Predicting patient readmission

Jack Etheredge

What’s the problem?

Many patients are readmitted early to the hospital

What’s the problem?

Many patients are readmitted early to the hospital **(<30 days from discharge, possibly due to being discharged too soon)**

Releasing from the hospital too early (or too late) decreases patient health and satisfaction

vectorstock.com/royalty-free-vector/cancer-patient-on-wheelchair-with-sad-happy-face-vector-16429517

nytimes.com/2016/01/05/upshot/the-hidden-financial-incentives-behind-your-shorter-hospital-stay.html

And it can also (now) lose the hospital money

“The federal government has created several new programs that penalize hospitals for readmissions. Under Medicare’s Hospital Readmissions Reduction Program, hospitals now lose up to 3 percent of their total Medicare payments for high rates of patients readmitted within 30 days of discharge.”

nytimes.com/2016/01/05/upshot/the-hidden-financial-incentives-behind-your-shorter-hospital-stay.html

How will we address the problem?

Treat diabetic patients as a (hopefully generalizable) case study in avoiding patient early release

Data from this publication:

BioMed Research International
Volume 2014, Article ID 781670, 11 pages
http://dx.doi.org/10.1155/2014/781670

Research Article

Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records

Beata Strack,[1] Jonathan P. DeShazo,[2] Chris Gennings,[3] Juan L. Olmo,[4] Sebastian Ventura,[4] Krzysztof J. Cios,[1,5] and John N. Clore[6]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA
[2]Department of Health Administration, Virginia Commonwealth University, Richmond, VA 23298, USA
[3]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298, USA
[4]Department of Computer Science and Numerical Analysis, University of Cordoba, 14071 Cordoba, Spain
[5]IITiS Polish Academy of Sciences, 44-100 Gliwice, Poland
[6]Department of Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA

Received 19 November 2013; Revised 25 February 2014; Accepted 7 March 2014; Published 3 April 2014

Academic Editor: Ali Rizvi

hindawi.com/journals/bmri/2014/781670/

Data:

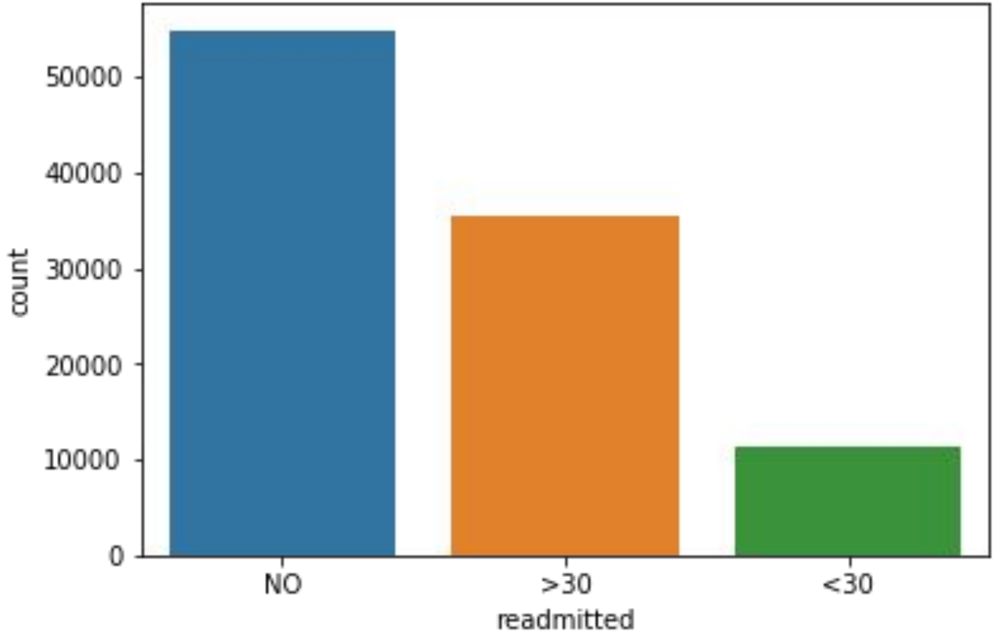
The dataset represents 10 years (1999-2008) of clinical care
130 US hospitals and integrated delivery networks.
Over 50 features representing patient and hospital outcomes.

Example features (not exhaustive):

1. Number of Laboratory tests
2. Which medications
3. Specialty of the physician

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

What does the data look like?



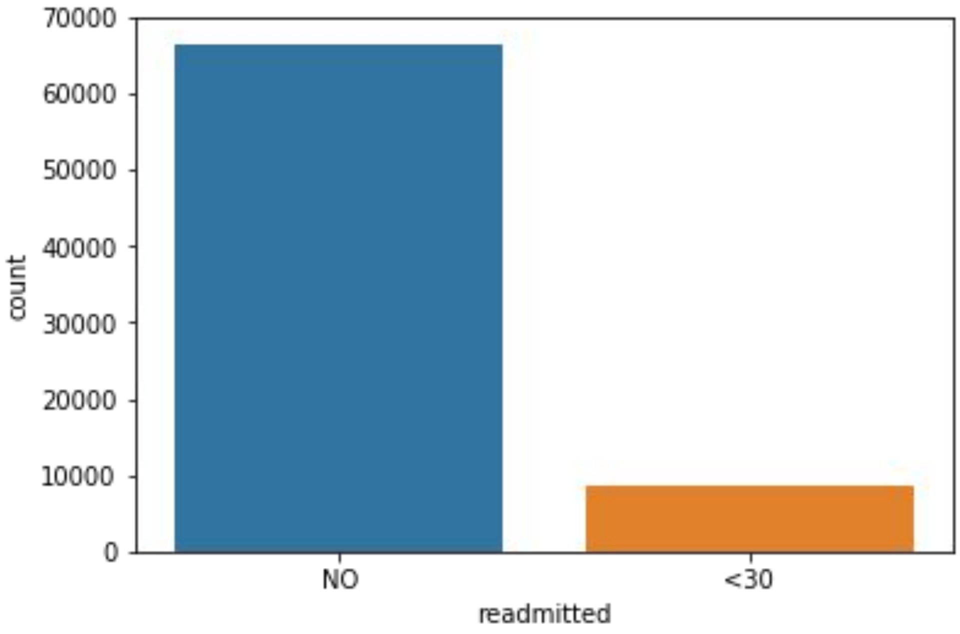
Any special considerations?

Biased classes

I undersampled and oversampled

throughout to combat this

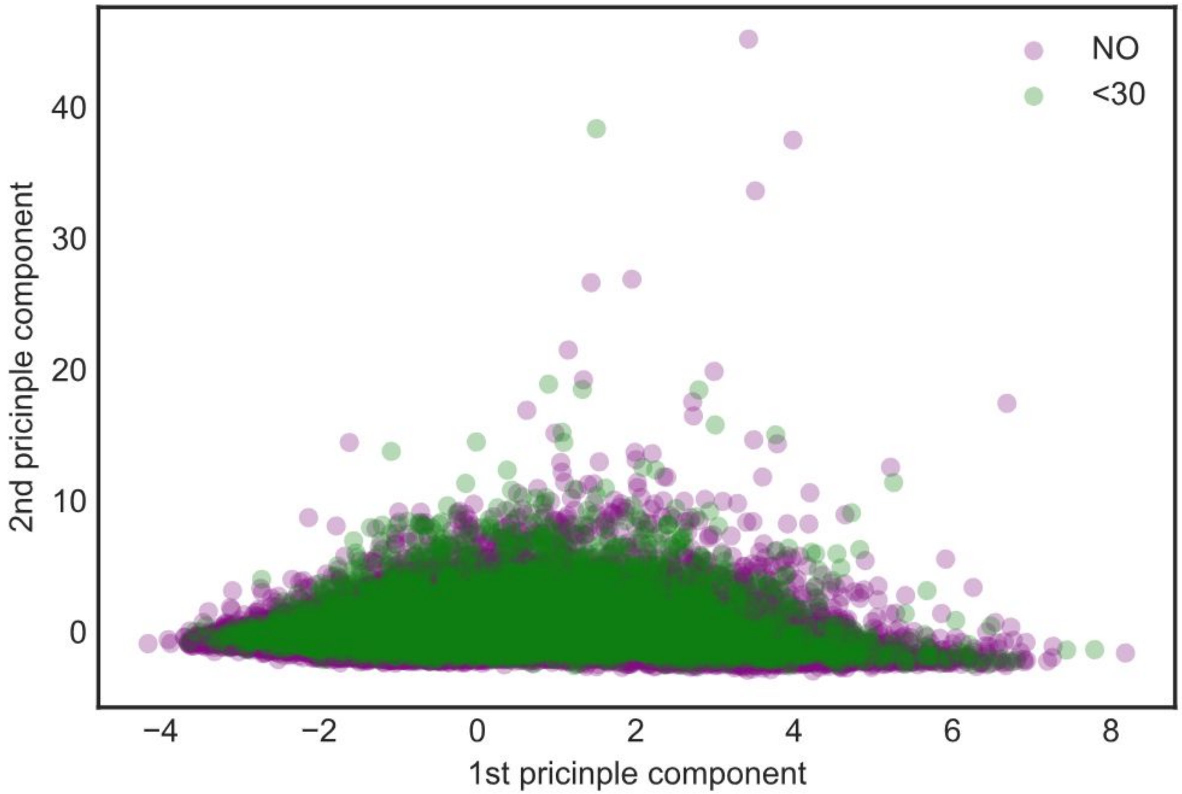
Making the problem simpler and better



Reducing the classes to readmitted within <30 days or not

This makes the data better suited to address both patient health and hospital cost savings

These populations are not easily separable



Data cleaning (reducing model performance, but giving us more realistic insights):

1. Removed expired patients (since they are not readmitted and are irrelevant to the current problem)

2. Removed return visits (repeated patient numbers)

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

Living only, binary class (<30 vs >30, no readmission):

Model performance (recall)

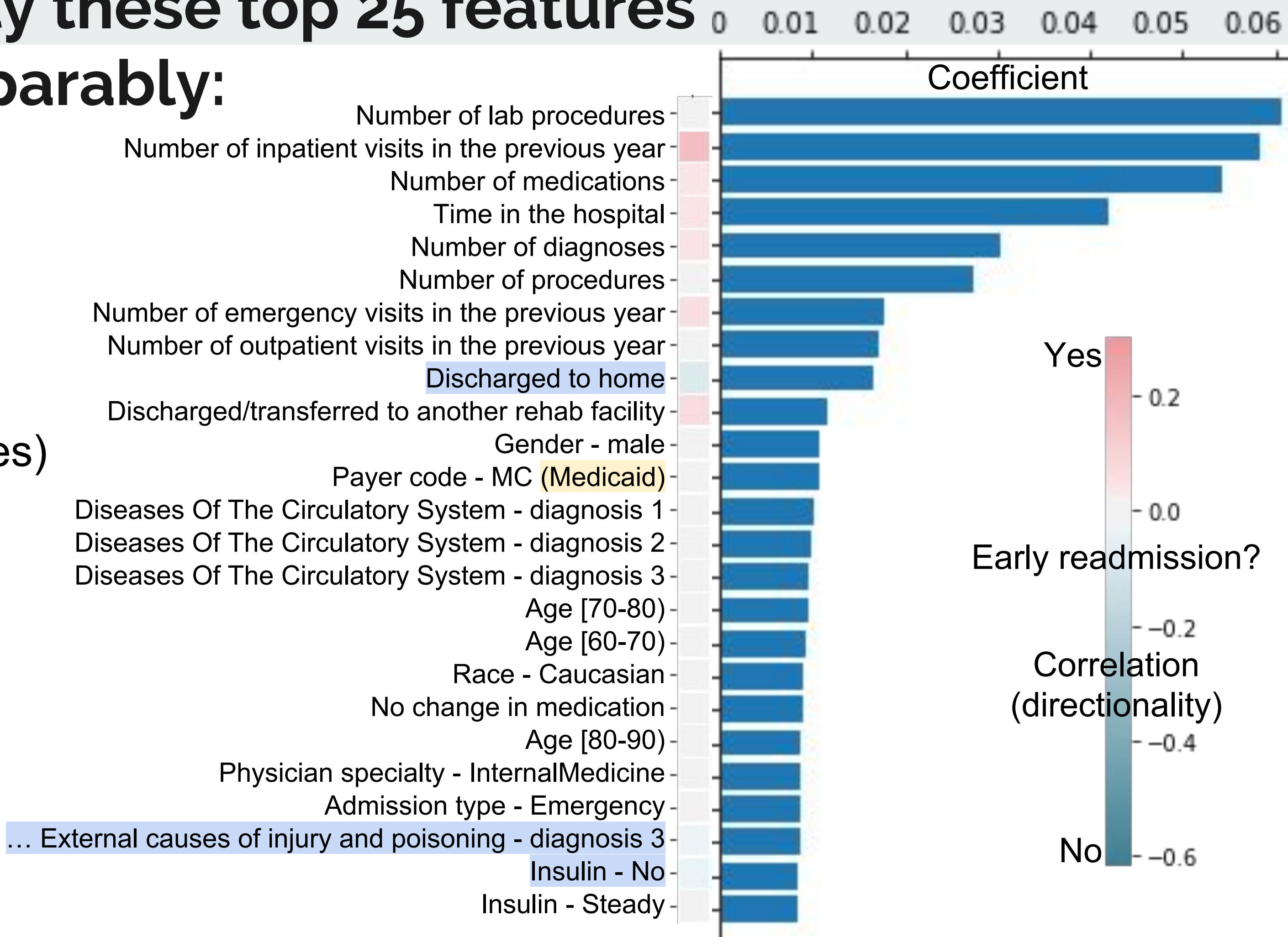
Model	SMOTE oversampling	random undersampling
Logistic Regression	58%, 65%	58%, 68%
KNN (k=5)		
Linear SVM	58%, 65%	58%, 67%
SVM (RBF)	50%, 72%	56%, 69%
Random forest	5%, 99%	63%, 63%
Boosted trees	2%, 100%	58%, 67%
Bernoulli Naive Bayes	96%, 6%	64%, 60%
Gaussian Naive Bayes		

Test recall
Early readmission
Yes, No

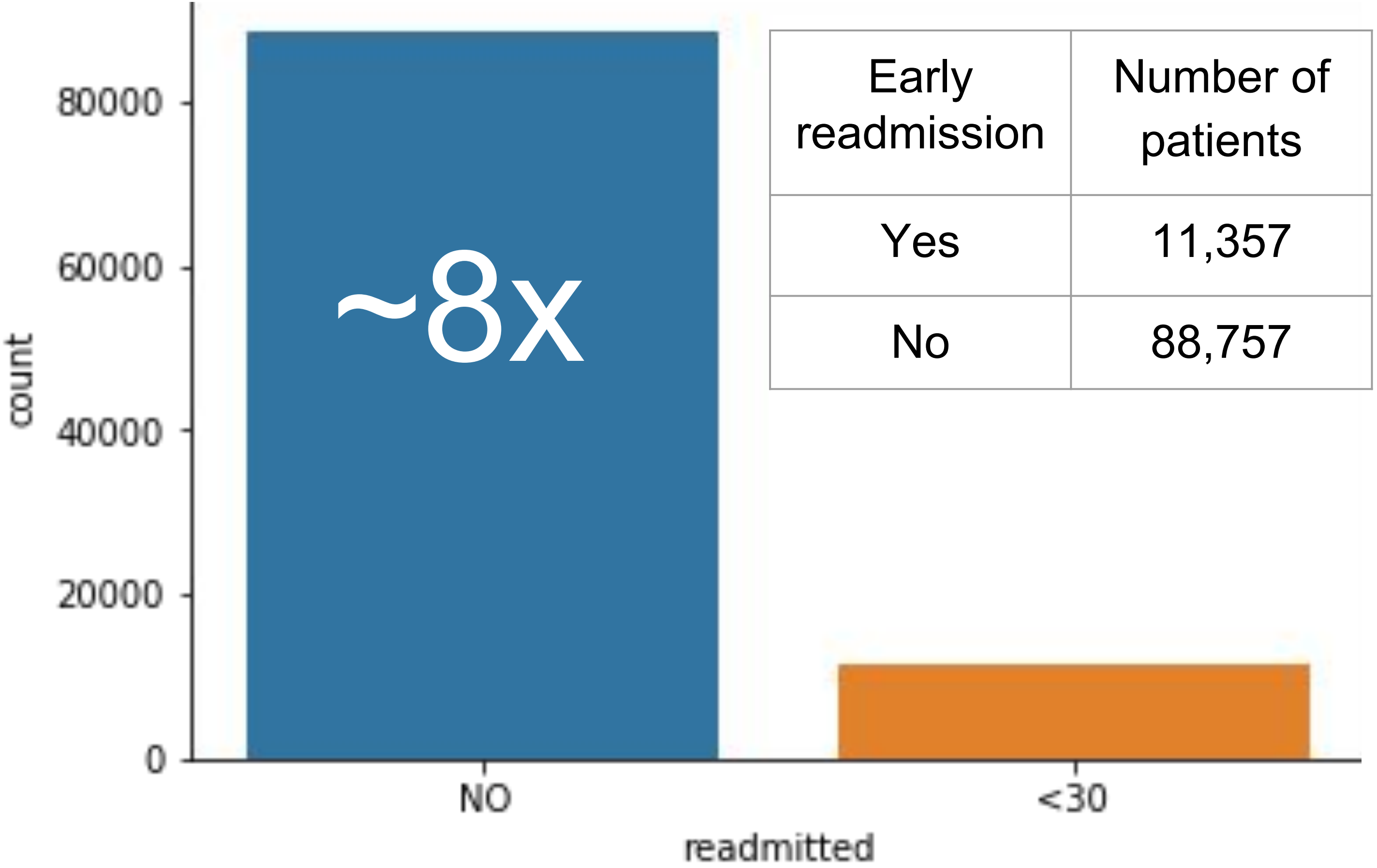
Model with only these top 25 features performs comparably:

Accuracy and recall of 62% vs 63% for all features

(Originally 262 features)



8x as many patients not readmitted within 30 days

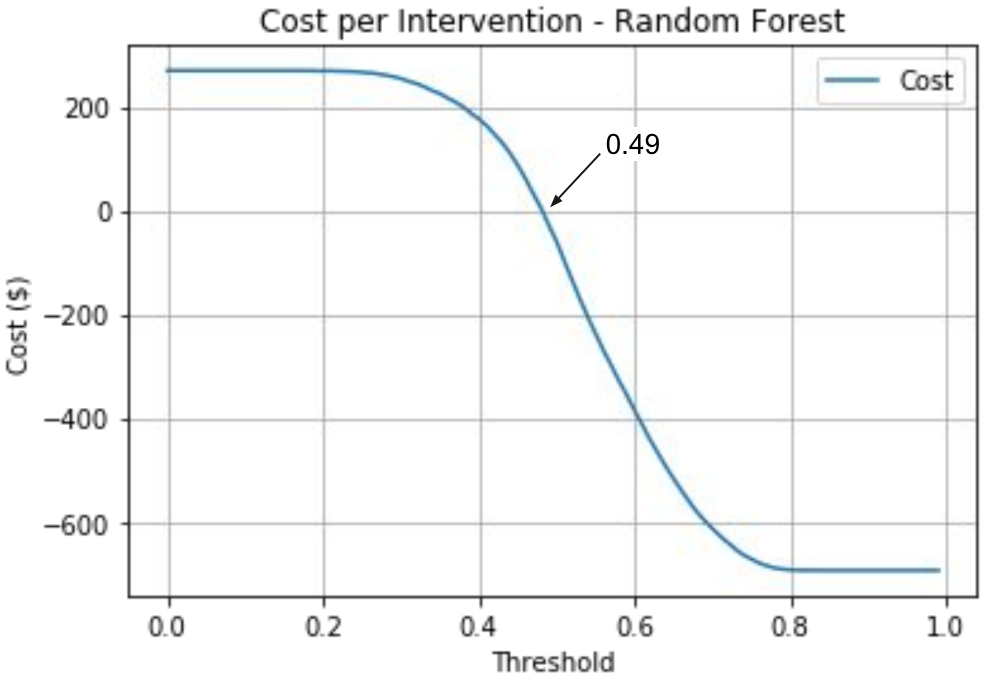


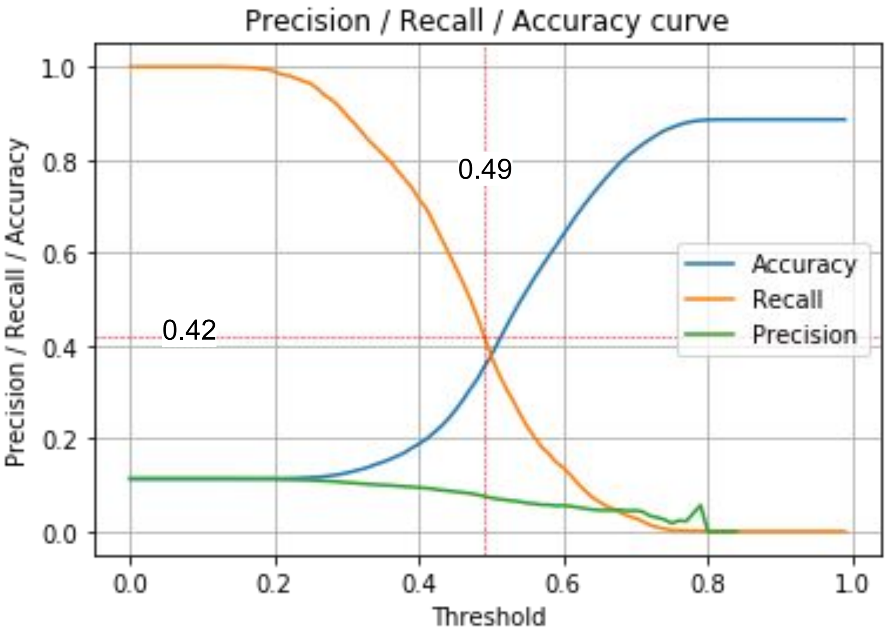
Early readmission	Number of patients
Yes	11,357
No	88,757

True positive
1765
False negative
1074
False positive
8395
True negative
13795
Actual
Early readmission
Not
Predicted
Early readmission
Not
12500
10000
7500
5000
2500

A threshold of 0.49 is the break-even point for cost, which gives us a recall of 0.42

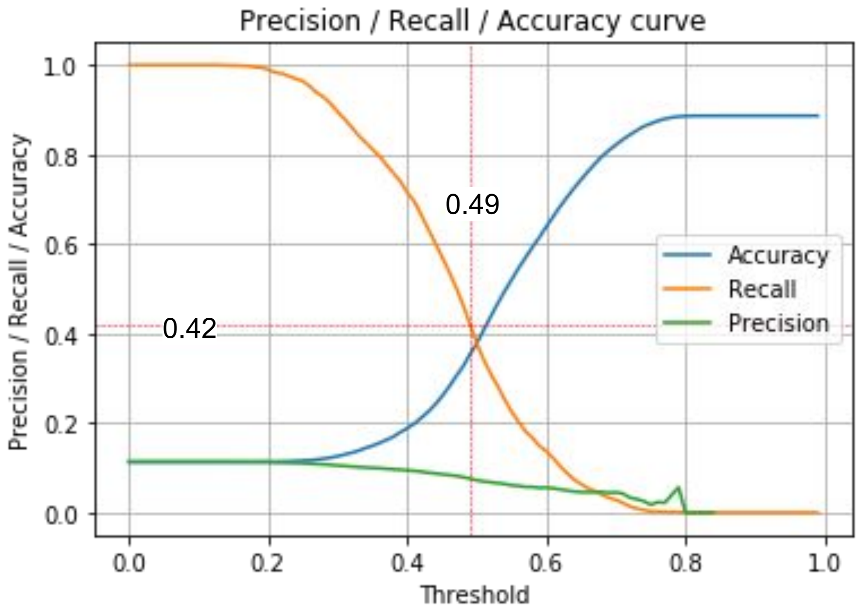
University of Michigan study showed that ~$304 last day for extended stay vs ~$1246 for first day





What are the actionable insights?

Using my 25-feature model with a threshold of 0.49 could prevent up to 42% of early readmissions and minimize early readmissions before costing the hospital additional money



What are the actionable insights?

Using my 25-feature model with a threshold of 0.49 could prevent up to 42% of early readmissions and minimize early readmissions before costing the hospital additional money

For high risk patients (identified by the features), follow up with phone calls

What are the actionable insights?

Using my 25-feature model with a threshold of 0.49 could prevent up to 42% of early readmissions and minimize early readmissions before costing the hospital additional money

For high risk patients (identified by the features), follow up with phone calls

What are the actionable insights?

Using my 25-feature model with a threshold of 0.49 could prevent up to 42% of early readmissions and minimize early readmissions before costing the hospital additional money

For high risk patients (identified by the features), follow up with phone calls

Some patients have a disproportionately high rate of early readmission

ex: Many lab procedures and medication, primary diagnosis of cardiovascular disease

For patients that are at high risk of readmission within 30 days, make sure that home care and instructions are adequate if keeping them longer in the hospital is not advisable

Phone-call follow-ups reduce readmission (and increase physician office visits, for net savings): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771544/

Future aims:

Ask for more data from the VCU researchers

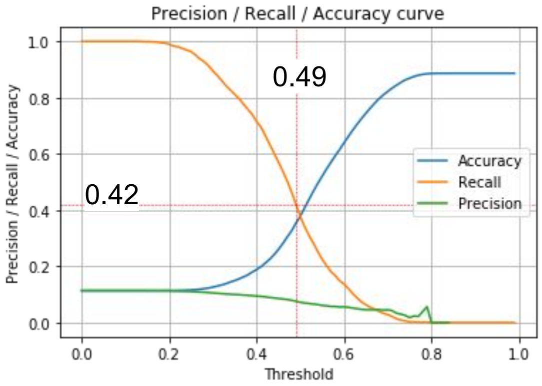
Web visualizations with D3

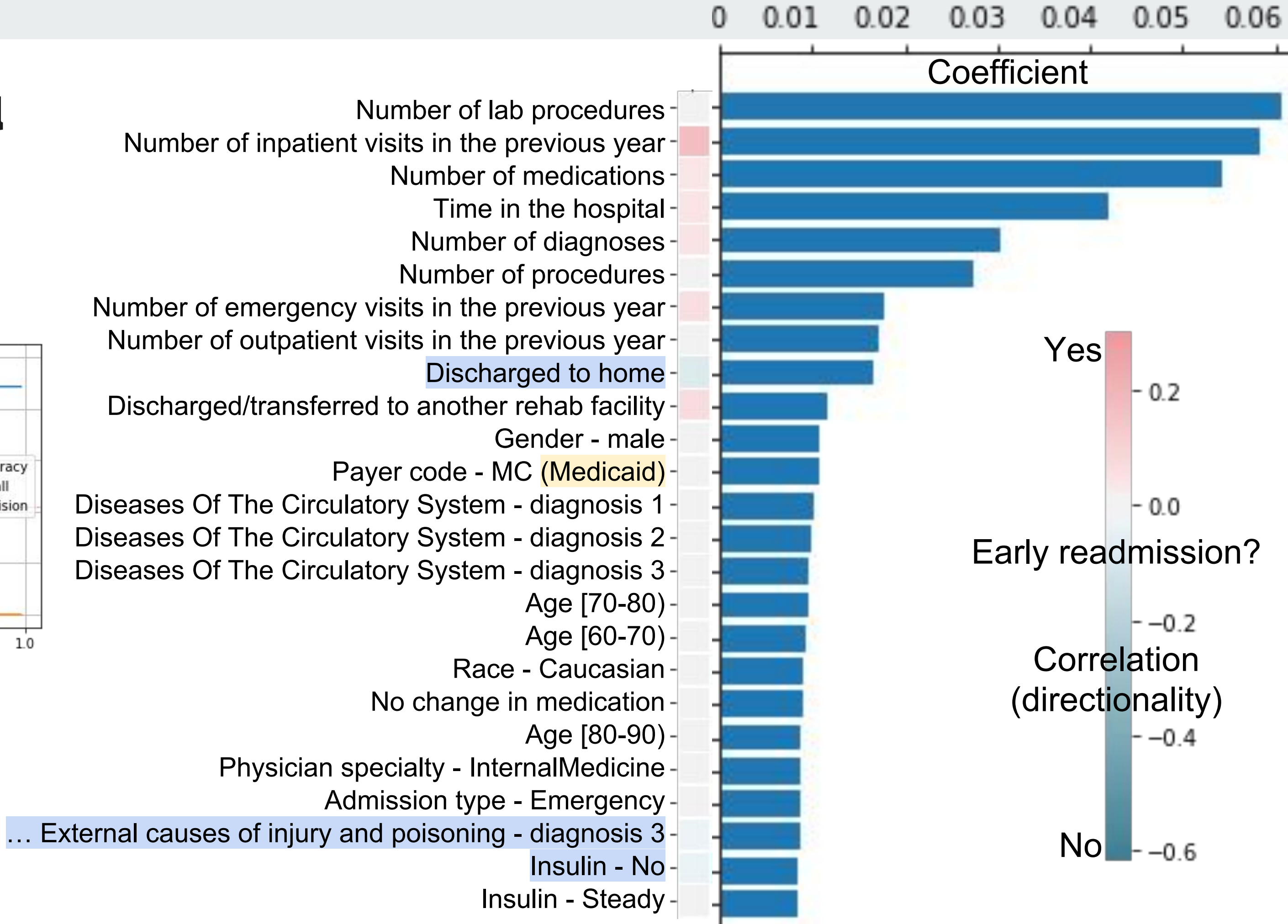
Improve my Flask prediction app (not shown for time, but working)

Take all my best models (of different types) and ensemble them

Thank you

Questions?





Citations/sources:

Variable costs per day of hospital stay:
Median cost first day $1246
Median cost last day $304
https://www.journalacs.org/article/S1072-7515(00)00352-5/fulltext

$2,289 nonprofit hospital average/day:
https://www.beckershospitalreview.com/finance/average-cost-per-inpatient-day-across-50-states.html

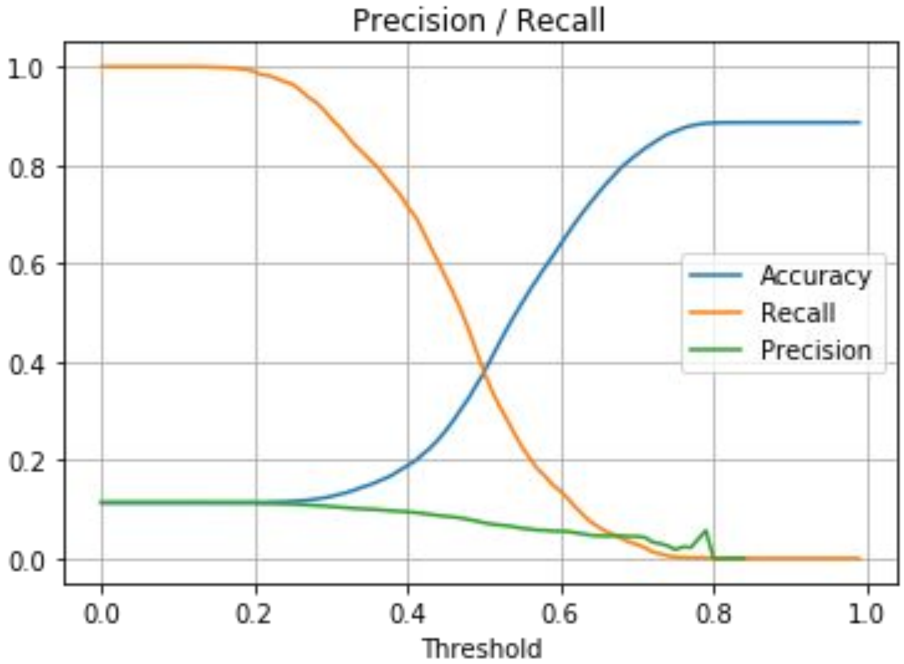
Phone-call follow-ups reduce readmission (and increase physician office visits, for net savings):
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771544/

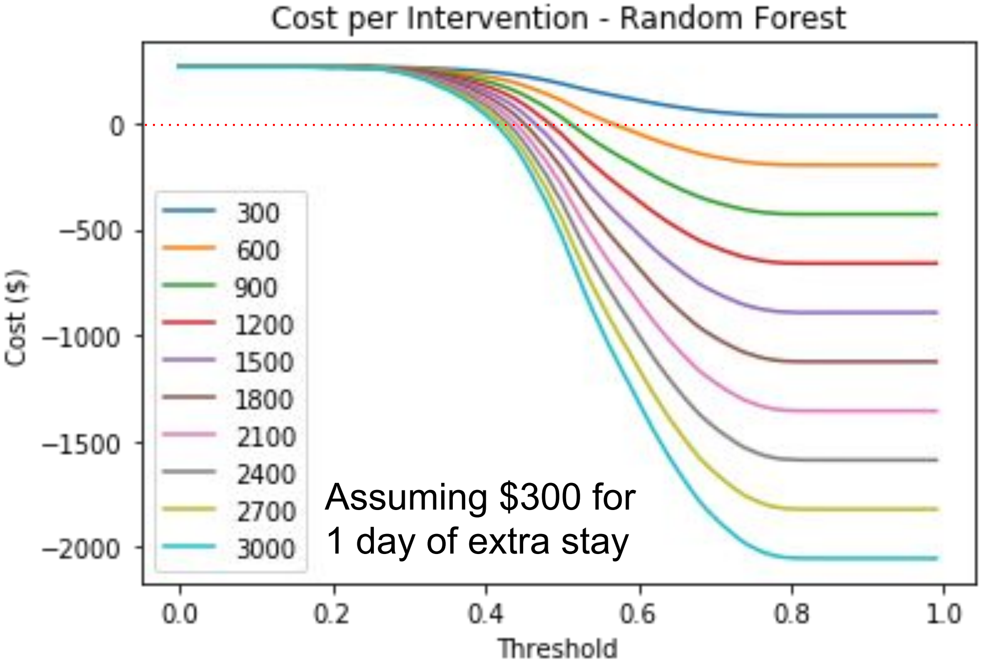
One extra day in the hospital reduces deaths and readmissions:
https://www8.gsb.columbia.edu/newsroom/newsn/3251/one-extra-day-in-the-hospital-cuts-readmission-rates-and-reduces-patient-deaths

Hospital stay duration vs readmission:
https://www.surgjournal.com/article/S0039-6060(17)30395-1/pdf

Supporting slides:

Threshold optimization for different cost ratios between readmission and extra day in hospital





Assuming $300 for
1 day of extra stay

Cost function explanation

```
cost_ben_log[i] = (FPsum*day_cost + TPsum*(day_cost-readmit_cost) + FNsum*(readmit_cost))/len(y_test_num)
# FP = stay in hospital, but didn't need to, costs $304/day
# TP = stay in hospital and needed to, costs $304/day but saves $1246
# FN = readmitted, but should have stayed in hospital, saved $304, but costed $1246
# TN = left hospital, not readmitted, no cost
```

“Manual” optimization of the 25-feature model

In [144]:

```
randomforest = RandomForestClassifier(n_estimators=300, min_samples_split=70)
randomforest.fit(x_train_undersampled, y_train_undersampled)
randomforest.score(x_test, y_test)
y_pred = randomforest.predict(x_test)
print("Accuracy: %.3f"% metrics.accuracy_score(y_test, y_pred))
print(metrics.classification_report(y_test, y_pred))
print(metrics.confusion_matrix(y_test, y_pred))
```

executed in 5.25s, finished 13:11:17 2018-05-15

```
Accuracy: 0.622
             precision    recall  f1-score   support

        <30       0.17      0.62      0.27      2839
         NO       0.93      0.62      0.74     22190

avg / total       0.84      0.62      0.69     25029

[[ 1765  1074]
 [ 8395 13795]]
```

Grid search optimization of the 25-feature model

```
randomforest = RandomForestClassifier(bootstrap=False, class_weight=None, criterion='gini',
            max_depth=70, max_features='sqrt', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=2, min_samples_split=10,
            min_weight_fraction_leaf=0.0, n_estimators=1200, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

Presentation notes

Insurance companies and patients are the target for the cost-benefit part

Nuances

Nuances

Optimizing length of stay (LOS) is its own problem

Longer isn't necessarily better

Length of stay vs readmission is both a health and economic challenge

Many papers and posts have been written about both of these problems

Nuances

Optimizing length of stay (LOS) is its own problem

Longer isn’t necessarily better

Length of stay vs readmission is both a health and economic challenge

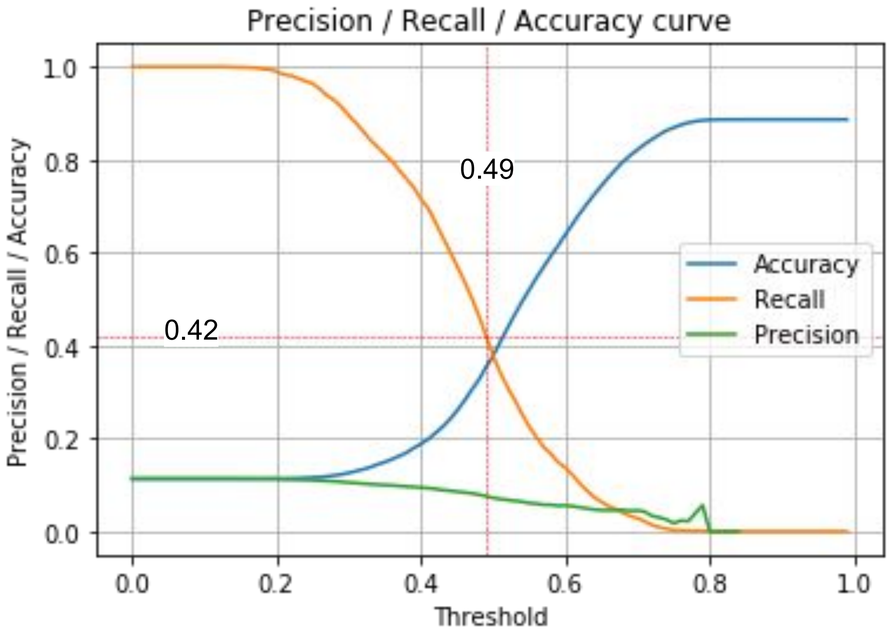
Many papers and posts have been written about both of these problems

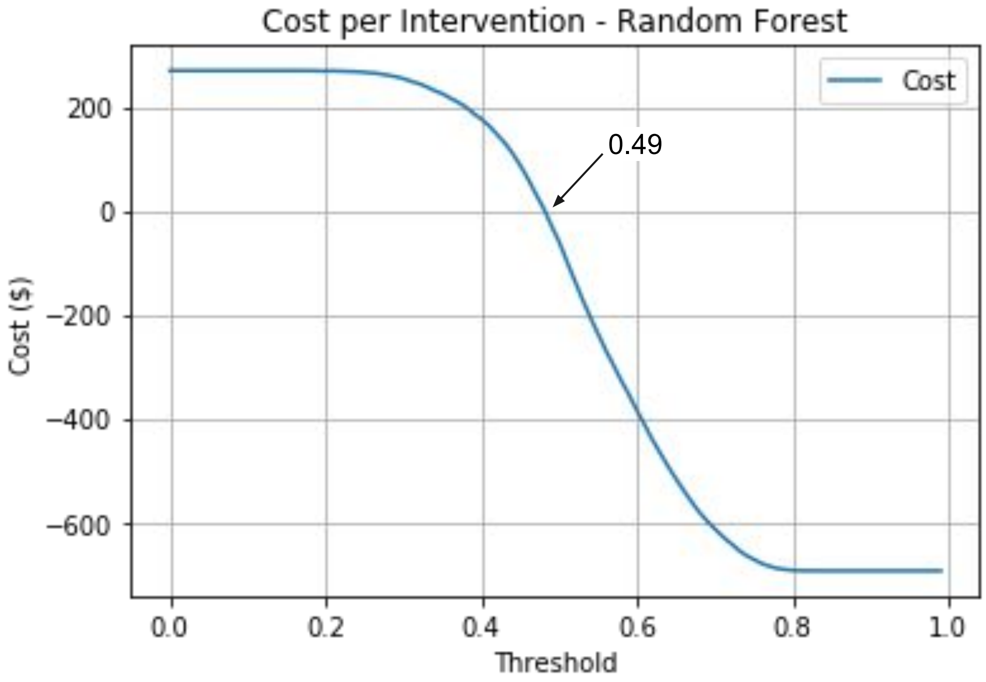
Some readmissions may not be avoidable by an increased duration of stay

False positive rate is rather costly here, particularly since there are ~8 times as many patients that aren’t readmitted within 30 days

A threshold of 0.49 is the break-even point for cost, which gives us a recall of 0.42

University of Michigan study showed that ~$304 last day for extended stay vs ~$1246 for first day





Model performance (accuracy)

	Initial performance	Scaled	Test/Train split (25/75)	10-fold cross-validation
Logistic Regression	54%	60%	58%	58%
KNN (k=5)	66%			
Linear SVM			56%	
SVM (RBF)				
Random forest				
Bernoulli Naive Bayes			57%	
Gaussian Naive Bayes			15%	

Model performance (recall)

	Test/Train split (25/75)	10-fold cross-validation
Logistic Regression	3, 37, 83	
KNN (k=5)		
Linear SVM	20, 44, 70 (balanced class weight)	
SVM (RBF)	38, 43, 61 (balanced class weight)	
Random forest	5, 44, 71	
Boosted trees	3, 36, 84	
Bernoulli Naive Bayes	4, 45, 77	
Gaussian Naive Bayes	96, 2, 10	

Model performance (SMOTE oversampling, recall)

	Test/Train split (25/75)	10-fold cross-validation
Logistic Regression	45, 38, 57	
KNN (k=5)		
Linear SVM	45, 37, 57	
SVM (RBF)		
Random forest	7, 43, 69	
Boosted trees	7, 45, 70	
Bernoulli Naive Bayes	6, 48, 72	
Gaussian Naive Bayes	94, 4, 11	

Model performance (random undersampling, recall)

	Test/Train split (25/75)	10-fold cross-validation
Logistic Regression	44, 39, 57	
KNN (k=5)		
Linear SVM	45, 38, 57	
SVM (RBF)	40, 37, 63	
Random forest	48, 37, 57	
Boosted trees	27, 44, 63	
Bernoulli Naive Bayes	45, 37, 56	
Gaussian Naive Bayes	18, 87, 10	

Notes to myself:

Future plans: updated Naive Bayes

Try Gaussian first, then try ensemble with Gaussian/Bernoulli/Multinomial for the three different types of columns I have (float, categorical, count)

Future work:

Ask for more data from the VCU researchers

Web visualizations and predictions with Flask and D3

Try at least implementing flask prediction?

Try dropping variables in random forest and seeing what happens

Try creating some interaction terms in random forest and seeing what happens (?)

Take all my best models (of different types) and ensemble them (Bagging?)

Votingclassifier / do it manually

Things implemented from “to do”

Transform x_test after doing fit_transform on x_train, or use a pipeline, since the pipeline fits on train, but not on test, when you’re predicting

Remove duplicate patients

> For each “Patient_nbr”, take the lowest “encounter_id”
> Feature engineering - add whether they were entered multiple times? Difficult if we’re keeping the first or last, but second to last could work

Cost benefit analysis

Older versions of slides

What’s the problem?

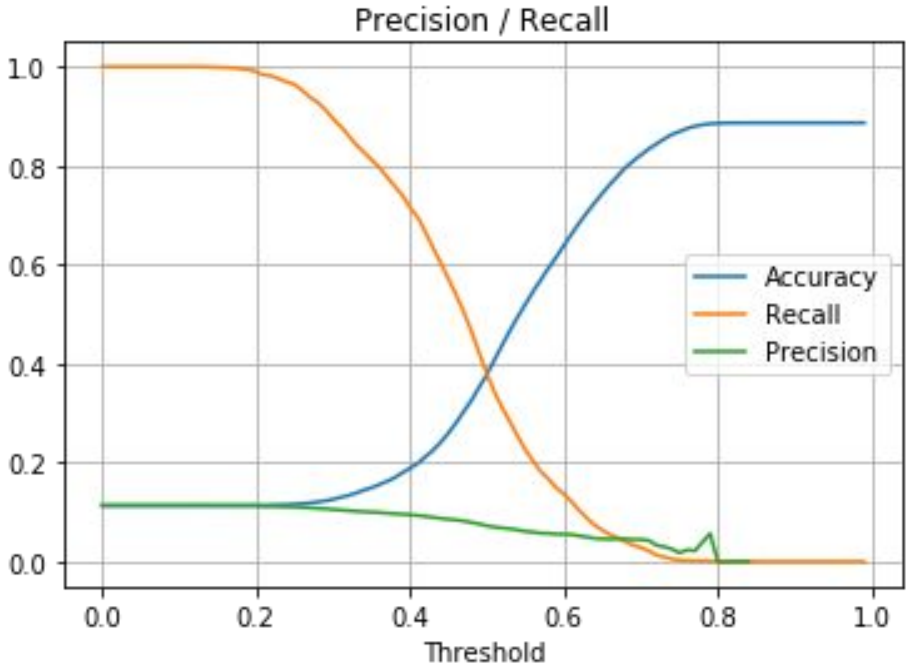
Many patients are readmitted early to the hospital

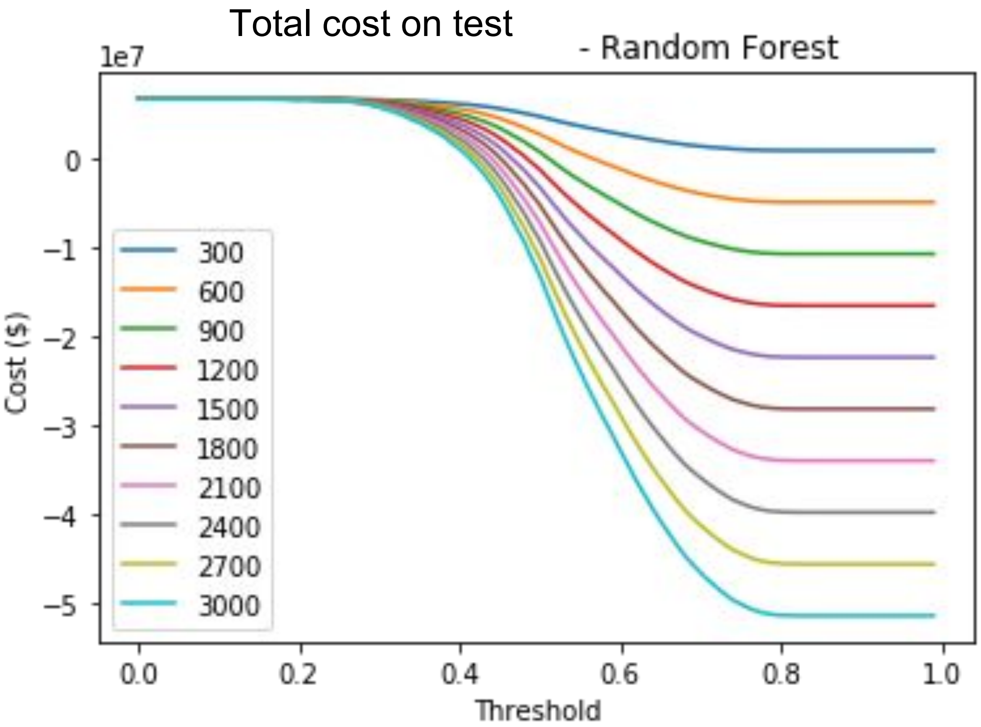
Kernel RBF SVM (slightly worse accuracy, but FAR better recall in the classes we care about):

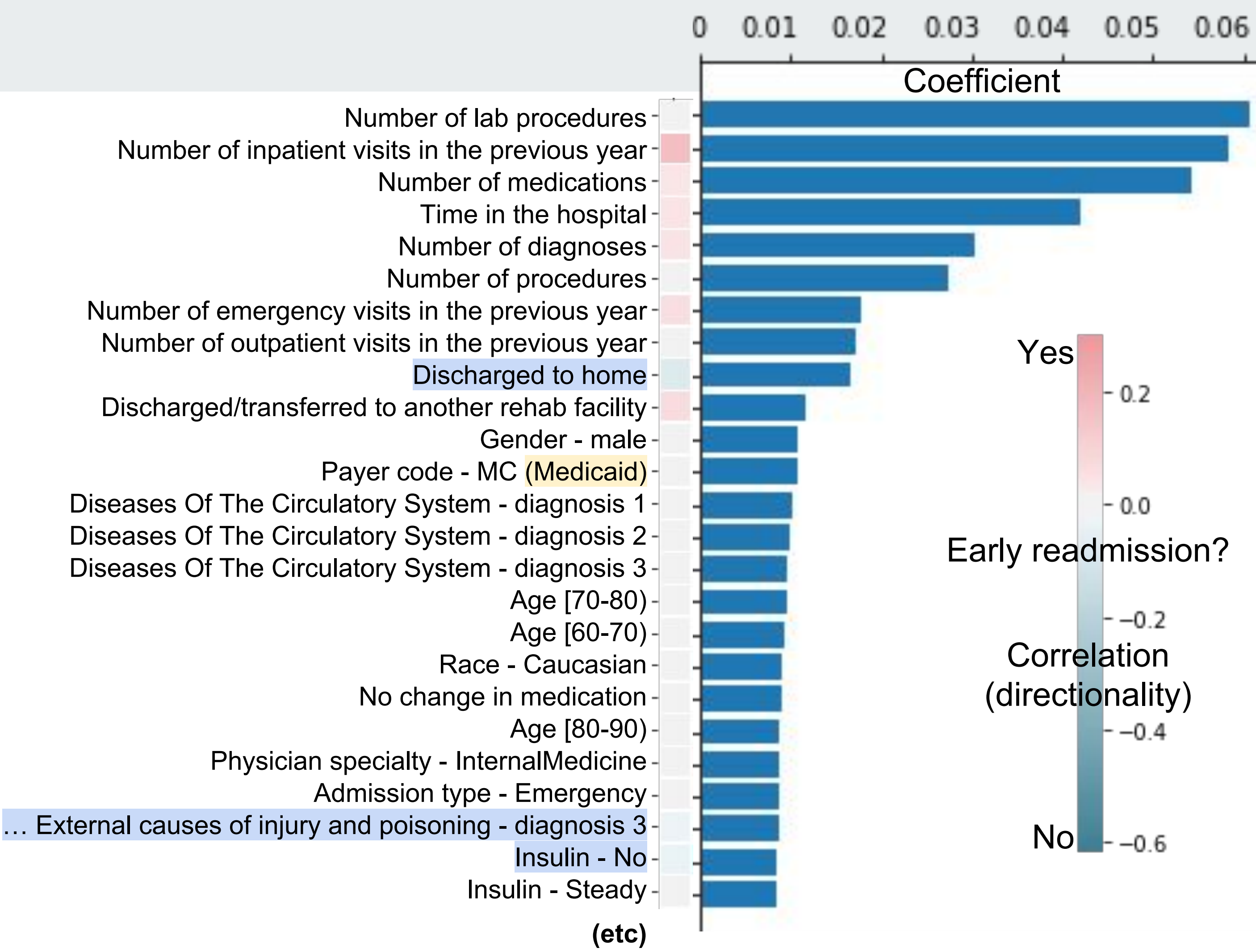
Accuracy: 0.520

	precision	recall	f1-score	support
<30	0.21	0.38	0.27	2839
>30	0.48	0.43	0.45	8887
NO	0.68	0.61	0.64	13716
avg / total	0.56	0.52	0.53	25442

Threshold optimization for different cost ratios between readmission and extra day in hospital

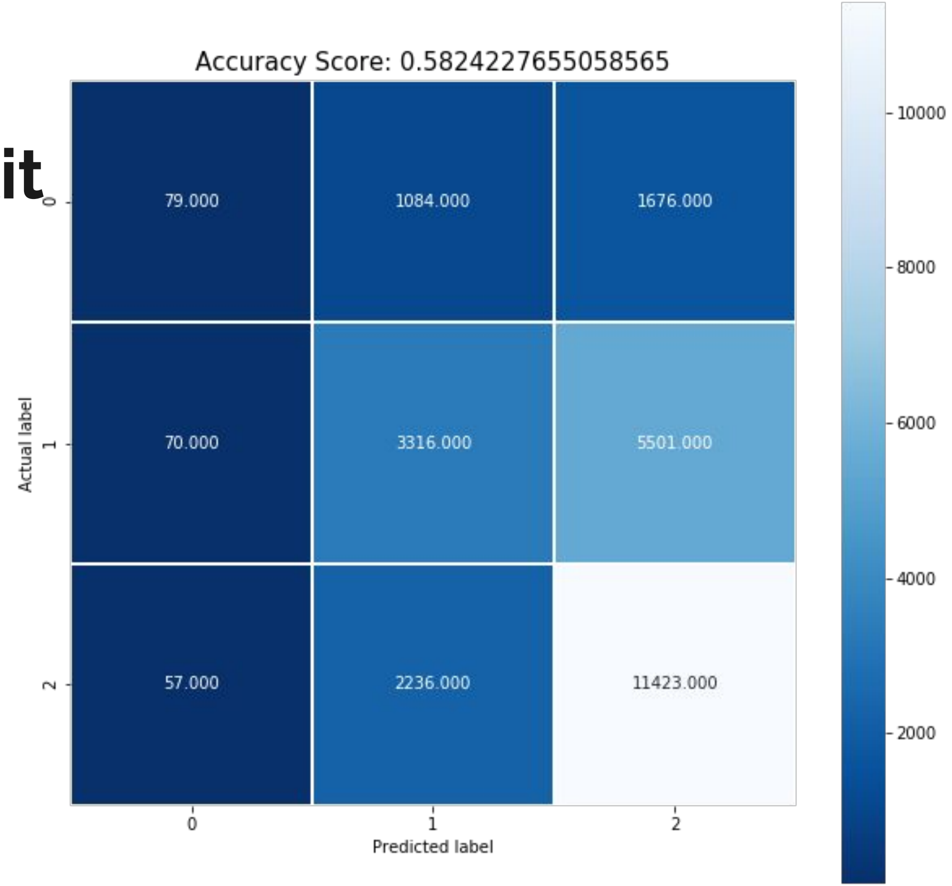





0
0.01
0.02
0.03
0.04
0.05
0.06
Coefficient
Number of lab procedures
Number of inpatient visits in the previous year
Number of medications
Time in the hospital
Number of diagnoses
Number of procedures
Number of emergency visits in the previous year
Number of outpatient visits in the previous year
Discharged to home
Discharged/transferred to another rehab facility
Gender - male
Payer code - MC (Medicaid)
Diseases Of The Circulatory System - diagnosis 1
Diseases Of The Circulatory System - diagnosis 2
Diseases Of The Circulatory System - diagnosis 3
Age [70-80)
Age [60-70)
Race - Caucasian
No change in medication
Age [80-90)
Physician specialty - InternalMedicine
Admission type - Emergency
… External causes of injury and poisoning - diagnosis 3
Insulin - No
Insulin - Steady
(etc)
Yes
0.2
0.0
Early readmission?
−0.2
Correlation
(directionality)
−0.4
No
−0.6

Logistic Test/Train Split

Confusion matrix



Model performance

Model vs accuracy:

KNN

SVM

Random forest

Logistic Regression

Gaussian Naive Bayes

Data source:

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Diabetes 130-US hospitals for years 1999-2008 Data Set

Download: Data Folder, Data Set Description

Abstract: This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes.

Data Set Characteristics:	Multivariate	**Number of Instances:**	100000	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	55	**Date Donated**	2014-05-03
Associated Tasks:	Classification, Clustering	**Missing Values?**	Yes	**Number of Web Hits:**	169416

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

Data:

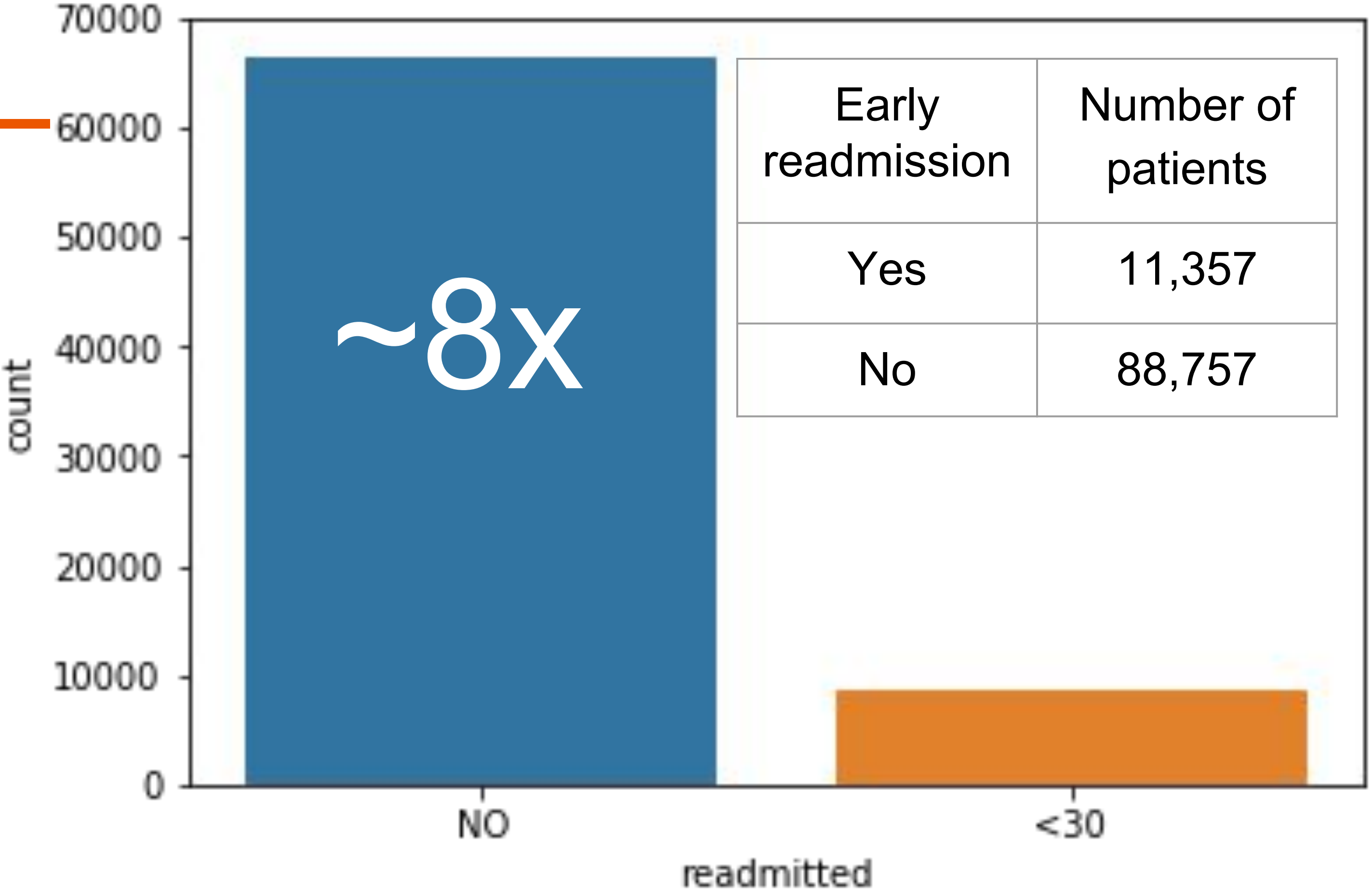
The dataset represents 10 years (1999-2008) of clinical care
130 US hospitals and integrated delivery networks.
Over 50 features representing patient and hospital outcomes.

(1) It is an inpatient encounter (a hospital admission).
(2) Some kind of diabetes was entered as a diagnosis.
(3) The length of stay was at least 1 day and at most 14 days.
(4) Laboratory tests were performed during the encounter.
(5) Medications were administered during the encounter.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

8x as many patients not readmitted within 30 days



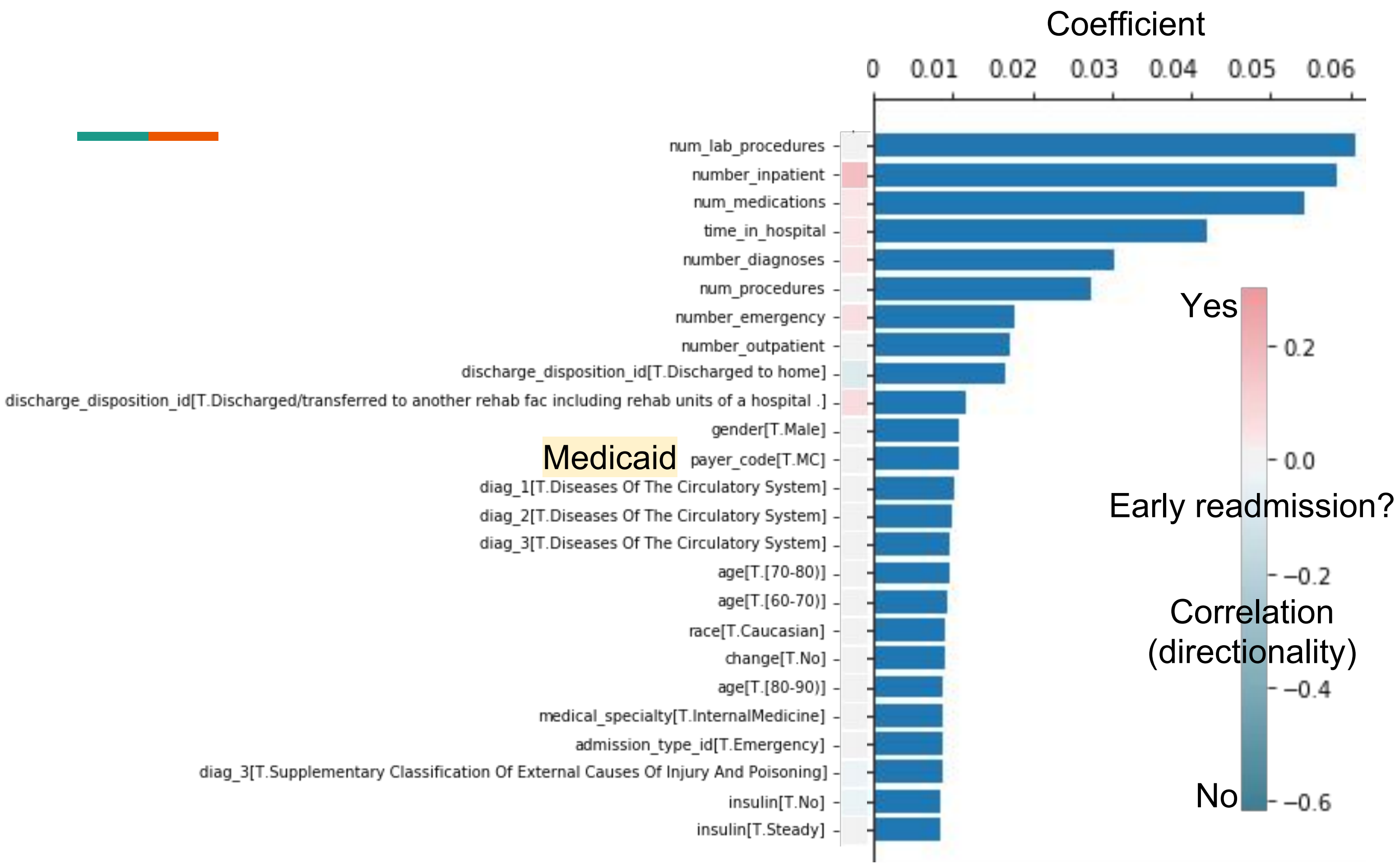
Early readmission	Number of patients
Yes	11,357
No	88,757

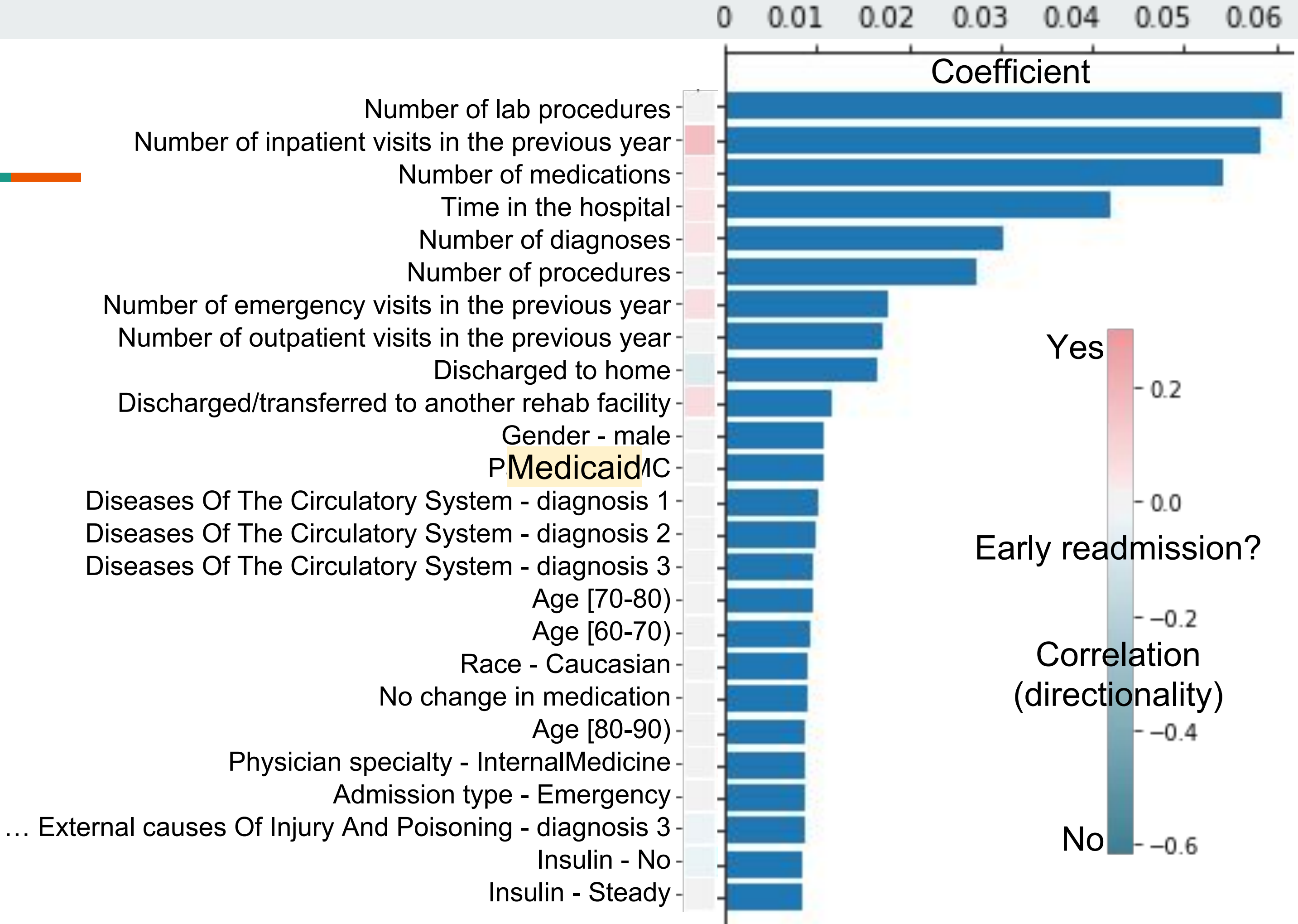
Payer code MC = medicaid

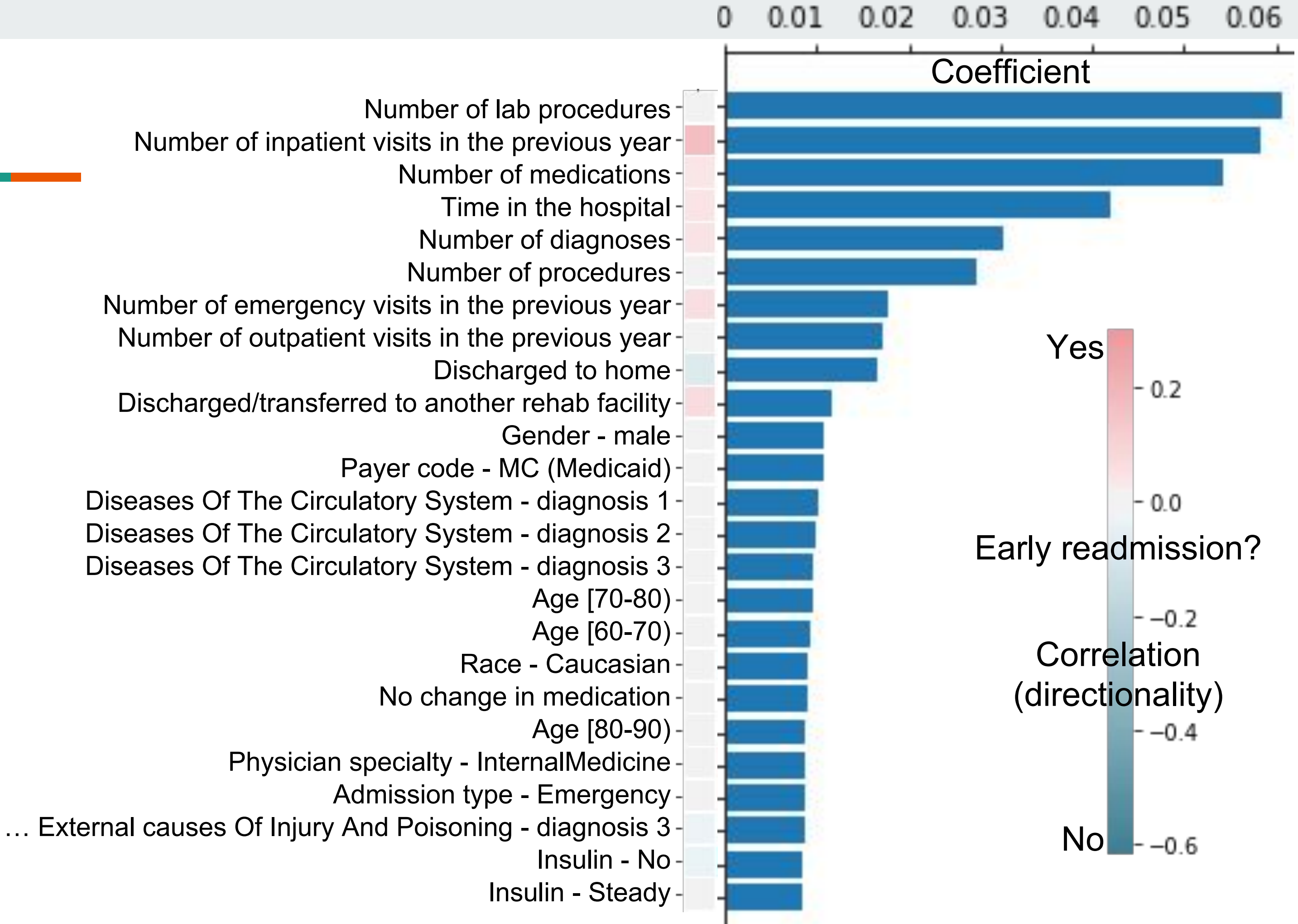
0.061	num_lab_procedures
0.058	number_inpatient
0.056	num_medications
0.040	time_in_hospital
0.029	number_diagnoses
0.028	num_procedures
0.018	number_emergency
0.017	discharge_disposition_id[T.Discharged to home]
0.017	number_outpatient
0.011	gender[T.Male]
0.010	discharge_disposition_id[T.Discharged/transferred to another rehab fac including rehab units of a hospital .]
0.010	payer_code[T.MC]
0.010	diag_2[T.Diseases Of The Circulatory System]

0.009	diag_1[T.Diseases Of The Circulatory System]
0.009	age[T.[70-80]
0.009	age[T.[60-70]
0.009	diag_3[T.Diseases Of The Circulatory System]
0.009	medical_specialty[T.InternalMedicine]
0.009	race[T.Caucasian]
0.009	insulin[T.No]
0.009	change[T.No]
0.009	admission_type_id[T.Emergency]
0.008	age[T.[80-90]

0.008	diag_3[T.Supplementary Classification Of External Causes Of Injury And Poisoning]
0.008	admission_type_id[T.Urgent]
0.008	age[T.[50-60]
0.008	diag_2[T.Supplementary Classification Of External Causes Of Injury And Poisoning]
0.008	race[T.AfricanAmerican]
0.008	insulin[T.Steady]
0.008	A1Cresult[T.None]
0.008	admission_source_id[T.Emergency Room]
0.008	discharge_disposition_id[T.Discharged/transferred to SNF]


Coefficient
0
0.01
0.02
0.03
0.04
0.05
0.06
num_lab_procedures
number_inpatient
num_medications
time_in_hospital
number_diagnoses
num_procedures
number_emergency
number_outpatient
discharge_disposition_id[T.Discharged to home]
discharge_disposition_id[T.Discharged/transferred to another rehab fac including rehab units of a hospital .]
gender[T.Male]
Medicaid
payer_code[T.MC]
diag_1[T.Diseases Of The Circulatory System]
diag_2[T.Diseases Of The Circulatory System]
diag_3[T.Diseases Of The Circulatory System]
age[T.[70-80)]
age[T.[60-70)]
race[T.Caucasian]
change[T.No]
age[T.[80-90)]
medical_specialty[T.InternalMedicine]
admission_type_id[T.Emergency]
diag_3[T.Supplementary Classification Of External Causes Of Injury And Poisoning]
insulin[T.No]
insulin[T.Steady]
Yes
0.2
0.0
Early readmission?
−0.2
Correlation
(directionality)
−0.4
No
−0.6


0
0.01
0.02
0.03
0.04
0.05
0.06
Coefficient
Number of lab procedures
Number of inpatient visits in the previous year
Number of medications
Time in the hospital
Number of diagnoses
Number of procedures
Number of emergency visits in the previous year
Number of outpatient visits in the previous year
Discharged to home
Discharged/transferred to another rehab facility
Gender - male
Medicaid
Diseases Of The Circulatory System - diagnosis 1
Diseases Of The Circulatory System - diagnosis 2
Diseases Of The Circulatory System - diagnosis 3
Age [70-80)
Age [60-70)
Race - Caucasian
No change in medication
Age [80-90)
Physician specialty - InternalMedicine
Admission type - Emergency
… External causes Of Injury And Poisoning - diagnosis 3
Insulin - No
Insulin - Steady
Yes
0.2
0.0
Early readmission?
−0.2
Correlation
(directionality)
−0.4
No
−0.6


0
0.01
0.02
0.03
0.04
0.05
0.06
Coefficient
Number of lab procedures
Number of inpatient visits in the previous year
Number of medications
Time in the hospital
Number of diagnoses
Number of procedures
Number of emergency visits in the previous year
Number of outpatient visits in the previous year
Discharged to home
Discharged/transferred to another rehab facility
Gender - male
Payer code - MC (Medicaid)
Diseases Of The Circulatory System - diagnosis 1
Diseases Of The Circulatory System - diagnosis 2
Diseases Of The Circulatory System - diagnosis 3
Age [70-80)
Age [60-70)
Race - Caucasian
No change in medication
Age [80-90)
Physician specialty - InternalMedicine
Admission type - Emergency
… External causes Of Injury And Poisoning - diagnosis 3
Insulin - No
Insulin - Steady
Yes
0.2
0.0
Early readmission?
−0.2
Correlation
(directionality)
−0.4
No
−0.6