**INTRODUCTION:**

The attached “SALES” excel file contains two tabs, “TRAIN” and “TEST”. These tabs are of fictitious sales numbers for the business.

* The “TRAIN” tab contains records pertaining to sales for each day. Use this tab as training data for your models.
* The “TEST” tab contains two months of 2022 sales. Use the “TEST” tab to measure how well your model predictions compare to the actual sales in the “TEST” set.
* Please do not use any other time period for Train or Test.

There are 16 features / independent variables (Columns A through P) and one dependent variable: SALES (column Q).

**INSTRUCTIONS:**

Use this data to build AT LEAST two (2) models designed to predict SALES:

1. Build AT LEAST one univariate timeseries model in which you ONLY use the SALES variable in Column Q. Fit the model using the TRAIN set data and then evaluate your model’s predictions against the data in the TEST set.
2. Build AT LEAST one additional model (preferably using a different algorithm than the first model) in which you use all (or any combination you choose) of the independent variables in Columns A through P. In other words, the second model should be a multivariate model using more than one predictor variable from Columns A through P to predict SALES.

**DATA:**

* Do not change the values in “year” (column A) or “forecast\_month” (column D).
* For example, you may notice that 2019-12-30 is coded as the year 2020 and as the month of January, as opposed to year 2019 and month December. This is intentional!
* This is because the majority of the days in that week fall in January, not December. And the majority of the days in that week (5 out of 7) fall in 2020, not in 2019.

**DELIVERABLES:**

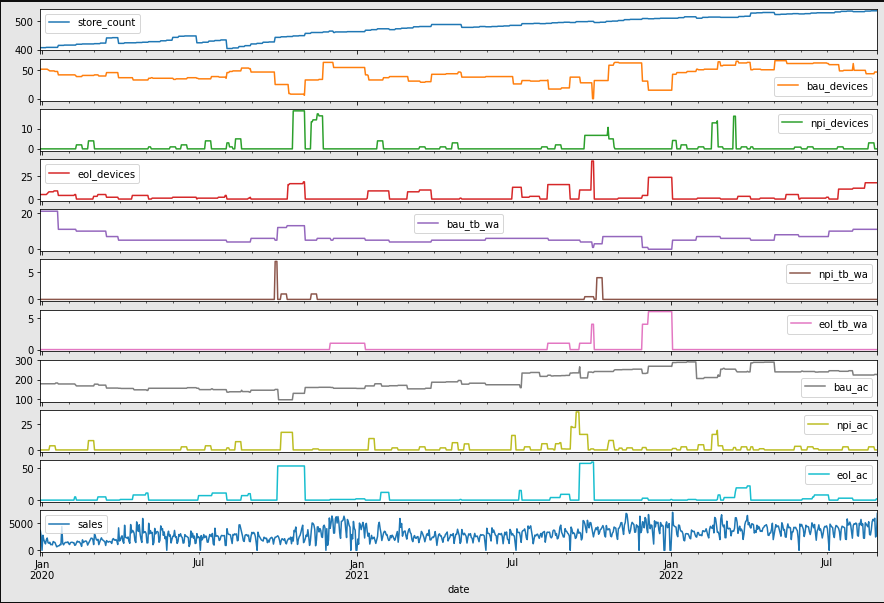
1. Attach a jupyter notebook that will allow us to view your modeling process and replicate your analysis, if needed.
2. Answer the questions in the next page either directly in this word document or in the jupyter notebook itself. You may also attach a powerpoint, if you want.
3. Please submit these deliverables within 5 business days after receiving this email.

*Questions? Email* [*albi\_dhimitri@comcast.com*](mailto:albi_dhimitri@comcast.com)*, if you are unclear about these instructions.*

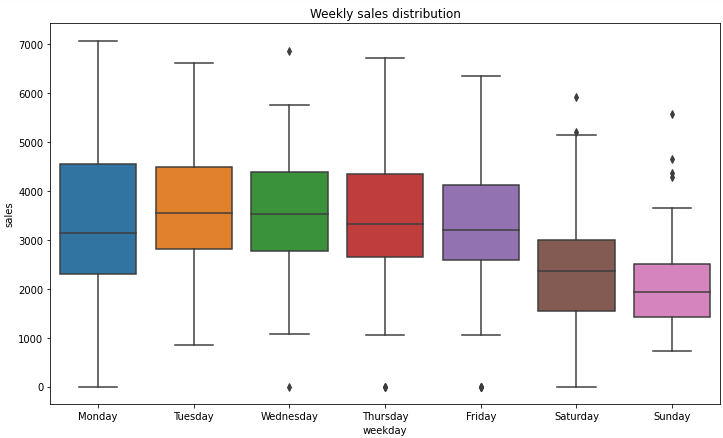
**QUESTION 1:** Explain any findings you made during the data exploration phase that you think are interesting or worth noting. Provide any charts and graphs needed to illustrate your findings.

**Distribution of data:**

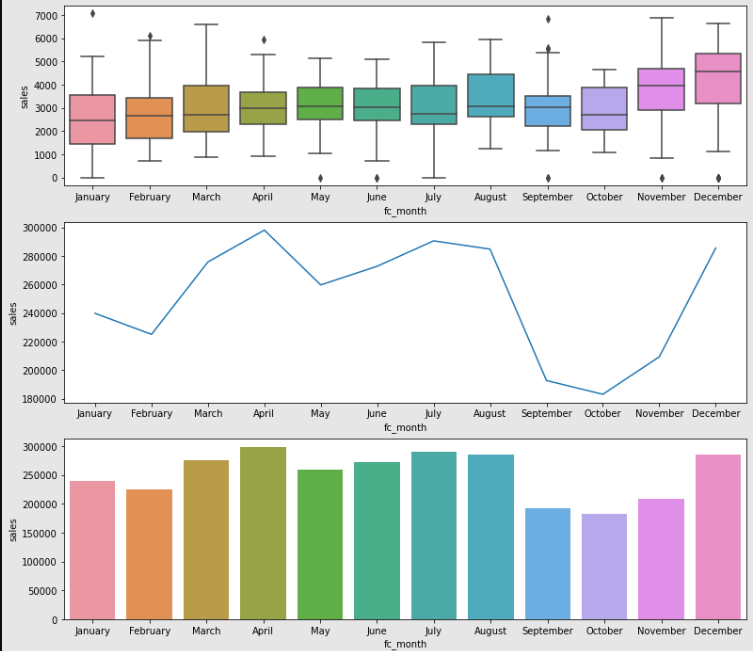
* store\_count increases YOY but sales does not increase at the same rate as store\_count.
* Business As Usual (BAU) is high across the devices, tablet & watches and accessories, means the Business is going usual.
* New Part Introduction (NPI) is comparatively high for devices and accessories, comparatively to tablet & watches, as sales of newly introduced product might not be regular.
* End Of Life (EOL) is comparatively high for devices and accessories, comparatively to tablet & watches, as sale of end of life product might occur once in a while.



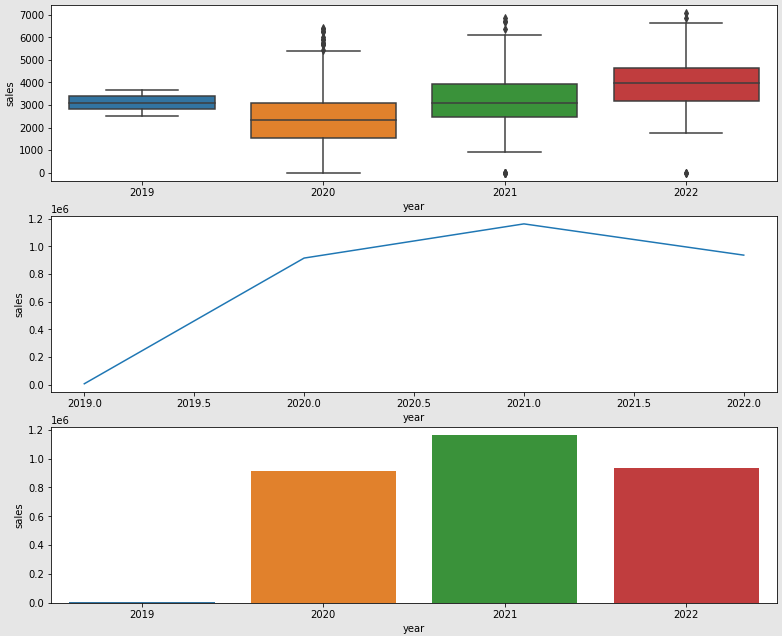
* The average number of sales increases over the week, and takes a sharp fall on Saturday.



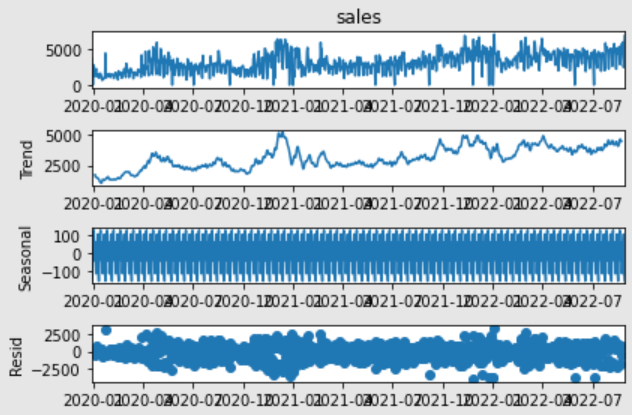
* The average number of sale gradually increases over the months, gets dropped in July and October. The sale is high in December.



* The average number of sales increases over the years

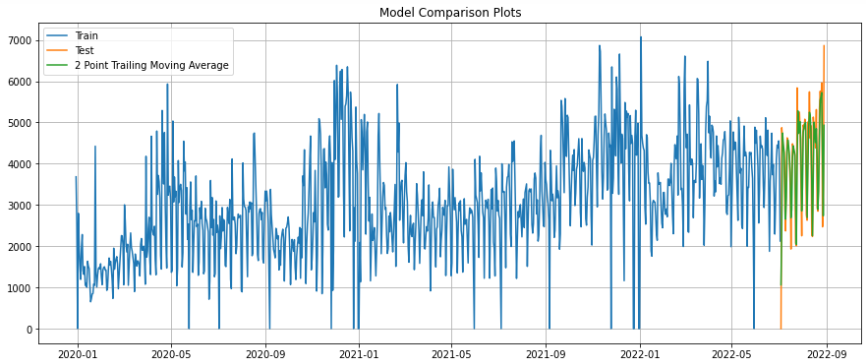


* Decomposition of the Time Series data in Additive Model.
* Trend is visible, Seasonality is not clear. Moreover deseasonal plot appears similar to the original plot.

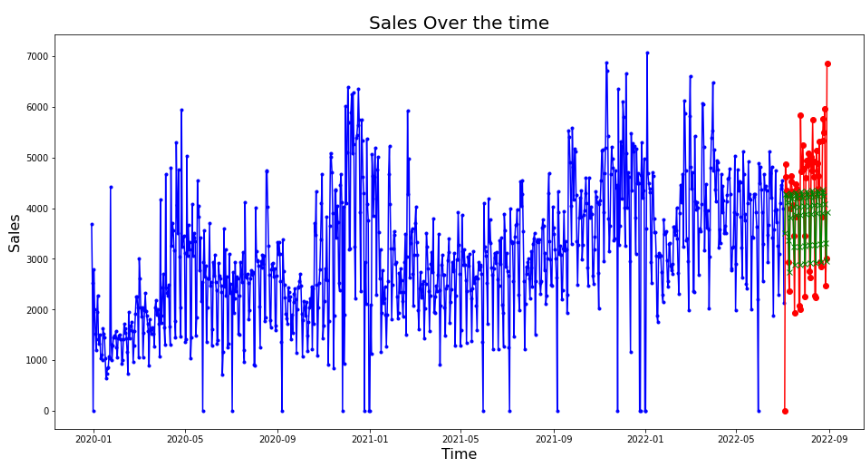


**QUESTION 2:** As your first model, which univariate timeseries model did you choose to forecast sales in 2022? Why did you choose that particular algorithm? How did your predicted sales compare against the actual sales provided in the TEST set? Did you build any additional univariate timeseries models?

* The **Moving Average (MA)** model is taken as a univariate timeseries model to forecast sales in 2022.
* It does not use the past forecasts to predict the future values whereas it uses the errors from the past forecasts.
* MA has the property to reduce the amount of variation present in the data, i.e., to eliminate fluctuations. The process is called **smoothing** of time series, which help to see patterns, trends.
* We can use ACF plot to figure out the order of MA model or try out manually with different Trailing MA points.
* I have tried with different trailing point MA - 2 point trailing MA, 4 point trailing MA, 6 point trailing MA, 9 point trailing MA.
* Here, we get 2 point trailing MA as the best model, which has ***least RMSE and MAE***.



* I have tried with different **Exponential Smoothing** models – Simple Exponential Smoothing, Double Exponential Smoothing/Holt model, Triple Exponential Smoothing/ Holt Winter’s model.
* I have also tried the **ARIMA and SARIMA (Seasonal ARIMA)** models is the used for univariate time series model to forecast sales in 2022.
* Checking the stationarity of the time series by ***Augmented Dicky Fuller Test*** and taking the first difference (d=1) for this data in order to be stationary***.***
* Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)
* The selected SARIMA (p,d,q) (P,D,Q,m) is as below. Setting the seasonality as 7 of the auto SARIMA model.
* We can also build the manual ARIMA and SARIMA model based on the **ACF and PACF plots**.

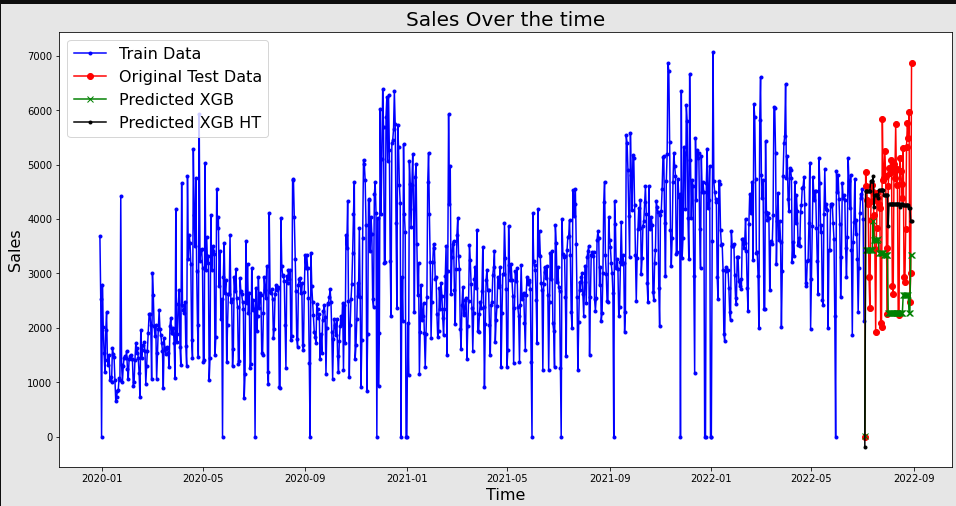


**QUESTION 3:** Which multivariate model did you choose to forecast sales in 2022? Why did you choose that particular algorithm? Did you try any other models or algorithms? How did your predicted sales for 2022 compare against the actual sales provided in the TEST set?

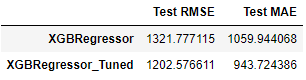
* The **gradient boosting models**, which creates a prediction model as an ensemble of other, weak prediction models.
* The gradient boosting models can be used for the time series forecasting, but it requires that the time series data to be transformed into a supervised learning problem.

**XGBRegressor:**

* The **Extreme Gradient Boosting Regressor (XGBRegressor)** model is taken for the multivariate timeseries model to forecast sales in 2022.
* XGBoost model tree grows depth-wise and can speed up the modelling procedure as well as we can get better results.
* Build the basic XGB model and further tune the model with hyperparameters to improve the model performance.

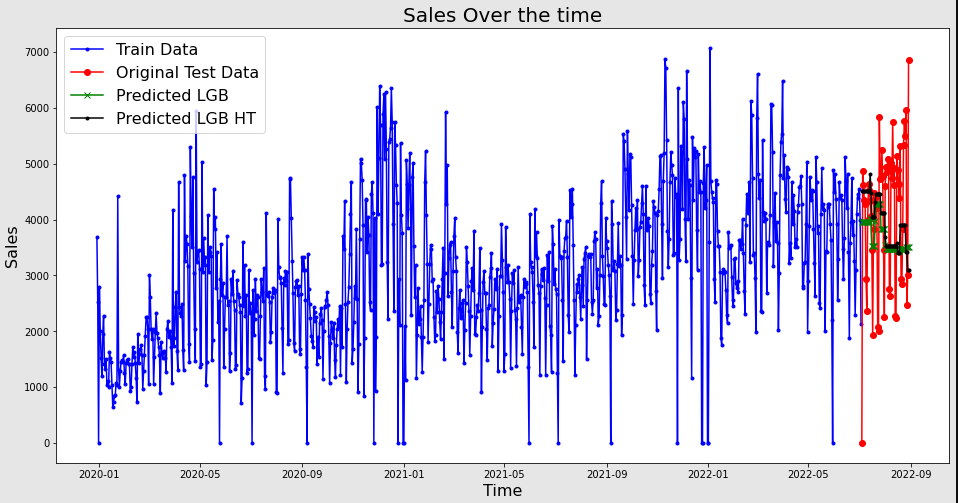


* Comparing the prediction results with least RMSE and MAE value.

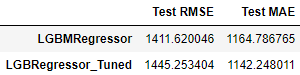


**LGBRegressor:**

* The **Light Gradient Boosting Regressor (LGBRegressor)** model is also used to forecast sales in 2022.
* LGBM tree grows leaf-wise and is light weighted.
* It requires fewer resources than other gradient booster models, thus making it slightly faster and more efficient.
* Build the basic LGB model and further tune the model with hyperparameters to improve the model performance.

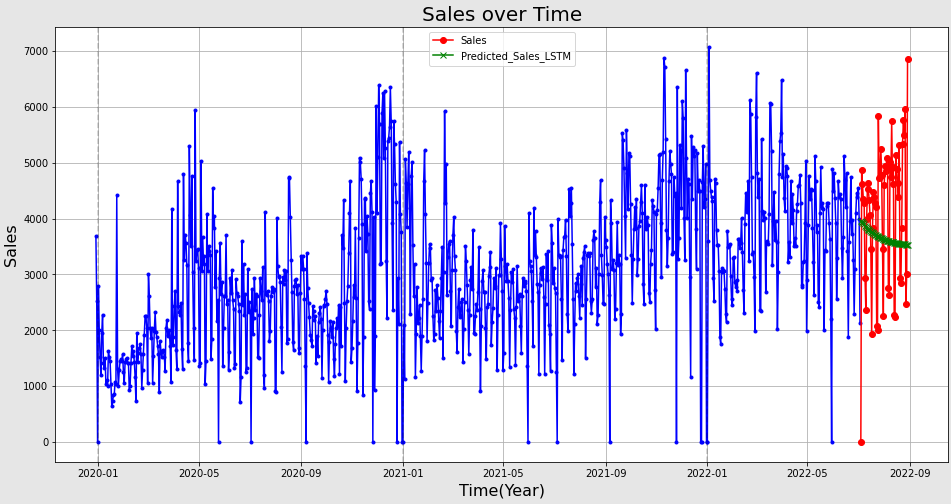


* Comparing the prediction results with least RMSE and MAE value.



**LSTM:**

* The **Long Short Term Memory (LSTM)** model is also chosen for the multivariate time series model to forecast sales in 2022.
* They are good at finding relationships between continuous data points, often over varying lengths of timeframes and can predict future values based on previous sequential data.
* They have an advantage over other regression models as they always look at recent past data. They are able to identify patterns resulting from seasonality as well.
* Scale the data with Min-Max scaler.
* Create a Time Series Generator object to generate batches of temporal data.
* Build a Sequential model and add the necessary layers with required parameters – LSTM layer, Dropout layer is added to avoid over-fitting, Dense layer is added to make the model more robust
* Fit the data to the model with EarlyStopping, to stop training when a monitored metric has stopped improving.
* Predict and evaluate on the test data.

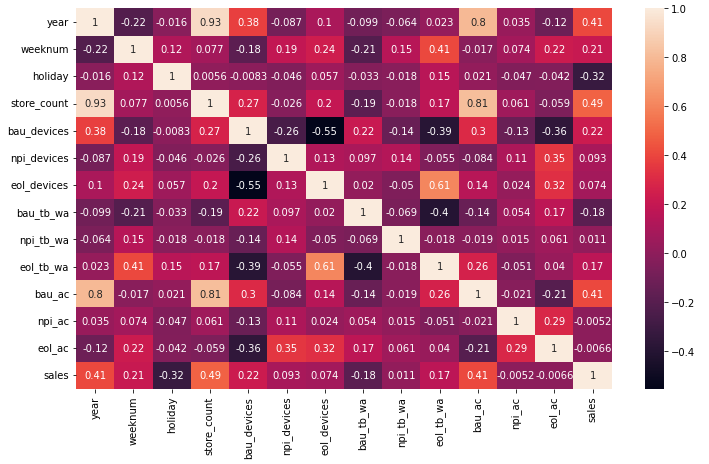


* Comparing the prediction results with RMSE and MAE value.



**QUESTION 4:** For the multivariate model(s), which group of features did you use to predict SALES? Did you use all features provided or did you select only a subset of them? Why or why not?

* For the multivariate model, only a subset of features has been used.
* The set of features considered in the multivariate models are ‘holiday’,’store\_count’,’bau\_devices’,’npi\_devices’,’eol\_devices’,’bau\_tb\_wa’,’npi\_tb\_wa’,’eol\_tb\_wa’,’bau\_ac’,’npi\_ac’,’eol\_ac’.
* The ‘holiday’ feature is not considered by the model, as it gives some sales value even when holiday flag is 1. But we can see the negative correlation between ’holiday’ and ‘sales’ in HeatMap.
* Other features in the data – ‘year’, ‘weeknum’,’fc\_month’,’weekday’ are considered only for the purpose of data exploration, but excluded in the model building as these features are derived from the ‘date’ feature.

****

**QUESTION 5:** Did you try to fine-tune the hyperparameters of any of your models? If not, how did you come up with the set of parameters you ultimately used? Or did you just go with the default values?

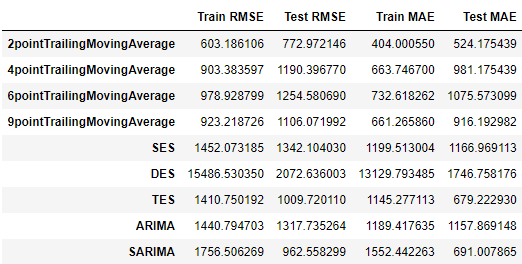
* In the univariate time series model, tried for different approaches for each model.
* In **Moving Average** (MA) model, tried with **various sliding windows** – 2 point Trailing window, 4 point Trailing window, 6 point Trailing window, 9 point Trailing window.
* In **ARIMA and SARIMA,** build an **automated version** and selecting the best parameters (p,d,q,P,D,Q) having **lowest Akaike Information Criteria (AIC)** value.
* In the multivariate time series models with boosting ML models – **XGBRegressor and** **LGBMRegressor** models, tuned with the hyperparameters.

**QUESTION 6:** Are your fitted models predicting non-zero values in cases where “holiday\_flag”==1 in the TRAIN data? How should we deal with this in the future when we are trying to predict sales for Holidays?

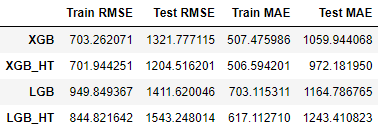
* Yes, the fitted models predict non-zero values, where “holiday\_flag”==1 in the train data.
* We can see in Heatmap, there is negative correlation (-0.32) between the holiday\_flag and sales column. Make our model understand this and train with Facebook Prophet model, which considers the holidays or recurring events.
* Include each occurrence of the holiday, both in the past and in the future. The model can identify if the holiday won’t repeat and not include the in future.

**QUESTION 7:** Are your models overfitting or underfitting? Or are they generalizing well enough? How can you tell?

* From the below figure, we can see different univariate models used for forecasting the test data.
* The 2 point Trailing MA model seems to have good result (lower RMSE & MAE), compared to other models.
* Next to that, the 4 point Trailing MA and the 9 point Trailing MA models seem to have better prediction.
* The 6 point Trailing MA model is slightly overfitting.
* The Simple Exponential model (SES), Triple Exponential model (TES), ARIMA models seems to be underfitting, as they have high error in both train and test data.
* In Double Exponential model and SARIMA model, the test data is having way less loss/error than the train data. This might be because the model might require more data with seasonality.

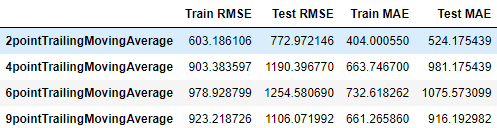


* From the below figure, we can see different multivariate models used for forecasting the test data.
* The XGB Regressor model with hyperparameter tuning is giving good result, with lower RMSE and MAE. The XGB model before tuning is overfitting.
* The LGB Regressor model with the hyperparameter tuning is overfitting. The LGB Regressor model before (with base parameters) seems not to overfit/underfit.



**QUESTION 8:** Which of your models performed the best? What metric did you use to measure the performance? What could be some ways in which we can increase the predictive accuracy of the models?

* For the evaluation of the time series model forecast, I have considered the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics.
* The Moving Average models performed the best overall with lower RMSE and MAE.



* This data seems to have trend but no seasonality, based on decomposition of data.
* We can use **rolling k-fold cross validation** on the data.

**QUESTION 9:** What difficulties did you encounter during this process?

* Understanding the domain (Supply Chain) and the features given in this dataset.
* Exploring how each feature affects the sales values and trying the check the same in ML model prediction (feature importance).
* Tried to recall the concepts used in the Univariate Time series – checking the presence of trend & seasonality, checking if the time series is stationarity or not, ADF Test, creating & understanding the ACF & PACF plots, building ARIMA and SARIMA models.
* Understanding the LSTM model functioning and building on the time series data as multivariate model.
* Trying to incorporate the “holiday” feature to the sales value, still not implemented for that case.