Prof. James MacKinnon QUEEN'S UNIVERSITY Department of Economics

Economics 882 Winter, 2025
Empirical Project

Due: May 2, 2025

This project is worth 60% of the final mark. It is deliberately open-ended, so you could spend a great deal of time working on it. However, you probably don't want to do that.

1. Pick a dataset suitable for performing supervised machine learning. The objective is to use it for either regression or classification of one, or perhaps more than one, output variable. Discuss why the dataset you chose is suitable for whatever you will be using it for.

Important Note. Two students should not use the same dataset unless the output variable(s) are different. In order to prevent this happening by accident, please tell Mehtab what dataset you are planning to use as soon as you make even a tentative decision. It will also be helpful to consult with Mehtab and/or with me before you make a final decision.

[15 marks]

2. Randomly divide the dataset into a training set and a test set. Explain how you did this, and submit the code. You will use the training set for fitting the models and the test set for evaluating how well they predict. If the sample is very small, or there is some other good reason for not splitting the sample, then explain why you did not split the sample and how you are going to evaluate model performance without a test set.

[5 marks]

3. Obtain several sets of predictions on the test dataset using at least two different machine-learning models, and compare their performance. Each model may, or may not, be used to generate more than one set of predictions. If possible, the models should be entirely different ones, and none of them should be a model that you expect to perform poorly. However, it may be interesting to compare predictions from these models with those from one or more models that you do not expect to perform well, such as naive linear regression models.

Explain why you chose to use the models you did. Compare the predictions in whatever ways seem appropriate. Please submit the code for whatever results you choose to discuss.

[70 marks]

4. Explain what you have learned from what you did in question 3. Is there more that you would like to do if you had the time?

[10 marks]

Choosing a Dataset

Ideally, you will find a dataset that is of interest to you and has not already been studied extensively using supervised learning methods. There are many places to find data on the web.

In some cases, journal articles that used conventional econometric methods could have used statistical learning methods instead. It would often be of interest to see how the results change.

You can easily find the data for many articles in journals such as the *AER*, *CJE*, *JAE*, *JPE*, and *QJE*. However, not all of these data will be complete, well-organized, and easy to read into R. These data may be behind pay walls, but the walls should be easy to breach if you are on campus or go through the library. Some authors also make data available on their own websites; Josh Angrist at MIT is a famous example.

There is an enormous number of datasets available through IPUMS at

`https://ipums.org`

These include both census and survey data, including the Current Population Survey and the American Community Survey, as well as health data for many countries. Even if you do not end up using data from IPUMS for this assignment, it might be worthwhile to spend some time on the website finding out about the data that are available.

If you cannot find a dataset that interests you on a journal website or at IPUMS, here are a few other suggestions:

[1] The Dataverse Project `https://dataverse.org/installations` provides access to an enormous number of academic data repositories around the world.

[2] Statistics Canada provides a variety of datasets. Go to

`https://www150.statcan.gc.ca/n1/en/type/data`

Some of them must be suitable for methods we have discussed.

[3] There are individual data on mortgage applications for a representative sample from the United States at

`https://www.consumerfinance.gov/data-research/hmda/historic-data/`

The objective would be to forecast mortgage approvals.

[4] Google provides Google Trends Data at:

`https://trends.google.com`

This gives an index for the popularity of terms searched on Google by region. It is intended to be used along with other data.

[5] There are various housing datasets on the web. In particular, there is a dataset for King County, Washington (Seattle) that is available from Kaggle. But see the discussion of Kaggle below.

[6] The Substance Abuse and Mental Health Services Administration in the United States (SAMSHA) provides a number of datasets concerning, among other things, drug abuse and emergency room visits here:

`https://www.samhsa.gov/`

[7] All the datasets for *Elements of Statistical Learning* are available here:

`https://web.stanford.edu/~hastie/ElemStatLearn/data.html`

and all the ones for ISLR/ISLP are available here:

`https://www.statlearning.com/resources-second-edition`

If you simply repeat the analysis in *ESL* or *ISLR/ISLP*, you will get a very bad mark. However, in some cases, it may be possible to employ methods on some of these datasets that are quite different from the ones used in the books. If you do this, you will need to convince me that what you did is interesting.

[8] I already mentioned Kaggle:

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

It provides a great many datasets. However, many of these have already been studied extensively. If you use a dataset from Kaggle, you will need to convince me that at least part of what you are doing with it is new and interesting.