Final Project

MGTF 495

Kaggle Deadline: June 4, 2021 11:59 PM PST

Report Due Date: June 4, 2021 11:59 PM PST

1. Instructions

The report and the code for the final should be submitted on Canvas. To secure full marks both the report and the code should be in sync and logically correct. Please only submit relevant and legible code. You will need to include your kaggle id, and score on the public leader board in the report. We will not be able to grade you otherwise. You are also required to mention the kaggle ids and names of your teammates in the report.

2. Overview

For the take-home final of MGTF 495, we combine the concepts we have learned so far and apply them to another housing price prediction problem. The problem is open-ended and you can use any method / library you like. You will find the data and the data description in the kaggle link provided below. Solutions will be graded on Kaggle. Please follow the link - https://www.kaggle.com/c/mgtf495/overview to view the webpage, you may signup using your UCSD email id (@ucsd.edu and not @eng.ucsd.edu). Note that the times reported on Kaggle are in UTC and not PST. Your grades will be determined by your performance on the regression task as well as your written report listing the approaches you took.

3. Files

train.csv - contains 2,051 house listings.
test.csv - contains 879 house listings. You will need to predict Sale Prices on this data.
sample submission.csv - Your solution file needs to be of this format to be acceptable.

You can download the files on the Data tab in kaggle. The description of the columns can also be found in the same tab. Since we are asking for your code, we will check for the originality and legitimacy of the code. You will need to cite any code or snippets referenced for this project in your report. Any unfair practices could end up earning you a zero on your final project.

4. Task

Sale Price Prediction - Students will need to predict the Sale Price on the test.csv. Recall from your HW that you can train models on a train dataset to predict values on the test set. The accuracy of your submission will be measured in terms of the root mean squared error.

The public leaderboard will show your results on half of your submitted test data, but a majority of your score will depend on the performance on the private leaderboard which will be visible to you only after the kaggle competition ends. (We advise you to not tune your models to overfit for the public leaderboard.)

5. Grading and Evaluation

You will be graded based on your ability to obtain a solution which outperforms the benchmarks on the test data (the unseen portion). You will be entitled to bonus marks if you perform substantially better than benchmarks. Following are your total marks for beating a benchmark.

Baseline: 88372
Benchmark 1: ≤ 50,000 25 marks

Benchmark 2: ≤ 40,000 30 marks

Benchmark 3: ≤ 38,000 35 marks

Benchmark 4: ≤ 35,000 40 marks

Benchmark 5: ≤ 33,000 45 marks

Benchmark 6: ≤ 30,000 50 marks

Benchmark 7: ≤ 25,000 55 marks

Bonus Benchmark 8: ≤ 22,000 60 marks Bonus

Obtain a solution which outperforms the baseline on the seen portion of the test data (i.e., the public leaderboard) to obtain 20 marks. This is a consolation prize in case you overfit to the leaderboard.

The report accounts for 30 marks. It should describe the approaches you took to perform the task. **Make sure that the methods you describe in the report include all the aspects of your final model including pre-processing, feature engineering etc. (if any).** The aim is to enable anyone with your report to be able to recreate your results. **Even if your model doesn't perform well, you can obtain marks in this section for the comprehensiveness of your analysis.** You can obtain a maximum of 100 marks + 10 bonus marks in this project, which will be scaled down to 40% of the total course assessment. To obtain good performance, you need not invent new approaches (though you are more than welcome to!).

6 Kaggle

We have set up a Kaggle page to help you evaluate your solution. You should be able to access the competition via your UCSD address(@ucsd.edu, not @eng.ucsd.edu

You can submit only 20 submissions per day to Kaggle. This is to ensure that you don't learn from the test data. Please form a validation set for measuring the performance of your model. Before the competition ends, you need to select two top submissions on which you want us to evaluate you at the end of the competition. If you do not select, your two best submission (based on public leaderboard) will be chosen for you.

7. Baselines

A simple baseline solution has been provided for the task. This is included in 'Baseline.ipynb' among the file available on Canvas. This contains the code for reading the train data, predicting on the test data and also for generating a submission file. The jupyter notebook shows a simple prediction. It always predicts the average of the SalePrice from the train.csv. Reach out to either of your TAs for any clarifications.