IE5202 Project 1 Report

Yang Xiaozhou, A0113538

October 5, 2017

1 Data Exploration

To visualize the linear relationship between target variable and the predictors, various scatter plots with regression lines are examined. In Figure 1, it can be seen that essential features have weak positive features with target variable. However, among themselves, several features have strong positive correlation, as seen in Figure 2, Figure 3 and Figure 4.

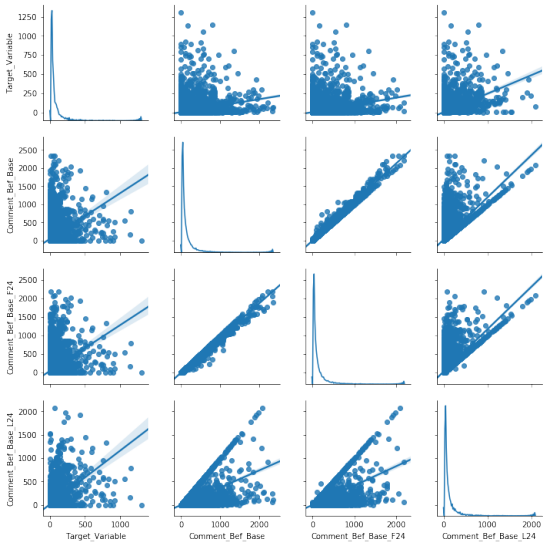

Figure 1: Relationship between target variable and essential features.

While other variables do not show strong linear relationship with the target variable, the box-plot of Page Category (4) does show that the difference in page category have some influence on the target variable, see Figure 5.

Another thing that is worth noting is the many variables, including the target variable empirically has the form of a power distribution that is heavily right-tailed. This can be seen in Figure 6.

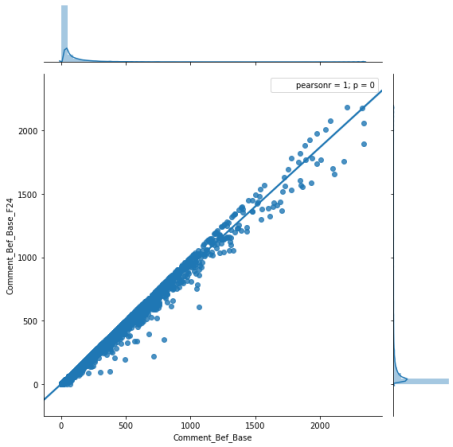

Figure 2: Comment before base time and comment before base time (first 24 hours).

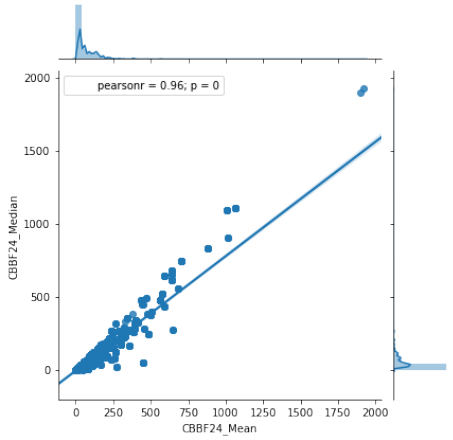

Figure 3: Comment before base time (mean of first 24 house) and comment before base time (median of first 24 hours).

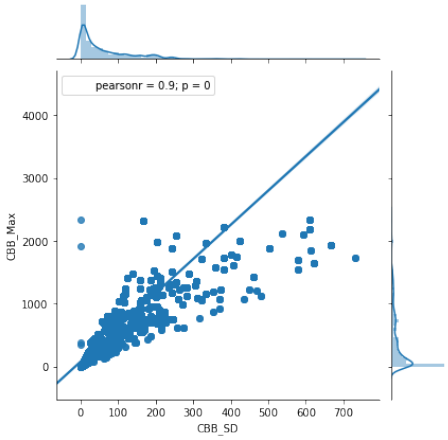

Figure 4: Comment before base time (standard deviation) and comment before base time (maximum).

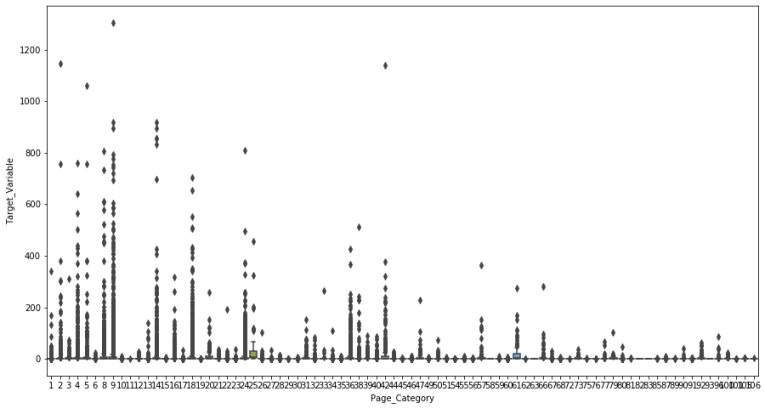

Figure 5: Boxplot of page category and target variable.

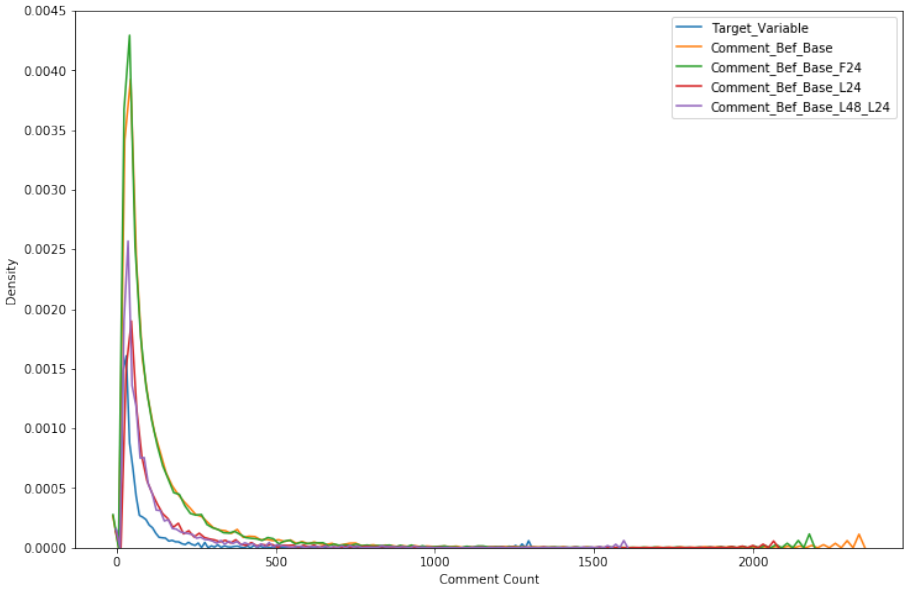

Figure 6: Emprical density of variables ensemble power distributions.

2 Simple Regression Model

2.1 Model Building

Without any interaction terms, the candidates used in this regression model is:

- 12 Mean of comment count in last 24 hours
- 26 Maximum of comment difference between C2 and C3
- 31 Comment of the last 24 hours but before base time
- 33 Comment of the first 24 hours but before base time
- 34 Comment difference between C2 and C3
- 35 Time gap

Also, both the target variable and non-categorical variables are transformed through cube root function:

$$\mathbf{Y} = \sqrt[3]{\mathbf{Y}} \tag{1}$$

$$\mathbf{X} = \sqrt[3]{\mathbf{X}} \tag{2}$$

Dep. Variable:	Target Variable	**R-squared:**	0.679
Model:	OLS	**Adj. R-squared:**	0.679
Method:	Least Squares	**F-statistic:**	1.442e+04
Date:	Sat, 30 Sep 2017	**Prob (F-statistic):**	0.00
Time:	16:57:42	**Log-Likelihood:**	-41766.
No. Observations:	40949	**AIC:**	8.355e+04
Df Residuals:	40942	**BIC:**	8.361e+04
Df Model:	6		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.3103	0.017	77.656	0.000	1.277	1.343
Comment_Diff_C2_C3	0.0434	0.002	21.095	0.000	0.039	0.047
CBBL24_Mean	0.2412	0.007	33.442	0.000	0.227	0.255
Comment_Diff_Max	-0.0490	0.003	-18.366	0.000	-0.054	-0.044
Comment_Bef_Base_F24	0.1302	0.005	28.297	0.000	0.121	0.139
Time_Gap	-0.4748	0.005	-94.776	0.000	-0.485	-0.465
Comment_Bef_Base_L24	0.2573	0.005	47.253	0.000	0.247	0.268

Omnibus:	13471.617	**Durbin-Watson:**	1.847
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	167947.876
Skew:	1.227	**Prob(JB):**	0.00
Kurtosis:	12.613	**Cond. No.**	41.9

Table 1: OLS Regression Results - Simple Model

	MAE	MSE
0	4.242216	512.076063
1	4.380490	498.156699
2	4.994586	1001.783356
3	5.284419	1054.551894
4	4.618331	670.193012
5	5.210752	1228.869105
6	5.063661	820.874563
7	5.853072	1125.894034
8	4.385391	573.695776
9	5.127996	1223.085040
Mean	**4.916092**	**870.917954**

Table 2: Cross Validation Result

2.2 Result

Regression model summary can be found in Table 1 and the 10-fold cross-validation score can be found in Table 2.

This regression model with 6 predictors yield an adjusted R^2 score of 0.679, which indicates that the model is not a strong representation of the relationship between the target variable and predictors. Since the F-statistic is large and the related p-value is very close to zero, there is strong evidence at 95% significant level to reject the null hypothesis that none of the predictors need to be in the model.

In the predictor table, coefficients of each predictor and the intercept value are reported. P-value of each of the predictor coefficients and the intercept is really close to 0 (reported as 0), this means that each of them have a value that is significantly different from 0 at 95% significant level.

Durbin-Watson statistic is 1.847, which is close to 2. This indicates insignificant level of autocorrelation in the residuals produced by this regression model. Also, the condition number here is 41.9, which may suggest that there is no strong collinearity between the covariates. However, upon further investigation, there is actually strong correlation between predictor 26 and 31, as shown in this scatter plot (Figure 7).

To measure the prediction performance of the regression model, a 10-fold cross-validation technique is used to collect the result in Table 2. Mean score of AIC and BIC are smaller than that reported in the summary table by about 8000. This table also reports the mean absolute error (MAE) and mean square error (MSE), which measures the prediction performance of the model on unseen data. In my model selection process, however, adjusted R^2 is used as the selection criterion.

The QQ-plot plots the empirical quantiles of residuals and the quantiles of a standard normal distribution. In Figure 8, the sample quantiles do not form a straight line with the theoretical quantiles at both the left and right tails. This

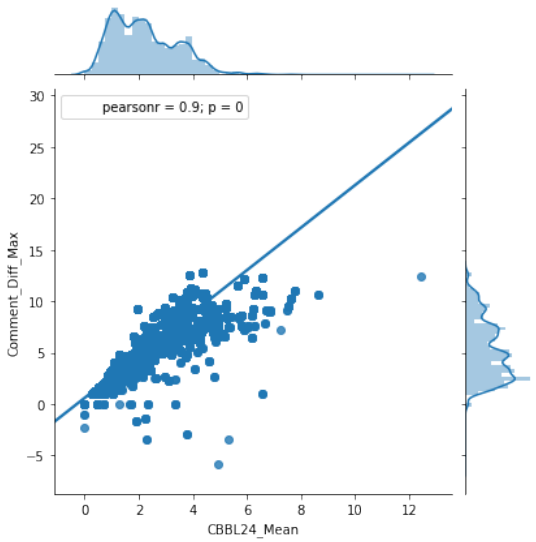

Figure 7: Strong collinearity between two of the predictors.

suggest that the normality assumption of the error term ϵ_i is violated.

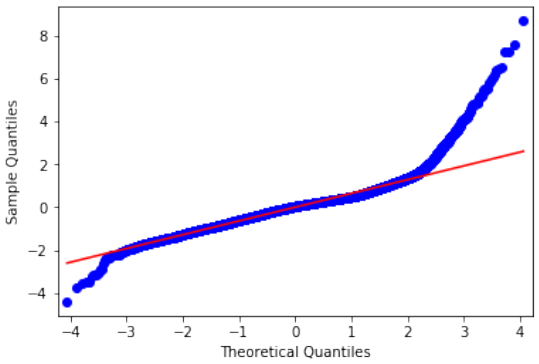

Figure 8: QQ-plot to check residual normality assumption.

By examining the plot of studentized residuals and fitted values, several assumptions of the linear regression model could be verified. Figure 9 shows the plot for this simple regression model. It seems that the residuals have a mean that is larger than 0 because more points lie above the zero line. Also, the variance of the residuals appear to increase with the fitted value, this suggests a violation of the constant variance assumption. Finally, there are plenty of residuals which have very large values (i.e. deviation from 0 with more than 3 standard deviations). This may indicate that the model is not adequate enough

to fit the data well.

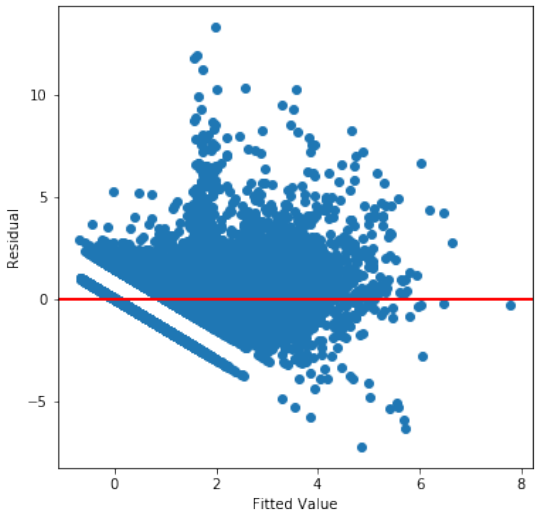

Figure 9: Residual plot of the simple regression model.

Two plots with partial regression are reported here for illustration with

- Post share count (Figure 10),
- Mean of comment count in last 24 hours (Figure 11).

From both of the top left plots, it can be seen that the model tends to underestimate the target value as the bulk of fitted values lie below the target values. This may have to do with the fact that the target variable follows a heavily right-values empirical distribution. Both of the partial regression plots on the bottom left show a moderate positive relationship between the target variable and the individual predictor.

3 Complex Regression Model

3.1 Model Building

In the complex regression model, interaction terms between two variables are added to the model. Based on the feature importance score (Figure 12) extracted from an ensemble boosting tree regression model, 20 predictors with the highest importance are used as the basis for creating interaction terms. A complete list of the predictors can be found in Table 3. The same cube root transformation in simple regression model has been applied here as well.

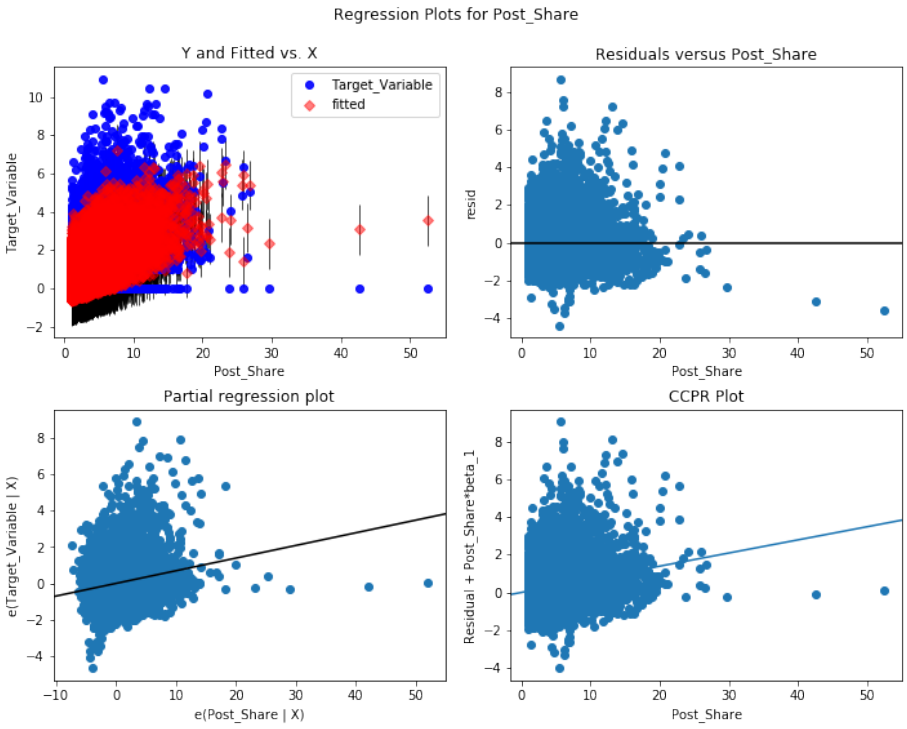

Figure 10: Plots on predictor: Post share.

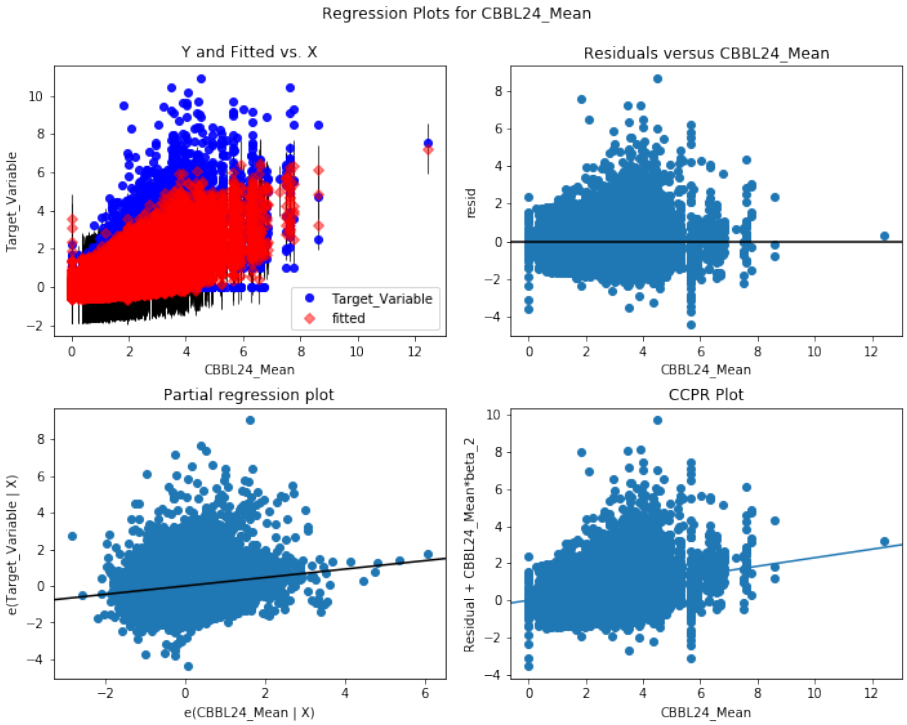

Figure 11: Plots on predictor: Mean of comment count in last 24 hours.

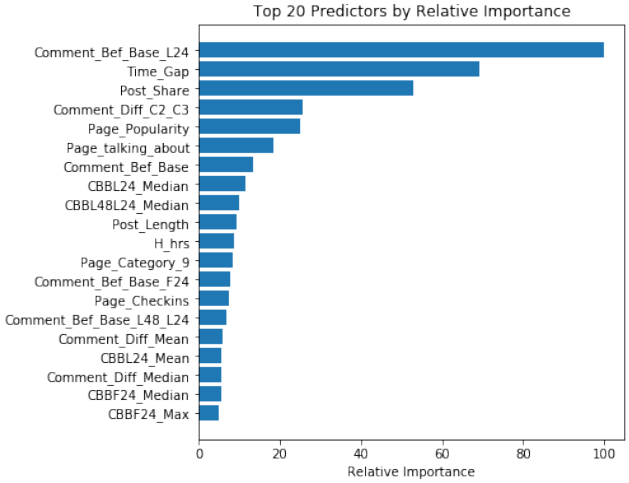

Figure 12: Top 20 predictors' relative feature importance generated from fitting a boosting tree regression model.

	Predictors
0	Comment_Bef_Base_L24
1	Time_Gap
2	Post_Share
3	Comment_Diff_C2_C3
4	Page_Popularity
5	Page_talking_about
6	Comment_Bef_Base
7	CBBL24_Median
8	CBBL48L24_Median
9	Post_Length
10	H_hrs
11	Page_Category_9
12	Comment_Bef_Base_F24
13	Page_Checkins
14	Comment_Bef_Base_L48_L24
15	Comment_Diff_Mean
16	CBBL24_Mean
17	Comment_Diff_Median
18	CBBF24_Median
19	CBBF24_Max

Table 3: Predictors used for creating interaction terms.

Dep. Variable:	Target Variable	**R-squared:**	0.799
Model:	OLS	**Adj. R-squared:**	0.799
Method:	Least Squares	**F-statistic:**	1522.
Date:	Sun, 01 Oct 2017	**Prob (F-statistic):**	0.00
Time:	13:04:18	**Log-Likelihood:**	-32116.
No. Observations:	40949	**AIC:**	6.445e+04
Df Residuals:	40841	**BIC:**	6.538e+04
Df Model:	107		

Omnibus:	5945.914	**Durbin-Watson:**	1.888
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45015.854
Skew:	0.481	**Prob(JB):**	0.00
Kurtosis:	8.046	**Cond. No.**	2.98e+05

Table 4: OLS Regression Results - Complex Model

3.2 Result

Regression result for this model is reported in Table 4. While the size of regression model has increased significantly from the simple model, the overall performance has also improved. Adjusted R^2 score is 0.799, hence the variation explained by the model is about 80%, up from 67.9% of the simple model. Test for all predictors are needed is successful with a close to zero p-value. Both AIC and BIC scores are smaller than those reported in the simple model. The Durbin-Watson statistic is still close to 2, which cause no alarm in residual autocorrelation. However, because of the high number of predictors deployed, the condition number of this model is very high. This suggests the presence of strong multicollinearity among certain predictors. This is not surprising, though, having seen strong correlation between predictors from the initial data exploration analysis.

Again, to accurately assess the prediction performance of the model to unseen data and prevent over-fitting on the available data, a 20-fold cross-validation technique is used to obtain performance metric values and the mean of these 20 values. They are reported in Table 5 and in a plot of the cross-validated mean square error (MSE) value (Figure 13). The mean value of MSE is 3.65893.

The QQ-plot of residuals shows deviation of the points from the straight line at the tails. There seems to be slight improvement over the simple regression model. However, the obvious deviation suggests that the normality assumption of residuals is violated and the model might still need revision.

The plot of residuals against fitted value shows a similar pattern. While the number of large residuals decreased relative to the simple model, there are still many residuals outside the 3 standard deviations region. The variance of residuals seem to be non-constant as well, with smaller variance at the two ends of fitted value and larger variance in the middle.

	MAE	MSE
0	2.902373	146.374018
1	3.560299	362.600952
2	3.768702	412.997428
3	3.051363	149.435547
4	3.548811	275.254989
5	3.942151	947.329354
6	3.410386	306.653846
7	4.319872	876.712590
8	2.993184	168.879596
9	3.731502	491.642907
10	4.235156	1032.601620
11	3.438691	217.428396
12	3.308620	295.118419
13	4.192867	391.431910
14	4.747604	906.253668
15	3.699576	215.585194
16	2.972562	206.520001
17	3.755930	478.474729
18	3.969618	829.406963
19	3.629519	215.318129
Mean	**3.658939**	**446.301013**

Table 5: Cross Validation Result

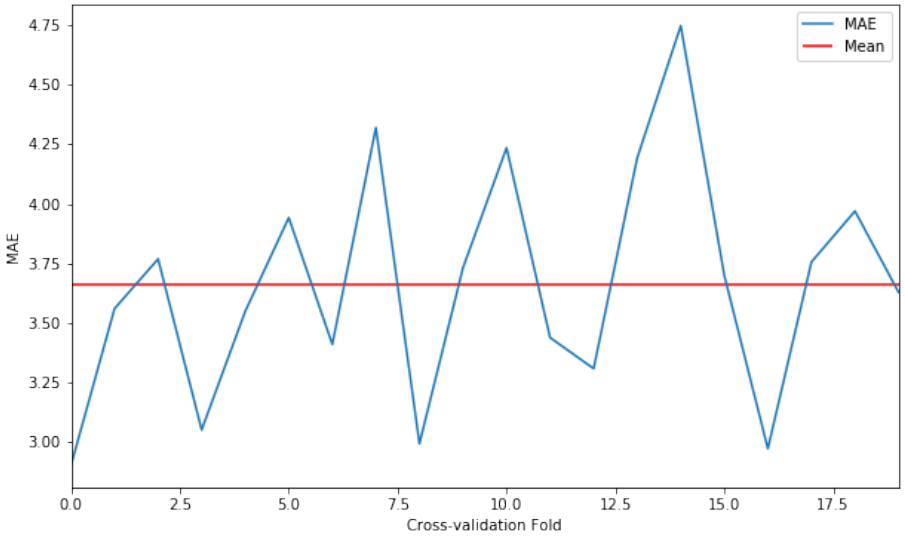

Figure 13: MAE score for 20-fold cross validation and their mean.

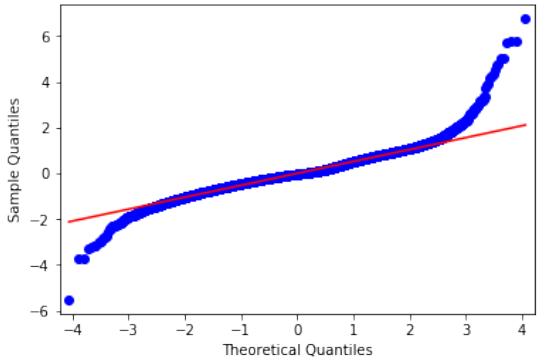

Figure 14: QQ-plot to check residual normality assumption.

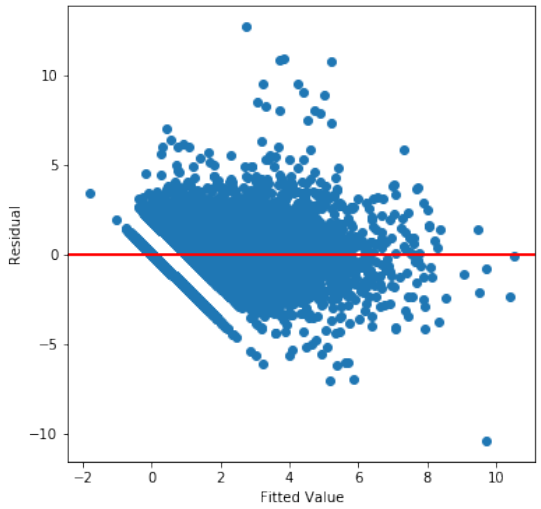

Figure 15: Residual plot of the simple regression model.

Plots of 5 individual predictors are also reported in Figure 16, 17, 18, 19, and 20. From each of the top left plot, the issue of underestimation is not so severe now compared to the simple model. And the residuals seem to have mean 0 and constant variance for each of the individual predictor. One particular observation can be made from the 5 partial regression plots. The variance of residuals around value 0 of predictor regression error (x-axis) is significantly higher than at other values.

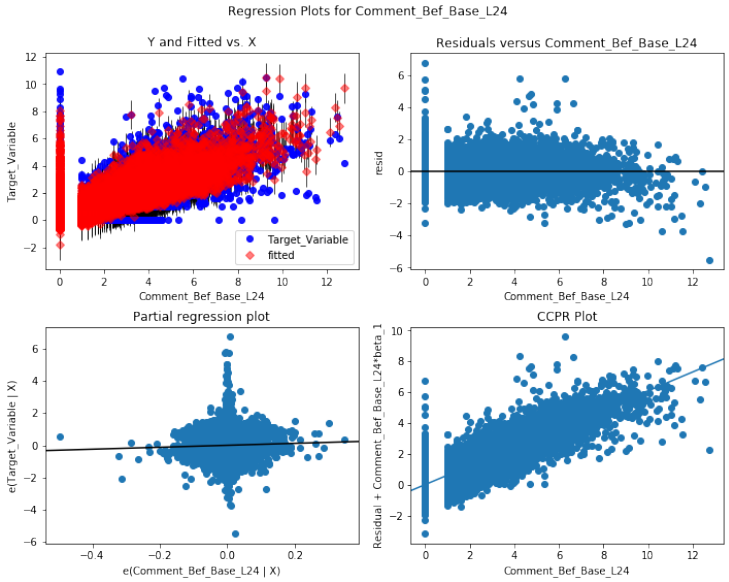

Figure 16: Plots on predictor: Comment count of last 24 hours before base time.

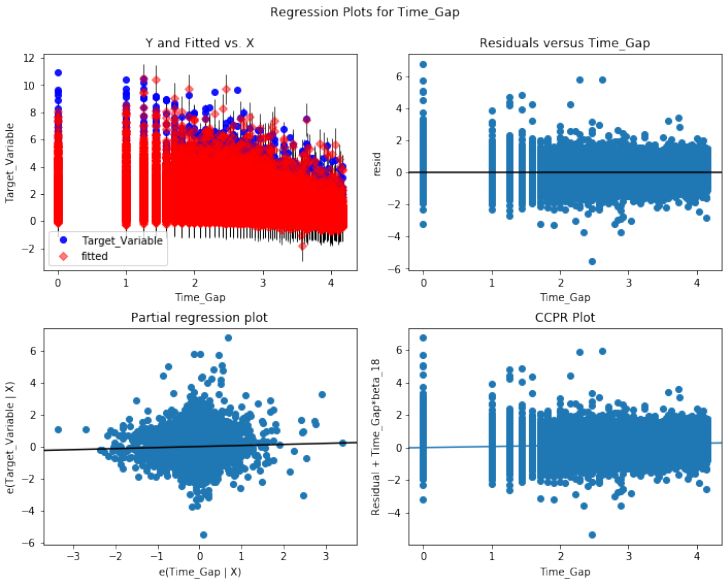

Figure 17: Plots on predictor: Time gap.

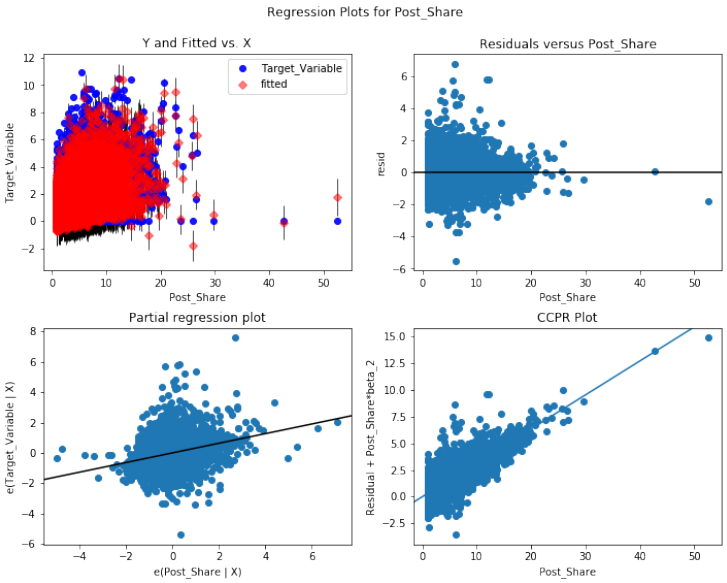

Figure 18: Plots on predictor: Post share count.

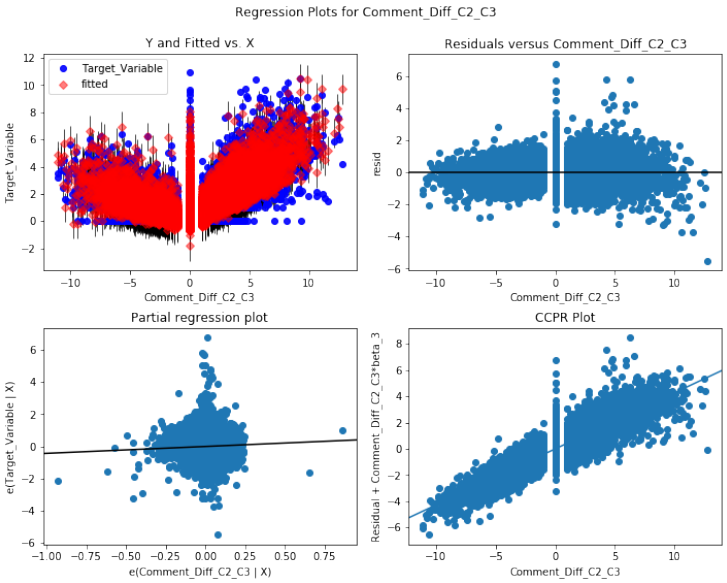

Figure 19: Plots on predictor: Comment count difference between C2 and C3.

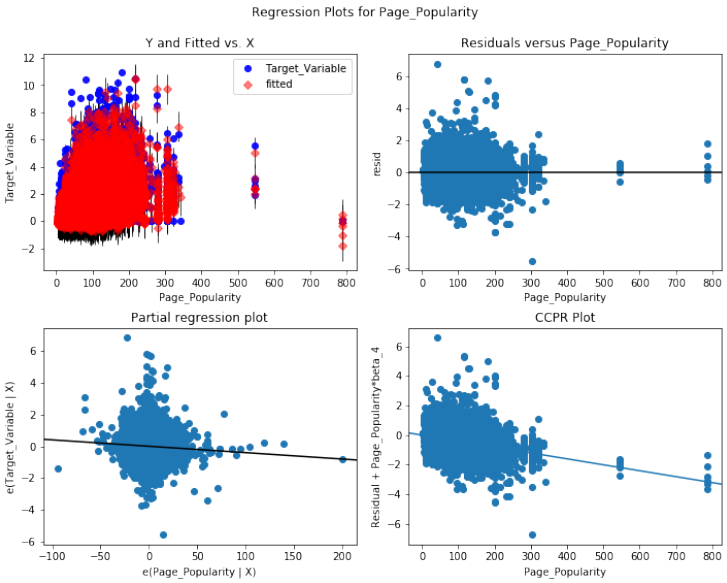

Figure 20: Plots on predictor: Page popularity.