Machine Learning for Public Policy - Mini-Project 2

The University of Chicago - Harris School of Public Policy
PPHA 30545 - Professors Clapp and Levy
Winter 2025

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Thursday, February 13th.** There will be separate Gradescope assignments for R and Python students. Please be sure to submit to the version that matches the coding language of the lab section you are enrolled in.

You are welcome (and encouraged!) to form study groups (of no more than 2 students) to work on the problem sets and mini-projects together. But you must write your own code and your own solutions. Please be sure to include the names of those in your group on your submission.

You should format your submission in one of two ways:

1. As a single PDF containing BOTH a write-up of your solutions that directly integrates any relevant supporting output from your code (e.g., estimates, tables, figures) AND your code appended to the end of your write up. You may type your answers or write them out by hand and scan them (as long as they are legible). Your original code may be a R (*.rmd) or Python (*.py) file converted to PDF format. OR

2. As a single PDF of an R Markdown (*.rmd), Jupyter Notebook (*.ipynb), or Quarto (*.qmd) document with your your solutions and explanations written in Markdown.[1]

Regardless of how you format your answers, be sure to make it clear what question you are answering by labeling the sections of your write up well and assigning your answers to the appropriate question in Gradescope. Also, be sure that it is immediately obvious what supporting output from your code (e.g., estimates, tables, figures) you are referring to in your answers. In addition, your answers should be direct and concise. Points will be taken off for including extraneous information, even if part of your answer is correct. You may use bullet points if they are beneficial. Finally, for your code, please also be sure to practice the good coding practices you learned in Data and Programming and comment your code, cite any sources you consult, etc.

You are allowed to consult the textbook authors' website, R/Python documentation, and websites like StackOverflow for general coding questions. If you get help from a large language model (LLM) or other AI tool (e.g., ChatGPT), you must provide in the query string you used and an explanation of how you used the AI tool's response as part of your answer. You are not allowed to consult material from other classes (e.g., old problem sets, exams, answer keys) or websites that post solutions to the textbook questions.

[1] Converting a Jupyter Notebook to PDF is not always straightforward (e.g., some methods don't wrap text properly). Please ensure that your PDF is legible! We will deduct points if we cannot read your PDF (even if you have the correct answers in your Notebook).

Table 1: Variable Names and Definitions

Variable	Definition
`Risk`	1 if firm found to have evaded taxes after audit; 0 otherwise
`Sector`	Historical risk score for the industry sector of a firm
`PARA A`	Discrepancy found in planned expenditure (in crore)
`Risk A`	Risk score computed from Para A and firm traits
`PARA B`	Discrepancy found in unplanned expenditure (in crore)
`Risk B`	Risk score computed from Para B and firm traits
`Money Value`	Firm revenue in past 2 years
`Risk D`	Risk score computed from some firm traits
`Score`	Comprehensive risk score
`Inherent Risk`	Firm's historical risk score
`Audit Risk`	Total discrepancy score computed from examining firm tax returns

Note: 1 crore = 10 million rupees = 130,000 USD.

1 Motivation

The capacity of the government to collect taxes is pivotal to long-run economic growth because without tax revenue, the state cannot provide public goods. One way the government can increase tax revenue is by increasing the tax rate. Another way to increase tax revenue is by reducing the probability of successful tax evasion; as probability of success decreases, the incentive to cheat gets weaker. The government can reduce the probability of successful tax evasion by simply increasing the number of audits it performs. However, increasing the number of audits also increases government expenditure, which may offset the increase in tax revenue. Another way to reduce the probability of successful tax evasion is to better target audits. That is, the government can increase the probability of catching tax evasion by reducing the number of audits performed on firms that paid their taxes and increasing the number of audits performed on firms that evaded their taxes. How might the government go about such an effort?

This is a prediction problem, so the government can approach this effort using the machine learning techniques we're covering in class! For instance, one way the government can become more adept at going after firms that were dishonest on their taxes is by using a Linear Probability Model (LPM) or k-Nearest Neighbors (KNN). Firms that evade taxes might display similar characteristics, allowing the government to predict whether a firm has evaded taxes with a low classification error rate.

The dataset you'll use for this project contains information on firms that the government of India suspected of tax evasion and subsequently the Comptroller and Auditor General (CAG) of India performed audits on. Table 1 contains the variable names and their definitions. The outcome variable is whether the auditor found that the firm evaded taxes as a result of the audit (`Risk`). The predictors include various quantitative measures about the firms.[2]

[2]The dataset has been modified slightly to facilitate the assignment (i.e., the definitions of some variables have been changed for simplicity).

2 Forest-for-the-Trees Questions[3]

1. Since it's important to use theory/intuition/common sense in concert with our data driven approaches, what factors do you suspect will affect the true, underlying model of whether or not a firm will commit tax evasion? Briefly explain.[4]

2. Assume that in addition to some combination of the predictors listed in Table 1, the interaction of two predictor variables also enters the true model. If the appropriate interaction is not explicitly included as a predictor in the fitted model, what advantage does KNN enjoy over the LPM if the interaction is indeed important to the true relationship?

3 Data Analysis Questions

3. Drop any observations with missing information. Split the sample set into a training set and a validation set. Please use a 50/50 training/validation set split. Python students should set `random_state=13`, and R students should set `set.seed(13)` when splitting the data. Use the training set to fit a linear probability model (LPM). Please use all the variables as predictors in your model. Apply the model to the validation set to predict the probability a firm cheated.

   (a) For firms with a predicted probability of tax evasion greater than 0.5, construct the confusion matrix.

   (b) For firms with a predicted probability of tax evasion greater than 0.6, construct the confusion matrix.

   (c) For each of the two thresholds, report the error rate. Which results in more accurate overall predictions?

   (d) For each of the two thresholds, what proportion of the firms predicted to evade their taxes actually evaded taxes?

   (e) Plot the ROC curve that illustrates the performance of your classifier and report the associated AUC score. Briefly explain how the ROC curve is an extension of the comparison in the previous parts of this question .

4. In measuring performance in this context, should a false negative matter as much as a false positive? Briefly explain why or why not and how changing the threshold for classifying a firm as a tax evader (as in the previous question) affects this trade-off.

5. Using the training set from the previous question, fit a KNN model with $k = 5$, then use it to predict outcomes in the validation set with a threshold of 0.5.

   (a) Construct the confusion matrix.

[3] Note that your responses in both this and the next section should be submitted and will be graded.

[4] This is an open-ended question designed to get you to think about the task at hand in general. You are not limited to the list of available predictors in answering this question.

(b) Report the error rate. How accurate are the overall predictions?

(c) What proportion of the firms predicted to evade their taxes actually evaded taxes?

6. Repeat the previous question with $k = 5$ after scaling your predictors.[5]

   (a) Construct the confusion matrix.

   (b) Report the error rate. How accurate are the overall predictions?

   (c) What proportion of the firms predicted to evade their taxes actually evaded taxes?

7. Which KNN model performs better: with or without the predictors normalized? Briefly explain how you make this determination and why you think this is the case.

8. For KNN, which k yields the lowest error rate? By 5-fold cross-validation (5FCV), find the k with the lowest classification error rate. Briefly explain.[6]

9. In the long run, what problem might arise from the nature of the sample if the government heavily uses your best KNN model to target audits? Hint: the firms in the data are all firms that were audited.

[5] There are several ways to scale or normalize variables. Scaling here refers to transforming each predictor so that the mean is 0 and the variance is 1. There is a trade-off between scaling before splitting into training and validation sets or after. The former uses more precise estimates of the population mean and variance. The latter does not bring information from the validation set into the training set (which could result in inadvertent overfitting). Given the sample size in this question, you should scale *before* splitting your data.

[6] Use the entire dataset for 5FCV, shuffle the data randomly for splitting, and set `random_state=13` (Python) or set `set.seed(12)` (R). You do not need to normalize the data for this question.