Assignment #2

By Kun Li, Karan Modi

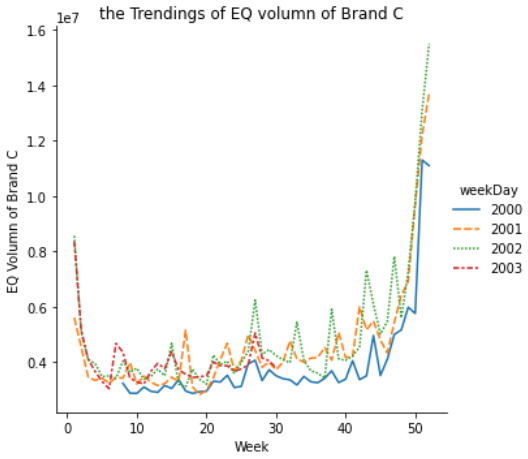
1. Univariate analysis

(a) Statistics Summary

	weekDay	eq_volum	disacv_c	bonusacv	price_c	price_e	price_p	tvgrp_c	tvgrp_u	trustad	fsi_holi	fsi_non	fsi_comp	itemstor	walmart
min	2000-02-26	2829145.00	5.42	0.00	0.84	0.70	0.54	0.00	0.00	False	0.00	0.00	0.00	8.50	0.22
max	2003-07-26	15500000.00	33.29	69.00	0.96	1.12	0.79	346.00	398.00	True	41590.00	41676.00	92896.00	10.00	0.45
mean	2001-11-10	4430421.47	15.14	33.07	0.93	1.02	0.71	48.14	36.62	0.26257	1014.83	3216.59	7239.80	9.24	0.32
sd	NaT	1982087.85	6.13	22.36	0.02	0.05	0.04	82.01	81.79	0.441265	6180.89	10883.52	17776.87	0.38	0.06

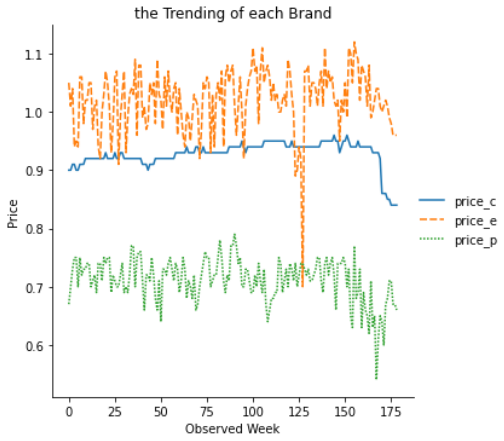
Based on the table above, the price range of our product (Brand C) is 0.12 compared to our competitors' price range – Brand E's range is 0.42, and Brand P's range is 0.25. Thus, it implies that our product does not offer big discounts like our competitors. Also, the minimum price of our product is the highest among the brands, which can put Brand C at a disadvantage during promotion week.

(b) Time-series analysis



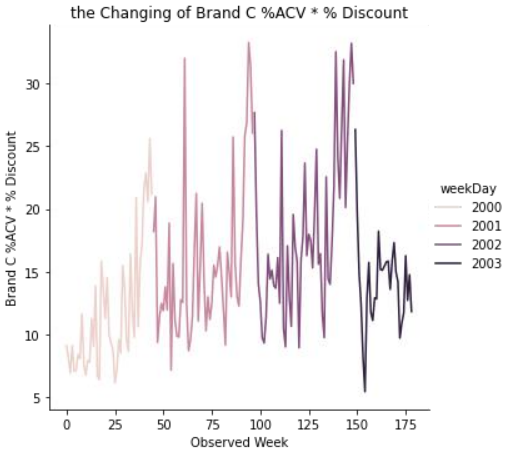
Analysis:

The EQ Volume sales of Brand C follow a similar pattern over the four years. The recorded sales were highest in the last couple of weeks of the year, which can result from heavy discount/promotion or salespeople meeting their yearly targets.



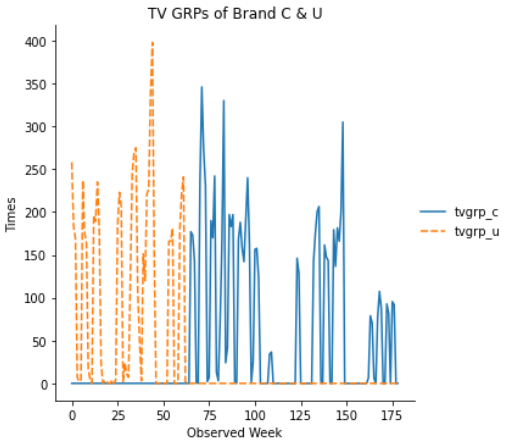
Analysis:

Brand C's price was stable over the given period whereas our competitors' prices vary significantly, indicating frequent discounts and promotions.



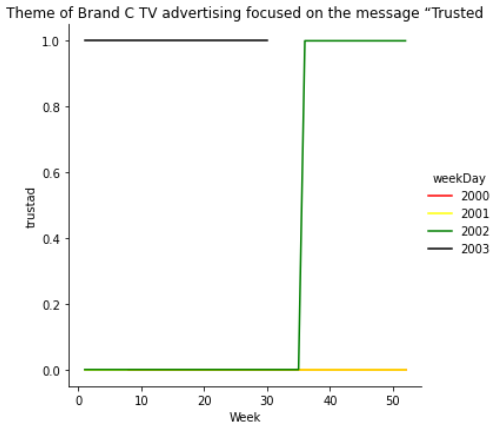
Analysis:

There's seasonality of promotions in the first three years of the dataset, highlighting that Brand C offers more promotions towards the end of the year. Since we have data till July of 2003, the same effect is not observed for the year 2003.



Analysis:

Brand C & U belong to the same company, which launches commercials for Brand U first and then Brand C.

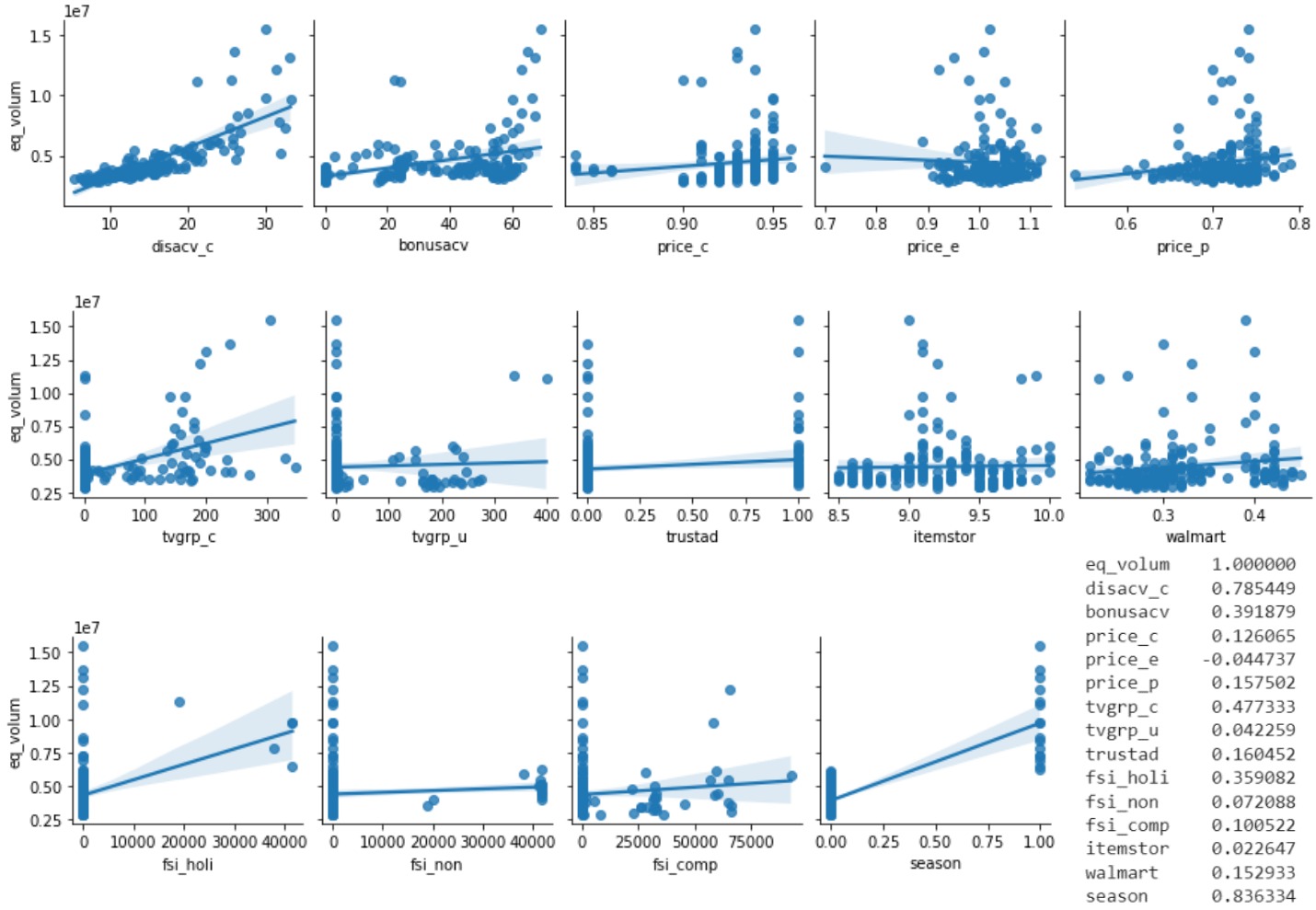


Analysis:

The theme of Brand C TV advertising focused on the message "Trusted" started in the latter half of 2002.

2. Bivariate analysis

a) X-Y plots between "eq_volum" and the independent variables.

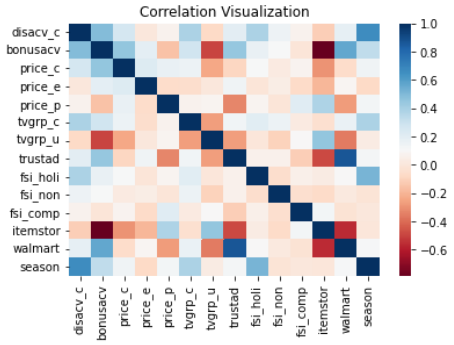



```
eq_volum     1.000000
disacv_c     0.785449
bonusacv     0.391879
price_c      0.126065
price_e     -0.044737
price_p      0.157502
tvgrp_c      0.477333
tvgrp_u      0.042259
trustad      0.160452
fsi_holi     0.359082
fsi_non      0.072088
fsi_comp     0.100522
itemstor     0.022647
walmart      0.152933
season       0.836334
```

According to the plots, Equivalent unit sales volume has high positive association with the variables "disacv_c", 'tvgrp_c', "bonusacv", which means that the value of equivalent unit sales increases as the value of either of these variables "disacv_c", 'tvgrp_c', "bonusacv" increases.

b) Do any pairs of the independent variables exhibit a high association with each other?

	disacv_c	bonusacv	price_c	price_e	price_p	tvgrp_c	tvgrp_u	trustad	fsi_holi	fsi_non	fsi_comp	itemstor	walmart	season
disacv_c	1.000000	0.491101	0.263706	0.007074	0.069647	0.390845	-0.069687	0.203785	0.393185	0.149893	0.066665	-0.115683	0.185108	0.668881
bonusacv	0.491101	1.000000	0.459770	0.191218	-0.152847	0.286355	-0.496794	0.454275	0.171479	0.109371	-0.006840	-0.785133	0.573048	0.335398
price_c	0.263706	0.459770	1.000000	0.227181	0.170993	0.158502	-0.237166	-0.085249	0.107475	0.027626	0.078172	-0.296946	-0.055367	0.143513
price_e	0.007074	0.191218	0.227181	1.000000	-0.054021	-0.050048	0.008430	0.147507	-0.007385	0.042639	-0.062295	-0.189451	0.091299	-0.067738
price_p	0.069647	-0.152847	0.170993	-0.054021	1.000000	0.072233	0.081495	-0.328522	0.084288	-0.009813	0.212704	0.379783	-0.270693	0.132422
tvgrp_c	0.390845	0.286355	0.158502	-0.050048	0.072233	1.000000	-0.264330	0.136592	0.206919	0.158131	0.026098	-0.035409	0.168593	0.391308
tvgrp_u	-0.069687	-0.496794	-0.237166	0.008430	0.081495	-0.264330	1.000000	-0.267915	-0.002769	-0.099715	0.103113	0.459060	-0.359007	0.035791
trustad	0.203785	0.454275	-0.085249	0.147507	-0.328522	0.136592	-0.267915	1.000000	0.065455	0.062609	-0.114208	-0.484284	0.871297	0.124551
fsi_holi	0.393185	0.171479	0.107475	-0.007385	0.084288	0.206919	-0.002769	0.065455	1.000000	-0.048798	0.056946	0.033497	0.102327	0.525523
fsi_non	0.149893	0.109371	0.027626	0.042639	-0.009813	0.158131	-0.099715	0.062609	-0.048798	1.000000	-0.045158	-0.059957	0.020777	-0.017664
fsi_comp	0.066665	-0.006840	0.078172	-0.062295	0.212704	0.026098	0.103113	-0.114208	0.056946	-0.045158	1.000000	0.134973	-0.018945	0.009568
itemstor	-0.115683	-0.785133	-0.296946	-0.189451	0.379783	-0.035409	0.459060	-0.484284	0.033497	-0.059957	0.134973	1.000000	-0.561871	0.001528
walmart	0.185108	0.573048	-0.055367	0.091299	-0.270693	0.168593	-0.359007	0.871297	0.102327	0.020777	-0.018945	-0.561871	1.000000	0.115371
season	0.668881	0.335398	0.143513	-0.067738	0.132422	0.391308	0.035791	0.124551	0.525523	-0.017664	0.009568	0.001528	0.115371	1.000000



- "bonusacv" and "itemstor"
- "trustad" and "walmart"
- "bonusacv" and "tvgrp_c"

3. Develop a regression model to explain drivers of Eq. Unit Sales.

OLS Regression Results

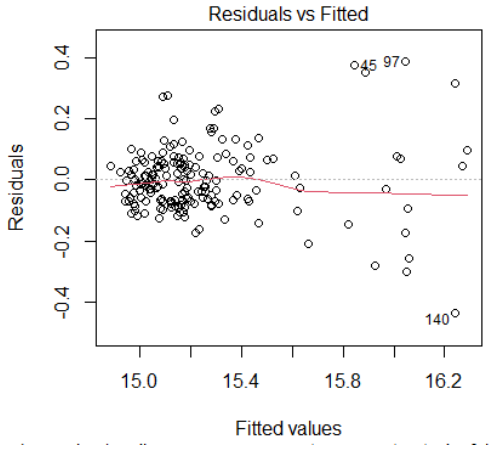
Dep. Variable:	eq_volum	R-squared:	0.879
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	176.9
Date:	Sun, 27 Mar 2022	Prob (F-statistic):	7.43e-75
Time:	01:15:05	Log-Likelihood:	140.95
No. Observations:	179	AIC:	-265.9
Df Residuals:	171	BIC:	-240.4
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	13.7800	0.244	56.485	0.000	13.298	14.262
season[T.True]	0.4485	0.044	10.195	0.000	0.362	0.535
trustad[T.True]	0.0491	0.022	2.200	0.029	0.005	0.093
disacv_c	0.0300	0.002	15.538	0.000	0.026	0.034
tvgrp_c	0.0005	0.000	4.541	0.000	0.000	0.001
fsi_holi	-5.344e-06	1.61e-06	-3.312	0.001	-8.53e-06	-2.16e-06
fsi_comp	1.213e-06	4.84e-07	2.508	0.013	2.58e-07	2.17e-06
itemstor	0.1003	0.026	3.902	0.000	0.050	0.151

Omnibus:	21.074	Durbin-Watson:	1.311
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65.048
Skew:	0.369	Prob(JB):	7.50e-15
Kurtosis:	5.859	Cond. No.	5.59e+05

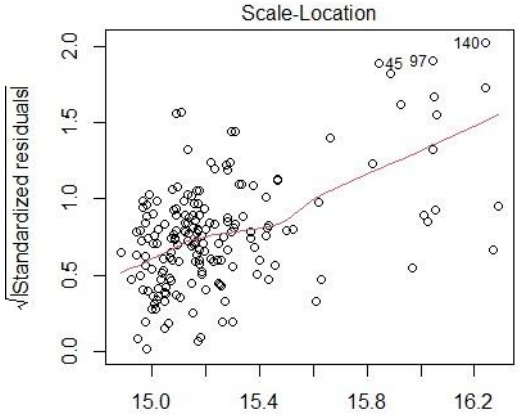
Model:

eq_volum = exp(13.78 + 0.4485 * season + 0.0491 * trustad + 0.03 * disacv_c + 0.0005 * tygrp_c + (-5.344) * 10^{-6} * fsi_holi + 1.213 * 10^{-6} * fsi_comp + 0.1003 * itemstor)



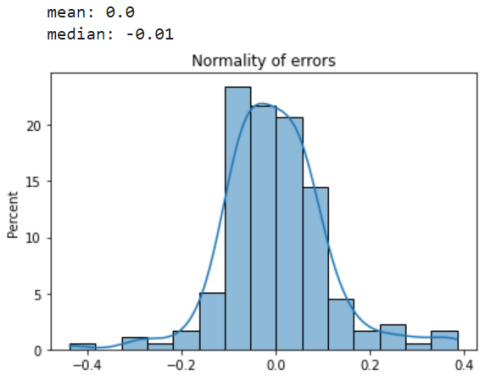
1) linearity

We find there is discernable pattern, like a lied funnel, of residualsfor each possible fitted value, the relationship is non-linear.



2) homoscedasticity of errors

Horizontal line with equally spread points is a good indication of homoscedasticity. However, the line here is sloped. Thus, we have a heteroscedasticity problem.



3) Normality of errors

According to the mean and median, the distribution of residuals is normal distribution.

4) Independence of residuals

Durbin-Watson:	1.311

Durbin-Watson score is 1.311. Thus, there's a positive autocorrelation

4. Justify your choice of independent variables in the model.

a) Why did you choose to include or exclude certain variables?

- We used log-transfrom on "eq_volum" because of non-linear relationship between dependent and independent variables.
- We excluded time-related variables "weekDay", "month", "weeknumber" and "year" from the final model since there is no big year-to-year differences in data.
- Because of multicollinearity(> 5), variables "walmart", "bonusacv" are deleted.
- Because of insignificant P-value(> 0.1), variables "tvgrp_u", "price_e" are deleted.
- For business decision, "ROI" should be considered since data collection needs cost. Therefore, significance level is made more stringent to 0.05. The variables "fsi_non", "price_c", "price_p" are deleted based on P-value (>0.05). The R^2 of the final model decline by 0.004.
- We decided to exclude interaction terms based on strong multicollinearity or insignificance
- We added a new dummy variable "season" to identify anomalies in the data collection (extreme data points).

b) How did your account for seasonality in your final model?

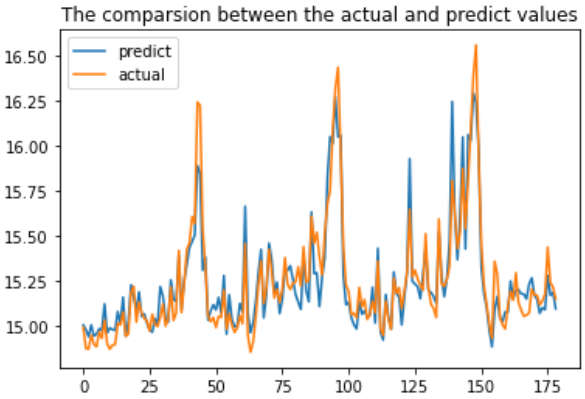
We created a new independent dummy variable, "season" to account for seasonality in the data set. We set up the cutoff between 0 and 1 based on the outliner definition of box plot, which represent values beyond IQR *1.5.

c) Briefly comment on multicollinearity and how we went about addressing it in your final model. (Generated by R)

```
disacv_c   season  tvgrp_c  trustad fsi_holi fsi_comp itemstor
1.966362 2.221914 1.231311 1.361511 1.395008 1.037804 1.340664
```

Since the VIF values are within limits (<5), there is no multicollinearity issue in the final model. We decided to keep all the variables.

5. Comment on the in-sample model fit (R2) and significance level of the independent variables for your final model. Include a time series plot that reports the actual as well as predicted (i.e., model based) eq_volum_c.

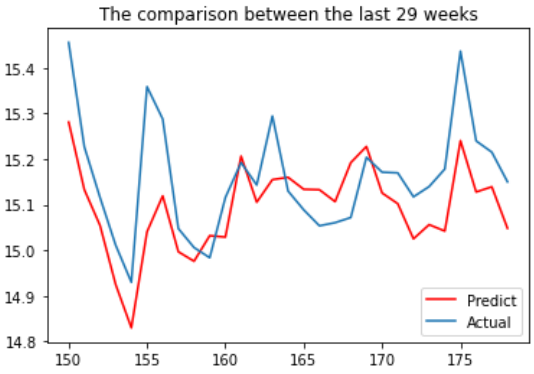


Adj. R^2 is 0.8737.

The independent variables are significant with significance level of 0.01.

According to Adj. R^2 and the plot, we are satisfied with this model.

6. Out of sample prediction



Finding:

a) The model almost catches every elbow points. Most of time, the predicted values are lower than actual values.

```
MSE for Train set: 0.0126448682532683
MSE for Test set:  0.012357539377568734
```

b) We checked the MSEs, which show there is no overfitting happening.

Comment on How well: It's good!

7. Using the final model, what can you say about the effectiveness of the different marketing activities for Brand C (e.g., price discounts vs. couponing vs. advertising)? Which marketing activities seem to have a stronger effect on sales? What can you say about the effect of competitive activities on Brand C sales?

The variable "disacv_c" corresponds to price discounts, "fsi_holi" corresponds to couponing, "trustad" and "tvgrp_c" corresponds to advertising. According to the respective coefficients of these variables in the final model, advertising is the most efficient method because it has highest coefficient. Couponing is the least effective one because of lowest coefficient. Thus, advertising has a stronger effect on sales.

Competitive activities have a unique effect on Brand C sales. Based on the bivariate analysis, as the price of Brand P increases, the sales of Brand C increases. The price increase of Brand E doesn't have significant impact on the Brand C's sales. Also, we observed that higher the FSI of competitors ("fsi_comp"), higher the sales of Brand C.

8. What if analyses:

a) The way to manage Brand C by final marketing mix model

eq_volum = exp(13.78 + 0.4485 * season + 0.0491 * trustad + 0.03 * disacv_c + 0.0005 * tygrp_c + (-5.344) * 10-6 * fsi_holi + 1.213 * 10-6 * fsi_comp + 0.1003 * itemstor)

First, we should invest in the variable with positive coefficients because of direct positive relationship (it will result in more sales if we invest more in them), but except for the dummy variable "season", and variable "itemsotr" since they're not under our control. Second, given the efficiency of investment for each variable, we prioritize the variables under control i.e., variable "trustad", "disacv_c", "tygrp_c", "fsi_comp" according to their coefficients. Third, according to the definition of each variable, we should strengthen the promotions by offering more and timely discounts, commercials and coupons to increase EQ volume.

b) one specific marketing mix related question that could be answered by your model and then answer that question.

Question: What is the incremental sales volume when Brand C TV advertising focused on the message "Trusted", keeping all the other variables same?

```
with trusted ad       4906161.70
without trusted ad    4671000.01
dtype: float64
The difference between "with trusted ad" and "without trusted ad" is 235161.69
```

Answer: The sales volume with the theme of Brand C TV advertising focused on the message "Trusted" is 235161.69 more than the sales volume without it.

9. What are some of the limitations on the inference you can draw from the above model? How could you overcome these limitations?

We tried more than 15 models to finally narrow down to one. Based on the analysis and discussion, we identified some of the limitations in the model.

1) The cutoff for the outliers' values is calculated by IQR * 1.5, which is a rough estimate. We can more accurately identify outliers by using "Structural Break" method.

2) It is difficult to tell if there is overfitting in the model. This issue can be solved by "learning curve" method.