Who returns to hospitals?

MLE

Skoltech

March 20, 2017

Overview

1. Data description
2. Problem statement
3. Visualization
4. Multiclass classification
5. Feature importance
6. Binary classification

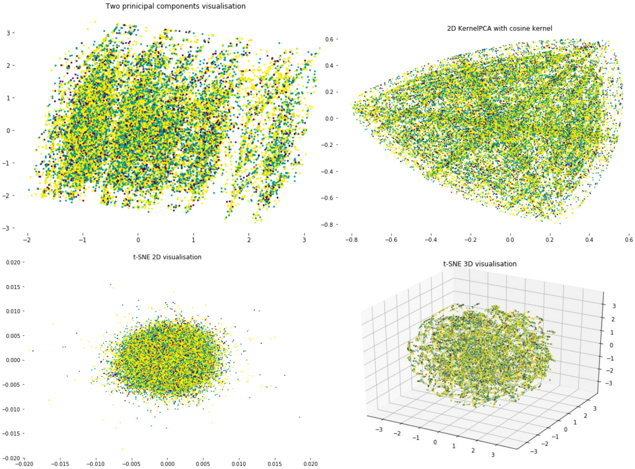
Data description

- This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes.
- Data Set Information:
  1. Some general features (age, weight, race etc.)
  2. A hospital admission
  3. A diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
  4. The length of stay (1 - 14 days)
  5. Laboratory tests were performed during the encounter
  6. Medications were administered during the encounter
- The database contains incomplete, redundant, and noisy information.

Problem statement

- **Problem:**
  We want to predict the probability of patient readmission based on the known features.
- We have tried different approaches: we considered
  1. all data
  2. only relevant medical data (e.g. medications, number of lab procedures)
  3. drug data separately

Visualization



Two prinicipal components visualisation

2D KernelPCA with cosine kernel

t-SNE 2D visualisation

t-SNE 3D visualisation

Multiclass classification

- Readmission labels:

  the patient was readmitted within 30 days, was readmitted in more that 30 days and was not readmitted $[< 30, > 30, \text{No}]$
- Methods:
  1. Random Forest Classifier
  2. One Vs Rest Classifier(estimator=Logistic Regression)
  3. Output Code Classifier(estimator=Logistic Regression)
- Accuracy

Methods	**All data**	**Medical data**	**Drug data**
Random Forest	0.58	0.57	0.53
One Vs Rest	0.57	0.57	0.54
Output Code	0.57	0.57	0.54

Table: Accuracy table

Feature importance

- 5 most important features for all data: admission source id, number emergency, number inpatient, primal diagnosis, number of lab procedures
- 5 most important features for medical data: encounter id, discharge disposition id, admission source id, time in hospital, race

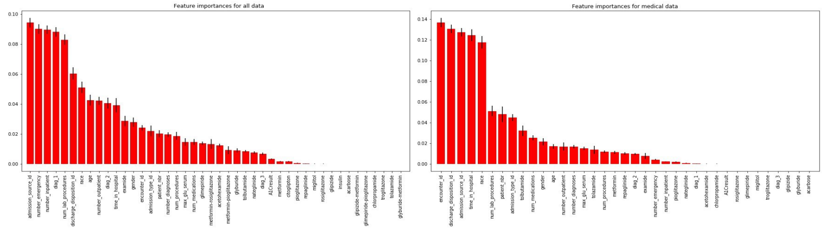


Figure: Feature importance for all and medical datasets

Binary classification

- We reduced the problem to two binary classification problems:
  1. The first class is readmission within 30 days and in more than 30 days, the second one – no readmission
  2. The first class is readmission within 30 days, the second one – in more than 30 days and no readmission
- Methods:
  1. Ada Boost Classifier (estimator=Logistic Regression)
  2. MLP Classifier (logistic activation function)
  3. Naive Bayes
  4. Linear Discriminant Analysis

Binary classification

Here we considered all dataset.

- First binarization – readmission within 30 days
- Second binarization – readmission within and more than 30 days

Methods	First binarization	Second binarization
Ada Boost	0.59	0.60
MLP	0.59	0.60
Naive Bayes	0.59	0.60
LDA	0.59	0.60

Table: ROC-AUC score table for binary classification

Binary classification

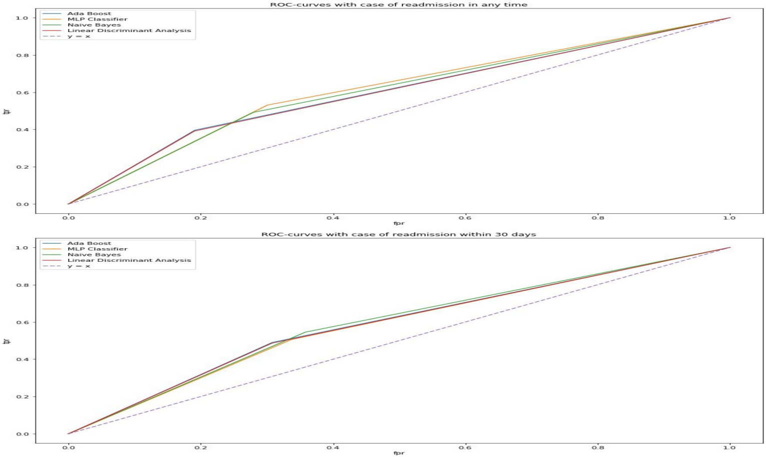


Figure: ROC-AUC curves

The End