D208 Performance Assessment (NBM3) Helpful Tips (Task 1 and Task 2)

Important Reminders!

1. Follow appropriate APA formatting guidelines for the cover page, table of contents, headers, etc.

2. (*) Indicates an area of the PA that often presents the most challenging to students. Utilize the Performance Assessment Rubric to ensure you respond to all requirements accurately and thoroughly.

3. Allow a 3-day turnaround for the evaluation of your performance assessment.

4. Remember, the goal for this course is to adhere to the process of creating a regression model, not to create a perfect model.

5. Only include your code when asked.

6. The use of appendices for code, visualizations, etc. is appropriate and helpful. You may find it helpful to include visualizations within the document, therefore, remember to refer to graphs and charts within your narratives (to provide context) and ensure that your graphs, tables, and charts are always labeled.

7. Write your performance assessment for a “non-technical” audience. The aim is to assess your ability to communicate “technical and statistical findings” to a broad business-oriented audience.

8. Refer to the verbs of the requirement. When asked to “state”, “explain” or “discuss”, there is an expectation that a narrative will be present. Be concise, yet thorough and comprehensive.

Part I: Research Question

A1. State your research question

- Your research questions should be a question and not as a statement (hence research question).

- Remember for Task 1 (Multiple Linear Regression), your target variable should be continuous and for Task 2 (Logistic Regression), your target variables should be categorical.

A2. State Objectives and Goals for Analysis

- For this requirement, briefly **state** the goal of your analysis as it pertains to your research question. I usually share with students and turn my research questions into a statement.

  For example, if my research questions were, "What student factors correlated to GPA (grade point average)? My objective and goal could state: The objective of my analysis is to gain greater insight to determine what student factors directly correlate to the GPA.

Part II: Method Justification

B1. Assumptions

- State only four assumptions of a multiple linear regression model (task 1) / logistic regression (task 2).

B2. Programming Language and Benefits

- Discuss what programming language you used to clean your data and at least two reasons why you are using this language.
- Also, discuss what libraries and packages and why.

Note: I would encourage you to use the following link to assist you when justifying the use (providing benefits of Python or R https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html]

B3. Justification of using Regression

- Explain why multiple linear regression (task 1) / logistic regression (task 2) is an appropriate technique to use for analyzing the research question summarized in part I.
- For example, why would you need to use either Multiple Linear Regression or Logistic Regression to answer your research question.

Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling)

C1. Data Cleaning

- **Describe** your **data cleaning goals** and **the steps** used to clean the data to achieve the goals that align with your research question.

  Note: You must still engage in the activities of detection and treatment of nulls and outliers.

- Note: If any variables within your dataset contain values of “NONE” as an acceptable value (according to the data dictionary), you must ensure that the values of “NONE” remain after importing. To avoid these values getting converted to nulls, for those who are using python, follow this helpful tip:

  - In python: Set the “keep_default_na” parameter to false.
    For example:
    df=pd.read_csv('C:/Users/keiona.middleton/Documents/DataFile/D208_churn_clean.csv', keep_default_na=False) #import data file to be cleaned

- **<u>Include</u>** a copy of your annotated cod**e** used for cleaning the data (nulls, outliers, etc. Remember, re-expression of categorical variables is not a data cleaning activity, but a data wrangling (transformation) activity. If you are uploading a copy of your code (which is highly recommended), state, “see code/script attached” in addition to providing your code within the document itself).

C2. Data Exploration (EDA)

- Perform summary statistics on your dataset

- **<u>Provide</u>** a screenshot of the output of the summary statistics>

  Note: The summary statistics is akin to “descriptive statistics” which summarizes or describes the characteristics of a data set (i.e., median, mean, IQR, count, etc.). This can be achieved by using the “describe” function (PYTHON) or the “summary” function (R Studio)

- **<u>Provide a brief discussion</u>** of your summary statistics results. In your discussion of the summary statistics, assume that the reader has no knowledge or understanding of mean, median, etc. (*)

  IMPORTANT NOTE: Summary statistics is usually a data exploration activity. Therefore, it is customary to perform summary statistics on your dataset prior to data wrangling (transformation), meaning your summary statistics may only include the quantitative variables you have decided to use for your initial model.

  Therefore, you should include a statement as to why these variables may not be visible in the summary statistics, however, you should include a brief summary of your categorical variables using a table or other visualization.

  This is an example of a summary for categorical variables:

  Variable Marital Status (# of observations: 10,000)
  - Married - 38%
  - Divorce - 12%

- Widowed – 25%
- Single – 25%

C3. Visualizations

- **Provide univariate** visualization for all independent (predicting) variables and your dependent (target) variable.

- **Provide bivariate** visualizations which includes **both** the dependent (target) and independent (predicting)variables.

NOTE: Remember, if you have 10 predicting variables and 1 target variable, you will have a total of 11 univariate visualizations, and 10 bivariate visualizations. *(*)*

Note: You should ensure that the visualizations, at minimum, include the variables included in the in the initial model.

Note: Include title names and axis names for your univariate and bivariate visualizations. Ensure they are easy to read.

C4. Data Transformation (Data Wrangling)

- **Describe** all data wrangling activities you performed on your dataset.

- **Include** a copy of your annotated cod**e** used to wrangle (transform) your data. If you are uploading a copy of your code (which is highly recommended), state, “see code/script attached” in addition to providing your code within the document itself).

NOTE: For example, if your variables contain categorical variables, describe why you need to re-express your categorical variables, and the steps used to complete this re-expression. If you decide to perform log-transformation (which is not required) on the dataset, discuss why and steps used.

Note: If no categorical variables were used explicit state that due to variables selected, re-expression was not necessary, and explain specifically why no categorical variables were utilized from the provided dateset.

C5. Prepared Dataset

- Provide (attached) a copy of the prepared data set as a CSV file.

Note: Your prepared dataset should reflect all the variables you decided to use for the initial regression model. The data should be cleaned and wrangled (transformed).

Part IV: Model Comparison and Analysis

D1. Initial Model

- **Construct** an initial multiple linear regression model (task 1) / logistic regression model (task 2) from *all* independent variables that were identified in part C5. Provide a screenshot of the initial model summary.

D2. Model Reduction Method and Justification

- Discuss and describe what statistically based feature selection procedure or a model metric procedure you will use to reduce the initial model in a way that aligns with the research question. *(Hint: Discussions regarding model reduction methods can be found in Getting Started with D208 Part I).* (*)

- Explain (justify) why this method is being used to reduce your model.

D3. Reduced Model

- After you reduce the model (using the methodology discussed in D2), **provide a screenshot** of the reduced model.

E1. Model Comparison

- Explain your data analysis process by comparing the initial multiple linear regression model and reduced linear regression model **using a model evaluation metric**. *(Hint: Discussions of evaluation metric can be found in Getting Started with D208 Webinar Part II).* (*)

E2. Provide the following below.

Multiple Linear Regression (Task 1)

a) a residual plot for the **reduced** model (Note: Any residual plot would suffice – i.e., qqplot, histogram of the model's residuals, etc.)
b) the model's residual standard error for the **reduced model**

Logistic Regression (Task 2)

a) confusion matrix
b) accuracy calculation

Note: Perform the confusion matrix on the reduced model; Splitting the data is not required.

E3. Code

- Provide an executable error-free copy of the code used to support the implementation of the linear regression models using a Python or R file. **Attached is a copy of your code (i.e., ipynb file, R studio file).**

Part V: Data Summary and Implications

F1. Regression Equation, Coefficients, etc.

a) **Provide a regression equation** for the variables within reduced model. *(Hint: Discussions of equations for both multiple linear regression and logistic regression can be found in Getting Started with D208 Webinar Part II).*

NOTE: Remember, you are not "solving" the equation but providing an illustration of the equation based on the variables remaining in your reduced model. (*)

b) **Provide an interpretation** of the coefficients of the reduced model. *(Hint: Discussions of coefficient interpretation can be found in Getting Started with D208 Webinar Part II).* (*)

NOTE: Remember, if any of the variables remaining in your reduced model are dummy variables, please ensure that you are interpreting the dummy variables appropriately.

NOTE: For logistic regression (task 2), ensure that you interpret the coefficient appropriately based on log odd or change odds.

c) **Provide a discussion** regarding the **statistical significance** and **practical significance** of your reduced model.

- First, you will need to discuss IF your model is/are statistically significant and why or why not.
- Second, you will need to discuss IF your model is practically significant and why or why not.

(Hint: Discussions of statistical and practical significance can be found in Getting Started with D208 Webinar Part II). (*)

d) Discuss the (implications) disadvantages of the methods you used to conduct your regression model (i.e., data preparation/manipulation, model reduction methodology, etc.) (*)

F2. Recommendations

- Based on your regression results / analysis / findings, discuss recommended action(s) would you suggest an organization.

NOTE: Your response should include very specific.
NOTE: Utilize the information discuss in E1, E2, F1.b, c as a start (*)

Part VI: Demonstration

G. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of:

- functionality of the code used for the analysis of the programming environment.

- Discussion of the version of the programming environment

- Discussion of comparison of the initial multiple linear regression model you used and the reduced linear regression model you used in your analysis

- Discussion of the interpretation of the coefficients of the reduced model

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the "WGU" option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto's website.

To submit your recording, upload it to the Panopto drop box titled "Regression Modeling – NBM3 | D208." Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

Sources

H. List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

In this section, cite the sources you used to assist with the **CODE** of your work.

Note: Any reference entry listed here must have an **in-text citation**. Use APA Citation. If no additional sources were used, please state that as such.

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

In this section, cite the source used to assist you with the **CONTENT** of your work.

Note: Any reference entry listed here must have an **in-text citation** in the report above. Use APA Citation. If no additional sources were used, please state that as such.

Professional Communications

J. Demonstrate professional communication in the content and presentation of your submission.