Predicting the Weather using Machine Learning

Myron Keith Gibert Jr

2022-11-02

Contents

Introduction

Assignment Scenario

Congratulations! You have just been hired by a US Weather forecast firm as a data scientist.

The company is considering the weather condition to help predict the possibility of precipitations, which involves using various local climatological variables, including temperature, wind speed, humidity, dew point, and pressure. The data you will be handling was collected by a NOAA weather station located at the John F. Kennedy International Airport in Queens, New York.

Your task is to provide a high level analysis of weather data in JFK Airport. Your stakeholders want to understand the current and historical record of precipitations based on different variables. For now they are mainly interested in a macro-view of JFK Airport Weather, and how it relates to the possibility to rain because it will affect flight delays and etc.

The Data

This project relates to the NOAA Weather Dataset - JFK Airport (New York). The original dataset contains 114,546 hourly observations of 12 local climatological variables (such as temperature and wind speed) collected at JFK airport. This dataset can be obtained for free from the IBM Developer Data Asset Exchange.

For this project, you will be using a subset dataset, which contains 5727 rows (about 5% or original rows) and 9 columns. The end goal will be to predict the precipitation using some of the available features. In this project, you will practice reading data files, preprocessing data, creating models, improving models and evaluating them to ultimately choose the best model.

Import required modules

Below, install "tidymodels", additionally "rlang" should be updated in order to properly run "tidymodels".

```r
knitr::opts_chunk$set(echo = TRUE)

if (!require("tidyverse")) install.packages("tidyverse")
library("tidyverse")

if (!require("ggplot2")) install.packages("ggplot2")
library("ggplot2")

if (!require("ggthemes")) install.packages("ggthemes")
library("ggthemes")

if (!require("rlang")) install.packages("rlang")
library("rlang")

if (!require("tidymodels")) install.packages("tidymodels")
library("tidymodels")
```

Understand the Dataset

Understand the Dataset The original NOAA JFK dataset contains 114,546 hourly observations of various local climatological variables (including temperature, wind speed, humidity, dew point, and pressure).

In this project you will use a sample dataset, which is around 293 KB:

Link to the sample dataset

The sample contains 5727 rows (about 5% or original rows) and 9 columns, which are:

- DATE
- HOURLYDewPointTempF
- HOURLYRelativeHumidity
- HOURLYDRYBULBTEMPF
- HOURLYWETBULBTEMPF
- HOURLYPrecip
- HOURLYWindSpeed
- HOURLYSeaLevelPressure
- HOURLYStationPressure

The original dataset is much bigger. Feel free to explore the original dataset:

Link to the original dataset

For more information about the dataset, checkout the preview of NOAA Weather - JFK Airport:

Link to the preview

1. Download NOAA Weather Dataset

Use the download.file() function to download the sample dataset from the URL below. Then untar it.

URL = https://dax-cdn.cdn.appdomain.cloud/dax-noaa-weather-data-jfk-airport/1.1.4/noaa-weather-sample-data.tar.gz

```
download.file("https://dax-cdn.cdn.appdomain.cloud/dax-noaa-weather-data-jfk-airport/1.1.4/noaa-weather-

untar("noaa-weather-sample-data.tar.gz")
```

2. Extract and Read into Project

We start by reading in the raw dataset. You should specify the file name as "noaa-weather-sample-data/jfk__weather__sample.csv".

Then, display the first few rows, and use glimpse to confirm its integrity (5727 rows x 9 columns)

```
data <- read.csv("noaa-weather-sample-data/jfk_weather_sample.csv")

head(data)

glimpse(data)
```

```
## Rows: 5,727
## Columns: 9
## $ DATE                   <chr> "2015-07-25T13:51:00Z", "2016-11-18T23:51:00Z",~
## $ HOURLYDewPointTempF    <chr> "60", "34", "33", "18", "27", "35", "4", "14", ~
## $ HOURLYRelativeHumidity <int> 46, 48, 89, 48, 61, 79, 51, 65, 90, 94, 79, 37,~
## $ HOURLYDRYBULBTEMPF     <int> 83, 53, 36, 36, 39, 41, 19, 24, 54, 73, 83, 44,~
## $ HOURLYWETBULBTEMPF     <int> 68, 44, 35, 30, 34, 38, 15, 21, 52, 72, 78, 35,~
## $ HOURLYPrecip           <chr> "0.00", "0.00", "0.00", "0.00", "T", "0.00", "0~
## $ HOURLYWindSpeed        <int> 13, 6, 13, 14, 11, 6, 0, 11, 11, 5, 21, 7, 17, ~
## $ HOURLYSeaLevelPressure <dbl> 30.01, 30.05, 30.14, 29.82, NA, 29.94, 30.42, 3~
## $ HOURLYStationPressure  <dbl> 29.99, 30.03, 30.12, 29.80, 30.50, 29.92, 30.40~
```

3. Select Subset of Columns

The end goal of this project will be to predict HOURLYprecip (precipitation) using a few other variables. Before you can do this, you first need to preprocess the dataset. Section 3 to section 6 focuses on preprocessing.

The first step in preprocessing is to select a subset of data columns and inspect the column types.

The key columns that we will explore in this project are:

- HOURLYRelativeHumidity
- HOURLYDRYBULBTEMPF
- HOURLYPrecip
- HOURLYWindSpeed
- HOURLYStationPressure

Data Glossary:

- 'HOURLYRelativeHumidity' is the relative humidity given to the nearest whole percentage.
- 'HOURLYDRYBULBTEMPF' is the dry-bulb temperature and is commonly used as the standard air temperature reported. It is given here in whole degrees Fahrenheit.
- 'HOURLYPrecip' is the amount of precipitation in inches to hundredths over the past hour. For certain automated stations, precipitation will be reported at sub-hourly intervals (e.g. every 15 or 20 minutes) as an accumulated amount of all precipitation within the preceding hour. A "T" indicates a trace amount of precipitation.
- 'HOURLYWindSpeed' is the speed of the wind at the time of observation given in miles per hour (mph).
- 'HOURLYStationPressure' is the atmospheric pressure observed at the station during the time of observation. Given in inches of Mercury (in Hg).

Select those five columns and store the modified dataframe as a new variable. Then, show the first ten rows.

```
data_subset <- data %>%
  dplyr::select(HOURLYRelativeHumidity,HOURLYDRYBULBTEMPF,HOURLYPrecip,HOURLYWindSpeed,HOURLYStationPres

head(data_subset,10)
```

```
##    HOURLYRelativeHumidity HOURLYDRYBULBTEMPF HOURLYPrecip HOURLYWindSpeed
## 1                      46                 83         0.00              13
## 2                      48                 53         0.00               6
## 3                      89                 36         0.00              13
## 4                      48                 36         0.00              14
## 5                      61                 39            T              11
## 6                      79                 41         0.00               6
## 7                      51                 19         0.00               0
## 8                      65                 24         0.00              11
## 9                      90                 54         0.06              11
## 10                     94                 73         <NA>               5
##    HOURLYStationPressure
## 1                  29.99
## 2                  30.03
## 3                  30.12
## 4                  29.80
## 5                  30.50
## 6                  29.92
## 7                  30.40
## 8                  30.35
## 9                  30.03
## 10                 29.91
```