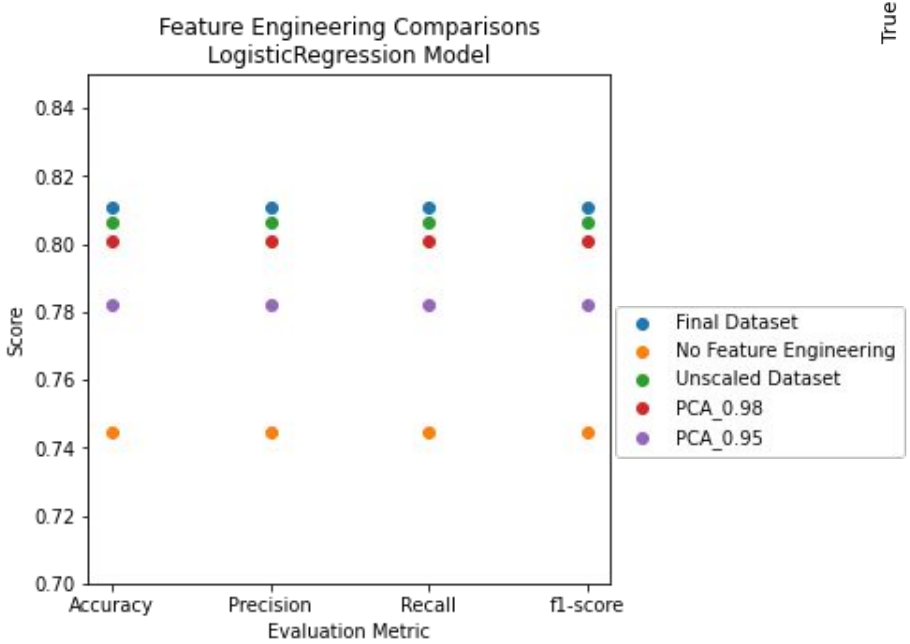


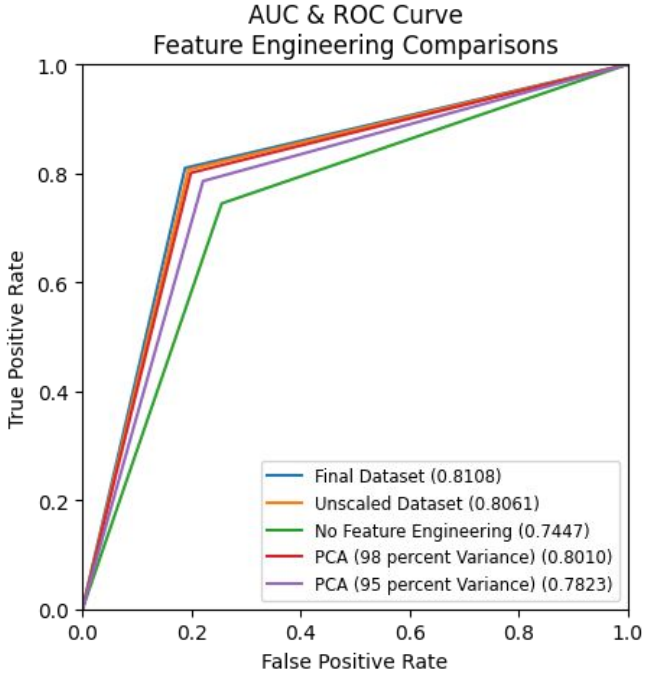
Logistic Regression for Predicting Game Wins: EDA



Feature Engineering

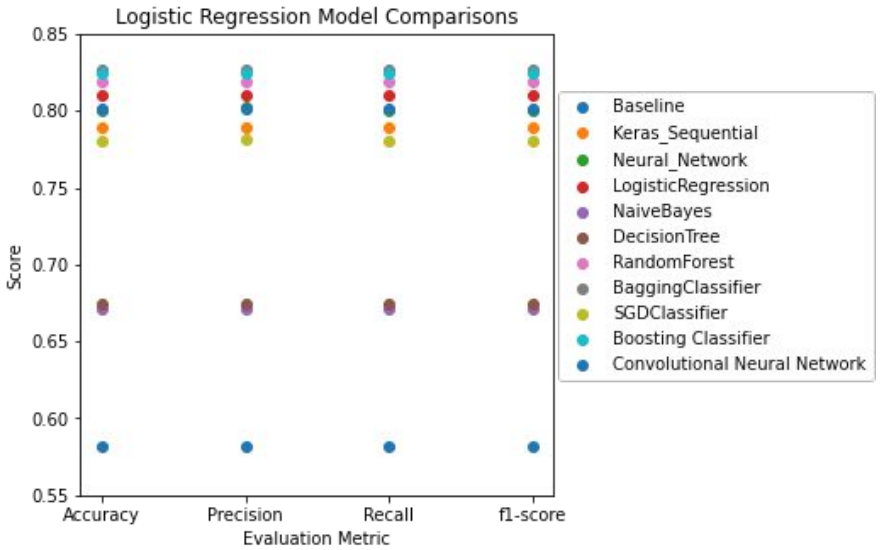
- Splitting test/train sets by season progression
- Adding Last 10 Game Data
- Scaling Features
- Tested out PCA

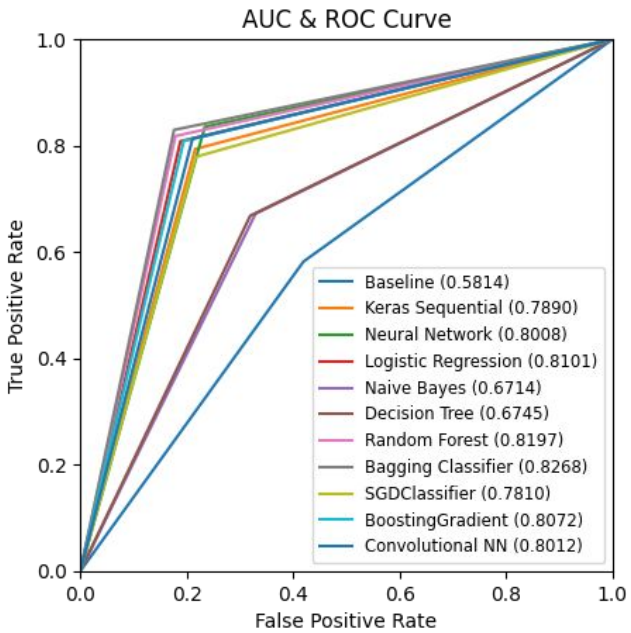




Model Comparison Results

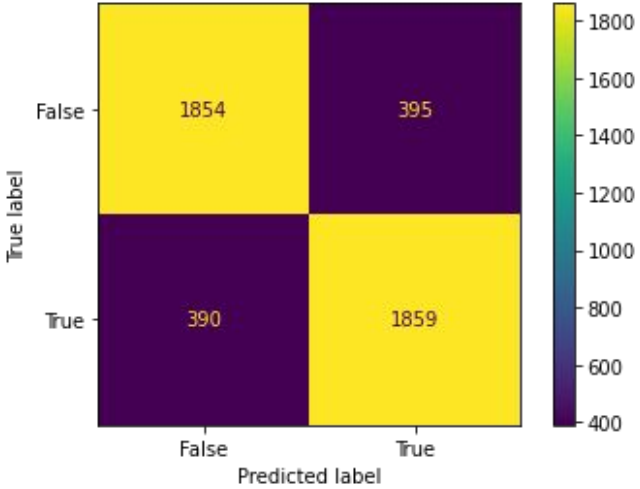
- Baseline: Whichever team won the previous matchup is predicted to win the current matchup
- Evaluation Parameters: Accuracy, precision, recall, f1-score, AUC

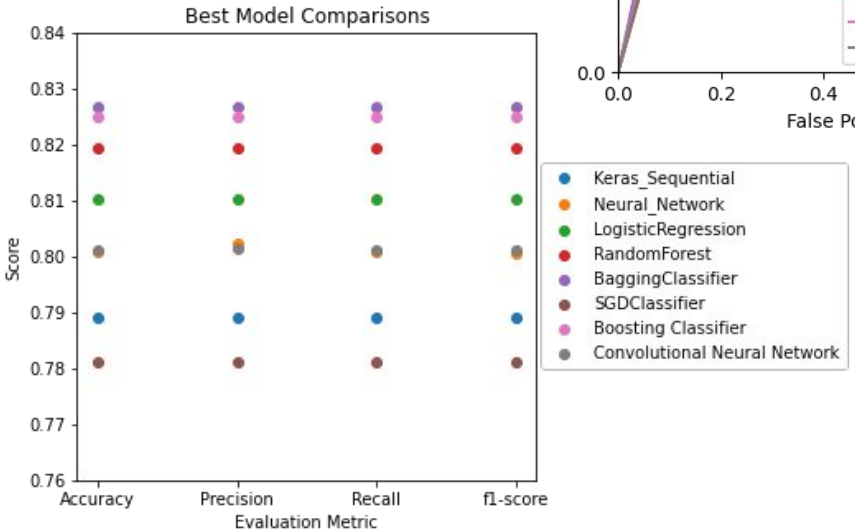


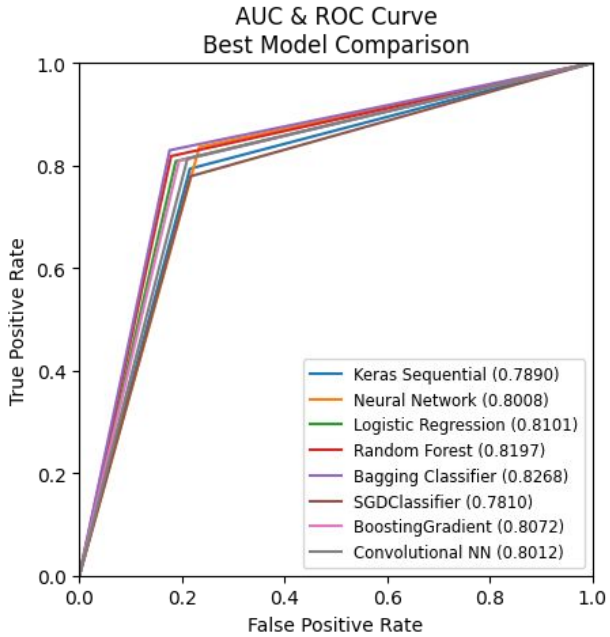


Best Model Results

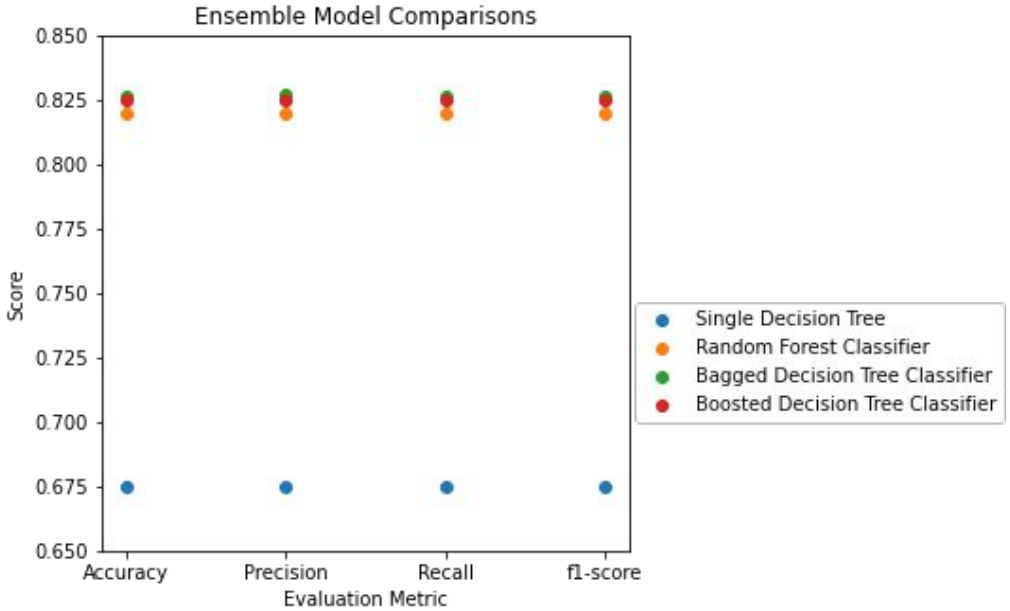
- Ensemble Models performed the best
- NN and CNN performed similarly

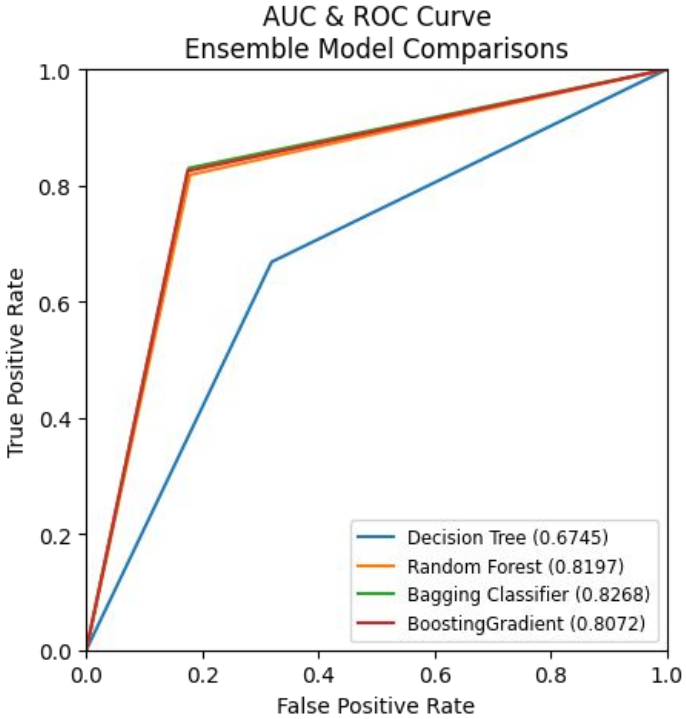






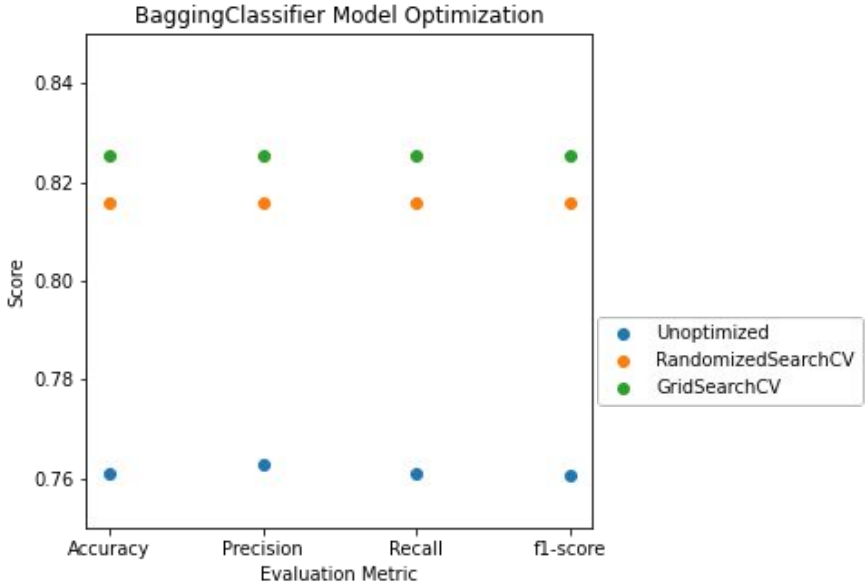
Ensemble Model Comparisons

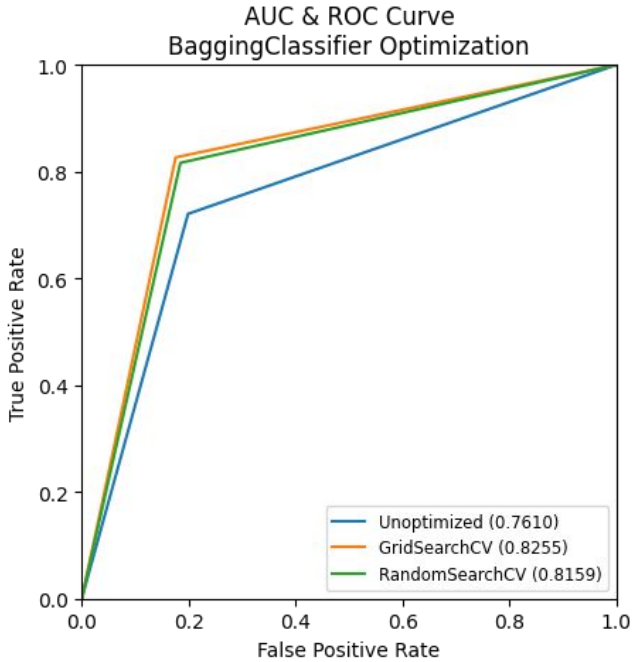




Optimization

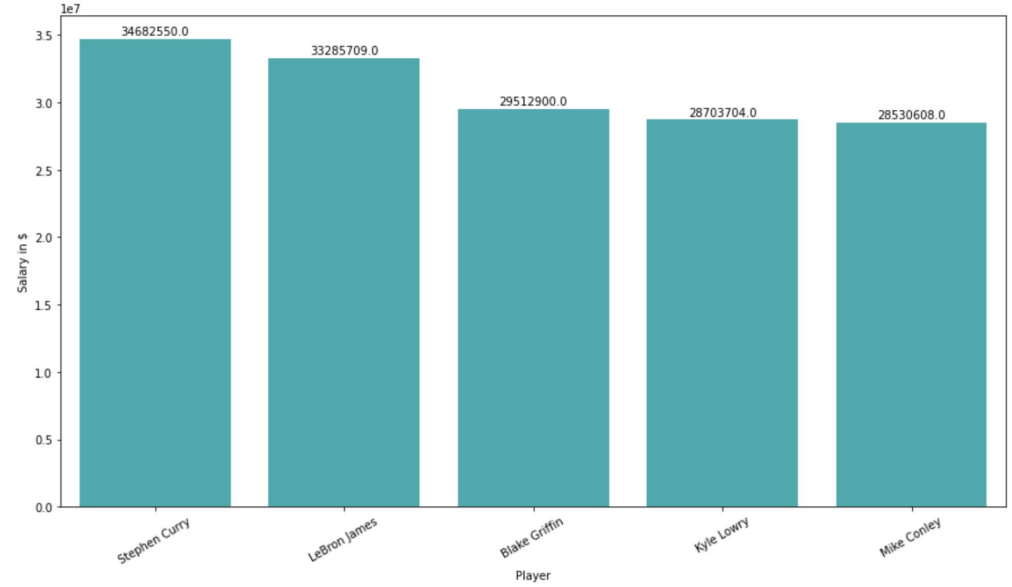
- Ensemble Models were optimized using RandomSearchCV to find appropriate ranges of hyperparameters, followed by GridSearchCV
- Bagging Classifier showed the most dramatic improvement



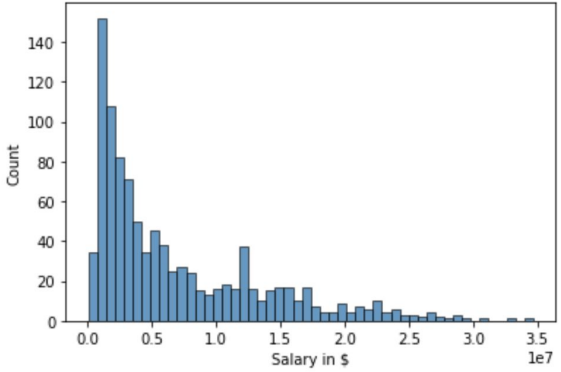


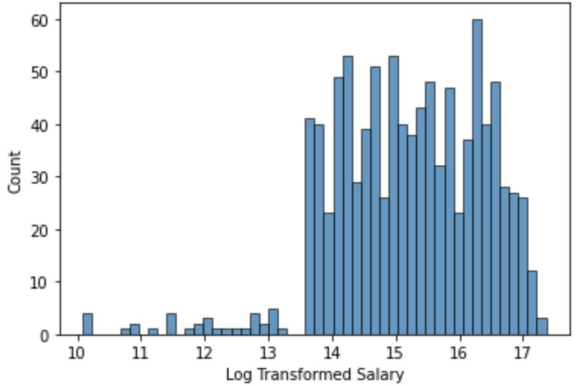
Linear Regression for Predicting Salaries: EDA

Highest Paid Players in 2017

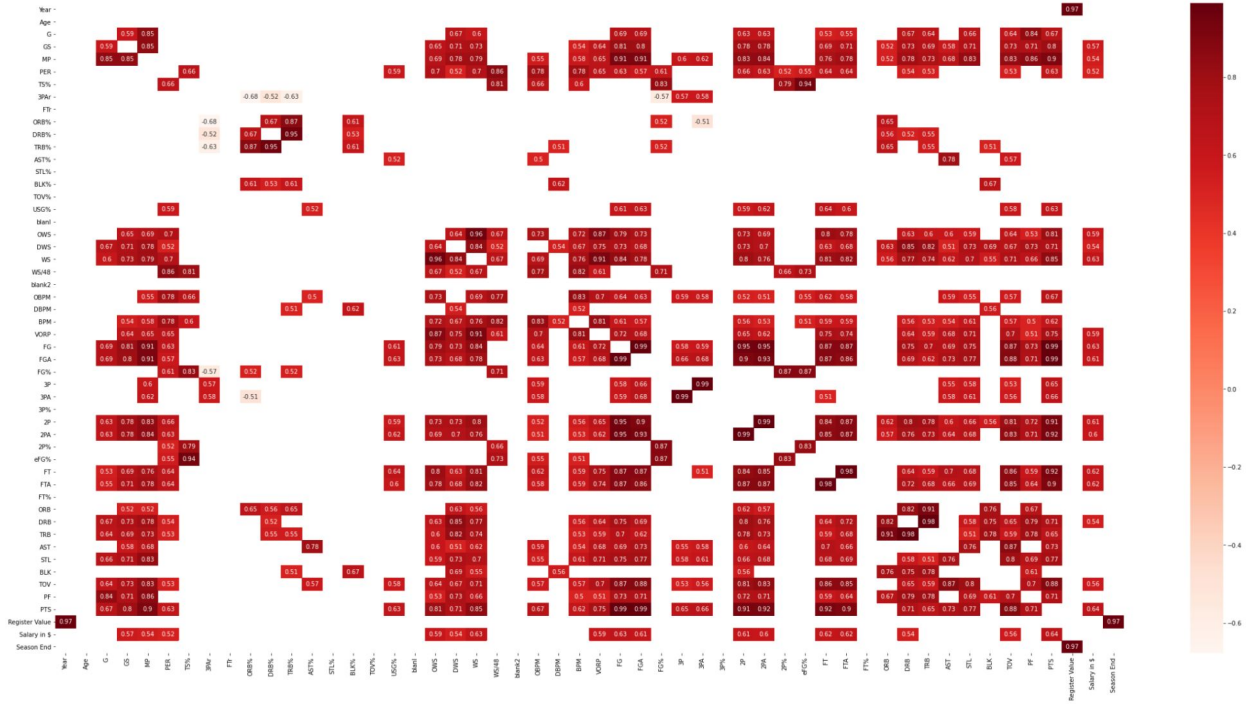


2014– 2017





Linear Regression for Predicting Salaries: Initial Models and Visualizations



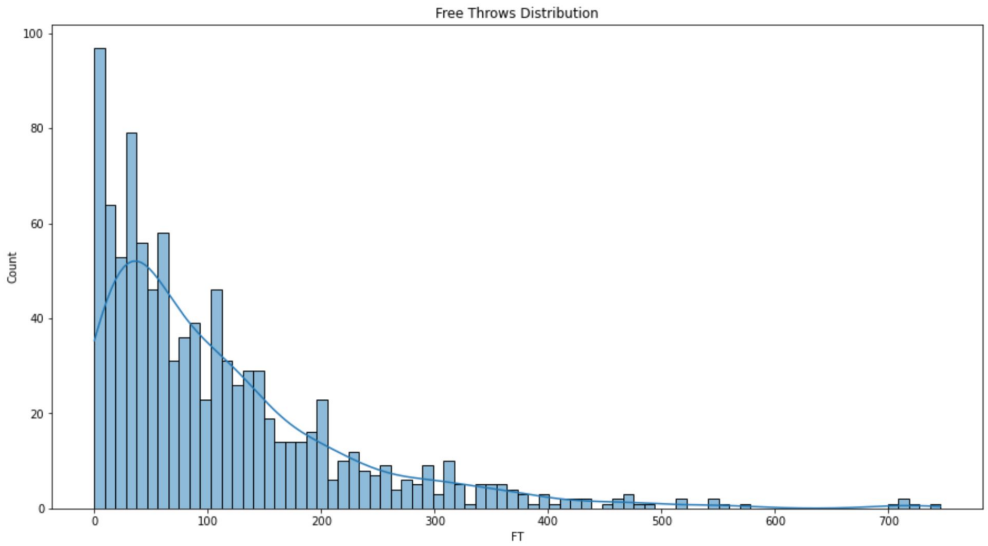
After standardizing and using **PCA to reduce dimensionality** to 42 columns (95% explained variance) :

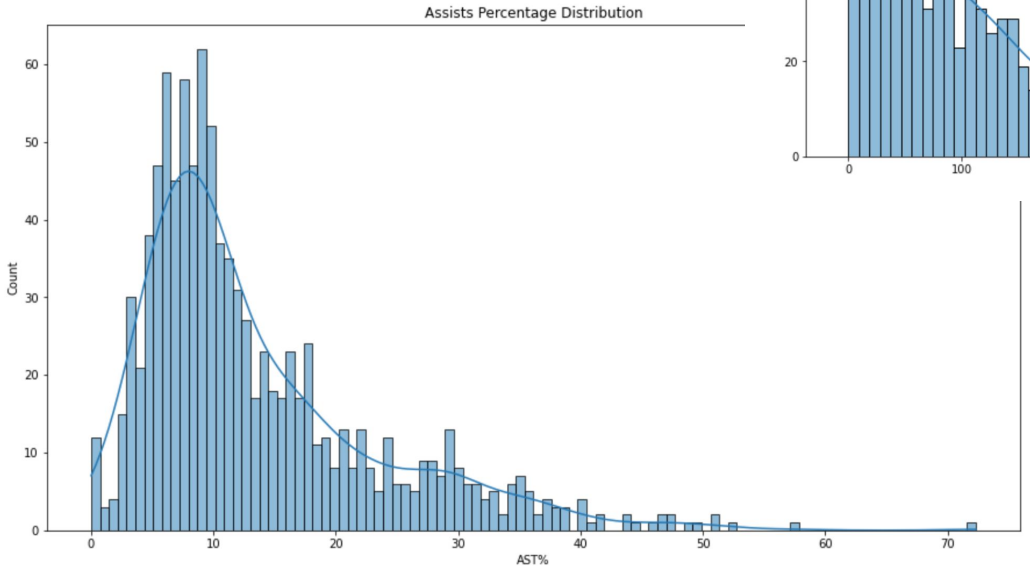
RMSE of Baseline Model (Average Salary) : 1.25
MSE of Baseline Model: 1.56
MAE: 0.99
MAPE:0.07

Experiments and Results Before Feature Engineering

	Linear Regression	SGD Regressor	Gradient Boosting Regressor	Random Forest Regressor
Spearman	0.75	0.745	0.714	0.732
RMSE	0.932	0.938	0.943	0.914
MSE	0.868	0.880	0.889	0.836
MAE	0.657	0.662	0.685	0.669
MAPE	0.046	0.046	0.0473	0.046

Feature Engineering



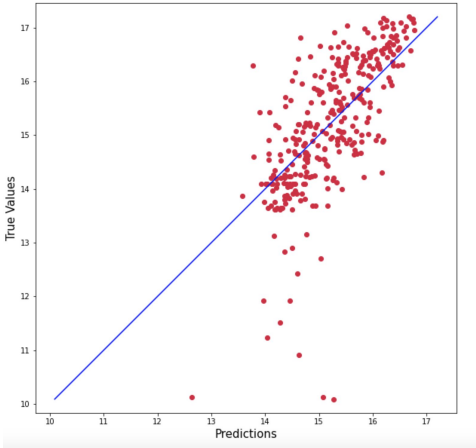


Applying Log Transformations on Skewed Features

	Linear Regression	SGD Regressor	Gradient Boosting Regressor	Random Forest Regressor
Spearman	0.751	0.755	0.752	0.759
RMSE	0.906	0.897	0.892	0.876
MSE	0.821	0.805	0.795	0.767
MAE	0.650	0.647	0.631	0.637
MAPE	0.045	0.045	0.043	0.044

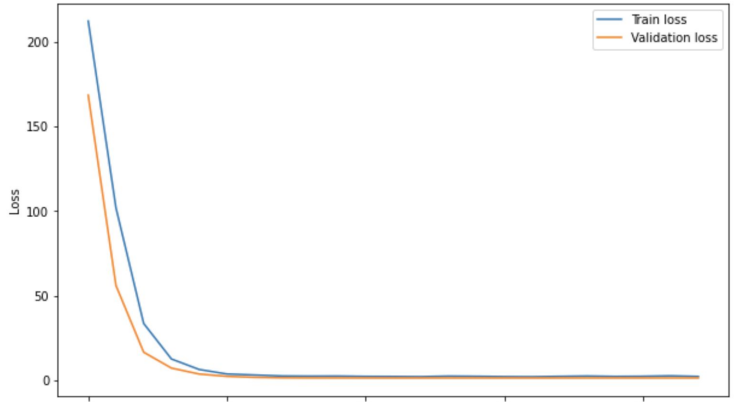
Optimization on Random Forest Regressor using GridSearch

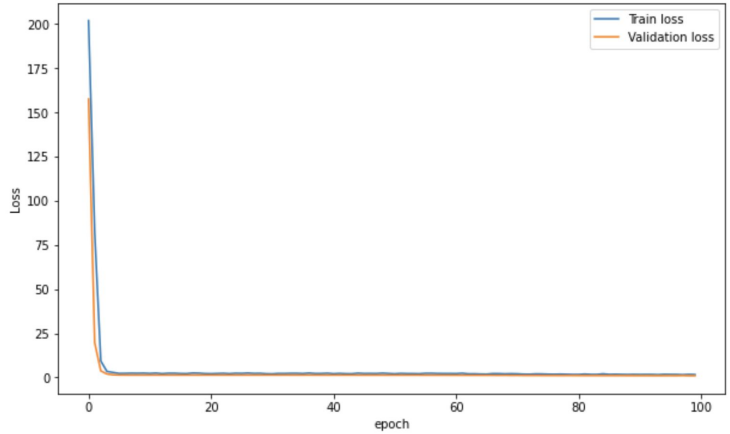
```
GridSearchCV
GridSearchCV(estimator=RandomForestRegressor(),
             param_grid={'criterion': ['squared_error'],
                         'max_features': ['sqrt', 'log2', None],
                         'min_samples_leaf': [1, 2, 3],
                         'min_samples_split': [2, 4, 6, 8],
                         'n_estimators': [100, 200, 300, 500, 600],
                         'oob_score': [True, False]},
             verbose=3)
estimator: RandomForestRegressor
RandomForestRegressor()
RandomForestRegressor
RandomForestRegressor()
```



	Random Forest Regressor
Spearman	0.764
RMSE	0.871
MSE	0.759
MAE	0.622
MAPE	0.043

Recurrent Neural Networks (Salary Prediction)



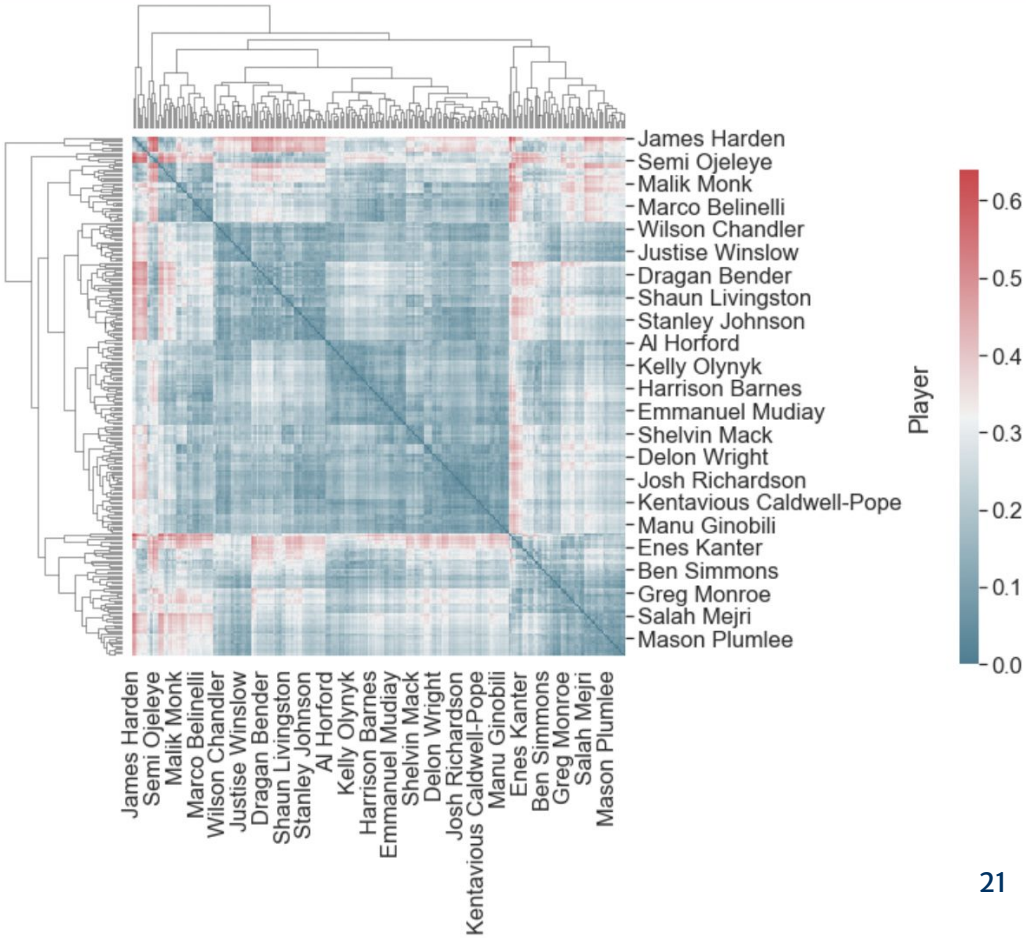


	LTSM	GRU
RMSE	1.227	1.225
MSE	1.506	1.500
MAE	0.978	0.977
MAPE	0.066	0.066

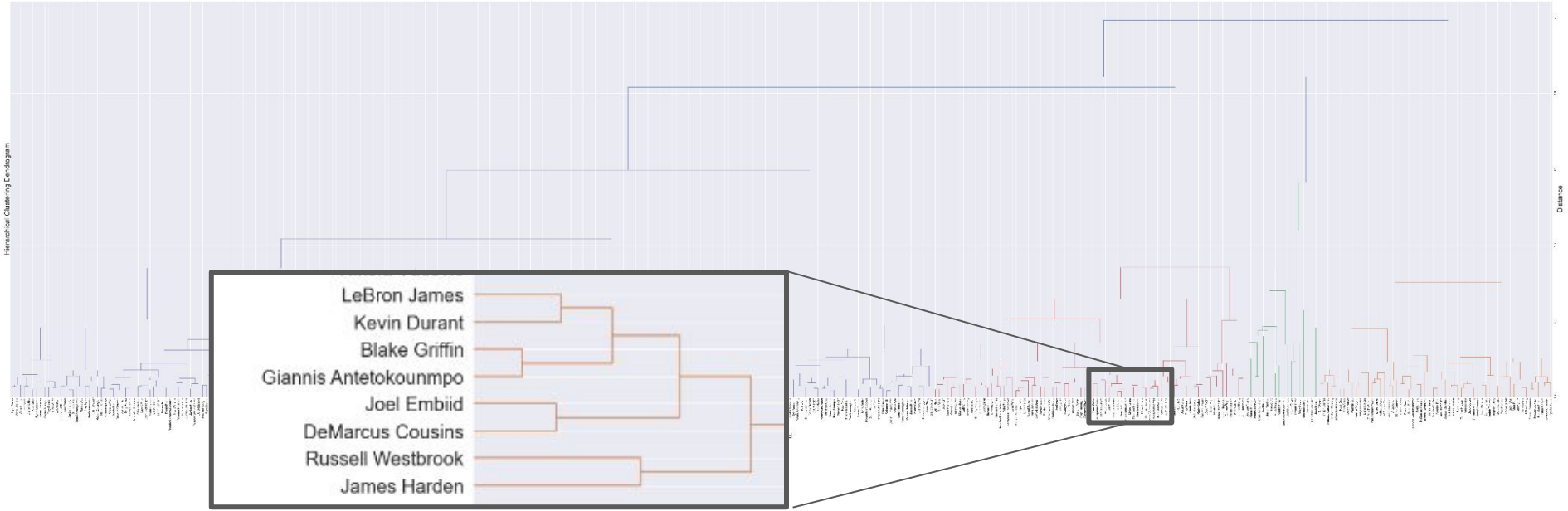
Clustering NBA Players: EDA

Euclidean Distance between Players

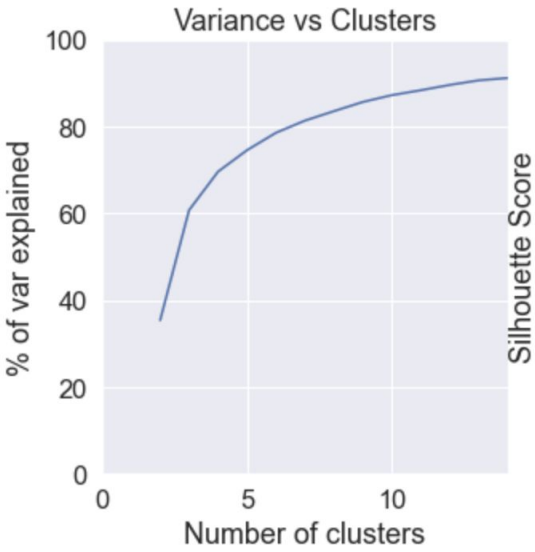
	PLAYER	CLOSEST_PLAYER
120	Anthony Davis	[Enes Kanter]
139	Kevin Durant	[LeBron James]
201	James Harden	[Damian Lillard]
253	LeBron James	[Kevin Durant]
269	Nikola Jokic	[Karl-Anthony Towns]
296	Kawhi Leonard	[Damian Lillard]
396	Chris Paul	[Jeremy Lin]
474	Isaiah Thomas	[Damian Lillard]
518	Russell Westbrook	[DeMarcus Cousins]

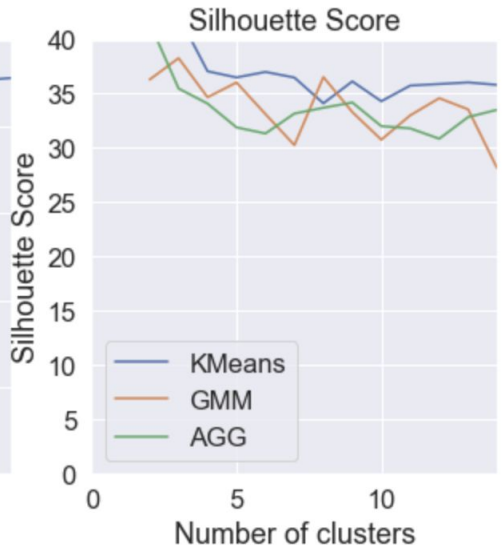


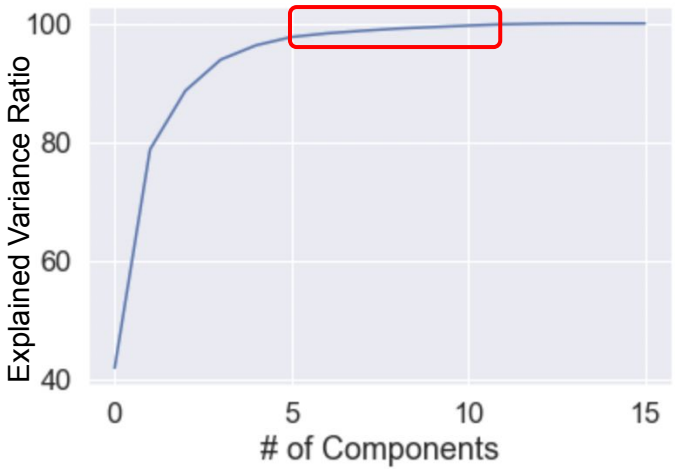
Hierarchical Dendrogram

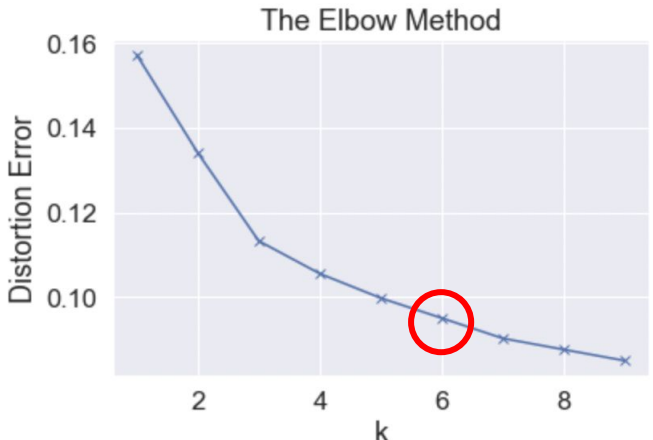


Clustering NBA Players: Initial Models and Visualizations

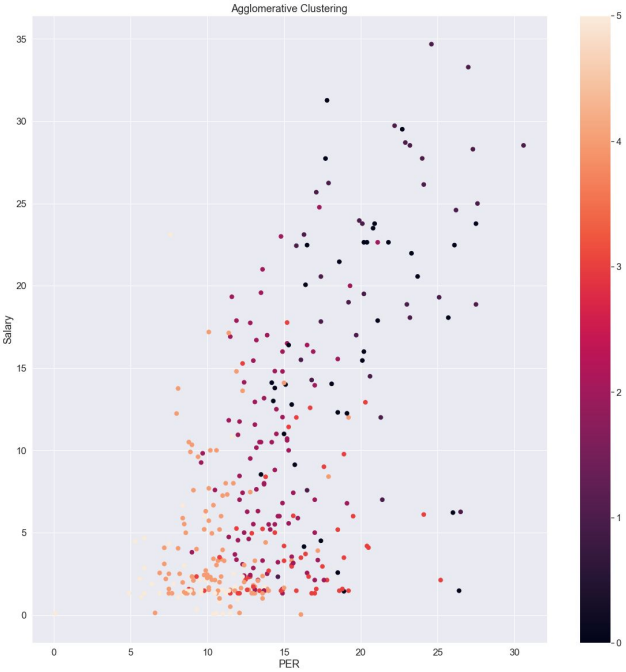


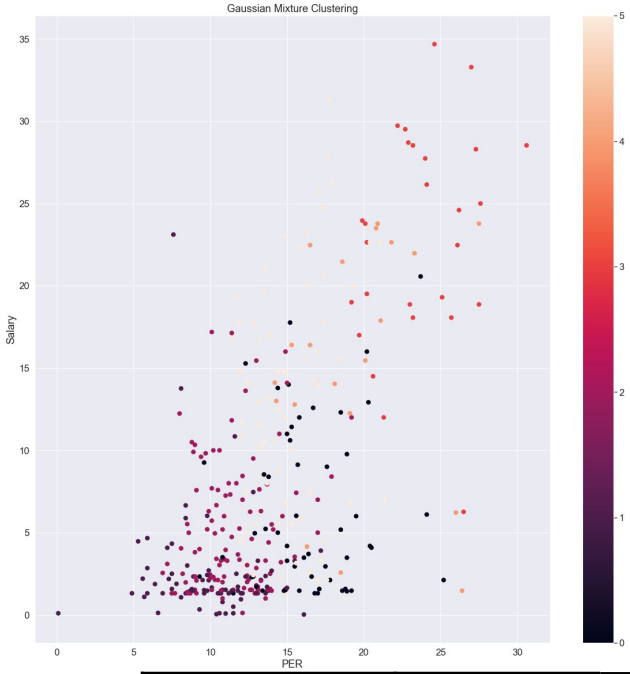


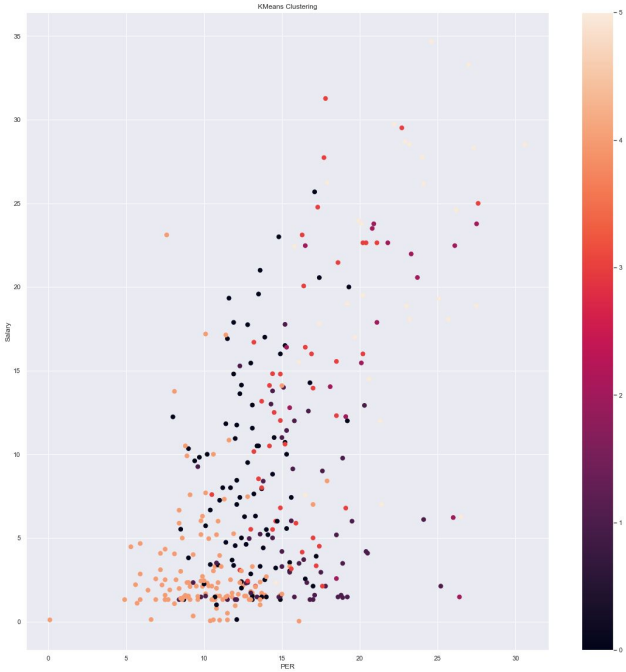




Cluster Model Comparison

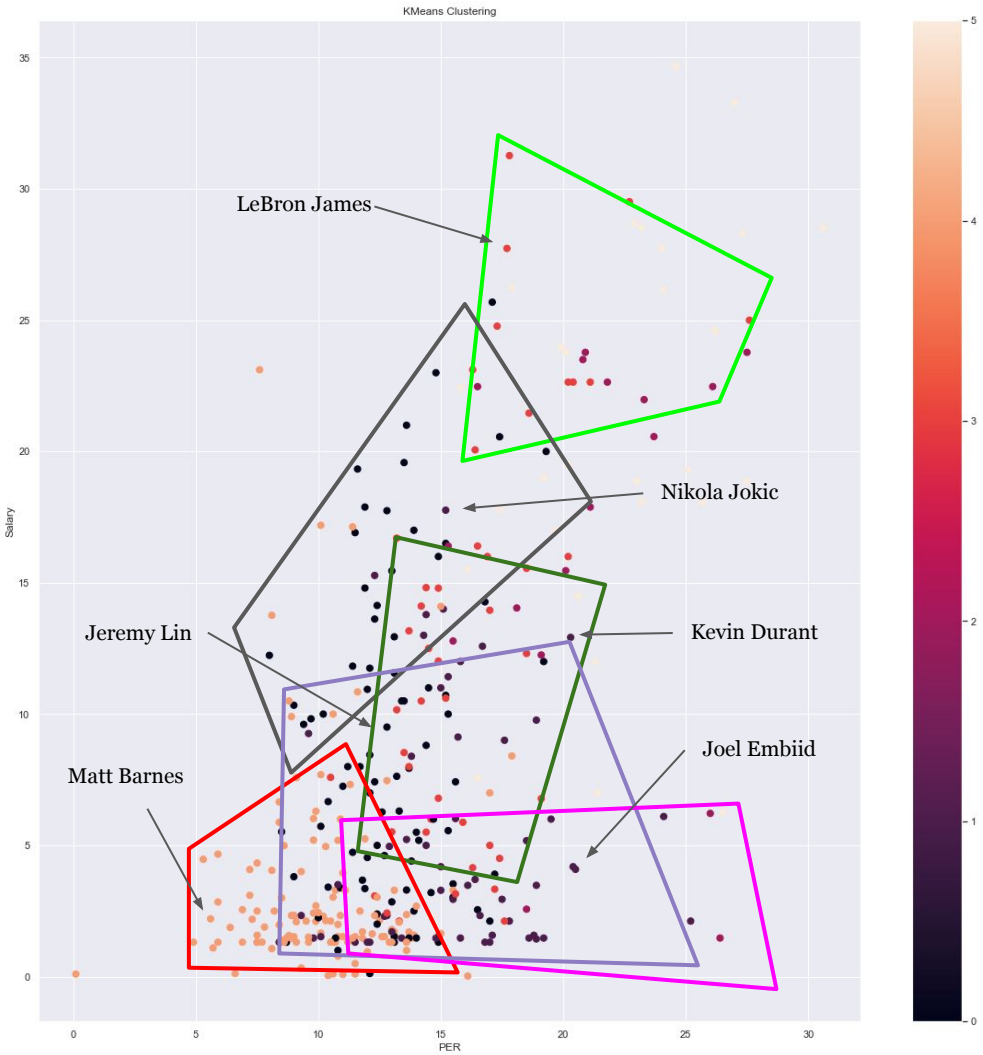






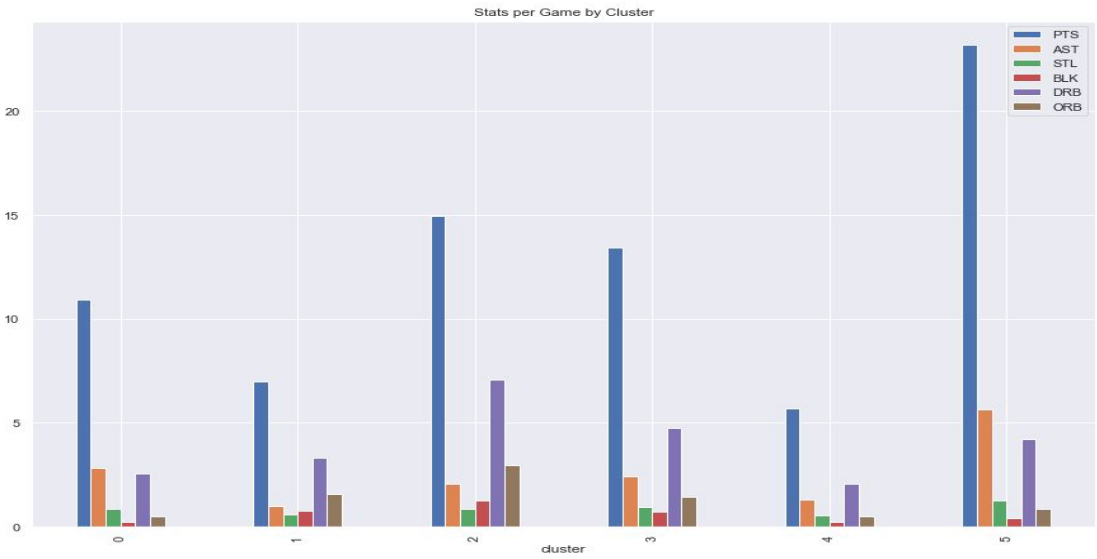
Model	Results		
	Dunn Index	Davies-Bouldin	Silhouette
KMeans	0.11	1.69	0.37
Gaussian Mixture	0.13	1.66	0.28
Agglomerative	0.11	1.77	0.26

Berkeley
UNIVERSITY OF CALIFORNIA

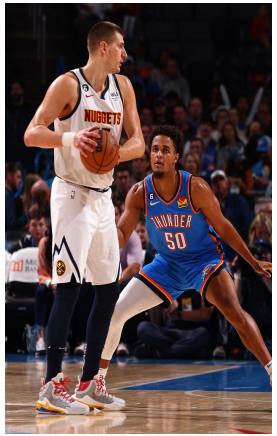

KMeans Clustering
LeBron James
Nikola Jokic
Jeremy Lin
Kevin Durant
Joel Embiid
Matt Barnes
Salary
PER
0
5
10
15
20
25
30
35
0
1
2
3
4
5

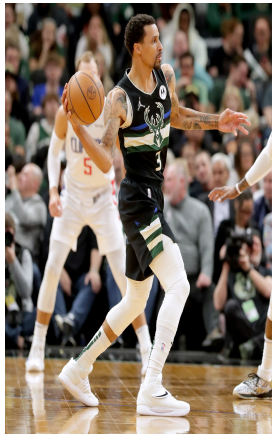
Player Archetypes

	cluster	G	MP	PER	PTS	AST	STL	BLK	Salary_17_18
5	5	75.20	34.31	0.29	23.16	5.65	1.26	0.39	20.61
2	2	78.65	29.99	0.27	14.93	2.06	0.88	1.28	16.49
3	3	72.78	29.02	0.22	13.44	2.41	0.95	0.72	12.66
0	0	69.89	26.48	0.19	10.94	2.82	0.85	0.25	8.53
1	1	57.64	18.43	0.27	7.00	0.98	0.57	0.76	5.24
4	4	54.33	17.12	0.19	5.71	1.31	0.53	0.23	3.34



Player Archetypes (cont.)





Cluster 5: Super Stars
Salary: $20,610,000
PER: 29
Points: 23.15/game
Min Played: 35/game

LeBron James
Kawhi Leonard
Stephen Curry

Cluster 2: Big Men
Salary: $16,490,000
PER: 27
Blocks: 1.28/game
Rebounds: 10.1/game

Nikola Jokic
Anthony Davis
DeAndre Jordan

Cluster 3: Shooters
Salary: $12,660,000
PER: 22
Shots: 10.92/game
3P %: 35%

Kevin Durant
Kevin Love
Blake Griffin

Cluster 0: Average Joes
Salary: $8,530,000
PER: 19
PTS: 10.94/game
Min Played: 26.48/game

George Hill
Jeremy Lin
JJ Barrea

Cluster 1: Bench Players
Salary: $5,240,000
Games: 57 (of 80)
Min Played: 18.2/game
PTS: 7game

JaVale McGee
Joel Embiid
Kyle O'Quinn

Cluster 4: Subpar
Salary: $3,240,000
PER: 10.52
Shooting: 48%
PTS: 5.74/game

Rudy Gay
Al-Farouq Aminu
Matt Barnes

Biases

- Salary Bias
  - Model trained on more examples of lower salaries than higher salaries because of imbalance
  - Log transformation

- Time Selection Sampling Bias
  - Time-frame chosen for modeling could affect model's outputs
  - Clustering one season of players

Limitations and Future Work

- Gather data on player's entire gaming stats for their whole career to cluster against other players
  - More accurate representation of that player's game history (beginning, middle and end of career)
- Salary prediction model could be trained on more data across a bigger time period and account for the difference in time period using time series analysis
  - Could be extended to other leagues outside the NBA
- Clustering data sets were extremely overlapped causing lower than anticipated Dunn / Silhouette Scores
- Game Outcome prediction may be specific to current era of NBA play and not generalize to previous or future eras.

Standards

- Scikit-Learn 1.1.2
- TensorFlow 2.10.0
- Python 3.9.7

Time/Budget:

- 10 hours/week
- No budget

Questions