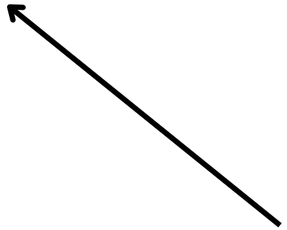
TEAM SATURN

Exploring the Relationship between Hospital Stays and the Readmission of Diabetic Patients

Anastasia Wei, Amy Wang, Kaitlyn Hung, Lila Wells

What is the relationship between a patient's initial hospital stay and their likelihood of readmission?

What is the relationship between a patient's initial hospital stay and their likelihood of readmission?



Inference

Identifying the relationship between the predictors (factors associated with one's hospital stay) and the response (hospital readmittance)

The Dataset

Data Sourcing and Timespan

Information from over 130 U.S. hospitals between 1999 - 2008

Entries and Variables

Over 10,000 entries & ~50 predictors related to patient demographics and facts about their initial hospital stay

Response Variable & Model Choice

Response Variable

Whether a patient had been readmitted (0 or 1)

Model Choice: Logistic Regression

The response is a categorical variable.

Stakeholders

Stakeholders	How they benefit from the analysis
Diabetic patients and their loved ones	Insights into health risks
Physicians and healthcare workers	Insights into altering patient healthcare plans
Hospital administration	Insights to reduce costly readmissions

Stakeholders

Stakeholders	How they benefit from the analysis
Diabetic patients and their loved ones	Insights into health risks
Physicians and healthcare workers	Insights into altering patient healthcare plans
Hospital administration	Insights to reduce costly readmissions

Stakeholders

Stakeholders	How they benefit from the analysis
Diabetic patients and their loved ones	Insights into health risks
Physicians and healthcare workers	Insights into altering patient healthcare plans
Hospital administration	Insights to reduce costly readmissions

TEAM SATURN

Developing the Model: Procedure

Developing the Model: Procedure

Procedure:

(1) Data cleaning and preparation

(2) Exploratory data analysis

(3) Variable selection and model training

(4) Model optimization with variable selection

Data Cleaning and Preparation

Data cleaning procedure

Variable transformations

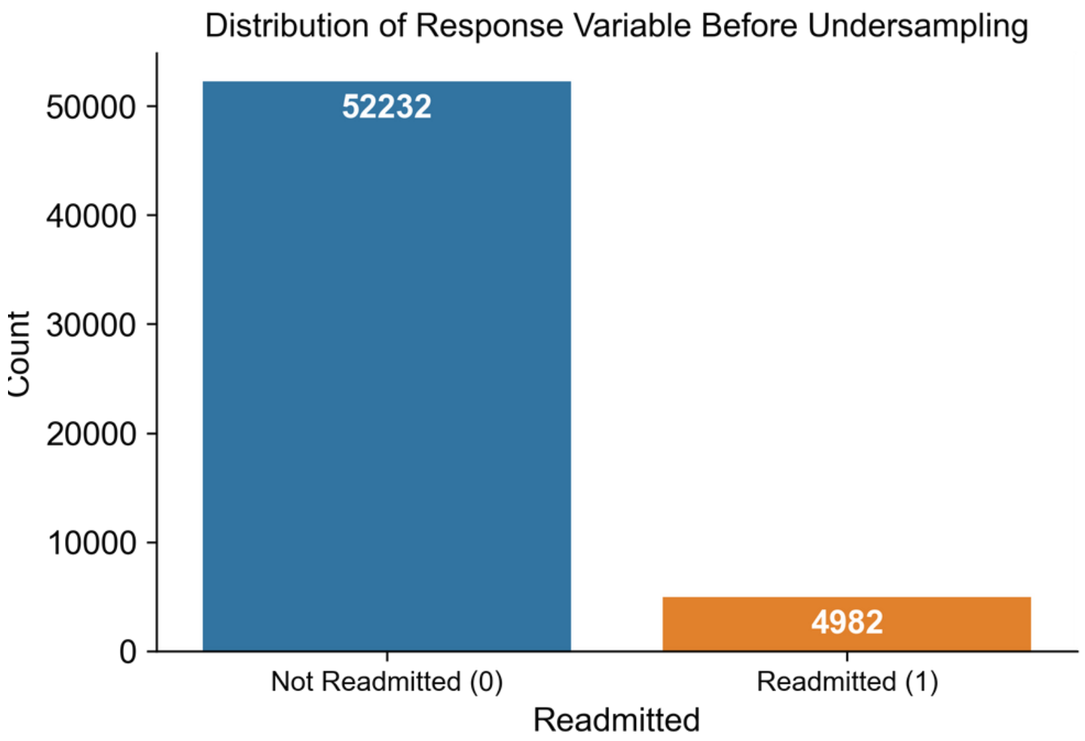
(1) Dropped variables with more than half of the data missing (weight & medical specialty)

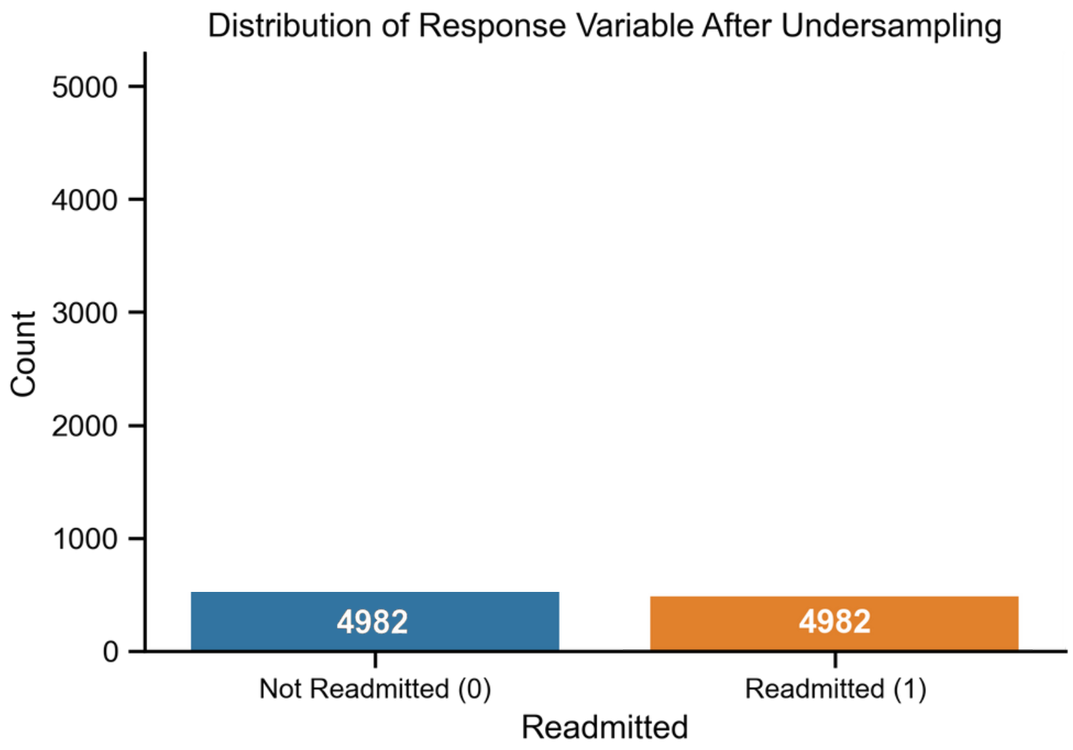
(2) Removed duplicate records (~10,000 to ~70,000 observations)

(3) Used domain knowledge to bin diagnosis (> 900 distinct values), admission type, discharge disposition, and admission source (> 60 distinct values)

(4) Created a number of changes variable based on changes in medications

Undersampling: correcting for an uneven response distribution





Creating and Training the Model

Building the Model

```
logit_model = sm.logit(formula = 'readmitted ~ num_of_changes*time_in_hospital
+ number_inpatient + age', data = train1).fit()
```

Logit Regression Results

Dep. Variable:	readmitted	**No. Observations:**	10023
Model:	Logit	**Df Residuals:**	10017
Method:	MLE	**Df Model:**	5
Date:	Sun, 26 Feb 2023	**Pseudo R-squ.:**	0.02193
Time:	10:12:37	**Log-Likelihood:**	-6794.9
converged:	True	**LL-Null:**	-6947.2
Covariance Type:	nonrobust	**LLR p-value:**	9.641e-64

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.9796	0.096	-10.191	0.000	-1.168	-0.791
num_of_changes	0.3339	0.082	4.069	0.000	0.173	0.495
time_in_hospital	0.0602	0.008	7.287	0.000	0.044	0.076
num_of_changes:time_in_hospital	-0.0415	0.013	-3.150	0.002	-0.067	-0.016
number_inpatient	0.3754	0.032	11.680	0.000	0.312	0.438
age	0.0086	0.001	6.503	0.000	0.006	0.011

Building the Model

```
logit_model = sm.logit(formula = 'readmitted ~ num_of_changes*time_in_hospital
+ number_inpatient + age', data = train1).fit()
```

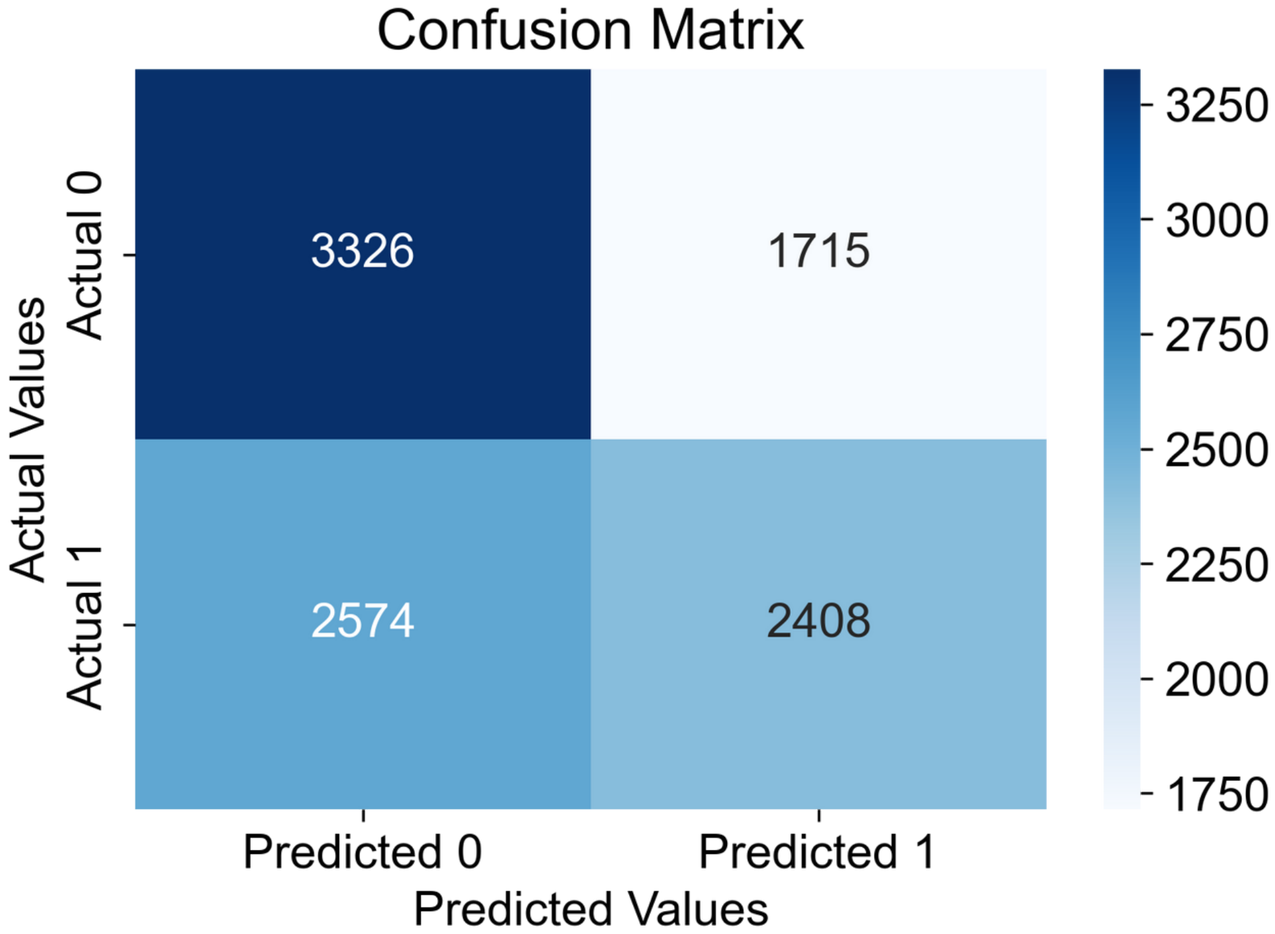
Logit Regression Results

Dep. Variable:	readmitted	**No. Observations:**	10023
Model:	Logit	**Df Residuals:**	10017
Method:	MLE	**Df Model:**	5
Date:	Sun, 26 Feb 2023	**Pseudo R-squ.:**	0.02193
Time:	10:12:37	**Log-Likelihood:**	-6794.9
converged:	True	**LL-Null:**	-6947.2
Covariance Type:	nonrobust	**LLR p-value:**	9.641e-64

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.9796	0.096	-10.191	0.000	-1.168	-0.791
num_of_changes	0.3339	0.082	4.069	0.000	0.173	0.495
time_in_hospital	0.0602	0.008	7.287	0.000	0.044	0.076
num_of_changes:time_in_hospital	-0.0415	0.013	-3.150	0.002	-0.067	-0.016
number_inpatient	0.3754	0.032	11.680	0.000	0.312	0.438
age	0.0086	0.001	6.503	0.000	0.006	0.011

- Number of medication changes
- Time spent in the hospital
- Interaction between medication changes and time spent in hospital
- Age of the patient
- Number of inpatient visits

Confusion Matrix on Train Data



Classification accuracy = 57.2%

Precision = 58.4%

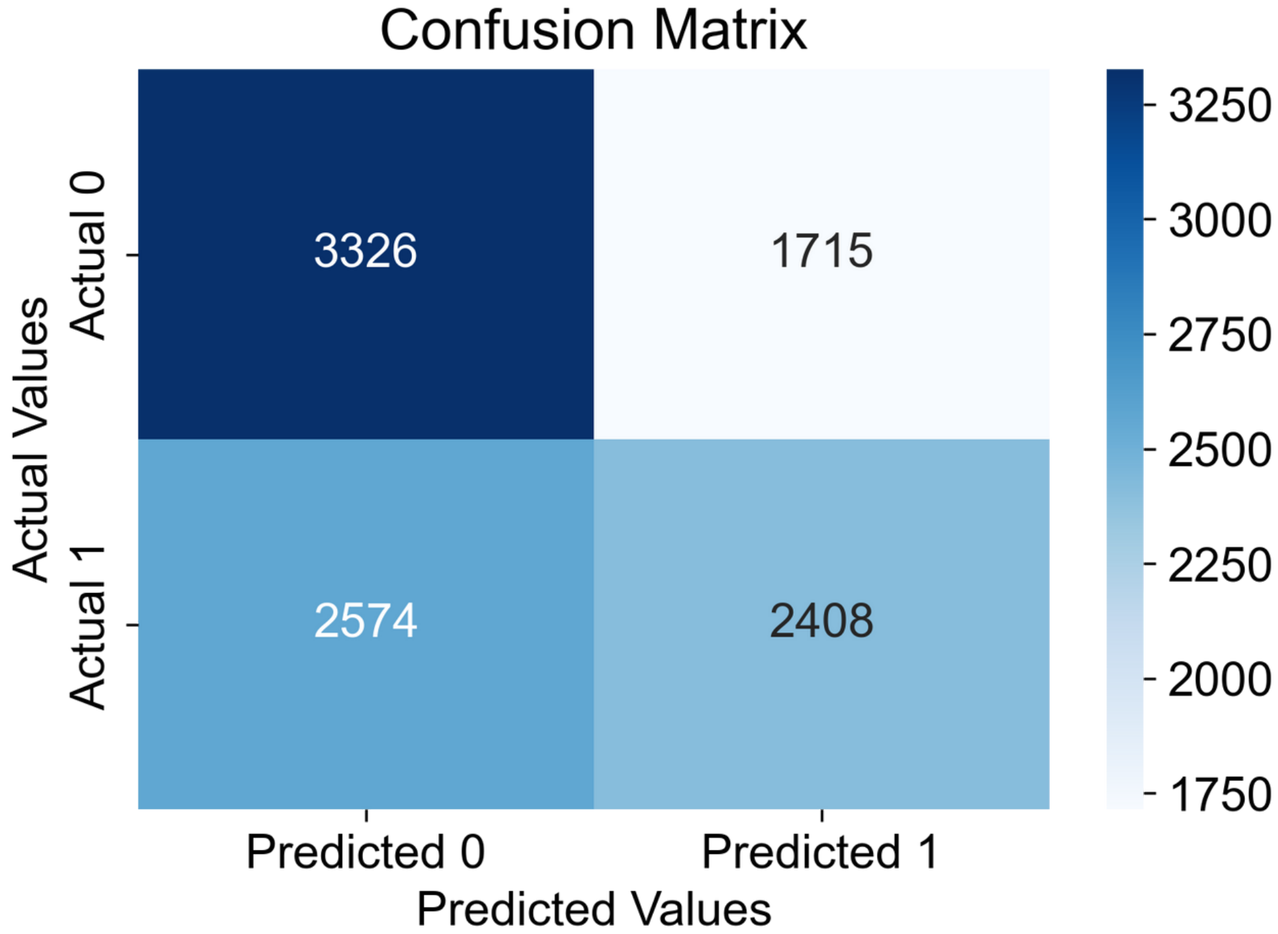
TPR or Recall = 48.3%

FNR = 51.7%

FPR = 34.0%

ROC-AUC = 59.9%

Confusion Matrix on Train Data



Classification accuracy = 57.2%

Precision = 58.4%

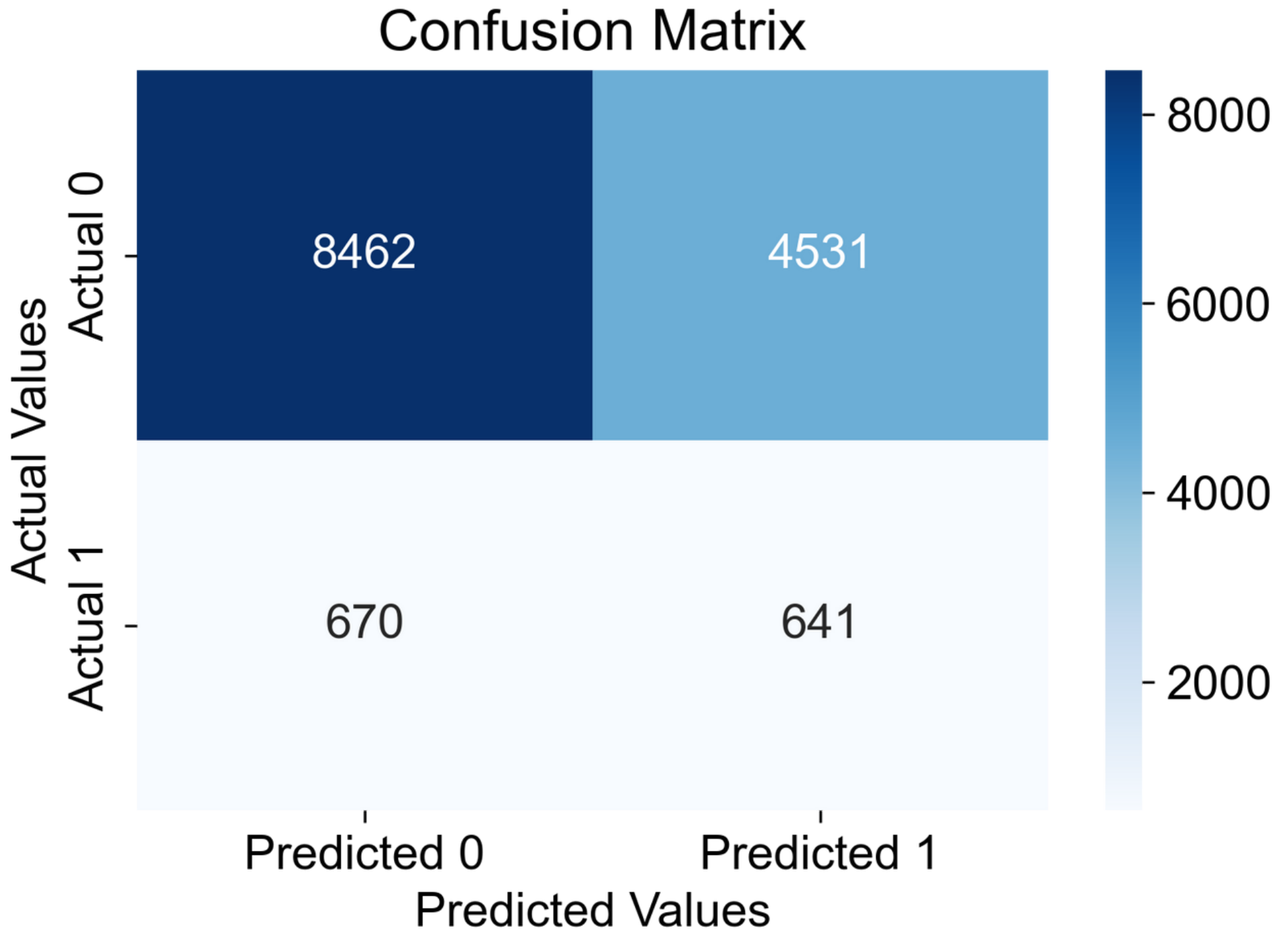
TPR or Recall = 48.3%

FNR = 51.7%

FPR = 34.0%

ROC-AUC = 59.9%

Confusion Matrix on Test Data



Classification accuracy = 63.6%

Precision = 12.4%

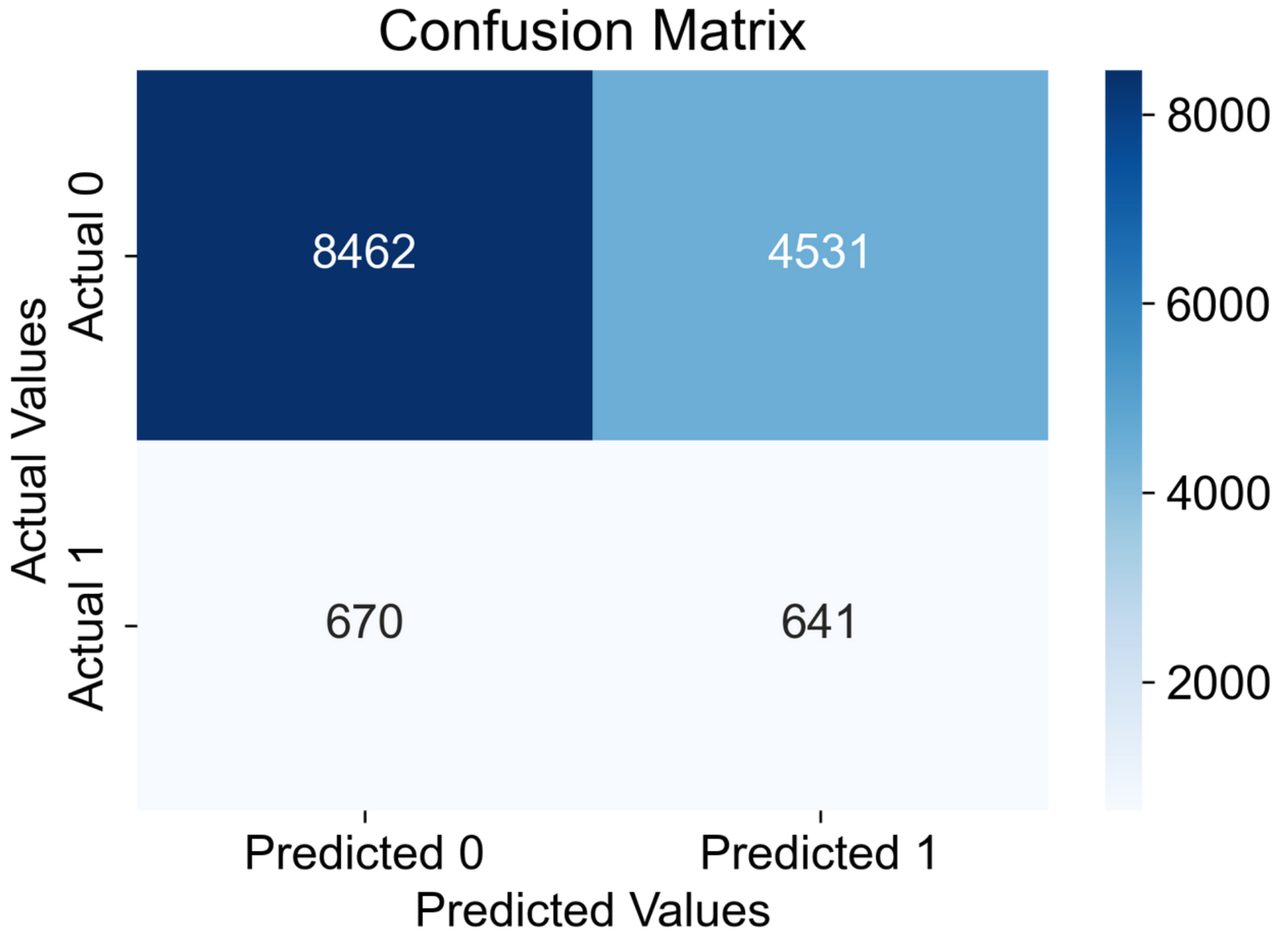
TPR or Recall = 48.9%

FNR = 51.1%

FPR = 34.9%

ROC-AUC = 60.0%

Confusion Matrix on Test Data



Classification accuracy = 63.6%

Precision = 12.4%

TPR or Recall = 48.9%

FNR = 51.1%

FPR = 34.9%

ROC-AUC = 60.0%

Model Optimization

Optimizing the Model

1

SMOTENC: correct uneven response distribution in train data

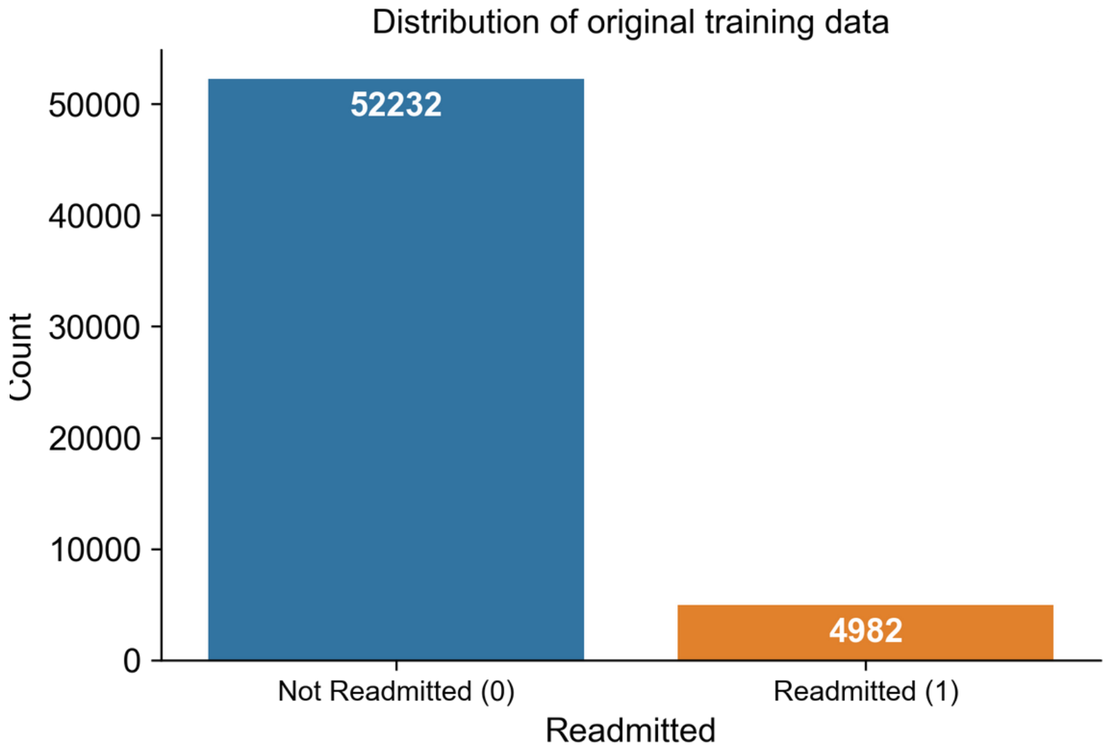
2

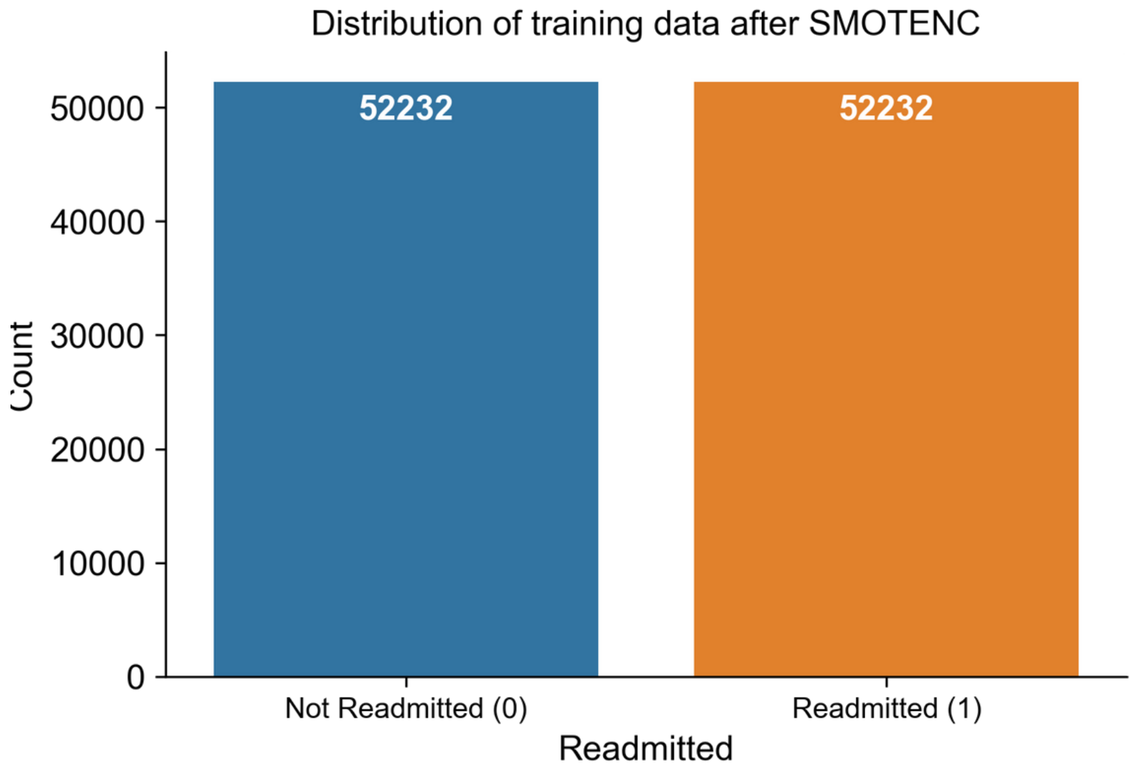
Remove multicollinear predictors

3

Variable selection: Create separate predictors for binned variables (age, time in hospital, discharge disposition, and diagnoses)

SMOTENC


Distribution of original training data
50000
40000
30000
20000
10000
0
Count
52232
4982
Not Readmitted (0)
Readmitted (1)
Readmitted


Distribution of training data after SMOTENC
50000
40000
30000
20000
10000
0
Count
52232
52232
Not Readmitted (0)
Readmitted (1)
Readmitted

Optimizing the Model

1

SMOTENC: correct uneven response distribution in train data

2

Remove multicollinear predictors

3

Variable selection: Create separate predictors for binned variables (age, time in hospital, discharge disposition, and diagnoses)

Optimizing the Model

1

SMOTENC: correct uneven response distribution in train data

2

Remove multicollinear predictors

3

Variable selection: Create separate predictors for binned variables (age, time in hospital, discharge disposition, and diagnoses)

Updated Model Equation

```
logit_model = sm.logit(formula = 'readmitted ~ age5 + age15 + age25 + age45 +
age55 + age65 + age75 + age85 + diag_1circulatory + diag_1diabetes +
diag_1digestive + diag_1injury + diag_1musculoskeletal + diag_1neoplasms +
diag_1other + diag_1pregnecy + time_in_hospital1 + time_in_hospital2 +
time_in_hospital3 + time_in_hospital4 + time_in_hospital5 + time_in_hospital6
+ time_in_hospital7 + time_in_hospital8 + time_in_hospital9 +
time_in_hospital10 + time_in_hospital11 + time_in_hospital12 +
time_in_hospital13 + discharge_disposition_id7 + discharge_disposition_id18 +
admission_type_id1 + admission_type_id3 + num_of_changes + number_inpatient +
diag_2circulatory + diag_2diabetes + diag_2digestive + diag_2injury +
diag_2musculoskeletal + diag_2other', data = train1).fit()
```

(Link) Coefficients on GitHub

Performance Improvement (Train Data)

Original Model	Optimized model
Classification accuracy = 57.2%	Classification accuracy = 65.4%
Precision = 58.4%	Precision = 64.3%
TPR or Recall = 48.3%	TPR or Recall = 69.2%
FNR = 51.7%	FNR = 30.8%
FPR = 34.0%	FPR = 38.4%
ROC-AUC = 59.9%	ROC-AUC = 70.7%

Performance Improvement (Train Data)

Original Model	Optimized model	
Classification accuracy = 57.2%	Classification accuracy = 65.4%	
Precision = 58.4%	Precision = 64.3%	↑ 5.9%
TPR or Recall = 48.3%	TPR or Recall = 69.2%	
FNR = 51.7%	FNR = 30.8%	
FPR = 34.0%	FPR = 38.4%	
ROC-AUC = 59.9%	ROC-AUC = 70.7%	

Performance Improvement (Train Data)

Original Model	Optimized model	
Classification accuracy = 57.2%	Classification accuracy = 65.4	
Precision = 58.4%	Precision = 64.3%	
TPR or Recall = 48.3%	TPR or Recall = 69.2%	↑ 20.9%
FNR = 51.7%	FNR = 30.8%	
FPR = 34.0%	FPR = 38.4%	
ROC-AUC = 59.9%	ROC-AUC = 70.7%	

Performance Improvement (Train Data)

Original Model	Optimized model
Classification accuracy = 57.2%	Classification accuracy = 65.4
Precision = 58.4%	Precision = 64.3%
TPR or Recall = 48.3%	TPR or Recall = 69.2%
FNR = 51.7%	FNR = 30.8%
FPR = 34.0%	FPR = 38.4%
ROC-AUC = 59.9%	ROC-AUC = 70.7%

↓ 20.9%

Inference Results and Preliminary Conclusions

Preliminary Model Conclusions

Significant predictors for readmission

Patient age

Days spent in hospital during their initial stay

Primary and secondary diagnoses

Number of inpatient hospital visits

Admission type (classification of patient hospital entrance, i.e., emergency, elective, etc.)

Discharge disposition (classification of patient hospital exit)

Number of changes to patient medications

Preliminary Model Conclusions

Significant predictors for readmission

Patient age

Days spent in hospital during their initial stay

Primary and secondary diagnoses

Number of inpatient hospital visits

Admission type (classification of patient hospital entrance, i.e., emergency, elective, etc.)

Discharge disposition (classification of patient hospital exit)

Number of changes to patient medications

← All p-values < 0.05

Preliminary Model Conclusions

Significant predictors for readmission

Patient age

Days spent in hospital during their initial stay

Primary and secondary diagnoses

Number of inpatient hospital visits

Admission type (classification of patient hospital entrance, i.e., emergency, elective, etc.)

Discharge disposition (classification of patient hospital exit)

Number of changes to patient medications

← All p-values < 0.05

Connecting results to recommendations

Connecting results with recommendations

Odds ratio calculations to determine the most salient predictors for readmission

What patient demographics are most at risk

Inform patients and their healthcare administrators about elevated risk levels to alter healthcare treatment plan

Next Steps

Model making

Cross-validation with scikit-learn

Attempt Lasso methods

Add interaction terms between significant predictors

Thank you!

Let us know if you have questions or clarifications.