CSC 599.70: Course Project

Grant Long*
itds.ccny@gmail.com

The City College of New York— October 1, 2018

Project Goal

The goal of the CSC 599.70 Course Project is to give students the opportunity to explore and analyze one or more data sets of their choosing with the goal of telling a compelling narrative using data. Through the project, students should :

- Apply Python to load, clean, and process data sets.
- Identify key patterns in a data set using computational analysis and statistical methods.
- Apply principles of statistical modeling and machine learning to data.
- Effectively explain, visualize, and communicate empirical findings.
- Demonstrate effective team collaboration.

The final product for the project should include data visualization, modeling – through regression, classification, or unsupervised learning – and a written and verbal presentation of findings that conveys the broader importance of the findings beyond the analytical exercise. In other words, students should be prepared not only to demonstrate their expertise and creativity in analyzing data, but also *convince an external audience that their findings are important*.

Project Deliverables

The project should include the following deliverables:

- **Due October 19 :** A project proposal that includes:
  (1) A brief paragraph identifying the major theme the project aims to cover.
  (2) A list of at least three potential questions the analysis will address.
  (3) Two or more data sets that address those questions.
  (4) A list of independent and dependent and variables to be included in the modeling portion of the project.
  (5) A 200-400 word explanation explaining why the questions and data are relevant to the broader theme the project proposes to address.
  (6) A team or project name with a link to the project repo on GitHub.
- **Due November 5:** A preliminary project update including:
  (1) A brief section addressing feedback from the project proposal.
  (2) A 400 to 600 word summary of findings from the data exploration.
  (3) At least two charts from exploratory data analysis.

*This project rubric borrows heavily from on from a previous iteration of the same class taught by Dr. Michael Grossberg as well as a similar project designed by Harvard University's CS109.

- **Due November 25:** A second project update including:
  (1) A brief section addressing feedback from first project update.
  (2) Two additional charts.
  (3) A summary of and performance metrics from the modeling phase of the project.
- **Due December 10:**
  (1) A write up of 1,200 to 1,600 words and at least 2 data visualizations on Medium. Students can either create a Medium account for their team or alternatively post the write up on one more team members individual accounts. Those who do not have an account can create one for free.
  (2) A fully documented repository of code on GitHub.
  (3) An 8-10 minute class presentation with slides and visuals.
  (4) A short project review from each team member summarizing each teammates contributions to the group effort and lessons learned through the project. Further details on the project review will be released later in the semester.

Teams

Team 1	Angelica	David	Excel	Sam
Team 2	Bashir	Shofi	Michelle	Thierno
Team 3	Carlos	Victor	Emma	Kenny
Team 4	James	Jane	Pooneet	Tahsin
Team 5	Oleks	Hanna	Vitaly	Shaf
Team 6	Boris	Chris	Shuchuan	Trace

Grading

The project makes up 40 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 received their final grade for the project based on his or her 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.

The team portion of the project grade will follow the project rubric below. Some questions below may not be relavant to some projects but all parts I-VIII are required.

I. Data Science Questions 20 pts
  (a) Is the background context for the question stated clearly (with references)?
  (b) Is the hypothesis/problem stated clearly ("The What")
  (c) Is it clear why the problems are important? Is it clear why anyone would care? ("The Why")
  (d) Is it clear why the data chosen should be able to answer the question being asked?
  (e) How new, creative, and significant are your problems? Do you go beyond checking the easy and obvious?

II. Item Data Cleaning/Checking/Data Exploration: 20pts

(a) Did you check for outliers?

(b) Did you check the units of all data points to make sure they are in the right range?

(c) Did you identify the missing data code?

(d) Did you reformat the data properly with each instance/observation in a row, and each variable in a column?

(e) Did you keep track of all parameters and units?

(f) Do you have specific code for reformating the data that does not require information not documented (eg. magic numbers)?

(g) Did you plot univariate and multivariate summaries of the data including histograms, density plots, boxplots?

(h) Did you consider correlations between variables (scatterplots)?

III. Transformation and Modeling: 20pts

(a) Did you transform, normalize, filter the data appropriately to solve your problem? Did you divide by max-min, or the sum, root-square-sum, or did you z-score the data? Did you justify what you did?

(b) Did you pick an appropriate set of models to solve the problem? Did you justify why these models and not others?

(c) Are you appropriately callibrating your model? For example, can you just

IV. Metrics, Validation and Evaluation 20pts

(a) Are you using appropriate choice of metrics? Are they well justified? If you are doing classification do you show an ROC curve? If you are doing regression are you justifying the metric least squares vs. mean absolute error? Do you show both?

(b) Do you validate your choices of hyperparameters? For example, if you use KNN or K-means do you use cross validataion to optimize your choice of parameters?

(c) Do you separately evaluate testing and training error? Do you estimate the uncertainty in each? Are you also making sure that your validatation for hyperparameter optimization is kept separate from your testing?

(d) Have you avoided overfitting?

V. Visualization 20pts

(a) Do you provide visualization summaries for all your data and features?

(b) Do you use the correct visualization type, eg. bar graphs for categorical data, scatter plots for numerical data, etc?

(c) Are your axes properly labeled?

(d) Do you use color properly?

(e) Do you use opacity and dot size so that scatterplots with lots of data points are not just a mass of uninterpretable dots?

(f) Do you write captions explaining what a reader should conclude from each figure (not just saying what it is but what it tells you)?

VI. Code 20pts

(a) Is code provided, and can it reproduce the entire work?

(b) Is the data included or at least linked (externally) with instructions on how to download it?

(c) Do you use markdown cells to explain every step of your code similar to homeworks and some example notebooks?

(d) Does the code demonstrate considerable work given the number of people on the project?

VII. Write Up and Presentation 20pts

(a) Do you tell a coherent story with a beginning, middle and end?

(b) Do you have good clear visuals with axis and data labeled?

(c) Are your slides relevant to your story and solving your problem or are they only vaguely relevant "padding"? Is each slide justifed in your narration?

VIII. Timely and Meaningful Submission of Intermediate Products 20pts

(a) Does your project proposal demonstrate thoughtful review of available data?

(b) Does your first project update adequately demonstrate meaningful progress in exploring the data?

(c) Does your second project update adequately demonstrate meaningful toward addressing modeling questions?