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. Our analysis

2 Introduction

Background. Emergency departments (EDs) across the United States are facing a crisis: chronic overcrowding. While most patients eventually receive care, the wait times are lengthy and 3/4ths of patients who left before receiving care cited these wait times as the reason for leaving early[1]. The national benchmarks of care set the wait times at an ED to be 30 minutes or less , yet the 56% of patients waited almost an hour before being seen[2]. This is not merely an inconvenience for patients; studies have shown that patients are harmed by these longer wait times for reasons including morbidity and mortality related to consequential delays of treatment for both high- and low-acuity patients, ambulance diversion, increased adverse events, and preventable error [3]. EDs use a triage system to filter patients into the ED, determining who needs to be seen urgently and who can wait for another hour. However, there is little research on the factors contributing to the outwards flow from the emergency department to the hospital, and studying the factors that lead to patients leaving the emergency department could contribute to faster movement of patients out of the ED, which in turn would ease the traffic within the ED and help reduce patient wait times at the emergency department.

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

[1] Kelen, G., Wolfe,R. D'Onofrio, G., et al. Emergency Department Crowding: The Canary in the Health Care System (2021). https://catalyst.nejm.org/doi/full/10.1056/CAT.21.0217.

[2] Odorczyk, K. Reinventing the Emergency Department at the Hospital of the University of Pennsylvania's New Pavilion (2021). https://www.pennmedicine.org/news/news-blog/2021/june/reinventing-the-emergency-department-at-the-new-pavilion

[3] Morley, C., Unwin, M., Peterson, GM., Stankovich, J., Kinsman, L. Emergency department crowding: A systematic review of causes, consequences and solutions.PLoS One2018; 13:e0203316 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203316.

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 comes 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.

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.

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%

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.

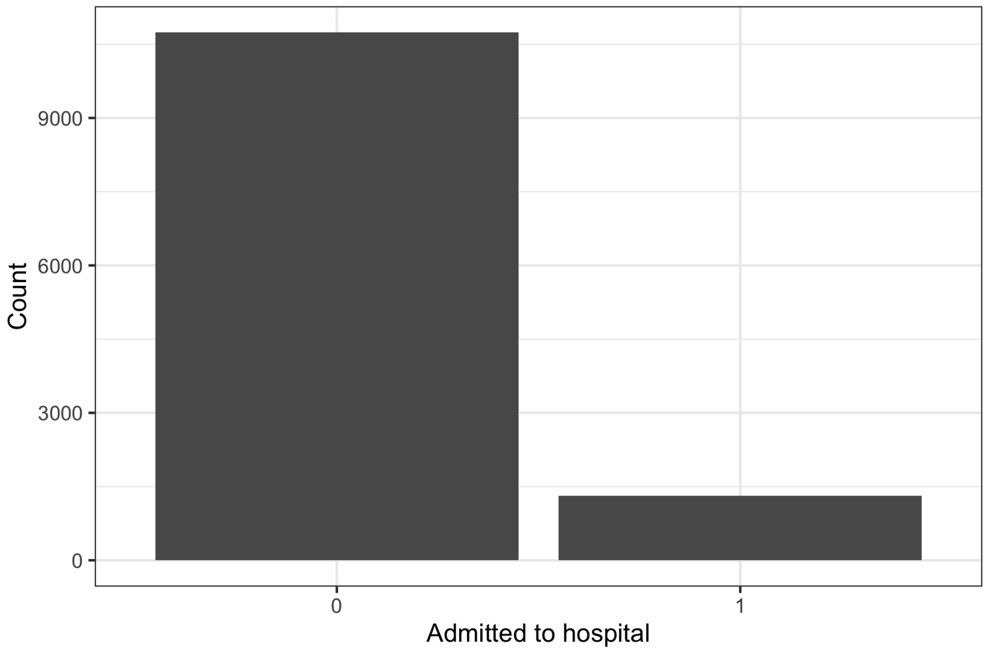

Figure 1: Distribution of case-fatality rate; vertical dashed line indicates the median.

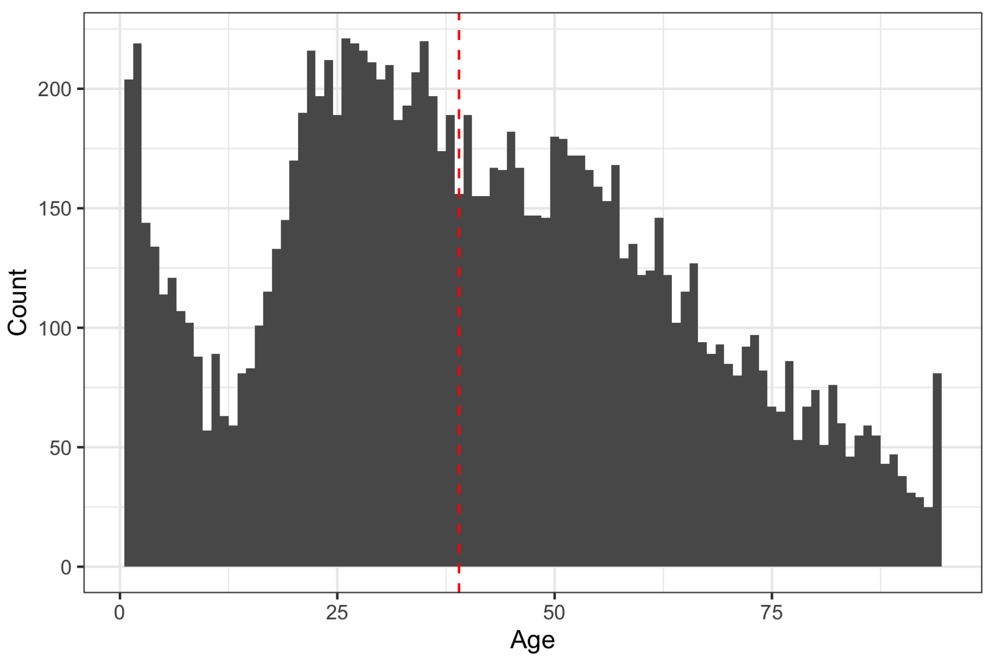

Figure 2: Distribution of age; vertical dashed line indicates the median.

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.

We proceeded to evaluate our classifier with an ROC Curve in Figure 3. The area under the curve is 0.6812, making the baseline classifier better than a naive classifier.

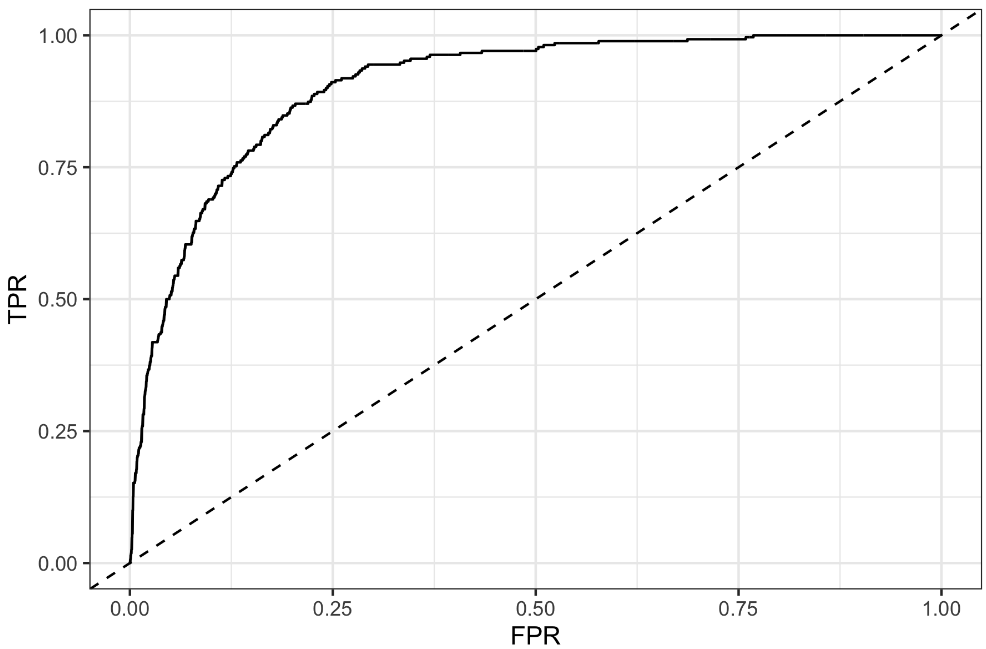

Figure 3: ROC curve for logistic regression, area under the curve of 0.6812.

4.1.2 Penalized regression

4.1.2.1 Lasso Logistic Regression For the lasso, Figure 4 shows the CV plot, Figure 5 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%.

4.1.2.2 Ridge Logistic Regression For the ridge regression, Figure 6 shows the CV plot, Figure 7 shows the trace plot, and Table 3 shows the selected features and their coefficients.

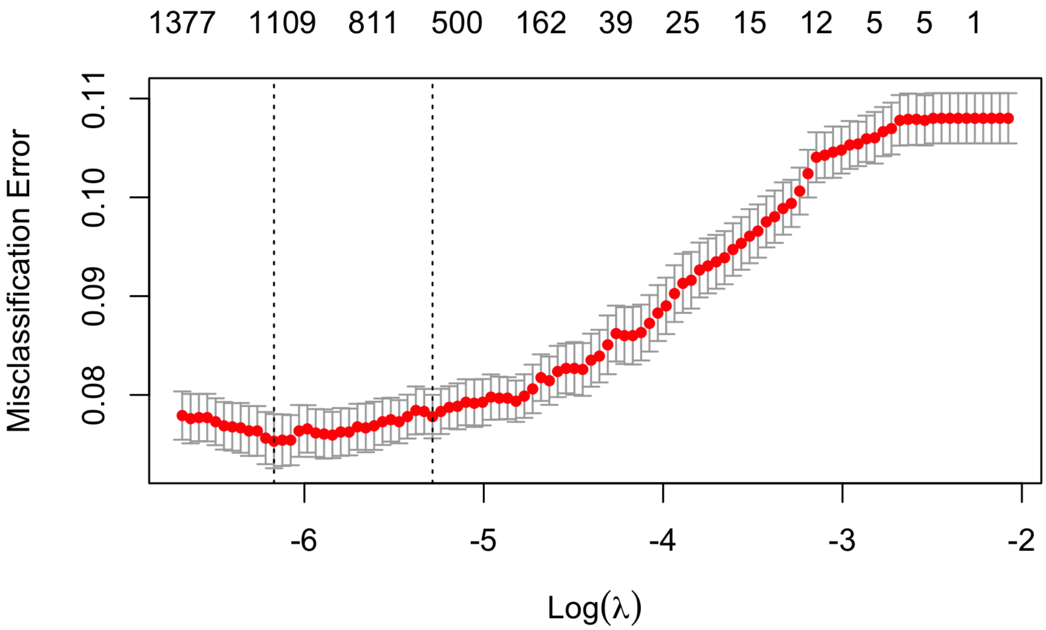

Figure 4: Lasso CV plot.

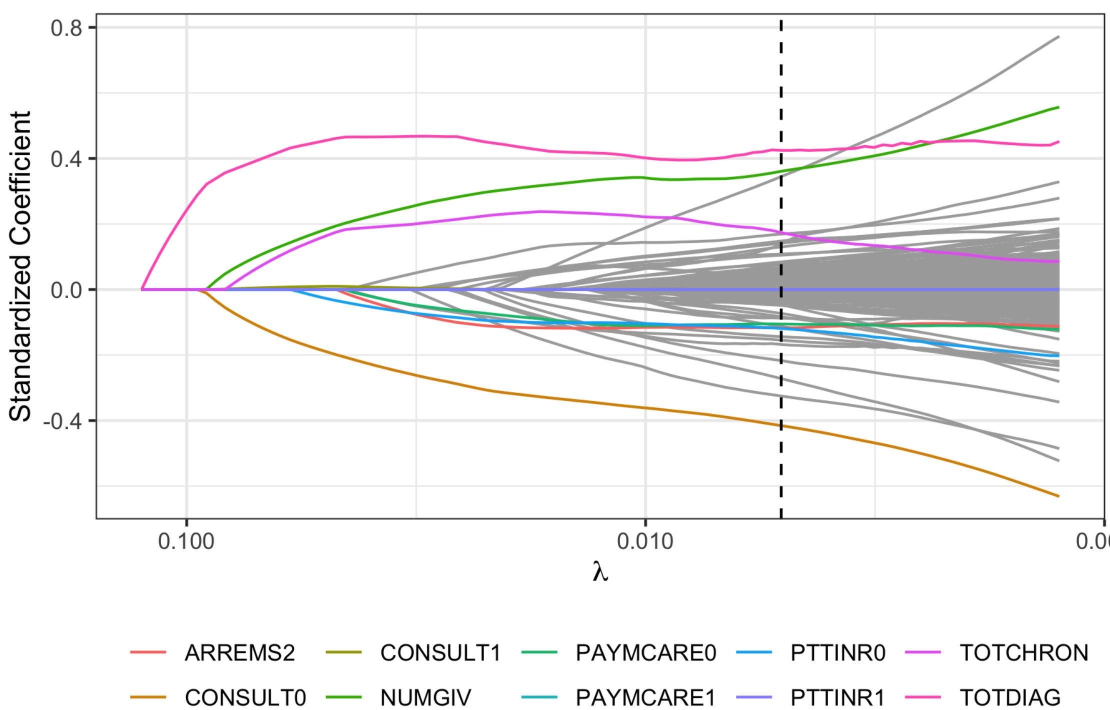

Figure 5: Lasso trace plot.

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

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%.

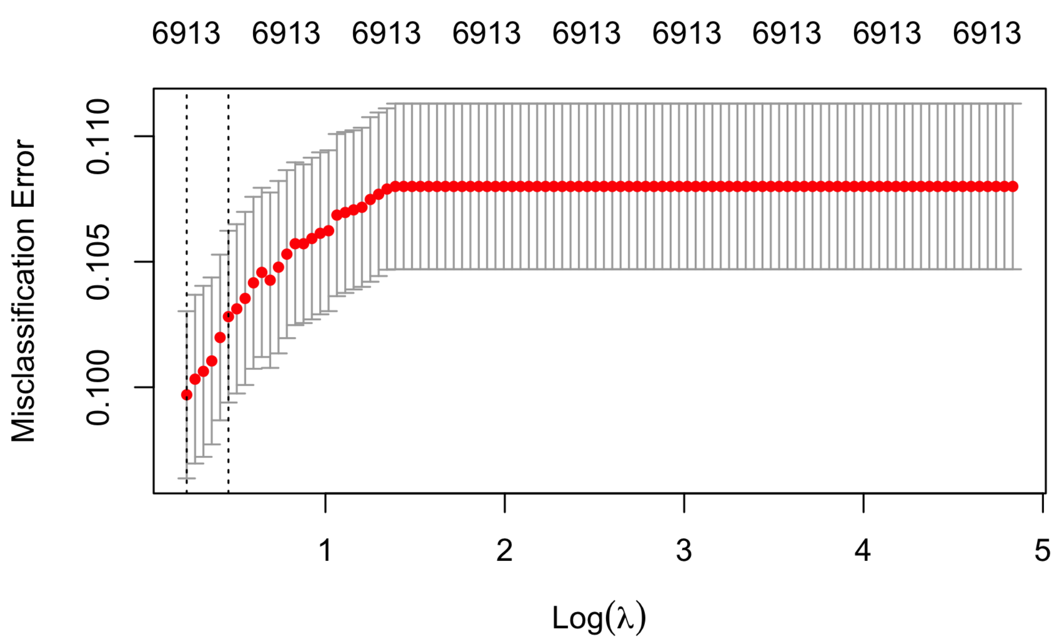

Figure 6: Ridge CV plot.

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 8 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%

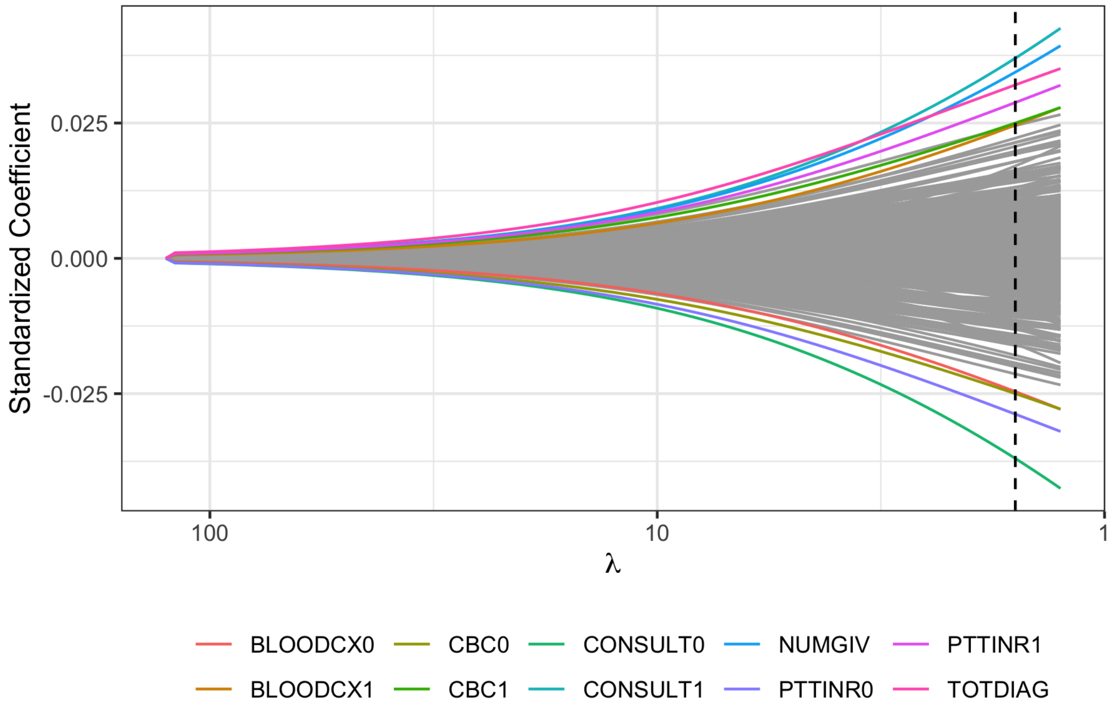

Figure 7: Ridge trace plot.

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

missclassification error of the lasso classifier is the lowest one out of the regression based methods, this suggests that few features have large effects.

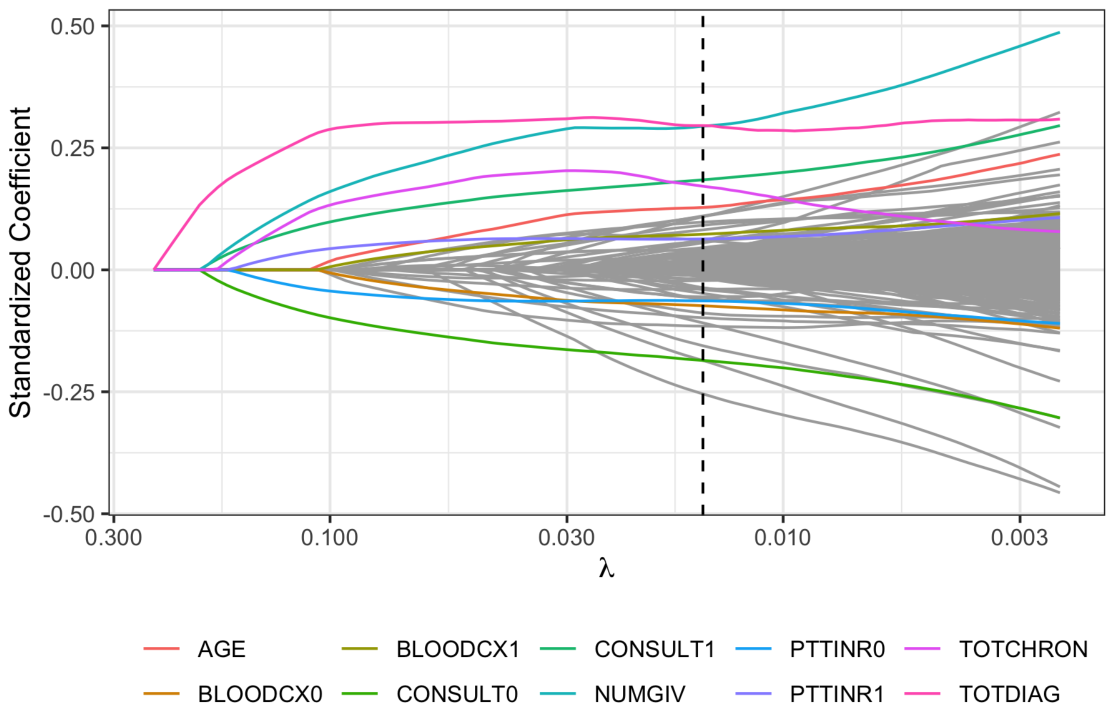

Figure 8: Elastic Net trace plot.

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 9, "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%.

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 10 shows that the optimal number of terminal nodes is 3.

4.2.3 Random Forest

The next classifier that we used was a random forest. The OBB error for the random forest can be found in Figure 11, the OBB errors for varying mtry or number of features to consider at each split point can be found in Figure 12 and the variable importance plot can be found in Figure 13.

We were able to achieve a misclassification rate of 7.9% which is a significant improvement from previous models.

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

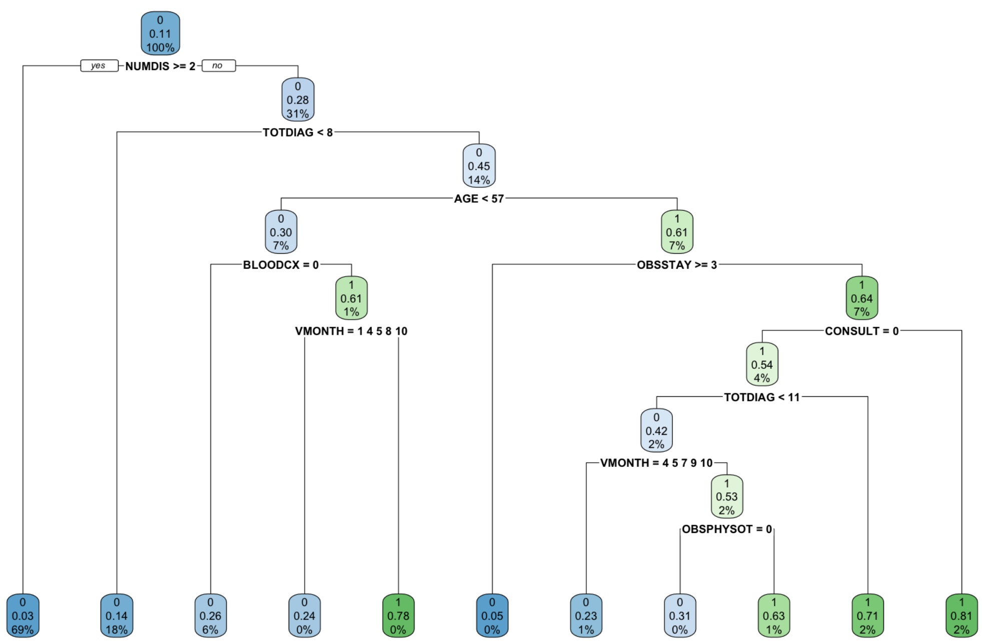

Figure 9: Unpruned Tree Plot.

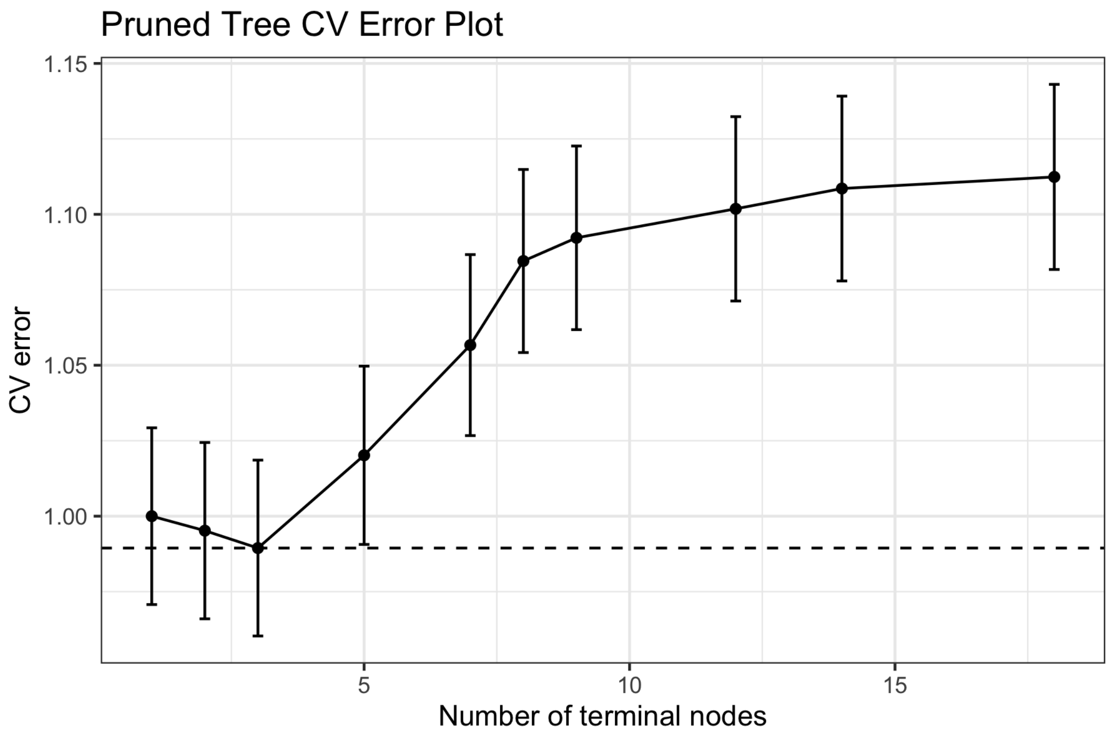

Figure 10: Pruned Tree CV Error Plot.

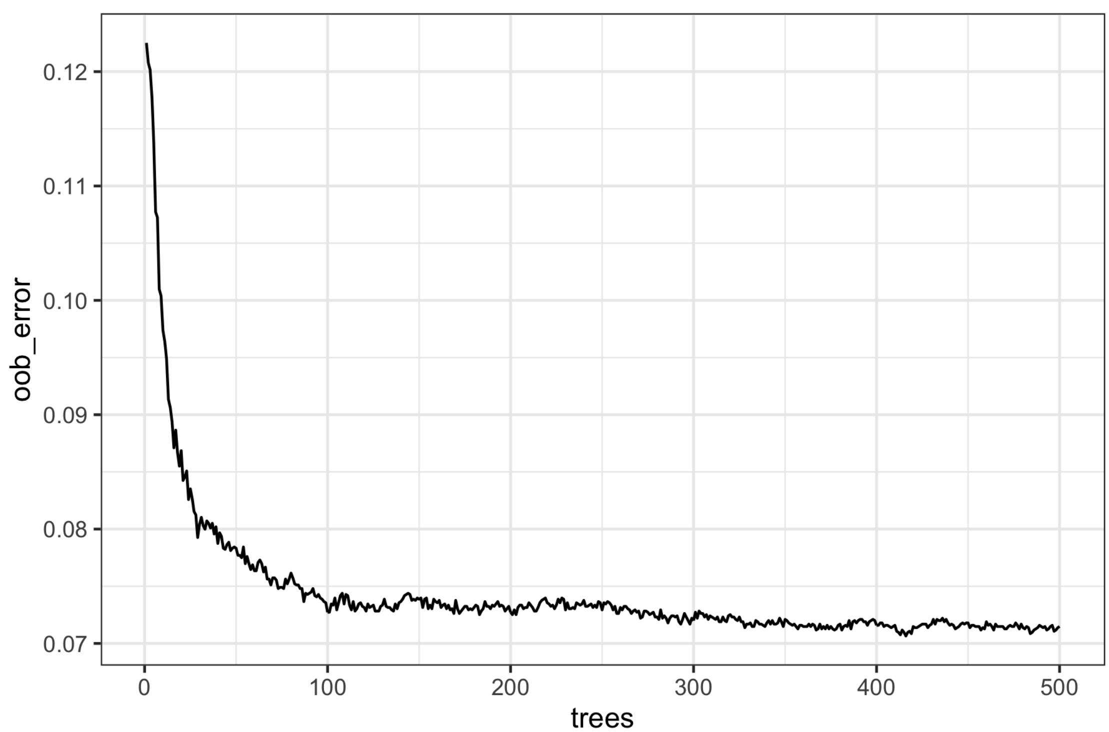

Figure 11: Random forest OBB Error Plot.

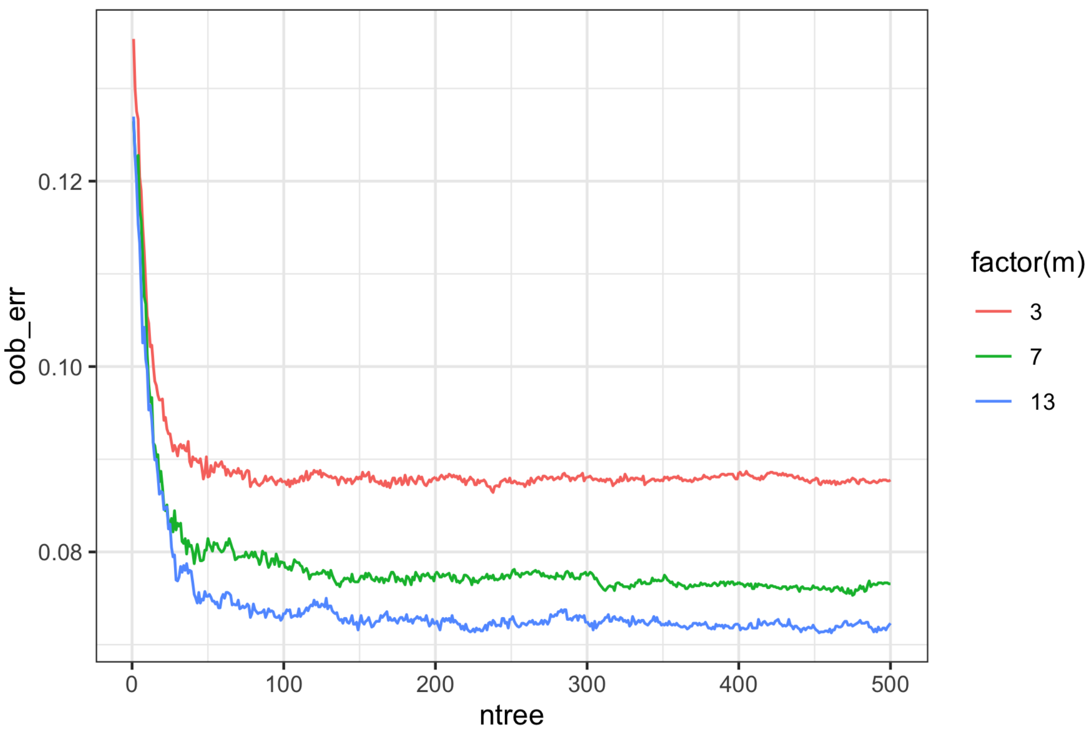

Figure 12: Random forest OBB Error Plot, varying mtry.

4.2.4 Boosting

Finally, we used boosted decision trees to predict our response variable. Figure 14 shows the CV errors for different tree depths, helping us to tune this hyperparameter. Based on this plot, we determined that the optimal model had depth of 2. Using that model we created predictions and were able to achieve a misclassification error of 9.9%.

Table 5 shows the different relative importances of the variables. Figure 15 shows the partial dependence plot for age and Figure 16 shows the partial dependence plot for total number of diagnostic services ordered or provided.

5 Conclusions

5.1 Method comparison

Table 6 shows the misclassification for all the methods considered. The random forest classifier has the best performance with a misclassification error of 8%, followed by logistic regression and lasso which each have a roughly 9% misclassification error. Elastic Net and Boosting have a 10% misclassification error each, followed by ridge and pruned trees which perform worst with misclassification errors of 11%.

Regardless of these differences in the misclassification error, 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 logistic regression model: CONSULT, TOTDIAG and NUMGIV. The random forest and boosting models both include AGE, CONSULT, TOTDIAG in the top 10 most important variables, as measured by their contributions to node purity.

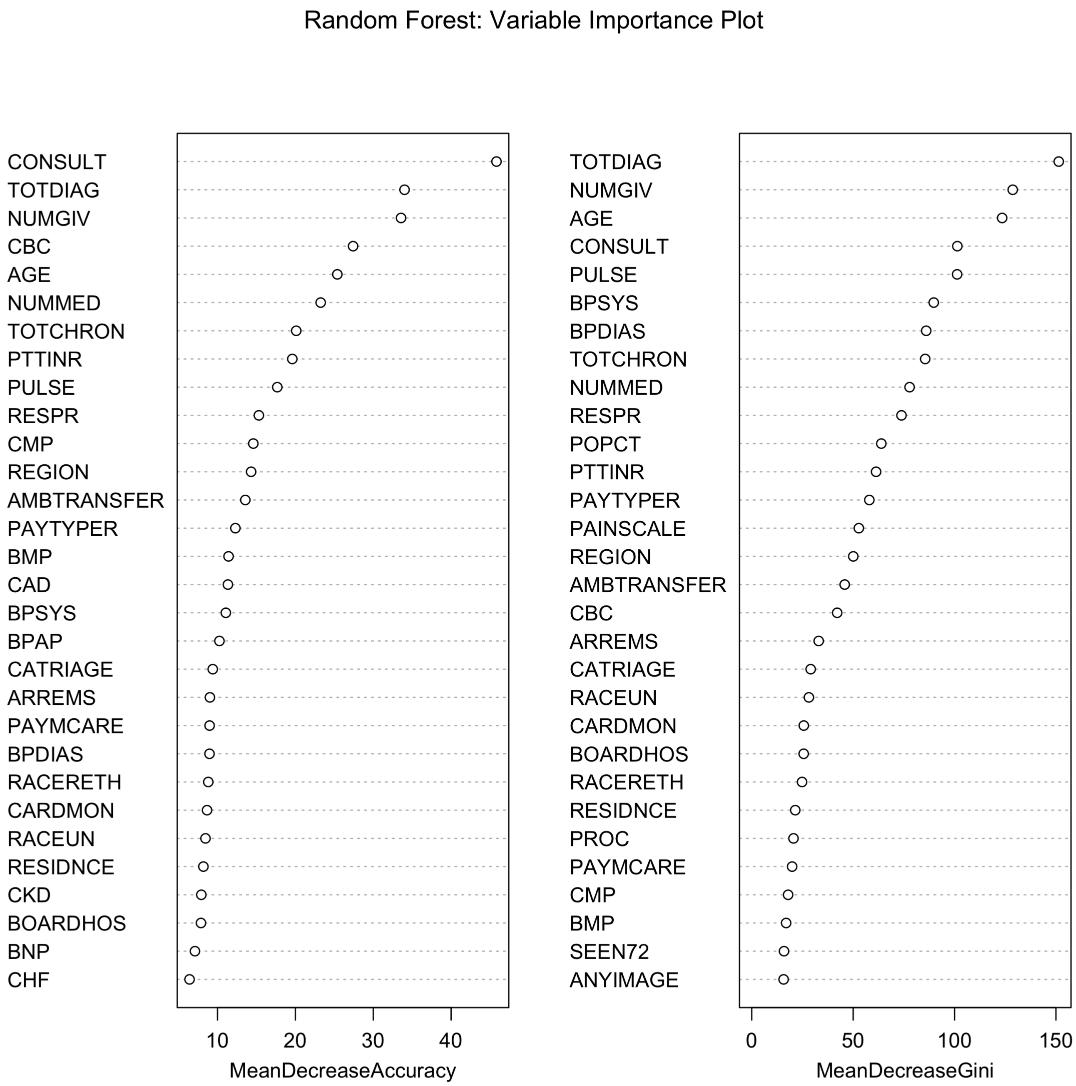

Figure 13: Random forest Variable Importance Plot.

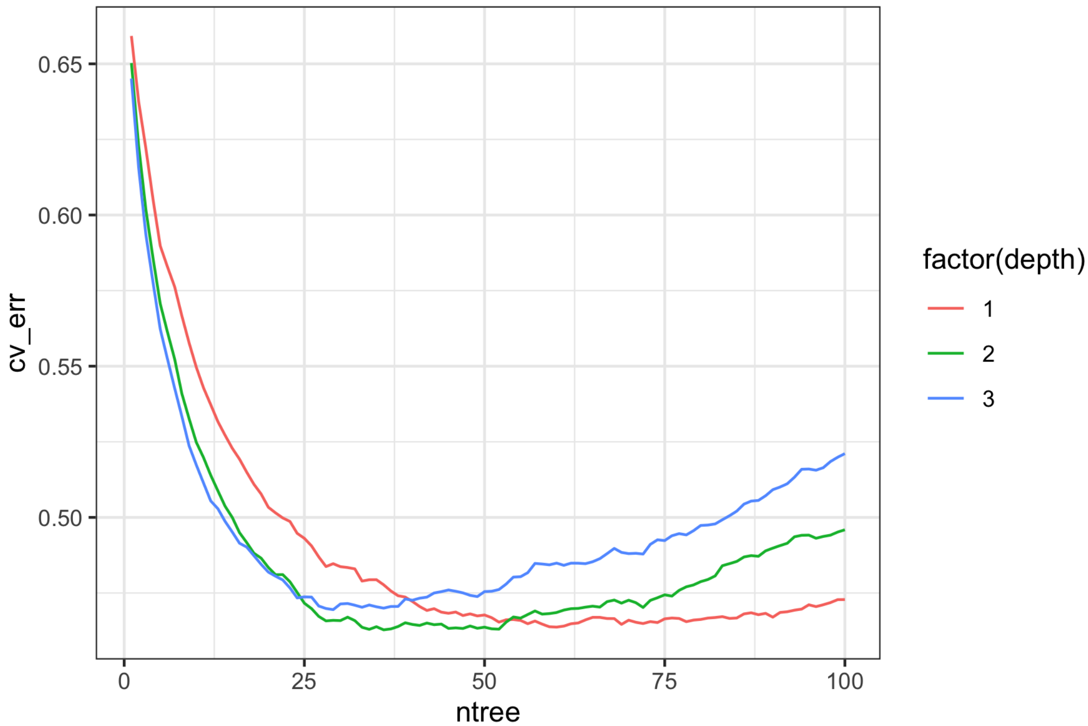

Figure 14: Boosting CV Errors for varying tree depths.

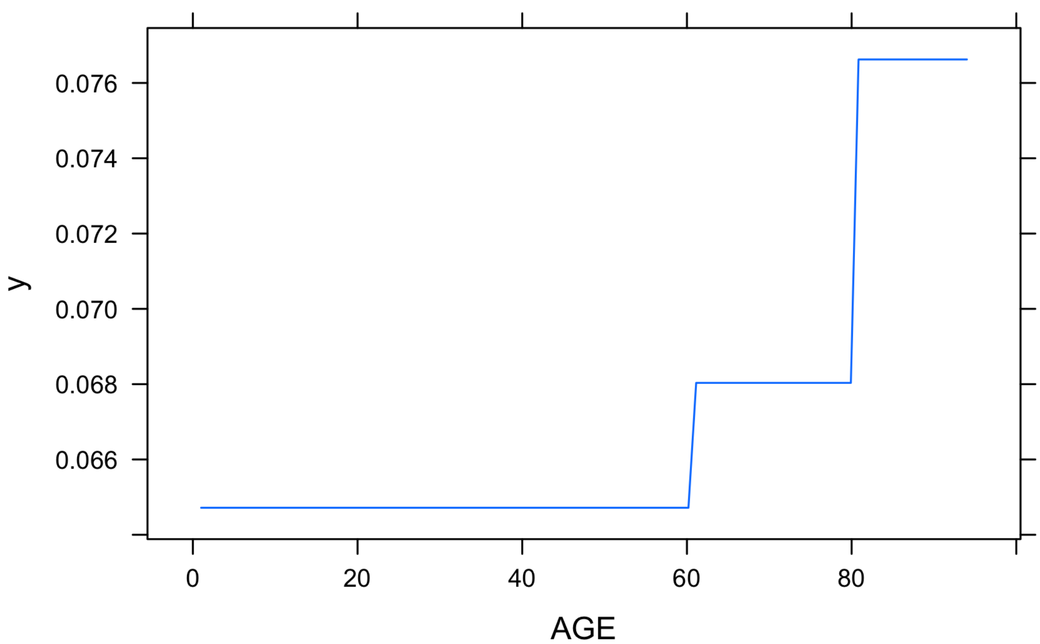

Figure 15: Partial dependence plot for age.

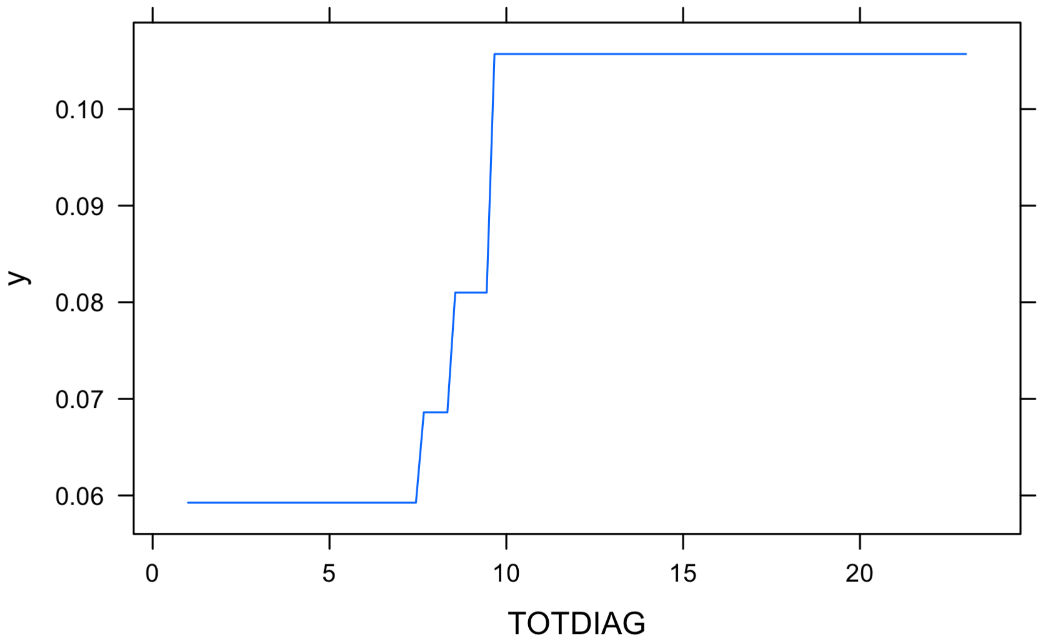

Figure 16: Partial dependence plot for total number of diagnostic services ordered or provided

Table 5: Relative importance of variables in boosted model.

Variable	Relative Importance
RFV13D	24.89
NUMDIS	16.37
RFV2	16.08
TOTDIAG	10.10
RFV3	9.52
CONSULT	7.89
NUMGIV	7.02
PTTINR	2.67
AGE	2.45
TOTCHRON	1.72
NUMMED	0.69
BLOODCX	0.59

Table 6: Misclassification errors for regression and tree based methods.

Model	Misclassification Error
Logistic Regression	0.09
Lasso	0.09
Ridge	0.11
Elastic	0.10
Unpruned Tree	0.11
Pruned Tree	0.11
Random Forest	0.08
Boosting	0.10

5.2 Takeaways

Among all initial patient characteristics, blood diagnostic tests, such as the complete blood count (CBC), the blood clotting test (PTTINR), and blood cultures (BLOODCX), seem to be the most indicative signs of whether a patient will be admitted to the hospital. This makes sense since our blood carries toxins, bacteria, and indicators of cancers. The other main features that were consistently important across multiple models indicated that sicker patients were more likely to be admitted to the hospital, which we expected due to common sense. For example, a visit from a consulting doctor (`CONSULT`) indicates that the patient has a specialty ailment that needs greater investigation, which typically indicates greater severity in their condition. Altogether, none of the coefficients that were indicated as important were surprising. This explorative research could aid future studies, but not a lot of new information was found. It is furthermore not surprising that the random forest performed the best.

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 visualization, some of our features included more than 1054 factors levels, for example features related to diagnosis codes or medicine names, and we were forced to exclude those variables from our training dataset for the tree visualization, since the function will not allow more than 1054 factor levels as an input. Because of this, we lost valuable visual insight but were still able to train the underlying model and receive a true misclassifiaction error that was calculated with the inclusion of all of our features. We are additionally missing the features excluded during the cleaning stage of our analysis and we cannot be sure that these features do not also exert influence of our response variable, so this is a weakness in our analysis. Lastly, the exclusion of survey weighting during our exploration of the response variable distribution and characterization may cause a misrepresentation of the data and should be looked into with further analysis.

5.4 Follow-ups

The inner machinations of any emergency departments are quite complex and our findings in this paper only graze the surface of the beginnings of a potential flow optimization. We saw trends pointing towards the importance of blood cultures, and a next step would be to gain more information about which specific blood tests are most likely to cause a patient to be admitted into the hospital. Another next step would be to find a way to keep the variables that we excluded during the cleaning stage to see if these features contribute to probability of hospital admittance. We would also like to update these models after data from 2020 and 2021 are released, to see if and how the COVID-19 pandemic has influenced the characteristics of who presents to the emergency departments and if the same features were relevant during such an unusual time.

A Appendix: Descriptions of features

Below are the 155 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 - Continue to admit elective surgery cases when on ambulance diversion (`ADMDIV`) - 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`)