MTH522 Project2: Forecasting Passenger

Utilizing Supervised Machine Learning for Survivability

College of Engineering
University of Massachusetts Dartmouth
Dartmouth, United States of America

By

Jeevan Kumar Banoth

Hanuma Venkata Vijay Kamal Konduri

Vikas Reddy Kothapalli

Sai Charan Pulugam

Abstract:

ML approaches are all the rage these days, and for good reason—they can be used to identify patterns and solve several (previously) inexplicable issues. Supervised and Unsupervised are the two basic categories into which ML techniques can be separated. This research's objective is to estimate the survivorship status of the Titanic passengers using supervised techniques, specifically Logistic Regression (LR) and Classification Trees (CTs). First and foremost, the correctness of the approaches will be evaluated. Furthermore, we'll be examining metrics for LR such as Specificity, Sensitivity, and the like, as well as the accuracy of the CTs that have been and have not been pruned.

Keywords:

Prune, Accuracy, Specificity, Sensitivity, Identification Trees, Logistic Regression, Supervised Learning, Unsupervised Learning, Machine Learning (ML).

I. MOTIVATION

The following goals are intended to be accomplished by this project:

a) The aim is to enhance comprehension of the machine learning topics covered in MTH 522.

b) Gaining and refining expertise in data wrangling, with an emphasis on feature engineering, will help you either make the data more analyzed or add new variables.

c) Learn about the submission formats accepted by Kaggle and adjust the final product accordingly.

II. INTRODUCTION – DATA SOURCE

A. Collection of data

The Titanic dataset from Kaggle will be used for this project. Participating in the challenge to forecast the survivorship status of Titanic passengers after their ship capsizes is the dataset. The data will be used for my study; it was pre-split into sets for the competition, Training (891 records) and Testing (418 records).

B. Specifications for the Dataset

Along with some extra comments and keys, the variable definition is provided in the table below:

Variable	Definition	Key
survival	Survival Status	0 = Died, 1 = Survived
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Gender of the passenger	
Age	Age in years	
sibsp	# of siblings/spouses aboard the Titanic	
Parch	# of parents/children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

TABLE 1: DATASET SPECIFICATIONS

C. Important Information

As stated in II-A, the entire dataset has been divided into Train and Test sets. However, the objective variable—survival—is absent from the test set. This is done by Kaggle to compare participant contributions to the actual data. Therefore, for most of this project, I will make a clone of the original Test set and add a pseudo-random Survival field for my analysis.

III. Approach

Let's define what a supervised machine learning approach is in brief before moving on. Properly labeled datasets are used in supervised machine learning models to train algorithms for data classification or precise outcome prediction. The cross-validation procedure modifies the model's weight when data is added to it. Several real-world issues, such as classifying emails in an inbox as spam, are solved at scale by supervised machine learning techniques.

i) Constraints and Acronyms
ii) For the sake of brevity, the following acronyms will be used often throughout the document:
iii) Machine Learning (ML)
iv) Classification Trees -> CTs
v) Logistic Regression (LR)
vi) Ordinary Least-Squares (OLS)
vii) Classification and Regression Trees (CART -> These can be used in place of (ii))

Methods:

Classification trees and logistic regression will be used separately by me to categorize and forecast which passengers survived the Titanic shipwreck. All the models will be run on the Training set first, and then, for the final classification, on the Test set, like with Supervised ML. We will next assess the accuracies of the test set and the train. At the same time, the CTs will examine the performance of the trees, both trimmed and unpruned.

Regression using Logistic Regression:

It is mostly used for modeling the conditional probability of a binary outcome variable and is also referred to as logit regression. Utilizing a non-linear link function, LR limits the fitted values to a range of 0. and 1. A linear combination of predictors is the model used to represent the log chances shown on the S-Curve.

An LR can be executed in R using the glm() command. To be modeled as a dichotomous outcome variable, the command's "family" option needs to be set to "binomial."

Mathematically,

$$\log(P(\text{survived}_i = 1)/1-P(\text{survived}_i = 1))=\beta_0+\beta_1 * X_1+\beta_2 * X_2 +\ldots\ldots.+\beta_i * X_n + \varepsilon_i \qquad - (1)$$

In Eq(1), X_1, X_2, …… X_n -> The independent variables on which the dependent variable "survived" will be modeled

ε_i -> The error/residual term

Before we continue, let us clarify that the target/dependent variable "survived" is absent from the original Test dataset (supplied by Kaggle), as stated in II-C. Therefore, I created a clone of the Test dataset, dubbed "test_copy," which would be utilized for all R studies. I also added a dependent variable that is pseudo-randomly filled with discrete 0–1 values. To understand the randomization that was done in test_copy ->, please see the conditions that were executed.

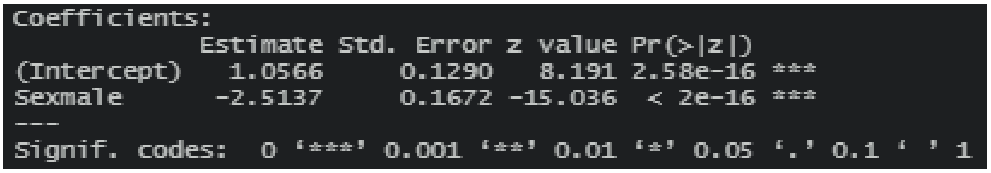
a.) *When sex = "male" then "survived" = 0*
b.) *When pclass != 1 then "survived" = 0*
c.) *When sex = "female" then "survived" = 1*

NOTE: The Specificity evaluation metric will show bias due to the above randomization in the "test_copy" dataset. It is crucial to remember that this type of exercise is not performed on the training set. As a result, although the Test dataset, which is used to make the prediction, is biased, the model is trained on an impartial dataset.

To examine the likelihood of surviving the shipwreck by gender, let's run a univariate model. Eq(1) will become the following for this project ->

$\log(P(\text{survived}_i = 1)/1 - P(\text{survived}_i = 1)) = \beta_0 + \beta_1 * \text{Sex} + \varepsilon_i$ - (2)

After running Eq(2) using the glm() function on R, we get the following result:



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.0566     0.1290   8.191 2.58e-16 ***
Sexmale       -2.5137     0.1672 -15.036  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Substituting the above values in Eq(2), we get the following equation ->

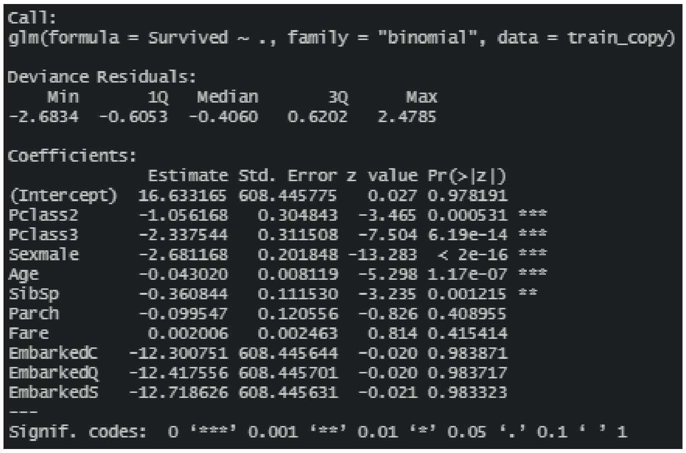
$\log(P(\text{survived}_i = 1)/1 - P(\text{survived}_i = 1)) = 1.05 + (-2.51) * \text{Sex} = \text{Male} + \varepsilon_i$ - (3)

The above equation shows that a Male passenger is more likely to die in this shipwreck by a factor of 2.51. The results of Eq(3) are visualized in Fig-2 in IV-A.

Let us now look at the effect of all the remaining variables. We’ll run the glm() command again by including all the remaining variables. Mathematically,

$\log(P(\text{survived}_i = 1)/1 - P(\text{survived}_i = 1)) = \beta_0 + \beta_1 * \text{Pclass} + \beta_2 * \text{Sex} + \beta_3 * \text{Age} + \beta_4 * \text{SibSp} + \beta_5 * \text{Parch} + \beta_6 * \text{Fare} + \beta_7 * \text{Embarked} + \varepsilon_i$ - (4)

After running Eq(4) using the glm() function on R, we get the following result:

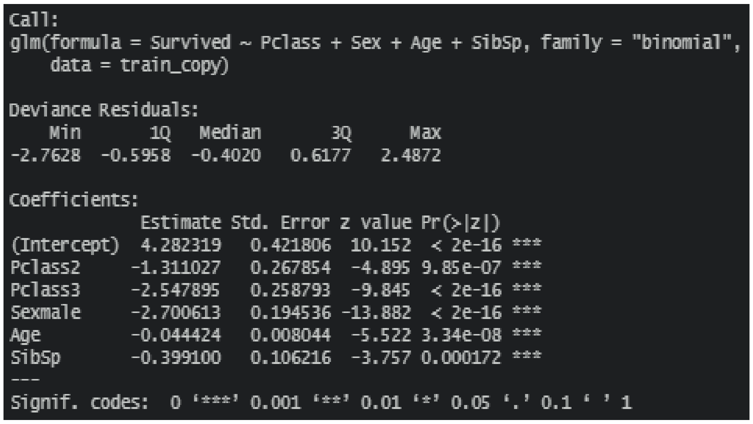


```
Call:
glm(formula = Survived ~ ., family = "binomial", data = train_copy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.6834  -0.6053  -0.4060   0.6202   2.4785  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  16.633165 608.445775   0.027 0.978191    
Pclass2      -1.056168   0.304843  -3.465 0.000531 ***
Pclass3      -2.337544   0.311508  -7.504 6.19e-14 ***
Sexmale      -2.681168   0.201848 -13.283  < 2e-16 ***
Age          -0.043020   0.008119  -5.298 1.17e-07 ***
SibSp        -0.360844   0.111530  -3.235 0.001215 ** 
Parch        -0.099547   0.120556  -0.826 0.408955    
Fare          0.002006   0.002463   0.814 0.415414    
EmbarkedC   -12.300751 608.445644  -0.020 0.983871    
EmbarkedQ   -12.417556 608.445701  -0.020 0.983717    
EmbarkedS   -12.718626 608.445631  -0.021 0.983323    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We observe that certain factors are insignificant, meaning they have no statistical significance in terms of capturing the variation in the data and have no discernible impact on the dependent variable. For instance, in this case, based on the statistics above, we can say that the chance of a passenger surviving was unaffected by the location of embarkation.

Now let's utilize the R "MASS" package's stepAIC approach to determine which independent variables are important. Because the AIC value penalizes utilizing extra variables, we can be certain that this will set us on the proper path.

When we apply the stepAIC approach to Eq(4) in R, we obtain the subsequent outcomes:



```
Call:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = "binomial",
    data = train_copy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7628  -0.5958  -0.4020   0.6177   2.4872

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.282319   0.421806  10.152  < 2e-16 ***
Pclass2     -1.311027   0.267854  -4.895 9.85e-07 ***
Pclass3     -2.547895   0.258793  -9.845  < 2e-16 ***
Sexmale     -2.700613   0.194536 -13.882  < 2e-16 ***
Age         -0.044424   0.008044  -5.522 3.34e-08 ***
SibSp       -0.399100   0.106216  -3.757 0.000172 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We see that the likelihood of surviving is influenced by age, parents/siblings, and socioeconomic class. In Section IV, we will examine the outcomes and the precision of our forecast.

Trees of Classification

Using CTs, records are categorized according to their likelihood as a binary outcome variable. The model will be identical to Eq. (4) mathematically. Based on the first run, the tree will be pruned if needed by looking at the matching "nsplit," which shows the ideal number of splits, and the minimal complexity parameter (CP).

Using R's rpart() command to run Eq(4) produces the classification tree shown below:

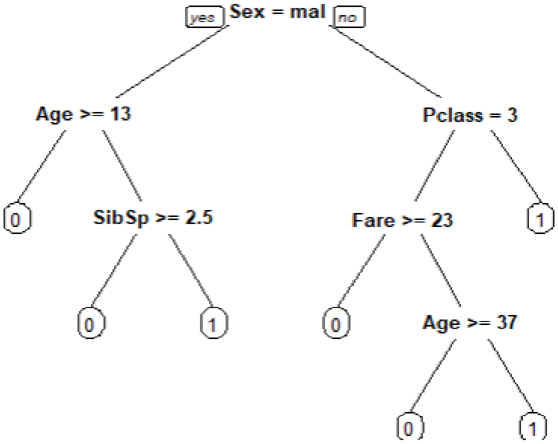
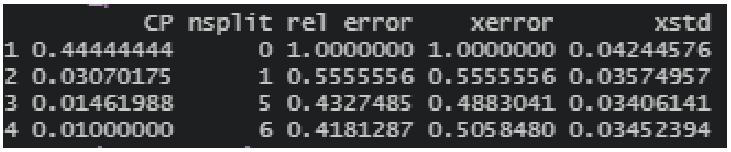


Fig-1: Classification Tree

The tree is seven feet tall with six splits. The outcomes appear to be comparable to LR. Nonetheless, it appears that "Fare" influences the likelihood of survival. Thus, based on the above tree, we may infer that, unless they are older than 37, travelers who purchased cheaper tickets—particularly those traveling in third class—generally have a poor chance of surviving.

To decide whether to further prune the tree, let's take a closer look at the complexity parameter. The following outcomes are obtained when we extract the CART model's complexity parameter:



```
          CP nsplit rel error    xerror       xstd
1 0.44444444      0 1.0000000 1.0000000 0.04244576
2 0.03070175      1 0.5555556 0.5555556 0.03574957
3 0.01461988      5 0.4327485 0.4883041 0.03406141
4 0.01000000      6 0.4181287 0.5058480 0.03452394
```

The above results show that a split of six has the lowest complexity parameter value and the lowest relative error. Hence, our original tree is the optimal tree that can be obtained. The examination of the model accuracy will be done in Section IV.

IV. EVALUATION & RESULTS

Logistic Regression

Let us look at the probability of survival basis Eq(2) in Section III-A. From Fig-2, Women had nearly three times more chance of surviving the shipwreck than men.

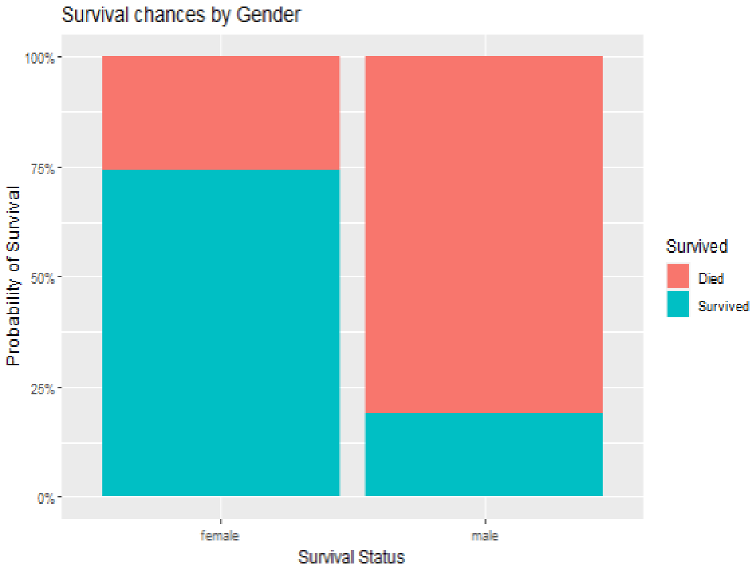


Figure – 2: Probability of Survival basis Gender

Plots such as the Receiver Operator Characteristic (ROC) curve help us to see the model's performance and provide a more thorough knowledge of how well it matches the empirical data. The ROC curve provides us with a decent idea of the model fit by illustrating the model's sensitivity and specificity. In a nutshell, sensitivity is the model's capacity to forecast a "positive" result when one really occurs. Comparably, specificity refers to the model's capacity to predict a "negative" effect in the event of a bad outcome. In my situation, a positive outcome will result from "survived" = 1, meaning the passenger lives, and a negative outcome from "survived" = 0, meaning the passenger passes away.

The ROC curves for the Training and Testing set are provided below:

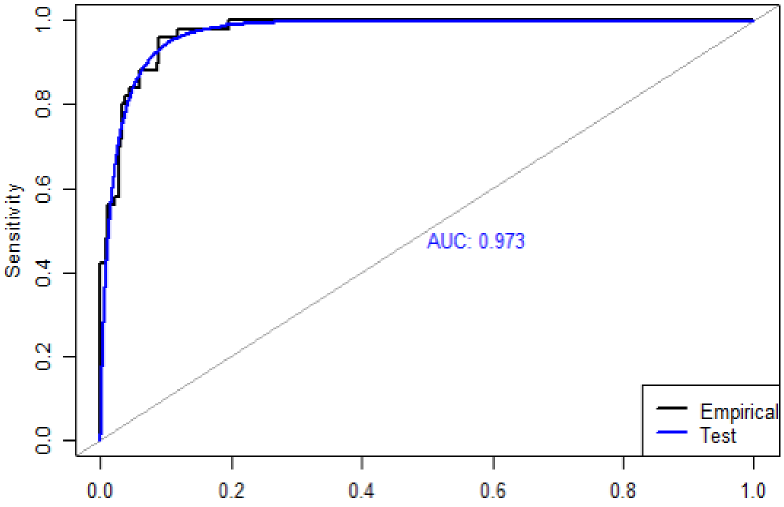


Fig -3: ROC curve for Test Set

The Area Under the Curve (AUC) from Fig. 3 is relatively high and extremely close to one. Accordingly, the model is performing well compared to the observed data. The model has a better match the closer the plot follows the leftmost corner. Let's examine the Training set ROC in a similar manner.

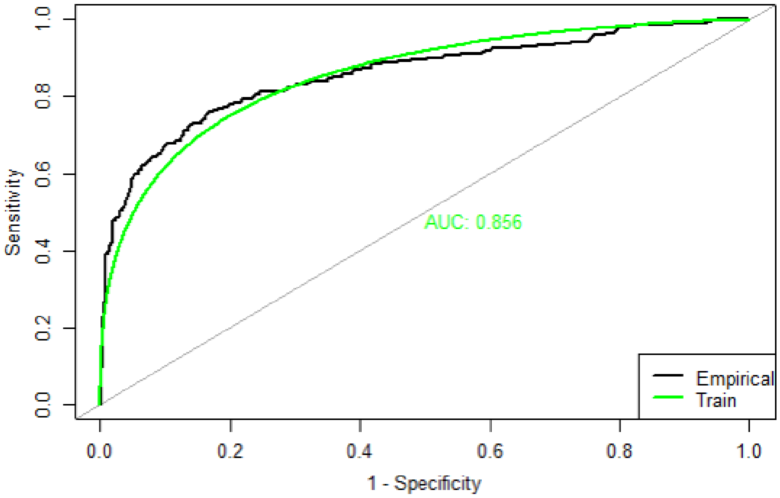


Fig 4: ROC curve of Training Set

Since LR works with a binary target variable, we must map the category probabilities to observe their distribution and force the result to be a binary of 0 or 1. The Sigmoid Function is represented by the S-Curve, which helps to visualize this. The following formula, when expressed mathematically, yields the Sigmoid function:

$$P(Yi) = 1/1+e^{-(b0+b1 * Xi)} \qquad - (5)$$

In our case, Yi -> Survived

$$x = -(b0 + b1 * Xi) \text{ will be Eq(4)}$$

The S-Curve plotting the probabilities of survival basis gender is given below:

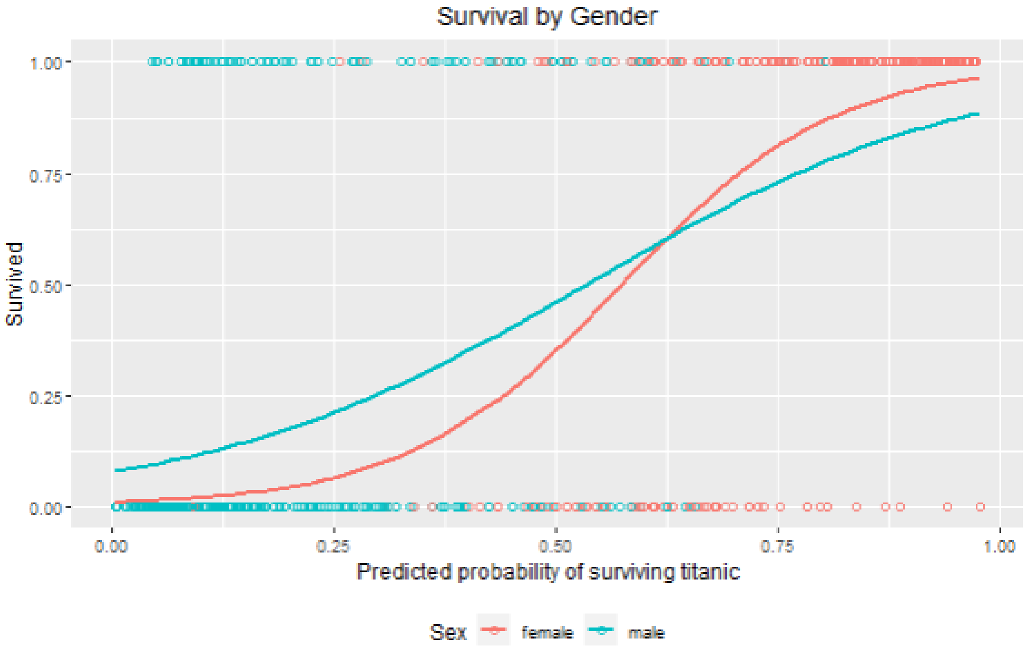


Fig 5: S-Curve

It is evident that Figs. 2 and 5 show the same results. As a result, we can declare that our model is coherent.

Trees of Classification

We need to see which depth has the least relative loss/error to determine the pruning of CRTs. R's matplot() function is used for this. The plot's result is displayed below:

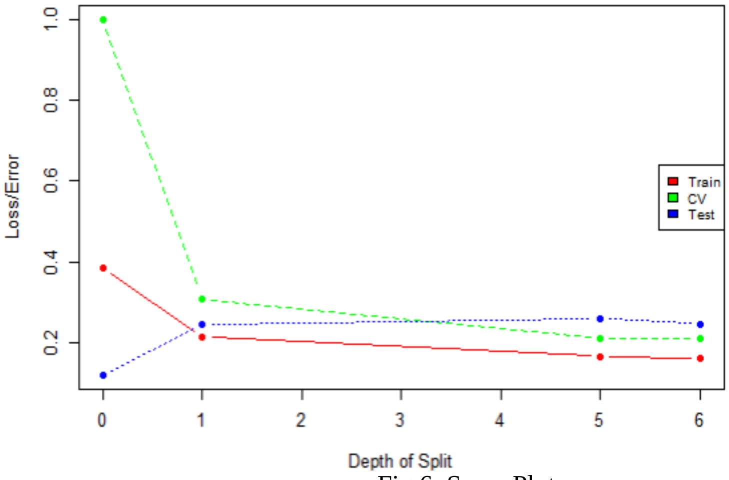
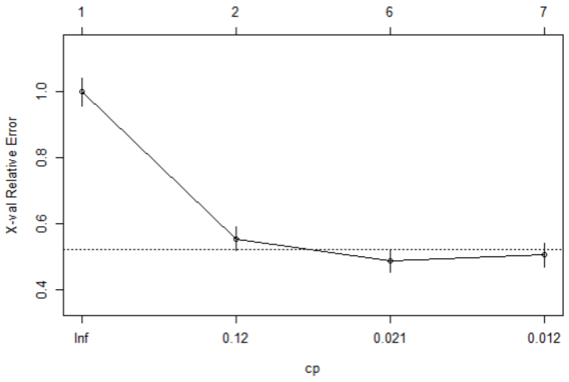


Fig 6: Scree Plot

Where the relative error of the training, testing, and cross-validated sets is lowest is indicated by the Scree plot. Usually, the "elbow point" is employed to calculate the ideal depth. As we can see, though, this runs counter to what we discovered in Section III B's CP Table. In addition, a split of one will not be able to forecast anything. I will therefore continue to use my original tree.

Similarly, we may use relative errors to determine the ideal split post-cross-validation. Below is the plot for the same:



Tables

Logistic Regression:

Metric	Testing	Training
Accuracy	0.746	0.805
Sensitivity	0.711	0.865
Specificity	1	0.710
PPV	1	0.827
NPV	0.320	0.766
AUC	0.973	0.856

TABLE:1 LOGISTIC REGRESSION EVALUATION

Classification Tree:

		Unpruned Tree	Pruned Tree
Tree Size		6	6
Accuracy	Train	0.839	0.839
	Test	0.753	0.753
	All (Cross-validated)	0.810	0.810

TABLE II: CLASSIFICATION TREE EVALUATION

RESULTS & FUTURE SCOPE

The Logistic and Classification Tree models function almost identically. However, as can be shown from the Test set accuracies in both caret and non-caret models, classification trees are doing marginally better for this dataset. More sophisticated techniques like XGBoost, Random Forest, and more relevant feature engineering can be used to enhance the model. The potential application of unsupervised techniques such as Support Vector Machines.

Even if a lot of the features I designed ended up being superfluous, I still had fun making them and am happy that I was able to.

REFERENCES

[1] *https://www.r-bloggers.com/2021/01/machine-learning-with-r-a-complete-guide-to-logistic-regression/*
[2] *https://rpubs.com/zheshuen/596809*
[3] https://www.kaggle.com/competitions/titanic/discussion
[4] https://www.r-bloggers.com/2013/09/roc-curves-and-classification/
[5] https://online.stat.psu.edu/stat462/node/207/
[6] https://www.wolframalpha.com/input?i=sigmoid%28x%29
[7] https://stats.stackexchange.com/questions/105501/understanding-roc-curve/105577#105577
[8] https://towardsdatascience.com/predicting-whos-going-to-survive-on-titanic-dataset-7400cc67b8d9
[9] https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/
[10] https://www.datacamp.com/community/tutorials/decision-trees-R
[11] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

--------x------