CS210 - Introduction to Data Science
Spring 2020

Project Description

The purpose of the project is to increase your knowledge about data science and get hands-on practical experience. You will work in groups of 2. You are expected to assemble your teams and let us know the team members by ~~22 March 2020~~. Feel free to write to the discussion board if you are looking for teammates.

All groups will be working on the same dataset, New York City Airbnb Open Data. The dataset contains registered Airbnb hosts and their associated attributes, such as location, price, # of reviews and availability. Instead of a rigid project structure, you will ask questions to get insights from data and answer them quantitatively using the tools and techniques in this class. You may define a prediction task, discover structure in the data, and/or ask questions and answer with statistical hypothesis tests, visualizations. It is up to your creativity and curiosity. You may download the dataset from this link.

In addition to the provided dataset, you are highly encouraged to include supplementary datasets that may help you extract further insights and increase the prediction performance. Weather, traffic, taxi trips and activities that attract masses are some examples that you may think of. In order to find such datasets, you may begin with NYC Open Data, a data sharing platform by New York City agencies.

Grading: All deliverables contribute to the project grading, we will be conducting random interviews with some of the project groups. The interview will ensure if you have done the project without any external help. The interviews will be announced and conducted in the week of May 4. The Interview will be graded in base-2 numeral system.

Grading Algorithm: Grade = (Proposal*0.1+Progress*0.4+Final*0.5)*Interview

The deliverables (ALL DATES ARE FINAL) are:

(i) M1: a half page proposal outlining your ideas of what you are planning to do and your teammates - 28 March 2020 (%10)

(ii) M2: a progress report- 17 April 2020 (%40)

(iii) M3: a final report - 5 May 2020 (%50)

Proposal: At most a half-page, preferably a pdf, proposal write up. It should include:

(1) group members

(2) additional datasets to be utilized

(3) a high-level description of the problem you are trying to solve

Progress Report: The progress stage will significantly affect the final grade of the project. You are expected to have results and exploratory visuals by the progress date. You will submit a notebook that contains your code, results, visuals and markdown comments. It must include:

(1) a high-quality introduction on what you are doing, why you are doing it

(2) a clear description of the datasets you have used

(3) methods you have used, data preprocessing, feature generation, and the machine learning models you have tried

(4) the next steps you are planning to take

Final Project Report: A final write up of the project. Again, you will submit a notebook. The final notebook should should include

(1) Introduction: A summary of the problem, methods, and results.

(2) Problem description: Detailed description of the problem. What question are you trying to address?

(3) Methods: Description of methods and datasets used.

(4) Results: The results of applying the methods to the data set. Include the list of questions your experiments are designed to answer. Details of the experiments; observations

(5) Discussion: Interpretation and discussion of the results.

(6) Conclusions: What is the answer to the question?

(7) Mention any future directions of interest.