Discovering Factors Contributing to Admittance into the Hospital from the Emergency Department

Leontij Potupin and Rachel Wu

December 19th, 2021

Contents

The code to reproduce this report is available on Github.

1 Executive Summary

Problem. With the exception of the hiatus during the peak of the COVID-pandemic, emergency departments in the United States have become increasingly crowded and play a central role in the healthcare of U.S. patients. Many patients use the emergency department as a substitute for primary care, especially patients without health insurance, and many patients show up to the department as a precaution, as many of us without medical training have a difficult time assessing the gray area between feeling ill and needing immediate medical attention. Due to a combination of these factors, emergency departments are congested and are often unable to meet the needs of the community. We explore the patient characteristics that affect the likelihood for a patient to be admitted to the hospital after presenting to the emergency department. With our exploration, we hope to elucidate the factors that contribute most to a patient's chance of being admitted into the hospital, so patients might be able to better select the severity of their own condition and be more selective about visiting the emergency department.

Data. Our data comes from the publicly available National Hospital Ambulatory Medical Care Survey (NHAMCS). It is compiled by the Centers for Disease Control and Prevention and a nationally representative survey of emergency departments in the United States when used with the weights included in the data set. However, we did not use the survey weights for our analysis. Each of the observations is a discrete patient visit to an emergency department. The features we use in our analysis are measurements taken upon entry to the emergency department and during the patient's stay in the emergency department. We faced challenges with selecting variables, classifying categorical variables as factors, and dealing with missing data.

Analysis.

Conclusions.

2 Introduction

Background. Coronavirus disease (COVID-19) has had a devastating global impact, with a cumulative total of 149,987,772 confirmed cases and 3,157,594 deaths worldwide as of April 28, 2021.[1] About a fifth of these cases have been in the United States, with a recent count of 32,551,440 cases and 582,668 deaths.[2] With these staggering numbers still increasing despite recent large-scale vaccine rollouts, it is of vital importance to utilize various data sources to understand both the progression of COVID-19 thus far as well as the highest risk factors for contracting COVID-19. Furthermore, a thorough analysis of COVID-19 rates and predictive factors may help inform strategies to improve public health policies that could mitigate the negative impact of a future pandemic, which many scientists say is not a matter of if but of when.[3]

Past research has shown that infectious diseases are influenced by a variety of factors. Obesity, for instance, is associated with a higher likelihood of contracting influenza A, and seasonal temperature changes have shown to be predictive of the 2003 severe acute respiratory syndrome (SARS).[4] The CDC is currently in the process of identifying potential risk factors for severe COVID-19 illness,[5] and some that have already been identified include heart disease, diabetes, and pregnancy.[6] Yet despite these efforts, there is still much to be learned. Specifically, there is still insufficient research to explain the differences in COVID-19 susceptibility and mortality that exist not just on the individual level but also on broader population levels.

[1] Coronavirus Cases: Worldometer. (n.d.). https://www.worldometers.info/coronavirus/.

[2] Ibid.

[3] Robbins, J. (2021, January 4). Heading Off the Next Pandemic. Kaiser Health News. https://khn.org/news/infectious-disease-scientists-preventing-next-pandemic/.

[4] Tian, T., Zhang, J., Hu, L., Jiang, Y., Duan, C., Li, Z., . . . & Zhang, H. (2021). Risk factors associated with mortality of COVID-19 in 3125 counties of the United States. Infectious diseases of poverty, 10(1), 1-8.

[5] Centers for Disease Control and Prevention. (n.d.). Assessing Risk Factors for Severe COVID-19 Illness. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html.

[6] Centers for Disease Control and Prevention. (n.d.). Certain Medical Conditions and Risk for Severe COVID-19 Illness. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

Analysis goals. As our analysis goal is to be able to return key factors that are most likely to predict a patient's admission to the hospital or discharge from the emergency department, success is defined as consistent agreement among models as to the features that contribute to a hospital admission.

Significance. We hope that our analysis will contribute to the understanding of key drivers in healthcare management. Emergency departments are chronically overloaded and aiding the flow of patients in and out of the department through optimizing and anticipating a patient's disposition will allow the emergency department to aid more patients.

3 Data

3.1 Data sources

Our data came from the National Hospital Ambulatory Medical Care Survey (NHAMCS). NHAMCS is compiled by the CDC to "meet the need for objective, reliable information about the provision and use of ambulatory medical care services in the United States", according to the NHAMCS home page, and is primarily used by researchers and policy makers. We uploaded the data from the CDC website and concatenated the datasets from years 2015-2019, inclusive. Although NHAMCS has been produced every year since 1992, the variables change from year to year and even with five years of data yielded management problems in regards to having to identify which columns were added or deleted between years. The 2019 data set is the most recent data set to date. Changes in emergency department management due to the COVID-19 pandemic are not reflected in our analysis.

3.2 Data cleaning

After merging five datasets together, the main data cleaning task became researching and eliminating variables that were indicative of a patient's disposition. It was important that we did not have any features that had a 1-to-1 correspondence with the response variable, ADMITHOS, a binary feature indicating whether the patient had been admitted to the emergency department's hospital or not. Given that we were investigating the factors that contribute to a patient disposition of an "admit" to the hospital, many of the 1058 features had to be removed from the cleaned data set because they were taken after a disposition decision was made.

Data cleaning also involved fixing the problem of copious missing data entries. NHAMCS changes the features collected from year to year, so there were some variables that were collected in 2015, but not in 2019, for example. Some features are impossible to impute, such as the prescription status codes or the controlled substance status codes, and have greater than 90% missing entries across all observations, such as the feature that records the code for the 20th prescription medication taken by the patient, since most patients do not ingest 20 pills on a daily basis. We removed those variables from our dataset and will speak to the limitations caused by the removal of variables in our conclusion.

For features that included missing data but were numerical or were factorable, we calculated the means to impute data into the "blank" and "unknown" slots and turned the features with categorical responses into factors. Blank is coded as -9 in this dataset, so factored features have a level of -9. Finally, we removed any rows with NA values to allow our models to run smoothly.

3.3 Data description

3.3.1 Observations

Our dataset has a total of 935 observations, corresponding to each of the counties included in our analysis.

Table 1: Hospital admittance by race.

Race	Hospital Admittance Rate
Non-Hispanic White	11.77%
Non-Hispanic Black	8.04%
Hispanic	10.79%
Total	10.90%

3.3.2 Response Variable

Our response variable is the ADMITHOS feature, which is a categorical variable assigned a 1 if the patient was admitted to this hospital and 0 if they were not admitted. All datasets 2015-2019 included this feature.

We have several similar features that tell us information on patient disposition (where a patient went after appearing at the emergency department). OBSHOS (indicating patients were admitted to the hospital after observation in the emergency department), TRANPYSC (patient transferred to a psychiatric hospital), and TRANOTH (patient was transferred to another hospital) were also variables that indicate the patient experienced additional care after being seen in the emergency department. However, we decided to focus only on ADMITHOS because of the clarity of the variable compared to the other potential response variables. TRANPYSC could indicate the patient went to a psychiatric hospital but may not have been admitted to that hospital; the same concern plagues the TRANOTH variable. OBSHOS is also binary and for every positive OBSHOS observed, there is also a positive ADMITHOS value for that given patient, so we eliminated OBSHOS as a feature so as to avoid a direct linear correspondence in our regression analysis.

3.3.3 Features

We include 150 total features. For a detailed specification of these variables, refer to Appendix A.

3.4 Data allocation

After cleaning the data, we randomly selected 80% of our dataset into a training set and reserved the remaining 20% for our testing datasets.

3.5 Data exploration

3.5.1 Response

We found that 10.9% of patients were admitted to the hospital after being seen in the emergency department, enhancing our understanding of the response variable's distribution within our dataset. We created histograms depicting the overlay of some relevant features with ADMITHOS, our response variable, to get a better feel for the distribution of our response variable.

3.5.2 Features

Looking at the race feature revealed that hospital admittance rates vary greatly across ethnic groups. Table 1 shows that 11.8% of white people get admitted to the hospital whereas only 8.0% of black people get admitted which poses the question of fairness and equity in the healthcare system.

As shown in Figure 2, the AGE variable in the cleaned dataset has two humps around 0 and 25 with an overall median age of 39. This is approximately consistent with the age distribution of the United States where there the age distribution declines approximately linearly with increasing age.

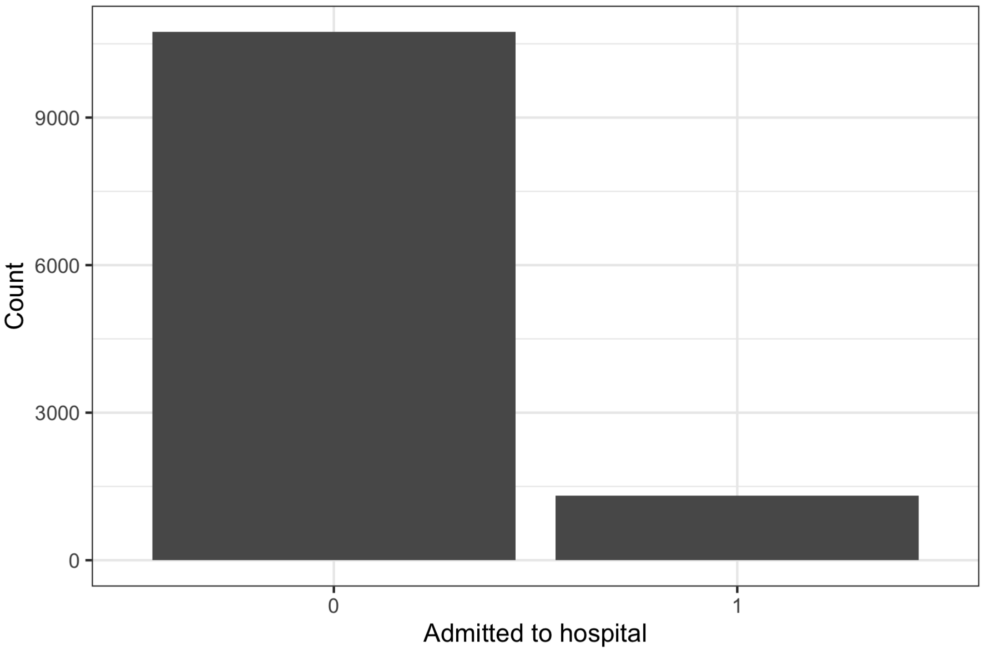

Figure 1: Distribution of case-fatality rate; vertical dashed line indicates the median.

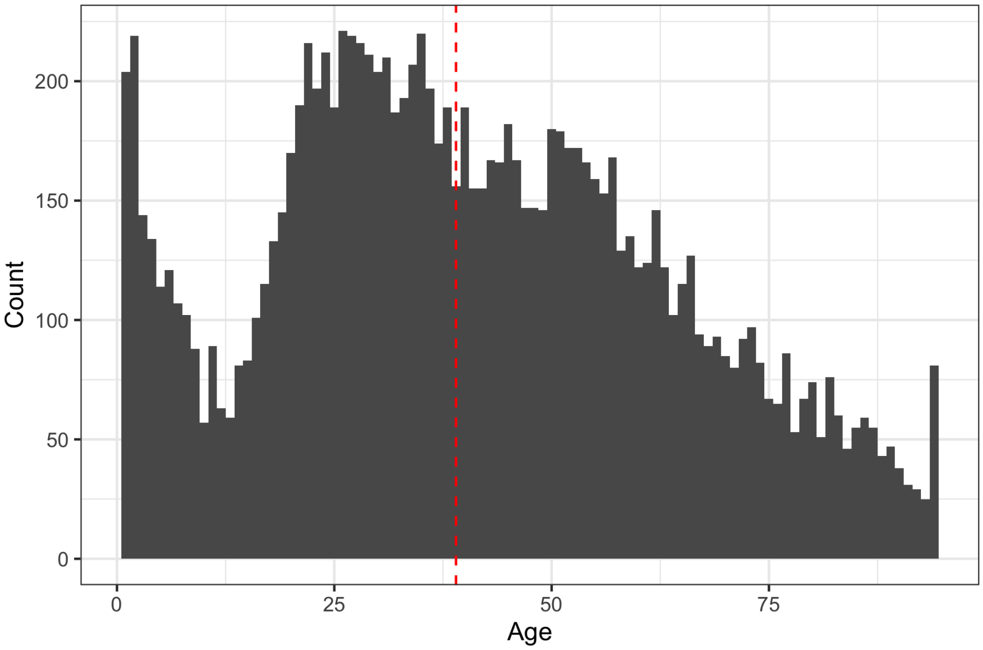

Figure 2: Distribution of age; vertical dashed line indicates the median.

4 Modeling

4.1 Regression-based methods

4.1.1 Logistic Regression

We started our analysis using a logistic regression based on a limited subset of 50 features due to the high complexity of a logistic regression classifier with 150 variables that would be prone to overfitting.

We were able to achieve a baseline misclassification error of 9.21%. Our approach was to improve on that performance by using other learning methods and expanding the feature space.

ROC CURVE Here

4.1.2 Penalized regression

4.1.2.1 Lasso Logistic Regression For the lasso, Figure 3 shows the CV plot, Figure 4 shows the trace plot, and Table 2 shows the selected features and their coefficients.

It is noteworthy that the variable “TOTDIAG”, i.e. the total number of diagnostic services ordered or provided seems to be important and indicative of whether or not a patient gets admitted to the hospital. “CONSULT0” has the largest negative coefficient which makes sense since not consulting a physician might increase the likelihood of not being admitted to the hospital.

With the lasso classifier we were able to achieve a misclassification error of 8.71%.

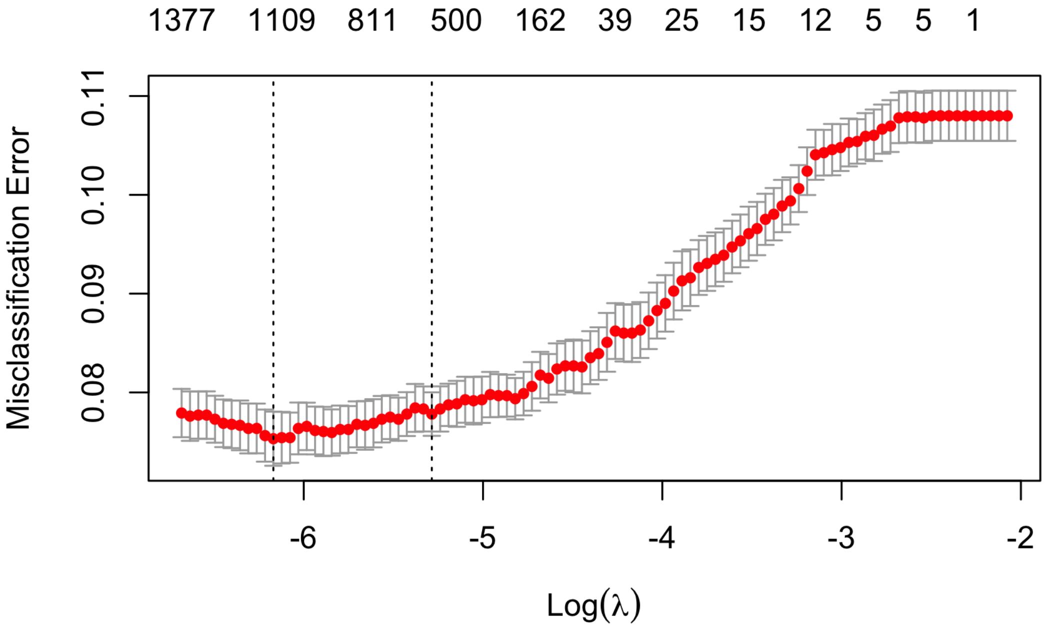

Figure 3: Lasso CV plot.

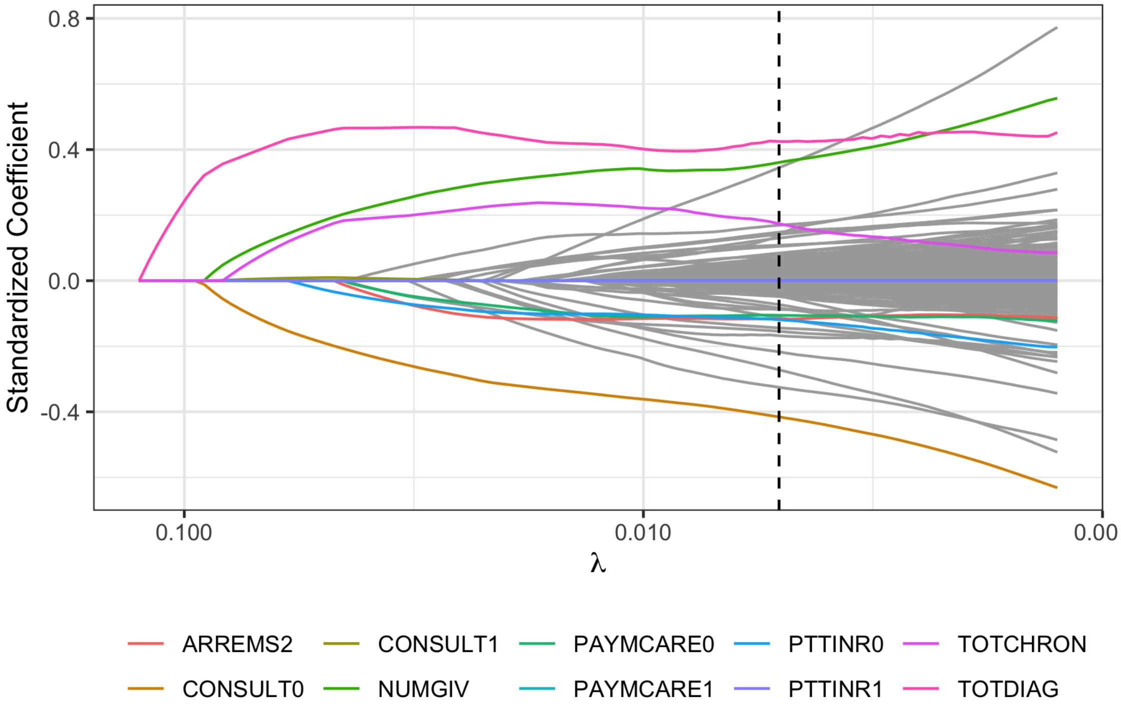

Figure 4: Lasso trace plot.

Table 2: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
TOTDIAG	0.41
CONSULT0	-0.40
NUMGIV	0.34
NUMDIS	-0.31
RETRNED0	0.30
OBSSTAY	-0.24
ZONENURS2	-0.20
TOTCHRON	0.19
CBC0	-0.17
AGE	0.16

4.1.2.2 Ridge Logistic Regression For the ridge regression, Figure 5 shows the CV plot, Figure 6 shows the trace plot, and Table 3 shows the selected features and their coefficients.

It is interesting to see that ridge also gave the "CONSULT" feature a high weight, both in the negative and positive range.

With the ridge classifier our missclassification error was 10.7%.

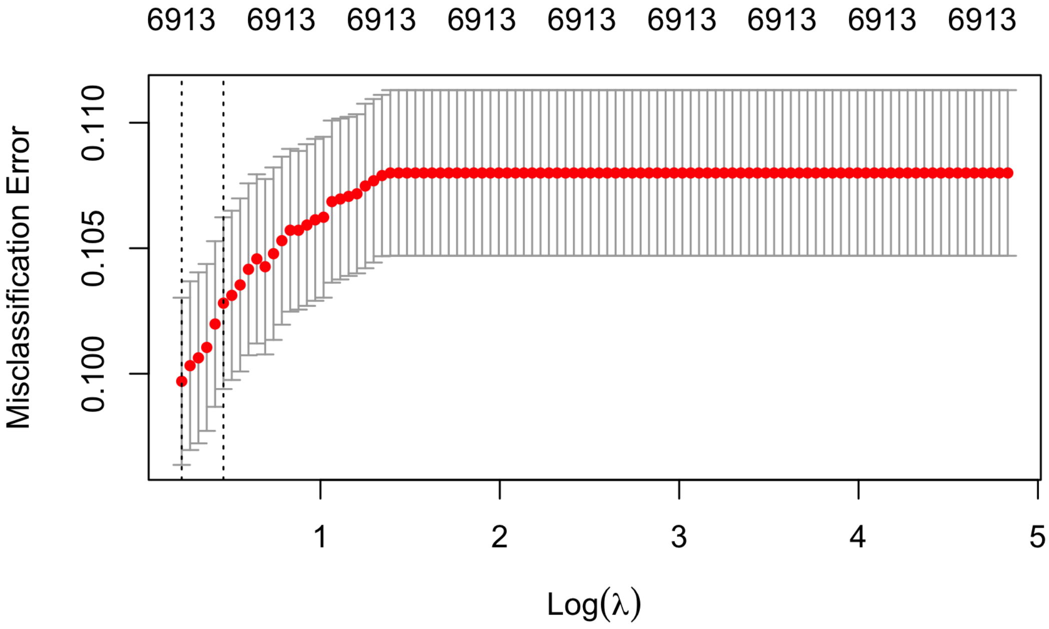

Figure 5: Ridge CV plot.

4.1.2.3 Elastic Net Regression As a next step, we used an elastic net regression model to get the benefits from ridge-like shrinkage as well as lasso-like selection.

For the elastic net regression, Figure 7 shows the trace plot, and Table 4 shows the selected features and their coefficients.

We were able to achieve a 9.54% misclassification error using the elastic net classifier. Since the 8.71% missclassification error of the lasso classifier is the lowest one out of the regression based methods, this suggests that few features have large effects.

4.2 Tree-based methods

4.2.1 Unpruned Decision Tree

For our first tree-based method we used an unpruned decision tree to predict the admittance to the hospital. As seen in 8, "NUMDIS" or the number of medications prescribed at discharge is a good feature to split on.

With the unpruned decision tree we were able to achieve a misclassification error of 11.4%.

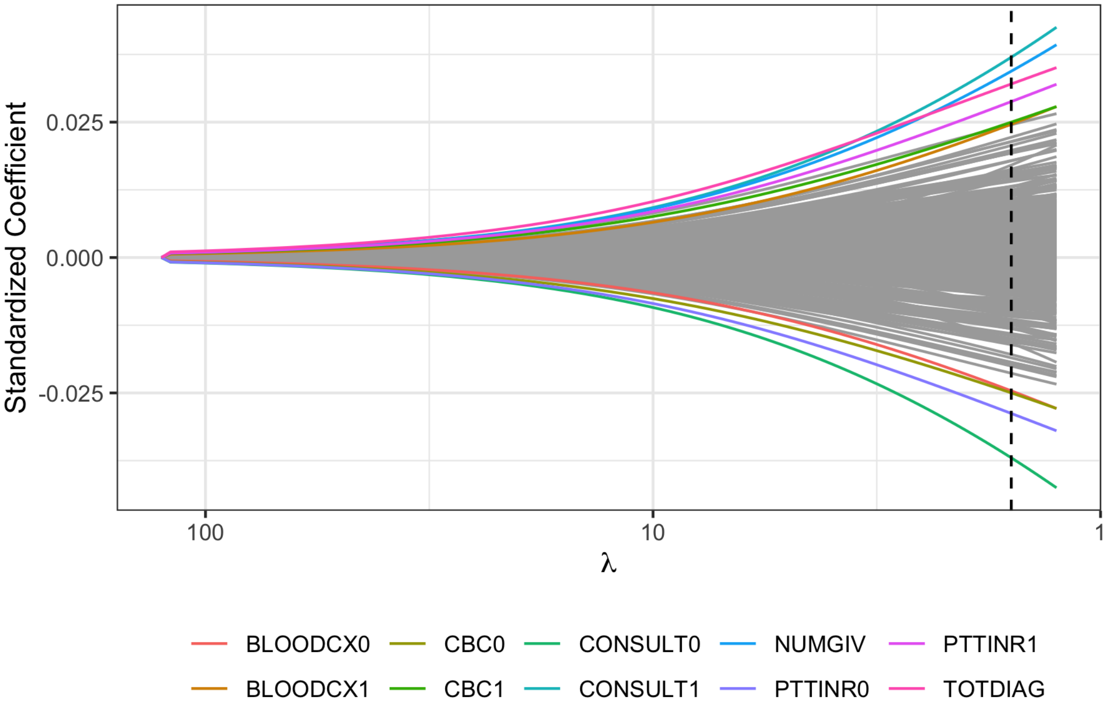

Figure 6: Ridge trace plot.

Table 3: Standardized coefficients for features in the ridge model based on the one-standard-error rule.

Feature	Coefficient
CONSULT0	-0.04
CONSULT1	0.04
NUMGIV	0.03
TOTDIAG	0.03
PTTINR0	-0.03
PTTINR1	0.03
CBC0	-0.02
CBC1	0.02
BLOODCX1	0.02
BLOODCX0	-0.02

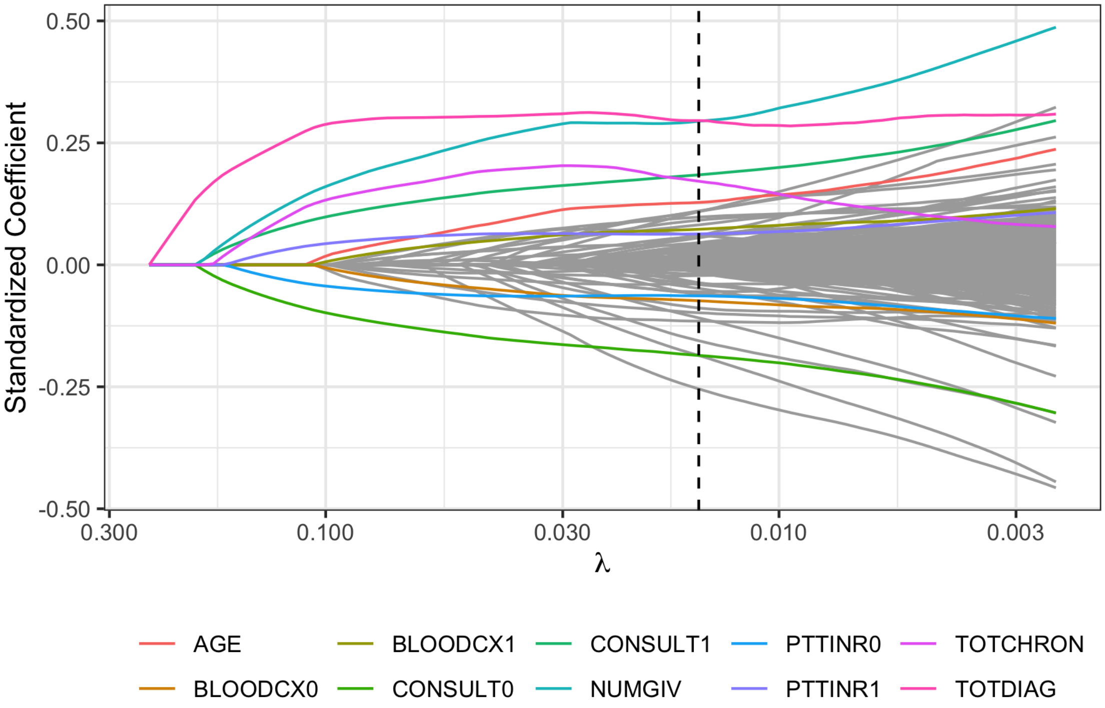

Figure 7: Elastic Net trace plot.

Table 4: Standardized coefficients for features in the Elastic Net model based on the one-standard-error rule.

Feature	Coefficient
TOTDIAG	0.30
NUMGIV	0.29
NUMDIS	-0.25
OBSSTAY	-0.19
CONSULT0	-0.19
CONSULT1	0.18
TOTCHRON	0.17
ZONENURS2	-0.16
AGE	0.13
ARREMS2	-0.12

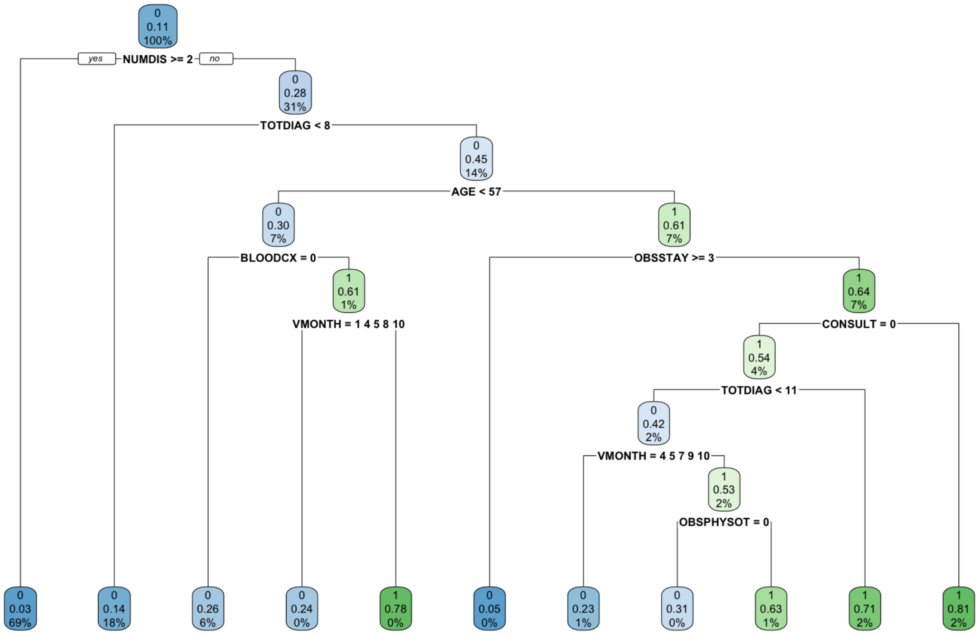

Figure 8: Unpruned Tree Plot.

4.2.2 Pruned Decision Tree

After pruning the decision tree, the misclassification error improves slightly to 11.2%. The pruned Tree CV error plot in Figure **??** shows that the optimal number of terminal nodes is 3.

4.2.3 Random Forest

4.2.4 Boosting

5 Conclusions

5.1 Method comparison

Table **??** shows the misclassification for all the methods considered. Except for the OLS, the random forest and the boosted model have the lowest test errors. This is reasonable given these models' tendencies to have high predictive accuracy. Between the two, the boosted model has the lowest test error, with a mean squared error of 0.000139, but it is closely followed by random forest, which has a mean squared error of 0.00141. Notably, however, the ridge, LASSO, and elastic net regressions perform about as well, with test MSEs of 0.000158, 0.000164, and 0.000161, respectively. Although OLS has the lowest training and test error, its adjusted R-squared value was only about 0.3, and there were too many features given the number of observations.

Regardless of these differences in test MSE, the methods overlap significantly in their identification of important variables from the larger set. For instance, the elastic net regression selects the following variables, which are also selected by LASSO and deemed significant in the OLS model: other providers ratio, unemployment, income inequality, housing overcrowding, residential segregation—non-White/White, homeownership,

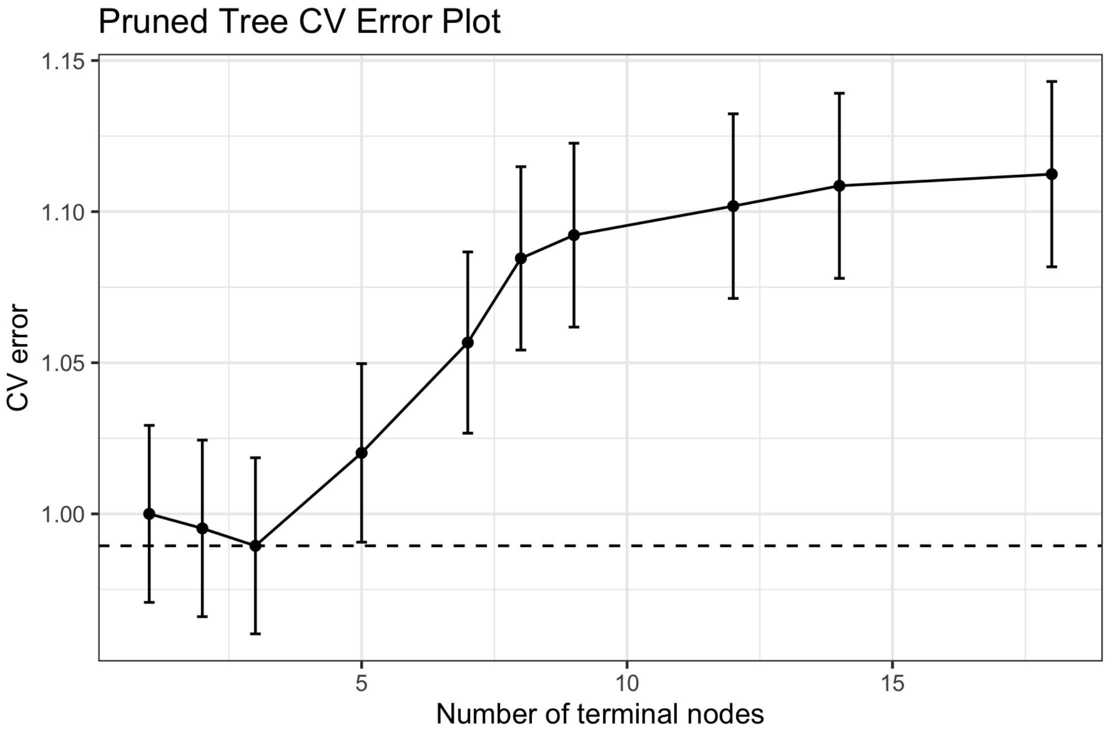

Figure 9: Pruned Tree CV Error Plot.

(#fig:tree-pruned- plot)

and physical inactivity. The random forest and boosting models both include low birthweight percentage, median income, and unemployment percentage in the top 10 most important variables, as measured by their contributions to node purity.

5.2 Takeaways

5.3 Limitations

5.3.1 Dataset limitations

NHAMCS changes the recorded variables every year, which makes it difficult to maintain consistency when concatenating five years of data. Additionally, looking at ADMITHOS as our response variable forced us to remove features that had a direct relationship with our response variable. For example, we removed any features that were measured at the time of discharge, since the presence of a value for these variables indicates that the patient is not being admitted to the hospital and is being released from the hospital.

Missing data in general caused us to lose valuable observations, when we dropped the rows with NA values, and it also cost us valuable features, which had to be left out due to less than 90% of observations containing values for these particular features. We lost about 90% of our total data due to problems with missing data and also imputed means for some missing values, which increases bias and lowers the variance in our results.

Furthermore, this data set precedes the COVID-19 pandemic and does not illustrate any admission changes due to COVID. For example, it is likely that certain demographics sorted by age and race will be admitted to the hospital in the NHAMCS 2020 and 2021 datasets more frequently than in the 2015-2019 datasets, given the evidence that COVID fatalities were higher among populations of color and that one might reasonably presume that these populations would be visiting the hospitals more often due to COVID.

5.3.2 Analysis limitations

For the unpruned tree output, we ran into problems creating the random forests and boosting models. Some of our features included more than 1054 factors levels, features related to diagnosis codes or medicine names, and we were forced to exclude those variables from our training dataset.

5.4 Follow-ups

A Appendix: Descriptions of features

Below are the 150 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are categorical.

Date of Visit - Month of visit (VMONTH): 1-12, January-December - Day of the week (VDAYR): 1-7, Sunday-Saturday

Patient's Reason for Visit - Reason for visit #1 (RFV1): coded 1005.0-8999.0 - Reason for visit #2 (RFV2): coded 1005.0-8999.0 - Reason for visit #3 (RFV3): coded 1005.0-8999.0 - Reason for visit #1 - broad (RFV13D): coded 0-1260 - Reason for visit #2 - broad (RFV23D): coded 0-1260 - Reason for visit #3 - broad (RFV33D): coded 0-1260

Patient Medical History - Alzheimer's/Dementia (ALZHD) - Asthma (ASTHMA) - Cancer (CANCER) - Cerebrovascular disease/History of stroke (CEBVD) - Chronic kidney disease (CKD) - Chronic obstructive pulmonary disease (COPD) - Congestive heart failure (CHF) - Coronary artery disease, ischemic heart disease, or hx of MI (CAD) - Depression (DEPRN) - Diabetes type 1 (DIABTYP1) - Diabetes type 2 (DIABTYP2) - Diabetes type unspecified (DIABTYP0) - Obesity (OBESITY) - Obstructive sleep apnea (OSA) - Osteoporosis (OSTPRSIS) - Substance dependence or abuse (SUBSTAB)

- None of the above (NOCHRON)
- Total number of chronic conditions (TOTCHRON): range 0-14

Diagnostic Services: - Were diagnostic services provided at this visit? (DIAGSCRN)

- Any imaging (ANYIMAGE)
- Arterial blood gases (ABG): laboratory test
- Blood alcohol concentration (BAC):
- Basic metabolic panel (BMP):
- Blood culture (BLOODCX)
- Brain natriuretic peptide (BNP)
- Cardiac Enzymes (CARDENZ):
- Cardiac Monitor (CARDMON)
- Complete blood count (CBC):
- Comprehensive metabolic panel (CMP):
- Creatinine/Renal function panel (BUNCREAT)
- CT abdominal/pelvic scan (CTAB)
- CT chest scan (CTCHEST)
- CT head scan (CTHEAD)
- CT with IV contrast (CTCONTRAST)
- CT scan (CATSCAN)
- CT scan other (CTOTHER)
- CT scan site unspecified (CTUNK)
- Liver enzymes/Hepatic function panel (LFT)
- MRI (MRI)

- Other blood test (OTHRBLD)
- Other culture (OTHCX)
- Other imaging (OTHIMAGE)
- Other test/service (OTHRTEST)
- Pregnancy test (PREGTEST)
- Prothrombin time (PTTINR)
- Throat culture (TRTCX)
- Total number of diagnostic services ordered (TOTDIAG): 0-20 range
- Toxicology screen (TOXSCREN)
- Ultrasound (ULTRASND)
- Urine dipstick (URINE)
- Urine culture (URINECX)
- Wound culture (WOUNDCX)
- X-ray testing (XRAY)

Procedures - Bilevel positive airway pressure device (BPAP) - Bladder catheter (BLADCATH) - Cast, splint, wrap (CASTSPLINT) - Central line (CENTLINE) - CPR (CPR) - Lumbar puncture (LUMBAR) - Nebulizer therapy (NEBUTHER) - Pelvic exam (PELVIC) - Skin adhesives (SKINADH) - suturing/staples (SUTURE) - Other procedure (OTHPROC) - Were procedures provided at this visit? (PROC) - Total number of procedures provided (TOTPROC): range of 0-6

Medications - Were medications given at this visit? (MED) - Medication #1 (MED1) - Number of medications given in ED (NUMGIV): range of 0-30 - Number of medications prescribed at discharge (NUMDIS): range of 0-30 - Number of medications coded (NUMMED)

Providers seen - Consulting physician (CONSULT) - ED attending physician (ATTPHYS) - ED resident or intern (RESINT) - Mental health provider (MHPROV) - Nurse practitioner (NURSEPR) - Physician assistant (PHYSASST) - Other provider (OTHPROV) - RN or LPN (RNLPN) **Supervisors for Observation Unit** - ED physicians (OBSPHYSED) - Hospitalists (OBSHOSP)

Hospital Management and Equipment Available - Computer assisted triage (CATRIAGE) - Electronic dashboard displaying updated patient info and status (DASHBORD) - Zone nursing (ZONENURS): all nurse's patients located in one area

Miscellaneous - Length of stay in observation unit in minutes (OBSSTAY)