Data Mining – D209

Task 2: Predictive Analysis

Western Governor’s University

Performance Assessment

Matthew Morgan

Student ID: 010471280

12/04/2022

**Part I: Research Question**

A.  Describe the purpose of this data mining report by doing the following:

1.  Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following prediction methods:

•  decision trees

•  random forests

•  advanced regression (i.e., lasso or ridge regression)

2.  Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

**Part II: Method Justification**

B.  Explain the reasons for your chosen prediction method from part A1 by doing the following:

1.  Explain how the prediction method you chose analyzes the selected data set. Include expected outcomes.

2.  Summarize **one** assumption of the chosen prediction method.

3.  List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

**Part III: Data Preparation**

C.  Perform data preparation for the chosen data set by doing the following:

1.  Describe **one** data preprocessing goal relevant to the prediction method from part A1.

2.  Identify the initial data set variables that you will use to perform the analysis for the prediction question from part A1, and group each variable as continuous or categorical.

3.  Explain the steps used to prepare the data for the analysis. Identify the code segment for each step.

4.  Provide a copy of the cleaned data set.

**Part IV: Analysis**

D.  Perform the data analysis and report on the results by doing the following:

1.  Split the data into training and test data sets and provide the file(s).

2.  Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

3.  Provide the code used to perform the prediction analysis from part D2.

**Part V: Data Summary and Implications**

E.  Summarize your data analysis by doing the following:

1.  Explain the accuracy and the mean squared error (MSE) of your prediction model.

2.  Discuss the results and implications of your prediction analysis.

3.  Discuss **one** limitation of your data analysis.

4.  Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

**Part VI: Demonstration**

F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

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 “Data Mining I – NVM2.” 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.

G.  Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

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

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