Homework Assignment 2

Readmission Prediction

November 22, 2018

1 Motivation

Healthcare has become one of the largest industries globally, and as such, it consumes a large amount of resources. In recent years hospital readmission has become a major topic of discussion in the U.S. healthcare system due to significant unnecessary costs associated with it. In 2004 about one-fifth of the Medicare beneficiaries were readmitted to hospitals within 30 days of discharge. It was estimated that the unplanned readmission of Medicare patients cost $17.4 billion.

Many of the preventable readmissions were related to low quality of care during patient stays in the hospital, as well as to poor arrangements of the discharge process.

It is advantageous for hospitals to reduce their readmission rates by using effective and efficient interventions during patient stays and the discharge process.

Considering that healthcare resources (including physicians, nurses, and other medical resources) are very costly and limited, it is impractical and inappropriate for hospitals to provide equal efforts and interventions for all patients. Therefore, a prediction model that can be used to identify high-risk patients in advance could greatly benefit healthcare providers by enabling them to target resources on risky patients and, by extension, reduce the overall readmission rate.

2 Task Details

In this assignment, we will try to find an advanced data mining technique to predict hospital readmission in our given dataset.

2.1 Dataset Description

The dataset we provide can be downloaded from `https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#`. The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

- It is an inpatient encounter (a hospital admission).
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter.

The dataset contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. The detailed description can be found in Table 1.

2.2 Task Description

It can be seen as a multi-class classification task. Which the attribute readmitted is our label, the readmitted feature is divided into three kinds: "NO"(which means, the patient never return to hospital), < 30(which means the patient return to hospital in 30 days) and > 30(which means the patient return to hospital in more than 30 days). You can use other about forty features to predict the label. The dataset should be divided into train

Table 1: Attribute Information

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: $[0;10);[10;20);\cdots;[90;100)$	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the en- counter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: "$>$ 200," "$>$ 300," "normal," and "none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: "$>$ 8" if the result was greater than 8%, "$>$ 7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: "$<$ 30" if the patient was readmitted in less than 30 days, "$>$ 30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.	0%

set, test set(if you wanna use validation set, just use it!), and final result should be calculated under the condition of ten fold cross validation.

To prove effectiveness of your model you should calculate overall accuracy, MACRO-F1 score on the test set, and corresponding confusion matrix needed be provided. Next we will intro the computing method of MACRO-F1, for a multi-class problem, TP_i is class i's True Positive, FP_i is class i's False Positive, TN_i is class i's True Negative, FN_i is class i's False Negative. Then, we can calculate every class's *precision*,

$$precision_i = \frac{TP_i}{TP_i + FP_i}$$

Macro precision is the average of all classes' precision,

$$precision_{macro} = \frac{\sum_{i=1}^{N} precision_i}{N}$$

Analogously, we can calculate each class's recall respectively,

$$recall_i = \frac{TP_i}{TP_i + FN_i}$$

Macro recall is the average of all classes' recall,

$$recall_{macro} = \frac{\sum_{i=1}^{N} recall_i}{N}$$

And finally, MACRO-F1's computational formula is :

$$F1_{macro} = 2\frac{recall_{macro} \times precision_{macro}}{recall_{macro} + precision_{macro}}$$

And for more detailed information, make a good use of SEARCH ENGINE, some websites like Google Scholar or DBLP always are useful and pay attention to most recent published papers in KDD, IJCAI, AAAI, NIPS .etc.

3 What You Should Do

Of course, you have to build your own models first, before it, you may need to preprocess your data, handle with missing values, convert the

attributes which can't be feed into specific machine learning methods directly, and feature engineering is always useful for this kind task. You can try different methods Prof. Wang has taught in classes or other state of art methods, don't forget BETTER PERFORMANCE, BETTER GRADE. When you submit your assignment, you should provide a runnable source code(a detailed README.MD or README.txt is needed) which TA can easily run(if not, reduction of points is inevitable), a final report and class presentation. Good luck.

4 Challenges and Bonus

As we know, different features contribute differently to the results, the importance of different features is also valuable for doctors in our task. If you can find a good way to explain the importance of different features, you can get some points up to 10 points.