Stat 486 Fall 2023 Final Project Description

1 Project Overview

The goal of the project is to perform in-depth statistical analysis and prediction on a dataset. You will, in a group of four, choose a dataset and analyze it via appropriate machine learning methods related to the class.

You will first conduct exploratory analysis, with helpful visualizations, to describe important features of the data, to give a high level idea of the relationships between the important features, and to identify any outliers or missing data in the dataset.

Next, you will use supervised learning algorithms to predict certain important features in the data. You should compare at least two different predictive algorithms. You may also apply variable selection/transformation, dimensionality reduction, or clustering methods as appropriate. You can use any R/Python commands or packages you would like.

You can choose a dataset in the **Data Ideas** part of this project description or find an interesting dataset of your own, subject to the instructor's approval. *At most 3 groups can use the same dataset in their project*; if more than 3 groups choose the same dataset, the assignment will be made on a first come first serve basis.

2 Group

You will work in a group of four. Groups of three may be acceptable with permission from the instructor. If you do not have a group, then send an email to the instructor with the heading *486: random group request* and you will be placed in a random group.

3 Deliverables

There will be three main deliverables: (a) 6-8 minutes presentation. The presentations will be given on the last two lectures (Dec 7 and Dec 12). (b) The slides used in the presentation. (c) A report that summarizes your findings, (d) All the codes used in the analysis.

The presentation and the slides should contain the following sections:

- An **introductory description** of the dataset and the predictive problems that you are considering.
- **Exploratory data analysis** results showing what the data looks like.
- A **description of the models, methods, algorithms** that you are using.
- The **results** of predictive analysis, variable selection, dimensionality reduction, or clustering.

- A qualitative **discussion** of your findings. You can also discuss any future directions: what you would do if you had more data, more time, and more observed features.

4 Deadlines

- **By Thursday, November 9th**. Submit on Canvas (one per group) a text file that gives your group members and your chosen dataset. If you choose a dataset not in the **Dataset Ideas** section, be sure to include a link so that we can see if it is appropriate for the project. [5% of total project grade]
- **By Thursday, November 23rd**. Submit on Canvas (one per group) a text file containing one or two paragraphs that succinctly describe any preliminary work that you have done. [5% of total project grade]
- **By Thursday, December 7th**. Project presentation. You must submit your finished slides and report on Canvas before Midnight on this day. [90% of total project grade]

5 Dataset Ideas

1. Kaggle Movies

   `https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv`

2. UCI bone marrow

   `https://archive.ics.uci.edu/ml/datasets/Bone+marrow+transplant%3A+children`

3. UCI Breast Cancer

   `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`

4. Kaggle football

   `https://www.kaggle.com/jeffgallini/college-football-team-stats-2019?select=cfb20.csv`

5. Kaggle KC housing

   `https://www.kaggle.com/harlfoxem/housesalesprediction`

6. Boston Housing

   `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`

7. Kaggle Wine Review

   `https://www.kaggle.com/zynicide/wine-reviews`

8. UCI Census income data

   `https://archive.ics.uci.edu/ml/datasets/Census+Income`

9. Kaggle Student Alcohol Consumption

   `https://www.kaggle.com/uciml/student-alcohol-consumption`