Time Series Forecasting: Predicting the temperature in La Paz, Mexico

Jon Reid

https://github.com/JonR45

Overview

Data collection

Exploratory data analysis

Data preparation

Modelling

Evaluation

Next steps

Overview

Time-series model for predicting the average daily temperature in La Paz, Mexico

'Business' Problem

- The people of La Paz don't know what the temperature will be months in advance

Impact

- Agricultural, energy management, forecasting storms, tourism, outdoor events

Aim

- Create a model that can accurately predict the average daily temperature for the next 365 days

Limitations

- Temperature is highly complex
- Model only takes one feature into account
- El Niño La Niña

Type of Data Science problem

- Time-series forecasting with historical data
- Supervised learning

Process

- Collected daily average temperature data from 2012-2022
- Created and compared several models
- Used XGBoost to turn traditional time series problem into a supervised learning problem

Results

- Best model: XGBoost with walk-forward validation
- RMSE 1.12°C

Data Collection

Source

https://www.visualcrossing.com/weather/weather-data-services

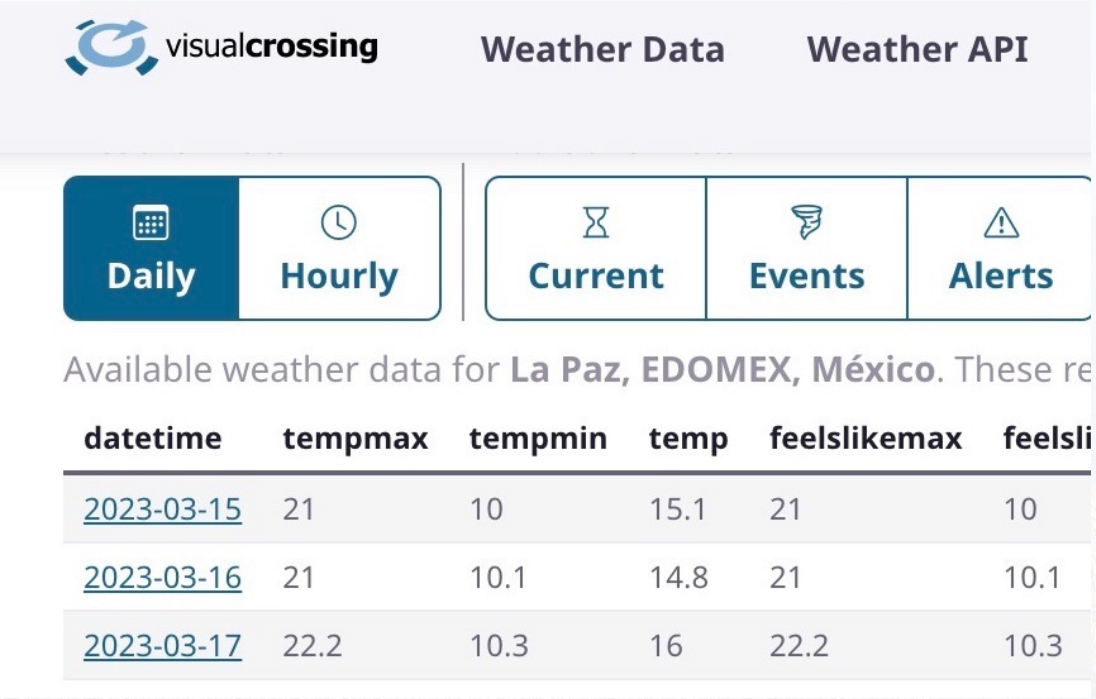


Fig.: Website where the data was available either to download or access via the API

	A	B	C	D	E	
1	name	datetime	tempmax	tempmin	temp	feelsl
2	la paz mexicc	01/12/2010	28	4.1	13.7	
3	la paz mexicc	02/12/2010	25.2	4.1	14	
4	la paz mexicc	03/12/2010	23.5	-0.9	12.1	
5	la paz mexicc	04/12/2010	24.3	-1.9	11.4	
6	la paz mexicc	05/12/2010	24.2	-1.9	10.9	
7	la paz mexicc	06/12/2010	20.7	-1.9	9.3	
8	la paz mexicc	07/12/2010	20.8	-2	10.1	
9	la paz mexicc	08/12/2010	20.5	-1.9	10.5	
10	la paz mexicc	09/12/2010	20.5	-5.9	8.9	
11	la paz mexicc	10/12/2010	24.1	-4.9	9.8	
12	la paz mexicc	11/12/2010	22.6	3.1	11.8	
13	la paz mexicc	12/12/2010	21.2	-4.9	11.1	
14	la paz mexicc	13/12/2010	21	-2.9	9.4	
15	la paz mexicc	14/12/2010	20.4	-1.9	10.4	

Fig.: Example of csv file

[7]:

	datetime	temp
0	2014-12-01	16.0
1	2014-12-02	14.1
2	2014-12-03	13.3
3	2014-12-04	13.2
4	2014-12-05	15.2

Fig.: Several csv files merged into one dataframe with all the data

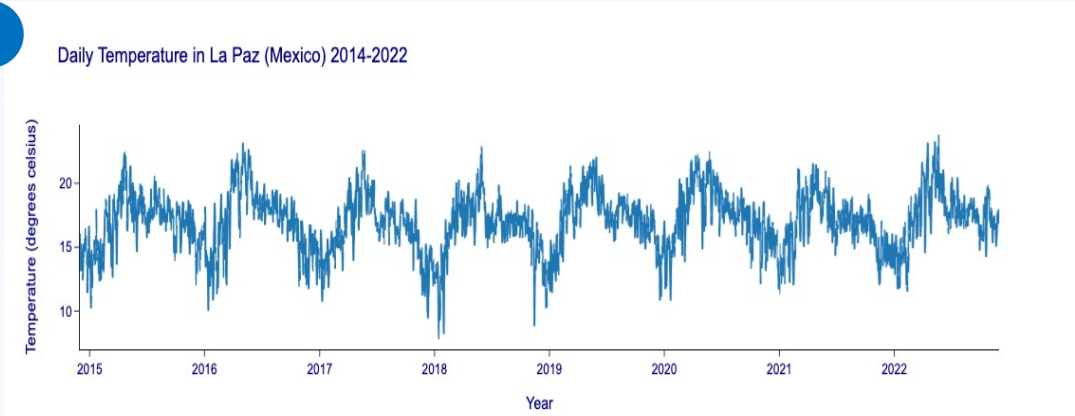
Exploratory Data Analysis

Date range	2014-2022
No. of. observations	2,922
Mean (avg. daily temp.)	17.0°C
Median (avg. daily temp.)	17.1°C
Max (avg. daily temp.)	23.7°C
Min (avg. daily temp.)	7.9°C

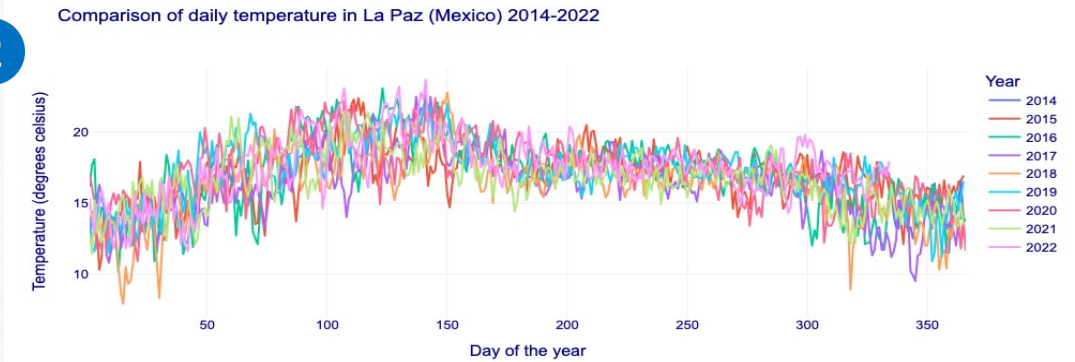
Insight

- **Graph 1** showed there is a repeating trend year on year
- **Graph 2** showed the highest average daily temperatures occur around May
- **Graph 3** drew attention to many outliers in 2018, 2019 and 2022 (both highs and lows).
  - Further investigation revealed the occurrence of el niño/la niña in 2018-2019 (an irregular complex series of hot/cold climatic changes)

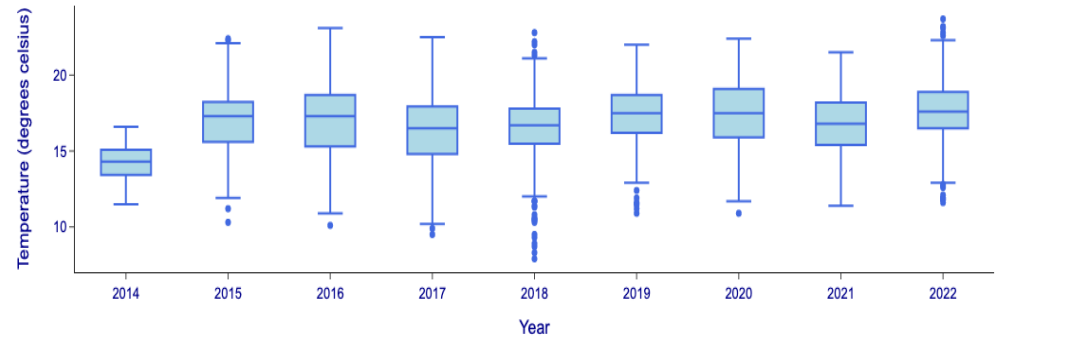
1



2



3



Data Preparation

Baseline, Prophet, DARTS Prophet & XGBModel

- Datetime data type as index, 'ds' and 'y'
- Train-test split: 87.5% training (1 year = 12.5%)

[7]:

	datetime	temp
0	2014-12-01	16.0
1	2014-12-02	14.1
2	2014-12-03	13.3
3	2014-12-04	13.2
4	2014-12-05	15.2

Fig.: Dataframe prepared for Baseline, Prophet and DARTS models

XGBoost – Supervised learning

- Series to supervised, XGBoost forecast and walk forward validation functions
- Train-test split: 87.5% training (365 days to test on)
- Walk forward validation with 7 inputs – avoid data leakage

	var1(t-7)	var1(t-6)	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
0	16.0	14.1	13.3	13.2	15.2	14.8	13.4	13.2
1	14.1	13.3	13.2	15.2	14.8	13.4	13.2	13.3
2	13.3	13.2	15.2	14.8	13.4	13.2	13.3	12.5
3	13.2	15.2	14.8	13.4	13.2	13.3	12.5	13.5
4	15.2	14.8	13.4	13.2	13.3	12.5	13.5	14.0
...	...	...	...	...	...	...	...	...
2910	16.2	15.1	15.4	15.9	16.5	17.5	17.6	17.8
2911	15.1	15.4	15.9	16.5	17.5	17.6	17.8	17.0
2912	15.4	15.9	16.5	17.5	17.6	17.8	17.0	17.5
2913	15.9	16.5	17.5	17.6	17.8	17.0	17.5	16.9
2914	16.5	17.5	17.6	17.8	17.0	17.5	16.9	17.9

Fig.: Dataframe prepared for XGBoost supervised learning with walk-forward validation

Modelling: Model Selection

Baseline

- NaiveSeasonal model using the value from the previous year's equivalent date as a prediction

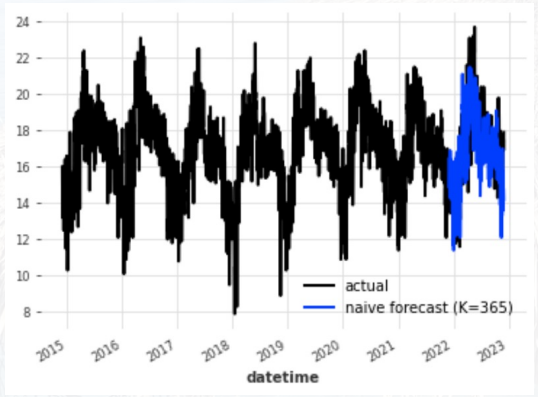


Fig.: NaiveSeasonal model

FB Prophet

- Additive regression model that works best with time series data that has strong seasonal effects
- Robust to outliers, missing data, and noisy data
- Easy to use but also tuneable

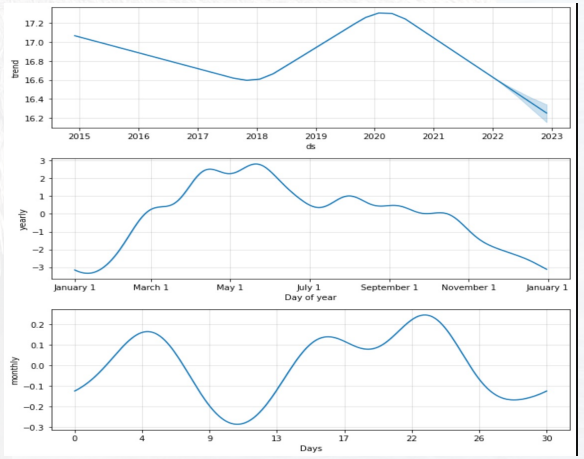


Fig.: FB Prophet trend, yearly, monthly

DARTS XGBModel

- DARTS' built in XGBModel
- Has different hyperparameters than XGBoost
- Create a TimeSeries object from dataframe

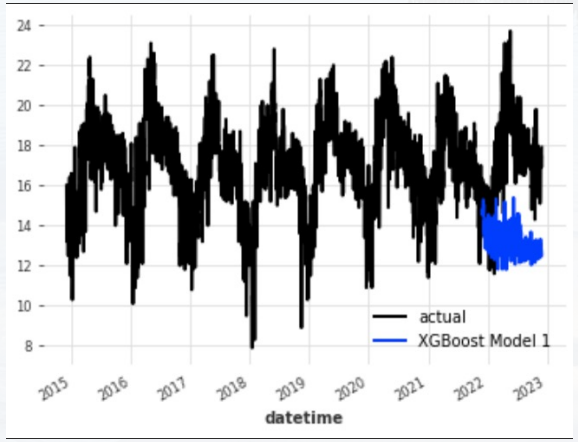


Fig.: DARTS XGBModel, lags=7

XGBRegressor

- Used functions to turn the time series problem into a supervised learning problem and fit an XGBRegressor model
- Avoid data leakage
- Ever expanding training dataset (expanding window)

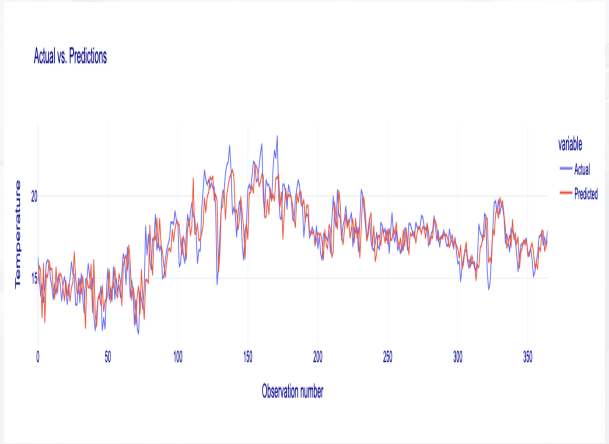


Fig.: XGBRegressor, supervised learning

Modelling: FB Prophet

Prophet Model 1

- Train-test split of 87.5% - 12.5%
- Fourier order = 5
- Simply fit on training data and tested on testing set

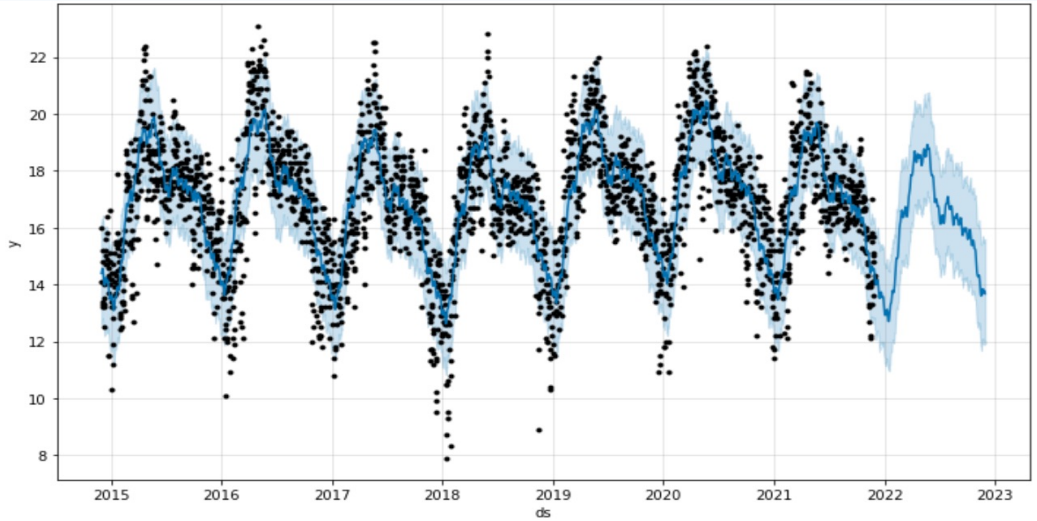


Fig.: Prophet Model 1

Prophet Model 2 (Tuned)

- Hyperparameters tuned: changepoint_prior_scale', 'seasonality_prior_scale'
- Prophet **built in cross validation**: define the model, initial training period, horizon, size of next training window, and the number of folds.
- Hyperparameter search took ~40 mins
- **best_params** = 'changepoint_prior_scale': 0.01, 'seasonality_prior_scale': 0.1

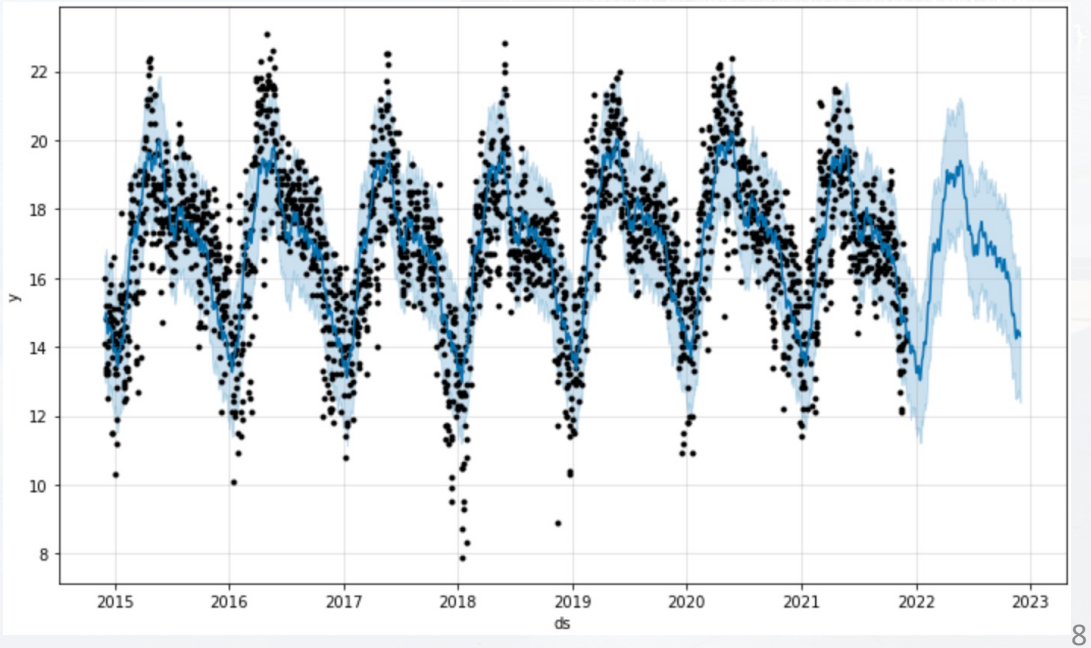


Fig.: Prophet Model 2 - Tuned

Modelling: DARTS XGBModel

- Used built in TimeSeries to split data into train and test sets
- lags=7, n_estimators=100, max_depth=6.
- Only 'tuned' no. of lags: lags = 120 returned RMSE -2.16°C vs. lags =7)

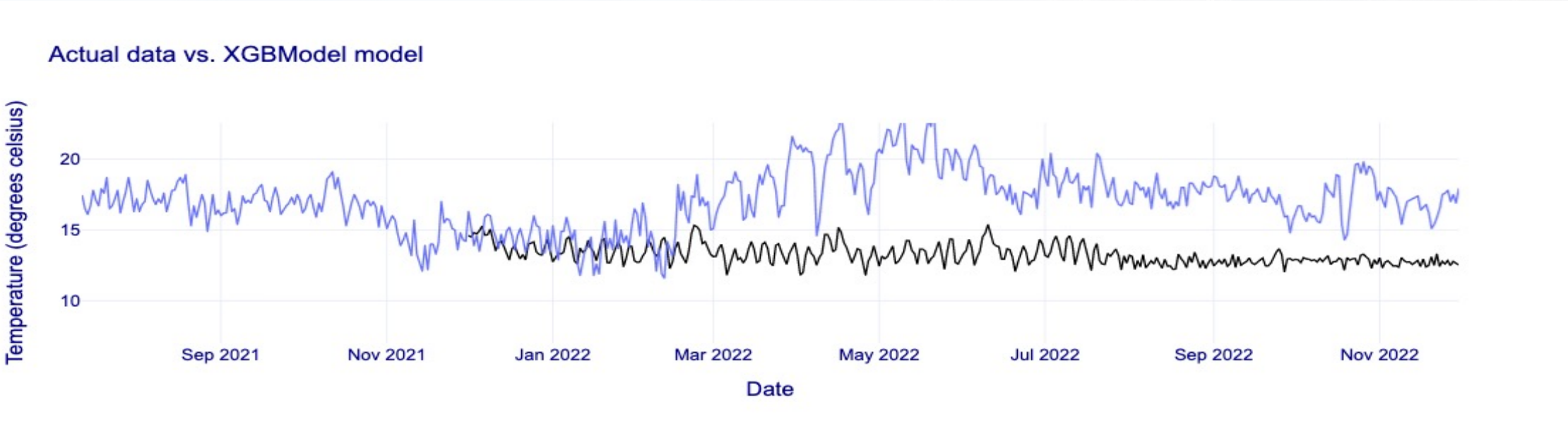


Fig.1: DARTS XGBModel, lags=7, did not perform well

Modelling: XGBRegressor

- Walk forward validation
- Untuned: n_estimators = 1,000
- Optuna: define objective, create study session
- Best_params = 'n_estimators': 1900, 'max_depth': 12, 'learning_rate': 0.2532767133910654

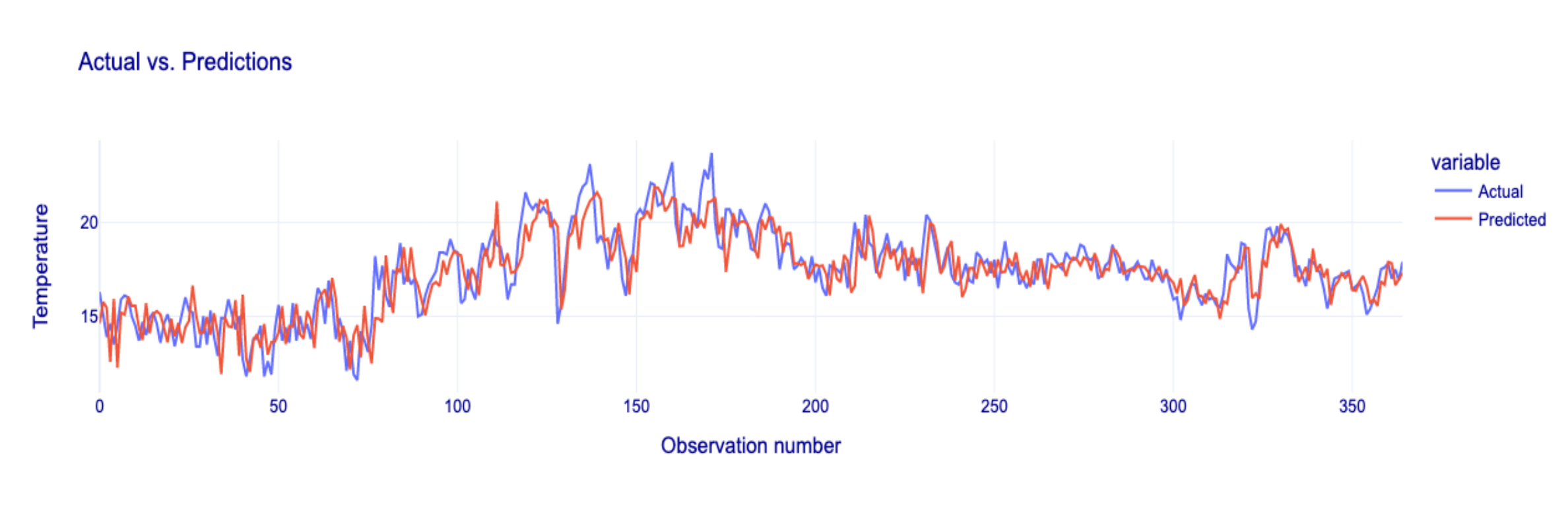


Fig.: XGBRegressor actuals vs. predictions

Modelling: Results

Model	RMSE
NaiveSeasonal	2.17°C
Prophet Basic	2.0°C
Prophet Tuned	1.68°C
DARTS XGBModel Basic	4.74°C
DARTS XGBModel lags=120	1.54°C
XGBRegressor Basic	1.25°C
XGBRegressor Tuned	**1.18°C**

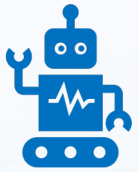
NaiveSeasonal was ~1°C less accurate than the best model

Tuning the lags of DARTS XGBModel made a big difference

XGBRegressor using walk-forward validation was the best performing model

Next Steps

Explore hyperparameter tuning for prophet, DARTS XGBModel, and XGBRegressor



Explore other models

Productionise – develop code to create a pipeline that connects to API, cleans data and outputs to model.

How would I change the model given knowledge of El Niño?

Work with hourly data, predict the hottest hour? Predict maximum temperature on a given day?