Weather prediction model learning based on synthetic weather dataset

Mihhail Batura

1. Business understanding

Background

Canadian businesses, facing ever-higher energy prices, are suffering losses from winter heating. Initially, the problem was thought to be a gas leak, but inspection and partial repairs failed to resolve the issue.

Because winters in southern Canada are unstable, alternating between cold and warm, some businesses have turned to meteorological services for more accurate forecasts to enable timely heating system modeling. Many of them lack automated heating systems, believing it's cheaper to rely on a reliable weather forecasting source. However, these businesses losses subsequently only increased. They then turn to a company that can develop a weather forecasting model that can also identify weather patterns.

Business goals and criterias

The goal of this project is to train a model to predict North American weather based on typical weather patterns for each season. The model development itself is a pilot project, so it allows for the use of synthetic data for testing. However, this data must cover at least the period from October to April. Since meteorological processes are often temporally related, data from as early as late summer, such as August, may be needed. Therefore, a year-long analysis is preferable.

Requirements

This model is necessary to reduce heating costs in winter. Ideally, this model would provide accurate hourly weather forecasts for up to 90% of all hours in a month. This would reduce heating costs during transitional periods by 10% per year. A 95% accurate weather forecast would reduce heating costs for the same period by up to 15%, etc. The primary parameters for evaluation will be temperature and humidity, with other parameters being secondary. Humidity as an indicator is especially

important for companies storing goods in warehouses. For temperature, the evaluation criterion will be RMSE < 1.5°C, and for humidity, RMSE < 5%.

Budget

The project's budget limits the ability to pay for access to paid data, so the project is limited to using open data or APIs. The project is also limited to using only free software for coding. The budget only covers project participants' work hours and other expenses associated with joint activities with businesses.

Data mining goals and criteria

The project requires defining training and test data. It's also important to understand which locations and meteorological parameters comprise the training data, and to base the test data preparation on these. These should definitely be cities in North America above 30°N latitude. The data should include temperature in Celsius, relative humidity, and other key weather forecasting parameters (wind speed, air pressure, etc). The dataset should be hourly, as the facilities are heated 24/7.

2. Data understanding

Data requirements

The training data should cover the entire year and have no significant gaps. If there are any gaps, these rows will need to be removed. The data should be hourly and date-time-based. Ideally, air temperature should be in degrees Celsius, and pressure in millimeters of mercury. The test data should also cover the entire year. If the test data has a different format and column names, they must be adjusted to match the training data. The training data can be either synthetic or real, but the test data must be real.

Describing data

The training weather data was obtained from the Kaggle environment. The data is synthetic and is intended as the basis for weather forecasting models, climate research, and educational projects. The dataset contains parameters or columns such as location, date_time, temperature_C, humidity_ptc, precipitation_mm, and

wind_speed_km/h. The sample includes 10 US cities: Chicago, Dallas, Houston, Los Angeles, New York, Phoenix, Philadelphia, San Antonio, San Diego, and San Jose. Each city has approximately 100,000 generated observations from January to May. The total data volume is 1,000,000 rows and 6 columns.

Three cities were selected for further analysis. Two cities were selected near southern Canada — Chicago and New York. Another city (Dallas) was chosen for contrast, to see if there was a visible difference in the data from a different (more warmer) climate zone. A general data assessment was conducted to identify any gaps and correlations (temperature - wind speed, precipitation - humidity). Average values for all meteorological parameters were calculated.

Training data quality

During the initial analysis, it became clear that the training data only covered the period from January to May. The table included key project parameters such as humidity and air temperature, but lacked other important parameters such as air pressure and wind direction. No correlation could be found between temperature and wind speed, precipitation, and humidity. Based on the average values, it became clear that the data was completely random — there was no correlation between temperature and part of day.

Based on the above analysis, it can be concluded that this data is suitable for training a model, but completely unsuitable for learning weather patterns and relationships. Therefore, another training data option is being considered — real historical weather data from North American cities for 2012-2017, Kaggle environment. The sample size is larger — 36 cities. Among the weather parameters, air pressure and wind direction are included. This dataset actually needs more time for data preparation for analysis.

Test data quality

The test data was taken from the Open-meteo environment. The sample includes the same three cities: Chicago, New York, and Dallas. The current time range is January to May 2024 and the table contains hourly weather data.

The initial analysis showed no missing data. There is a weak relationship between precipitation and humidity, meaning the more precipitation, the higher the humidity.

3. Planning your project

The following tasks are required in the future:

1. **Evaluation of real historical weather data for North America from 2012 to 2017**
This involves assessing whether the data is suitable for studying real-world weather patterns and relationships.

2. **Final selection of training data**
If other training data is selected, the choice of locations and time range for the test data will likely need to be reconsidered.

3. **Data preparation**
This includes selecting locations and time ranges, denoising the data, splitting the date_time column into two or more parts, and establishing identical column names for the training and test data tables.

4. **Selection of classifiers and machine learning methods**
It is planned to use three to five different machine learning methods. (need to clarify)

5. **Application of machine learning methods**

6. **Evaluation of forecasting accuracy**
First, the accuracy of air temperature and humidity forecasting needs to be assessed, followed by the remaining parameters.

7. **Analysis and presentation of results (poster)**

F10-Weather-prediction

https://github.com/MihBat/F10-Weather-prediction