Web Scraping and Linear Regression Project

Identify relationship between stock prices and financial statistics

Kristy Yang 2021 Oct

Introduction

- Motivation: The goal of this project is to find out how much a public company’s stock price can be explained by its financial numbers. The research will help potential investors, traders, and whoever interested to make better informed decisions based on the contents the project offers.

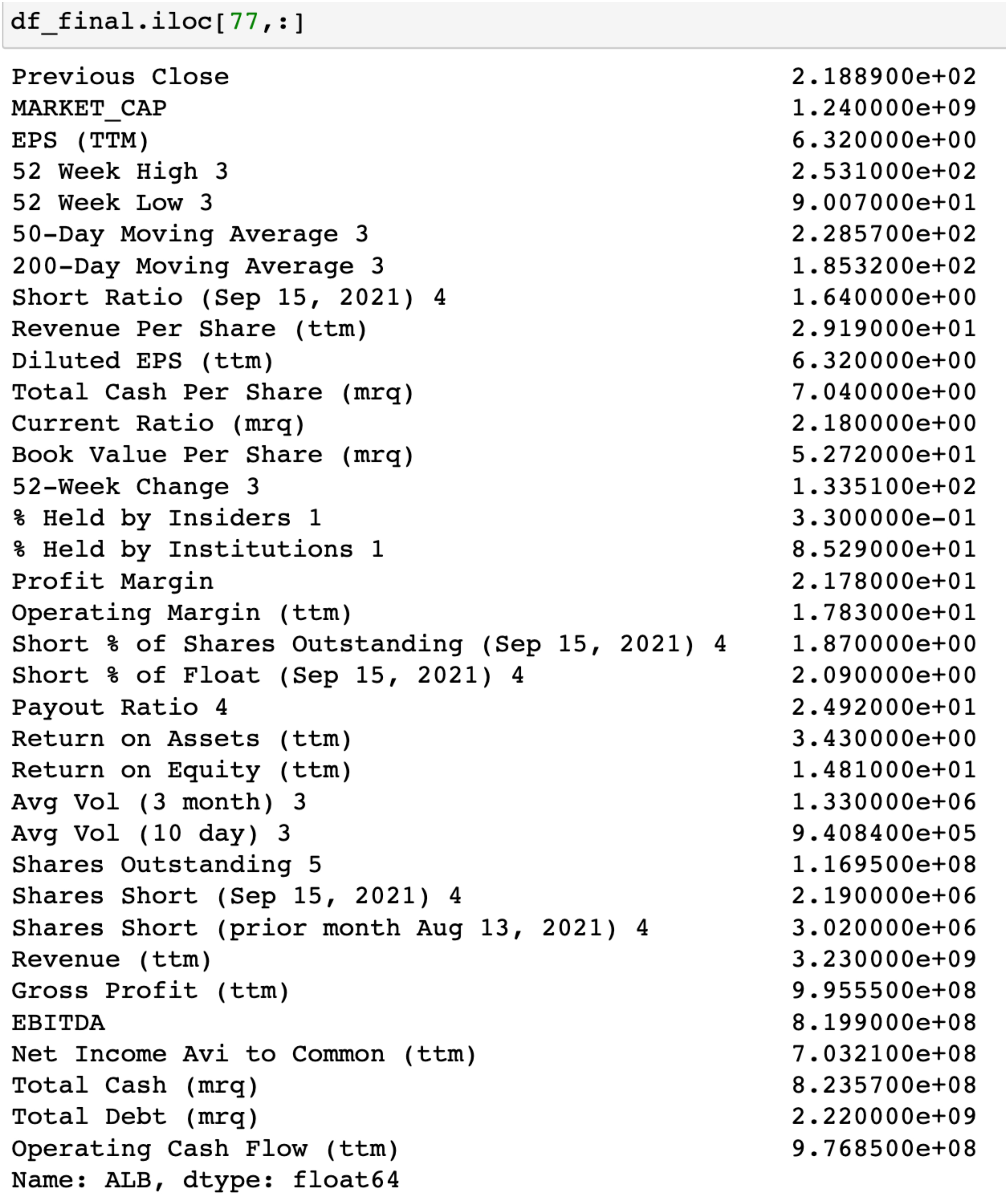
- Objectives and goals: To build up a regression model to explain the relation between stock price and financial statistics of a company

- Final result: Throught operating linear regression and related regulation / future engineering skills, an ElasticNet model is obtained as our final model.

Methodology

Data

- Data scraped from website https://finance.yahoo.com and https://stockanalysis.com/stocks/.
- Close price and 69 financials statistic numbers obtained for 3057 companies
- After cleaning and dropping N/As, a dataset with shape (1453, 35) is for our analysis. Each row represents a company.



```
df_final.iloc[77,:]
```

```
Previous Close                                    2.188900e+02
MARKET_CAP                                        1.240000e+09
EPS (TTM)                                         6.320000e+00
52 Week High 3                                    2.531000e+02
52 Week Low 3                                     9.007000e+01
50-Day Moving Average 3                           2.285700e+02
200-Day Moving Average 3                          1.853200e+02
Short Ratio (Sep 15, 2021) 4                      1.640000e+00
Revenue Per Share (ttm)                           2.919000e+01
Diluted EPS (ttm)                                 6.320000e+00
Total Cash Per Share (mrq)                        7.040000e+00
Current Ratio (mrq)                               2.180000e+00
Book Value Per Share (mrq)                        5.272000e+01
52-Week Change 3                                  1.335100e+02
% Held by Insiders 1                              3.300000e-01
% Held by Institutions 1                          8.529000e+01
Profit Margin                                     2.178000e+01
Operating Margin (ttm)                            1.783000e+01
Short % of Shares Outstanding (Sep 15, 2021) 4    1.870000e+00
Short % of Float (Sep 15, 2021) 4                 2.090000e+00
Payout Ratio 4                                    2.492000e+01
Return on Assets (ttm)                            3.430000e+00
Return on Equity (ttm)                            1.481000e+01
Avg Vol (3 month) 3                               1.330000e+06
Avg Vol (10 day) 3                                9.408400e+05
Shares Outstanding 5                              1.169500e+08
Shares Short (Sep 15, 2021) 4                     2.190000e+06
Shares Short (prior month Aug 13, 2021) 4         3.020000e+06
Revenue (ttm)                                     3.230000e+09
Gross Profit (ttm)                                9.955500e+08
EBITDA                                            8.199000e+08
Net Income Avi to Common (ttm)                    7.032100e+08
Total Cash (mrq)                                  8.235700e+08
Total Debt (mrq)                                  2.220000e+09
Operating Cash Flow (ttm)                         9.768500e+08
Name: ALB, dtype: float64
```

Methodology

Tools

- request, selenium for data web scraping ;
- BeautifulSoup for HTML syntax parsing ;
- pandas and numpy for data manipulation ;
- Matplotlib and Seaborn for data visualization;
- Pickle for data object serialization and de-serialization
- sklearn and statsmodels for regression

Metric:

Choose the model with highest Mean R squared value with relatively mild complexity

Methodology

Work flow and Algorithm

1.Clean data: check/transfer data types, deal with missing value.

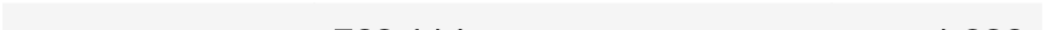
2. Create a baseline OLS linear regression model on all features.

3. Check assumptions and evaluation: get rid of outliers, check VIF and move away features with multi-collinearity, remove non-significant features, and transform dependent feature with log function to correct residual heteroskedasticity.

4. Also Compare R squared values for regressions with / without historical price feature

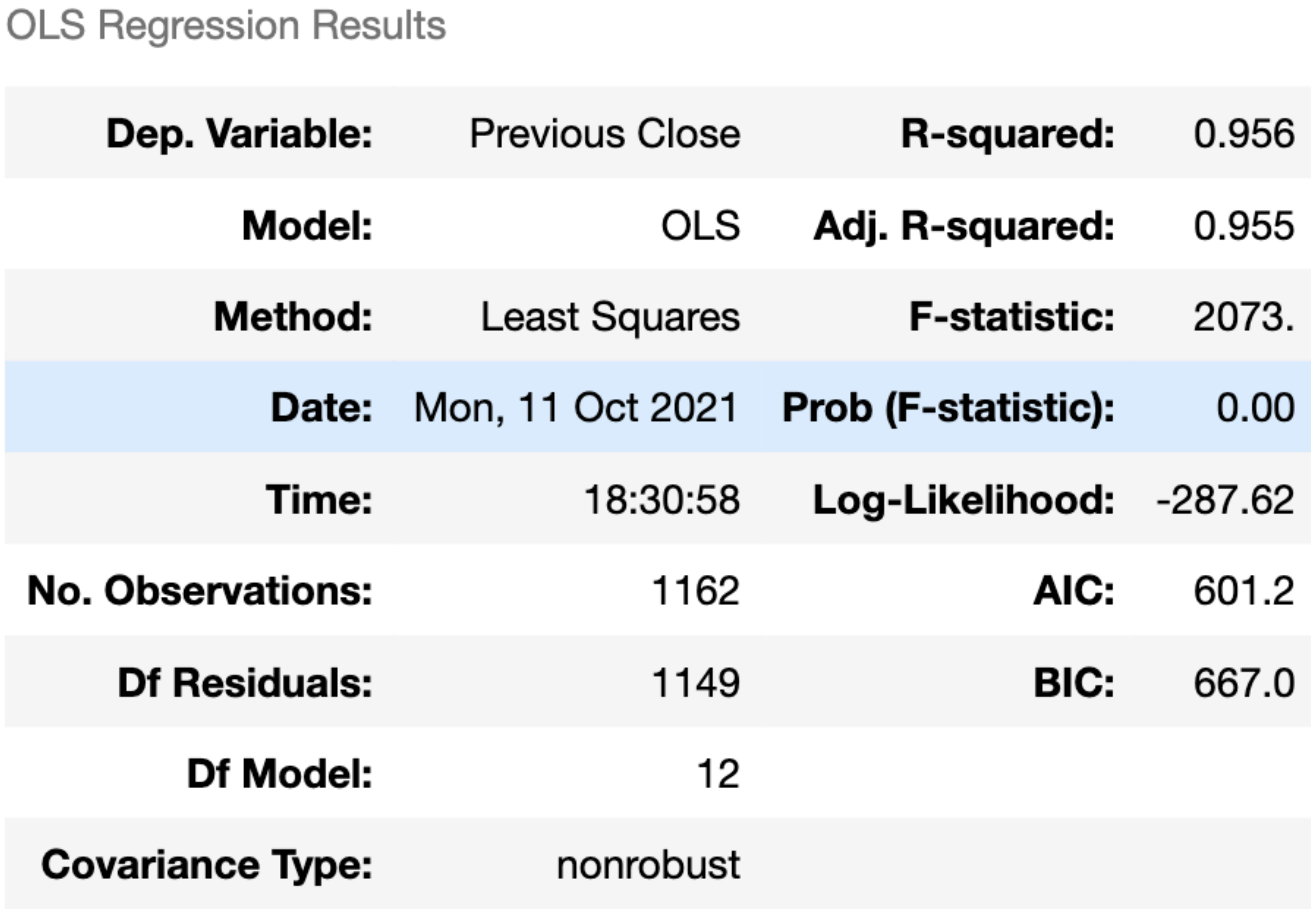
5. Add Polynomial features and try different regulation methods with cross validation: Lasso, Ridge and ElasticNet, and choose the model with best mean R squared value.

6. Try to add available dummy variable "Industry" to improve model.



Result

1: Linear regression with historical price related feature: “52 Week High 3”



OLS Regression Results

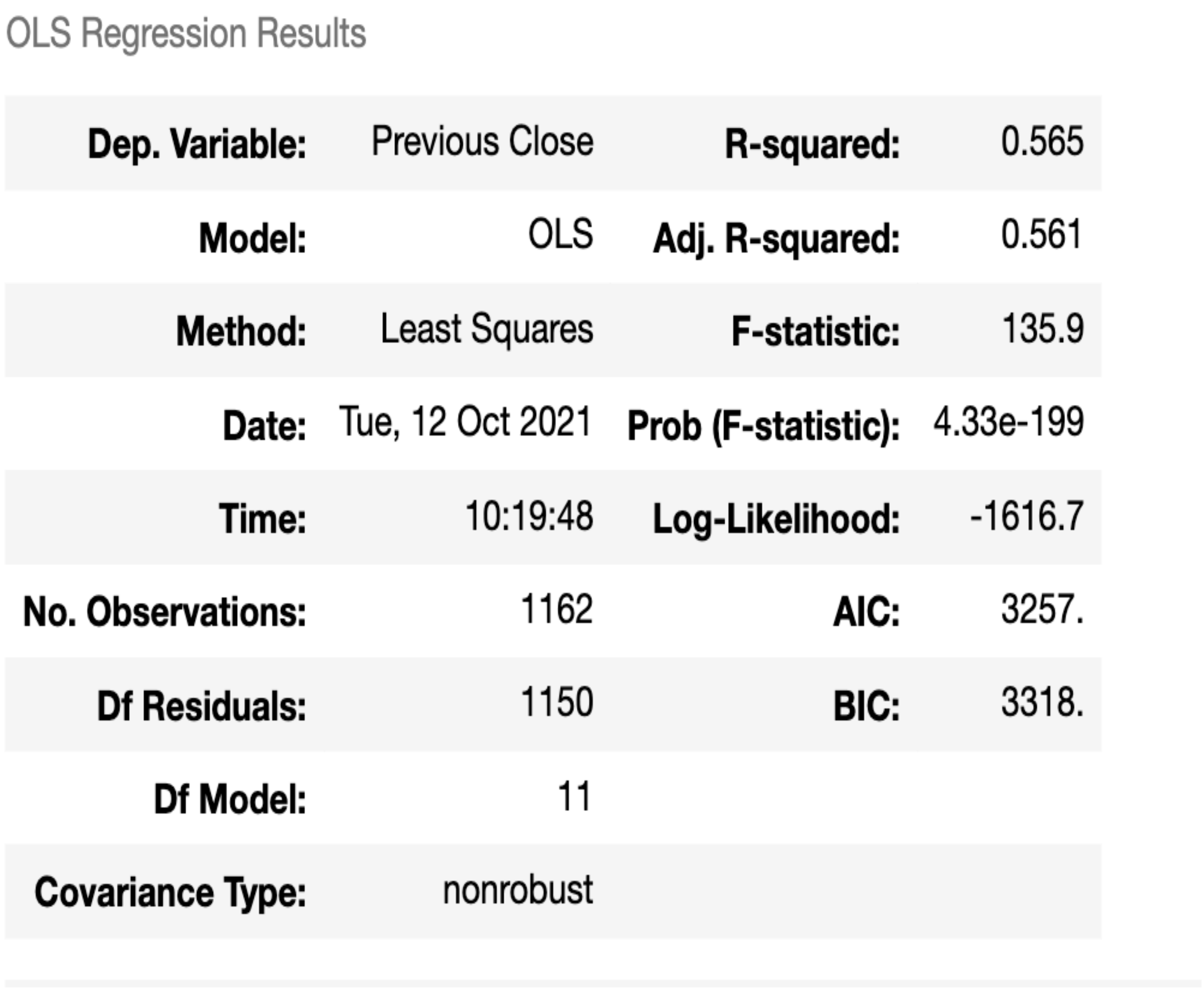
Dep. Variable:	Previous Close	**R-squared:**	0.956
Model:	OLS	**Adj. R-squared:**	0.955
Method:	Least Squares	**F-statistic:**	2073.
Date:	Mon, 11 Oct 2021	**Prob (F-statistic):**	0.00
Time:	18:30:58	**Log-Likelihood:**	-287.62
No. Observations:	1162	**AIC:**	601.2
Df Residuals:	1149	**BIC:**	667.0
Df Model:	12		
Covariance Type:	nonrobust		

Insights:

The R-squared and Adj. R-sqaured are both terrifically high, it is caused by feature “52 Week High 3”. “52 Week High 3” is defined as the highest price at which a security, such as a stock, has traded during previous 52 weeks. It is an important factor in the analysis of a stock's current value, so it is very reasonable to have a high R squared value with “52 Week High 3” as one of our feature.

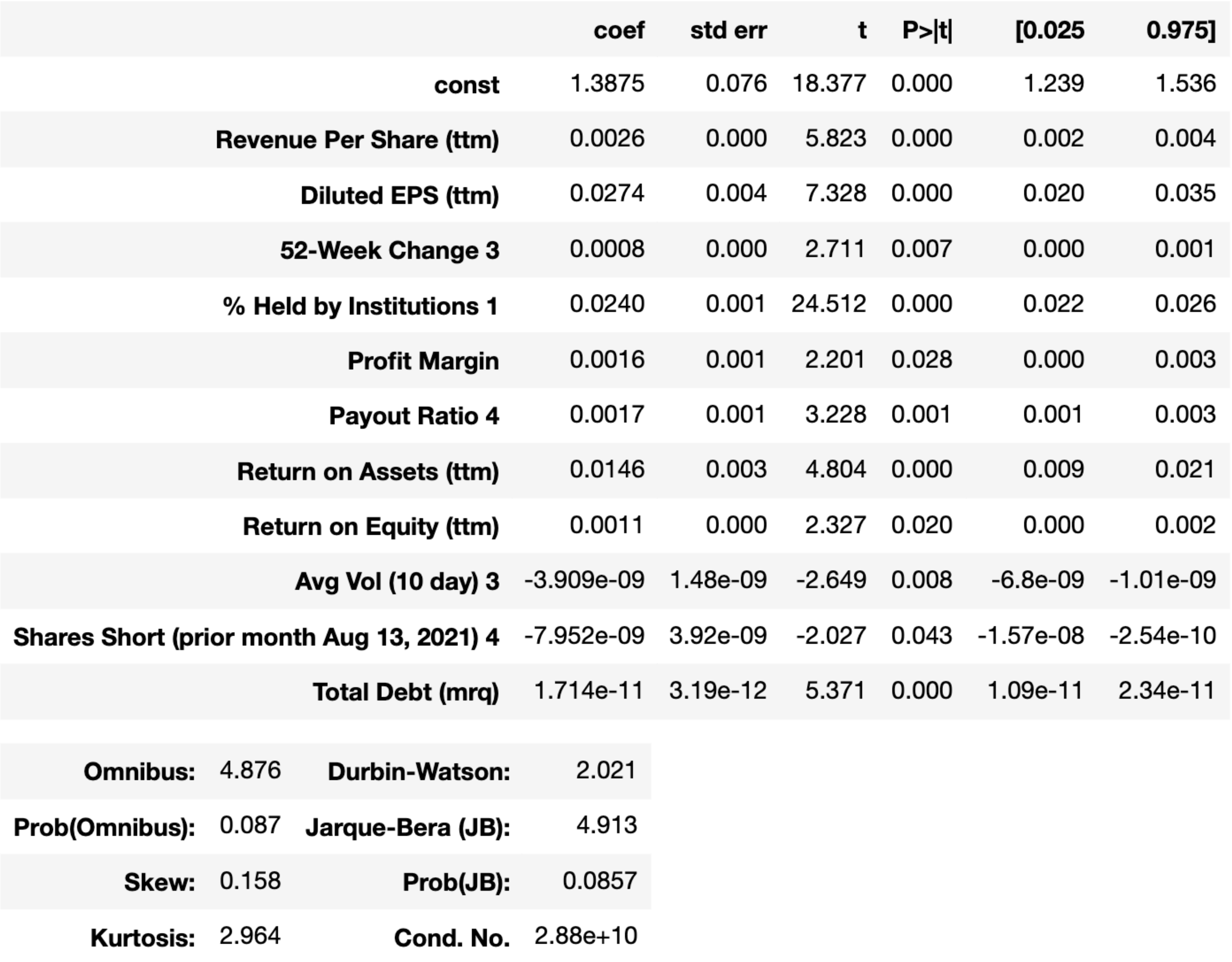
Result

2. Linear regression without historical price related feature.



OLS Regression Results

Dep. Variable:	Previous Close	**R-squared:**	0.565
Model:	OLS	**Adj. R-squared:**	0.561
Method:	Least Squares	**F-statistic:**	135.9
Date:	Tue, 12 Oct 2021	**Prob (F-statistic):**	4.33e-199
Time:	10:19:48	**Log-Likelihood:**	-1616.7
No. Observations:	1162	**AIC:**	3257.
Df Residuals:	1150	**BIC:**	3318.
Df Model:	11		
Covariance Type:	nonrobust		



	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3875	0.076	18.377	0.000	1.239	1.536
Revenue Per Share (ttm)	0.0026	0.000	5.823	0.000	0.002	0.004
Diluted EPS (ttm)	0.0274	0.004	7.328	0.000	0.020	0.035
52-Week Change 3	0.0008	0.000	2.711	0.007	0.000	0.001
% Held by Institutions 1	0.0240	0.001	24.512	0.000	0.022	0.026
Profit Margin	0.0016	0.001	2.201	0.028	0.000	0.003
Payout Ratio 4	0.0017	0.001	3.228	0.001	0.001	0.003
Return on Assets (ttm)	0.0146	0.003	4.804	0.000	0.009	0.021
Return on Equity (ttm)	0.0011	0.000	2.327	0.020	0.000	0.002
Avg Vol (10 day) 3	-3.909e-09	1.48e-09	-2.649	0.008	-6.8e-09	-1.01e-09
Shares Short (prior month Aug 13, 2021) 4	-7.952e-09	3.92e-09	-2.027	0.043	-1.57e-08	-2.54e-10
Total Debt (mrq)	1.714e-11	3.19e-12	5.371	0.000	1.09e-11	2.34e-11

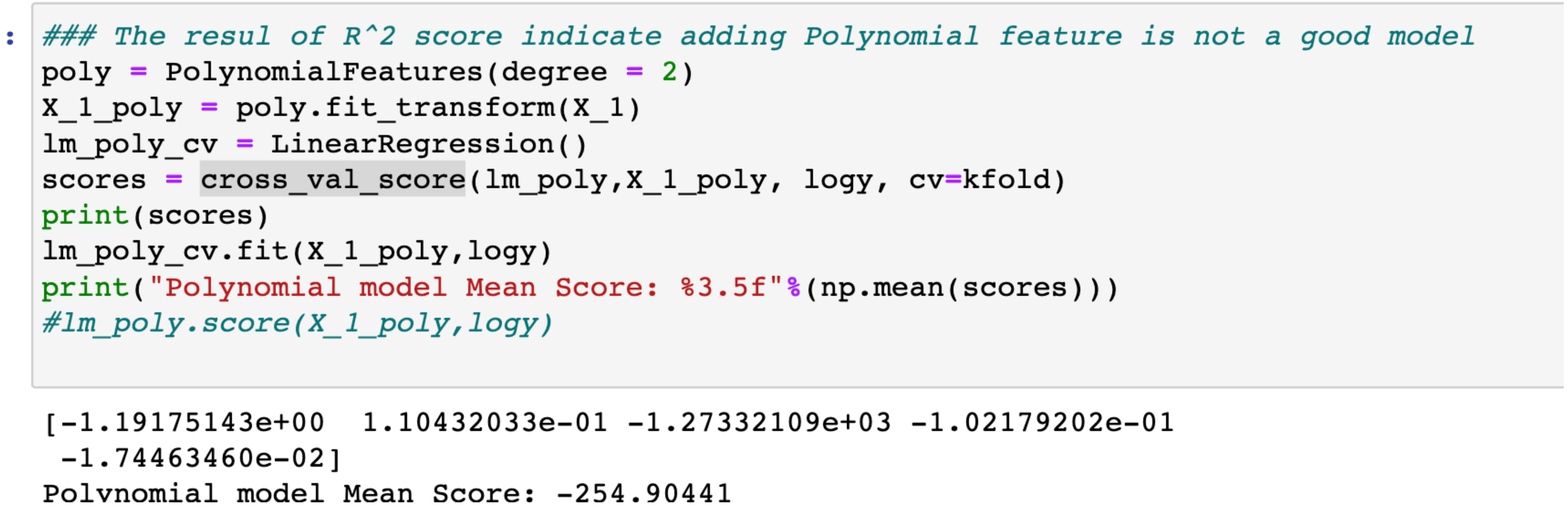
Omnibus:	4.876	**Durbin-Watson:**	2.021
Prob(Omnibus):	0.087	**Jarque-Bera (JB):**	4.913
Skew:	0.158	**Prob(JB):**	0.0857
Kurtosis:	2.964	**Cond. No.**	2.88e+10

Insight:

- Generally speaking, the financial numbers of a public company explains 56.5% of its stock price. Those features reflect how healthy the company's financials, operations, returns and etc, and have no direct relation with stock price.
- Besides financials, there are many outside elements with have important impact on stock price. eg, future of the industry/ product, leadership, policy change influence, law suits and etc.

Result

3. Linear regression with polynomial features.



```
### The resul of R^2 score indicate adding Polynomial feature is not a good model
poly = PolynomialFeatures(degree = 2)
X_1_poly = poly.fit_transform(X_1)
lm_poly_cv = LinearRegression()
scores = cross_val_score(lm_poly,X_1_poly, logy, cv=kfold)
print(scores)
lm_poly_cv.fit(X_1_poly,logy)
print("Polynomial model Mean Score: %3.5f"%(np.mean(scores)))
#lm_poly.score(X_1_poly,logy)
```

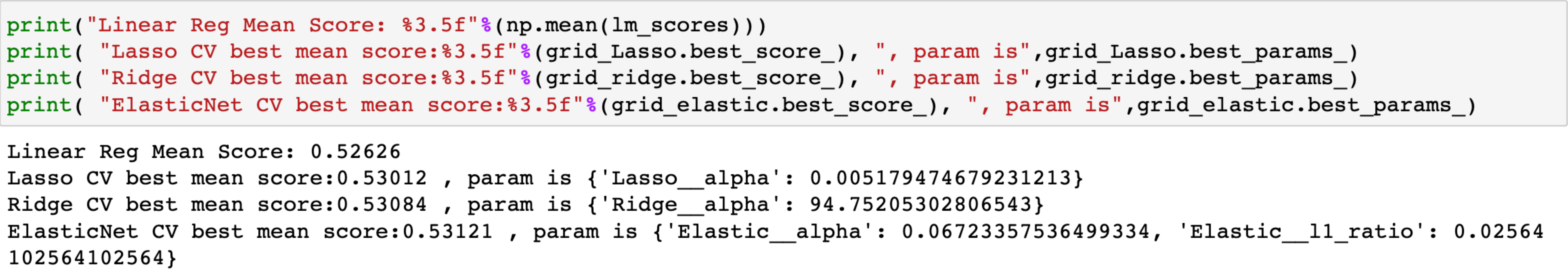
```
[-1.19175143e+00  1.10432033e-01 -1.27332109e+03 -1.02179202e-01
 -1.74463460e-02]
Polynomial model Mean Score: -254.90441
```

Insight:

- Polynomial features with Degree of 2 shows overfitting problem.

Result

4. Regression with Regulations: Choose best mean score — ElasticNet model



```
print("Linear Reg Mean Score: %3.5f"%(np.mean(lm_scores)))
print( "Lasso CV best mean score:%3.5f"%(grid_Lasso.best_score_), ", param is",grid_Lasso.best_params_)
print( "Ridge CV best mean score:%3.5f"%(grid_ridge.best_score_), ", param is",grid_ridge.best_params_)
print( "ElasticNet CV best mean score:%3.5f"%(grid_elastic.best_score_), ", param is",grid_elastic.best_params_)
```

```
Linear Reg Mean Score: 0.52626
Lasso CV best mean score:0.53012 , param is {'Lasso__alpha': 0.005179474679231213}
Ridge CV best mean score:0.53084 , param is {'Ridge__alpha': 94.75205302806543}
ElasticNet CV best mean score:0.53121 , param is {'Elastic__alpha': 0.06723357536499334, 'Elastic__l1_ratio': 0.02564
102564102564}
```

Definitions:

Revenue per share :Amount of revenue over common shares outstanding.

Diluted EPS : Earnings per share (EPS) if all convertible securities were exercised.

Profit margin: The difference between sales and the cost of goods sold divided by revenue.

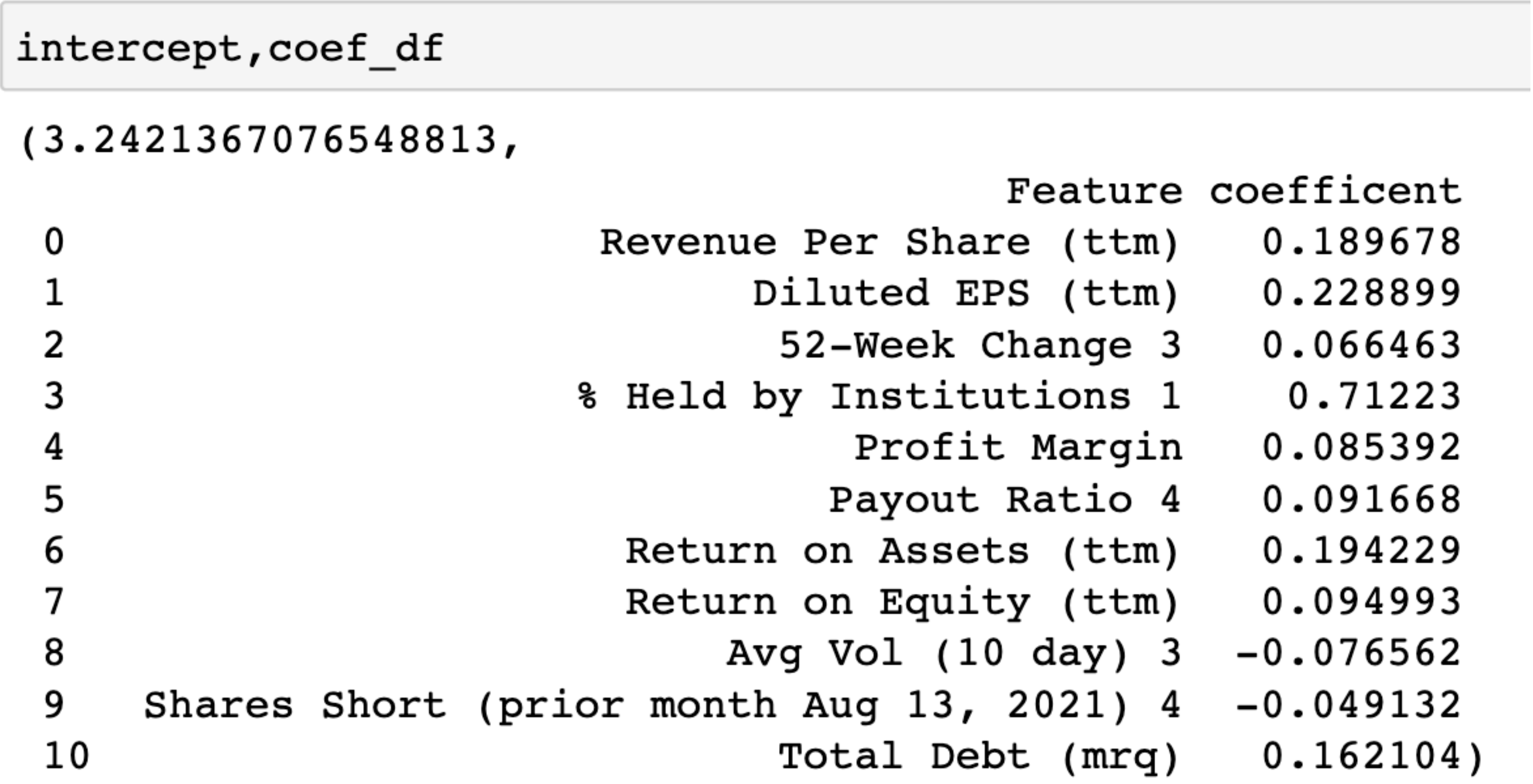
Payout Ratio: The percentage of net income that a company pays out as dividends to common shareholders.

Return on Asset: The rate of return (after tax) being earned on all of the firm's assets regardless of financing structure.

Return on Equity: Rate of return on the money invested by common stock owners and retained by the company thanks to previous profitable years.

Avg vol(10days): The average number of shares traded within a day in a given stock.

Shares Short: Number of shares that investors don't own but selling to other investors.



```
intercept,coef_df
```

```
(3.2421367076548813,
                                        Feature  coefficent
 0                         Revenue Per Share (ttm)    0.189678
 1                               Diluted EPS (ttm)    0.228899
 2                                52-Week Change 3    0.066463
 3                       % Held by Institutions 1     0.71223
 4                                   Profit Margin    0.085392
 5                                  Payout Ratio 4    0.091668
 6                         Return on Assets (ttm)    0.194229
 7                         Return on Equity (ttm)    0.094993
 8                             Avg Vol (10 day) 3   -0.076562
 9   Shares Short (prior month Aug 13, 2021) 4   -0.049132
 10                               Total Debt (mrq)    0.162104)
```

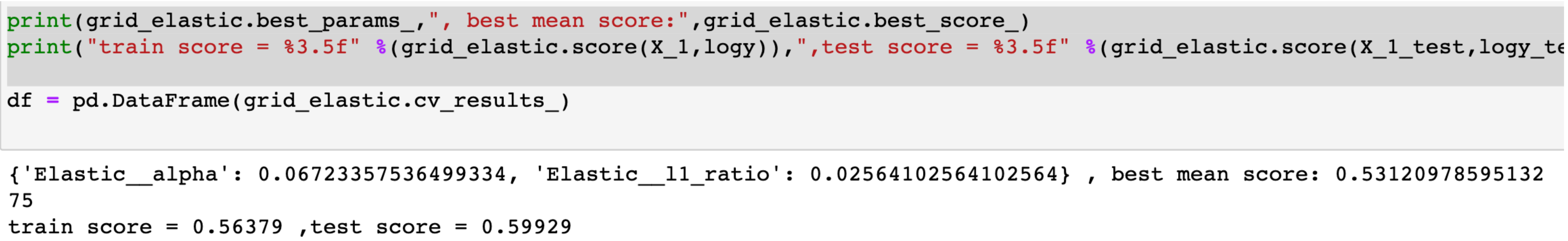
Result

5. Regression with Regulations:

Insight:

1. 'Revenue Per Share (ttm)', 'Diluted EPS (ttm)', '52-Week Change 3', '% Held by Institutions 1', 'Profit Margin', 'Payout Ratio 4', 'Return on Assets (ttm)', 'Return on Equity (ttm)''Total Debt (mrq)' have a Positive relationship with stock price. For example: For every dollar increase in Diluted EPS, the stock price will increase e^(0.229) = 1.26 times

2.'Avg Vol (10 day) 3', 'Shares Short (prior month Aug 13, 2021) 4' have negative relationship with stock price. For example, for every share Short by investors, the stock price decrease by (1-np.e**-0.049132) = 4%



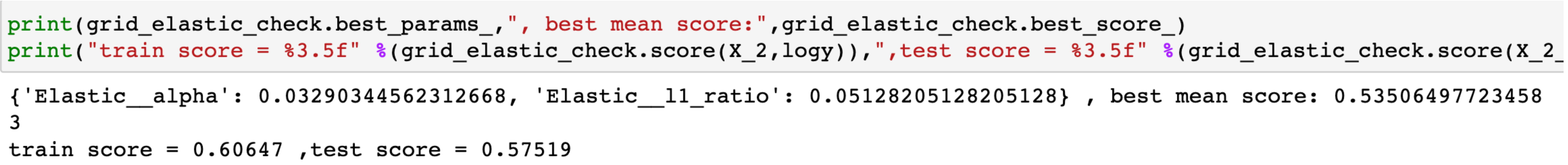
```
print(grid_elastic.best_params_,", best mean score:",grid_elastic.best_score_)
print("train score = %3.5f" %(grid_elastic.score(X_1,logy)),",test score = %3.5f" %(grid_elastic.score(X_1_test,logy_te

df = pd.DataFrame(grid_elastic.cv_results_)
```

```
{'Elastic__alpha': 0.06723357536499334, 'Elastic__l1_ratio': 0.02564102564102564} , best mean score: 0.53120978595132
75
train score = 0.56379 ,test score = 0.59929
```

Yes, no overfitting !

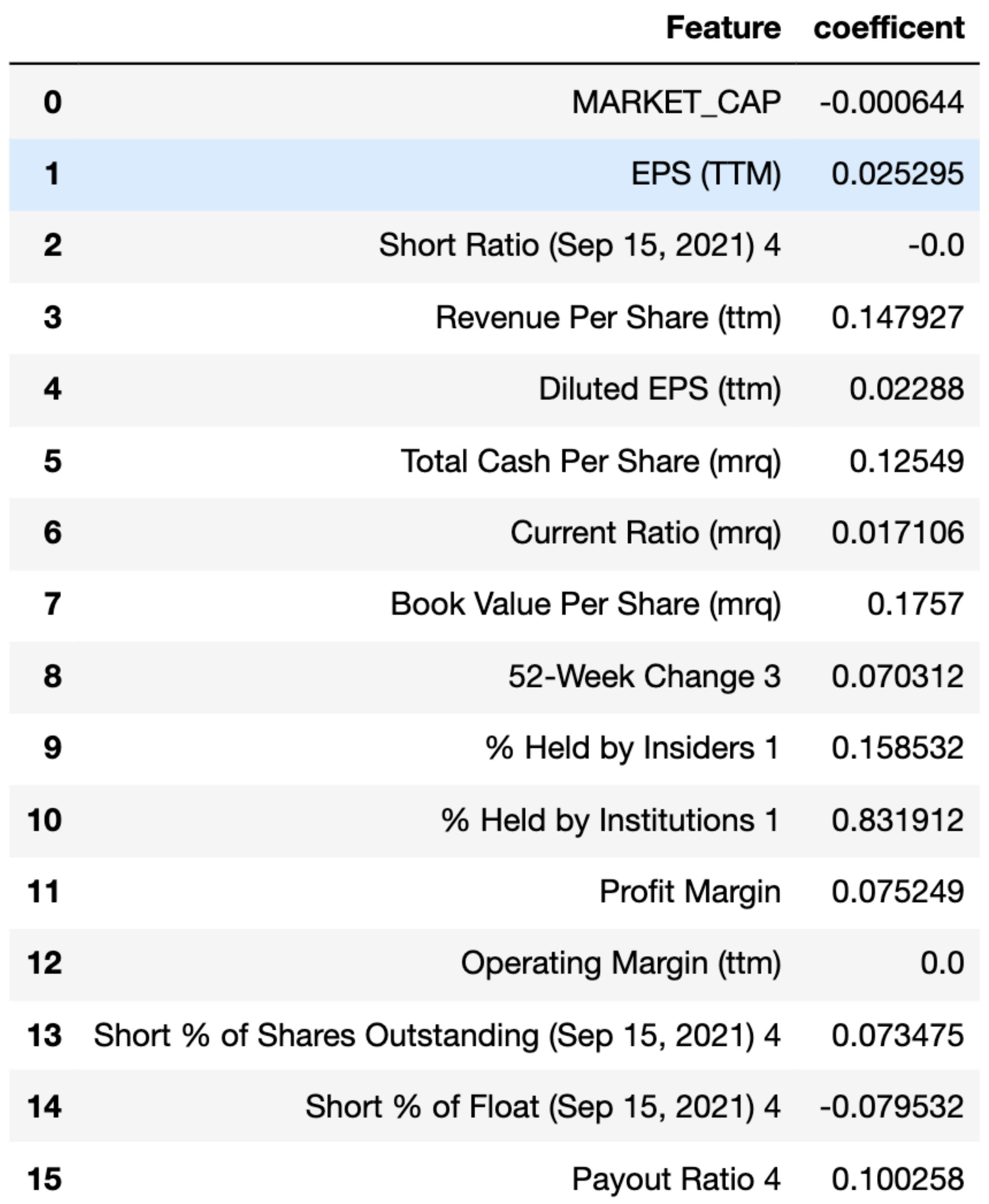
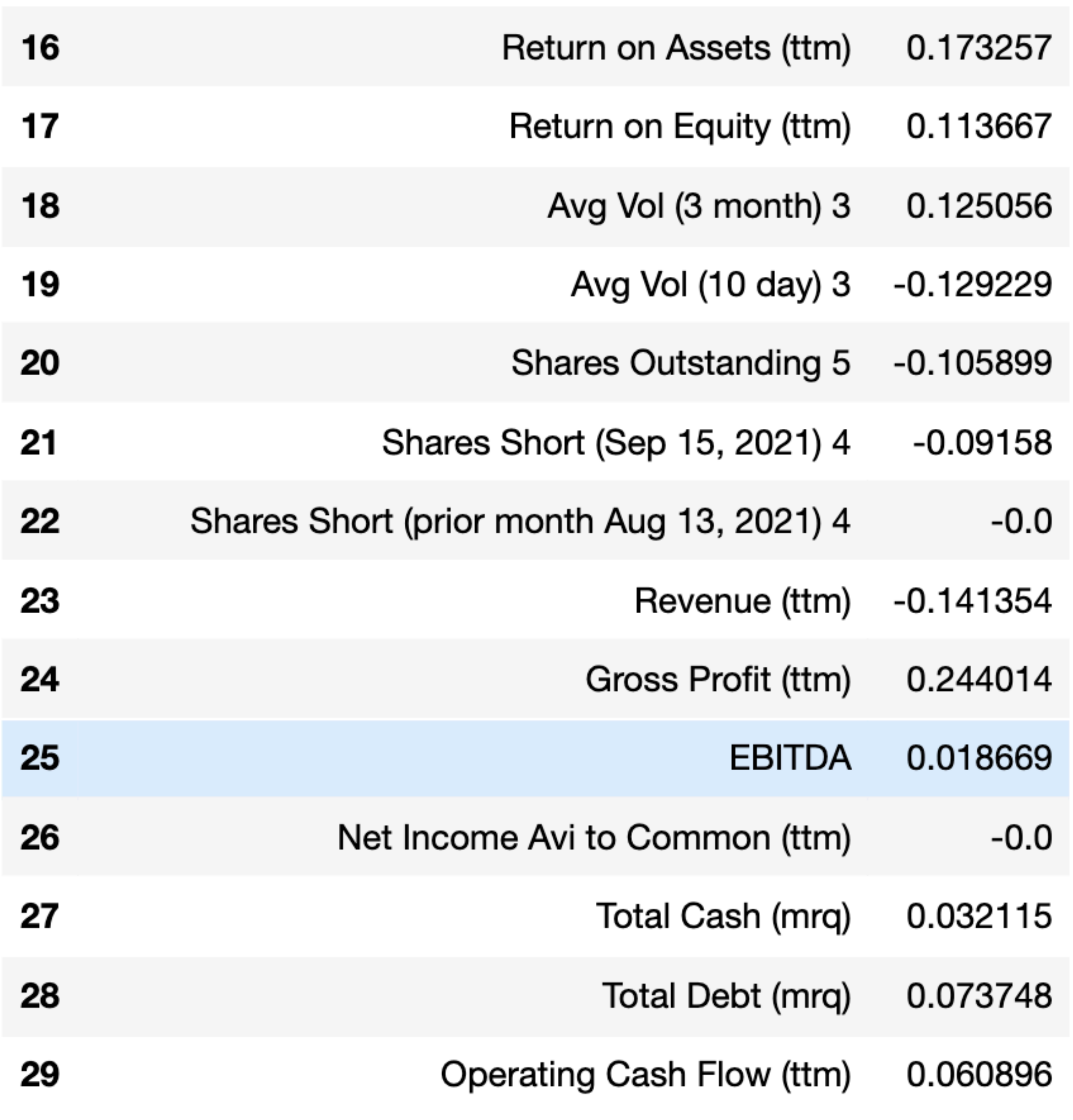
Result

6. What if feed all non-price-related features to ElasticNet model



```
print(grid_elastic_check.best_params_,", best mean score:",grid_elastic_check.best_score_)
print("train score = %3.5f" %(grid_elastic_check.score(X_2,logy)),",test score = %3.5f" %(grid_elastic_check.score(X_2_
```

```
{'Elastic__alpha': 0.03290344562312668, 'Elastic__l1_ratio': 0.05128205128205128} , best mean score: 0.53506497723458
3
train score = 0.60647 ,test score = 0.57519
```

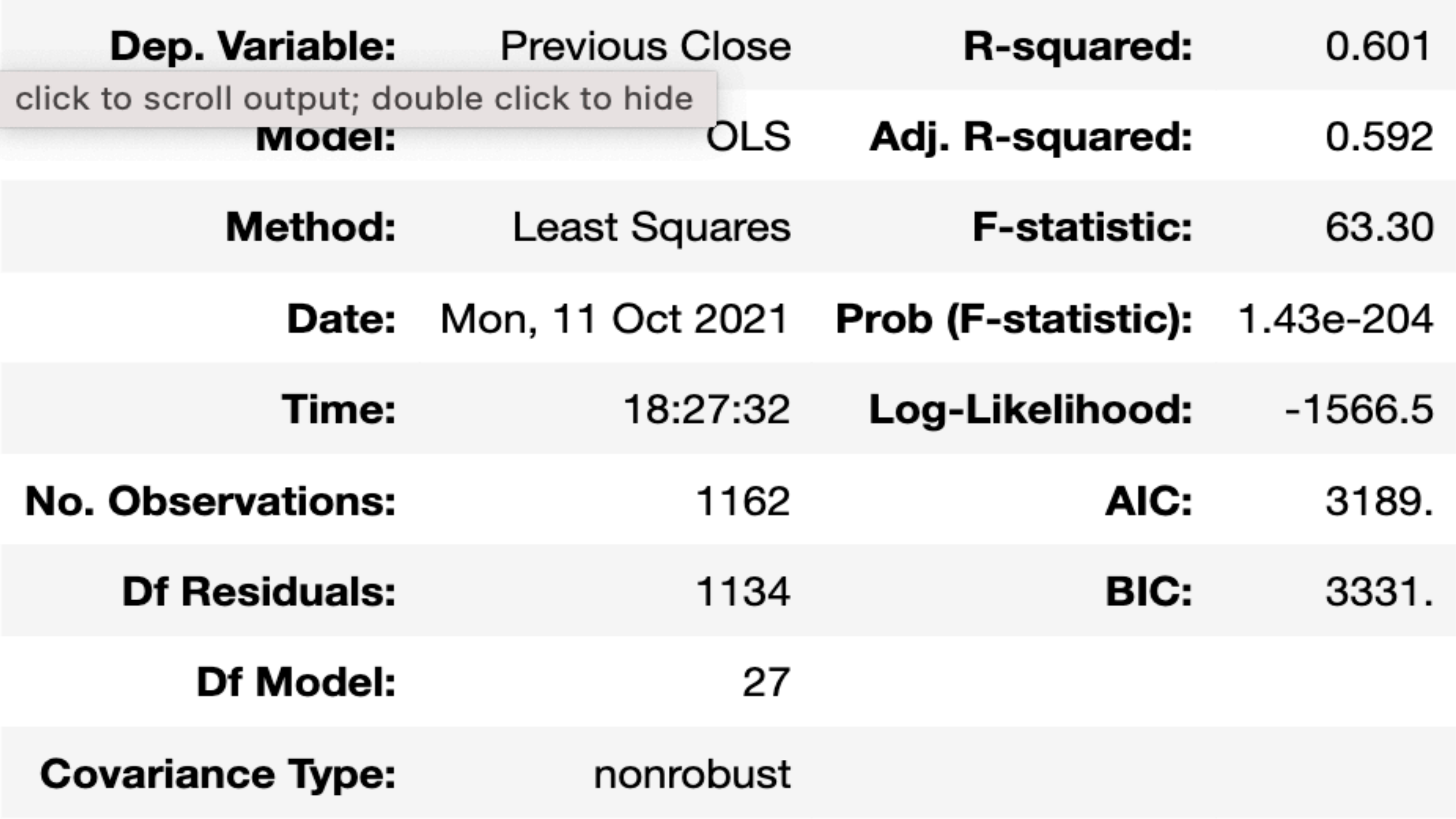



	Feature	coefficent
0	MARKET_CAP	-0.000644
1	EPS (TTM)	0.025295
2	Short Ratio (Sep 15, 2021) 4	-0.0
3	Revenue Per Share (ttm)	0.147927
4	Diluted EPS (ttm)	0.02288
5	Total Cash Per Share (mrq)	0.12549
6	Current Ratio (mrq)	0.017106
7	Book Value Per Share (mrq)	0.1757
8	52-Week Change 3	0.070312
9	% Held by Insiders 1	0.158532
10	% Held by Institutions 1	0.831912
11	Profit Margin	0.075249
12	Operating Margin (ttm)	0.0
13	Short % of Shares Outstanding (Sep 15, 2021) 4	0.073475
14	Short % of Float (Sep 15, 2021) 4	-0.079532
15	Payout Ratio 4	0.100258
16	Return on Assets (ttm)	0.173257
17	Return on Equity (ttm)	0.113667
18	Avg Vol (3 month) 3	0.125056
19	Avg Vol (10 day) 3	-0.129229
20	Shares Outstanding 5	-0.105899
21	Shares Short (Sep 15, 2021) 4	-0.09158
22	Shares Short (prior month Aug 13, 2021) 4	-0.0
23	Revenue (ttm)	-0.141354
24	Gross Profit (ttm)	0.244014
25	EBITDA	0.018669
26	Net Income Avi to Common (ttm)	-0.0
27	Total Cash (mrq)	0.032115
28	Total Debt (mrq)	0.073748
29	Operating Cash Flow (ttm)	0.060896

Insight: The mean score increased slightly, but our model will become much more complex. not recommend.

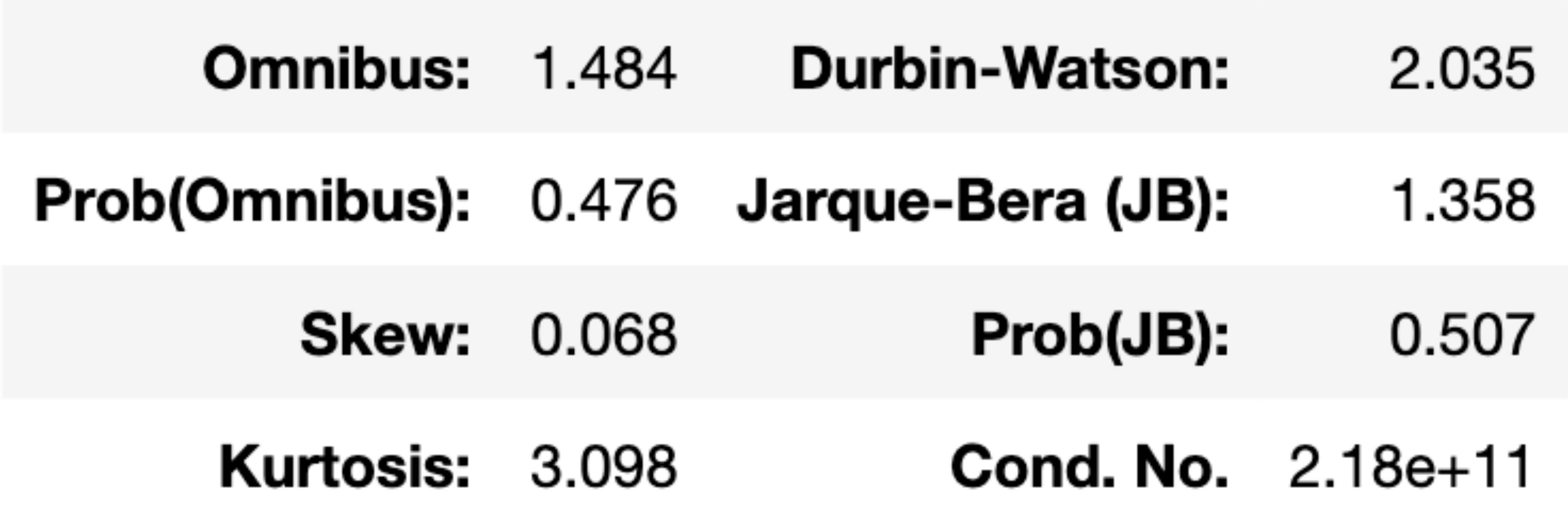
Result

7. Linear regression with dummy variables “Industry"

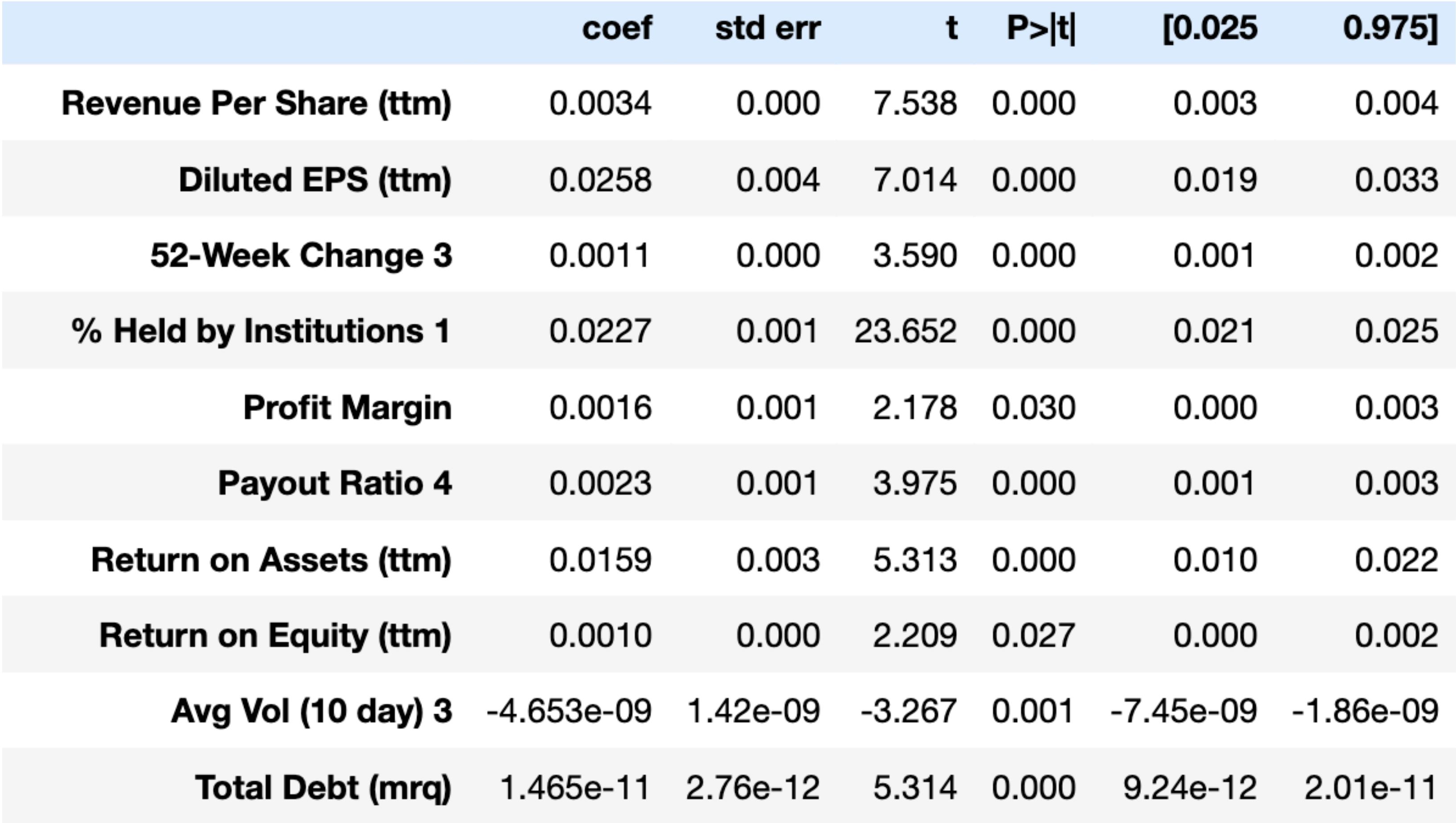


click to scroll output; double click to hide

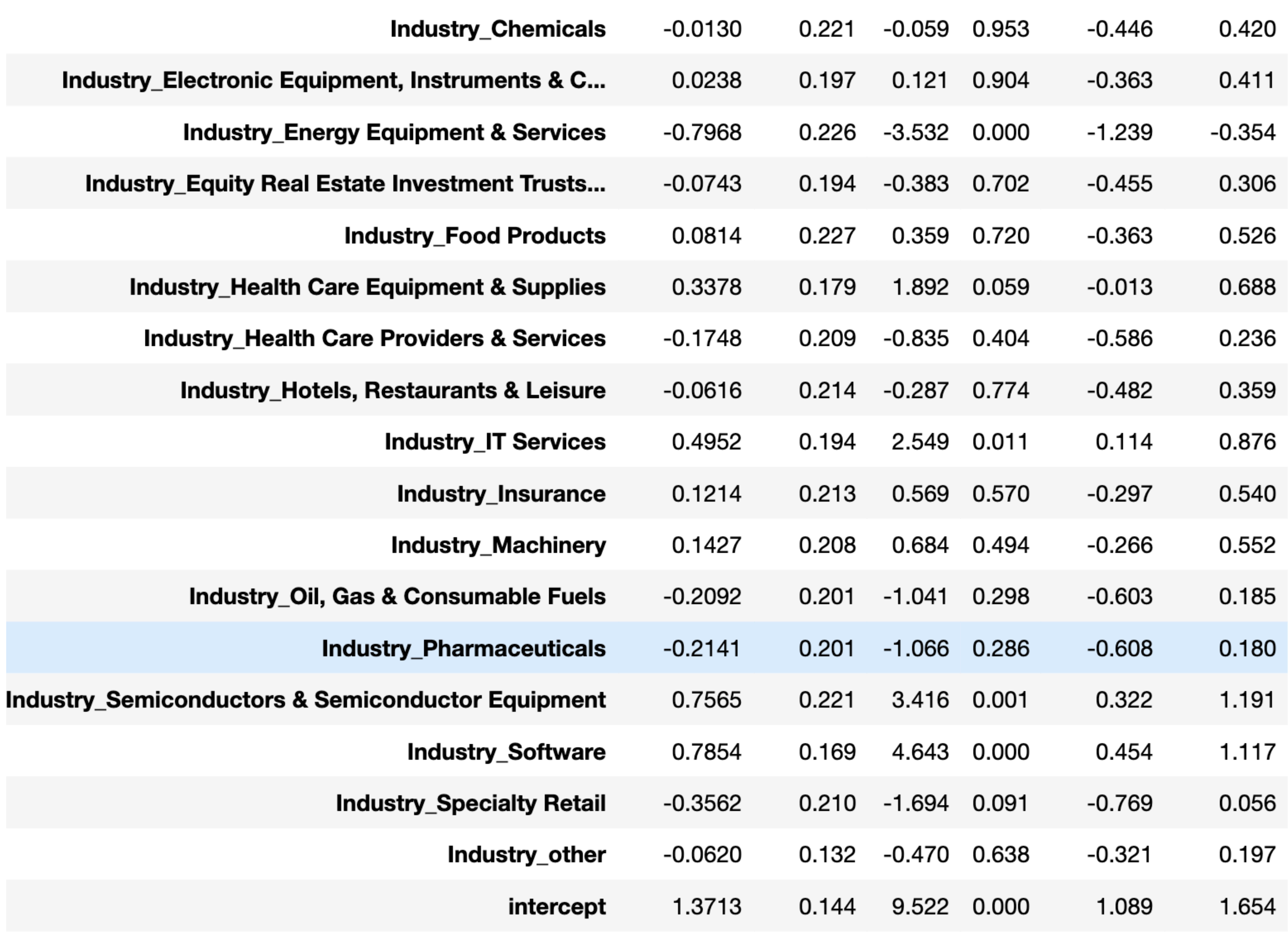
Dep. Variable:	Previous Close	**R-squared:**	0.601
Model:	OLS	**Adj. R-squared:**	0.592
Method:	Least Squares	**F-statistic:**	63.30
Date:	Mon, 11 Oct 2021	**Prob (F-statistic):**	1.43e-204
Time:	18:27:32	**Log-Likelihood:**	-1566.5
No. Observations:	1162	**AIC:**	3189.
Df Residuals:	1134	**BIC:**	3331.
Df Model:	27		
Covariance Type:	nonrobust		



Omnibus:	1.484	**Durbin-Watson:**	2.035
Prob(Omnibus):	0.476	**Jarque-Bera (JB):**	1.358
Skew:	0.068	**Prob(JB):**	0.507
Kurtosis:	3.098	**Cond. No.**	2.18e+11



	coef	std err	t	P>|t|	[0.025	0.975]
Revenue Per Share (ttm)	0.0034	0.000	7.538	0.000	0.003	0.004
Diluted EPS (ttm)	0.0258	0.004	7.014	0.000	0.019	0.033
52-Week Change 3	0.0011	0.000	3.590	0.000	0.001	0.002
% Held by Institutions 1	0.0227	0.001	23.652	0.000	0.021	0.025
Profit Margin	0.0016	0.001	2.178	0.030	0.000	0.003
Payout Ratio 4	0.0023	0.001	3.975	0.000	0.001	0.003
Return on Assets (ttm)	0.0159	0.003	5.313	0.000	0.010	0.022
Return on Equity (ttm)	0.0010	0.000	2.209	0.027	0.000	0.002
Avg Vol (10 day) 3	-4.653e-09	1.42e-09	-3.267	0.001	-7.45e-09	-1.86e-09
Total Debt (mrq)	1.465e-11	2.76e-12	5.314	0.000	9.24e-12	2.01e-11



Industry_Chemicals	-0.0130	0.221	-0.059	0.953	-0.446	0.420
Industry_Electronic Equipment, Instruments & C...	0.0238	0.197	0.121	0.904	-0.363	0.411
Industry_Energy Equipment & Services	-0.7968	0.226	-3.532	0.000	-1.239	-0.354
Industry_Equity Real Estate Investment Trusts...	-0.0743	0.194	-0.383	0.702	-0.455	0.306
Industry_Food Products	0.0814	0.227	0.359	0.720	-0.363	0.526
Industry_Health Care Equipment & Supplies	0.3378	0.179	1.892	0.059	-0.013	0.688
Industry_Health Care Providers & Services	-0.1748	0.209	-0.835	0.404	-0.586	0.236
Industry_Hotels, Restaurants & Leisure	-0.0616	0.214	-0.287	0.774	-0.482	0.359
Industry_IT Services	0.4952	0.194	2.549	0.011	0.114	0.876
Industry_Insurance	0.1214	0.213	0.569	0.570	-0.297	0.540
Industry_Machinery	0.1427	0.208	0.684	0.494	-0.266	0.552
Industry_Oil, Gas & Consumable Fuels	-0.2092	0.201	-1.041	0.298	-0.603	0.185
Industry_Pharmaceuticals	-0.2141	0.201	-1.066	0.286	-0.608	0.180
Industry_Semiconductors & Semiconductor Equipment	0.7565	0.221	3.416	0.001	0.322	1.191
Industry_Software	0.7854	0.169	4.643	0.000	0.454	1.117
Industry_Specialty Retail	-0.3562	0.210	-1.694	0.091	-0.769	0.056
Industry_other	-0.0620	0.132	-0.470	0.638	-0.321	0.197
intercept	1.3713	0.144	9.522	0.000	1.089	1.654

Conclusion

- ElasticNet regression Model
- Future work if time allows:
    1. Time Series model to predict stock prices
    2. Explore the relationship between stock price and other categorical variables.