Predicting Major League Baseball (MLB) Player Salaries

Jeff Hoffman

I. Introduction

Introduction

Goal: Build a model using XGBoost to predict Major League Baseball (MLB) player salaries.

Purpose: Determine player value and provide insights on what drives value creation.

Application: Salary negotiations, team budgets, overpaid/underpaid players

Client: MLB teams and their organizations, fantasy baseball

II. Data

Data

http://www.seanlahman.com/baseball-archive/statistics/

- Batting data
- Salary data
- All Star data
- Pitching data

https://www.bls.gov/cpi/

- Consumer Price Index (CPI) data

Batting Features	Description (http://m.mlb.com/glossary/)
playerID	Player ID code
yearID	Year
stint	player's stint: Order of appearances within a season
teamID	Team - a factor
lgID	League - a factor with levels AA AL FL NL PL UA
G	Games Played - A player is credited with having played a game if he appears in it at any point -- be it as a starter or a replacement.
AB	At-bat - An official at-bat comes when a batter reaches base via a fielder's choice, hit or an error (not including catcher's interference) or when a batter is put out on a non-sacrifice.
R	Run - A player is awarded a run if he crosses the plate to score his team a run.
H	Hit - A hit occurs when a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice.
2B	Double - A batter is credited with a double when he hits the ball into play and reaches second base without the help of an intervening error or attempt to put out another baserunner.
3B	Triple - A triple occurs when a batter hits the ball into play and reaches third base without the help of an intervening error or attempt to put out another baserunner.
HR	Home Run - A home run occurs when a batter hits a fair ball and scores on the play without being put out or without the benefit of an error.
RBI	Runs Batted In - A batter is credited with an RBI in most cases where the result of his plate appearance is a run being scored.
SB	Stolen Bases - A stolen base occurs when a baserunner advances by taking a base to which he isn't entitled.
CS	Caught Stealing - A caught stealing occurs when a runner attempts to steal but is tagged out before reaching second base, third base or home plate.
BB	Walk - A walk occurs when a pitcher throws four pitches out of the strike zone, none of which are swung at by the hitter. After refraining from swinging at four pitches out of the zone, the batter is awarded first base.
SO	Strikeout - A strikeout occurs when a pitcher throws any combination of three swinging or looking strikes to a hitter.
IBB	Intentional Walk - An intentional walk occurs when the defending team elects to walk a batter on purpose, putting him on first base instead of letting him try to hit.
HBP	Hit-by-pitch - A hit-by-pitch occurs when a batter is struck by a pitched ball without swinging at it.
SH	Sacrifice Bunt - A sacrifice bunt occurs when a player is successful in his attempt to advance a runner (or multiple runners) at least one base with a bunt.
SF	Sacrifice Fly - A sacrifice fly occurs when a batter hits a fly-ball out to the outfield or foul territory that allows a runner to score.
GIDP	Ground Into Double Play - A GIDP occurs when a player hits a ground ball that results in multiple outs on the bases.

III. Data Wrangling

Data Wrangling

Step 1: Remove the pitchers from the batting data.

Step 2: Remove all years before 1985 from the batting data.

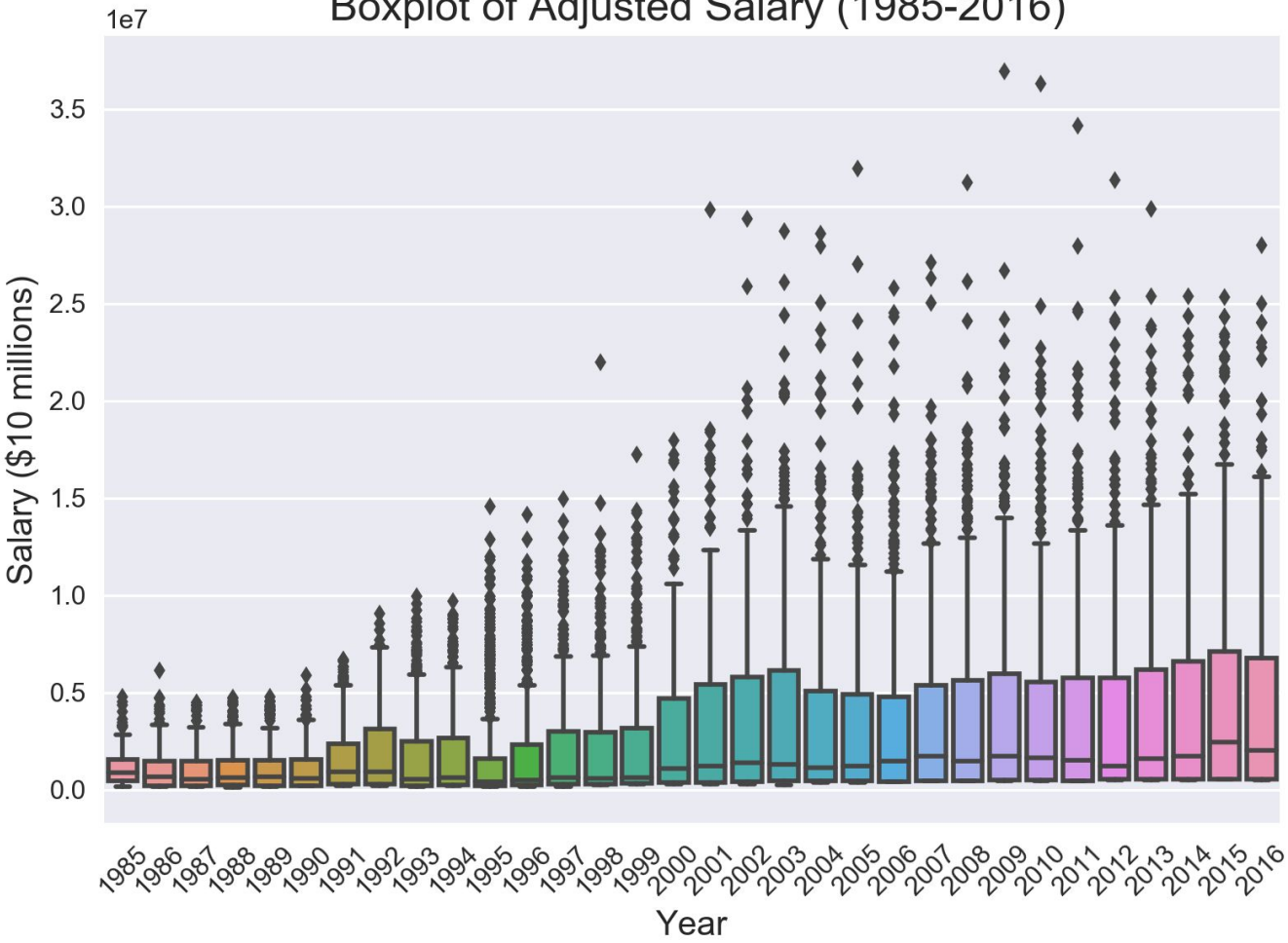
Step 3: Merge the batting data with the salary data.

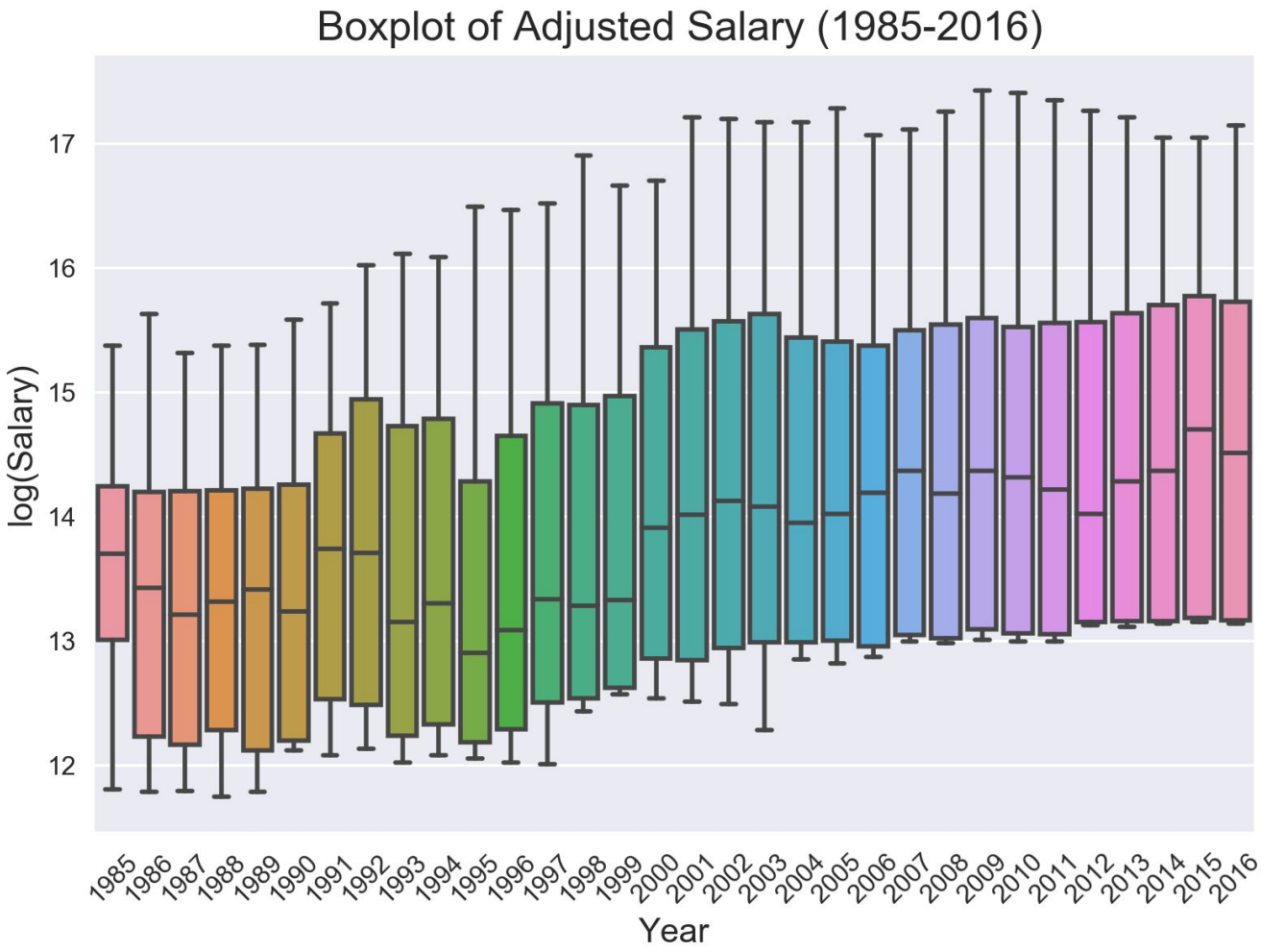
Step 4: Remove data where salary is below the set minimum salary in 1985.

Step 5: Add Experience and All Star features.

Step 6: Adjust salary for inflation.

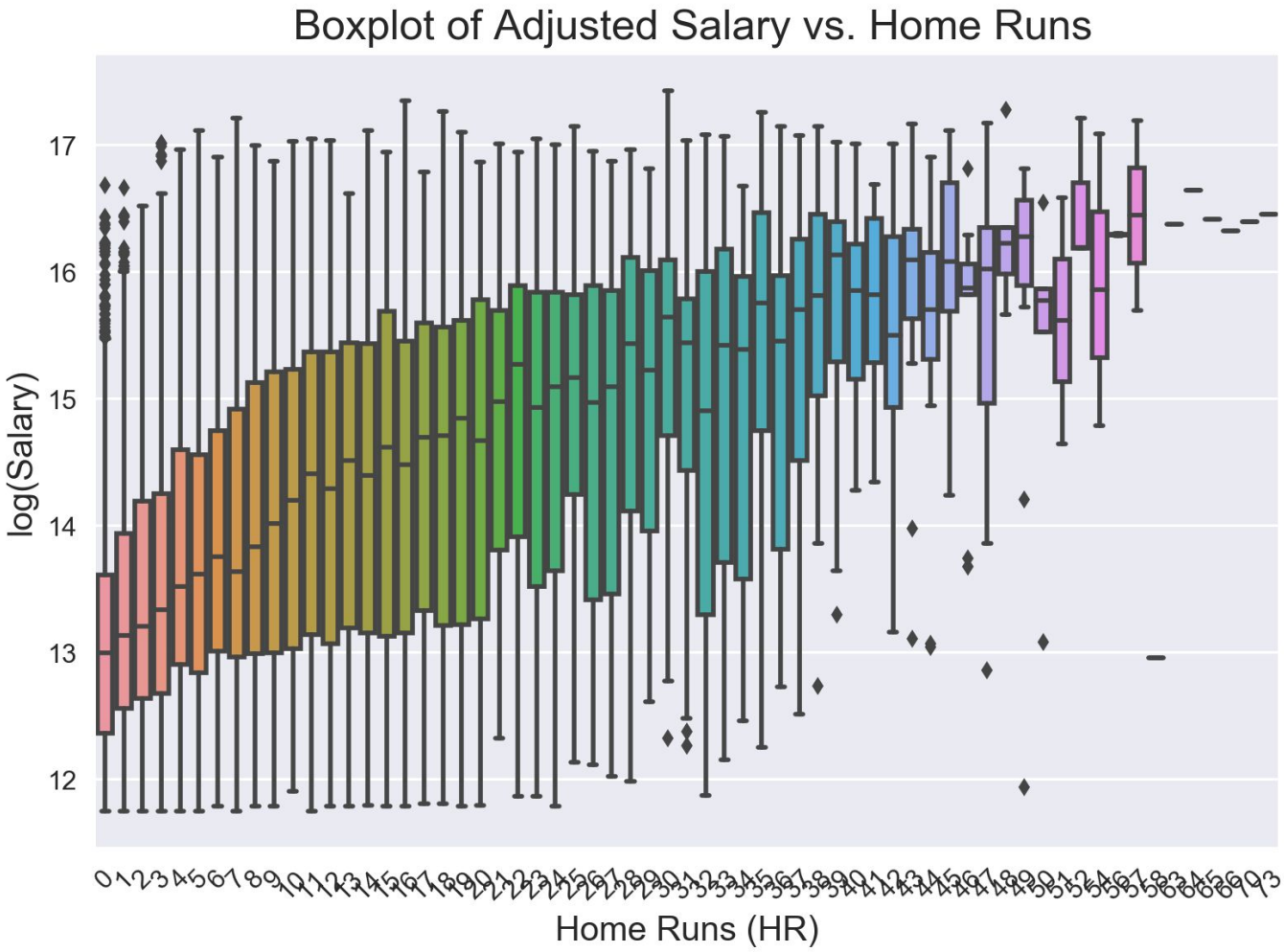
IV. Exploratory Data Analysis

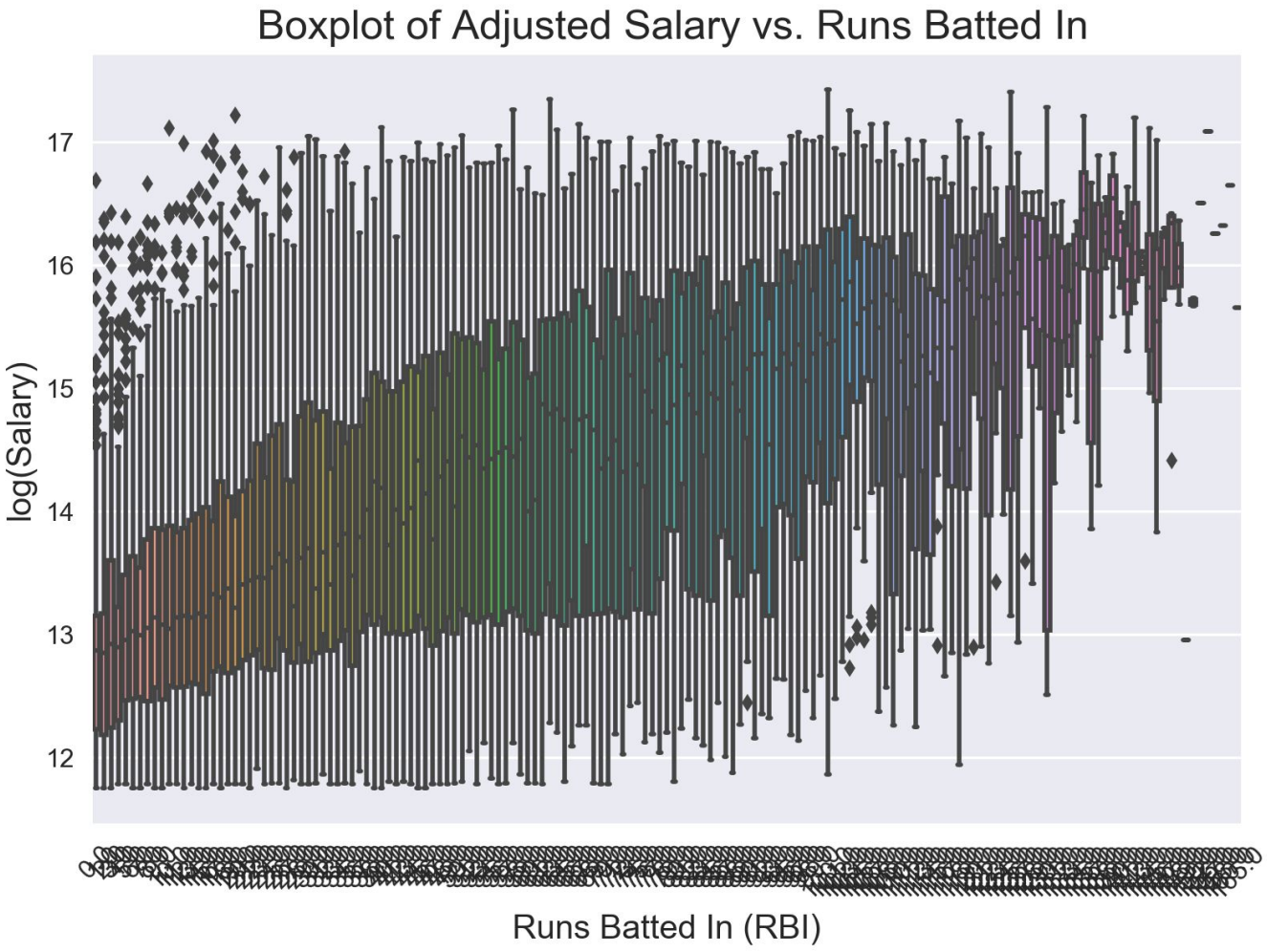

Boxplot of Adjusted Salary (1985-2016)
1e7
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Salary ($10 millions)
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year

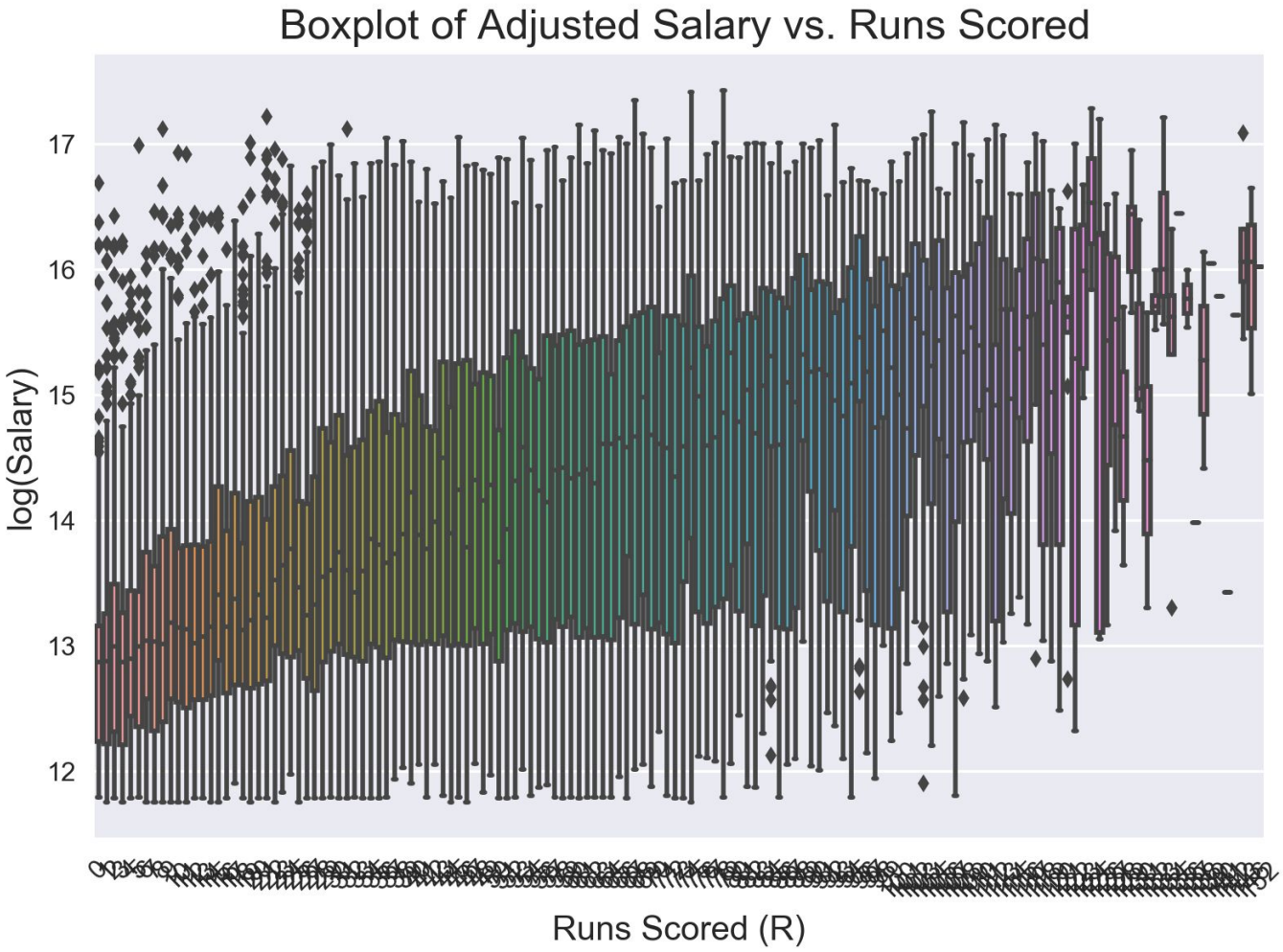

Boxplot of Adjusted Salary (1985-2016)
log(Salary)
17
16
15
14
13
12
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year

Let’s look at some plots between salary and what the MLB calls “standard stats”:

- batting average (AVG)
- **home runs (HR)**
- **runs batted in (RBI)**
- **runs (R)**
- stolen bases (SB)


Boxplot of Adjusted Salary vs. Home Runs
log(Salary)
17
16
15
14
13
12
Home Runs (HR)

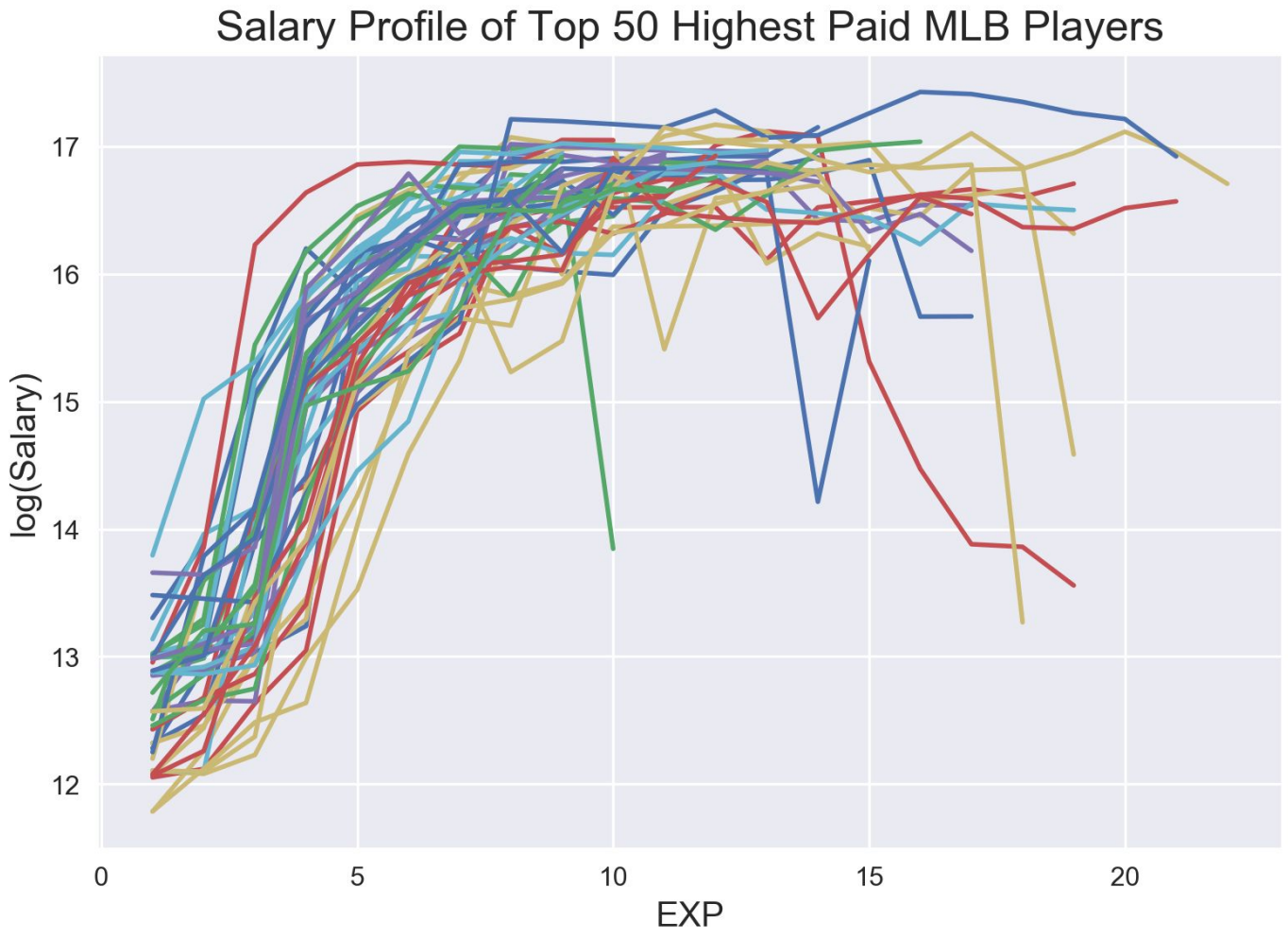

Boxplot of Adjusted Salary vs. Runs Batted In
log(Salary)
17
16
15
14
13
12
Runs Batted In (RBI)

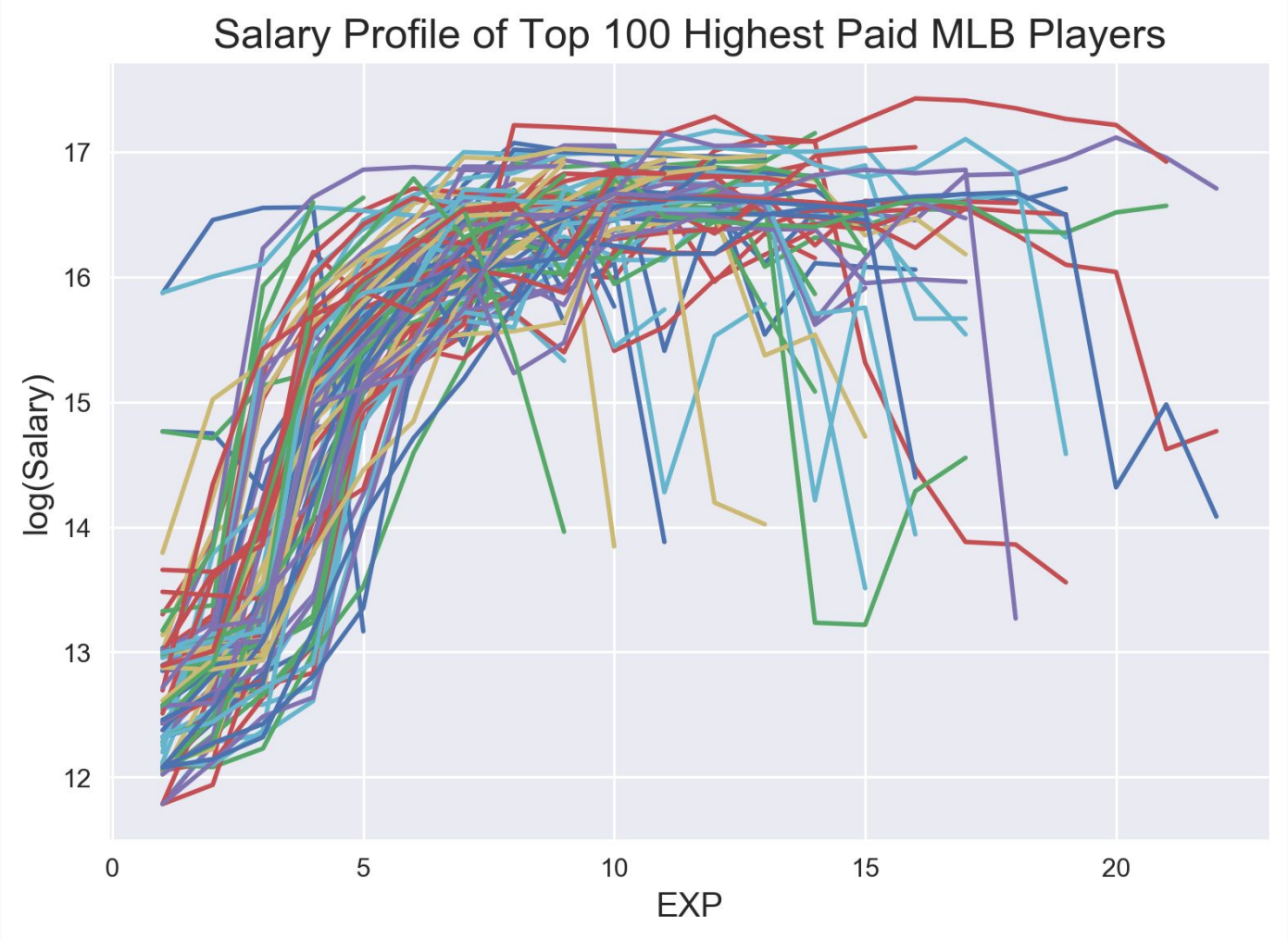

Boxplot of Adjusted Salary vs. Runs Scored
log(Salary)
17
16
15
14
13
12
Runs Scored (R)

...and how each of the features relate to each other and the target variable.

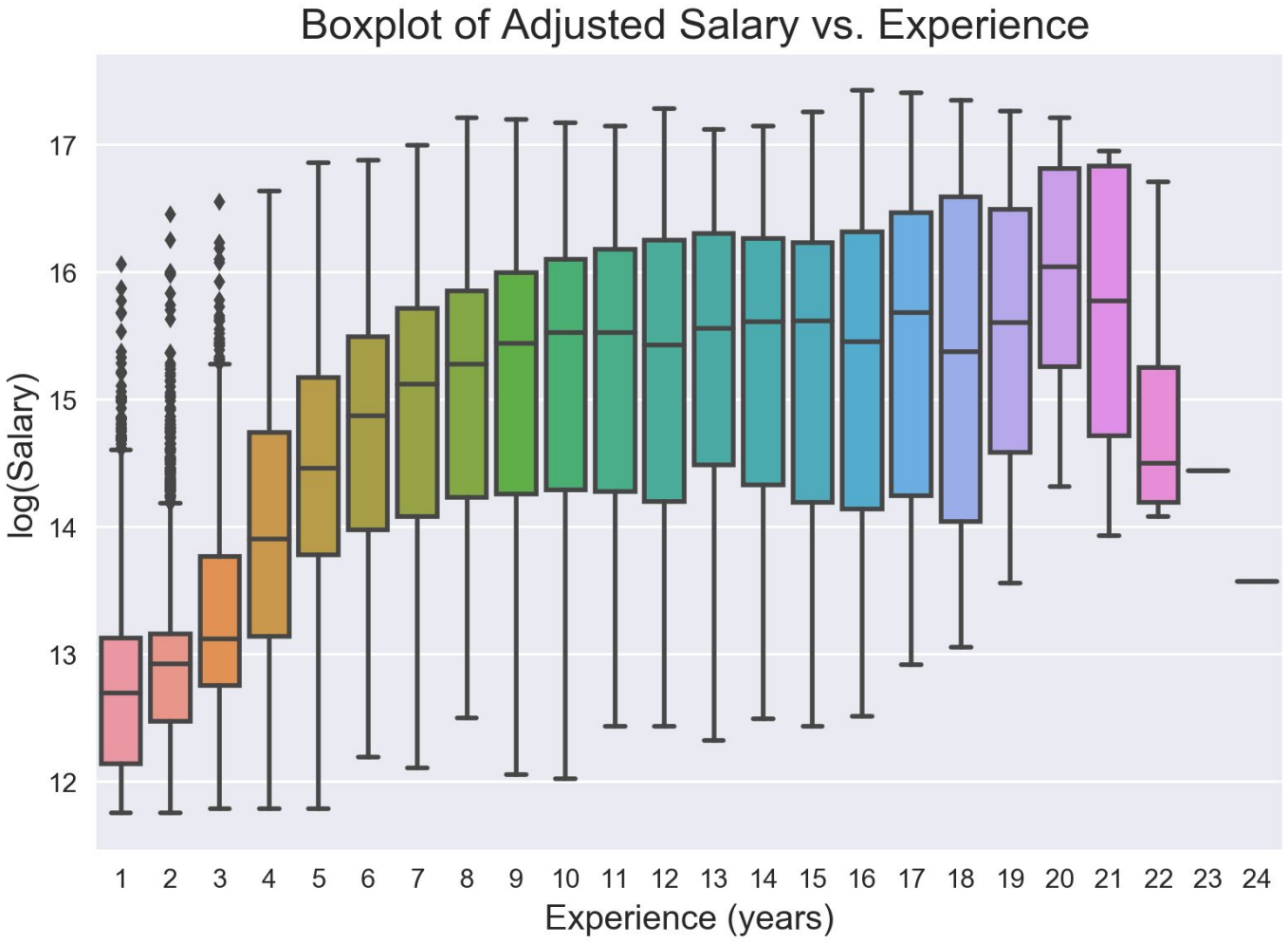
	log_salary2016	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SF	SH	GIDP	AVG
log_salary2016	1	0.39	0.46	0.46	0.46	0.45	0.11	0.45	0.49	0.13	0.1	0.47	0.37	0.35	0.28	0.37	-0.084	0.42	0.24
G	0.39	1	0.95	0.88	0.92	0.85	0.5	0.66	0.83	0.44	0.5	0.77	0.79	0.46	0.49	0.65	0.26	0.72	0.47
AB	0.46	0.95	1	0.93	0.98	0.91	0.54	0.7	0.88	0.49	0.53	0.78	0.81	0.47	0.51	0.69	0.23	0.76	0.48
R	0.46	0.88	0.93	1	0.95	0.89	0.55	0.78	0.89	0.54	0.54	0.84	0.78	0.5	0.52	0.65	0.17	0.67	0.51
H	0.46	0.92	0.98	0.95	1	0.92	0.55	0.71	0.89	0.49	0.53	0.78	0.76	0.49	0.5	0.69	0.2	0.76	0.55
2B	0.45	0.85	0.91	0.89	0.92	1	0.44	0.7	0.87	0.38	0.41	0.74	0.73	0.48	0.5	0.67	0.11	0.71	0.51
3B	0.11	0.5	0.54	0.55	0.55	0.44	1	0.22	0.36	0.59	0.56	0.36	0.39	0.16	0.22	0.29	0.29	0.24	0.29
HR	0.45	0.66	0.7	0.78	0.71	0.7	0.22	1	0.9	0.16	0.18	0.72	0.75	0.57	0.45	0.57	-0.18	0.57	0.37
RBI	0.49	0.83	0.88	0.89	0.89	0.87	0.36	0.9	1	0.28	0.32	0.79	0.78	0.58	0.49	0.73	-0.023	0.73	0.48
SB	0.13	0.44	0.49	0.54	0.49	0.38	0.59	0.16	0.28	1	0.78	0.38	0.33	0.12	0.21	0.23	0.32	0.16	0.25
CS	0.1	0.5	0.53	0.54	0.53	0.41	0.56	0.18	0.32	0.78	1	0.4	0.37	0.13	0.23	0.26	0.37	0.23	0.27
BB	0.47	0.77	0.78	0.84	0.78	0.74	0.36	0.72	0.79	0.38	0.4	1	0.72	0.62	0.43	0.58	0.066	0.56	0.4
SO	0.37	0.79	0.81	0.78	0.76	0.73	0.39	0.75	0.78	0.33	0.37	0.72	1	0.39	0.48	0.52	0.057	0.56	0.31
IBB	0.35	0.46	0.47	0.5	0.49	0.48	0.16	0.57	0.58	0.12	0.13	0.62	0.39	1	0.23	0.4	-0.11	0.39	0.28
HBP	0.28	0.49	0.51	0.52	0.5	0.5	0.22	0.45	0.49	0.21	0.23	0.43	0.48	0.23	1	0.34	0.078	0.37	0.25
SF	0.37	0.65	0.69	0.65	0.69	0.67	0.29	0.57	0.73	0.23	0.26	0.58	0.52	0.4	0.34	1	0.06	0.57	0.35
SH	-0.084	0.26	0.23	0.17	0.2	0.11	0.29	-0.18	-0.023	0.32	0.37	0.066	0.057	-0.11	0.078	0.06	1	0.073	0.08
GIDP	0.42	0.72	0.76	0.67	0.76	0.71	0.24	0.57	0.73	0.16	0.23	0.56	0.56	0.39	0.37	0.57	0.073	1	0.38
AVG	0.24	0.47	0.48	0.51	0.55	0.51	0.29	0.37	0.48	0.25	0.27	0.4	0.31	0.28	0.25	0.35	0.08	0.38	1

Let’s look at the salary profile of the top 50 and top 100 highest paid MLB players.


Salary Profile of Top 50 Highest Paid MLB Players
log(Salary)
17
16
15
14
13
12
0
5
10
15
20
EXP


Salary Profile of Top 100 Highest Paid MLB Players
log(Salary)
17
16
15
14
13
12
0
5
10
15
20
EXP

Notice the nonlinearity of the salary profile. Let’s take a look at the boxplot.


Boxplot of Adjusted Salary vs. Experience
log(Salary)
17
16
15
14
13
12
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Experience (years)

V. Feature Engineering

Feature Engineering

1. Quadratic term for experience (EXP-squared).
2. Lags of the target and feature variables.
3. On base percentage (OBP) - a measure of how often a batter reaches base.
4. Interactions between features - EXP with OBP, HR with OBP.

VI. Models and Results

Linear Regression Models

45%

Of the total variation in salary is explained by only using the feature variables given to us, lagged one year.

Baseline OLS

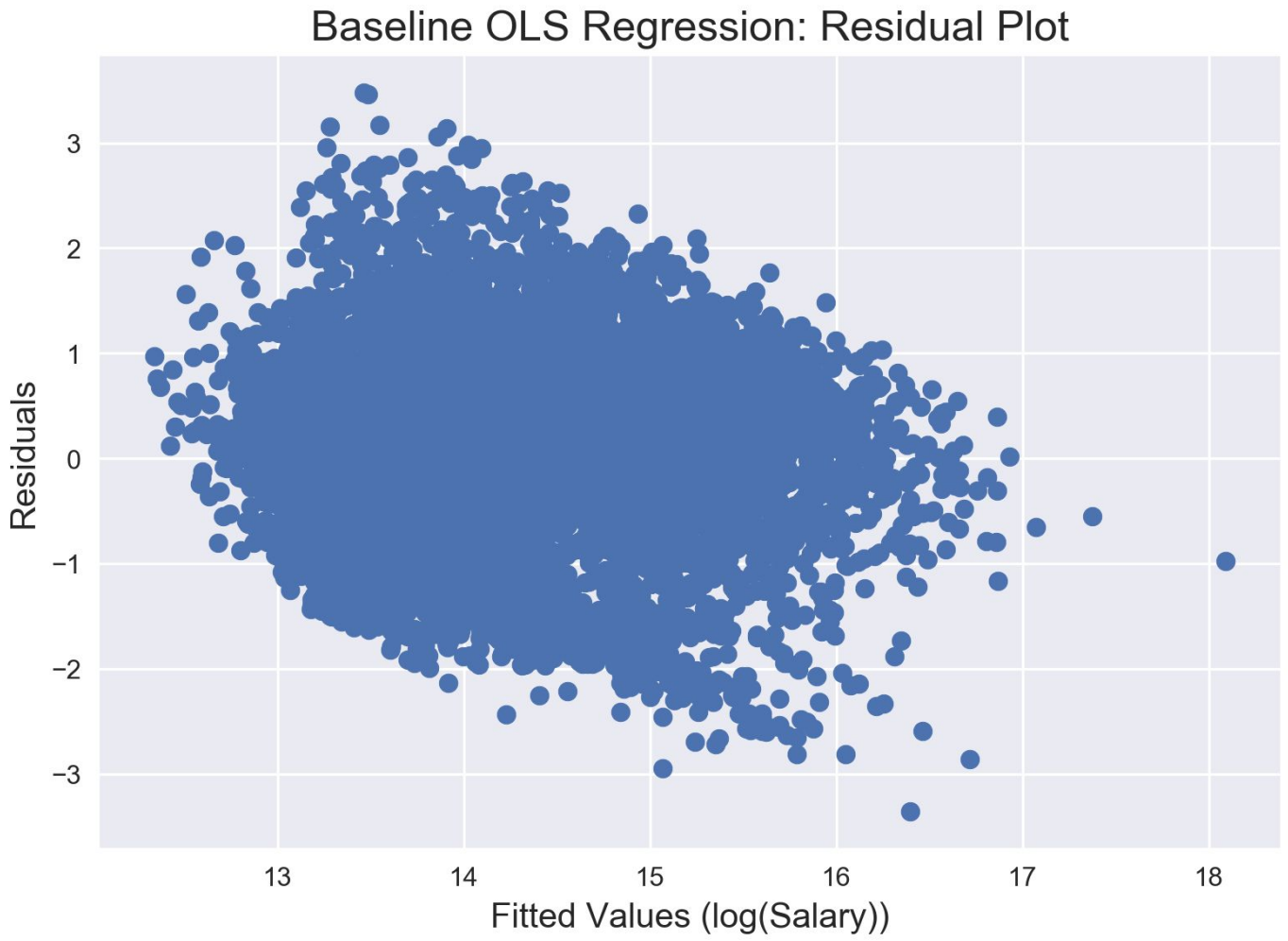
- “Lazy” model
- Just lag what you are given by one year.

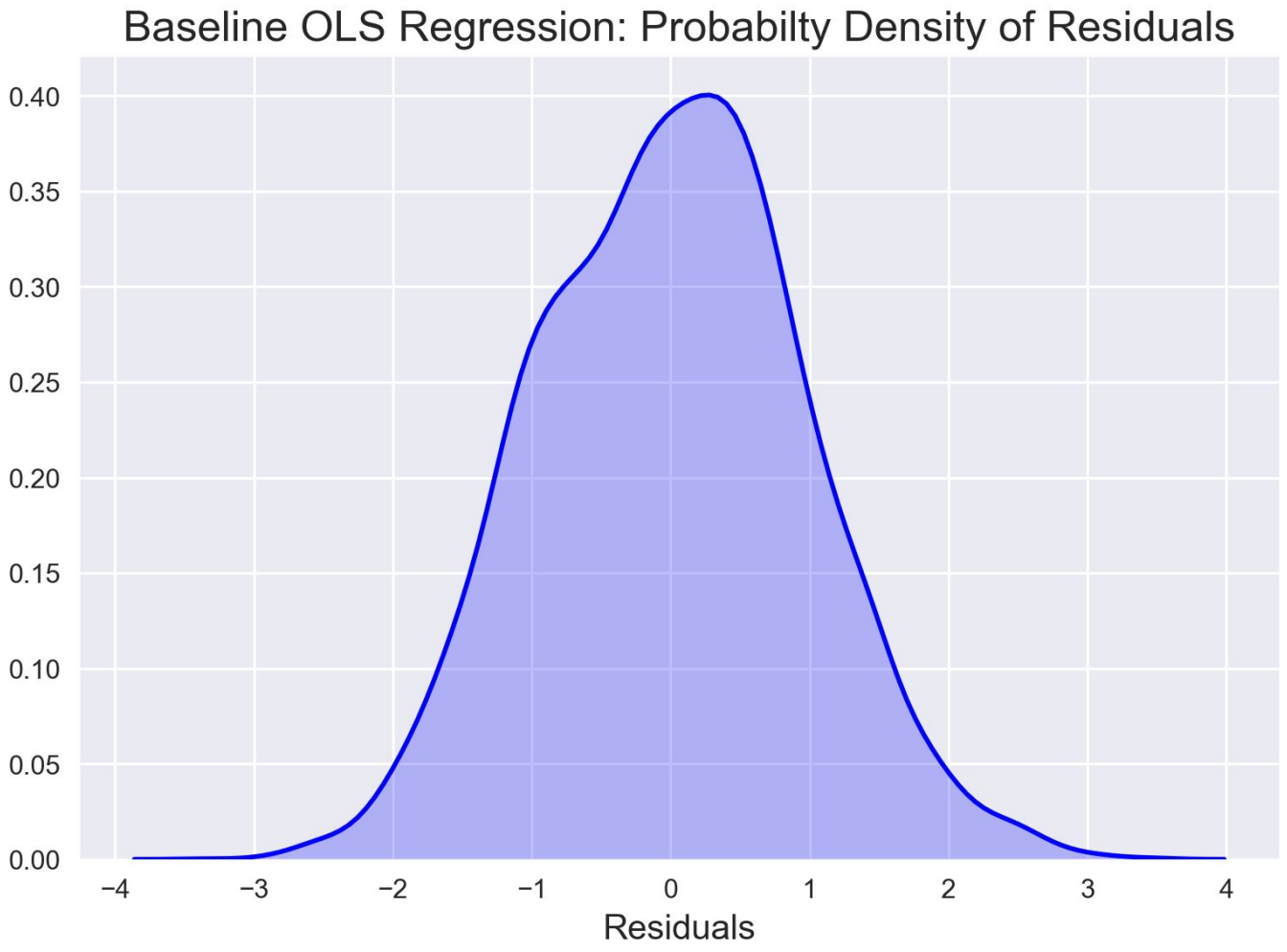
OLS Regression Results

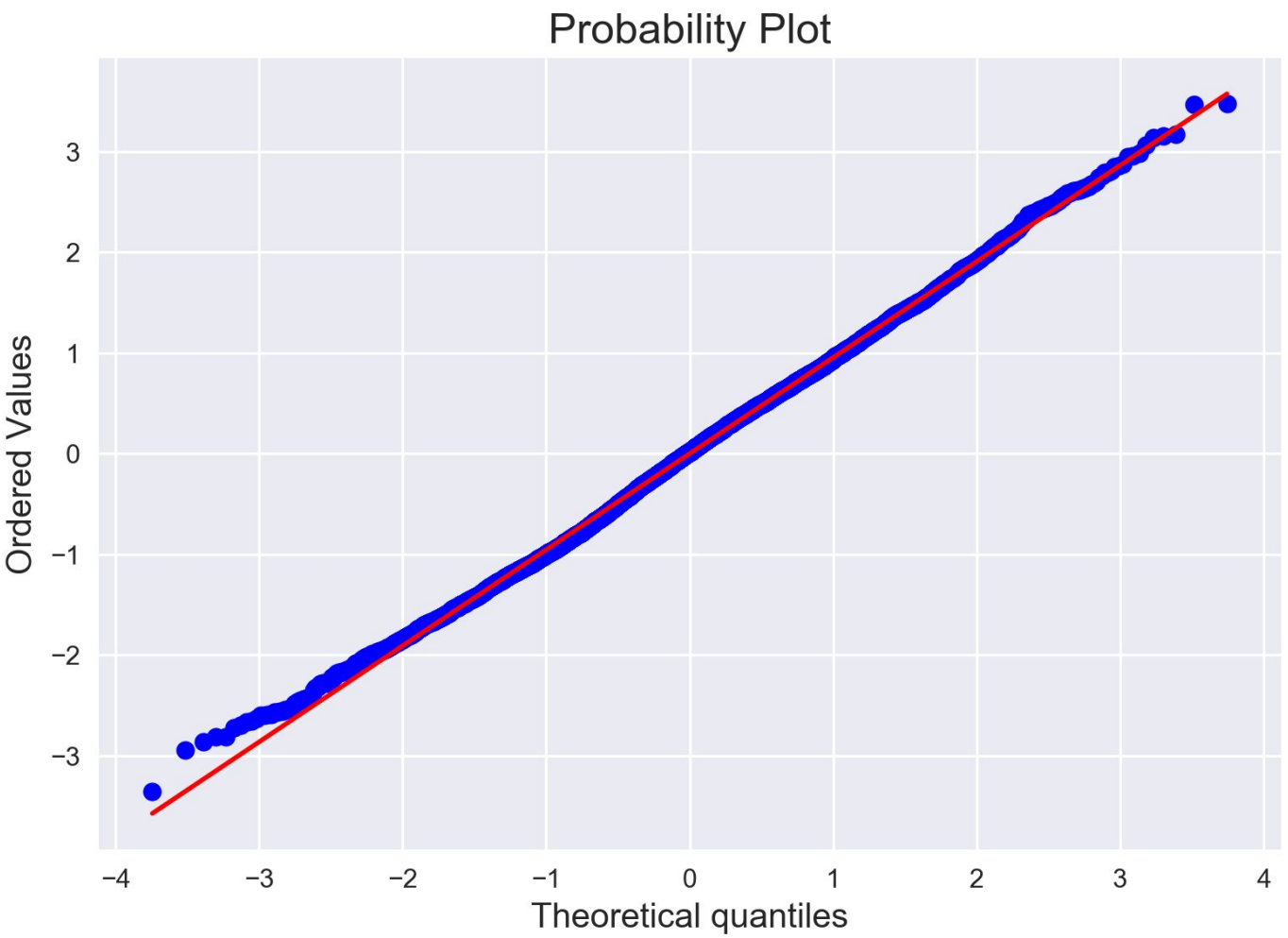
Dep. Variable:	log_salary2016	R-squared:	0.451
Model:	OLS	Adj. R-squared:	0.449
Method:	Least Squares	F-statistic:	367.6
Date:	Sun, 15 Jul 2018	Prob (F-statistic):	0.00
Time:	16:17:10	Log-Likelihood:	-10481.
No. Observations:	7638	AIC:	2.100e+04
Df Residuals:	7620	BIC:	2.112e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
G_t_1	-0.0126	0.001	-14.611	0.000	-0.014	-0.011
AB_t_1	0.0044	0.000	9.790	0.000	0.004	0.005
R_t_1	0.0014	0.002	0.801	0.423	-0.002	0.005
H_t_1	0.0035	0.001	2.383	0.017	0.001	0.006
2B_t_1	-0.0005	0.002	-0.198	0.843	-0.005	0.004
3B_t_1	-0.0427	0.006	-6.689	0.000	-0.055	-0.030
HR_t_1	0.0113	0.004	3.183	0.001	0.004	0.018
RBI_t_1	0.0003	0.002	0.211	0.833	-0.003	0.003
SB_t_1	0.0081	0.002	4.283	0.000	0.004	0.012
CS_t_1	-0.0401	0.005	-7.662	0.000	-0.050	-0.030
BB_t_1	0.0109	0.001	11.526	0.000	0.009	0.013
SO_t_1	-0.0021	0.001	-3.330	0.001	-0.003	-0.001
IBB_t_1	0.0109	0.003	3.206	0.001	0.004	0.018
HBP_t_1	0.0167	0.004	4.709	0.000	0.010	0.024
SH_t_1	-0.0309	0.004	-6.932	0.000	-0.040	-0.022
SF_t_1	0.0029	0.006	0.461	0.645	-0.010	0.015
GIDP_t_1	0.0160	0.003	4.911	0.000	0.010	0.022
constant	13.2443	0.032	416.150	0.000	13.182	13.307

Omnibus:	14.287	Durbin-Watson:	2.006
Prob(Omnibus):	0.001	Jarque-Bera (JB):	12.921
Skew:	0.059	Prob(JB):	0.00156
Kurtosis:	2.836	Cond. No.	1.27e+03


Baseline OLS Regression: Residual Plot
Residuals
3
2
1
0
−1
−2
−3
13
14
15
16
17
18
Fitted Values (log(Salary))


Baseline OLS Regression: Probabilty Density of Residuals
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
Residuals


Probability Plot
Ordered Values
Theoretical quantiles
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

79%

Of the total variation in salary is explained by adding in the engineered features, lagged one year.

Feature Engineered OLS

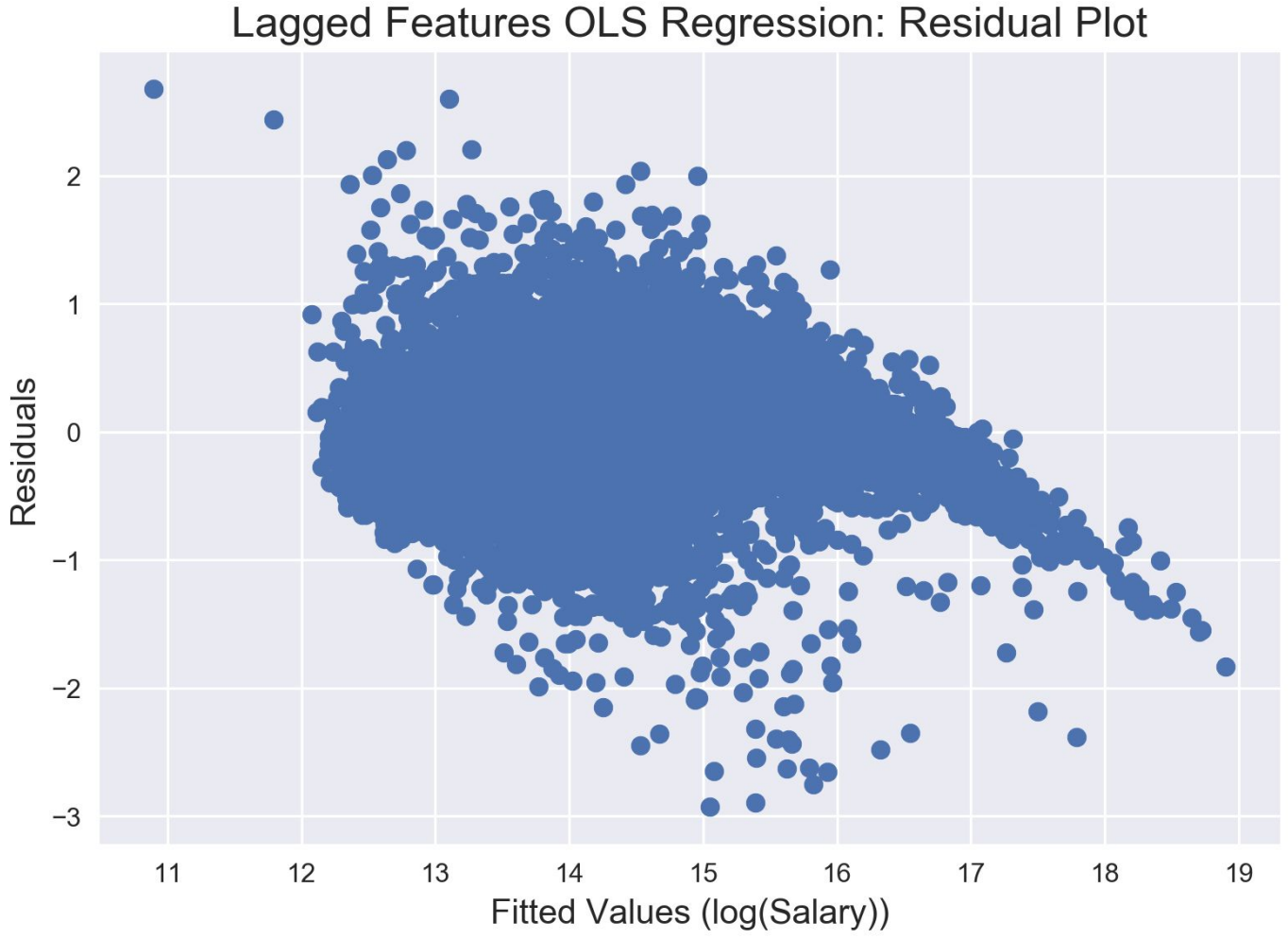
- Much better!

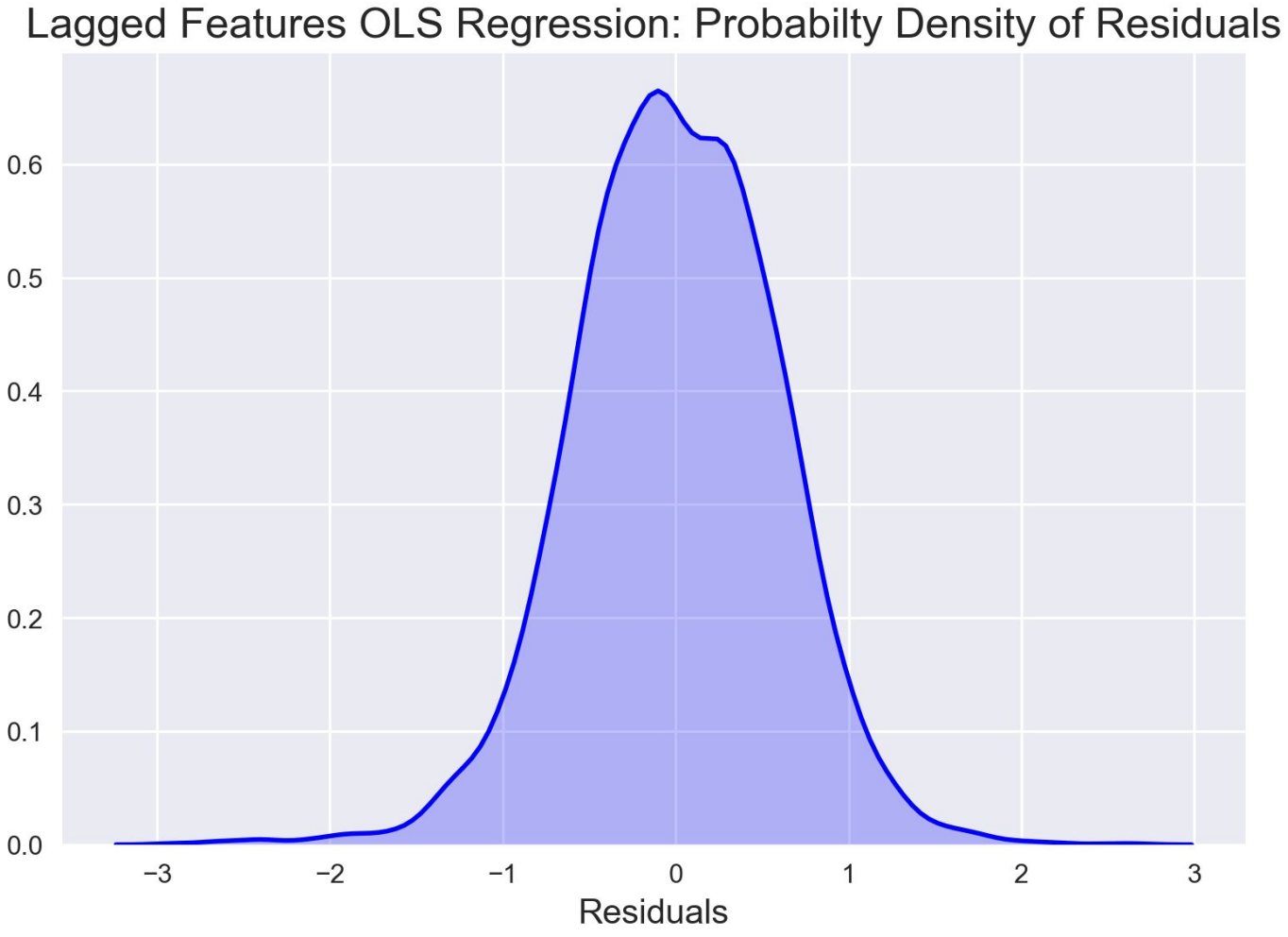
OLS Regression Results

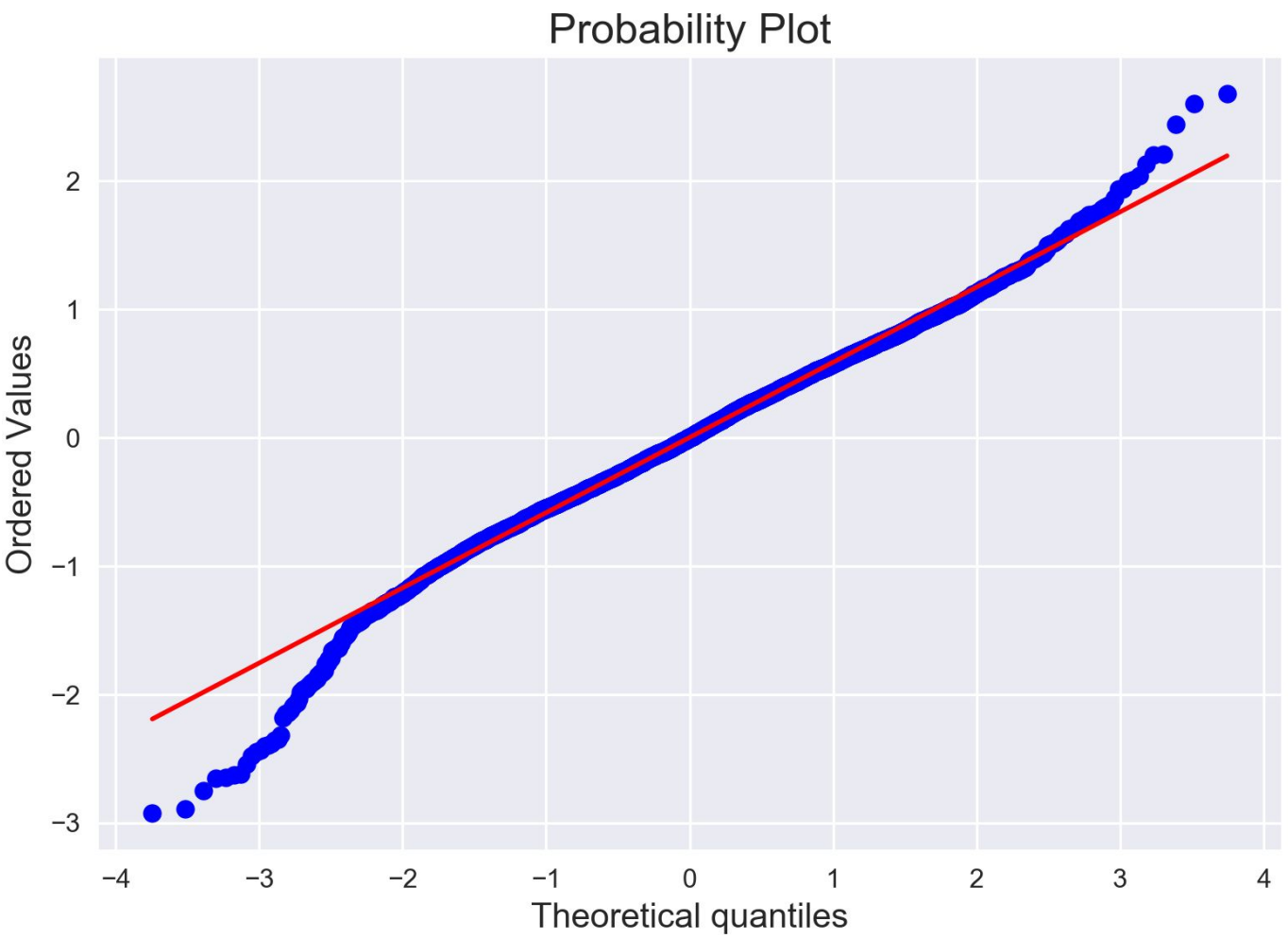
Dep. Variable:	log_salary2016	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.791
Method:	Least Squares	F-statistic:	1113.
Date:	Sun, 15 Jul 2018	Prob (F-statistic):	0.00
Time:	16:17:12	Log-Likelihood:	-6776.0
No. Observations:	7638	AIC:	1.361e+04
Df Residuals:	7611	BIC:	1.379e+04
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sal_t_1	1.05e-07	2.37e-09	44.249	0.000	1e-07	1.1e-07
G_t_1	-0.0058	0.001	-10.515	0.000	-0.007	-0.005
AB_t_1	0.0022	0.000	5.219	0.000	0.001	0.003
R_t_1	-0.0015	0.001	-1.374	0.170	-0.004	0.001
H_t_1	0.0032	0.001	2.489	0.013	0.001	0.006
2B_t_1	-0.0021	0.002	-1.383	0.167	-0.005	0.001
3B_t_1	-0.0020	0.004	-0.518	0.605	-0.010	0.006
HR_t_1	0.0439	0.008	5.244	0.000	0.028	0.060
RBI_t_1	0.0019	0.001	1.910	0.056	-4.95e-05	0.004
SB_t_1	0.0025	0.001	2.116	0.034	0.000	0.005
CS_t_1	-0.0012	0.003	-0.376	0.707	-0.008	0.005
BB_t_1	0.0063	0.001	6.727	0.000	0.004	0.008
SO_t_1	-0.0018	0.000	-4.498	0.000	-0.003	-0.001
IBB_t_1	0.0085	0.002	3.740	0.000	0.004	0.013
HBP_t_1	-0.0001	0.002	-0.060	0.952	-0.005	0.004
SH_t_1	-0.0050	0.003	-1.778	0.075	-0.011	0.001
SF_t_1	-0.0038	0.004	-0.952	0.341	-0.012	0.004
GIDP_t_1	0.0020	0.002	1.012	0.312	-0.002	0.006
AVG_t_1	-4.023e-06	0.000	-0.010	0.992	-0.001	0.001
OBP_t_1	0.0005	0.000	1.350	0.177	-0.000	0.001
EXP	0.3668	0.013	28.829	0.000	0.342	0.392
EXP_SQ	-0.0187	0.000	-45.660	0.000	-0.019	-0.018
allStar_t_1	0.1101	0.026	4.262	0.000	0.059	0.161
EXP_OBP_t_1	5.635e-05	3.68e-05	1.530	0.126	-1.59e-05	0.000
OBP_HR_t_1	-0.0001	2.26e-05	-4.518	0.000	-0.000	-5.79e-05
min_salary2016	1.347e-06	5.82e-08	23.135	0.000	1.23e-06	1.46e-06
constant	11.3207	0.078	145.073	0.000	11.168	11.474

Omnibus:	181.978	Durbin-Watson:	2.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	318.479
Skew:	-0.200	Prob(JB):	6.97e-70
Kurtosis:	3.917	Cond. No.	6.15e+07


Lagged Features OLS Regression: Residual Plot
Residuals
2
1
0
−1
−2
−3
11
12
13
14
15
16
17
18
19
Fitted Values (log(Salary))

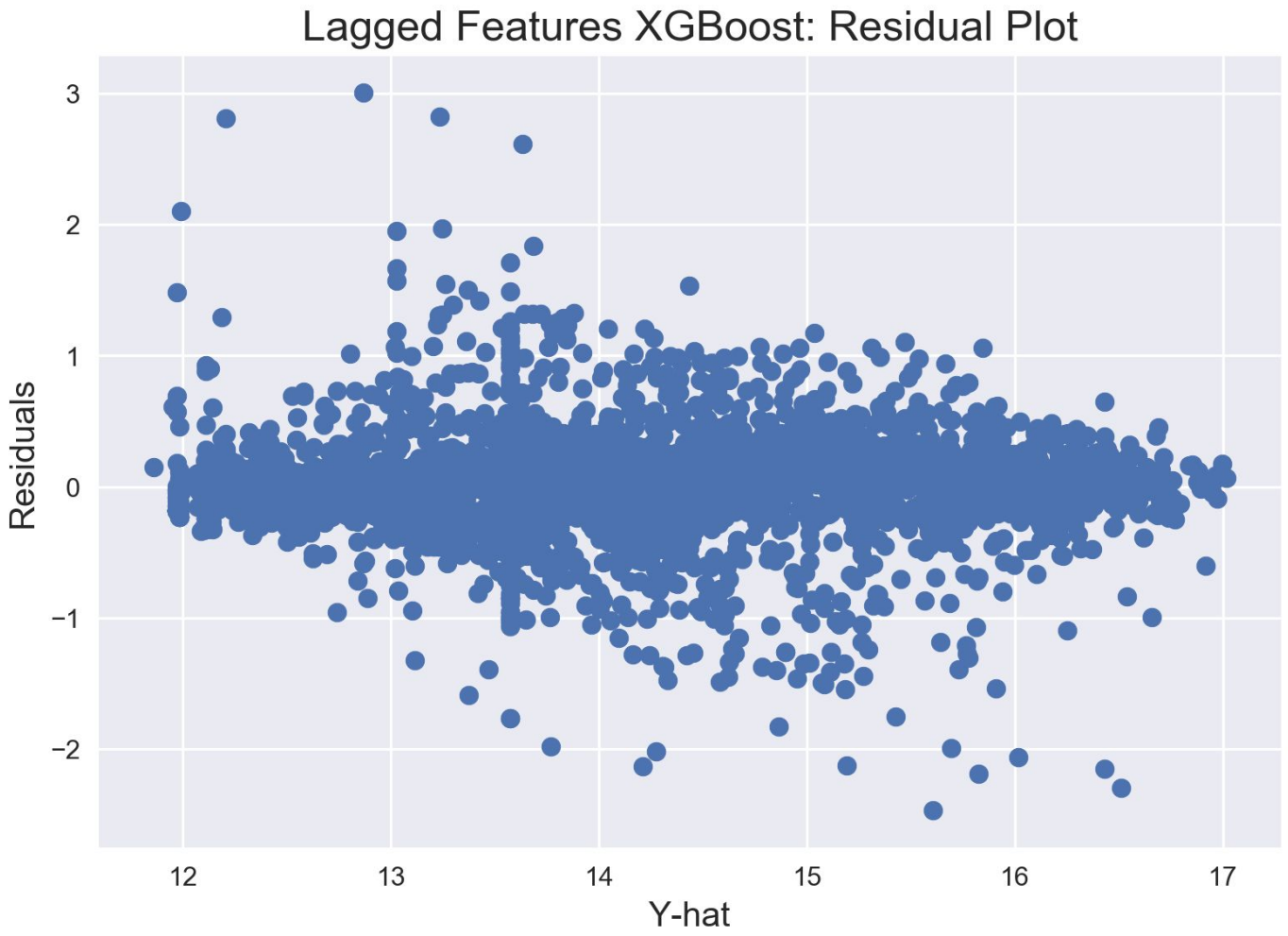

Lagged Features OLS Regression: Probabilty Density of Residuals
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
Residuals

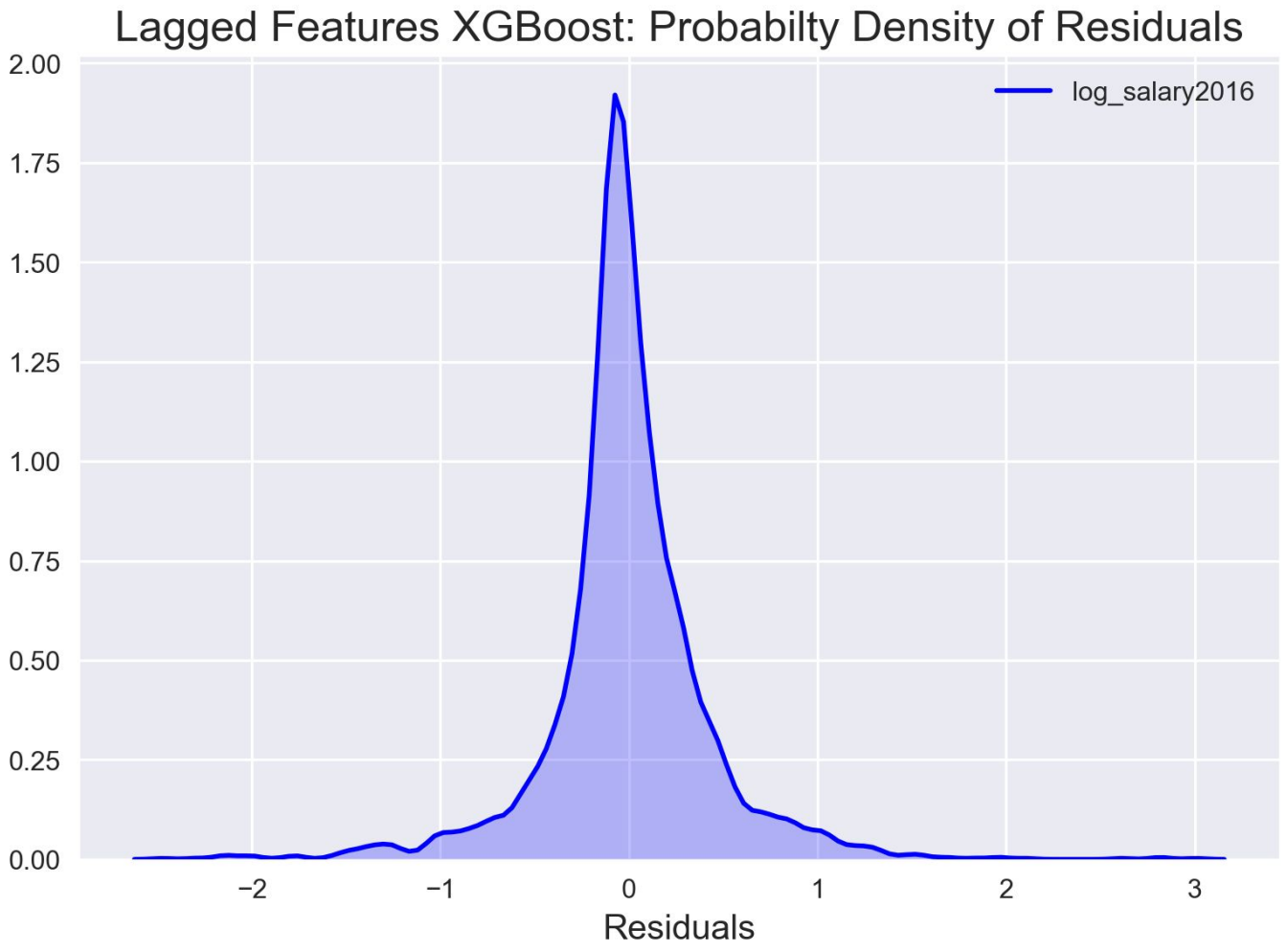

Probability Plot
Ordered Values
Theoretical quantiles
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

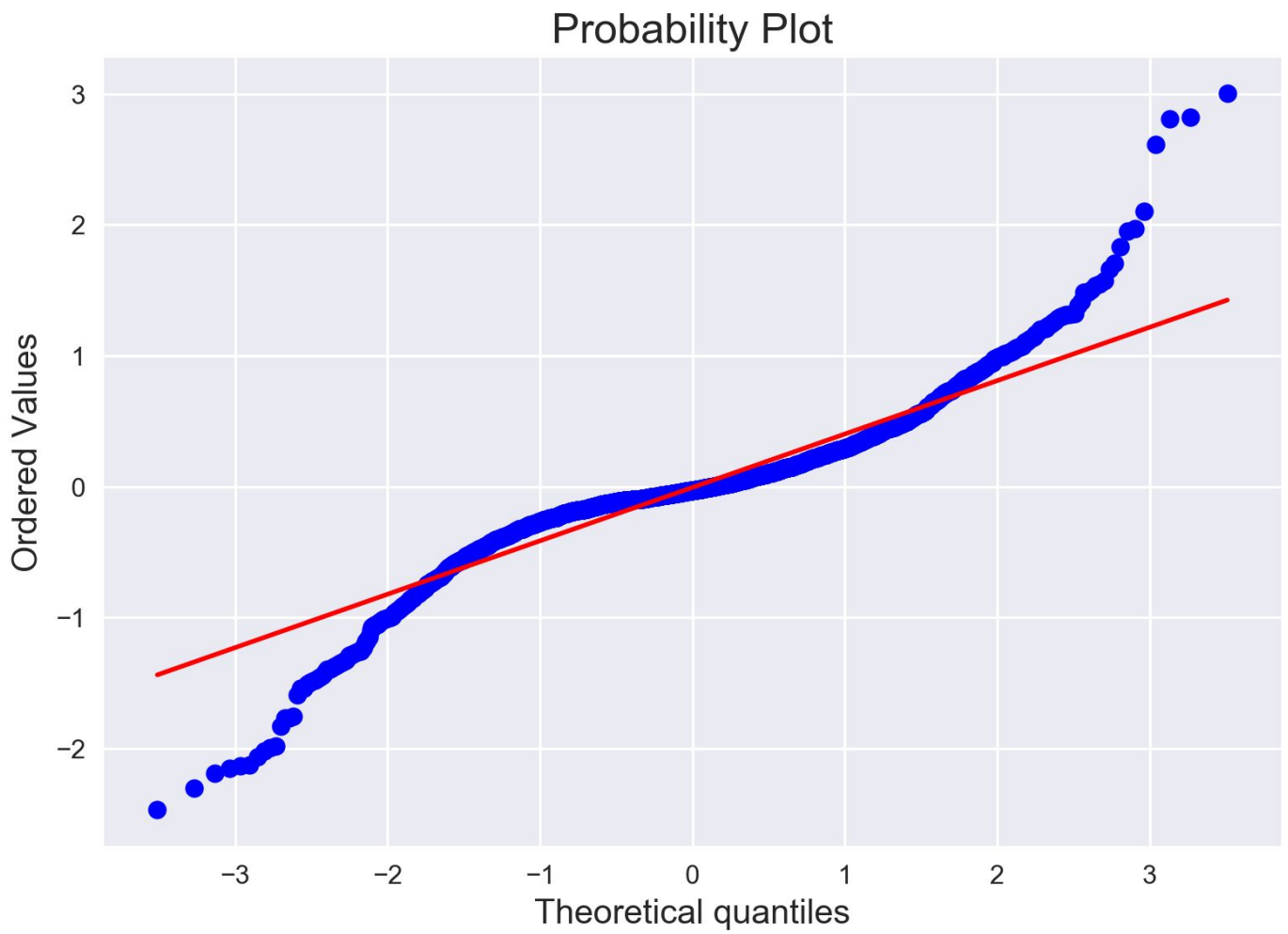
XGBoost Models

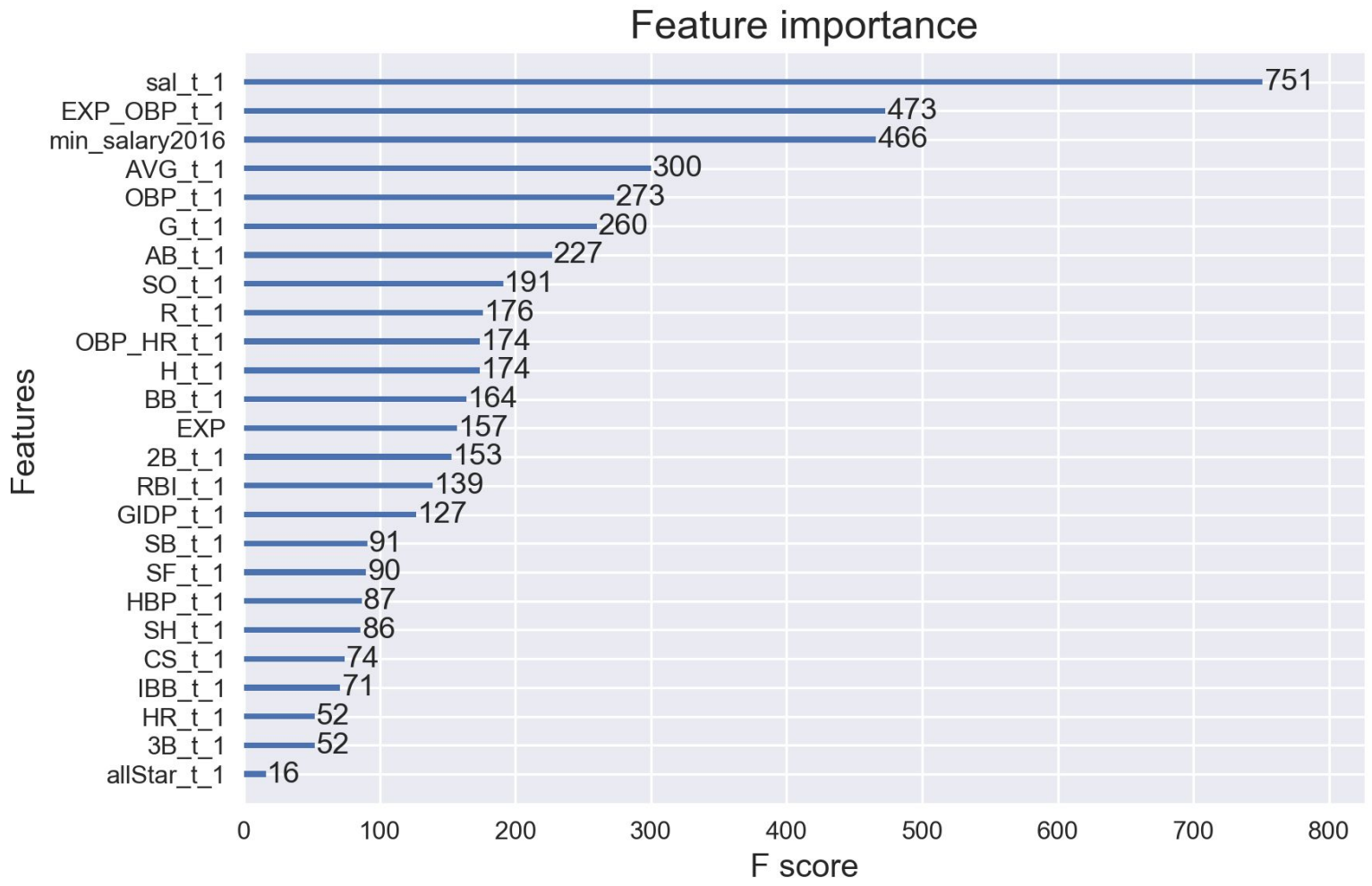
89%

Of the total variation in salary is explained by using the engineered features, lagged one year.


Lagged Features XGBoost: Residual Plot
Residuals
3
2
1
0
−1
−2
12
13
14
15
16
17
Y-hat


Lagged Features XGBoost: Probabilty Density of Residuals
log_salary2016
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
3
Residuals

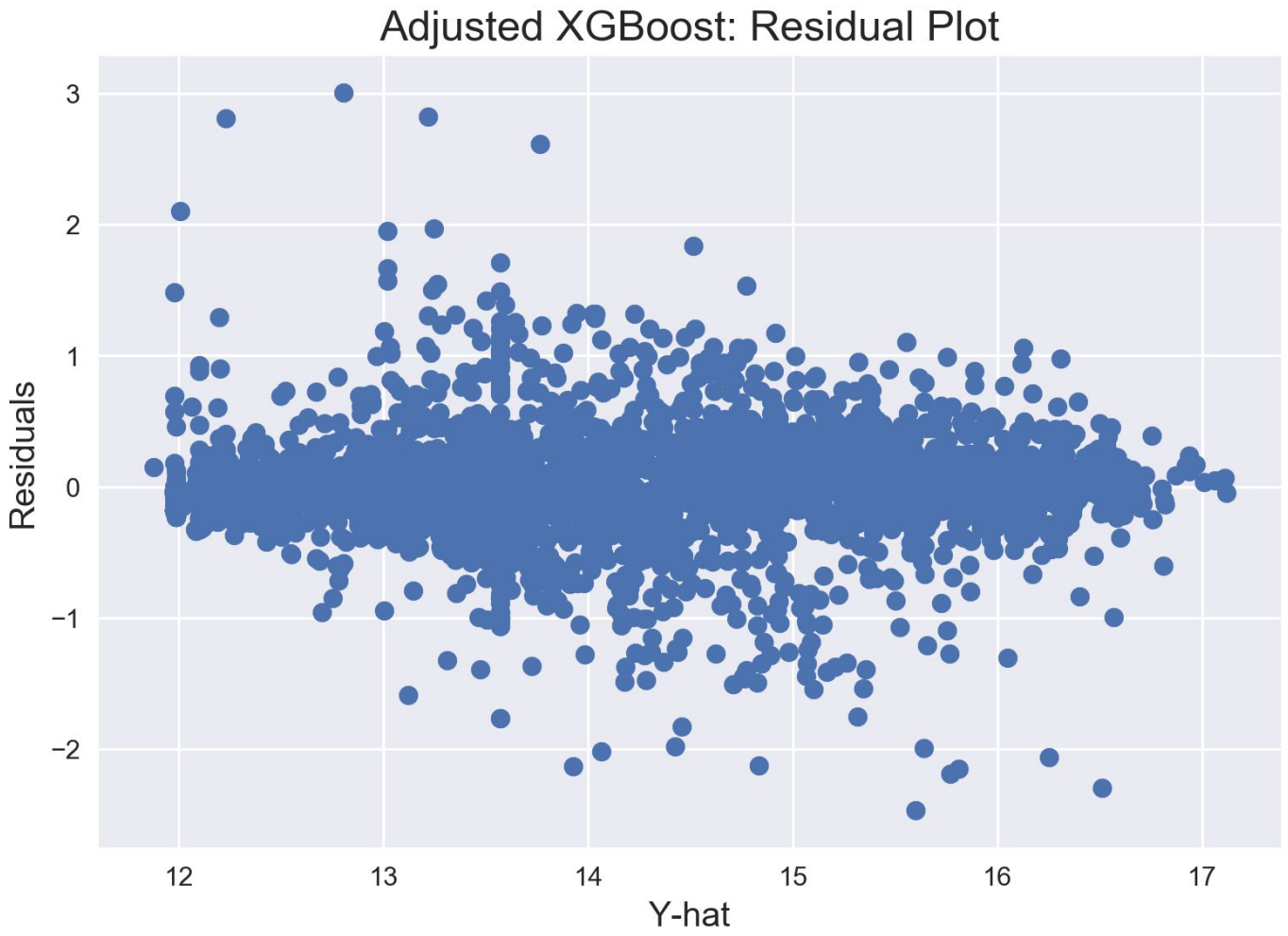

Probability Plot
Ordered Values
Theoretical quantiles
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

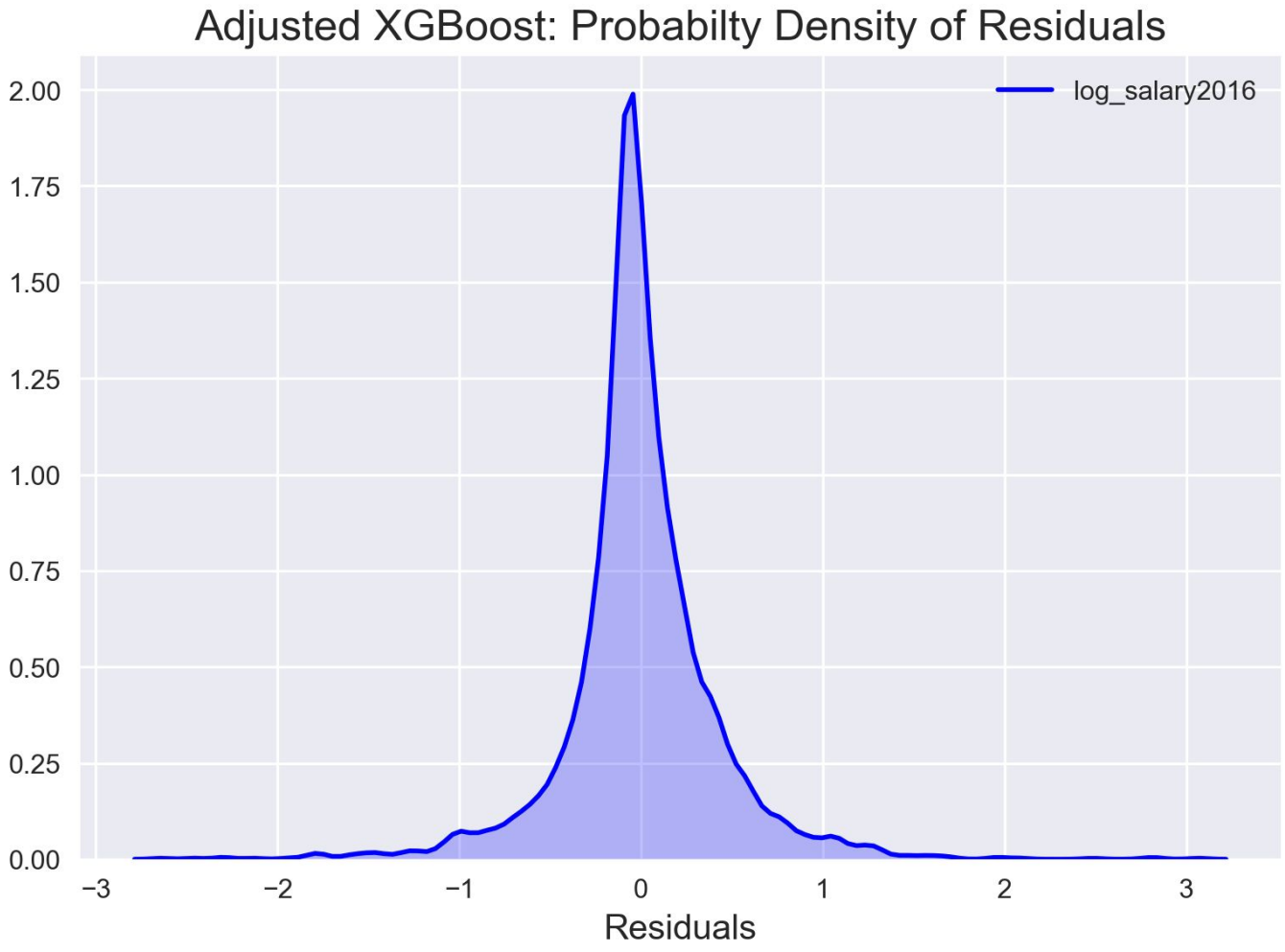

Feature importance
sal_t_1
EXP_OBP_t_1
min_salary2016
AVG_t_1
OBP_t_1
G_t_1
AB_t_1
SO_t_1
R_t_1
OBP_HR_t_1
H_t_1
BB_t_1
EXP
2B_t_1
RBI_t_1
GIDP_t_1
SB_t_1
SF_t_1
HBP_t_1
SH_t_1
CS_t_1
IBB_t_1
HR_t_1
3B_t_1
allStar_t_1
751
473
466
300
273
260
227
191
176
174
174
164
157
153
139
127
91
90
87
86
74
71
52
52
16
Features
0
100
200
300
400
500
600
700
800
F score

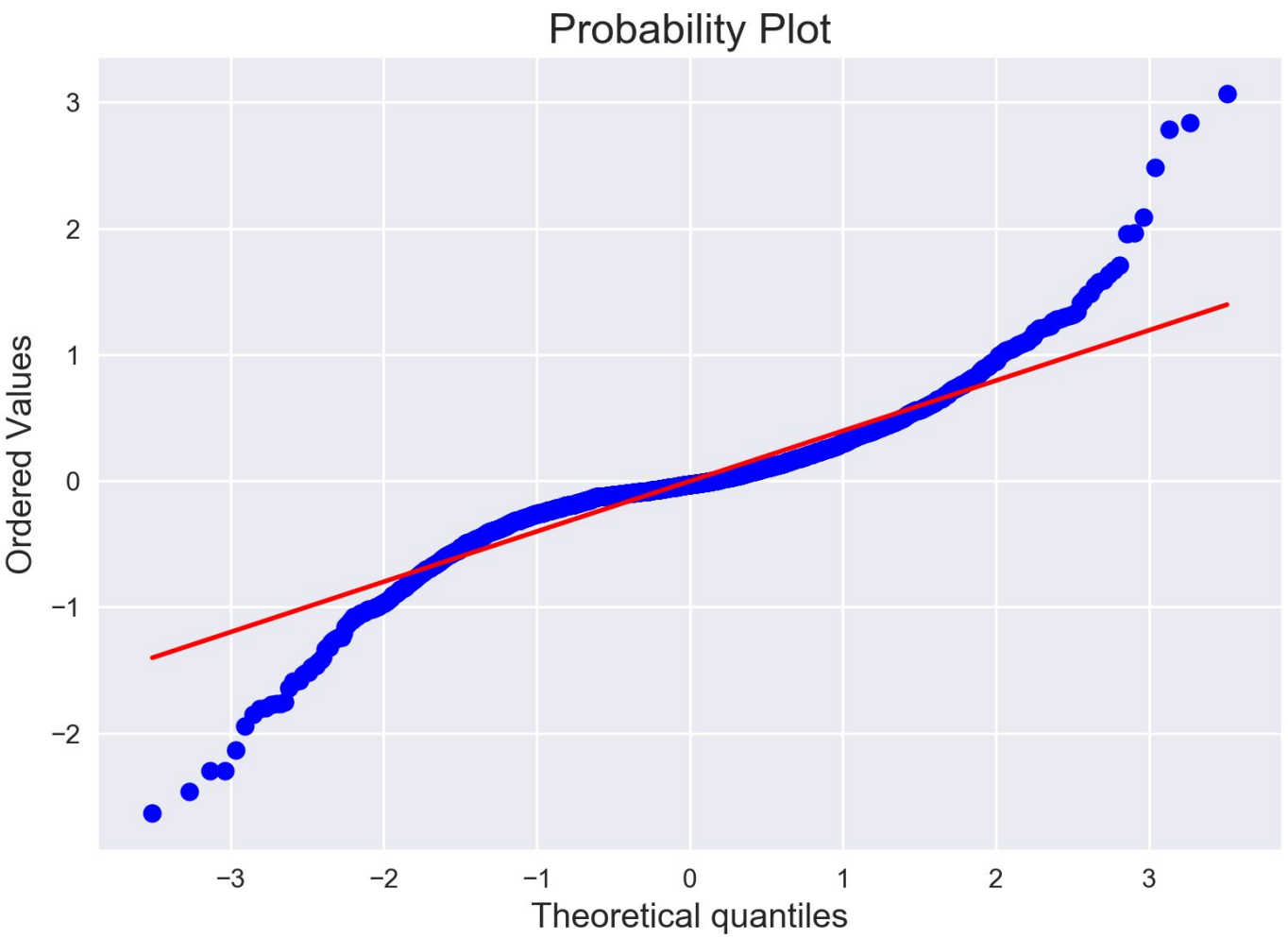
Now let's use the feature importance graph from above to try and increase our model performance. Let's try removing all the features that have a F score of less than 100. Let's also add in two-year lags for the features that have an F score of above 100. Let's see if this makes any difference in our model performance.

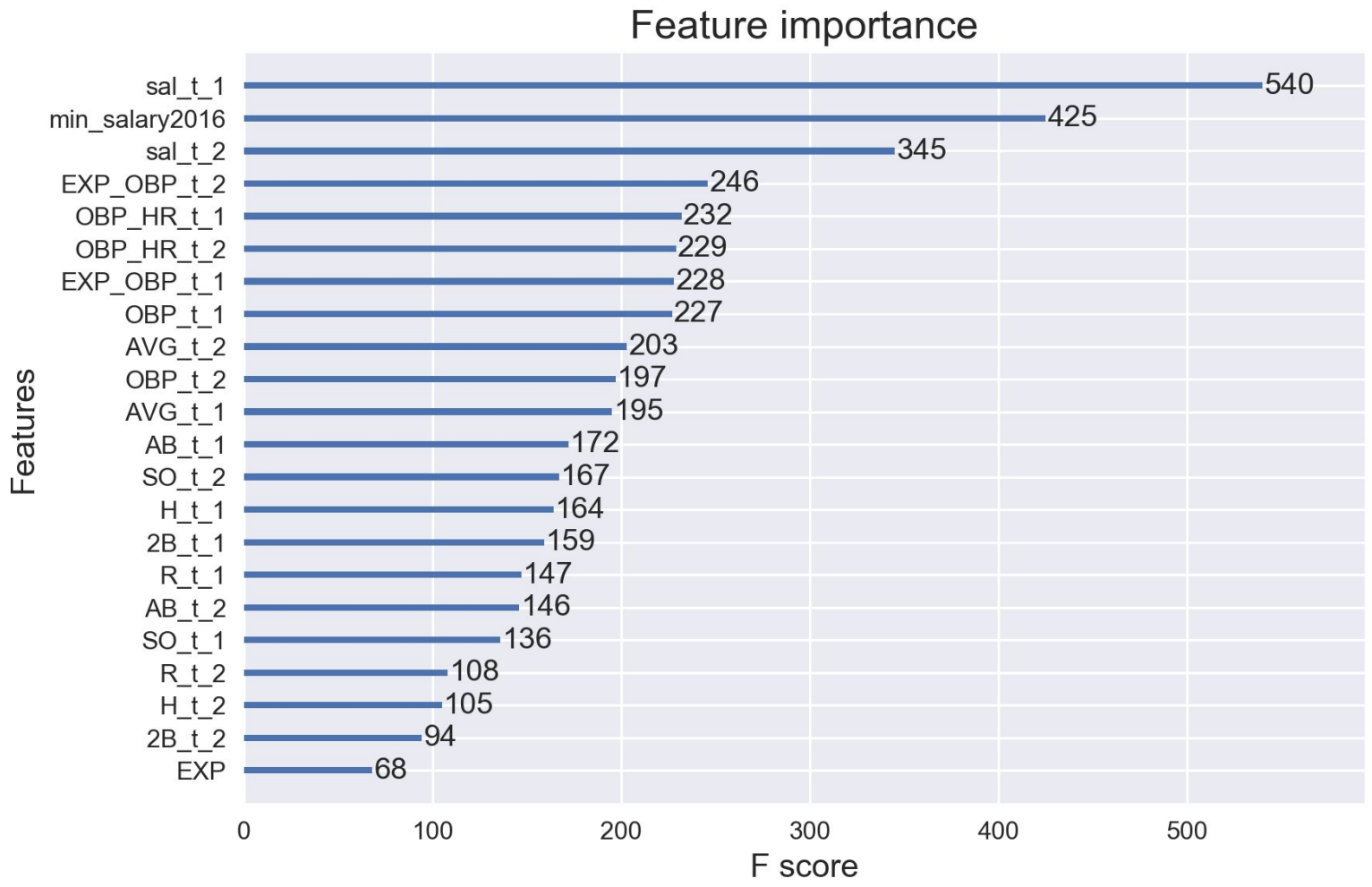
90%

Of the total variation in salary is explained by the adjusted XGBoost model.


Adjusted XGBoost: Residual Plot
Residuals
3
2
1
0
−1
−2
12
13
14
15
16
17
Y-hat


Adjusted XGBoost: Probabilty Density of Residuals
log_salary2016
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−3
−2
−1
0
1
2
3
Residuals


Probability Plot
Ordered Values
Theoretical quantiles
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


Feature importance
sal_t_1
min_salary2016
sal_t_2
EXP_OBP_t_2
OBP_HR_t_1
OBP_HR_t_2
EXP_OBP_t_1
OBP_t_1
AVG_t_2
OBP_t_2
AVG_t_1
AB_t_1
SO_t_2
H_t_1
2B_t_1
R_t_1
AB_t_2
SO_t_1
R_t_2
H_t_2
2B_t_2
EXP
540
425
345
246
232
229
228
227
203
197
195
172
167
164
159
147
146
136
108
105
94
68
Features
0
100
200
300
400
500
F score

The most important predictors...

Lagged Salary
Current year minimum salary
Lagged interaction of experience and on-base %
Lagged Interaction of home runs and on-base %
Lagged on-base %
Lagged batting average

Thanks!

Contact me:

jeffreywhoffman@gmail.com