Stats 131 Final Project Requirements and Guidelines

Purpose:

The purpose of the project is to give students some real-life experience using Python to analyze a data set. The project encourages students first to explore and think about the data before attempting to fit a predictive model. The project will also challenge students to clearly communicate to a general audience.

Grading:

The total project will be worth 40% of the final course grade.

The project will be graded out of 400 points.

The project has two parts:

1. A written technical report – 200 points
2. A presentation for a general audience consisting of slides and a video – 200 points

The Written Report

The report will be written as a Jupyter notebook. Explanations will be provided using Markdown and all supporting code and diagrams will be included.

The audience of the report is someone with technical and theoretic knowledge of data analysis, and machine learning. The assumptions made should be clearly explained.

The report will have three sections:

1. Context and description of the data – 50 points
2. Exploratory Data Analysis – 90 points
3. Data modeling – 60 points

Context and description of the data

In this section of the report, students will provide information about the data set.

They should include:

1. Background information on the subject and field of study.
    a. Eg: “The data explores measurements of three species of iris flowers. Iris is a plant genus with an estimated 260-300 species. … more …
2. Information about data collection.
    a. Who collected the data
    b. When the data was collected
    c. How the data was collected
    d. Any implications this may have on analysis

e. Eg: Dr. Edgar Anderson collected the data for a paper published in 1935. Two of the three species were collected in the Gaspe Pensinsula all from the same pasture, and picked on the same day, and measured at the same time by the same person with the same apparatus. Being a convenience sample, the data cannot be used to …

Exploratory Analysis of the Data

This portion of the report is weighted the most heavily as it is the most important. The report should do a thorough exploration of potential relationships within the data.

If data needs to be 'cleaned up,' it should be performed in this part.

The exploratory analysis should investigate the features present in the data, including, but not limited to:

1. Summary statistics and the distributional shape of variables in the data
2. Unusual features or outliers present in the data
3. Potential relationships that may exist in the data, including, but not limited to:
    a. two-way tables and side-by-side bar charts for relationships between categorical data
    b. scatter plots for relationships between numeric data
    c. side-by-side histograms or boxplots for relationships between numeric and categorical data
4. Findings should be reported with readable tables or clearly labeled graphs.
5. There must also be text to explain the findings and the included tables.

The exploratory data analysis should be guided by a series of guiding questions or curiosities. Each question need not uncover a significant relationship, but should reflect a reasoned approach.

> For example, we might be curious if there is a difference between the weights of male and female babies, and we may find that there is a difference. We may further explore to see if there is a difference between the weights of babies for different ethnicities or races and may find that there is not a significant difference. Both sets of findings should include tables, graphs, and commentary.

Data Modeling

In this section, students will fit a statistical model to the data for the purpose of insight and/or prediction.

Students are allowed to fit any model they choose, ranging from simple models like linear regression to more complex ones like random forests, or neural networks.

Students should explain their decisions regarding the choice of model, and if appropriate, the reasoning behind the inclusion of predictive features. (Probably backed up by the findings in the exploratory data analysis.)

If students fit a model to gain insight to the data, then explanations of the findings are necessary. Students should explain the relationship between variables when possible. For example, explanations of linear regression coefficients need to be included, or an explanation of what significance the principal components have.

If students fit a model for predictive purposes, then care should be taken to separate training and testing data. Students should perform some form of cross-validation to ensure that the model is not overfitting features unique to the training data. A metric will need to be selected to show the predictive performance of the model.

Slide and Video presentation:

Students should imagine they are making a slide presentation for the management team in an imaginary company interested in the chosen data.

While the target audience of the written report is someone with technical and theoretic knowledge, the target audience of the slide presentation and video is someone who has general knowledge and is familiar with the data. This portion of the project will evaluate students' ability to communicate to a general audience.

Students can assume the audience has general math understanding, but lack deep knowledge of theory or programming. Students should not include code, or math equations.

Students can assume the audience is familiar with the data, and do not need to spend time or slides explaining the different variables in the data.

Students should create a short slide presentation to summarize their findings – between 2 to 5 slides of content (plus a title slide). Students should not try to summarize the entire report. Students will need to choose on which portions of the report to focus and summarize those findings.

Slides will be graded on:

1. Clarity (will a person with general but not technical knowledge understand what the slide is communicating)
2. Concision (does the slide avoid unnecessary information)
3. Content (does the slide presentation accurately summarize the important content of the larger report)

Video presentation

The video presentation goes together with the slide presentation. As such, the video presentation is for a general audience.

Videos will be graded on clarity, concision, content, as well as professionalism. Videos should be about 5 minutes long. Videos longer than 6.5 minutes will not be accepted.

Professionalism simply means that the student(s) speaking in the video should appear to have practiced the presentation. Students should not be stumbling over their words or struggle in explaining the content. The video should have no (or extremely little) background noise or distracting background elements.

Videos do not need to have "high production value," as that is not part of the grading criteria. You do not need animations or multi-camera edits. A video shot with a phone of a student going through the slide presentation next to a computer screen is perfectly adequate. Of course, students are welcome to edit their video and add whatever flair they wish, but that will not factor into the final grade.