Traffic Volume Price Prediction Challenge

Steps Followed

i. Loading various library required for analysis, manipulation and model building.
ii. Loading data and Basic Sanity Check
iii. Data Analysis with various plot and inferences
iv. Feature engineering
v. Model Building
vi. Model Evaluation
vii. Finalizing Model

Loading various library required for analysis, manipulation and model building

Various Library used in this case Study: -

i. NumPy (1.18.5)
ii. Pandas (1.1.3)
iii. Matplotlib (3.3.1)
iv. Seaborn (0.11.0)
v. SKlearn (0.23.2)
vi. Xgboost (1.2.0)

Numpy

NumPy is used here for performing various mathematical operation and Pandas is heavily dependent on NumPy.

Pandas

Pandas is used here loading data and converting data to data frame, so data manipulation become easy.

Matplotlib and Seaborn

Both Matplotlib and seaborn is used for plotting various plot like bar plot, box plot ,scatter plot ,heatmap etc. Visualizing data make easier to understand when we have lot of data in data frame.

SKlearn

Sklearn comes with various machine learning algorithm, preprocessing and model evaluation, which make easier to implement various machine learning algorithm and evaluation of them. In our case study we used sklearn to implement Lasso, Ridge regularization and ensemble based machine learning algorithm such as Random Forest. Apart from them we have used sklearn for some preprocessing such as scaling data. For performing cross-validation.

XGboost

Xgboost is used for applying xgboost algorithm.

Loading data and Basic Sanity Check

With the help of pandas we have loaded data as “df_train” and “df_test”. Df_train is used for training data and part of data is also used for evaluation of model as well.

In basic Sanity check, We have found that:

Based on train data

i. shape - 38563x9
ii. No null value is found.
iii. 1 Column is Date-Time
iv. 3 Columns are of Object Data type
v. 5 Columns are of numerical Data type.
vi. **Traffic_volume** is our dependent column.

Based on test data

i. shape - 9641x8
ii. No null value is found.
iii. 1 Column is Date-Time
iv. 3 Columns are of Object Data type
v. 4 Columns are of numerical Data type.
vi. Dependent feature is not present.

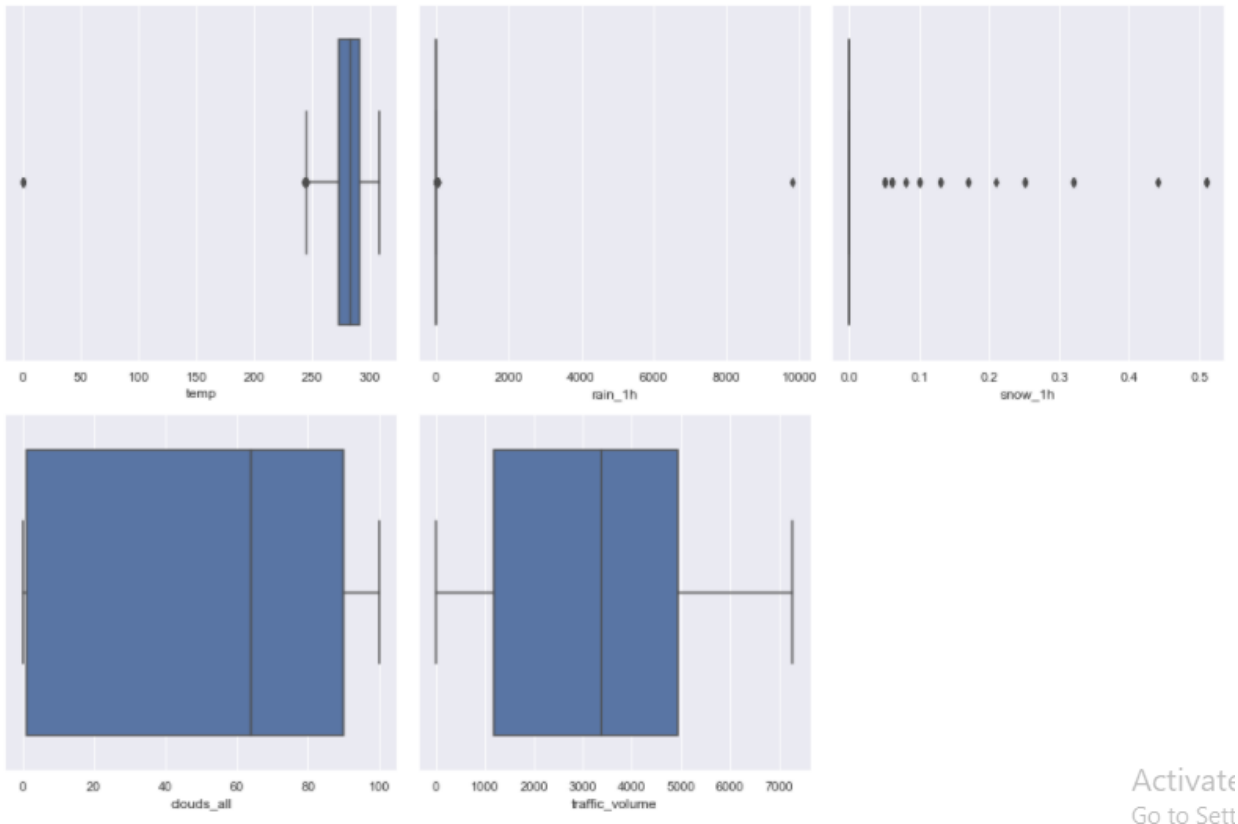
With the help of date-time column we derived some features like weekday, hour, month and year.

	temp	rain_1h	snow_1h	clouds_all	traffic_volume	Month	Year	Hour	Weekday
count	38563.000000	38563.000000	38563.000000	38563.000000	38563.000000	38563.000000	38563.000000	38563.000000	38563.000000
mean	281.351757	0.392733	0.000278	49.920364	3260.940409	6.580894	2014.934393	11.397635	2.985712
std	13.216927	50.075055	0.009131	38.849106	1991.628329	3.394472	1.665401	6.949958	2.004087
min	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	2012.000000	0.000000	0.000000
25%	272.858000	0.000000	0.000000	1.000000	1186.500000	4.000000	2013.000000	5.000000	1.000000
50%	282.750000	0.000000	0.000000	64.000000	3378.000000	7.000000	2015.000000	11.000000	3.000000
75%	291.540000	0.000000	0.000000	90.000000	4939.000000	10.000000	2016.000000	17.000000	5.000000
max	308.240000	9831.300000	0.510000	100.000000	7280.000000	12.000000	2017.000000	23.000000	6.000000

If we observe above table, we can say that rain_1h, snow_1h. have outlier as there is huge difference between 75 percentile and max value.

Data Analysis with various plot and inferences.

Checking For outlier



We can see temperature(temp) has outlier in lower quantile near 0, Which is seem error 0 kelvin is very low temperature.

Rain_1h and snow_1h have outlier in upper quantile. Both value is given in mm. we don't have enough evidence or knowledge to say anything about them. We can make inferences that 0 in rain_1h and snow_1h means that no rain or snowfall happen on particular time.

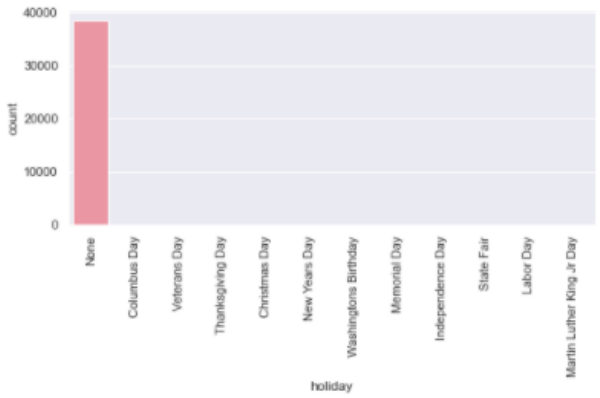
If we see clouds_all, we can say that major of data is below the average or mean, no outlier is present.

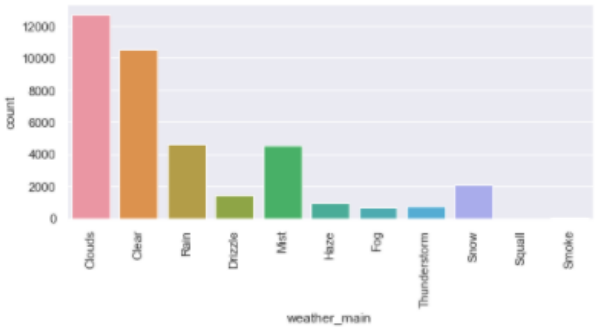
Traffic_volume is dependent or target variable, doesn't have any outlier, mean is around 3100-3500.

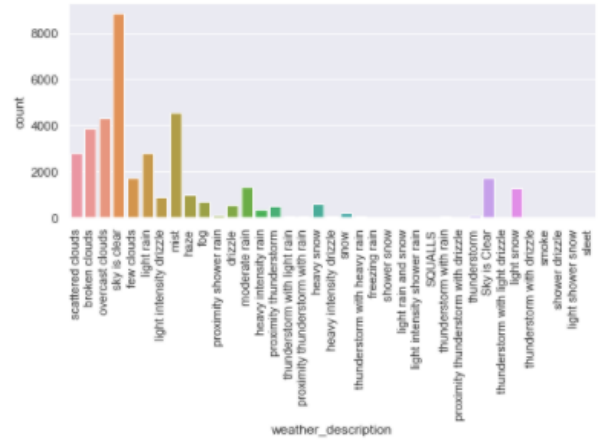
Mean is slightly above the median.

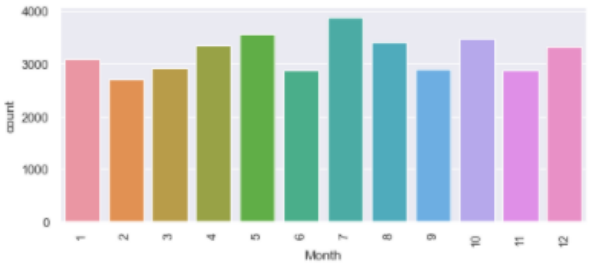
Action taken:-

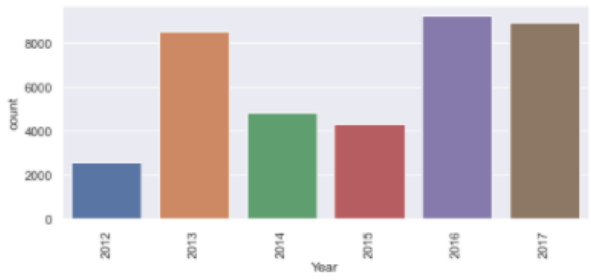
We used IQR method and soft range of 0.01 and 0.99 to handle outlier below .01 and above .99 is considerd as outlier, We don't want to loose much data that's why used soft range.

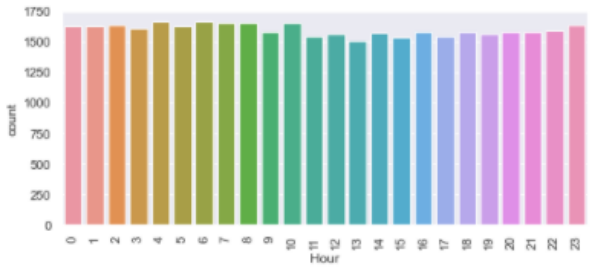


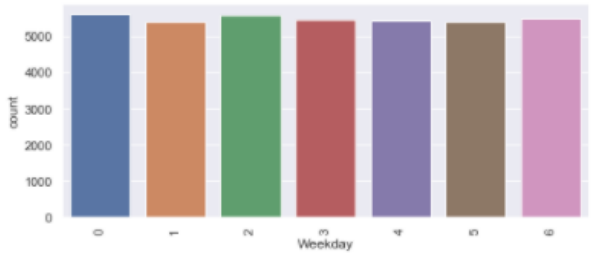










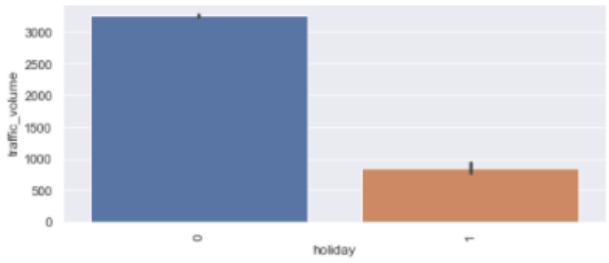


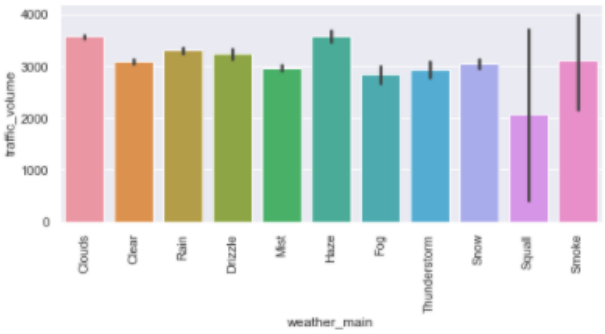
We can see count plot of various categorical features

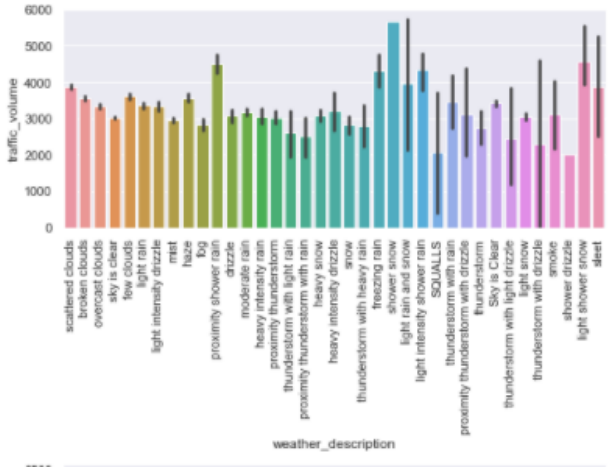
We can see that in holiday we have almost same value . We will convert holiday as binary column If no holiday 1 else 0.

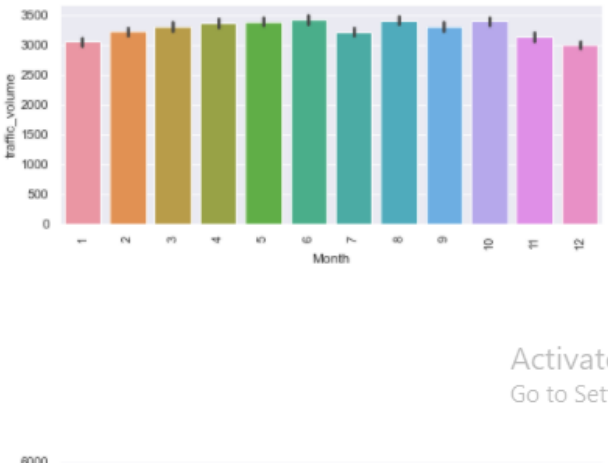
We cannot see any skewness in categorical columns

With Respect to Target Column









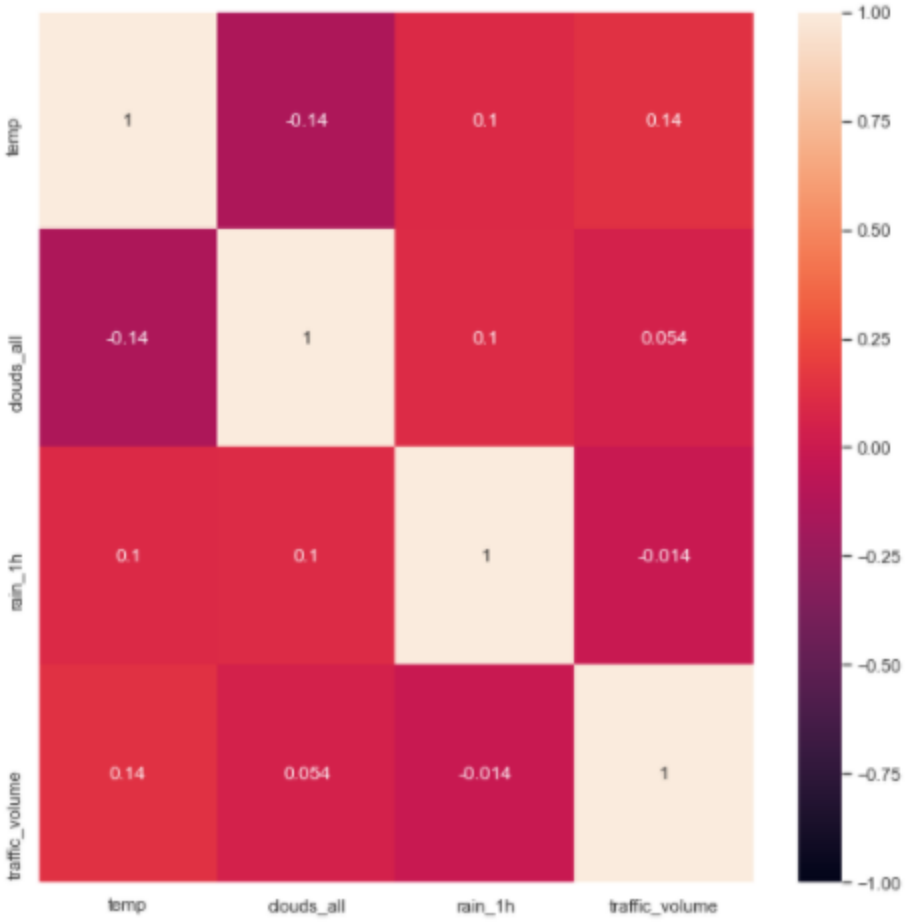
If we see plot between traffic volume and holiday, we can see that traffic is less on holiday. But we cannot rely on this as above we have found(count plot) that there is only 47 holiday and 37144 value in non-holiday which can be reason.

If we hour vs traffic volume traffic is less between(0-4) i.e. midnight.

Traffic volume start increasing from 5 in morning reach peak at 16 and again start decreasing

If We see weekday vs traffic, traffic is high in normal days compare to weekend

If We see weather_main vs traffic, fog, mist, snow, squall effect traffic.



We can see rain1h has negative impact on traffic volume but not so much.

Apart from that we can see that there is no so much correlation between features

Feature engineering

Generally hours, weekday ,days are cyclic in nature and they are need to treated as cyclic.

A common method for encoding cyclical data is to transform the data into two dimensions using a sine and cosine transformation. We have done cos and sine transformation on weekday, month etc.

There are some categorical column , we need to do encoding in this case study we have used

i. Mean encoding
ii. One-hot encoding

Column such as weather description, wether detail description have so many categories in it using one hot encoding will increase dimension and they are nominal categorical column so we used mean encoding to handle them. In mean encoding we replace value with mean of target value for particular categories.

One- hot encoding is used for year column.

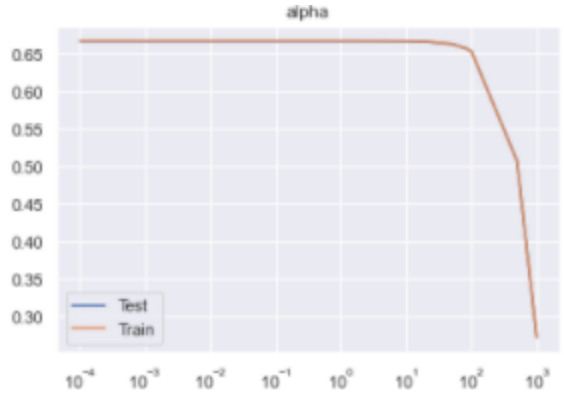
Model Building and Model Evaluation

We Started with Lasso and Ridge Regularization

In lasso

we got adjsusted r2 score of 66 in both train and test set after performing hyper parameter tuning .

we tuned alpha.



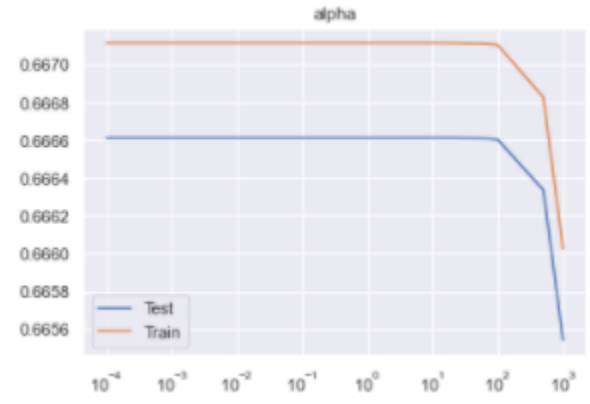
Alpha value tuned .00001 to 1000

We found that model start underfitting after 100

Possible Regression equation can be

traiff_volume = 3260.6 + (- 11.5 * holiday)+ (24.59 * temp)+ (-14.9 * rain_1h)+(-4.1 * clouds_all)+(-649 * hour_sin)+(-1441.3* hour_cos)+(116.68 * wd_sin)+(-314.39 * wd_cos)+(0 * month_sin)+(-74.02 * month_cos)+(-14.1 * weather_main_encoded)+(25.47 * weather_desc_encoded)+ (17.1 * Year_2013)+ (-6.58 * Year_2014)+ (-3.19 * Year_2015) + (-24.27 * Year_2016)+(48.62 * Year_2017)+(0* Year_2018) .

In ridge regression,

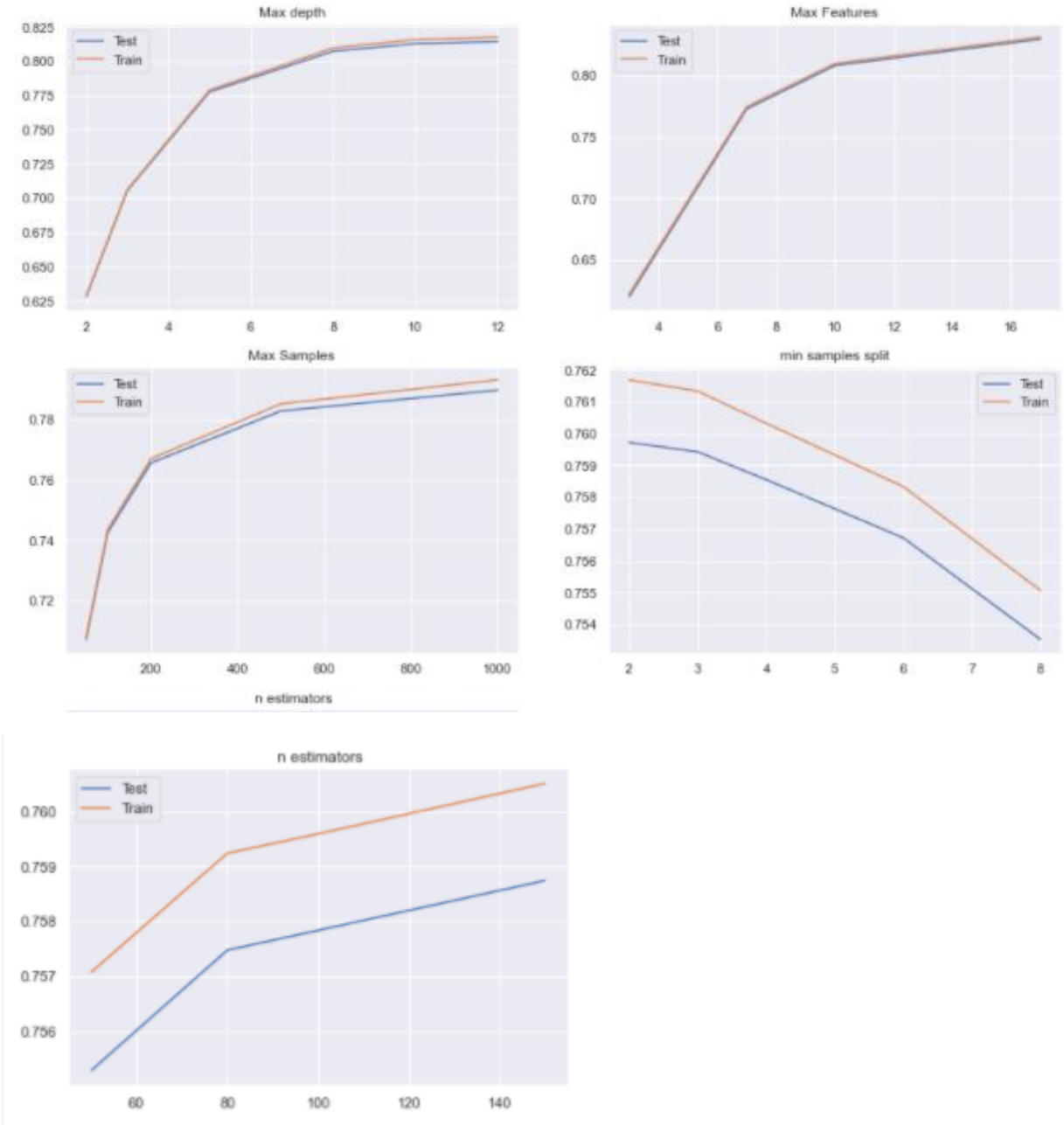


We can see same behaviors as lasso

Ensemble, Boosting based technique

Random Forest with hyper-parameter tuning

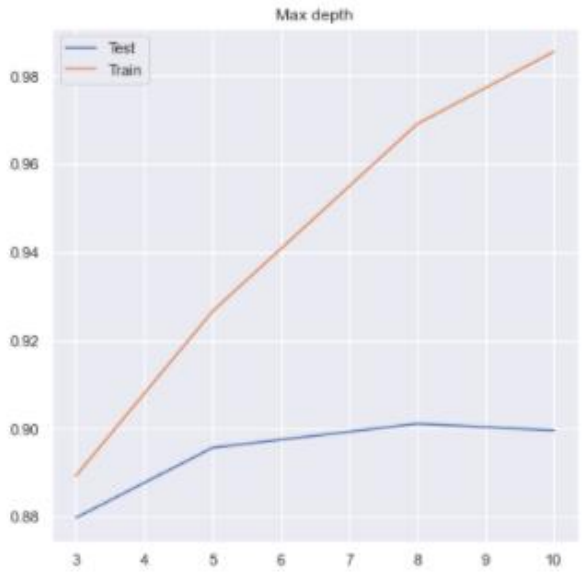
We got 88 r2 score on holdout set, which is good improvement from lasso and ridge.

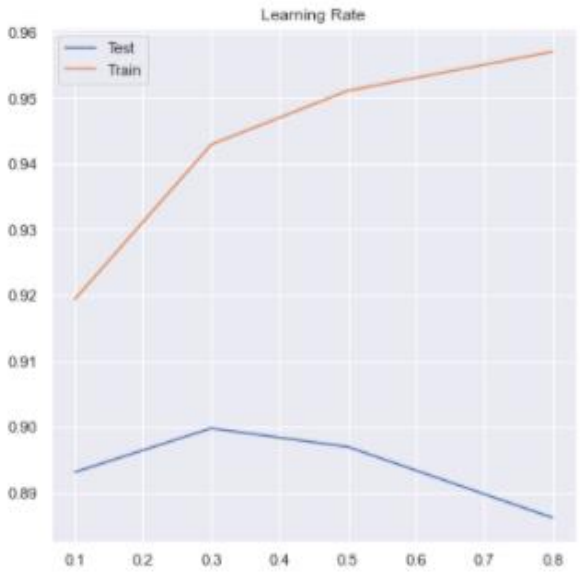


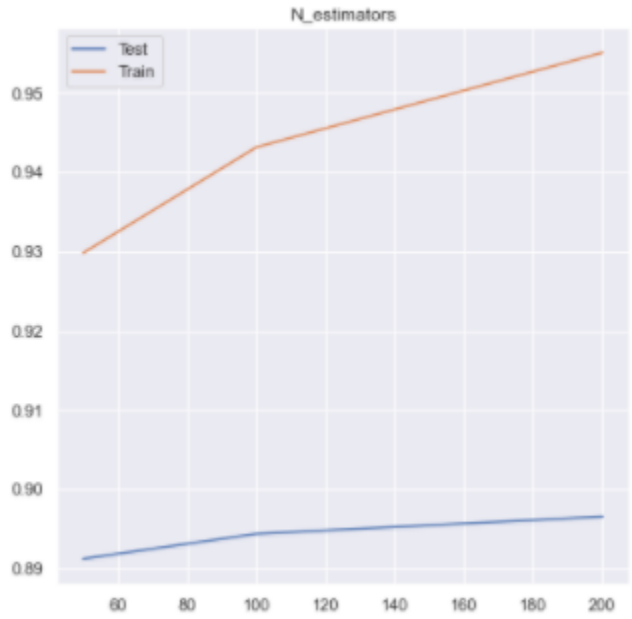
We can see that as maximum depth, number of estimator maximum number of features increases r2 score also increases but after some limit, gap between test and train score also increases which means model start overfitting.

XG Boost Regressor With Hyparameter Tuning.

We got 91 r2 score in hold out set.







We can clearly see that train score keep increasing but test score start decreasing as parameter increases which leads to overfitting of model .

Finalizing Model

We have made total of 4 model random Forest, extreme gradient Boosting L1 and L2

We have Found that model With Xgb algorithm performed very well in holdout set

If We compare Standerd Deviation of test score we r getting lowest in xgb, 0.003426 whch means our model stable, with the above discussion we can finalize xgboost regressor model.