MMA 869
Individual Prediction Competition 2: Classification, Feature Construction and Feature Importance

June 11, 2024

Submissions are due on Tuesday June 25, 5pm.

- 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 (such as "BellKor97"; please do NOT write your own name anywhere – I can see it on Course Website).
  2. The prediction accuracy in the training set (Percentage of predictions correct (number between 0.00 and 1.00; please do not include % character.).
  3. Confusion matrix in the training data.
  4. Graph for Q2.
  5. Screenshot of an example from ChatGPT/GPT4 interaction.
  6. The rest of PDF must include **code** for Q1 and Q2 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: Prediction accuracy in the training data (typically 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.
  - line 4: Name of algorithm used (this also does not influence grade; only accuracy matters).
  - lines 5 through 50,004: one prediction for every observation in the test set, **in the same order as the observations are in the "test data without response variable" file.** Each prediction must 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 50,000 observations, the CSV file must have $4 + 50,000$ 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. An LLM may be used to evaluate student answers.

You can use any programming language/statistical software package.

Collaboration is encouraged, even sharing code is okay. 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 sets.

There are two training data sets posted on Course Website on used car prices and car characteristics: "small" (100,000 observations) and "large" (500,000 observations. **You can take advantage of either or both datasets.** There are also "only6features" versions; they are not recommended but can be helpful if very little time to do the assignment. The data are comma separated.

The test data without response variable have also been posted. These data have 50,000 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 50,000 observations in the test set** (and for all observations in the training set you utilize).

Q1. [8 points] This question is a prediction competition.

Predict whether price of used car is less than 19,500 for the observations in the test set. Prediction "1" means price is equal to or less than $19,500$. Prediction "0" means price is more than $19,500$.

Important constraint:: You can use any of the following algorithms: Decision Tree, Naive Bayes, or KNN. You can try multiple approaches, but final model must include only one of these algorithms. Random forests/boosting/bagging/ensemble models/ are **not** allowed.

The data include a number of variables that can be used to construct features. Use domain knowledge to prioritize feature construction efforts to the right variables.

Performance of your model will be evaluated based on the **prediction accuracy (percentage of correct predictions) in the test set.**

Q2) [1 point] Draw a figure that demonstrates the relative importance of different variables in terms of prediction accuracy (in the training data).

Q3) [1 point] Demonstrate how GPT/Other LLM can be potentially useful in answering Q1 or Q2 either in coding or in designing the approach.