Using Logistic Regression to Predict Readmission in Diabetes Patients

I. The Problem

Diabetes is a metabolic disorder in which the body is incapable of producing any or enough of the hormone insulin to absorb blood glucose efficiently.

There are two types of diabetes, both of which can lead to a condition called hyperglycemia, or high blood sugar. Type II is more common and occurs when the body cannot use insulin properly, whereas Type I is an autoimmune disease in which the body cannot produce insulin. If left untreated, hyperglycemia can turn severe and cause complications affecting the heart, nerves, and kidneys. Self-management is necessary and possible by tracking blood sugar levels on a routine basis, and through medication.

Patient readmissions to hospitals are generally unplanned. Though not always preventable, readmission can be considered a marker of negligent care and is a waste of both hospital resources and spending. Patients with diabetes are known to have a high risk of 30-day readmission following initial hospitalization for hyperglycemia (American Diabetes Association 2019). To avoid readmission, hospitals must work to implement a structured protocol for the general admission and discharge of said patients, with focus on treating hyperglycemia and avoiding hypoglycemia (low blood sugar).

Initial evaluation of the patient should determine the type of diabetes or no previous history of diabetes, and an assessment of diabetes self-management knowledge. “Because inpatient insulin use and discharge orders can be more effective if based on an A1C level on admission, perform an A1C test on all patients with diabetes or hyperglycemia admitted to the hospital if the test has not been performed in the prior 3 months”. The A1C test (also known as the glycated hemoglobin test) gives an overall view of a patient’s average blood sugar level over several weeks or months. A1C levels can help assess the efficacy of current medication/treatment and allow the medical team to make changes (ADA 2019).

Reducing hospital readmissions—especially those resulting from poor inpatient or outpatient care—has long been a health policy goal because it represents an opportunity to lower health care costs, improve quality of care, and increase patient satisfaction. Stakeholders in the case of readmission are defined as “... those entities that are integrally involved in the healthcare system and would be substantially affected by reforms to the system” (Saint Joseph’s University 2011).
A stakeholder in this example would be a hospital that must decide between the following two options: discharge a patient with the risk of potential readmission in less than 30 days, or treat them as an inpatient for a longer period of time, increasing costs and using more resources (possibly unnecessarily), all in an effort to avoid readmission.

My focus in this project is to determine which factors cause a diabetes patient to be at risk for early readmission (readmission in less than 30 days), and to use these factors to build a successful predictive model.

II. The Data

The dataset can be found in the UC Irvine Machine Learning Repository, a collection of databases used for the empirical analysis of machine learning algorithms. It is available as an Excel file with 101,766 entries, representing 10 years (1999-2008) of clinical care at 130 hospitals and integrated delivery networks across the United States. The dataset contains both inpatient and outpatient data, in which hyperglycemia management took place.

III. Methodology

A. Data Wrangling

The columns in the dataset are as follows:

Table 1. Column Names and Descriptions

Column Name	Column Description
encounter_id	Unique identifier of an encounter
patient_nbr	Unique identifier of a patient
race	Race: Caucasian, Asian, African American, Hispanic, and Other
gender	Gender: Male, Female, and Unknown/Invalid
age	Age (grouped in 10-year intervals)
weight	Weight (in lbs)
admission_type_id	Integer identifier corresponding to 9 distinct types of admission
discharge_disposition_id	Integer identifier corresponding to 29 distinct types of discharge
admission_source_id	Integer identifier corresponding to 21 distinct types of admission source
time_in_hospital	Number of days between admission and discharge
payer_code	Integer identifier corresponding to 23 distinct types of insurance payers
medical_specialty	Integer identifier corresponding to 84 distinct specialties of admitting physicians
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 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
diag_1, diag_2, diag_3	Primary, secondary, and additional secondary diagnosis of patient (coded as the first three digits of the *International Classification of Diseases, 9th Revision*)
number_diagnoses	Number of diagnoses entered to the system
max_glu_serum	">200," ">300," "normal," and "none" if not measured
A1Cresult	● ">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% ● "none" if not measured
24 Features for Medications	Feature indicates whether the drug was prescribed or there was a change in the dosage: ● "up" if the dosage was increased during the encounter ● "down" if the dosage was decreased ● "steady" if the dosage did not change ● "no" if the drug was not prescribed
change	Indicates if there was a change in diabetic medications (either dosage or generic name)
diabetesMed	Indicates if any diabetic medication was prescribed
readmitted	Days to inpatient readmission: ● "<30" if the patient was readmitted in less than 30 days ● ">30" if the patient was readmitted in more than 30 days ● "No" for no record of readmission

An initial exploration of the dataset included observing the value counts for all variables. Columns "encounter_id", "weight", "payer_code", "medical_specialty", and "max_glu_serum" were dropped for missing too many values. Next, all duplicate "patient_nbr" entries were dropped, and the first recorded encounter for each patient was retained.

The nominal target "readmitted" was converted to a binary variable by dropping all instances of readmissions ">30", and attributing a value of 0 to no record of readmission, and 1 if a patient was

readmitted in "<30" days. Major class imbalances were found for many features, as well as for the target.

There were three columns for diagnosis - "diag_1", "diag_2", and "diag_3". A preliminary run of Random Forest Feature Importance indicated that the secondary and additional secondary diagnoses did not have a significant impact on classifying readmission. As such, the two columns were dropped. The values in "diag_1" were converted to integers, and then binned as follows:

Group name	icd9 codes
Circulatory	390–459, 785
Respiratory	460–519, 786
Digestive	520–579, 787
Diabetes	250.xx
Injury	800–999
Musculoskeletal	710–739
Genitourinary	580–629, 788
Neoplasms	140–239
Other (17.3%)	780, 781, 784, 790–799
	240–279, without 250
	680–709, 782
	001–139
	290–319
	E–V
	280–289
	320–359
	630–679
	360–389

Figure 1. Primary Diagnoses Values with Reference to Binned Categories. Adapted from "Impact of HbA1c on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records" by B.Strack, 2014, April 3, *BioMed Research International, https://www.hindawi.com/journals/bmri/2014/781670/tab2/*

After binning and removing any missing values, there were a total of 38,911 remaining entries. It was found that a majority of patients were initially admitted for circulatory system issues, followed by digestive and respiratory concerns.

B. Graphical Exploratory Data Analysis

Visual EDA followed the data wrangling process, and box plots were created for the following continuous variables in the dataset:

- time_in_hospital
- num_lab_procedures
- num_procedures
- num_medications
- number_outpatient
- number_emergency
- number_inpatient
- number_diagnoses

The variables were plotted against the target.

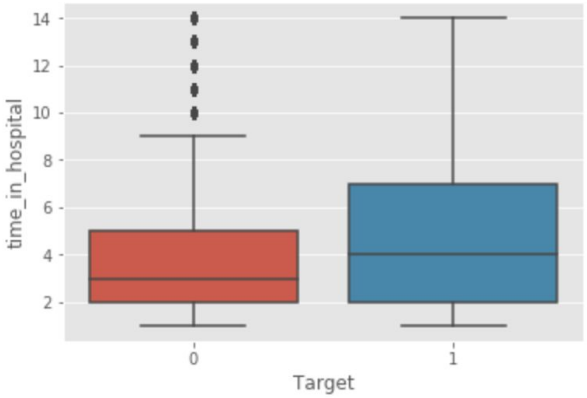

Figure 2. Number of Days Spent in Hospital and Its Effect on Readmission

There appears to be a significant difference in distribution between the positive and negative classes of patients regarding the number of days spent in the hospital between admission and discharge (Fig. 2). The median number of days spent for both groups is similar - 3 days for the negative class, and 4 days for the positive class. In general, patients who are not readmitted spend between 2 - 5 days in the hospital, with the maximum days being at 9, with outliers ranging all the way up to 14. For patients readmitted however, the number of days is between 2 - 7, with the maximum days topping off at 14. Thus, “time_in_hospital” would be a good feature for classifying the binary outcome.

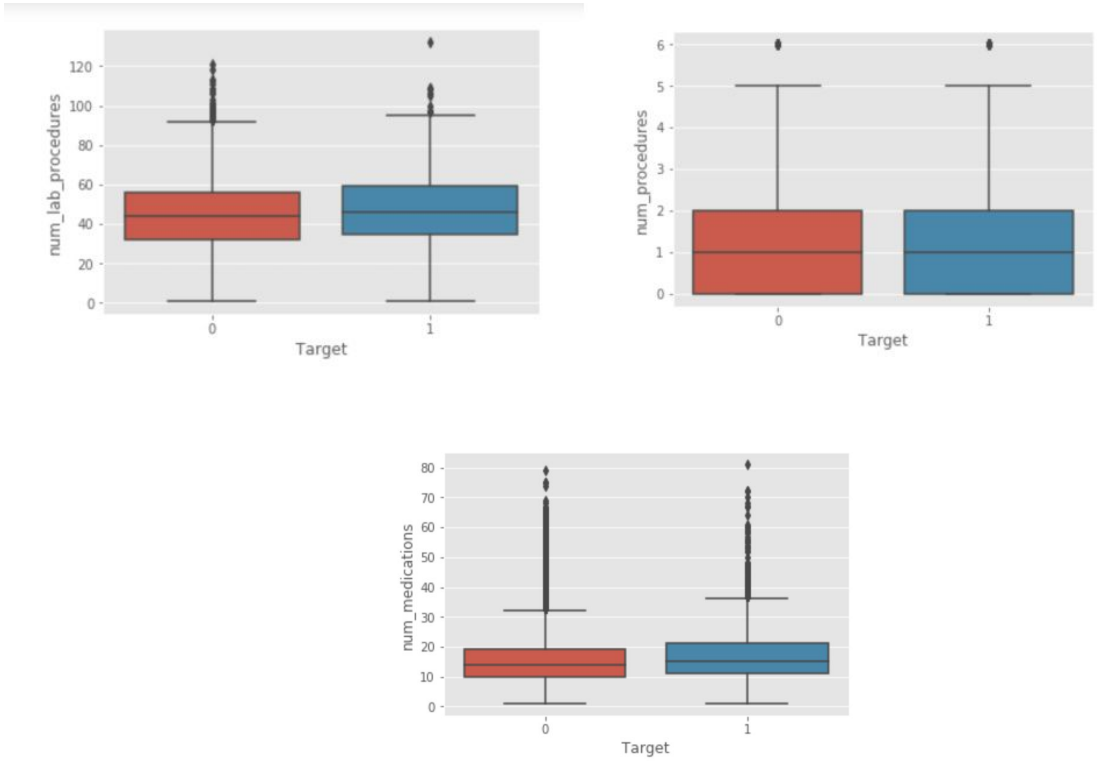

Figure 3. The Numbers of Lab/Other Procedures and Medications Administered and their Effects on Readmission

The distributions for the number of lab procedures and non-lab procedures performed, and the number of medications administered during an encounter are very similar for both classes (Fig. 3). These features would not classify the binary outcome well.

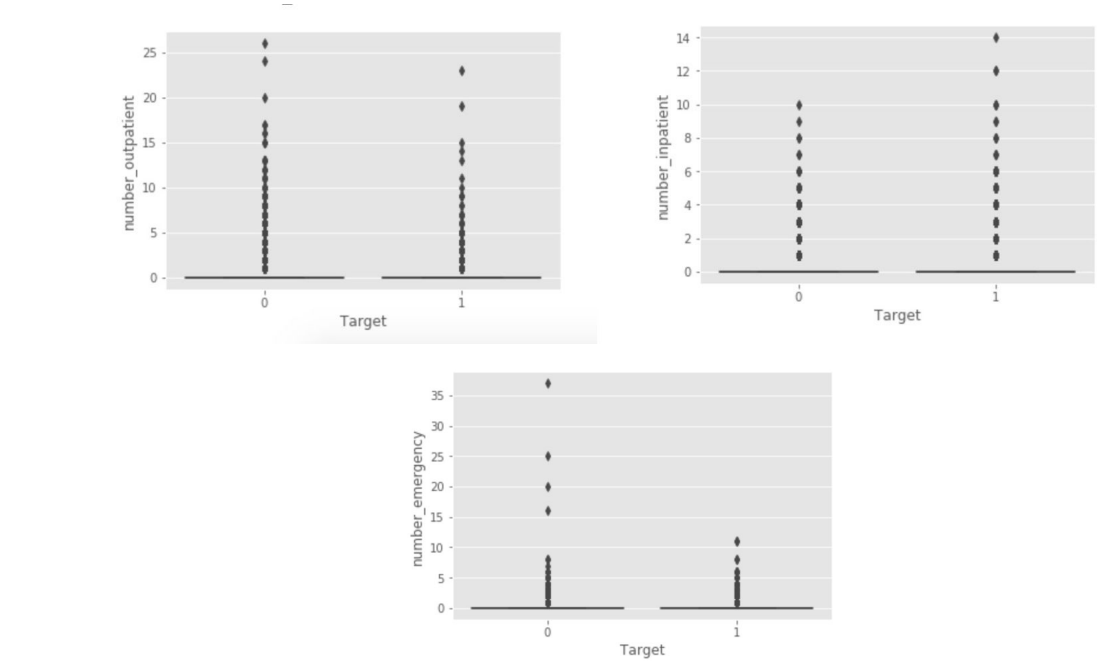

Figure 4. Number of Visits in the Prior Year and their Effects on Readmission

The distributions for the number of outpatient and inpatient visits in the year prior to a patient's current encounter are similar for both classes. The spread for the number of emergency visits is slightly different however. Whereas both classes show similar frequencies for 0 - 10 visits, there are multiple outliers present for the negative class of patients, where the range goes all the way up to 25 - 35 visits (Fig. 4).

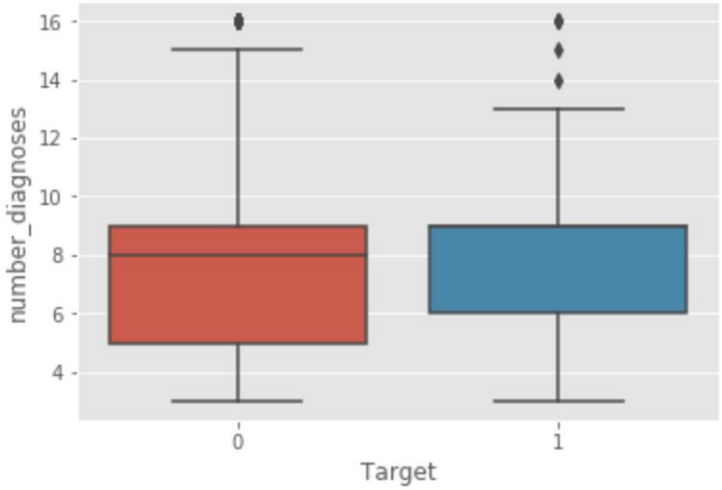

Figure 5. Total Number of Diagnoses Entered into System and its Effect on Readmission

The distributions for readmission in relation to the total number of diagnoses made is quite interesting. Whereas the maximum number of diagnoses for the negative class is higher than that of the positive class (15 diagnoses vs. 13), there is only a small overlap between the two boxplots, making the feature a good one for classifying the binary outcome (Fig. 5).

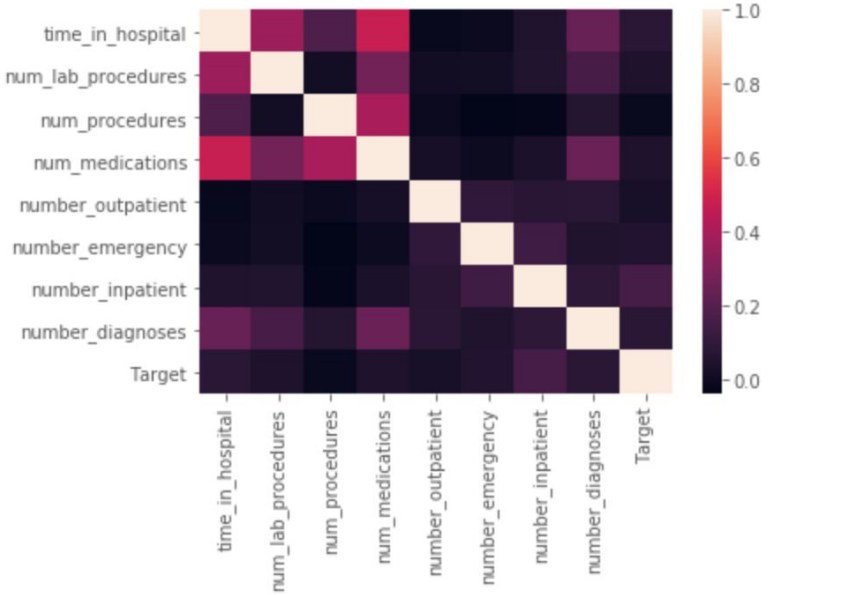

Figure 6. Standard correlation matrix for all continuous variables

A standard correlation matrix was graphed in an effort to find noticeable correlations and check for collinearity (Fig. 6). One immediate correlation that is seen is between the number of medications the patient is prescribed and the time they spend in the hospital.

This was followed by inferential statistical analysis to determine whether any perceived correlations were statistically significant.

C. Inferential Statistics

Two-sample t-tests were conducted for three categorical variables - "gender", "change", and "diabetesMed". A p-value of 0.053 was calculated for gender with regard to readmission, leading to the rejection of the alternate hypothesis at the 0.05 significance level, signifying that no difference in the rates of readmission can be proven for males and females.

Conversely, the null hypothesis was rejected for "change" (p-value = $1.7e^{-08}$), suggesting that there is a significant difference in the rate of readmission for patients with medication changes vs. no changes.

The null hypothesis was also rejected in favor of the alternate for "diabetesMed" (p-value = $4.15e^{-19}$), implying that there is a significant difference in the rate of readmission for patients with diabetes prescriptions vs. no prescriptions.

IV. Feature Selection

Random Forest Feature Selection

	0	1
4	num_lab_procedures	0.113606
6	num_medications	0.104233
3	time_in_hospital	0.068583
5	num_procedures	0.051845
10	number_diagnoses	0.044851
1	discharge_disposition_id	0.043070
9	number_inpatient	0.033549
0	admission_type_id	0.026938
2	admission_source_id	0.025386
16	gender_Male	0.022841
7	number_outpatient	0.018667
64	insulin_Steady	0.017864
72	change_No	0.017618
82	Binned_Circulatory System	0.017511
24	age_[70-80)	0.015910
8	number_emergency	0.015360
23	age_[60-70)	0.015120
63	insulin_No	0.014209
13	race_Caucasian	0.013891
11	race_AfricanAmerican	0.013854
28	A1Cresult_None	0.011607
30	metformin_No	0.011162
25	age_[80-90)	0.011123
31	metformin_Steady	0.010822
65	insulin_Up	0.010791
45	glipizide_No	0.010531
22	age_[50-60)	0.010438
46	glipizide_Steady	0.010334
84	Binned_Digestive System	0.010080
83	Binned_Respiratory System	0.010009

Figure 7. Random Forest Feature Importance

It was found that the variables listed in the figure above were most important in classifying readmission, based on this form of feature selection, which relies on impurity-based ranking. It is important to note that impurity reduction is known to be biased towards variables with more categories.

Once the most pertinent variables were accounted for, a new DataFrame was created with the features and target. As a majority of the features were categorical, they were converted to dummy variables before performing the regression. This method was used to limit the feature set to 56 features.

V. Model Fitting & Validating Results

Logistic Regression (with train-test split)

The first logistic regression run generated an accuracy score of 0.876, implying an 87% probability that a patient is correctly classified as being readmitted or not. Accuracy, however, is not the best metric for classification - it works well on balanced data. To account for the aforementioned class imbalances in the data, balanced accuracy proved to be a more valuable metric.

Balanced Accuracy

Balanced accuracy is a metric used in binary and multiclass classification problems to deal with an imbalanced dataset. It is defined as the average recall obtained in each class. The balanced accuracy score for the first logistic regression = 0.511.

Logistic Regression (with balanced "class weight")

To further account for the imbalanced dataset, a second logistic regression was run in which the default argument for class_weight = "none" was replaced by class_weight = "balanced". The following accuracy scores were calculated:

- Accuracy Score = 0.663
- Balanced Accuracy Score = 0.614

Precision-Recall

- **True Positive (TP): readmitted and predicted as 1**
- **False Positive (FP): not readmitted but predicted as 1**
- **False Negative (FN): readmitted but predicted as 0**
- **True Negative (TN): not readmitted and predicted as 0**

Precision-Recall is a very useful measure for evaluating the success of a classification model, especially when classes show imbalances.

Precision can be interpreted as the accuracy of positive predictions and is defined by the following formula:

$$precision = \frac{TP}{TP + FP}$$

It essentially answers the question - "what proportion of positive predictions were correct?".

Recall is the ability of a classifier to find all positive instances and answers the question - "what proportion of positive cases were caught?". It is defined as the following:

$$recall = \frac{TP}{TP + FN}$$

Ideally, a combination of high precision with high recall rates is desired as it implies that the model is highly specialized with the ability to catch most, if not all, positive cases accurately. However, the relationship between the two metrics is that of a tradeoff - one metric is usually maximized at the expense of the other. Classification reports can break down these metrics on a per-class basis, and were used to evaluate the success of both logistic regression models - with balanced and unbalanced class weights.

- **Logistic Regression - Run #1 (class_weight = "none")**

```
Accuracy Score: 0.8762206612986123
Balanced Accuracy Score: 0.5117282526633672
[[10191    27]
 [ 1418    38]]
              precision    recall  f1-score   support

           0       0.88      1.00      0.93     10218
           1       0.58      0.03      0.05      1456

    micro avg       0.88      0.88      0.88     11674
    macro avg       0.73      0.51      0.49     11674
weighted avg       0.84      0.88      0.82     11674
```

Figure 8. Classification Report for Logistic Regression Run #1

- **Precision for (1)** = 0.58
  - 58% of predictions were correct

- **Recall for (1)** = 0.03
  - Only 3% of positive cases were caught, and 97% of those readmitted were predicted as not readmitted.

- **Logistic Regression - Run #2 (class_weight = "balanced")**

```
Accuracy Score: 0.6634401233510365
Balanced Accuracy Score: 0.6139777843022118
[[6947 3271]
 [ 658  798]]
              precision    recall  f1-score   support

           0       0.91      0.68      0.78     10218
           1       0.20      0.55      0.29      1456

    micro avg       0.66      0.66      0.66     11674
    macro avg       0.55      0.61      0.53     11674
weighted avg       0.82      0.66      0.72     11674
```

Figure 9. Classification Report for Logistic Regression Run #2

- **Precision for (1)** = 0.20
  - 20% of predictions were correct

- **Recall for (1)** = 0.55
  - 55% of positive cases were caught, and 45% of those readmitted were predicted as not being readmitted.

This is far better as we are able to capture 55% of positive cases. To improve this further, the next step was GridSearching and thresholding.

GridSearching & Thresholding

GridSearching, a form of hyperparameter optimization, was performed in an effort to determine the ideal parameters for logistic regression, i.e. C = 0.1 and penalty = "l1". These optimal parameters were used to then loop over a list of thresholds lower than 0.5 in an effort to maximize recall.

```
 ** Threshold = 0.15 **
Accuracy Score = 0.7424190508823025
Balanced Accuracy Score = 0.597254844392249
[[8079 2139]
 [ 868  588]]
/Users/springboard/anaconda3/envs/ProjectEnv/lib/python3.7/site-packages/ipykernel_launcher.py:9: FutureWarning: Meth
od .as_matrix will be removed in a future version. Use .values instead.
  if __name__ == '__main__':
              precision    recall  f1-score   support

           0       0.90      0.79      0.84     10218
           1       0.22      0.40      0.28      1456

   micro avg       0.74      0.74      0.74     11674
   macro avg       0.56      0.60      0.56     11674
weighted avg       0.82      0.74      0.77     11674
```

Figure 10. Classification Report with Optimal Thresholding Results for Logistic Regression Run #1 (unbalanced)

```
 ** Threshold = 0.45 **
Accuracy Score = 0.5447147507281137
Balanced Accuracy Score = 0.6153575273327182
[[5326 4892]
 [ 423 1033]]
              precision    recall  f1-score   support

           0       0.93      0.52      0.67     10218
           1       0.17      0.71      0.28      1456

   micro avg       0.54      0.54      0.54     11674
   macro avg       0.55      0.62      0.47     11674
weighted avg       0.83      0.54      0.62     11674
```

Figure 11. Classification Report with Optimal Thresholding Results for Logistic Regression Run #2 (balanced)

ROC/AUC

To evaluate the model further, the AUC (Area Under the Curve) of the ROC (Receiver Operating Curve) was also measured. The ROC curve looks at the performance of a model across all positive thresholds. The AUC measures the degree of separability, i.e. how capable the model is of predicting 0's as belonging to the negative class, and predicting 1's as belonging to the positive class. An AUC of 1

indicates perfect prediction at all thresholds, whereas a measure of .5 indicates the model does no better than a random guess.

The following curves were generated for the two regression models:

```
0.6598232702900936
```

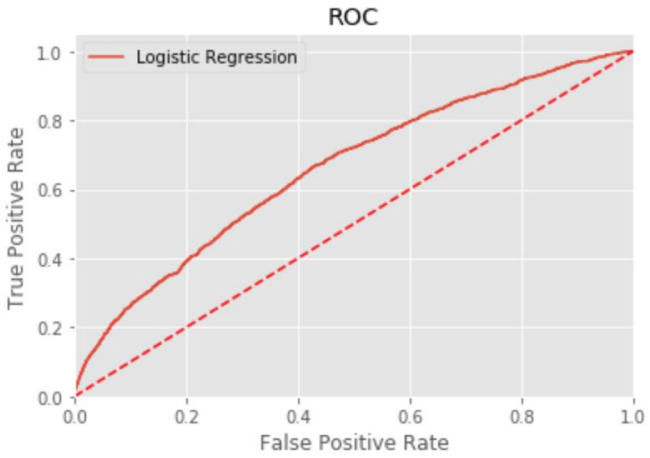

Figure 12. ROC Curve for LR #1 (unbalanced) with AUC = 0.659

```
0.661502258995653
```

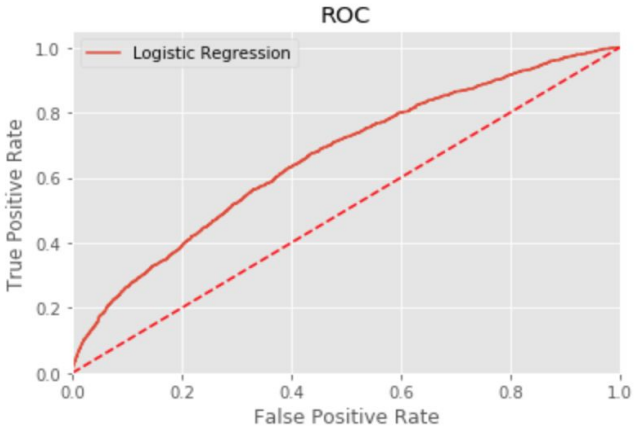

Figure 13. ROC Curve for LR #2 (balanced) with AUC = 0.662

VI. Conclusion & Further Research

A flaw of the dataset is the presence of major class imbalances for multiple features and the target. For example, after wrangling, only 12% of the data was found to belong to the positive class “readmission” for the target variable. Unbalanced features included race, admission type, discharge type, numbers of outpatient/inpatient/emergency visits, A1C result, and diabetes medication prescription. Nevertheless, a simple logistic regression algorithm was chosen to run this classification problem with a binary outcome.

First, feature selection via Random Forest was performed in an effort to optimize scores, followed by GridSearching to find the ideal parameters for running two logistic regression models - with balanced and unbalanced class weights. Thresholding was then performed, with a focus on optimizing for balanced accuracy - the key metric.

For this problem, although balanced accuracy was our primary metric, recall was also of interest as it answered the question - of all readmissions, how many were correctly classified? As such, the threshold was kept under .5 so as to ensure high enough recall.

Recall for the initial regression run was very low (3%), but increased when a threshold of 0.15 was selected. Recall for the positive class was maximized to 40%. Recall for the negative class decreased significantly however, and the rate of false positives increased steadily, thereby lowering the overall precision of the model, which was now over-predicting readmission.

The two AUC values generated ranged from 0.659 - 0.662, indicating that both models have about a 66% prediction capability.

The model could be improved through future research. Feature selection was implemented in this analysis due to the high-dimensionality of the dataset, and further feature sets could be tested. A possible new study could explore building logistic regression models with all features included in the train-test split. Performance of the models would be evaluated using the metrics described in this report. A second study could implement under- and over-sampling techniques to train the imbalanced dataset. Finally, a cost-benefit analysis could be applied to this analysis to determine which decision is more costly - for a patient to receive inpatient services for longer than necessary in an effort to prevent readmission, or to allow an at-risk patient to be discharged, with the possibility of readmission.

VII. Client Recommendations

First, the American Diabetes Association emphasizes the importance of a patient's A1C result during hospital admission - a very important feature for classifying readmission. The dataset failed to provide enough information on glycated hemoglobin levels however as about 81% of the patients fell into the "None Reported" category for results. As such, I would strongly recommend that HbA1c tests become mandatory, with a care and discharge plan then tailored to each patient's levels.

Secondly, diabetes and hyperglycemia can be managed very well by individuals suffering from the disease. Thus, hospitals should focus on impactful ways to educate patients regarding care in the outpatient setting, including information on diet, exercise, medication, and the usage of blood glucose meters.

VIII. References

American Diabetes Association. “15. Diabetes Care in the Hospital.” *Standards of Medical Care in Diabetes*, vol. 42, no. 1, Jan. 2019, https://doi.org/10.2337/dc19-S015.

“Health Care Reform: Duties and Responsibilities of the Stakeholders.” *Institute of Clinical Bioethics*, Saint Joseph's University, 6 Sept. 2011, https://sites.sju.edu/icb/health-care-reform-duties-and-responsibilities-of-the-stakeholders/.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

Strack, Beata, and Jonathan P DeShazo. “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.” *BioMed Research International*, vol. 2014, 3 Apr. 2014, http://dx.doi.org/10.1155/2014/781670.

World Health Organization. (1996). *ICD-9 : International Classification of Diseases and Related Health Problems, 9th Revision,* https://www.cdc.gov/nchs/icd/icd9.htm.