Healthcare Analytics Assignment -- Readmission of Patients with Diabetes[1]

Problem Context:

Hospital readmission is a high-priority health care quality measure and target for cost reduction, particularly within 30 days of discharge (30-day readmission is typically concerning). Reducing readmission rates among patients with diabetes has the potential to greatly reduce health care costs while simultaneously improving care.

Assignment Goal:

Predicting 30-day readmission of patients with diabetes using data on prior medical encounters and patients historic clinical data.

Dataset[2]:

The dataset includes multiple encounters of over 70 thousand diabetes patients. It contains 101766 rows and 50 features including patient demographics, patient characteristics, medical history, drugs, and previous admission related information. See data dictionary for detailed information (Appendix A).

Data Preprocessing:

o   Missing values are removed from the data
o   Rows with missing values in the diag1, diag2, diag3, and gender are dropped.
o   Features with missing values are dropped from the dataset
o   The target variable is reduced into 2 classes (1- readmission within 30 days, 0- no readmission within 30 days) to indicate whether the patient had a readmission within 30 days of the previous encounter
o   Admission type, Admission source, and Discharge disposition Ids are mapped into their respective categories as mentioned in the data dictionary
o   A1Cresult values are translated to indicate levels
o   Service_utilization - Create a new feature which captures the sum of outpatient, emergency, and inpatient admission values
o   Drug_change - Another new feature that captures the total number of medications that were changed (up/down) in an encounter. It is the total count of all the drugs whose dosage were modified (either up or down) in the encounter
o   Multiple encounters of patients are removed (Only the first encounter of each patient is retained)
o   Final dataset has 67580 encounters of patients with 47 features
o   After one hot encoding - 63 features

---

[1] Prepared by Professor Ranganathan of the University of Illinois at Chicago, with research assistance from graduate students Abhilash Nagaraja and Rahul Gera.
[2] Original dataset can be obtained from UCI Machine Learning repository

Check-list:

Target label distribution post data processing

o   Readmitted – 6129 records

o   No readmission – 61451 records

Additional Steps:

- Values for the feature age are grouped into three categories – "<50", "50-70", and "70+" based on the data distribution
- All the diagnosis codes to be mapped to categories based on the conditions given below

Diagnosis Codes	Group Name	ID	Description
390–459, 785	Circulatory	1	Diseases of the circulatory system
460–519, 786	Respiratory	2	Diseases of the respiratory system
520–579, 787	Digestive	3	Diseases of the digestive system
250.xx	Diabetes	4	Diabetes mellitus
800–999	Injury	5	Injury and poisoning
710–739	Musculoskeletal	6	Diseases of the musculoskeletal system and connective tissue
580–629, 788	Genitourinary	7	Diseases of the genitourinary system
140–239	Neoplasms	8	Neoplasms
780, 781, 784, 790–799	Other	0	Other symptoms, signs, and ill-defined conditions
240–279, without 250			Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
680–709, 782			Diseases of the skin and subcutaneous tissue
001–139			Infectious and parasitic diseases
290–319			Mental disorders
E–V			External causes of injury and supplemental classification
280–289			Diseases of the blood and blood-forming organs
320–359			Diseases of the nervous system
630–679			Complications of pregnancy, childbirth, and the puerperium
360–389			Diseases of the sense organs
740–759			Congenital anomalies

- The primary diagnosis codes are used for modeling purposes i.e. diag_1
- Perform Exploratory data analysis to capture the relationship with the readmission rate of patients
- All the categorical features are one-hot encoded

Questions:
- After preprocessing the data, focus on exploring the dataset to understand the basic relationships among the predictor variables.
- Develop a machine learning model that can predict the readmission of patients with diabetes.
- Use multiple modelling methods (eg: regression-based approaches, ensemble approaches etc) and compare the model performance. Use cross-fold validation,
- Use Alteryx for entire exercise.

Appendix A – Data Dictionary

Column Name	Feature	Description
encounter_id	Encounter ID	Unique identifier of an encounter
patient_nbr	Patient Number	Unique identifier of a patient
race	Race	Values: Caucasian, Asian, African American, Hispanic, and other
gender	Gender	Values: male, female, and unknown/invalid
age	Age	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)
admission_type_id	Admission Type	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
discharge_disposition_id	Discharge Disposition	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
admission_source_id	Admission Source	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
time_in_hospital	Time in hospital	number of days between admission and discharge
num_lab_procedures	Number of lab procedures	Number of lab tests performed during the encounter
num_procedures	Number of procedures	Number of procedures (other than lab tests) performed during the encounter
num_medications	Number of medications	Number of distinct generic names administered during the encounter
number_outpatient	Number of outpatient visits	Number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Number of emergency visits	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Number of inpatient visits	Number of inpatient visits of the patient in the year preceding the encounter
diag_1	Diagnosis 1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
diag_2	Diagnosis 2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
diag_3	Diagnosis 3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
number_diagnoses	Number of Diagnoses	Number of diagnoses entered to the system
max_glu_serum	Glucose serum test result	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured
A1Cresult	A1c test result	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.
metformin		
repaglinide		
nateglinide	24 features for medications	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
chlorpropamide	Readmitted	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.
glimepiride	Drugs	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"
acetohexamide		
glipizide		
glyburide		
tolbutamide		
pioglitazone		
rosiglitazone		
acarbose		
miglitol		
troglitazone		
tolazamide		
insulin		
glyburide-metformin		
glipizide-metformin		
glimepiride-pioglitazone		
metformin-rosiglitazone		
metformin-pioglitazone		
change		Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"

diabetesMed	Diabetes medications	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
service_utilization	Utilization	Total sum of inpatient, Emergency visits
drug_change	Change of medications	Total number of drug changes (up/down) prescribed for the patient
readmitted	Readmission	1- readmitted within 30 days, 0 - no readmission within 30 day