SDS 322E - Project 2 Instructions

2023-04-15

Overview

The goal of this project is to build a prediction model that can predict ambient air pollution concentrations across the continental United States. You have previously done this using linear and logistic regression and a smaller dataset. In this project, you will have an expanded dataset and will need to

1. Build a series of at least 3 models (or 4 models, if working in a group) using the collection of available predictors;
2. Compare the performance of your models to find the optimal approach;
3. Use the output of your models to answer some policy questions.

Working independently or in a group

For the project, you can work independently or in a group of at most 3 students (including yourself). You can choose the members of your group (i.e. you do not have to use your lab group or the group you used for Project 1). If working in a group, you will be asked to build **4 different models** (instead of 3). But still only produce one unique report per group.

Download the RStudio Project

1. Open RStudio and click on "New Project…" in the drop down menu in the upper right.
2. In the New Project Wizard, select Version Control.
3. In the next menu titled "Create Project from Version Control", select Git.
4. Under "Repository URL", enter the web site URL for this lab: https://github.com/SDS322E/Project2

5. Under "Project directory name", enter Lab10. (RStudio may automatically put this into the box for you.)
6. You may choose to select a directory to store the project or you can use the default.
7. Click "Create Project".

Once you are in the RStudio project for **Project2** you can:

- Edit the **Project2_Report.Rmd** file included in the RStudio Project and edit the file according to the instructions it contains.
- Knit the **Project2_Report.Rmd** file into an HTML file and then print as a PDF using your web browser.

Data

The dataset contains annual average concentrations of fine particulate matter (PM2.5) across the U.S. Environmental Protection Agency's monitoring network in the continental U.S. In addition, the dataset You can read the data into R using the following code:

```
library(tidyverse)

dat <- read_csv("pm25_data.csv.gz")
```

Outcome

The outcome is in a column called `value` and it represents the annual average PM2.5 concentration at a monitor. It is measured in micrograms per cubic meter. The measurements here were taken using what is known as a gravimetric method and it is considered the gold standard for outdoor air pollution concentrations.

Predictors

Information about each of the predictor variables found in the dataset can be found in the table at the end of these instructions. Please go through this table carefully as it will help you in developing your prediction models.

Some predictors of note are:

- `CMAQ` - the values in this column represent predictions from a numerical computer model developed by the EPA. This model simulates pollution in the atmosphere and does not require any monitoring data to run.

- `aod` - this measurement represents the "aerosol optical depth", which is measured from satellites and is related to the amount of pollution near the surface.

Objectives

The main goal of this project is to build 3 different prediction models (or 4 models, if working in a group) and compare their performance to find the best model. You are welcome to use any of the models discussed in class (e.g. linear regression, k-nearest neighbors, random forest) or any other model that you learn about elsewhere. You can search the full list of models supported by Tidymodels to find other modeling approaches. Please note that for some of the models listed on that web site, you may need to install additional R packages.

When building your models, make sure to split the data into training and testing datasets so that the prediction metrics can be properly evaluated in an unbiased manner. Because the outcome (`value`) is continuous, the primary prediction metric you will be using to compare your models is **root mean-squared error (RMSE)**.

NOTE: You will **not** be doing classification in this project.

Use the training dataset to identify important predictors, do feature extraction or transformations of any of the predictors, and to explore the performance of each of your models. If your modeling approach has tuning parameters, make sure to tune your model to find the optimal set of tuning parameters. Compare the RMSEs across your different modeling approaches and determine which is the **best** model according to RMSE. For the "best and final" model, evaluate its RMSE on the test dataset.

Additional notes:

1. You do NOT have to use all of the predictors in the dataset to build your prediction model. However, using more predictors may result in a better-performing model.
2. You do not need to use the same set of predictors in each of your models.

Primary Questions

As part of your final report, you must answer the following questions using the model that you chose as your "best and final" model.

1. Based on test set performance, at what geographic locations in the test set does your model give predictions that are closest and furthest from the observed values? What do you hypothesize are the reasons for the good or bad performance at these locations?

2. What variables might predict where your model performs well or not? For example, are their regions of the country where the model does better or worse? Are there variables that are not included in this dataset that you think might improve the model performance if they were included in your model?
3. There is interest in developing more cost-effect approaches to monitoring air pollution on the ground. Two candidates for replacing the use of ground-based monitors are numerical models like CMAQ and satellite-based observations such as AOD. How does the prediction performance of your best and final model change when `CMAQ` or `aod` are included (or not included) in the model?
4. The dataset here did not include data from Alaska or Hawaii. Do you think your model will perform well or not in those two states? Explain your reasoning.

Report

The text of your report will provide a narrative structure around your code and outputs with R Markdown. Answers without supporting code will not receive credit and outputs without comments will not receive credit either: write full sentences to describe your findings. All code contained in your final project document must work correctly (knit early, knit often)! **(If a group project, only submit one report**.

Guidelines for the report:

Introduction

Write a narrative introduction containing:

- A description of the modeling approaches you have chosen for your comparison;
- An explanation of how the predictor variables were chosen for your model;
- Any exploratory analysis that was done with the data (e.g. correlations, scatterplots, boxplots, histograms, etc.) in order to learn about relationships in the data;
- Your expectation for what the RMSE performance of your best and final model should be.

Wrangling

This section should include any wrangling or transformations of the data that were done prior to modeling. Code should be included throughout and a textual explanation should be included to explain any wrangling operations.

Results

Describe the development of your 3 prediction models (or 4 models, if working in a group) and how you compared their performance. Be sure to describe the splitting of training and testing datasets and the use of cross-validation to evaluate prediction metrics. Remember that the primary metric for your prediction model will be root mean-squared error (RMSE).

Your results should include a scatterplot of predicted values vs. observed values for the best and final model in the test dataset. You should also include a table summarizing the prediction metrics (i.e. RMSE) across all of the models that you tried.

Discussion

Putting it all together, what did you learn from your data and your model performance?

- Answer the Primary Questions posed above, citing any supporting statistics, visualizations, or results from the data or your models.
- Reflect on the process of conducting this project. What was challenging, what have you learned from the process itself?
- Reflect on the performance of your final prediction model. Did it perform as well as you originally expected? If not, why do you think it didn't perform as well?
- Include acknowledgements for any help received. If a group project, report the contribution of each member (i.e. who did what?).

Formatting

Create the report using R Markdown, with headers for each section; include comments to the R code; include references (datasets, context). The final report should be no more than 20 pages (the number of pages can vary greatly depending on the cleaning process). It is extremely important that you **select pages** when submitting on Gradescope.

Appendix: Predictor Variables Summary

Variable	Details
id	Monitor number – the county number is indicated before the decimal – the monitor number is indicated after the decimal **Example**: 1073.0023 is Jefferson county (1073) and .0023 one of 8 monitors
fips	Federal information processing standard number for the county where the monitor is located – 5 digit id code for counties (zero is often the first value and sometimes is not shown) – the first 2 numbers indicate the state – the last three numbers indicate the county **Example**: Alabama's state code is 01 because it is first alphabetically (note: Alaska and Hawaii are not included because they are not part of the contiguous US)
Lat	Latitude of the monitor in degrees
Lon	Longitude of the monitor in degrees
state	State where the monitor is located
county	County where the monitor is located
city	City where the monitor is located
CMAQ	Estimated values of air pollution from a computational model called **Community Multiscale Air Quality (CMAQ)** – A monitoring system that simulates the physics of the atmosphere using chemistry and weather data to predict the air pollution – ***Does not use any of the $PM_{2.5}$ gravimetric monitoring data.*** (There is a version that does use the gravimetric monitoring data, but not this one!) – Data from the EPA
zcta	Zip Code Tabulation Area where the monitor is located – Postal Zip codes are converted into "generalized areal representations" that are non-overlapping – Data from the 2010 Census
zcta_area	Land area of the zip code area in meters squared – Data from the 2010 Census
zcta_pop	Population in the zip code area – Data from the 2010 Census
imp_a500	Impervious surface measure – Within a circle with a radius of 500 meters around the monitor – Impervious surface are roads, concrete, parking lots, buildings – This is a measure of development

Variable	Details
imp_a1000	Impervious surface measure – Within a circle with a radius of 1000 meters around the monitor
imp_a5000	Impervious surface measure – Within a circle with a radius of 5000 meters around the monitor
imp_a10000	Impervious surface measure – Within a circle with a radius of 10000 meters around the monitor
imp_a15000	Impervious surface measure – Within a circle with a radius of 15000 meters around the monitor
county_area	Land area of the county of the monitor in meters squared
county_pop	Population of the county of the monitor
Log_dist_to_prisec	Log (Natural log) distance to a primary or secondary road from the monitor – Highway or major road
log_pri_length_5000	Count of primary road length in meters in a circle with a radius of 5000 meters around the monitor (Natural log) – Highways only
log_pri_length_10000	Count of primary road length in meters in a circle with a radius of 10000 meters around the monitor (Natural log) – Highways only
log_pri_length_15000	Count of primary road length in meters in a circle with a radius of 15000 meters around the monitor (Natural log) – Highways only
log_pri_length_25000	Count of primary road length in meters in a circle with a radius of 25000 meters around the monitor (Natural log) – Highways only
log_prisec_length_500	Count of primary and secondary road length in meters in a circle with a radius of 500 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_1000	Count of primary and secondary road length in meters in a circle with a radius of 1000 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_5000	Count of primary and secondary road length in meters in a circle with a radius of 5000 meters around the monitor (Natural log) – Highway and secondary roads

Variable	Details
log_prisec_length_10000	Count of primary and secondary road length in meters in a circle with a radius of 10000 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_15000	Count of primary and secondary road length in meters in a circle with a radius of 15000 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_25000	Count of primary and secondary road length in meters in a circle with a radius of 25000 meters around the monitor (Natural log) – Highway and secondary roads
log_nei_2008_pm25_sum_10000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 10000 meters of distance around the monitor (Natural log)
log_nei_2008_pm25_sum_15000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 15000 meters of distance around the monitor (Natural log)
log_nei_2008_pm25_sum_25000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 25000 meters of distance around the monitor (Natural log)
log_nei_2008_pm10_sum_10000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 10000 meters of distance around the monitor (Natural log)
log_nei_2008_pm10_sum_15000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 15000 meters of distance around the monitor (Natural log)
log_nei_2008_pm10_sum_25000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 25000 meters of distance around the monitor (Natural log)
popdens_county	Population density (number of people per kilometer squared area of the county)
popdens_zcta	Population density (number of people per kilometer squared area of zcta)

Variable	Details
nohs	Percentage of people in zcta area where the monitor is that **do not have a high school degree** – Data from the Census
somehs	Percentage of people in zcta area where the monitor whose highest formal educational attainment was **some high school education** – Data from the Census
hs	Percentage of people in zcta area where the monitor whose highest formal educational attainment was completing a **high school degree** – Data from the Census
somecollege	Percentage of people in zcta area where the monitor whose highest formal educational attainment was completing **some college education** – Data from the Census
associate	Percentage of people in zcta area where the monitor whose highest formal educational attainment was completing an **associate degree** – Data from the Census
bachelor	Percentage of people in zcta area where the monitor whose highest formal educational attainment was a **bachelor's degree** – Data from the Census
grad	Percentage of people in zcta area where the monitor whose highest formal educational attainment was a **graduate degree** – Data from the Census
pov	Percentage of people in zcta area where the monitor is that lived in **poverty** in 2008 - or would it have been 2007 guidelines??https://aspe.hhs.gov/2007-hhs-poverty-guidelines – Data from the Census
hs_orless	Percentage of people in zcta area where the monitor whose highest formal educational attainment was a **high school degree or less** (sum of nohs, somehs, and hs)
urc2013	2013 Urban-rural classification of the county where the monitor is located – 6 category variable - 1 is totally urban 6 is completely rural – Data from the National Center for Health Statistics

Variable	Details
urc2006	2006 Urban-rural classification of the county where the monitor is located – 6 category variable - 1 is totally urban 6 is completely rural – Data from the National Center for Health Statistics
aod	Aerosol Optical Depth measurement from a NASA satellite – based on the diffraction of a laser – used as a proxy of particulate pollution – unit-less - higher value indicates more pollution – Data from NASA