WEATHER FORECASTING APPLICATION

BUILDING DS PRODUCT – STAGE 2

Course:
Data Science Product Development
(50742)

Group members:
Arsalan Shaikh - 23388
Ramsha Arif - 23395
Sarosh Aamir - 23390

BUSINESS OBJECTIVE

Weather forecasting is the application of scientific techniques and technology to predict the conditions of the atmosphere at a certain location and time. Weather Forecasting in old time is carried out by hand, using changes in barometric pressure, current weather conditions, and sky condition or cloud cover, weather forecasting now relies on computer-based models that take many atmospheric factors into accounting now relies on computer-based models that take many atmospheric factors into account.

Our main goal is to analyze how the different machine learning techniques will perform in the forecasting of weather combining time series with the neural network.

This project is divided into two parts, namely a research component and an application component, aiming to provide users with weather **predictions** using different machine learning models along with providing a good user experience with **interactive interface**.

UNDERLYING DATA SCIENCE PROBLEM

In this case, a software system can learn from data for improved analysis. Compared to traditional demand forecasting methods, machine learning helps to:

→ Accelerates data processing speed
→ Provides a more accurate forecast
→ Automates forecast updates based on the recent data
→ Analyzes more data
→ Identifies hidden patterns in data
→ Creates a robust system
→ Increases adaptability to changes

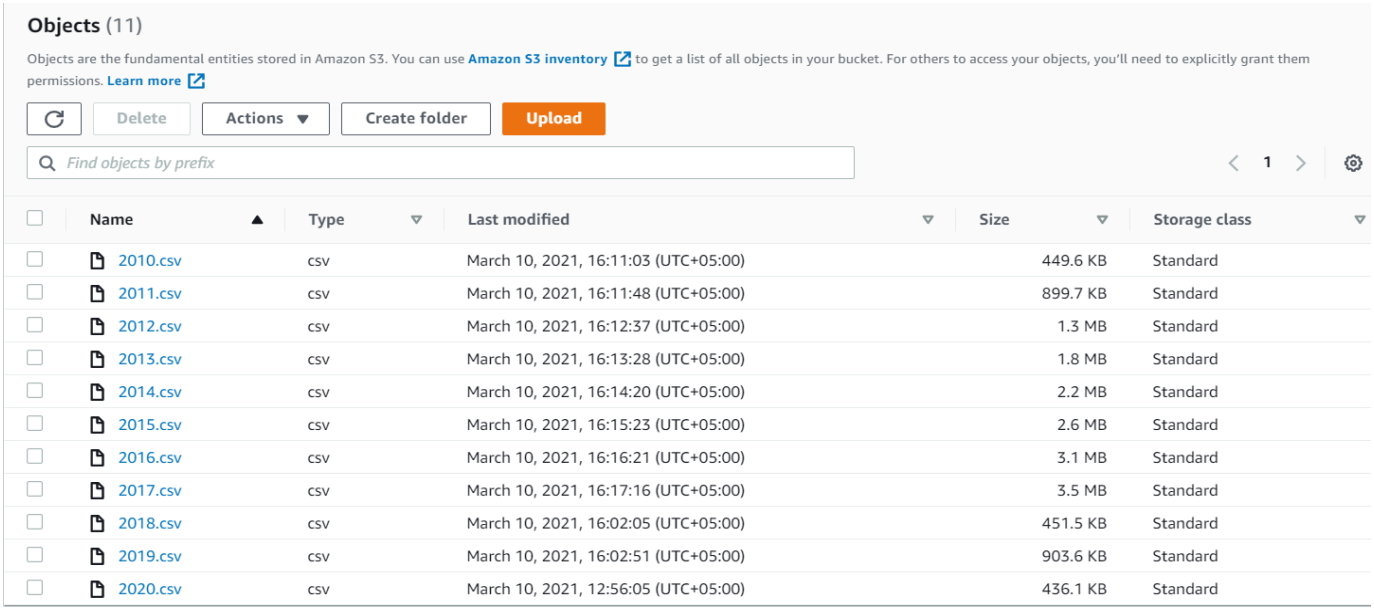
TOOLS & TECHNIQUES

Following are the details that have been achieved successfully and has been implemented practically:

- → **Dataset-** hosted on S3 bucket (AWS).
- → **Python-** is being used as a programming language.
- → **Boto3-** has been used to create, configure, and manage AWS services into colab.
- → **Fast API-** is used for building APIs.
- → **Docker Container –** is be used to manage application separately**.**
- → **Other Libraries-** Pandas, Scikit-Learn, TensorFlow, Keras etc.
- → **Regression Models-** Random Forest and Decision Tree
- → **Time-Series Based-** Fb Prophet and LSTM

DATA SET

- The dataset that has been picked from World Weather Online website.
- It provides us historical data as well as gives us information with respect to location that we choose. Here, “Karachi, Pakistan” data has been collected.
- The source data contain various features that can help to find optimum prediction for the data. Moreover, time-series data is used to obtain accuracy and apply some advance Machine Learning and Deep Learning algorithms.
- Through their API, last 11 years historical data has been fetched.
- After which we have converted the JSON response into .csv format, segregating it into each year’s file.
- All of these files have been **stored in Amazon S3 Bucket**. By using boto3, the csv data was called from S3 and converted it into our dataset.



Name	Type	Last modified	Size	Storage class
2010.csv	csv	March 10, 2021, 16:11:03 (UTC+05:00)	449.6 KB	Standard
2011.csv	csv	March 10, 2021, 16:11:48 (UTC+05:00)	899.7 KB	Standard
2012.csv	csv	March 10, 2021, 16:12:37 (UTC+05:00)	1.3 MB	Standard
2013.csv	csv	March 10, 2021, 16:13:28 (UTC+05:00)	1.8 MB	Standard
2014.csv	csv	March 10, 2021, 16:14:20 (UTC+05:00)	2.2 MB	Standard
2015.csv	csv	March 10, 2021, 16:15:23 (UTC+05:00)	2.6 MB	Standard
2016.csv	csv	March 10, 2021, 16:16:21 (UTC+05:00)	3.1 MB	Standard
2017.csv	csv	March 10, 2021, 16:17:16 (UTC+05:00)	3.5 MB	Standard
2018.csv	csv	March 10, 2021, 16:02:05 (UTC+05:00)	451.5 KB	Standard
2019.csv	csv	March 10, 2021, 16:02:51 (UTC+05:00)	903.6 KB	Standard
2020.csv	csv	March 10, 2021, 12:56:05 (UTC+05:00)	436.1 KB	Standard

- Dataset consists of following files:
  - **Elements:** date, time, temperature, feels like, humidity, dew point, wind speed, wind direction etc.
  - **Format:** .csv format.
  - **Date Range:** 2010-2020

MODELS / TECHNIQUES

Regressor Models

1) Random Forest:

A **random forest** is a meta estimator that fits a number of classifying **decision** trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Analysis:

The difference between the actual and predicted values is the very minor, so this model could be used for the out of sample predictions.

2) Decision Tree:

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

Analysis:

This model mapped the predicted values accurately and hence the difference between actual and predicted values was almost equal to zero. The model seems to be prefect to predict the out of sample values, but this may have an overtraining issue as this is perfectly mapping the training data.

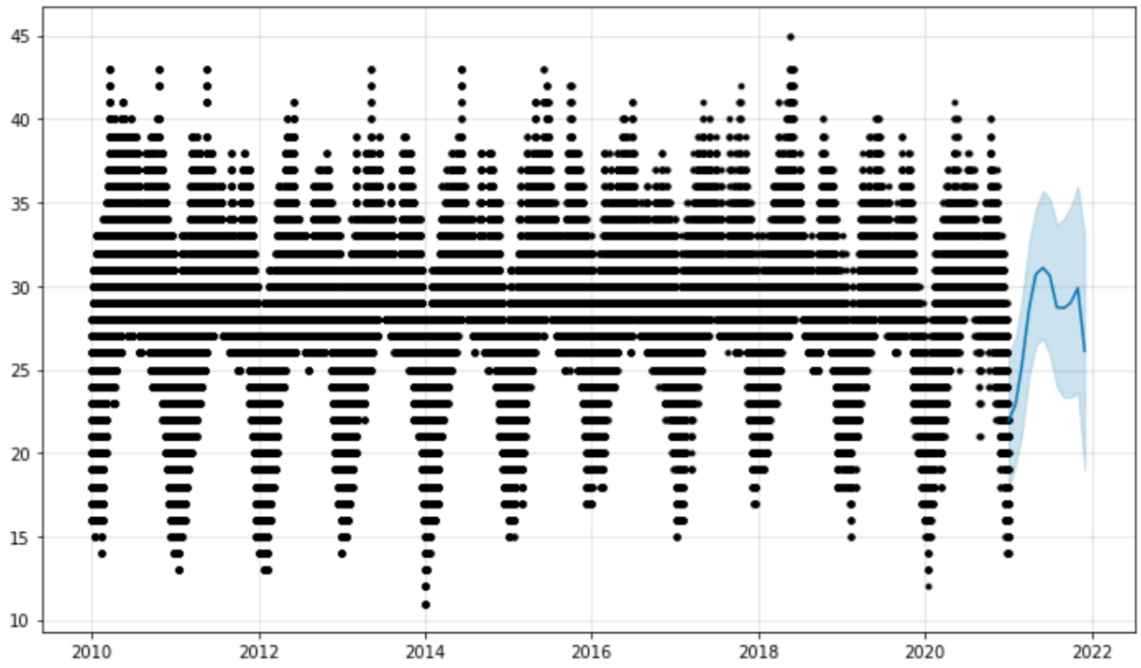
Time-Series Models

3) Prophet: Automatic Forecasting Procedure

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

Analysis:

There is an uncertainty in prediction, it is not completely accurate. Further analysis is needed to optimize the model.



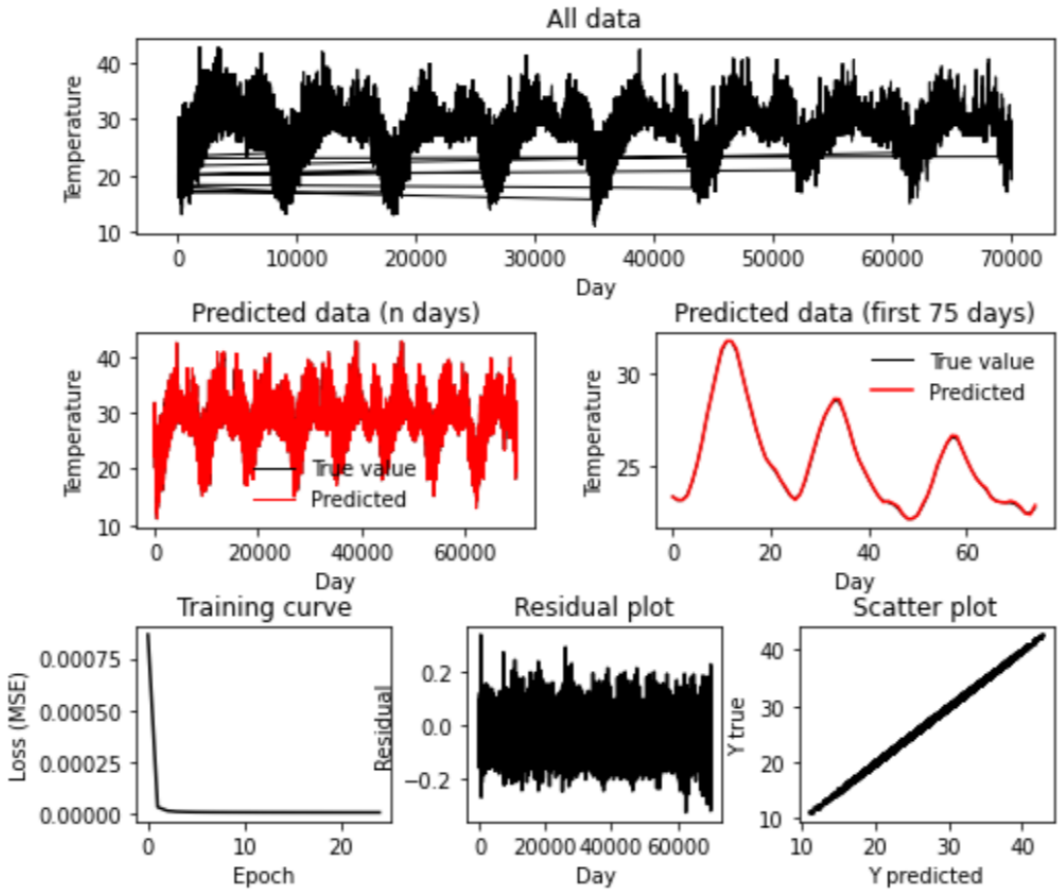
2-years prediction using Prophet

4) Long Short-Term Memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. An LSTM is generally used favorably for time-series data (such as our precipitation datasets) as it is able to 'remember' long-term dependencies/information from a feature called cell state. This cell state runs across the networks and passes down (or up) information such that earlier information from a sequence, for instance, could influence prediction of later input. This passing of information is crucial for time-series weather data where each input is dependent on other input from varying time points. The information that is stored, removed, modified, and passed around is controlled by gates.

Analysis:

An initial effort has been made to implement LSTM in Weather App. However, there is a need to further dig in and perform multivariate analysis.



For next 75-days prediction using LSTM

To perform weather forecasting in a timely and efficient manner, it is critical to understand the model and examine it in different ways to achieve highest accuracy without overfitting the model. Hence, in order to improve the predicted outcome, further analysis will be done.

API SETUP

- → **FastAPI** is a modern, fast (high-performance), web framework for building APIs with Python.
- → It is used in this project as the model is exported and wrapped around Fast API.
- → Through which, using Fast API's endpoint will be able to receive date for which prediction is required and will return output to the user.

DOCKERIZATION

- → Docker is an open platform for developing, shipping, and running applications. Dockerization enables you to separate your applications from your infrastructure so you can deliver software quickly.
- → As per the requirement, a docker image has been created in which all the necessary libraries are mentioned to be installed under the dockerfile.
- → Currently, forecast analysis that is done using Prophet is only dockerized.