Data Mining – D209

Task 1: Classification Analysis

Western Governor’s University

Performance Assessment

Matthew Morgan

Student ID: 010471280

12/04/2022

**Part I: Research Question**

**A1. Research Question**

**Can I predict which patients are at risk of re-admission using the k-nearest neighbor’s or Naïve Bayes method, so the hospital can take appropriate steps to reduce re-admissions?**

**A2. Goal of the Data Analysis**

**The primary goal of this data analysis is to develop a machine learning model using KNN or NB to help the company identify customers who are at re-admission risk.**

**Based on EDA identify key contributors/features causing customer churn.**

**Based on the results and implications of the analysis, recommend a course of action.**

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 classification methods:

•  k-nearest neighbor (KNN)

•  Naive Bayes

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**

**B1. Explain Method from Part A1**

**The basic idea of k-NN is to predict the label of any data point by looking at the ‘k’ closest labeled data point, for example, 5 closes labeled data points, and getting them to vote on what the unlabeled point should have. The new data point is then assigned the label of the majority of the ‘k’ closest points. k-NN uses distance, and the standard Euclidean distance is the most common and default choice. See the example below:**

**OR**

**Naïve Bayes classifier classification method is more suitable for a large dataset, and all the categorical parameters are independent. A dataset of 10,000 observations and 50 variables is large enough for NB. When the calibration sample sizes increase, Naïve Bayes classifiers will outperform other highly advanced classification methods in terms of classification accuracy . (Rudner, 2016: Vomlel, 2004).**

**Naïve Bayes classification is based on Bayes Theorem which combines previous knowledge with new information obtained from observed data. Bayes Theorem can be used to identify which class (re-admission or no re-admission) is most likely for the provided predictor variables. (Larose, C. D., & Larose, D. T. 2019)**

**B2. Summarize one assumption of your chosen classification model.**

**Assumption of KNN**

* **The k-NN classification method assumes that similar things exist in proximity to each other.**
* **If a data point is far away from another group, it’s dissimilar to those data points.**
* **The algorithm depends on this assumption being true enough for the algorithm to be useful.**
* **The algorithm classifies new data points based on how the neighbors are classified (Medium, 2019)**

**Assumption of NB**

* **Naïve Bayes classification assumes the variables are independent; in other words, a specific feature in a class is unconnected to the occurrence of other elements. (Draper & Smith, 1998).**
* **Naïve Bayes model is assumed to be easy to build and particularly useful for enormous data sets and is known to surpass high advanced classification method. (Vomlel J, 2004).**

**B3. List Packages or Libraries Chosen**

|  |  |
| --- | --- |
| **Packages** | **Usage** |
| **Pandas** | **Importing data and data manipulation** |
| **Numpy** | **Provides array objects for calculations** |
| **Seaborn** | **For visualizations like correlation matrix** |
| **Matplotlib.pyplot** | **For visualizations like ROC curve** |
| **Sklearn,preprocessing** | **To scale features** |
| **Skleanr.feature\_selection** | **For feature selection** |
| **Sklearn.model\_selection** | **For splitting data into train and test sets** |
| **Sklearn.pipeline** | **To assemble several steps that can be performed together while setting different parameters** |
| **Scipy.stats** | **To run statistical calculations** |
| **Statsmodels.formula.api import ols** | **To calculate linear regression** |

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

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

2.  Summarize **one** assumption of the chosen classification 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**

**C1. One Data Preprocessing Goal from A1**

Data preprocessing covers a variety of operations used by data scientists to get the data in a form to be ready for analysis and including the following:

1. For KNN and NB, removing white spaces, imputing missing data, binding, one-hot encoding, etc.
2. For k-NN, scaling and standardizing are preprocessing methods used to improve model accuracy. (Code example for scaling X=scale(X)
3. k-NN uses the distance between data points (Euclidean Distance) to label new data points accurately.
4. k-NN finds a predefined number of training samples closest in the distance to the new point and predicts the label from these. If the feature variables have different scales, then the k-NN will be less accurate.
5. For NB, feature selection and the removal of data not utilized in the analysis.
6. Converting character fields to factors.

**C2. Identify Initial Data Set Variables Used for Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable #** | **Independent Variable** | **Data Type** | **Data Class** |
| **1** | **Initial\_days** | **Continuous** | **Quantitative** |
| **2** | **Services\_CT\_Scan** | **Categorical** | **Qualitative** |
| **3** | **Children** | **Continuous** | **Quantitative** |
| **4** | **Services\_Intravenous** | **Categorical** | **Qualitative** |
| **5** | **Population** | **Continuous** | **Quantitative** |
| **6** | **Initial\_Admin\_Emergency\_admission** | **Categorical** | **Qualitative** |

**C3. Explain Each of the Steps Used to Prepare the Data for Analysis**

1. **Load data into dataframe**
2. **View the data**
3. **Evaluate data structures and types**
4. **Check for null values, missing data, and outliers**
5. **Convert all data to quantitative**
6. **Rename columns from pd.get\_dummies**
7. **Select variables/features using SelectKBest from scikit-learn**
8. **Look for correlation in data. Correlated features will carry extra weight on the distance calculation than desired which could negatively impact the model.**
9. **Convert categorical variables into dummy variables, and this also takes care of datatype change to int for all the object datatypes variables.**
10. **Scale the features, the k-NN algorithm relies on distance for classification, k-NN accuracy can severely degrade if feature scales are not consistent with their importance.**

**C4. Provide a Copy of the Cleaned Data Set**

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

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

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

3.  Explain each of 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**

**D1. Split Data into Training and Test Data Sets and Provide the File(s)**

**D2. Describe Analysis Technique Used**

**\*\*If you are using classification to predict readmission, you would provide some discussion of what the total number of neighbors would be.**

**D3. Provide the Code Used to Perform Classification Analysis from D2**

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 classification 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 area under the curve (AUC) of your classification model.

2.  Discuss the results and implications of your classification 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.