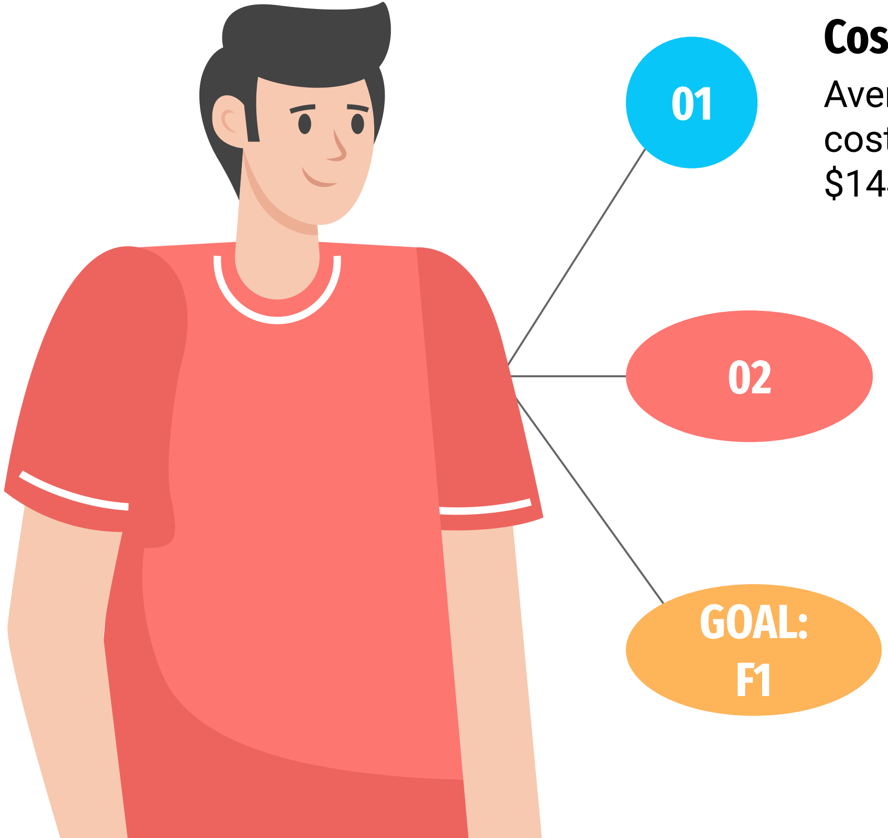
Predicting Rehospitalization

Classification Project

Zoe Liao
1/26/22

Goal: Prevent rehospitalization



Cost savings

Average readmission cost (any diagnosis): $14400

Centers for Medicare & Medicaid Penalty

Excess readmission ratio factors into hospital performance and can reduce payment

Guiding Question

Which factors (patient or encounter-related) during an inpatient (non-ICU) visit contribute to predicting readmission?

Data

VCU

C. Kenneth and Dianne Wright Center
for Clinical and Translational Research

Original CSV

- 130 hospitals and integrated delivery networks
- 1999-2008
- All patients have diabetes
- 100,000+ rows/encounters and 50+ features

Preprocessing

- Only kept 1 encounter per patient
- Target: Rehospitalization
  - **0 if None**
  - **1 if <30 days, OR >30 days**

Baseline Model

- Focused on numerical columns
- Logistic regression model (default):
  - F1: 0.2519
  - Accuracy
    - Training: 62.36%
    - Test: 62.39%

Feature Engineering & Modeling

- Exclude: patients with discharge status as hospice/expired/missing
- Dummy variables:
  - Admission type
  - A1C, Glucose
  - Medications (diabetes)
    - Made a change (start, increase, decrease = 1, steady/not on = 0)
    - Grouped med classes together
  - Diagnoses (1, 2, 3)
    - Top 5 for each, with rest as “Other”
- Tried interaction: heart failure and TZD class medication

<u>60/20/20 split w/ 5-fold CV</u>

- **Random Forest**
  - F1: 0.443
  - Accuracy:
    - Train: 97.87%, Test: 60.15%
- **Logistic Regression**
  - F1: 0.5218
  - Accuracy:
    - Train: 59.93%, Test: 59.78%
- **XGBoost**
  - F1: 0.4346
  - Accuracy:
    - Train: 96.45%, Test: 57.26%

Final Model

Final data

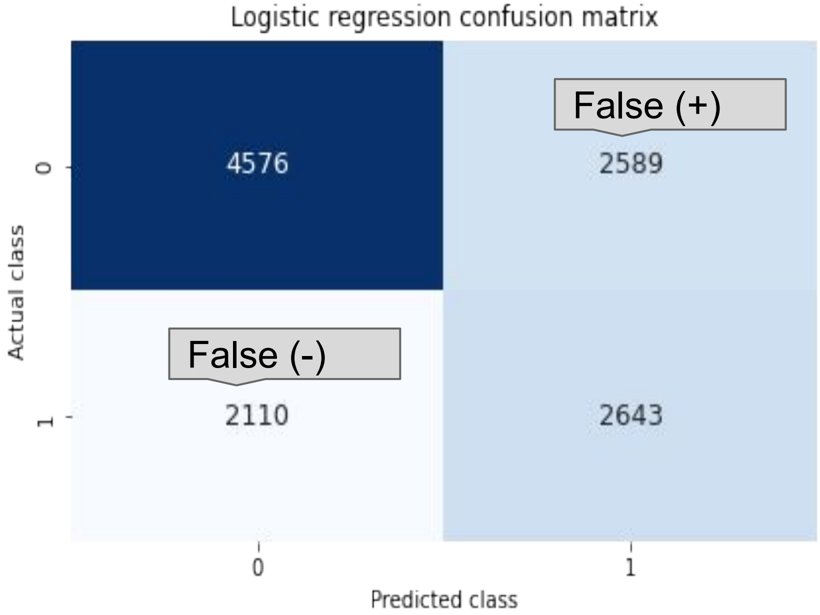
- 50 features
- Target class:
  0: 35615
  1: 23974

Logistic Regression

- C = 0.7
- Class weight: balanced

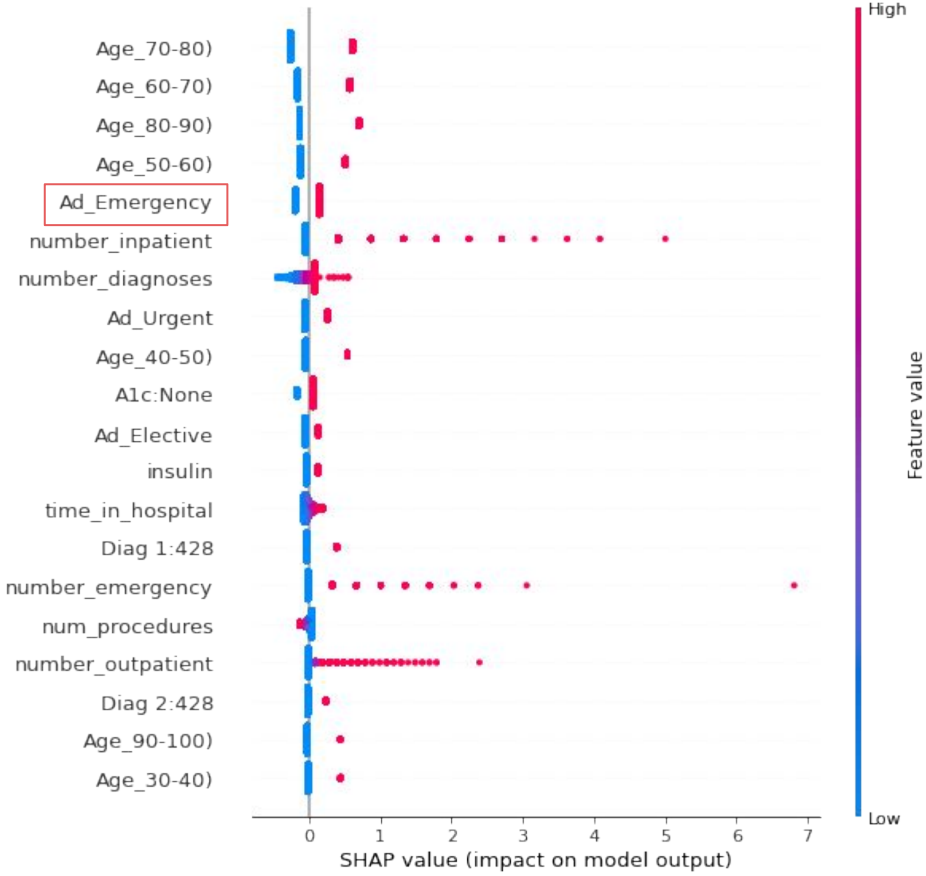
Results

- F1: 0.5294
- Accuracy: 60.57%



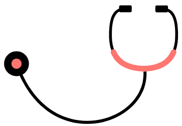
Top Features in Final Model

Feature	Coefficient (Odds)
Age 70-80	1.452
Age 60-70	1.356
Number of inpatient visits in preceding year	1.310
Age 80-90	1.359
Age 50-60	1.271



Future Work

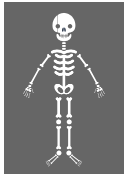
01



Features

Interactions, Unused (race, sex, payer code, weight)

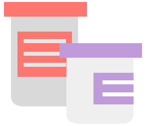
02



Modeling

Hyperparameter tuning, class imbalance strategies

03



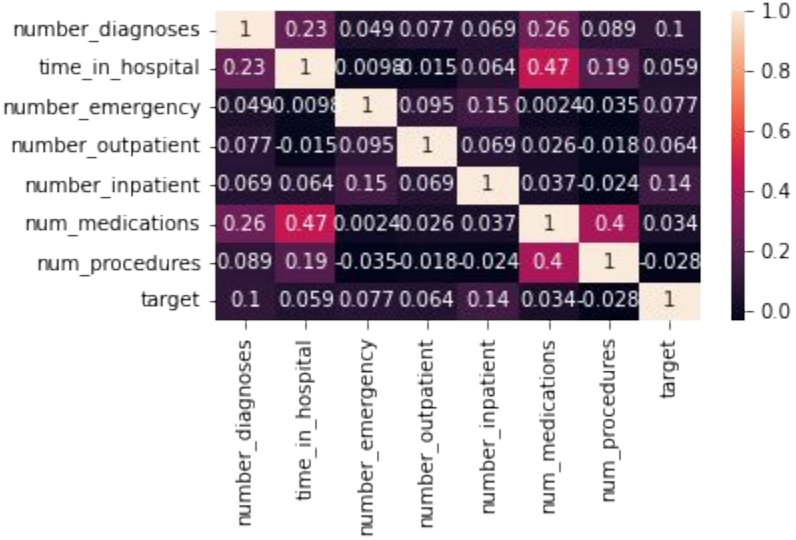
Limitations

Older data, T1DM vs T2DM, heart meds, low glucose

Appendix

Data exploration

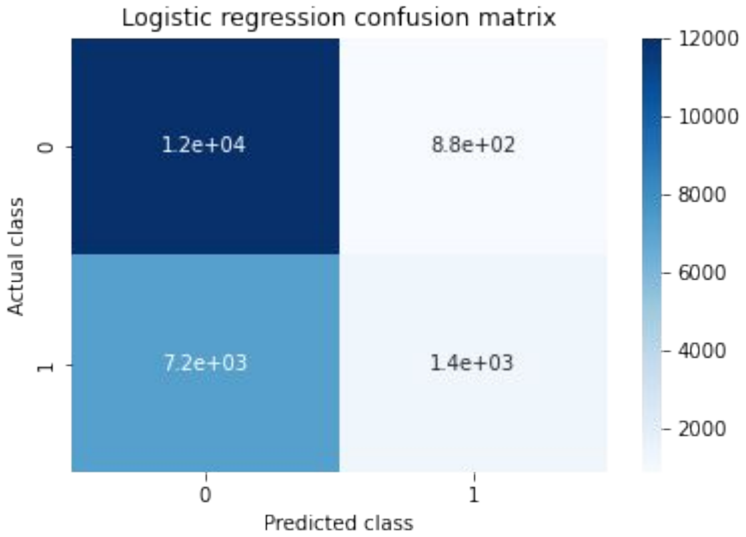
Correlation Heatmap of numerical features (baseline)



Baseline model

- Features (# of):
  - Diagnoses
  - Time in hospital
  - Emergency visits
  - Outpatient visits
  - Inpatient visits
  - Medications
  - Procedures

- Target Distribution:
  - 0: 42985
  - 1: 28533
- Other Stats:
  - Precision: 0.6062
  - Recall: 0.1590



One-hot encoding details

- Discharge status (dropped: discharge to healthcare)
- Admission (dropped: Newborn)
- Age group (dropped: [0-10))
- A1c (droppped: Norm)
- Max Glu Serum (dropped: Norm)
- Diabetes medications (kept same dose or not on med = 0, increase/decrease dose = 1) -> grouped by classes (sum)
    a. meglitinide_class: repaglinide, nateglinide
    b. sulfonylurea_class: chlorpropamide, glimepiride, glipizide, glyburide
    c. tzd_class: pioglitazone, rosiglitazone
    d. agi_class: miglitol, acarbose
- Diagnoses (1,2, and 3) - categorizing just top 5 for each and grouping rest as "Other", then binarize (dropped 'Other' category for each)

Hyperparameters Used in Model Comparison

Random Forest

Class_weight: "balanced"

Max_depth: 30

Logistic Regression

C = 0.7

Class_weight = 'balanced'

Tried: L1 vs. **L2**

XGBoost

Max_depth = 31

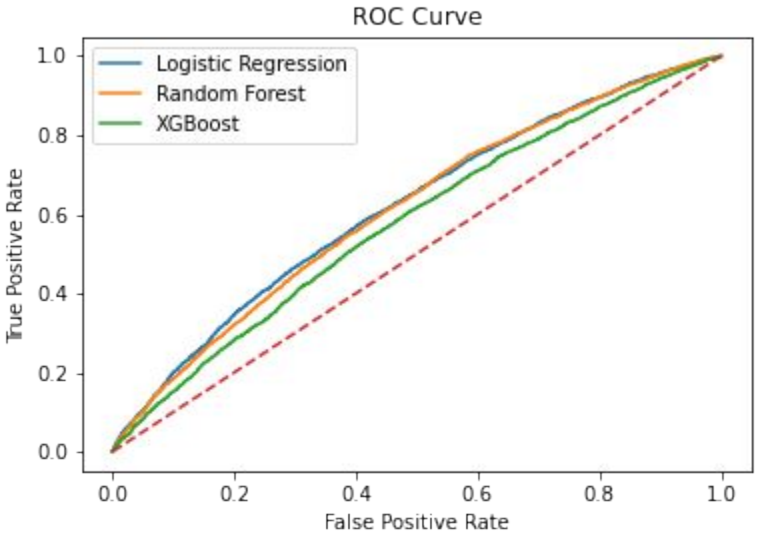
Learning_rate = 0.2

Subsample = 0.4

Min_child_weight = 1

Colsample_bytree = 0.9

Comparing 3 models



Final model: Precision vs. Recall based on threshold