EES mini Project

Xiaoning Nie. (cshb26)

April 14, 2022

1 Introduction

1.1 Description of the project

In this project, some weather data from 1901 to 2019 were provided to predict the average daily temperature in 2020.

There are many ways to complete this project, but here I have chosen to use `prophet`, an open source Python package from Facebook, which is a temporal prediction package that helps us to deal with time-related problems easily.

1.2 Description of the dataset

Analysis of the dataset shows that it has a total of 43,464 sample data, with 8 features per sample and no missing values. The eight of these features are:

- `Year`, `Month`, `Day` and `Date` describe the specific date of the temperature data;
- `PPT.` describes the daily rainfall;
- `Tmax` describes the daily maximum temperature;
- `Tmin` describes the daily minimum temperature;
- `Av temp` describes the daily average temperature, which is also the target.

2 Processing & Results

2.1 Data preprocessing

According to the question, we need to predict the daily average temperature in 2020, so feature `Date`, `Tmax` and `Tmin` from previous years are not important to us, thus, they are discarded.

The relationship between daily rainfall and average daily temperature over the previous 10 years shows that they are not relevant, so the `PPT.` feature is deleted directly.

2.2 Prediction & results

In this project, forecast is made using the Prophet model from the prophet package. Visualisation of the predicted data gives all the average daily temperature data from 1901 to 2020 as shown in Figure 1. The average daily data for 2020 required by the project is shown in Figure 2.

At the same time, the data obtained can be further extracted to obtain the trend of the average annual temperature as shown in Figure 3. From the graph we can easily see that the

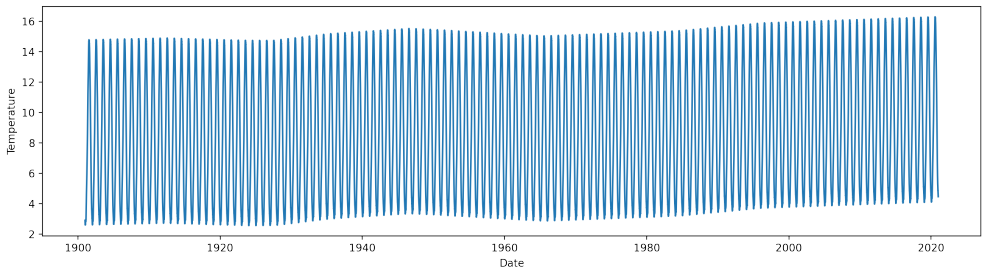

Figure 1: Temperature of 1901-2020

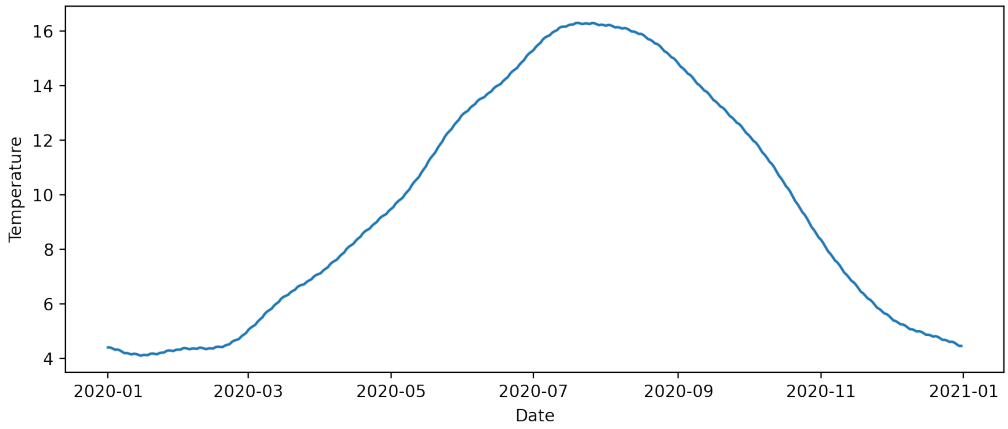

Figure 2: Predicted temperature of 2020

average annual temperature is gradually increasing over the last 120 years, which reflects the current global warming problem.

3 Discussion

3.1 Why I chose prophet

prophet is an open source time series data prediction tool developed by facebook, which is based on time and variable values combined with time series decomposition and machine learning. Its powerful prediction ability can solve most of the practical scenarios of the prediction of single values.

The advantage of prophet as a time series tool over advanced machine learning algorithms is that it does not require a lot of feature engineering to obtain the future trends, seasonal factors and holiday factors. Also, prophet is suitable for data with clear intrinsic patterns and is able to predict future movements of time series almost fully automatically with high efficiency, which is very appropriate for the problem of predicting temperatures in this project.

For the convenience of statisticians and machine learning learners, prophet provides interfaces to both the R and Python languages. Meanwhile, prophet follows the application programming interface of the sklearn library. That is, we can create an instance of the Prophet class and then use `fit` to fit the model and use `predict` to perform prediction calculations. The prophet library itself also comes with parameters relating to seasonality and holidays, allowing for more accurate

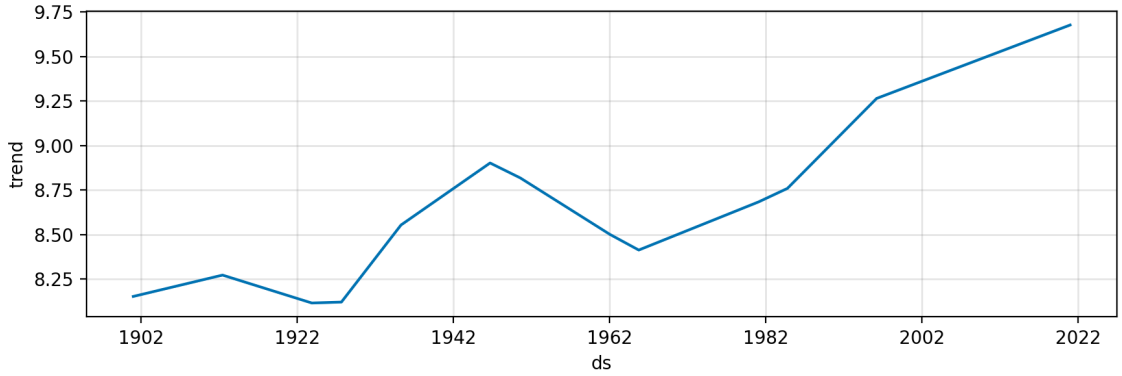

Figure 3: Trend of average annual temperature

forecasting by season or separating holidays from weekdays, making it more user-friendly.

3.2 What are the limitations of prophet

Throughout the project, I found several prophet shortcomings:

1. The iterative computing was slow and the project took more than 20 minutes to run on my PC.

2. High requirements for data quality. Prophet will be somewhat delayed to the emergence of new turning points so it requires a large accumulation of data for forecasting, and generally these data volumes often take more than one year to accumulate so that they are suitable for long-term forecasting.

3. Prophet is not very suitable for time series that are not very cyclical or trendy. Emerging activities cannot be predicted if they are not supported by previous data.

4. It cannot make use of more information, such as information about the merchandise, the shops, the promotions, etc., when forecasting the sales of the products.

3.3 What could I do differently with more time and resources

LSTM

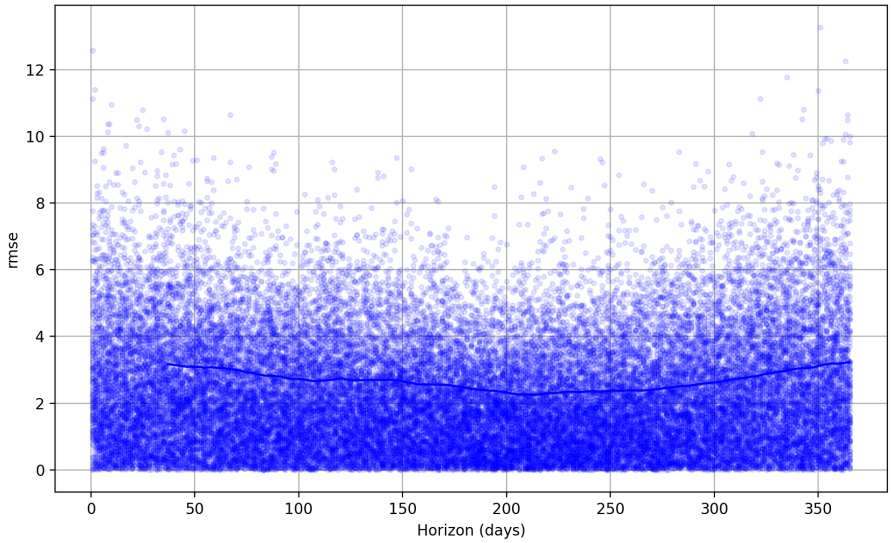

Figure 4: RMSE of the prediction