TIME SERIES FORECAST OF SYDNEY WEATHER (TEMPERATURE)

ADVANCED DATA SCIENCE WITH IBM – CAPSTONE PROJECT

TONY THEKKEDATH

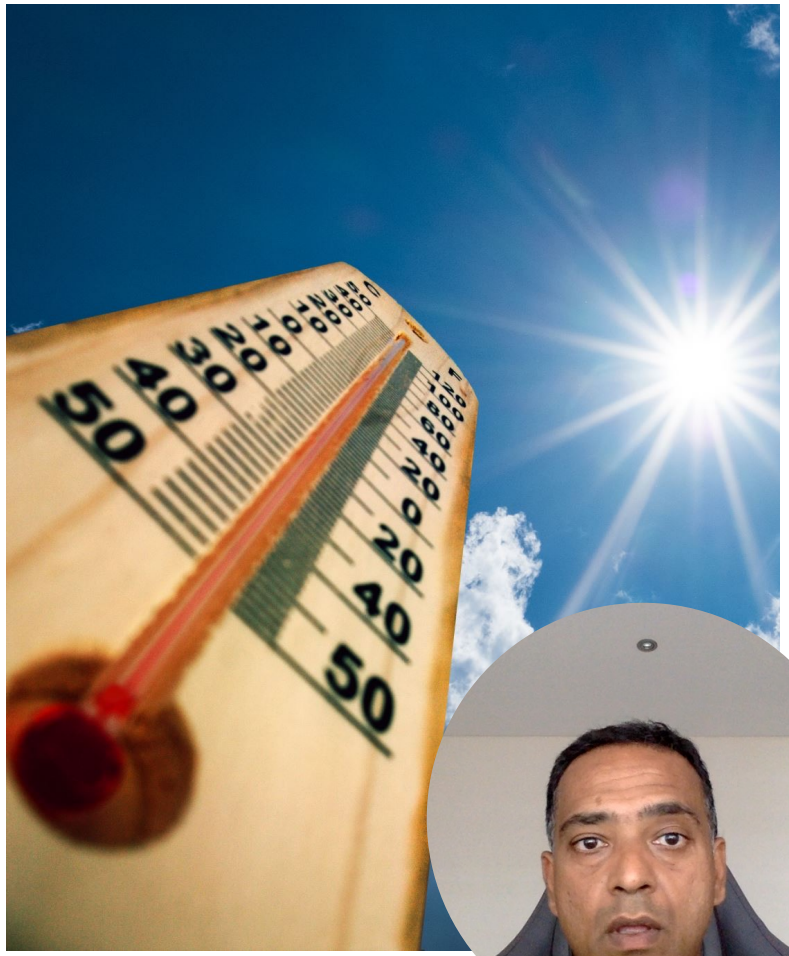
7 FEB 2024

AGENDA

- Data set
- Use Case
- Architectural choices
- ETL
- EDA
- Architectural choices
- Feature Engineering
- Modelling and Performance
- Stakeholder Deliverable & Future work

USE CASE

- Demonstration of whether machine learning can be applied to climate indicators – while the data is monthly – potential to explore whether this can be extended to forecast on a weekly, daily basis
- Target variable is maximum temperature – possible to extend this to daily, weekly temperature. As well extend this to other weather conditions such as rainfall, solar exposure – **all of which may have applicability in navigation, agriculture, travel advisory, energy demand predictions**. While climate forecasting methods are already established, machine learning can be another approach to provide confirmation for the forecasts, in addition machine learning has an advantage in being able to improve prediction accuracy using features – for example improve energy demand prediction utilizing temperature, seasonal demand, time of year (month feature) etc to provide more more accurate prediction.

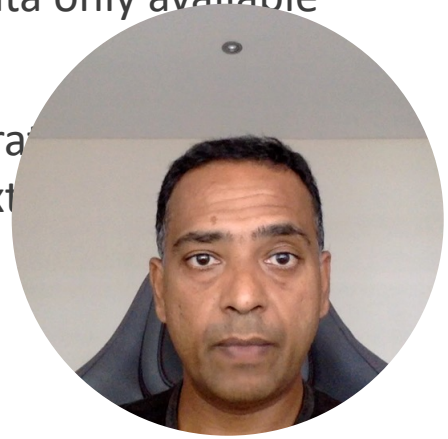


DATA SET

Sydney Airport TEMP

Product code	Bureau of Meteorology station number	Year	Month	Mean maximum temperature (°C)	Quality
IDCJAC0002	66037	1939	4	22.6	Y
IDCJAC0002	66037	1939	5	20.6	Y
IDCJAC0002	66037	1939	6	17.5	Y
IDCJAC0002	66037	1939	7	15.3	Y
IDCJAC0002	66037	1939	8	18.5	Y
IDCJAC0002	66037	1939	9	20.0	Y

- Data Source is a CSV extract – this extract is originally sourced from the http://www.bom.gov.au/climate/data/?ref=ftr
- Sydney Airport – Station number 066037 – historical temperature (Mean – Maximum temperature).
- Data Range from 1939 onwards. Data only available on a monthly basis
- Target variable is Maximum tempera
  location to be predicted for the next



ARCHITECTURAL CHOICES

Model is not resource / compute intensive due to low amount of data used for training and forecasting

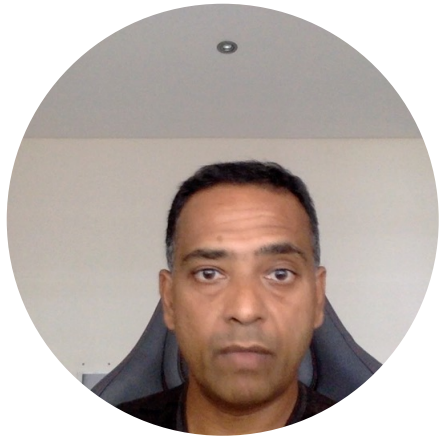
Extracts data from a website which is locally downloaded before training on Jupyter notebooks. The output is a single prediction plot showing predicted values for the next month.

Language – Python

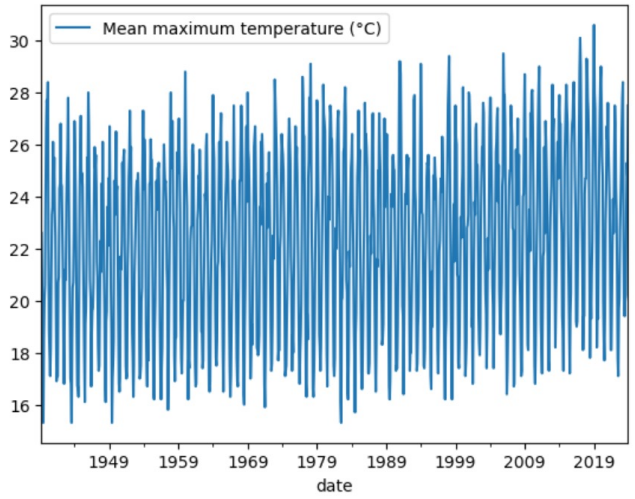
Machine learning

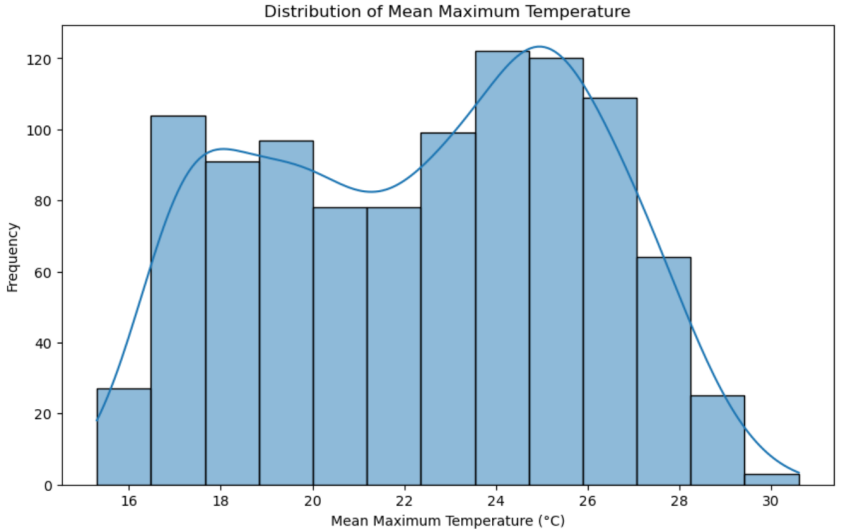
XGBoost

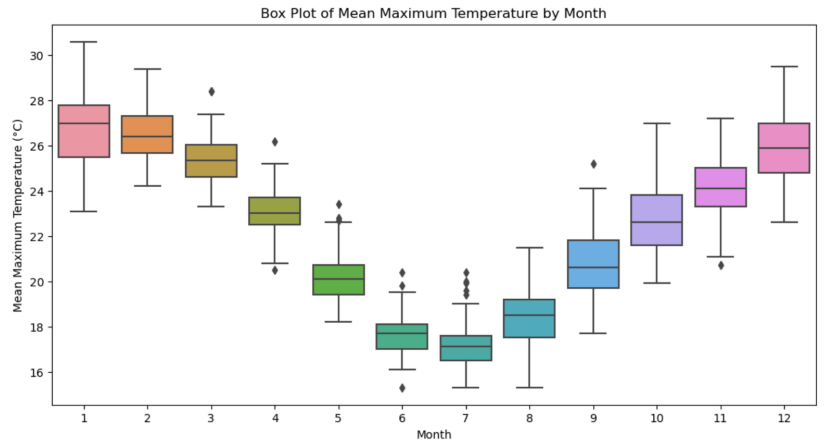
LSTM



EXPLORATORY DATA ANALYSIS







Summary Statistics:

	Bureau of Meteorology station number	Year	Month
count	1017.0	1017.000000	1017.000000
mean	66037.0	1981.123894	6.513274
min	66037.0	1939.000000	1.000000
25%	66037.0	1960.000000	4.000000
50%	66037.0	1981.000000	7.000000
75%	66037.0	2002.000000	10.000000
max	66037.0	2023.000000	12.000000
std	0.0	24.477460	3.449876

	Mean maximum temperature (°C)	date
count	1017.000000	1017
mean	22.394592	1981-07-31 23:38:45.663716800
min	15.300000	1939-04-01 00:00:00
25%	19.300000	1960-06-01 00:00:00
50%	22.700000	1981-08-01 00:00:00
75%	25.300000	2002-10-01 00:00:00
max	30.600000	2023-12-01 00:00:00
std	3.562814	NaN

Data Types:

Product code	object
Bureau of Meteorology station number	int64
Year	int64
Month	int64
Mean maximum temperature (°C)	float64
Quality	object
date	datetime64[ns]

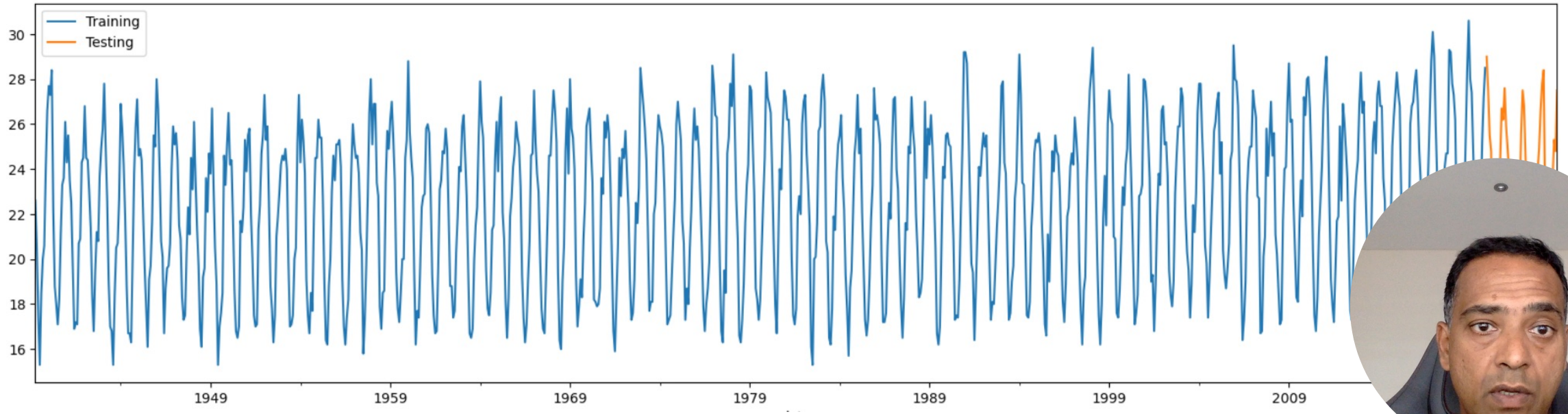
dtype: object

Discussion

- Data shows high seasonality – month [illegible] could be useful features to enab[illegible]
- Data format and structure is [illegible] no need to clean data furth[illegible] values, handle missing valu[illegible] data

DATASET – TRAINING VS TEST DATA SETS

- Data available from 1939 (monthly)
  - Training Data – all data till end 2019
  - Test Data – all data from 2020 onwards



FEATURE ENGINEERING

- Features used initially
  - Seasonality
    - Quarter
    - Month
    - Day
  - To further optimize model performance
    - Lagging – by 1 month, 3 months and 12 months

	Product code	Bureau of Meteorology station number	Year	Month	Mean maximum temperature (°C)	Quality	date	month	quarter	day_of_week	...	lag_3	lag_4	lag_5	lag_6	lag_7	lag_8	lag_9	lag_10	lag_11
12	IDCJAC0002	66037	1940	4	22.2	Y	1940-04-01	4	2	0	...	27.7	26.6	24.0	20.6	20.0				20.6
13	IDCJAC0002	66037	1940	5	18.8	Y	1940-05-01	5	2	2	...	27.3	27.7	26.6	24.0	20				
14	IDCJAC0002	66037	1940	6	18.0	Y	1940-06-01	6	2	5	...	28.4	27.3	27.7	26.6					
15	IDCJAC0002	66037	1940	7	17.1	Y	1940-07-01	7	3	0	...	22.2	28.4	27.3	27					
16	IDCJAC0002	66037	1940	8	18.2	Y	1940-08-01	8	3	3	...	18.8	22.2	28.4	2					

MODELLING

- XGBoost and LSTM were used to predict the temperature
  - XGBoost is an open-source gradient boosting algorithm that performs well in a variety of machine learning tasks
  - Long Short-Term Memory (LSTM) is a specialized recurrent neural network architecture designed for sequential data processing, featuring memory cells and gating mechanisms to effectively capture and retain long-term dependencies.
  - Model accuracy is below

Model	PARAMETERS	FEATURES	MAE	MSE	R2
XGBoost	Grid Search Optimisation, Max depth - 10	Date, Year, Month	1.34	2.48	0.77
LSTM	LSTM layer-100, EPOCH-100	Date, Year, Month	2.7	10.4	0.08
LSTM	LSTM layer-200, EPOCH-200	Date, Year, Month	1.1	1.7	0.84

MODELLING OPTIMISATION

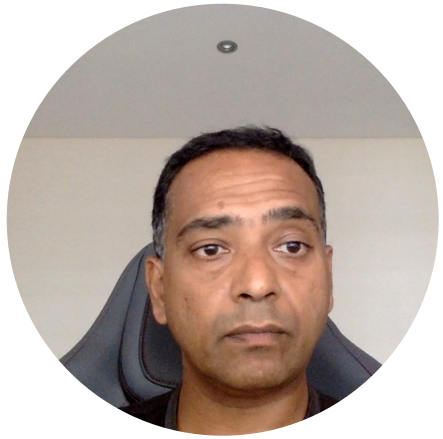
- Models were further optimized by varying model parameters and adding additional features such as Lags or parameters such as LSTM layers, EPOCHs
  - Model accuracy is below

Model	PARAMETERS	FEATURES	MAE	MSE	R2
XGBOOST	Grid Search Optimisation, Max depth - 10	Date, Year, Month, Lag 12	1.34	2.58	0.77
XGBOOST	Grid Search Optimisation, Max depth - 10	Date, Year, Month, Lag 1	1.37	2.59	0.76
XGBOOST	Grid Search Optimisation, Max depth - 15		1.4	2.73	0.75
LSTM	LSTM layer-300, EPOCH-300	Date, Year, Month,	1.0	1.6	0.75
LSTM	LSTM layer-300, EPOCH-300	Date, Year, Month, Lag 12	1.0	1.6	25
LSTM	LSTM layer-300, EPOCH-300	Date, Year, Month, Lag 1	1.1		
LSTM	LSTM layer-500, EPOCH-500	Date, Year, Month, Lag 12	0.99		

- Changing parameters in the model further resulted in degraded R2 scores, eg. Improving LTSM layers, EPOCH
- Lags did not improve prediction accuracy and in XGBoost case resulted slightly worse R2 scores

MODELLING - DISCUSSION

- Further optimization in the algorithms
  - Adding lags did not improve the performance of the XGBoost algorithm but increased the performance of the LSTM layer
  - LSTM: Performance did improve from increasing the LSTM layers to 300 and EPOCH stages to 300, higher values above this did not yield material improvements and slight degradation in some cases
  - XGBoost: Performance was flat varying the parameters in the XGBoost model
    - Increasing batch sizes, adding more lags information did not improve performance; increasing Depth contributed mostly to the performance – increasing above 15 provided R2 score improvement of ~0.01 and tapering off thereafter
  - Other optimizations' that could yield improvements but could not be addressed due to time constraints
    - Features
      - Adding other weather data such as rainfall, solar exposure
      - Adding data from adjacent sites
    - Other algorithms



STAKEHOLDER DELIVERABLE

- Prediction plot available to stakeholders / the output of the model