Data Science for AAE

AAE 718 – Summer 2025

Project 04

Models

Due: Monday, June 16

1 Portfolio

Similar to Project 3 you will be creating a GitHub repo with your code and write up. Give your repository an informative name and structure it similar to the previous project.

You will, again, submit a PDF to Canvas with a link to your GitHub repo. Make sure that either the repo is public or I have access.

2 Project Description

The goal of this project is to find some data and perform a modeling task. You can use any data you want and ask any question, but you must use either a regression model or a decision tree/forest[1].

What ever data you find you should write a function to pull the data and clean it. That means connecting to the API and downloading the data. You should have no data files in your repo, be sure to use gitignore to exclude them.

If you require an API key, you should store it in a separate file and use gitignore to exclude it. This isn't exactly best practice, but it's better than sharing your key. Include in the README how to register for an API key and where to place it.

For this project you will want to use the train_test_split function to split your data into a training and testing set. This is important because you want to test your model on data it hasn't seen before. Just be sure to train your data on the training set and test it on the testing set. There are a TON of examples online using this methodology.

Include in your report your training and testing score. Make predictions and graphs and tell me why you are considering this problem. What are the implications of your findings? What are the limitations of your model? How could you improve it?

2.1 Potential Data Sources

- The Census - You're already familiar with the API.
- Kaggle - There are many datasets on Kaggle. You can search for datasets by topic or type.
- Data.gov - This is a repository of government data. You can search by topic or agency.
- World Bank - The World Bank has a lot of data on development, poverty, and other topics. You can search by country or topic.
- Google Dataset Search - This is a search engine for datasets. You can search by topic or type.

2.2 Backup Idea

If you are overwhelmed with choice and would like an idea, this section is for you.

Your goal is to compete in this Kaggle competition. Your write up with introduce the problem and detail the methods you used to solve the problem.

[1] Or you can use a different method not covered in class, want to experiment with a neural network? Go for it

There are many techniques you can employ. In this class we covered linear modelling and decision trees, but there are so many more methods, the book details options. You are not required to use something new, but you're allowed to.

I would expect you to get a score of at least .75 to be successful. The higher your score, the better. Be sure to report your score in your write up.