Predicting Early Readmission in Diabetic Patients

Determining whether a diabetic patient will readmit within 30 days

What is diabetes?

- Diabetes mellitus is a collection of conditions that affects the pancreas' ability to produce insulin to regulate blood glucose
- Having too much glucose in a person’s blood can cause damage to blood vessels and nerves and reduce blood flow
- Diabetes increases the risk of cardiovascular disease and organ damage

What is diabetes?

- Diabetes mellitus is a collection of conditions that affects the pancreas' ability to produce insulin to regulate blood glucose
- Having too much glucose in a person’s blood can cause damage to blood vessels and nerves and reduce blood flow
- Diabetes increases the risk of cardiovascular disease and organ damage

What is diabetes?

- Diabetes mellitus is a collection of conditions that affects the pancreas' ability to produce insulin to regulate blood glucose
- Having too much glucose in a person’s blood can cause damage to blood vessels and nerves and reduce blood flow
- Diabetes increases the risk of cardiovascular disease and organ damage

What is diabetes?

- Diabetes mellitus is a collection of conditions that affects the pancreas' ability to produce insulin to regulate blood glucose
- Having too much glucose in a person’s blood can cause damage to blood vessels and nerves and reduce blood flow
- Diabetes increases the risk of cardiovascular disease and organ damage

Diabetes Epidemic in the U.S.

- In 2018, the CDC reported that:
  - 10.5% of Americans had diabetes
  - More than 26% of those aged 65 or older had diabetes

- Diabetes is the 7th leading cause of death in the United States

Diabetes and Increased Risk of Re-hospitalization

- A 30-day unplanned re-hospitalization is likely reflective of the quality of care the patient received during their initial visit.
- Readmission is expensive. In 2016, the average cost of readmission was about $1900 more expensive than that of the initial visit. [3]
- Diabetic patients account for about one-fifth of 30-day unplanned re-hospitalizations, generating about 123 billion dollars in U.S. hospital costs in 2017 alone. [4]

Diabetes and Increased Risk of Re-hospitalization

- A 30-day unplanned re-hospitalization is likely reflective of the quality of care the patient received during their initial visit.
- Readmission is expensive. In 2016, the average cost of readmission was about $1900 more expensive than that of the initial visit. [3]
- Diabetic patients account for about one-fifth of 30-day unplanned re-hospitalizations, generating about 123 billion dollars in U.S. hospital costs in 2017 alone. [4]

Diabetes and Increased Risk of Re-hospitalization

- A 30-day unplanned re-hospitalization is likely reflective of the quality of care the patient received during their initial visit.
- Readmission is expensive. In 2016, the average cost of readmission was about $1900 more expensive than that of the initial visit. [3]
- Diabetic patients account for about one-fifth of 30-day unplanned re-hospitalizations, generating about 123 billion dollars in U.S. hospital costs in 2017 alone. [4]

What is currently being done to identify early readmitters?

- Identifying potential early readmitters can give hospitals an opportunity to intervene, thereby reducing hospital costs
- Risk assessment tools like the HOSPITAL or LACE index score have been used in practice to identify at-risk patients [5]

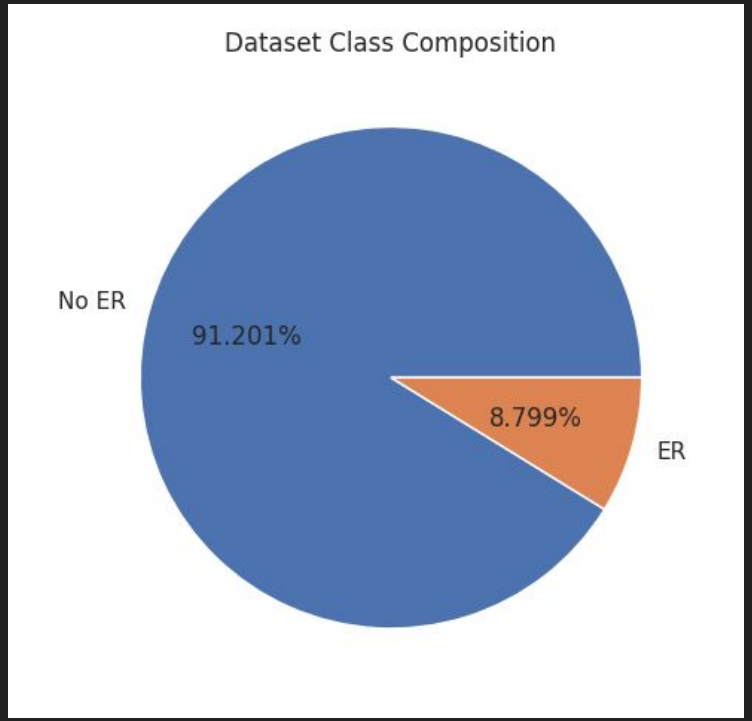
Data

- Data was acquired from the UCI Machine Learning Repository, uploaded on behalf of the Center for Clinical and Translational Research at Virginia Commonwealth University
- 50 features representing over 100k diabetic patient outcomes from 130 U.S. hospitals between 1999 and 2008
- Contains features like demographics, glucose and hemoglobin A1C test results, type of admission and discharge, days stayed in the hospital, medication information

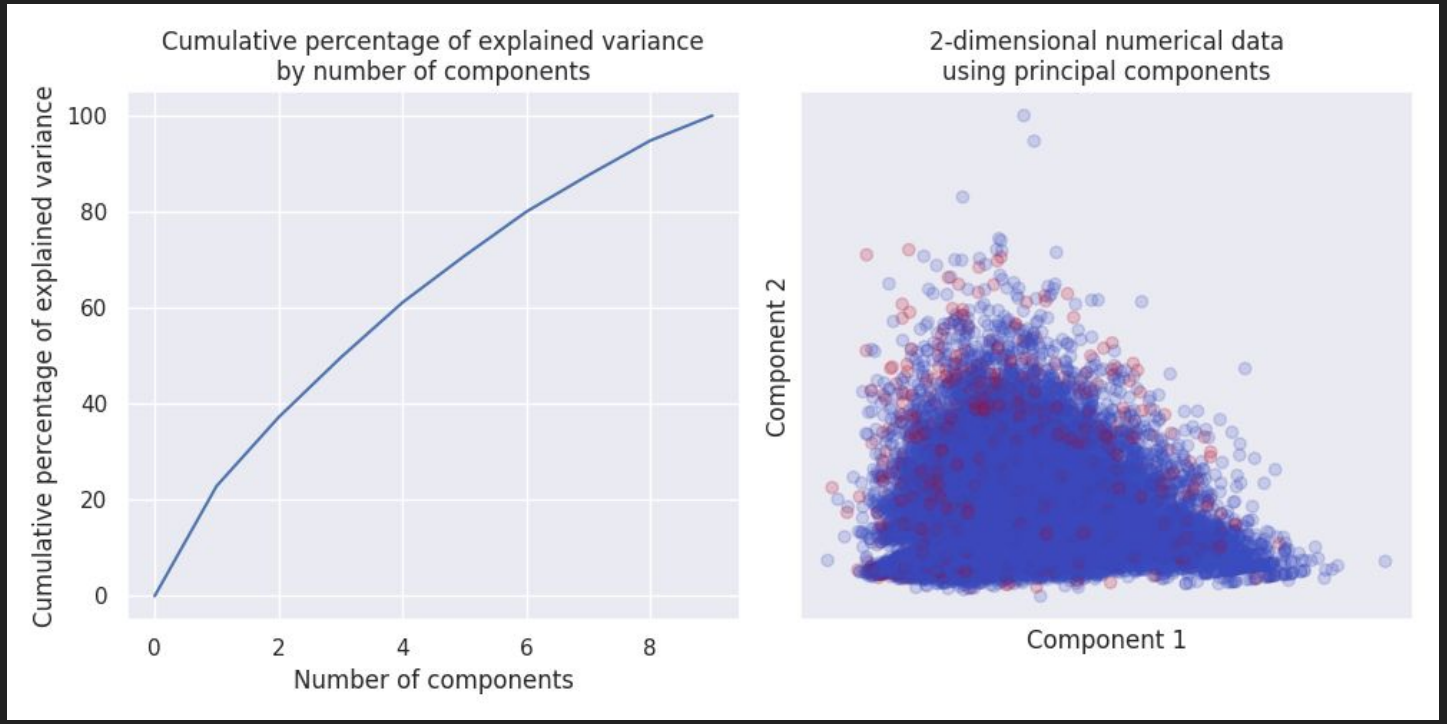
Exploratory Data Analysis

Class Imbalance

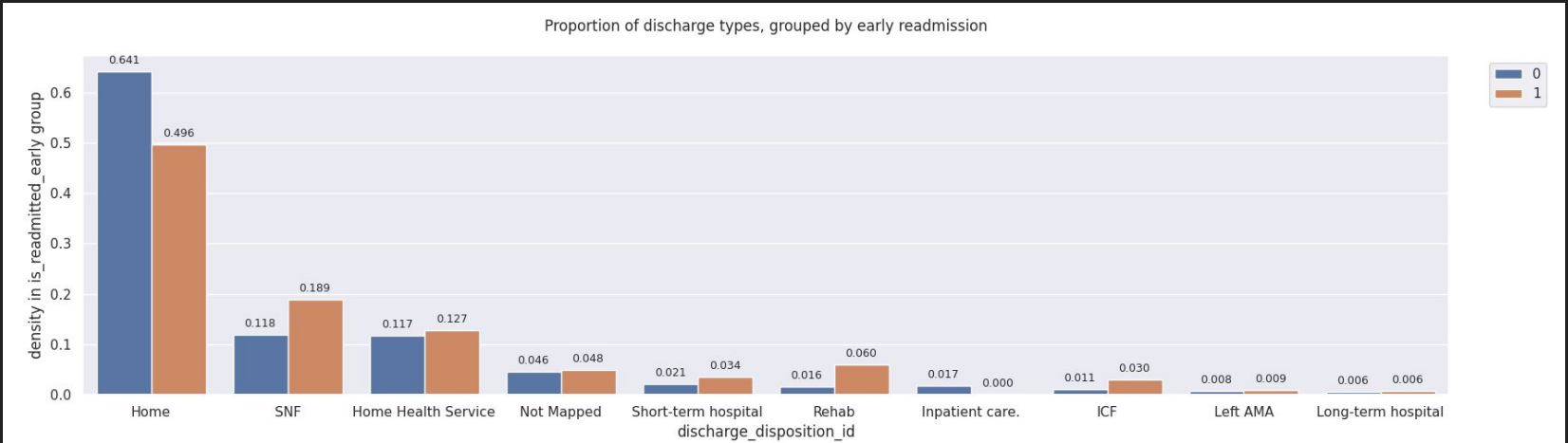
- About a 9:1 ratio of classes between non-early readmission (No ER) and early readmission patients (ER)
- Forced to choose evaluation metrics that are resistant to imbalanced data



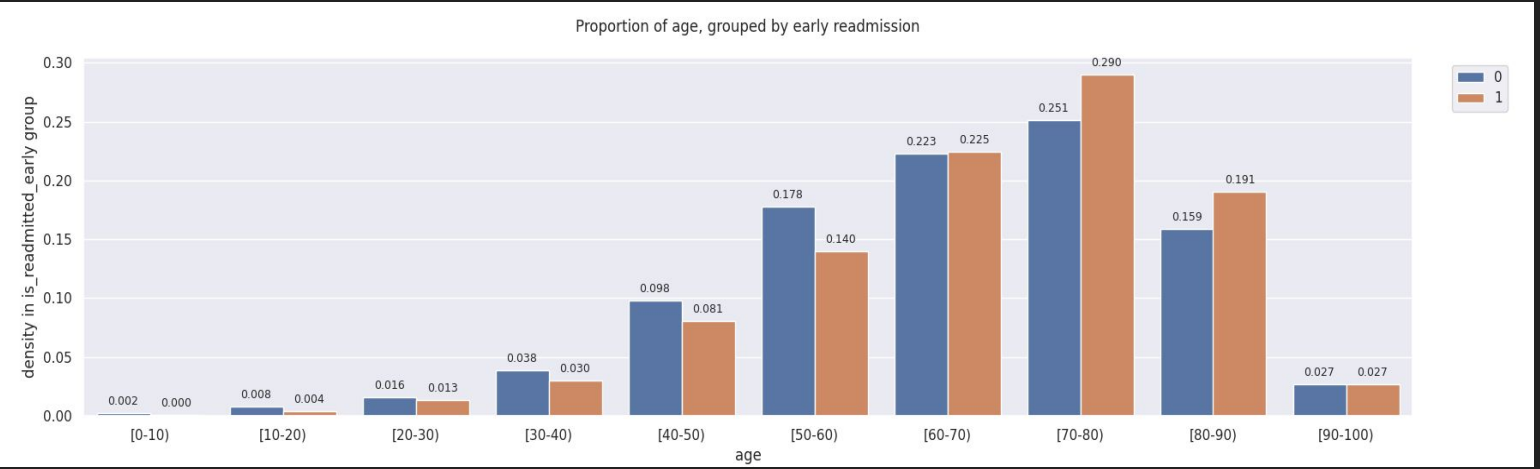
Visualizing the dataset



Differences in discharge distributions across readmission types



Differences in age distributions across readmission types



Numerical features

- Numerical predictors included:
  - days in the hospital
  - number of lab procedures,
  - number of procedures
  - umber of emergency visits
  - number of inpatient encounters,
  - number of outpatient encounters
  - number of medications
  - number of diagnoses

Numerical features

- Numerical features were generally uncorrelated with exception to the number of medications and days in the hospital (r = 0.47) and number of medications and number of procedures (r = 0.4)
- Most numerical features had left or right-skewed distributions with many outliers
- We decided to keep the outliers as they did not appear to be erroneous or unreasonable values

Other interesting findings

- Caucasian patients accounted for about 75% of the dataset. The next largest racial or ethnic group was African Americans at about 18%.
- Patients with hemiplegia or paraplegia, mild liver disease, peripheral vascular disease, cerebrovascular disease, diabetes with chronic complications, and metastatic tumors had a noticeably higher proportion of early readmission than those who do not.
- Patients who were readmitted early had a higher incidence of circulatory issues as a primary diagnosis.

Other interesting findings

- Caucasian patients accounted for about 75% of the dataset. The next largest racial or ethnic group was African Americans at about 18%.
- Patients with hemiplegia or paraplegia, mild liver disease, peripheral vascular disease, cerebrovascular disease, diabetes with chronic complications, and metastatic tumors had a noticeably higher proportion of early readmission than those who do not.
- Patients who were readmitted early had a higher incidence of circulatory issues as a primary diagnosis.

Other interesting findings

- Caucasian patients accounted for about 75% of the dataset. The next largest racial or ethnic group was African Americans at about 18%.
- Patients with hemiplegia or paraplegia, mild liver disease, peripheral vascular disease, cerebrovascular disease, diabetes with chronic complications, and metastatic tumors had a noticeably higher proportion of early readmission than those who do not.
- Patients who were readmitted early had a higher incidence of circulatory issues as a primary diagnosis.

Modeling

Generating useful features

- Prior to modeling, we generated features through scaling, transforming, imputing, and removing redundancies in the data
- Summary of this feature engineering can be viewed here

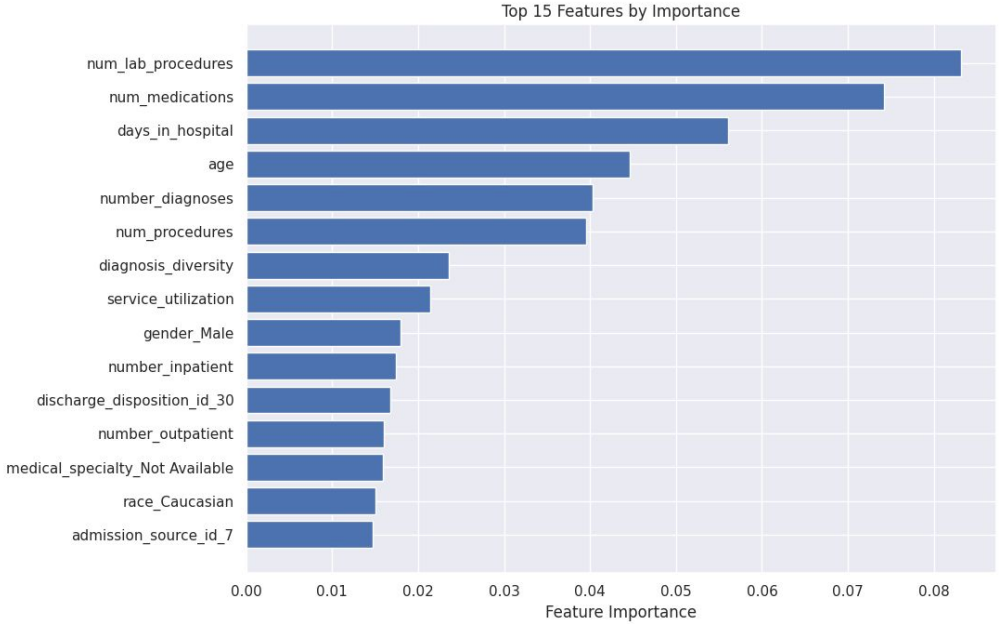
How do we evaluate a model?

- A false negative (a patient who fails to be classified as someone who will likely readmit within 30 days) has much more dire health and financial implications than those of a false positive.
- To minimize false positives, we evaluate models based on with precision, recall, f2-score, and Receiver Operating Characteristic Area Under Curve (ROC AUC)

Can we reduce our number of features?

- Training on a large number of features risks overfitting
- We perform recursive feature elimination (RFE) to iteratively pare our features down to an optimal feature set

Can we reduce our number of features?



How do we handle our imbalanced data?

- Class imbalance risks biasing our models by exposing them to a disproportionate number of observations per class
- We explore three methods of resampling to counteract this:
  - SMOTE (Synthetic Minority Oversampling Technique)
  - Random Oversampling
  - Random Undersampling

How do we handle our imbalanced data?

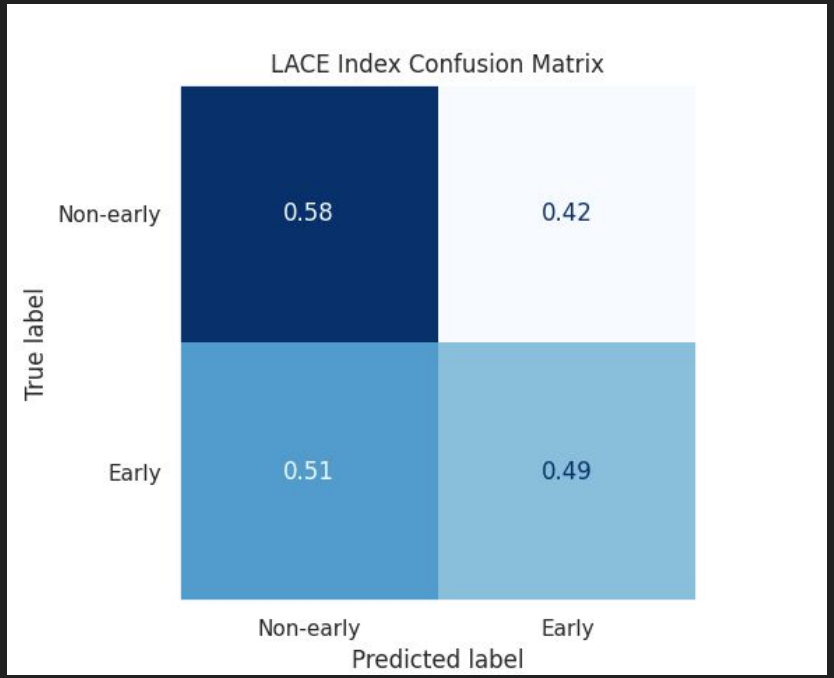
- Training on randomly undersampled data yielded higher scores for all evaluation metrics compared to all other resampling techniques
- Therefore, random undersampling was used to train all candidate models

Experimentation

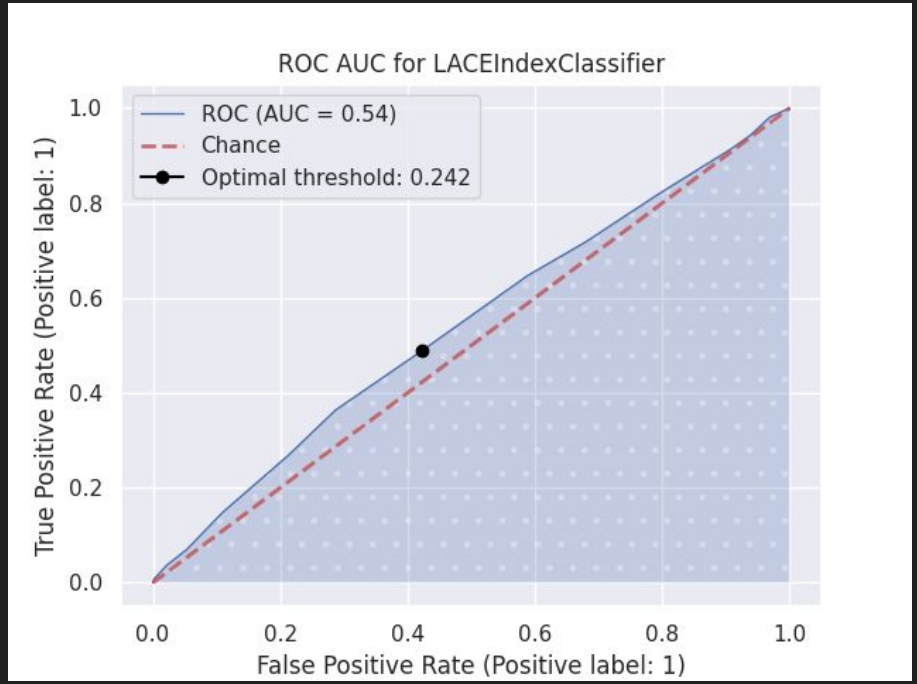
- We conducted 48 studies, varying:
  - The models that performed well during model selection (logistic regression, random forest, LightGBM)
  - The type of dataset
  - The evaluation metrics to maximize

Results

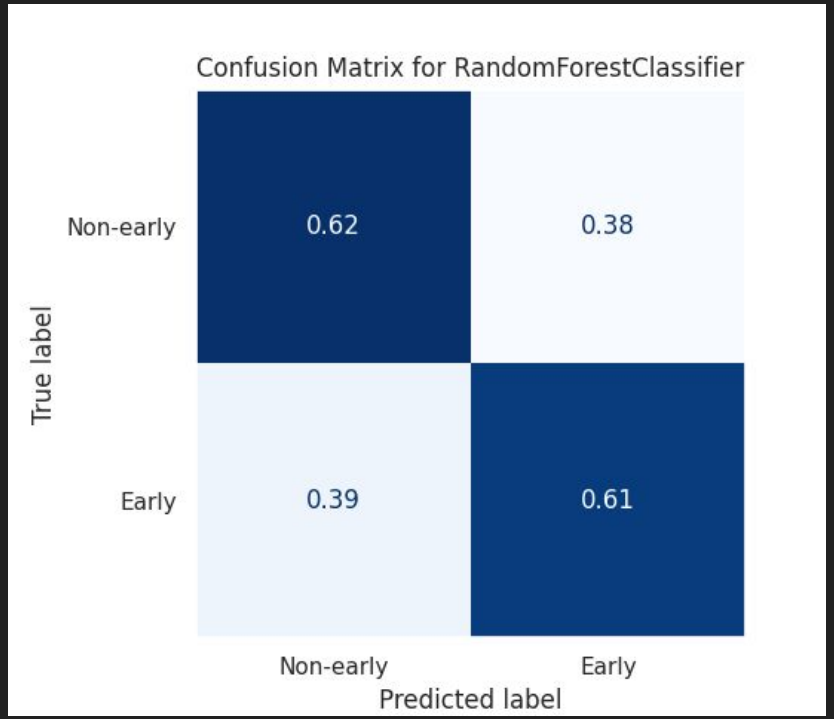
Evaluating the LACE Index



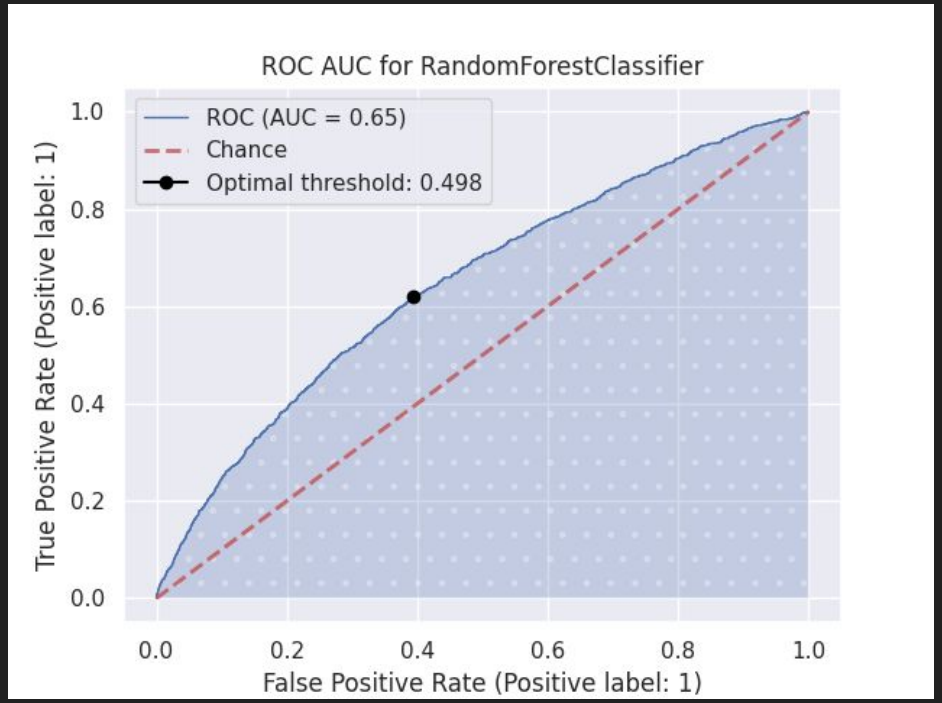
Evaluating the LACE Index



Evaluating the our model



Evaluating the our model

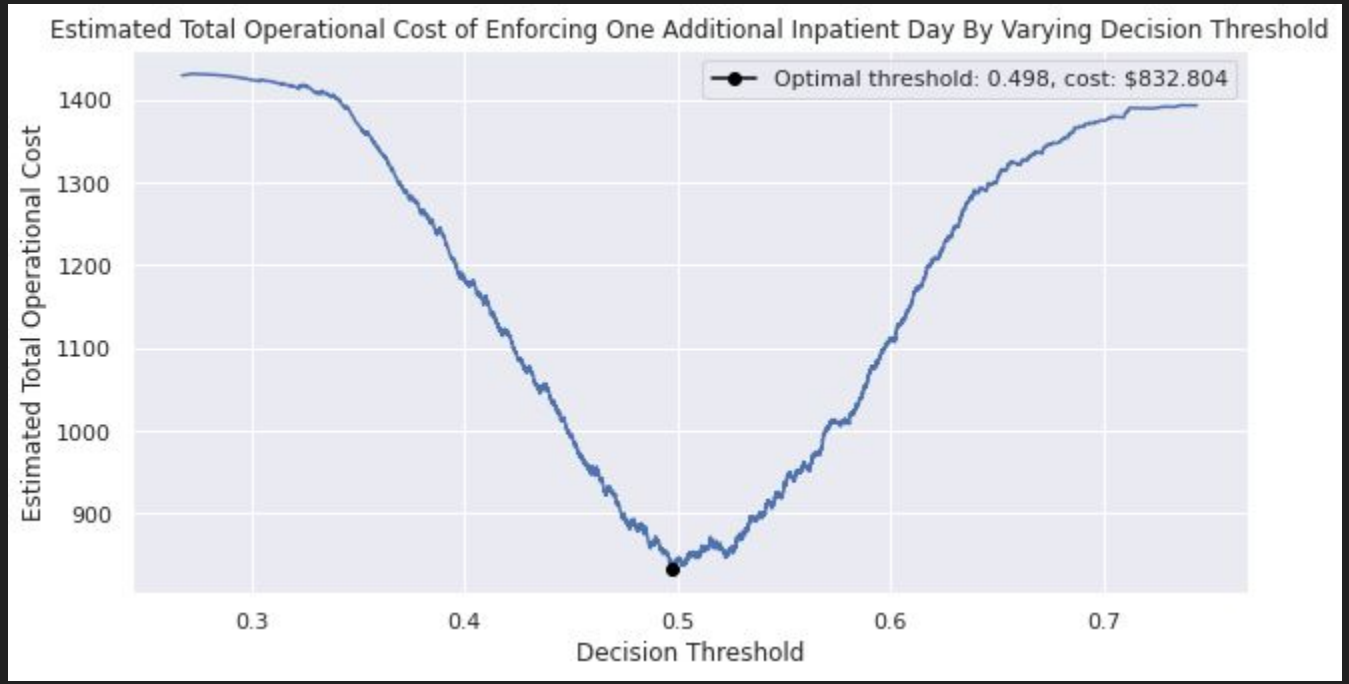


Recommendations

Costs and Intervention Strategies

- Hospitals should research the trade-off between costs of intervention strategies and costs of early re-hospitalization to determine how sensitive they can afford the classifier to be
- If intervention costs are relatively inexpensive, they may consider a higher sensitivity, accepting a higher rate of false positives and their associated intervention costs if it means saving on more expensive re-hospitalization costs.

Costs and Intervention Strategies



Costs and Intervention Strategies

- We recommend that hospitals aggregate racial and socioeconomic data specific to their hospital in order to retrain this model
- As we learned earlier, symptoms and indicators present themselves differently across varying racial and socioeconomic groups. Re-training the model and possibly generating different models for different strata of patients allow the model to resist bias from irrelevant data populations. Setting up an online learning environment, where the model can intermittently learn from new batches of patient encounters, is ideal.

Racial and Socioeconomic Representation

- We recommend that hospitals aggregate racial and socioeconomic data specific to their hospital in order to retrain this model
- Symptoms and indicators present themselves differently across racial and socioeconomic groups. Re-training the model and possibly generating different models for different strata of patients allow the model to resist bias from irrelevant data populations.
- Setting up an online learning environment, where the model can intermittently learn from new batches of patient encounters, is ideal.

Multi-faceted Approach

- Because the model does not identify all potential early readmits with 100% accuracy, it should be used within a multi-faceted context, along with a physician's approval prior to making critical decisions.