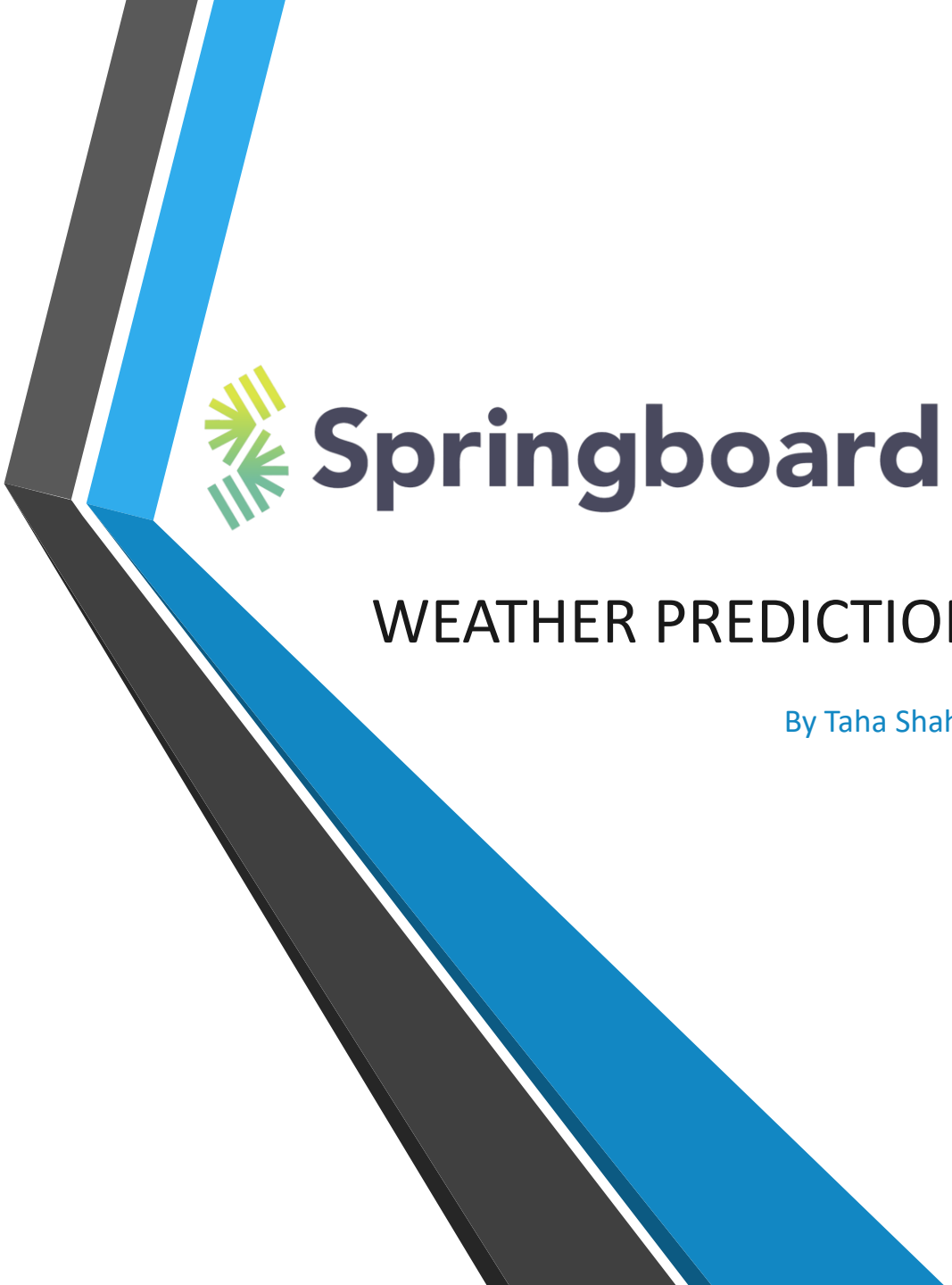


WEATHER PREDICTION

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.

Data Science Career Track - Springboard

1. Introduction

- The data set selected is called “Weather in Szeged 2006-2016” and it is a Time-Series Dataset which has hourly recording of weather for 10 years.
- Predicting weather conditions and temperature with multiple techniques and manipulating the dataset for better algorithms and neural networks.
- Client will be weather solutions companies in Szeged because these companies will be utilizing the analysis and machine learning models to better prepare and advise their clientele regarding what safety measure would be necessary due to the weather conditions.

2. Motivation

- Weather patterns have always been really sparse in the Midwest region of USA.
- There was a weather vortex that occurred in the Midwest couple months ago.
- Predicting weather conditions and temperature changes was then decided as the project for analysis.

3. Project Description

- The problem statement was prediction of weather conditions and temperature changes with other changing criteria's.
- The dependent variable in this project was the Temperature in Degree Celsius.
- Some important variables that played a part in the weather prediction consisted of:
  - Hourly and Daily Summary of the weather.
  - Humidity
  - Wind Speed
  - Pressure

4. Data Acquisition and Management

- This dataset was acquired from the Kaggle repository.
- Dataset was available in the csv file format.
- The dataset had 96453 rows of observation.
- The dataset consisted of 12 columns:
  - 3 categorical columns
  - 8 quantitative columns
  - 1 datetime column

5. Dataset Description

- Categorical columns for the dataset:
- Formatted Date – Datetime column
- Hourly Summary – As string object
- Precipitation Type – As string object
- Daily Summary – As string object

5. Data Description

- Quantitative columns for the dataset:
  - 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

6. Data Cleaning

- This dataset had some missing values in the quantitative columns such as:
  - Loud Cover consisted only of zeros and was removed from the dataset for analysis
  - There were some zeros in the Pressure column of the Dataset and was replaced with the medians as we know that pressure never takes zero value in millibars.
- For the sake of running the classifiers with no issues a float to integer conversion was taken place for the Temperature Column.

6. Data Cleaning CONT'D

- Converted string columns to integer column by assigning unique numbers to a particular hourly and daily summary.
- 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.
- Cleaned the data column of summary with respect to visibility and compared the data to find similar trend of visibility to replace the zeros of clouds type.

7. Statistical Analysis

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

7. Statistical Analysis (CONT'D)

Most Common Categories per Daily Summary

- Mostly cloudy throughout the day

  21%

- Partly cloudy throughout the day

  10%

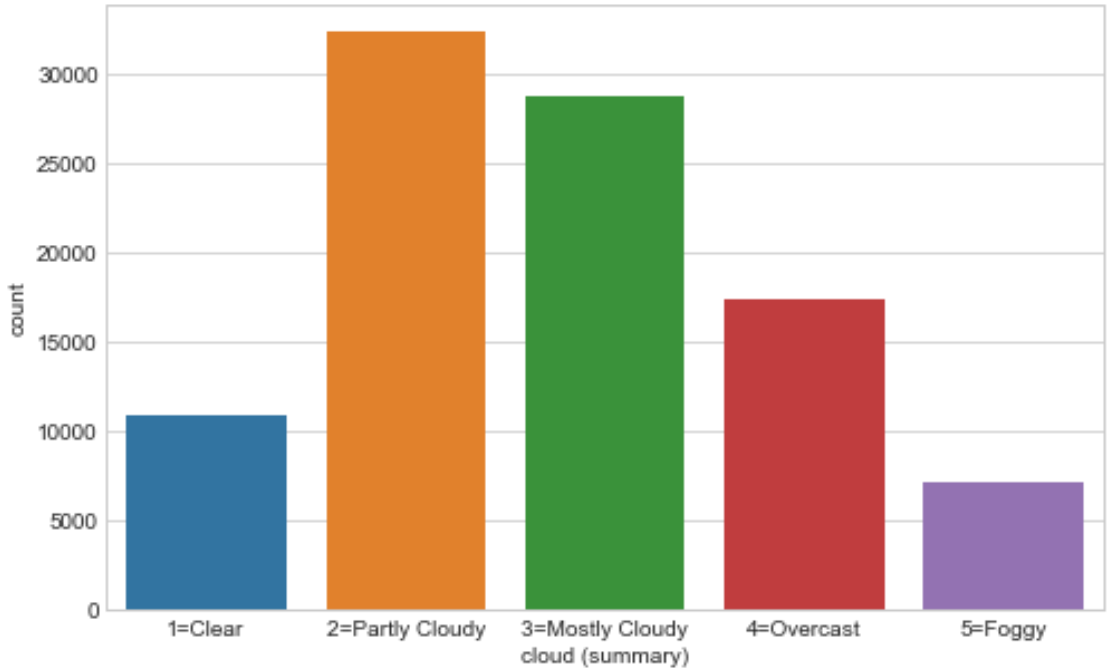
- There were a total of 214 different

  Daily summaries.

col_0 Daily Summary	count
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

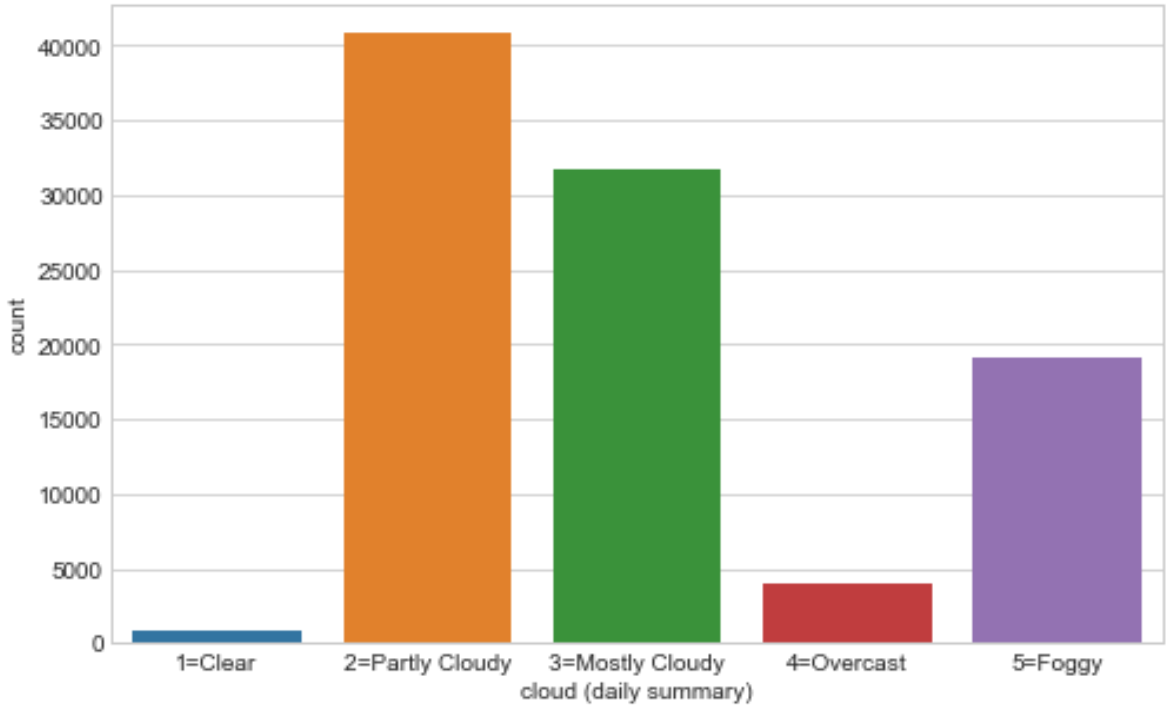
8. Feature Engineering

- Feature 1:
  - Cloud (summary)
- Used the hourly summary for this
- Replaced the zeros of the clouds in the summary and created a new column



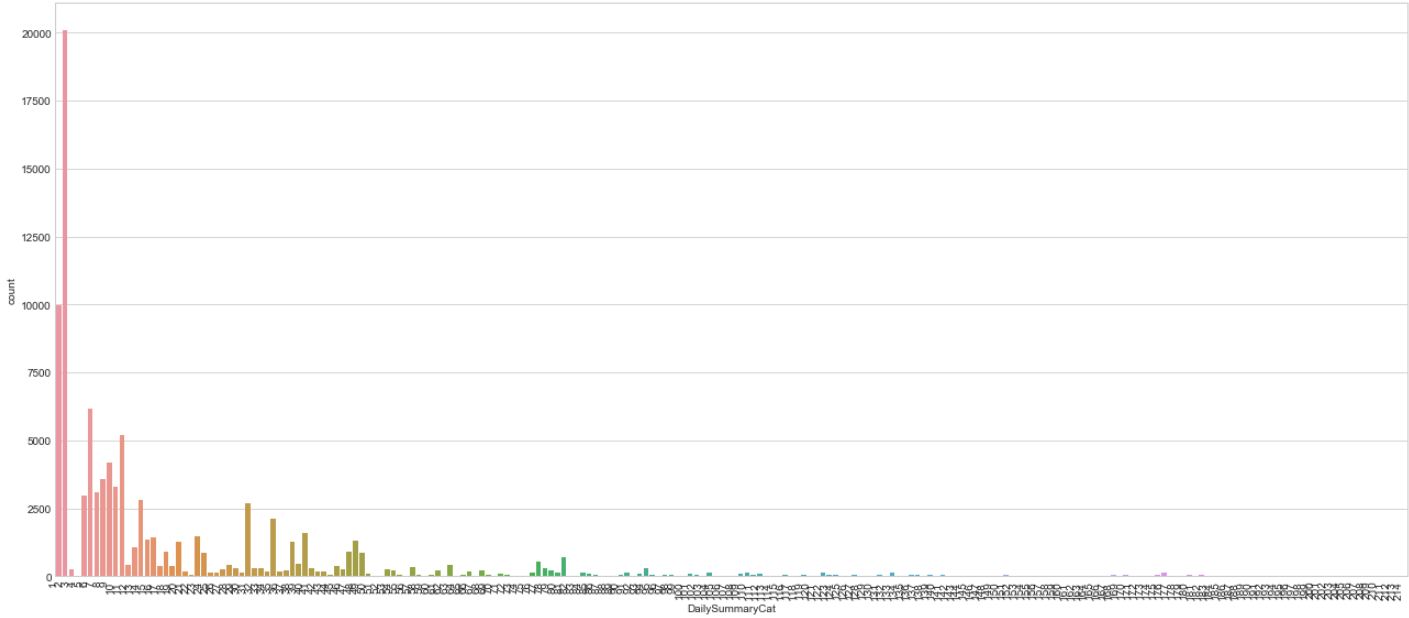
8. Feature Engineering

- Feature 2:
  - Cloud (daily summary)
- Used the daily summary for this



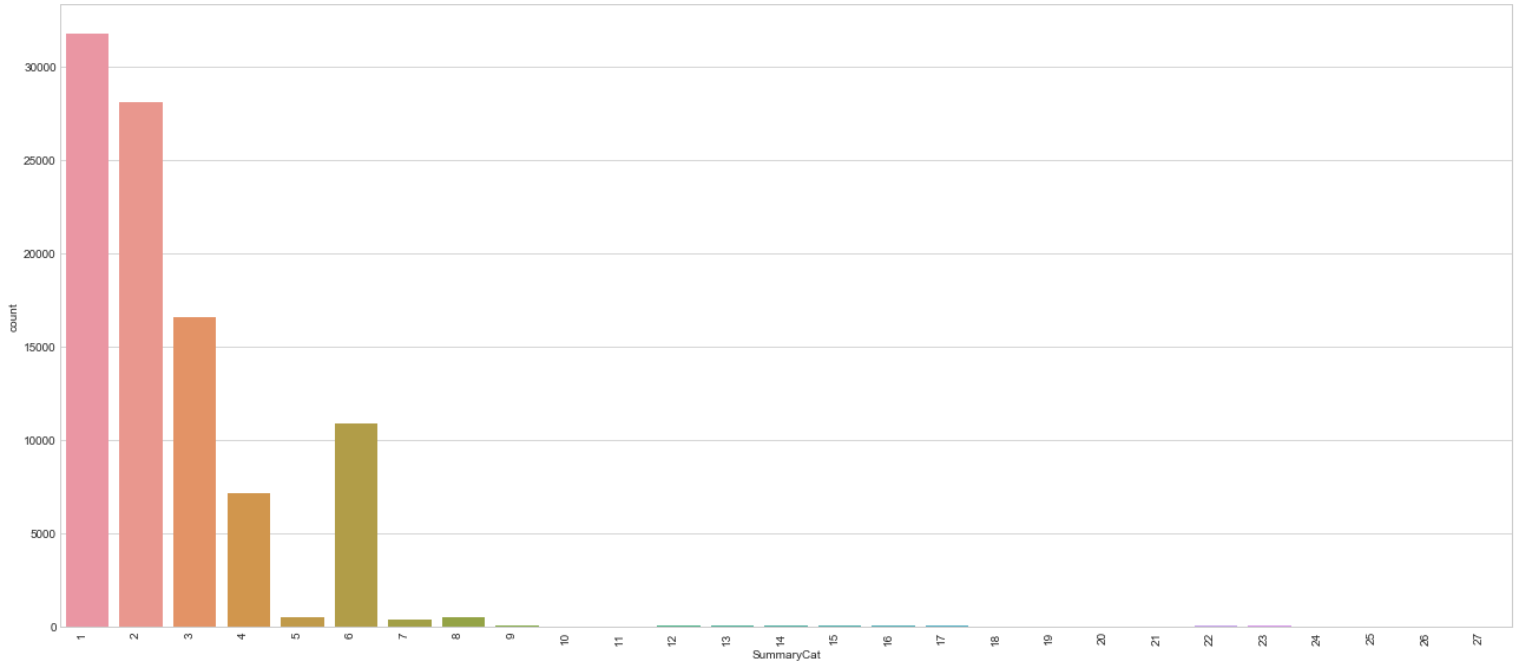
8. Feature Engineering

- Feature 3:
  - DailySummaryCat
- Used the daily summary for this to create a new feature that converted the Daily summary to a unique integer value



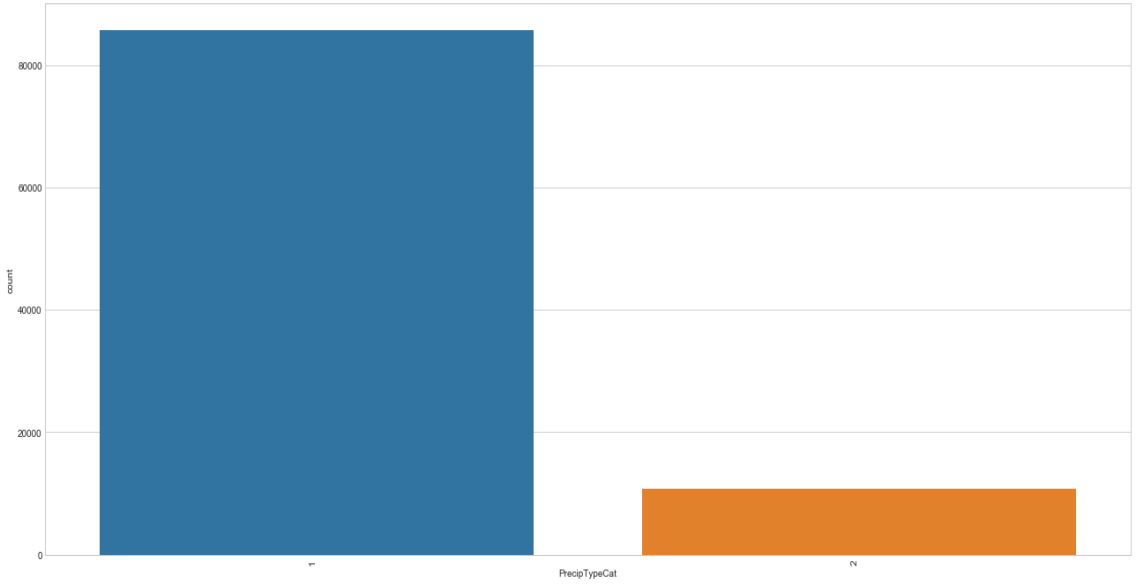
8. Feature Engineering

- Feature 4:
  - SummaryCat
- Used the hourly summary for this to create a new feature that converted the hourly summary to a unique integer value



8. Feature Engineering

- Feature 5:
  - PrecipTypeCat
- Used the Preicitation type for this to create a new feature that converted it to a unique integer value



9. Classifiers List

- Linear Regression
- Deep Neural Network Regressor
- LSTM

10. Machine Learning Results

Linear Regression

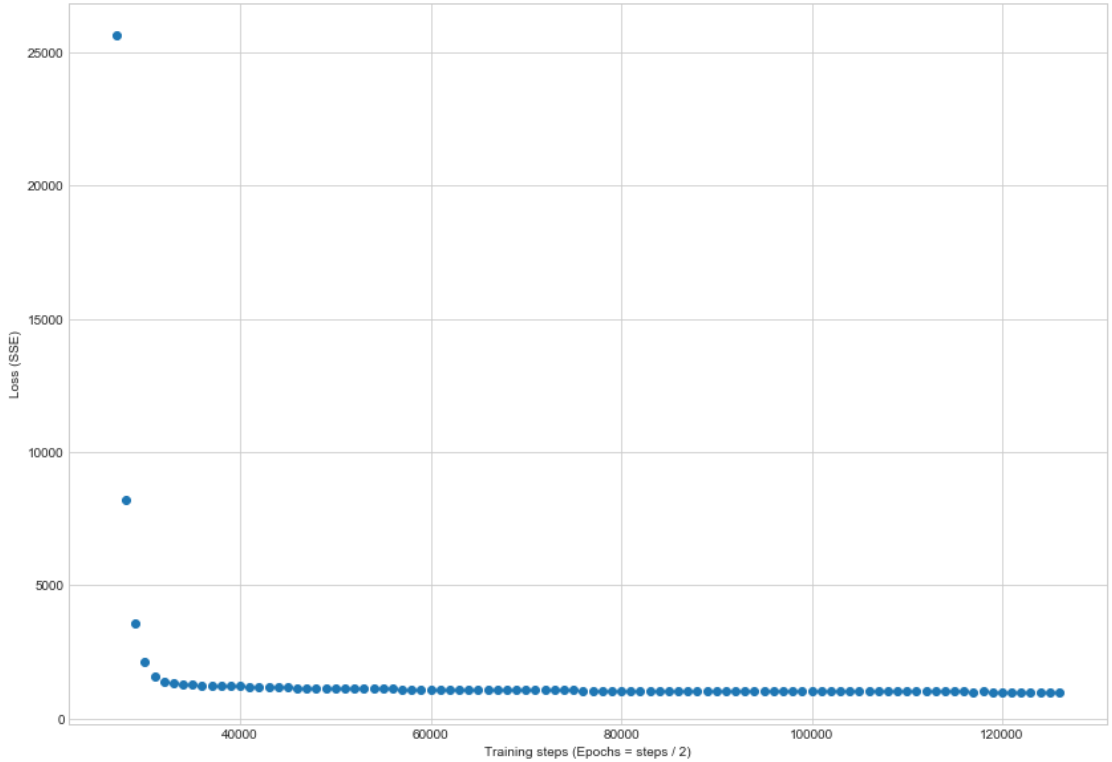
- Linear regression on Temperature as a function of Humidity
- RMSE: 7.41

- Linear regression on Temperature using feature engineered dataset
- RMSE: 0.933
- Test Score: 99.05%

10. Machine Learning Results

Deep Neural Network Regressor

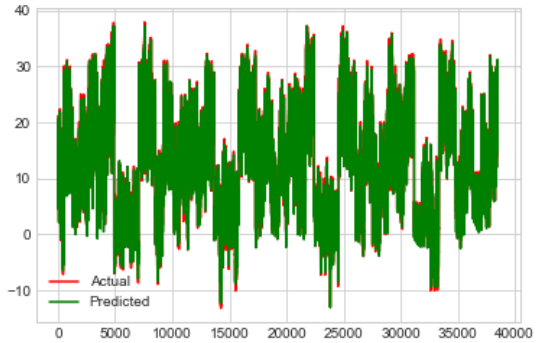
- MSE: 26.559624
- RMSE: 5.153
- Label/mean: 11.48189
- SSE: 25618.418
- Prediction/mean: 11.390016
- Global step: 27000
- The Explained Variance: 0.99
- The Mean Absolute Error: 0.80 degrees Celcius
- The Median Absolute Error: 0.66 degrees Celcius



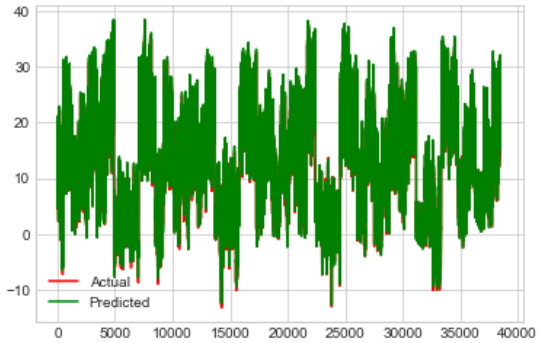
10. Machine Learning Results

Long Short-Term Memory

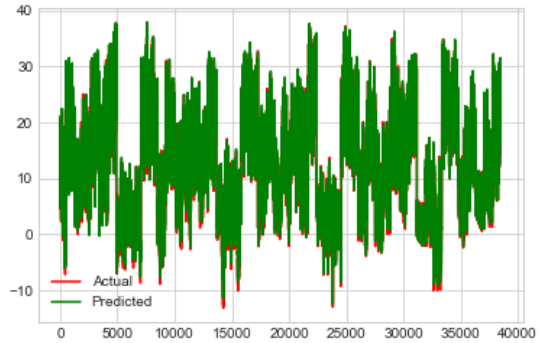
- LSTM RMSE: 1.136



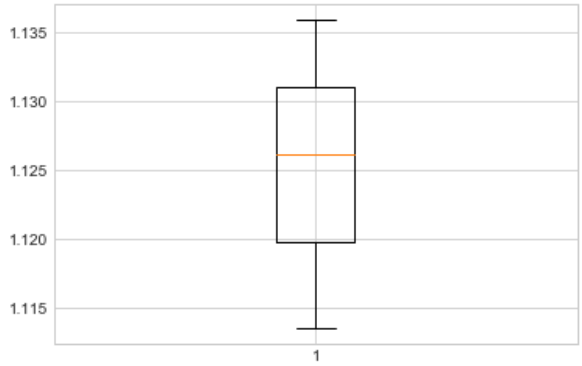
- LSTM RMSE: 1.126



- LSTM RMSE: 1.113



- LSTM RMSE: 1.125 (+/- 0.009)



10. Machine Learning Results

LINEAR REGRESSION	RMSE: 7.41
LINEAR REGRESSION (FE)	RMSE: 0.933, TEST SCORE: 99.05%
DECISION TREES (FE)	RMSE: 4.38, TEST SCORE: 31.79%
LOGISTIC REGRESSION (FE)	RMSE: 3.13, TEST SCORE: 21.71%
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

Thank You!

By Taha Shahid