CO3093 - Big Data and Predictive Analytics

CW Assignment

Assessment Information

Assessment Number	2
Contribution to overall mark	70%
Submission Deadline	27/03/2024 at 15:00

Assessed Learning Outcomes

This second assessment aims at testing your ability to

- carry out data cleansing and visualization
- develop a classifier and evaluate its performance
- perform appropriate and justified clustering of the data
- communicate your findings on the data

How to submit

For this assignment, you need to submit the followings:

1. A short report (about 8 pages in pdf including all the graphs) on your findings in exploring the given dataset, a description of your model and its evaluation, a description of your clusters and its justification, as well as your recommendations (any decisions or actions that may be taken following your analyses).
2. The Python source code written in order to complete the tasks set in the paper. You should submit the Python code file, (e.g. for group 1 submit a file called `group1_solution.py` or `group1_solution.ipynb`). Note that even if you decide to work on your own, you must enrol yourself into a group.
3. A signed coursework cover – this should include the names of all the students involved in the work submitted.

Please put your source code, report and signed coursework cover into a zip file `CW2_GroupID.zip` (e.g., `CW2_Group1.zip`) and then submit your assignment through the module's Blackboard site by the deadline. Note that to submit, you need to click on the Coursework link on Blackboard and then upload your zipped file. Remember it is **1 submission per group**!

Problem Statement

In this coursework, you will create a classification model that can predict whether a patient will readmit to the hospital within 30 days using 130 US hospitals diabetes dataset. This dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row contains hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. ***The dataset `diabetic_data.csv` can be downloaded from Blackboard.***

Objective

Using the given dataset, you will develop a predictive model to predict which hospitalized diabetic patients will be readmitted for their condition at a later date and use a K-Means approach to propose a non trivial set of patients' clusters that may make business sense to the healthcare industry.

Exploring the Data

Your first task is to prepare the data and carry out data munging or cleansing, bearing in mind the question you would like to answer. For example, what is the impact of age, number of hospital visits, or various other medical conditions in getting readmitted to the hospital?

Part 1: Building up a basic predictive model

Load the dataset `diabetic_data.csv` into a pandas dataframe and carry out the following tasks. Organise your code bearing in mind robustness and maintainability.

Data cleaning and transformation

If you have a closer look at the dataset, you will see that there are missing values. We need to treat them and in this first model, we are going to follow a basic strategy, which you will improve for a better predictive model later on:

- Show the shape of the dataframe.
- You will find that some missing values are marked with other symbols (such as ?). Replace all missing values of this type with the `numpy.nan`.
- Drop all columns that have more than 50% of missing values. You can also drop columns for which over 95% of their values are the same.
- Ages are given in a 10 years range (i.e. [10-20)). Transform these ages to be the middle value in each given range (i.e. in the case of [10-20), change the range to 15).
- Replace possible missing values in the columns `diag_1`, `diag_2`, and `diag_3` by the number 0.
- Drop all rows with missing values.
- Identify all numerical features and form a list of numerical features and another for the remaining categorical features.
- Identify outliers in the numerical columns and remove them. To keep it simple, you may decide to only keep values that are within 3 standard deviations away from the mean for each feature of the dataset.
- Remove duplicates in the column `patient_nbr` and show the shape of the resulting dataframe.

Data exploration

Carry out a data exploration using appropriate plots to identify patterns or trends in the data. Bearing in mind our objective, we need to assess the impact of the predictors e.g. age, race, gender, or diagnosis type on the outcome (readmitted). Use graphs to prove or disprove the following hypotheses:

- Age has a higher impact on readmission.
- African Americans are more likely to be re-admitted than other ethnic groups.
- Women patients are more likely to be re-admitted than men.
- Diagnosis types have a higher impact on re-admission rates. For this purpose, you need to take into account the classification of disease (icd_codes) in the diagnosis columns (`diag_1`,`diag_2`,`diag_3`) and plot, for example, `diag_1` vs `readmitted`.

Hint 1: Check for distinct values in categorical data and their frequencies. If there are too many distinct values (levels), then you may want to reduce the number of levels by grouping some of the detailed levels. This could be the case for race or diagnosis types.

Hint 2: You may want to transform the readmitted column values to a binary feature. Replace 'NO' with 0, and '<30' and '>30' with 1 respectively.

Hint 3: You may want to join both datasets `diabetic_data.csv` and `icd_codes.csv`.

- **Further information on ICD codes**

  You can find more guidelines and documentation on the ICD codes on the website for the Centers for Disease Control (CDC) here: https://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm. As there are over 1000 codes in the icd documentation, you will want to group certain diagnosis into more general categories. You can find an example of such grouping under the 'Primary diagnosis' section here: https://www.hindawi.com/journals/bmri/2014/781670/tab3/.

Model building

Consider the sub-dataset for the following columns:

['num_medications', 'number_outpatient', 'number_emergency', 'time_in_hospital', 'number_inpatient', 'encounter_id', 'age', 'num_lab_procedures', 'number_diagnoses','num_procedures', 'readmitted']

Build up a model that predicts whether a diabetic patient will be re-admitted or not. Ensure you transform the readmitted column values to be 0 if patients were not readmitted, and 1 if patients were both readmitted within and after 30 days. Split the data into a training and test sets; build up the model; and then show the confusion matrix. Evaluate your model by using a cross-validation procedure.

Part 2: Improved model

This is an open-ended question and you are free to push your problem-solving skills in order to build up a useful model with higher performance.

1. Consider the entire datasets given in this assignment. Develop an improved predictive model that predicts the likelihood for a given diabetic patient to be re-admitted in hospital. Make sure to validate your model. You should aim for a model with a higher predictive accuracy or with results that are easy to explain/interpret.
2. Use the K-Means algorithm to cluster your cleansed dataset and compare the obtained clusters with the distribution found in the data. Justify your clustering and visualise your clusters as appropriate.
3. Include in your report any decisions or actions that should be taken from your improved classification model as well as the obtained clusters on this application.

Marking Criteria

The following areas are assessed:

1. Cleansing, visualizing, and understanding the data **[30 marks]**
2. Building up and evaluating the predictive model **[20 marks]**
3. Improved model, clustering and evaluating/justifying cluster-based model **[20 marks]**
4. Coding style **[10 marks]**
5. Writing the report (up to 8 pages) interpreting the results. **[20 marks]**

Indicative weights on the assessed learning outcomes are given above and can be found in the marking rubric on Blackboard. The following is a guide for the marking:

- **First++ (≥ 90 marks):** As in **First+** plus a classification model with excellent performance, excellent justification and visualisation of the clusters, great insights from the data, and a report of professional standards.
- **First+ (≥ 80 marks):** As in **First** plus a comprehensive coverage of data cleansing techniques leading demonstrating an excellent understanding of the data, a classification model of high performance and a well-structured, maintainable, and robust code usefully using functions.
- **First (≥ 70 marks):** As in **Second Upper** plus a well-justified predictive model by the data cleansing with sound evaluation techniques; well-justified clusters and a concise report containing any decisions that may be recommended.
- **Second Upper (60 to 69 marks):** A good coverage of data cleansing techniques exploring the dataset, a good visualisation of the clusters, a predictive model with an appreciable accuracy with a rationale behind it, a working code and a wellstructured report on the results obtained from the dataset.
- **Second Lower (50 to 59 marks):** Some techniques used for data cleansing are overlooked, a predictive model partially justified with an appreciable accuracy, a working clustering, a partially commented code with very few functions, and a narrative of the findings about the dataset with few deficiencies.
- **Third (40 to 49 marks):** Essential data cleansing techniques are covered, a predictive model is given with some justification, a working but basic block code with no clustering, and a written report describing some of the work done.
- **Fail (≤ 39 marks):** Doesn't satisfy the pass criteria and will still get some marks in most cases.
- **No submission:** A mark of 0 will be awarded.

Marking Group Work

Normally, a group will be given the same mark unless some members made little or no contributions. Any group can be called for an interview during marking. All group members **must attend**, explain their contributions, and defend the work submitted.