Predicting early readmissions for diabetes patients using clinical records from 130 US hospitals

Capstone Project 1 for Springboard Program

Problem statement

- Although diabetic patients represent about 8% of the US population, they account for 23% of hospitalizations (8.8 million) each year.
- Hospitalization is major cost driver for diabetes patients.
- Readmission after a hospital discharge is a high-priority healthcare quality measure and target for cost reduction.

Who might care?

- Physicians, healthcare researchers, and health plan insurance companies might be interested in using a model to predict the readmission outcome for diabetic patients.
- Based on the predicted probability of early readmission, early interventions can be done to patients such as
  - discharge education;
  - further lab testing to ensure stabilized glucose level; and
  - home care reminders of medications.

Data Source

- The dataset for this study was obtained from the Health Facts database, which was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States.
- It was later adjusted to suit for a diabetes research and was released on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University on UCI Repository.

Description of the dataset

Table 1. Meta Descriptions for 50 Variables (N=101,766)

Number of Variables	Class	Example
2	identifications	encouter id
5	patient demographics	race
3	admission/discharge status	admission scource
7	numeric values of events	number of lab tests
24	medications	insulin taken or not
8	others	payer code
1	target outcome	readmission or not

Data Wrangling Steps

- **Missing Values**
  - 2 variables are filtered out due to too many missing observations:
    - Weight (numeric; 97% values missing); and
    - Payer code (17 categories; 40% values missing).

Data Wrangling Steps

- **Outliers**
  - Majority of the features are categorical, and
  - No extreme values were seen for numeric variables after checking their range.

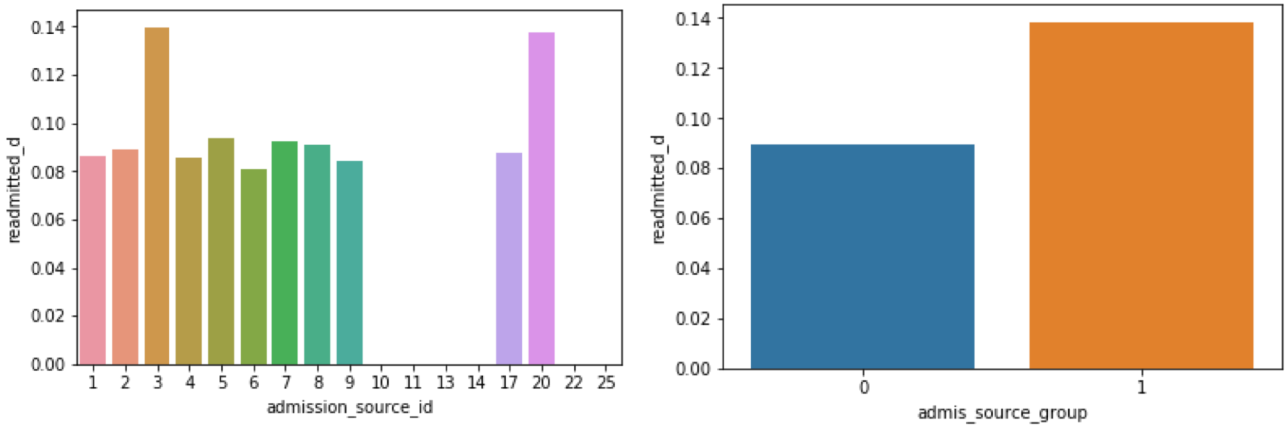
Data Wrangling Steps

- **Duplicate Patients and Deaths**
  - There are duplicate patient ids so the same patient may have multiple encounters in the dataset. Using results of same individuals tend to bias the true association so duplicates were dropped.
  - Additionally, patients who were discharged to hospice or discharged to a death were also deleted.
  - Finally, 69,970 encounters were left for the study.

Data Wrangling Steps

- **Transformation of Categorical Features**
  - Dummy features were created for categorical variables such as race and gender.
  - Several categorical variables were re-grouped based on their readmission rates visualized by bar plots (see the next slide).

Figure 1. Readmission Rates among Patients with Different Admission Sources (before and after regrouping)



Data Wrangling Steps

- **Creation of Features**
  - Medication changed or not based on 24 medication records
  - Charlson Comorbidity Score (CCI) to take into account the clinical characteristics (disease severity) of the patients, by grouping and mapping of ICD-9 codes to a CCI scores.

Inferential Statistics and Feature Selection

- **Dummy features**
  - 2 sample proportion z-tests were applied using the *proportions_ztest* from *statsmodels.stats.proportion*.
  - Features with p-value<0.05 were retained.

Inferential Statistics and Feature Selection

- **Categorical features**
    - Chi-squared test was chosen since we can compare if distribution of the observed counts in each class is significantly different than the expected counts (proportional to class counts).
    - Features with p-value<0.05 were retained.

Inferential Statistics and Feature Selection

- **Continuous features**
  - To test for significance impact of continuous features on binary outcomes, Logit model from *statsmodels.discrete.discrete_model* was applied.
  - Features with p-value<0.05 were retained.

Table 2. Included Significant Features and Test P-values

Features	p-values	Features	p-values
race_?	0.012	glipizide_Down	0.007
race_Caucasian	0.001	glipizide_No	0.006
race_Other	0.012	dischar_group	0.000
max_glu_dummy	0.014	ad_type_group	0.021
A1Cresult_dummy	0.011	med_spec_group	0.000
diabetesMed_d	0.000	admis_source_group	0.006
Alc_nonch	0.008	diag_group	0.000
med_up	0.004	time_in_hospital	0.000
med_down	0.000	num_lab_procedures	0.000
insulin_Down	0.000	num_procedures	0.000
insulin_No	0.000	num_medications	0.000
insulin_Up	0.008	number_emergency	0.008
metformin_No	0.001	number_inpatient	0.000
metformin_Steady	0.003	number_diagnoses	0.000
metformin_Up	0.040	CCI	0.000

Machine Learning and Prediction

- **Overview**
  - A supervised classifier is needed to train the data and predict the binary outcome.
  - In case the imbalanced data would affect the model performance especially in this study we are more interested in the minority outcome (<10% of patients were readmitted), over sampling will be applied.

Machine Learning and Prediction

- **Metrics selection**
  - People in healthcare industries would prefer a prediction model that can identify most of the early-readmission cases, and tolerate some extent of low prediction score. E.g., high recall and moderate precision.
  - Therefore, AUC score will be used to train models and recall will be evaluated among different probability thresholds

Machine Learning and Prediction

- **Different machine learning models**
  - Cross-validation scores and test score were returned for each model fit.
  - For the first *DecisionTree* model, the results are as followed:
    CV AUC Scores: 0.66091814, 0.64988475, 0.63991324, 0.64551209, 0.63448754
    Test AUC Score: 0.60903298
  - *RandomForest* did better as both the CV AUC and test AUC increased.

Machine Learning and Prediction

- **Parameter Tuning**
    - Parameter tuning was applied for the *logistic* regression and *Xgboost* regression models with the *GridSearchCV* approach.

Machine Learning and Prediction

- **Exploring the threshold**
  - Explore the probability threshold of the model to see if we can have more early readmission included without reducing too much AUC value.
  - Some threshold value with their AUC and confusion matrix are shown in the next slide

Exploring the threshold

******** For p = 0.46 ********
Our AUC is 0.614
[[9005 6930]
[523 1035]]

******** For p = 0.47 ********
Our AUC is 0.619
[[9455 6480]
[552 1006]]

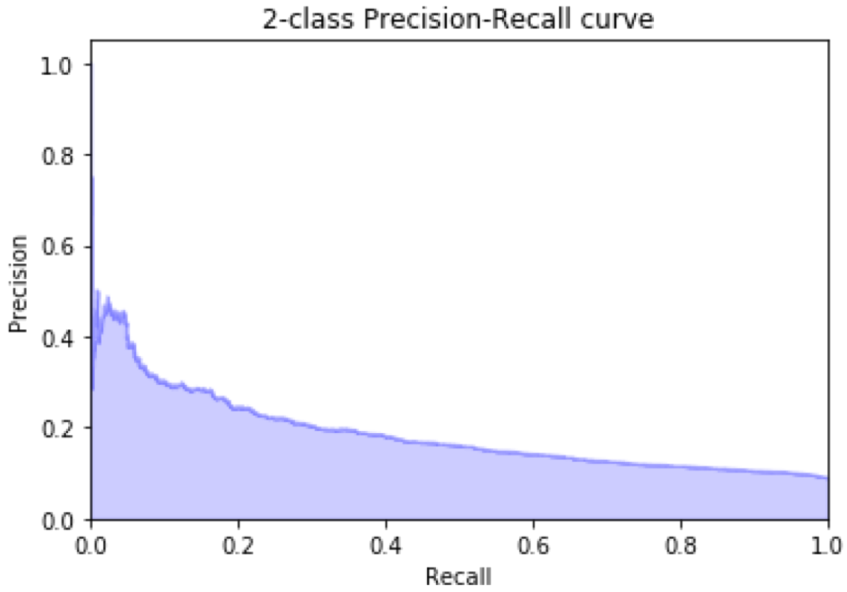
******** For p = 0.48 ********
Our AUC is 0.622
[[9912 6023]
[588 970]]

******** For p = 0.49 ******
Our AUC is 0.621
[[10358 5577]
[635 923]]

******** For p = 0.5 ******
Our AUC is 0.621
[[10784 5151]
[676 882]]

******** For p = 0.51 ******
Our AUC is 0.619
[[11207 4728]
[722 836]]

Figure 2. Precision-Recall Curve for Different Thresholds



Conclusions

- **Recall vs. Precision:** even the recall is 0, the precision cannot be improved much. So we can bear to lower the precision, which means we can bear with predicting more patients as early-readmitted who actually would not, so that the recall will be higher.
- **A lower threshold** (<0.5), e.g., a threshold of 0.45 would identify around **2/3** of the early-readmitted patients with this model;
- Among all the patients predicted with this model as early-readmitted in the future, **7/8** of them will actually not be early-readmitted.
- **Further screening process** can be applied to these patients to filter out the right person that will have early-readmission.