# NBM2 — NBM2 TASK 2: LOGISTIC REGRESSION FOR PREDICTIVE MODELING

PREDICTIVE MODELING — D208

PRFA — NBM2

TASK OVERVIEW SUBMISSIONS EVALUATION REPORT

## COMPETENCIES

**4030.5.1 : Logistic Regression**
The graduate employs logistic regression algorithms in describing phenomena.

**4030.5.3 : Regression Implications**
The graduate makes assertions based on regression modeling.

## INTRODUCTION

As a data analyst, you will assess continuous data sources for their relevance to specific research questions throughout your career.

In your previous coursework, you have performed data cleaning and exploratory data analysis on your data. You have seen basic trends and patterns and now can start building more sophisticated statistical models. In this course, you will use and explore both multiple regression and logistic regression models and their assumptions.

For this task, you will select **one** of the Data Sets and Associated Data Dictionaries from the following link:

Data Sets and Associated Data Dictionaries

You will then review the data dictionary related to the raw data file you have chosen, and prepare the data set file for logistic regression modeling. The organizations connected with the given data sets for this task seek to analyze their operations and have collected variables of possible use to support decision-making processes. You will analyze your chosen data set using logistic regression modeling, create visualizations, and deliver the results of your analysis. It is recommended that you use the cleaned data set from your previous course.

Help

*Note: The link to the data files can also be found below in the web links section. If you have trouble accessing the link, copy and paste the link directly into your web browser.*

# REQUIREMENTS

*Your submission must be your original work. No more than a combined total of 30% of the submission and no more than a 10% match to any one individual source can be directly quoted or closely paraphrased from sources, even if cited correctly. The originality report that is provided when you submit your task can be used as a guide.*

*You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.*

*Tasks may **not** be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc., unless specified in the task requirements. All other submissions must be file types that are uploaded and submitted as attachments (e.g., .docx, .pdf, .ppt).*

**Part I: Research Question**

A. Describe the purpose of this data analysis by doing the following:
   1. Summarize **one** research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.
   2. Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

**Part II: Method Justification**

B. Describe logistic regression methods by doing the following:
   1. Summarize the assumptions of a logistic regression model.
   2. Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.
   3. Explain why logistic regression is an appropriate technique to analyze the research question summarized in Part I.

**Part III: Data Preparation**

C. Summarize the data preparation process for logistic regression by doing the following:
   1. Describe your data preparation goals and the data manipulations that will be used to achieve the goals.
   2. Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.
   3. Explain the steps used to prepare the data for the analysis, including the annotated code.

4. Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.
5. Provide a copy of the prepared data set.

**Part IV: Model Comparison and Analysis**

D. Compare an initial and a reduced logistic regression model by doing the following:

1. Construct an initial logistic regression model from *all* predictors that were identified in Part C2
2. Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.
3. Provide a reduced logistic regression model.

*Note: The output should include a screenshot of each model.*

E. Analyze the data set using your reduced logistic regression model by doing the following:

1. Explain your data analysis process by comparing the initial and reduced logistic regression models, including the following elements:
   - the logic of the variable selection technique
   - the model evaluation metric
2. Provide the output and *any* calculations of the analysis you performed, including a confusion matrix.

*Note: The output should include the predictions from the refined model you used to perform the analysis.*

3. Provide the code used to support the implementation of the logistic regression models.

**Part V: Data Summary and Implications**

F. Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following elements:
   - a regression equation for the reduced model
   - an interpretation of coefficients of the statistically significant variables of the model
   - the statistical and practical significance of the model
   - the limitations of the data analysis
2. Recommend a course of action based on your results.

**Part VI: Demonstration**

G. Provide a Panopto video recording that includes *all* of the following elements:

- a demonstration of the functionality of the code used for the analysis
- an identification of the version of the programming environment
- a comparison of the **two** logistic regression models you used in your analysis

- an interpretation of the coefficients

*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 "Logistic Regression Modeling – NBM2 | 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.*

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.

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

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

### File Restrictions

File name may contain only letters, numbers, spaces, and these symbols: ! - _ . * ' ( )
File size limit: 200 MB
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z

## RUBRIC

**A1: RESEARCH QUESTION:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| A summary of 1 research question is not provided. | The summary includes 1 research question, but the research question is not relevant to a realistic organizational situation or cannot be | The summary includes 1 research question that is relevant to a realistic organizational situation and can be addressed using the selected data set and logistic regression. |

addressed using the selected data set and logistic regression.

A2: OBJECTIVES AND GOALS:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not define the objectives or goals of the data analysis. | The submission defines the objectives or goals of the data analysis, but 1 or more of the objectives or goals are not reasonable for the scope of the scenario or are not represented in the available data. | The submission clearly defines the objectives or goals of the data analysis, and the objectives or goals are reasonable for the scope of the scenario and are represented in the available data. |

B1: SUMMARY OF ASSUMPTIONS:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not summarize the assumptions of a logistic regression model. | The submission summarizes the assumptions of a logistic regression model, but 1 or more of the assumptions contain inaccuracies. | The submission accurately summarizes the assumptions of a logistic regression model. |

B2: TOOL BENEFITS:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not describe the tool(s) chosen in support of various phases of the analysis. | The submission describes the tool(s) chosen in support of various phases of the analysis, but the description does not include the benefits of using the tool(s) for logistic regression, or the benefits do not logically align with the goal of the analysis. | The submission describes the benefits of using the tool(s) chosen in support of the various phases of the logistic regression analysis, and the benefits logically align with the goal of the analysis. |

B3: APPROPRIATE TECHNIQUE:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not explain why logistics regression is an appropriate technique. | The submission explains why logistic regression is an appropriate technique, but the explanation does not relate to the research question from Part I, or the explanation contains inaccuracies. | The submission accurately explains why logistic regression is an appropriate technique to analyze the research question from Part I. |

C1: DATA GOALS:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not include a description of the data preparation goals. | The submission describes the data preparation goals but not the data manipulations that will be used to achieve the goals. Or the manipulations are not in alignment with the data preparation goals, logistic regression analysis, or the research question. | The submission describes the data preparation goals and the data manipulations that will be used to achieve the goals. The goals and manipulations align with each other and with logistic regression analysis and the research question. |

C2: SUMMARY STATISTICS:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not discuss the summary statistics. | The submission discusses summary statistics but does not discuss the target variables or *all* predictor variables that need to be gathered from the data set to answer the research question, or the discussion contains inaccuracies. | The submission accurately discusses the summary statistics and discusses the target variable and *all* predictor variables that need to be gathered from the data set to answer the selected research question. |

C3: STEPS TO PREPARE THE DATA:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not explain the steps used to prepare the data for the analysis. | The submission explains the steps used to prepare the data for the analysis, but is miss- | The submission explains *all* necessary steps used to prepare the data for the analysis. The |

| | | |
|---|---|---|
| | ing 1 or more necessary steps, or the steps do not include the annotated code, or are not related to a logistic regression analysis. | steps include the annotated code and relate to preparing for a logistic regression analysis. |

**C4: VISUALIZATIONS:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not include *both* univariate and bivariate visualizations of the distributions of variables in the cleaned data set. | The submission generates *both* univariate and bivariate visualizations of the distributions of variables in the cleaned data set, but *any* of the visualizations contain inaccuracies, or the bivariate visualizations do not include the target variable. | The submission accurately generates *both* univariate and bivariate visualizations of the distributions of variables in the cleaned data set. The bivariate visualizations include the target variable. |

**C5: PREPARED DATA SET:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not provide a data set. | The submission provides a copy of a prepared data set, but the data set is not fully prepared or is incomplete. | The submission provides a copy of the fully prepared data set. |

**D1: INITIAL MODEL:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not provide an initial logistic regression model. | The submission provides an initial logistic regression model from *some*, but not *all*, predictors identified in Part C2, or the model contains inaccuracies. | The submission provides an accurate initial logistic regression model from *all* predictors identified in Part C2. |

**D2: JUSTIFICATION OF MODEL REDUCTION:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission neither justifies a variable selection procedure nor a model evaluation metric to reduce the initial model. | The submission justifies a variable selection procedure or a model evaluation metric to reduce the initial model but does not justify *both*, or the justification is not in alignment with the research question, or the variable selection procedure is not statistically based. | The submission justifies a statistically based variable selection procedure and a model evaluation metric to reduce the initial model. The justification is in alignment with the research question. |

**D3: REDUCED LOGISTIC REGRESSION MODEL:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not provide a reduced logistic regression model. | The submission provides a reduced logistic regression model, but the reduced model is not in alignment with the justification from part D2. | The submission provides a reduced logistic regression model, and the model is in alignment with the justification from part D2. |

**E1: MODEL COMPARISON:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not explain the data analysis process by comparing the initial and reduced logistic regression models. | The submission explains the data analysis process by comparing the initial and reduced logistic regression models but does not include *all* of the given elements, or the explanation contains inaccuracies. | The submission accurately explains the data analysis process by comparing the initial and reduced logistic regression models, including *all* of the given elements. |

**E2: OUTPUT AND CALCULATIONS:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|

| | | |
|---|---|---|
| The submission does not provide the output or calculations of the analysis performed. | The submission provides the output and calculations of the analysis performed, but not all of the output is included or 1 or more of the calculations are missing or contain inaccuracies, or the submission does not include a confusion matrix. | The submission provides the accurate output and calculations of the analysis performed, including a confusion matrix. The submissions includes *all* necessary output and calculations. |

**E3: CODE:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not provide the code used to support the implementation of the logistic regression models. | The submission provides the code used to support the implementation of the logistic regression models, but the code is incomplete or contains inaccuracies. | The submission provides the code used to support the implementation of the logistic regression models, and the code is complete and accurate. |

**F1: RESULTS:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not discuss the results of the data analysis, or the discussion is not in alignment with the research question and the data analysis. | The submission discusses the results of the data analysis, but the discussion does not address *all* of the given elements, or the discussion contains inaccuracies. | The submission accurately discusses the results of the data analysis, and the discussion addresses *all* of the given elements and is in alignment with the research question and the data analysis. |

**F2: RECOMMENDATIONS:**

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not recommend a course of action based on results. | The submission recommends a course of action based on results, but the recommendation is not appropriate based on the research question or the results of the data. | The submission recommends an appropriate course of action based on the results as they relate to the research question. |

G: PANOPTO DEMONSTRATION:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| A Panopto video is not provided. | A Panopto video recording is provided, but it does not include *all* of the given elements, or the video does not capture *both* the presenter and the functioning code for the duration of the video. | A Panopto video recording is provided that includes *all* of the given elements. For the duration of the presentation, the video captures *both* the presenter and the functioning code in a Panopto video recording. |

H: SOURCES OF THIRD-PARTY CODE:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not record *any* web sources. | The submission lists only *some* of the web sources used to acquire data or segments of third-party code. Or the web sources are not reliable. | The submission lists *all* web sources used to acquire data or segments of third-party code, and the web sources are reliable. |

I: SOURCES:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|
| The submission does not include both in-text citations and a reference list for sources that are quoted, paraphrased, or summarized. | The submission includes in-text citations for sources that are quoted, paraphrased, or summarized and a reference list; however, the citations or reference list is incomplete or inaccurate. | The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available. |

J: PROFESSIONAL COMMUNICATION:

| NOT EVIDENT | APPROACHING COMPETENCE | COMPETENT |
|---|---|---|

| | | |
|---|---|---|
| Content is unstructured, is disjointed, or contains pervasive errors in mechanics, usage, or grammar. Vocabulary or tone is unprofessional or distracts from the topic. | Content is poorly organized, is difficult to follow, or contains errors in mechanics, usage, or grammar that cause confusion. Terminology is misused or ineffective. | Content reflects attention to detail, is organized, and focuses on the main ideas as prescribed in the task or chosen by the candidate. Terminology is pertinent, is used correctly, and effectively conveys the intended meaning. Mechanics, usage, and grammar promote accurate interpretation and understanding. |

## WEB LINKS

Data Sets and Associated Data Dictionaries
If you have trouble with the link, copy and paste the link directly into your web browser.

Panopto Access
Sign in using the "WGU" option. If prompted, log in with your WGU student portal credentials, which should forward you to Panopto's website. If you have any problems accessing Panopto, please contact Assessment Services at assessmentservices@wgu.edu. It may take up to two business days to receive your WGU Panopto recording permissions once you have begun the course.

Panopto How-To Videos