Determining Titanic Shipwreck Survival Through Classification

Andras Horvath

STAT 6310 Master’s Report

Abstract

The prediction of categorical responses is a major area of practical and theoretical study. Probabilistic classification is a supervised learning procedure by which the probability that an individual observation belongs to a certain class of a categorical response variable is used to establish rules to classify future observations. This paper seeks to explore the use of probabilistic classification methods such as logistic regression, naive bayes, different methods based on assuming Normal populations, and other more general methods such as classification trees to determine the survival of individuals from the Titanic shipwreck.

Contents

1 Background

Whether you are a physician trying to determine if a patient has a malignant tumor based on an MRI or similar scan, or whether an email should be flagged and sent to spam or reach your inbox, or whether an autonomous vehicle perceives a traffic light image as a green light or red light, different classification techniques can be used to build rules to determine the outcome. I begin by making a distinction between classification and discrimination as the lines between the two concepts very often become blurred in practice. The key difference is in the goal of the project. "Discriminant analysis is rather exploratory in nature [and its main goal is to] try to find 'discriminants' whose numerical values are such that the collections [i.e. populations] are separated as much as possible" (Johnson and Wichern 575). The goal of classification on the other hand is "to sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes" (Johnson and Wichern 575). Therefore, the goal of discrimination is separation whilst the goal of classification is allocation. As such, the emphasis in this paper will be the process of defining rules which will allow for allocation of future observations. There are countless ways in which these rules can be created but the one which will mainly be focused on in this paper is a probabilistic classification approach. This means that we will "first predict the probability that [an] observation belongs to each of the categories of a qualitative variable, [and use this] as the basis for making the classification" (James et al. 129). To achieve this end methods such as logistic regression, naive bayes, different methods using normal distributions, and classification trees will be explored.

2 Methodology

We will now discuss some of the different probabilistic classification models at our disposal. This section will include the strengths and weaknesses of each method in addition to a short theoretical overview of the algorithms and how they are used.

Logistic regression

The idea of using linear regression to use predictors to determine an outcome makes a great deal of intuitive sense. There are, however, some challenges with this approach. Suppose that the response variable, Y, which is a qualitative variable takes on the values 0 or 1. Also assume, without loss of generality, that we are interested in using only a single predictor in our linear regression model. If the goal is to model the probability that a 1 occurs for an individual observation using a linear model, we can write the model as follows:

$$p(X) = E(Y = 1|X) = \beta_0 + \beta_1 X$$

One of the major drawbacks of this approach is that the predicted values that Y can take are unbounded and therefore it is possible to obtain a probability that is negative or greater than 100 percent. To highlight this phenomenon and compare it to a curve produced from logistic regression refer to Figure 1 below:

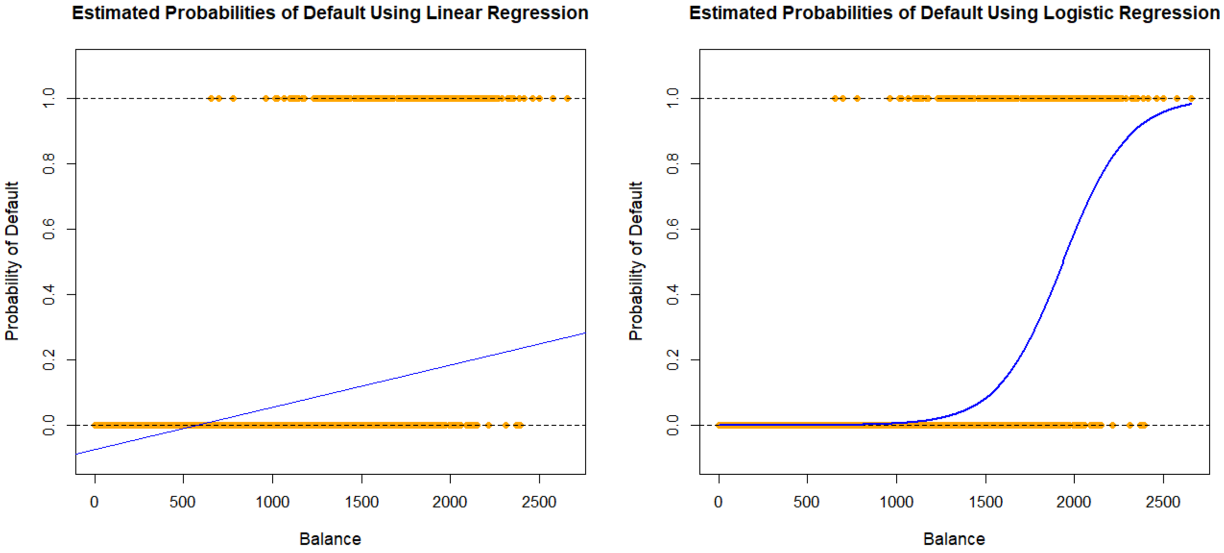


Figure 1: Comparison of Linear and Logistic Regression

Figure 1 is a replication of a plot that appears on page 133 of “An Introduction to Statistical Learning with Applications in R” (James et al.). The dataset contains information about whether a person defaulted and other factors such as their balance that can be used to help determine this. Notice that some of the estimated probabilities of defaulting in the linear regression approach can be negative. In contrast, the logistic regression approach does not have this problem as all estimated probabilities are between 0 and 1. Additionally, notice that the chance that a person will never default vs will certainly default, while not impossible, is very small. The logistic function for p predictors is defined as follows:

$$p(X) = \frac{\mathrm{e}^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + \mathrm{e}^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

This can be used to classify which population a given observation should belong to. The choice of cutoff for $p(X)$ is up to the research scientist. If no prior information is known about the populations then a good default cutoff is $p(X) = 0.5$ where we assign an observation to the population for $Y = 1$ if $p(X) > 0.5$.

Additionally, we can show why logistic regression is a linear model. It is the log odds, or logit as it is often referred to, that follows a linear model. The log odds for p predictors is as follows:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Here odds are used which are another way to view the same information. Faraway explains that "odds are more convenient from a mathematical perspective and do allow us to express the effects of [the predictors] in a compact way. One disadvantage [however] is that they only express relative difference" (32).

Finally, what are some of the benefits and potential pitfalls of logistic regression? Some of the obvious benefits include the ease of implementation, interpretation, and that it gives rise to estimated probabilities that can be used to create straightforward rules for classification. There are however some downsides. As logistic regression is a linear model, it assumes that the relationship between the response and the covariates will be linear. This, however, is not guaranteed and is of particular concern if more complex nonlinear relationships would better describe an association between the variables. Another problem occurs "when there is substantial separation between the two classes [as] the parameter estimates for the logistic regression model are surprisingly unstable" (James et al. 141). Some of the future methods discussed will alleviate this problem. However, even with the downsides of logistic regression it is still a powerful and widely used tool as it serves as a good baseline to compare other more complex classification algorithms that a researcher might develop.

Naive Bayes

While logistic regression can directly model the probabilities $p(X)$ using the logistic function, there are other approaches to this problem. One such approach seeks to "model the distribution of the predictors X separately in each of the response classes (i.e. for each value of Y)" (James et al. 141). For our purposes since Y can only take on two values there are only K = 2 classes. Next, Bayes' Theorem is then applied to this approach in order to get estimates for $p_k(X) = P(Y = k|X = x)$. James et al. defines Bayes Theorem as

$$p_k(X) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where the notation $p_k(X)$ denotes the posterior probability that an observation $X = x$ belongs to the kth class, π_k denotes the overall or prior probability, and $f_k(X)$ denotes the density function of X for a given observation. Where this method differs from logistic regression is that now the posterior probability can be calculated not by using the logistic function but rather by using estimates for the prior probability, π_k, and density function $f_k(X = x)$. One of the benefits of using this approach is that "estimating π_k is easy if we have a random sample from the population: we simply compute the fraction of the training observations that belong to the kth class" (James et al. 142). The downside of using this approach is that it can often be difficult to estimate the density function. The reasoning for this is that "estimating a p-dimensional density function is challenging because we must consider not only the marginal distribution of each predictor […] but also the joint distribution of the predictors" (James et al. 155). One method to tackle this problem is to introduce assumptions that help to simplify estimation of the density function. The Naive Bayes classifier is just one such method. The Naive Bayes classifier assumes that within each particular class that the p predictors are independent. Therefore, because the predictors are assumed to have no association with one another this simplifies the task of estimating the density function as there are no joint distributions that need to be considered. This Naive Bayes assumption leads to the following transformation of the Bayes Theorem:

$$p_k(X) = P(Y = k|X = x) = \frac{\pi_k * f_{k1}(x_1) * f_{k2}(x_2) * \dots * f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l * f_{l1}(x_1) * f_{l2}(x_2) * \dots * f_{lp}(x_p)}$$

The estimation of the density functions can be done in several different ways. A parametric approach "if X_j is quantitative [is to] assume that within each class, the jth predictor is from a (univariate) normal distribution" (James et al. 155). One nonparametric approach is to use a kernel density estimator. Alternatively, we can create "a histogram for the observations of the jth predictor within each class. Then we can estimate $f_{kj}(x_j)$ as the fraction of the training observations in the kth class that belong to the same histogram bin as x_j" (James et al. 156). Finally, "if X_j is qualitative, then we can simply count the proportion of training observations for the jth predictor corresponding to each class" (James et al. 156). Additionally, it is not necessary, strictly speaking, to calculate the denominator of the transformed Bayes Theorem above. This simply serves to keep the posterior probability in [0,1]. The numerator alone could be calculated and the class which has the largest numerator value is the class that the observation will be assigned to.

What are some of the strengths and weaknesses of Naive Bayes? One of the strengths is the ease of estimating the prior probability. A data free prior can be constructed based on preexisting knowledge which can later be updated to a new prior based on data. The foremost weakness of the Naive Bayes classifier lies within its assumption. The assumption that within each class the predictors are independent is not necessarily true and will not be valid in plenty of cases. However, even with this large downside Naive Bayes may still be reasonable to implement. This is because Naive Bayes "often leads to pretty decent results, especially in settings where n is not large enough relative to p for us to effectively estimate the joint distribution of the predictors within each class… The Naive Bayes assumption introduces some bias but reduces variance" (James et al. 155).

Normal Population Methods

The methods discussed in this section will rely on the assumption that the populations from which the different classes come from all have Normal probability density functions. As Johnson and Wichern state, "classification procedures based on normal populations predominate in statistical practice because of their simplicity and reasonably high efficiency across a wide variety of population models" (584). However, as with any method there will be a price to pay for this simplicity. For one, the boundary between classification and discrimination, or separation, will become blurred in this section. This is due to the related nature of the two topics in the formulations of the classification rules. Johnson and Wichern state the following:

> The farther apart the groups [are] the more likely it is that a useful classification rule can be developed…[Allocation] rules appropriate for the case involving equal prior probabilities and equal misclassification costs correspond to functions designated to maximally separate populations. (606)

Finally, how the methods in this section will be distinguished from one another will be based on the additional assumptions of their covariance matrices. We will first discuss the case where there are only two classes or populations, and then later generalize the classification rules to any number of classes.

We begin with the simplest case where there are only two populations, which we assume each come from a multivariate normal distribution, and that their covariance matrices are equal. We will refer to methods that use the equal covariance assumption as Linear Discriminant Analysis (LDA). In the upcoming classification rule we will need to use an estimation to the population covariance matrix which will be the following:

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + \ (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + \ (n_2 - 1)}\right] S_2$$

The first method here considers that "an optimal classification procedure should, whenever possible, account for the costs associated with misclassification" (Johnson and Wichern 579). Therefore, when trying to minimize the expected cost of misclassification we can allocate observation x_0 to the first population if:

$$(\bar{x}_1 - \bar{x}_2)^T * S^{-1}{}_{pooled} * x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T * S^{-1}{}_{pooled} * (\bar{x}_1 + \bar{x}_2) \geq \log\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Otherwise, the observation x_0 will be allocated to the second population. Here c(1|2) refers to the cost when an observation belongs to the second population but it is misclassified as belonging to the first population, c(2|1) is the cost when an observation belongs to the first population but was misclassified as belonging to the second population, and p_1 and p_2 are the prior probabilities. The costs can be found from a confusion matrix and will be the counts that the misclassification occurs.

The second method, which is an alternative method proposed by Fisher, "does not assume that the populations are normal. It does, however, implicitly assume that the population covariance matrices are equal" (Johnson and Wichern 590). Here the allocation rule proposed by Fisher says to allocate an observation x_0 to the first class or population if:

$$\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)^T * S^{-1}{}_{pooled} * x_0 \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T * S^{-1}{}_{pooled} * (\bar{x}_1 + \bar{x}_2)$$

Otherwise, the observation x_0 will be assigned to the second class or population. It should be mentioned that "Fischer's idea was to transform the multivariate observations $\boldsymbol{x}$ to univariate observations y such that the y's derived from [the first and second population] were separated as much as possible" (Johnson and Wichern 590). Based on the above formulation it can be seen "provided that the two normal populations have the same covariance matrix, Fisher's classification rule is equivalent to the minimum ECM rule with equal prior probabilities and equal costs of misclassification" (Johnson and Wichern 592).

The next case that is to be considered is the more general case where the covariance matrices are not equal to one another. This leads to the following Quadratic classification rule what allocates observation x_0 to the first class if:

$$-\frac{1}{2}x_0{}^T\left(S_1{}^{-1} - S_2{}^{-1}\right)^T x_0 + \left(\bar{x}_1{}^T S_1{}^{-1} - \bar{x}_2{}^T S_2{}^{-1}\right) x_0 - k \geq \log\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

where $k = \frac{1}{2}\log\left(\frac{|S_1|}{|S_2|}\right) + \frac{1}{2}\left(\bar{x}_1{}^T S_1{}^{-1}\bar{x}_1 - \bar{x}_2{}^T S_2{}^{-1}\bar{x}_2\right)$. Otherwise, the observation is classified as belonging to the second class.

We now expand the previous allocation rules to the more general case where there can be more than two groups which are being classified. Much like before there are two distinct cases: one assumes the equal covariances between the two normal populations whereas the other does not assume this equality and is thus more general. When assuming equality of these covariance matrices for the $i = 1,2, \ldots, g$ different groups we will utilize the following estimate for the linear discriminant score $\widehat{d_\iota}(x)$:

$$\widehat{d_\iota}(x) = \bar{x}_i^{\ T} * S_{pooled}^{\ \ -1} * x - \frac{1}{2}\bar{x}_i^{\ T} * S_{pooled}^{\ \ -1} * \bar{x}_i + \log(p_i)$$

where $S_{pooled} = \frac{1}{n_1 + n_2 + \cdots + n_g - g}((n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + \left(n_g - 1\right)S_g)$. Once the linear discriminant scores are calculated for a given observation x, we allocate x to the class which has the largest linear discriminant score. If we do not assume equality of the covariance matrices for the $i = 1,2, \ldots, g$ different groups, we will utilize the following estimate for the quadratic discriminant score $\widehat{d_\iota}^{\,Q}(x)$:

$$\widehat{d_\iota}^{\,Q}(x) = -\frac{1}{2}\log(|S_i|) - \frac{1}{2}(x - \bar{x}_i)^T * S_i^{\ -1} * (x - \bar{x}_i) + \log(p_i)$$

Once the quadratic discriminant scores are calculated for a given observation x, we allocate x to the class which has the largest quadratic discriminant score.

In the Application section of this paper a different, though equivalent formulation will be utilized which bases the classification on probabilities. Recall in the Naive Bayes formulation that the Bayes Theorem can be used to classify objects based on posterior probabilities. There the major assumption was that the p predictors were all independent which simplifies estimating the density functions. The same process can now be used but the assumption is that the density functions are all multivariate normal and slightly different approaches will be taken based on the covariances of the different classes. This will give us the same result as using the score functions but now allows us to return to using probabilities for classification purposes.

What are some of the benefits and downsides of using LDA and Quadratic Discriminant Analysis (QDA)? The single largest benefit is that they provide a very simple, closed form solution to the classification problem at hand. However, the downside of using these methods is that they may not be appropriate to use as the normality assumption may be violated. This weakness is especially problematic for QDA. However, even if there are qualitative predictor variables present, which would clearly violate the normality assumption, LDA may still be used as it is "often remarkably robust to model violations" (James et al. 148). So, if normality cannot be established, can we still use these methods? When this occurs, there are two possible options as highlighted by Johnson and Wichern:

> First, the nonnormal data can be transformed to data more nearly normal, and a test for the equality of covariance matrices can be conducted […] to see whether the linear rule […] or the quadratic rule […] is appropriate…Second, we can use a linear (or quadratic) rule without worrying about the form of the parent populations and hope that it will work reasonably well. (595)

Having established the differences between LDA and QDA one other question that could be asked is which method would we prefer to use for a given situation? James et al. states the following:

> LDA tends to be a better bet than QDA if there are relatively few training observations and so reducing variance is crucial. In contrast, QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern, or if the assumption of a common covariance matrix for the [g] classes is clearly untenable. (153)

Given everything that has been discussed in this section we can see that not everything is black and white. Even if the proposed methods may not be appropriate, if normality is violated, there may still be avenues to use them to provide the researcher with more information and to compare the results with other models.

Classification Trees

The previous methods could be used to define rules for classification based on probabilities. However, this is not the only metric by which classification rules can be built. We now turn to one of the many other possible methods and observe this approach's advantages and disadvantages over using probabilities to define classification rules. The alternative method which will be discussed in this paper is Decision Trees, which are a very computer intensive approach hence why they only became popular in more recent years. They are also the first nonparametric method discussed here as decision trees are not tied to any underlying assumptions as was previously the case. Johnson and Wichern describe the general Decision Tree procedure as follows:

> Initially, all objects are considered as a single group. The group is split into two subgroups using, say, high values of a variable for one group and low values for the other. The two subgroups are then each split using the values of a second variable. The splitting process continues until a suitable stopping point is reached. (644)

An example of a typical decision tree can be found in Figure 2.

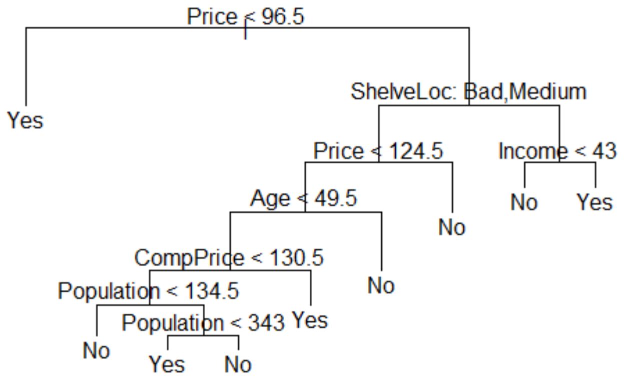


Figure 2: Classification Tree for Car Seat Sales

This decision tree was produced to classify whether the child car seat sales in one of 400 stores were high or low based on various predictor variables. Figure 2 is the result after cross validation was used to determine optimal tree complexity and pruning was implemented. The code to produce this plot in R can be found on pages 353-355 in James et al. We can see that the plot appears to be an upside-down tree where the leaves or branches grow downwards. Each branch which terminates and has no more splitting branches is referred to as a terminal node or leaf. We can see that after the first split at the top of the tree the left branch does not split any further and so it can be called a terminal node or leaf. Each point where a split originates from is referred to as an internal node. The initial split Price < 96.5 at the top of the decision tree is one such internal node. This tree has 8 internal nodes and 9 terminal nodes.

The plot describes not only the binary splitting rules but also the most important indicators. For example, the first split occurs at the top of the tree with the price variable. This indicates that the most important factor in determining high or low sales is the price of the seat. The first split also indicates that we should pass to the left branch if the price of the seat is less than $96.50. The end result is classifying the sales for that particular store as high. It also shows us that if the price of a seat is less than $96.50 then the other factors play little to no role in determining the sales. If on the other hand a store's prices are greater than or equal to $96.50 then we pass through the right branch and continue following the branches until a terminal node is reached. While this tree's leaves/terminal nodes are categorical, as this is a classification tree, if instead a regression tree is used then the terminal nodes will be a number which will correspond to the mean value of the response for an observation that falls there.

Depending on the type of response data there are several different kinds of trees that can be used. If the response is quantitative then a regression tree is often used. Regression trees perform recursive binary splitting of the predictor space such that the chosen cutoff at each node leads to the greatest reduction in the residual sums of squares. This idea can be extended to other types of responses as well. Faraway mentions some examples including "binomial, multinomial, Poisson, and survival data by using a deviance, instead of the RSS, as a criterion" (354). Classification trees are instead used when the response is qualitative. One other crucial difference is that while RSS is pivotal to regression trees, it is not used for classification trees. Instead, a new measure must be used so that "the splits [...] divide the observations within a node so that the class types within a split are mostly one of a kind" (Faraway 354). Three such measures that are commonly used are the Deviance, Entropy, and Gini Index as shown below:

1. Deviance: $D_i = -2\sum_k n_{ik} \log(p_{ik})$
2. Entropy: $D_i = -\sum_k p_{ik} \log(p_{ik})$
3. Gini Index: $D_i = 1 - \sum_k {p_{ik}}^2$

Here, Faraway defines n_{ik} as "the number of observations of type k within terminal node i and p_{ik} [as] the observed proportion of type k within node i. [Finally,] let D_i be the measure for node i so that the total measure is $\sum D_i$" (354). Notice that "all these measures share the characteristic that they are minimized when all members of the node are of the same type" (Faraway 355). It is "for this reason [that] the Gini index is referred to as a measure of node purity – a small value indicates that a node contains predominantly observations from a single class" though it should be stated that the Entropy is quite similar numerically (James et al. 336).

What are some of the benefits and downsides of using tree methods? By far the greatest benefit of Decision Trees is that they can easily be displayed graphically and their "structure is easier for nontechnical people to understand" (Faraway 344). One of the great challenges in creating new statistical methods is the power of the method as well as its ease of explanation to a nontechnical audience. If the method performs extremely well but is absurdly difficult to grasp by an audience, then the method likely won't be implemented on a large scale. Classification trees strike a great balance between their general level of accuracy and their ease of interpretation even by individuals of a nontechnical background. But this isn't the only benefit that can be seen from their graphical display. It is easy to see that "tree models are well suited to finding interactions. If we split on one variable and then split on another variable within the partitions of the first variable, we are finding an interaction between the two variables" (Faraway 344). To a certain extent we can find higher order interactions as we continue to follow the splits in a decision tree. Another benefit of Decision Trees is their ability to handle missing data. Faraway mentions that "missing values can be handled quite easily by tree methods...If we believe the fact of being missing expresses some information, we might choose to treat missingness as an additional level of a factor" (344). This could therefore even be used for the original Titanic dataset for this paper to try to solve some of the problems caused by missing data.

Despite all the great benefits of decision trees, as one might expect, there are some serious downsides as well. For one, when compared to linear models, decision trees "lack the inferential apparatus of prediction intervals. They have discontinuities in the prediction at the partition boundaries but are constant elsewhere" (Faraway 350). So, unless a complex, nonlinear relationship exists between a response and its predictor variables a decision tree is often not the best method to use. The next issue is that trees are often "very non-robust. In other words, a small change in the data can cause a large change in the final estimated tree" (James et al. 340). This largely stems from the fact that much of the splitting and rule creating comes from using local means which will be heavily affected by outliers. The final major downside is that "trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches" (James et al. 340). To cope with the prediction problem, different ensemble approaches have been proposed including bagging, random forests, and boosting. The general idea behind these methods is instead of using a single tree we create a large number of trees, often through the use of bootstrapping. A deeper explanation can be found in section 8.2 of James et al. Even these methods have their own problems and deciding whether to use them depends on the goal of your project or investigation. Faraway, who mentions the use of random forests here, says that "if the goal of your analysis is prediction, then the random forest is a good choice. In contrast, if your goal is to explain the relationship between the predictors and the response, random forests may provide some insight but lack the full power of linear or additive models" (354). Whatever the purpose of your investigation is, most models can potentially provide additional insight.

A final topic discussed in this section is about the use of pruning to create better decision trees. Producing larger and more complex trees often leads to better predictions on a training set but suffers from potential overfitting. In contrast, "a smaller tree with fewer splits…might lead to lower variance and better interpretation at the cost of a little bias" (James et al. 331). The process of pruning is somewhat similar to that of backward elimination in linear regression. The general idea is to start with a large tree and then make it smaller through pruning the branches until the optimal tree is found. The major question that comes about is "how do we determine the best way to prune the tree" in order to reach this optimal model (James et al. 331)? One idea would be to look at every possible subtree that could come from pruning a single large tree and then calculate the testing error using cross validation. The downside of this approach is that there could be a very large number of subtrees that could exist and implementing cross validation on all of them could take a very long time. Another method would be to find a way to only look at a subset of the possible subtrees and then implement cross validation on those trees. This is more feasible and from which cost complexity pruning comes about. The idea is when starting from a large tree T_0 "rather than considering every possible subtree, we consider a sequence of trees indexed by a nonnegative tuning parameter α. For each value of α there corresponds a subtree $T \subset T_0$ such that

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

is as small as possible" (James et al 332-333). For the above summation, James et al. define the following:

> $|T|$ indicates the number of terminal nodes of the tree T, R_m is the rectangle (i.e. the subset of predictor space) corresponding to the mth terminal node, and $\hat{y}_{R_m}$is the predicted response associated with R_m…The tuning parameter α controls a tradeoff between the tree's complexity and its fit. (333)

To implement this, we would begin by finding the value of α through cross validation. Once the value of α has been found we can then prune the initial large tree and obtain the optimal tree that would correspond to this value of α.

Evaluation of Classification Performance

When evaluating the classification performance for a procedure there are numerous metrics that can be used. The first method that will be used is a confusion matrix. Confusion matrices are a convenient way to display the number of observations that were classified correctly, how many were misclassified, and how the classifier allocated them. The second method will be the overall prediction accuracy. This will be found by taking $1 - prediction\ error\ rate$. The prediction error rate is defined as:

$$\frac{Total\ count\ of\ off\ diagonal\ elements\ in\ the\ confusion\ matrix}{Total\ number\ of\ classified\ observations}$$

As such, the prediction accuracy will come directly from the confusion matrix. Prediction accuracies can be found for both a training and testing set. A classical approach would be to use different methods to classify the observations in a dataset, and then to compare the models using their testing set accuracies or errors. Alternatively, the prediction accuracy for a training set can also be used to compare different models. However, this is not as commonly used because a researcher is often more interested in how well their classification model works for new data. Therefore, comparing accuracies or errors for the data that was used to create a particular model is not as meaningful. With this said, it may still be a useful metric especially if it is not possible to obtain a testing set accuracy or error. Finally, when trying to compare the error rates across different choices of thresholds one typical classification error metric is to use a Receiver Operating Characteristics (ROC) curve. For a given classification model, the ROC curve plots the two types of error for every possible value of threshold. To evaluate the performance from an ROC curve the Area Under the Curve (AUC) can be used. The closer the area is to its maximum, one, the better the performance.

3 Application

The dataset used for this paper gives characteristics for individuals who were on the 1912 maiden voyage of the Titanic. The dataset was a part of a challenge posted on the Kaggle website for which the overarching theme was to give individuals practice with implementing different machine learning algorithms for classification. However, more specifically the goal of the competition was to create a model which can best classify an individual on the Titanic's maiden voyage as either surviving the trip or not surviving. Our goal, however, will be to determine survival by using probabilistic classification techniques and learn about their strengths and weaknesses. The full description of the challenge as well as where the dataset can be obtained on Kaggle's website is available in the following link: https://www.kaggle.com/c/titanic. The data is made available in two files. The first is a training dataset which contains 891 out of the approximate 2200 passengers aboard the Titanic as well as several characteristics about each passenger. The most notable characteristic is whether they survived or not. In contrast, the test dataset has only 418 passengers and does not include whether the passenger survived or not. Because the techniques used in this paper will need to know whether a passenger survived in order to help determine the cutoff for the probabilities to improve the accuracy of the method, only the first file i.e. training dataset was used. A detailed understanding of the variables is outlined in Table 1.

Variable	Description
Survival	Whether the passenger survived.
Ticket Class (Pclass)	The socioeconomic status of the passenger. This is broken into the following: upper class, middle class, and lower class.
Name	The passenger's name.
Sex	Whether the passenger was Male or Female.
Age	The age of the passenger.
Number of siblings/spouses (SibSp)	The number of siblings/spouses aboard the Titanic. Sibling is defined to mean brother, sister, stepbrother, and/or stepsister. Spouse is defined to mean husband or wife (mistresses and fiancés are ignored).
Number of parents/children (Parch)	The number of parents/children aboard the Titanic. Parent is defined to mean mother or father. Child is defined to mean daughter, son, stepdaughter, and/or stepson.
Ticket Number	A passenger's ticket number.
Passenger Fare (Fare)	The amount of money a passenger's ticket cost (taken to be in U.S. dollars)
Cabin Number	The cabin number of a passenger.
Port of Embarkation (Embark)	One of three ports which the Titanic left after a passenger came aboard. This is broken into the following: Cherbourg, Queenstown, and Southampton.

Table 1: Description of Variables

Not all of these original variables or observations were used in the analysis. First, passenger's name was deemed to not be a great indicator to determine whether that passenger survived or not and was thus removed. Second, because almost all of the ticket numbers were unique this was also deemed as an indicator that would not be helpful in determining whether a passenger survived or not. Third, because 687 of the 891 passengers did not have a cabin number listed, this variable was also removed. Finally, there were 177 passengers whose age was not listed and 2 observations whose embarkment was not listed. Since both may be deemed as important characteristics that could help determine whether a passenger survived those 179 observations were dropped leaving 712 passengers to conduct the classification procedures. A discussion about the impact of removing these observations will be discussed in the application section using logistic regression.

Different sources give slightly different estimates for the exact number of deaths during this voyage, but all of them estimate just over 1500 deaths of the roughly 2200 passengers aboard. While the likelihood of survival may in large have depended on random luck, there are characteristics that led some passengers to have a higher chance of survival. For example, it is well documented that "as attempts were made to contact nearby vessels, the lifeboats began to be launched, with orders of women and children first" (Britannica). So, whether a passenger was a woman or a child may have largely impacted the individuals who escaped.

As there are functionally only two quantitative variables, age and ticket price, when trying to determine if potential associations exist between the variables, boxplots are a valuable tool to compare the measures of center between different classes. Each box plot below shows a qualitative variable vs age or fare for both the group that survived and for the group who did not survive. We begin by observing a set of boxplots detailing the survival rate of men vs women based on their age in Figure 3.

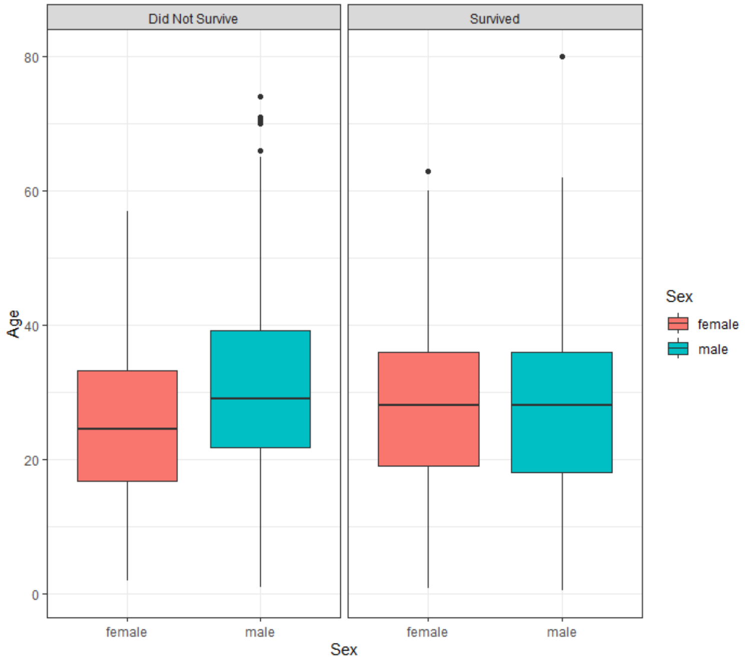


Figure 3: Boxplots of Survival Based on Age for Males and Females

We can see from the Figure 3 that there does not appear to be any substantial difference in the ages of males and females who survived, but there does appear to be a slight difference between the ages of males and females who did not survive. Figure 3 does not tell us the whole story however, as 195 out of 259 females in our dataset survived while only 93 out of 453 males survived. This gives us a difference of almost 50 percentage points between the survival rates of each group. This just shows us that not all of the variables may be helpful in predicting the survivability of an individual passenger. Next, Figure 4 details the survival rate of lower, middle, and upper class passengers based on their age.

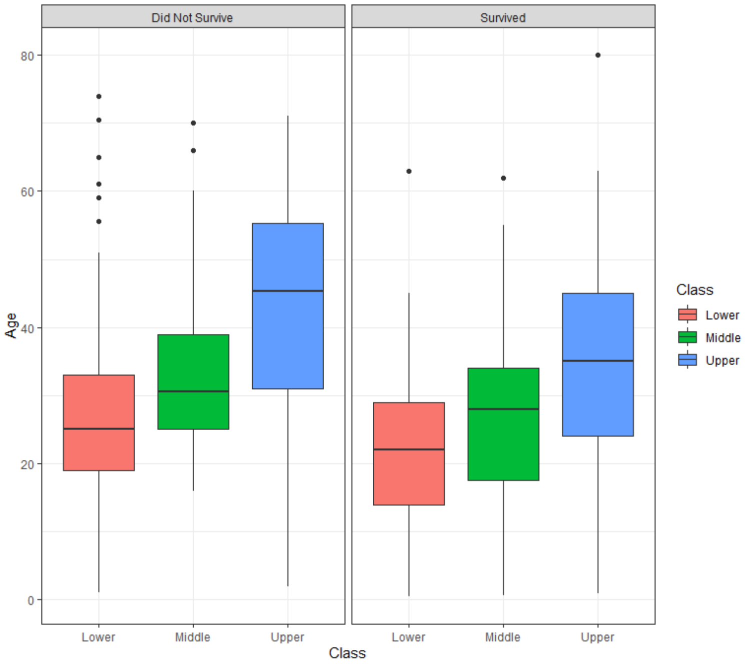


Figure 4: Boxplots of Survival Based on Age for Economic Class

We can see from Figure 4 that there does appear to be an association between age and economic class. The general trend is that as economic class increases age also increases. Even more telling, however, is that this pattern is seen in both the survival and nonsurvival group, but that ages were lower across all economic classes in the survival group. Figure 5 compares the age of passengers based on their port of embarkment.

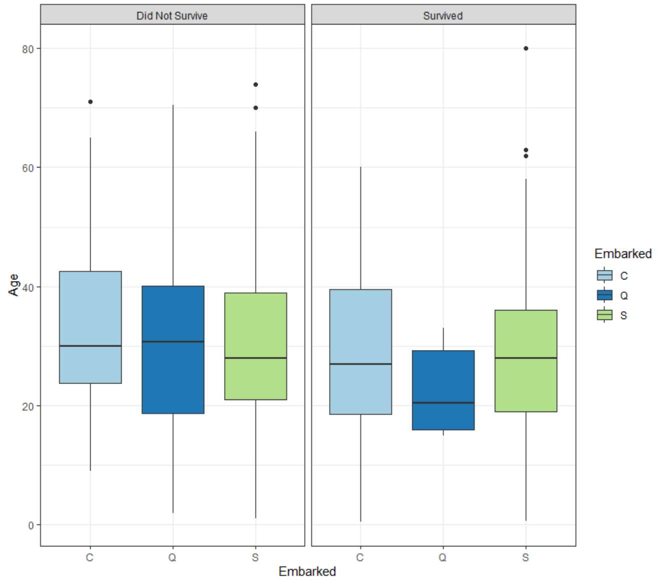


Figure 5: Boxplots of Survival Based on Age for Port of Embarkment

Figure 5 does not seem to suggest any strong association between age and the ports of embarkment. The only real thing of note is that the passengers who embarked from Queenstown did have a difference in the ages of people who survived and did not survive. Next, Figure 6 compares the age of passengers and the number of siblings or spouses they had, and Figure 7 shows age vs number of children or parents a passenger had.

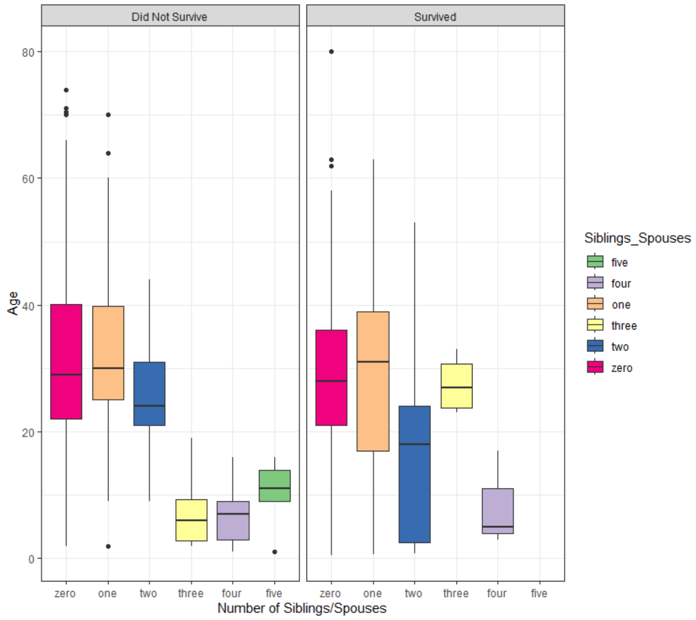


Figure 6: Boxplots of Survival Based on Age and Number of Siblings/Spouses

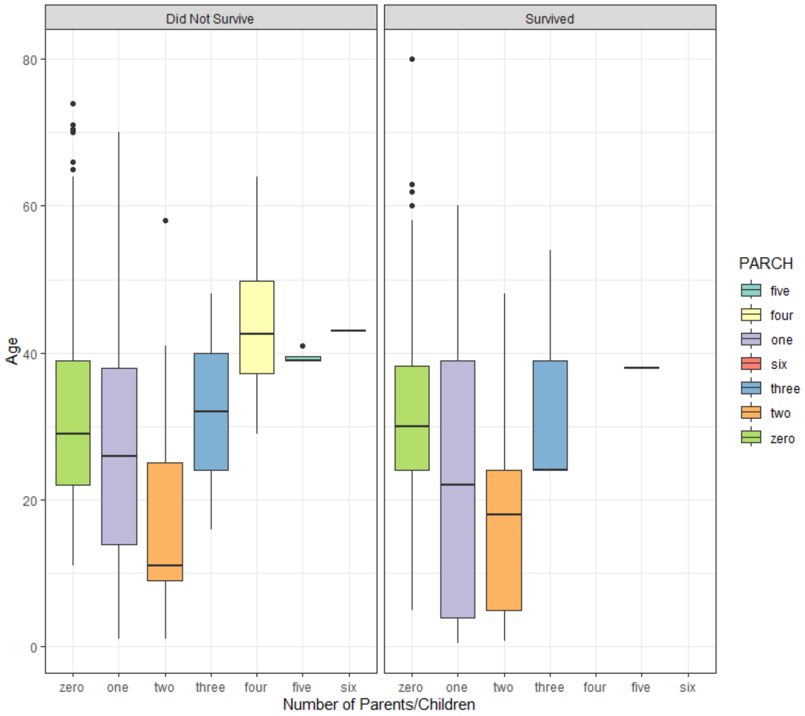


Figure 7: Boxplots of Survival Based on Age and Number of Parents/Children

Figures 6 and 7 in general does show a difference in the ages across different groups. This seems to suggest that there is an association between age and number of parents/children or number of siblings/spouses. However, this could largely be anticipated as the number of passengers who had a larger family is likely smaller than that of the other groups. What must also be taken into consideration is that most families would only have one or two parents and so these observations are mixed in with families who have one or two children. The number of passengers who also only had one sibling and those with a single spouse would also be contained within the same group. This likely accounts for why the spread of the data is so much larger for the first couple of groups in both Figures 6 and 7.

The following boxplots compare the ticket fare of a passenger along with the other qualitative variables. We begin with Figure 8 which details the survival rate of men vs women based on their ticket fare price.

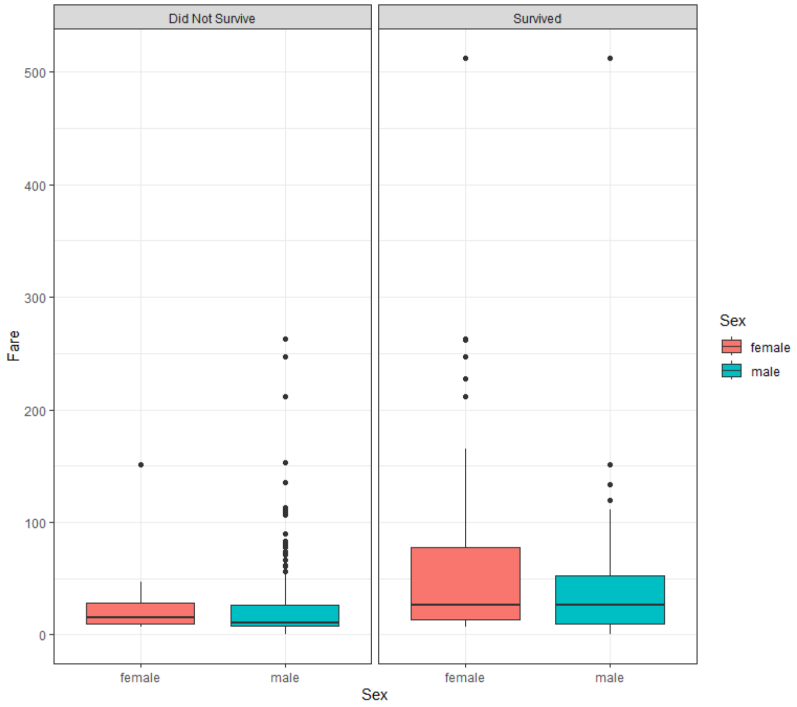


Figure 8: Boxplots of Survival Based on Ticket Fare for Males and Females

Both plots in Figure 8 seem to suggest that there is an association between sex and ticket fare. For the group that did not survive, there are a number of outliers for the males' group while the center is roughly the same between males and females. On the other hand, for those individuals who did survive it appears that females spent more on their ticket fares. Next Figure 9 explores the relationship between the survival rate of lower, middle, and upper class passengers based on their ticket fares.

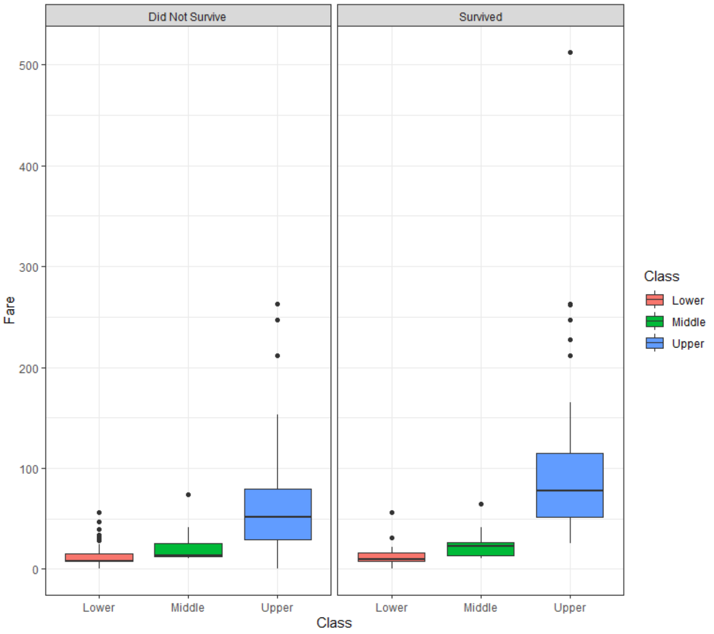


Figure 9: Boxplots of Survival Based on Ticket Fare for Economic Class

As one might largely anticipate, the higher an individual's economic class the higher their ticket fare was. What is more interesting is that the ticket fares for each economic class was largely the same for both the survival and nonsurvival groups. Figure 10 explores the relationship between a passenger's fare and the port from which they embarked.

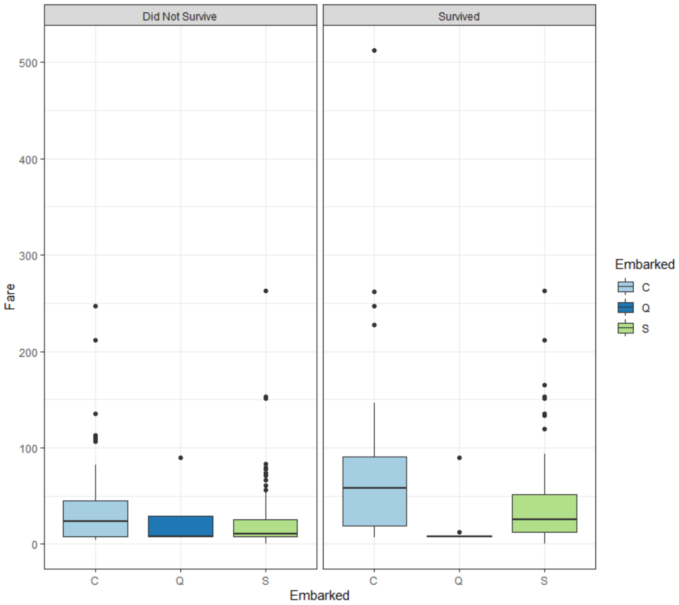


Figure 10: Boxplots of Survival Based on Fare and Port of Embarkment

Figure 10 shows a small association between port of embarkment and ticket fare. However, this is mostly due to the lower fare for those leaving Queenstown. Also interesting is that the fares for the survival group and non-survival group were largely the same between embarkment groups with the exception of the Queenstown group. The fares for those from Queenstown were somewhat smaller than that for the group that did not survive. It should be noted that this result may largely be due to other factors. Finally, Figure 11 compares the fare of passengers and the number of siblings or spouses they had, and Figure 12 shows fare vs number of children or parents a passenger had.

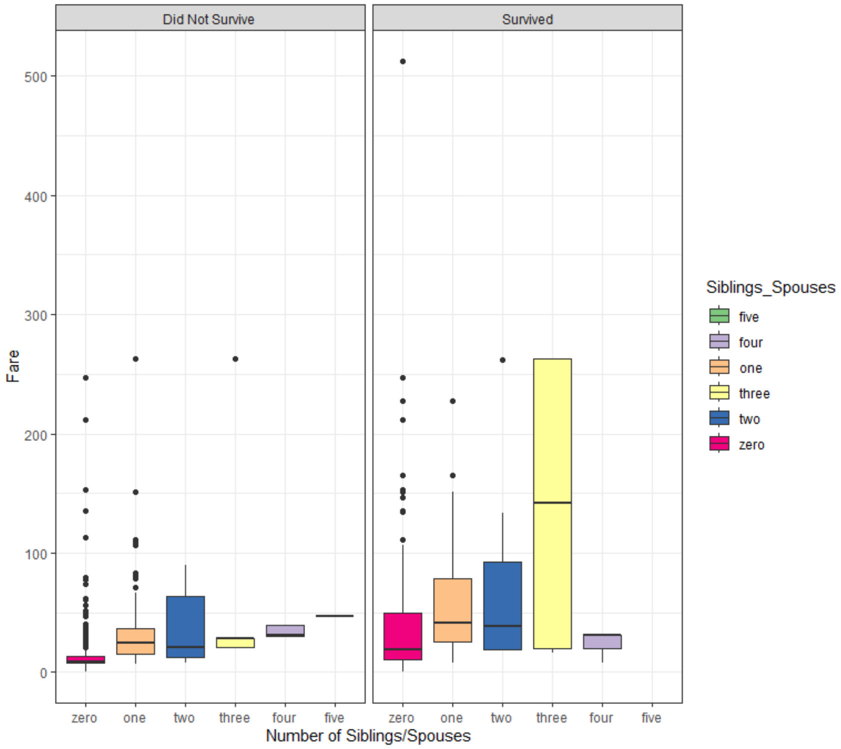


Figure 11: Boxplots of Survival Based on Fare and Number of Siblings/Spouses

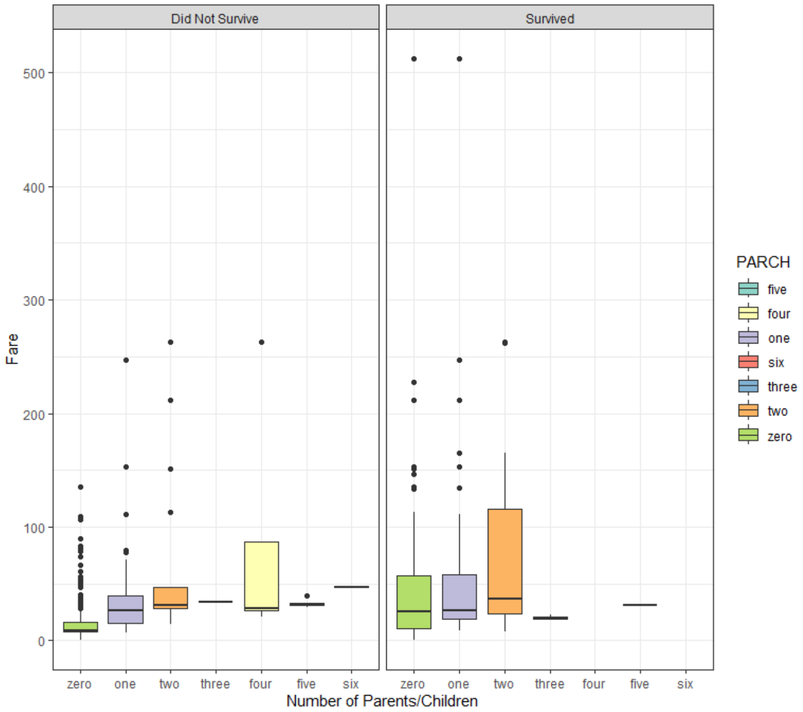


Figure 12: Boxplots of Survival Based on Fare and Number of Parents/Children

In general, as the number of siblings, spouses, parents, or children increases the median fare also increases. Additionally, the fares of the survivors were roughly equal to or greater than that of the nonsurvival group in both Figures 11 and 12. All of this sets us up for the analysis as this shows us that many of the variables are associated with one another.

Logistic regression

First, we implement logistic regression on the Titanic dataset as a baseline for our other models. The full logit model is given below:

$$\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = 2.73 - 1.24 * Ticket\ Class - 0.04 * Age - 0.37 * Number\ of\ Siblings/Spouses - 0.06 * Number\ of\ Parents/Children + 0.002 * Fare + 2.62 * Sex + 0.10 * Embark$$

From the model we can see that as a person's ticket class, age, number of siblings/spouses, and number of parents/children increases a person's log odds of survival decreases. All of these make intuitive sense as it was documented that children were among the first allowed onto the lifeboats, and those with more family would mean a smaller chance for that individual to have a space on a lifeboat. Also of note is that ticket class is recorded as:

$$\begin{cases} Upper\ Class = 1 \\ Middle\ Class = 2 \\ Lower\ Class = 3. \end{cases}$$

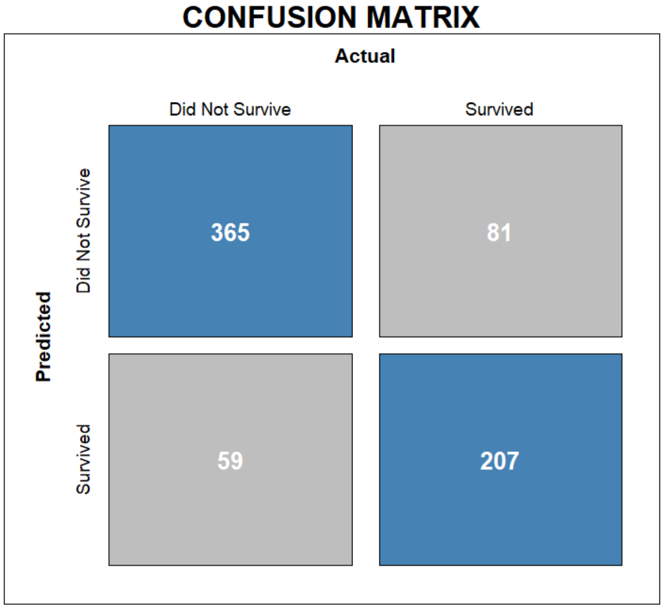
Therefore, this tells us that passengers considered upper or middle class had a higher chance of surviving. Finally, we can see that both sex and port of embarkation have positive slopes while the fare essentially causes little to no effect. This indicates that the more a person spent on a ticket and if the person was a woman then their log odds of survival would increase. Using this we can create a logistic function for $p(X)$ to determine each passenger's probability of survival. The equation using the shortened covariate names is shown below.

$$\hat{p}(X) = \frac{\mathrm{e}^{2.73-1.24*Pclass-0.04*Age-0.37*SibsSp-0.06*Parch+0.002*Fare+2.62*Sex+0.10*Embark}}{1+\ \mathrm{e}^{2.73-1.24*Pclass-0.04*Age-0.37*SibsSp-0.06*Parch+0.002*Fare+2.62*Sex+0.10*Embark}}$$

Now that $p(X)$ has been obtained the threshold for the allocation rule must be determined. How to determine which threshold is best? This in large part depends on the costs of misclassification and background knowledge one has of the subject matter. If we are concerned about incorrectly predicting the survival for individuals who actually survived the Titanic shipwreck, then we may choose to use a lower threshold to account for this. We could also use preexisting knowledge to help determine the threshold. In this case we know from historical records that roughly 1500 of the 2200 passengers did not survive. Thus, it would make sense to use a threshold of roughly $\hat{p}(X) = 0.68$ where any observation's probability exceeding this value would be classified as having survived. Another argument can be made for $\hat{p}(X) = 0.5$ since there were missing values in the data. These missing values are likely not missing at random. The cabin variable for example, which was deleted, is probably more likely to be listed for an individual who survived. As such, this variable's removal potentially loses valuable information. Additionally, several observations were removed due to missing values for their age. This too could either lower or increase the likelihood of survival in the dataset. Therefore, since there is some uncertainty on how the probability of survival was affected from deletion of missing data a natural and conservative choice of threshold to allocate each $\hat{p}(X)$ would be $\hat{p}(X) = 0.5$. In other words,

$$\begin{cases} Classify\ passenger\ as\ not\ surviving\ if\ \ \hat{p}(X) \leq 0.5 \\ \ \ Classify\ passenger\ as\ surviving\ if\ \hat{p}(X) > 0.5. \end{cases}$$

A confusion matrix for this choice of probability threshold is shown in Figure 13.



	Did Not Survive	Survived
Did Not Survive	365	81
Survived	59	207

Figure 13: Confusion matrix for $\hat{p}(X) = 0.5$ threshold

This decision rule leads to an accuracy of 80.34%. The next evaluation tool that will be used is an ROC curve. For this dataset, the true positive rate, also known as the sensitivity, is taken to mean the fraction of survivors who were correctly identified as survivors for varying thresholds. The false positive rate, which is 1 – specificity, is taken to mean the fraction of nonsurvivors who were mistakenly classified as survivors. The linear gray line represents a classifier that randomly guesses. In other words, if survival was in no way associated with the age, sex, or other covariates then we would obtain this line. The ROC curve is shown in Figure 14 below.

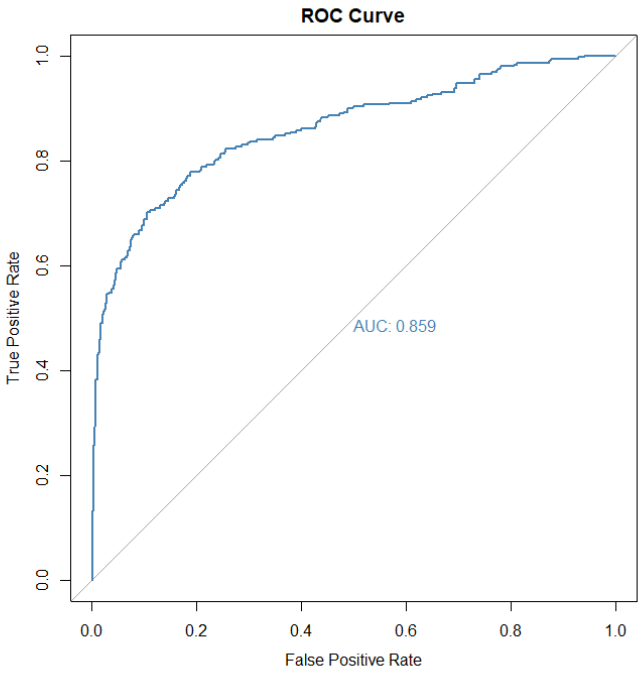


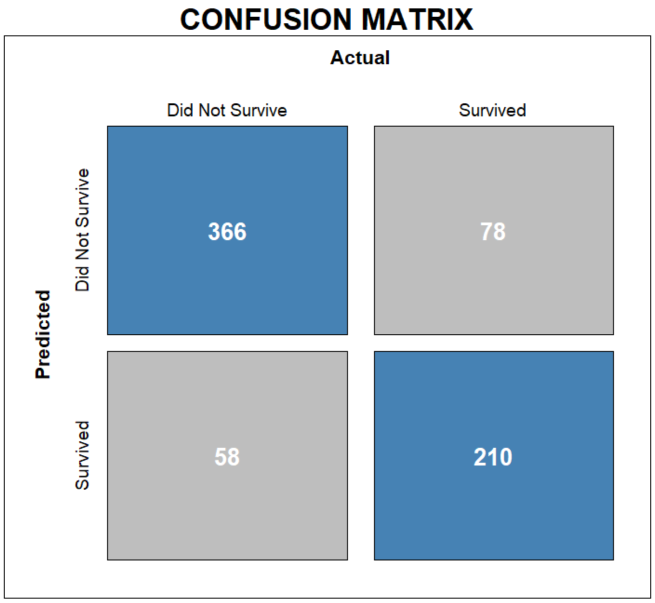
Figure 14: ROC Curve for Logistic Regression

For this particular curve the AUC is found to be approximately 0.859 which is reasonable.

Next, one additional tool that is afforded by logistic regression is model building. In the full model only the ticket class, age, number of siblings/spouses, and the passenger's sex were significant predictors to determine if a passenger survived. As such, a refined model is created using backward elimination with an $\alpha = 0.05$ significance level. The corresponding logistic function is given below:

$$\hat{p}(X) = \frac{\mathrm{e}^{2.98-1.31*Ticket\ Class-0.04*Age-0.37*Number\ of\ Siblings/Spouses\ +2.61*Sex}}{1+\ \mathrm{e}^{2.98-1.31*Ticket\ Class-0.04*Age-0.37*Number\ of\ Siblings/Spouses\ +2.61*Sex}}$$

The corresponding AIC for the logit model is 646.18 which is smaller than 651.05 for the full model. Next, using a threshold of $\hat{p}(X) = 0.5$ to allocate each observation we obtain the following confusion matrix:

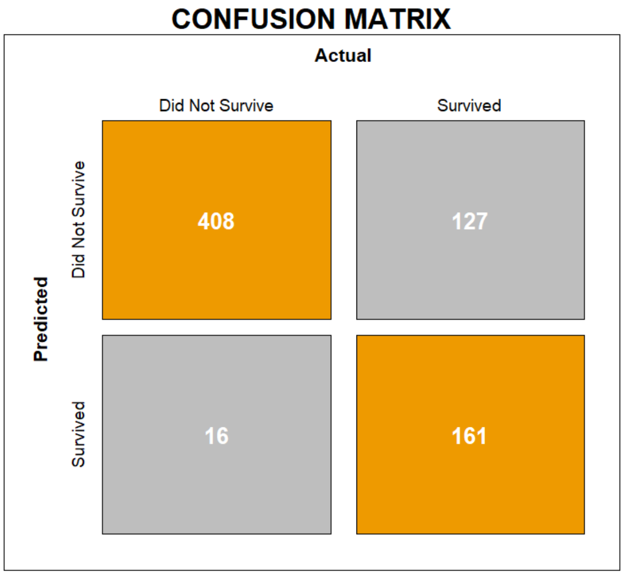


	Did Not Survive	Survived
Did Not Survive	366	78
Survived	58	210

Figure 15: Confusion Matrix for Refined Logistic Regression Model

A slight increased accuracy of 80.9% is obtained for the refined model which comes from the additional 4 passengers who were correctly classified. Additionally, the resulting ROC has an AUC almost identical to that of the full model. Therefore, the refined model leads to slightly better results.

Finally, for completeness we can also check a decision rule that uses $\hat{p}(X) = 0.68$ which uses preexisting knowledge of the population of Titanic passengers and compare the results. The confusion matrix for the refined logistic regression model using $\hat{p}(X) = 0.68$ threshold is shown in Figure 16.



CONFUSION MATRIX

		Actual	
		Did Not Survive	Survived
Predicted	Did Not Survive	408	127
	Survived	16	161

Figure 16: Confusion Matrix for Refined Logistic Regression Model Using $\hat{p}(X) = 0.68$

The accuracy of this model is approximately 79.9% which is slightly lower than that of the decision rule using $\hat{p}(X) = 0.5$. By increasing the threshold this decision rule leads to a higher number of observations being correctly classified as having not survived. In fact, only 16 observations were misclassified for individuals who did not survive in this dataset. However, the tradeoff for doing this is that a larger number of individuals who actually survived were misclassified. This then leads to the question of which threshold is better? The answer lies in the interest of the researcher. If there was a larger cost associated with wrongly classifying a passenger as surviving, then $\hat{p}(X) = 0.68$ decision boundary is best. Otherwise, the conservative $\hat{p}(X) = 0.5$ is best.

Naive Bayes

Next, we implement the Naive Bayes classifier on the Titanic dataset. First the prior probability is estimated. A general data free prior is constructed first and then updated based on the data. As was discussed before, the true probability of nonsurvival was approximately 0.68 so it would be reasonable to start off our prior as $\pi_{survived} = 0.32$. However, upon closer inspection of the data itself, the prior is found and updated to be $\pi_{survived} = 0.40$. This tells us that the probability of survival in our dataset is higher than for the general population of Titanic passengers. This could be due to the deletion of certain observations with missing ages. It is not unreasonable to think that passengers with missing ages were more likely to not have survived the accident and tell others. While this is not central to the analysis provided by Naive Bayes it is nonetheless an important fact to consider its implications and a benefit of the Naive Bayes flexibility when estimating the prior. Next the density functions for the quantitative predictors are estimated using a Gaussian distribution though other options are available as previously discussed. Finally, the conservative threshold of $\hat{p}(X) = 0.5$ is used to allocate each observation to the survival or nonsurvival class. This is equivalent to allocating an observation to the class whose probability is highest. The resulting confusion matrix is shown in Figure 17:

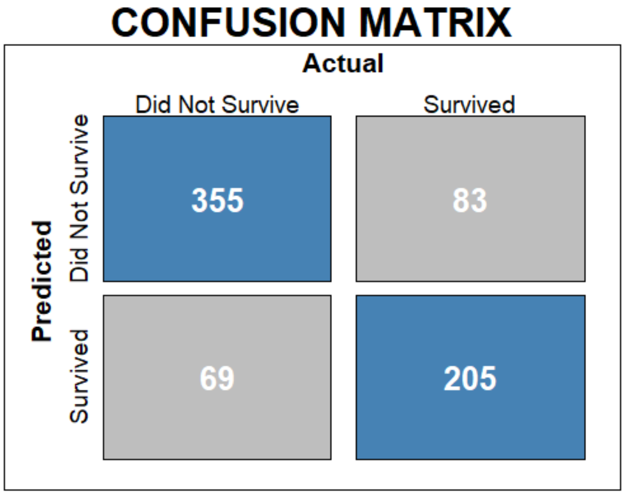


Figure 17: Confusion Matrix for Naive Bayes Model

This decision rule leads to an accuracy of 78.65% for the training data. When compared with the results of the refined logistic regression model it is apparent that the Naive Bayes performs slightly worse than the logistic regression.

Another method to compare the results of Naive Bayes with the other methods is to use an ROC curve as shown in Figure 18.

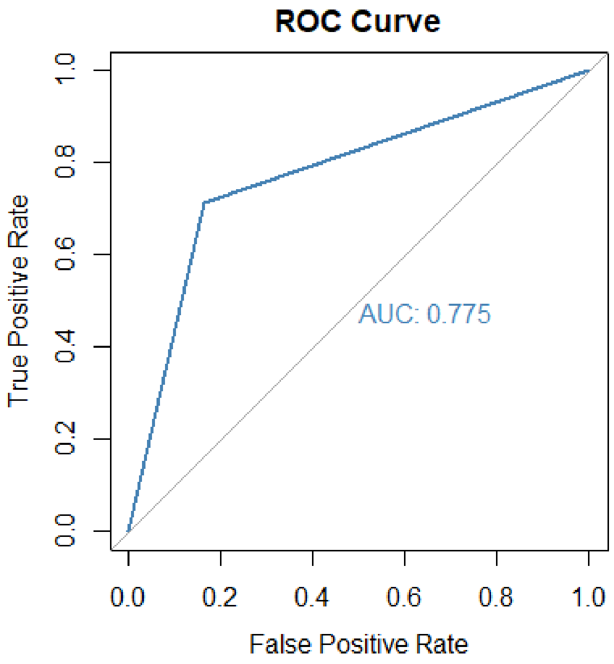


Figure 18: ROC Curve for Naive Bayes

When compared with the ROC of the logistic regression it is apparent that the curve produced by the Naive Bayes does not hug the left corner as well and thus directly leads to a smaller AUC value of 0.775.

It is not all that unsurprising that Naive Bayes does not outperform logistic regression. For one, the assumption that the covariates are independent is likely not justified here. An example of two variables which might not be independent is a passenger’s socioeconomic status and the price they paid for their ticket. It would not be unreasonable to think that someone with a higher socioeconomic status would pay for a higher ticket price to ensure additional luxuries and benefits on the trip, as opposed to someone of a lower socioeconomic class. Another potential reason why Naive Bayes underperforms logistic regression stems from the fact that there are n = 712 observations and p = 7 covariates present in the data. As is mentioned by James et. al, “we expect to see a greater pay-off to using Naive Bayes […] in instances where p is larger or n is smaller, so that reducing the variance is very important” (158).

Normal Population Methods

Before LDA and QDA are implemented in this section the normality assumption is formally tested for completeness using the Mardia’s Test. The hypothesis test is outlined below.

$$\begin{cases} H_0: X \sim MVN \\ H_1: X \nsim MVN \end{cases}, \alpha = 0.05$$

The Mardia's test statistic for multivariate skewness (corrected for small samples) is found to be approximately 7.3466×10^3, and Mardia's test statistic for multivariate kurtosis is approximately 76.6801. The corresponding p-values for each of these results are approximately zero. Therefore, the null hypothesis (H_0) is rejected at a significance level of $\alpha = 0.05$ and it can be inferred that the predictor variables do not follow a multivariate normal distribution $(X \nsim MVN)$.

Recall that Johnson and Wichern's second recommendation for dealing with nonnormal data is to simply move forward with the normality assumption even if the data is not normal. Very often the results can give more information to the researcher and see if the method works reasonably well. We will therefore proceed with this and implement both LDA and QDA on the Titanic dataset and see if the results are reasonable. First, the prior probability is estimated. In the Naive Bayes implementation, it was found that the prior for the survival group is approximately $\pi_{survived} = 0.40$. Next the density functions for the predictors are estimated assuming multivariate normal distributions and equal covariance matrices. Finally, the conservative threshold of $\hat{p}(X) = 0.5$ is used to allocate each observation to the survival or nonsurvival class. The resulting confusion matrix is shown in Figure 19:

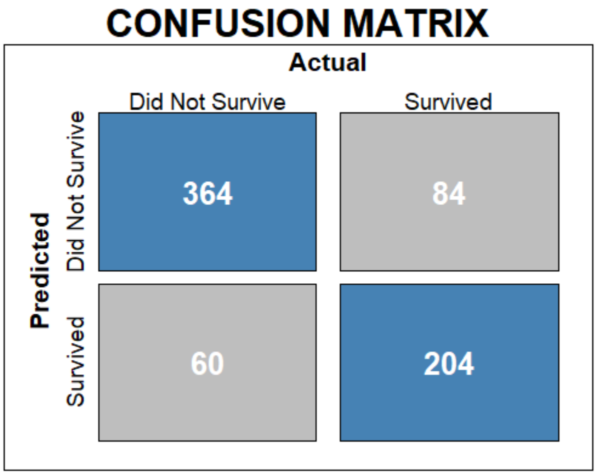


Figure 19: LDA Confusion Matrix

This decision rule leads to an accuracy of 79.77% for the training data. When compared with the previous results we can see that LDA performs rather similarly to the refined logistic regression and Naive Bayes models. Next, the ROC curve is given in Figure 20.

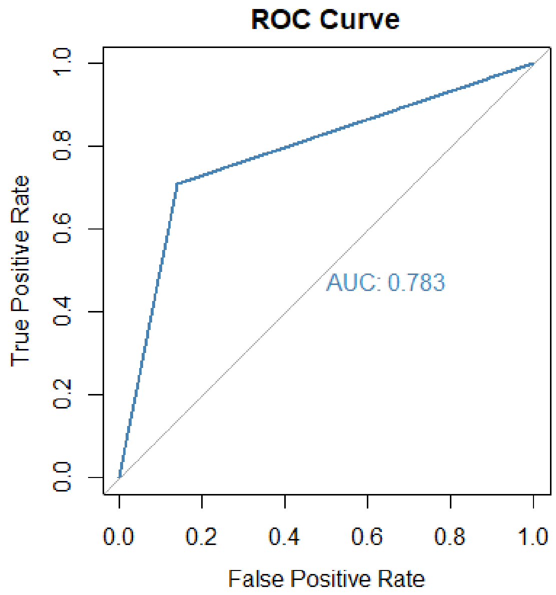
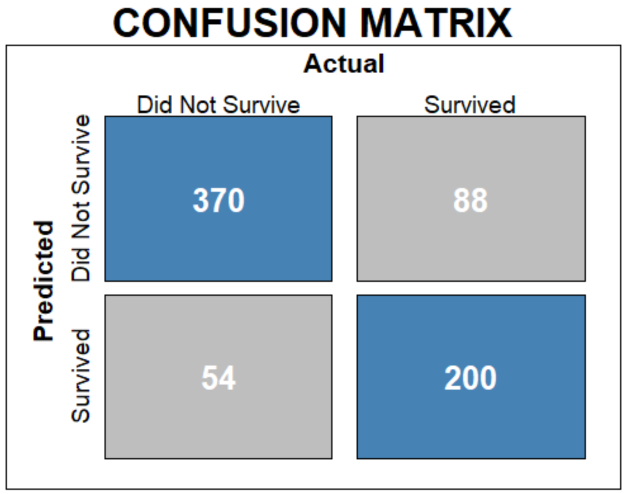


Figure 20: LDA ROC Curve

The AUC is approximately 0.783 which tells us that LDA performs worse than logistic regression and slightly better than Naive Bayes for an ROC curve. Next the results for QDA are given below in Figure 21. Again, recall that now the assumptions are that the density functions for the predictors are multivariate normal, and we do not assume equal covariance matrices.



CONFUSION MATRIX

Actual

Predicted	Did Not Survive	Survived
Did Not Survive	370	88
Survived	54	200

Figure 21: QDA Confusion Matrix

Using the conservative threshold of $\hat{p}(X) = 0.5$ to allocate each observation to the survival or nonsurvival class gives an accuracy of 80.06% on the training data. As with LDA, this is very similar to the accuracy for the other methods. Finally, the ROC curve for QDA is investigated in Figure 22 and an AUC of 0.784 is attained which is almost identical to LDA.

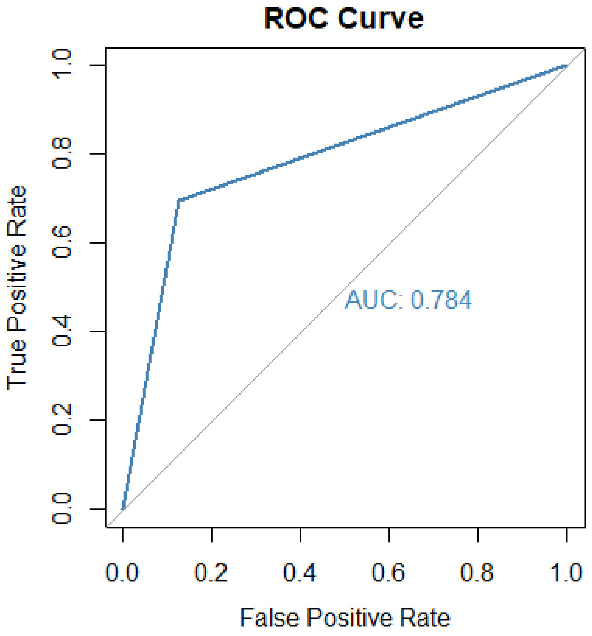


Figure 22: QDA ROC Curve

Overall, there does not seem to be a big difference between the results of LDA and QDA for the training data.

Additionally, based on the assumptions it is apparent that the results are reasonable and not drastically different from the other methods. Therefore, even though the data do not fit the critical assumptions the results are still reasonable.

Classification Trees

Even though classification trees allow for easy handling of missing data for both simplicity and ease of comparison with the other methods in this paper this was not done here. Instead, the same transformed dataset was used as in the previous methods. The resulting classification tree is shown in Figure 23.

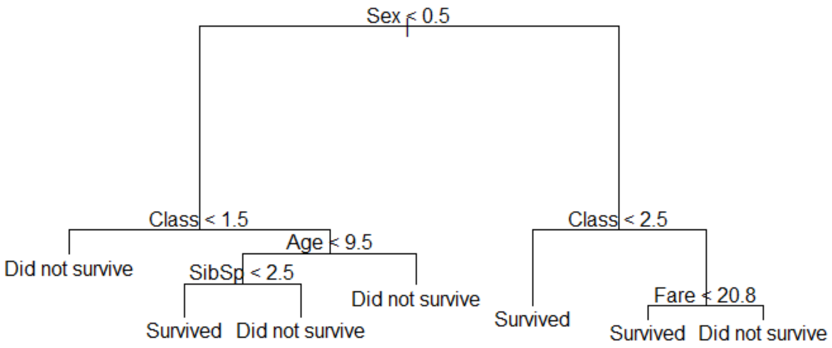
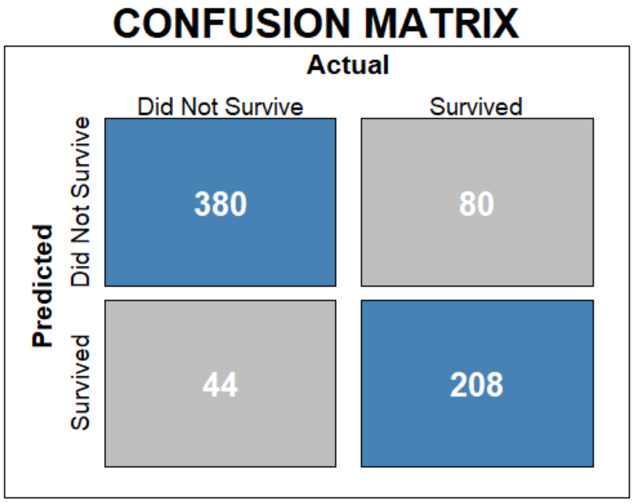


Figure 23: Classification Tree for Titanic Survival

The first split occurs at the top of the tree with the Sex variable. This indicates that the most important factor in determining a passenger's survival is the sex of the person. The first split also indicates that we should pass to the left branch if the person's sex is less than 0.5. Since males are encoded as zeros and females are encoded as ones, this tells us that if the individual is a male that they should pass to the let branch and pass to the right branch if they are a female. Also of interest is the fact that the tree only deems a person's sex, ticket class, age, and number of siblings/spouses is important in the classification. Notice that these are the same predictors that were found to be significant and used in the refined logistic regression model.

Next, the confusion matrix for the classification tree is shown below in Figure 24.



CONFUSION MATRIX

Predicted	Actual: Did Not Survive	Actual: Survived
Did Not Survive	380	80
Survived	44	208

Figure 24: Confusion Matrix for Classification Tree Model

The decision rules from the classification tree leads to an accuracy of 82.58% for the training data. This outperforms the accuracy of the previous methods. However, one should remember that if we wanted to properly evaluate the performance of the classification tree, or any of the other methods for that matter, the correct approach would be to compute the error for a testing set. While we do not have a testing set this is alright as the central focus of this paper is the classification procedure and creation of the rules. Another method to compare the results of the previous methods is with the use of an ROC curve as shown in Figure 25.

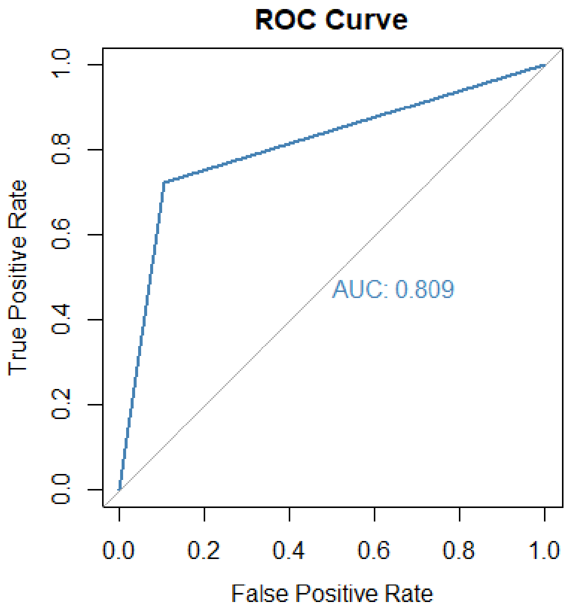


Figure 25: ROC Curve for Classification Tree

When compared with the ROC of the logistic regression it is apparent that the curve produced by the classification tree does not hug the left corner as well and thus directly leads to a smaller AUC value of 0.809. However, it does outperform the AUC of the Naive Bayes, LDA, and QDA.

Finally, we can attempt to prune the given tree in order to see if there is a more optimal model. A plot of the cross-validation errors vs the number of terminal nodes is given below.

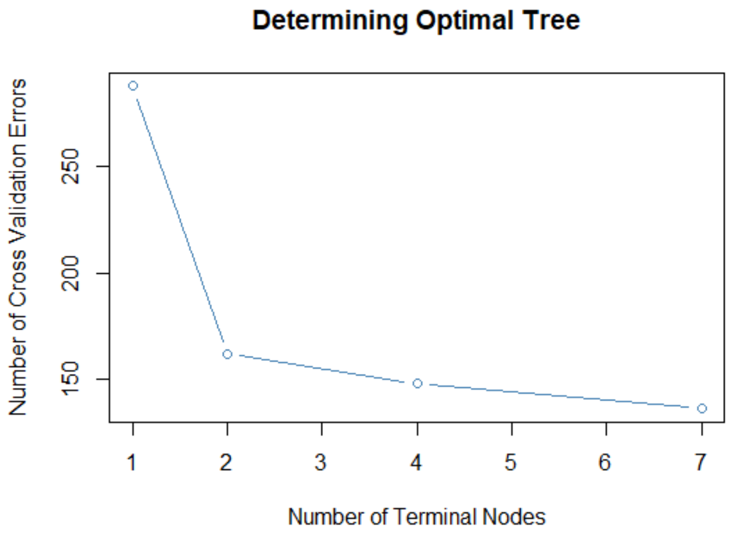


Figure 26: Determining Optimal Tree

As it turns out our original tree size is optimal and therefore does not need to be pruned. The minimum number of cross validation errors occurs for the classification tree which was previously obtained.

4 Conclusions and Future Research

The results of the different method prediction accuracies and their AUC are provided in Table 2 below.

Model	Prediction Accuracy of Training Set	Area Under Curve
Full Logistic Regression $\hat{p}(X) = 0.5$	0.8034	0.859
Refined Logistic Regression $\hat{p}(X) = 0.5$	0.809	0.858
Refined Logistic Regression $\hat{p}(X) = 0.68$	0.799	0.858
Naive Bayes	0.7865	0.775
Linear Discriminant Analysis (LDA)	0.7977	0.783
Quadratic Discriminant Analysis (QDA)	0.8006	0.784
Classification Tree	0.8258	0.809

Table 2: Summary of Results

We have now discussed four of the many possible approaches to tackling classification problems. The major difference between the suggested parametric and nonparametric approaches was the use of a probability to create the classification rules. Based on the results for this dataset it appears that the nonparametric approach of classification trees produces slightly better accuracies for the training dataset. However, while the assumptions for Naive Bayes, LDA, and QDA likely do not hold for this classification problem and thus likely explain why they performed slightly worse, none of the methods stood out as being significantly better than the others. It should also be noted that "it is a mistake to generalize from a single dataset. Different classification methods are likely to do well for certain kinds of data but perform poorly for others. It's always a good idea to try more than one" (Faraway 360). Additionally, Faraway states that "the insistence on specifying a model, right from the start, does limit statistics. It is often difficult to specify a model, particularly for larger and more complex datasets" (343). This helps shed light on the battle between classical statistical approaches and the more modern machine learning approaches. This paper lacks the formal use of a testing dataset to compare accuracies of the different methods. This, however, is not a major issue as the focus here is on the model building and classification rule creation rather than the importance of prediction accuracies so typical for machine learning algorithms. However, other approaches can be taken, and some possible ideas will be discussed in the future research ideas below.

The first major issue that was encountered at the beginning of this paper was the missing data. Using the classification trees to handle the missing data could be one approach to tackling this problem. However, there are other more formal approaches that could also be implemented. Another component that was missing was the use of a test or validation set to help build the best classification model. To accommodate this one idea would be to split the original training dataset into a new training and testing dataset. Alternatively, since the goal of the competition was to correctly predict the responses in a testing dataset which did not have the true labels whether an individual survived or not, an unsupervised learning approach can also be taken by using k-means, hierarchical clustering, or other clustering techniques. This would therefore become a separation problem rather than a classification problem. Other ideas could include introducing other methods, and perhaps even opening the door to machine learning classification methods, such as neural networks, even though many of these focus on prediction accuracy rather than the model and inference. While these examples are not exhaustive, they do provide a good starting point in ways to improve the methods examined in this paper and other new approaches which might help better classify whether or not a person survived the Titanic shipwreck.

5 References

Faraway, Julian. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Second ed., CRC Press Taylor & Francis Group, 2016.

James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. Second ed., Springer, 2021.

Johnson, Richard, and Dean Wichern. *Applied Multivariate Statistical Analysis*. Sixth ed., Pearson Education, 2019.

Tikkanen, Amy. “Titanic”. *Britannica*, Britannica, 17 March 2024, https://www.britannica.com/topic/Titanic. Accessed 18 March 2024.