Final project

Due April 30, 2021

Introduction

The National Impatient Sample (NIS) data, collected by the Healthcare Cost and Utilization Project (HCUP), is the largest publicly available dataset that contains information on inpatient healthcare in hospitals throughout the United States. The NIS is used by policymakers and health officials to make national estimates of healthcare utilization, and observe key features of inpatient care. The NIS was first started in 1998 by the Healthcare Cost and Utilization Project, and contains information such as patient demographics, classification of diseases, total hospital bill, length of stay, and many other features that characterize hospital care. The goal of this assignment will be to build a model to predict impatient mortality an determine what factors contribute a increased risk of death during hospitalization.

The data that will be used in this assignment consists of a random subset of 200,000 patients from the 2012 National Impatient Sample. The data was taken from the Healthcare Cost and Ultilization Project (HCUP), which is the largest collection of hospital care data in the United States. The data was taken from discharge records from all hospitals that are participating with the HCUP, and use state guidelines to help identify the hospitals that qualify for the data collection process. 47 states and the District of Columbia participate in the NIS, and data is available for hospitals in those states. The outcome of interest is the inpatient mortality, of whether the patient died during the period of hospitalization. Features such as patient demographic, severity of disease, risk of mortality, and comorbidities were incorpated to determine if a patient was likely to die during hospitalization. This can be used to identify features that increase the risk of patient mortality in hospitals and seek to prevent such deaths in the future.

1. Describe the problem explaining in particular why prediction is of primary interest (inference could also be of interest but there has to be a good reason for wanting to predict a particular outcome)
2. Describe the data (e.g. data source, data collection, outcome of interest, available features, sample size, missing data, etc.)

Methods

First the relevant features to the outcome of interest was sorted out from the 175 original features that were present.

Then then data was then reevaluated and factors were added when necessary.

1. Describe any data pre-processing steps (e.g. cleaning, recoding, variable transformation, dealing with missing data, selection of features to be included in your models, etc)

Out of the 175 possible features that were present in the original dataset, only 44 variables were selected to be included in analysis and model building. These 44 include data regarding patient demographics (age, race, gender), comorbidities (such as alcohol abuse and COPD), and the risks of patient mortality. Each variable was examined and was made into factor variables as was appropriate. A majority of the features were converted into dummy variables, however some remained as strings and integers. In examining the missing data, there was less than 1% of the total sample size that was missing from the target variable, whether the patient died. Because the sample was small compared to the dataset, the missing values of the target variable were removed before the analysis.

2. Briefly describe the Machine learning methods you will be using and why they are appropriate for your data (e.g. given the sample size and dimensionality of your training data, are you more concerned about bias or variance?) You should try and compare at least 3 distinct appropriate methods.

3. Describe how you are splitting the data into testing and training and any resampling strategy used for comparing methods, tuning parameters, and/or model/feature selection.

Logistic Regression

I will be comparing 3 different methods to build a predictive model for patient mortality. The first will be logistic regression model. The logistic regression model is one of the most commonly used and basic binary classifiers. Because the desired goal is to determine if a patient died during their hospitalization, the outcome is a binary outcome. Given the extremely large sample size of the data with around 200,000 observations, both the training and testing sets will be large enough to ensure an accurate prediction model.

Forward selection was used to determine the features that will be included in the logistic regression model. According to the forward selection process, only the **APRDRG_Risk_Mortality**, a factor variable that characterizes the risk of patient mortality, was determined to be significant in the data. However, the race variable was also included to determine the effect of patient demographics on mortality. There will only be a couple of features included in the actual logistic prediction model, therefore the model will be a simpler one indicating that the model will have a higher bias. However, the large sample size of the data and the use of cross validation will be used to determine the accuracy of the results. I will be using the misclassification error and the AUC as performance metrics to determine the effectiveness of the logistic model and compare it to other models that I wiil be using.

K-fold cross validation will be used to reduce the error that comes from different training/testing splits. The resulting misclassification error from the k-fold cross validation will be compared with the misclassification error from the initial training/testing split.

Balanced Random Forests

The data itself is very unbalanced, with 184,598 patients that were successfully discharged compared to the 3,412 that died in the hospital. This could lead to an optimistically low misclassification error. Therefore, balanced random forests will be used to help balance the two binary outcomes and correct over optomistic misclassification errors.

I will be using all of the 44 variables that were selected form the original dataset in the balanced random forest model. All of the variables are included because the random forest model will tune the parameters and adjust the model according to which variables are considered important. The variable importance plot generated from the balanced random forest model will be compared to the variables that were considered important in the forward selection algorithm used in the logistic regression. The AUC will be used as a performance metric as the goal of the model is to correctly predict patient mortality.

Lasso Regression

Lasso regression will also be used to build a predictive model for determining patient mortality. Lasso regression was chosen over ridge regression because it is likely that only a small number of predictors will be significant in determining patient mortality than the all of the features that we have available. The regression model itself will chose which variables are important and act similarly to a feature selection algorithm. Therefore, we will continue to increase the tuning parameter to determine which parameters are important in determining patient mortality.

Results

1. Present key summaries (table and/or plots, but plots preferred when both available) of your data (e.g. class frequencies if a classification problem)
2. Report training, validation/cross-validation, and test errors. Present cross-validation plots for tuning parameters if available. Report variable importance (e.g. p-values, model coefficients, Random forest and boosting variable importance).

Logistic Regression:

Below is the AUC and associated 95% Confidence Interval for the logistic regression on the training data.

```
## Area under the curve: 0.9362
```

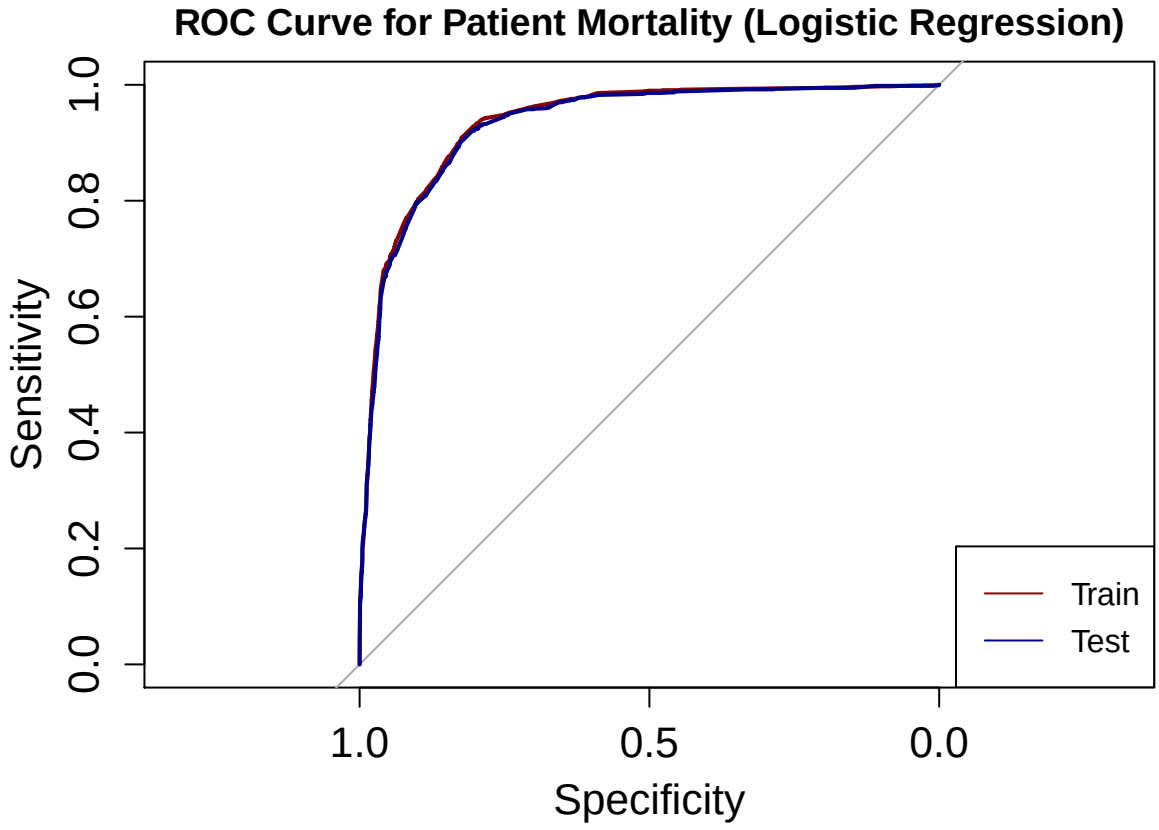
```
## 95% CI: 0.9317-0.9408 (DeLong)
```

Below is the AUC and associated 95% confidence interval for the logistic regression on the testing data.

```
## Area under the curve: 0.9324
```

```
## 95% CI: 0.9257-0.9391 (DeLong)
```

The plot below shows the ROC curve for both the training and testing data for the logistic regression model on patient mortality.



The table below shows the parameter estimates and the associated p-values for the logistic regression. The risk of mortality variable is statistically significant across almost all of its factor levels, and the length of stay variable is also statistically significant. The race variable is statistically significant only if the individual is Black. However, the parameter estimates for Race and Length of Stay are not very large compared to that of the Risk Mortality, indicating that the Risk Mortality has the highest influence in determining the probability of patient mortality.

```
##                                            Estimate     Pr(>|z|)
## (Intercept)                              -2.715474760 9.069181e-11
## APRDRG_Risk_MortalityMinor Likelihood    -4.212201722 5.005929e-21
## APRDRG_Risk_MortalityModerate Likelihood -1.602078731 1.622046e-04
## APRDRG_Risk_MortalityMajor Likelihood     0.209901893 6.175816e-01
## APRDRG_Risk_MortalityExtreme Likelihood   2.347800631 2.188817e-08
## RACEBlack                                -0.266228485 1.814111e-04
## RACEHispanic                             -0.087103289 3.181046e-01
## RACEAsian                                 0.006590652 9.662414e-01
## RACENative American                      -0.269350057 4.107313e-01
## RACEOther                                -0.030889503 8.115125e-01
## LOS                                      -0.009580439 2.065340e-79
```

Balanced Random Forests:

```
## Area under the curve: 0.9492
```

```
## 95% CI: 0.9457-0.9527 (DeLong)
```

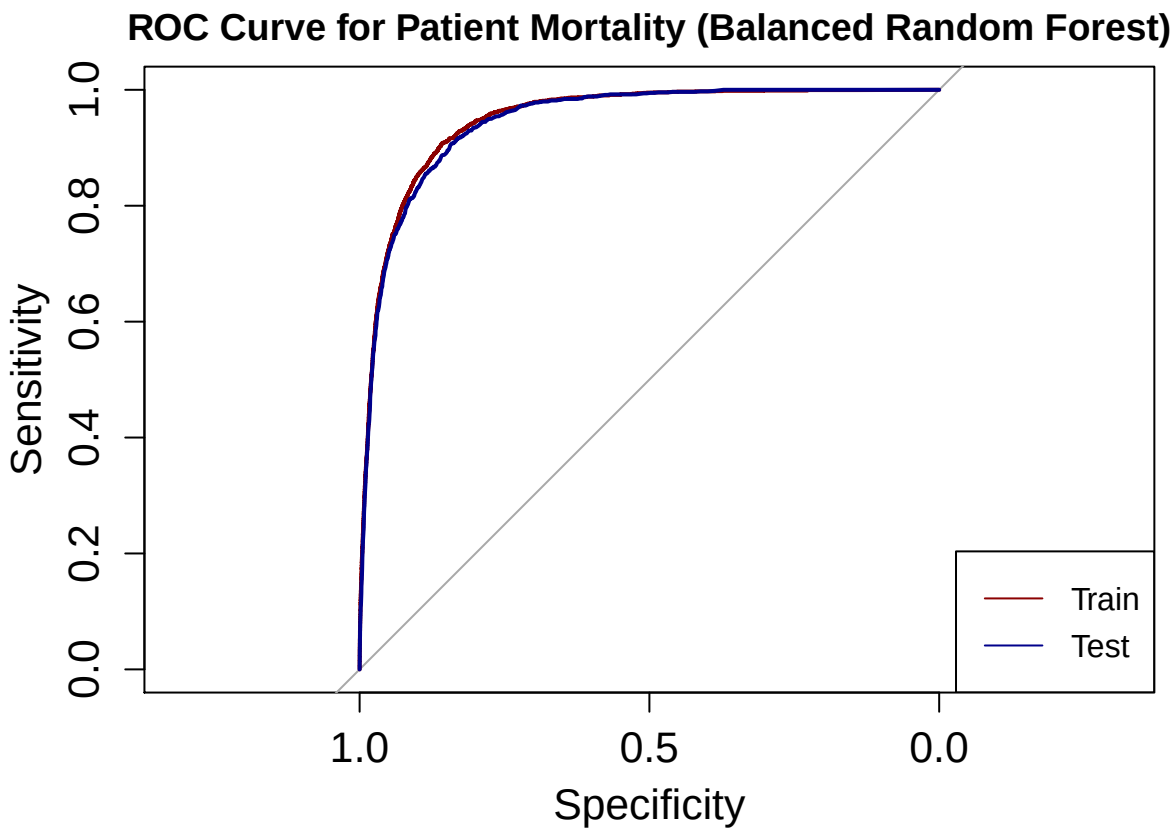
```
auc(rf_roc_test)
```

```
## Area under the curve: 0.9459
```

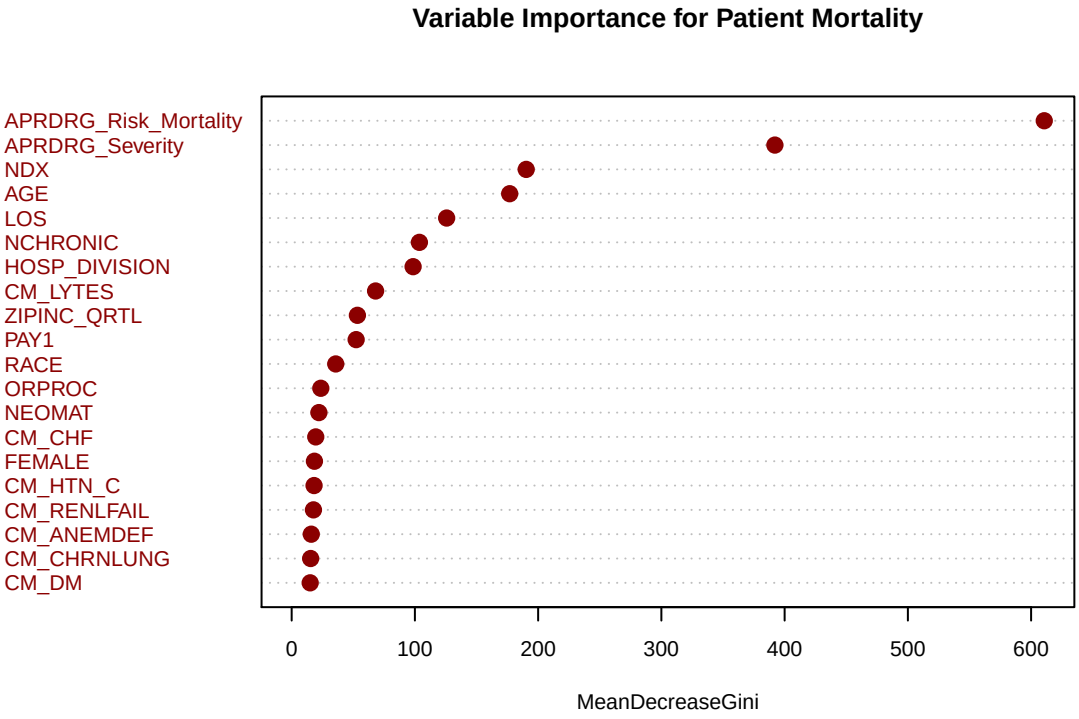
```
ci(rf_roc_test)
```

```
## 95% CI: 0.9409-0.9509 (DeLong)
```

The plot below shows the ROC curve for both the training and testing data for the balanced random forest method in predicting patient mortality.



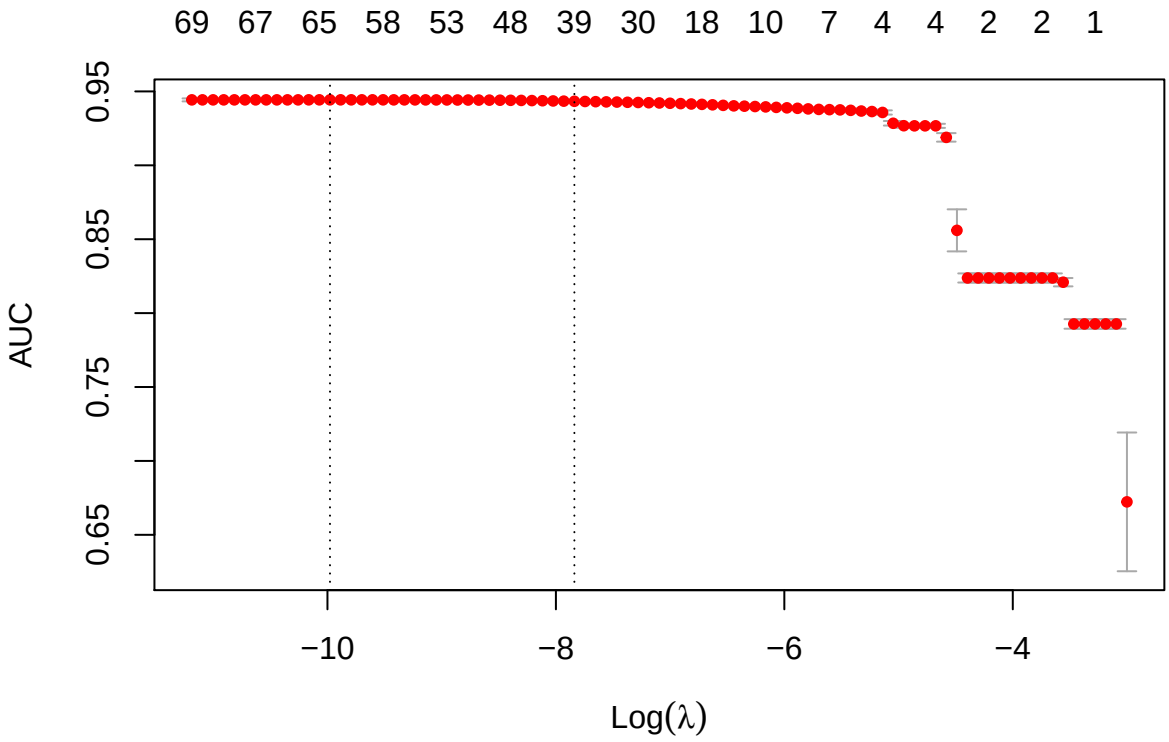
Below is the variable importance in predicting Patient Mortality using the balanced random forests method. This indicates that the Risk Mortality and the Severity of the disease are the most important variables in predicting patient mortality. Other significant variables include age and the number of diagnoses coded on the patient's health record.



Lasso Regression

```
## [1] 0.9443041
```

```
##       mmce
## 0.01800968
```



4. If applicabble, describe any model/feature selection used.
5. If applicabble, describe any tuning parameters and how you will be tuning them.
6. Describe what performance metric(s) you will be using and why.

Conclusions/discussion

Discuss whether and why the prediction model(s) developed achieved sufficient high accuracy to be usefully deployed to predict new observations.

#Additional notes for those using the NIS data The data provided consists of a random subset of 200,000 patients from 2012 from the National Impatient Sample (NIS) data collected by the Healthcare Cost and Utilization Project (HCUP). You can find information on the HCUP database at https://www.hcup-us.ahrq.gov. You can choose to develop a model to predict death during hospitalization also known as impatient mortality (variable DIED in the dataset) or hospital length of stay (variable LOS in the dataset). For extra credit, you can also choose to predict both. The dataset has a relatively large number of variables. In the provided data dictionary I preselected variables (highlighted) which are both availabe (not all variables in the dictionary are available for 2012) which might be relevant for predicting impatient mortality and/or hospital length of stay. Based on their description and additional info from the HCUP site you should choose which variables among the preselected ones you will consider as features/predictors. You don't have to use them all. There maybe variables that are redundant (capture pretty much the same info others already capture), variables that are too complex (e.g. categorical with way too many levels), or that based on your judgment are unlikely to be important. Be aware that the data is real and has not been pre-processed in any way and you will have to do some data cleaning. For example, you should carefully check the variables you consider as possible predictors for correctness of type (e.g. many numeric variables will be read in as factor variables when you use `read.csv`), outliers, missing observations, nonsensical values, etc.