CS584
Final Project

Short-Term Energy Consumption Forecasting using Machine Learning

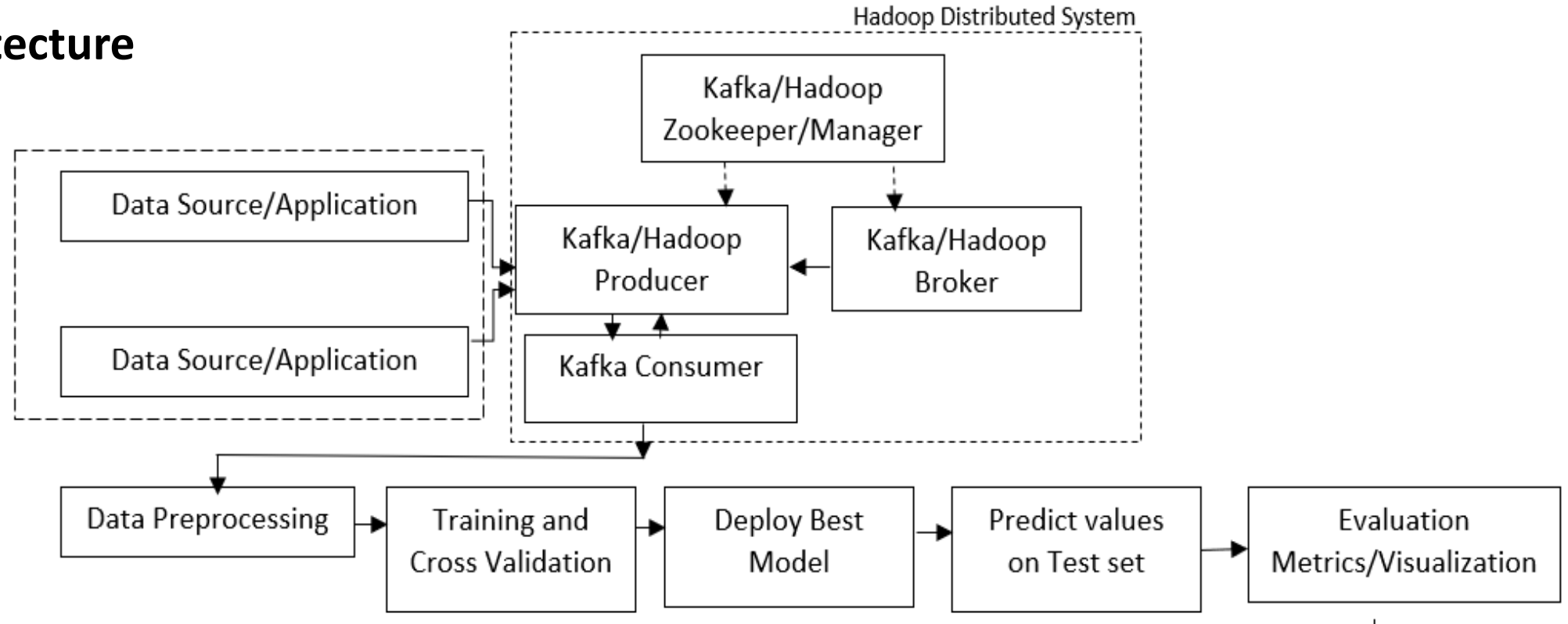
(Chaithanya Pramodh Kasula and Aishwarya Varala)

(Team-11)

Dataset Source and Description:

- https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption
- The current data set titled, 'Individual household electric power consumption Data Set' consists of 2,075,259 rows and 8 columns.
- Each row represents the active energy consumed every minute (in watt) between December 2006 and November 2010 of a single household.
- Features: Date, Time, global_reactive_power, Voltage, global_intensity , sub_metering_1, sub_metering_2, sub_metering_3.

Architecture



Data Preprocessing

- Initially 1.2% of the rows in the data had missing values. These null values are replaced by the previous rows' values.
- The data consists of readings collected for every minute. The values of every minute have been aggregated to a single day. This is performed by adding the values of rows for each column belonging to a single day.
- The size of the data set by date is 1443.

Data Normalization

- Normalization is performed to accommodate all the algorithms being used.
- It has been performed for each column in the data set.
- Algorithms such as SVM perform better when normalized.
- Additionally, the RMSE values are better understood when all the features have common results in the range of [0,1].

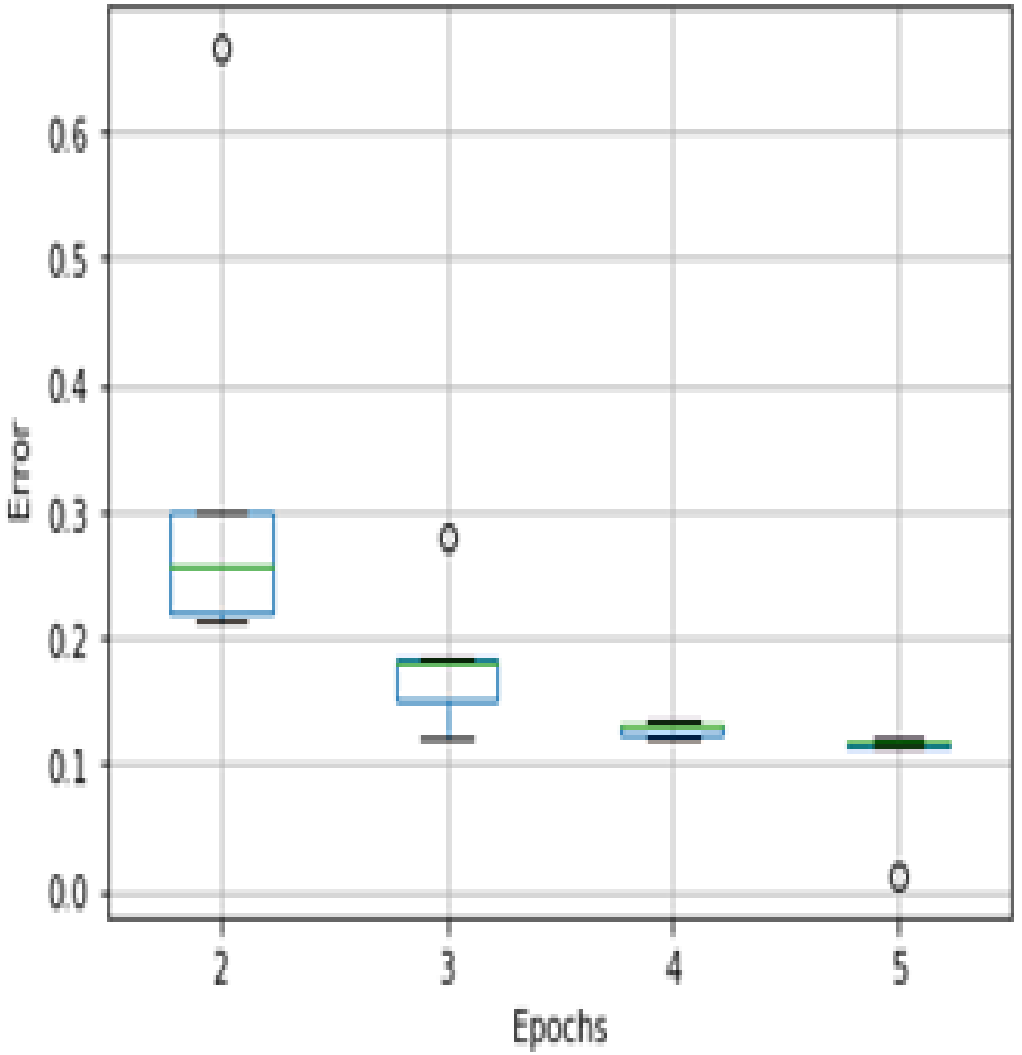
Models Used

- Random Forest Regression and 0.632 Bootstrap
- Support Vector Regression
- Deep Learning Framework – LSTM

Preparing the data for Training

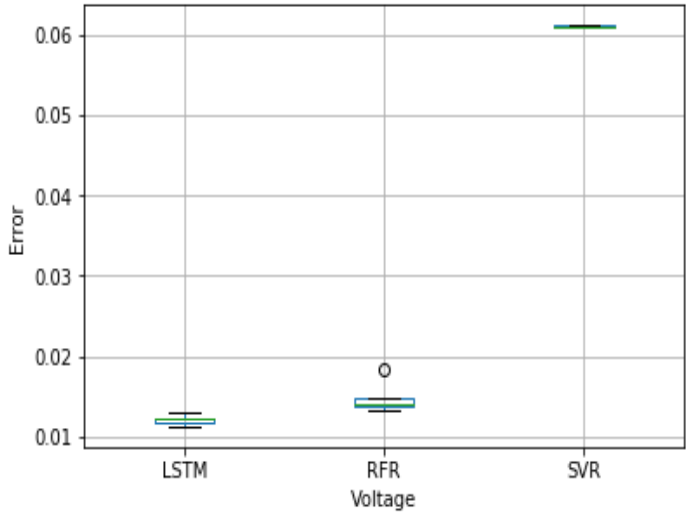
- Walk forward Validation
- Recursive Multi-Step Forecast
- **Hyper-parameter Tuning**
- **Performance Metric:**

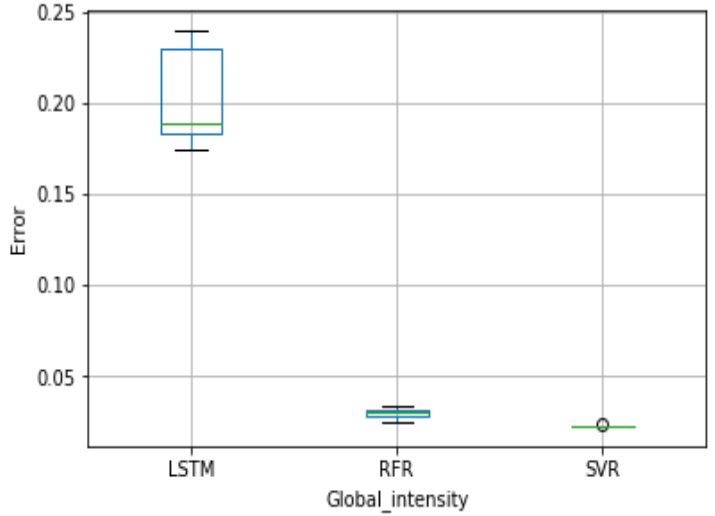
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

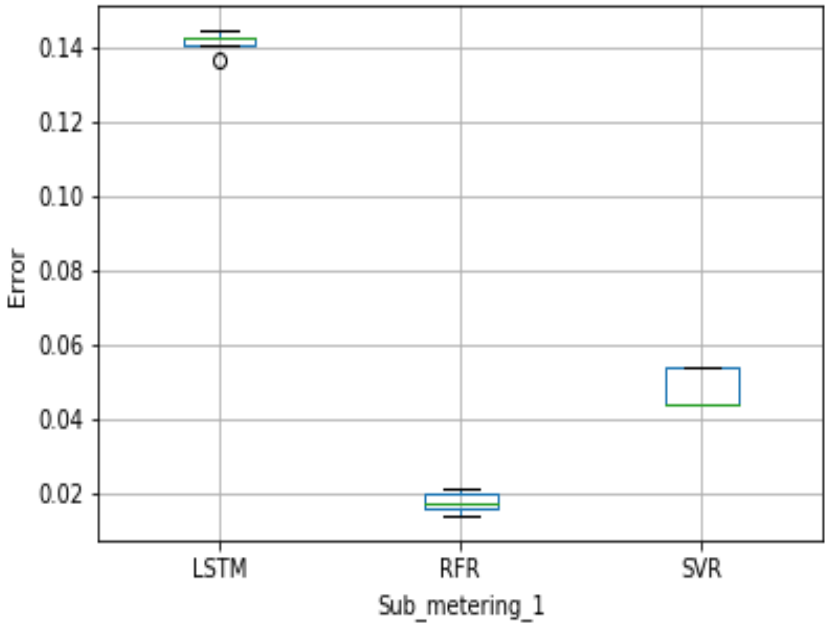


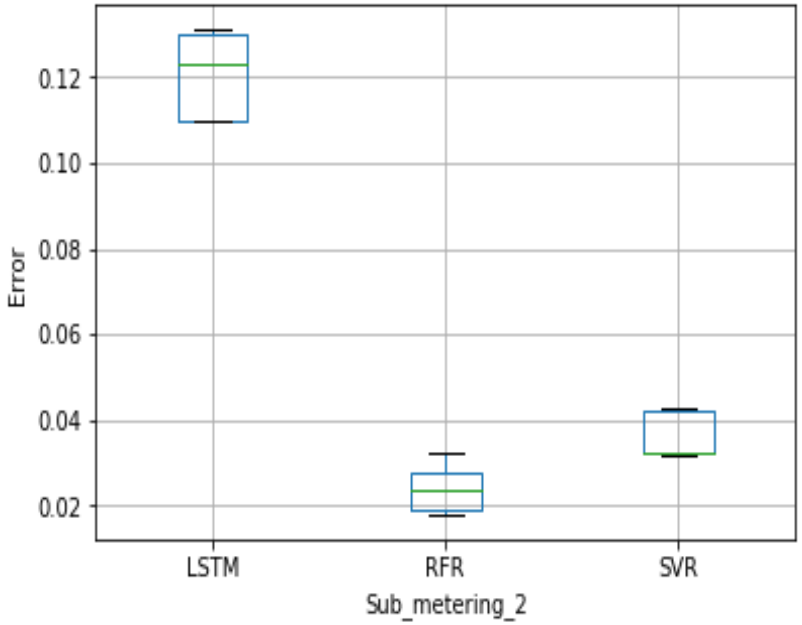
5 – fold Cross-Validation:

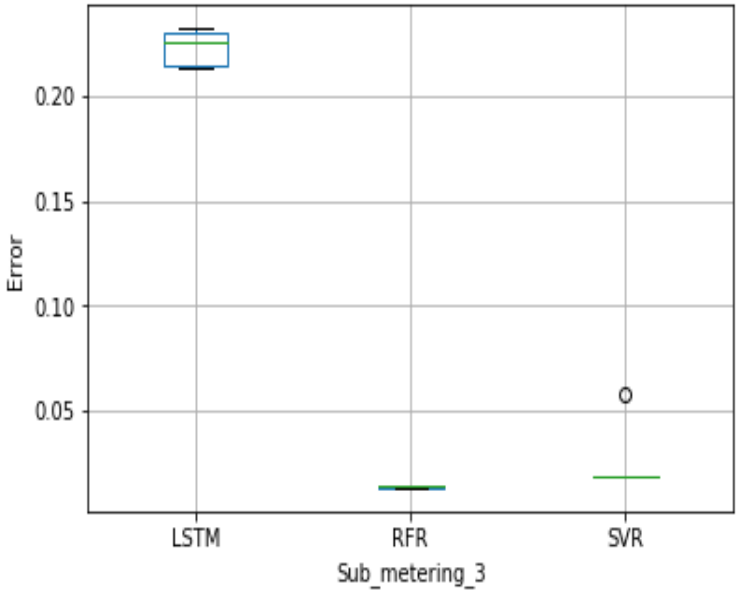
- As in classification, generic procedure of splitting/shuffling data for cross-validation is not possible with time-series data.
- This is because shuffling rows disturbs the sequence or the temporal order and results in data leakage.
- Instead the data is divided into parts without disturbing the sequence.











Model Selection

- Voltage – LSTM
- Global_intensity – SVR
- Sub_metering_1 – RFR
- Sub_metering_2 – RFR
- Sub_metering_3 – RFR
- In time series data, when there is a **concept drift**, an ensemble of all models can be used.

Training

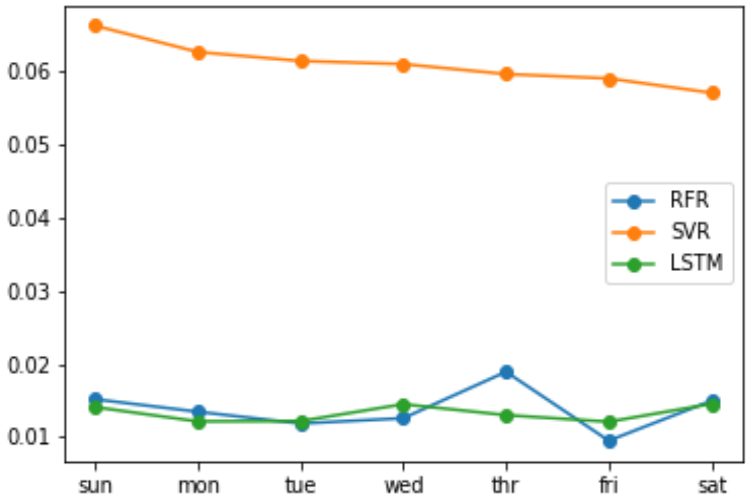
- After cross validation and model selection, the train and the test set are divided in the ratio of 80:20.
- The train set consists of 1142 records whereas the test set consists of 301 records.

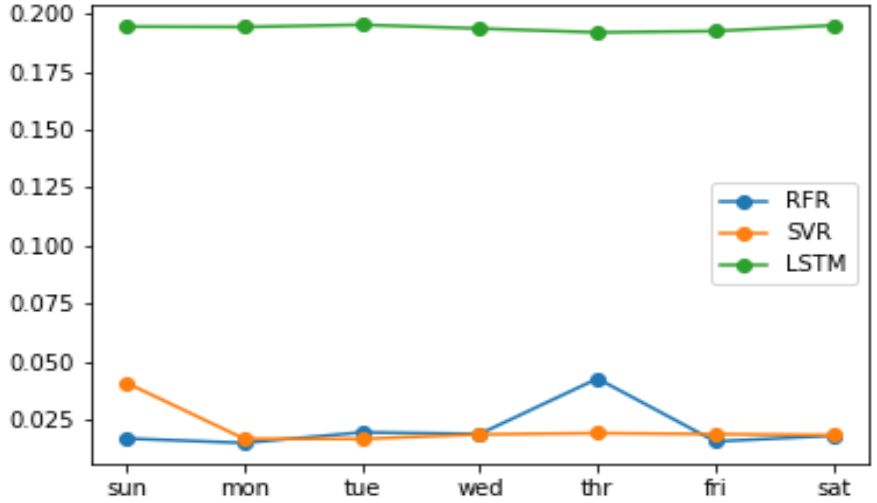
Testing:

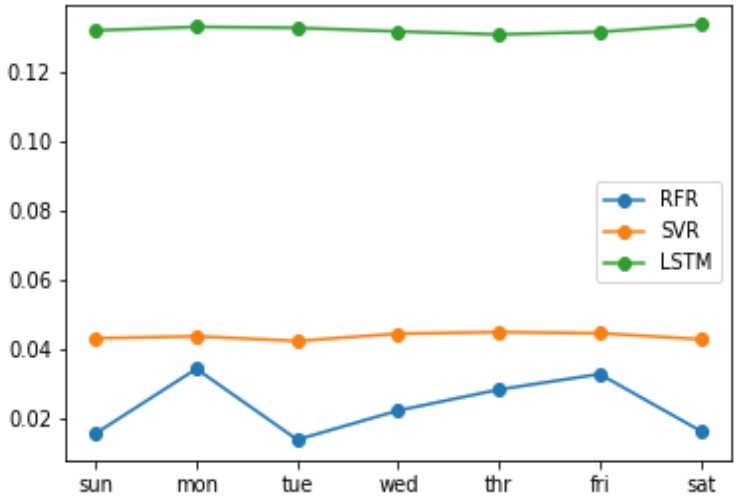
- The trained model is used to perform testing through walk forward validation process.
- A multi-step recursive forecast is selected to predict the test instances as stated above.

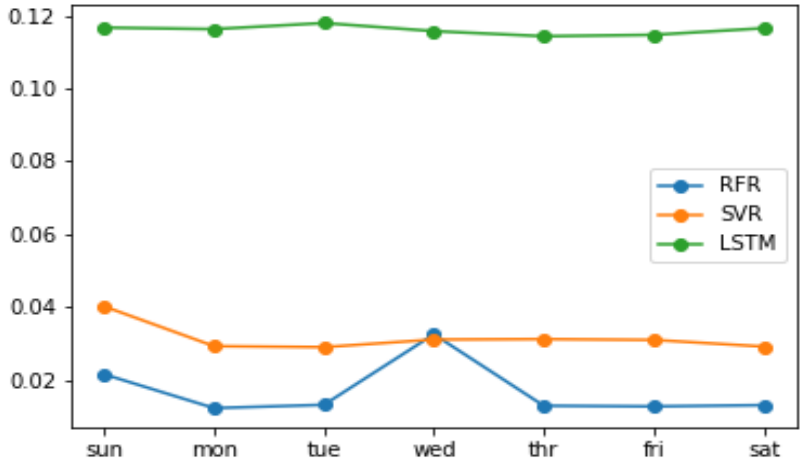
Performance Evaluation

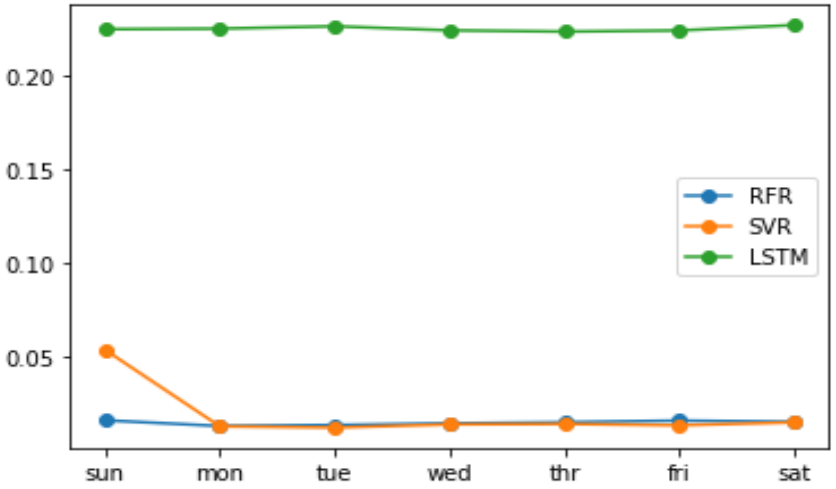
- In relation to the ensemble concept, the plots below include the performance of all the models over all the features.











Average RMSE Values by feature			
	RFR	**SVR**	**LSTM**
Voltage	0.013707694451809646	0.01690740582729737 3	0.01113446515422513
Global Intensity	0.0226040467716400 1	0.02256988645323157 8	0.20337612894998616
Sub_metering_1	0.0246102436930369 3	0.0436913521209787 9	0.137077533134143
Sub_metering_2	0.0183047852702643 8	0.0317578775031278 5	0.1313617268273868 4
Sub_metering_3	0.0144377907334941 06	0.0236186128338155 3	0.2248516081664869 2

Conclusion:

- With the availability of more training data, LSTM could perform better. However, one of the downsides of LSTM is that it takes a lot of time for training.
- Random Forest Regression is quick and offers performance very close to that of LSTM. From the cost and time perspective, in the given conditions, we can use Random Forest Regression instead of LSTM.
- The models are also not overfitted as the train and the test errors for all the models are minimum.