STAT207 – Data Science Exploration – Mini-Project #2 – 30 Points

Due: Friday, March 29 11:59pm CST on Canvas *(you can submit 2 days late penalty free if you need to, but not encouraged).*

This is an individual project. To receive full credit, you should follow the steps and answer the questions given in this document for your project.

Text that appears in the oranges boxes below is new and unique to this particular project #2. The rest was discussed and requested in project #1.

Primary Research Goal of Analysis: [Prediction]

The primary research goal of this project is to **build a predictive model** that will perform the best when predicting your chosen response variable for *new datasets.*

Secondary Research Goal of Analysis: [Interpretation]

Ideally, we would like for our chosen model to also **yield reliable interpretative insights** about the nature of the relationship between the variables in the dataset.

Qualitative Assessment of Report

In addition to being graded for correctness and completion, 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 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, without looking at the code!*
    - *Know why you made the choices that you did in your analyses.*
- ***Clarity about Motivation (ie. the "so what?") of your Analyses***
  - *Beginning of the Report:*
    - *Someone who is **about to** read your report and watch your presentation should be able to clearly answer the questions.*
      - *"Why should I (or someone else) care about the report that I am about to read/listen to?"*
      - *"What research questions do they intend to answer?"*
      - *"How do these research questions relate to their motivation?"*
    - *Therefore, in the introduction of your report and presentation you should make this clear.*
  - *Middle of the Report:*
    - *While **in the middle of** your report and presentation, 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, for 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 and watching your presentation should be able to clearly answer the questions:*
        - *"Why should I (or someone else) care about the analysis that I just read/listened to?"*
        - *"Did their analyses and conclusions answer the research questions that they stated at the beginning of the report/presentation? 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 and presentation you should make this clear.*

- ***Professionalism***
    - *Your report and findings should be well-explained and written in **paragraphs** and **complete sentences** and in the **markdown cells (not in code blocks or in comments).***
    - *Do not just spit out code and expect your reader to automatically know:*
        - *Why you chose to use this code, what its purpose is, what you're doing in the code block, and what you want them to notice in the result.*
        - *Why the output of your code is important.*
        - *How your code answers any relevant questions.*
    - *Any paragraphs, sentences, and explanations that you write should be considered satisfactory to, say, your high school writing teacher.*

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 STAT207 classmates. **Theoretically, you should be able to send your report to one of your classmates and they should be able to understand everything that you did and the claims that you are making.***

Project Format

Project Report [25 pt]

***Deadline*: Friday, March 29 11:59pm CST on Canvas. (Can submit 2 days late if you need to. Not encouraged)**

Should contain*:** *Everything stipulated in the* ***Project Report Specifications *discussed below.*

***Format*:**

- *Jupyter notebook.*
- *This should look like a clean data analysis report that you would theoretically submit to an employer (not a homework assignment). Thus, at the very least, your report should have:*
  - *a title*
  - *headings for each of your sections*
  - *You should write paragraphs and in complete sentences.*
- *You can use and modify the attached project* ***Mini_Project_2_YOURNAMEHERE.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.*

***Graded*:**

- *See "Project Report Specifications" section below for point breakdown.*

Peer Evaluation [5 pts]

***Deadline*: Friday, April 5 11:59pm CST on Canvas.**

- ***Purpose***:
  - *For report writers:*
    - *The purpose of this final part of the project* ***report writers*** *is to give constructive feedback on:*
      - *how* ***clearly*** *you were able to communicate and answer your research questions with your analyses*
      - *how well you were able to* ***motivate*** *your research to a peer, and*
      - *how* ***reproducible*** *your analysis was.*
  - *For readers:*
    - *The purpose of this final part of the project* ***for report readers*** *is to* ***get ideas*** *as to how to make your own report delivery better.*
- ***Steps***:
  - *Ater you submit your report, you will be randomly assigned to read another person's report.*
  - *After reading their report you will fill out a survey form on* ***Canvas****, which will ask you the following questions (see last page of this document).*
  - *The person that you evaluated be able to see the constructive feedback and your summarization.*
  - *If you are unclear about how to answer the questions in this document, you are encouraged to reach out to the person that you were assigned to for clarification.*
- ***Graded***:
  - *For completeness*

Dataset Options

You can choose your own dataset or you can use the supplied dataset discussed in the next page. *The csv for this dataset is located in the same folder that this document is in. There is more information about each of these datasets below.*

There are several places you can go to to find interesting datasets, but here are some places you can start.

https://www.kaggle.com/datasets

https://corgis-edu.github.io/corgis/csv/

https://archive.ics.uci.edu/ml/datasets.php

https://data.world/datasets/regression

https://github.com/fivethirtyeight/data

For students interested in sports data:

- *NFL: https://www.nflfastr.com/*
- *MLB and other baseball: https://billpetti.github.io/baseballr/*
- *CFB: https://saiemgilani.github.io/cfbfastR/index.html*
- *More sports stuff: https://sportsdataverse.org/*

Choosing your Own Dataset

If you decide to choose your own dataset, it must meet the following specifications.

1. It must have **at least 50 rows.**
2. It must have at least **6 meaningful variables**. That is, six variables that are either categorical or numerical.
3. You will build predictive models that will predict a **numerical response variable** with **at least 5 explanatory variables.**
    a. Thus, **at least one of your variables should be numerical**.
    b. Your explanatory variables can be categorical or numerical.
4. Your categorical variables should not be something with a distinct value for every row (like name/userid/etc).

Pre-Selected Dataset Option

1. ***Video Games Dataset*** *Originally collected by Dr. Joe Cox, 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".*
   Some more information can be found here.

   This is where Dr. Ellison downloaded this csv file from on 9/8/2023.

Project Report #2 Specifications

Your report should include the analyses, code, and explanations detailed in each of the following sections. **These specifications are completely new for project #2. You should read over the whole thing.**

1. Introduction You should write an introduction (about a paragraph) for your report. Your introduction paragraph should include/incorporate the following things.	
Professionalism * Paragraph, written in complete sentences. * Written in a markdown cell, not a code cell.	0.5
Research Goal Statement * Clearly state the primary research goal that you are pursuing. That is: "Build a predictive model that will effectively predict INSERT_NUMERICAL_RESPONSE_VARIABLE for new datasets". * You should consider **at least 5 explanatory variables** * Clearly state your secondary research goal (see the beginning of this document).	0.5
Research Motivation * Clearly state the motivation for why someone might want to build a predictive model that predicts YOUR PARTICULAR NUMERICAL RESPONSE VARIABLE for NEW DATASETS? * Describe at least one person (or type of person) who may find your predictive model useful and how they might use it.	1
2. Dataset Discussion You should write a paragraph in your report discussing your dataset(s) that you will be using to answer these research questions. This paragraph should include/incorporate the following things.	
Professionalism *Written in complete sentences. * Written in a markdown cell, not a code cell.	0.5
Dataset Display * Read your csv file and display the first 5 rows of your dataframe. * How many rows are in your dataframe (originally before any data cleaning)?	0.25
Dataset Source * State where YOU got this csv file (dataset) from. * Provide a link/reference to where it came from. * State when you downloaded this csv file.	0.5
3. Dataset Cleaning You should show and discuss any dataset cleaning decisions that you made in this section.	
Professionalism * Your written discussion in this section should be written in complete sentences. * Written in a markdown cell, not a code cell.	0.5

General Data Cleaning * When experimenting with your datasets and models in your "scratchsheet" jupyter notebook you may have determined that your dataset should be cleaned in various ways in order to more effectively pursue your research goals (before splitting it into a training and test dataset). * Discuss and show any data cleaning steps taken in this section. * Be sure to discuss WHY you choose to perform each step of your data cleaning and how this might impact the results of your analysis.	1.5
4. Preliminary Analysis These analyses may help your linear regression models achieve a better fit in the next section.	
Variable Transformations * Show the pairsplot for every pair of numerical variables in your cleaned dataset. * Show a fitted values vs. residuals plot for the linear regression model that predicts your response variable given ALL of your 5+ explanatory variables that you intend to explore. * Do you have any reason to believe that some of your linear regression models may achieve a better fit (of either the training or the test dataset) if you were to first transform one or more of your variables in your CLEANED dataset? Explain. * Create this transformed variable(s) in your dataset and refit your "full model" that uses this transformed variable(s). * Reevaluate the linearity assumption of this transformed model. Did transforming these variables help your linearity assumption become closer to being met?	1.5
Interaction Terms * For every (numerical explanatory variable x1, categorical explanatory variable x2) pair, create a scatterplot with x=x1, y=YOUR RESPONSE VARIABLE, and color code by x2. Create a best fit line for every distinct value of x2. * Do any of these plots suggest that there is a large interaction between how a given x1 and x2 impact your predicted response variable? Explain.	1.5
5. Predictive Models You should build and test your linear regression models in this section. Your linear regression model should predict your chosen numerical response variable. You should have **at least 5 explanatory variables** that you are considering	
Professionalism * Your written discussion in this section should be written in complete sentences. * Written in a markdown cell, not a code cell.	0.5
Important Note about Creating Indicator Variables for Regularization * IF your **dataset has categorical explanatory variables** AND you plan to use **regularization**, then you will need to create your own 0/1 indicator explanatory variables here FIRST (before the train-test-split). You can do so with the code below. `df_with_ind = pd.get_dummies(df, drop_first=True)`	
Train-Test Split * For this project, we'll use the train-test split method to help us select our best predictive model. * So create a training and test dataset from your cleaned dataset.	0.5

Scaling * Ideally, we'd like to TRY to interpret the magnitude of our slopes as representing how important the explanatory variable is when predicting the response variable. * So z-score scale your TRAINING dataset NUMERICAL VARIABLES using the TRAINING dataset column means and standard deviations. * And z-score scale your TEST dataset NUMERICAL VARIABLES using the TRAINING dataset column means and standard deviations.	1
5.1. Non-Regularized Linear Regression Full Model	
* Fit a non-regularized linear regression model that uses ALL of the 5+ explanatory variables that you intend to explore. * If you decided to transform any of your variables, make sure you use these transformed variables. * (You may also decide to ADDITIONALLY explore fitting a model with the untransformed variables) to see if this gives you a better test R^2. This is optional for this project, but would be good practice to test.) * Calculate the test dataset R^2 for this model.	1.25
5.2. Non-Regularized Linear Regression Full Model with Interaction Terms	
* Fit a non-regularized linear regression model that uses ALL of the 5+ explanatory variables that you intend to explore AND any interaction terms that correspond to the strong interactions that you identified in section 4. * Calculate the test dataset R^2 for this model.	1.5
5.3. Feature Selection You can choose to use one of three feature selection algorithms (or you can do all three if you'd like).	
Feature Selection OPTION A: Backwards Elimination Algorithm	
* Because we have decided to split our dataset into a training and test dataset, we can use a backwards elimination algorithm to TRY to find the linear regression model with the highest possible TEST R^2 (instead of the adjusted R^2 which we have explored in the lectures). The template of the algorithm works the same way. The only difference is you are using TEST R^2 instead of the adjusted R^2. * Conduct this backwards elimination algorithm with your 5+ explanatory variables that you are considering. * Show the test R^2 for your final model.	4.5
Feature Selection OPTION B: Forward Selection Algorithm	
* Because we have decided to split our dataset into a training and test dataset, we can use a forward selection algorithm to TRY to find the linear regression model with the highest possible TEST R^2 (instead of the adjusted R^2 which we have explored in the lectures). The template of the algorithm works the same way. The only difference is you are using TEST R^2 instead of the adjusted R^2. * Conduct this forward selection algorithm with your 5+ explanatory variables that you are considering. * Show the test R^2 for your final model.	4.5
Feature Selection OPTION C: Regularization	

* Select at least one type of regularization model (ie. LASSO, ridge regression, or elastic net). *(Note that a complete analysis would try all 3, but this is optional for this report).* * In this section you should train MANY regularization models using MANY different values of lambda with your training dataset. And you should evaluate each of these models by calculating the corresponding test R^2. * Your goal here is to try to find the lambda value in your chosen regularization model which will yield the HIGHEST possible test R^2. * Try out at least 100 evenly spaced lambda values within a certain range and examine how the test R^2 changes (see individual lab assignment 8) in a line plot. * If you don't see a "peak" test R^2 value in your line plot, then keep broadening your range of lambda values until you do see a peak. However, it is possible that your "peak" is at lambda = 0. It's also possible that the true test R^2 peak happens in the "gap" between two of your lambda values that you tried out. * Show the highest test R^2 that you found and the lambda value that corresponds to it.	4.5
6. Best Model Discussion You'll discuss the "best model" that you found here (ie. the one with the highest test R^2).	
Professionalism * Your written discussion in this section should be a paragraph written in complete sentences. * Written in a markdown cell, not a code cell.	0.5
Equation * Write out the equation for your model that had the highest test R^2. * If you used a regularization model that had some of your slopes "zeroed out", then you can leave these corresponding explanatory variables out. * Make sure to use the appropriate notation discussed in class.	0.75
Test Dataset Fit * How *good* is the overall fit of this "best" model for the test dataset?	0.75
Overfitting Explanatory Variables * Does the fact that this is your "best model" (after performing feature selection) suggest that some of your original 5+ explanatory variables were overfitting the model? If so, which ones?	0.75
Multicollinearity * Do the remaining explanatory variables (or explanatory variables with non-zero slopes) in this model exhibit an issue with multicollinearity? Explain.	0.75
Slope Interpretations * Are you able to interpret the magnitudes of the slopes as indicating how important the corresponding explanatory variable are when it comes to predicting your response variable in a linear regression model? Explain. * If you are able to, state explanatory variables are the most important based on the slope magnitudes.	0.75
7. Conclusion You should answer your research question B here. "How does the Relationship between `x` and `y` Change based on Different Values of `z` in the Dataset?"	
Professionalism * Your written discussion in this section should be a paragraph written in complete sentences. * Written in a markdown cell, not a code cell.	0.5
Recommendation * Would you recommend your best model to be used by the person that you mentioned in your motivation? Why or why not?	0.75

Shortcomings/Caveats * Do you know FOR SURE that your chosen best model will yield the HIGHEST possible test R^2 out of all possible models that you could make with this dataset? * What are some other techniques and steps that a more "complete" analysis would have also tried in search of a model with the highest test R^2? * Discuss any other shortcomings to your analysis here (all analyses have SOME shortcomings).	1.5
Future Work * Based on what you observed in your analysis, what is one idea you might have for future work?	0.5
8. Peer Evaluation Feedback After submitting your report, you will be randomly assigned to another student (and vice versa) to provide feedback on their report.	
See the Canvas quiz questions.	5
Total	**30**

STAT207 Peer Evaluation Questions [5 points]

Deadline: **Friday, April 5 11:59pm CST** on **Canvas**.

These questions will be posted on a Canvas quiz for you to submit.

1. What was their **research goal** and to what extent were they able to meet this research goal?

2. What is the **motivation** for the research goal that this person pursued in this report? How would the results of this analysis be **useful** to someone?

3. How easy would it be to **reproduce** this person's entire analysis on your own in Python *without looking at their code*? If it was not extremely easy (or straightforward), what was not straightforward about it?

4. Name at least two **steps/decisions/interpretations** that this person made in their report in which you could envision another data scientist doing something different. Why do you think that this other data scientist might have done something different?

5. **Shortcomings:** Name at least two other things that could have been done in this analysis to more thoroughly and effectively pursue this research goal? Try to come up with something that they have not already mentioned.