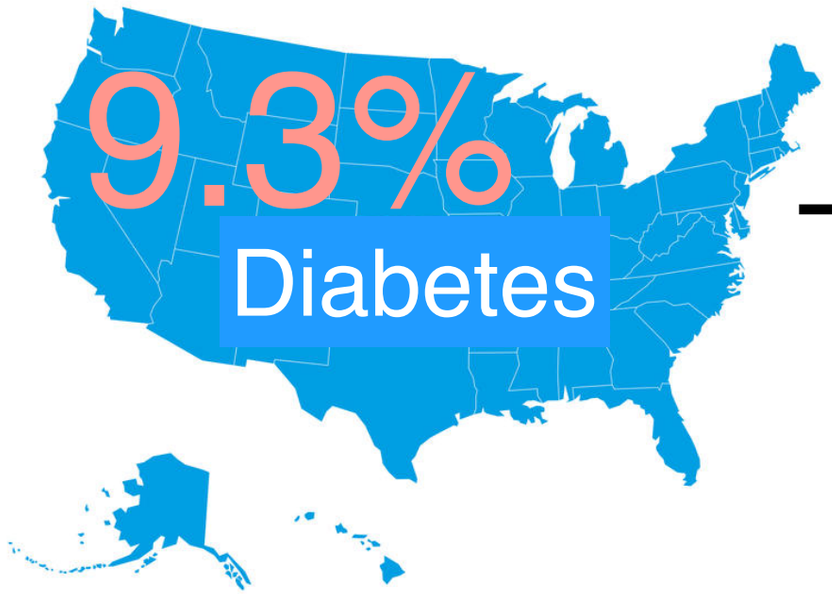
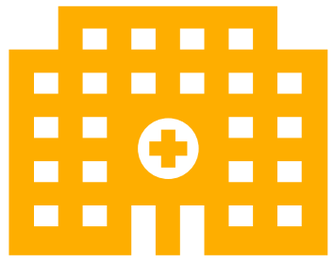
Predicting hospital readmission within 30 days of discharge for diabetic patients

Ignasi Sols

Introduction





~14-22%

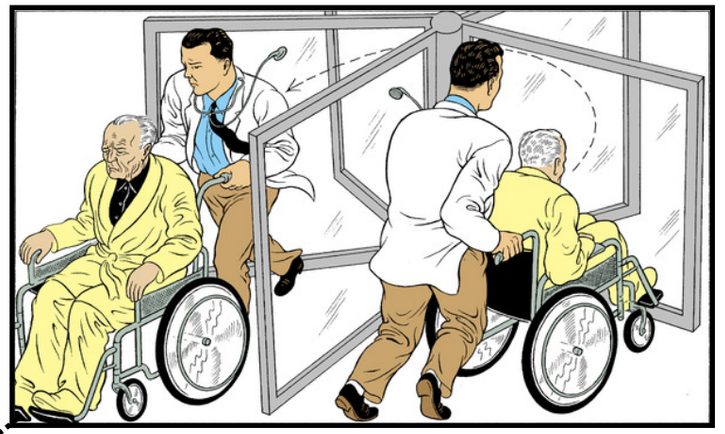
Readmissions < 30 days

Medicare

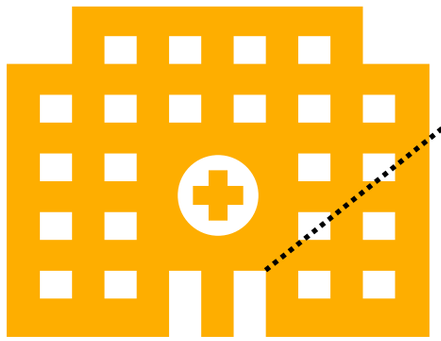
Hospitals with excessive readmission rates
are **penalized** **(not reimbursed)**

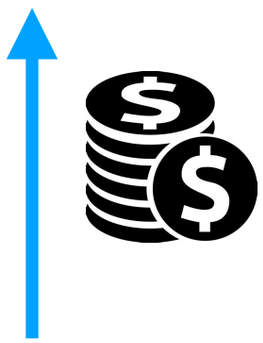
American Diabetes Association
Ostling et al., 2017

Business impact:

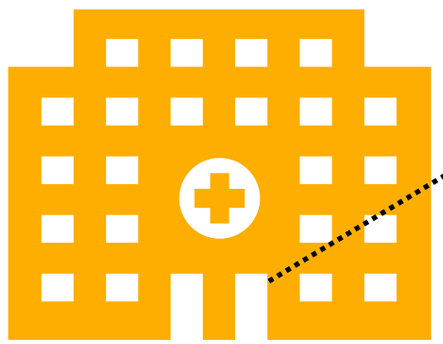


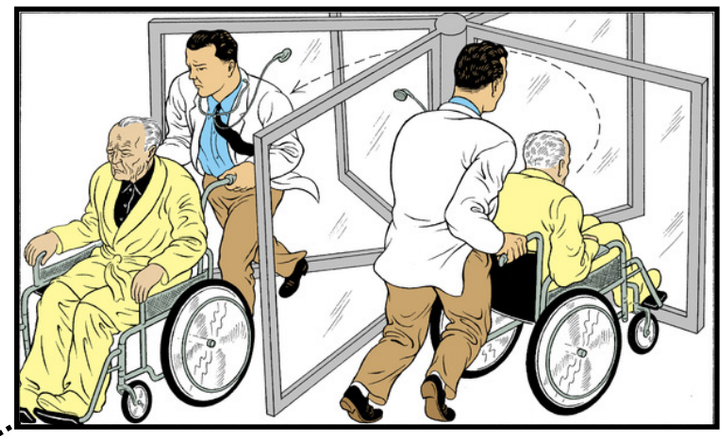
Reduce claims
from readmissions





Impact hypothesis:



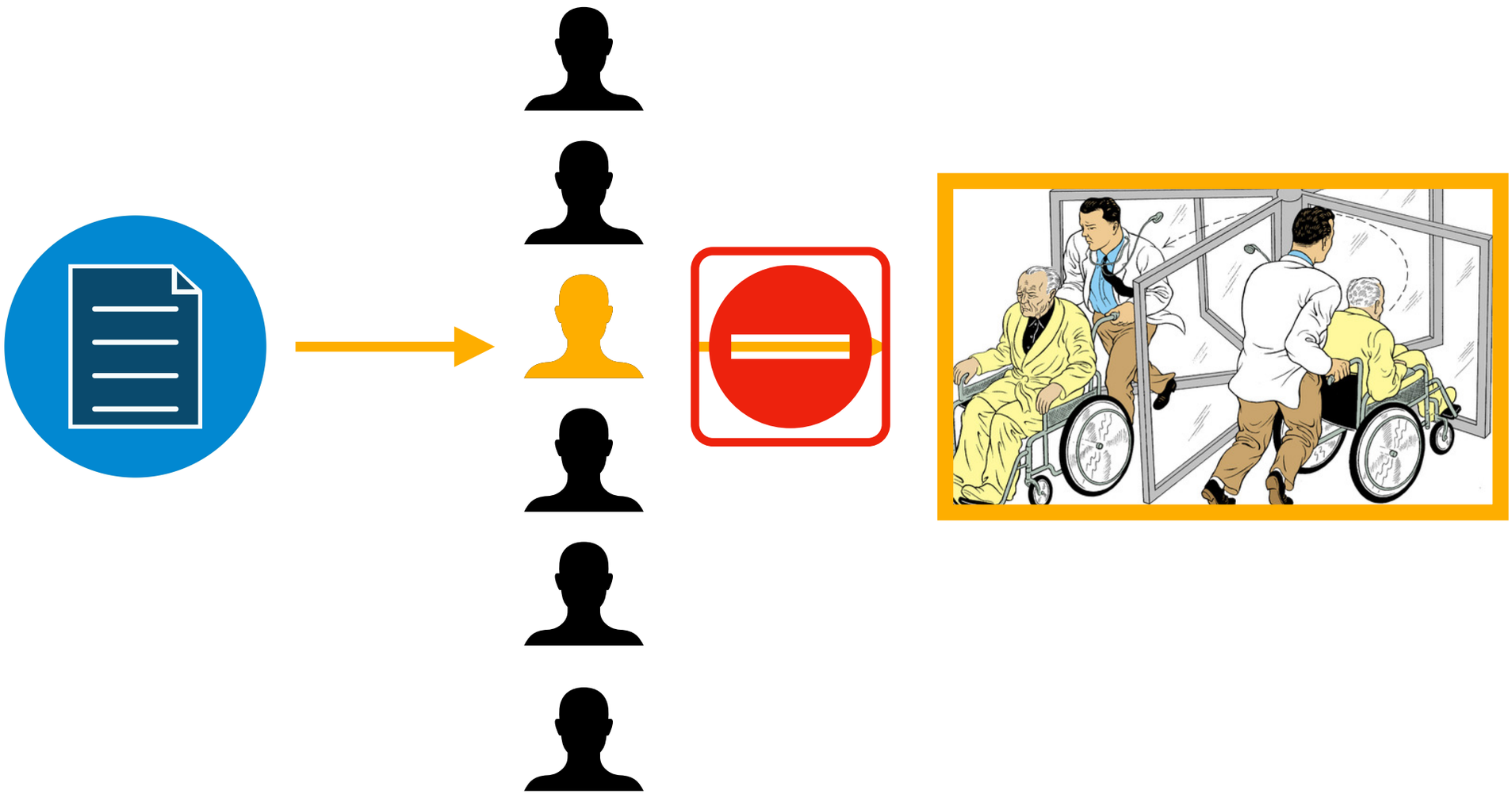


Reduce claims
from readmissions

A **specialized** diabetic and lifestyle **education program** led by **nurse educators** for patients at higher risk of readmission will reduce the readmission rates.

Data Science Path:

Develop a **classification** model that **predicts** which hospitalized patients will later be readmitted



Data / Tools

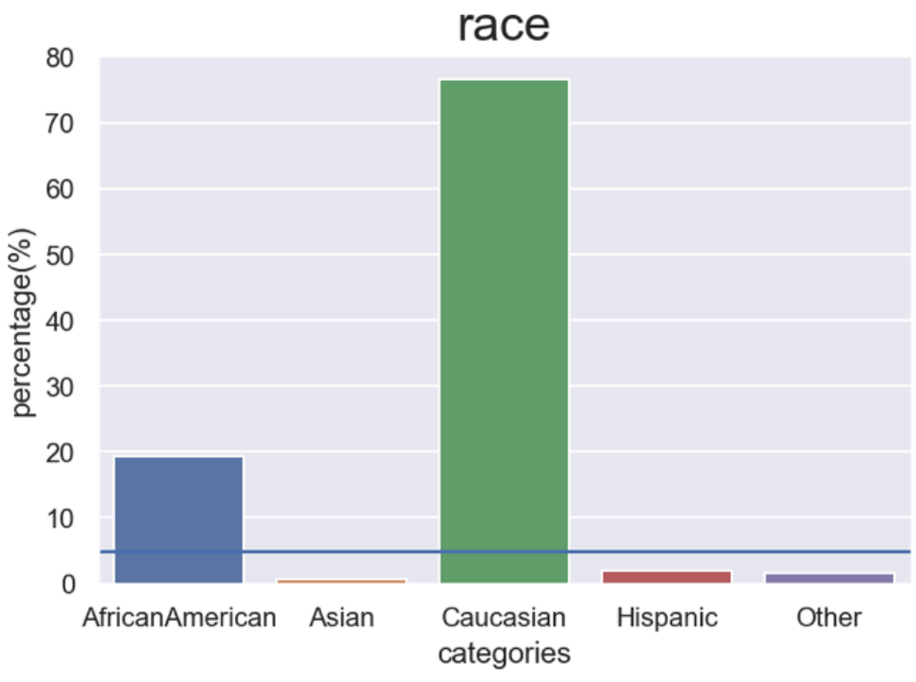
kaggle

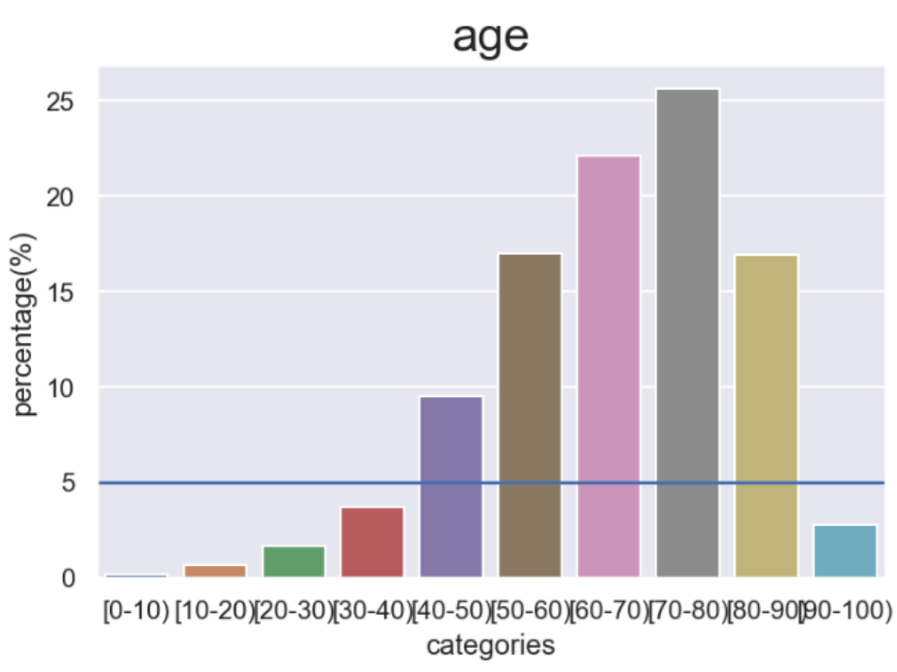
UC Irvine /
Diabetes 130 US
Hospitals for years
1999-2008

~ **100K** rows, one
Readmission record
per row.

Assumptions/Risks

(1) Different **historical** context.
(2) Might not generalize to **other hospitals** not included.
(3) Might not generalize well to non-caucasians and < 50 y.

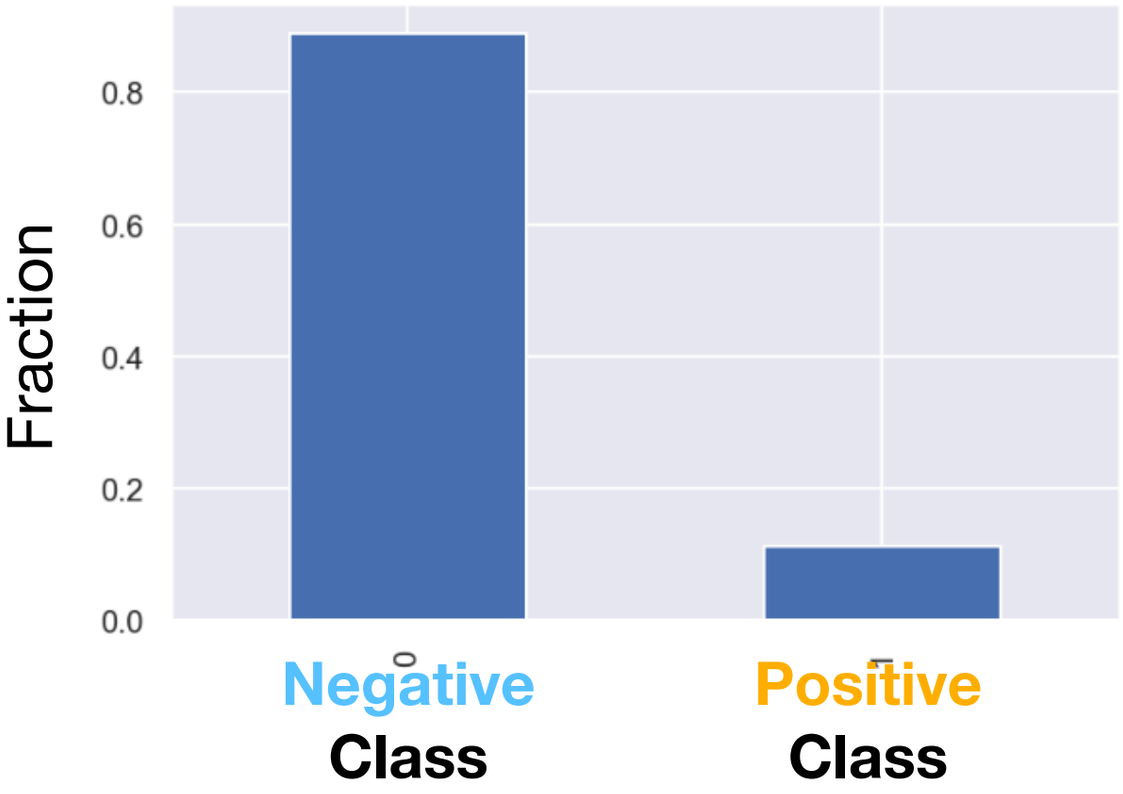




Target: readmitted record

3 classes: I **binarized** them:
Positive class: <30 days
Negative class: >30 days **&** ‘no record of readmission’

Highly **imbalanced** data.



Implemented in the models a method to **handle** it

Modeling

-Classes with < 5% observations were grouped, one-hot encoding.

-**74** features in total.

-**Logistic Regression**, **Random Forests**.

-10 fold cross-validation - GridSearchCV
Some patients appeared on more than one row. **Data bleeding prevented**

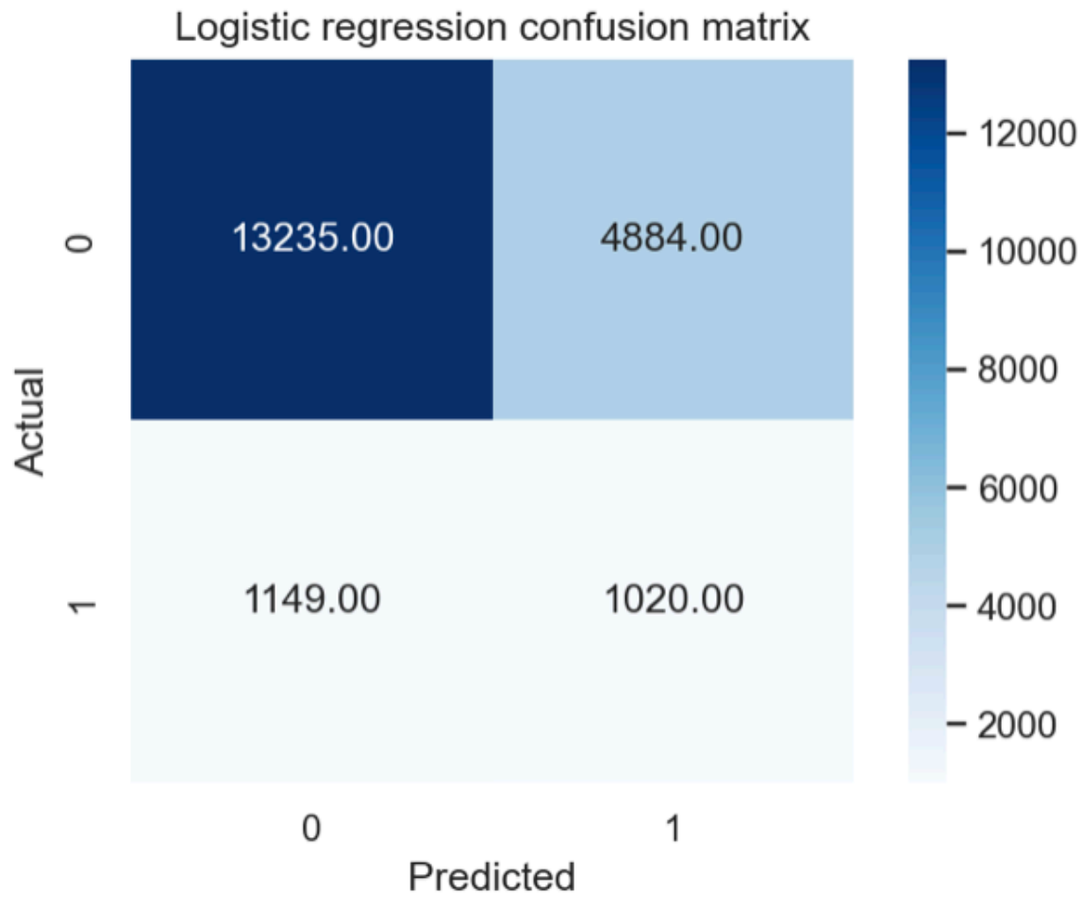
-Scoring: **F1**

Results:

*Baseline (Logistic Regression, no imbalance handling)
F1 score = 0.023

*Logistic Regression
F1 score = 0.248

*Random Forests model
F1 score = 0.253



Confusion matrix translated into metrics relevant to the hospitals:

For every 100 patients that are readmitted per month

Money 

We detect **47** readmissions

If the education program had a 50% efficacy

↓

~**24**+ claims could be reimbursed per month

Huge savings!

Time 

Nurse educators don't have the time to train all the patients. If each patient takes **30 min**:

Time invested to reach the 100 patients that would be readmitted:

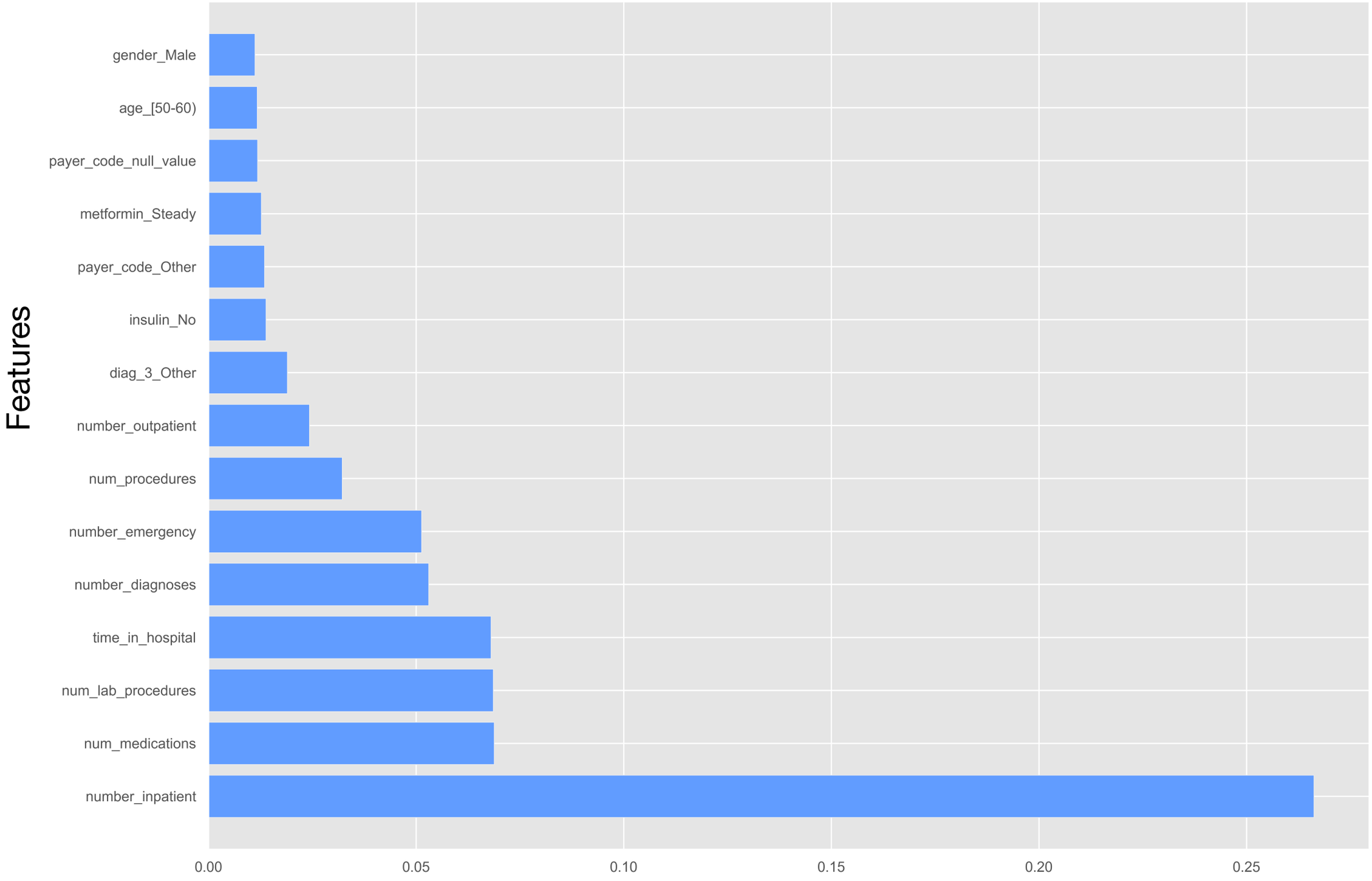
↓

Reach to all diabetic patients: **468h**/month

↓

Our model **~136h**/month

Feature importance (top 10)



Future directions:

-Run XGBoost.
-Improved feature engineering and feature selection.
-Treat it as a multi-class problem?

Thanks!