Project 3
STAT 207 - Data Science Exploration

Due: Tuesday, April 23 by 11:59 pm on GitHub

Main Goal of Analysis

The main goal of this project is to use a dataset to fit models. We will use these models to make predictions and to understand and quantify relationships between variables. A secondary goal will be to build a classifier to the data.

You are required to perform two main analytical tasks:

1. Linear Regression Analytical Task
2. Logistic Regression Analytical Task

Additional descriptions for these tasks can be found later in this document. If you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is required in this document.

Project Format

Your project will be submitted as a written report to GitHub. This report is worth 40 points. Additional details for the report can be found below.

Group Structure

You can work in groups of up to 3 people, or you can work individually.

- If you work with a group of 3, you must do at least 25% of the work in order to get full credit.
- If you work with a group of 2, you must do at least 33% of the work in order to get full credit.

Dataset Options

You can choose your own dataset, or you can choose from one of the three supplied datasets below. There are several places you can go to find interesting datasets, but here are some places to start:

- https://www.kaggle.com/datasets
- https://corgis-edu.github.io/corgis/csv
- https://archive.ics.uci.edu/ml/datasets.php
- https://github.com/fivethirtyeight/data
- For sports data, you may choose to explore (nflfastr.com for NFL, billpetti.github.io/baseballr for MLB, cfbfastr.sportsdataverse.org/index.html for CFB, and sportsdataverse.org for more sports data).

Choosing your own data:

If you choose your own data, it must meet the following specifications.

1. It must have at least three variables total (five for later projects)
    a. Variables that have uninformative information don't count and won't be useful. Examples of uninformative variables include those that provide the observational unit (a row name or row id) or are a linear combination of other variables in the datasets. If you aren't sure, come ask!
2. It must have at least two categorical variable
    a. For at least one of these categorical variables, you will use only two levels; that is, only two of the possible options that it contains. You can filter your data at the time to accommodate this need
3. It must have at least one quantitative variable
4. It must have at least 50 rows

You may continue using the same dataset from Project 1, or you could change to a different dataset.

Provided dataset:

1. Video Games Data (video_games.csv)
    a. This dataset has information about the sales and playtime of over a thousand video games released between 2004 and 2010. The playtime information was collected from crowd-sourced data on "How Long to Beat"
    b. This was originally collected and curated by Dr. Joe Cox
    c. This data was originally downloaded on 2/8/2024 from here: https://researchportal.port.ac.uk/en/datasets/video-games-dataset
    d. Read more about this data here: https://researchportal.port.ac.uk/en/publications/what-makes-a-blockbuster-video-game-an-empirical-analysis-of-us-s

Project Report Specifications

Deadline: Tuesday, April 23 by 11:59 pm on GitHub
Format:

- Jupyter notebook
- This should be a clean data analysis report that you could submit to an employer or client (not a homework assignment). At the very least, your report should have a title, headings for each section, and be written in paragraphs and with complete sentences.
- You can use and modify the attached project_03_template.ipynb file as a template for this report if you'd like. You can add and delete as many cells as you'd like in this file.

1. **Introduction [6 points]**
   Goal: In your introduction, you should orient the reader to what they are about to read. This will help to prepare the reader, so that they can figure out how to connect the different components that they will read.
   While working towards this goal, you should complete/address the following:
   a. Title: Give your research report a title
   b. Dataset introduction: You should briefly introduce the reader to the context of your dataset and the available data.
   c. Motivation: Describe the motivation for why you or someone else would want to explore your dataset (or a dataset of this type). Describe one person or the type of person who would find these insights useful. You can give background research (with citations) if this would help back up your motivation.
   d. Research Questions: You will answer two **sets** of research questions. In your introduction, you should briefly describe your research questions and why you (or someone else) would be interested in answering these research questions. How could the answers to these research questions be used?

2. **Linear Regression Analytical Task [20 points]**
   Goal: For your linear regression analytical task, you should pick a quantitative response variable and at least four explanatory variables that you suspect might affect your response variable. Explore whether there is a linear relationship between the explanatory variables and the response variable.
   While working towards this goal, you should complete/address the following:
   a. State your research questions: Your research question should have two parts. For example, you might ask "what is the relationship between x1 and y after controlling for x2, x3, and x4 (both in the sample and in the underlying population)? How does my model perform on new data?"
   b. First, split your data into a training and testing set. Be sure that this split is reproducible (can be replicated by another). You may pick what proportion of observations are in the training and what proportion are in the testing sets.
   c. Fit a linear regression model to the training data. Include the following:
      i. Show the summary output for your linear regression.
      ii. Write out the linear regression equation for your model. Use appropriate notation. (Note: if you use a categorical explanatory variable that has a

large number of levels, you may pick a specific level and write the model for that specific level. Be sure to clearly state what specific fitted model you are writing.)

d. Examine the relationship between your selected x and y in the context of this model. Include the following:
    i. Interpret your selected slope coefficient (not the intercept) for your linear model (descriptive analytics), describing the relationship between your two variables of interest.
    ii. Provide and interpret a 95% confidence interval for the slope coefficient, and in doing so perform inference for the underlying population.
    iii. Check all linear regression conditions for your model. Do the conditions support that your answers for the two questions above (descriptive analytics and inference) are valid?
e. Evaluate your model performance. Include the following:
    i. Discuss what percent of variability in your response variable is explained by this model in the training data. Is this high? Is this low?
    ii. Select one observation in your test data. Make a prediction with your model on this observation, and calculate its residual.
    iii. Calculate the RMSE on your training data and your testing data. Interpret the RMSE for the testing data. Is this high? Is this low?
f. Finally, discuss how your approach and conclusions address your research questions.

3. Logistic Regression Analytical Task [16 points]

Goal: For your logistic regression analytical task, you should pick (or make) a categorical response variable with two levels and at least three explanatory variables that you suspect might affect your response variable. Explore a logistic regression model and a classifier based on this model.

While working towards this goal, you should complete/address the following:

a. State your research question: For instance, you could ask "Is there a linear relationship between the log-odds of the success level of y and x1, x2, and x3 in the sample? How does a classifier built on this model perform on new data?"
b. Fit a logistic regression model to the data. Include the following:
    i. Show the summary output for your model.
    ii. For each explanatory variable, indicate what characteristics (lower or higher values) are associated with higher probability of success.
    iii. Report two measures of model strength for your logistic regression model: the pseudo-R^2 and the AUC. What do these measures indicate about the strength of your model?
c. Build a classifier:
    i. Define the type of error that you might want to minimize, both the specific term and in the context of your data.
    ii. Use an ROC curve on the training data to pick a good predictive probability threshold. Be sure to report your predictive probability

threshold. Explain how you picked your predictive probability threshold, given your research goals.

d. Evaluate your classifier:
    i. Use the predictive probability threshold to classify your data. What is the accuracy rate, sensitivity, and specificity of your classification?
    ii. Are you satisfied with the performance of your current classifier?
e. Finally, discuss how your approach and conclusion address your research question

4. **Conclusion [6 points]**
    a. Summarization: Summarize your confidence interval and hypothesis test tasks in the conclusion. Provide at least a paragraph. (This will likely be a restatement of what you have already included in your report).
    b. Limitations: What limitations did you face in your analysis, results, or interpretations? What challenges did you face in your data analysis? What contextual information is important before you make strong claims from these results? How might these affect how the person you described in the introductory motivation paragraph uses these results?
    c. Future work: If you (or someone else) were to conduct future work based on these analyses, what kind of research questions or analyses might that entail?

The remaining 2 points of this project will be graded on writing quality, clarity, conciseness, and professional and neat formatting of the report.

Intended Audience/Reader of your Project

The intended audience of your report/presentations should be someone who has the same level statistical/python knowledge as you and your STAT 207 classmates. **Theoretically, you should be able to send your report to one of your classmates (who is not on your team), and they should be able to understand everything that you did and the claims that you are making.**

Grading

In addition to being graded for correctness and completion (as noted), this project will be graded on a qualitative basis. Qualitatively, we will be looking for the following things:

- **Clarity about Analyses, Algorithms, and Data Choices**
    - Someone who has taken a STAT207-level class should be able to read through your report and easily be able to do the following:
        - Replicate what you did in your analyses.
        - Know why you made the choices that you did in your analyses.
- **Clarity about Motivation (i.e. the "so what?") of your analyses**
    - Beginning of the Report

        - Someone who is **about to** read the body of your report should be able to clearly answer the questions:
            - Why should I (or someone else) care about the report that I am about to read?
            - What research questions do they intend to answer?
            - How do these research questions relate to their motivation?
        - Therefore, in the introduction of your report you should make this clear.
    - Middle of the Report:
        - While **in the middle of** your report, your audience should be able to clearly answer the question:
            - How do each of these analyses/algorithms/data choices that they're making/using tie back into the overarching motivation of this whole analysis?
        - Therefore, each new analysis/model/algorithm/data choice that you make, you should explain this and make it clear to your audience.
    - End of the Report:
        - Someone who has **just finished** reading your report should be able to clearly answer the questions:
            - Why should I (or someone else) care about the analysis that I just read?
            - Did their analyses and conclusions answer the research questions that they stated at the beginning of the report? If so, how? What were the answers to these research questions?
            - How would the results/answers to these research questions be useful to someone?
        - Therefore, in the conclusion of your report you should make this clear.