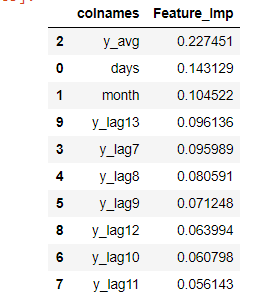
# Q3 :- Model Comparison and Prediction

## Random Forest Model

**Feature Importance**



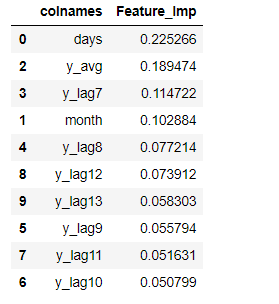
y\_avg – Monthly average y. This is computed on a month basis (1,12) disregarding year, to take seasonality into account.

The Cross Validated Mean Squared Error for Random Forest Model was 5941539 (RMSE = 2437.527)

The features y\_lag 7 is a lag feature that gives the value of y (volume) 7 weeks ago. Lag features are created from 7 to 14 days ago (y\_lag7 to y\_lag 14)

## Gradient Boosting Model

**Feature Importance**



The Cross Validated Mean Squared Error for Gradient Boosting Model was 6354150 (RMSE = 2520.74).

## Conclusion

The Random Forest Model performs marginally better under Cross Validation. The Cross validated error is computed on the validation data set. A methodology of 3-fold cross validation is followed in the project for both Random Forest and Gradient Boosting. The model is trained on 2 folds and evaluated on the third fold. The Mean Squared error given above is calculated on the third (validation) fold.

The Feature importances under both models is different. In the Random Forest model the Monthly Average volume – y\_avg is the most important feature while in the Gradient Boosting Model, number of days (days) is the most important feature. The Feature importances of lag features decreases as the lag increases (ex.- lag 7 is more important than lag 11) in both models.

**Performance on Training Data**

