108 Advanced data analysis Final project

Description

This data represents ten years of clinical are at one hundred thirty hospitals and integrated delivery network. It includes over fifty features representing patient and hospital result. Because patients have the opportunity to be re-admitted after being discharged again, it’s helpful that if we can predict this patient whether be re-admitted and when the patient will be re-admitted. Therefore, your task is to predict each patient whether be re-admitted.

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	Object	Values: Caucasian, Asian, African American, Hispanic, and other, total: 6	2%
Gender	Object	Values: male, female, and unknown/invalid	0%
Age	Object	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Object	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Object	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Object	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	Object	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical	Object	Integer identifier of a specialty of the	53%

specialty		admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	
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 encounter	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	Object	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Object	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Object	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	Object	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	Object	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	Object	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Object	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Object	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	Object	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%

File descriptions

- sample_submission.csv – a sample submission file in the correct format.
- test_dataset.csv – the feature of test dataset.
- train_dataset.csv – the feature of train dataset.
- train_label.csv – the label of train dataset.

Performance metric

$$Accuracy = \frac{Number\ \ of\ \ correct\ \ predictions}{Total\ \ of\ \ number\ \ of\ \ predictions}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Submission

- Submit 5 times / day.
- Finally, you should choose the best 5 csv to be your final result.
- 1st checkpoint is 12/17 (Tue) 0:00

  2nd checkpoint is 12/24 (Tue) 0:00

  3rd checkpoint is 12/31 (Tue) 0:00

  Final checkpoint is 1/7 (Tue) 0:00

  The performance of the first three checkpoints is based on 30% test data called public score; The final result is based on the other 70% test data called private score.

Reference

Introduction of caret package

http://topepo.github.io/caret/index.html

Model Training and Tuning in caret

http://topepo.github.io/caret/model-training-and-tuning.html

Available Models in caret

h1ttp://topepo.github.io/caret/available-models.html