MTH522 Project2: Predicting Passenger Survivability using Supervised Machine Learning

Anubhav Shankar

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

ashankar@umassd.edu

Abstract **— Machine Learning (ML) methodologies are the talk of the town these days, and rightfully so, as they can be used to decipher patterns and, thus, solve a lot of (previously) incomprehensible problems. ML methods can be divided into two main groups -> Supervised and Unsupervised. In this project, I'll be using Supervised methods, namely Logistic Regression(LR) and Classification Trees (CTs), to predict the survival status of the passengers on the Titanic. The methodologies will primarily be judged basis of their accuracy. However, we'll also be looking at metrics like Specificity, Sensitivity, etc., for LR and the Accuracy of the pruned and unpruned CTs.**

Keywords *— Machine Learning (ML), Supervised Learning, Unsupervised Learning, Logistic Regression, Classification Trees, Prune, Accuracy, Specificity, Sensitivity*

I. MOTIVATION

This undertaking aims to achieve the following objectives:

a.) To develop a more thorough understanding of the ML concepts learned in MTH 522.
b.) Develop and hone skills regarding data wrangling with a particular focus on Feature Engineering to make the data either analyzable or introduce new variables for analysis.
c.) Get familiar with Kaggle's submission formats and, thus, modify the final output accordingly.

II. INTRODUCTION – DATA SOURCE

A. *Dataset*

For this project, I'll use the Titanic dataset from Kaggle. The dataset is part of the competition to predict the passengers' survival status on board the Titanic post its capsizing. The data has been pre-split into Training(891 records) and Testing(418 records) sets for the competition, and I'll be using the same for my analysis.

B. *Dataset Specifications*

The table below gives the variable definition and some additional notes/keys regarding the same:

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 Disclaimer

The overall dataset has been split into Train and Test sets, as mentioned in II-A. However, the Test set does not have the target variable, i.e., Survival. Kaggle does this to judge the participant's submissions against the ground truth. Hence, for my analysis, I'll create a copy of the original Test set and introduce a pseudo-random Survival field for the bulk of this undertaking.

III. METHODOLOGY

Before we proceed, let us briefly describe what we mean by a Supervised Machine Learning method. A Supervised ML model utilizes properly labeled datasets to train the algorithms to classify the data or accurately predict outcomes. As the data is fed into the model, the cross-validation process adjusts its weight accordingly. Supervised ML methods solve many real-world problems at scale, like spam classification of emails in an inbox.

A. Abbreviations and Acronyms

The following abbreviations will be used extensively throughout the document for brevity:

i.) ML -> Machine Learning

ii.) CTs -> Classification Trees

iii.) LR -> Logistic Regression

iv.) OLS -> Ordinary Least-Squares

v.) CART -> Classification and Regression Trees (can be used interchangeably with (ii))

B. Approach

I'll be using Logistic Regression and Classification Trees individually to classify/predict whether a particular passenger survived the Titanic shipwreck. As with Supervised ML, all the models will be executed on the Training set and then run on the Test set for the final classification. The train and test set accuracies will then be analyzed. Concurrently, the performance of the pruned and un-pruned trees will be analyzed for the CTs.

C. Logistic Regression

Also known as Logit Regression, it is primarily used for modeling the conditional probability of a binary outcome variable. LR uses a non-linear link function that restricts the fitted values to 0 and 1. The *log odds* visualized on the S-Curve are modeled as a linear combination of the predictors.

In R, the glm() command can be used to run an LR. The "family" argument in the command must be set to "binomial" to be modeled as a dichotomous outcome variable.

Mathematically,

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

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

ε_i -> The error/residual term

*** Before we proceed: as mentioned in II-C, the original Test dataset (provided by Kaggle) does not have the target/dependent variable "survived." Hence, I have made a copy of the Test dataset, called "test_copy," which will be used for all the analyses in R, and introduced the dependent variable filled with discrete 0-1 values in a pseudo-random form. Please go through the conditions executed to understand the randomization done in test_copy ->*

a.) When sex = "male" then "survived" = 0

b.) When pclass != 1 then "survived" = 0

c.) When sex = "female" then "survived" = 1

NOTE: The above randomization will introduce bias in the "test_copy" dataset, reflected in the Specificity evaluation metric. It is important to note that such an exercise is not done on the Training set. Hence, the model is trained on an unbiased dataset, but the Test dataset on which the prediction must be made is biased.

Let us run a univariate model to analyze the probability of surviving the shipwreck broken down by Gender. For this undertaking, Eq(1) will change to the following ->

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

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 ->

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

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,

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

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 see that some variables are insignificant, i.e., they don't have any statistical significance regarding capturing the variance in the data and are not affecting the dependent variable in any measurable way. For example, in this context, by looking at the results above, we can conclude that point of embarkation did not affect the likelihood of a passenger surviving.

Let's use the stepAIC method from the "MASS" package in R to deduce which independent variables are significant. We can be sure that this will put us on the right track as the AIC value penalizes using extra variables.

Upon running the stepAIC method on Eq(4) in R, we get the following results:

```
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 Socio-Economic Class, Age, and Siblings/Parent affect the probability of survival. We'll see the results and our prediction accuracy in Section IV.

D. Classification Trees

CTs are used to classify a record basis the likelihood into a binary outcome variable. Mathematically, the model will be the same as Eq(4). Basis the initial run, the tree will be pruned if necessary by observing the minimum complexity parameter(CP) and the corresponding "nsplit" that indicates the optimal number of splits.

Running Eq(4) using the rpart() command on R yields the following classification tree:

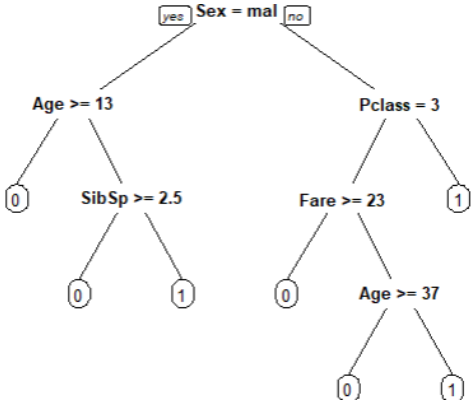


Fig-1: Classification Tree

The tree has six splits and a height of seven. The results seem to be similar to LR. However, "Fare" seems to play a role in determining the probability of survival. So, from the above tree, we can deduce that passengers, especially third-class passengers, who had bought lower-priced tickets tend to have a low probability of survival unless they are over 37.

Let's now look at the complexity parameter to determine if we must prune the tree further. Upon extracting the complexity parameter of the CART model, we get the following results:

	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

A. Logistic Regression

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

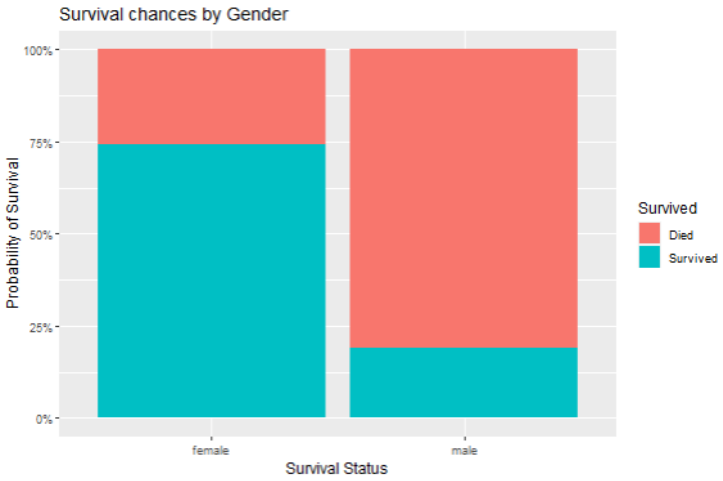


Figure – 2: Probability of Survival basis Gender

To have a more comprehensive understanding of the model performance, we need to look at plots like the Receiver Operator Characteristic (ROC) curve to understand how closely the model is tacking to the empirical observations. By plotting the Sensitivity and Specificity of the model, the ROC curve gives us a good idea of the model fit. Briefly, Sensitivity is the ability of the model to predict a "positive" outcome when the outcome is indeed positive. Similarly, Specificity is the ability of the model to expect a "negative" effect when the outcome is negative. In my case, the case of "survived" = 1, i.e., the passenger survives, will be a positive case vice-versa for "survived" = 0, i.e., the passenger dies.

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

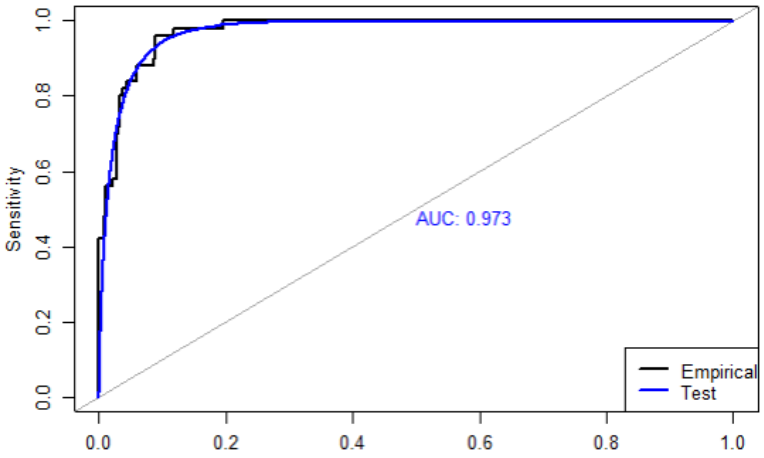


Fig -3: ROC curve for Test Set

From Fig-3, the Area Under the Curve (AUC) is pretty high and very near to one. Thus, the model is functioning pretty well vis-à-vis the observed data. The more the plot hugs the leftmost corner, the better fit the model has. Similarly, let us look at the Training set ROC.

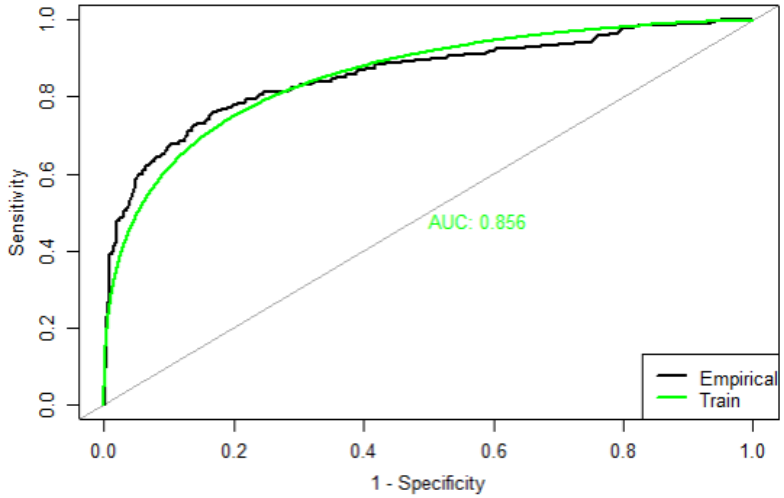


Fig 4: ROC curve of Training Set

As LR operates on a binary target variable, we need to map the probabilities of the categories to see their distribution and coerce the output to a 0-1 binary. This is visualized through the S-Curve, which represents the Sigmoid Function. Mathematically, the Sigmoid function is given by the formula below:

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

In our case, Yi -> Survived

$x = -(b0 + b1 * Xi)$ will be Eq(4)

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

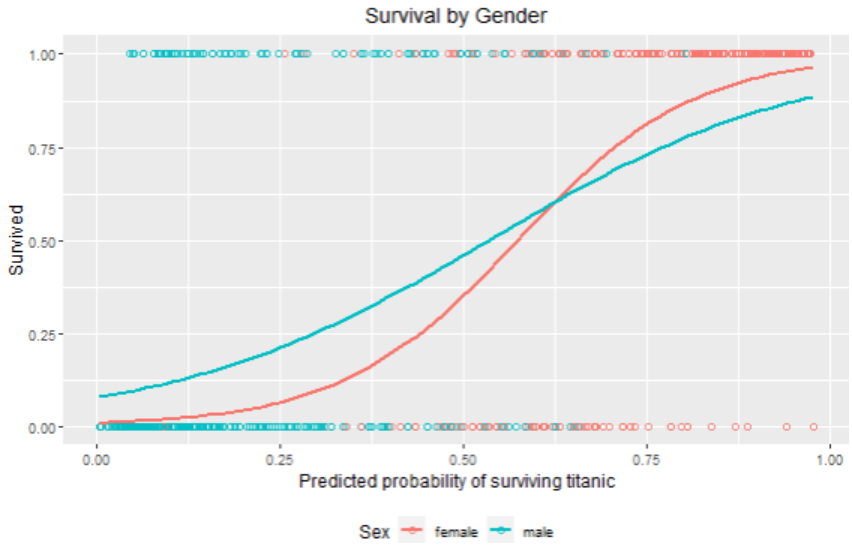


Fig 5: S-Curve

We see that Figs 2 & 5 convey the same outcomes. Hence, we can conclude that our model is consistent.

B. *Classification Trees*

To ascertain CRTs' pruning, we need to visualize which depth has the least relative loss/error. This is done using the matplot() function in R. The output of the plot is shown below:

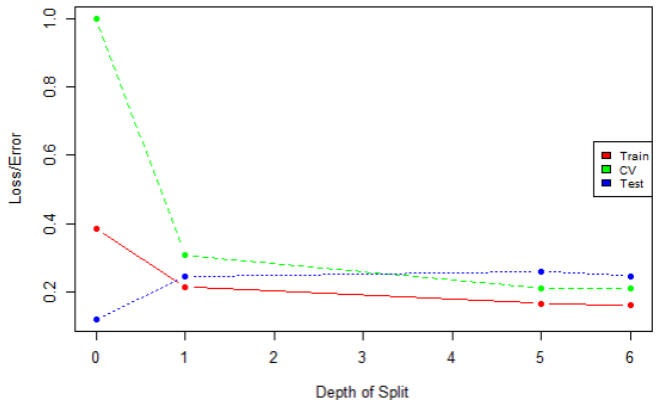
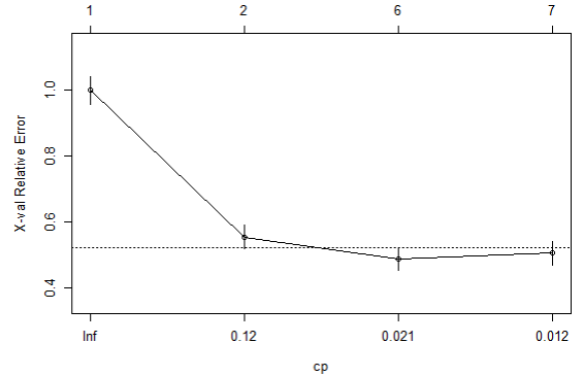


Fig 6: Scree Plot

The Scree plot tells us where the relative error of the training, testing, and cross-validated sets is minimum. The "elbow point" is generally used to determine the optimal depth. However, as we can see here, this contradicts our findings from the CP Table in Section – III B. Concurrently, a split of one will have no predictive power. Hence, I'll be sticking to my original tree. Similarly, we can decide the optimal split post-cross-validation with relative errors. The plot for the same is shown below:



C. Tables

a) Logistic Regression:

TABLE I. LOGISTIC REGRESSION EVALUATION

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

[a] All values obtained from the test_copy dataset on R

b) Classification Tree:

TABLE II: CLASSIFICATION TREE EVALUATION

		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

[a] All values obtained from the test_copy dataset on R

RESULTS & FUTURE SCOPE

The performance of both the Logistic and Classification Tree models is nearly similar. Classification trees, though, are working a tad bit better for this dataset, as can be seen from the Test set accuracies in both caret and non-caret models. The model can be improved using more advanced methods like XGBoost, Random Forest, and more pertinent feature engineering. The possible scope of using unsupervised methods like Support Vector Machines.

While many of the features I engineered turned out to be useless in the long run, I still enjoyed creating them and am proud of being able to do so.

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.