Machine Learning - Mini-Project 2

PPHA 30545 - Professor Clapp
Winter 2023

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Wednesday, February 8th.** 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 submit your answers 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 Python (*.py) or Jupyter Notebook (*.ipynb) file converted to PDF format. OR

2. As a single PDF of a Jupyter Notebook (*.ipynb) file with your your solutions and explanations written in Markdown.[1]

Regardless of how you submit your answers, be sure to make it clear what question you are answering by labeling the sections of your write up well. 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 PPHA 30535/6 and comment your code, cite any sources you consult, etc.

You are allowed to consult the textbook authors' websites, Python documentation, and websites like StackOverflow for general coding questions. 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 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

[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.

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 KNN (*k*-Nearest Neighbors). 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. Naturally, 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 Forest-for-the-Trees Questions[3]

1. (5 points) 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

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

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

whether or not a firm will commit tax evasion? Briefly explain.[4]

2. (5 points) Assume that in addition to some combination of the predictors listed in Table 1, the interaction of two independent variables also enters the true model. Without explicitly having the interaction term as a predictor in the fitted model, what advantage does KNN enjoy over the LPM if the interaction variable is indeed important to the true relationship?

3 Data Analysis Questions

3. (10 points) Use the first half of the data (the first 388 units of observations henceforth) to train your linear probability model (LPM).[5] Apply the model to the second half of the data to predict the probability a firm cheated.

   (a) For firms with a predicted probability of tax evasion greater than 0.5, what proportion of the firms evaded taxes?

   (b) For firms with a predicted probability of tax evasion greater than 0.8, what fraction of the firms evaded taxes?

   (c) Construct the confusion matrices for both.

4. (10 points) In measuring performance, should a false negative rate matter as much as a false positive rate? 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. (10 points) Using the first half of the data as training data, fit a KNN model with $k = 1$, then use it to predict outcomes in the testing data.

   (a) Construct the confusion matrix.

   (b) With firms that are predicted for tax evasion, what fraction actually evaded taxes?

   (c) What fraction of the firms that evaded taxes were predicted to have evaded taxes?

   (d) Determine whether KNN performs better with or without the attributes normalized.[6]

6. (10 points) Repeat the previous question with $k = 5$.

   (a) Construct the confusion matrix.

   (b) With firms that are predicted for tax evasion, what fraction actually evaded taxes?

   (c) What fraction of the firms that evaded taxes were predicted to have evaded taxes?

   (d) Determine whether KNN performs better with or without the attributes normalized.

[4] Note: 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.

[5] Normally, you would randomize when using a validation set approach, but then everyone's answers would be slightly different. So we're asking you do to this to facilitate grading.

[6] There are several ways to scale or normalize variables. Normalization here refers to scaling each attribute so that the variance becomes 1. The variance of each attribute can be normalized by dividing each observation of the attribute by the standard deviation (as calculated from the training set).

7. (10 points) Which KNN model performs better? Briefly explain how you make this determination and why you think this is the case.

8. (15 points) For KNN, which k yields the lowest error rate? By 5-fold cross-validation (5FCV), find the k with the lowest classification error rate. (Use the entire dataset for 5FCV, shuffle the data randomly for splitting, and set `random_state=13`.)

9. (20 points) Compare the optimal KNN from Problem 8 with the LPMs from Problem 3. Which is better? Follow these steps to answer. First, find how many firms are predicted by the optimal KNN to have committed tax fraud. Second, find the threshold q where you would "predict" the same number of firms as having committed tax fraud using the LPM.[7] Then within the pool of firms that are classified as having evaded taxes, identify which has a lower proportion of firms that did not commit tax evasion.

10. (5 points) 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.

[7] Recall, Problem 3 had thresholds of 0.5 and 0.8.