ML and AI

MCSBT2024S-1 Group Project

Teacher: Nir Ailon

FIRM DEADLINE: 4/4/2025 11:59pm

Introduction:

Diabetes (type II), a chronic metabolic disorder characterized by elevated blood sugar levels, affects millions of people worldwide and imposes a substantial economic burden on healthcare systems, particularly in the United States. According to the Centers for Disease Control and Prevention (CDC), diabetes-related healthcare costs in the US totaled over $327 billion in 2017, encompassing direct medical expenses and indirect costs such as reduced productivity and disability. Amidst this healthcare landscape, predicting the readmission of diabetic patients becomes paramount. Hospital readmissions not only contribute significantly to the financial strain on healthcare systems but also indicate potential gaps in patient care, treatment effectiveness, and disease management. Early identification of patients at high risk of readmission allows healthcare providers to implement targeted interventions, such as personalized treatment plans, care coordination, and patient education, aimed at reducing readmission rates, improving patient outcomes, and ultimately mitigating the economic impact of diabetes on the healthcare system.

In this work, you will help predict readmission rates at hospitals. A readmission occurs when an inpatient is discharged from the hospital, and then readmitted again within a short period of time (a few weeks to a few months), thus indicating possible inadequate treatment.

Data Description:

The dataset contains tabular information for more than 100k patients, with the “readmission” endpoint as a label for each patient. This is a 3 class classification problem, with the following classes:

1. “NO” No readmission (53.9% of the cases)
2. “<30”: A readmission in less than 30 days (this is the worst situation, because it may be a result of malpractice at current admission) (34.9% of the cases)
3. “>30”: A readmission in more than 30 days (this is bad, but not as bad as “<30”, because it is more likely to happen due to the patient’s advanced disease progression, and not malpractice. Nevertheless, we would like to try to predict this case as well, because readmissions are a huge burden to the system. (11.2% of the cases).

The baseline accuracy is 53.9% (obtained by simply prediction no readmission). The goal is, of course, to improve this. Small percentage improvements can result in significant money saving for the healthcare system in the United States. The dataset will be provided on Blackboard. Documentation of the different columns is provided in the links below.

Modeling:
You are free to choose any strategy. For example, you can also decide to perform binary classification (YES/NO readmission), then perform another binary classification of the two types of YES (“<30”, “>30”).

Note that there are many columns with many missing values. In some cases, you may want to create a “is_missing” column.

The columns diag_1 diag_2 and diag_3 specify 3 diagnoses done during the patient’s visit at the hospital. The diagnoses are encoded using ICD-9, with hundreds of categories. It is important to find a good strategy to deal with these features. For ideas, you can refer to notebooks submitted by Kaggle competitors.

Expected Outcome:
Submit a concise ~8 minute video summarizing your analysis and solution. Be sure to include the following in your presentation:

- A business oriented introduction. You may include numbers and data (e.g. costs of readmissions, diabetes related economic burden etc) as you find on the web or by asking your expert friends and colleagues. Make sure to cite your references.

- Initial data analysis. Display histograms that seem to be insightful, as well as possible correlations between features and targets that are easy to visualize.
- Data cleaning and manipulation strategy, categorical feature encoding (if applicable), scaling, imputation, outlier detection, feature engineering, possible clustering. Justify your choices whenever there is an intuitive explanation.
- Explanation of your data split (train/val/test) strategy. The dataset is quite large, and your test set (which you are allowed to use once) should be of at least 10k rows. Aside from that, you can choose any splitting strategy you want, with or without k-fold cross validation.
- Explanation of your ML algorithm, including what steps you took to fine tune hyperparameters, if applicable.
- Business interpretation of your solution, including suggesting possible action items for hospitals, together with cost-saving analysis.

In all steps, you are also welcome to include ideas that did not work, if you think that there is insight to gain from these failed efforts.

You are also invited to learn about solutions people have developed and shared on Kaggle or other websites, and use them in your own solution. As always, give credit by citing your references whenever such credit is due. If you do use someone else's solution as reference, you must contribute to that solution with at least one idea of your own, and report it. If your contributed idea didn't work, try to explain why that happened in your report.

Submission:

One group member should submit the video and python code (including jupyter notebooks) on Blackboard by the date announced in due time.

Grading:

Each of the following elements is assessed on a scale of 0 to 10, contributing to the final grade, which is determined by the weighted average of these components:

1. **Data Understanding (20%):**
   - Evaluation of completeness
   - Exploration of data
   - Implementation of data cleaning
   - Execution of feature engineering

2. **Machine Learning Modeling (20%):**
   - Appropriateness of selected algorithms
   - Demonstration of technical understanding of the employed algorithms

3. **Technical Results and Interpretation (20%):**
   - Presentation of results in a clear and organized manner
   - Interpretation of results in the context of the real-world problem
   - Discussion of any limitations or assumptions made during the analysis

4. **Business Modeling (20%):**
   - Integration of modeling insights into the broader business context
   - Demonstration of practical applicability
   - Illustration of the practical use of machine learning algorithms in a business context, emphasizing relevance

5. **Communication and Presentation Skills (20%):**
   - Clarity and effectiveness of oral communication
   - Quality of visual aids
   - Engagement in presenting findings

References

Kaggle data page:
https://www.kaggle.com/datasets/brandao/diabetes

Kaggle community solutions:
https://www.kaggle.com/datasets/brandao/diabetes/code

Original publication introducing the dataset:

https://pubmed.ncbi.nlm.nih.gov/24804245/