STAT207 – Data Science Exploration – Final Group Project – 150 Points

Due: Wednesday, December 6 11:59pm CST on Canvas.

Group Work

- You should work in groups of 3-4.
- There is a penalty for working in a group with less than 3 people.
- As an individual, you must do at least 25% of the work in your team to get full credit.

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 final project. The rest was discussed and requested in mini-project #1 and #2.

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 **categorical response variable (with two levels)** in *new datasets.*

2. 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.

3. Secondary Research Goal of Analysis: [Descriptive Analytics]

You should also thoroughly **describe** the nature of the variables as well as the **relationship** between the variables that you would like to use in your model. These descriptive analytics techniques that you use should be interpreted in the context of your primary research goal.

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 [5 components]

Project Report [100 pt]

Deadline: Wednesday, December 6 11:59pm CST on Canvas.

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 **Final_Project_YOURNAMESHERE.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.

Project Presentation [25 pt]

Presentation Date:

- During your final lab section time **Wednesday December 6 in-person.**
- If you foresee an issue in presenting on 12/6 let me know asap!

Format:

- **Your presentation should be no more than 9 minutes.**
- You must present some part of the presentation in order to get full presentation credit.
- Presentation should be presented in **slides** (not the Jupyter notebook).

Graded:

- See attached **presentation rubric** for what you should present and how you will be graded.

Report Peer Evaluation [10 pts]

Deadline: Friday, December 13 11:59pm CST on Canvas.

- **Steps**:
  - You will be randomly assigned to **read** another group's report *(as an individual).*
  - After reading their report you will fill out a survey form on **Canvas**, which will ask you the following questions (see last pages of this document).
- **Graded**:
  - For completeness

Presentation Peer Evaluation [10 pts]

***Deadline*: Friday, December 13 11:59pm CST on Canvas.**

- **Steps**:
  - You will be randomly assigned to **watch** another group's presentation in your lab *(as an individual).* It will not be the same group that you read the report for.
  - After watching their presentation, you will fill out a survey form on **Canvas**, which will ask you the following questions (see last pages of this document).
- **Graded**:
  - For completeness

Individual Research Impact Questions [5 pts]

***Deadline*: Friday, December 13 11:59pm CST on Canvas.**

- **Steps**:
  - **As an individual, you must do at least 25% of the work in your team to get full credit.**
  - You will be asked a few questions about the work that you *individually* contributed to your group.
  - You should have an understanding as to how your individual contributions influenced and were influenced by the insights and decisions made by your teammates. (See questions in last pages of this document)
- **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 **categorical response variable (with 2 distinct values)** with **at least 5 explanatory variables.**
    a. Your **response variable** should be **categorical (2 distinct values).**
    b. You should have at least one **numerical explanatory variable.**
    c. You should have at least one **categorical explanatory variable.**
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.

Final Project Report Specifications

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

1. Introduction You should write an introduction (1-2 paragraphs) for your report. Your introduction should include/incorporate the following things.	**General Point Breakdown**
Report Professionalism * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1.5
Research Introduction and Motivation * ~~Clearly state the motivation for why someone might want to build a predictive model that predicts YOUR PARTICULAR BINARY 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.~~ * ~~Do you think this person would desire a classifier that that was better at classifying the "positives" or better at classifying the "negatives" of your response variable? Or would they prefer equally high accuracy for both "positives" and "negatives"? Explain.~~ * ~~You should use at least TWO CITATIONS that support your motivation/answers in this section.~~ * ~~Make sure that your citations are referenced and cited appropriately in this document.~~	6
Research Goal Statement * ~~Clearly state the primary research goal that you are pursuing. That is: "Build a predictive model that will effectively predict INSERT_BINARY_RESPONSE_VARIABLE for new datasets".~~ * ~~You should consider **at least 5 explanatory variables (one should be categorical, one should be numerical)**~~ * ~~Clearly state your secondary research goals (see the beginning of this document).~~	2
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.	
Report Professionalism * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1.5
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)?~~	1.5

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.~~	1.5
Original Dataset Information In the place where you found this dataset, try to answer the following questions. If the source does not give the answer to these questions, say so. * ~~What do the rows (ie. observations) represent in this dataset?~~ * ~~How was this dataset collected?~~ * ~~Is this dataset inclusive of ALL possible types of observations that could have been considered in this dataset? If not, what types of observations might be left out?~~ * ~~How does your answer to the question above impact the types of actions that the person in your research motivation might take based on the answer to your research questions?~~	3
Selected Variables * ~~Briefly describe the variable you intend to use as your response variable~~. * ~~Briefly describe the 5+ explanatory variables you intend to use in your model~~. * ~~Why did you choose to focus on these 5+ explanatory variables?~~	1.5
3. Dataset Cleaning You should show and discuss any dataset cleaning decisions that you made in this section.	
Report Professionalism * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1.5
Missing Value Detection and Cleaning * ~~Does your dataset have any IMPLICIT missing values? Demonstrate in the code whether it does or does not~~. * ~~If it does have IMPLICIT missing values, what strings represent these missing values?~~ * ~~Deal with these missing values using one of the techniques we discussed in class. If you dropped rows, how many rows did you drop?~~ * ~~Evaluate the pros and cons of using this missing values cleaning technique that you just used~~.	2
Sample Size Cleaning * ~~If your categorical explanatory variable(s) has some levels that only have a few observations, you may consider dropping all rows that correspond to these particular levels. Situations like this can lead to overfitting.~~	2.5
Outlier Cleaning - Two Variable Outlier Inspection *For every pair of numerical explanatory variables that you're using, create a scatterplot. * If you detect any outliers in these scatterplots, evaluate the pros and cons of dropping these outliers, when it comes to pursuing your research goals. * If you chose to drop the outliers, do so. * If you dropped rows, how many did you drop?	3

Other Data Cleaning * When you tried to answer your research questions below, did you discover any other data cleaning ideas that might help make the answer to your research question more clear? What were they? Why did you choose to perform this additional data cleaning? * If there are, do so here. * If you dropped rows, how many did you drop?	1.5

4. Preliminary Analysis

Report Professionalism * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1.5
Relationships between the Response Variable and the Explanatory Variables * Visualize the relationship between your explanatory variable and the response variable with the appropriate plot. Do this for every explanatory variable. * Which explanatory variables have strong relationships with the response variable? * Which explanatory variables have weak relationships with the response variable?	4
Relationships between Explanatory Variable Pairs * Visualize the relationship between each pair of your explanatory variables with the appropriate plot. * Are there any pairs of explanatory variables that have strong associations with each other?	4
Interaction Effects * For every (numerical explanatory variable, categorical explanatory variable) pair, determine if there is an interaction between how these two explanatory variables impact the predicted response variable. * Use the appropriate visualizations and models that we discussed in Unit 14 to show this.	4

5. Model Data Preprocessing

We will be using the scikit-learn LogisticRegression() function to build our logistic regression models, because we will be using cross-validation techniques to select our best model in section 6.

0/1 Response Variable * Make sure you have created your 0/1 response variable if you haven't done so already.	1.5
Features Matrix and Target Array * Make sure to create a features matrix and target array.	1.5
Explantory Variable Scaling * Because of our secondary research goal which focuses on model interpretability, you should scale your numerical explanatory variables.	1.5
Indicator Variables * Be sure to translate your categorical explanatory variable(s) into indicator variables.	1.5

6. Feature Selection with k-Fold Cross-Validation

In order to make more robust estimates as to how a given model might perform when classifying observations in NEW datasets, we will use k=5 fold cross-validation to assess a given models performance (rather than the train-test-split method). Specifically, we will measure the **average test AUC** of a given model.

Thus, we would like to select a logistic regression model that has the highest average test AUC in the k=5 fold cross-validation.

You can choose from one of three feature selection techniques below which will attempt to find this logistic regression model with the highest average test AUC.

OPTION A: Backwards Elimination with Cross-Validation

Perform a backwards elimination algorithm that tries to select the logistic regression model with the **highest** average test AUC in the k=5 fold cross-validation.	10

OPTION B: Forward Selection with Cross-Validation

Perform a forward selection algorithm that tries to select the logistic regression model with the **highest** average test AUC in the k=5 fold cross-validation.	10

OPTION C: Regularization with Cross-Validation

1. Select at least one type of logistic regularization model (ie. LASSO, ridge regression, elastic net) that you would like to use. 2. Discuss WHY you chose this particular type of regularization model (if you only looked at one like LASSO). 3. In this section you should train MANY regularization models using MANY different values of lambda with your features matrix and target array. And you should evaluate each of these models by calculating the average test AUC in the cross-validation. 4. Your goal here is to try to find the lambda value in your chosen regularization model which will yield the HIGHEST possible average test AUC in the cross-validation. 5. Try out at least 100 evenly spaced lambda values within a certain range and examine how the average test AUC changes (see individual lab assignment 8) in a line plot. 6. If you don't see a "peak" average test AUC 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 average test AUC peak happens in the "gap" between two of your lambda values that you tried out. 7. Show the highest average test AUC that you found and the lambda value that corresponds to it.	10

7. Best Model Discussion

You'll discuss the "best model" that you found from #6 here.

<u>Report Professionalism</u> * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1.5
<u>Train-Test-Split</u> * Randomly split your dataset into a training dataset and a test dataset.	1.5

Fit the Chosen Model * Actually fit your best model from #6 with your training features matrix and target array.	1.5
Equation * Write out the equation of the best logistic regression equation that you selected in #6. * If you used a regularization model had some of your slopes have been "zeroed out", then you can leave these explanatory variables out. * Make sure to use the appropriate notation discussed in class.	2.5
Multicollinearity * Do the remaining explanatory variables (or explanatory variables with non-zero slopes) in this model exhibit an issue with multicollinearity? Explain.	2.5
Slope Interpretations * Are you able to interpret the magnitudes of the slopes as indicating how important the corresponding explanatory (or indicator) variable is when it comes to predicting your response variable in a logistic regression model? Explain. * If you are able to, which explanatory (or indicator) variables are the most important?	2.5
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? * Let's explore why these overfitting variables may have been overfitting: - Describe the strength of the relationship bewteen any of these overfitting explanatory variables and the response variable. - Were any of these overfitting explanatory variables strongly associated with another explanatory variable that is still left in the model?	4
Test ROC and AUC * Plot the ROC curve for this logistic regression model and the test dataset. * Calculate the AUC of this ROC curve. * Interpret this test ROC and AUC. What does the ROC and AUC tell us about our best logistic regression model's ability to classify the observations in this test dataset in general?	4
Best Predictive Probability Threshold * Use this ROC curve to select a predictive probability threshold that best meets the particular research goals of the person that you described in your research motivation. * What is the test FPR and TPR of the classification that would be created using this predictive probability threshold? * Put this FPR and TPR into words in the context of your research analysis.	4
8. Additional Analysis/Insight	
Report Professionalism * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1

1. Perform one additional analysis with this dataset that is outside of the scope of the questions that I have specifically asked in this document. 2. This "analysis" can involve creating another visualization, using another analysis (learned in this class), using another analysis (learned in another class, like k-means), or any other technique that you think would be interesting and helps advance one of your three research goals. 3. How do your insights from this additional analysis advance one of the three research goals that you're pursuing in this report?	6
9. Conclusion	
Report Professionalism * Write in complete sentences * Write text in the markdown files (not code blocks). * You are not copy-pasting the prompts/questions from this rubric and answering them. Rather, you should incorporate the requirements in this rubric naturally into a paragraph.	1.5
Recommendation * Would you recommend your best model to be used by the person that you mentioned in your motivation? Why or why not? (At the very least, you should discuss and interpret the average test AUC of your best model here).	3
Shortcomings/Caveats * Do you know FOR SURE that your chosen best model will yield the HIGHEST possible average test AUC 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 average test AUC? * Discuss any other shortcomings to your analysis here (all analyses have SOME shortcomings).	3
Future Work * Based on what you observed in your analysis, what is one idea you might have for future work?	3
Total	**100**

Project Presentation Rubric [25 points]

PRESENTATION

Length [3 points]

- **The presentation is no longer than 9 minutes.** ***(-1 point for every 20 seconds over 9 minutes).***

Clarity and Motivation [2 points]

- Clearly states their research goals.
- Clearly states and "sells" their motivation for the analysis. The motivation is believable.
- Clearly states the extent to which they achieved their research goals.

Presenting [3 points]

- All team members speak and present some portion of the material.
- Team members speak loud enough for everyone to hear.
- Team members understand the material, they **are not reading directly from a notecard or script.**

SLIDES

Content [10 points] You should present *some* content on each of these topics.

- Motivation and introduction
- *Very* brief display/discussion of the dataset (no more than 30 seconds).
- *Briefly* discusses any data cleaning decisions made (no more than 30 seconds).
- Show noteworthy visualizations from your descriptive analytics section. Why were these visualizations that you showed noteworthy with respect to your research goals?
- Discusses the model selection technique used.
- Shows the best model found by the feature selection technique and shows the improvement in average test AUC, compared to the full model.
- Writes out the best logistic regression model equation.
- Discusses insights from the "best model" discussion section.
- Conclusion (including shortcomings)

Correctness [3 points]

- Analyses are appropriate for the data, results are interpreted correctly.

Layout [4 points]

- **No code shown on the slides!**
- **No irrelevant code output is shown on the slides.**
- Content is well organized.
- Fonts are easy to read.
- Visualizations are not messy and are easy to see and interpret.
- Slides are engaging.
- Slides are not too wordy.
    - Your slides should not contain paragraphs of text.
    - Should use bullet points.
    - Complete sentences are often not needed and can be visually burdensome.

Report Peer Evaluation Questions [10 points]

Your assigned group will see your responses.

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 group'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 group 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.

Presentation Peer Evaluation Questions [10 points]

Your assigned group will see your responses.

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

1. What were their **research goals**? How **clear/upfront** were they when it came to articulating their research goals?

2. What is the **motivation** for the research goals that this group pursued in this report? How **effective** were they at "selling" this motivation?

3. To what extent were they able to **meet their research goals**? How **clear/upfront** were they when it came to articulating this?

4. **Shortcomings:** Name at least one other thing 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.

Individual Research Impact Questions [5 points]

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

1. What were your individual **contributions** to this group project?

2. Select one of your teammates. How might a **decision that you made** in your assigned part of the report have **influenced** the **results, insights, and/or decisions made by this teammate** (or vice versa)?

3. Select *another one* of your teammates. How might a **decision that you made** in your assigned part of the report have **influenced** the **results, insights, and/or decisions made by this teammate** (or vice versa)?