Econ 424. ML for Economists
Prediction Competition 3: Classification

September 26, 2023

Answers are due on Thursday Oct 5, 6pm.

- Your submission must consist of two parts: CSV file and PDF file.
- **The top of first page** of PDF must include (1) anonymized name to be shown on class leaderboard, (2) the prediction accuracy in the training set (percentage of predictions correct), (3) **the Confusion Matrix** for the training set, (4) answer to Q2-Q3 (two figures), and (5) code for your Q1 and Q2 answers.
- The CSV file must include one prediction (either 0 or 1) 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 100,000 observations, the CSV file must have 1-0,000 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.** As always, you can utilize ChatGPT/other LLMs in any way you wish.

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 a subset of the latter data set. You can build predictions either based on the small or the large data set. 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 $18,400**. That is, first construct a binary variable that is 1 for cars with price below $18,400, 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.

Utilize any features already included (such as Year and mileage) or any features that can be constructed. Obviously, you cannot construct features from the price variable.

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 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.

Q3) [1 point, **very challenging**] Utilizing the training data only, 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.