CIS 400: Course Project

Instructions: This is a group project. You will work in teams of 3-4 students. There are multiple due dates corresponding to various activities. As a final document you need to upload a project report as a PDF file, screenshot of your Kaggle score, your final predictions as a CSV file, as well as a PDF version of the python code you used including the outputs generated. As in the assignments, make sure the reports are readable so that we can grade without having to look into the code.

- **Team formation**: By Wednesday November 1 11:59pm send an email with the names of all students in your team to me ngautam@syr.edu, Cc the TA (Minmin Yang: myang47@syr.edu), as well as ALL the team members. Only ONE member of the team needs to send an email but they MUST copy the others. Teams should have 3 to 4 students each.
- **Problem Description**: There is an ongoing Kaggle competition `https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview` from where you can read up about the problem and download data. Kaggle has provided training data and testing data. Only the training data has the dependent variable of `'SalePrice'`.
- **Data**: To access the data on Kaggle, you need to register (just one person needs to). Alternatively you can go to Blackboard where we have uploaded the data. This includes the training and testing data. Also, a sample submission file, and a description of the columns are provided.
- **Process**: First just use the training data to create a python code and divide that into a train and test. This way you can compare all your models. You can choose the right variables, code appropriately, and also decide what columns to use non-linear. Once you like a model or a combined model, use the test data to make predictions. Upload the predictions and obtain a score from Kaggle. You are only allowed a few submissions per day. Keep a screenshot of your best score.
- **Deadlines and Expectations**: All are 11:59pm on
  1. Wednesday November 1, 2023: Email ngautam@syr.edu (with Cc to team and TA) team information (names of team members)
  2. Wednesday November 8, 2023: Email ngautam@syr.edu (with Cc to team and TA) project status information (in one paragraph explain what you have done so far and what you have observed)
  3. Wednesday November 15, 2023: Email ngautam@syr.edu (with Cc to team and TA) initial results in terms of the accuracy of the various models using training data
  4. Wednesday November 29, 2023: Email ngautam@syr.edu (with Cc to team and TA) your best score according to Kaggle (include screenshot as proof)
  5. Wednesday December 6, 2023: Upload on blackboard a project report (also describing what was done in the project), final predictions, Kaggle score screenshot, and PDF file of the code with output. **Every student must submit all four files on Blackboard so we can assign a grade. It is not enough if one member of a team uploads it.**

Out of the 25 points for this assignment, each of the November deadlines are worth 1 point (0.5 points for late submission until the next deadline, and 0 points after that). Mark your calendars and instructions to be sure expectations and deadlines are not missed.

- **Methodology**: While we do not want to prescribe a method, our only requirement is that it is **not** be based on neural networks or deep learning. Any approach using regression that we did in class, as well as any statistical analysis would be reasonable choices.

- **Report**: Your report must be well written so that you can use it to show what you did for the class, say to a prospective employer. You must explain in some detail so the graders can understand what you did without looking into the code. Also, you must cite sources of data and make references to any information you used (such as journal articles, if any). Even though each team member submits a separate report, the reports for students in a team can be identical.

- **Grading**: Grading will be based on the creativity used in the approach taken and the justification provided in the report. Any obvious errors in terms of the code will be penalized. While it is important to get accurate predictions (based on Kaggle), that will not be the main focus. However, as an incentive, *we will take the winning team to lunch.*