Capstone Two Proposal [Rayees Ahamed B]

Title: Diabetes Patients' Readmission Prediction

Problem statement:
What are the factors leading to high readmission of diabetes patients' to hospital within 30 days of discharge?

Context:
Diabetes is a chronic health disease caused due to excess glucose in the blood and failure of our pancreas to produce insulin properly. This leads to several health complications to patients such as cardiovascular risk, frequent infections, blindness (retinopathy) etc. The data from 130 US hospitals from 1999-2008 shows there is a high percentage of diabetes patients getting readmitted within 30 days of their last discharge. This poses the question what factors are responsible that makes patients' getting readmitted within a short time. Finding these factors will help in improving the quality of care in diabetes patients', lowering the medical cost involved and identifying medicines effective in treating diabetes.

Criteria for success:
1. Identifying one or more attributes that caused diabetes patients' getting readmitted
2. Understanding the relationship between different attributes
3. Key medicines used in treatment that helped patients' not getting readmitted

Scope of solution space:
Finding the attributes highly correlated in the patients' who are frequently admitted. Finding the features that correlated in the patients' who are not frequently admitted. Understanding the epidemiology of different diabetes patients (age, gender, race) that are prevalent to readmission. Medicines that are successfully treated and helped patients' in avoiding hospital readmission.

Constraints within solution space:
Some attributes may lack information in all the patients. Due to the large number of categorical variables, it may be difficult to use those variables in prediction. Available data may not be sufficient to arrive at a conclusion, and may require additional data to make a clinical decision. Final model can only be used for research purposes and should not be used for diagnostic or treatment purposes.

Stakeholders to provide key insight:

1. Dipanjan Sarkar - Data Scientist & Mentor at Springboard
2. Rayees Ahamed B - Springboard Data Science Student

Key data sources:
https://www.kaggle.com/saurabhtayal/diabetic-patients-readmission-prediction

Data can be downloaded from the above kaggle page. Data originally deposited in UCI Machine Learning repository. This data derived from the published article named 'The relationship between diabetes mellitus and 30-day readmission rates' from Clinical Diabetes and Endocrinology journal.

Dataset Name: Diabetes 130-US hospitals for years 1999-2008 Data Set

Feature 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), . . ., [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 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	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%

Dataset information:

- Number of instances: 101767
- Number of attributes: 50
- Number of independent variables: 49
- Number of dependent variable: 1
- Number of numerical variables: 17
- Number of categorical variables: 33

Variable identification:
***1. Independent variables (49)*:** encounter*id, patient*nbr, race, gender, age, weight, admission*type*id, discharge*disposition*id, admission*source*id, time*in*hospital, payer*code, medical*specialty, num*lab*procedures, num*procedures, num*medications, number*outpatient, number*emergency, number*inpatient, diag*1, diag*2, diag*3, number*diagnoses, max*glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed.

***2. Dependent variable (1)*:** readmitted (Categorical)

Initial problem solving approach:

1. **Data wrangling**
    - Exploring the dataset
    - Grouping attributes into categories:
        - Epidemiology data - e.g. age, weight, gender, race
        - Hospital admission data - e.g. admission, discharge, time in hospital, readmitted
        - Diagnosis data - e.g. number of laboratory procedures, diagnosis data
        - Treatment data - e.g. medicines administered
    - Finding and filling missing values
    - Exploring numerical and categorical variables
    - Converting data types of attributes

2. **Exploratory Data Analysis (EDA)**
    - Exploring each attribute of dataset
    - Identify relationship between different features
    - Visualize the features
    - Visualize the high dimensional data
    - Feature engineering and correlation maps

3. **Preprocessing and training**
    - Split dataset into train and test sets
    - Metrics and summary statistics
    - Testing initial models
    - Develop pipelines
    - Assess different models

4. **Modelling, testing and deployment**

Deliverables:

- Slide deck containing each step summary
- Project report in a PDF format
- Github repository with code notebooks