Homework 6 Report

Jae Young Kim

Applied Data Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: April 29, 2021)

Abstract

Through this report, ps6_trainvalid.csv and ps6_test.csv data set were given and used to predict the future temperature based on the previous temperature and other features such as humidity and wind speed. Linear regression model was used as base line models and RNN model was used to predict multiple future temperatures. As a result, linear model showed the best MAE score 0.348 in the test data. After training models with the features, some feature were selected based on the feature importance. 5 features were selected and trained again with the same models. Finally, there was significant performance increase and LSTM model scored 0.080 MAE score.

I. INTRODUCTION

To predict future temperature, various features as temperature, wind speed, direction, humidity were used. To get the best model, I tried linear regression baseline model, baseline repeat last model, baseline repeat 24 hours model, linear model, dense model, conv1d model, LSTM model and Feedback model. Then based on the feature importance, these models were trained one more time with the selected features. As feedback model seems to be the most reliable model, it was expected to score the highest score. However, the result showed different scores.

II. DATA EXPLORATION AND PREPROCESSING

To preprocess data, missing data was collected. 152 humidity, 252 pressure and 3 temperature were found as missing data. Outliers were not found. There were 45013 rows in train and validation set. Since the proportion of missing data was small, I filled the missing values by using linear interpolation.

To use the seasonality, datetime was changed into Day sin, Day cos, Year sin and Year cos. The result is shown in the Figure 1.

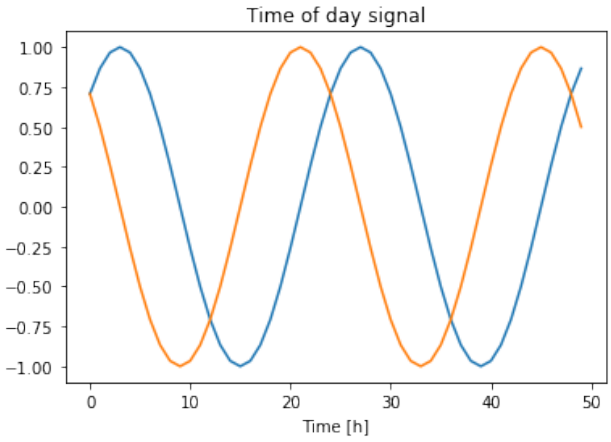

FIG. 1. Expressed Seasonality of Day sin and Day cos

To preprocess feature "weather" which is expressed in words, 6 features were created. Clear, haze, cloud, rain, dust and thunderstorm are those features. Each of them expresses in number based on the weather expression.

To check the seasonality, each feature was visualized in Figure 2. It is possible to check strong seasonality in temperature, year and day.

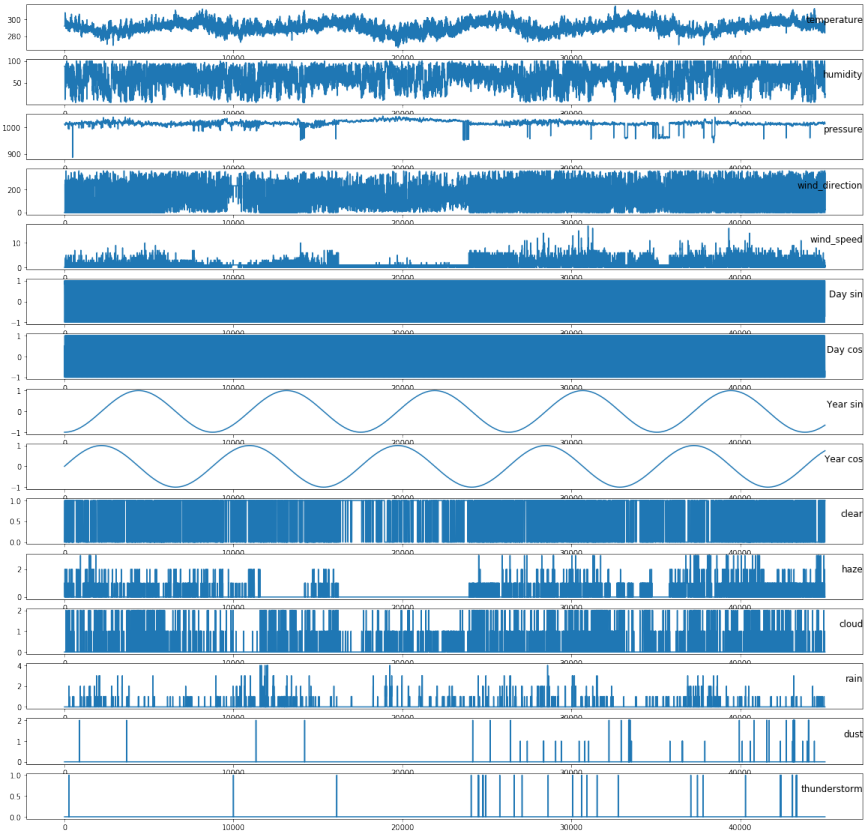

FIG. 2. Vizualization of features

Finally, preprocessed train data, validation data and test data were formed 36009 rows * 15 columns, 9003 rows * 15 columns and 240 rows * 15 columns. The split data set are normalized to use it as an input of deep learning models.

III. MODEL SELECTION AND EVALUATION

As a baseline model, linear regression model is chosen. It is because linear regression is the simplest model that can be tried. To predict the temperature after 24 hours, temperature, humidity and pressure of previous 24 hours were used. As a result, 72 columns are used as train data to predict the temperature after 24 hours. Finally, calculated MAE on test data was 0.996.

Then feature importance in linear regression model is plotted in Figure 3. Figure 4 shows top important features and it show that previous temperatures have largest impact on the

temperature prediction. Especially, the temperature of 23 hours ago showed the largest feature importance.

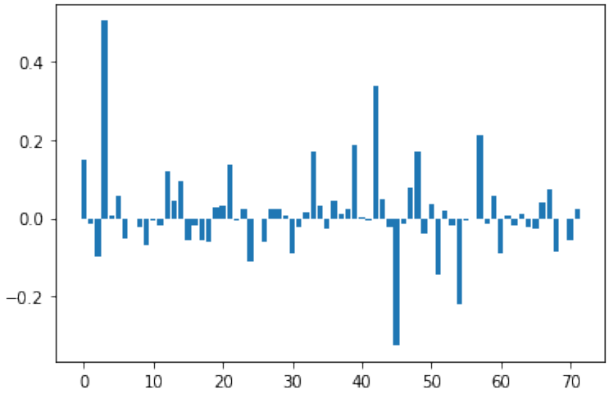

FIG. 3. Feature importance of baseline model(Order : temperature 0, humidity 0, pressure 0, temperature 1, ...)

```
('temperature 1', 0.5033259893548194),
('temperature 14', 0.3388233932333411),
('temperature 15', -0.3238918190877497),
('temperature 18', -0.22032815940242423),
('temperature 19', 0.21319524016601907),
('temperature 13', 0.18805639851213013),
('temperature 16', 0.17096951101751548),
('temperature 11', 0.16878792076376295),
('temperature 0', 0.1477025092227703),
('temperature 17', -0.1458521259411389),
('temperature 7', 0.13732843860144212),
('temperature 4', 0.11988916923234702),
('temperature 8', -0.1098206374506196),
('pressure 0', -0.09820597356804675),
```

FIG. 4. Sorted Feature importance of baseline model

To predict multiple timestamps with previous 24 hours data, at first I made another two baseline models. First one is baseline model which repeats the last temperature for the output. The other one is baseline model which repeats the past 24 hours temperature. Baseline model repeating 24 hours data was expected to have better result since this model cares about the time. The sample result is plotted in Figure 5 and 6. As the trend of the temperature is similar from the last day, the second baseline model showed better result. The two models showed MAE 0.605 and 0.386 for the validation data and 0.605, 0.366 for the test data.

The next model tried is linear model. This model only uses the last time step. It applies linear function to the last time step. (Similar to linear regression model). The sample result

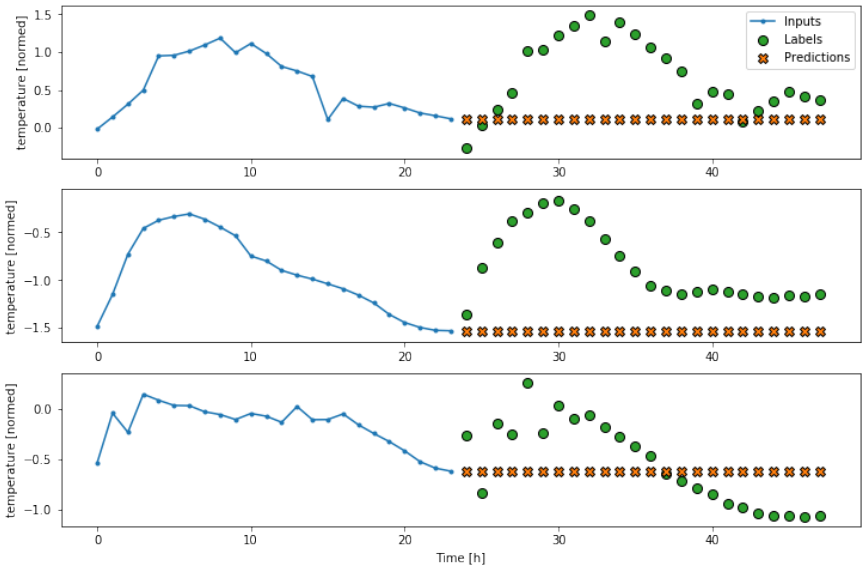

FIG. 5. Baseline model repeating last

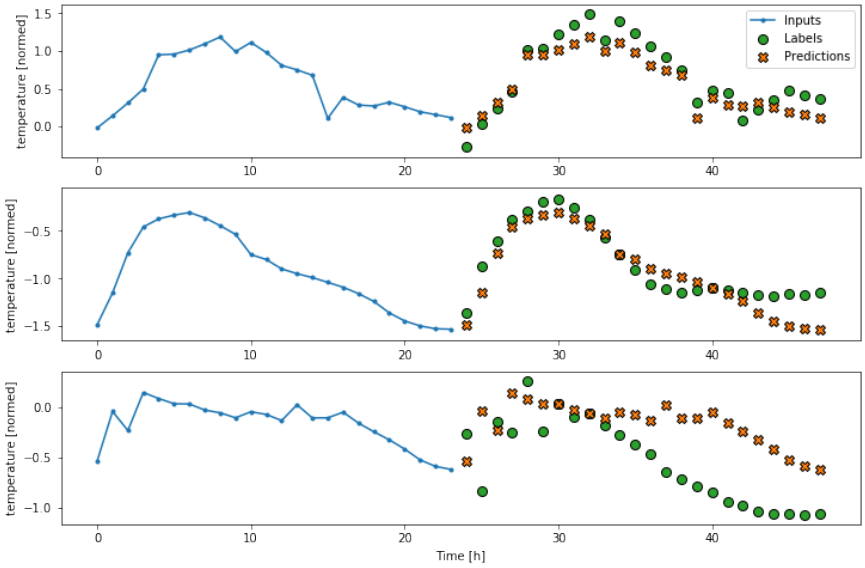

FIG. 6. Baseline model repeating 24 hours

of this model is shown in Figure 7. As it only cares the last time step, it has limitation. Linear model showed MAE 0.397 on validation set and 0.328 on test set. Even though it has limitation, it showed better result then two baseline models.

To use nonlinear function, in the next dense model, relu activation function is used. The sample result is shown in Figure 8. As dense layer with activation function can express nonlinear relationship, better performance was expected. However, this model scored 0.350 MAE in test data. This score is slightly lower than the score of linear model.

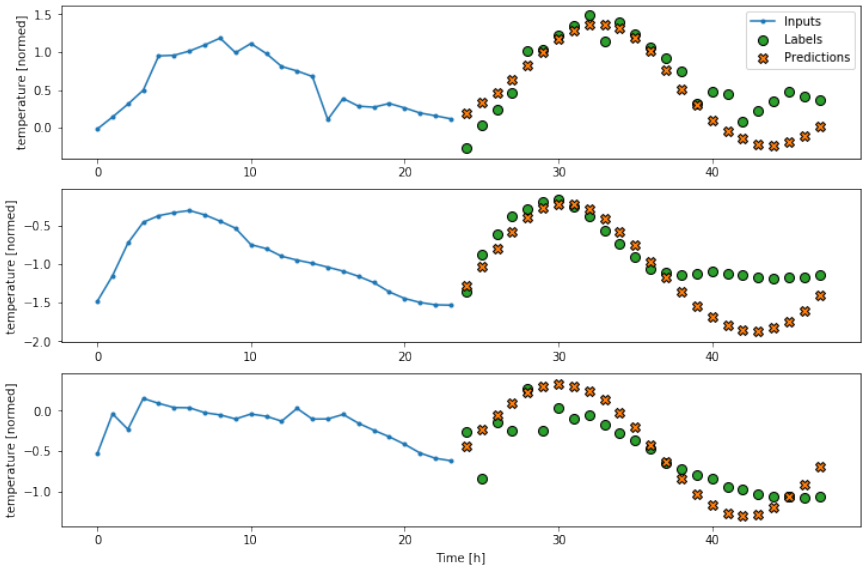

FIG. 7. Linear model

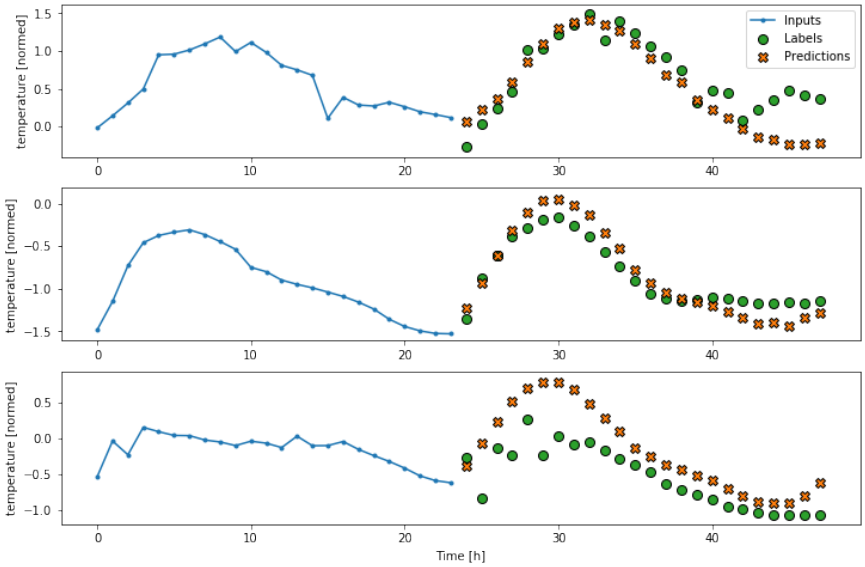

FIG. 8. Dense model with relu activation function

Previous models used only the data of last timestamp. To use the data of previous timestamps, convolutional layer and LSTM layers are used. In Conv1d model, 3 timestamp window was used. The sample result is shown in Figure 9 and this model showed similar MAE score to dense model. It showed 0.396 MAE for validation set and 0.353 MAE for test set. On the other hand, in case of LSTM model, it showed worse result. it showed 0.464 MAE in validation set and 0.396 MAE in test set. The sample result is shown in Figure 10.

Final model I used is LSTM self feedback model. In this model, with the data of previous

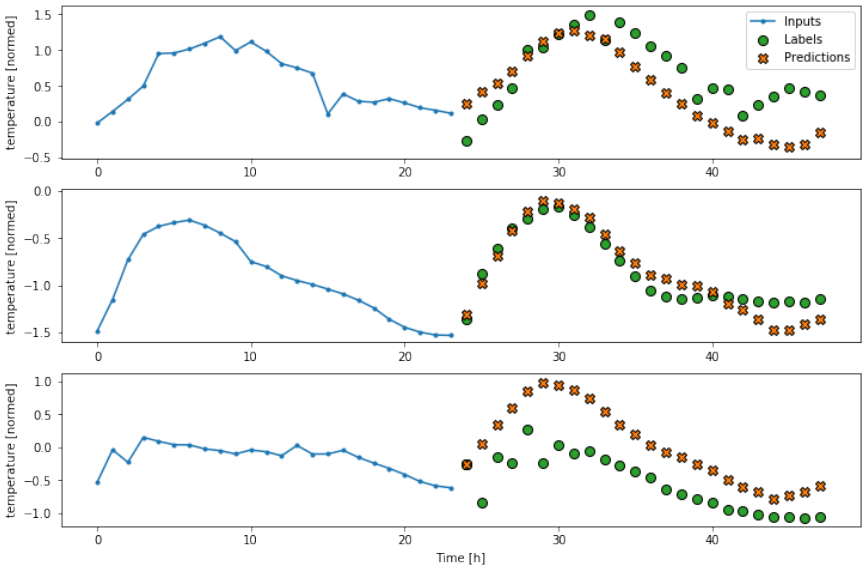

FIG. 9. Conv1d model

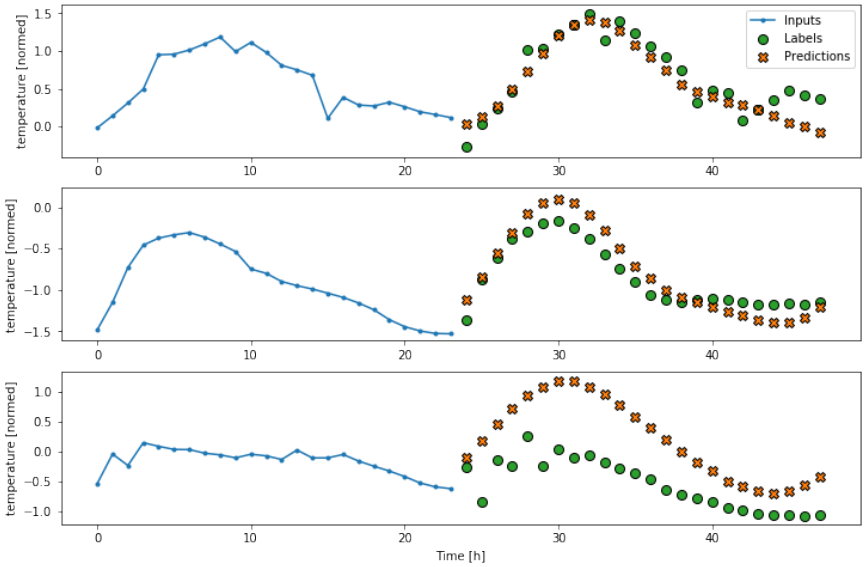

FIG. 10. LSTM model

timestamps, the right next timestamp temperature is predicted. Then the result is used as an input to predict the next timestamp temperature. As it uses this feedback approach, better result is expected. However, the result showed 0.424 and 0.396 MAE for validation and test set.

After training these models, the same models were trained with reduced number of features. Feature [temperature, Day sin, Day cos, Year sin, Year cos] were selected. These features were selected from the feature importance in Figure 12. When the models were

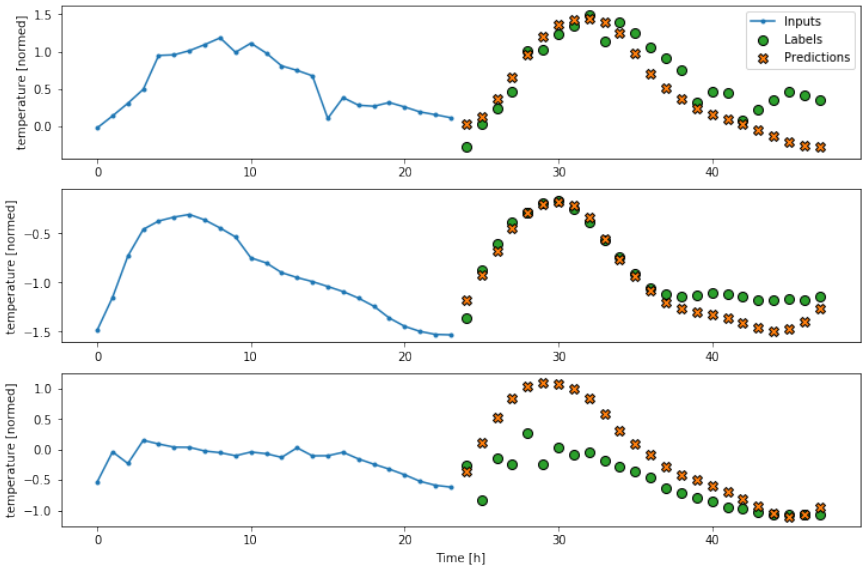

FIG. 11. LSTM feedback model

trained again, there was significant performance increase. It shows that the other features were not informative in predicting the future temperature. LSTM model showed the best MAE score 0.080.

The result is summarized in Figure 12, Figure 13 and the tables below.

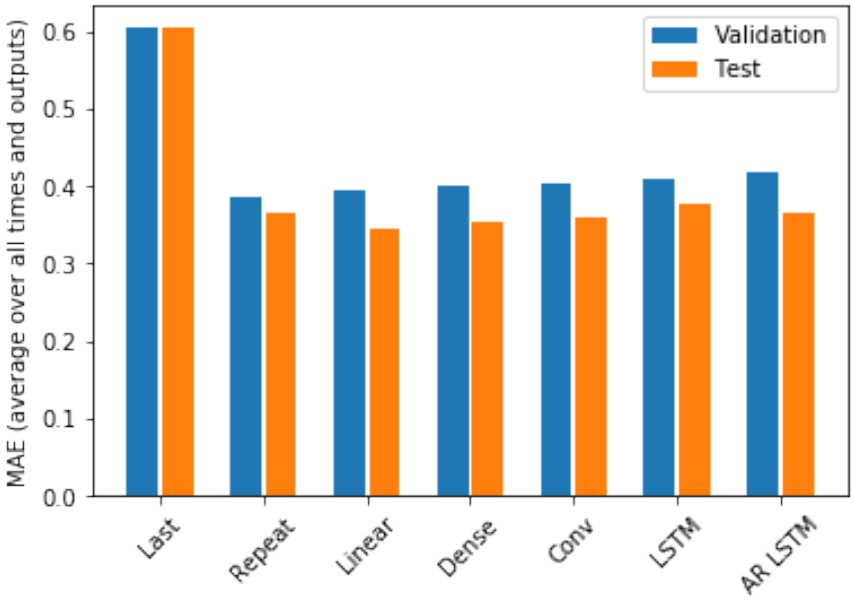

FIG. 12. Summary of MAE score of models with full features

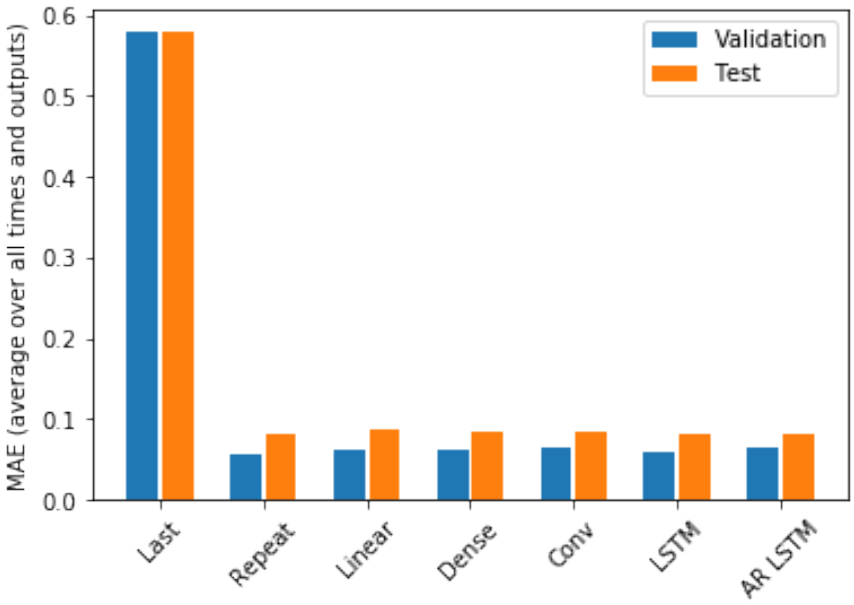

FIG. 13. Summary of MAE score of models with selected features

Model with full features	Validation MAE	Test MAE
Baseline linear regression	-	0.996
Baseline Repeat Last	0.605	0.605
Baseline Repeat 24hours	0.386	0.366
Linear Model	0.397	0.348
Dense Model	0.397	0.350
Conv1d Model	0.396	0.353
LSTM Model	0.464	0.396
Feedback Model	0.424	0.396

Model with selected features	Validation MAE	Test MAE
Baseline linear regression	-	0.722
Baseline Repeat Last	0.580	0.580
Baseline Repeat 24hours	0.055	0.082
Linear Model	0.062	0.086
Dense Model	0.062	0.085
Conv1d Model	0.066	0.084
LSTM Model	0.058	0.080
Feedback Model	0.064	0.082

IV. FEATURE IMPORTANCE

To check the feature importance from the 15 columns of preprocessed data, single shot linear model was used. As the coefficients of the model represents the feature importance of the linear model, the coefficients were plotted in Figure 12.

It showed temperature of the previous timestamp affects the most. Since Day sin and Day cos showed next largest feature importance, the seasonality affects a lot in predicting future temperature.

V. CONCLUSIONS

To sum up, to predict future temperature, 8 models were used. Baseline linear regression model predicted one timestamp which is 24 hours later from the input timestamp but the other models predicted multiple timestamps in 24 hours after input timestamp. As a result, Linear model showed the best MAE score, 0.348. Based on the feature importance, 5 features were selected and trained same model. Finally, LSTM model showed the best result, 0.080 MAE score on the test set.

DATA AVAILABILITY

Data is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps6-jeayoung114/data

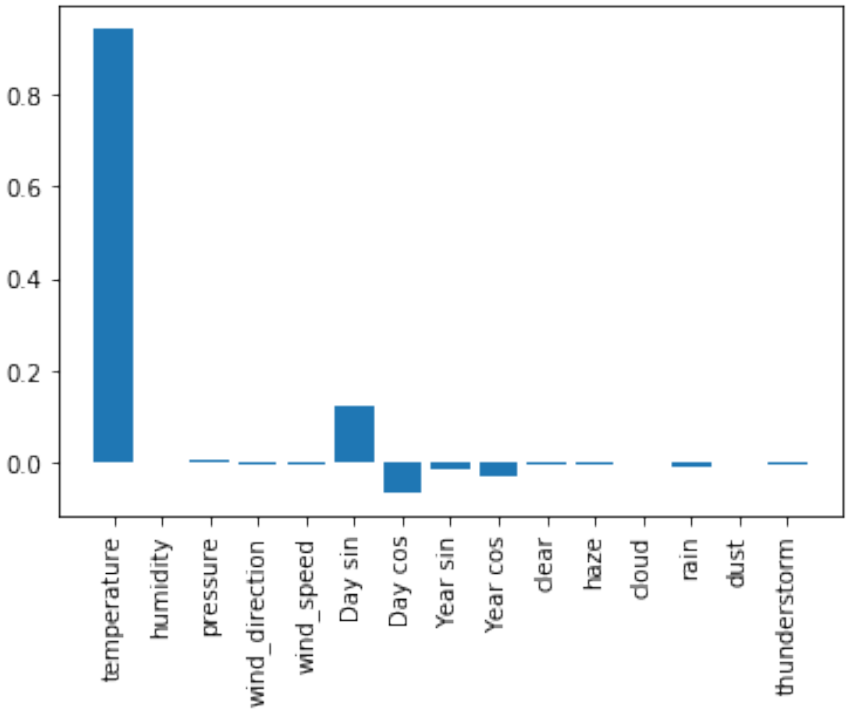

FIG. 14. Feature importance from single shot linear model

CODE AVAILABILITY

Code is available at

https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps5-jeayoung114/code