Econ 484. ML for Economists
Final Prediction Competition

December 7, 2020

Submission 1 is due on Wednesday Dec 16, 5pm. Submission 2 is due on Friday December 18, 5pm.

- Submission 1 must include the algorithms for Q1 and Q2 (and the Bonus Question, if you choose to do it).
- **THE FIRST PAGE of Submission 2** must include
  - the R^2 and MSE for both the test set and the training set for Q1,
  - the prediction accuracy (percentage of predictions correct) and the Confusion Matrix for both the test set and the training set for Q2,
  - anonymized name to be shown on class leaderboard.
- **THE REST of Submission 2** must include the code for your Q1 and Q2 answers.

Best answers to Q1 and Q2 will be distributed to class. Students whose answer is selected will receive 0.25 bonus points for each answer.

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.**

There are three training data sets posted on learn: "small" (30,000 observations), "large" (300,000 observations) and "additional text" (300,000 observations). The data are comma separated.

The test data set that will be distributed later (Dec 16, 5.01pm) consists of 30,000 observations (again there will be a file without the additional text variables, and a file with them).

The data set "small" is a subset of the data set "large". You can merge the data set "large" and the data set "additional text" based on the variable "id".

The original data set was downloaded from Kaggle. The observations in these data are built from used car listings on Craigslist.

Many variables have missing values. As discussed earlier, in the prediction competitions you cannot skip observations with missing values on some features.

You can build predictions either based on the small or the large data set. Use of the "additional text" data set is optional. The "small" and "large" files already include four variables constructed from one of the variables ("description") in the "additional text" file. These variables are (1) description length, (2) whether description mentions "owner", (3). whether description mentions "credit", and (4) whether description mentions "bad credit".

Q1) [5 points] **Prediction Competition 1**. Utilize the training datasets to train a model that predicts the **logarithm of car prices**. You are free to use any algorithm and can utilize boosting, bagging, or random forests in building your ML prediction. You can add non-linear and interaction terms in the model. The main challenge is, I believe, to **construct a set of relevant features** from the data.

Accuracy of your model will be evaluated **based on R^2 in the test datat set** (distributed later).

Please include one (no more!) picture that captures variable importance (if possible)

Please remember to produce a prediction for every observation in the data sets that you utilize. Thus in the test set, there must be 30,000 predictions, in the small training set 30,000 predictions, and in the large training set 300,000 predictions.

Please remember that here you are predicting the **logarithm** of the car price variable that is included in the data.

Q2) [5 points] **Prediction Competition 2**. Utilize the training datasets to train a model that predicts the **whether car price is less than $10,000**. This, first construct a binary variable that is 1 for cars with price below $10,000, and zero otherwise, and then use the available features to predict this constructed binary variable.

In this question too, you are free to use any algorithm.

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

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

Please again include one picture that captures variable importance (if possible).

Bonus Question) [1 point] The paper "Heuristic Thinking and Limited Attention in the Car Market" (https://www.aeaweb.org/articles?id=10.1257/aer.102.5.2206; may require campus access but free versions available on the web too) documents an interesting pattern in used car prices and provides a behavioral explanation for the pattern.

Examine whether this pattern holds for the Craigslist data. Please provide a brief explanation of any potential differences between your result and the result provided in the paper.