TEAM 35

STOCK MARKET PREDICTION

Work Distribution:

The initial data collection and sources of information were provided by Ronak, who surfed online to compare different algorithms and ways to optimize the models.

The practical implementation of several models and approaches were done by Aniket to get a brief idea about different models and their implementation hardships.

The moving average on the VTI dataset was done by Aanshul, hyperparameter tuning and optimization was also done by Aanshul.

Deep studied new approaches involving boosting techniques such as XG Boost and how to optimize every single hyperparameter. Mathematical learning method - linear regression was also implemented by him to find the motion of stocks in recent history.

Documentation in terms of maintaining GitHub repo and project proposals, etc was done by Ronak and Aanshul.

Post mid evaluations, different research papers were collected and studied by Aniket who explained the team novel approaches and helped set up the pipeline for the work.

Aanshul performed the feature selection task using Lasso and Correlation. Linear regression and Lasso as well as Ridge regression models were implemented by him.

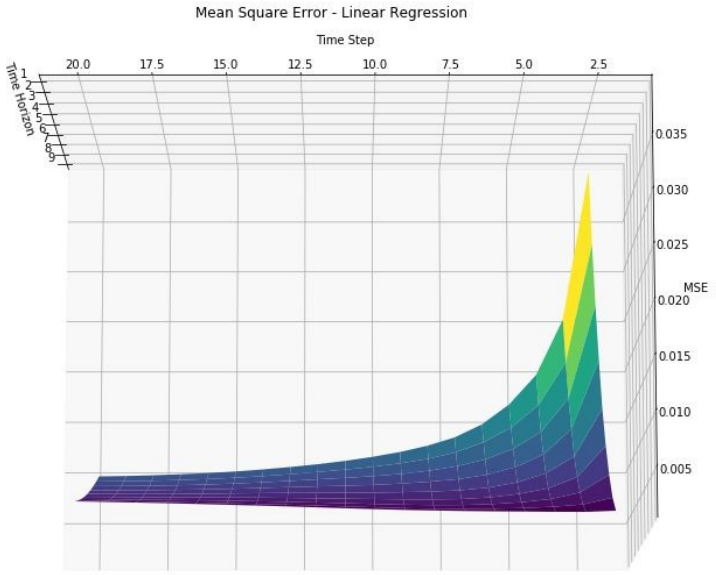
The data collection was done by Ronak using the API key and extracting intraday and historical data. Data preprocessing and data cleaning was done by him.

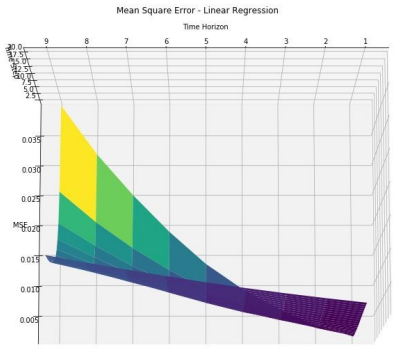
XG Boost implementation was done by Deep and Shrimal and he later modified and optimized the hyperparameters to build the best model suited for the problem. LSTM technique was implemented by

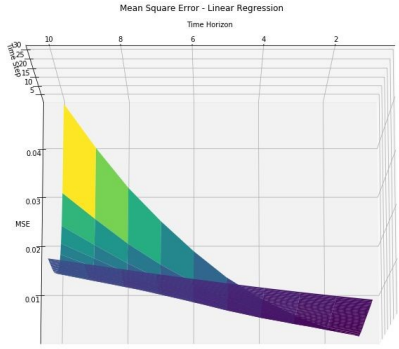
Aanshul and Ronak performed the task of hyperparameter optimizations. A similar distribution of work was done for the HMM part as well.

Linear Regression:

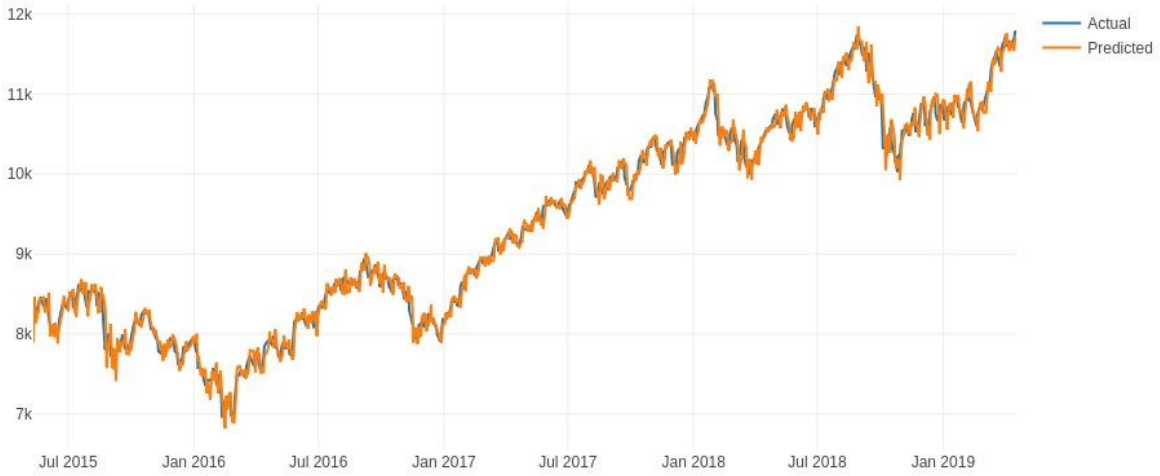
A quick display of all the work done is presented because it resembles the same content as the presentation slides.







This shows that for higher time horizon prediction, higher time steps is beneficial. But overall, the prediction for the time horizon of 1 day gives the best result.



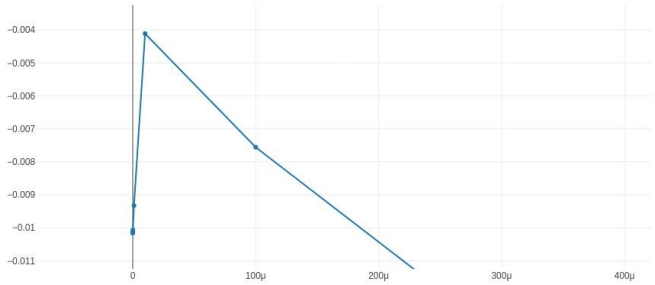
This is the time series prediction graph for linear regression with a time step of 3 and time horizon of 1 day.

This model helps to understand the motion of the stocks.

Lasso and Ridge Regression:

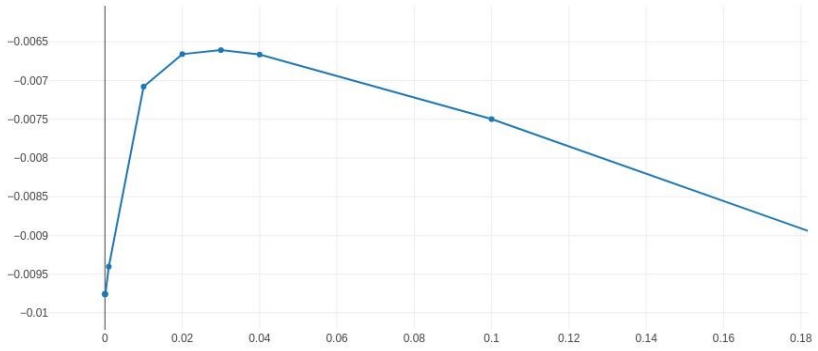
Lasso model was used as an embedded method for feature selection. Ridge regression was used to reduce the weight of coefficients which didn't help to improve the predicted value of the stock price.

The hyper tuning parameter steps are mentioned in the slides, hence we will directly proceed to results and observation.

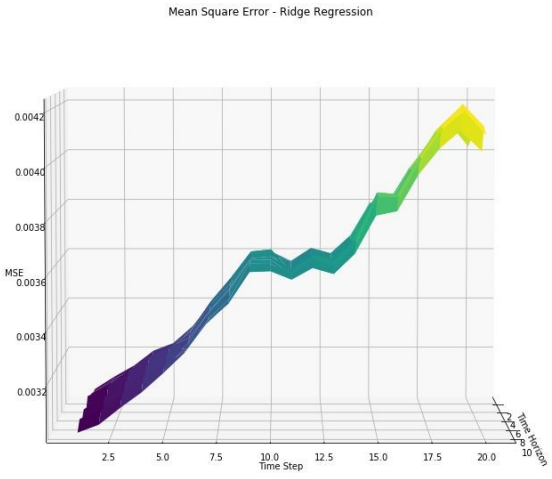


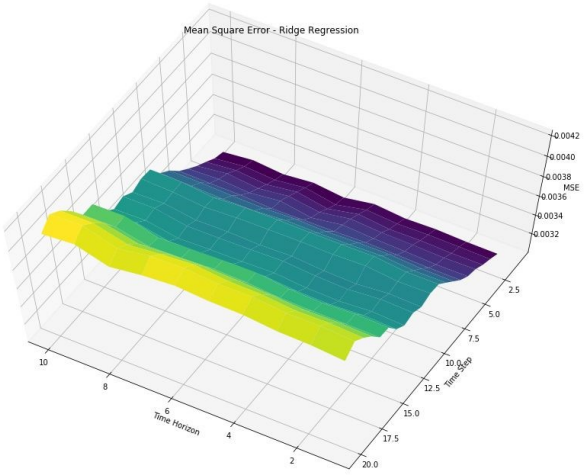
This is the curve for negative mean square values, hence higher the curve point and lowers the mean square error.

Following is the alpha selection curve for ridge regression.



The dataset was trained with optimized hyperparameters and following are the respective graphs of lasso testing and ridge testing.





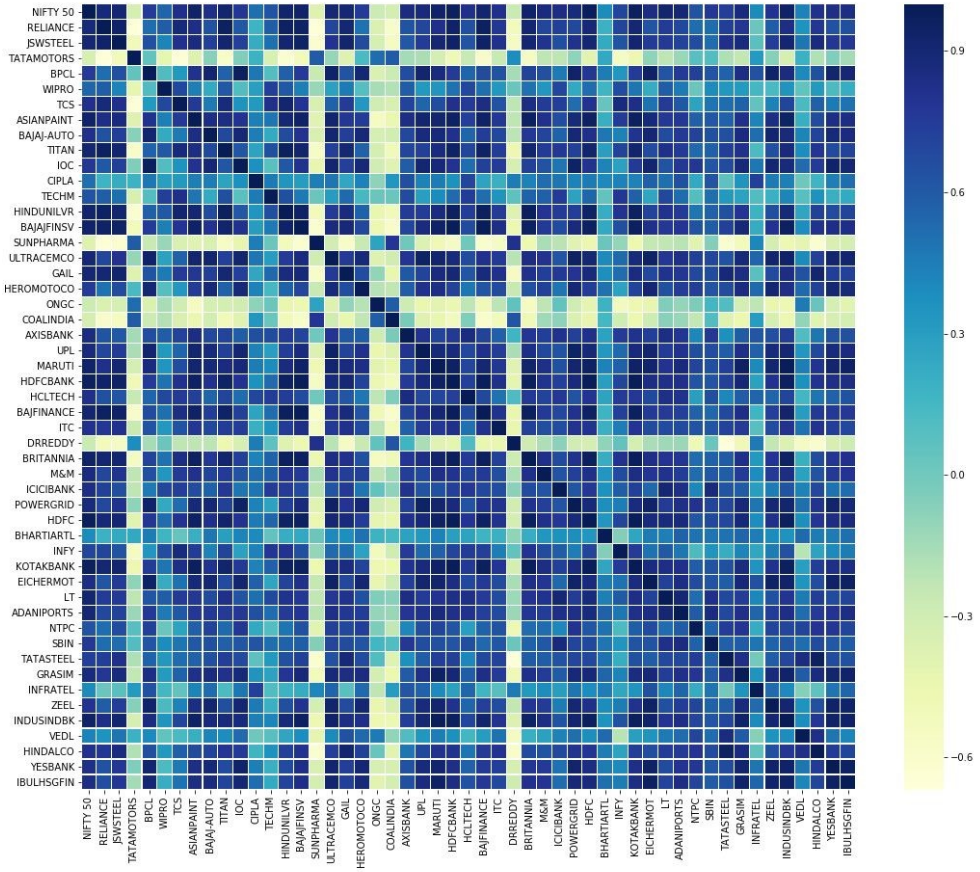
This shows the efficiency of a time step for a particular time horizon prediction.

The last curve for ridge regression takes in previous 10 values and predicts the next 20 values and has a low MSE of 0.0027. Hence, this proves vital if we want to predict values with the higher time horizon.

LSTM and HMM:

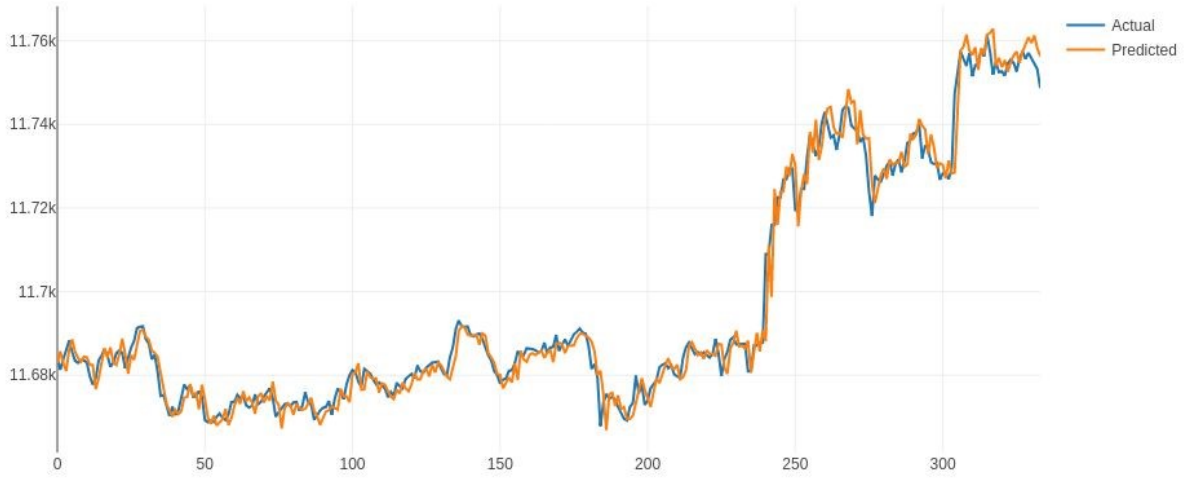
These models were implemented as a part of Assignment 10, hence skipping out the general introduction of the models.

The features used in LSTM uses the extracted features of Lasso regression. Also, Pearson's Correlation was used to find optimal features having the highest correlation coefficient. The threshold value was set to be 92.5% and still, we got 9 features as a part of this wrapper method feature selection.



Here, LSTM and HMM are used to predict the intraday stock values of NIFTY 50. LSTM uses the past values as a part of learning, whereas HMM uses the relative changes in the open price, close price, and the difference between high and low prices for the minute.

Hence, LSTM and HMM try to look out for historical presence, one based on learning and the other based on probability distribution.



The curve above is the test predicted values of LSTM model, which is used to recognize the change in trends.

XGBOOST:

Since XGBOOST gave a pretty good performance in the previous models for performing stock predictions. We are using it once again for performing NIFTY-50 predictions.

Approach:

- Preprocessing the data in a somewhat same manner as in the previous case.
- Now Using the XGBOOST model for training and fitting the data.
- Making predictions for the next day closing value of NIFTY-50
- Performing Hyperparameter Tuning.

But what different we did this time:

Last time we only did a prediction of the closing price for the next day only using a fixed set of features only.

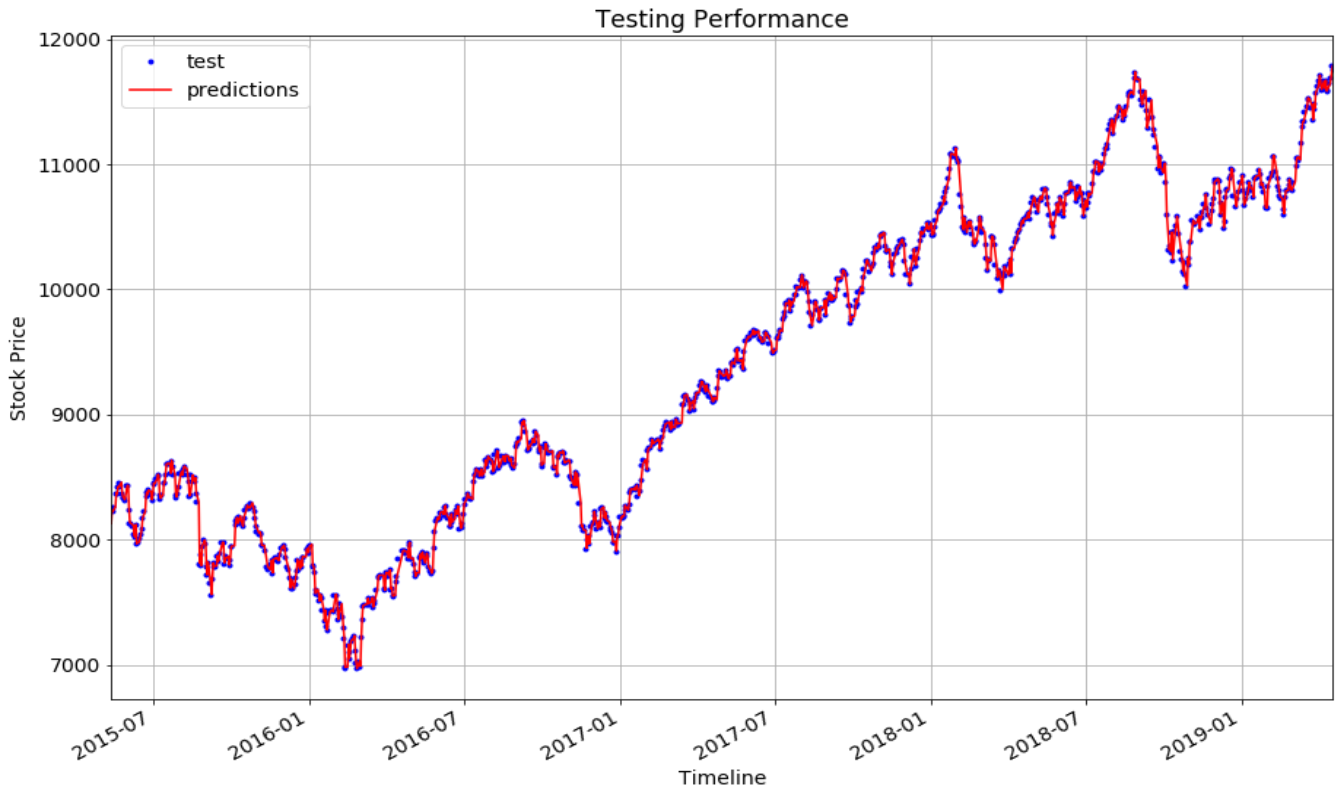
But this time we tried a few different things and tried multiple sets of features for predicting the next value:

- Intraday trading: using Opening price, High, Low and closing price of the previous N days from predicting the closing price of NIFTY.
- Intraday trading: using only the current opening, high and low values of the stock to predict the value for the next minute.
- Intraday trading: Using the same features as above but also including the features of previous N minutes also.
- For the above-mentioned approaches, we used the value of Nifty-50 stock prices only. So now we decided to take some k companies out of the 50 companies whose stock prices decide the value of Nifty. Using the stock price of that k companies, along with the nifty stock prices we are predicting the next minute's nifty value.

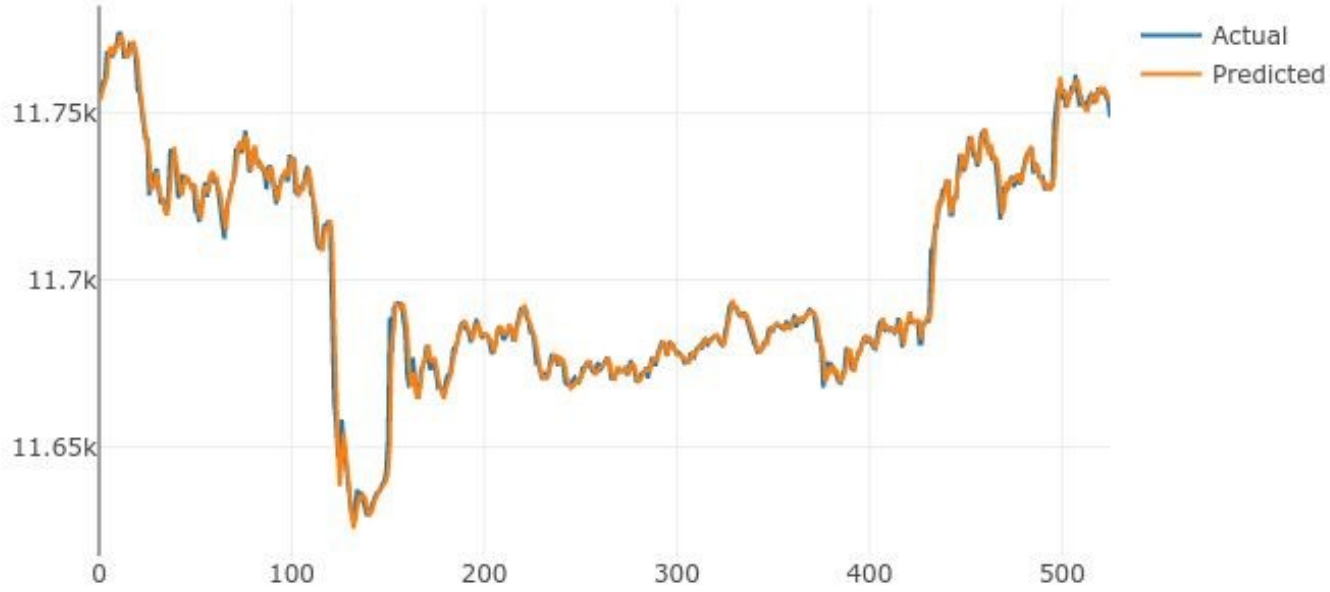
Performance, results, and Outputs:

Interday trading predictions:

Dataset	MAPE	RMSE
Testing	0.613%	74.705

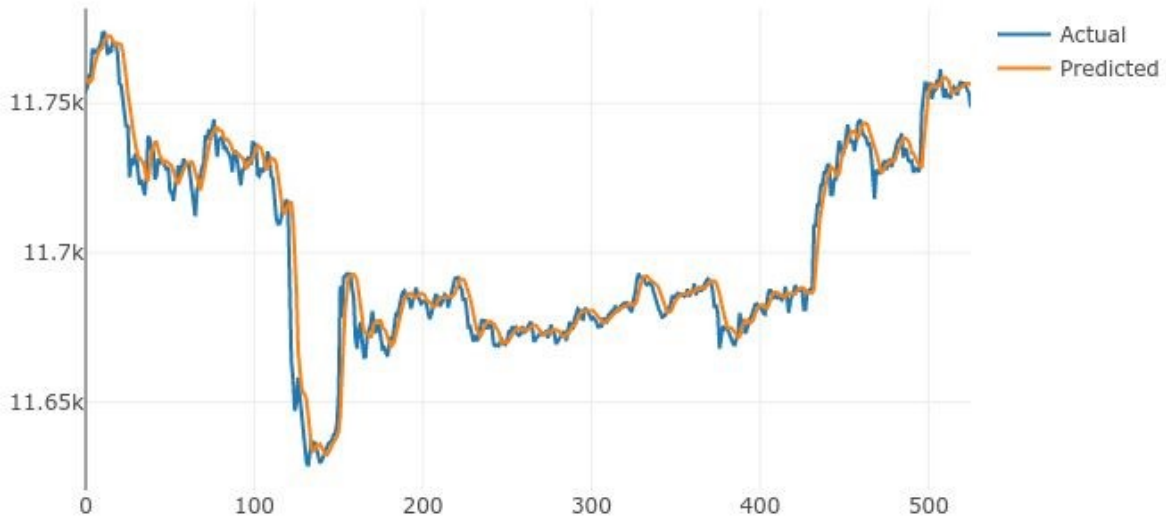


Intraday without lag:



Dataset	MAPE	RMSE
Testing	0.018%	3.789

Intraday with lag:



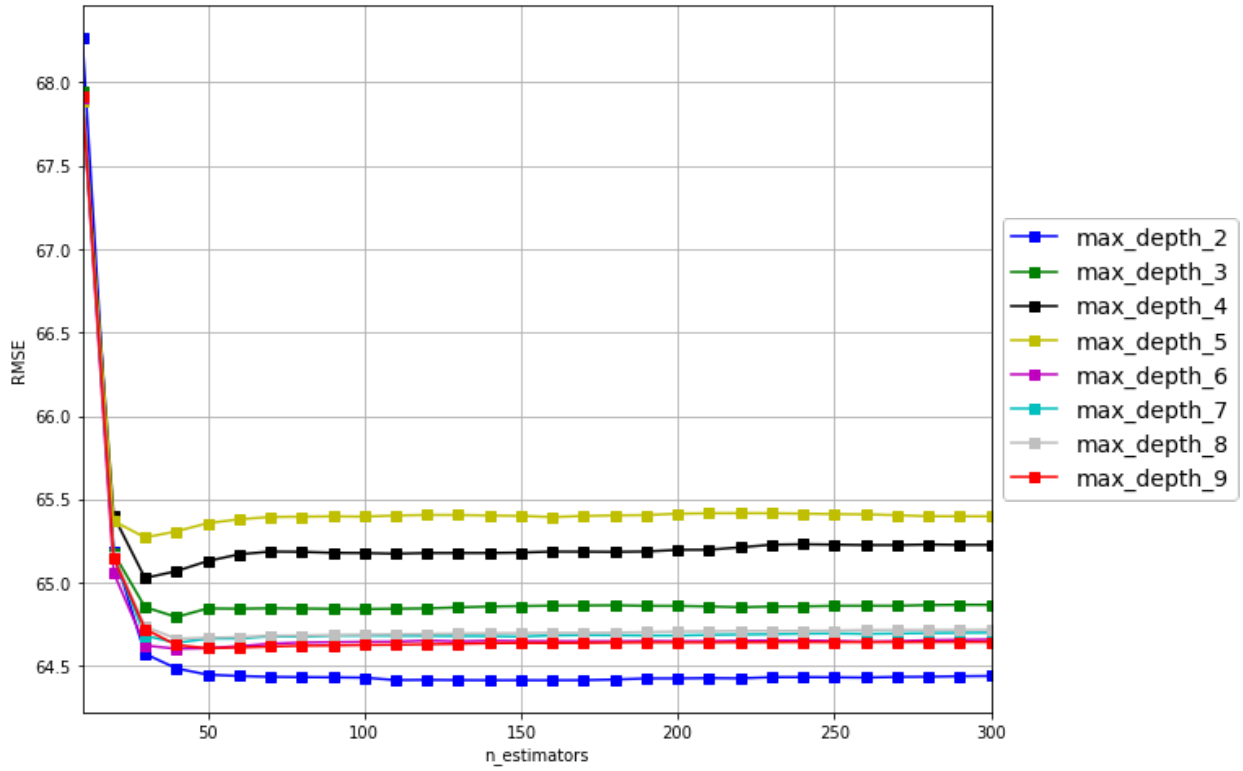
Dataset	MAPE	RMSE
Testing	0.39%	7.69

Hyperparameter Tuning:

For each model selecting a range of some specific values for each parameter and fine-tuning for the same is performed to get the best set of hyper-parameters.

Below are the graphs for Error measure vs value of hyperparameters for interday model:

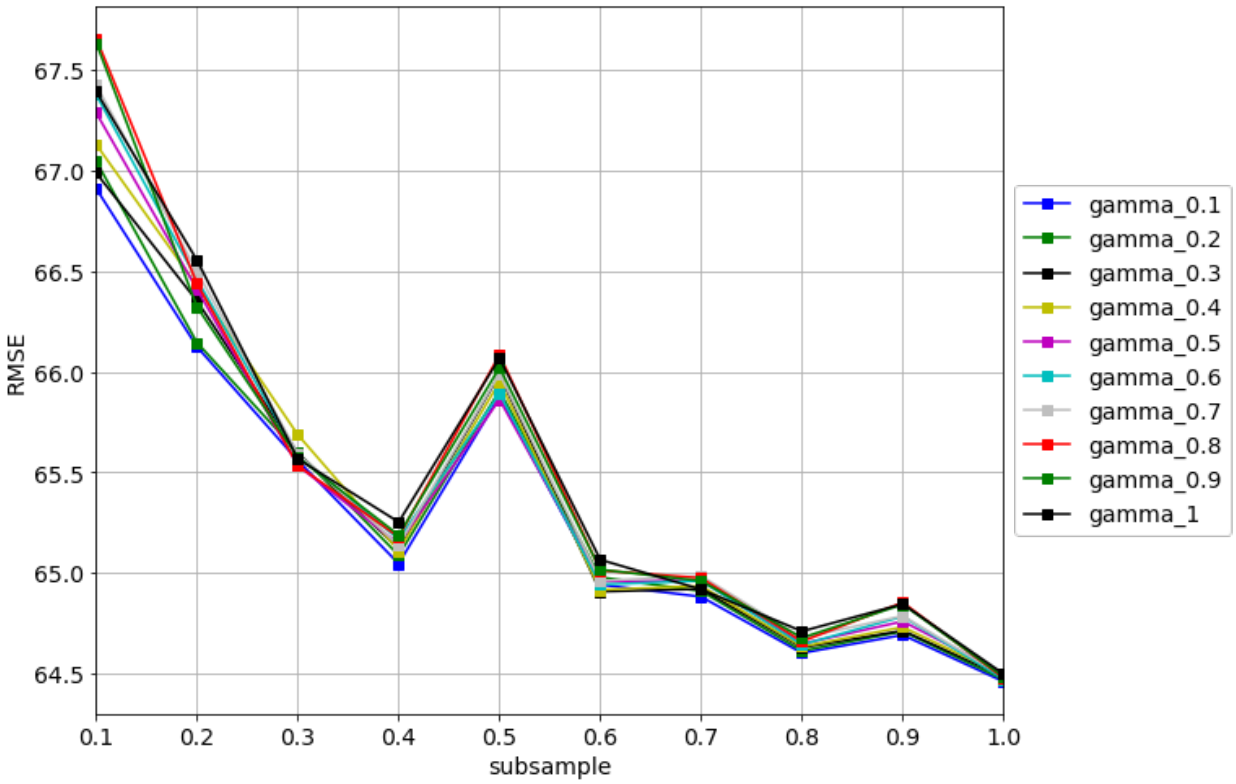
- Max_depth of tree and no_of_estimators vs RMSE



Parameters value selected after tuning:

- Max_depth = 2
- N_estimators = 160

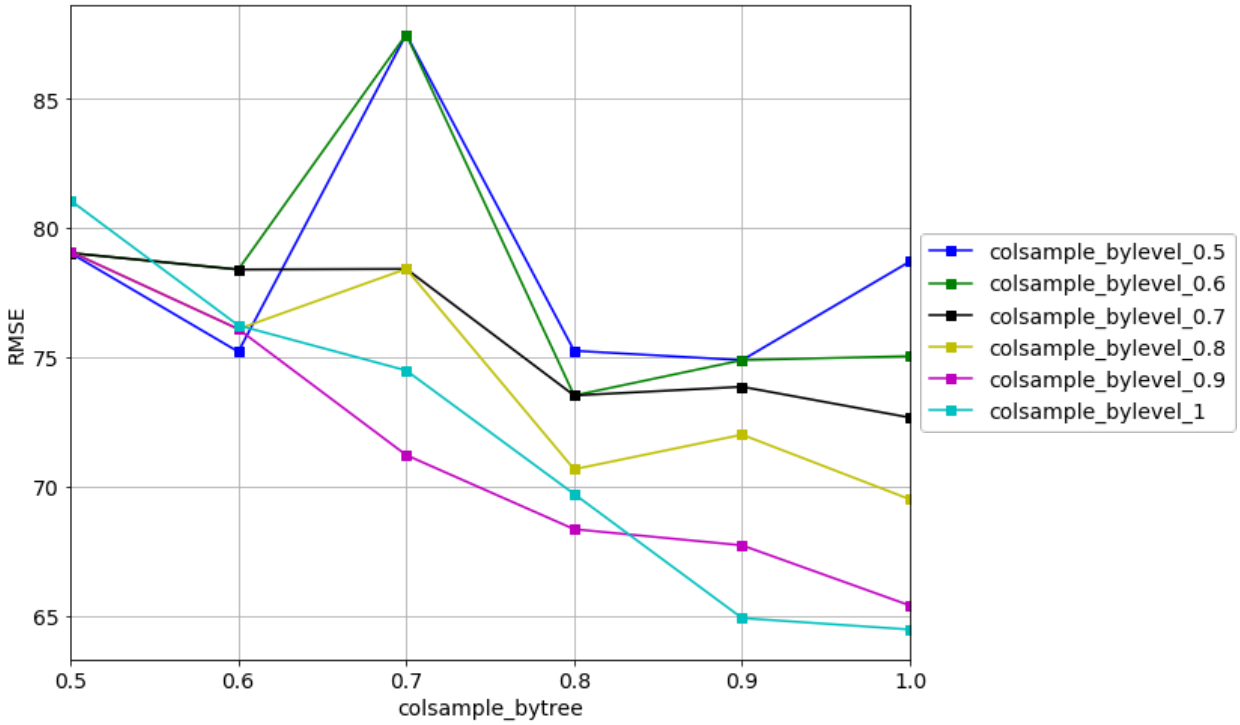
- Subsample and gama value vs RMSE



Parameters value selected after tuning:

- subsample = 1
- N_estimators = 0.1

- Comsample_by_tree and colsample_by_level VS RMSE.



Parameters value selected after tuning:

- colsample_by_tree = 1
- colsample_by_value = 0.1