Analysis of the readmission of patients with diabete

Violaine Bellée

Machine Learning Course

June 5, 2017

1. Introduction

Diabetic patients admitted in hospitals need specific care and medical actions in order to avoid readmission. In the context of the machine learning course given at EPFL in spring 2017, an analysis of the factors linked to readmission of diabetic patients within less than 30 days is conducted. The goal is to understand which measurements and medical actions are correlated with a lower probability of readmission.

2. Description of the data set

The data used for this study is extracted from the Health Facts database (Cerner Corporation, Kansas City, MO). It contains originally the record of 10 years (between 1998 and 2008) of medical care in 130 hospitals in the United States. A detailed description is given in an analysis made public along with the published article analysing it [1]. The data set used in the present analysis (link) is a subset of the above and has been selected following the following criteria:

- the patient has been diagnosized with diabete, whatever the reason of the hospital admission
- the length of the stay is within 1 and 14 days
- laboratory tests have been performed during the stay
- medications have been administered during the stay

After this first selection, the data set publicly available contains 101766 entries. The features reported are:

- the ID of the encounter (or stay in the hospital)
- the patient number
- the race (can be AfricanAmerican, Asian, Caucasian, Other or unknown)
- the gender (male or female)
- the age category (from '$[0-10)$' to '$[90-100)$')
- the weight
- the type of admission (description to add)
- the discharge disposition (description to add)
- the admission source (description to add)
- the time spent in hospital (between 1 and 14 days)
- the payer code corresponding to the organism that paid for the patient (can be the patient him- or her-self)
- the medical specialty of the admitting physician

- the number of laboratory tests performed
- the number of procedures (other than laboratory tests) performed
- the number of distinct generic names of medications administered
- the number of outpatient visits of the patient in the year preceding the encounter
- the number of emergency visits of the patient in the year preceding the encounter
- the number of inpatient visits of the patient in the year preceding the encounter
- the primary, secondary and supplementary secondary diagnoses
- the total number of diagnoses
- the glucose serum test result (indicated as none, normal, '> 200' or '> 300')
- the measurement of HbA1c (called A1Cresult in the data set) which is a measurement of the glucose control that can either be not performed ("None"), '$< 7\mu$g' or '$< 8\mu$g'.
- the administration of several drugs during the encounter. For each of them, the value indicated is 'No' if the medication has not been prescribed, 'Steady', 'Up' or 'Down' if the dosage is kept the same, increased or decreased.
- the change in the diabetic medication (either change of medication or change of the dosage). Values are 'Ch' if the medication was changed, 'No' otherwise.
- the prescription of diabetic medications ('No' or 'Yes')
- the readmission of the patient ('NO', '> 30' or '< 30' depending of if the patient has been readmitted after or before 30 days after the discharge)

In this data set, some of the categories are not valid or not filled. This is the case for most of the information about the weight, the payer code and the specialty of the admitting physician. Some data is also missing for the race and the supplementary secondary diagnosis.

3. Goal of the analysis and strategy

The goal of this analysis will be to understand the factors preventing readmission in hospital. An admission is counted as such when it occurs less than 30 days after the discharge (ref to the article). This criterion has been chosen because it aligns with the definition of readmission used by funding agencies. Hence the case under study falls down to a classification problem in which two classes are considered: 'Readmission' (when the patient is readmitted in less than 30 days) and 'No readmission' (when the patient is not readmitted, or readmitted after more than 30 days). The strategy will consist in three steps. First, the data is transformed in order to adapt to the current problem, and it is visualized. Then, a classifier is built to estimate the feature importance in the readmission of patients in hospitals. Finally, a logistic regression is performed to estimate the odds that a patient is readmitted depending on whether the HbA1c has been measured.

4. Data set transformation

The first operation on the available data has been to keep only the independent encounters. In the available data set, some patients had several encounters registered and these encounters can not be considered as independent. In such cases, only the first encounter of a patient has been kept. After this operation, the data set contains 69,973 independent encounters.

In order to study this data set using the tools available in the SciKit learn package, some features had to be transformed. In the case of the admission source ID for example, the ID is just an indication of the category of the admission source, and it shouldn't be interpreted as a continuous variable by the classifiers available in SciKit learn. Each category of admission source ID has hence been considered a feature (for example "is admission source ID 2") that can take either the value 1 (if the admission source ID is 2) or 0 (otherwise). The same has been done for the following features:

- race (split between 'African American', 'Asian', 'Caucasian' or 'Other')
- age (split between 10 age categories)
- admission type (6 categories)
- discharge disposition (19 categories because 6 discharge disposition identifiers were not used in the data set, so no category has been created for them)
- admission source (17 categories because 9 admission source IDs were not used in the data set)
- specialty of the admitting physician (8 categories)

Some features have also been simplified for the needs of this study:

- the measurement of HbA1c is either performed ('1') or not performed ('0')
- the glucose serum test is either performed ('1') or not performed ('0')
- each of the diabetic medications are either given ('1') or not ('0') without taking into account the changes in dose

5. Feature reduction

The data set under study is characterized by a large number of available features (93 after the operations described above). The first step in this analysis is to determine the most relevant features for the readmission rate.

The first attempt of dimensionality reduction has been performed using a principal component analysis (PCA). The goal was to find the 2 most discriminating dimensions and to visualize the separation in two dimensions. The output of this analysis can be seen on Fig. 1 and it shows that the PCA can not offer a satisfying separation. The same strategy has been used with more dimensions, without any conclusive results. A linear treatment does not enable to separate the data so a more complex approach is adopted by training a random forest classifier on the data.

In order to avoid any correlations between the estimators of this classifier, each tree is grown on a bootstrap sample. The data is split between a training sample and a test samples in the proportions 70-30%. Two classes are used for the result of the classification: readmitted within 30 days or not. 80 estimators are used and the score on the test sample is 0.91 with a very low correlation between trees (0.04 on average). The random forest classifier allows to rank the features by order of importance in

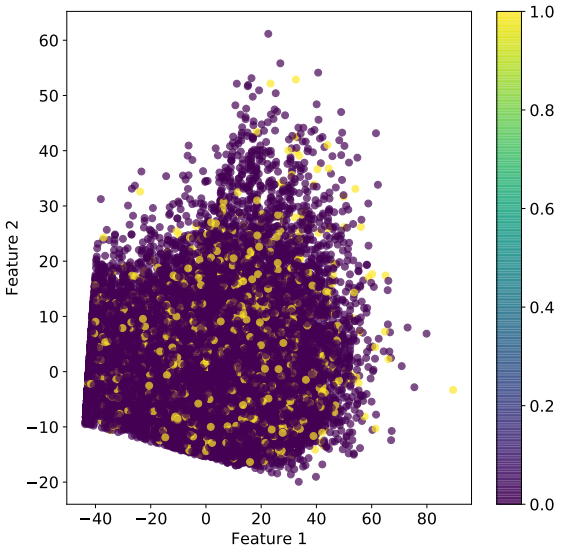


Figure 1 – 2D representation of the data set along its 2 most discriminating components.

the final decision, as shown in Fig. 2. This shows that the ten most important features related to the readmission of a patient are the number of laboratory procedures, the number of medications given, the time spent in hospital, the number of procedures, the number of diagnoses, the gender of the patient, the change in medication during the stay, the measurement of HbA1c, the administration of metformin and the administration of insulin.

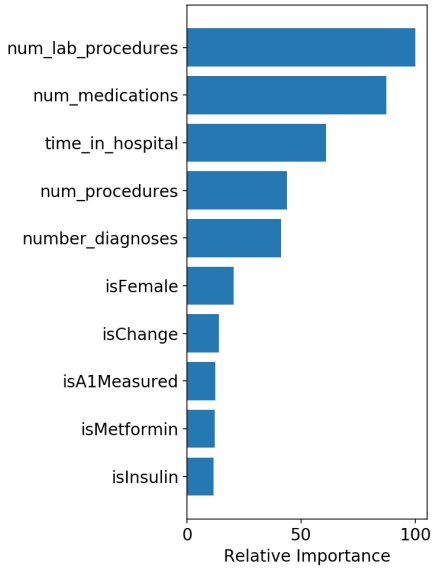


Figure 2 – Relative importance for the 10 most important features in a random forest classifier trained on 70% of the data using 2 output classes (readmitted or not).

6. Data visualization

The correlation between the 10 most important features determined previously is shown in Fig. 3. It appears that the most important correlations are between the number of medications given and the time spent in hospital, which is expected, and between the change in medication and the administration of insulin.

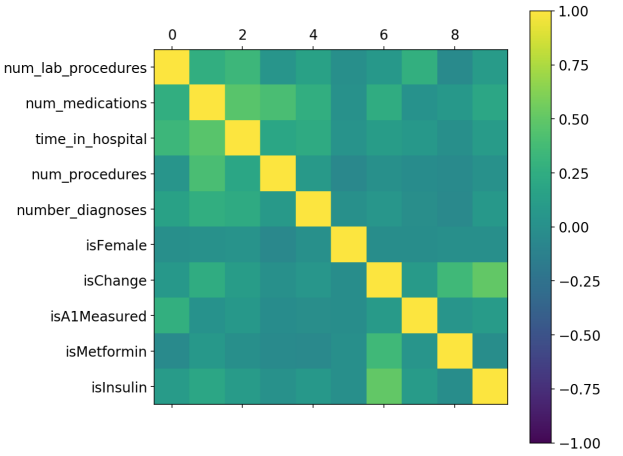


Figure 3 – Correlations between the 10 most important features in determining whether a patient is readmitted or not (as determined by a random forest classifier).

The histograms showing the distributions of the main features in the total data set are shown on Fig. 4. We can see that most of the time, only one lab procedure is conducted, and that the second most probable value is 43. The most probable value for the number of medications given is 13, for the time in hospital is 3 days, and for the number of non laboratory procedure is 0.

The ten most important features will be used (along with the age categories) to determine the probability that a patient is readmitted.

7. Logistic regression

The goal of this study is to model the probability of readmission of the patients depending on the values of the ten features extracted above, as well as on the age categories of the patients. A logistic regression is thus performed on a training sample representing 70% of the data. The score of this logistic regression on the test sample (30% of the data) is of 0.91. This logistic regression allows to obtain the probabilities of readmission as a function of the features described above. The distributions of each of the features shown in Fig. 5 and 6 are obtained by integrating over the values of all the other features. It tends to show that readmission is linked to:

- A high number of laboratory procedures
- A high number of medications given
- A long time spent in the hospital
- A high number of diagnoses
- A higher age

– A level of HbA1c not measured

– Metformin not administered to the patient

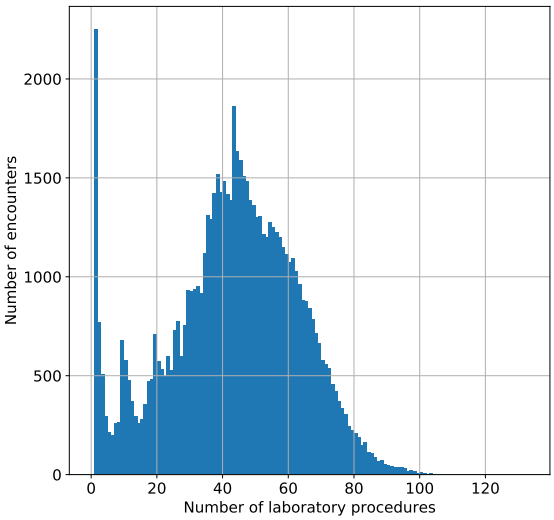
From these results, it seems that the measurement of HbA1c is linked to a lower rate of readmission, but the confidence interval is not given in the usual tools available in SciKit learn. To access the level of confidence given by the model, the module 'statsmodels' has been used and another logistic function has been trained on the same training data set. The results of this regression indicate that the odds that a patient for whom the HbA1c has been measured to control the glucose level is 88% as likely to be readmitted as a patient for whom the HbA1c has not been measured, with a 95% confidence interval between 82% and 94%. Similar intervals of odds of readmission are found for the other parameters and would deserve further studies, but the procedure is exactly similar.

8. Conclusion

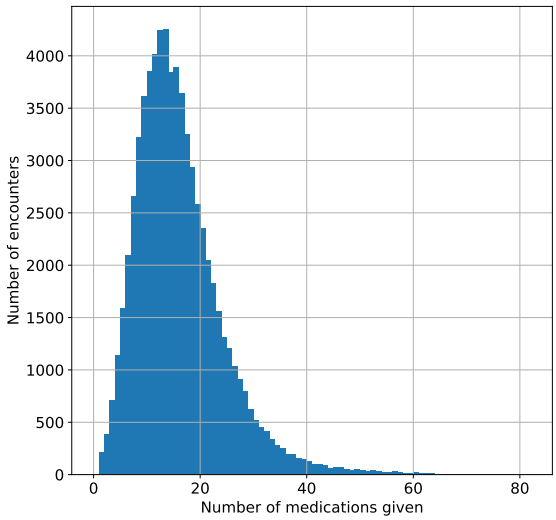
The data concerning diabetic patients from 130 US hospitals has been analysed to determine the factors contributing to the readmission of diabetic patients within less than 30 days. After a feature selection using a random tree classifier, a logistic regression has been trained to model the data. In particular, the results of this logistic regression indicates that when the HbA1c is measured, the patients are significantly less likely to be readmitted in hospital. This suggests that this measurement should be performed more consistently in the case of admission of diabetic patients. More studies could be performed but go out of the scope of this time limited exercise, which still allowed to put in place the procedure for this analysis in pedagogical steps.

References

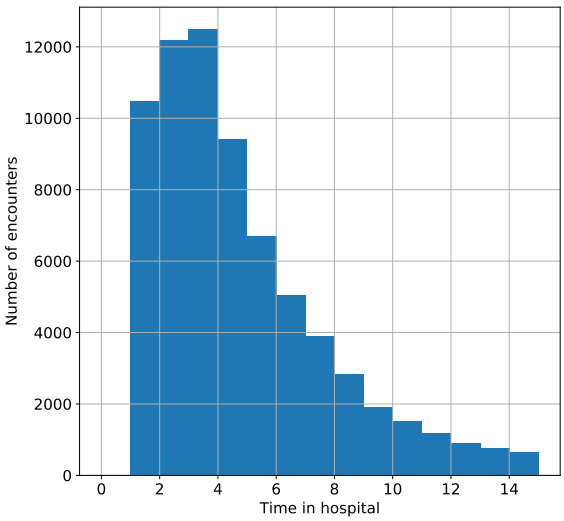
[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. doi:10.1155/2014/781670



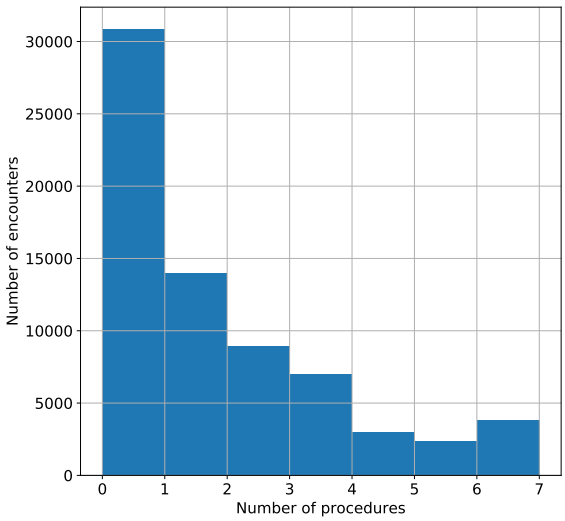
(a) Number of laboratory procedures



(b) Number of medications

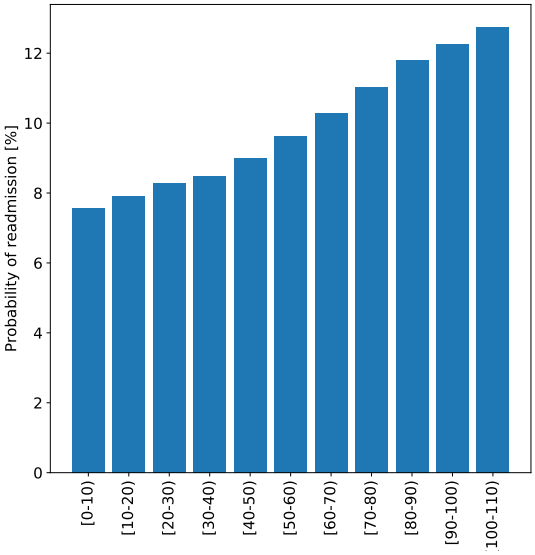


(c) Time in hospital (in days)

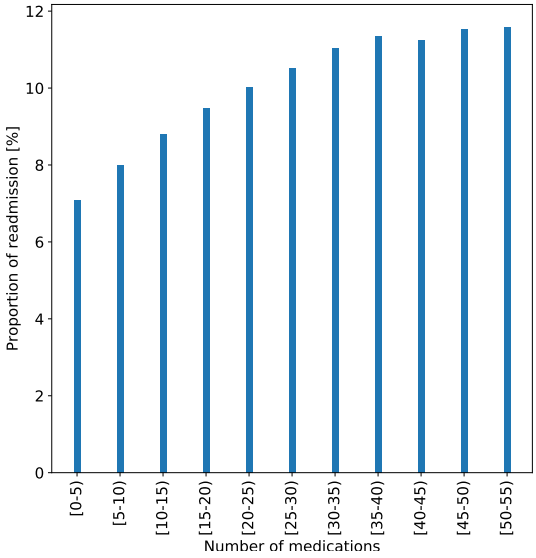


(d) Number of medical procedures

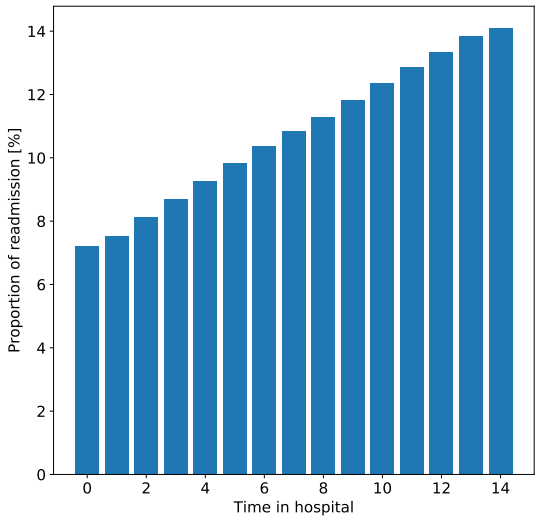
Figure 4 – Distributions in data of the four most important features.



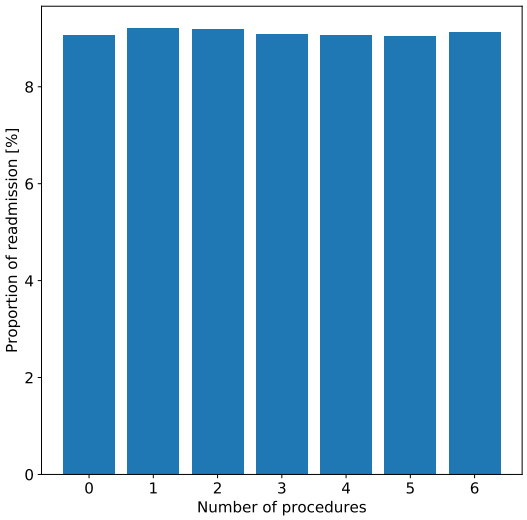
(a) Number of laboratory procedures



(b) Number of medications

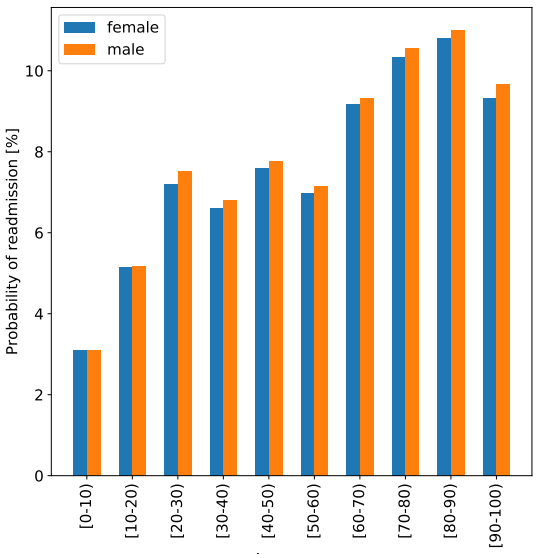


(c) Time in hospital (in days)

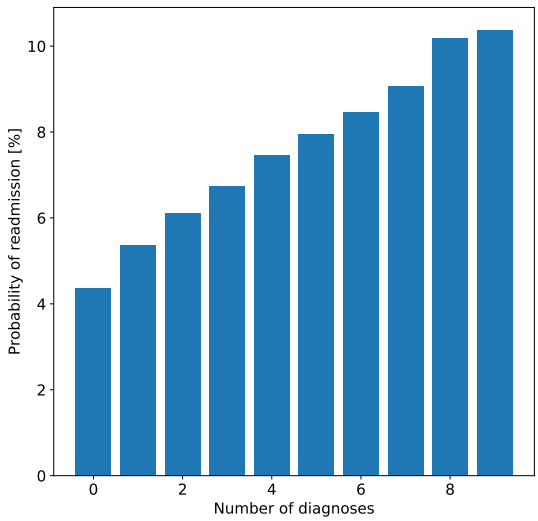


(d) Number of medical procedures

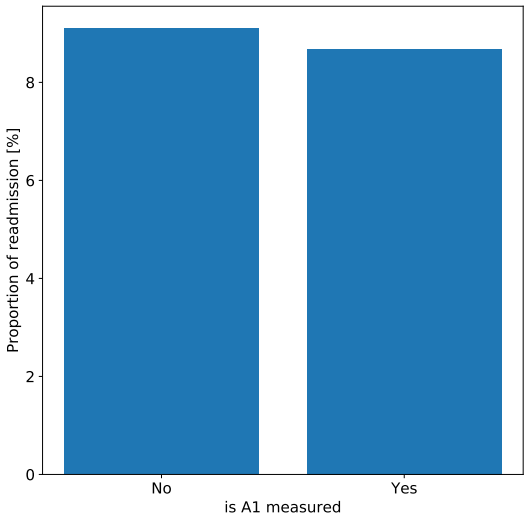
Figure 5 – Probability of readmission as a function of encounter features as given by the logistic function fitted on data (the projection is made by integrating over the value of the remaining features).



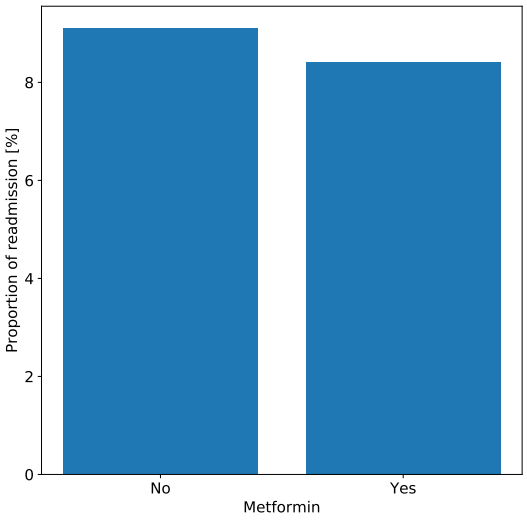
(a) Age and gender



(b) Number of diagnoses



(c) Measurement of HbA1c



(d) Administration of metformin

Figure 6 – Probability of readmission as a function of encounter features as given by the logistic function fitted on data (the projection is made by integrating over the value of the remaining features).