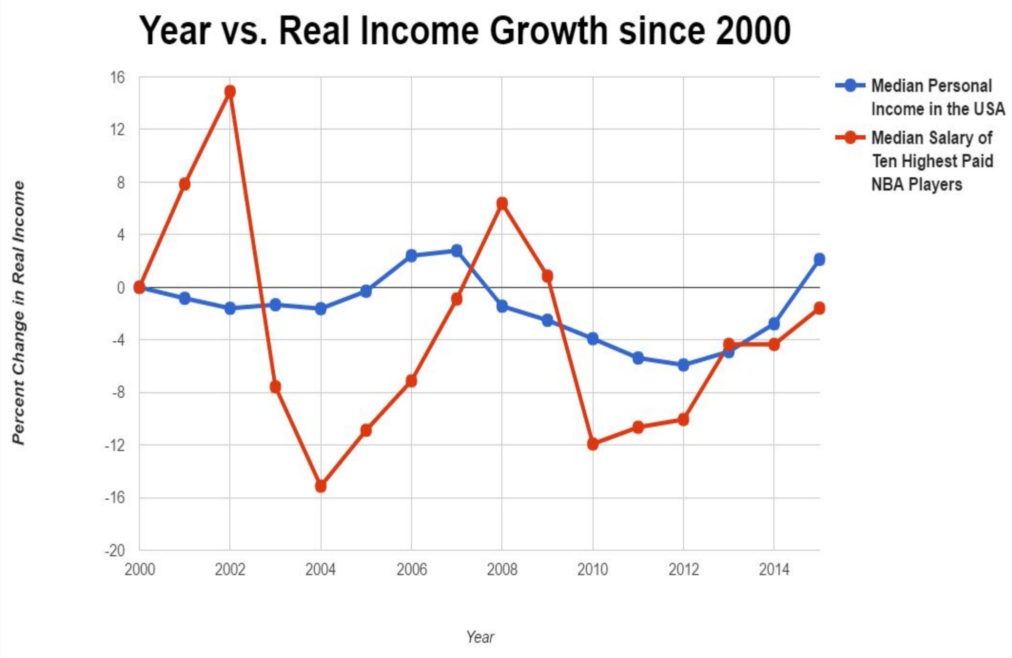
Predicting Basketball Salaries

CSP 571 Final Project Data Preparation and Analysis Final Project



A Breakdown of NBA Salaries
by Adam Wesolowski-Mantilla. [1]

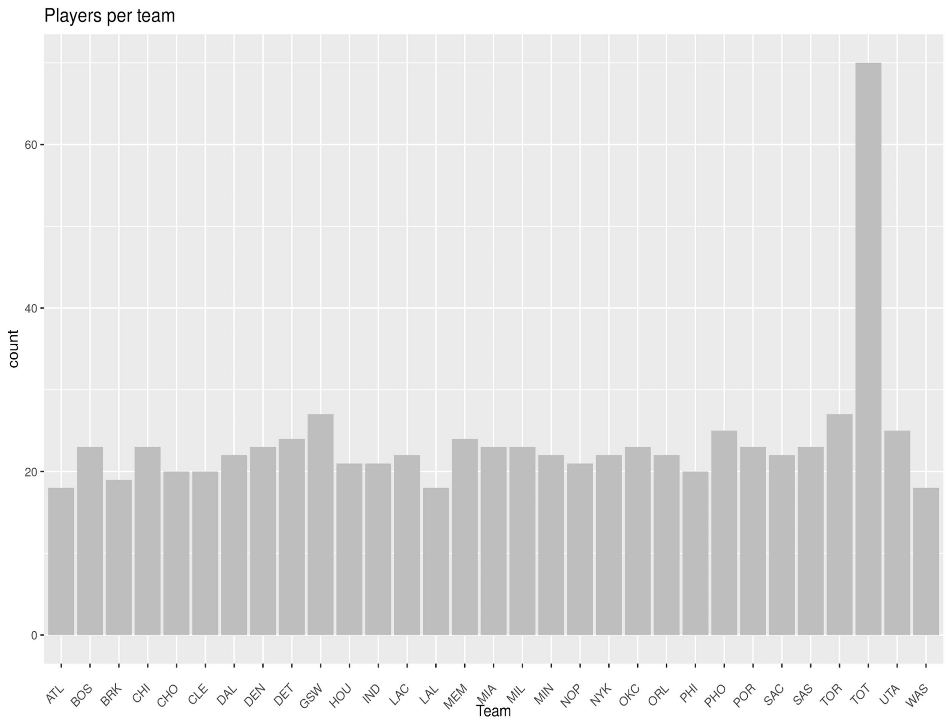
Project Objective

- Aims to explore a wide variety of NBA statistics
- Discover which statistics are the best predictors of an NBA players’ salary
- Predict professional basketball player salaries based on statistics
- Determine which players have been overvalued and undervalued based on the model

Primary Data Set

NBA Player Statistics

- Kaggle[2], CSV file
- 1950 - 2017 Seasons
- 3 Factor Predictors
  Year, Position, Team
- 46 Numeric Predictors
  Age, Games, Games Started...



Secondary Data Set

NBA 2K Player Ratings

- MTDB[3], Scraped
- 2016 - 2020 Seasons
- 2 Factor Predictors
    - Year, Position,
- 7 Numeric Predictors
    - Overall, Outside, Inside, Playmaking, Athleticism, Defense, Rebounding

Combined Data Set

Choose 2016 & 2017 Seasons

Combine 3 factor & 53 numeric predictors

Total 734 Players

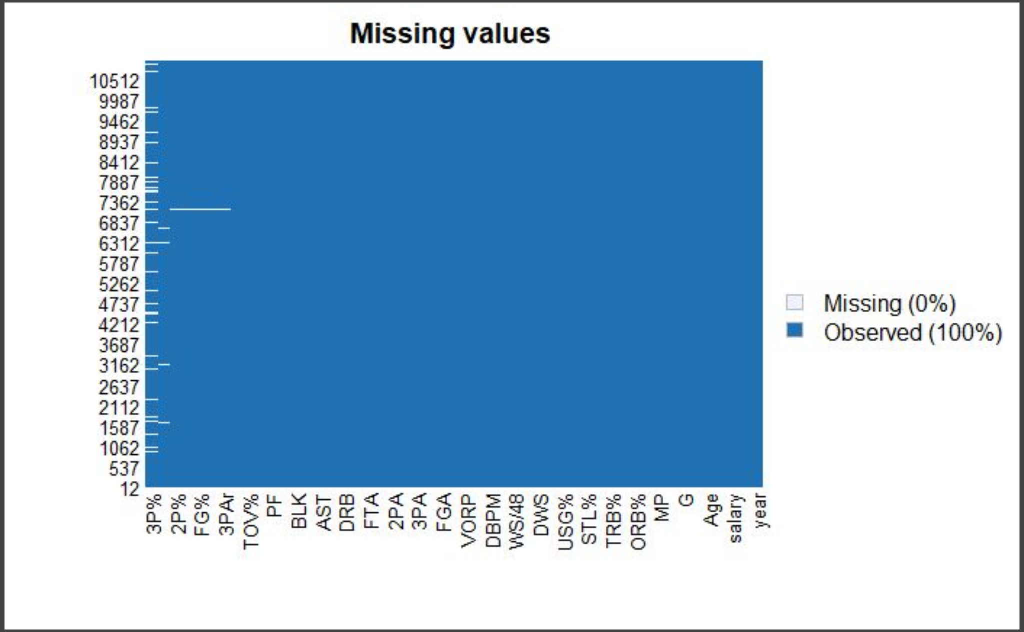
Data Preparation and analysis

- **<u>Data cleaning</u>**
- Missing values
- Exploratory data analysis
- Correlations
- Outliers
- Variable selection
- Train/test split

- Traded player ⟶ keep aggregated stats
  - ‘TOT’ team
- Multiple versions of a player ⟶ keep highest overall rating
- Used levenshtein distance
- `luke babbit` -> 'luke babbitt',
- `patrick beverly` -> 'patrick beverley'

Data Preparation and analysis

- Data cleaning
- **Missing values**
- Exploratory data analysis
- Correlations
- Outliers
- Variable selection
- Train/test split



3P% has the most missing values.

Data Preparation and analysis

- Data cleaning
- **Missing values**
- Exploratory data analysis
- Correlations
- Outliers
- Variable selection
- Train/test split

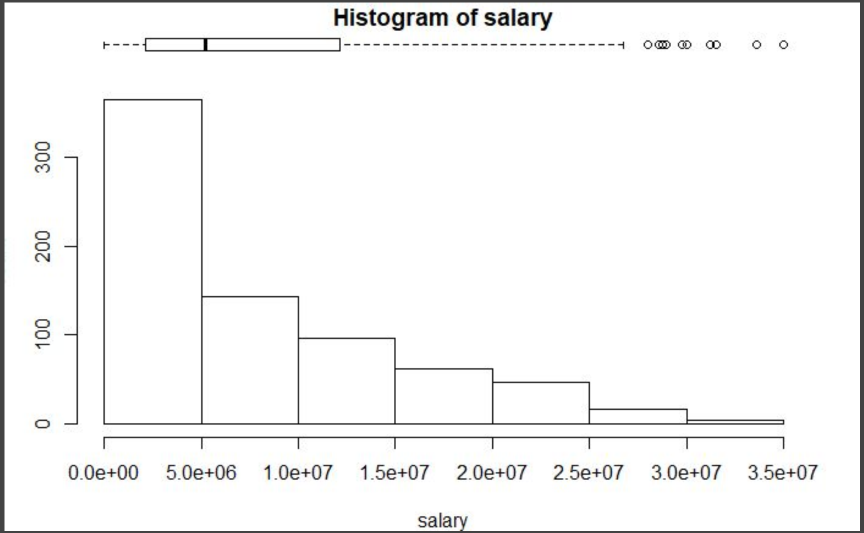
year	name_p	salary	Pos	Age	Tm	G	GS	MP	PER	TS%	3PAr	FTr
0	0	0	0	0	0	0	0	0	0	24	30	30
ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM
0	0	0	0	0	0	19	0	0	0	0	0	0
DBPM	BPM	VORP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%
0	0	0	0	0	30	0	0	1562	0	0	40	30
FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	
0	0	233	0	0	0	0	0	0	0	0	0	

- Total of 1998 values were missing.
- Missing values were replaced by the mean of their columns.

```
df_primary$`3P%`[is.na(df_primary$`3P%`)] <- mean(df_primary$`3P%` , na.rm = T)
df_primary$`FT%`[is.na(df_primary$`FT%`)] <- mean(df_primary$`FT%` , na.rm = T)
df_primary$`TOV%`[is.na(df_primary$`TOV%`)] <- mean(df_primary$`TOV%` , na.rm = T)
df_primary$`FG%`[is.na(df_primary$`FG%`)] <- mean(df_primary$`FG%` , na.rm = T)
df_primary$`2P%`[is.na(df_primary$`2P%`)] <- mean(df_primary$`2P%` , na.rm = T)
df_primary$`eFG%`[is.na(df_primary$`eFG%`)] <- mean(df_primary$`eFG%` , na.rm = T)
df_primary$`TS%`[is.na(df_primary$`TS%`)] <- mean(df_primary$`TS%` , na.rm = T)
df_primary$`3PAr`[is.na(df_primary$`3PAr`)] <- mean(df_primary$`3PAr` , na.rm = T)
df_primary$`FTr`[is.na(df_primary$`FTr`)] <- mean(df_primary$`FTr` , na.rm = T)
```

Data Preparation and analysis

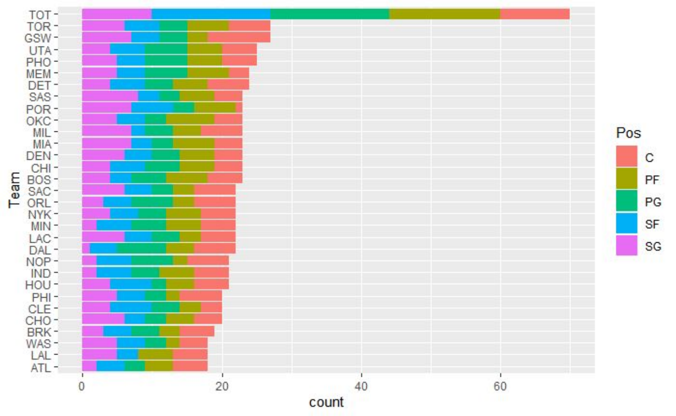
- Data cleaning
- Missing values
- **Exploratory data analysis**
- Correlations
- Outliers
- Variable selection
- Train/test split

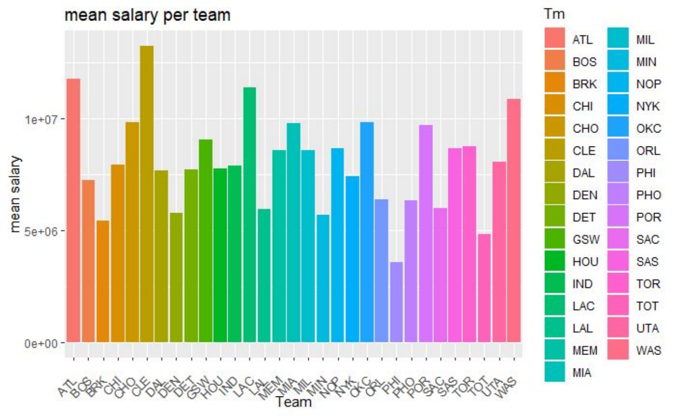


- Mean = $7,837,816
- Median = $5,200,000
- Mode = $1,312,611
- Outliers in salary variable.

Data Preparation and analysis

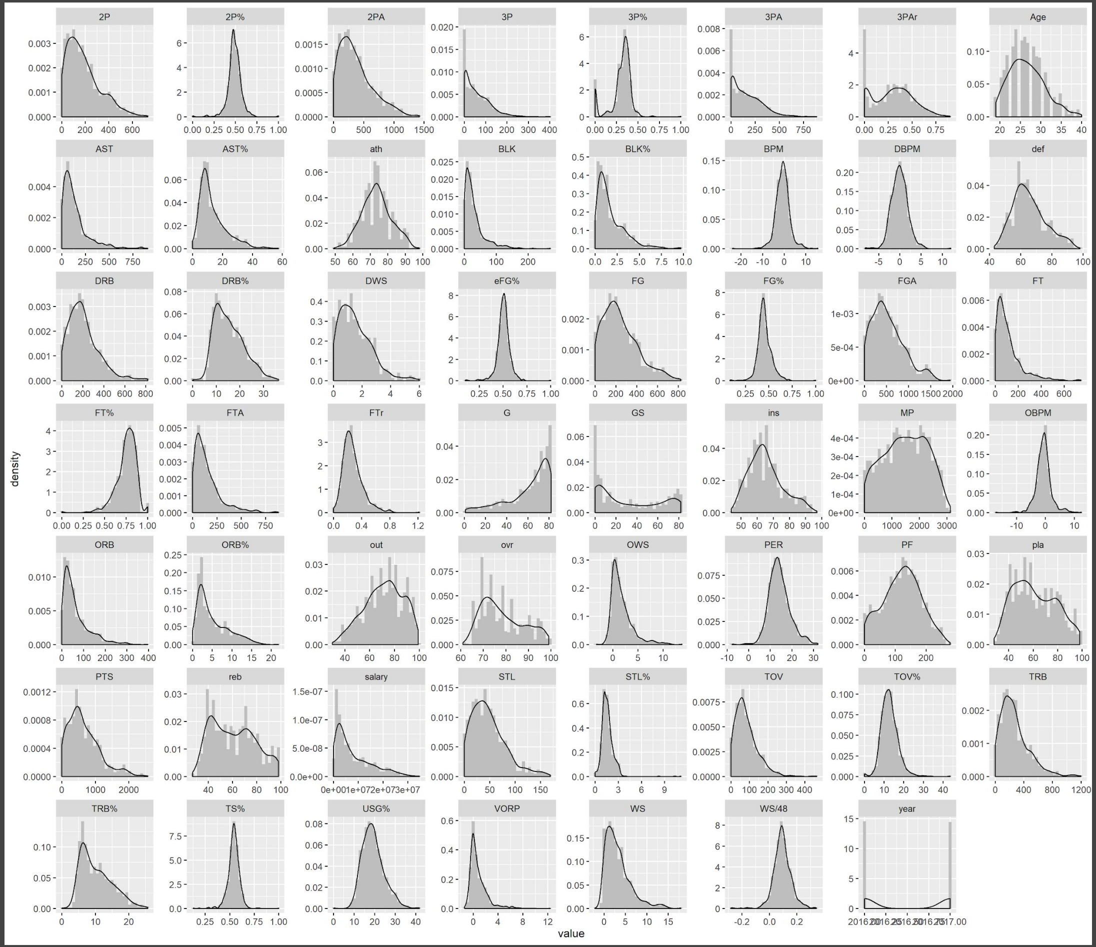
- Data cleaning
- Missing values
- **Exploratory data analysis**
- Correlations
- Outliers
- Variable selection
- Train/test split





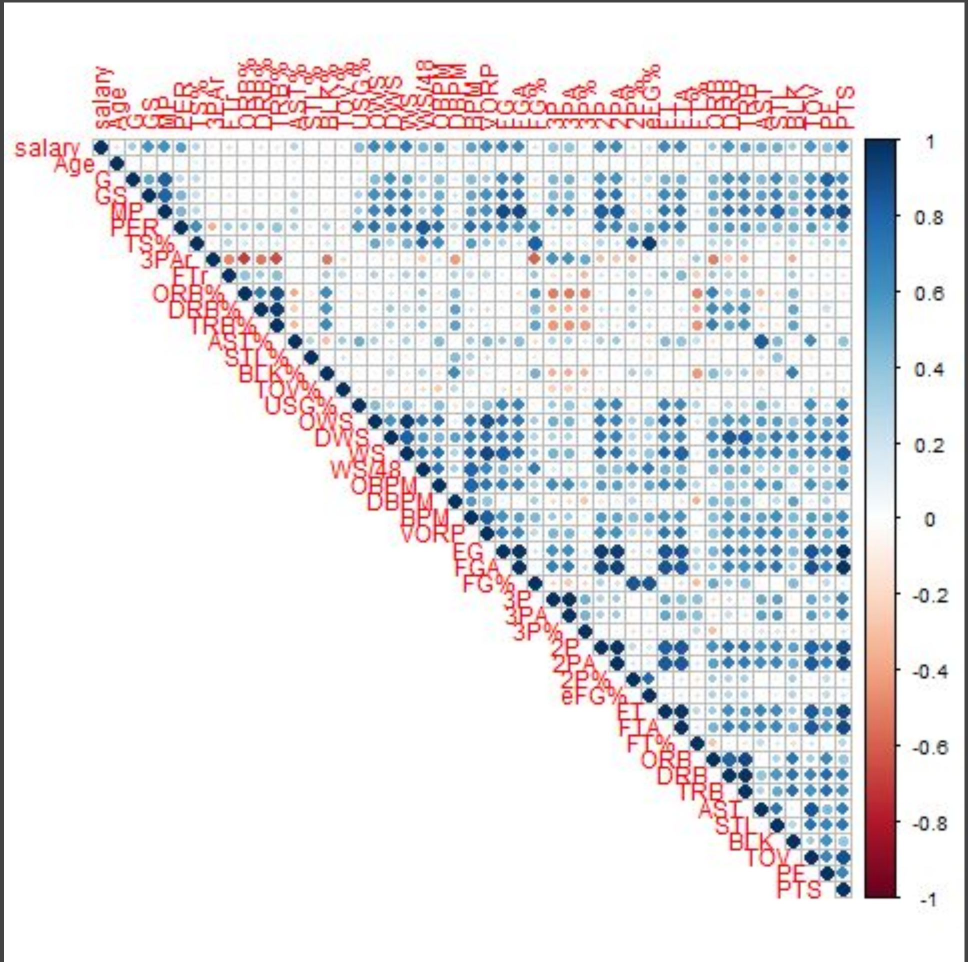
Data Preparation and analysis

- Data cleaning
- Missing values
- **Exploratory data analysis**
- Correlations
- Outliers
- Variable selection
- Train/test split



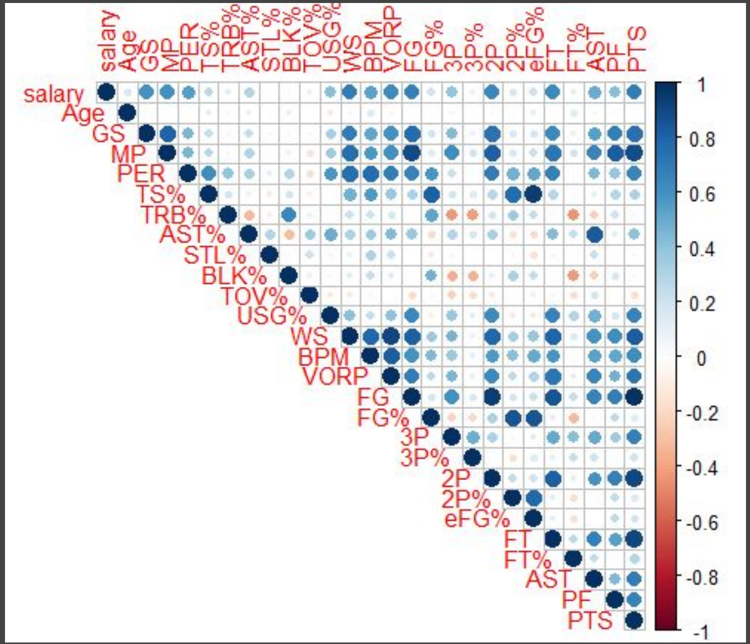
Data Preparation and analysis

- Data cleaning
- Missing values
- Exploratory data analysis
- **<u>Correlations</u>**
- Outliers
- Variable selection
- Train/test split



Data Preparation and analysis

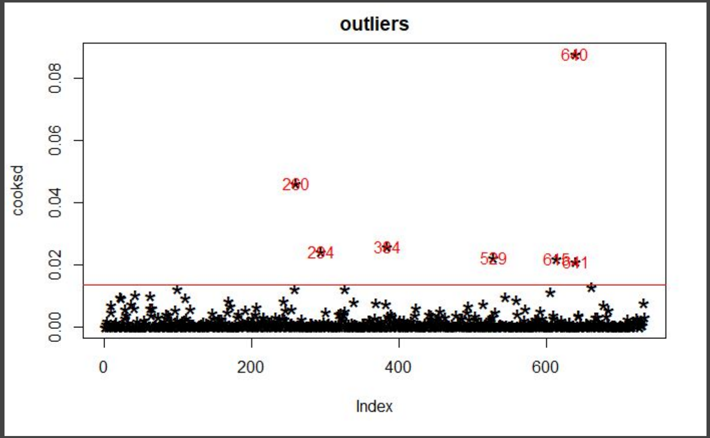
- Data cleaning
- Missing values
- Exploratory data analysis
- **<u>Correlations</u>**
- Outliers
- Variable selection
- Train/test split



- Removed variables that were highly correlated with each other and variables that were explaining the same thing like 3P and 3PA
- Reduced number of variable from 51 to 38

Data Preparation and analysis

- Data cleaning
- Missing values
- Exploratory data analysis
- Correlations
- **Outliers**
- Variable selection
- Train/test split



- Multivariate approach using cook's distance.
- It computes the influence exerted by each data point on the predicted outcome.

Data Preparation and analysis

- Data cleaning
- Missing values
- Exploratory data analysis
- Correlations
- **Outliers**
- Variable selection
- Train/test split

	name	year	salary	Pos	Age	Tm	GS	MP	AS...	WS	FG%	3P%	2P%	FT%	PF	PTS
	<chr>	<fctr>	<dbl>	<fctr>	<dbl>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
260	isaiah thomas	2017	6261395	PG	27	BOS	76	2569	32.6	12.6	0.463	0.3790000	0.528	0.909	167	2199
294	jarrett jack	2017	2328652	PG	33	NOP	0	33	20.3	0.0	0.667	0.0000000	1.000	1.000	4	6
384	karl anthony towns	2017	6216840	C	21	MIN	82	3030	13.2	12.7	0.542	0.3670000	0.582	0.832	241	2061
530	nikola mirotic	2016	5782450	PF	24	CHI	38	1646	9.4	3.9	0.407	0.3900000	0.430	0.807	151	777
616	sam dekker	2017	1794600	SF	22	HOU	2	1419	7.7	3.1	0.473	0.3210000	0.591	0.559	83	504
641	stephen curry	2017	34682550	PG	28	GSW	79	2638	31.1	12.6	0.468	0.4110000	0.537	0.898	183	1999
642	stephen zimmerman	2017	1312611	C	20	ORL	0	108	5.3	0.0	0.323	0.2711419	0.323	0.600	17	23

- On closer inspection, outliers are players who have really good stats or bad stats for that year.

Data Preparation and analysis

- Data cleaning
- Missing values
- Exploratory data analysis
- Correlations
- Outliers
- **Variable selection**
- Train/test split

Method

- Automated F-test backward selection

Selected primary dataset predictors

- year, Pos, Age, Tm, GS, TS%, AST%, WS, VORP, FG%, 3P, FT, PF

Selected complete dataset predictors

- Age, GS, MP, USG%, WS, PF, out, ovr

Data Preparation and analysis

- Data cleaning
- Missing values
- Exploratory data analysis
- Correlations
- Outliers
- **Variable selection**
- Train/test split

WS	`FG%`	VORP	`TS%`	TmTOT	PosPG	FT	`AST%`	`3P`	PosSG
14.914773	9.299794	8.728394	7.144946	5.086995	4.190961	4.167082	3.650485	3.456991	3.092757
TmTOR	GS	TmPHO	TmGSW	TmUTA	TmMEM	TmCHI	TmSAS	TmLAC	TmOKC
3.037526	2.863495	2.858784	2.812127	2.793899	2.720141	2.692851	2.678805	2.637327	2.577475
TmMIN	TmBOS	TmDET	TmNYK	TmPOR	PF	PosSF	TmORL	TmCHO	TmHOU
2.573613	2.557200	2.547414	2.535575	2.526887	2.499607	2.498992	2.488734	2.472843	2.410641
TmCLE	TmIND	TmPHI	TmDEN	TmNOP	TmMIL	TmDAL	TmSAC	TmMIA	TmLAL
2.398189	2.392428	2.383919	2.327086	2.320728	2.256420	2.255879	2.252197	2.238935	2.203021
TmBRK	TmWAS	PosPF	Age	year2017					
2.172056	2.134109	1.908324	1.280955	1.049395					

MP	PF	GS	WS	ovr	out	`USG%`	Age
7.984500	3.929993	3.197309	3.078852	2.832498	2.563631	1.613312	1.136827

Method

- Variance inflation factor(VIF) was used to check for multicollinearity in selected variables.
- All selected variables have low VIF values so there is little to no multicollinearity.

Data Preparation and analysis

- Data cleaning
- Missing values
- Exploratory data analysis
- Correlations
- Outliers
- Variable selection
- **Train/test split**

Stratified Sampling

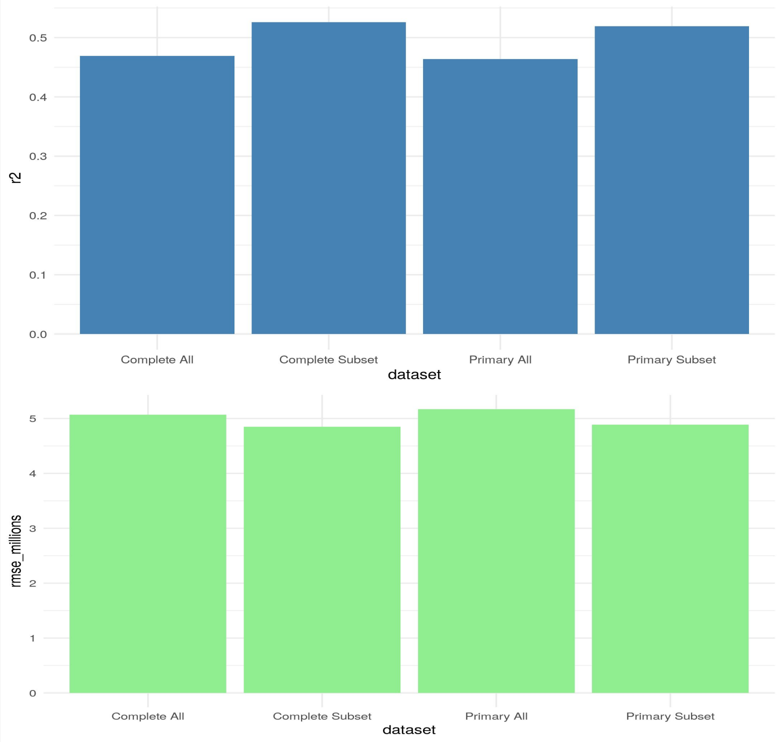
- Caret package

Train / Test Split

- 80% / 20%
- 590 records / 144 records

Linear Regression Models

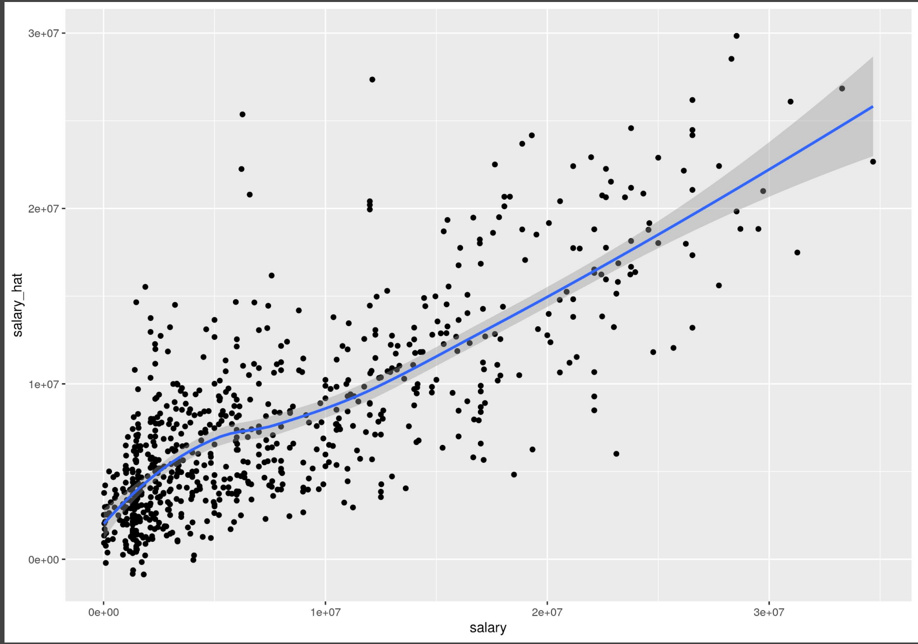
R-squared Test		
	Primary	Combined
All Var	0.478	**0.544**
Subset Var	0.440	0.472
RMSE Test		
All Var	5.13M	**4.82M**
Subset	5.32M	5.16M



SLR : Complete Dataset, All Variables

- Significant Coefficients by 90% confidence
  - Primary Dataset
    - Intercept, Year, Position: 'PG', Position 'SG', Age, Games, Turnover %, Value Over Replacement Player, Steals, Personal Fouls,
  - NBA 2K Dataset
    - Outside, Overall, Rebounding

- Training Metrics
  - R-squared = 0.66
  - Adj R-squared = 0.60
  - F-statistic = 11.69
    - F* ~ F(84, 503)
    - P-value < 2.2e-16



Elastic Net Models

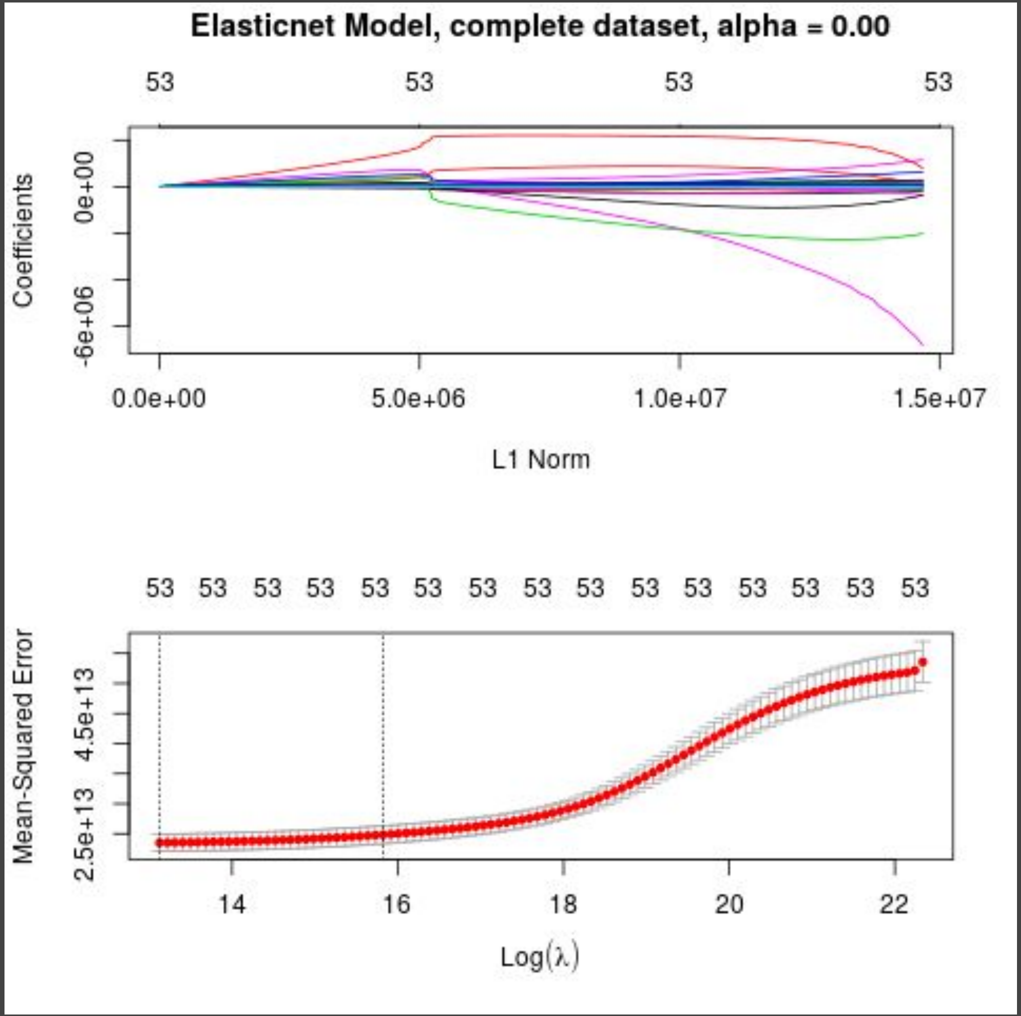
- Generalizes Ridge & Lasso Regression
- Penalty Function
  - $// y\text{-}XB //^{2} + \lambda_1 // B //^{2} + \lambda_2 // B //$
- glmnet [5] Gaussian Objective Function
  - $// y\text{-}XB //^{2}/(2n_{obs})+\lambda[(1 - \alpha) // \beta //^{2} /2 + \alpha // \beta //$
  - α = 1⟶Lasso
  - α = 0⟶Ridge
- Models Compared
  - α = 0, 0.05, 0.10, 0.15, …, 1.00
  - Cross Validation (to optimize λ)
  - Primary & Combined datasets

Model: fit_alpha_0.00	Dataset: primary	R^2 Test: 0.471	MSE: 2.668e+13
Model: fit_alpha_0.05	Dataset: primary	R^2 Test: 0.470	MSE: 2.673e+13
Model: fit_alpha_0.10	Dataset: primary	R^2 Test: 0.471	MSE: 2.672e+13
Model: fit_alpha_0.15	Dataset: primary	R^2 Test: 0.470	MSE: 2.672e+13
Model: fit_alpha_0.20	Dataset: primary	R^2 Test: 0.470	MSE: 2.673e+13
Model: fit_alpha_0.25	Dataset: primary	R^2 Test: 0.470	MSE: 2.673e+13
Model: fit_alpha_0.30	Dataset: primary	R^2 Test: 0.470	MSE: 2.673e+13
Model: fit_alpha_0.35	Dataset: primary	R^2 Test: 0.470	MSE: 2.672e+13
Model: fit_alpha_0.40	Dataset: primary	R^2 Test: 0.470	MSE: 2.672e+13
Model: fit_alpha_0.45	Dataset: primary	R^2 Test: 0.469	MSE: 2.681e+13
Model: fit_alpha_0.50	Dataset: primary	R^2 Test: 0.468	MSE: 2.683e+13
Model: fit_alpha_0.55	Dataset: primary	R^2 Test: 0.468	MSE: 2.685e+13
Model: fit_alpha_0.60	Dataset: primary	R^2 Test: 0.468	MSE: 2.687e+13
Model: fit_alpha_0.65	Dataset: primary	R^2 Test: 0.467	MSE: 2.687e+13
Model: fit_alpha_0.70	Dataset: primary	R^2 Test: 0.467	MSE: 2.688e+13
Model: fit_alpha_0.75	Dataset: primary	R^2 Test: 0.468	MSE: 2.686e+13
Model: fit_alpha_0.80	Dataset: primary	R^2 Test: 0.468	MSE: 2.686e+13
Model: fit_alpha_0.85	Dataset: primary	R^2 Test: 0.467	MSE: 2.687e+13
Model: fit_alpha_0.90	Dataset: primary	R^2 Test: 0.467	MSE: 2.688e+13
Model: fit_alpha_0.95	Dataset: primary	R^2 Test: 0.467	MSE: 2.688e+13
Model: fit_alpha_1.00	Dataset: primary	R^2 Test: 0.467	MSE: 2.688e+13
Model: fit_alpha_0.00	Dataset: complete	R^2 Test: 0.619	MSE: 1.947e+13
Model: fit_alpha_0.05	Dataset: complete	R^2 Test: 0.613	MSE: 1.977e+13
Model: fit_alpha_0.10	Dataset: complete	R^2 Test: 0.614	MSE: 1.974e+13
Model: fit_alpha_0.15	Dataset: complete	R^2 Test: 0.615	MSE: 1.967e+13
Model: fit_alpha_0.20	Dataset: complete	R^2 Test: 0.609	MSE: 1.997e+13
Model: fit_alpha_0.25	Dataset: complete	R^2 Test: 0.606	MSE: 2.011e+13
Model: fit_alpha_0.30	Dataset: complete	R^2 Test: 0.602	MSE: 2.034e+13
Model: fit_alpha_0.35	Dataset: complete	R^2 Test: 0.603	MSE: 2.029e+13
Model: fit_alpha_0.40	Dataset: complete	R^2 Test: 0.601	MSE: 2.036e+13
Model: fit_alpha_0.45	Dataset: complete	R^2 Test: 0.600	MSE: 2.041e+13
Model: fit_alpha_0.50	Dataset: complete	R^2 Test: 0.600	MSE: 2.045e+13
Model: fit_alpha_0.55	Dataset: complete	R^2 Test: 0.599	MSE: 2.049e+13
Model: fit_alpha_0.60	Dataset: complete	R^2 Test: 0.599	MSE: 2.051e+13
Model: fit_alpha_0.65	Dataset: complete	R^2 Test: 0.598	MSE: 2.053e+13
Model: fit_alpha_0.70	Dataset: complete	R^2 Test: 0.597	MSE: 2.057e+13
Model: fit_alpha_0.75	Dataset: complete	R^2 Test: 0.599	MSE: 2.047e+13
Model: fit_alpha_0.80	Dataset: complete	R^2 Test: 0.599	MSE: 2.049e+13
Model: fit_alpha_0.85	Dataset: complete	R^2 Test: 0.598	MSE: 2.051e+13
Model: fit_alpha_0.90	Dataset: complete	R^2 Test: 0.598	MSE: 2.053e+13
Model: fit_alpha_0.95	Dataset: complete	R^2 Test: 0.598	MSE: 2.053e+13
Model: fit_alpha_1.00	Dataset: complete	R^2 Test: 0.598	MSE: 2.054e+13

Best Elastic Net Models

- Primary Dataset
  - α (optimal) = 0
    - Ridge regression
  - ƛ ≈ 52k
  - R^2 (test) = 0.471
  - MSE (test) = 2.67e13
    - RMSE ≈ 5.2M

- **Combined Dataset**
  - α (optimal) = 0
    - Ridge regression
  - ƛ ≈ 50k
  - R^2 (test) = 0.62
  - MSE (test) = 1.95e+13
    - RMSE ≈ 4.4M



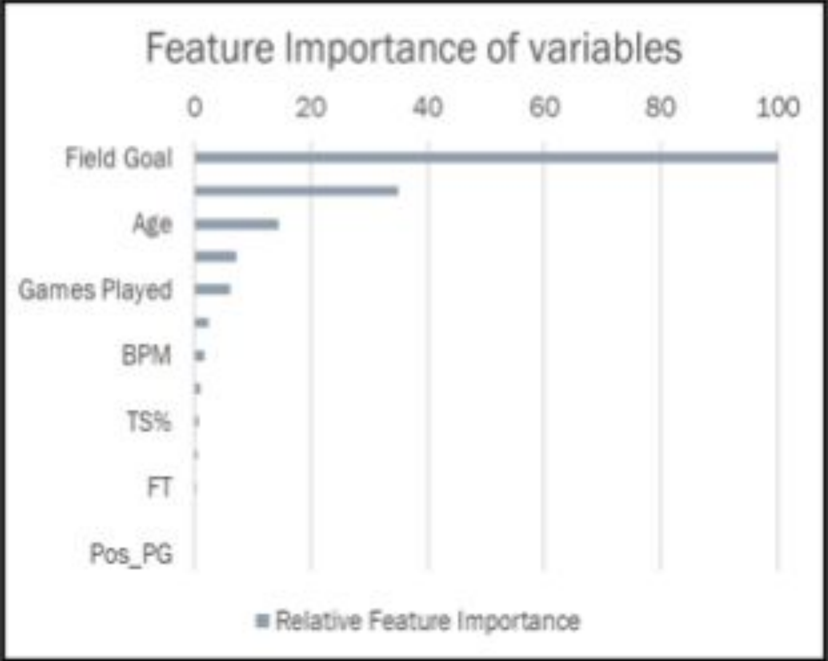
Model Performance with Decision Trees

Decision tree visualization

Accuracy Metrics

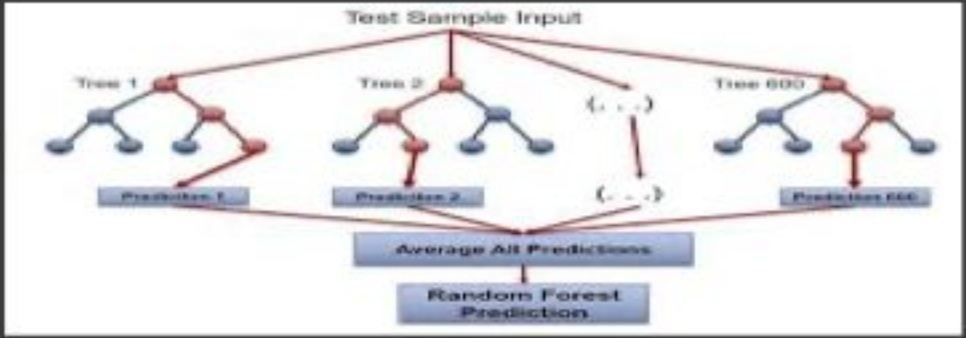
Metric	Train Data	Test Data
R-squared	64.55%	37.86%
RMSE	4361.63 k	5599.45 k
MAE	3255.67 k	4293.70 k

Feature importance plot



Model Performance with Random Forests

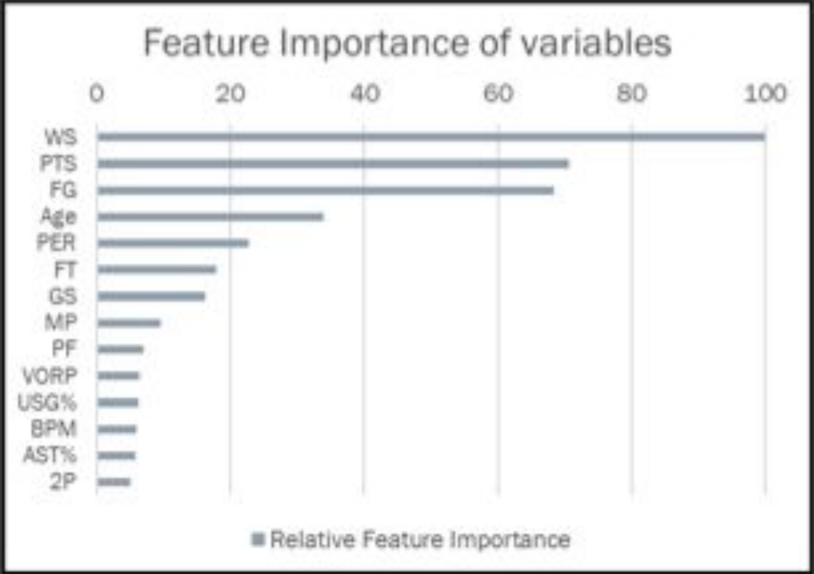
Illustrative visualization



Accuracy metrics

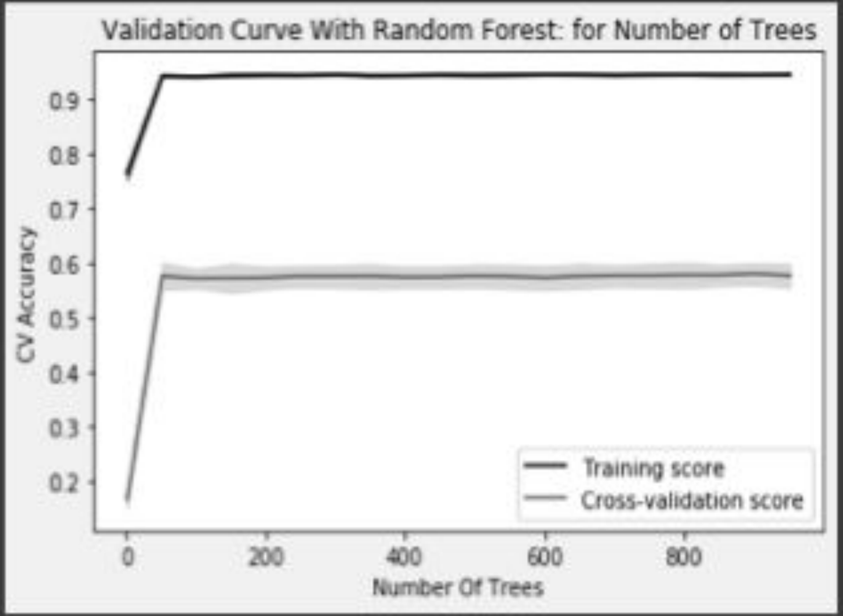
Metric	Train Data	Test Data
R-squared	80.82%	49.41%
RMSE	3207.71 k	5052.41 k
MAE	2398.16 k	3877.85 k

Feature importance plot

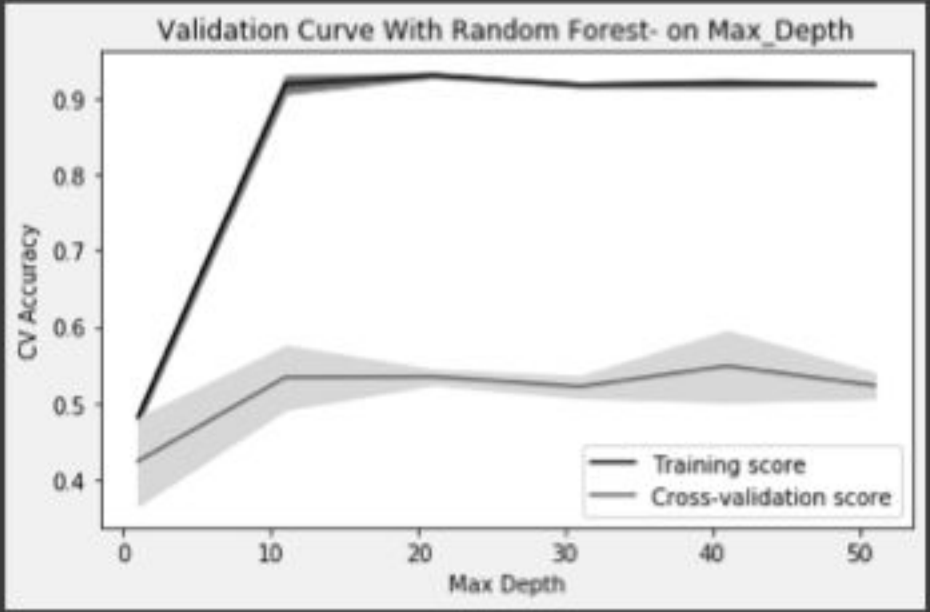


Hyper-parameter tuning for Random Forests

Grid search: number of trees

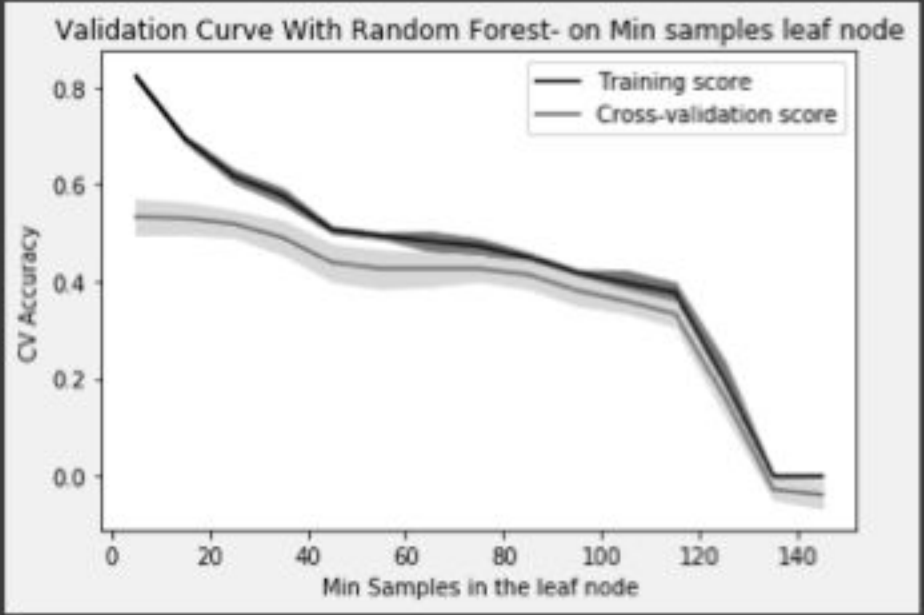


Grid search: Max Depth



Hyper-parameter tuning for Random Forests

Grid search: Min Samples leaf node

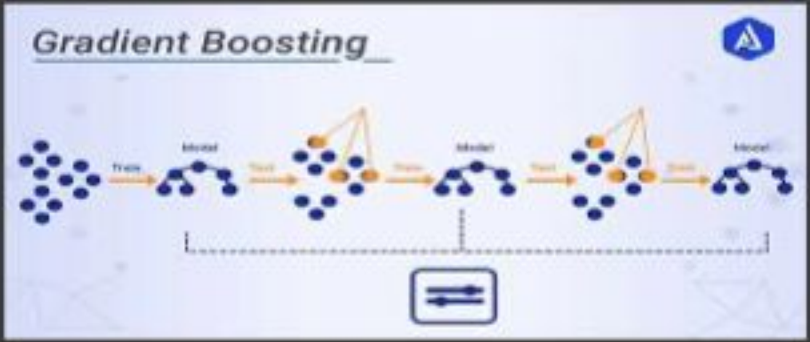


Best Hyperparameters

HYPERPARAMETER	*VALUE*
'max_depth'	6
'max_features'	46
'min_samples_leaf'	5
'n_estimators'	400

Model Performance with GBM

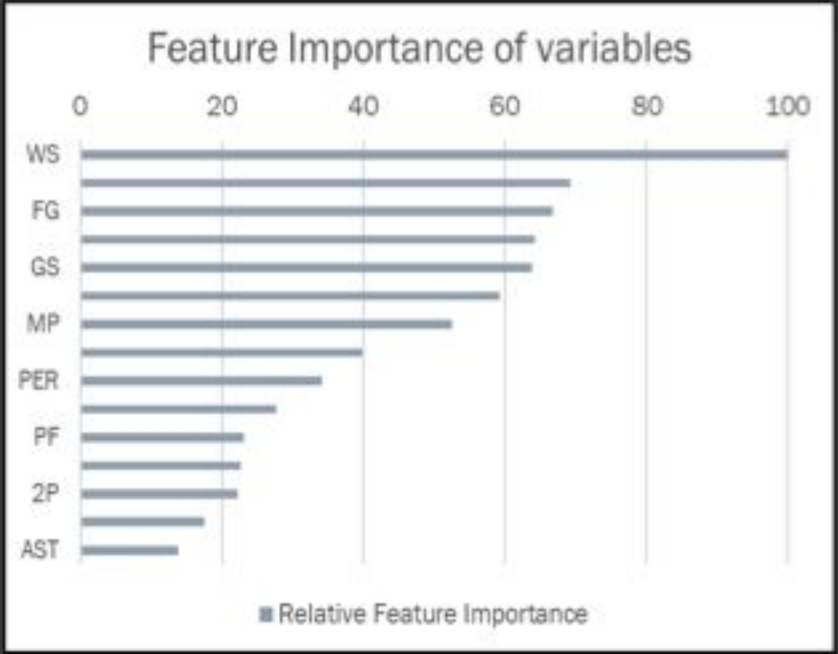
Illustrative visualization



Accuracy metrics

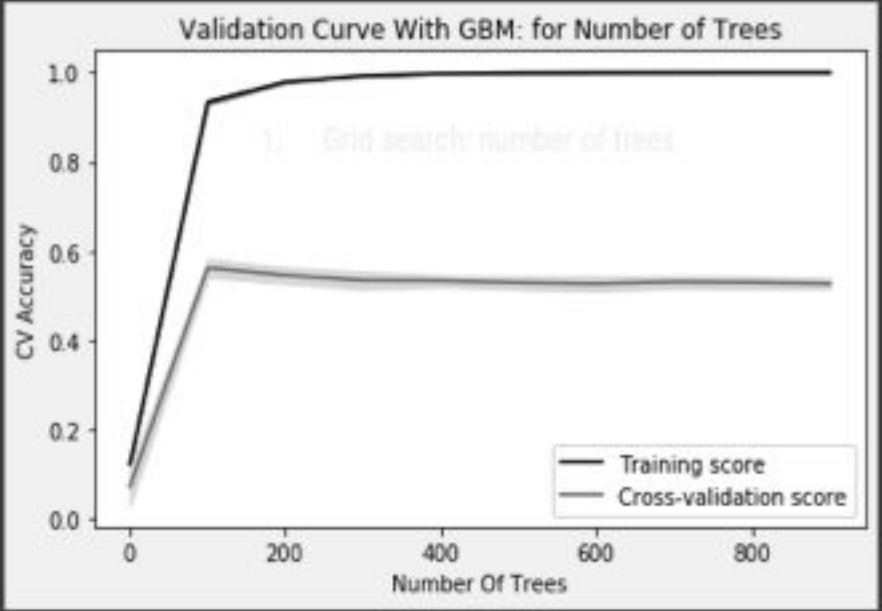
Metric	Train Data	Test Data
R-squared	93.88%	48.30%
RMSE	1811.82 k	5107.31 k
MAE	1356.15 k	3861.93 k

Feature importance plot

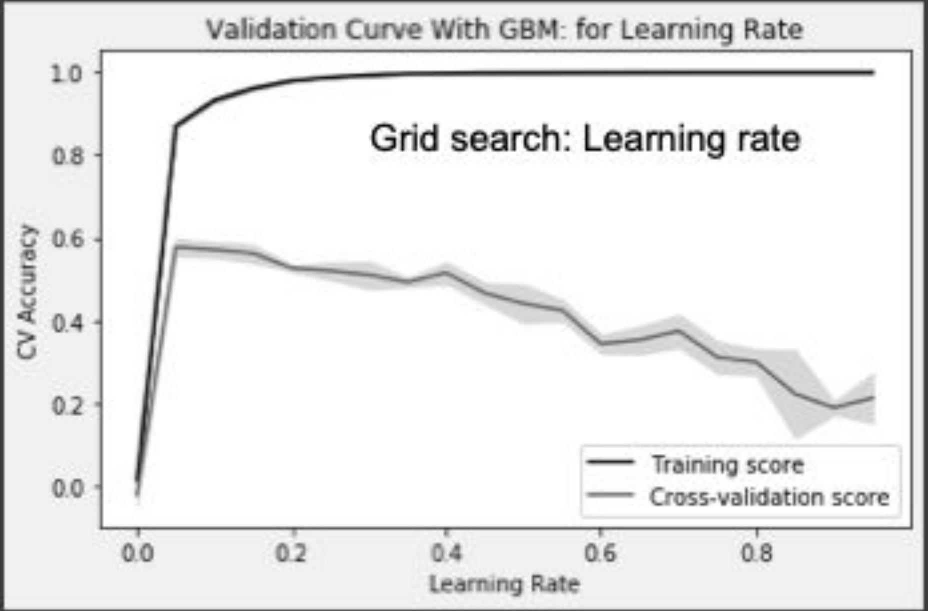


Hyper-parameter tuning for GBM

Grid search: number of trees

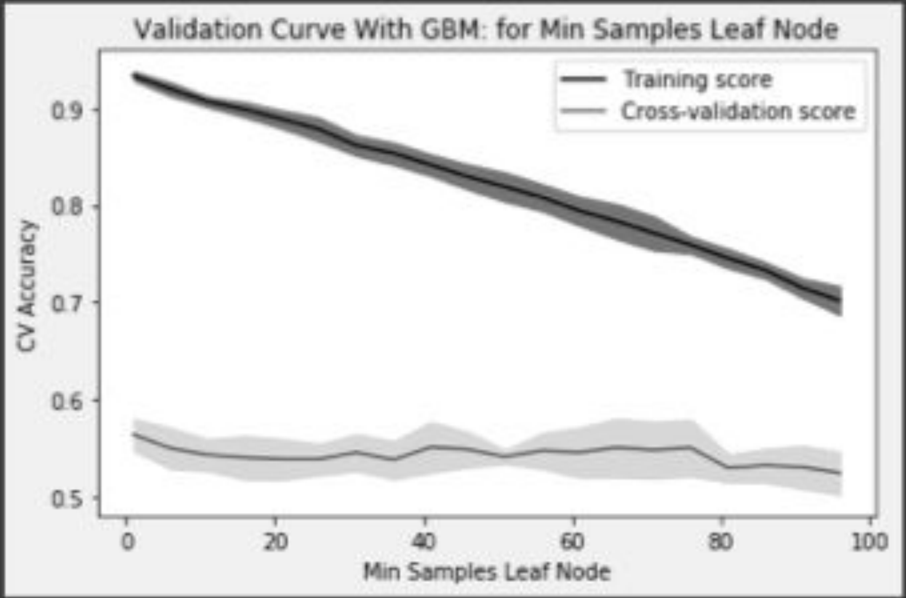


Grid search: Learning rate



Hyper-parameter tuning for GBM

Grid search: Minimum samples in the leaf node



Best Hyperparameters

HYPERPARAMETER	***VALUE***
'learning_rate'	0.01
'max_depth'	6
'min_samples_leaf'	5
'n_estimators'	400
'max_features'	10

Underrated and Overrated Players By Optimal Elastic Net Model

Underrated players

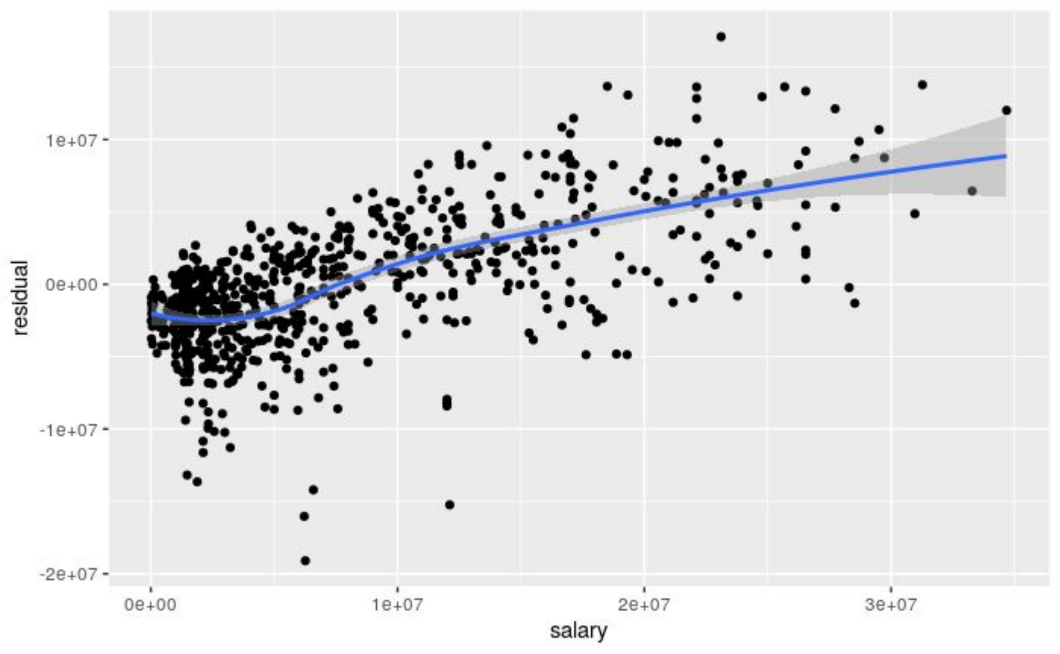
- Model predicts to make > 2x as much
- 92 players in 2016
  - 62% saw > 25% increase for 2017

Overrated players

- Model predicts to make < ½ as much
- 31 players in 2016
  - 29% saw decrease for 2017

Model tends to

- Overestimate players with lower salaries
- Underestimate players with higher salaries



Model Deployment

- Plumber [6]
  - Swagger API
- Access
  - Query from web
  - Query from terminal

LeBron James 2016 Query

```
curl -X POST
'http://localhost:8000/predict_salary?Age=31&G=76&GS=76&MP=2709&PER=27.5&TS.=.588&X3PAr=0.199&FTr=0.347&ORB.=4.7&DRB.=18.8&TRB.=11.8&AST.=36.0&STL.=2.0&BLK.=1.5&TOV.=13.2&USG.=31.4&OWS=9.6&DWS=4&WS=13.6&WS.48=0.242&OBPM=6.9&DBPM=2.3&BPM=9.1&VORP=7.6&FG=737&FGA=1416&FG.=0.520&X3P=87&X3PA=282&X3P.=0.309&X2P=650&X2PA=1134&X2P.=0.573&eFG.=0.551&FT=359&FTA=491&FT.=0.731&ORB=111&DRB=454&TRB=565&AST=514&STL=104&BLK=49&TOV=249&PF=143&PTS=1920&out=94&ovr=99&ins=89&pla=91&ath=92&def=91&reb=91'
```

Future Work

- Primary Dataset
  - Web scrape ESPN
- Secondary Dataset
  - Contracts
    - Restricted vs unrestricted
  - Salary cap
  - Years in the league
    - Scaled min/max player salary

Max Player Salary [9]			
Years in league	0-6	7-9	10+
Max Salary (as % of Salary Cap)	25	30	35

Small a NBA Salary Cap Between 2020 and 2021 [11]

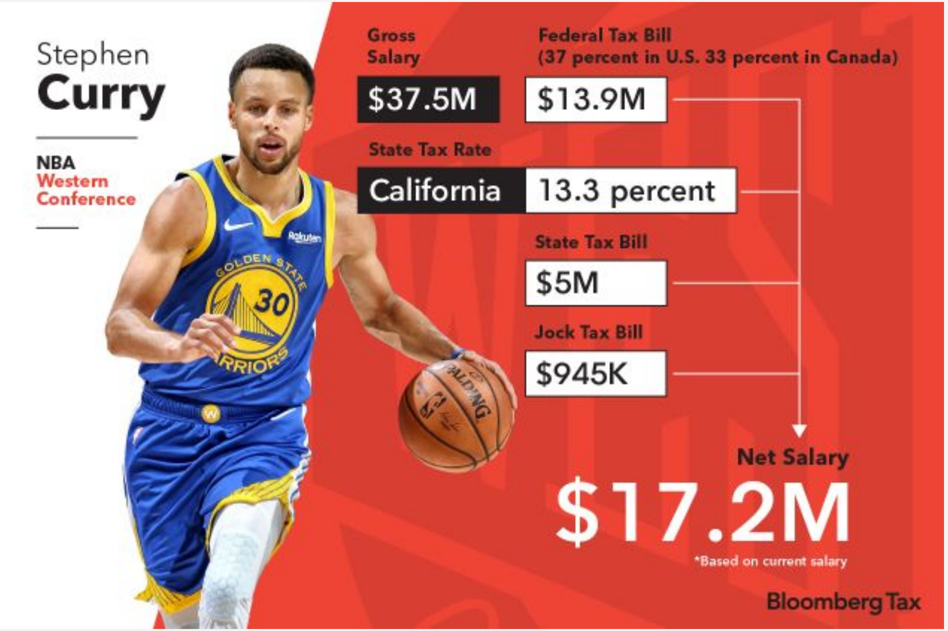
References

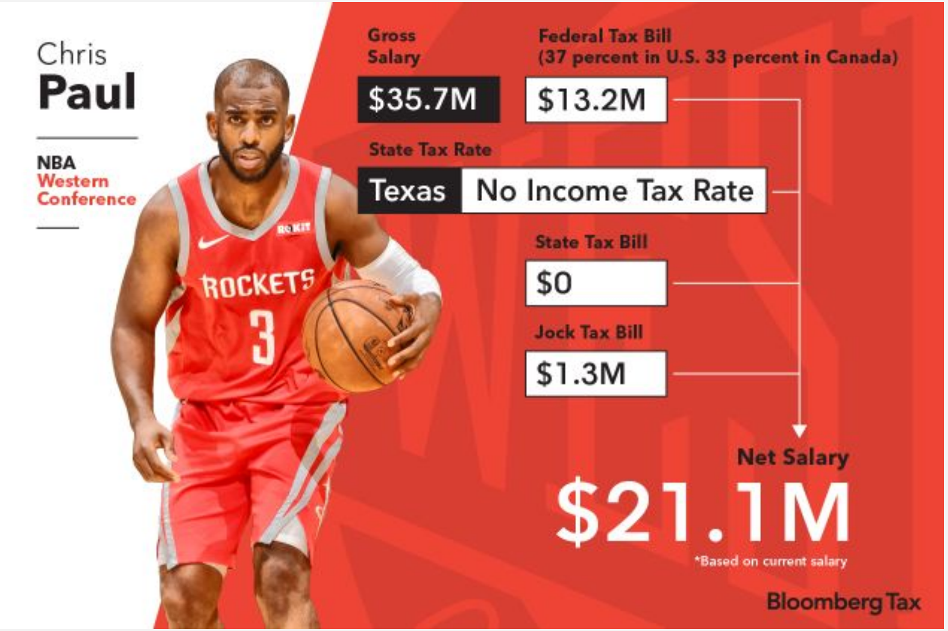
[1] Wesolowski-Mantilla, A. (2017). A Breakdown of NBA Salaries. Northwestern Business Review.
https://northwesternbusinessreview.org/a-breakdown-of-nba-salaries-d8c37c56a4ee
2] Blanco, F. (2018). NBA - Advanced & Basic Season Stats (1950-2017). Kaggle.
https://www.kaggle.com/whitefero/nba-players-advanced-season-stats-19782016
[3] My Team Database (MTDB). (2015).
http://mtdb.com/20
[4] Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86.
https://CRAN.R-project.org/package=caret
[5] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.
http://www.jstatsoft.org/v33/i01/
[6] Trestle Technology, LLC (2018). plumber: An API Generator for R. R package version 0.4.6.
https://CRAN.R-project.org/package=plumber
[7] Ryan Prete. NBA Superstars Face Tax Bills in the Millions. (2018). Bloomberg Tax.
https://news.bloombergtax.com/daily-tax-report-state/nba-superstars-face-tax-bills-in-the-millions

References

[8] csp471_fp GitHub repository. GitHub.
https://github.com/alegresor/csp571_fp
[9] Kerry Miller. How NBA Free Agency, Salary Cap Work. (2018). Bleacher Report.
https://bleacherreport.com/articles/2787871-how-nba-free-agency-salary-cap-work
[10] Josh Rosson. NBA Salary Predictions using Data Science and Linear Regression. (2019). Towards Data Science.
https://towardsdatascience.com/nba-salary-predictions-4cd09931eb55
[11] Ben Glover. NBA announced record salary cap for 2016-17 after historic climb. (2016). Sports Illustrated.
https://www.si.com/nba/2016/07/03/nba-salary-cap-record-numbers-2016-adam-silver

Thank You for Listening. Questions?





2018 Player Taxes [7]

GitHub [8]