SDS 322E: Project 2 Instructions

Overview

The goal of this project is to build a prediction model and comparing the performance of different prediction modeling methods. You will do this in a dataset of your choosing (there is a list of suggested datasets below). 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 independently or in a group

For the project, you will work in a group of 4 students (including yourself). See the Canvas site for the group allocation.

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. The due date is 11/20/2025 11:59PM.

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. The due date is 12/08/2025 11:59PM. **Make sure to mark the pages that correspond to each part of the project**.

Suggested Datasets

The following datasets are suggestions. 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 that you 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 Oct, 1st 2013 to present. 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 the suggested datasets are in CSV format and can be read using the read_csv() function in the tidyverse.

You are welcome to find and use your own dataset if you do not want to use one of the suggested datasets. If you choose to use another dataset, you will need to download it separately and copy it into the project folder.

Download the RStudio Project

Create the RStudio project for Project 2 by running the following command in R:

```
usethis::create_from_github("https://github.com/SDS322E-2025Fall/Project2", fork = FALSE)
```

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 suggested dataset files