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.

## 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 | [ | No |
| F4 | Date | Date with less frequently updated time |  |  |
| F5 | temp | Temperature in Fahrenheit |  | No |
| F6 | dewpoint | Dew Point |  | No |
| F7 | RH | Relative Humidity |  | No |
| F8 | windDIR | Wind Direction in Degrees |  | No |
| F9 | windMPH | Wind Speed |  | No |
| F10 | precip | Precipitation |  | No |
| F11 | mslp | Mean Sea Level Pressure |  | No |
| F12 | visibility | Visibility |  | No |
| F13 | gust | Gust Speed |  | No |
| F14 | wxcodes | Weather Codes |  | Yes |
| F15 | (top) Snow Depth (in) | Top Snow Depth |  | No |
| F16 | (middle) Snow Depth (in) | Middle Snow Depth |  | No |
| F17 | (bottom) Snow Depth (in) | Bottom Snow Depth |  | No |
| F18 | (top) Snow Temp. (deg. F) | Top Snow Temperature |  | No |
| F19 | (middle) Snow Temp. (deg. F) | Middle Snow Temperature |  | No |
| F20 | (bottom) Snow Temp. (deg. F) | Bottom Snow Temperature |  | No |
| F21 | (top) Snow Density (%) | Top Snow Density |  | No |
| F22 | (middle) Snow Density (%) | Middle Snow Density |  | No |
| F23 | (bottom) Snow Density (%) | Bottom Snow Density |  | No |
| F24 | Date w/o Time | Date |  | Yes |
| F25 | Hour | Hour |  |  |
| F26 | Forecasted from 0 UTC | Forecasted from 0 UTC |  | Yes |
| F27 | FEW | Few Cloud Layer |  |  |
| F28 | SCT | Scattered Cloud Layer |  |  |
| F29 | BKN | Broken Cloud Layer |  |  |
| F30 | OVC | Overcast Cloud Layer |  |  |
| F31 | VV | Vertical Visibility |  |  |
| F32 | Clouds | Cloud Coverage |  |  |
| F33 | Clds1000 |  |  |  |
| F34 | Clds2000 |  |  |  |
| F35 | Clds3000 |  |  |  |
| F36 | Clds4000 |  |  |  |
| F37 | Clds5000 |  |  |  |
| F37 | Clds6000 |  |  |  |
| F38 | Clds7000 |  |  |  |
| F39 | Clds8000 |  |  |  |
| F40 | Clds9000 |  |  |  |
| F41 | Clds10000 |  |  |  |

### Data Cleaning

Replace missing values

### 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