MLB Salary

B02705013 蕭友量
B02705027 陳信豪
B02705028 冷俊瑩

Outline

- Introduction
- Related Research
- Data
- Regression
- Classification
- Ensemble
- Summary

Introduction

- Players with better performance deserve better salary.
- A player’s salary should be decided by his performance last year.

Related Research

- Using Regression to Predict Baseball Salaries.
- By Nate Reed.
- http://natereed.com/using-regression-to-predict-baseball-salaries/

Related Research - Modeling

- Clean useless data.
- Feature Engineering.
  - Normalizing input variables.
  - Rescale salary
- Creating Linear Regression Model.
  - $R^2 = 0.68$
  - 5-fold Cross Validation -> Accuracy = 0.65 ± 0.08

Related Research - Regularization

- Ridge : Performs L2 Regularization.
  - Did not eliminate any variable.
- LASSO : Performs L1 Regularization.
  - 38 variables eliminated.
- ElasticNet : Combines L1 and L2 Regularization.
  - 35 variables eliminated.

Related Research - Simplified Model

- Eliminating statistically insignificant variables.
- OLS Regression.

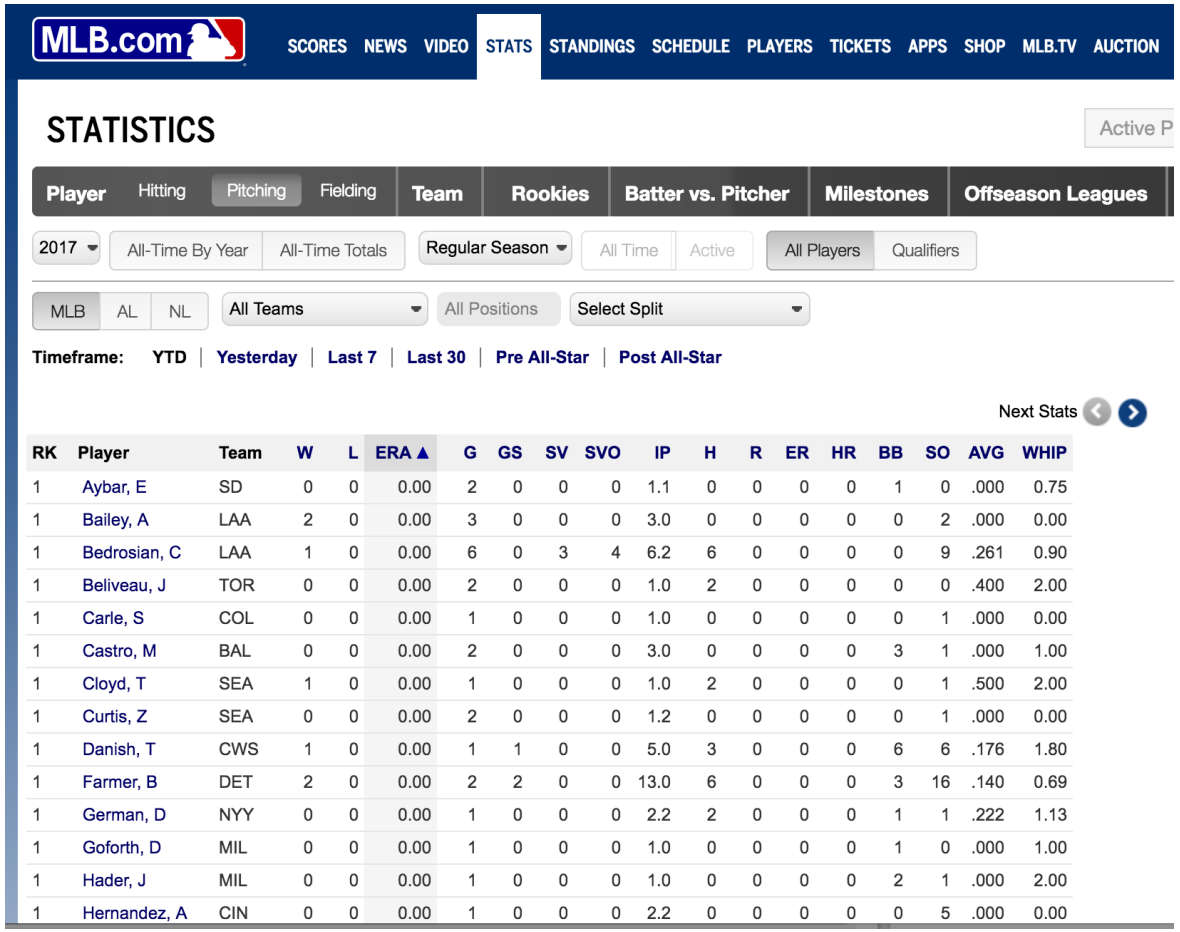
Adjusted Salary = 0.15 + 3.29 * Batting_Career_TB - 0.32 * Pitching_Career_IP + 6.3 * Pitching_Career_SO + 3.22 * Num_All_Star_Appearances - 0.34 * NO_POSITION + 0.2 * FIRST_BASE + 0.4 * SECOND_BASE

- $R^2 = 0.64$

Data - Source

- MLB Official Website – Player Statistics
- Spotrac – Player Salary
- 2011 ~ 2016
- 3834 observations, 85 features

Data - Source



RK	Player	Team	W	L	ERA ▲	G	GS	SV	SVO	IP	H	R	ER	HR	BB	SO	AVG	WHIP
1	Aybar, E	SD	0	0	0.00	2	0	0	0	1.1	0	0	0	0	1	0	.000	0.75
1	Bailey, A	LAA	2	0	0.00	3	0	0	0	3.0	0	0	0	0	0	2	.000	0.00
1	Bedrosian, C	LAA	1	0	0.00	6	0	3	4	6.2	6	0	0	0	0	9	.261	0.90
1	Beliveau, J	TOR	0	0	0.00	2	0	0	0	1.0	2	0	0	0	0	0	.400	2.00
1	Carle, S	COL	0	0	0.00	1	0	0	0	1.0	0	0	0	0	0	1	.000	0.00
1	Castro, M	BAL	0	0	0.00	2	0	0	0	3.0	0	0	0	0	3	1	.000	1.00
1	Cloyd, T	SEA	1	0	0.00	1	0	0	0	1.0	2	0	0	0	0	1	.500	2.00
1	Curtis, Z	SEA	0	0	0.00	2	0	0	0	1.2	0	0	0	0	0	1	.000	0.00
1	Danish, T	CWS	1	0	0.00	1	1	0	0	5.0	3	0	0	0	6	6	.176	1.80
1	Farmer, B	DET	2	0	0.00	2	2	0	0	13.0	6	0	0	0	3	16	.140	0.69
1	German, D	NYY	0	0	0.00	1	0	0	0	2.2	2	0	0	0	1	1	.222	1.13
1	Goforth, D	MIL	0	0	0.00	1	0	0	0	1.0	0	0	0	0	1	0	.000	1.00
1	Hader, J	MIL	0	0	0.00	1	0	0	0	1.0	0	0	0	0	2	1	.000	2.00
1	Hernandez, A	CIN	0	0	0.00	1	0	0	0	2.2	0	0	0	0	0	5	.000	0.00

Data - Source

2017 | 2017 Base Salaries | Filter by Team | - All Positions - | All Status Types | GO

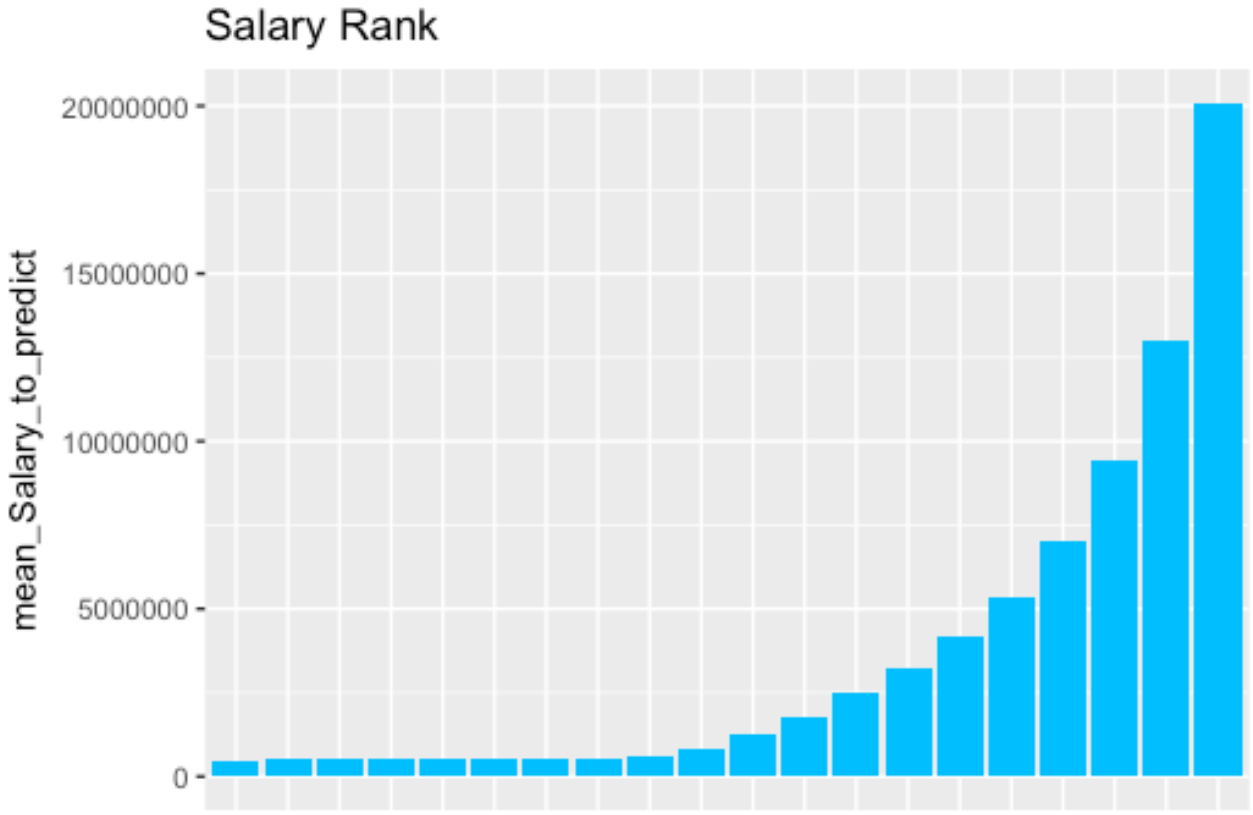
Reset

2017 Base Rankings

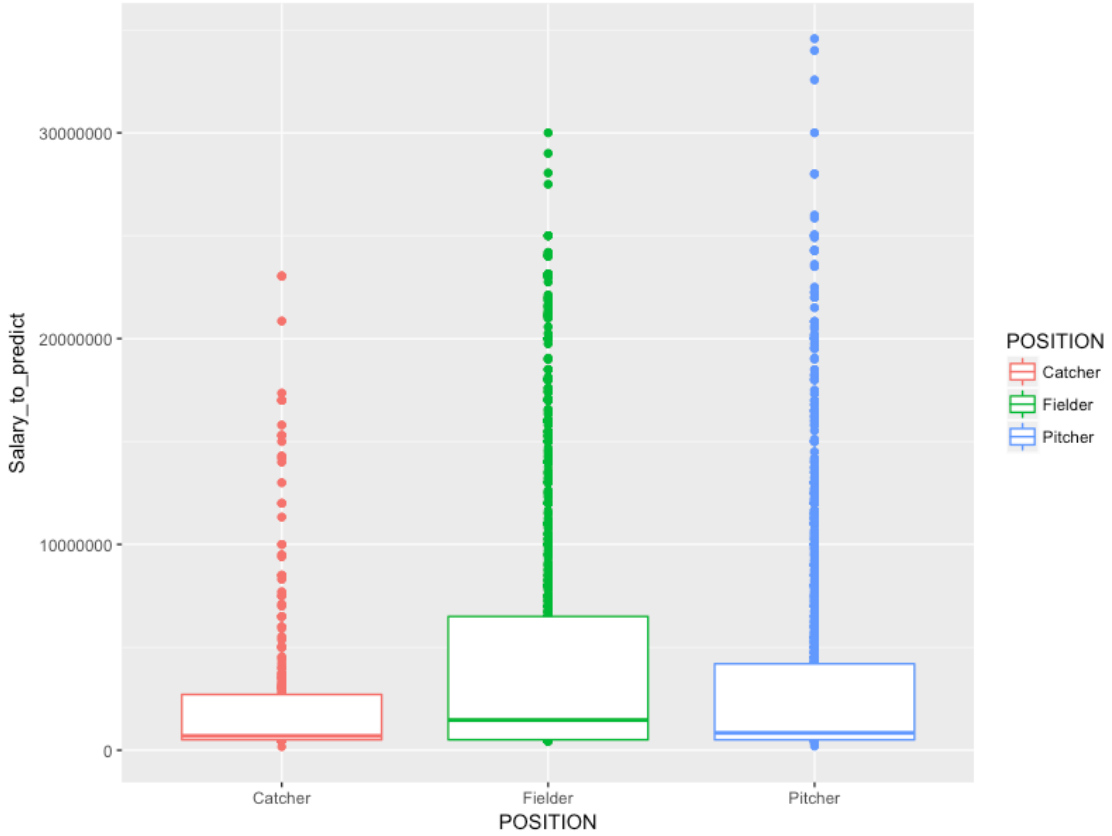
		PLAYER	2017 BASE
1		Clayton Kershaw STARTING PITCHER	$33,000,000
2		Zack Greinke STARTING PITCHER	$31,000,000
3		David Price STARTING PITCHER	$30,000,000
4		Miguel Cabrera 1ST BASE	$28,000,000
		Justin Verlander STARTING PITCHER	$28,000,000
6		Albert Pujols DESIGNATED HITTER	$26,000,000
		Felix Hernandez STARTING PITCHER	$26,000,000
8		C.C. Sabathia STARTING PITCHER	$25,000,000

spotrac

Data – Salary Distribution



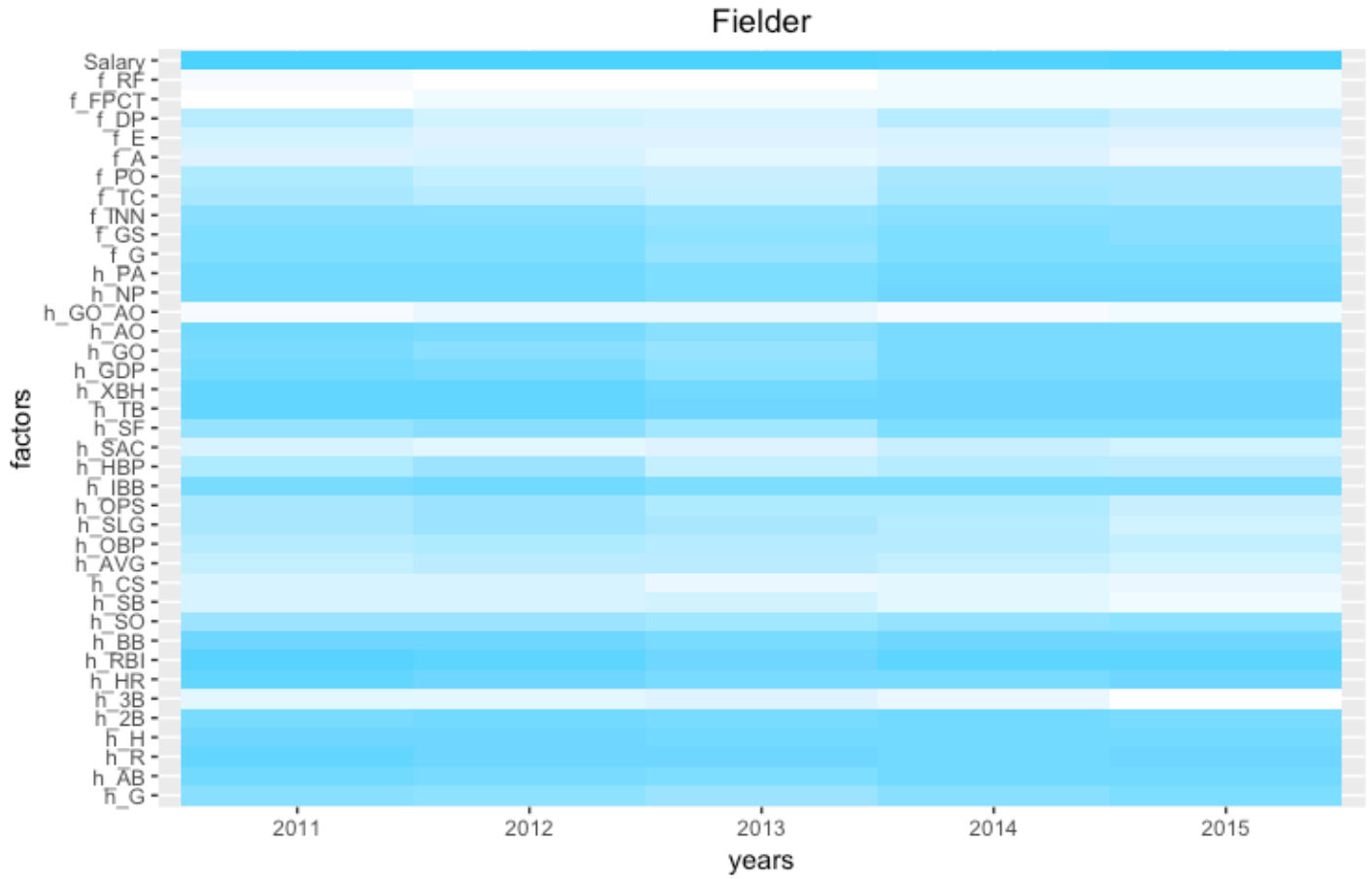
Data – Salary Distribution



Data - Feature

Basic	• Player Name, Team, Position, Salary
Hitting	• G, AB, H, 2B, 3B, HR, RBI, BB, SO, SB, CS, AVG, OBP, SLG, OPS ….
Fielding	• G, GS, INN, TC, PO, A, E, DP, SB, CS, SBPCT, PB, C_WP, FPCT, RF …
Pitching	• W, L, ERA, G, GS, SV, SVO, IP, H, R, ER, HR, BB, SO, AVG, WHIP …

Feature Selection – Correlation Heatmap



Feature Selection - Correlation

Pitcher	Catcher	Fielder
Salary	Salary	Salary
Pitching SO	Hitting BB	Hitting RBI
Pitching IP	Hitting NP	Hitting TB
Fielding INN	Hitting H	Hitting XBH
Pitching IBF	Hitting R	Hitting R
Pitching NP	Hitting RBI	Hitting BB
Pitching W	Hitting GO	Hitting HR
Pitching GO	Hitting PA	Hitting H
Pitching H	Hitting AB	Hitting NP
Pitching AO	Hitting AO	Hitting PA
Pitching GS	Fielding G	Hitting AB
Fielding GS	Fielding PO	Hitting 2B
Fielding TC	Fielding GS	Hitting AO
Fielding A	Fielding TC	Hitting GDP

Feature Selection – Chi2

	Pitcher	Catcher	Fielder
Original Feature	86	86	86
Selected Feature	61	38	39
Selection Ratio	69.77%	44.19%	45.34%

Feature Normalization

- Performance is relative in each year
- Scale each feature in each year

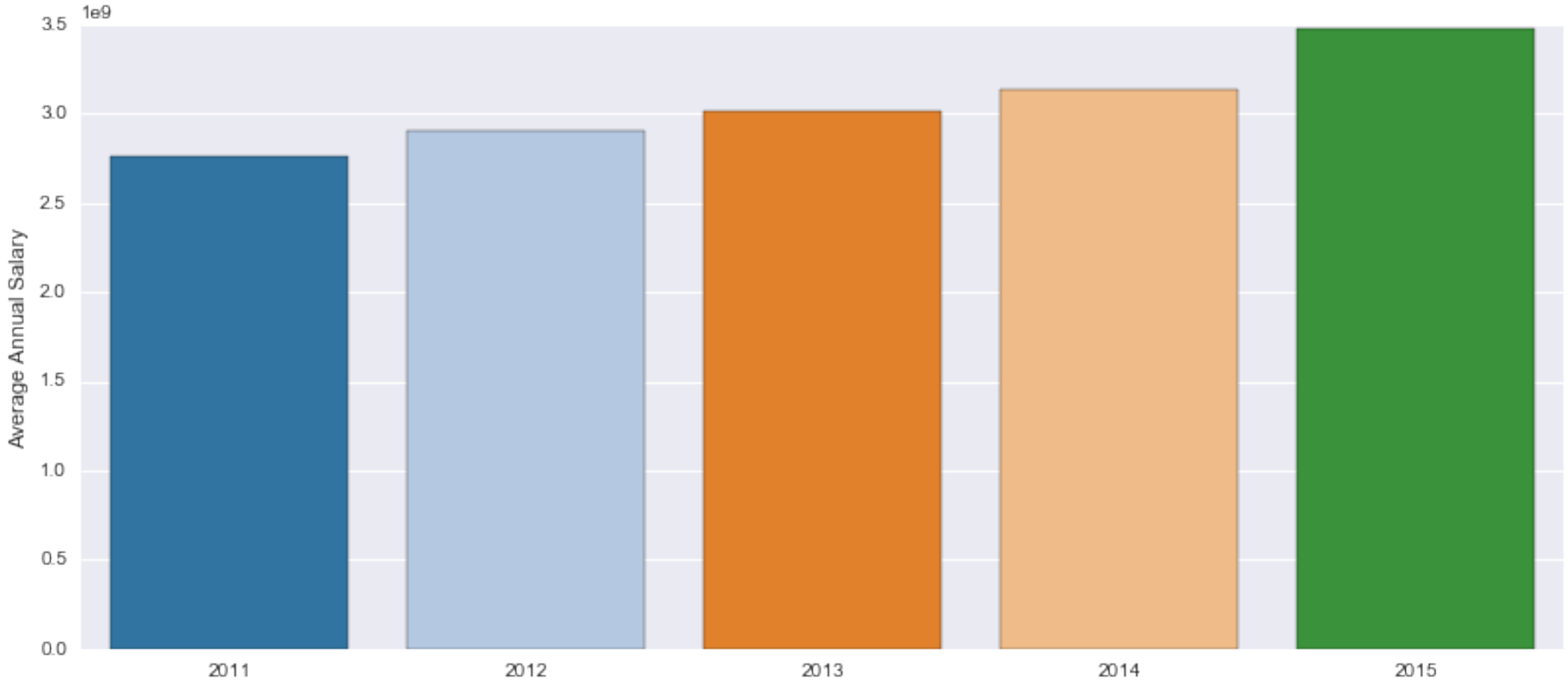
Regression

- RMSE : 2054935
- Relative Mean Abs Error ： 70%

Regression – Cause of Failure

- Contract Based Salary
- Average salary tends to get higher every year
- Disabled List

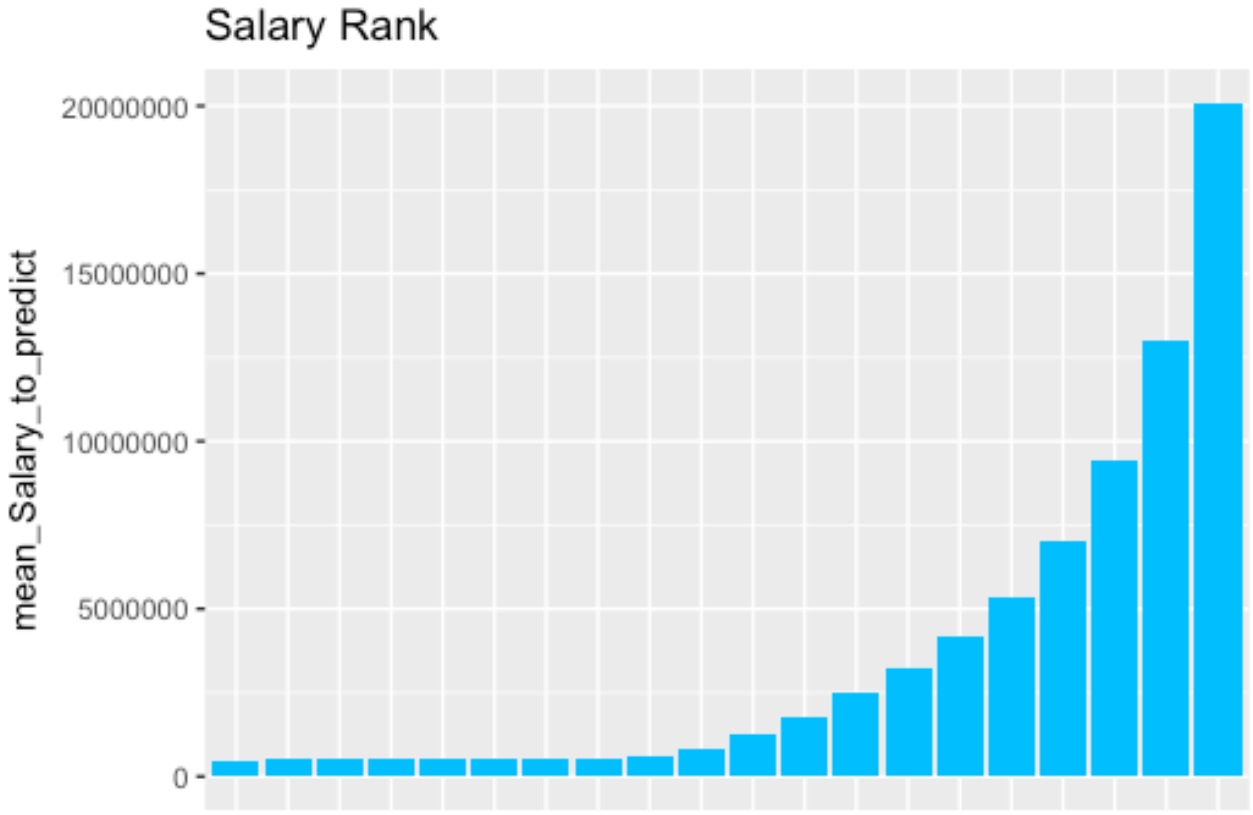
Regression – Cause of Failure



Classification

- Determine whether the player‘s performance deserves raise on salary.

Classification



Classifier

- Logistic Regression
- RandomForest
- Libsvm
- Xgboost

Classifier –Logistic Regression

- “improved” = 1
- “umimproved “= 0

Classifier - RandomForest

- Parameter
  - mtry
  - ntree

Classifier - Libsvm

- library e1071
- Parameter
  - kernel (linear, rbf, poly, sigmoid)

Classifier - Xgboost

- Gradient Boosting Machine
- Powerful and fast

$$Obj(\Theta) = L(\theta) + \Omega(\Theta)$$

- Parameter
  - max_depth

K fold Cross-Validation

(before train-test-spliting)

K = 10

1/10 tuning

9/10 training

Classifier – Pitcher Comparison

	LogisticRegr	RandomForest	Libsvm	Xgboost
Mean Train Acc	0.6859502173	0.7045790198	0.7049024569	0.7923430145
Mean Valid Acc	0.6422589628	0.706045945	0.6522902889	0.7113087365
Mean Test Acc	0.6239361702	0.6739361702	0.620212766	0.6734042553

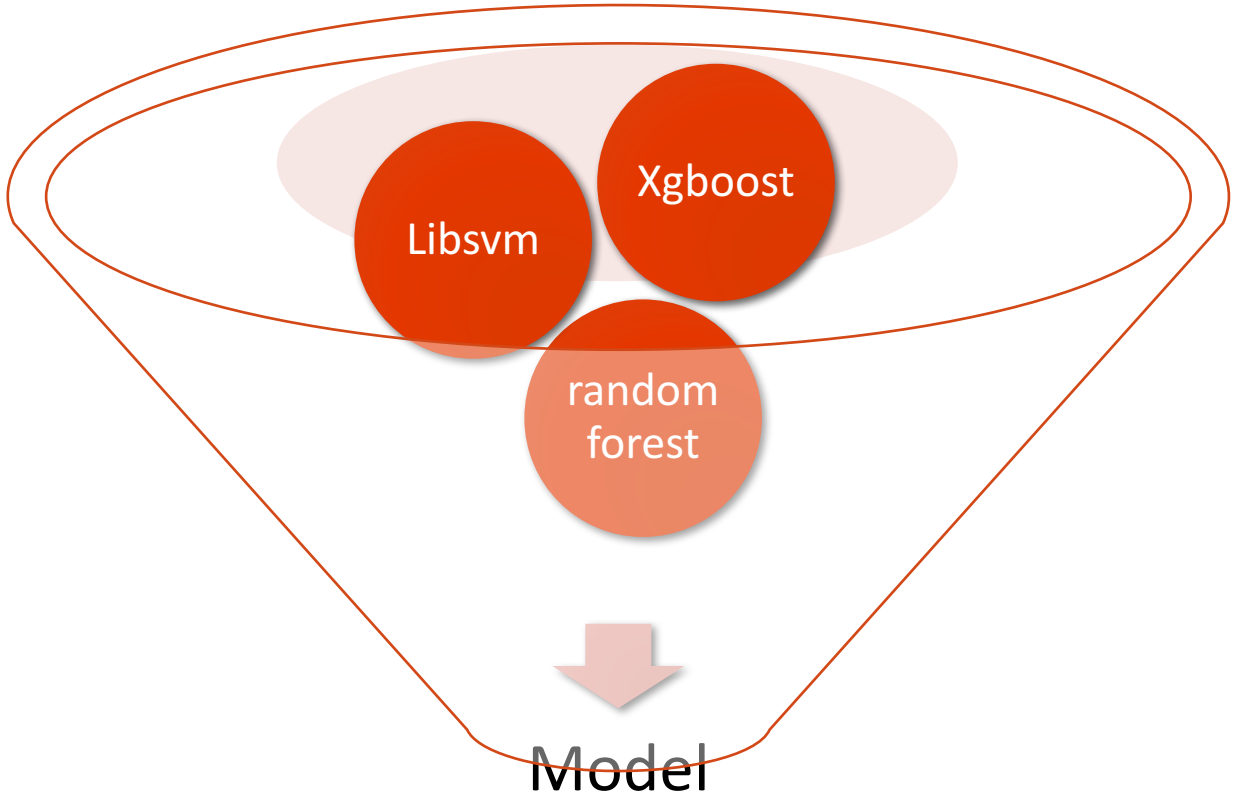
Classifier – Catcher Comparison

	LogisticRegr	RandomForest	Libsvm	Xgboost
Mean Train Acc	0.6744898897	0.6263602941	0.6373912377	0.8719852941
Mean Valid Acc	0.5280788177	0.6052955665	0.6349753695	0.7009852217
Mean Test Acc	0.7085714286	0.6657142857	0.6514285714	0.6471428571

Classifier – Fielder Comparison

	LogisticRegr	RandomForest	Libsvm	Xgboost
Mean Train Acc	0.6989419181	0.6808750417	0.7048879832	0.8125888984
Mean Valid Acc	0.6740821678	0.678962704	0.6768502331	0.7081390831
Mean Test Acc	0.6937106918	0.6993710692	0.6849056604	0.6597484277

Ensemble



Ensemble

	1	2	3	4
RandomForest	1	0	0	1
Libsvm	0	1	0	1
Xgboost	1	0	0	0
Voting Sum	2	1	0	2
Voting Mean	0.67	0.33	0	0.67
Voting Result	1	0	0	1
	improved	unimproved	unimproved	improved

Ensemble

	1	2	3	4
RandomForest	1 * 1.5	0 * 1.5	0 * 1.5	1 * 1.5
Libsvm	0 * 1	1 * 1	0 * 1	1 * 1
Xgboost	1 * 2	0 * 2	0 * 2	0 * 2
Voting Sum	3.5	1	0	2.5
Voting Mean	0.78	0.22	0	0.56
Voting Result	1	0	0	1
	improved	unimproved	unimproved	improved

Ensemble

- Taking various types of models to make it better.

- It may not be the best, but it is assumed to be good.

Ensemble

	Pitcher	Catcher	Fielder
Best Mean Test Acc	0.6739361702	0.7085714286	0.6993710692
Blending Test Acc	0.670212766	0.7285714286	0.7044025157

Summary

- Take the position into consideration.
- Using RandomForest and Xgboost gives us good result in this MLB Salary Case.
- Try Model Blending to get the best result.

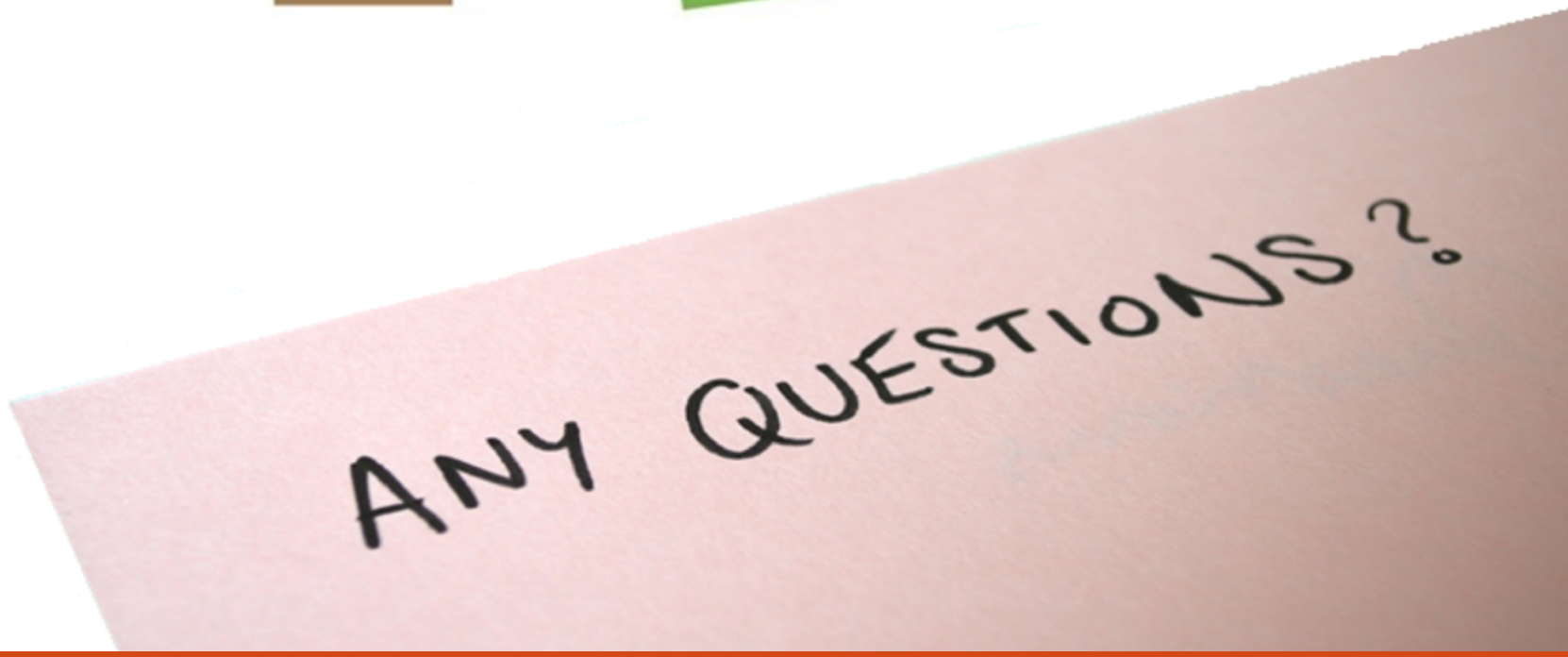
Summary – To be improved

- Our dataset is too small
- Lack of some features
  e.g. age, years of experience, team record, times appear in all star game
- Consider the correlation between different features.
  - OPS = OBP + SLG

Reference

- Ensemble method of machine learning 機器學習中的組合方法 https://read01.com/3zJOK.html
- Introduction to Boosted Trees http://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf
- MLB Sortable Player Stats http://mlb.mlb.com/stats/sortable.jsp
- MLB Salary Rankings http://www.spotrac.com/mlb/rankings/
- Pay for Play: Are Baseball Salaries Based on Performance? https://ww2.amstat.org/publications/jse/v6n2/datasets.watnik.html#denby
- A Critical Look at Some Analyses of Major League Baseball Salaries https://www.jstor.org/stable/2684201?seq=1#page_scan_tab_contents


THANK YOU


ANY QUESTIONS?