

## **Project 3**

### **STAT 207 - Data Science Exploration**

Due: Tuesday, April 29 by 11:59 pm on GitHub

#### **Main Goal of Analysis**

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

You are required to perform two main analytical tasks:

1. Understand the relationship between variables using a linear regression model
2. Evaluate a classifier built on a logistic regression model based on the data context

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 35 points. Additional details for the report can be found below.

#### **Group Structure**

You must work in a group of either 2 or 3 students, or your group should all be enrolled in the same lab section.

- 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 use the supplied dataset 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 be smaller than 25 MB (25 megabytes, or 25,000 kilobytes). If your file is measured in kilobytes, it should be small enough for the purposes of our projects. This file size restriction is to ensure that you can push your file to GitHub. You can stop by office hours if you find a larger dataset, and we can help you take a random sample of the data to fulfill this data size requirement.
2. It must have at least five variables total
  - a. Variables that have uninformative information don't count and won't be useful. Examples of uninformative variables include those that provide the unit of observation (a row name or row id) or are a linear combination (sum, product) of other variables in the datasets. If you aren't sure, come ask!
3. It must have at least two categorical variables
  - a. For at least one of these categorical variables, you will adjust it to be a logical variable; that is, you will only use two levels. There are many ways to create a logical variable, including:
    - i. The variable only having two values or levels possible by default
    - ii. Filtering your data to include only two levels of the categorical variable
    - iii. Exploring one level vs. all other levels of the categorical variable
    - iv. Combining similar levels of the categorical variable, resulting in two levels. For example, you could make "year in school" into two groups: "underclass" for first- and second-year students and "upperclass" for third-year or older students.
4. It must have at least three quantitative variables
5. 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>
  - e. Note: while reading in this data to Python, you will need to use the argument encoding = 'unicode-escape' with code like df = pd.read\_csv('video\_games.csv', encoding = 'unicode-escape')

## Project Report Specifications

Deadline: Tuesday, April 29 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 [4 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: In a couple of sentences, you should briefly introduce the reader to the context of your dataset and the available data.
- c. Research Questions: You will answer two **sets** of research questions. In your introduction, you should state your research questions. These should be the same research questions stated at the beginning of section 2 and at the beginning of section 3.
- d. Response Variables: Describe what your two response variables will be for your two research questions, respectively.
- e. Contextual Importance: You should describe why you (or someone else) would be interested in the answer to these research questions. For example, how could the answer be used? You can use creativity and imagination when describing a situation where this answer would be helpful.

### 2. Linear Regression Analytical Task [13 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. You will 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  $x_1$  and  $y$  after controlling for  $x_2$ ,  $x_3$ , and  $x_4$ , 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 test 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 test sets.
- c. Fit a linear regression model to the training data. Include the following:
  - i. Show the coefficient values or 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 slope coefficient for your  $x$  variable of interest from your research question, describing the relationship between your two variables of interest (descriptive analytics).
  - ii. Provide a 95% confidence interval for the slope coefficient, and in doing so perform inference for the slope in the underlying population.
  - iii. Create 2-3 graphical or numerical summaries that allow us to check the linear regression conditions for your model. Do these summaries support that the conditions are met and 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. Calculate the RMSE on your test data. Interpret the RMSE for the test data. Is this high? Is this low?

### 3. Logistic Regression Analytical Task [12 points]

Goal: For your logistic regression analytical task, you should pick (or make) a logical 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 “How do  $x_1$ ,  $x_2$ , and  $x_3$  relate to the log-odds of the success level of  $y$  in the sample training data? How does a classifier built on this model perform on new data?”
- b. Fit a logistic regression model to the same training data from the Linear Regression Task. Include the following:
  - i. Show the coefficients or summary output for your model.
  - ii. For each explanatory variable, indicate what characteristics (lower or higher values for quantitative predictors or a specific level for categorical predictors) 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 and evaluate a classifier:
  - i. Build a classifier to the test data with a predictive probability threshold of your choosing. You might consider 0.5 as a first option. (Note: you do

not need to search systematically to find a well-performing threshold; selecting one threshold is sufficient).

- ii. Calculate the accuracy rate, sensitivity, and specificity of your classifier.
- iii. Define the type of error that you might want to minimize, both the specific term and in the context of your data. Note: this question allows for some creativity in setting up the situation.
- iv. Are you satisfied with the performance of your current classifier, especially in consideration of the type of error that you might want to minimize?

#### **4. Conclusion [4 points]**

- a. Summarization: Summarize your linear regression and logistic regression tasks in the conclusion. Provide about 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 introduction 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.