Big Data & Predictive Analytics

Classification & Clustering

Learning Outcomes

This task 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 task, you need to submit the followings:

- 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).
- The Python source code written in order to complete the tasks set in the paper. Please put your source code, report into a zip file Task.zip.

Problem Statement

Consider this 'diabetic_data.csv' dataset. The given dataset contains records of diabetic patients admitted to US hospitals from 1999 to 2008. The goal is to monitor and prevent readmission of patients as this is a metric of potential poor care as well as a financial burden to patients, insurers, governments and health care providers.

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? Address the following questions:

1 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:

1. **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:

A. Show the shape of the dataframe. Replace all missing values with the numpy.nan.
B. 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.
C. Transform the age to be the middle value in each given range.
D. Replace possible missing values in the columns diag_1, diag_2, and diag_3 by thenumber 0.
E. Drop all rows with missing values.
F. Identify all numerical features and form a list of numerical features and another for the remaining categorical features.
G. 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.
H. Remove duplicates in the column patient_nbr and show the shape of the resulting dataframe.

2. **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:

1. Age has a higher impact on readmission.
2. African Americans are more likely to be re-admitted than other ethnic groups.
3. Women patients are more likely to be re-admitted than men.
4. Diagnose types have a higher impact on re-admission rates. For this purpose, you need to take into account the icd_codes and plot say diag_1 vs readmitted.

Hint 1: You may want to join both datasets diabetic_data.csv and icd_codes.csv.

Hint 2: 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 3: You may want to transform the readmitted column values to be 0 if the value is NO and 1 otherwise for a better exploration of the data.

3. **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 **[35 marks]**

2. Building up and evaluating the predictive model **[15 marks]**

3. Building up and justification of your clusters **[15 marks]**

4. Coding style **[15 marks]**

5. Writing the report (up to 8 pages) interpreting the results. **[20 marks]**