Assignment 2 : A Mathematical essay on Logistic Regression

Vishal Rishi MK
Department of Chemical Engineering
IIT Madras
ch18b013@smail.iitm.ac.in

***Abstract*—In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In this article, we use logistic regression to build a predictive model that helps us to determine the groups of people that were more likely to have survived the accident.**

Keywords—logistic regression, logits

I. Introduction

In many situations, the variables are qualitative. For example, eye color is qualitative, taking qualitative values blue, brown, or green. Often qualitative variables are referred to as categorical. Approaches for predicting qualitative responses is known as classification. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods. There are many possible classification techniques, or classifiers, that one might use to predict a qualitative response. In this article, we discuss the logistic regression approach for classification. Just as in the regression setting, in the classification setting we have a set of training observations ***(x1, y1),...,(xn, yn)*** that we can use to build a classifier. We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier.

Mathematically, a binary logistic model has a dependent variable with two possible values, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables or predictors. The function that converts log-odds to probability is the logistic function. The defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter. Outputs with more than two values are modeled by multinomial logistic regression. The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification.

We try to understand the factors that determined the survival status of passengers in the Titanic, when it sank on April 15, 1992. We employ logistic regression for our analysis. This is an interesting issue in itself as the probability of survival differs greatly between individuals. For example, according to our analysis, men traveling first class were much more likely to survive than men in second and third class, and nearly all women traveling in first class survived compared to women traveling in the other two classes. Yet, the Titanic disaster is also relevant in a more general context. It allows us to analyze behavior under extraordinary conditions, namely, in a life and death situation.

Section 2 of this article gives a mathematical background to logistic regression and the estimation of the parameters. Section 3 of the article describes the data processing pipeline and several observations, insights, visualisations. We also provide quantitative results after applying logistic regression analysis. Finally, section 4 draws conclusions from our explorations and analysis.

II. Logistic Regression

Consider a model with two predictors, ***x1*** and ***x2***, , and one binary (Bernoulli) response variable ***Y***, which we denote ***p = P(Y = 1 | x1, x2).*** We assume a linear relationship between the predictor variables and the log-odds (also called logit) of the event that ***Y = 1,*** given ***x1 and x2***. This linear relationship can be written in the following mathematical form

$$l = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where ***l*** is the log-odds and ***βi, i = 0, 1, 2*** are parameters of the model. By simple algebraic manipulation, the probability that ***Y = 1*** is

$$p = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

where ***σ(.)*** is the sigmoid function. The above formula shows that once ***βi*** are fixed, we can easily compute the log-odds that ***Y = 1,*** given ***x1 and x2*** for a given observation. The coefficients ***βi*** are unknown, and must be estimated based on the available training data. The more

general method of maximum likelihood is used, since it has better statistical properties. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for $\boldsymbol{\beta i}$ such that the likelihood of observing the data is maximised. Hence, the likelihood function becomes the objective function. Optimisation algorithms like gradient descent are employed to find the optimal parameters. In some instances, the model may not reach convergence. Non-convergence of a model indicates that the coefficients are not meaningful because the iterative process was unable to find appropriate solutions. A failure to converge may occur for a number of reasons: having a large ratio of predictors to cases, multicollinearity, sparseness, or complete separation.

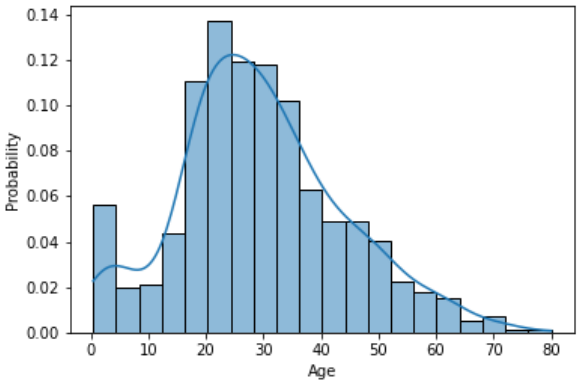

Figure 1: histogram of the feature Age. We can see some outliers in the left tail of the distribution

III. Application to the problem

The data that we have contains data records for 891 passengers who sailed in the Titanic. We have 11 features (categorical and numerical) that describe the survival status of the passengers. We have a target feature that indicates the survival status of each passenger. The 11 features are 1) Passenger Id, 2) Passenger class, 3) Name, 4) Sex, 5) Age, 6) Number of siblings and spouses, 7) Number of parents and children, 8) Ticket Id, 9) Fare, 10) Cabin Id, 11) Port of embarkation. The features Age, Cabin Id, and Port of Embarkation contain missing values. Specifically, feature Age has 177 missing values and feature Cabin Id has 687 missing values. These are too large to impute those data records with the mean, median or mode of the particular feature. If we do so, we might miss some of the interesting patterns. Moreover, not all the features might be relevant in predicting the probability of survival. Clearly, the factors that could have played a significant role in deciding the survival status of a passenger at the time of sinking are Age, Sex, Passenger class (and hence Ticket fare), Cabin Id, number of siblings, spouses, parents, and children. The other features like Name, Passenger Id, and Ticket Id could not have played a role in deciding the survival status. Hence, we can conveniently discard them from our analysis.

A. Exploratory Data Analysis

Though we have data records for 891 passengers, we only have 681 unique Ticket Id's. This could be due to the presence of family tickets, which share a common ticket id for all the members of the family. We define a family ticket as one which is shared by at least two passengers. From the data, we can see that there are 131 family ticket id's. Also, the passengers sharing a family ticket id have the same port of embarkation. Hence, we can easily assume that there were at least 131 families on the ship. Figure 1 shows the histogram of the feature Age. We can see some outliers in the left tail of the distribution. We need to see the effect of these outliers on the logistic regression model. Also, Figure 2 shows the histogram of the feature Fare. This is highly skewed and is not normally distributed. We might need to do some transformations to remove the skewness. Fortunately, the Box-Cox transformation of this feature is shown in Figure 2. Though it shows multimodal nature, it has removed the skewness. Also, the features Fare and Passenger class are highly correlated. All the 3rd class ticket fares are less than 70. For passengers who have bought tickets with fares greater than 70, 100 out of the 105 passengers belong to the first class. Another interesting fact to note is that 15 passengers have travelled with zero fares (First class: 5, Second class: 6, Third class: 4). This could be erroneous values. Instead of working with the feature Fare, we can use the feature Passenger class, which neither has erroneous values nor skewness. Figure 3 shows the barplots between Survival status and the features Sex and Passenger class. We can see that the features Sex and Passenger class induvidually played an important role in determining the survival status. But if we were to use both the features together in our logistic regression model, we need to check their combined significance. Figure 4 shows one such barplot. We can easily infer some points. Within a particular passenger class, females were given priority in an emergency situation. This results in a high survival rate. Compared to females, the survival rate of males is significantly lower. Also, for a given gender, the survival rate is the highest for passenger class 1. This proves the presence of economic bias at the time of sinking of the ship. We also need to analyse the combined effect of age, sex, and passenger class on the survival status. Since age is a continuous variable, we discretize it into three bins - less than 15 years, greater than 48 years, and between 15 and 48 years. Figure 5 shows the barplots showing the combined effect of sex and passenger class on survival rate for different age groups. We can see that the passengers less than 15 years of age have the highest survival rate and the passengers greater than 48 years of age have the least survival rate. Interestingly, though the overall survival rate for males is low, the survival rate of males less than 15 years of age (male children) belonging to passenger class 1 and 2 strikes 100 percent. Also, none of the male passengers greater than 48 years of age belonging to passenger class 3 have survived. These facts show the strong effect of the feature Age on survival rate, combined with the features Sex and Passenger class. These features are a good starting point to use in our logistic regression model.

B. Data Preparation

We have used only three features in our initial model - Age (continuous variable), Sex, and Passenger class. Since the feature Age has 177 missing values, we have dropped those data records with missing age values. Hence, the size of our dataset is reduced to 714. We now create two additional variants of this dataset. The first variant discretises the feature Age into four groups - less than 15 years, between 15 and 30 years, between 30 and 48 years, greater than 48 years. The second variant removes the outliers in the feature Age (shown in Figure 1). Hence, the second variant has 674 data points. In all the variants of the

dataset, we have used ordinal encoding for the feature Passenger class and one-hot encoding for the feature Sex. There is no standardisation applied to the feature Age, as it did not change the results drastically. In all the datasets, we have set aside 100 randomly selected data points as the test set. We have also made sure that the class representation in the train and test sets remain approximately the same.

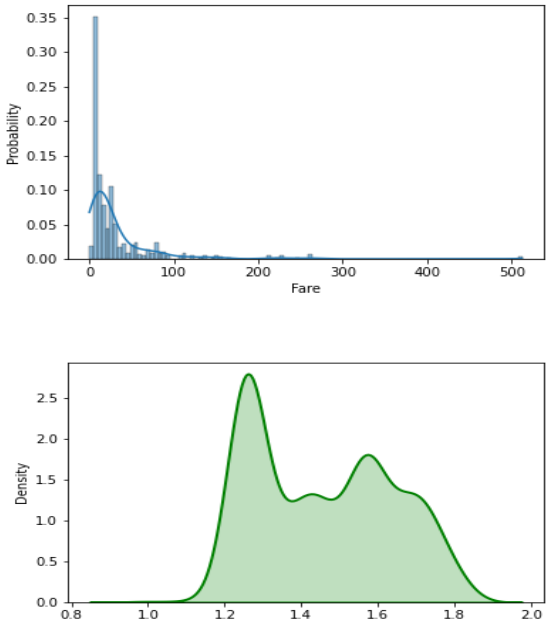

Figure 2: Top: Histogram of the feature Fare; Bottom: Box-Cox transformation of the feature Fare

C. *Logistic Regression Model*

We train logistic regression models for each of the dataset. The parameters of the model are tuned using 10-fold cross validation. The best results are obtained when we use L1 penalty (Lasso regression). The precision, recall, f1, and ROC AUC (Receiver Operator Characteristics Area Under Curve) scores on the test sets for each of the models is shown in Table 1. Figure 6 shows the ROC curves for each model. We can see that model 2 gives the best performance on the test set. This is due to the fact that we have removed the outliers in the feature Age. Model 1 is trained with these outliers and hence, it's performance is worse than that of model 2. Model 3 is trained with the discretized version of the feature Age. Because of the discretization, the model loses access to the granularity of the feature and hence underperforms.

D. *Permutation importance*

Permutation feature importance is a model inspection technique that can be used for any fitted estimator when the data is tabular. This is especially useful for non-linear or opaque estimators. The permutation feature importance is defined to be the decrease in a model score (in our case, it is the F1 score) when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature. This technique benefits from being model agnostic and can be calculated many times with different permutations of the feature. Permutation importance does not reflect the intrinsic predictive value of a feature by itself but how important this feature is for a particular model. We use permutation importance for model 2, since this is our best model. Figure 7 shows the relative importance of the features in both the training and test sets. Table 2 shows the decrease in F1 score in train and test sets when each of the features are permuted. We can easily see that the feature Sex plays the most important role in deciding the survival status of a passenger, followed by passenger class and age. Though the relative importance of the feature Sex remains the same in the train and test sets, the relative importance of the other features differs in the train and test sets. This indicates slight overfitting of the model. But still, the differences are small.

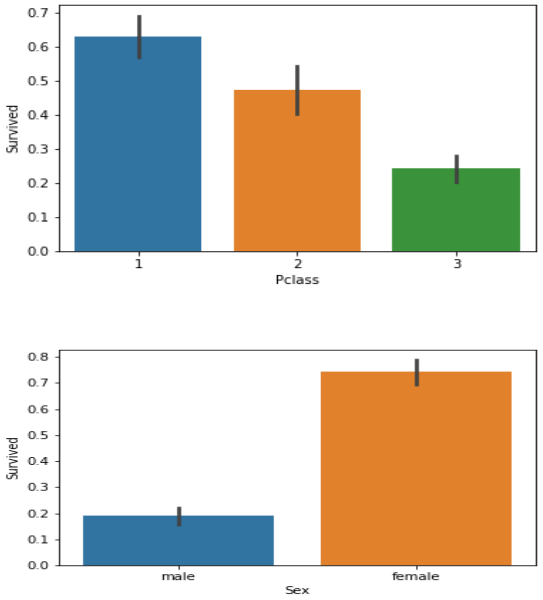

Figure 3: Top: Barplot between Survival rate and the feature Passenger class; Bottom: Barplot between Survival rate and the feature Sex

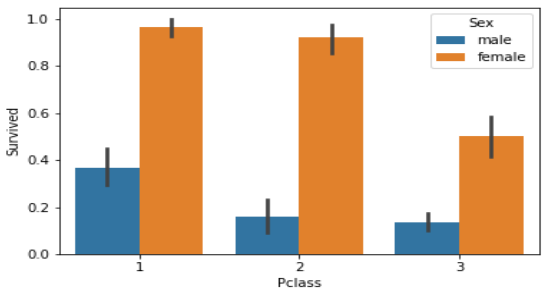

Figure 4: Barplot between Survival rate and the feature Passenger class conditioned on gender

E. *Confidence intervals for the predictions*

Our model outputs the probability of survival of a passenger given their sex, passenger class, and age. We need confidence intervals for our predictions. Hence, we turn to the method of bootstrapping. We resample the entire dataset with replacement several times. The number of times we resample is given by the bootstrap parameter. Every time we resample, we train a logistic regression model with the same hyperparameters used by model 2 and get the predicted probabilities on the test set. Hence, for each observation, we obtain a distribution for the predicted probability of survival. From this, we can construct confidence intervals. Figure 8 shows the distribution and the confidence intervals

for eight randomly selected observations from the test set for a bootstrap size of 2000. We can also use the bootstrapped probabilities as an ensemble technique. We obtain the sample mean of the bootstrapped probabilities for each observation and if this crosses a particular threshold, we classify the observation as "1" (survived). Otherwise, the observation is classified as "0" (not survived). This is called soft classification and the metrics we obtain are a function of the threshold we use. For all practical purposes, we impose a threshold of 0.5. Figure 9 shows the metrics - precision, recall, and F1 scores - as a function of threshold. We see that a threshold of 0.3 gives the best F1 score on the test set and it seems reasonable as well. In such an emergency, humans tend to put in huge effort to ensure survival. Hence, even if the situation indicates a low probability of survival, the passenger is able to survive the mishap.

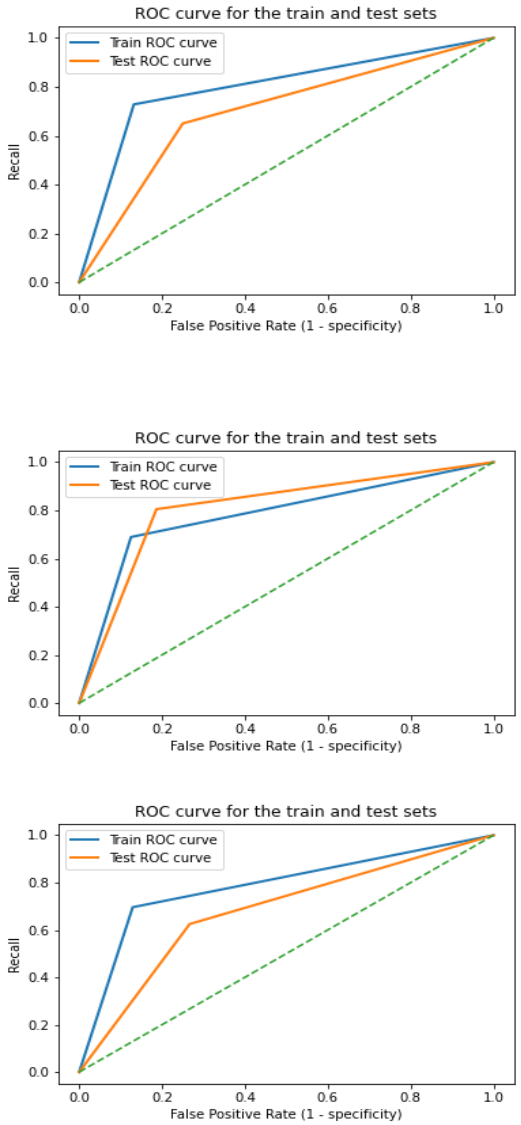

Figure 6: ROC curves for model 1 (top), model 2 (middle), model 3 (bottom)

IV. Conclusion

The estimates of the factors determining survival during the sinking of the Titanic produce a coherent story. While people in their prime were more likely to be saved, it was women—rather than men—who had a better chance of being saved. Children also had a higher chance of surviving. At the time of the disaster, the unwritten social norm of "saving women and children first" seems to have been enforced. Passengers with high financial means, traveling in first class, were better able to save themselves as were passengers in second class (compared to third class). The sinking of the Titanic represents a rare case of a well-documented and most dramatic life and death situation. However, even under these extreme situations, the behavior of human beings is not random or inexplicable, but can be accounted for by statistical analysis. There is also scope for including other factors. There is documentation stating that crew members who had access to better informational and relational resources managed to survive more often than others aboard. This applies in particular to the deck crew who were partly in charge of the rescue operations. Including this factor in our model can improve its performance and generalization ability. Also, we need to take into account the position of the cabin in which the passengers were present at the time of sinking. Passengers who were present in cabins that were in the lower deck would have a very low chance of survival.

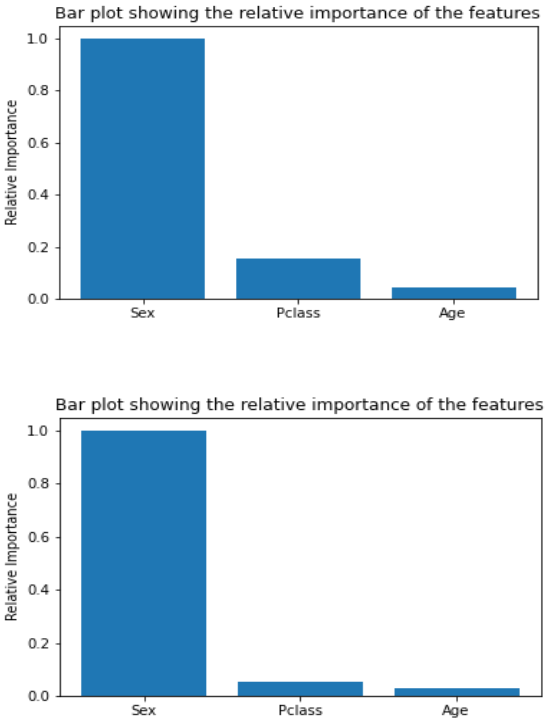

Figure 7: Relative importance of the features in both the training (top) and test (bottom) sets

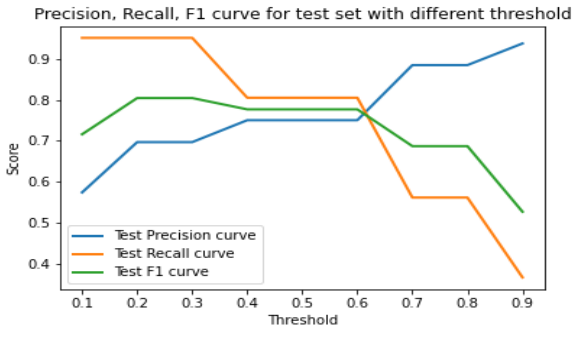

Figure 9: Precision, Recall, and F1 scores on the test set as a function of threshold. We can see that a threshold of 0.3 gives the best F1 score of 0.804 on the test set

References

[1] B. Frey, D. Savage, and B. Torgler, “Surviving the Titanic Disaster: Economic, Natural and Social Determinants”, 2009

[2] “The Journal of Economic Perspectives”, Vol. 25, No. 1 (Winter 2011), pp. 209-221

[3] W. Hall, “Social Class and the Survival on the S.S. Titanic”

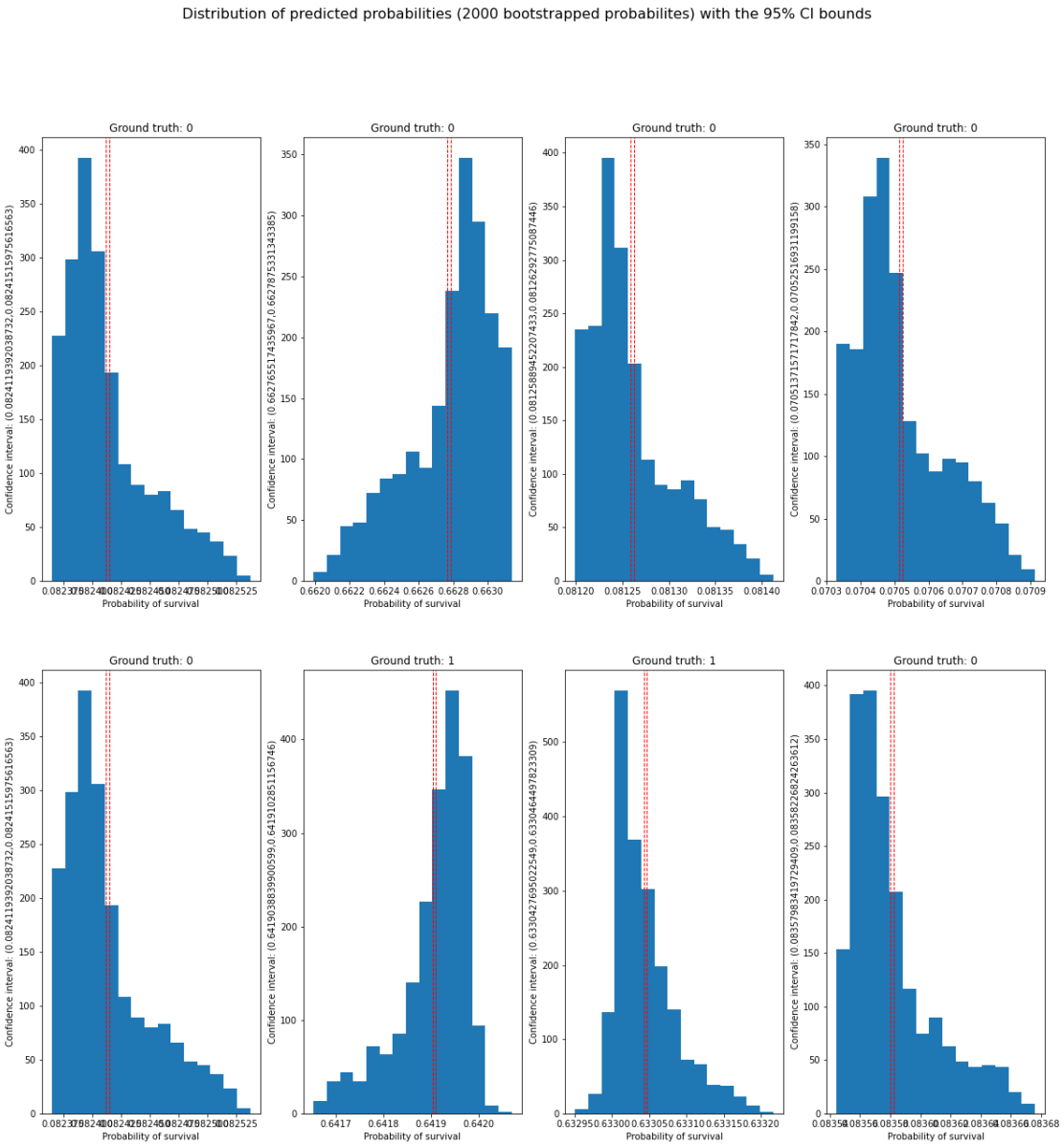

Figure 8: Distribution and the confidence intervals for the probability of survival for eight randomly selected observations from the test set for a bootstrap size of 2000. The ground truth is shown at the top of each plot. The red dotted lines indicate the confidence bounds for the probability of survival for that observation

Metrics	Model 1	Model 2	Model 3
Precision	0.63	0.75	0.61
Recall	0.65	0.81	0.63
F1 score	0.64	0.77	0.62
ROC AUC	0.70	0.81	0.68

Table 1: Model 1: Trained with outliers (L1 penalty; C = 1); Model 2: Trained without outliers (L1 penalty; C = 1); Model 3: Trained with Age feature being discretized (L1 penalty; C = 0.5)

Feature	Decrease in F1 score in Train set (standard deviation in brackets)	Decrease in F1 score in Test set (standard deviation in brackets)
Sex	0.349 (0.017)	0.298 (0.058)
Passenger class	0.054 (0.010)	0.016 (0.018)
Age	0.016 (0.008)	0.009 (0.032)

Table 2: Decrease in F1 score in train and test sets when each of the features are permuted. The feature Sex has the highest importance in predicting the probability of survival of a passenger

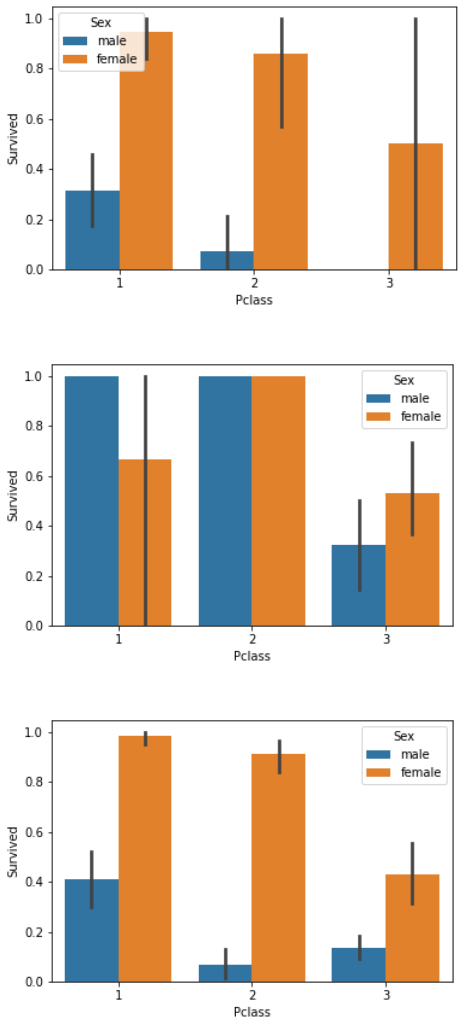

Figure 5: Barplots showing the combined effect of sex and passenger class on survival rate for different age groups. Top: greater than 48 years; Middle: less than 15 years; Bottom: between 15 and 48 years