DS7333 Quantifying the World: Case Study 2
Rick Fontenot, Feby Thomas Cheruvathoor

1 Introduction

The goal for this case study is to build a logistic regression model predict which hospital patients are likely to be readmitted within 30 days or longer term and evaluate which variables carry the most importance for determining readmission.

The insights generated could potentially help hospitals plan resources such as staffing and rooms, medical insurance companies in determining risk of increased costs, or patients and doctors plan for what is likely for future health risks.

Monetary data elements such as price details, cost of the medicine and procedures, the fee for certain procedures have not been included. Based on the type of information provided we are approaching this analysis with the intent of helping hospitals plan resources. While some information such as the patients race and gender have been provided and could present ethical issues if used to set conditional pricing or denial of treatment, we are utilizing this information since it does not bias decisions on hospital resource planning.

2 Methods

Data description

The raw data set contains 101,766 observations of patients with 47 features, two identifiers, plus the target variable including readmission outcomes. The features include age, race, gender, diagnoses, medications, limited test results, and information on length of stay, and the number of inpatients, outpatient and emergency room stays.

For a more detailed description of each feature, see appendix section 1.

Missing Values

Missing values in many features has been presented as "?" through out in the data set as the non-available data point. This has been replaced with "Nan" in python to summarize the issue and determine how best to impute missing values.

Feature	Count	% Missing
weight	98,569	97%
medical_specialty	49,949	49%
payer_code	40,256	40%
race	2,273	2%
diag_3	1,423	1 %
diag_2	358	<1%
diag_1	21	<1 %

Figure 1: Summary of features with missing values

We determined the best method of imputation for these variables prior to modeling were:

weight: With such a large percentage missing we dropped this feature. We explored imputing by using age but this would cause a multicollinearity issue and also observed that the small amount of weight records included seemed unreliable, for instance a large percentage of patients in the 60-80 year range had weight listed as less than 25 pounds

medical_specialty: With 73 unique values and no other features to help impute missing records, we filled the NA's with "missing" to preserve the rows from being dropped from modeling. Without know the importance of this feature we could still gain prediction possibilities from other features on these records. Including the "missing" category will also help make predictions on future unseen data that has this field missing

payer_code: Similar to medical specialty with 18 unique values and other obvious feature to assist in filling the blanks, we filled the NA's with "missing" to preserve the rows from being dropped.

race: while we considered replacing missing values with the mode or some sampling based on the distribution, in the end we handled the same as other features and filled with "missing" This could be an important feature indicating it is missing due to patient privacy, race other than what's available on the check boxes or other reasons that could cause readmission outcomes to be different for this subset.

diag_1, diag_2, diag_3: With >800 unique values and no other obvious feature to assist in filling the blanks, we filled the NA's with "missing" to preserve the rows from being dropped.

Duplicate Observations

There were no duplicates rows are observed in the data set.

Variable Type and Formatting

The raw data contains multiple categorical features, some of which are in numeric formats, and some which could benefit from level reductions through regrouping. We re-coded the following variables prior to modeling:

age: Originally binned in the format '[0-10)' Since age is ordered information, rather than one-hot encode this feature we used the upper bound of the bins such as 10, 20, 30, etc. and modeled this feature as numeric

diag_1 , diag_2, diag_3: with >800 unique categories including both numeric and text formats, we grouped these into their broader IC9 categories (International Statistical Classification of Diseases and Related Health Problems). In addition to the broad 18 IC9 categories we broke out diabetes which is of particular interest from the other diagnoses within the immunity disorder group. We also preserved the "missing" values rather than include them in another category. For more information on these groupings, see appendix section 2.

max_glu_serum: This is a test with feature values: ">200," ">300," "normal," and "none" if not measured. Rather than one-hot encode we treated this feature as numeric with the ordering 0=None, 1=Normal, 2=>200, and 3=>300 with the idea that severity of results should be ordered.

A1Cresult: This is a test with feature values: ">200," ">300," "normal," and "none" if not measured. Rather than one-hot encode we treated this feature as numeric with the ordering 0=None, 1=Normal, 2=>7, and 3=>8 with the idea that severity of results should be ordered.

Medications: All 24 of these features treated this feature as numeric with the ordering 0=Not taken, 1=Dosage Down, 2= Dosage Steady, and 3= Dosage Up, with the idea that change of dosage should be ordered.

encounter_id & patient_nbr: These identifier features were dropped from modeling set as they are unique and not predictors of health status

One-hot encoding: After all processing summarized above the following categorical variables were one-hot encoded for modeling: 'race', 'gender', 'admission_type_id', 'discharge_disposition_id', 'admission_source_id', 'payer_code',' medical_specialty', 'change',' diabetesMed',' diag_1_group', 'diag_2_group', 'diag_3_group'

Target Variable & Class Imbalance

The target variable "readmission" has the categories: patients readmitted in less than 30 days, patients readmitted in more than 30 days, and patients not readmitted. The time frame of data was not provided, so it is unclear if the non-readmissions or the more than 30 days are bound on the upper end by 60 days, 90 days, or multiple years etc.

There is significant class imbalance in the target with just 11% of the observations in the readmission less than 30 days category.

Readmission	Count	Percentage
NO	54,864	54%
> 30 days	35,545	35%
< 30 days	11,357	11%

Figure 2: Summary of observations in each readmission category

In order to deal with effects of this imbalance we used stratified splits so that the training and validation sets have approximately equal proportions of the three readmission classes. See our cross-validation method description below for more details.

In addition to using stratified splits, we evaluated additional models using SMOTE from the python imblearn.over_sampling package to resampling and balance the observations in each class of a training set.

Models of the original training set and the resampled training set will be compared for both performance and feature importance with particular attention to the validation set to ensure the resampled model is not overfitting

Expired Patients

Exploratory data analysis shows that 3 categories (11,19,20) had zero patients readmitted. These discharge categories represent "expired" patients. Since death is a 100% certainty of not being readmitted, these rows were deleted from our modeling data set so that the weights and importance of other features are not affected.

Upon this realization we discussed also removing records who were discharged to hospice as this is typically an end-of-life setting. (discharge codes 13,14). However the figure below shows that some hospice patients to get readmitted to the hospital, so the records were included in our models.

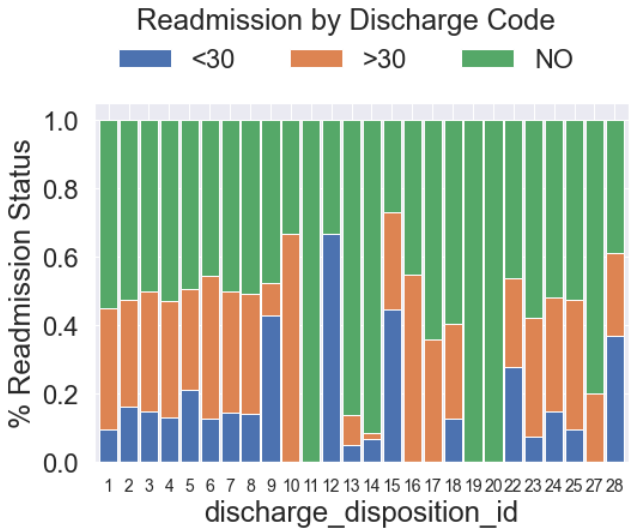

Figure 3: Readmission Rates by discharge type – expired patients 0%

Standardizing data

In order to determine which features are most important to predicting the critical temperature, we scaled the data so that the regression weights can be compared. We scaled all features to a mean of zero scaled to unit variance. Scikit learn StandardScaler function has been used to scale the data points.

Cross Validation

In order to compare the models for overfitting, we first set aside 10% of the data into a validation set (stratified shuffled data to get random observations with similar proportion of records in the target classes). After our model hyperparameters are tuned on the training set, we will compare performance metrics of the model on this unseen data.

For the training data used to build models, we used StratifiedKFold cross validation with 10-folds for cross validation and the mean accuracy across the folds for model tuning. In

addition to accuracy, we will also summarize the precision and recall for each class, for comparison.

Particular attention will be focused on the validation set results to check for overfitting on the re-sampled data model

Logistic Regression modeling

For prediction of classifications we used the sklearn LogisticRegression package with the "saga" solver so that elastic net regularization could be used to minimize risks of overfitting.

Grid searches were used to tune hyperparameters individually for models using the original records versus the SMOTE resampled records.

3 Results

The optimized base model using original observations for training was able to classify the 3 classes with an overall accuracy of 58%. As expected, the majority class had higher precision and recall than the minority classes. The readmitted less than 30 days class particularly had extremely poor recall at 2% and concerningly low 4% f1-score.

The second model with resampling to balance the observation counts by class, significantly improved the minority class recall to 42% and f-1 score to 26% however it came at the expense of overall accuracy which dropped to 49%.

Classification Performance Summary

	Original Raw Records				Resampled Records with class balancing			
	precision	recall	f1-score	support	precision	recall	f1-score	support
Not_Readmitted	0.6	0.87	0.71	5321	0.66	0.59	0.62	5253
More_than_30	0.52	0.32	0.4	3555	0.48	0.36	0.41	3550
Less_than_30	0.32	0.02	0.04	1136	0.19	0.42	0.26	1132
accuracy			0.58	10012			0.49	9935
macro avg	0.48	0.4	0.38	10012	0.44	0.46	0.43	9935
weighted avg	0.54	0.58	0.52	10012	0.54	0.49	0.51	9935

Figure 4: Classification Performance metrics for original and resampled training data models

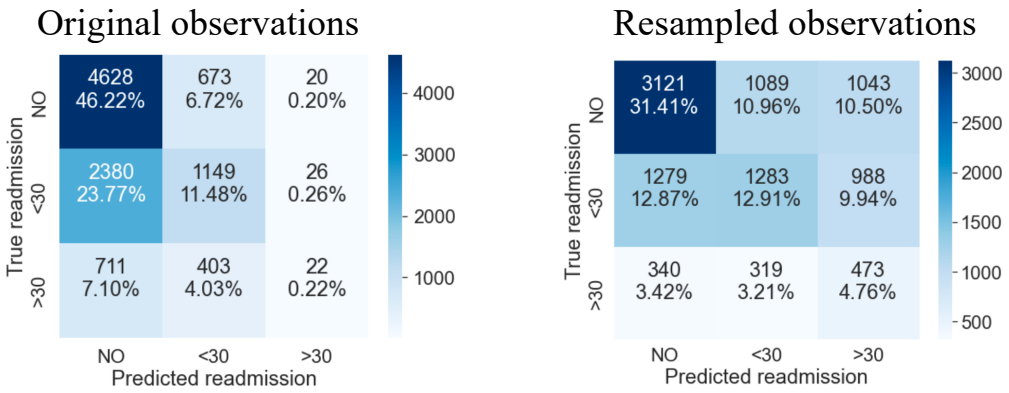

Figure 5: Confusion Matrix for original and resampled training data models

Based on the feedback from the hospitals resource planning team, we could choose the model more appropriate based on their priority for overall accuracy vs. the precision and recall for predicting the amount of patients expected to be readmitted in less than 30 days.

Both models have similar rankings of the most important features, so the driving factors can be understood with either model:

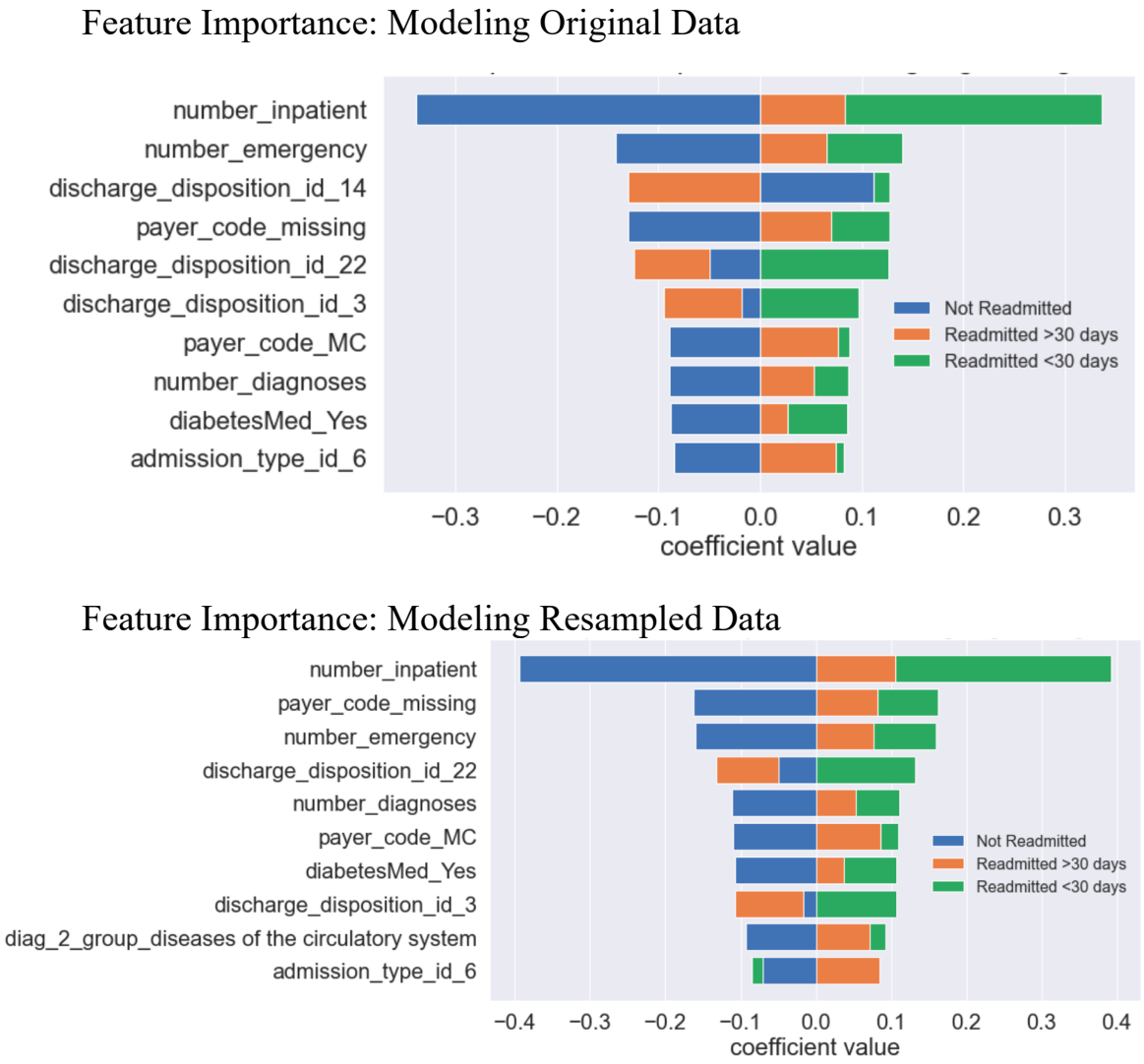

Figure 6: Ranking of most important features

The number of inpatient visits of the patient in the year preceding the encounter are the most important factor determining the likelihood of readmission, followed by the number of emergency visits of the patient in the year preceding the encounter

4 Conclusions

Based on this modeling, independent of the specific medical diagnoses, the biggest predictors of future hospital readmission are the amount of recent inpatient visits and recent prior emergency room visits.

Higher numbers of inpatient visits prior to hospitalization strongly correlate to the short term readmission rates (<30 days) but are not as strongly correlated to longer term future readmissions:

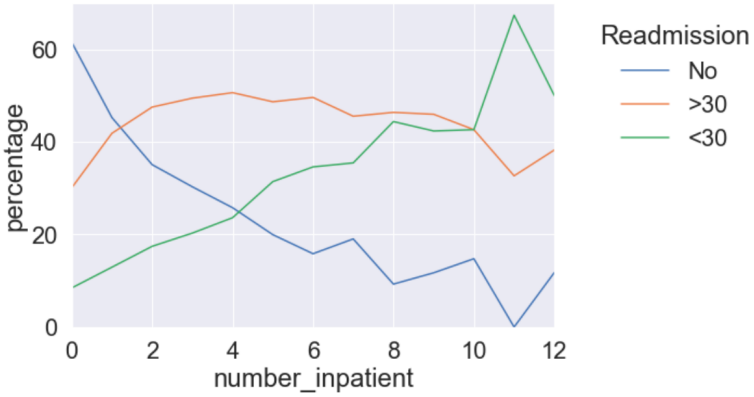

Figure 7: Rates of readmission vs amount of inpatient visits year prior

Higher numbers of emergency visits prior to hospitalization strongly correlate to both the short term and long term readmission rates:

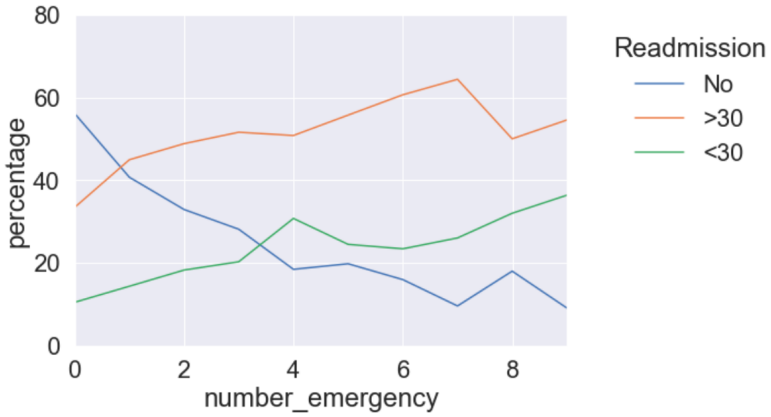

Figure 8: Rates of readmission vs amount of emergency visits year prior

While these insights are generally useful information, the poor accuracy of the model may provide challenges in guiding hospital resource planning. The modeling suggests that there are other significant factors affecting the causes of readmission beyond what is available in this feature set. Working with a team of domain experts to collect additional relevant information could be helpful in attempts to improve future modeling.

Appendix

I. Variable descriptions

Variable	Description
race	Values: Caucasian, Asian, African American, Hispanic, and other
gender	Values: male, female, and unknown/invalid
age	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
admission_type_id	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
discharge_disposition_	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
admission_source_id	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
payer_code	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
medical_specialty	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
diag_1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
diag_2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
diag_3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
max_glu_serum	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
A1Cresult	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	medication 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
repaglinide	medication 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
nateglinide	medication 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	medication 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
glimepiride	medication 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
acetohexamide	medication 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

Variable	Description
glipizide	medication 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
glyburide	medication 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
tolbutamide	medication 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
pioglitazone	medication 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
rosiglitazone	medication 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
acarbose	medication 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
miglitol	medication 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
troglitazone	medication 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
tolazamide	medication 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
examide	medication 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
citoglipton	medication 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
insulin	medication 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
glyburide-metformin	medication 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
glipizide-metformin	medication 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

Variable	Description
glimepiride-pioglitazor	medication 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
metformin-rosiglitazor	medication 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
metformin-pioglitazon	medication 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
change	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
diabetesMed	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
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.
encounter_id	Unique identifier of an encounter
patient_nbr	Unique identifier of a patient
time_in_hospital	Integer number of days between admission and discharge
num_lab_procedures	Number of lab tests performed during the encounter
num_procedures	Number of procedures (other than lab tests) performed during the encounter
num_medications	Number of distinct generic names administered during the encounter
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter
number_diagnoses	Number of diagnoses entered to the system

II. Diagnosis categories and descriptions

We utilized the diagnosis code and descriptions here https://en.wikipedia.org/wiki/List_of_ICD-9_codes to create the rules below for grouping and reducing the categories for features diag_1 , diag_2, and diag_3:

```
(modeling_df['diag_1']==-1)] = "missing"
(modeling_df['diag_1']==0)] = "external causes of injury and supplemental classification" #start with 'E' or 'V'
(modeling_df['diag_1']>=1) & (modeling_df['diag_1']< 140)] = "infectious and parasitic diseases"
(modeling_df['diag_1']>=140) & (modeling_df['diag_1']< 240)] = "neoplasms"
(modeling_df['diag_1']>=240) & (modeling_df['diag_1']< 280)] = "immunity disorders, without diabetes"
(modeling_df['diag_1']>=250) & (modeling_df['diag_1']< 251)] = "diabetes"
(modeling_df['diag_1']>=280) & (modeling_df['diag_1']< 290)] = "diseases of the blood"
(modeling_df['diag_1']>=290) & (modeling_df['diag_1']< 320)] = "mental disorders"
(modeling_df['diag_1']>=320) & (modeling_df['diag_1']< 390)] = "nervous system"
(modeling_df['diag_1']>=390) & (modeling_df['diag_1']< 460)] = "diseases of the circulatory system"
(modeling_df['diag_1']>=460) & (modeling_df['diag_1']< 520)] = "diseases of the respiratory system"
(modeling_df['diag_1']>=520) & (modeling_df['diag_1']< 580)] = "diseases of the digestive system"
(modeling_df['diag_1']>=580) & (modeling_df['diag_1']< 630)] = "diseases of the genitourinary system"
(modeling_df['diag_1']>=630) & (modeling_df['diag_1']< 680)] = "complications of pregnancy, childbirth, and the puerperium"
(modeling_df['diag_1']>=680) & (modeling_df['diag_1']< 710)] = "diseases of the skin and subcutaneous tissue"
(modeling_df['diag_1']>=710) & (modeling_df['diag_1']< 740)] = "diseases of the musculoskeletal system and connective tissue"
(modeling_df['diag_1']>=740) & (modeling_df['diag_1']< 760)] = "congenital anomalies"
(modeling_df['diag_1']>=760) & (modeling_df['diag_1']< 780)] = "certain conditions originating in the perinatal period"
(modeling_df['diag_1']>=780) & (modeling_df['diag_1']< 800)] = "symptoms, signs, and ill-defined conditions"
(modeling_df['diag_1']>=800) & (modeling_df['diag_1']< 1000)] = "injury and poisoning"
```

III. Code

A rendered notebook containing code for this analysis can be accessed at: https://nbviewer.org/github/rickfontenot/QTW/blob/main/Case%20Study%202/case2_rick.ipynb#LogRegClass