ECON861 Data Analysis(Spring 2022) Final Exam

Due: 8th May 2022 (11:59pm)

Name: _____

UID: _____

---

1. (70 points) Scenario 1 (part 1): Train a model of house price prediction.

   - Download the house price dataset (dataset.csv). It is a CSV file (`https://link.us1.storjshare.io/jv7newf7xy3veop6ezk4yyb3ue4a/class%2Fdataset.csv`) (approximately 1.2 MB)
   - This dataset contains the price range of 15400 houses and some characteristics of the houses.
   - Download the image files of the houses (image_files.zip). It is a zip file (`https://link.us1.storjshare.io/juoxppvnoqisi43whrypjt2gdb3a/class%2Fimage_files.zip`) (approximately 391 MB)
   - Pick a model from the models we have discussed in class. Using the house price dataset and the corresponding image files, train the model to predict the price of the houses.
   - Explain your choice of model and parameters, and explain why the model you trained is performing well.
   - You should NOT upload the dataset and the image files. Instead, you should upload the machine you trained using a pickle file, together with your python code files.

2. (30 points) Scenario 1 continued (part 2): Right after you finish your work, your colleague tells you that the dataset.csv was corrupted. The number of bedrooms for houses between row 10000 and 12000 in the dataset.csv is incorrect.

   - Your colleague suggests that you should drop all the observations from row 10000 to 12000 and re-train the model. Implement this solution and explain how does your answer change compare to what you did in part 1?
   - Can you suggest another way to fix this problem (other than dropping all the problematic rows)? Explain why your method is better than the method suggested by your colleague.

3. (? points) Scenario 2 (bonus question): In this scenario, the data for the number of bedrooms is correct. However, right after you finish your work, your colleague tells you that there is a new dataset which contains more detailed price information of the houses.

   - Download the house price dataset (detail_price_dataset.csv). (`https://link.us1.storjshare.io/ju77ji4oczzjpxuoxrau3wwh5lrq/class%2Fdetail_price_dataset.csv`) It is a CSV file (approximately 1.3 MB). This file is exactly the same as the dataset.csv in part 1 except that there is an extra column "price". The column contains the price of the house instead of just a price range.
   - Re-train the model using this new dataset. You probably should pick a new model and a new set of parameters. Explain your choice of model and parameters, and explain why the model you trained is performing well.

   Hints:

   - You do not need to use all the data. You may use only a subset of it. Some of the observations are obviously problematic, you may need to drop them before you do your analysis.

- The number of bathrooms is shown in "number with decimal point", but these are not the regular decimal numbers. In particular, "2.1" does not mean that there are two and one-tenth of a bathroom. Google for the meaning of the decimal point number of bathroom and make an appropriate transformation in your dataset before you perform your analysis.
- To improve the performance of the model, you may need to try:
  - Experimenting many different sets of parameters
  - Using only a subset of house characteristics
- You should hand in via Github. Please set your Github repository to a private repository and invite me to be your collaborator. DO NOT commit survey_dataset.csv into the repository. Use the .gitignore file to exclude the dataset from the repository.
- Do not print out your answers and hand in hardcopies. You will fail this class immediately if you do that.
- In your Github repository, there should be:
  - One (or more) python file, which contains your machine learning code
  - A pickle file which contain the machine you trained.
  - A README.md file, which contains instructions on how to train the model and how to use the trained model to make predictions. (The README.md file will automatically appear on the front page of your repository)
  - A document which contains a three pages (maximum) report. In the report, you can explain how you choose the parameters, how you subset your dataset, any limitations of your model and why the model you pick is better than other models.
- In the class, we discussed how to use the accuracy scores and other measures to evaluate the model's performance. However, do remember that it is also very important to keep your model simple. For example, your model is considered better if it uses only a small subset ofcharacteristics, but the accuracy score is just a little bit lower. You do not want to add a huge number of characteristics to your model just for a marginal increase in accuracy score. Choose the set of characteristics carefully. You may need to experiment a lot to determine which characteristics are actually useful.
- Even though the question only asks you to train one model, in order to compare its performance with other models, it would be very helpful if you use the dataset to train other models as well.
- There are a lot of ways to perform image analysis. In particular, using images to predict house prices is a very popular topic. However, for the purpose of this examination, I do expect that you can complete this exam by using the image analysis techniques that I mentioned in class.

Notice:

- The datasets used in this exam are modified datasets. The image files are obtained from Kaggle, but they are associated with a dataset which is modified for examination purposes. Therefore the house prices in this dataset are not necessarily the house prices of the actual houses shown in the image files. Do not use the dataset for research purposes.