Econ 626: ML for Economists

Prediction Competition 3: Classification and Domain Knowledge

January 30, 2024

Answers are due on Thursday Feb 8, 6pm.

- Your submission must consist of two parts: CSV file and PDF file.
- **The top of first page** of PDF must include (in this order!):
  1. Anonymized name (please do NOT write your own name anywhere – I can see it on Learn).
  2. The prediction accuracy in the training set (percentage of predictions correct).
  3. **The Confusion Matrix** for the training set.
  4. Graph for Q2 as calculated from the training data.
  5. Two graphs for Q3 as calculated from the training data.
  6. Screenshot of an example from ChatGPT/GPT4 interaction.
  7. The rest of PDF must include **code** for Q1, Q2 and Q3 answers.
- The CSV file must include the following:
  - line 1: student id number (so TA can connect your predictions to your name)
  - line 2: anonymized name (for the class leaderboard)
  - line 3: Accuracy rate in the training data (a number between 0.00 and 1.00; please do not include the percentage sign! Remember: it does NOT matter how high/low this number is.
  - lines 4 through 100,003: one prediction for every observation in the test set. **Note: every prediction has to be either 1 or 0.**

  Again, the CSV file must have one prediction for every observation in the test set, and nothing else (e.g. no variable names or other headers). Hence, given that the test set has 10,000 observations, the CSV file must have $3 + 10,000$ lines (not a line less, not a line more).

Best PDF answers will be distributed to class. Students whose answer is selected will receive 10 percentage point bonus.

You can use any programming language/statistical software package.

Collaboration is encouraged but **everyone must run their own code and write up their own answers.** You are always free to use ChatGPT/GPT4/Other LLMs in any way you consider useful (to help in writing, coding, analysis, etc.)

The following introduces the data set.

There are two training data sets posted on learn on used car prices and car characteristics: “small” (100,000 observations) and “large” (500,000 observations). The former is not a subset of the latter data set. You can build predictions based one or both training data sets. The data are comma separated.

The test data without response variable have also been posted. These data have 100,000 observations.

The original data was downloaded from Kaggle. You can use any resource you find on Kaggle, but please do not try to download more data from Kaggle to help with prediction.

Some feature variables may have missing values. In the prediction competition you cannot skip observations with missing values on some features: **you must give some prediction for all 100,000 observations in the test set** (and for all observations in the training set you utilize).

Q1. [7 points] This question is a prediction competition.

Utilize either training data set to train a model that predicts the **whether car price is less than $19,500**. That is, first construct a binary variable that is 1 for cars with price below $19,500, and zero otherwise, and then use the available features to predict this constructed binary variable.

Utilize only the KNN algorithm. Please do **not** use bagging or boosting. (And please do not use other ML algorithms to transform variables; though you can create new variables and use simple “averaging techniques” to transform and normalize variables).

Utilize any features already included (such as year and mileage) or any features that can be constructed from the available numerical and text data. Obviously, you cannot construct features from the price variable. Please try use **domain knowledge** to prioritize your feature construction efforts in an efficient (not too time consuming) manner.

Accuracy of your model will be evaluated **the prediction accuracy** (percentage of predictions correct) in the test data set.

As emphasized above, you must produce a prediction for every observation in the data sets that you utilize.

Q2) [1 point] Utilizing the training data only (large and/or small), draw a graph that replicates the pattern shown in Figure 2.17. That is, show that KNN estimation has the following general pattern (1) training error decreases as 1/K (model flexibility) increases, and (2) test error first decreases and then increases as 1/K (model flexibility) increases. Report both your code and the graph.

For this exercise, it is perfectly fine to utilize a a relatively simple set of features, such as only year and mileage.

Q3) [1 point, **very challenging**] Utilizing the training data only (large and/or small), draw a figure with (1) and (2) the two types of error rates as a function of the classification threshold, and (3) the overall error rate as a function of the classification threshold. (In other words, replicate ISRL figure 4.7 for your training set.) Draw also an additional figure of the ROC curve. (In other words, replicate also ISLR figure 4.8 for your training set.

Q4) [1 point] Demonstate how ChatGPT/Other LLM can be useful in answering Q1-Q3, either in coding or in designing the approach.