CO7093 - Big Data & Predictive Analytics

CW Assignment

Classification & Clustering

Assessment Information

Contribution to overall mark	**70%**
Submission Deadline	**27 March 2024, 15:00**

Overview

In this Coursework, you will apply what you've learned so far in the module. The dataset is provided in the CSV format. The purpose of the CW is to introduce you to various stages of a data sciences project such as how to answer questions about data, how to visualize data, and how to build predictive models using data that can make new predictions.

Learning Objectives

After this homework, students will be able to:

- Work with basic **Python data structures** such as dict, tuple, list etc.
- Use **Pandas** as the primary tool to process structured data in Python with CSV files. Handle extreme cases appropriately. Use appropriate methods to address missing data.
- Use **Pyplot** to make simple plots to investigate a specific phenomenon. Read plotting library documentation and use example plotting code to figure out how to create more complex **Seaborn** plots.
- Train a machine learning model and use it to make a prediction about the future using the **scikit-learn** library.

How to submit

For this assignment, you need to submit the followings:

1. A short report (about 8-10 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, local regressors based on your clusters as well as their evaluation.
2. The Python source code written to complete the tasks set in the paper. You should submit the Python code file, 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 page 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**. We encourage you to use **GitHub** versioning control system to store your code and report. Details about using GitHub will follow shortly.

Instructions for group work

This assignment may be attempted in groups of size up to 5. Group size of 1 (individual groups) are also allowed. The link to join a group will be available on the module’s Blackboard page under Assessment and Feedback Section. Please read the following instructions carefully before you join a group.

1. Please join one of the existing groups. Do not create new groups.
2. The maximum number of members in a single group are 5.
3. You should join a group, even if you intend to work individually.
4. The deadline to join a group is 26/02/2024 5:00 pm. After the deadline you will be automatically assigned to your own group of size 1 and you will not be allowed to change your group.
5. Before joining a group, please discuss it with other group members and make sure that you are happy working with each other.
6. Once the groups are finalised, you will not be able to leave or join another 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 concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. The dataset has 47 features and 101766 instances. The dataset is provided on the blackboard.

Objective

Using the given dataset, the goal is to determine the early readmission of the patient within 30 days of discharge. You will use appropriate performance metrics to evaluate the performance of your model. The data is not clean, and you will have to apply appropriate methods to clean the data. Additionally, using unsupervised clustering, you will have to implement cluster-based classification model that may improve the performance of the model.

Exploring the Data

Your first task is to prepare the data and carry out data cleansing, bearing in mind the question you would like to answer. For example, which factor is the most important factor in predicting the readmission of a patient.

Part 1: Building up a basic predictive model

Load the dataset `patients.csv` into pandas dataframe and carry out the following tasks. Perform the following task.

Data Cleaning and Transformation

If you have a closer look at the dataset, you will see that there are lots of inconsistencies and missing values data. You need to deal with those missing values appropriately but in the first instance, you should consider an aggressive approach to deal with them. Following is a list of some of the steps that you may consider.

- Show the shape of the data.
- Delete the column ‘`encounter_id`’.
- Identify missing values in the columns. You will notice that there are some columns with the missing values, but those values are represented by a different character such as ‘?’. Replace them by NaN.
- Show a summary of all missing values before and after applying the above operation.
- The response variable, ‘`readmitted`’, originally had three levels: ‘<30’, ‘>30’, and ‘NO’. Since we are predicting early readmission of the patient within 30 days, we need to convert this feature to a binary feature. Replace ‘<30’ with a 1 and ‘>30’ and ‘NO’ to 0 respectively.
- Check the datatype of each column.
- For each column calculate the percentage of missing values. Drop columns with more than 90% missing values.
- Some columns have no variations. The variables `'examide'` and `'citoglipton'` have only one value. These columns are not useful in prediction and can be deleted. Delete the following near zero-variance colums: `'repaglinide'`, `'nateglinide'`,`'chlorpropamide'`,`'glimepiride'`,`'acetohexamide'`,`'tolbutamide'`,`'acarbose'`,`'miglitol'`,`'troglitazone'`,`'tolazamide'`,`'examide'`,`'citoglipton'`,`'glyburide-metformin'`, `'glipizide-metformin'`,`'glimepiride-pioglitazone'`, `'metformin-rosiglitazone'`,`'metformin-pioglitazone'`.
- Drop rows with null values.
- Display the summary statistics of the numerical columns. Could you identify any outliers in the data? Remove outliers from the data.
- Perform feature normalisation, if required.
- Show the shape of resulting dataframe.

Data Visualisation

Consider the resulting Dataframe. This first aggressive cleaning should give a smaller dataset, which you can start by exploring relationships between the various features of the dataset.

- Plot the distribution of unique classes of the target variable, i.e., readmitted.
- Plot the count of number of readmitted cases against age.
- Plot a graph that displays the count of target variable against the number of medications.
- Show the scatter matrix plot and the correlation matrix. This should be a very large matrix and you might find it difficult to analyse. Which pair of features are highly correlated?
- Generate additional plots that demonstrate your understanding of the problem and the data. You are free to select the plot and features for visualisation. For better visualisation and understanding of data, consider using seaborn library.

Model Building

Consider the resulting Dataframe:

- Select the predictors that would have impact in predicting `readmission`.
- Build up a first linear model with appropriate predictors and evaluate it. Split the data into a training and test sets. Evaluate your model by using a cross-validation procedure.
- Use different performance metrics to evaluate the performance of your model. You might have noticed that the data is imbalanced. The number of positive examples are roughly 10% of the total dataset. Choose appropriate performance metrics to evaluate the performance of your model.
- Balance your data using undersampling or oversampling data balance technique. Train your model again and evaluate its performance. Did you achieve better prediction accuracies with more balanced data?

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.

- Consider the entire datasets again. Develop an improved classification model that predicts the patient readmission. Validate your model. You should aim for a model with a higher performance while using a maximum of data points. This implies treating missing values differently for example through imputation rather than dropping them. You may also consider retaining some of the near-zero variance columns.
- 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.
- Build up local classifiers based on your clustering and discuss how this clusters-based classification compares to your model obtained in the first part of Improved model.
- As in Part1, balance the data and train and test your model with the balanced data.

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 cluster-based model. **[20 marks]**
4. Coding style and use of appropriate data structures. **[10 marks]**
5. Writing the report and 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:

- **[90 – 100]:** A predictive model with excellent performance, excellent justification and visualisation of the clusters, great insights from the data, and a report of professional standards.
- **[80 – 90]:** A comprehensive coverage of data cleansing techniques demonstrating an excellent understanding of the data, a sound comparison of the global predictive model against the clusters-based model and a well-structured, maintainable, and robust code usefully using functions.
- **[70 – 80]:** As in Second Upper plus a well-justified predictive model by the data cleansing with good performance using sound evaluation techniques; a well-justified clusters and a concise report on the results obtained from the dataset.
- **[60 – 70]:** 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 well- structured report on the results obtained from the dataset.
- **[50 – 60]:** Essential data cleansing techniques are covered, 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.
- **[30 – 50]:** Data is not cleaned appropriately. Some important features are ignored. Dataset is too small, and the model is overfitting the data.
- **[0 – 30]:** Code is not complete and is not compiling.

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.