Final Project
DS160-01
Introduction to Data Science
Spring 2022

Exploring Machine Learning Models (100 points)

Goal: This project has two goals: 1) for you to apply the techniques you learned this semester to prepare and explore a dataset and 2) for you to implement a machine learning (ML) model for which there is a package in Python or R.

Instructions: This assignment will result in three deliverables:

1. Paper describing your analysis (template provided).
2. The well commented/documented notebook or R file containing your analysis code.
3. An 8 – 10 minute presentation of your analysis to the class (We will do these on final exam day)

Below are the specifications for this project.

1. For this project, you may choose your partner (teams of two only).

2. Once you have chosen a partner, **send an email to me. I will respond with a link to the data set that your team has been assigned**.

3. With your partner do a thorough **exploratory analysis** of the data set you have been assigned using techniques we discussed in class and your previous homework project. Use the **final_project_teamplate** document to guide you for what you need to include in your paper. Your notebook/ R script should contain all of your exploration. The paper will only include what you found unusual. You may include additional tables and plots in an appendix if you want.

4. Your data set is likely not in perfect shape to immediately run models. Perform whatever steps or transformations necessary to get the data set into a format that you can use with your machine learning algorithms.

5. Choose two of the following models. You may pick from the following ML models:
   a. Softmax Regression (Multiclass Logistic Regression)
   b. K-Nearest Neighbor
   c. Decision Tree
   d. Random Forest
   e. Naïve Bayes
   f. Support Vector Machine

6. Perform a **complete implementation of your model** including data preparation, experimental design (running the model with several different versions of the data), (see the **final_project_template** for further instructions). You will also provide information about the tools you used for your analysis.

7. Document the output of the experiments with different versions of your model, (see **final_project_template** for further instructions).

8. Write a conclusion that summarizes your work, (see template for further instructions).

9. Proofread and follow the specifications for the paper. I will be ruthless evaluating papers that do not follow the template.

10. Prepare an 8 – 10 minute presentation of your work that you will deliver on final exam day. Both team members must present part of the work. Other than that, you have free rein to design the presentation as you

want (e.g. PPT or not, live code, show PDF of paper, etc.) If you use PPT, I have posted a PPT template for you to use.

11. Push all materials to a repository on GitHub called **DS160_Final_Project_XXX_YYY** where **XXX** and **YYY** are your team member initials.. Includes your notebook/R files, your paper in Word format, **your data set**, presentation files (PPT, handouts, etc.) Ensure that your README.md **provides a summary of your project**. A good option would be to put your abstract in the README.md file, although you may do it a different way if you want.

This assignment must be completed in a teams of two. Please send me an email as soon as possible with your partner's name, so I know who is working together.

Project Submission: Upload a link to your GitHub repository for the project in the area provided in Moodle by the deadline specified.

Evaluation Criteria

Here are some of the aspects of this project I will be looking at for evaluation

1. Evidence the that you worked on the project over time (e.g. multiple commits, good commit messages).
2. Your paper follows the template as closely as possible (both in content and formatting).
3. Your README file contains a summary of your project.
4. Your notebook or R script contains good (and adequate commenting) so I can follow it when I review it.
5. That you demonstrated thoughtful consideration of your experimental design.
6. That you demonstrate thoughtful consideration of your results in your paper.
7. That you performed a thorough and complete exploratory analysis of the data set and provided an adequate summary of it in the appropriate section your paper.
8. That you adequately explain your project during the class presentation. This includes background of the data set, and your complete analysis from beginning to the end.
9. That both team members participate equally in the presentation and are able to answer questions as appropriate.