UNIVERSITY COLLEGE LONDON
Faculty of Engineering Sciences
Department of Biochemical Engineering

BENG0095: Group Assignment

Predicting Hospital Readmission Rates for Diabetic Patients

Eddie Edwards (eddie.edwards@ucl.ac.uk)

Overview

- **Assignment Release Date: Friday 24th November 2023**
- **Assignment Hand-in Date: Friday 19th January 2024 at 4pm**
- Weighting: 50% of module total
- Submission Format: A zip file containing a PDF file of the written report and a Juptyer Notebook .ipynb file of the code

Please note that this is a group assignment:

Please arrange yourself into groups of 5 or 6 (up to the specified group maximum capacity) and then register your group on the Moodle module page under the 'BENG0095 (2023/24) Group Assignment: Group Choice' section under the 'Assessment ' tab.

Assignment Description

One of the common application areas for machine learning is in healthcare. Many clinical decisions require accumulation of information from a large array of sources to make a diagnosis or decide patient treatment. Such problems are well-suited to machine learning.

This assignment asks you to predict readmission for diabetic patients admitted to hospital, i.e., what is the likelihood that a patient will need to come back to hospital given the data about this visit. This is based on a database of 101766 hospital visits with a set of 50 features covering patient information, treatment and prescribed medication during the hospital stay.

The database covers patients admitted to 130 hospitals in the US over a period of 10 years to 2008. The original data came from this source:

https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

You will need to build a model that predicts the readmission status of the patient. There are thre categories: No readmission; readmission after 30 days and readmission in less than 30 days.

You can use the training data along with a suitable evaluation method (e.g. splitting the training data into training and validation sets) to train and validate your own models.
You will be assessed primarily not so much on the final predictive accuracy that you obtain, but rather on your approach when attempting this task, and on the level of understanding that you demonstrate. The marking schedule describes the criteria you will be assessed on.

This is not an easy dataset to work with. There are missing values that you will need to account for. The prediction itself is difficult and I do not expect miraculous results. I just want to see that you have made a decent attempt to tackle this problem and shown understanding of machine learning in the methods you have chosen.

Be creative and ask questions of the data. Think about methods to generate features missing in the test data. Most of all, please reflect upon what you have learned during the term and seek to apply machine learning in a way that is appropriate for this task. You can use any of the algorithms in sklearn, Tensorflow or from any other source you may find.

Be sure that all members of the group participate. There are plenty of tasks to be done. Your contributions will be peer-assessed.

Data Description

The data consists of the following files:

- diabetic_data_training.csv (training data with 91589 instances)
- diabetic_data_test.csv (test data with 10177 instances)
- FeatureTable.csv (Descriptions of the features)
- IDS_mapping.csv (The meaning of the numeric IDs for some features)

Do not amend or recreate the test set. This is separate from the training set and must not be used in model selection. I would like all teams to evaluate with the same data. The test set should be used for final evaluation only.

The assignment submission will take the form of:

1. A PDF file containing a 10-page report of your results on the task.

2. A Jupyter notebook containing the **Python** source code of your approach as well as (brief) in-line documentation

Submission Format & Structure

The assignment submission will take the form of a zip file containing:

1. A **PDF file** containing a **report** of the task (this should be at most 10 A4 sides in length, including references). This PDF should be prepared using the LATEX files included on the module Moodle page under the ‘Group Assessment’ link, which provide a format similar in style to \preprint" publications such as arXiv.

2. A **Jupyter notebook .ipynb file** containing the **Python source code** of your approach as well as (brief) in-line documentation. The notebook should include an analysis of the performance of your classifier on the data from the test set file.

The PDF report should adopt the following structure:

1. **Introduction**
   A brief description of your approach to the problem and the results that you have obtained on the training data.

2. **Data Transformation & Exploration**
   Any transformations that you apply to the data prior to training. Also, any exploration of the data that you performed such as visualization, feature selection, etc.

3. **Methodology Overview**
   Start by describing in broad terms your methodology. Include any background reading you may have done and a step by step description of how you have trained and evaluated your model. Describe any feature engineering that you have applied. If you had attempted different approaches prior to landing on your final methodology, then describe those approaches here.

4. **Model Training & Validation**
   This contains a breakdown of how your model was trained and evaluated.

5. **Results**
   Here you show the results that you obtain using your model on the training data. If you have multiple variations or approaches, this is where you compare them.

6. **Final Predictions on Test Set**
   This is the section where you perform your final predictions on the test set using the model that you have trained in the previous section.

7. **Conclusion**
   This is the section where you consider your findings and suggest avenues for future research.

   References
   List of relevant academic papers or sources used in your coursework and cited in the main report (1/2 to 1 page max).

The Notebook should adopt the following structure:

1. **Introduction**
A brief precis of the equivalent section in your report.

2. **Data Import**
This section is how you import the data into the notebook. It should be written in such a way that I can modify it to run on my own machine by simply changing the location of the training data and any additional data sources that you have used.

3. **Data Transformation & Exploration**
Code for the equivalent section in your report, together with in-line documentation of that code.

4. **Methodology Overview**
Code for the equivalent section in your report, together with in-line documentation of that code.

5. **Model Training & Validation**
Code for the equivalent section in your report, together with in-line documentation of that code.

6. **Results**
Code for the equivalent section in your report, together with in-line documentation of that code.

7. **Final Predictions on Test Set**
Code for the equivalent section in your report, together with in-line documentation of that code.

Note:
• Your notebook need only contain brief in-line documentation, while the PDF should contain a more detailed description.
• You will be assessed primarily on the contents of your PDF report. The notebook is required so that we can check that your results are replicable.
• Keep in mind that your notebook should be written in such a way that we can modify the location of the data and then step through your notebook to obtain the same results as you have submitted.

Marking Guidelines

All reports will be marked against the marking rubric, which is available online via the module's Moodle page under the 'Group Assessment' link.

The mark weighting for each section is as follows:

• **Methodology (15%)**
How well is the methodology described? How appropriate is it to the task at hand? Have any extra data sources been used and if so are they useful? Have you done more than just apply a classifier to the training data?

• **Evaluation Strategy (15%)**
Has a suitable evaluation strategy been used so as to avoid any possible bias? If your methodology contains multiple parameters, how have the final parameter values been chosen? Have you used any form of cross validation?

• **Presentation of Results (15%)**
Have you presented results on the training data? Are the results presented appropriate and displayed in an easy to interpret manner? Do they reveal any extra insights about how your model performs?

• **Interest of Approach (40%)**
How interesting and novel is your approach (regardless of predictive accuracy)? Have you used any extra data sources, or transformed the training data in an interesting way? Have you done something that is beyond simply using a standard classifier on the training data?

• **Format, structure, referencing, and clarity of writing/code (15%)**
Are your final notebook and report well laid out and does the write-up follow a clear structure? Have you included any references to show background research/reading? Is your writing free from spelling, punctuation, and grammatical errors and is your code well commented?

For a more detailed breakdown of what constitutes a good (and bad) mark for each of these sections please refer to the marking schedule.