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 – 250 points
2. A presentation for a general audience consisting of slides and a video – 150 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. Background information, including context and description of the data – 75 points
2. Exploratory Data Analysis – 100 points
3. Data modeling – 75 points

Data Selection

You can use almost any dataset that you can find.

The minimum number of observations in the dataset will be 40. An ideal number is probably a couple hundred to a couple thousand. A link to the data must be included in the report, or the data must be uploaded to the project's github repository.

You are required to pick data from a subject about which someone in the group is knowledgeable or at the very least highly interested in. For example, if you choose to analyze the wine dataset in Kaggle, someone in the group should be knowledgeable about wine. For example, this person should be able to explain the difference between a Merlot and Cabernet.

Background information, including context and description of the data

In this section of the report, students will provide background information about the subject being studied. This is where you will demonstrate your knowledge and interest in the subject matter.

This section of the report should cover general information about the subject, so that someone who is unfamiliar with the topic will understand what the data covers. If there is 'common knowledge' that is known for people familiar with the subject, that should be explained and detailed in this section as well.

There should be a thorough discussion of each variable included in the dataset along with an explanation of what an observational unit is.

The section should provide a review of other relevant studies that have been done and summarize their findings.

This section of the report is very important and is weighted to be equally valuable as the Data Modeling section (75 points). A thorough background section cannot be completed in just a few paragraphs.

I will not perform a word count, but this section should probably be at least 700 words. I anticipate most will be over 1000 words.

For example:

1) Your dataset covers information about baseball players. Your background section should include information like:
    a. How baseball is played: teams take turns playing offense (batting) and defense (fielding)
    b. Baseball players are generally divided into two groups: batters and pitchers
        i. For batters, the players' ability to hit the ball is the key driver in a players' value to the team. There are two types of positions batters play: infield vs outfield.
        ii. For pitchers, the players' ability to prevent the other team from hitting the ball is key to the players' value to the team. There are two types of pitchers: starting pitchers and relief pitchers.
    c. A discussion of each of the variables included in the dataset, and what the observational unit is
        i. Explain what a hit is, how it differs from base-on-balls, etc.
        ii. Explain the difference between batting-average, and on-base-percentage, slugging, etc.
        iii. Each row is an observational unit: One row for each player for each year. Thus a player who has played for 6 years (seasons) will have 6 rows in the dataset.
    d. Discussion of how the data is collected:
        i. Baseball data is recorded officially at every game.
        ii. The data was compiled by …
    e. "Common knowledge" shared among baseball fans. For example:
        i. There are 30 teams
        ii. A batting average that is over 0.300 is considered good.
        iii. A pitcher with an ERA of 3 or under is good.
        iv. Average age of players is in their late twenties.
    f. A review of other baseball research or studies that have been done.

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 stats knowledge, the target audience of the slide presentation and video is someone who has a general knowledge background. 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 loosely familiar with the data, and do not need to spend more than one slide explaining the different variables in the data.

Students should create a short slide presentation to summarize their findings – approximately 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. The video should be uploaded to Youtube (either a public or unlisted video) and a link to the video must be included in the report.

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, and the speech must be clear.

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. Slides in the video must be large enough so that all text is legible.

Of course, students are welcome to edit their video and add whatever flair they wish, but that will not factor into the final grade.