Springboard

WEATHER PREDICTION

By Taha Shahid

Abstract

Machine learning techniques, as well as deep learning models can be used as an aid to predict trends of the weather

Abstract

Machine Learning is an emerging area of research that aims at extracting meaningful patterns from available data. This paper highlights the significance of classification in predicting new trends from voluminous data. Performance analysis of various algorithms' like Linear Regression, Deep Neural Network Regressor, and Long Short-Term Memory in predicting the weather is discoursed in this project. Dataset from the Kaggle repository comprising of 12 attributes and 96543 instances have been employed to analyze the performance of algorithms and deep learning models. Moreover, the effect of feature selection has also been identified with respect to each algorithm. It has been concluded from the experimental results that include new found features from the data and the original data both yield the information useful for prediction of Deep Neural Network Regressor method is highest in predicting weather.

Table of contents

Contents

1. Introduction

The data set selected is called “Weather History in Szeged 2006-2016” and as the name suggest we predict which machine learning technique and neural network system will give the best accuracy for the prediction of temperature.

The dataset is provided by the Kaggle repository and will be acquired through downloading csv format of the dataset from the website.

To analyze this problem, different machine learning algorithms such as Logistic Regression and Neural Networks (LSTM, MLP) will be used to see which method has the best accuracy for the prediction of temperature. Also, PCA (Principal component analysis) will be used to see if dimensionality can be reduced. Any or all methods learned in the machine learning algorithms will also be applied.

For the deliverables of the project, an IPYTHON notebook code will be provided along with the paper summarizing the findings as well as a PowerPoint presentation.

2. Literature review

There is much research on weather prediction, it is a widely researched subject. Many prediction models have been applied to safety measure weather prediction, such as linear regression, deep neural networks, and recurrent neural networks etc. Advanced machine learning methods including deep neural network regressors and long short-term memory have been applied as well. A short introduction to these techniques is provided here.

Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Deep Neural Networks (Deep Learning)

Deep Neural Network is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised. They use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input. They also learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.

Long short-term memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections that make it a "general purpose computer" (that is, it can compute anything that a Turing machine can. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

3. Dataset Description (Data Wrangling)

The weather of Szeged 2006-2016 dataset was downloaded in excel(csv) format from the Kaggle repository. For cleaning the data, multiple steps had to be followed such as: fixing missing values, float to integer conversion, and string to integer conversion.

3.1. Original Features

The dataset contains 1 Datetime feature. As per the data description these features include:

- Formatted Date – Datetime column

The dataset also contains 3 categorical features which are:

- Hourly Summary – As string object
- Precipitation Type – As string object
- Daily Summary – As string object

The dataset also contains 8 quantitative features. Each of these are hourly recordings of weather April 2006 to March 2016. These are:

- Temperature (C) – float
- Apparent Temperature (C) – float
- Humidity – float
- Wind Speed (km/h) – float
- Wind bearing (degrees) – float
- Visibility (km) – float
- Loud Cover – float
- Pressure (millibars) – float

3.2. Null Values

Missing values were found in both quantitative and categorical columns. To deal with the missing values in the quantitative column, Loud Cover consisted only of zeros and was removed from the dataset for analysis, and zeros in the Pressure column of the Dataset and was replaced with the medians of pressure close to them as we know that pressure never takes zero value in millibars. Also, there were some missing values in the categorical Precipitation Type column. Used the temperature to replace the zeros with possible precipitation type as it is dependent on the temperature.

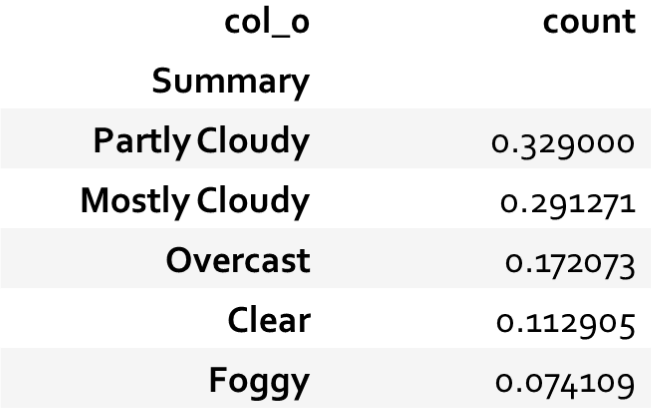
3.3. Outliers

The dataset did not contain any Outliers.

4. Exploratory Data Analysis (Data Story)

4.1. Proportion of most common categories per hourly summary.

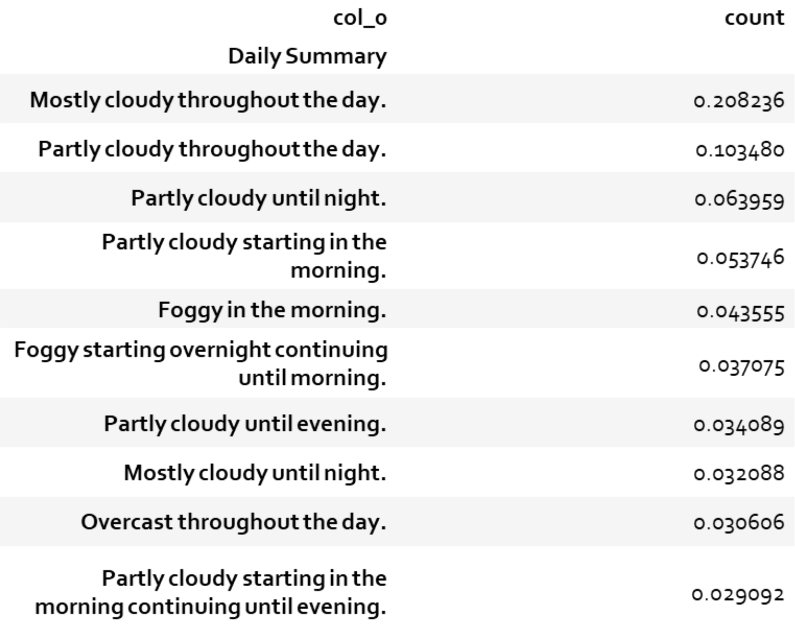
- Partly Cloudy (33%)
- Mostly Cloudy(29%)
- Overcast (17%)
- Clear (11%)
- Foggy(7%)
- Others(3%)



col_0	count
Summary	
Partly Cloudy	0.329000
Mostly Cloudy	0.291271
Overcast	0.172073
Clear	0.112905
Foggy	0.074109

4.2. Proportion of most common categories per daily summary.

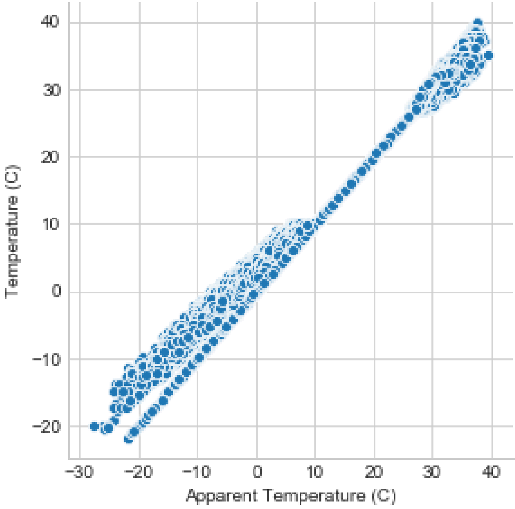
- Mostly cloudy throughout the day 21%
- Partly cloudy throughout the day 10%
- There was a total of 214 different Daily summaries.

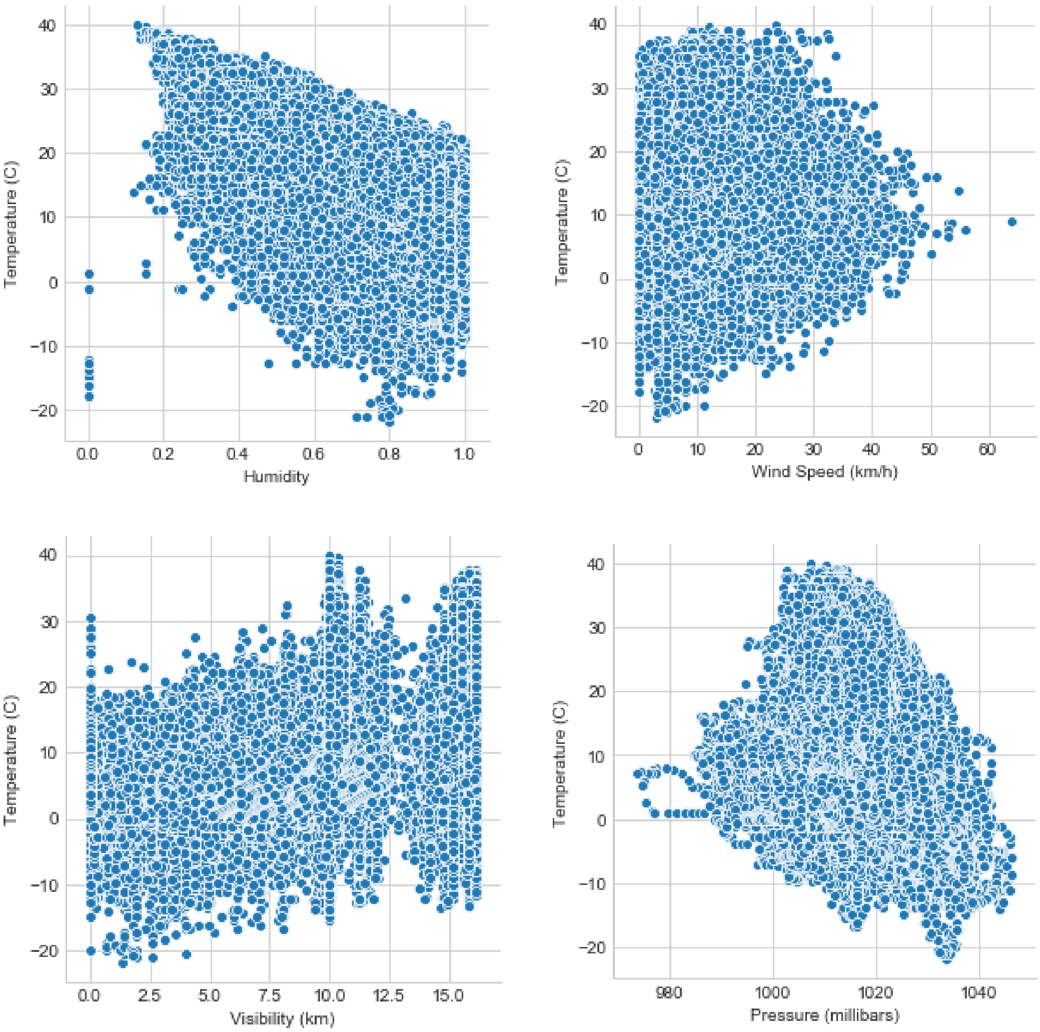


col_0	count
Daily Summary	
Mostly cloudy throughout the day.	0.208236
Partly cloudy throughout the day.	0.103480
Partly cloudy until night.	0.063959
Partly cloudy starting in the morning.	0.053746
Foggy in the morning.	0.043555
Foggy starting overnight continuing until morning.	0.037075
Partly cloudy until evening.	0.034089
Mostly cloudy until night.	0.032088
Overcast throughout the day.	0.030606
Partly cloudy starting in the morning continuing until evening.	0.029092

4.3. Data Visualization.

Graphs of Temperature with respect to other features plotted:

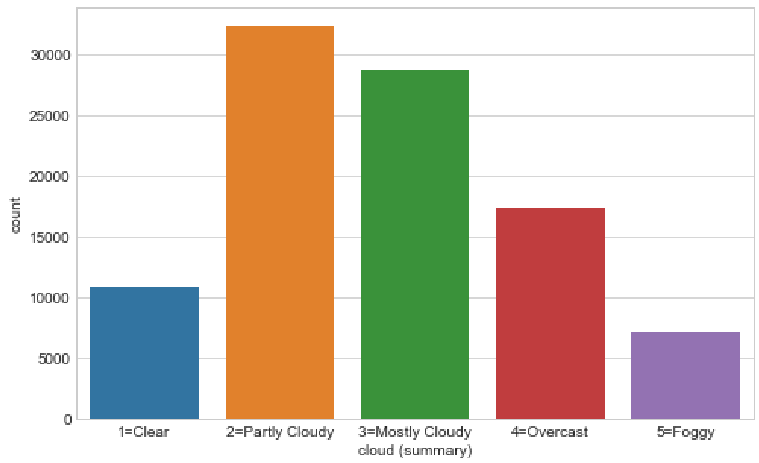



Temperature (C)
40
30
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0
Humidity
Temperature (C)
0
10
20
30
40
50
60
Wind Speed (km/h)
Temperature (C)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
Visibility (km)
Temperature (C)
980
1000
1020
1040
Pressure (millibars)

5. Feature Engineering

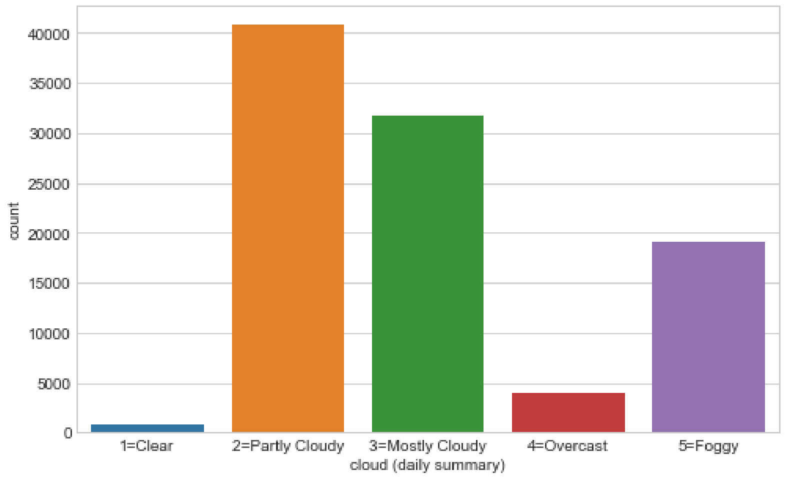
Feature 1:

- Cloud (summary)
- Used the hourly summary for this
- Replaced the zeros of the clouds in the summary and created a new column



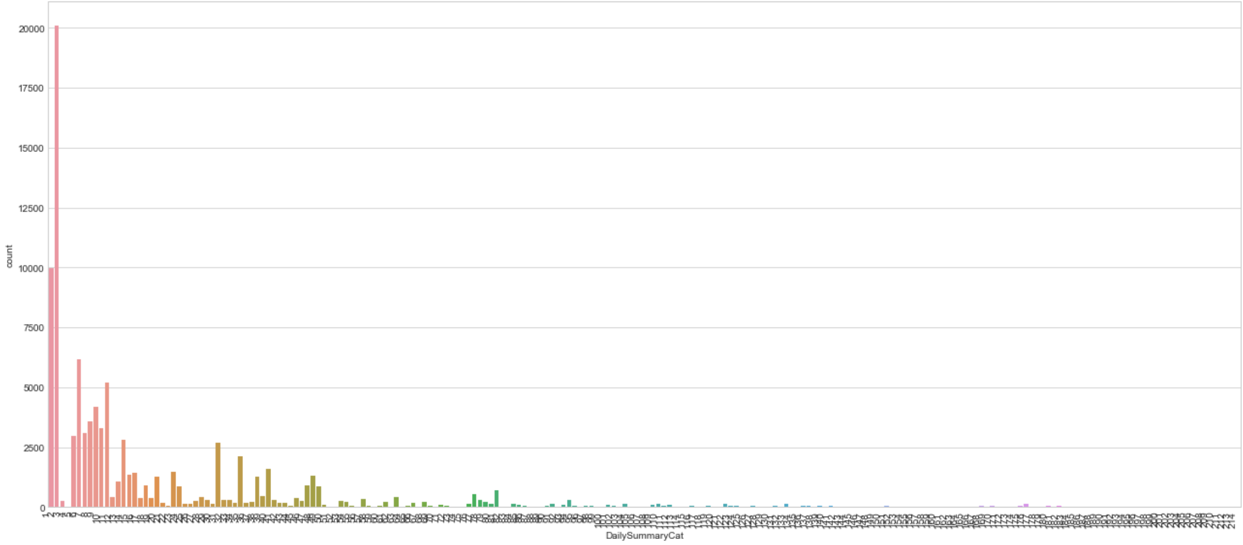
Feature 2:

- Cloud (daily summary)
- Used the daily summary for this



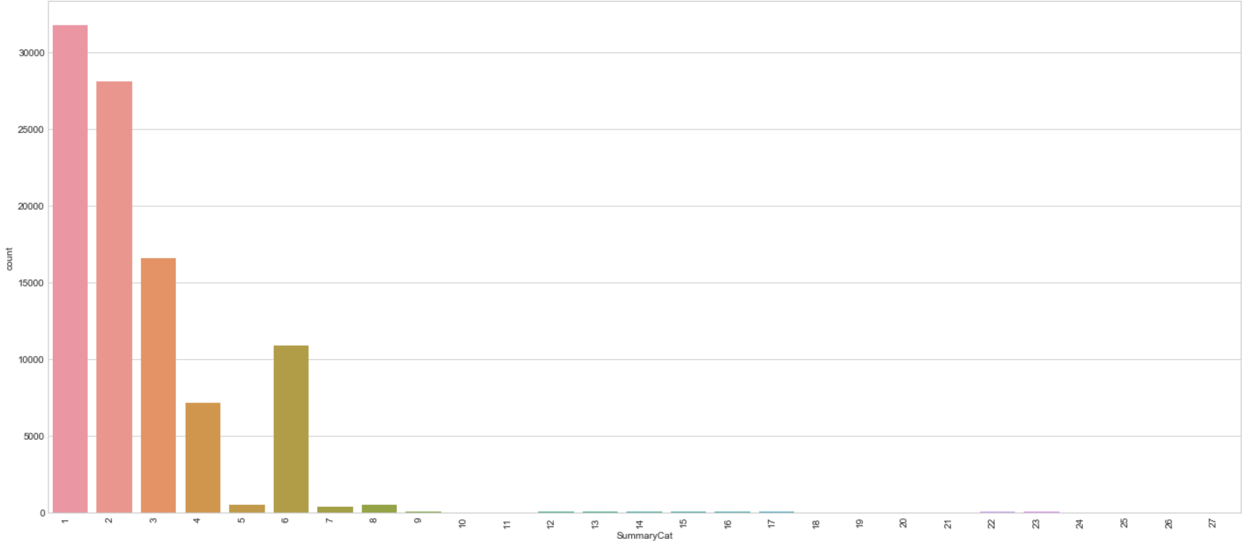
Feature 3:

- DailySummaryCat
- Used the daily summary for this to create a new feature that converted the Daily summary to a unique integer value



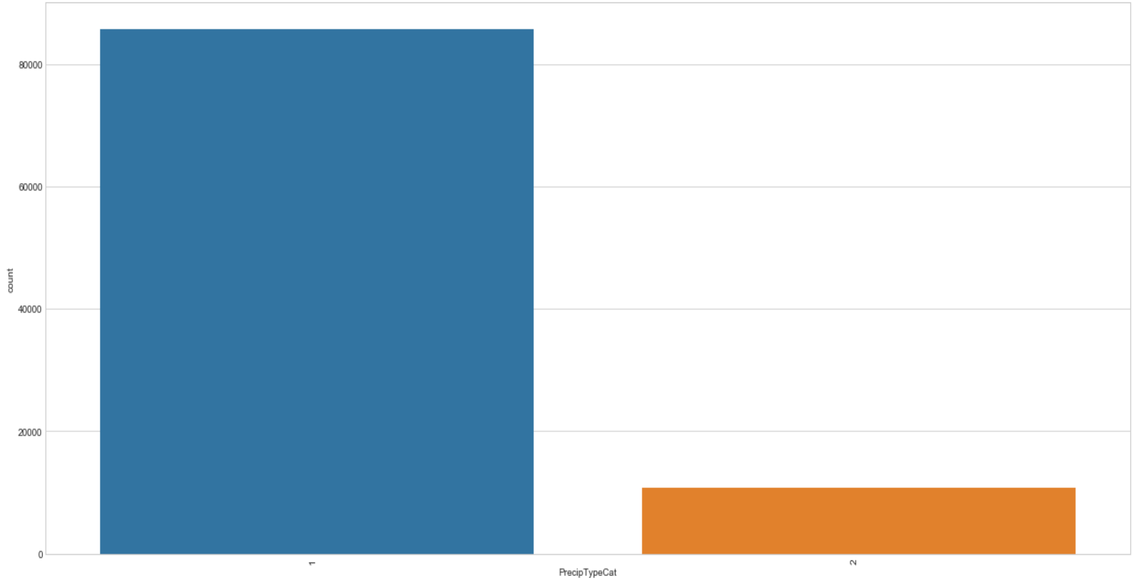
Feature 4:

- SummaryCat
- Used the hourly summary for this to create a new feature that converted the hourly summary to a unique integer value



Feature 5:

- PrecipTypeCat
- Used the Preicitation type for this to create a new feature that converted it to a unique integer value



6. Machine Learning and Results

For the in-depth analysis of the default of credit card client's dataset retrieved from uci repository following machine learning models and techniques were applied.

1. Logistic Regression
2. Logistic Regression after Feature engineering
3. Deep Neural Network Regressor after Feature engineering
4. Long Short-Term Memory after Feature engineering

In order to fit the machine learning models, first, the dataset, which was imported as a dataframe using pandas library, was used to make X(predictors) and y(target). Second, the data was split into test and training sets using the sklearn library.

6.1. Logistic Regression

Linear regression on Temperature as a function of Humidity gave:

- RMSE: 7.41.

6.2. Logistic Regression after Feature engineering

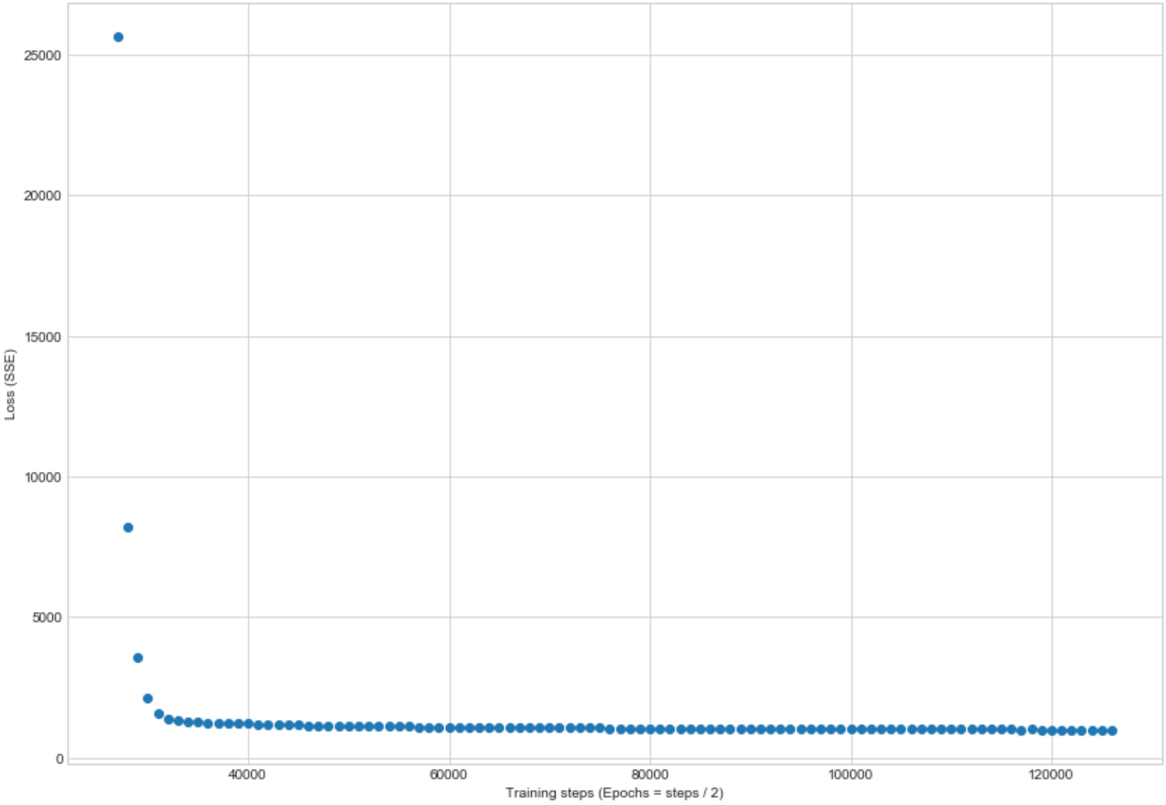
Linear regression on Temperature using feature engineered dataset gave:

- RMSE: 0.933

6.3. Deep Neural Network Regressor after Feature engineering

Deep Neural Network Regressor using feature engineered dataset gave:

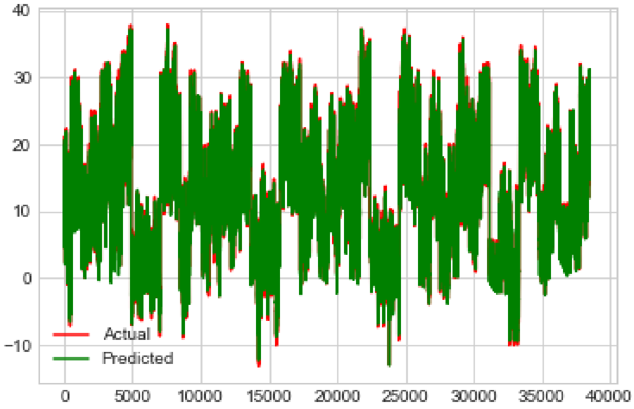
- MSE: 26.559624
- RMSE: 5.153
- Label/mean: 11.48189
- SSE: 25618.418
- Prediction/mean: 11.390016
- Global step: 27000
- The Explained Variance: 0.99
- The Mean Absolute Error: 0.80 degrees Celcius
- The Median Absolute Error: 0.66 degrees Celsius



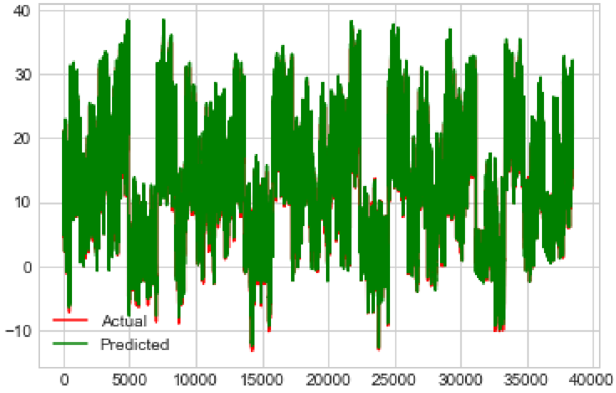
6.4. Long Short-Term Memory after Feature engineering

Long Short-Term Memory using feature engineered dataset gave:

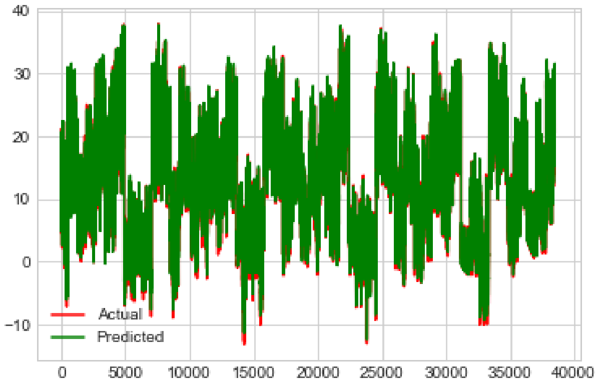
- LSTM RMSE: 1.136



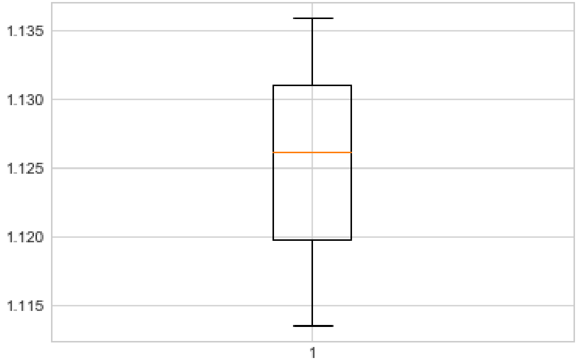
- LSTM RMSE: 1.126



- LSTM RMSE: 1.113



- LSTM RMSE: 1.125 (+/- 0.009)



6.5. Result Table

LINEAR REGRESSION	RMSE: 7.41
LINEAR REGRESSION (FE)	RMSE: 0.933
DEEP NEURAL NETWORK REGRESSOR (FE)	RMSE: 5.153, MEAN ABS ERROR: 0.80 °C, MEDIAN ABS EROR: 0.66 °C
LONG SHORT-TERM MEMORY (FE)	RMSE: 1.125

7. Conclusion

When it comes to weather prediction, both linear regression and functional regression were indicating that over longer periods of time, our models may result in more accurate prediction. Linear regression proved to be a low bias, high variance model whereas Deep Neural Network Regressor and Long Short-Term Memory proved decent. Linear regression is inherently a high variance model as it is unstable to outliers, so one way to improve the linear regression model is by collection of more data. However, this would require much more computation time along with retraining of the weight vector, so this will be deferred to future work.

8. Analysis Notebooks

All of the Python code and Jupyter notebooks used in this project can be found on GitHub:
https://github.com/taha-shahid/SpringBoard/tree/master/Capstone_2