SDS 322E: Project 2 Instructions

Overview

The goal of this project is to build a prediction model and compare the performance of different prediction modeling methods. You will do this with one of the four provided datasets, included in the Project 2 repository. In this project, you will need to

1. Build a base prediction model using linear or logistic regression (depending on the nature of the outcome)
2. Tune your base model to find the optimal version
3. Compare your prediction model to a nonparametric machine learning method
4. Tune the machine learning model's tuning parameters to find the optimal configuration
5. Determine the best performing model of the ones you have tried

Working in a Group

For the project, you must work in your assigned group. The groups have been randomly generated and assigned on Canvas. Note that it is a different group than your lab group and Project 1 group. Although you will be working in groups, you will produce one unique report per group.

Preliminary Report

The project will be done in two parts. The first part will be the **Preliminary Report** where you will need to answer some basic questions about your dataset and your project. Use the file `Project2_Preliminary.Rmd` in the Project 2 repository for your preliminary report and follow the prompts in that file.

The preliminary report will be graded on **completion only**. Furthermore, if you decided to make a change to your project after completing the preliminary report, that is okay.

The preliminary report will be submitted as a PDF in Gradescope.

Final Report

The final report will contain your main prediction model analysis. Use the file `Project2_Report.Rmd` in the Project 2 repository for your final report and follow the prompts in that file.

The final report will be submitted as a PDF in Gradescope. **Make sure to mark the pages that correspond to each part of the project**.

Dataset Options

Please choose **one** of the following datasets for your project. The data files are located in the Project 2 repository already, and their file names are listed below.

1. Fine Particulate Matter Air Pollution in the United States – This is an expanded version of the air pollution dataset used in Lab 11 (this version has more variables). You can learn more about the variables in this dataset at the Open Case Studies web site.
    - Repository file: `pm25_data.csv.gz`

2. Austin Crash Report Data – This dataset contains traffic crash records for crashes which have occurred in Austin, TX, in the last ten years.
    - Repository file: `Austin_Crash_Report_Data_-_Crash_Level_Records_20250407.csv.gz`
3. Austin Animal Center Intakes – Animal Center Intakes from October 2013 to April 2025. Intakes represent the status of animals as they arrive at the Animal Center.
    - Repository file: `Austin_Animal_Center_Intakes_20250407.csv.gz`
4. Barton Springs Salamanders DO and Flow – Data collected to assess water quality conditions in the natural creeks, aquifers and lakes in the Austin area.
    - Repository file: `Barton_Springs_Salamanders_DO_and_Flow_20250407.csv.gz`

All of these datasets are in CSV format and can be read in using the `read_csv()` function in the tidyverse.

Download the Project

1. Open RStudio.
2. Click File > New Project... > Version Control > Git.
3. Enter this information.

    Project URL: `https://github.com/SDS-322E-FA25/Project2`

    Project directory name: Project2
4. Create project as a subdirectory of: *You may choose to select a directory to store the project or use the default.*
5. Click Create Project.

Once you are in the RStudio project for **Project2** you will find

- `Project2-Instructions.pdf` – the instructions for the project
- `Project2_Preliminary.Rmd` – the template for the preliminary report
- `Project2_Report.Rmd` – use this file for your final report
- The four dataset files