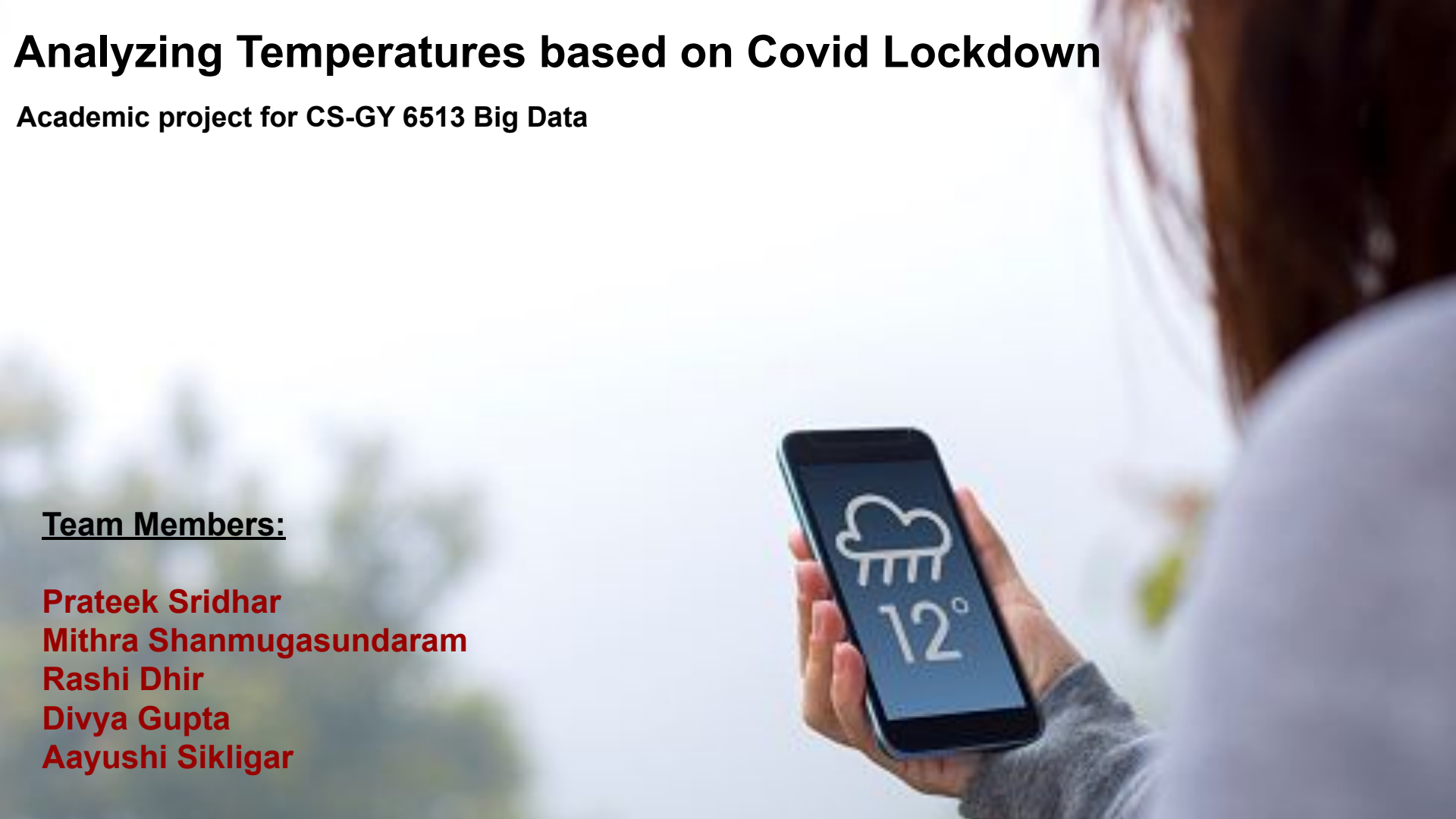
Analyzing Temperatures based on Covid Lockdown

Academic project for CS-GY 6513 Big Data



Team Members:

Prateek Sridhar
Mithra Shanmugasundaram
Rashi Dhir
Divya Gupta
Aayushi Sikligar

Problem Statement

- A very short lockdown can can trigger a chain reaction of interesting atmospheric effects and cause the everyday temperatures to plummet or rise.
- Analysis of the effect of covid lockdowns on everyday temperatures.
- We are specifically focusing on NYC’s weather data and applying the lockdown period info to get a pattern of temperature curve.
- We have years of temperature info that we are trying to work with, from 2015-2021.

Why is this a big data problem

Big Data has three main characteristics: Volume (amount of data), Velocity (speed of data in and out), Variety (range of data types and sources).

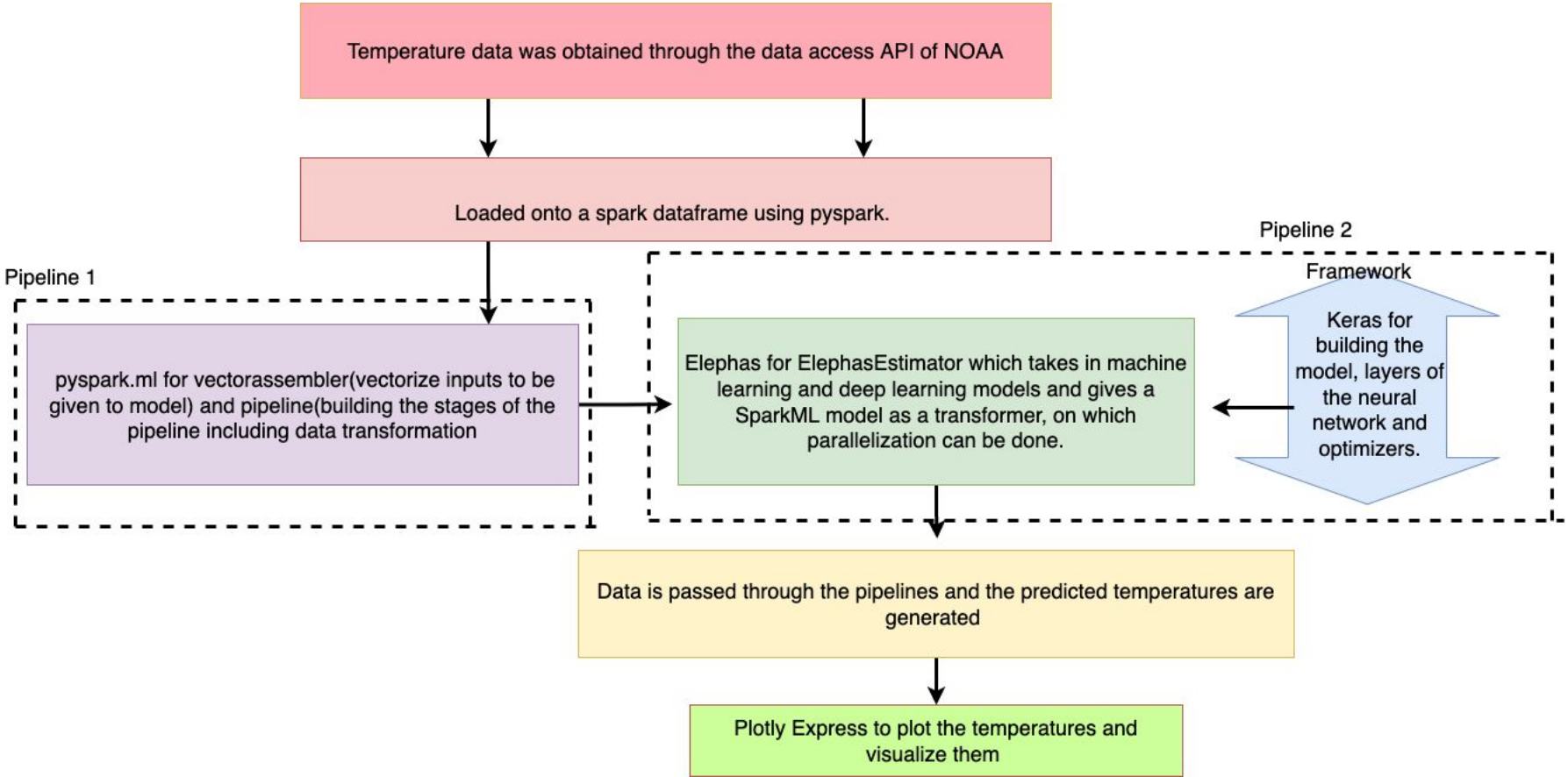
- **Volume**
- **Velocity**
- **Variety**

All the above criteria are met in our case.

Technology Stack

- **Framework/ETL Tool:** Apache Spark + Pyspark as an ETL tool to pre-process large scale data
- **Libraries:**
  - **Keras**: To build neural network models - preprocessing and loading data, defining a model, defining a loss function and optimizer and fitting the model with the data
  - **Elephas**: Elephas is an extension of keras to run distributed deep learning models at scale with spark
  - **Plotly**: To plot the results and predictions as an interactive graph

Architecture



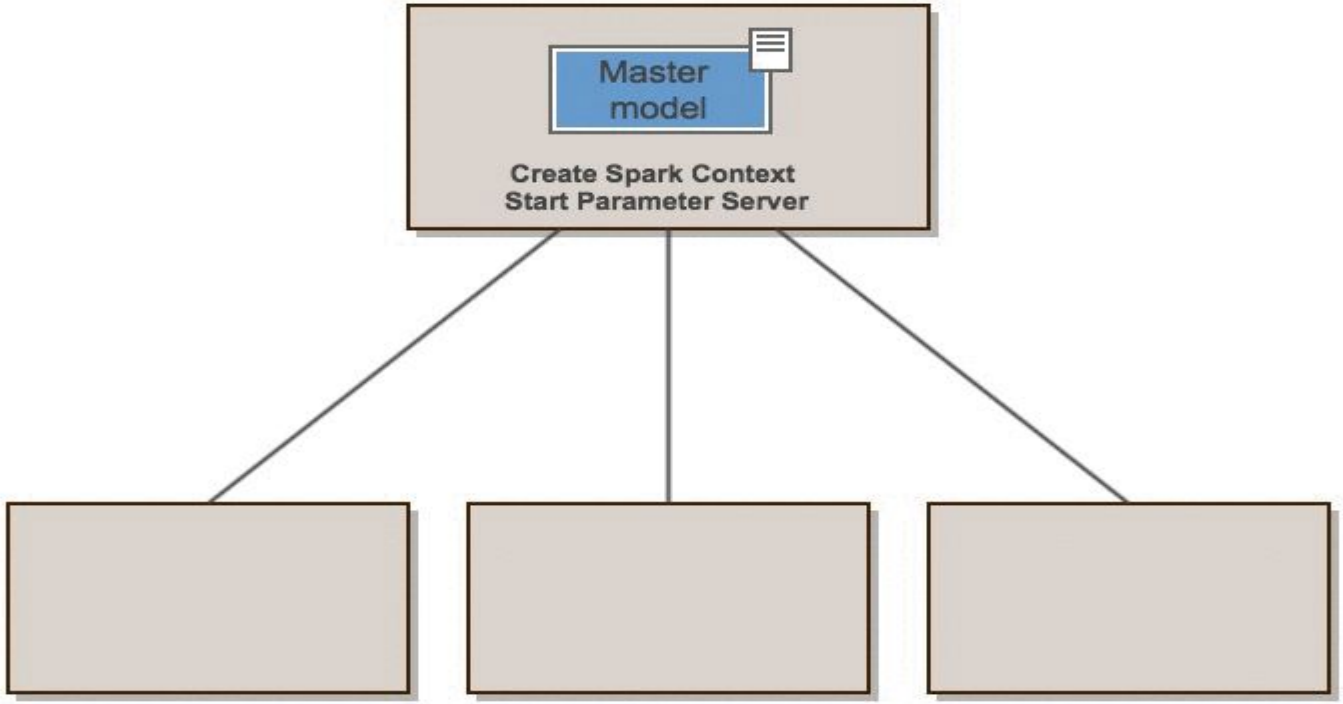
Sequence of steps in the project

- Data was acquired from https://www.ncdc.noaa.gov/cdo-web/ through its API
- The data was loaded into Apache spark, pre-processed and converted it to the required format, then loading it into a spark dataframe.
- Compile information from news articles to obtain lockdown start and end dates of different regions
- **Pipeline 1:** Created the Spark Data Pipeline(Transforming and vectorizing the data per feature; so that it is ready for modelling)
    - Identified the features and vectorized using VectorAssembler.
    - Assembled all our features into a single vector.
- Ran the data through the Spark Pipeline.

Sequence of steps in the project

- **Pipeline 2:**
- Built the Deep Learning Model with Keras
- Fed the Keras model through Elephas to create a distributed Deep Learning Architecture, to enable parallelization of model training.
- Visualizing the results by collecting the predictions and plotting using Plotly and Plotly Express.

Why did we choose this tech stack+architecture?



Data Loading and Pre-processing

- We are using a very reliable source for the temperature data: NOAA(National Oceanic and Atmospheric Administration)
- Using an access token and by specifying a station, we were able to build a simple API to gather data from ncdc.noaa.gov
- Few parameters that we had to specify to collect data from this API are as follows:
    - Datasetid
    - Datatypeid
    - Limit
    - Station ID
    - Start and end date

Data Loading and Pre-processing

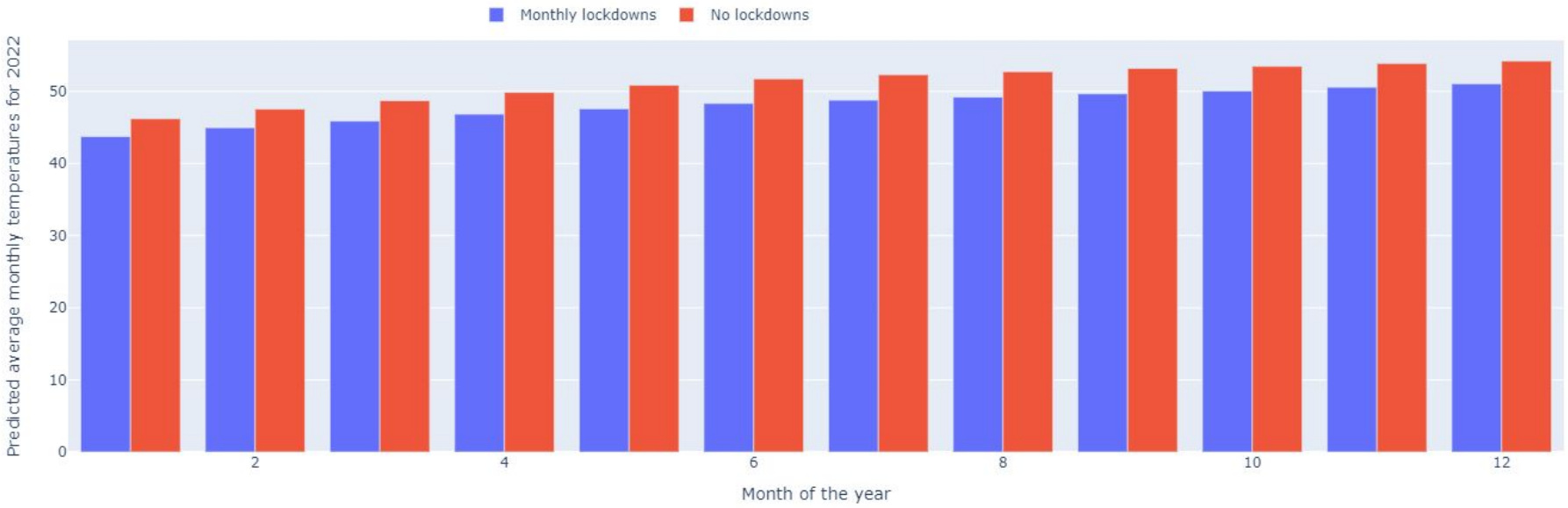
- To make features more compatible with the machine learning models, we transformed the data obtained from the API, by splitting the date into features such as month, year and date.
- To suit our requirements we had to compile information from news articles to obtain lockdown start and end dates of different regions and add them as features to the dataframe obtained from the API.

Deploying the ML Pipeline

- We built a sequential Neural Network model using Keras and Deep Learning to implement regression to predict temperature based on features(year, month, date and lockdown information)
- We initialized a spark ML Estimator using Elephas and set parameters to facilitate parallel training of the model using the spark instance.
- We created a deep learning pipeline (Pipeline 2) from the Elephas estimator which creates a transformer, to which we fed the transformed data from Pipeline 1.
- We used Pipeline 2 to train the model and obtain prediction results.

Prediction and Analytical Insights

Based on our model, we can clearly see that temperatures are lowered with a lockdown. The temperature for each monthly lockdown was calculated such that we considered all previous months to not have a lockdown.



Conclusion and future scope

- We successfully analyzed the temperature variation during the pandemic using pipelining approach in ML to generate a predictive model.
- The scope of the project can be expanded by incorporating more parameters into this project like Air Quality information, precipitation information to get a detailed analysis of the overall effects of the lockdown on the atmosphere
- From this project, we learned how to deal with a real world problem in the Big Data domain using the tools and techniques we learned during the course of this class.
- Finally, we understood how to implement a real world based Big Data problem and using ML and PySpark, we accomplished to obtain efficient results. .