HDSC Summer ‘22 Premiere Project Presentation: Weather Prediction

A Project by Team Seaborners

Predicting Weather Conditions In NewYork City(A Time-Series Forecasting Analysis)

Introduction

Weather prediction and forecasting is the application and combination of various scientific and technological concepts to predict the atmospheric condition of any place at a given time. Weather forecast remains an important factor that affects our everyday lives. Globally, meteorologists face an onerous task of predicting future weather conditions of a particular place, city, country, and region.

Weather forecasting plays an important role in our daily lives, from farmers in the agricultural industry, pilots in airspace navigation, athletes, artists, and understanding renewable energy. Weather forecasting technology is projected to grow at an estimated compounded annual growth rate of 4.25% from USD 2.29 billion in 2021 to USD 2.38 billion in 2022[1].

The weather prediction for this project is based on identified and collected past dataset (a five-year hourly record) of various weather attributes such as temperature, humidity, air pressure from about 36 cities across Canada, Israel, and the United States.

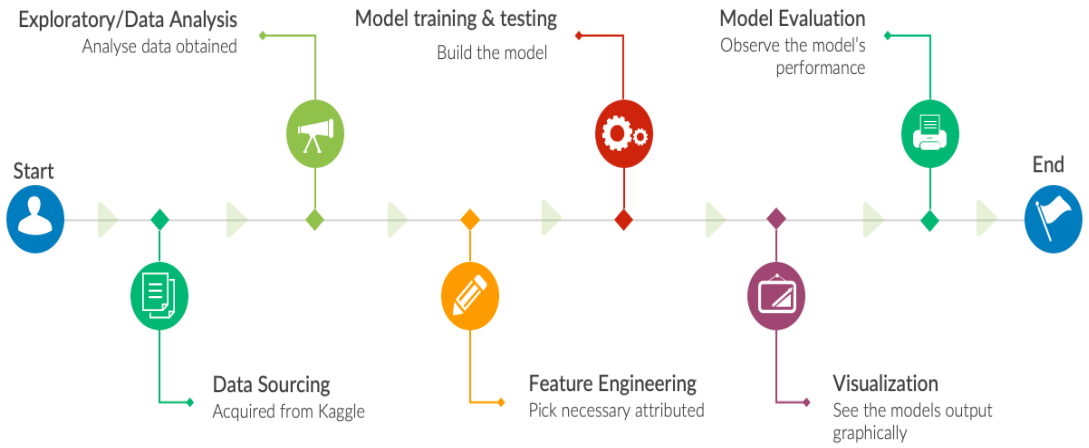
Objective/Problem Statement

The objective of this project is to build a machine learning model that predicts the future weather condition of a city. The model incorporates various features and is deployed to predict interactive future weather conditions of New York City. We aim to solve the problem of inaccurate weather predictions by forecasting the average temperature of New York City using traditional machine learning algorithms. To this end, we decided to train and test our model, using various permutations to illustrate signal processing concepts such as filtering, Fourier transform, auto-correlation, cross-correlation.

[1] ReportLinker

Methodology – Model Training and Evaluation

The model employed in this project will be represented by the following flow diagram in **Figure 1**:



Cleansing and Data Processing

The pre-processed data was downloaded from Kaggle. From the data, all the various weather attributes were studied, and the team decided to draw insights from the temperature attribute in our study for the impact it has on average New Yorkers. The temperature weather attribute had 45253 entries with 793 null values. For the dataset to remain constant, the null values were filled using forward and backward filling techniques. The dataset was then transformed to include days, months, and years.

Exploratory Data Analysis

Descriptive Statistics

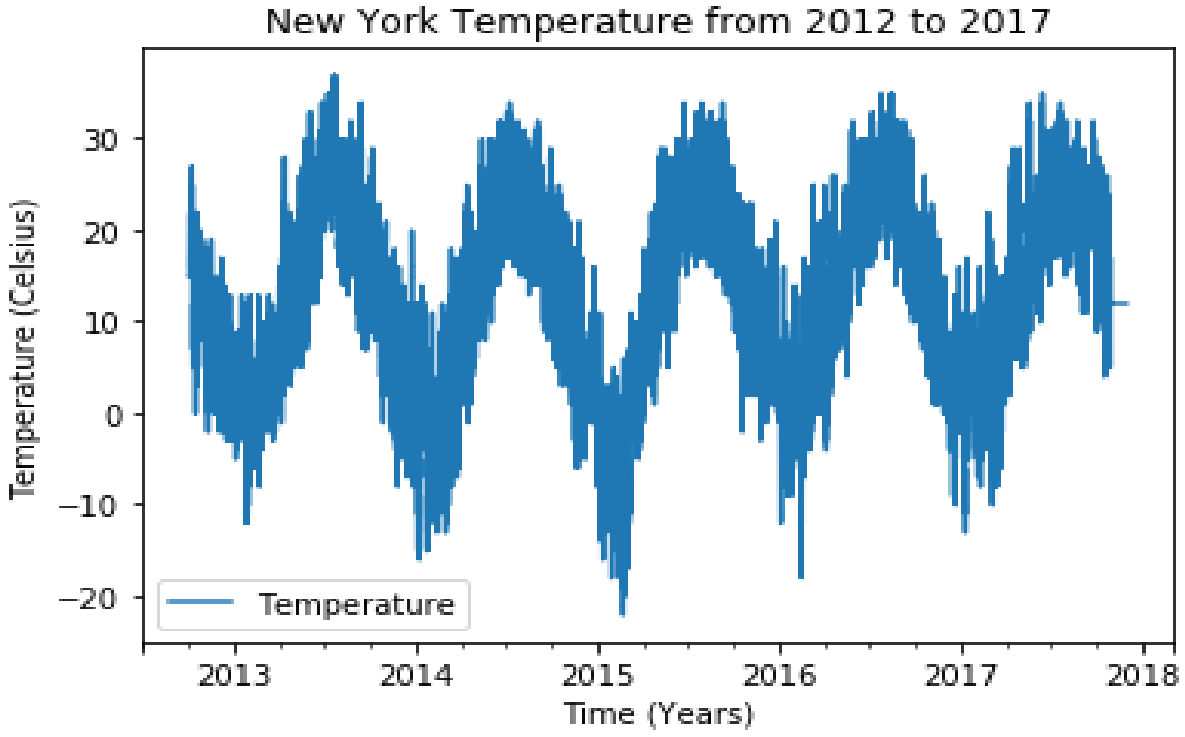
The minimum and maximum temperature in New York between 2012 and 2017 was 250.77k and 310.24k respectively. The average temperature was 285.40K, representing a low temperature in New York.

Figure 2 below represents a descriptive statistics of the dataset

```
count    44460.000000
mean       285.400406
std         10.220932
min        250.774000
25%        277.370000
50%        285.870000
75%        293.760000
max        310.240000
Name: Temperature, dtype: float64
```

Additionally, a plot of temperature readings between 2012 to 2017 was plotted to identify any trend, patterns, outliers, and insights from the dataset.

Figure 3 below represents the yearly data of temperature in New York City between 2012 and 2017.

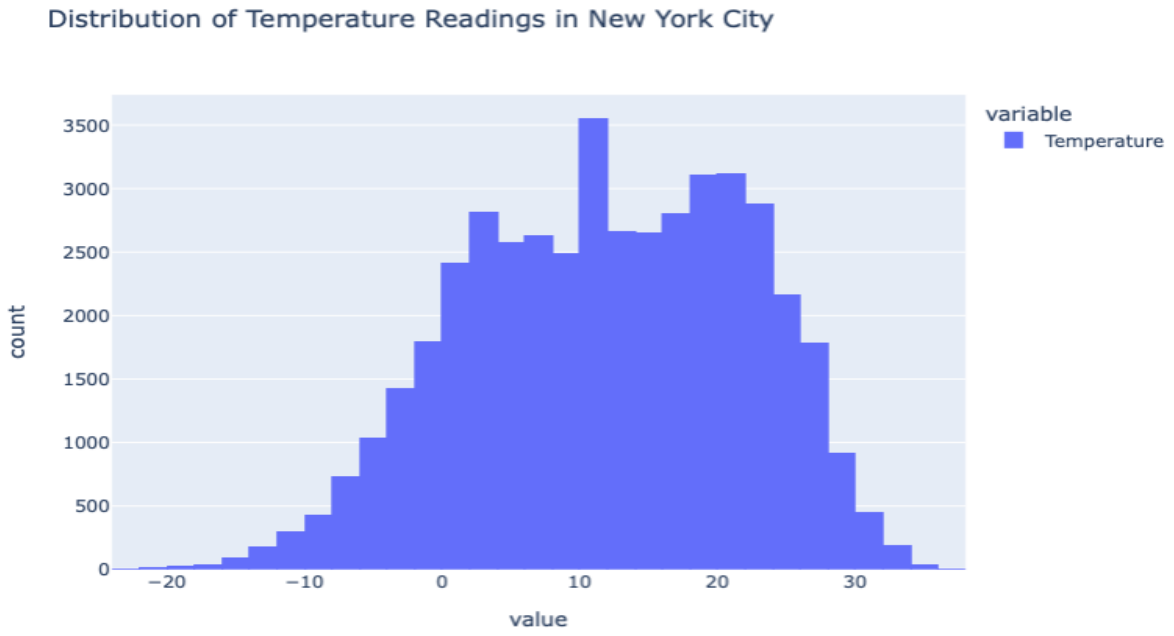


We observe that the lowest temperatures in the dataset was between Feb and March 2015, with an upward trend in the overall temperature. The NYC panel on climate change[2] records that the prevalence of high temperatures will likely continue and might triple by 2080. The trend in overall temperature conforms to the NASA (2016) report[3] that confirms the assertion that the earth's surface temperature was the warmest since modern record keeping began in 1880.

Normal Distribution

The shape of the temperature attribute follows a normal distribution curve, conveying that it possesses a uniform seasonal change, and confirming its time series properties.

Figure 4 below represents the normal distribution properties of our temperature attribute

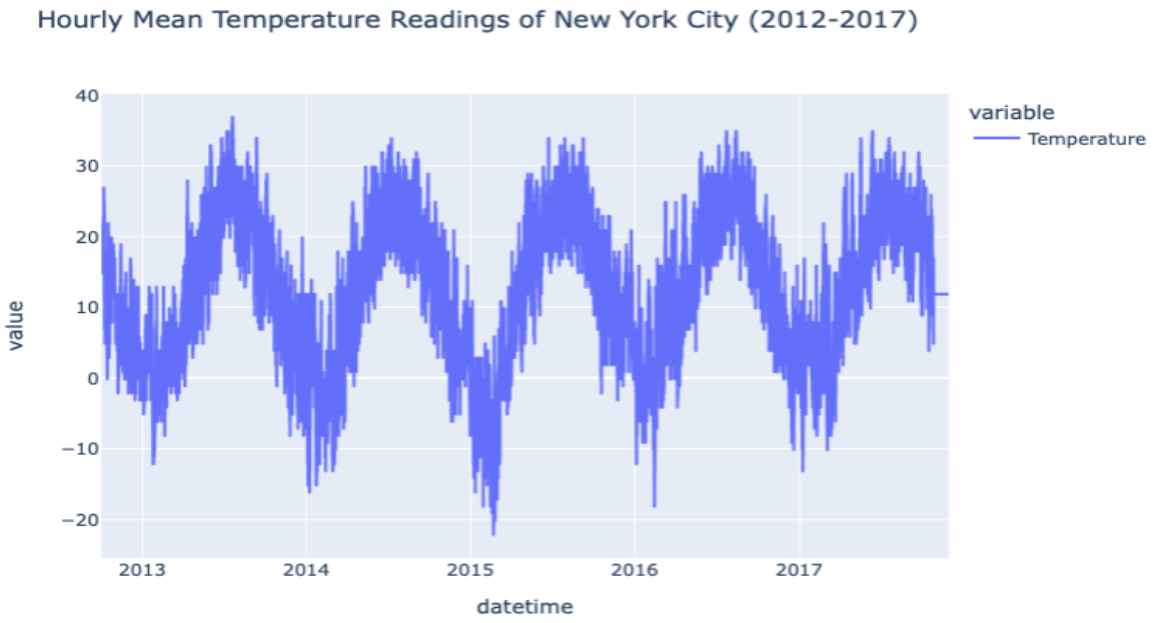


A quick representation of the hourly temperature (Figure 5) suggests that there is a trend with the temperature in New York City as we can see that the readings of the temperature over the years are in the form of a sine wave. It is very pertinent to note that there was an upward trend in temperature in New York City during this period. This could be because of the effects of global warming.

Figure 5 below represents the hourly mean temperature of New York City

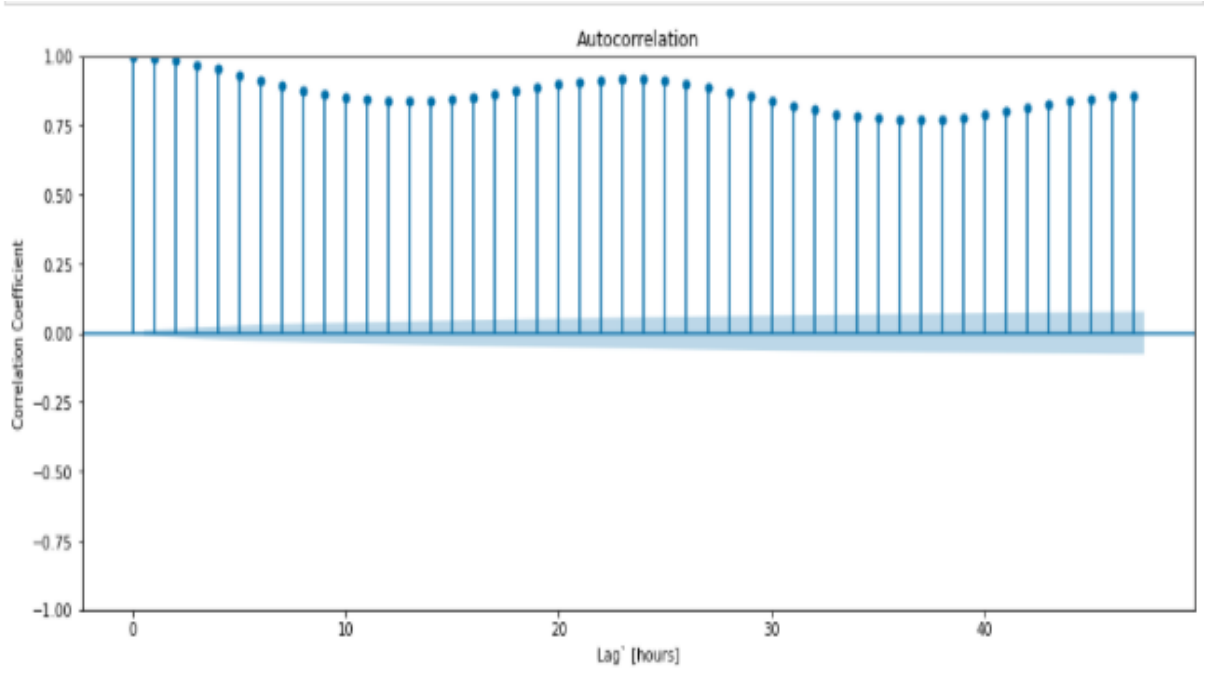
[2] https://nyaspubs.onlinelibrary.wiley.com/doi/epdf/10.1111/nyas.12591

[3] https://www.nasa.gov/feature/goddard/2016/climate-trends-continue-to-break-records



Autocorrelation

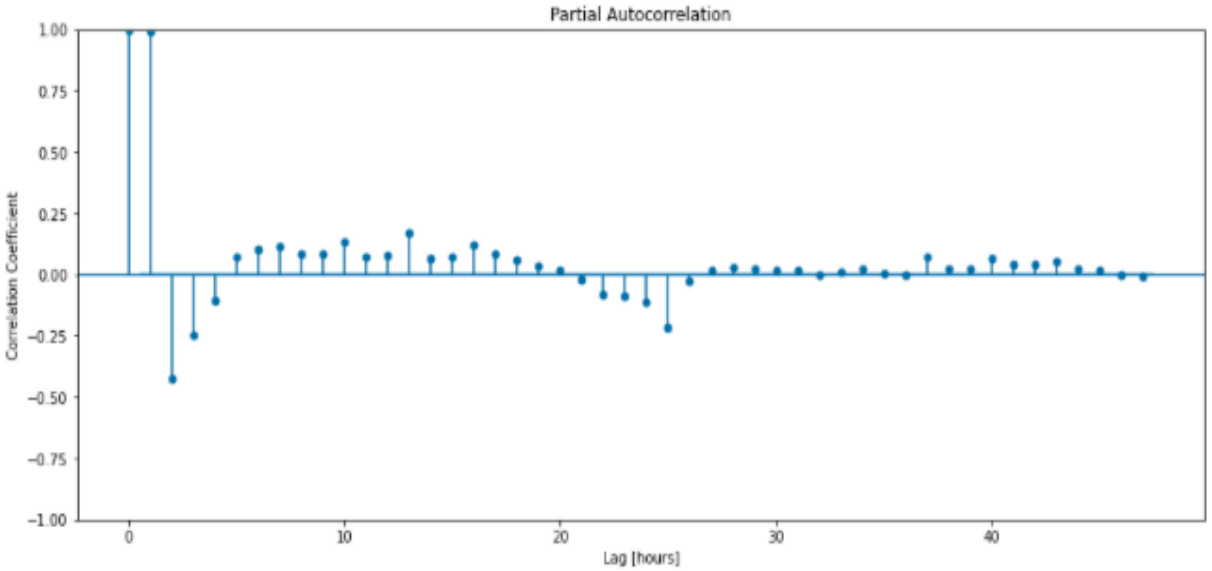
The model tests the autocorrelation of temperature of New York city and the autocorrelation function is very close to one, suggesting that it is statistically significant and autocorrelated against its own lag. Figure 6 below represents the autocorrelation of our temperature dataset with all our current data points hovering around its past values.



Partial Auto Correlation Coefficient

The model tests the partial correlation coefficient of temperature of New York city and reveals that the model is statistically significant. The partial correlation coefficient function after the first 2 lags is very low.

Figure 7 below represents the Partial Correlation Coefficient of our temperature attribute in New York City



Model Building and Tuning

The model used to predict the time series data was an Auto Regressive model, and we tuned the model's hyperparameters to find the best parameter for accurately predicting our data. Mean Absolute Error was the evaluation metric we used to evaluate our model based on the results of those hyperparameters. After deciding on the best hyperparameter, we used the walk forward validation technique to create our final model, which will aid us in predicting our data accurately.

Figure 8 below describes the range of values for the model hyperparameter tuning and evaluation. We got the lowest mean absolute error (best fit model) with the best set of hyperparameters after 44 iterations of hyperparameters tuning.

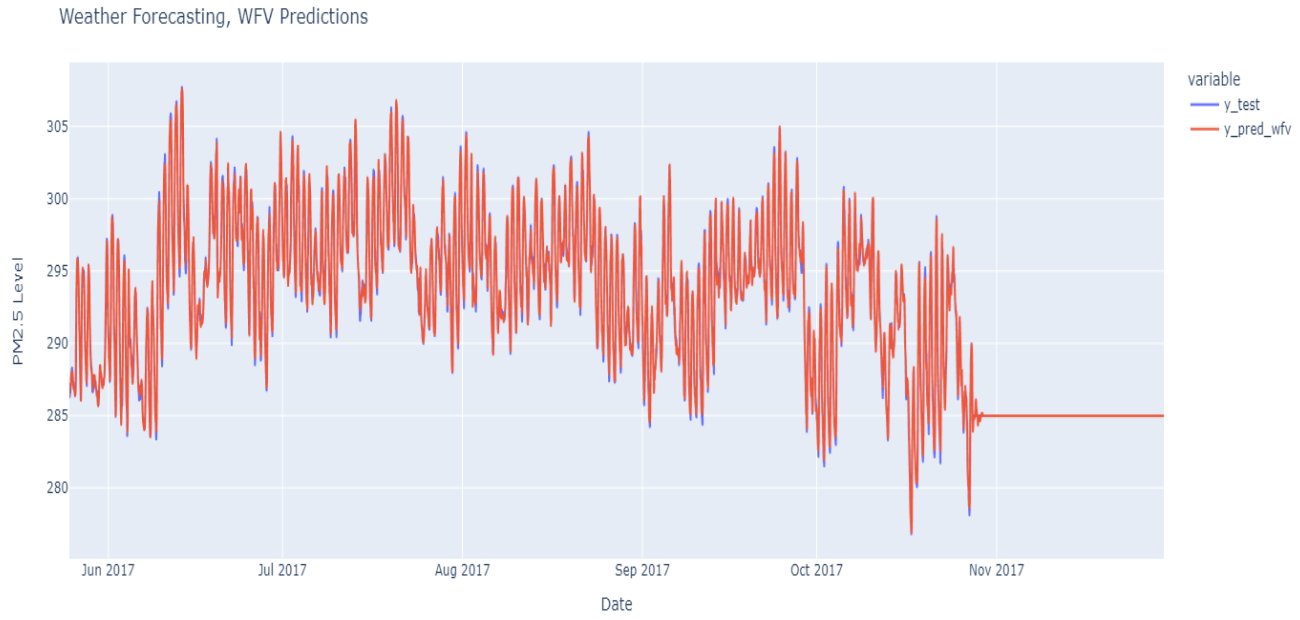
```
43    0.539295
44    0.539317
42    0.539919
41    0.540170
40    0.540445
Name: mae, dtype: float64
```

Model Evaluation

The model predicted the temperature for an extra hour by training a subset of the sample dataset and then using it to forecast for the next hour. This dataset contains temperature for New York City from 2012-2017, to ensure that individuals and meteorologists can predict accurate weather for the next hour and ensure that everyone is adequately prepared for either rain, sun, snow or fall. The mean absolute error of the model is calculated as 0.36939 and R2 score gave a 99% accuracy report on the model.

Visualization of Model Output

Figure 9 below represents the overall forecasting model



Conclusion

In sum, the dataset reveals that on the 20th of July 2013, the temperature was at its third hottest in the history of New York city, conversely the temperature in New York was the lowest on 20th of February 2015. It is of utmost importance that our collective actions and inactions continue to contribute to the global warming of the earth surface and a rapid departure from our usual day-day is needed to save the planet. The effects of global warming and climate change are seen all over the city of New York and by extension the World.