Keene Particulate Matter Project - Roadmap Document

Index

1. [Data Science Lifecycle Review](#_Data_Science_Lifecycle)
2. [Problem Definition and Domain Knowledge](#_Problem_Definition_and)
   1. Problem Definition
   2. Domain Knowledge
      1. Literature Review
3. [Data Collection and Sourcing](#_Data_Collection_and)
4. [Data Cleaning and Processing](#_Data_Cleaning_and)
5. [Exploratory Data Analysis](#_Exploratory_Data_Analysis)
6. [Model Building and Evaluation](#_Model_Building_and)
7. [Model Results](#_Model_Results)
8. [Model Deployment](#_Model_Deployment)

## Data Science Lifecycle Research/Review

A diagram of a process

Description automatically generated

## Problem Definition and Domain Knowledge

### Problem Definition

1. To forecast, with the highest possible accuracy, when PM 2.5 levels will be elevated in the Connecticut River Valley, specifically over the City of Keene, given meteorological data gathered by KSC’s Nora Traviss.
2. To evaluate each model to optimize for a variety of factors: error, compute, number of features, and more to be determined.

### Domain Knowledge

Ad Hoc Links-

Time Series Analysis

<http://mayoral.iae-csic.org/timeseries2021/hamilton.pdf>

<https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm>

<https://www.youtube.com/watch?v=R4tcKNJe3xw>

Weather Modeling

Particulate Matter Forecasting

Air Inversion Forecasting

## Data Collection and Sourcing

Data set is collected from a meteorological station on Water Street in Keene NH. Data was gathered and distributed by Dr. Nora Traviss for the purpose of forecasting PM 2.5 concentration in the Keene area.

TODO: Verify this information and contact Dr. Traviss or Dr. McGregor for updated data.

## Data Cleaning and Processing

Table 1: Data Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Code** | **Column** | **Descriptive Name** | **Type** | **Nulls** |
| F0 | Datapoint Number | Row Number | Int [0,+] | No |
| F1 | Datetime | Datetime Object | Datetime | No |
| F2 | Created At | Text Date and Time | Date + Time |  |
| F3 | PM25 | Particulate Matter <2.5 Microns Diameter | Float [0,+] | No |
| F4 | Date | Date with less frequently updated time | Datetime |  |
| F5 | temp | Temperature in Fahrenheit | Float [-,100] | No |
| F6 | dewpoint | Dew Point | Float [-,+] | No |
| F7 | RH | Relative Humidity | Float [0,100] | No |
| F8 | windDIR | Wind Direction in Degrees | Float [0,360] | No |
| F9 | windMPH | Wind Speed | Float [0,+] | No |
| F10 | precip | Precipitation | Float [0,+] | No |
| F11 | mslp | Mean Sea Level Pressure | Float [0,+] | No |
| F12 | visibility | Visibility | Float [0,10] | No |
| F13 | gust | Gust Speed | Float [0,+] | No |
| F14 | wxcodes | Weather Codes | String | Yes |
| F15 | (top) Snow Depth (in) | Top Snow Depth | Float [0,+] | No |
| F16 | (middle) Snow Depth (in) | Middle Snow Depth | Float [0,+] | No |
| F17 | (bottom) Snow Depth (in) | Bottom Snow Depth | Float [0,+] | No |
| F18 | (top) Snow Temp. (deg. F) | Top Snow Temperature | Float [0,+] | No |
| F19 | (middle) Snow Temp. (deg. F) | Middle Snow Temperature | Float [0,+] | No |
| F20 | (bottom) Snow Temp. (deg. F) | Bottom Snow Temperature | Float [0,+] | No |
| F21 | (top) Snow Density (%) | Top Snow Density | Float [0,+] | No |
| F22 | (middle) Snow Density (%) | Middle Snow Density | Float [0,+] | No |
| F23 | (bottom) Snow Density (%) | Bottom Snow Density | Float [0,+] | No |
| F24 | Date w/o Time | Date | Date | Yes |
| F25 | Hour | Hour | Float [0,24] |  |
| F26 | Forecasted from 0 UTC | Forecasted from 0 UTC | String | Yes |
| F27 | FEW | Few Cloud Layer | Int [0,12000] | Yes |
| F28 | SCT | Scattered Cloud Layer | Int [0,12000] | Yes |
| F29 | BKN | Broken Cloud Layer | Int [0,12000] | Yes |
| F30 | OVC | Overcast Cloud Layer | Int [0,12000] | Yes |
| F31 | VV | Vertical Visibility | Int [0,12000] | Yes |
| F32 | Clouds | Cloud Coverage | [0,8] | Yes |
| F33 | Clds1000 | Cloud Coverage at 1000 ft | [0,8] | Yes |
| F34 | Clds2000 | Cloud Coverage at 2000 ft | [0,8] | Yes |
| F35 | Clds3000 | Cloud Coverage at 3000 ft | [0,8] | Yes |
| F36 | Clds4000 | Cloud Coverage at 4000 ft | [0,8] | Yes |
| F37 | Clds5000 | Cloud Coverage at 5000 ft | [0,8] | Yes |
| F37 | Clds6000 | Cloud Coverage at 6000 ft | [0,8] | Yes |
| F38 | Clds7000 | Cloud Coverage at 7000 ft | [0,8] | Yes |
| F39 | Clds8000 | Cloud Coverage at 8000 ft | [0,8] | Yes |
| F40 | Clds9000 | Cloud Coverage at 9000 ft | [0,8] | Yes |
| F41 | Clds10000 | Cloud Coverage at 10000 ft | [0,8] | Yes |

### Data Cleaning

Replace missing values

Which features need this: WX Code, Clouds

Could I make the data more usable by regularizing the time delta between data readings. Perhaps just duplicate rows of data to more evenly space out the time of recording.

Remove erroneous duplicate values?

Ex. Dewpoint and RH

### Feature Engineering

Each feature \* Each feature

## Exploratory Data Analysis

<https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>

Summary stats

Histograms, boxplots, scatterplots

Correlation analysis

Hypothesis testing

## Model Building and Evaluation

## Model Results

## Model Deployment