Predicting 30-Day Early Readmission in Diabetic Patients

Brandon Fujii

1 Introduction

Diabetes mellitus describes a collection of conditions that affects the pancreas' ability to produce insulin to regulate blood glucose. Diabetes is prevalent in the United States: In 2018, the Center for Disease Control and Prevention reported that about 10.5% of Americans had diabetes, and more than 26% of those aged 65 or older had diabetes [6]. Combined with conditions like hypertension and other cardiovascular diseases, diabetes can be fatal, with it being the seventh leading cause of death in the United States [6].

Diabetic patients are commonly readmitted shortly after recent hospitalizations due to episodes related to hyperglycemia, hypoglycemia, or pre-existing cardiovascular disease. As a result, these 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 [9].

Given the meteoric rise of diabetes-related illness, there is an increasing need for a tool that automatically assesses the risk of a patient readmitting within 30 days of initial admission. We attempt to build a generalized model to solve a binary classification task of predicting whether or not a diabetes patient will readmit into the hospital within 30 days. We developed a model that was able to correctly identify 62% of patients who would not readmit early and 61% of patients who would readmit early. This improves upon existing assessment methods like the LACE index [8], which only identifies 58% and 49% of non-early and early readmission patients respectively. We then provide recommendations on using the model with respect to adjusted operational costs.

2 Data

The raw dataset was retrieved from the UCI Machine Learning repository. It was uploaded on behalf of the Center for Clinical and Translational Research at Virginia Commonwealth University [11].

It includes over 50 features representing patient and hospital outcomes from 130 U.S. hospitals between the years 1999 to 2008. All observations abide by the following criteria:

1. It is an inpatient encounter

2. It is a diabetic encounter

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 attributes such as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, hemoglobin A1C (hbA1c) test result [3], diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization.

3   Data Wrangling

Prior to exploring the data deeply, we perform basic operations like cleaning, imputation, and outlier handling.

The data contained 101,766 observations spanning 50 columns (47 independent variables, 1 encounter index, 1 patient index, 1 dependent variable. 9 features were numerical and the remainder were categorical. The dependent variable includes three values: readmits within 30 days, does not readmit, or readmits after 30 days. We transpose this into a binary classification task by conflating the last two values, only predicting whether or not a patient readmits within 30 days of initial hospital admission (early readmission).

The observations were not statistically independent because patients could have multiple inpatient encounters. Enforcing statistical independence reduced the size of the dataset to 71,518 observations, a 29.7% reduction.

9 features (race, weight, payer code, diagnoses (3), medical specialty, max glucose serum test, and hbA1c result) contained missing or null values:

- Weight was missing more than 96% of its data, so it was dropped completely.

- Only about 2.7% of the race feature were missing. Because other features in these observations were not missing, we opted to encode this unavailability as a Not Available category in the race feature.

- About 49% of medical specialties were not available. Because not needing to see a specialist could be indicative of certain health conditions that may be correlated with early or non-early readmission, we classified missing data as its own category.

- Most patient encounters contain a primary diagnosis. The primary, secondary, and tertiary diagnosis had 0.01%, and 0.4% and 1.7% missing values respectively. All missing values were classified as its own category.

- About 81% and 95% of hemoglobin A1c test results and glucose serum test results were missing respectively. Not having taken or having unavailable results may encode some amount of information, so we decided to classify missing data as its own category.

- The payer code, the ID that corresponds to the method that the patient used to pay for their treatment, had a 43% missing value rate. However, not having a payment method may be indicative of a particular socioeconomic status that may influence readmission. We classified missing data as Not Available.

- There were no missing data for numerical features.

Based on inter-quartile range, we found that the feature with the largest amount of outliers was the number of inpatient encounters. We see later that many of these numerical features are skewed left, where it is common for observations to contain 0 or 1 values. As a result, realistic values are viewed as outlying. With this in mind, we decide to keep these outliers and try to reduce its skew through transformation.

4 Exploratory Data Analysis

We visualized our interim data in order to identify latent patterns that can potentially inform our model.

First, we recognized the imbalanced nature of our dataset, where positive observations (patients who were readmitted early) only accounted for about 9% of our data (Figure 1).

By scanning the data, it was clear that this data was not easily linearly separable. As an illustrative sample, we projected the dimensionality of numerical features to a lower dimensional space via principal component analysis, plotting its cumulative percentage of explained variance by number of components and the first two principal components in a scatter plot (Figure 2).

4.1 Categorical Features

- There were discharge categories related to hospice or death, which means that these patients will definitely not be readmitted to the hospital. Observations with the following features were to be removed in the feature engineering phase:

  - Expired
  - Hospice / home
  - Hospice / medical facility
  - Expired at home. Medicaid only, hospice.
  - Expired in a medical facility. Medicaid only, hospice.
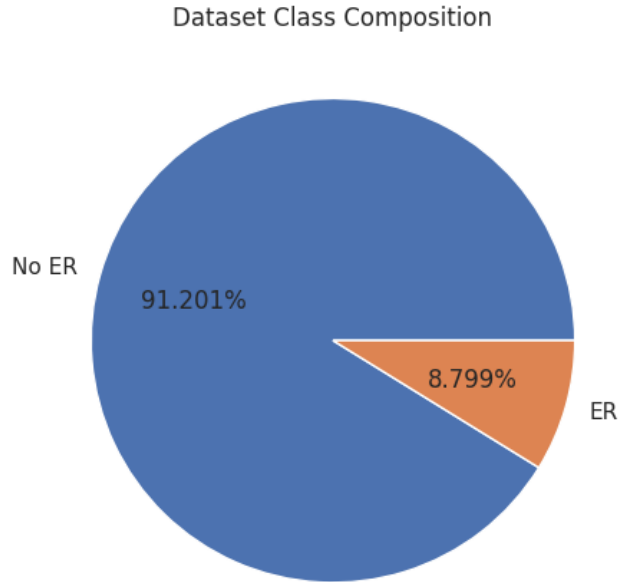
Dataset Class Composition

Figure 1: There is about a 91:9 % class imbalance in the dataset between non-early readmission patients (No ER) and early readmission patients (ER). This indicated that total accuracy was not a good measure of success, as an naive learner that predicts no early readmission 100% of the time will yield over 91% accuracy. Using a confusion matrix, precision, recall, and other metrics in combination with accuracy may help establish a better idea on how our learner is performing. This eventually led us to re-sample the dataset to restore balance in our training set.

- – Expired, place unknown. Medicaid only, hospice.

- Patients who were not readmitted early tended to be discharged to their home more often that the early readmission group (0.64 vs 0.50) (Figure 3).

- Patients who had been readmitted early tended to be older, and more likely to be in the 70-90 age range than those who were not readmitted early (Figure 4).

- 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,
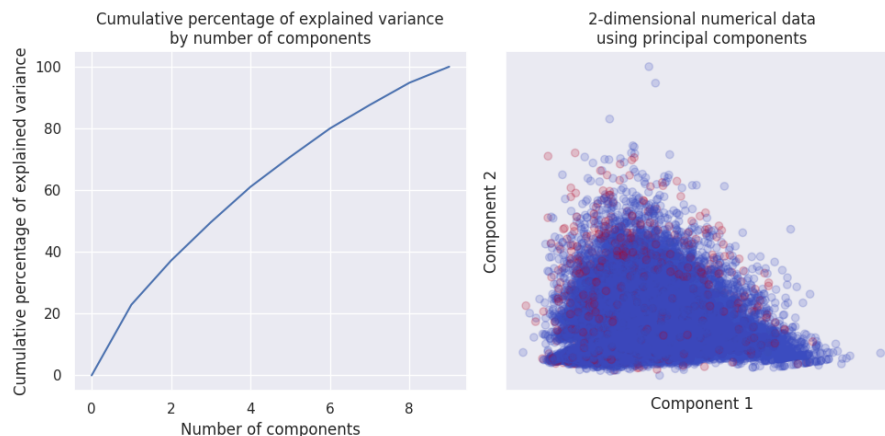
Figure 2: **2a.** According to the left-hand figure, there is not one numerical feature that is responsible for most of the variance in the dataset. **2b.** Although two components only explains about 40% of the data variance, we can see on the right-hand figure that the early readmission patients (red) and non-early readmission patients (blue) are similar positionally within a 2-dimensional Euclidean space.

and metastatic tumors have a noticeably higher proportion of early readmission than those who do not.

- Positive observations had a higher proportion of unavailable healthcare payer codes. This could indicate that (1) these patients tended to not have measures to pay for their treatment, that (2), circumstantially, these codes were neglected to be reported for a higher proportion of this group, or (3) some other reason unknown to us. Positive observations had a slightly higher proportion of Medicare payer codes, which may make sense in that Medicare, a national US healthcare program, is often given toward individuals 65 or older, or those with a pre-existing disability. Older patients may be more susceptible to severe illness, which may warrant earlier re-hospitalization. Overall, about 44% of all payer codes were unavailable so this may be an unreliable predictor of early readmission.

- Patients who were readmitted early had a higher incidence of circulatory issues as a primary diagnosis. Similarly, non-early readmission patients had a higher incidence of respiratory illness as a primary diagnosis (Figure 5). Circulatory diagnoses dominate all three diagnosis categories, but are not clearly favored by one group.

- Many of the medication features were fairly homogeneous. Examide, citoglipton and glimepiride-pioglitazone were completely homogeneous.

Exceptions to this were metformin and insulin, which both had the highest amount of variability in their prescription change status.
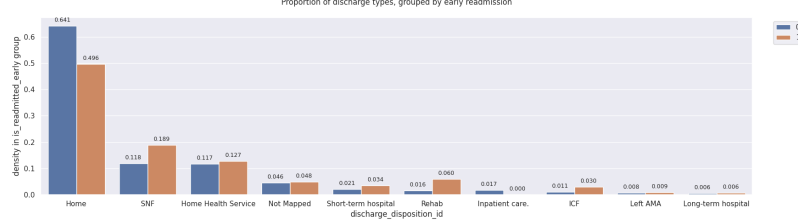

Figure 3: A grouped bar plot visualizing the proportion of discharge types within each readmission status group: early readmission (1) and non-early readmission (0). Non-early readmission patients tend to be discharged to their homes more often than early readmission patients, who tend to be discharged to a skilled nursing facility.
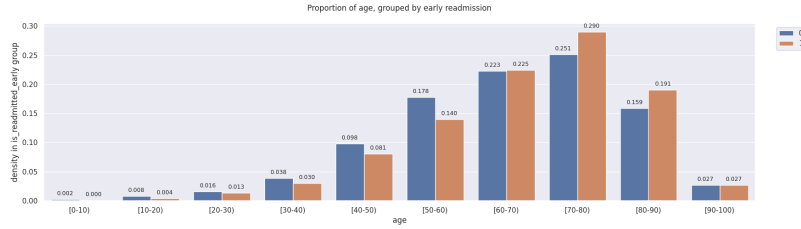

Figure 4: A grouped bar plot visualizing the proportion of age brackets within each readmission status group: early readmission (1) and non-early readmission (0). Both distributions are skewed a bit left, containing patients mainly in the 50-90 age range. We observe a higher incidence of 70-90 year old patients in the early readmission group than the non-early readmission group.

4.2 Numerical Features

The numerical features include days in the hospital, number of lab procedures, number of procedures, number of emergency visits, number of inpatient encounters, number of outpatient encounters, number of medications, number of diagnoses. These features were generally uncorrelated with exception to number of medications and days in the hospital ($r = 0.47$) and number of medications and number of procedures ($r = 0.4$). We can speculate that a confounding factor could be the severity of or susceptibility to illness, which could warrant a longer stay or higher number of medications or testing procedures.

We log-transformed and scaled our numerical features to have a center of 0 and a standard deviation of 1. Most of the resulting distributions were left or
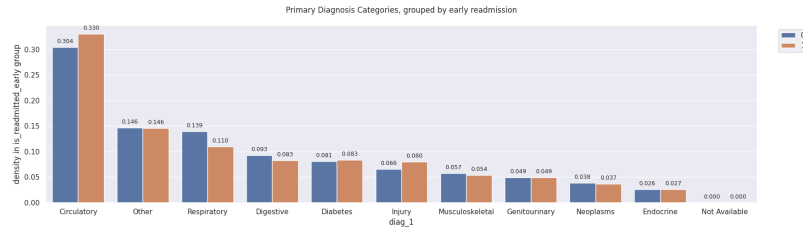
Figure 5: A grouped bar plot visualizing the proportion of primary diagnoses within each readmission status group: early readmission (1) and non-early readmission (0). Patients who were readmitted early had a higher incidence of circulatory issues as a primary diagnosis. Conversely, non-early readmission patients had a higher incidence of respiratory and digestive illness as a primary diagnosis.

right-skewed. Features like the number of inpatient, outpatient, and emergency encounters commonly had a zero value, so common that a value of 1 was considered an outlier. We examined outliers and realized that many were many were, in fact, valid or reasonable, and may encode some importance in distinguishing between the two target groups.

5 Features

We generated features by performing actions, including but not limited to:

- Log-transforming and standard scaling numerical features

- Removing homogeneous numerical features that have a variance lower than 0.1

- Encoding null or unavailable values as their own category for a given set of features, in situations when omission of values carry some useful information

- Imputing missing values in categorical features

- Combining features into simpler, compound features

- Converting ICD-9 diagnosis codes into a more general taxonomy of classifications

- Collapsing feature values if they only occurred in less than 5% of the dataset

- Removing redundant features whose absolute Pearson correlation coefficient with another feature was greater than 0.9
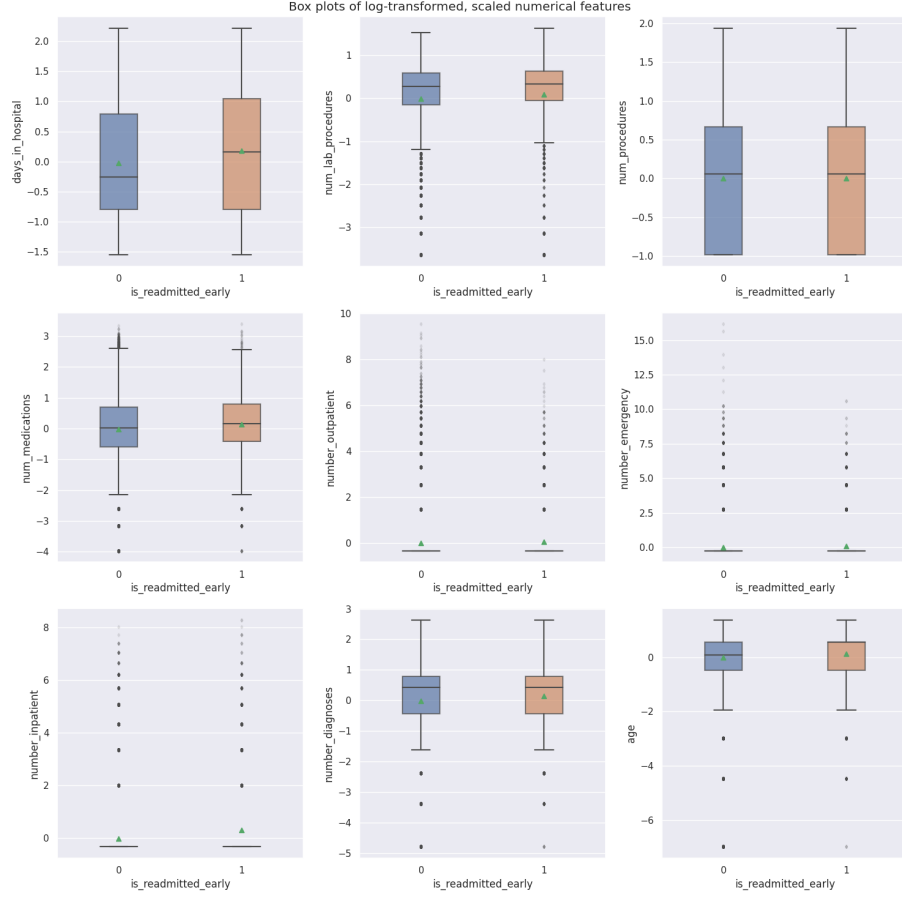
7

Figure 6: Grouped box plots, comparing distributions of various log-transformed, scaled numerical features in early readmission and non-early readmission groups. Mostly all distributions are skewed left (number of diagnoses, age, number of lab procedures) or right (number of inpatient and outpatient encounters, number of emergency visits, number of procedures).

These actions were automatically performed using a pre-processing pipeline, which is described in more detail in a project summary in our code repository [12].

We split our dataset into a training set and testing set, which contained 80% and 20% of our original dataset size respectively.

6 Modeling

6.1 Model Selection and Sampling Methods

We defined a suite of metrics that would be most useful in evaluating models with respect to our data. We decided to evaluate models with precision, recall, f2-score, and Receiver Operating Characteristic Area Under Curve (ROC AUC) due to the majority-minority class imbalance in our dataset.

A false negative (a patient who fails to be classified as someone who will likely re-admit within 30 days) has much more dire health and financial implications than those of a false positive. With that in mind, we prioritize the recall score and weighted metrics like the f2-score, as we would like to minimize false negatives.

Class imbalance risks biasing our models by exposing them to a disproportionate number of observations per class. We explored sampling methods to counteract this. We experimented with SMOTE (Synthetic Minority Oversampling Technique) [1], random oversampling the minority class, and random undersampling the majority class. We then performed 3-fold cross-validation on each of these generated datasets to determine which method yielded the best overall performance and to also get a general idea of which types of models are most effective in learning from our data. Training on randomly undersampled data yielded higher scores for all evaluation metrics compared to training on SMOTE, randomly oversampled, or initial data. Thus, random undersampling was used to train all candidate models.

The types of models we tested were logistic regression, naive bayes, decision trees, random forests, gradient boosted trees, Light GBM, and XGBoost. We also compared them to benchmark classifiers like dummy and LACE Index classifiers. Logistic regression and tree-based ensemble methods performed well on average. Logistic regression, random forest, and LGBM selected for hyperparameter tuning due to their flexibility and reasonable training speed.

6.2 Hyperparameter Tuning

We sought to improve our performance by tuning our model's hyperparameters. We focused on logistic regression, random forest, and LGBM classifiers. We first established a benchmark using a LACE index classifier, which only correctly identified 49% of early readmission patients and 58% of non-early readmission patients (Figure 7a, 7b).

We generated multiple datasets in an effort to reduce the dataset's original 80 features and reduce overfitting. We performed recursive feature elimination to select the optimal subset of features that yielded the highest average ROC AUC score over 5 cross-validated folds. We used random forest as our base classifier for this incremental feature elimination. The best performance required 79 features. The most relevant features were the log-transformed numerical features in the number of lab procedures, days in the hospital, number of medications, the number of procedures, number of diagnoses, and the diversity of the three
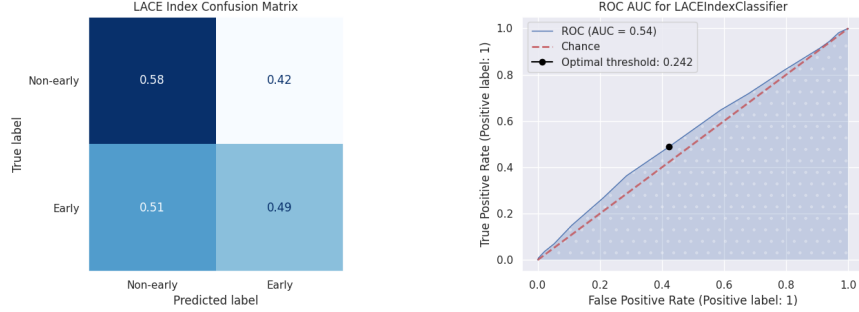
(a) The LACE index only yields a performance slightly better than chance in identifying non-early readmission patients. It performs worse than random at identifying early readmission patients, only successfully identifying 49% of these observations.

(b) The LACE index's mediocre performance can be observed in the ROC curve, which is an aggregate measure of performance across all classification thresholds, only yielding an AUC of 0.54.

Figure 7: LACE index classifier's resulting confusion matrix and ROC curve

given diagnoses. Age and whether the patient was discharged to their home were also important. A notable observation is that these feature importances are all extremely low in weight, meaning there did not exist a ubiquitous feature that proved to be a key indicator for all trees that analyzed various subsets of data.
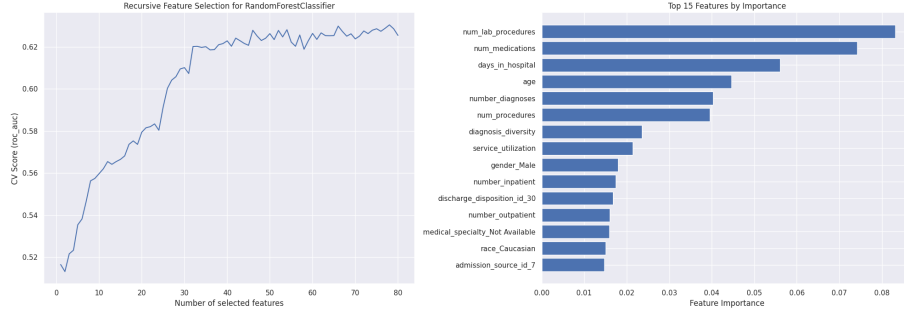

Figure 8: We performed recursive feature elimination (RFE) to select the optimal subset of features that yielded the highest average ROC AUC score over 5 cross-validated folds. The best performing classifier used 79 of 80 features. We observe that by 30-35 features, we already achieve an average cross-validated ROC AUC score of 0.62. The most relevant features were the log-transformed numerical features in the number of lab procedures, days in the hospital, and number of medications.

We conducted 48 studies, $M \times D \times O$, where:
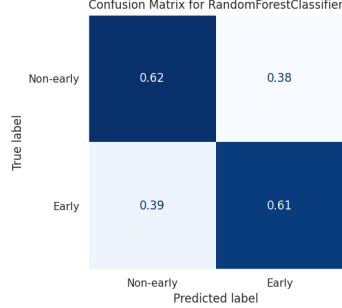
(a) The normalized confusion matrix for our final random forest classifier. The model is able to identify 61% of early readmission patients and 62% of non-early readmission patients.
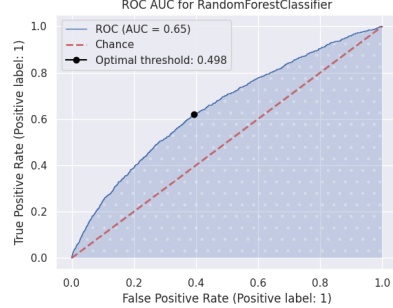
(b) The final model yields a ROC AUC of 0.65, which can be interpreted as the probability that our model labels a future early readmission as positive. We observe about a 20% difference in the ROC AUC, compared to the LACE Index.

Figure 9: Final model performance

- M is the model type (logistic regression, random forest, LGBM). These models yielded the consistent cross-validation results in our model selection phase.

- D is the dataset type (the original undersampled dataset, the undersampled dataset with only feature selecteds via RFE, the undersampled dataset with only features that had an importance above 0.05, top 35 most relevant features according to RFE)

- O is our maximization objective (recall, $f\beta$-score ($\beta = \{1, 2\}$), ROC AUC)

6.3 Final Performance

Our best classifier was on the RFE-reduced dataset, optimizing for the f2-score. It was able to identify 62% of non-early readmission patients and 61% of early readmission patients against a standard 0.5 classification threshold. Considering all classification thresholds, it yielded an ROC AUC of 0.65, with an optimal threshold at 0.5. We observe about a 20% difference in the ROC AUC, compared to the LACE Index.

7 Future Work

7.1 Patient Stratification

We attempted to create a generalized model for predicting early readmission. However, it did not yield decisive, satisfactory results. It may be more helpful
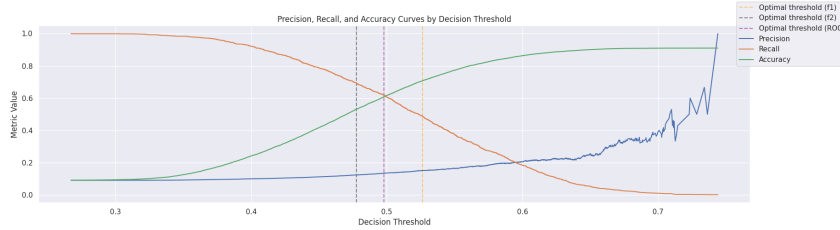
Figure 10: We visualize precision, recall, and accuracy as a function of varying decision threshold. Accuracy tends to increase the more we increase the decision threshold. This is because of the 9:1 class imbalance that is still present in our test set, as we do not perform any resampling on unseen data. For decision thresholds greater than 0.7, the recall (or the accuracy with which we classify early readmission patients) falls to about zero, meaning we generally aren't extremely confident when predicting the probability that an early readmission patient will readmit early.

to stratify across attributes known to increase the risk of readmission. For example, by controlling for age, we may discover key indicators in 70-90 year old diabetic patients that may otherwise be diluted by patient observations in younger age brackets that don't demonstrate these characteristics.

Using this approach, we would begin to notice the lack of representation in our dataset. For instance, non-Hispanic white patients accounted for about 75% of the data. This fact is significant, given the biological (glucose metabolism and insulin resistance, types of symptoms, responses to medications, comorbidities) and social (access to healthcare, access to proper nutrition, social cohesion) disparities across racial and ethnic groups [10].

7.2   Costs and Implementation Strategies

Our threshold sensitivity (i.e. whether we'd like to be more sensitive in detecting patients at risk for early readmission) is contigent on (1) the costliness of the practices that accommodate at-risk patients and (2) how much the hospital saves by avoiding patient readmission (also considering that the federal government penalizes hospitals over excess patient readmissions for certain conditions and procedures via the Hospital Readmissions Reduction Program [5]

As we increase the decision threshold, it becomes more difficult to classify a patient with an early readmission label. As a result, we are more likely to classify more future early readmission patients as not-at-risk, thereby yielding more false negatives and high costs if they eventually readmit. Decreasing the decision threshold and making us more likely to classify a patient in the early readmission group increases our false positive rate, which may generate additional costs as well, depending on what type of intervention strategy we decide to employ.

For example, if we ask at-risk patients to stay another day in the hospital, we

accept one more day of inpatient care expenses to hopefully reduce the likelihood that the patient will readmit, potentially saving long-term costs. We estimate the average inpatient care expenses per day as \$2827, adjusting for inflation, and the average readmission cost as \$15,556 based on the HCUP Statistical Brief [2] hosted by the Agency of Healthcare Research and Quality. Based on this strategy, we assess the costs of each of the following cases:

- Patients who will readmit and are identified by our model (True Positives) will stay another day in the hospital, but will save on readmission costs.

- Patients who will readmit and are missed by our model (False Negatives) will not stay another day in the hospital, but will cost the hospital readmission costs at some point in the future.

- Patients who will not readmit but are identified as at-risk by our model (False Positives) will stay another day in the hospital needlessly.

- Patients who will not readmit and are identified as not-at-risk by our model (True Negatives) do not incur any additional hospital costs.

We visualize the total operational cost of enforcing one additional inpatient day for patients who are identified as at-risk of early readmission, shown Figure 11. Here, the global minimum in cost is near our original decision threshold, but this may change depending on a hospital's specific costs and intervention strategies.

Comprehensive discharge planning presented in Naylor et al [7] describes a protocol where a nursing specialist contacts the patient 24-48 hours after the initial hospital admission, assesses the patient and caregiver needs and expectations, and creates a preliminary discharge plan for when the patient is eventually discharged to their home. The specialist visits the patient in the hospital at least every 48 hours to determine if it's necessary adjust the preliminary plan. The nursing specialist calls the patient or their caregiver at least twice within the first 24-48 hours after discharge to assess their condition, educate, and re-adjust the current plan. The specialist is available by phone for at least two weeks after discharge. This protocol demonstrates a combination of personalized discharge planning, patient and caregiver assessment, and ongoing patient-specialist communication in the days following discharge. It has been shown to reduce the number of readmissions in an intervention group of patients aged 70 or older.

The nursing specialists in the study spent, on average, about 3.59 hours on discharge planning and about 46.6 minutes, totalling an average of 4.363 paid hours. If nursing specialists are compensated for about 65 dollars an hour, we can visualize the operational cost as a function of our classifier's decision threshold and the average total number of worked hours by the nursing specialist.

In Figure 12, we visualize the estimated total operational cost of implementing comprehensive discharge planning, across varying levels of paid nursing hours. The fewer hours that nursing specialists must work per patient, the more financially viable it is to implement comprehensive discharge planning for
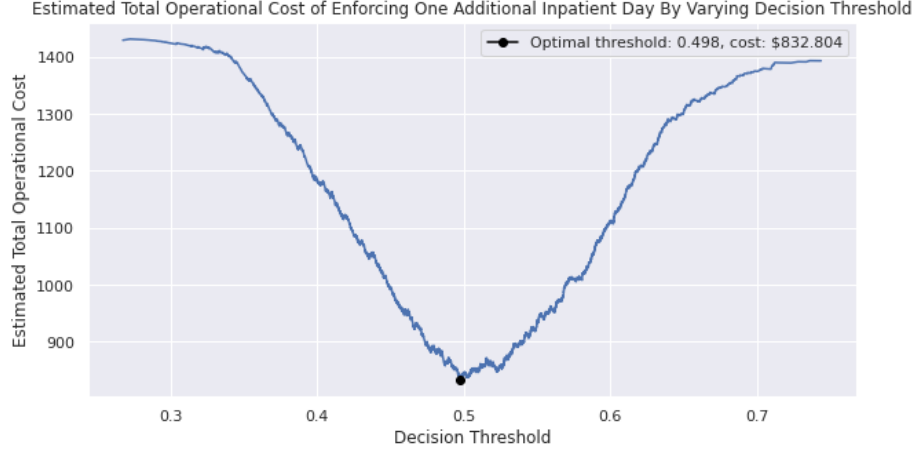
Figure 11: The estimated total operational cost of enforcing one additional inpatient day in response to high risk of early hospital readmission. We see that the global minimum of this curve is based on a decision threshold that's very similar to the optimal threshold that we found for our ROC. Increasing hospital stay may seem like a reasonable readmission reduction strategy in theory, but Rachoin et al showed that length of stay is associated with a higher likelihood of hospital readmissions. It's likely that there exists a confounding factor that warrants a longer length of stay while also increasing the likelihood of readmission (e.g. severity of principal diagnosis, number of comorbidities) but this may be sufficient in assuming that increasing length of stay alone does not adequately reduce hospital readmission rates [4].

all patients. Based on Naylor et al [7], the first figure is most similar to the nursing hours observed in a randomized clinical trial. Therefore, not only does this strategy appear to be a promising cost-saving approach, it may also displace the need to identify early readmission patients if it proves sufficiently cost-effective for all patients.

7.3 Intervention Domain Knowledge

Finally, the most useful predictive models can be leveraged early enough to allow for intervention. Therefore, using features like whether the patient was discharged to a skilled nursing facility may be too late for proper intervention. Future work may involve understanding how early prediction must occur to allow for intervention and build a model independent of features that do not meet this temporal criteria.
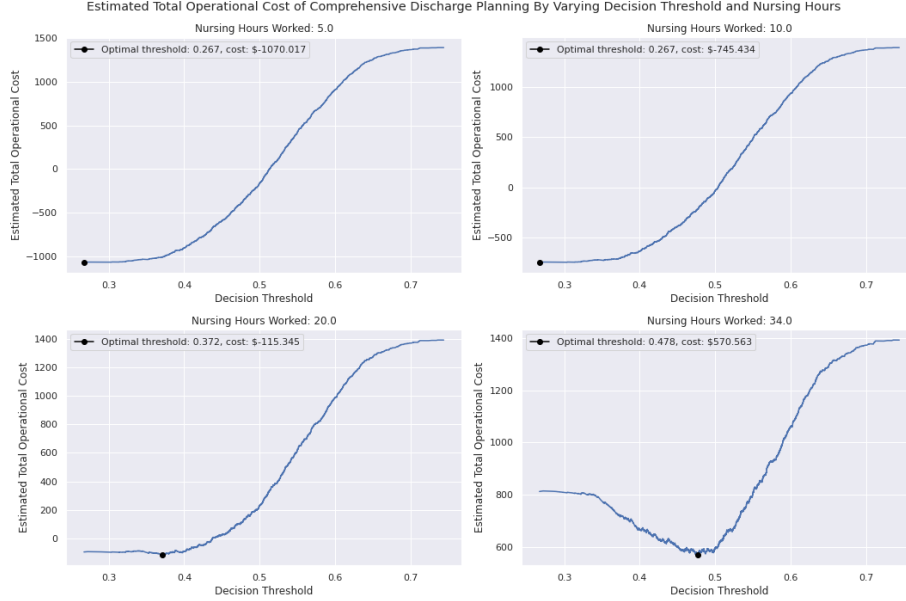
Figure 12: Estimated total operational cost of implementing comprehensive discharge planning, based on Naylor et al [7], across varying levels of paid nursing hours.

8 Recommendations

- Considering our section on Costs and Implementation Strategies, it is advantageous for hospitals to 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. Said differently, 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. After performing a cost-benefit analysis similar to 11, hospitals can set their a decision threshold adjusted for their specific use case, before applying it in practice.

- We recommend that hospitals aggregate racial and socioeconomic data specific to their hospital in order to re-train 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.

- 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. Even using a reasonably high sensitivity, this model reduces the amount of patients a physician would have to assess for risk of early readmission.

References

[1] *[1106.1813] SMOTE: Synthetic Minority Over-sampling Technique*. URL: https://arxiv.org/abs/1106.1813 (visited on 08/20/2021).

[2] Molly K Bailey and Marguerite L Barrett. "Characteristics of 30-Day All-Cause Hospital Readmissions, 2010-2016". en. In: (), p. 14.

[3] CDC. *All About Your A1C*. en-us. Aug. 2018. URL: https://bit.ly/2Nc2IA0 (visited on 08/20/2021).

[4] *Cureus — Impact of Length of Stay on Readmission in Hospitalized Patients*. URL: https://www.cureus.com/articles/40535-impact-of-length-of-stay-on-readmission-in-hospitalized-patients (visited on 08/20/2021).

[5] *Hospital Readmissions Reduction Program (HRRP) — CMS*. URL: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program (visited on 08/20/2021).

[6] "National Diabetes Statistics Report 2020. Estimates of diabetes and its burden in the United States." en. In: (2020), p. 32.

[7] Mary Naylor et al. "Comprehensive Discharge Planning for the Hospitalized Elderly". In: *Annals of Internal Medicine* 120.12 (June 1994). Publisher: American College of Physicians, pp. 999–1006. ISSN: 0003-4819. DOI: 10.7326/0003-4819-120-12-199406150-00005. URL: https://www.acpjournals.org/doi/10.7326/0003-4819-120-12-199406150-00005 (visited on 08/20/2021).

[8] Robert Robinson and Tamer Hudali. "The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital". In: *PeerJ* 5 (Mar. 2017), e3137. ISSN: 2167-8359. DOI: 10.7717/peerj.3137. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374974/ (visited on 08/20/2021).

[9] Jade Gek Sang Soh et al. "Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with meta-analysis". en. In: *BMJ Open Diabetes Research and Care* 8.1 (Aug. 2020). Publisher: BMJ Specialist Journals Section: Clinical Care/Education/Nutrition, e001227. ISSN: 2052-4897. DOI: 10.1136/bmjdrc-2020-001227. URL: https://drc.bmj.com/content/8/1/e001227 (visited on 08/20/2021).

[10]   Elias K. Spanakis and Sherita Hill Golden. "Race/Ethnic Difference in Diabetes and Diabetic Complications". In: *Current diabetes reports* 13.6 (Dec. 2013), 10.1007/s11892–013–0421–9. ISSN: 1534-4827. DOI: 10.1007/s11892-013-0421-9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3830901/ (visited on 08/20/2021).

[11]   *UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set.* URL: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 (visited on 08/20/2021).

[12]   uncanny-valley. *diabetes_readmission.* original-date: 2021-05-02T21:45:40Z. Aug. 2021. URL: https://github.com/uncanny-valley/diabetes-readmission (visited on 08/20/2021).