Final Project

Requirements

- Students work in a team with up to **4** students
  - You can do the project individually, but the expectation for the project delivery will not be reduced.
- Each group participates in any one of the Kaggle competitions
  - The Competitions filtered by Getting Started (such as Spaceship Titanic)
  - The Competitions filtered by Playground (such as San Francisco Crime Classification.
  - Or any other competitions
    * Please mainly focus on the tabular data analysis.
    * The competitions without specific evaluation metrics are NOT allowed, such as NFL Big Data Bowl 2022.
- You must try several different methods to solve the problem. The method you finally choose for the competition does not have to be a model taught in this course.
- You must compare your fancy methods with simple baselines, e.g., random guess, all-positive, all-negative, simple linear models, and beat the baselines. The evaluation metric must be the one required by the competition.

Milestone and Final Delivery

1. **Project information**: You will have your team and have registered for a Kaggle competition. Submit your team and competition at Canvas.
2. **Project checkpoint 1** (15%): For this milestone, you will need to have downloaded the data, and also run some initial pre-processing on it. You should also make **at least one** dummy submission (all-positive, all-negative, most-frequency) on Kaggle and submit the Jupyter Notebook including the code and a screenshot of the score on Canvas.
3. **Project checkpoint 2** (15%): For this milestone, you should do more data preprocessing and have made **at least two** non-dummy submissions on Kaggle. You should submit an updated Jupyter Notebook.
4. **Presentation** (35%): Give a presentation and show your explanation and results by the presentation date.
5. **Final results** (30%): By this time, you should have made **at least six** non-dummy submissions with different settings (models, preprocessing, hyper-parameters) on Kaggle totally. You should submit the Jupyter Notebook as the final report (including implementation, documentation,

explanation…). In the Jupyter notebook, you should clearly explain what experiments (data preprocessing, feature engineering, machine learning models) you have conducted.

Presentation

We will host the presentation sessions (35%) on Nov 29 - Dec 8. The following is a suggested structure for the presentation. You don't necessarily have to organize your presentation using these sections in this order, but that would likely be a good starting point for most projects.

- Overview: Briefly describe the competition your group is tackling.
- Describe the overall objective.
- Introduce the dataset as well as the exploratory data analysis.
- Data Preprocessing and Feature Engineering
- Machine learning models including the dummy and non-dummy approaches you have tried.
- Evaluation results.
- Conclusion and future work Students from Charleston will give the presentation online.

Key Dates

- Oct. 23: Have a group and choose a competition.
- Nov. 4: Checkpoint 1
- Nov. 25: Checkpoint 2
- Nov. 29 – Dec. 8: Project presentation
- Dec. 9: Submit final results (No late submission)

Tips

1. Start from a simple task, a simple model, and simple hyper-parameters.
2. A good prediction performance is desired, but the workflow of machine learning is more important.
3. This video is a tour for Kaggle.
4. You can directly write your code at Kaggle and submit results. Please see the tutorials Getting Started on Kaggle: Python coding in Kernels and Getting Started on Kaggle: Writing code to analyze a dataset.
5. You can also download the dataset and implement machine learning models offline. Then, you submit the prediction only. See the tutorial My First Kaggle Submission.
6. There is a tutorial for starting a new Kaggle project: Beginner Kaggle Data Science Project Walk-Through (Titanic).