Using Logistic Regression to Predict Readmission in Diabetes Patients

Rohini Lahiri

Springboard Data Science Career Track

August 2019

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.
- Diabetes – Types I and II – can lead to hyperglycemia, or high blood sugar, which if left untreated, 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.
- However, patients with diabetes are known to have a high risk of 30-day readmission following initial hospitalization for hyperglycemia (American Diabetes Association 2019).

Motivation

- In general, patient readmissions are unplanned events and can imply negligent care. High rates of readmission can also be considered a waste of resources and hospital funding.
- To avoid readmissions, hospitals must work to implement structured protocols for the general admission and discharge of said patients, with a focus on treating hyperglycemia and avoiding hypoglycemia (low blood sugar).
- The motivation for this project is to:
  - Build a logistic regression model which can accurately predict readmission by …
    - Finding major trends in the data
    - Identifying potential factors related to readmission

Data Collection & Preparation

- 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.

Methodology

- Visual Exploratory Data Analysis
- Statistical Inference
- Feature Selection
- Random Forest Feature Importance
- GridSearching & Thresholding
- Logistic Regression (with unbalanced & balanced class weights)
- Balanced Accuracy
- Precision/Recall
- ROC/AUC

Visual Exploratory Data Analysis & Statistical Inference

Visual EDA and statistical inference indicated that the following features were important in classifying readmission:

Table 1. Selected Column Names and Descriptions

Column Name	Column Description
time_in_hospital	Number of days between admission and discharge
number_of_emergency	Number of emergency visits of the patient in the year preceding the encounter
number_of_diagnoses	Number of diagnoses entered to the system
change	Indicates if there was a change in diabetic medications (either dosage or generic name)
diabetesMed	Indicates if any diabetic medication was prescribed

Feature Selection: Random Forest Feature Importance

- Random Forest Feature Importance indicated that the features listed in **Figure 1** were most important in classifying readmission.
- Dummy variables were used to limit the feature set to 56 features.

	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 1. Results of Random Forest Feature Importance

Logistic Regression #1: (class_weight = 'none')

- 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 is not the best metric for classification as it works well on balanced data.
- To account for major class imbalances found in the data, balanced accuracy proved to be a more valuable metric.
  - **Balanced Accuracy:** 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.
  - **Balanced accuracy score (for first logistic regression)** = 0.511.

Logistic Regression #2: (class_weight = ‘balanced’)

- 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

GridSearching & Thresholding

- GridSearching, a form of hyperparameter tuning, was performed in an effort to determine the ideal parameters for logistic regression:
  - **C** = 0.1
  - **Penalty** = “l1”.
- These optimal parameters were used to then loop over a list of thresholds lower than 0.5 to maximize recall.
- Thresholding was performed with a focus on optimizing for balanced accuracy - the key metric. The threshold was kept under .5 so as to ensure high enough recall.
- **Optimal Threshold (LR #1)** = 0.15
- **Optimal Threshold (LR #2)** = 0.45

Optimal Results:
Logistic Regression Run#1 (class_weight = 'none')

```
 ** 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 2. Classification Report for Logistic Regression Run #1

Optimal Results:
Logistic Regression Run#2 (class_weight = 'balanced')

```
 ** Threshold = 0.45 **
Accuracy Score = 0.54471475072811 37
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 3. Classification Report for Logistic Regression Run #2

ROC/AUC

- To evaluate the model further, the AUC (Area Under the Curve) of the ROC (Receiver Operating Curve) was also measured.
- 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
```

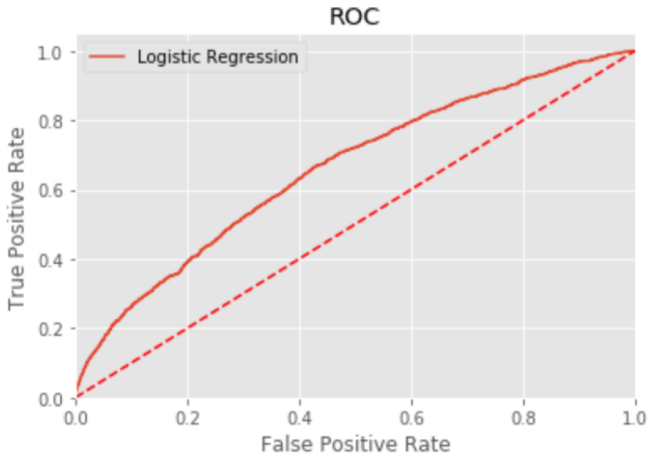

Figure 4. AUC for LR #1

```
0.661502258995653
```

Figure 5. AUC for LR #2

Conclusion & Further Research

- For this problem, recall was chosen as the metric of interest as it answered the question - of all readmissions, how many were correctly classified?
- 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.
- 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:
    1. 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.
    2. 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.

Client Recommendations

1. 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.
2. 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.

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.