CSC 599.70: Course Project

Grant Long*
itds.ccny@gmail.com

The City College of New York — Fall 2019

Goal

The goal of the CSC 599.70 Course Project is to give students a chance to apply data handling and modeling skills taught in class to a real world data set. Students shall work in teams of 2 to 4 individuals to predict asking rents for and answer several modeling questions pertaining to for New York City apartments posted on StreetEasy, an online marketplace for New York City homes. Predictions will be judged on the mean squared error of their estimated rents for the provided test sets.

Data

The data sets for the project come from a random selection of homes posted for rent on StreetEasy during the summer of 2018. A training set with a sample of 12,000 homes posted in May, June, and July of 2018, along with their respective asking rents and several details pertaining to their listing on StreetEasy, including publicly posted bedroom count, bathroom count, descriptions, and select building and unit amenities.

Students are required to generate predictions on a random set of listings posted on StreetEasy during August 2018. One full set, including observed rents, is provided with the project posting. Students are required to submit predicted rents on two additional sets, including test2 and test3, which do not include the observed rents.

Students are expected to attach at least one additional data set to the set provided. The data set includes several data points designed to facilitate attaching additional third party data sets to the StreetEasy data set. Examples of these include the street address, latitude and longitude, and New York City BIN and BBL numbers. Additional data could come from the U.S. Census Bureau, New York City open data, the NYC Geoclient or any number of other open sources.

Deliverables

The project should include the following deliverables:

- **Due October 14, 2019, 11:59pm:**
  (1) *Optional*: An emailed list of desired teammates. Students are expected to work in teams of 2-4 individuals. The professor has discretion to set teams as he sees fits, but will do his best to honor requests. Final teams will be announced in class on October 16, 2019.
- **Due November 23, 2019, 11:59pm:**
  (1) An email to the professor with:

*This project rubric borrows from a previous iteration of the same class taught by Dr. Michael Grossberg as well as a similar project designed by Harvard University's CS 109.

(i) A creative team name, the names of all members of the team, and a link to the team's repository on GitHub. The repository need not be public, but access must be granted to the professor (grantmlong).

(ii) An attached csv with predictions against test2.csv. The csv must consist exclusively of 2 columns, header rows with the titles *rental_id* and *predictions*, and the 2,000 rental ids and corresponding rent predictions. A example of the required formatting can be found here, with suggested methodology for creating the submission file here.

(2) A markdown file posted in the proejct Github repo entitled *initial_findings.md* containing:

(i) A 200-300 word explanation of the expected performance of the model in terms of mean squared error and the key features driving the team's modeling performance.

(ii) A 200-300 word summary outlining the team's intended strategy to improve the predictions for the final round.

- **Due December 7, 2019, 11:59pm:**

(1) An email to the professor with:

(i) The team name, the names of all members of the team, and a link to the team's repository on GitHub.

(ii) An attached csv with predictions against test3.csv. The csv must consist exclusively of 2 columns, header rows with the titles *rental_id* and *predictions*, and the 2,000 rental ids and corresponding rent predictions. A example of the required formatting can be found here, with suggested methodology for creating the submission file here.

(2) A markdown file posted in the proejct Github repo entitled *project_findings.md* containing:

(i) A markdown file entitled *project_findings.md* containing answers and supporting evidence for all of points in the *Questions and Tasks* section that follows.

(ii) A Jupyter notebook allowing for the complete replication of the modeling process.

- **Due December 9, 2019, 6:30pm:**

(1) Peer reviews from each team member individually emailed to the professor. For more on the peer reviews, see the following section.

Grading

The project makes up 30 percent of the grade for *Intro to Data Science*. Each student's final grade will consist of both team and individual assessments broken into components as follows.

- **Individual Contribution to Team Effort.** While the project is a team effort, each member is expected to pull his or her own weight. While only one grade will be assigned for each project's final work material, each student will receive their final grade for the project weighted by their contributions toward the team's efforts. Each workscore will be assessed based upon:

1 **Evidence of work on GitHub.** Each member of the team is expected to actively contribute to the teams coding by actively and regularly commiting code, logging issues, and commenting on GitHub. Note that I will not just be counting commits, but rather reviewing their progression over time to see which team members are making meaningful contributions. As a result, if you want to receive credit for your work, *you must commit it to GitHub*.

2 **Peer Assessment.** At the end of the semester each team member will be asked to grade their fellow teammates contributions via written assessment.

- **Team Performance.** Each team will be scored as follows:

(1) Accuracy of predictions for *test2.csv*. (10 points)

(i) Achieving a mean squared error of less than 4,226,362. (4 points)

(ii) Achieving a mean squared error in ranking in the top 50 percent of the class. (2 points)

(iii) Achieving a mean squared error in ranking in the top 25 percent of the class. (2 points)

(iv) Achieving a mean squared error in ranking in the top 10 percent of the class. (2 points)

(2) Preliminary evidence of progress toward the project goal in *initial_findings.md*. (10 points)
   (i) Do you show evidence of making meaningful progess beyond the modeling steps made in lecture 5?
   (ii) Have you applied techniques and strategies demonstrated over the course of the semester?
   (iii) Do you have a well thought designed strategy for improving your model performance before the final due date?

(3) Accuracy of predictions for *test3.csv*. (20 points)
   (i) Achieving a mean squared error of less than 4,226,362. (8 points)
   (ii) Achieving a mean squared error in ranking in the top 50 percent of the class. (4 points)
   (iii) Achieving a mean squared error in ranking in the top 25 percent of the class. (4 points)
   (iv) Achieving a mean squared error in ranking in the top 10 percent of the class. (4 points)

(4) Submission of well-documented and reproducible code. (20 points)
   (i) Is code provided, and can it reproduce the entire work?
   (ii) Is additional data included or at least linked (externally) with instructions on how to access it?
   (iii) Do you use comments and markdown cells to explain every step of your code, similar to the course data dives?
   (iv) Does the code demonstrate considerable work given the number of people on the project?

(5) Completion of each of the points outlined in the *Questions and Tasks* section. The grade for this section will include the degree to which each answer reflects evidence of creativity, effort, and expertise gained over the course of the semester. (100 points)

Note: no late work will be accepted. In order to receive full credit for all pieces of the project, work must be submitted on time.

Questions and Tasks

(1) Data Usage.
   (a) What outside data have you appended to the original data set? Why did you choose this data?
   (b) Does the inclusion of this additional data raise any ethical considerations?

(2) Data Exploration.
   (a) What outliers present issues for your analysis? How have you chosen to handle them? Why?
   (b) To what extent do missing values pose a challenge for your analysis? How have you chosen to handle them? Why?
   (c) Are there any other aspects of the data your exploration shows might be problematic?
   (d) Create at least one visualization that demonstrates the predictive power of your data.

(3) Transformation and Modeling.
   (a) Describe 5 features you think play the biggest role in your model.
      - How did you create these features?
      - How do you know these features are playing key roles?

      If your modeling process uses less than five features, explain why you think other features didn't add value.
   (b) Describe how you are implementing your model. Why do you think this works well?
   (c) Describe your methodology for selecting your model. Why do you think this type of model works well?

(4) Metrics, Validation, and Evaluation.

(a) How well do you think you model will perform on the hold out test set? How do you know?

(b) Is your model useful? Why or why not?

(c) Are there any special cases in which your model works particularly well or particularly poorly?

(d) Create at least one visualization that demonstrates the predictive power of your model.

(5) Conclusion

(a) How would you use this model?

(b) If you could have additional modeling features, what would they be?

(c) Would you rather have more data, or more features?

Words of Wisdom

This exercise is intended to be an opportunity for you to demonstrate the skills students have acquired over the course of the semester, but it is also intended to be fun. My hope is that predictive and competive elements of the project will make the exercise both focused and fun, but do note that generating the best predictions in the class will yield a minimal amount of points relative to the full project score. I will reward effort and creativity accordingly, and students would be well advised to focus on developing a sound modeling strategy, not just focus on the final metrics. In this spirit, students are also advised to be sure to submit well documented code to prove that their modeling strategies are the reason for their model performance, and not random chance or untoward reasoning.