Introduction:

The sinking of the Titanic was one of the most recognizable moments in our history. Although this was a tragic event with many lives lost, information based on certain characteristics can provide valuable insights associated with survival with the ill-fated voyage. These insights can be useful for historical analyses on other events. To evaluate characteristics associated with survival, using classification machine learning models to help determine if a passenger survived or not based on certain attributes can help historians narrow down the indicators of what leads to survival. In this study, three classification methods were employed: Logistic Regression, Gaussian Naïve Bayes, and Bernoulli Naïve Bayes. Based on the results, it is recommended that a Logistic Regression classification model is used until other classification models can be explored.

Data Preparation:

Data was imported via CSV, which contained 1309 observations (rows) and 11 explanatory variables (columns). CSV files contained two data sets, one for training and the other for testing, where the split was about 68% and 32% (891 and 418 observations, respectively). Additionally, a few non-values were discovered in the training data (tbl. 1). Data for the 177 missing age values were imputed by the mean age (rounded) by name title. Name titles were also added. The 687 missing cabin values were first imputed with the code U0, and a variable called 'Decks' was created with the letters in each cabin value, which indicated the deck they were on. From decks A thru G, letters were mapped to numbers (1 to 7) for modeling. Additionally, missing deck values were not imputed with 0, but were instead imputed with a value based on a random choice generator that was applied by passenger class. Though not recommended, as it comprised 77% of the observations, the imputation seemed to have helped as validation scores were acceptable. To confirm that the imputation was viable they were compared with the distribution percentages with the available data, and the percentages were close (tbl. 2). The remaining missing values were imputed with the most common values by passenger class, as there were not many.

Density plots showed that fares were highly skewed to the right with a distribution plot showing significantly high kurtosis of 29.95 (fig. 1, fig. 2). Fares were therefore split into bins and categorized from levels 0 to 5 to even out the distribution (fig 3). Though ages only showed a kurtosis of 3.68 and a skewness of 0.38 (fig.4), they were binned and categorized from levels 0 to 6 as well (fig. 5). Box plots showed significant outliers with passengers with parents and children (fig. 6). Parents and children were combined with siblings and spouses and were created into a variable indicating number of relatives. Variables were then dropped, and extreme outliers for relatives were 47 compared to 213 for parents and children (tbl. 3). An interaction term, age times passenger class was added as a two-way ANOVA showed that the interaction effect between the two in relation to survival rate was significant with a p-value of 0.025 (tbl. 4). The variable calculating fare per person by number of relatives was also added. Other variables that were dropped included names, cabin number, and ticket number were dropped as they could not be converted into useful categories, as it was assumed that they would add noise to the model. Name titles of similarity were combined such as males and females with titles of nobility or respect. 18 dichotomous variables were then created based on gender, passenger class, fare category, and name title.

Data Analysis

Initial exploratory analysis showed characteristics such as age, gender, passenger class, deck, and fares paid were associated with survival rates. Descriptive statistics showed that the mean age of passengers were about 30 years of age and the median fare paid was around $14.50 (tbl. 5). A closer look at the data showed that mean price paid for those that survived and died was around $48 and $22, respectively (fig. 13). Median statistics were computed by name title which showed that most passengers had the name title 'Mr.' with a passenger count of 512, and a median price paid of $9.35 (fig, 7, tbl. 6). Only 81 of those passengers survived with a median fare of $26.28 paid with a median of 2nd class (tbl. 7). The highest number that survived were women that had the name title 'Miss' with a count of 127, median fare of $19.50, and a median age of 22 (fig. 8). Additionally, females with titles of nobility survived, while males perished. All reverends went down with the ship. The highest percentage of passengers that died by age group with the least amount of survival were in the 30 to 35 range (fig. 9), with the highest proportional percentage being passengers under 15 (over 50% survived). Most males aboard the ship perished in comparison to females (fig. 10), with 74% of females surviving, and 81% of males died. Percentage by class shows that the highest proportion of survived were in first-class (fig. 11) where 63% survived and 75.8% of third-class passengers died. Looking at the port of embarkation the highest proportion of passengers that died were from Southhampton (fig. 12). Further examination of survival by fares using a swarm plot showed how significant the difference was (fig.14). A scatter matrix showed that survival rates were higher in the upper decks, which is consistent with fares and passenger class (fig.15). Characteristics such as fares, age, gender, passenger class, deck number, name title, and number of relatives all played a role in the classification models employed.

Binary Classifiers: Methods and Evaluation

Training data was tested utilizing three classification methods: Logistic Regression, Gaussian Naïve Bayes, and Bernoulli Naïve Bayes. Models were tested using a 10-fold cross-validation design using the ROC AUC, accuracy, precision, recall, and F1 scores as evaluation metrics to gauge the performance of each model. While results explained here only provide a brief summary of the methods employed, additional details are provided from page 21 and on, as well as in the appendix. Additionally, models were also tested using the ROC, precision versus recall, precision and recall versus threshold, and learning curves (training set sizes: 80, 260, 440, 620, 801) to assess and visualize validation scores. Confusion matrices were also used for comparison between each method.

Results and Recommendations:

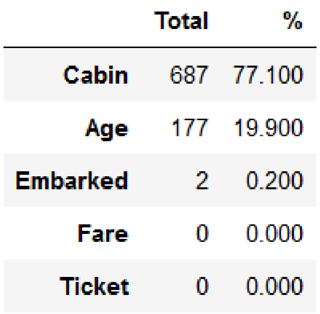
Results from the 10-fold cross-validation tests informed us that the Logistic Regression model using l_2 regularization (at its highest) performed the best with a mean test ROC AUC score of 0.871, accuracy score of 0.836, with precision and recall at 0.818 and 0.737. The average test F1 score was 0.773. The AUC value of 0.871 indicated that there were fewer type I and fewer type II errors, and that the model was a good measure of separability as it reproduced the data well. The coefficient magnitudes were not too large, they did not have a wide spread, and were not pushed too close to zero which may indicate that this model had good predictability (considering the scores) while treating most coefficients as having meaningful magnitude (fig.16). However, upon examination of the decision boundaries between age and fare, there seemed to have been some overlap, which may have made classification difficult (fig. 17). The confusion matrix showed that there were 493 true negatives, 56 false positives, 90 false negatives, and 252 true positives (fig.18). The ROC curve compared to a random classifier, confirmed that the classification method reproduced the data well as the curve stayed away from the random classifier line reaching towards the top-left corner indicating good separability (fig.19). As there were fewer positives (survived) than negatives (died) the PR (precision/recall) curve was examined, which showed that the curve was close to the top right corner indicating that the classifier fairly worked well (fig. 20). The PR versus threshold plot showed that at a threshold score of 0, there was a good tradeoff, which was inline with the average precision and recall score results (fig. 21). It is recommended that the threshold remains unchanged. Although the learning curve showed a pretty wide gap between cross-validation scores and training scores at the beginning, the gap narrowed as the training set sizes grew while maintaining high enough ROC AUC scores indicating that the model generalized well (fig. 22).

Cross-validation scores from the Gaussian Naïve Bayes classifier did not show better results as the Logistic Regression method. The mean test ROC AUC score was 0.829, accuracy score was 0.763, with a precision and recall score of 0.676 and 0.740, and an F1 score of 0.705. The confusion matrix showed higher false positives, 122, with 427 true negatives, 89 false negatives, and 253 true positives, which was significantly different than the Logistic

Regression model (fig. 23). While the ROC curve (fig. 24) showed that the model was a good measure of separability, due to the number of false positives, the PR curve showed that much improvement was needed, which is also supported by the precision and recall scores (fig. 25). Therefore, as the PR versus threshold plot shows, the threshold could be changed for better scoring (fig. 26). The learning curve showed similarity to the Logistic Regression model, yet the gap between cross-validation scores and training scores were much wider at the smaller training set sizes, then narrowed as the cross-validation scores rose, but the gap started to widen again as cross-validation scores started to fall (fig. 27). This indicated that the model would not generalize as well as the Logistic Regression model, and further indicated that a more complex model is needed.

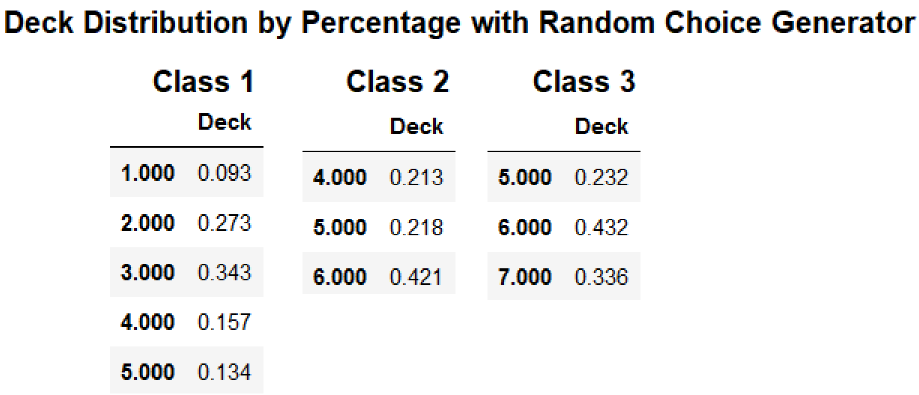
Initial results from utilizing the Bernoulli Naïve Bayes classifier showed improvement from the Gaussian Naïve Bayes classifier, yet did not do as well as the Logistic Regression model. The mean test ROC AUC score was 0.841, accuracy score of 0.773, with a precision and recall score of 0.687 and 0.752, and an F1 score of 0.717. The confusion matrix showed slightly less false positives, 117, than the Gaussian Naïve Bayes classifier, but was still significantly higher than the Logistic Regression model (fig. 28). Although the ROC curve (fig.29) looked similar to the Gaussian Naïve Bayes classifier, the number of false positives was still an issue, as the PR curve showed a significant drop in precision at around 0.60 recall (fig. 30). This is also evident upon examination of the PR versus threshold curve as recall is significantly higher than precision at the 0.50 mark (fig. 31). The learning curve was in stark contrast to the other models, as the cross-validation scores were higher than training scores for some training sets (fig. 32). The curve also shows that the scores cross at two points. This is a sign of underfitting which is evident when looking at the learning curve using the mean RMSE as the scores are fairly high, the gap between the curves are narrow, and they rise and plateau while remaining close (fig. 33). The ROC plot comparing all three curves (fig. 34) clearly shows that using Logistic Regression as a classification method is the better of the three. It stays furthest away for the random classifier line showing that the model had the best measure of separability, which is also evident in the confusion matrix, as it had significantly less false positives. In addition, the methods were further tested on the test data set, which produced an accuracy score of 0.80382 for Logistic Regression, while Gaussian Naïve Bayes scored a 0.75598, and Bernoulli Naïve Bayes scored 0.7461 (fig. 35). Therefore, from these results it is recommended that Logistic Regression is used over the other models that were evaluated, for further insights for historical analyses regarding characteristics associated with survival until other methods can be explored that provide better predictive accuracy.

Table 1 – Missing Data from Training Set:



	Total	%
Cabin	687	77.100
Age	177	19.900
Embarked	2	0.200
Fare	0	0.000
Ticket	0	0.000

Table 2 – Distribution Percentages of Deck Imputations by Passenger Class:



Deck Distribution by Percentage with Random Choice Generator

Class 1

	Deck
1.000	0.093
2.000	0.273
3.000	0.343
4.000	0.157
5.000	0.134

Class 2

	Deck
4.000	0.213
5.000	0.218
6.000	0.421

Class 3

	Deck
5.000	0.232
6.000	0.432
7.000	0.336

Actual Distributions:

Deck Distribution, Decks 1 to 5, By Percentage in Ascending Order by Deck Passenger Class 1:

1	0.091
2	0.267
3	0.335
4	0.165
5	0.142

Deck Distribution, Decks 4 to 6, By Percentage in Ascending Order by Deck Passenger Class 2:

4	0.250
5	0.250
6	0.500

Deck Distribution, Decks 5 to 7, By Percentage in Ascending Order by Deck Passenger Class 3:

5	0.250
6	0.417
7	0.333

Figure 1 – Density Plots:

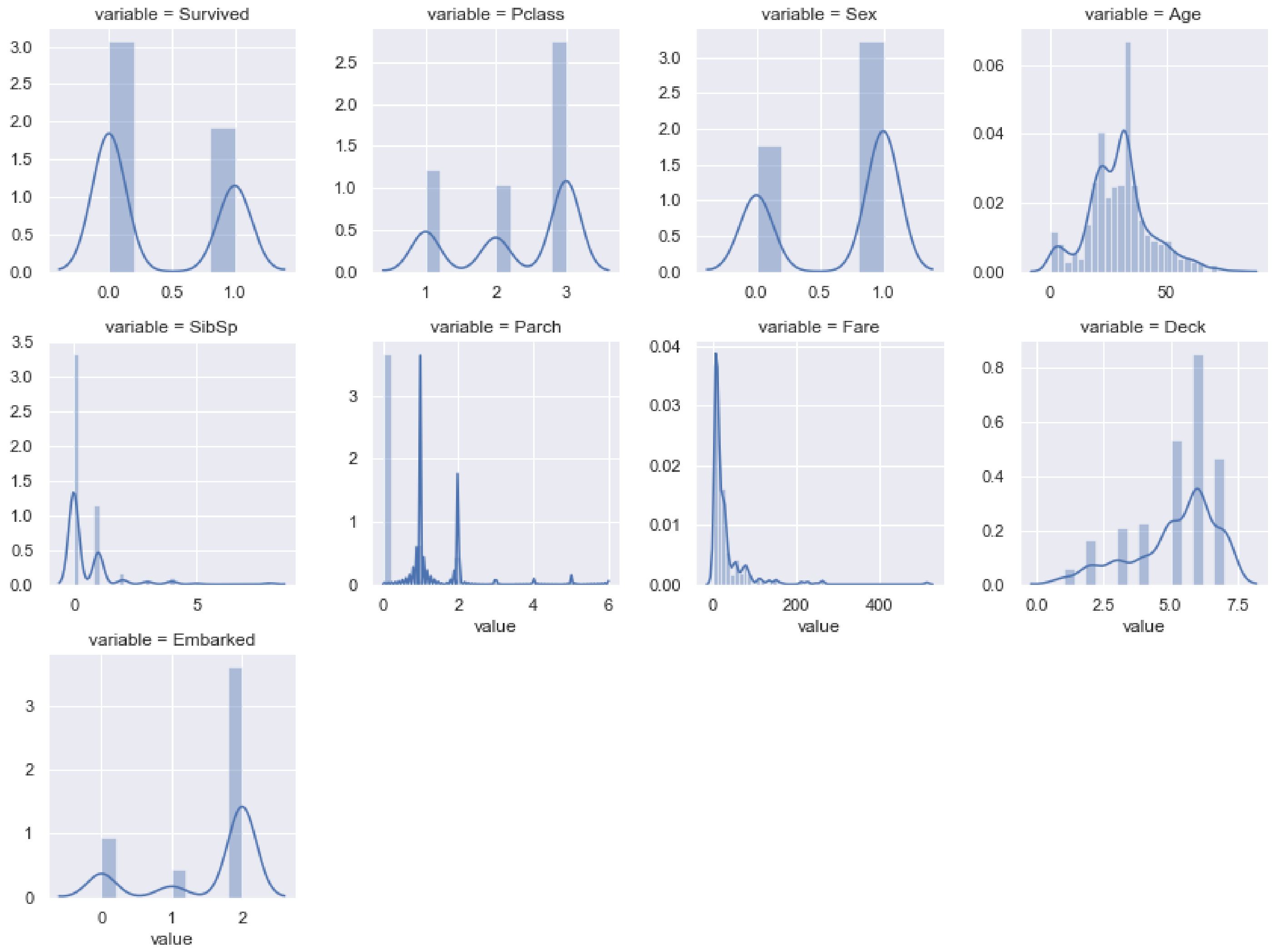
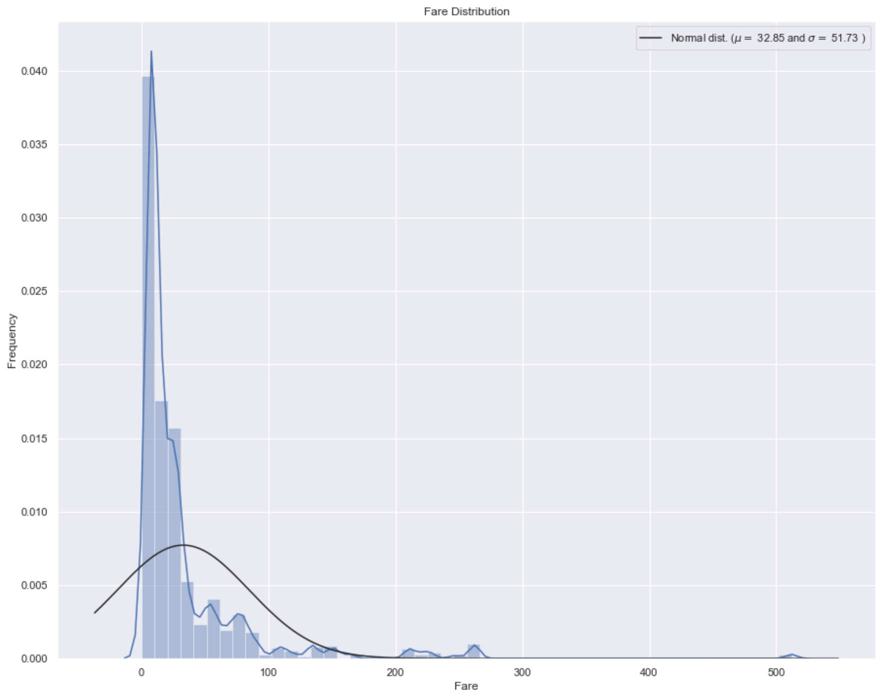


Figure 2 – Distribution and QQ Plot of Fares:



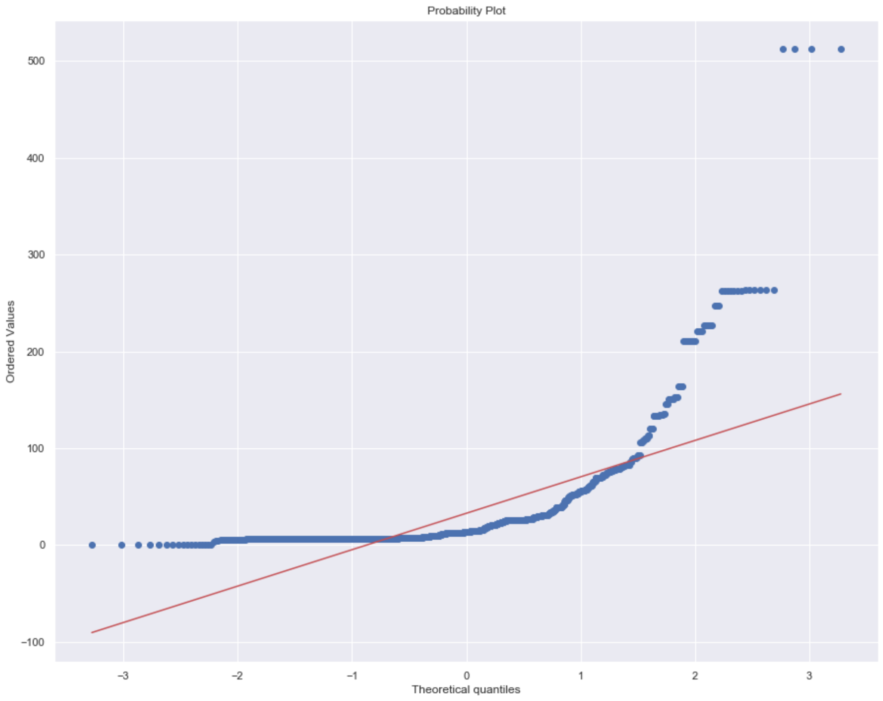


Figure 3 – Fare Distribution Plot After Transformation:

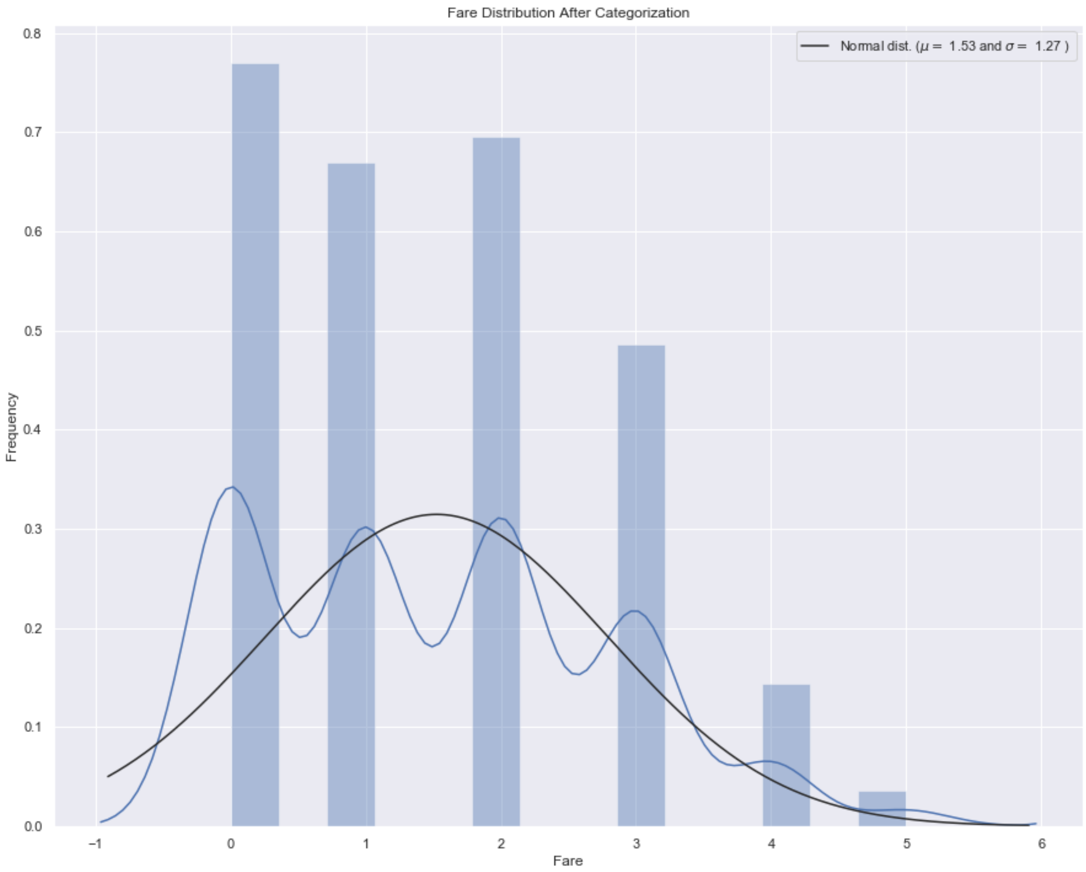


Figure 4 – Age Distribution:

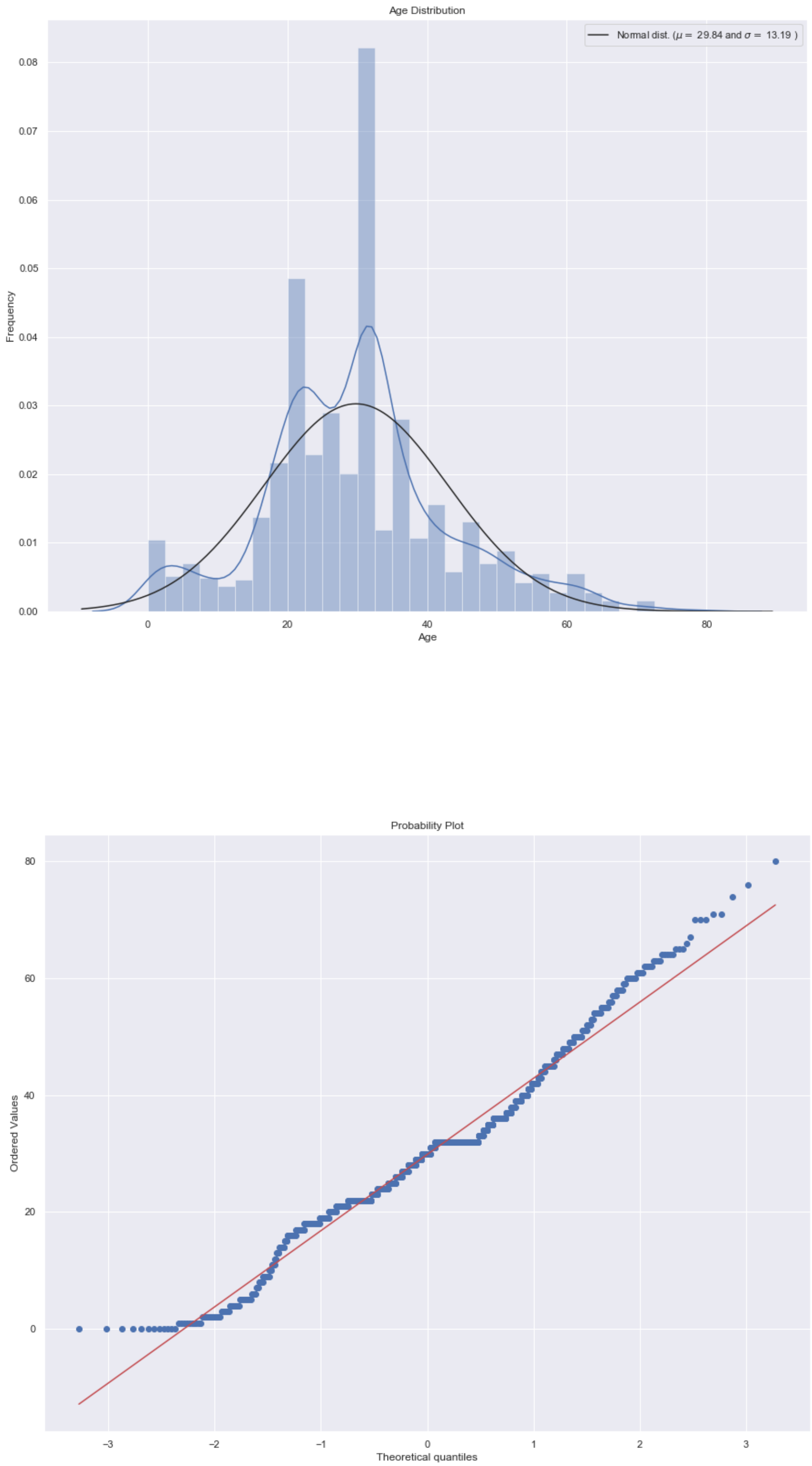


Figure 5 – Age Distribution After Transformation:

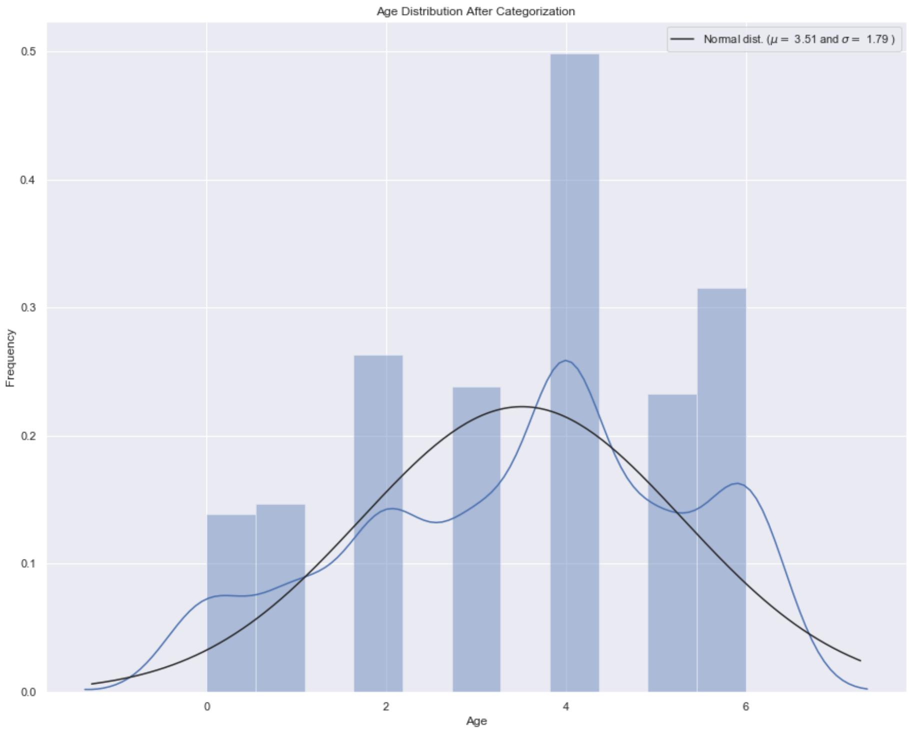


Figure 6 – Boxplots:

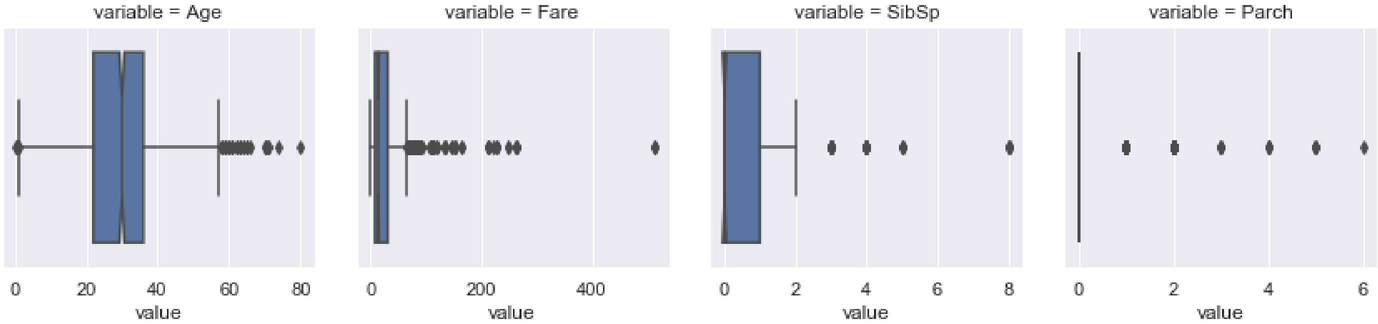
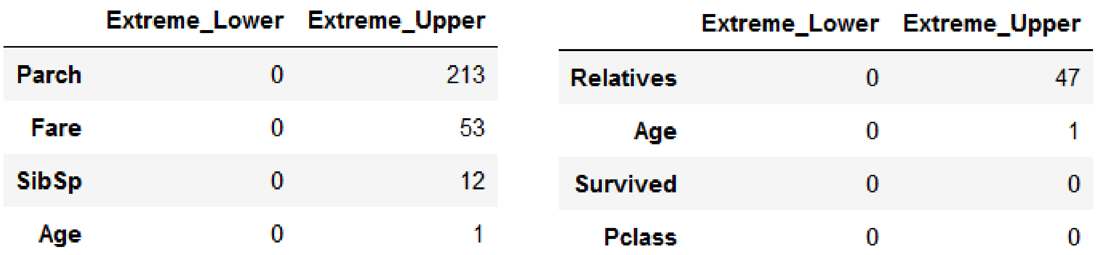


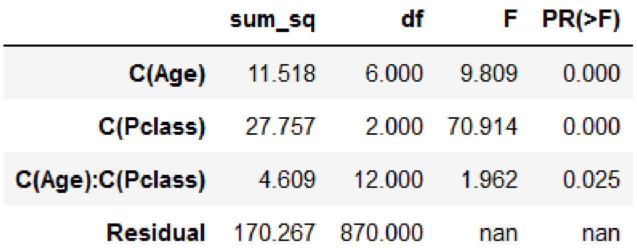
Table 3 – Extreme Outliers:



	Extreme_Lower	Extreme_Upper
Parch	0	213
Fare	0	53
SibSp	0	12
Age	0	1

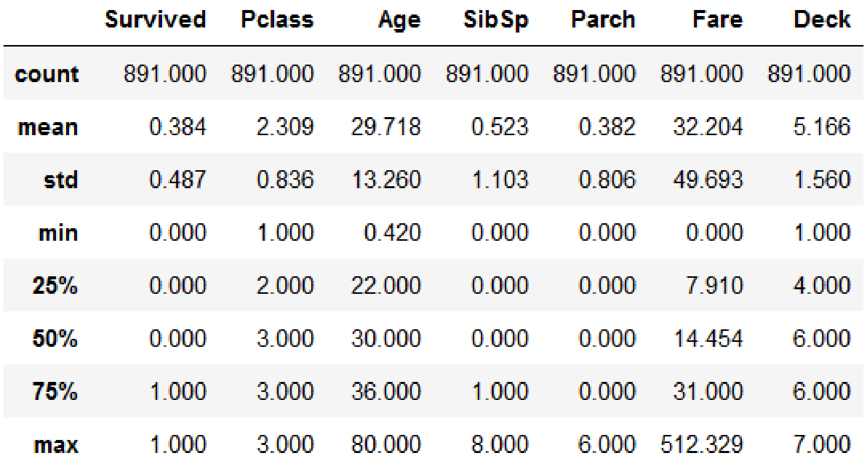
	Extreme_Lower	Extreme_Upper
Relatives	0	47
Age	0	1
Survived	0	0
Pclass	0	0

Table 4 – Two Way Anova, Age and Passenger Class:



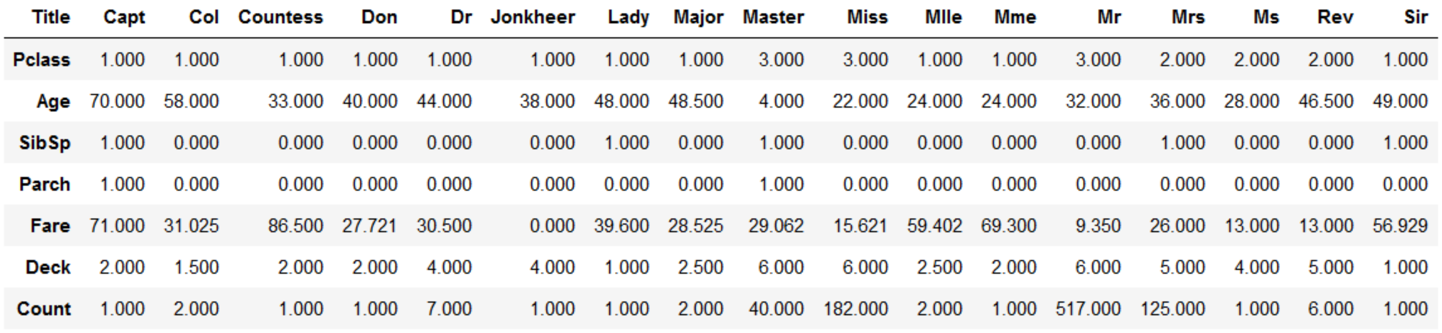
	sum_sq	df	F	PR(>F)
C(Age)	11.518	6.000	9.809	0.000
C(Pclass)	27.757	2.000	70.914	0.000
C(Age):C(Pclass)	4.609	12.000	1.962	0.025
Residual	170.267	870.000	nan	nan

Table 5 – Descriptive Statistics:



	Survived	Pclass	Age	SibSp	Parch	Fare	Deck
count	891.000	891.000	891.000	891.000	891.000	891.000	891.000
mean	0.384	2.309	29.718	0.523	0.382	32.204	5.166
std	0.487	0.836	13.260	1.103	0.806	49.693	1.560
min	0.000	1.000	0.420	0.000	0.000	0.000	1.000
25%	0.000	2.000	22.000	0.000	0.000	7.910	4.000
50%	0.000	3.000	30.000	0.000	0.000	14.454	6.000
75%	1.000	3.000	36.000	1.000	0.000	31.000	6.000
max	1.000	3.000	80.000	8.000	6.000	512.329	7.000

Table 6: Median Statistic by Name Title:



Title	Capt	Col	Countess	Don	Dr	Jonkheer	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Rev	Sir
Pclass	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	3.000	3.000	1.000	1.000	3.000	2.000	2.000	2.000	1.000
Age	70.000	58.000	33.000	40.000	44.000	38.000	48.000	48.500	4.000	22.000	24.000	24.000	32.000	36.000	28.000	46.500	49.000
SibSp	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000
Parch	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Fare	71.000	31.025	86.500	27.721	30.500	0.000	39.600	28.525	29.062	15.621	59.402	69.300	9.350	26.000	13.000	13.000	56.929
Deck	2.000	1.500	2.000	2.000	4.000	4.000	1.000	2.500	6.000	6.000	2.500	2.000	6.000	5.000	4.000	5.000	1.000
Count	1.000	2.000	1.000	1.000	7.000	1.000	1.000	2.000	40.000	182.000	2.000	1.000	517.000	125.000	1.000	6.000	1.000

Figure 7: Number of Passengers by Name Title:

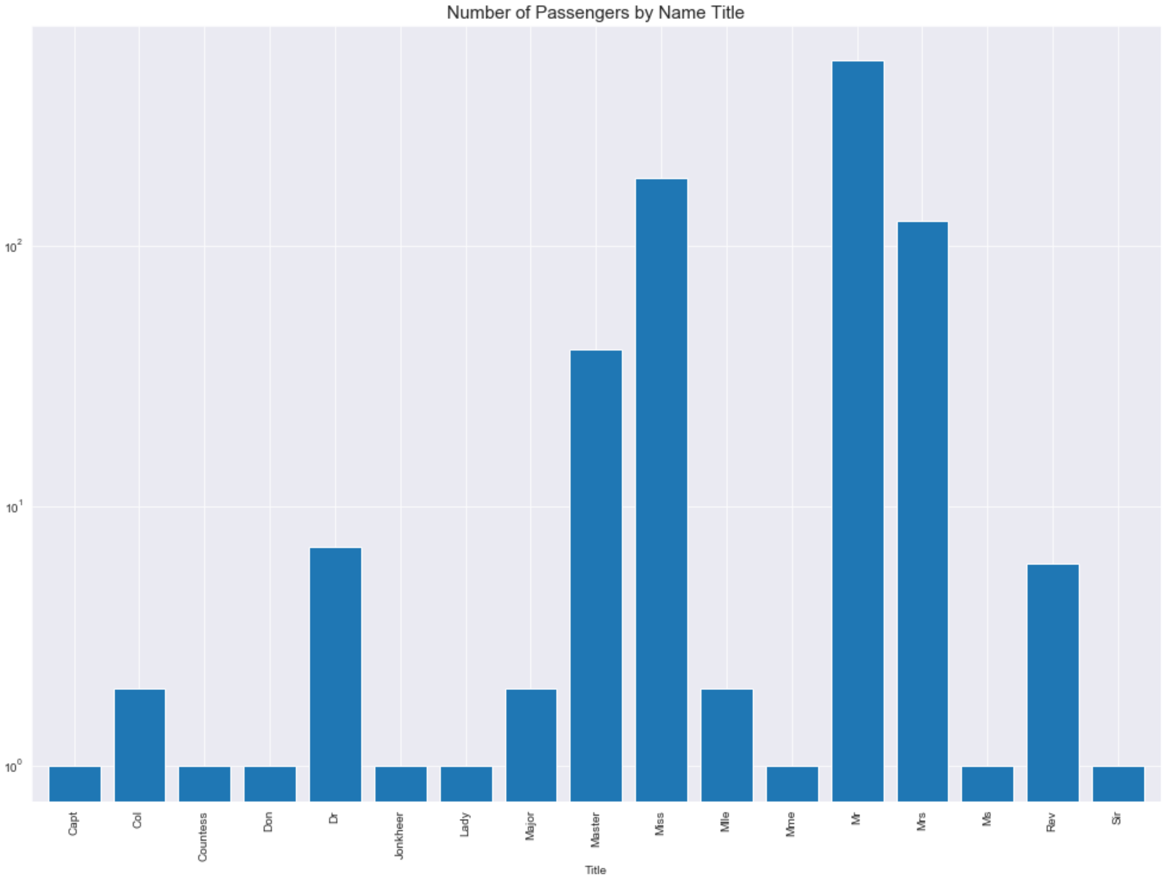
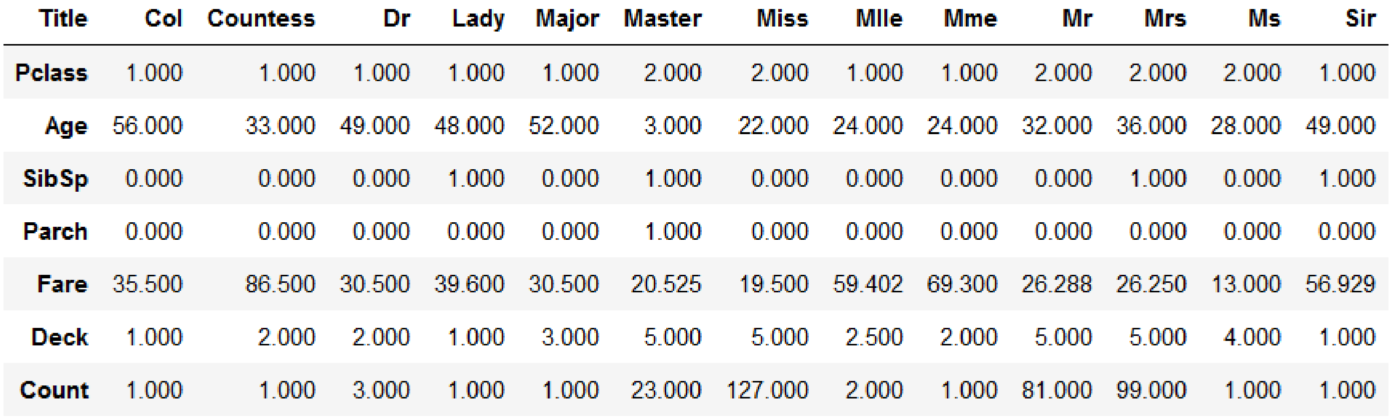
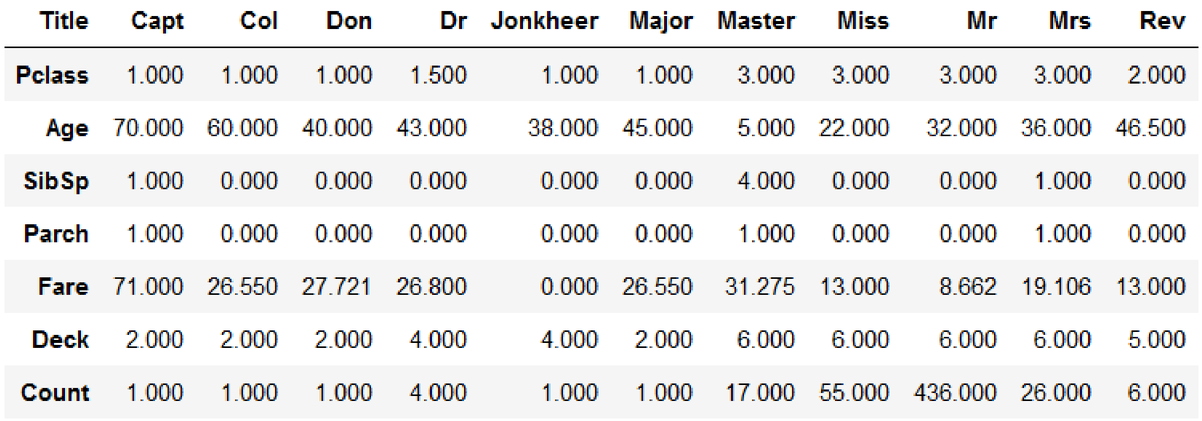


Table 7 – Median Statistics of Those that Survived by Name Title:



Title	Col	Countess	Dr	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Sir
Pclass	1.000	1.000	1.000	1.000	1.000	2.000	2.000	1.000	1.000	2.000	2.000	2.000	1.000
Age	56.000	33.000	49.000	48.000	52.000	3.000	22.000	24.000	24.000	32.000	36.000	28.000	49.000
SibSp	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000
Parch	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Fare	35.500	86.500	30.500	39.600	30.500	20.525	19.500	59.402	69.300	26.288	26.250	13.000	56.929
Deck	1.000	2.000	2.000	1.000	3.000	5.000	5.000	2.500	2.000	5.000	5.000	4.000	1.000
Count	1.000	1.000	3.000	1.000	1.000	23.000	127.000	2.000	1.000	81.000	99.000	1.000	1.000

Table 8 – Median Statistics of Those that Died by Name Title:



Title	Capt	Col	Don	Dr	Jonkheer	Major	Master	Miss	Mr	Mrs	Rev
Pclass	1.000	1.000	1.000	1.500	1.000	1.000	3.000	3.000	3.000	3.000	2.000
Age	70.000	60.000	40.000	43.000	38.000	45.000	5.000	22.000	32.000	36.000	46.500
SibSp	1.000	0.000	0.000	0.000	0.000	0.000	4.000	0.000	0.000	1.000	0.000
Parch	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
Fare	71.000	26.550	27.721	26.800	0.000	26.550	31.275	13.000	8.662	19.106	13.000
Deck	2.000	2.000	2.000	4.000	4.000	2.000	6.000	6.000	6.000	6.000	5.000
Count	1.000	1.000	1.000	4.000	1.000	1.000	17.000	55.000	436.000	26.000	6.000

Figure 8 – Survived and Died by Name Title and Passenger Class:

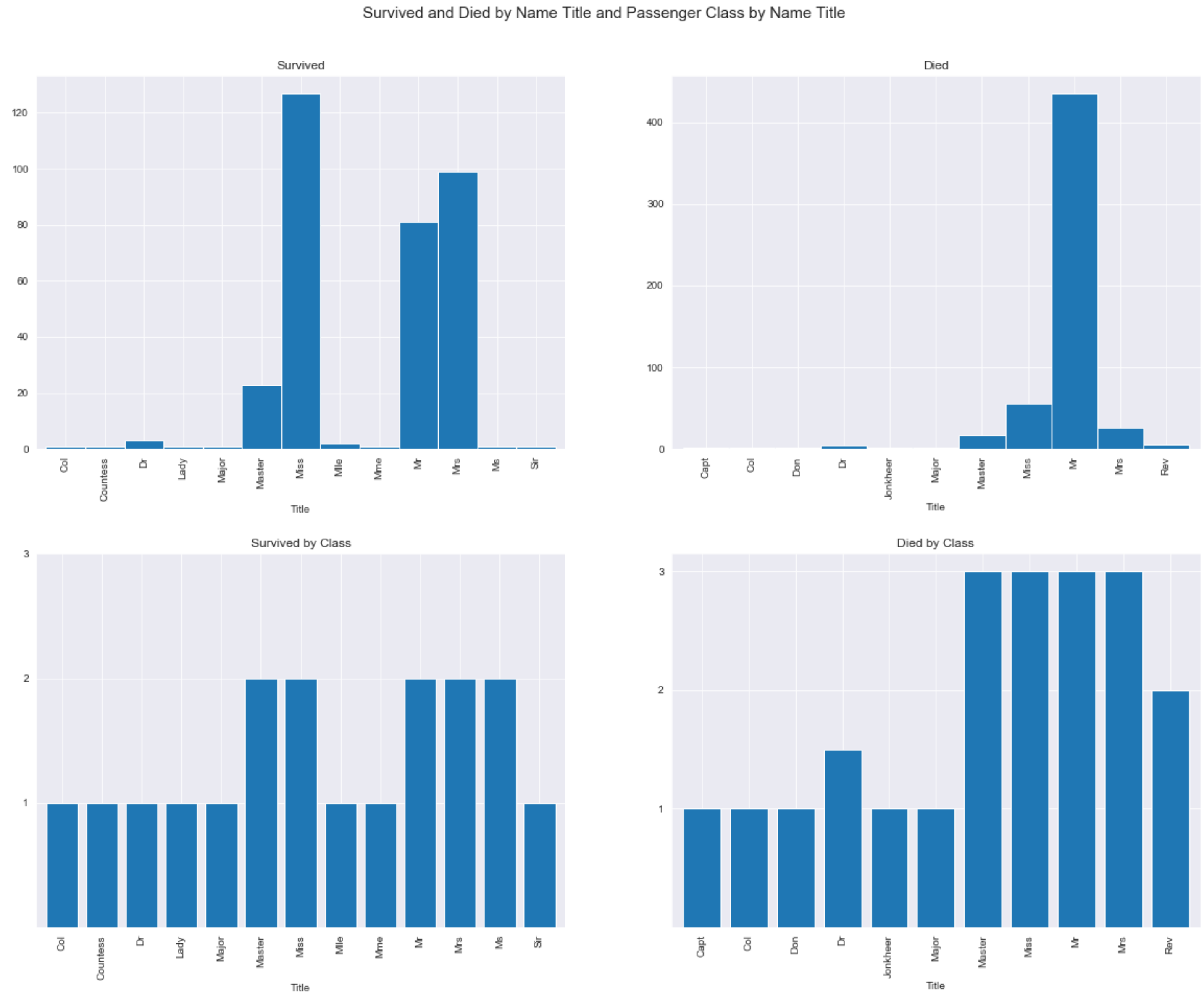


Figure 9 – Percentage of Passengers that Survived or Died by Age Group:

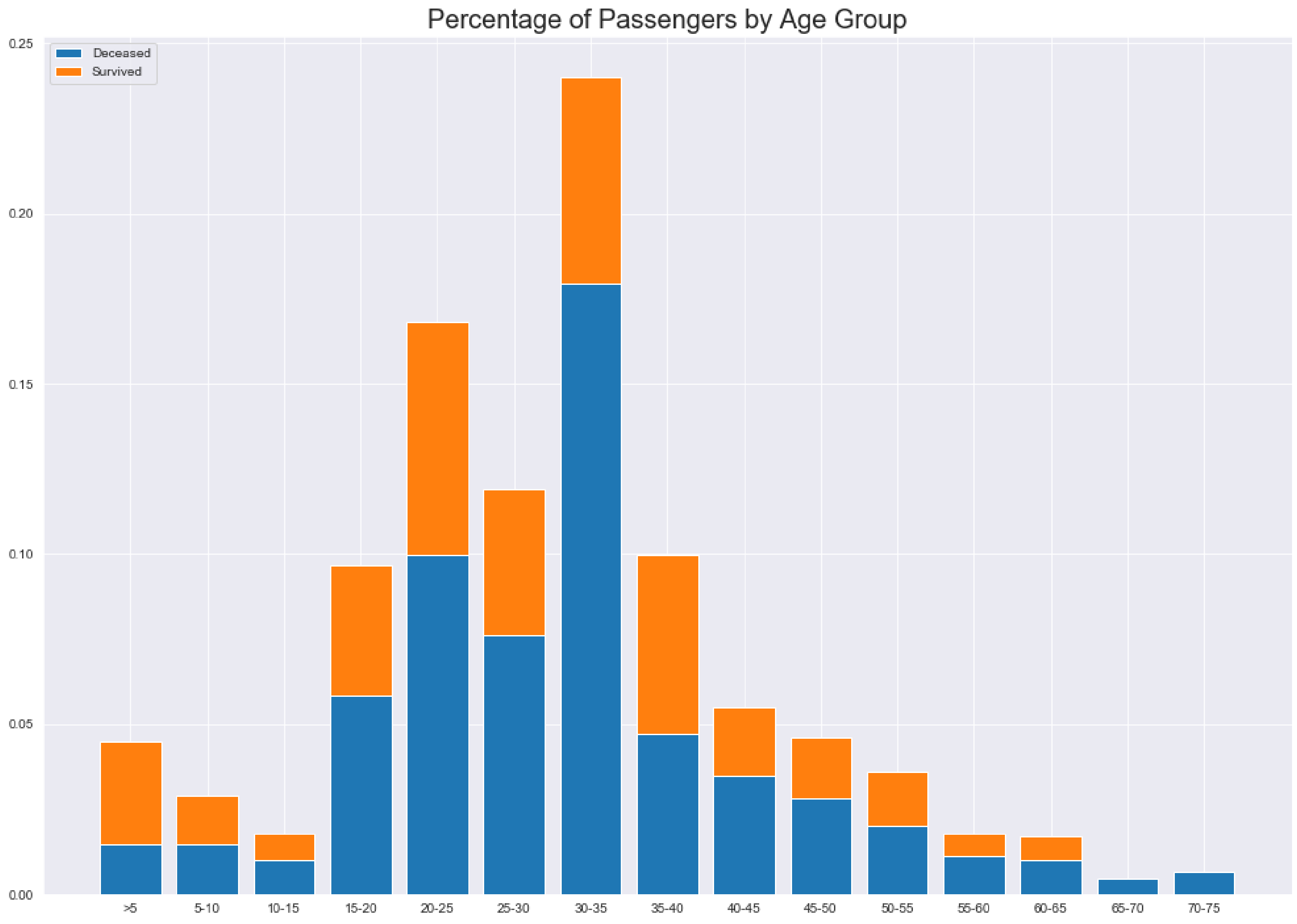


Figure 10 – Percentage of Passengers that Survived or Died by Gender:

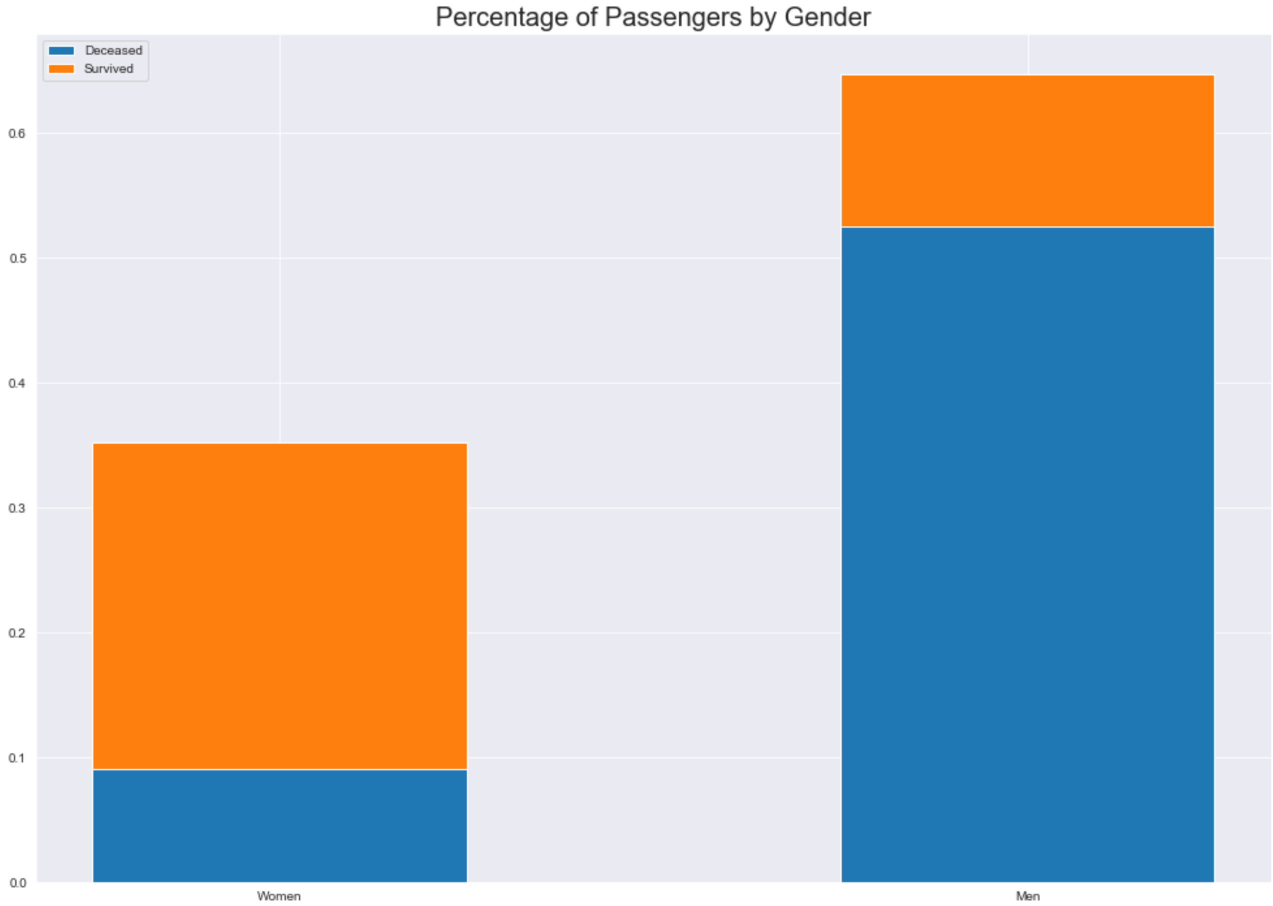


Figure 11 – Percentage of Passengers that Survived or Died by Class:

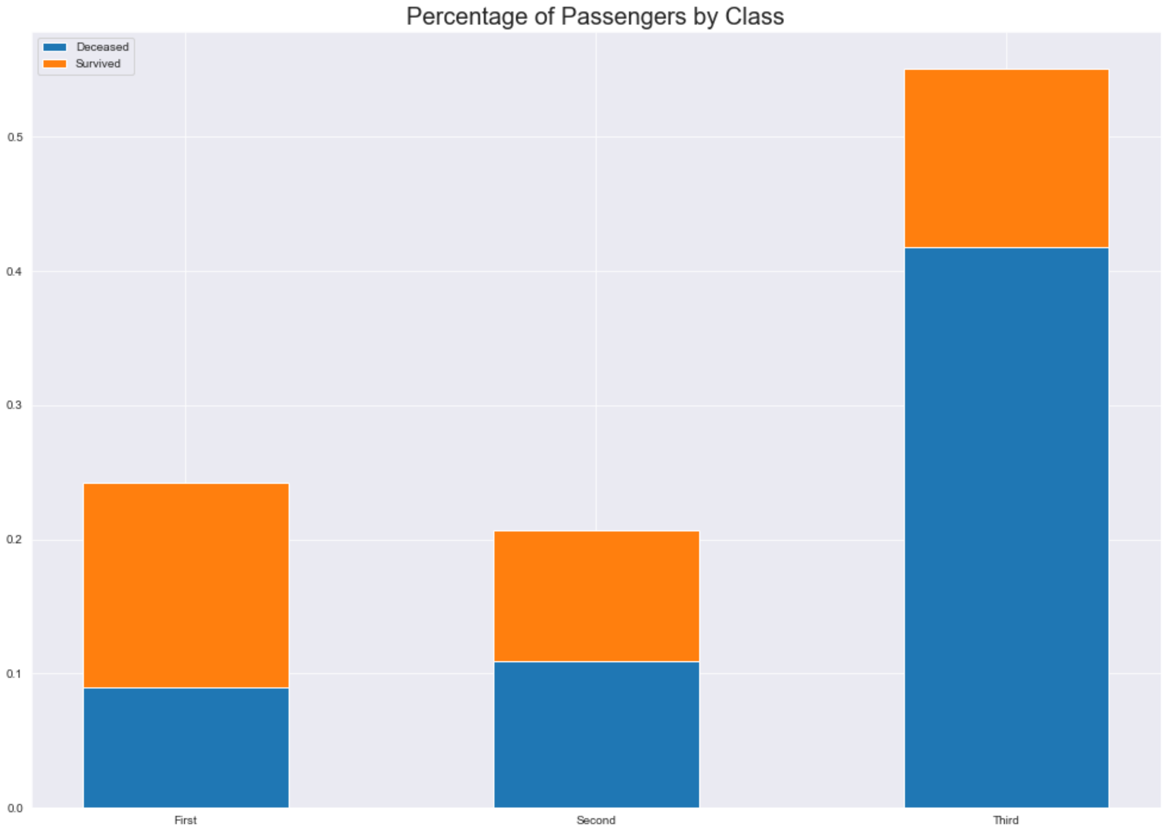


Figure 12 – Survived or Died by Port of Embarkation:

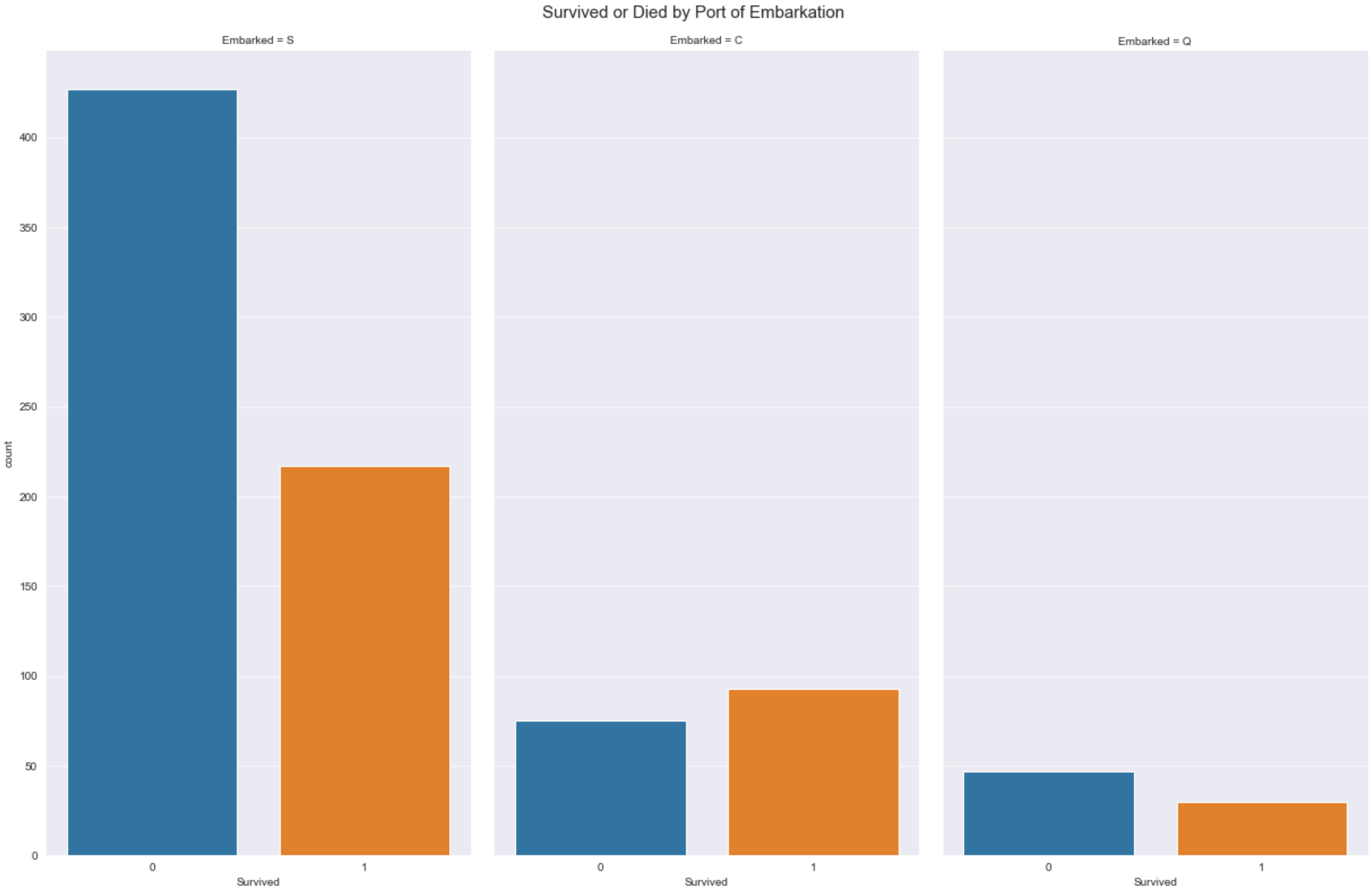
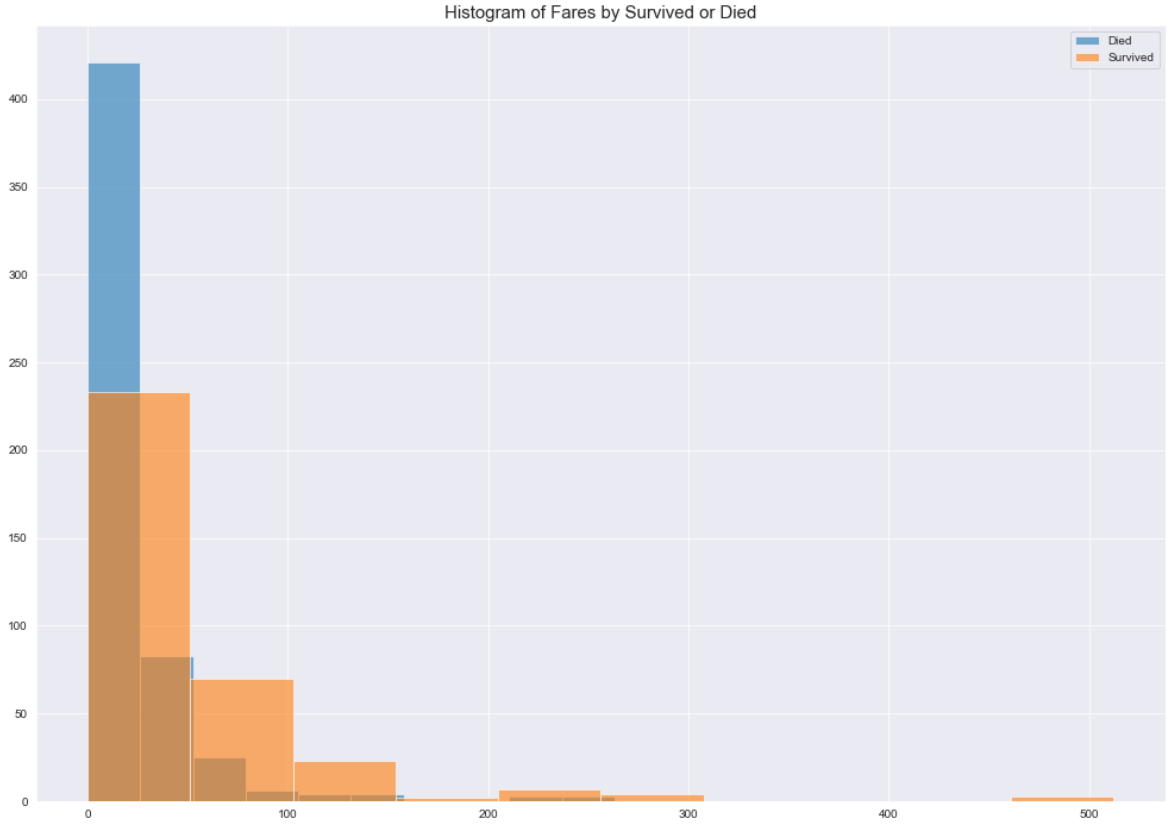
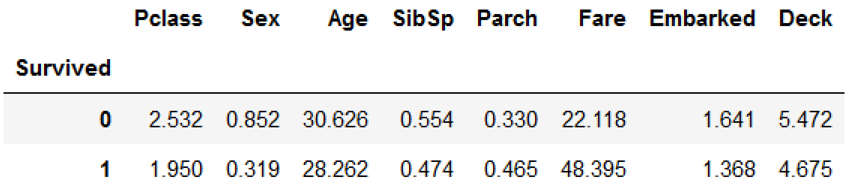


Figure 13 - Histogram of Fares by Survived or Died:





	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Deck
Survived								
0	2.532	0.852	30.626	0.554	0.330	22.118	1.641	5.472
1	1.950	0.319	28.262	0.474	0.465	48.395	1.368	4.675

Figure 14 – Swarm Plot of Survived or Died by Fare:

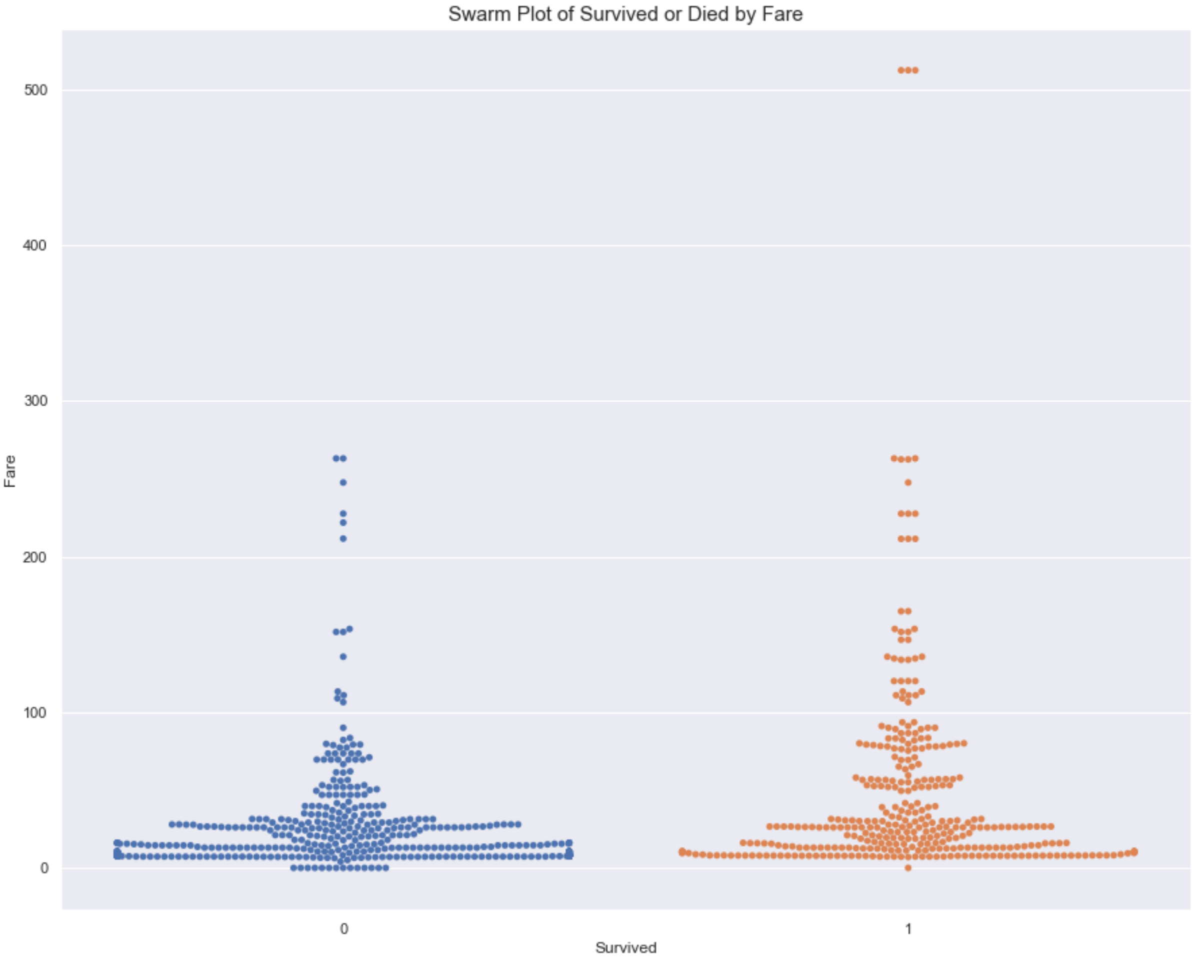
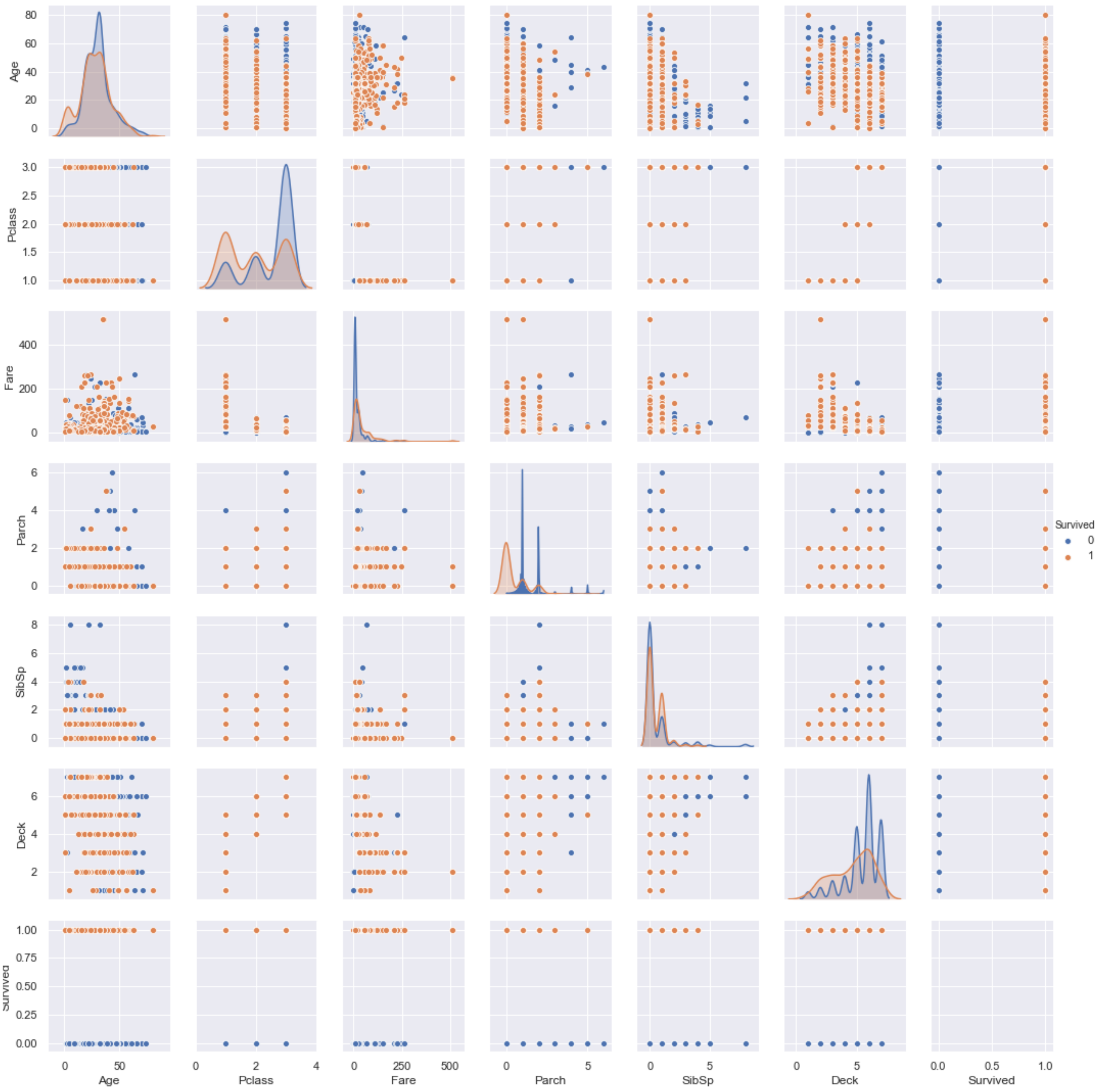


Figure 15 – Scatter Matrix of Characteristics of Passengers Based Off of Survival:



Modeling Using Different Binary Classifiers (Logistic Regression and Naïve Bayes):

The main objective of this study was to test data using two binary classifiers: Logistic Regression and Naïve Bayes. Two naïve Bayes methods were used: Gaussian and Bernoulli naïve Bayes. Cross-validation used was K-fold cross-validation, which randomly splits a training set into 10 distinct subsets called folds, then trains and evaluates the model n (i.e. 10) times, picking a different fold for evaluation every time and training on the other n (i.e. 9) folds. Results showed that using a logistic regression performed the best as it scored with an average ROC AUC of 0.871 in a 10-fold cross-validation design.

Scaling:

For scaling a robust scaler was used for Logistic Regression. The centering and scaling statistics of the robust scaler are based on percentiles and are therefore not influenced by a few number of very large marginal outliers. The robust scaler scales based off the quartiles of distributions, and it captures the outliers. It is not as sensitive to outliers such as in min-max and standard scaling. However, the min-max scaler was used for the naïve Bayes was used as performance was an issue using the robust scaler.

Models:

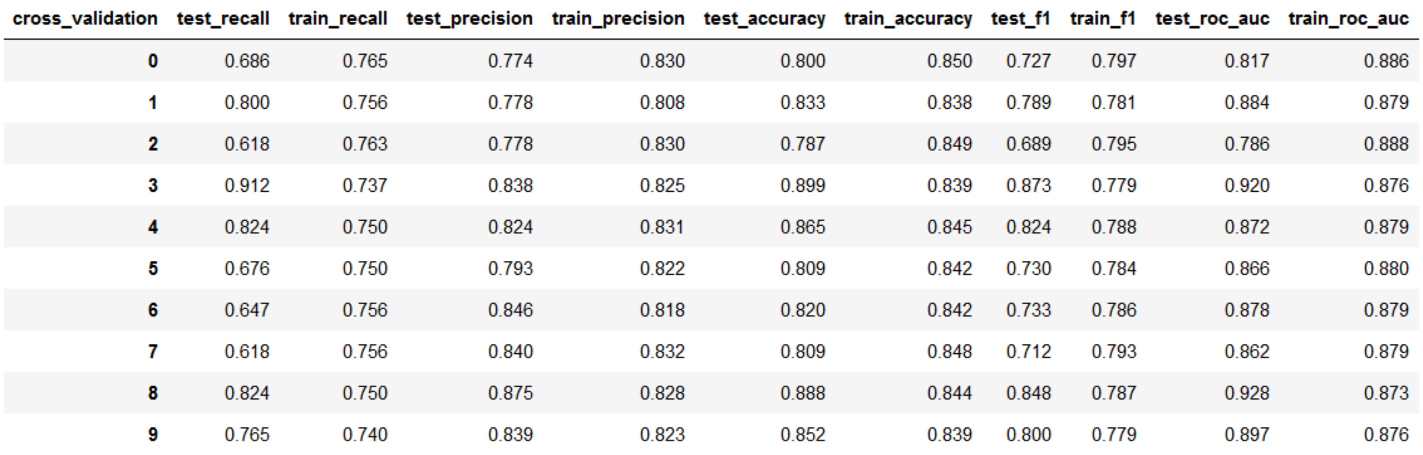
Logistic Regression:

- Commonly used to estimate the probability that an instance belongs to a particular class. If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (positive 1 or negative 0.
- Similar to a linear regression model, a logistic regression model computes a weighted sum of the input features (plus a bias term), but instead of outputting the result directly like in linear regression, it outputs the logistic of the results.
- The function is defined as: $\sigma(t) = 1 / 1 + \exp - t$
- The prediction is defined as: $\hat{y} = 0$ if $p < 0.5$, 1 if $p \geq 0.5$.
- For each instance it computes the prediction error and multiplies it by the jth feature value, and then it computes the average over all training instances.
- The objective of training is to set the parameter vector θ so that the model estimates high probabilities for positive instances ($y = 1$) and low probabilities for negative instances ($y = 0$).
- In a single training instance, the cost function is: $c(\theta) = - \log \hat{p}$ if $y = 1$, $- \log 1 - \hat{p}$ if $y = 0$
- $- \log(t)$ grows very large when t approaches 0, so the cost will be large if the model estimates a probability close to 0 for a positive instance, and it will also be very large if the model estimates a probability close to 1 for a negative instance. On the other hand, $- \log(t)$ is close to 0 when t is close to 1, so the cost will be close to 0 if the estimated probability is close to 0 for a negative instance or close to 1 for a positive instance, which is what is desired.
- Logistic regression models can be regularized using l1 or l_2 penalties. The hyperparameter controlling regularization strength is the inverse of alpha: C, the higher the value of C the less it is regularized.
- Generalization performance is better than naïve Bayes classifiers.

Logistic Regression Score Results:

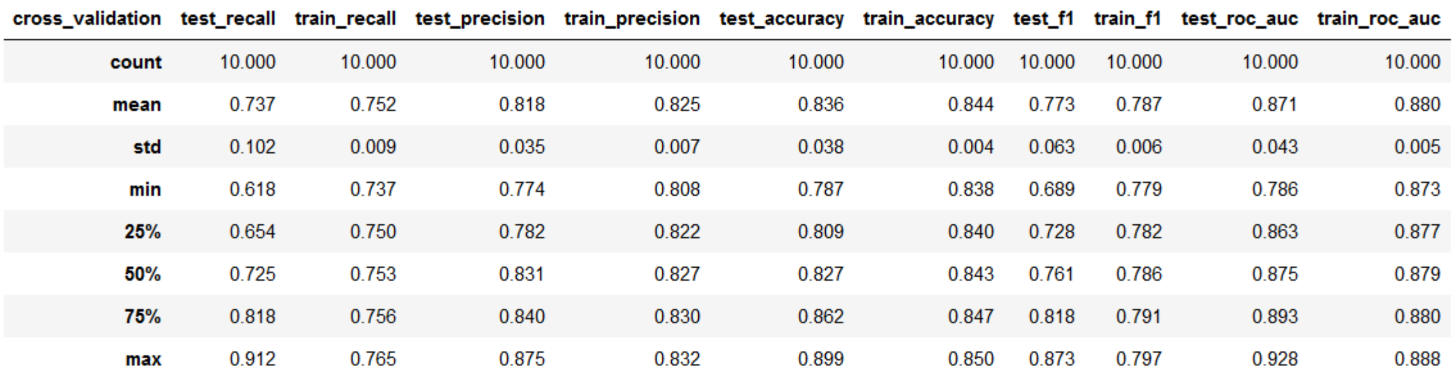
- Testing the training set using a 10-fold cross validation design showed a test mean ROC AUC score of 0.871, accuracy score of 0.836, with test precision and recall at 0.818 and 0.737. Test F1 score was 0.773
- The AUC value of 0.871 indicates that there were fewer type I and fewer type II errors, and that the model is a good measure of separability as it reproduces the data very well.

Cross Validation Scores:



cross_validation	test_recall	train_recall	test_precision	train_precision	test_accuracy	train_accuracy	test_f1	train_f1	test_roc_auc	train_roc_auc
0	0.686	0.765	0.774	0.830	0.800	0.850	0.727	0.797	0.817	0.886
1	0.800	0.756	0.778	0.808	0.833	0.838	0.789	0.781	0.884	0.879
2	0.618	0.763	0.778	0.830	0.787	0.849	0.689	0.795	0.786	0.888
3	0.912	0.737	0.838	0.825	0.899	0.839	0.873	0.779	0.920	0.876
4	0.824	0.750	0.824	0.831	0.865	0.845	0.824	0.788	0.872	0.879
5	0.676	0.750	0.793	0.822	0.809	0.842	0.730	0.784	0.866	0.880
6	0.647	0.756	0.846	0.818	0.820	0.842	0.733	0.786	0.878	0.879
7	0.618	0.756	0.840	0.832	0.809	0.848	0.712	0.793	0.862	0.879
8	0.824	0.750	0.875	0.828	0.888	0.844	0.848	0.787	0.928	0.873
9	0.765	0.740	0.839	0.823	0.852	0.839	0.800	0.779	0.897	0.876

Test Statistics:

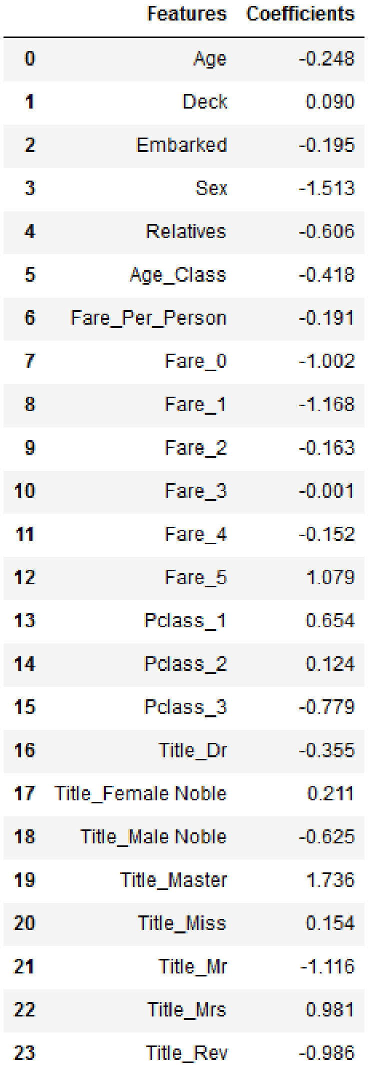


cross_validation	test_recall	train_recall	test_precision	train_precision	test_accuracy	train_accuracy	test_f1	train_f1	test_roc_auc	train_roc_auc
count	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000
mean	0.737	0.752	0.818	0.825	0.836	0.844	0.773	0.787	0.871	0.880
std	0.102	0.009	0.035	0.007	0.038	0.004	0.063	0.006	0.043	0.005
min	0.618	0.737	0.774	0.808	0.787	0.838	0.689	0.779	0.786	0.873
25%	0.654	0.750	0.782	0.822	0.809	0.840	0.728	0.782	0.863	0.877
50%	0.725	0.753	0.831	0.827	0.827	0.843	0.761	0.786	0.875	0.879
75%	0.818	0.756	0.840	0.830	0.862	0.847	0.818	0.791	0.893	0.880
max	0.912	0.765	0.875	0.832	0.899	0.850	0.873	0.797	0.928	0.888

Exploration and Data Visualization of Logistic Regression Model:

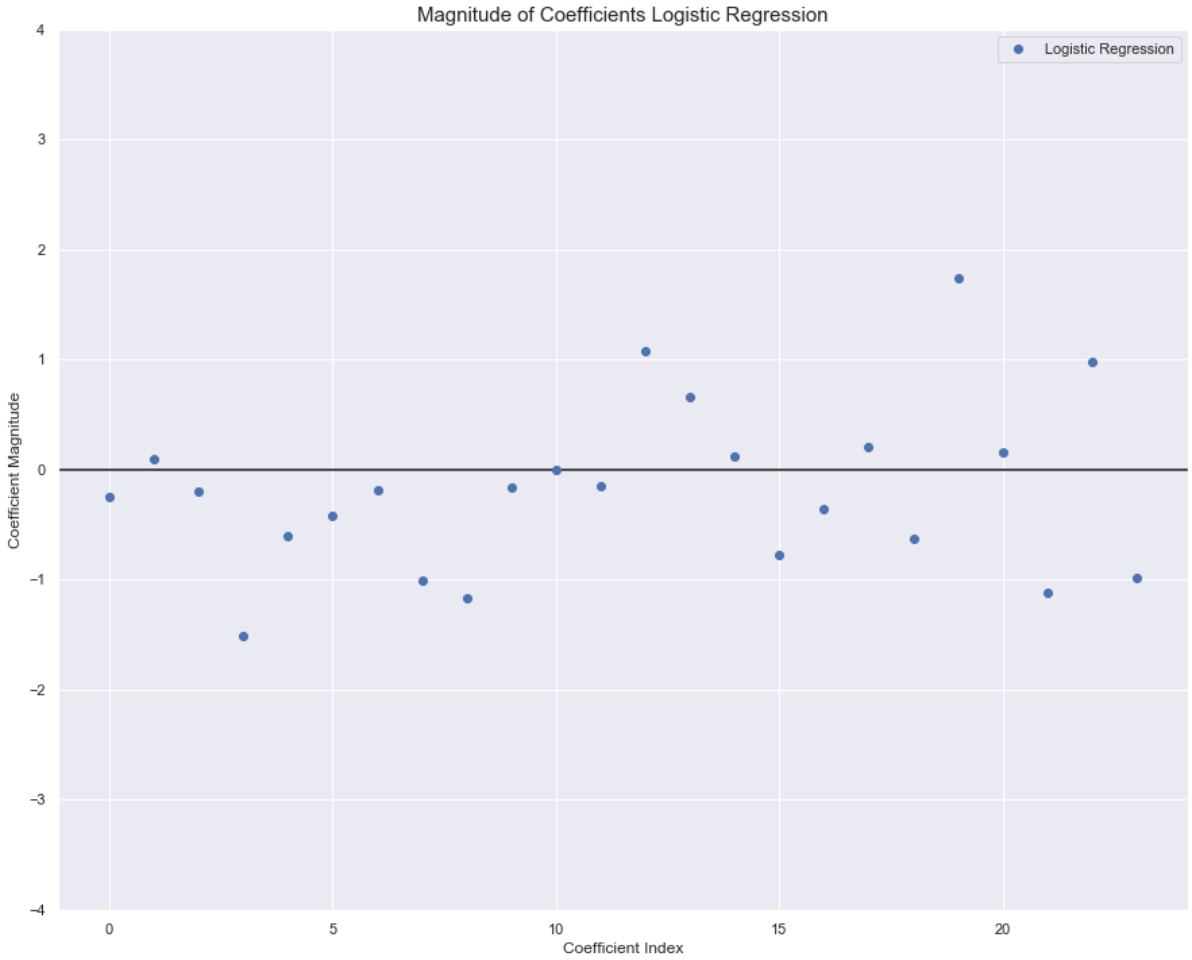
Coefficients:

- Although representations of coefficients should be taken with a grain of salt, it is interesting to see the magnitude of the coefficients.
- The coefficients below show how the model regularized the coefficients and many were pushed close to zero, but were not completely zero at all due to l_2 regularization.



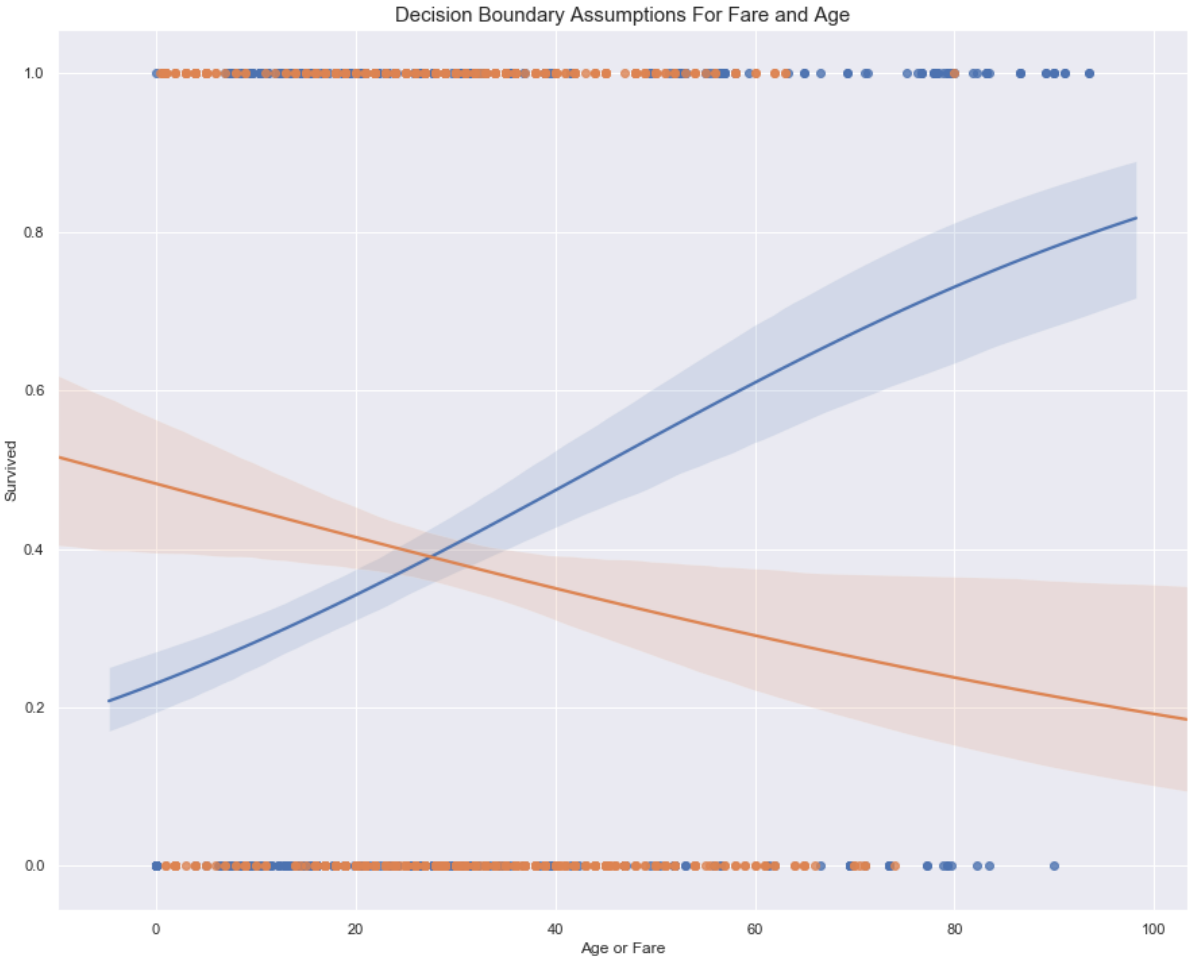
	Features	Coefficients
0	Age	-0.248
1	Deck	0.090
2	Embarked	-0.195
3	Sex	-1.513
4	Relatives	-0.606
5	Age_Class	-0.418
6	Fare_Per_Person	-0.191
7	Fare_0	-1.002
8	Fare_1	-1.168
9	Fare_2	-0.163
10	Fare_3	-0.001
11	Fare_4	-0.152
12	Fare_5	1.079
13	Pclass_1	0.654
14	Pclass_2	0.124
15	Pclass_3	-0.779
16	Title_Dr	-0.355
17	Title_Female Noble	0.211
18	Title_Male Noble	-0.625
19	Title_Master	1.736
20	Title_Miss	0.154
21	Title_Mr	-1.116
22	Title_Mrs	0.981
23	Title_Rev	-0.986

Figure 16 – Coefficient Magnitude (Logistic Regression):



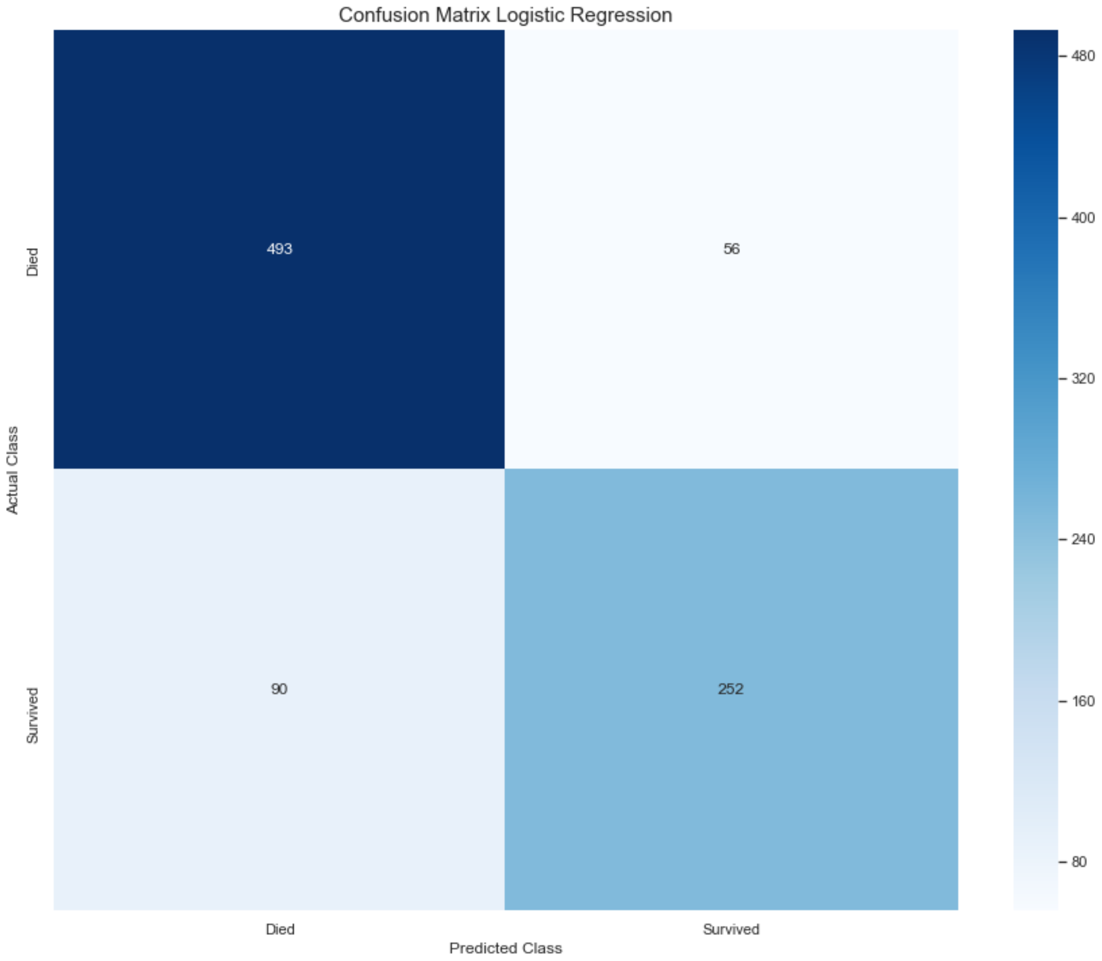
Coefficient magnitudes shows regularization of features as they are reasonably close to zero, but are not quite there indicating that the model still treated coefficients as having meaningful magnitude.

Figure 17 – Decision Boundary for Age and Fares (Logistic Regression):



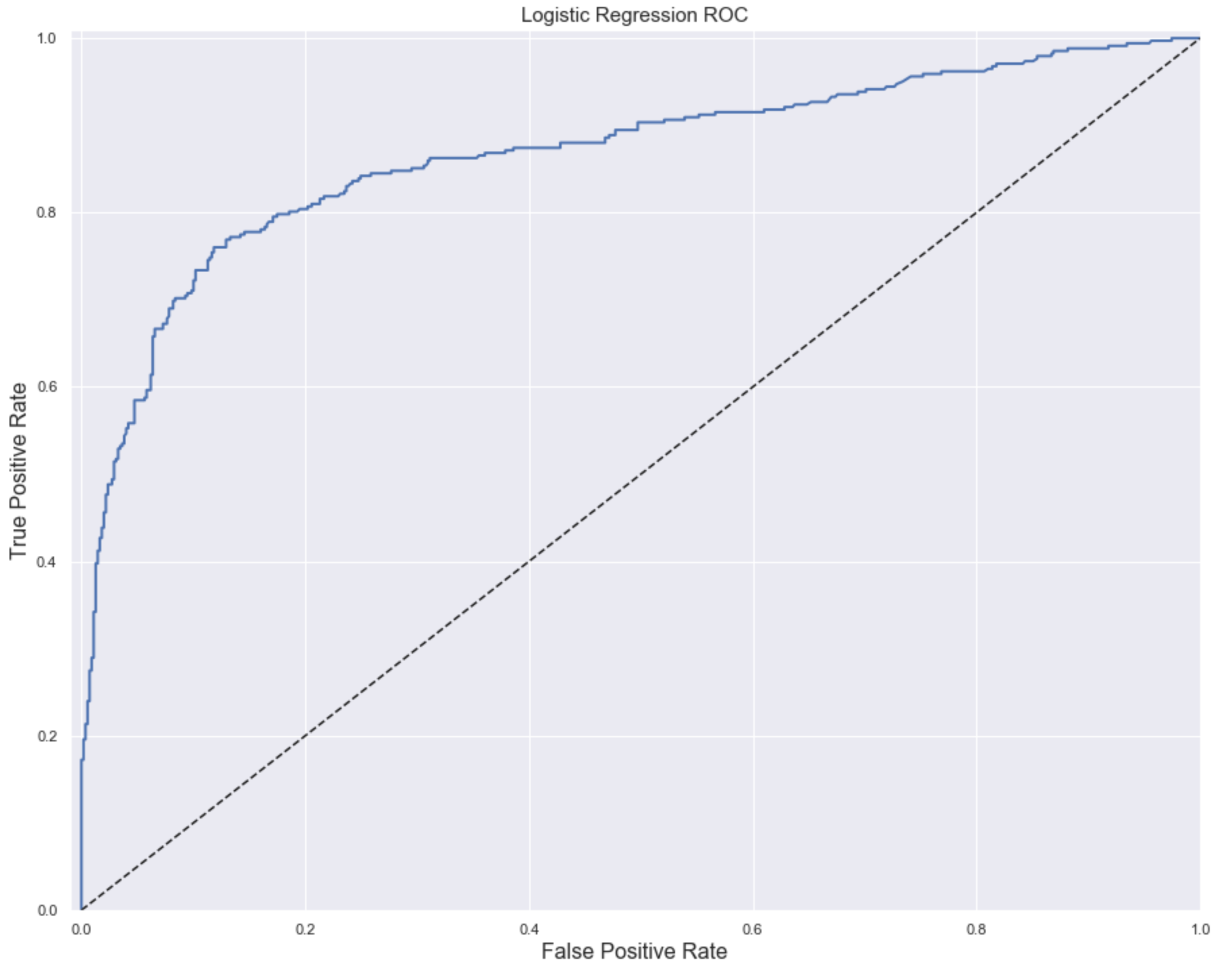
- The decision boundaries for Fare and Age show that at around the age of 30 and a fare of 30 dollars, the probability of survival is around 40% and as the lines cross and progress the survival rate gets lower for ages 30 and above (yellow line), and the survival rate gets higher for fares that increase above 30 mark.
- However, there is a lot of overlap with just these two attributes, which makes classification difficult without others.

Figure 18 – Confusion Matrix (Logistic Regression):



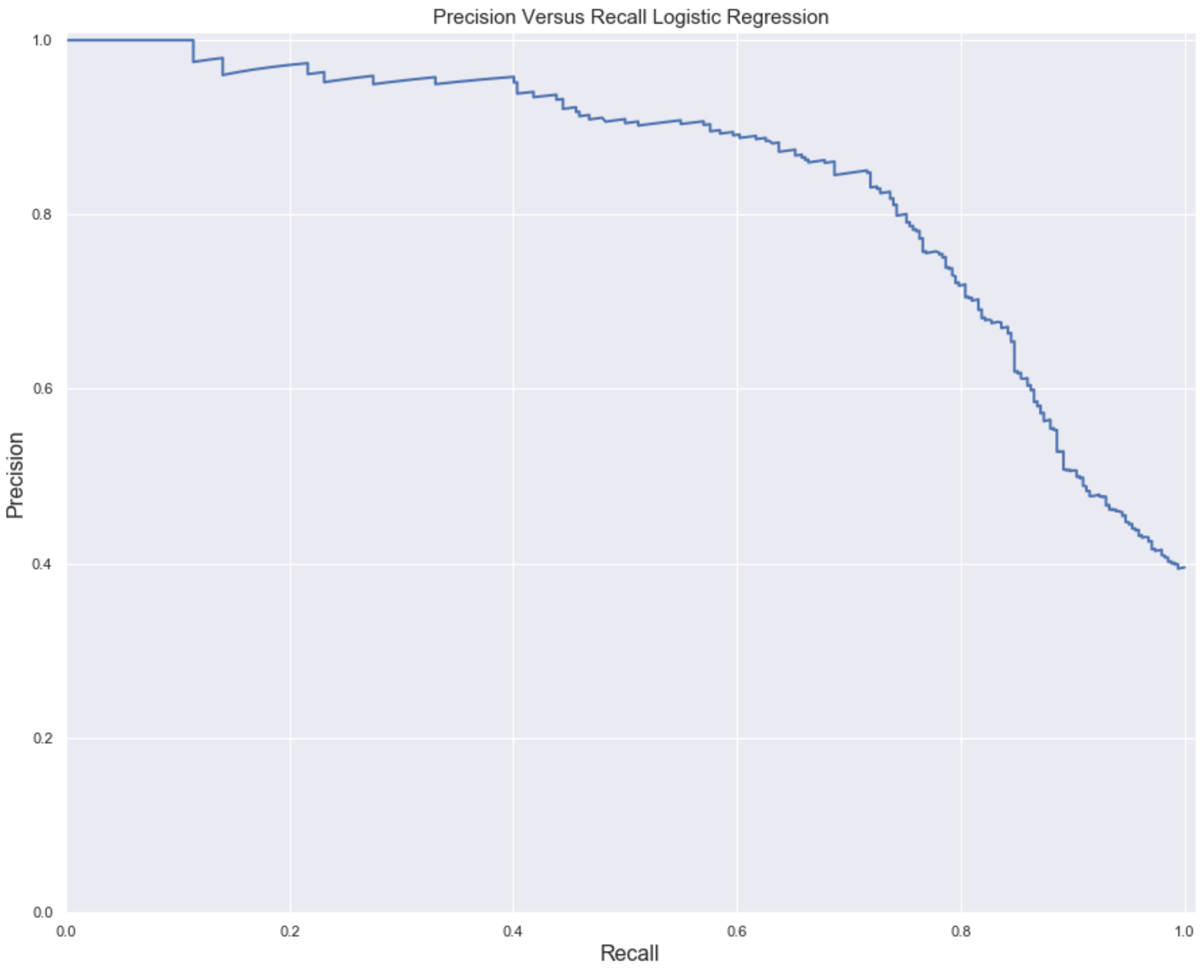
Confusion matrix shows that 493 passengers were correctly classified as died, and 56 passengers were wrongly classified as survived. 90 passengers were wrongly classified as died, and 252 were correctly classified as survived.

Figure 19 – ROC Curve (Logistic Regression):



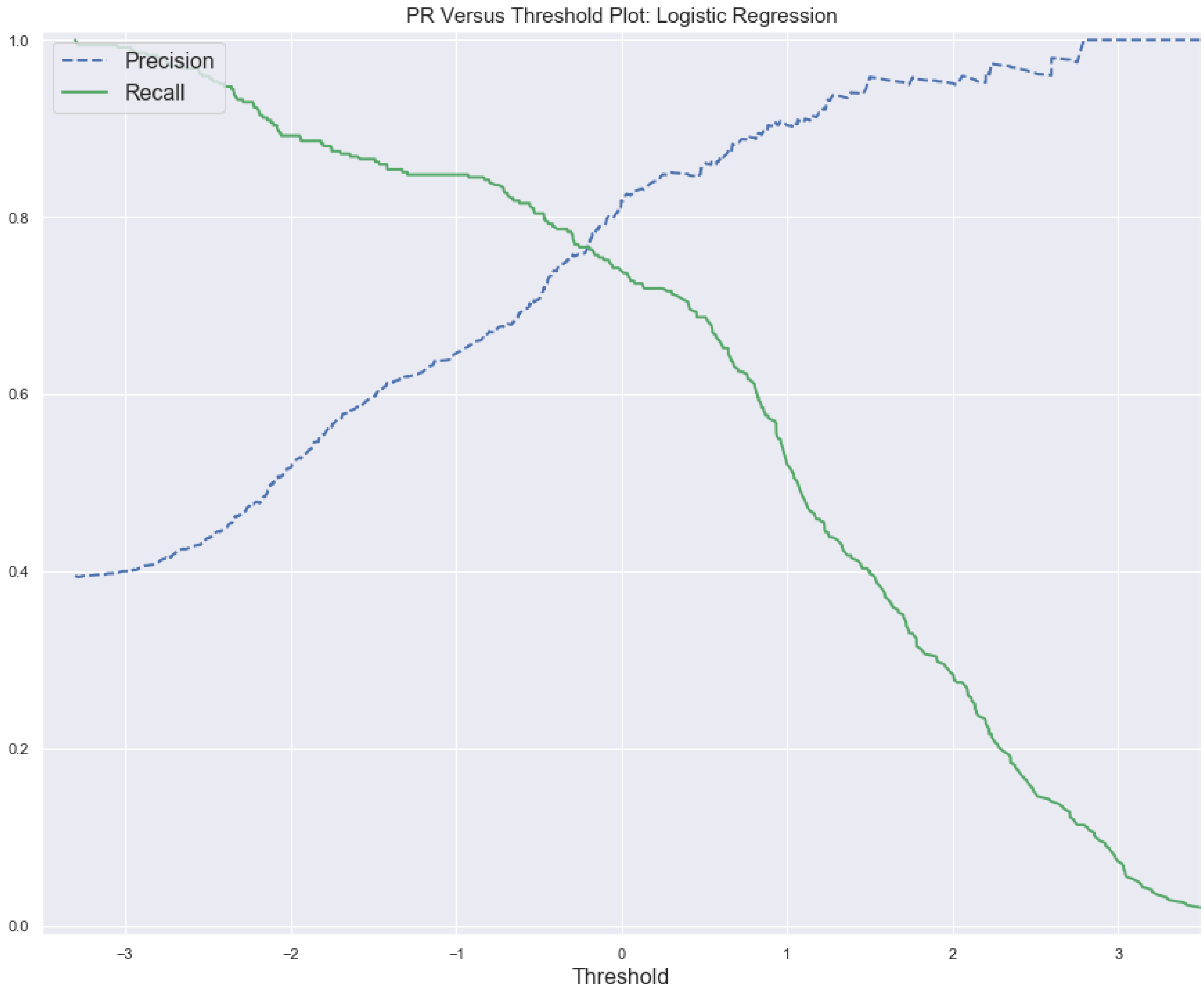
- The ROC curve indicates that there were fewer type I and fewer type II errors, and that the model is a good measure of separability as it reproduces the data very well.
- In summary, a ROC curve plots the true positive rate (recall) against the false positive rate (FPR), which is the ratio of negative instances that are incorrectly classified as positive. It is basically equal to one subtracted from the true negative rate, or the ratio of negative instances that are correctly classified as negative. Hence, the ROC curve plots the true positive rate versus one minus the true negative rate.
- The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). The model seems to do this pretty well. This was confirmed in the validation scores where the mean AUC was equal to 0.871.
- However, since there were fewer positives (survived) than negatives (died) the model is further examined using the precision recall curve.

Figure 20 – Precision Versus Recall Curve (Logistic Regression):



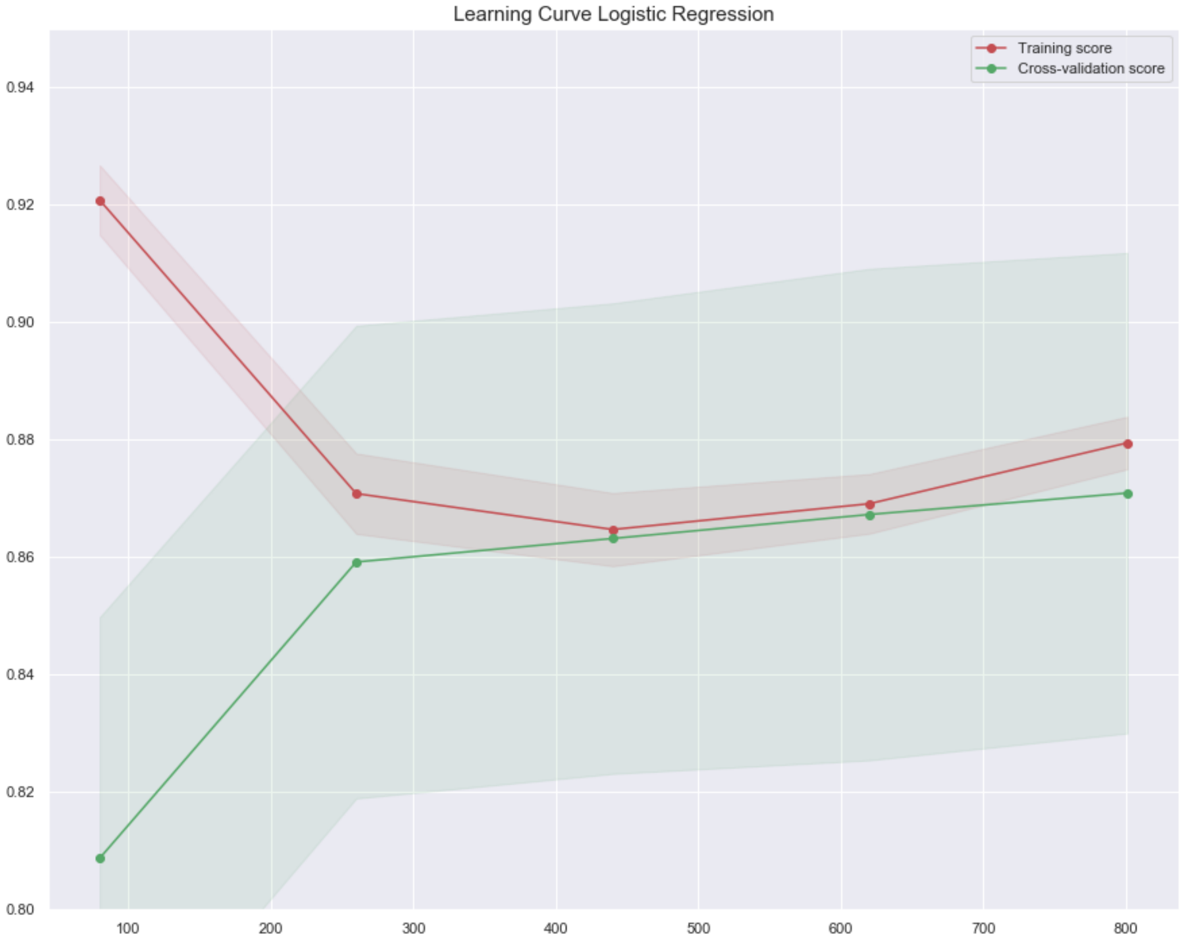
- The precision recall curve seems to support that the classifier is acceptable. The curve is close to the top right corner showing that precision does not drop sharply at the expense of higher recall. The mean f1 score of 0.773 seems to support this. Also, the mean score was at 0.818 while recall was at 0.737 indicating that the tradeoff was not terrible.
- Precision is basically defined as the ratio of the number of true positives over true positives plus false positives. Recall is defined as the ratio of the number of true positives over true positives plus false negatives. Essentially, precision is a measure of exactness, and recall is a measure of completeness. The PR curve shows the trade-off between the two, the more precision the less recall, and vice versa.

Figure 21 – Precision and Recall Versus Threshold (Logistic Regression):



The precision and recall versus threshold plot showed that the line between precision and recall crosses right before the threshold hits 0. At that threshold precision are recall are even at above 0.75 but below 0.80. At the threshold of 0 precision is a little above 0.80 and precision looks to be around 0.75. This is consistent with the mean precision score of 0.818, and recall score of 0.737. This tradeoff is significant when the threshold is set higher as recall drastically drops when it hits the threshold score of 1. However, at the threshold score of -1 precision is above 0.60 indicating that the tradeoff for recall is not as high as it is for precision. It is recommended that the threshold score remain unchanged.

Figure 22 – Learning Curve (Logistic Regression):



The learning curve for ROC AUC scores seems to show that the model may be severely underfitting the data as the gap between the curves are very narrow. However, the scores are not low which is a good sign. The gap at the beginning may also be an indicator of an overfitting model, but these were at training sizes of under 100. As the training sets get larger the model seems to generalize very well especially around the 600 mark. On the other hand, the gap seems to widen at the training set size of 800 which shows that the model may not be complex enough, or more data would be needed. As data from the Titanic is limited, a better evaluation model may be required.

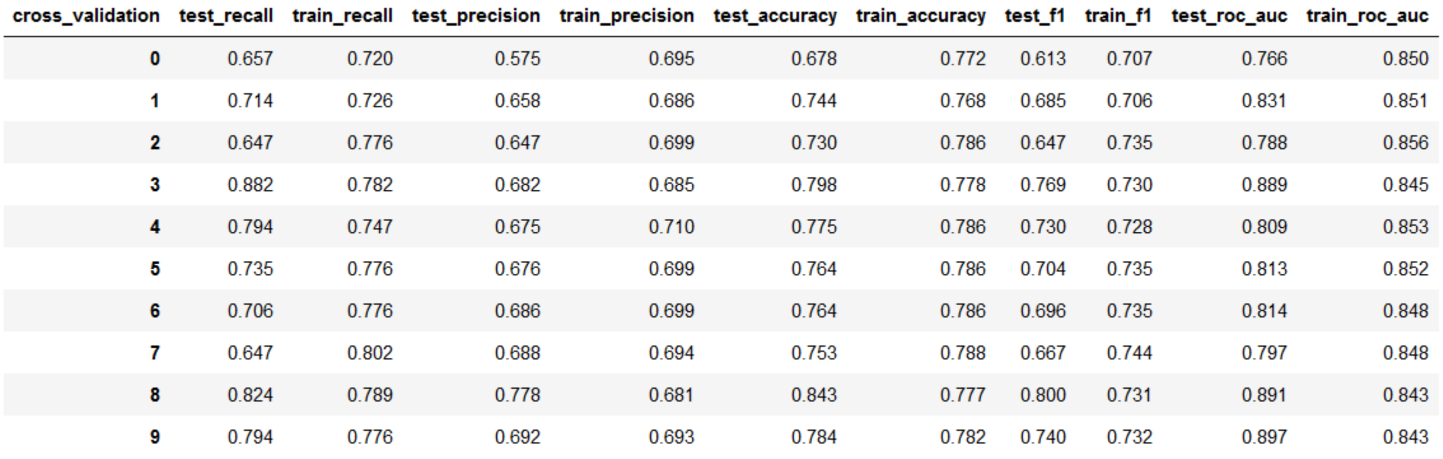
Naïve Bayes Using Gaussian NB:

- Naive Bayes classifiers are a family of classifiers that are similar to linear models. However, they to be faster in training. But it comes at a cost. The price paid for efficiency is the naive Bayes models often provide generalization performance that is slightly worse than linear classifiers like Logistic Regression.
- Naive Bayes models are efficient because they learn parameters by looking at each feature individually and collect simple per class statistics from each feature.
- The Gaussian model used below stores the average value and standard deviation of each feature for each class, and to make a prediction a data point is compared to the statistics for each of the classes, and the best matching class is predicted.
- The Gaussian model is mostly used on very high-dimensional data, it is very fast to train and predict, and the training procedure is very easy to understand.
- The model works very well with high-dimensional sparse data and are relatively robust to parameters. Naive Bayes models are great baseline models and are often used on very large datasets, where training even a linear model might take too long.
- Assumes data is Gaussian, and can be applied to any continuous data.

Gaussian NB Score Results:

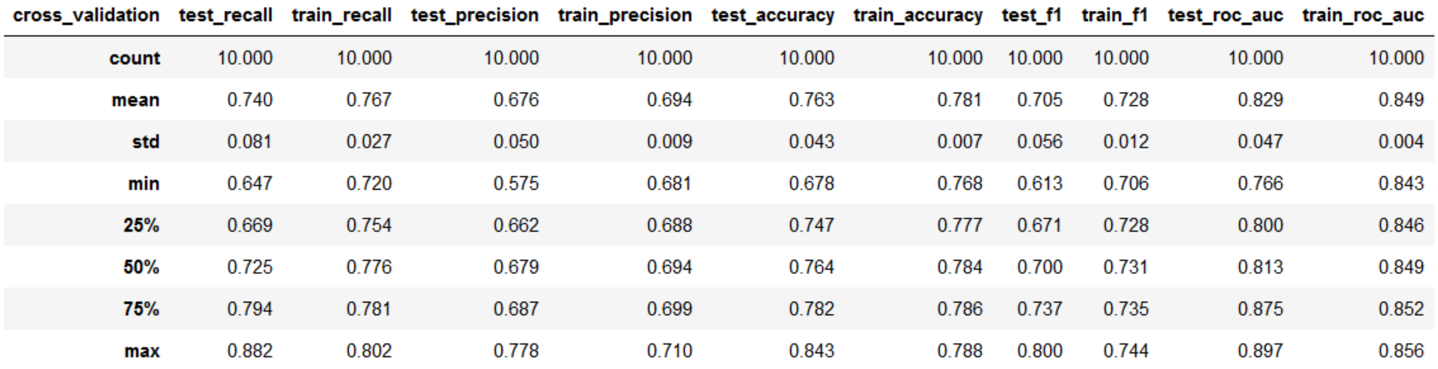
- Testing the training set using a 10-fold cross validation design showed a test mean ROC AUC score of 0.829, accuracy score of 0.763, with test precision and recall at 0.676 and 0.740. Test F1 score was 0.705.
- The AUC value of 0.829 indicated that there were few type I and few type II errors, and that the model is a good measure of separability as it reproduces the data well. However, the AUC score with Logistic Regression was significantly better.

Cross Validation Scores:



cross_validation	test_recall	train_recall	test_precision	train_precision	test_accuracy	train_accuracy	test_f1	train_f1	test_roc_auc	train_roc_auc
0	0.657	0.720	0.575	0.695	0.678	0.772	0.613	0.707	0.766	0.850
1	0.714	0.726	0.658	0.686	0.744	0.768	0.685	0.706	0.831	0.851
2	0.647	0.776	0.647	0.699	0.730	0.786	0.647	0.735	0.788	0.856
3	0.882	0.782	0.682	0.685	0.798	0.778	0.769	0.730	0.889	0.845
4	0.794	0.747	0.675	0.710	0.775	0.786	0.730	0.728	0.809	0.853
5	0.735	0.776	0.676	0.699	0.764	0.786	0.704	0.735	0.813	0.852
6	0.706	0.776	0.686	0.699	0.764	0.786	0.696	0.735	0.814	0.848
7	0.647	0.802	0.688	0.694	0.753	0.788	0.667	0.744	0.797	0.848
8	0.824	0.789	0.778	0.681	0.843	0.777	0.800	0.731	0.891	0.843
9	0.794	0.776	0.692	0.693	0.784	0.782	0.740	0.732	0.897	0.843

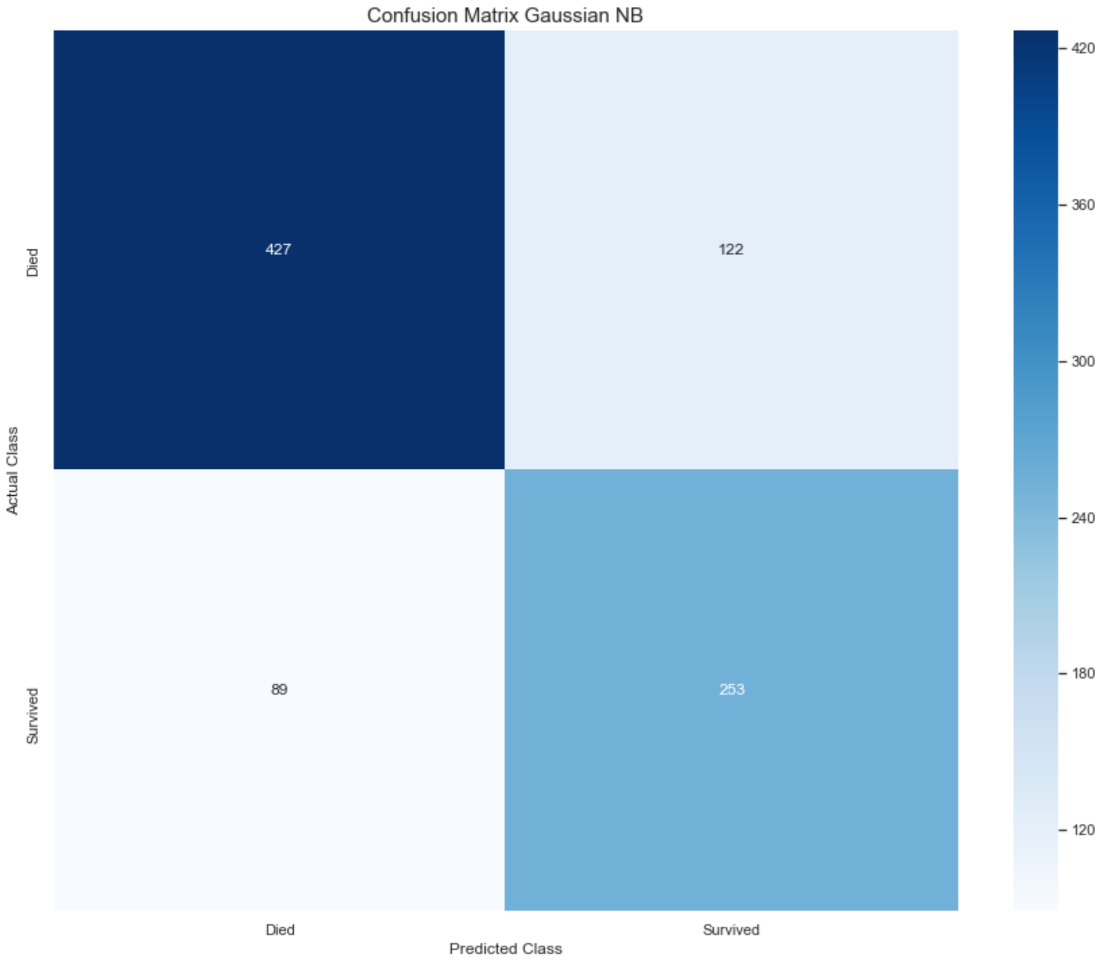
Test Statistics:



cross_validation	test_recall	train_recall	test_precision	train_precision	test_accuracy	train_accuracy	test_f1	train_f1	test_roc_auc	train_roc_auc
count	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000
mean	0.740	0.767	0.676	0.694	0.763	0.781	0.705	0.728	0.829	0.849
std	0.081	0.027	0.050	0.009	0.043	0.007	0.056	0.012	0.047	0.004
min	0.647	0.720	0.575	0.681	0.678	0.768	0.613	0.706	0.766	0.843
25%	0.669	0.754	0.662	0.688	0.747	0.777	0.671	0.728	0.800	0.846
50%	0.725	0.776	0.679	0.694	0.764	0.784	0.700	0.731	0.813	0.849
75%	0.794	0.781	0.687	0.699	0.782	0.786	0.737	0.735	0.875	0.852
max	0.882	0.802	0.778	0.710	0.843	0.788	0.800	0.744	0.897	0.856

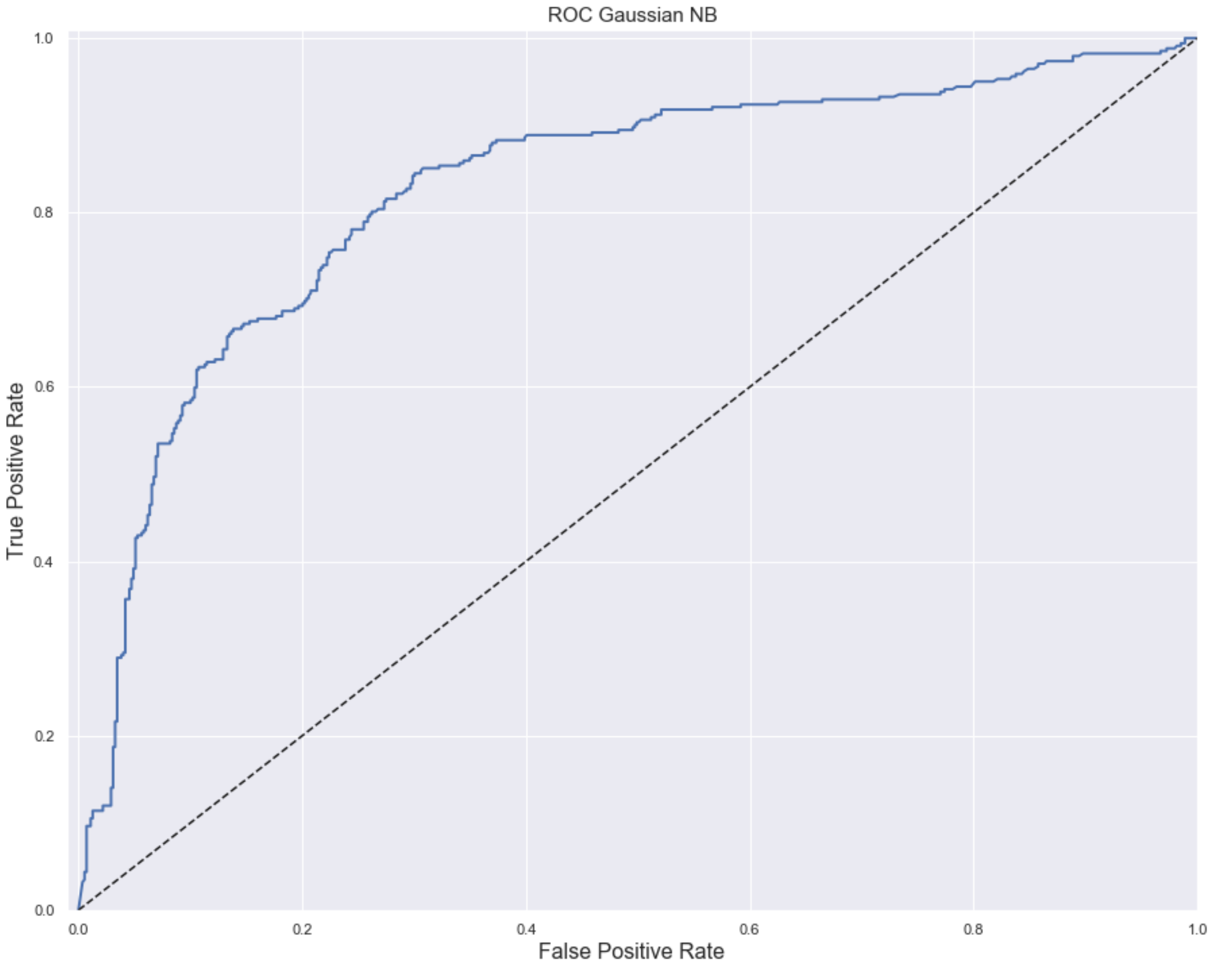
Exploration and Data Visualization of Gaussian NB:

Figure 23 – Confusion Matrix (Gaussian NB):



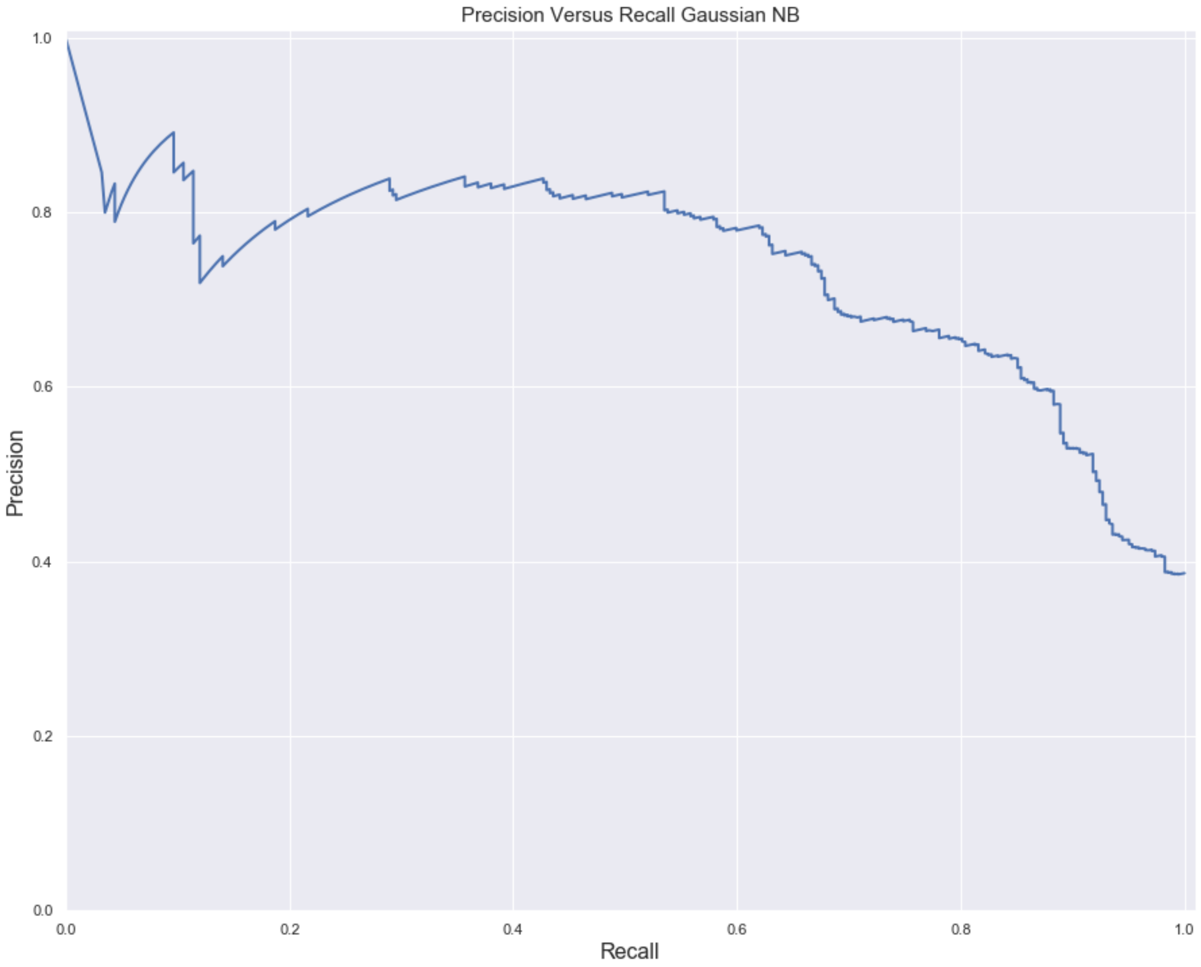
Confusion matrix shows that 427 passengers were correctly classified as died, and 122 passengers were wrongly classified as survived. 89 passengers were wrongly classified as died, and 253 were correctly classified as survived. The confusion matrix shows significantly more false positives than Logistic Regression.

Figure 24 – ROC Curve (Gaussian NB):



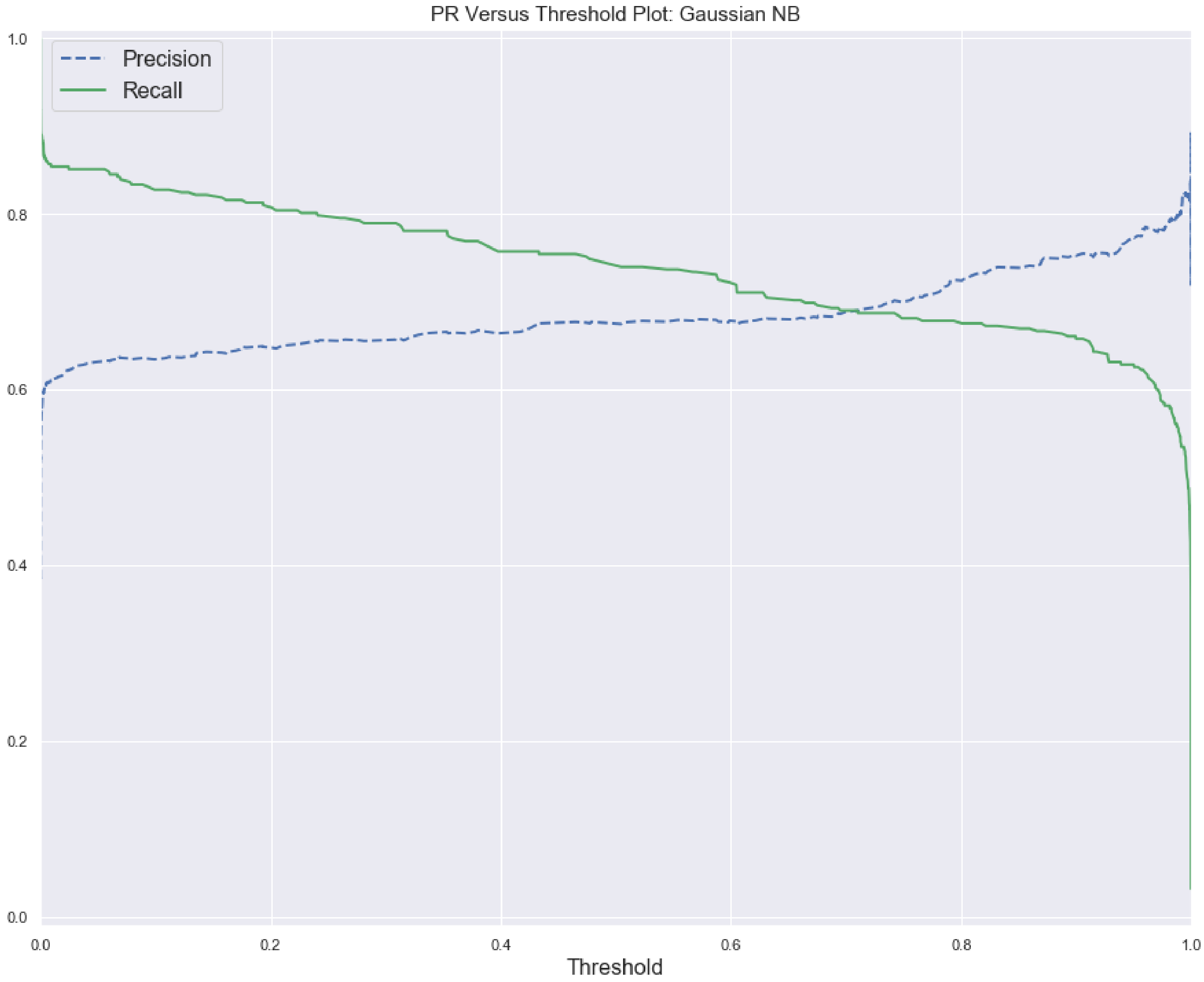
- The ROC curve indicates that there were fewer type I and fewer type II errors, and that the model is an acceptable measure of separability as it reproduces the data well.
- The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). The model seems to do this pretty well. This was confirmed in the validation scores where the mean AUC was equal to 0.829. However, the curve is not as good as the one shown for Logistic Regression.
- Since there were fewer positives (survived) than negatives (died) the model is further examined using the precision recall curve.

Figure 25 – Precision Versus Recall Curve (Gaussian NB):



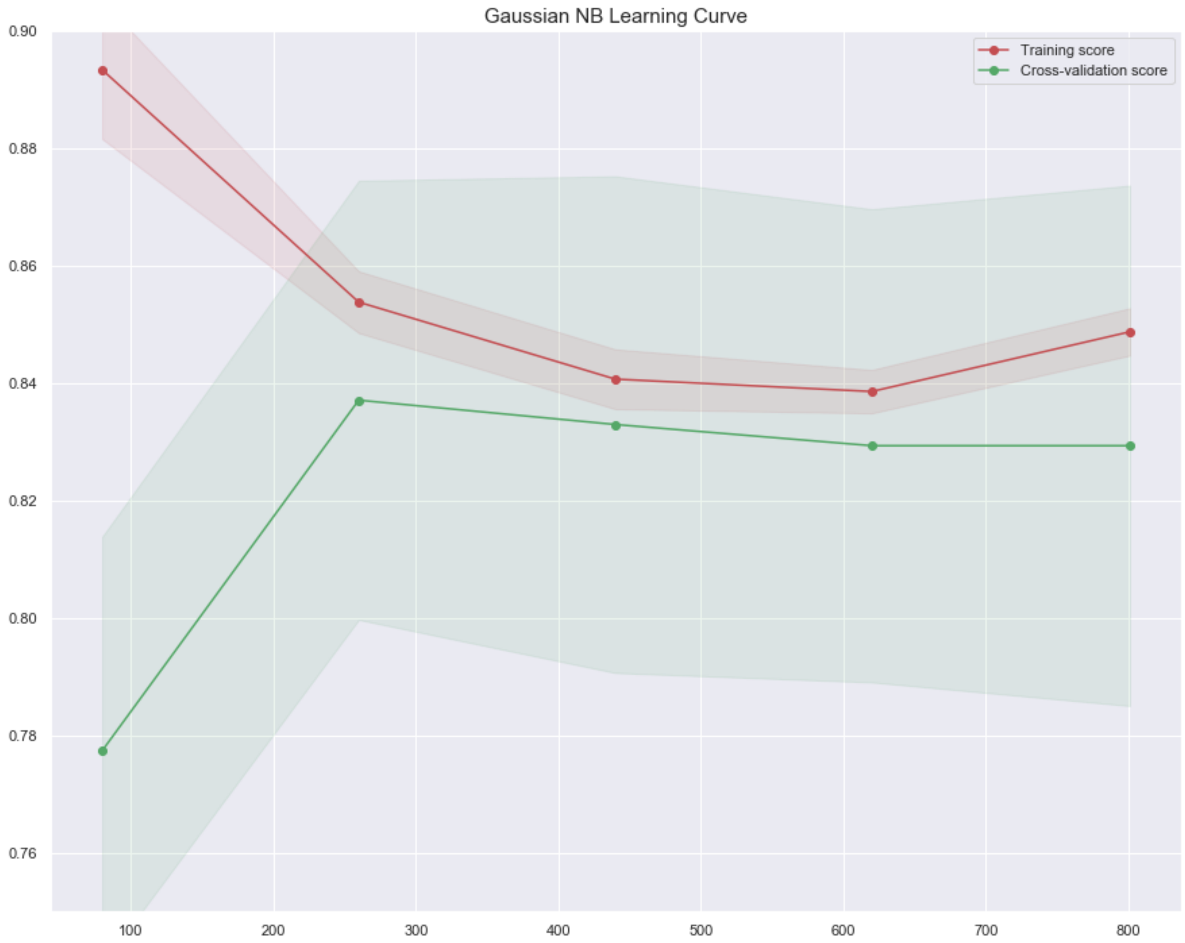
- The precision recall curve shows a significant difference from the curve for Logistic Regression. The curve is jagged and does not get close to the top right corner. The curve seems to look flatter as well, which indicates that precision does not fall as steep as recall increases. However, the mean f1 score of 0.705 is much lower than Logistic regression. Also, the model seems to focus more on recall than precision as the mean score was 0.676 and the mean recall score was at 0.740 indicating that the tradeoff was not ideal.
- Therefore, the curve shows that although the ROC curve and AUC score seemed okay, the PR curve shows that there is room for improvement. This is also supported by the confusion matrix as it showed a higher number of false positives than Logistic Regression.

Figure 26 – Precision and Recall Versus Threshold (Gaussian NB):



The precision and recall versus threshold plot shows that the line between precision and recall crosses around 0.8. At that intersection the scores look to be around 0.70. At the 0.5 point of the threshold, the lines separating the two seem to be in line with the precision and recall scores. It seems that if the threshold could be raised to around 0.8 the model may perform better with less false positives, yet recall would be affected. However, this is recommended if possible.

Figure 27 – Learning Curve (Gaussian NB):



The learning curve for Gaussian NB seems to show similarity with the learning curve for Logistic Regression. However, the gap is very wide in the beginning and though the cross-validation line rises as the training set increases it starts to drop after the 400 mark. The gap then widens again showing that the model does not generalize as well as Logistic Regression. This indicates that a more complex model may be required.

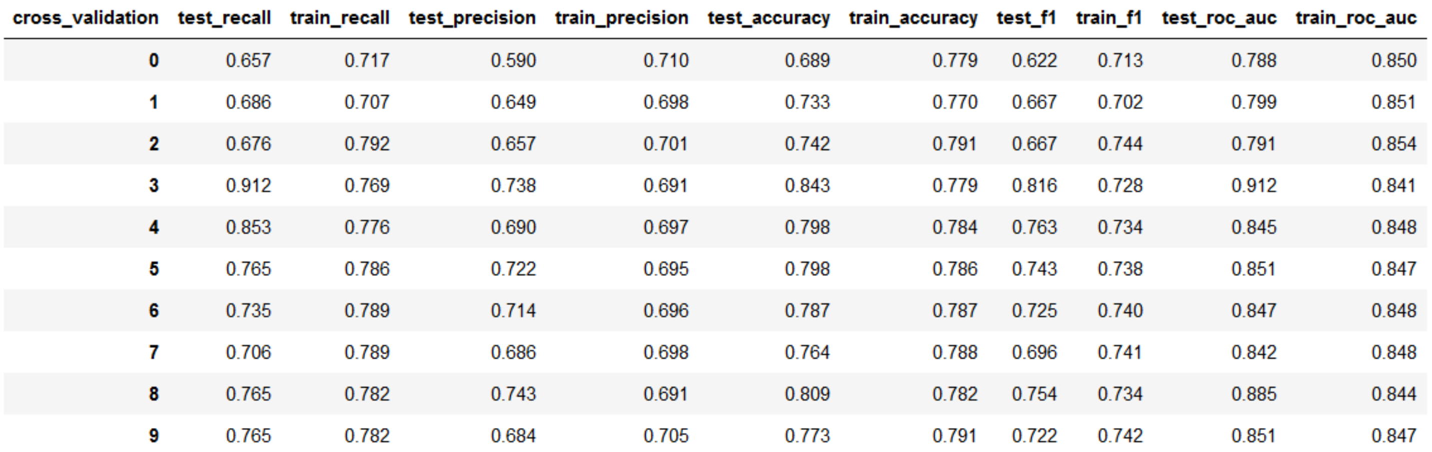
Naïve Bayes Using Bernoulli NB:

- Like Gaussian NB, Bernoulli NB is a classifier that is similar to linear models. It is faster in training. But the price paid for efficiency provides generalization performance that is slightly worse than linear classifiers like Logistic Regression.
- This model works very well with high-dimensional sparse data and is relatively robust to parameters. Bernoulli Naive Bayes models are great baseline models and are often used on very large datasets, where training even a linear model might take too long.
- Bernoulli NB assumes that the data is binary.
- Bernoulli NB is mostly used for sparse count data such as text for text classification.
- If handed any other kind of data, a Bernoulli NB instance may binarize its input, such as in this case as not all data was binarized in preprocessing.

Bernoulli NB Score Results:

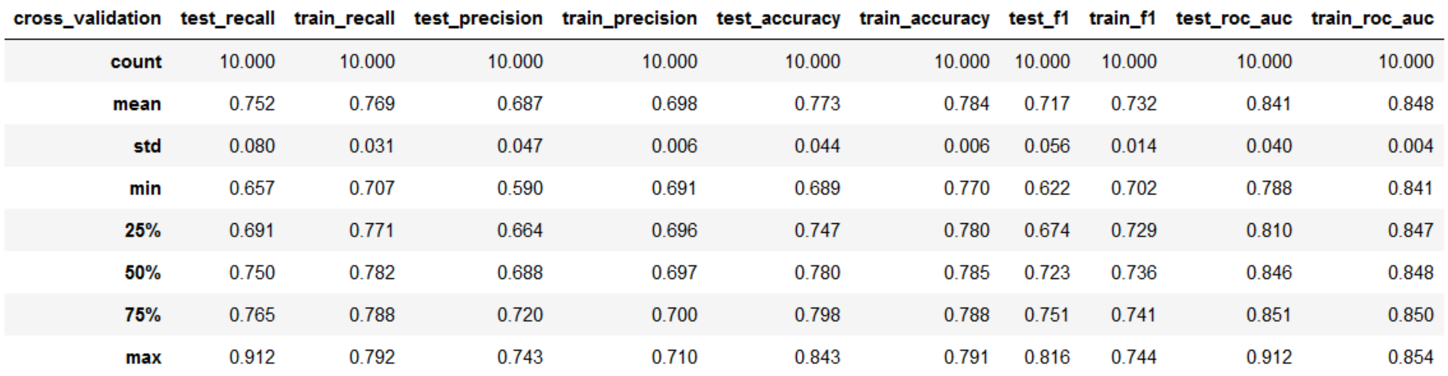
- Testing the training set using a 10-fold cross validation design showed a test mean ROC AUC score of 0.841, accuracy score of 0.773, with test precision and recall at 0.687 and 0.752. Test F1 score was 0.717.
- The AUC value of 0.841 indicates that there were few type I and few type II errors, and that the model is a good measure of separability as it reproduces the data very well. However, the AUC score with Logistic Regression was significantly better, but this method performed better than Gaussian NB.

Cross-validation Scores:



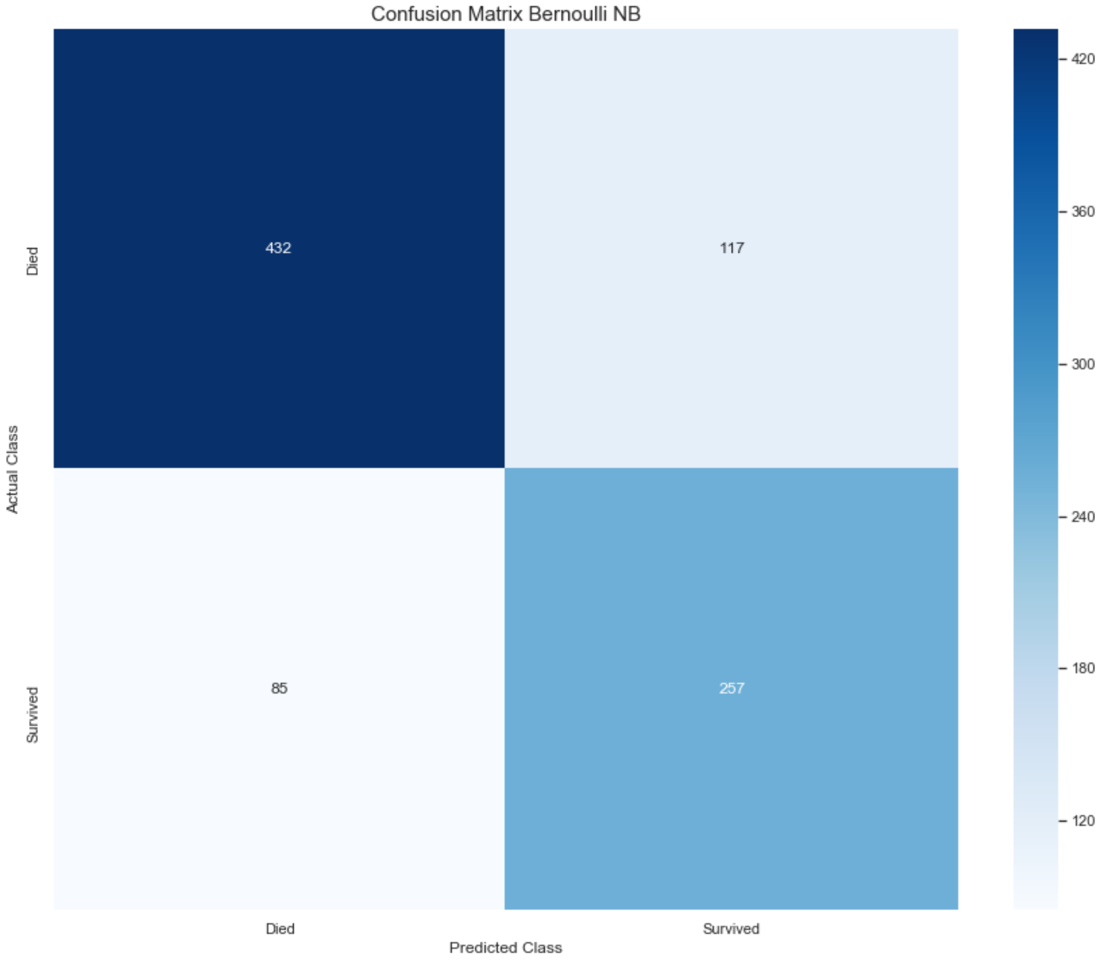
cross_validation	test_recall	train_recall	test_precision	train_precision	test_accuracy	train_accuracy	test_f1	train_f1	test_roc_auc	train_roc_auc
0	0.657	0.717	0.590	0.710	0.689	0.779	0.622	0.713	0.788	0.850
1	0.686	0.707	0.649	0.698	0.733	0.770	0.667	0.702	0.799	0.851
2	0.676	0.792	0.657	0.701	0.742	0.791	0.667	0.744	0.791	0.854
3	0.912	0.769	0.738	0.691	0.843	0.779	0.816	0.728	0.912	0.841
4	0.853	0.776	0.690	0.697	0.798	0.784	0.763	0.734	0.845	0.848
5	0.765	0.786	0.722	0.695	0.798	0.786	0.743	0.738	0.851	0.847
6	0.735	0.789	0.714	0.696	0.787	0.787	0.725	0.740	0.847	0.848
7	0.706	0.789	0.686	0.698	0.764	0.788	0.696	0.741	0.842	0.848
8	0.765	0.782	0.743	0.691	0.809	0.782	0.754	0.734	0.885	0.844
9	0.765	0.782	0.684	0.705	0.773	0.791	0.722	0.742	0.851	0.847

Test Statistics:



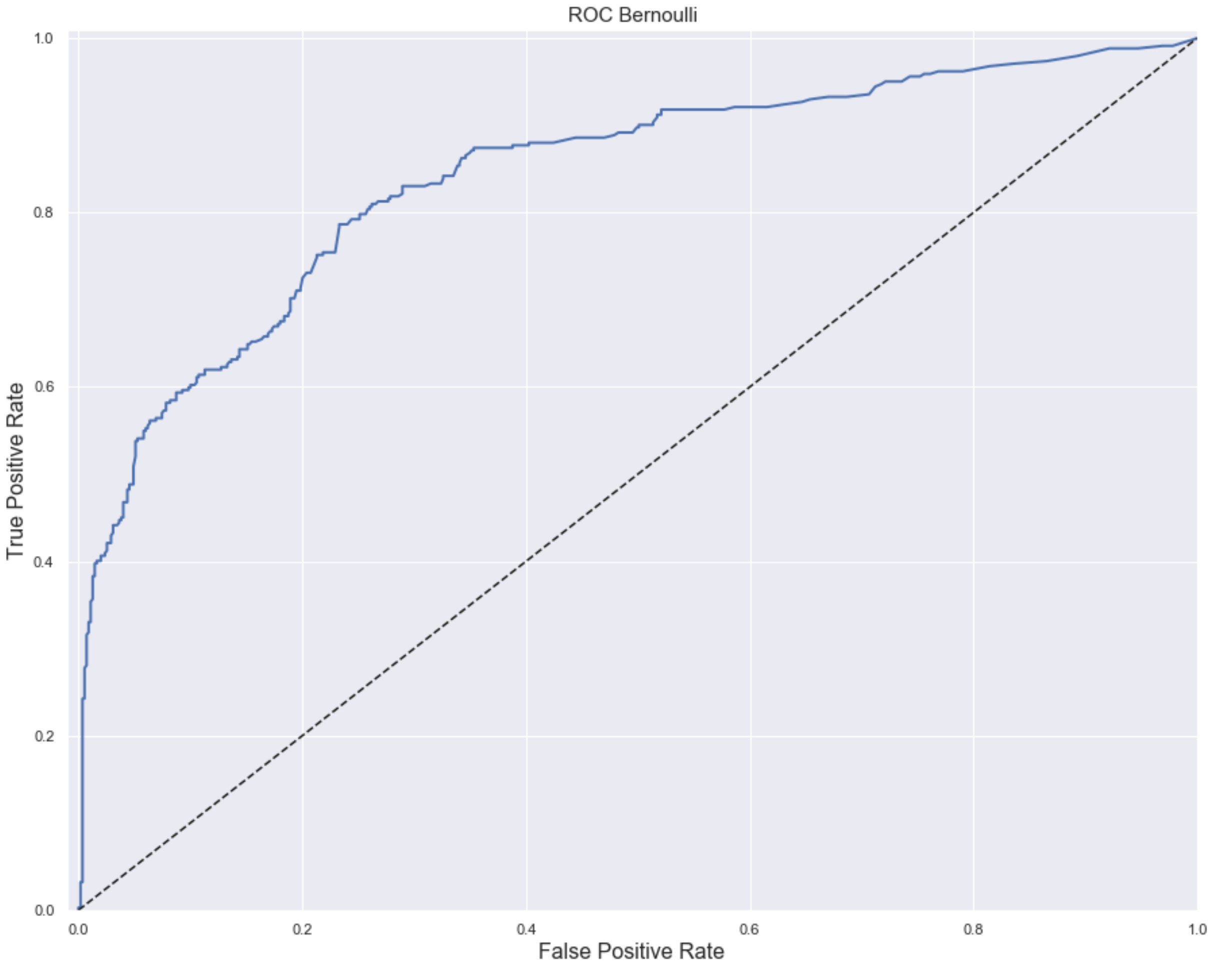
cross_validation	test_recall	train_recall	test_precision	train_precision	test_accuracy	train_accuracy	test_f1	train_f1	test_roc_auc	train_roc_auc
count	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000
mean	0.752	0.769	0.687	0.698	0.773	0.784	0.717	0.732	0.841	0.848
std	0.080	0.031	0.047	0.006	0.044	0.006	0.056	0.014	0.040	0.004
min	0.657	0.707	0.590	0.691	0.689	0.770	0.622	0.702	0.788	0.841
25%	0.691	0.771	0.664	0.696	0.747	0.780	0.674	0.729	0.810	0.847
50%	0.750	0.782	0.688	0.697	0.780	0.785	0.723	0.736	0.846	0.848
75%	0.765	0.788	0.720	0.700	0.798	0.788	0.751	0.741	0.851	0.850
max	0.912	0.792	0.743	0.710	0.843	0.791	0.816	0.744	0.912	0.854

Figure 28 – Confusion Matrix (Bernoulli NB):



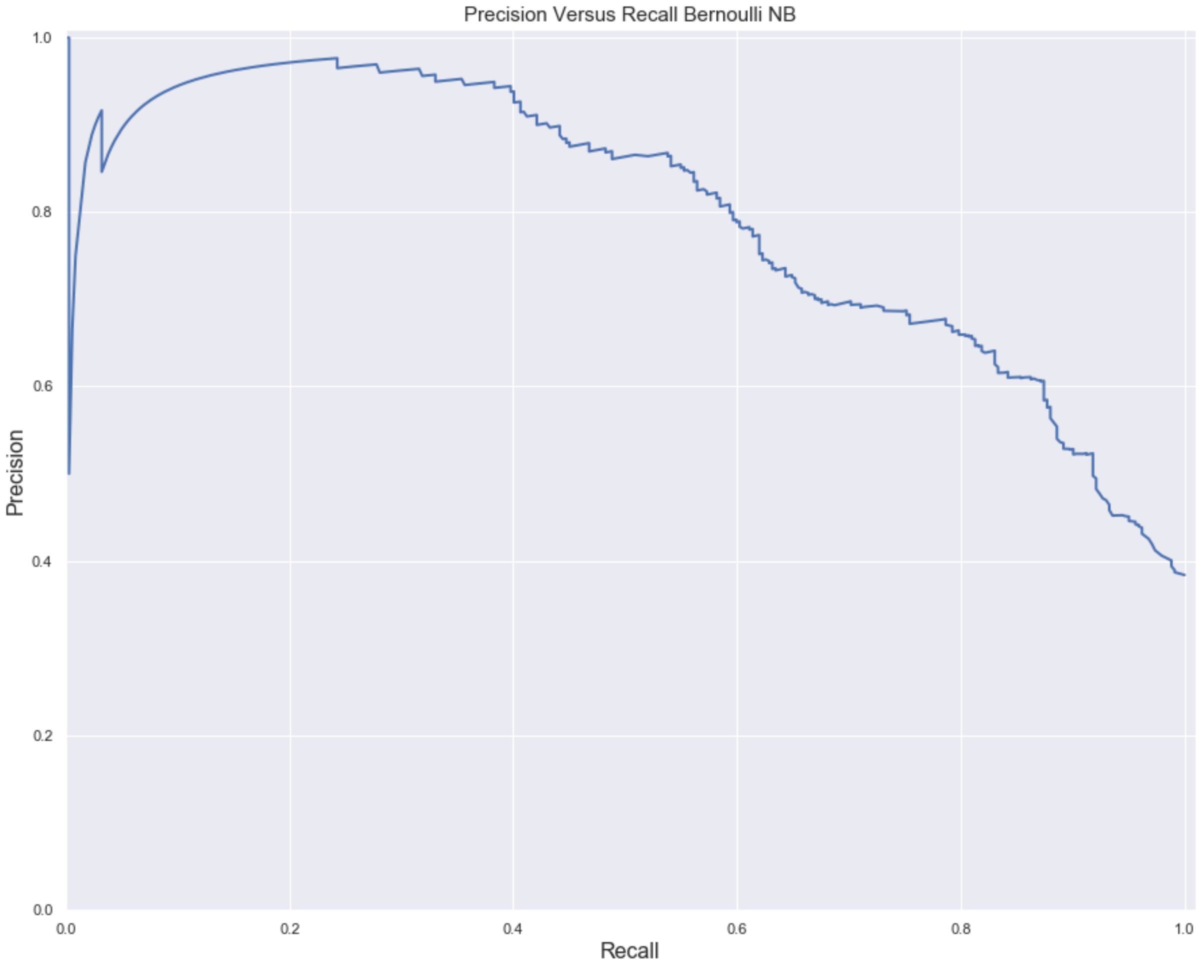
Confusion matrix shows that 432 passengers were correctly classified as died, and 117 passengers were wrongly classified as survived. 85 passengers were wrongly classified as died, and 257 were correctly classified as survived. Confusion matrix shows significantly more false positives than Logistic Regression, but fewer than Gaussian NB.

Figure 29 – ROC Curve (Bernoulli NB):



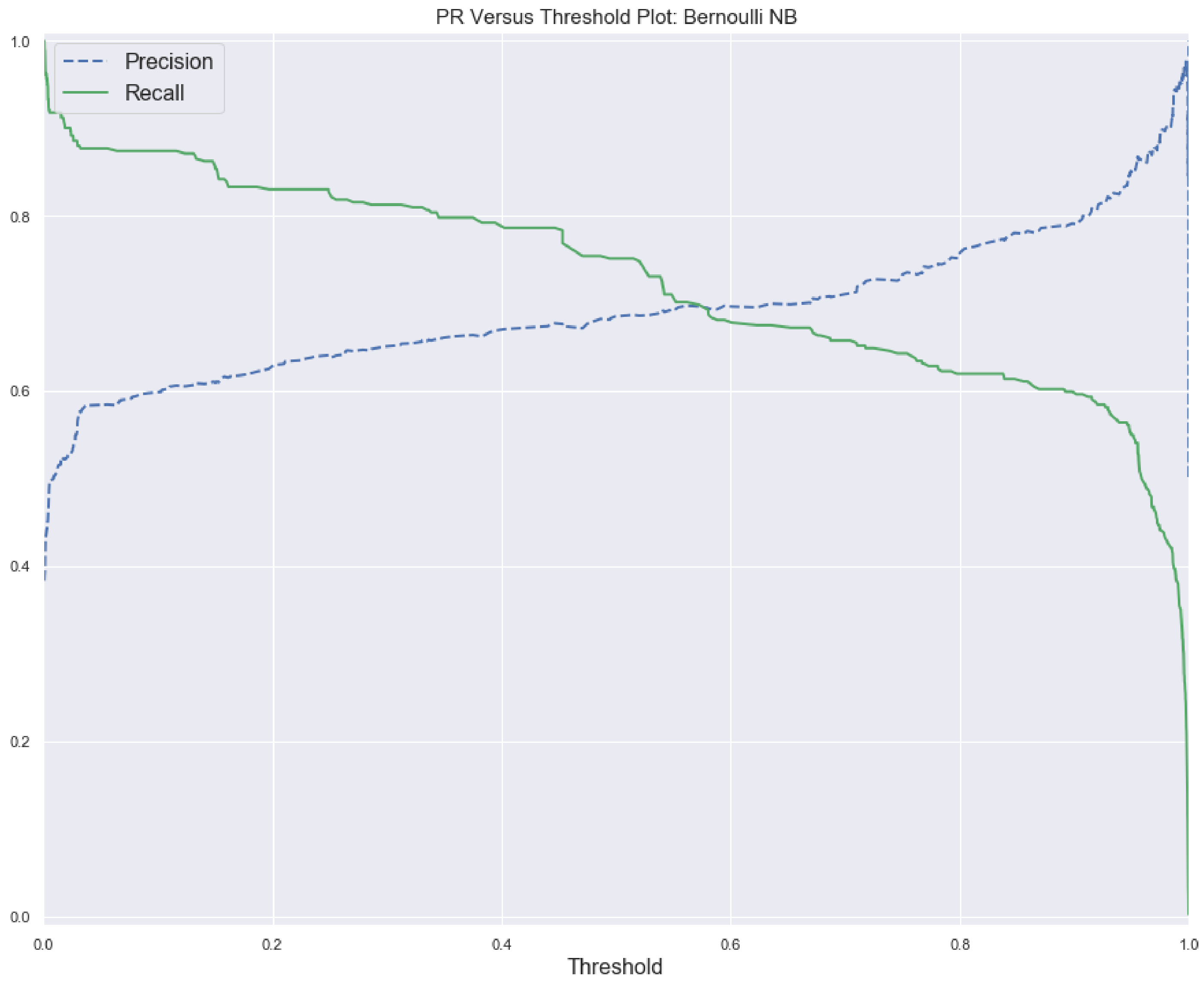
- The ROC curve indicates that there were fewer type I and fewer type II errors, and that the model is an acceptable measure of separability as it reproduces the data well.
- The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). The model seems to do this pretty well. This was confirmed in the validation scores where the mean AUC was equal to 0.841. However, the curve is not as good as the one shown for Logistic Regression and may be better than Gaussian NB.
- Since there were fewer positives (survived) than negatives (died) the model is further examined using the precision recall curve.

Figure 30 – Precision Versus Recall (Bernoulli NB):



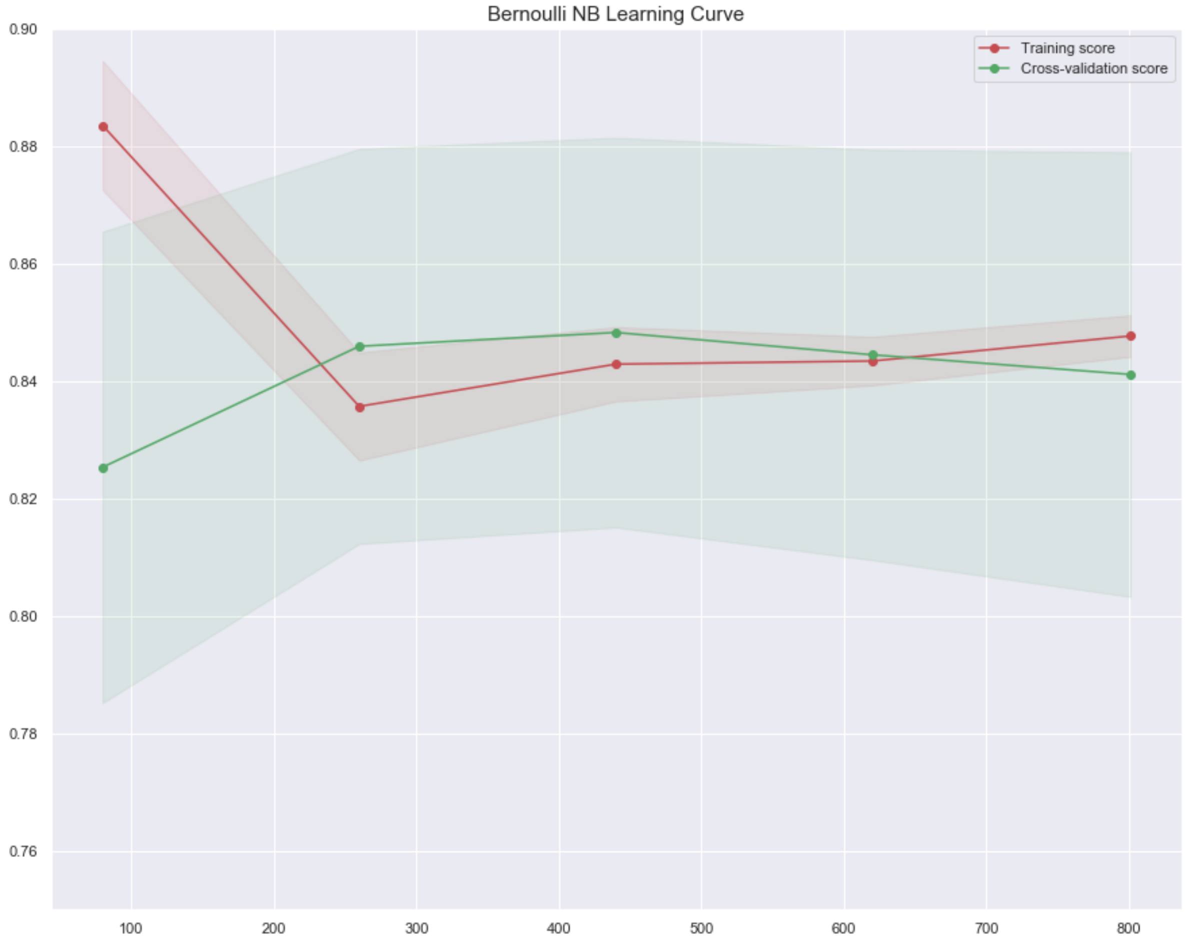
- The precision recall curve shows a significant difference from the curve for Logistic Regression. The curve is also jagged and does not get close to the top right corner, and it looks slightly worse than the curve for Gaussian NB as it significantly drops after a recall of 0.60. Also, like Gaussian NB, the model seems to focus more on recall than precision as the mean score was 0.687 and the mean recall score was at 0.752 indicating that the tradeoff was not ideal.
- Therefore, the curve shows that although the ROC curve and AUC score seemed okay, the PR curve shows that there is room for improvement. This is also supported by the confusion matrix as it showed a higher number of false positives that Logistic Regression, but less than Gaussian NB.

Figure 31 – Precision and Recall Versus Threshold (Bernoulli NB):



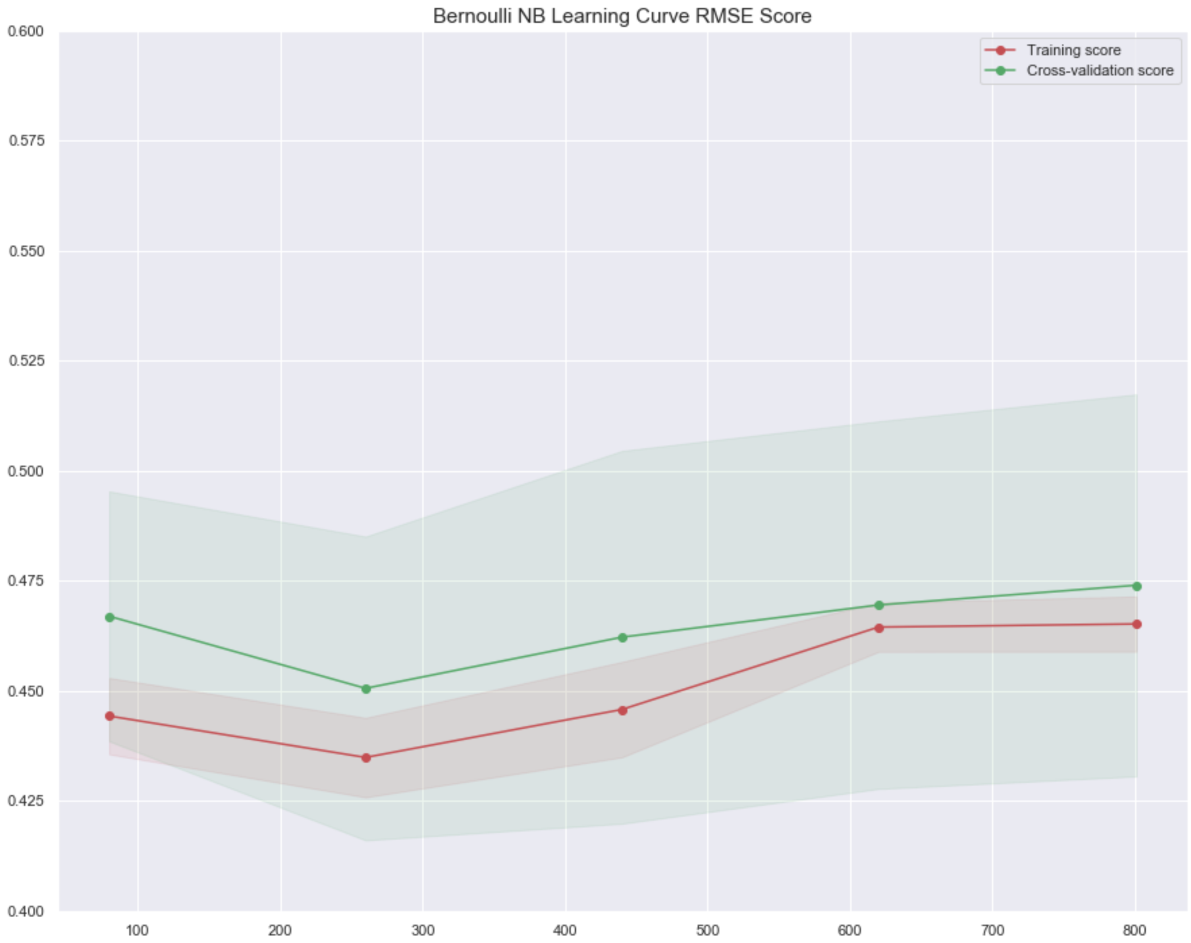
The precision and recall versus threshold plot shows that the line between precision and recall crosses around 0.6. At that intersection the scores look to be around 0.70. At the 0.5 point of the threshold, the lines separating the two seem to be in line with the precision and recall scores. The difference is significant, which was evident in the PR curve. It seems that if the threshold could be raised to around 0.65 the model may perform a little better with less false positives, yet recall would be affected. However, this is recommended if possible.

Figure 32 – Learning Curve (Bernoulli NB):



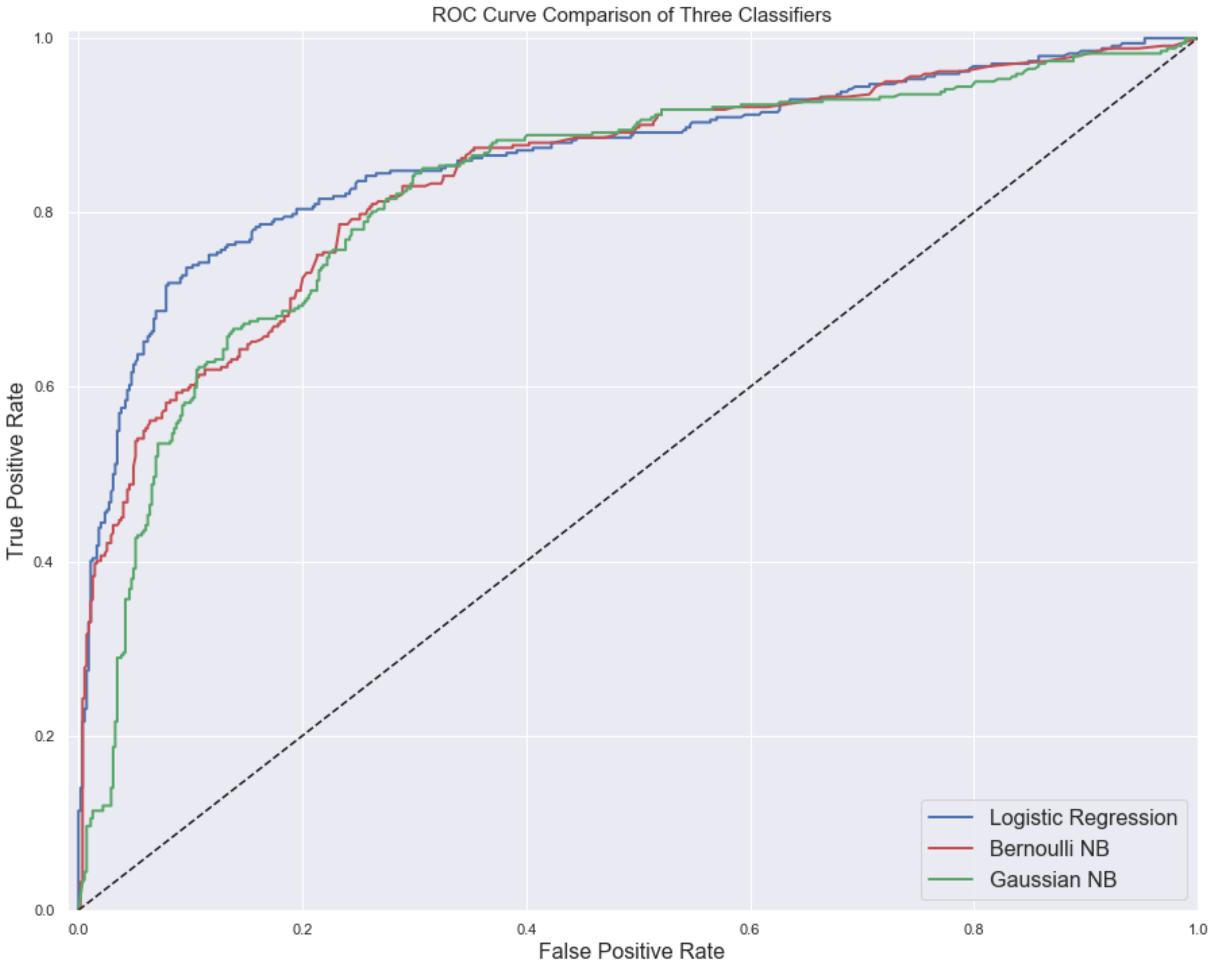
The learning curve for ROC AUC scores seems to be very different from the other models. the cross-validation scores rise as the training scores drop. They also cross at two points. This seems to be a sign of underfitting and the model may not generalize well to new data. This may indicate that a more complex model is required. Underfitting is further tested and visualized in the next figure.

Figure 33 – Learning Curve Using RMSE Score (Bernoulli NB):



The high RMSE score with the narrow gap that rises and plateaus is a classic sign of underfitting. A more complex model is therefore needed.

Figure 34 – ROC Curve Comparison of the Three Classifiers:

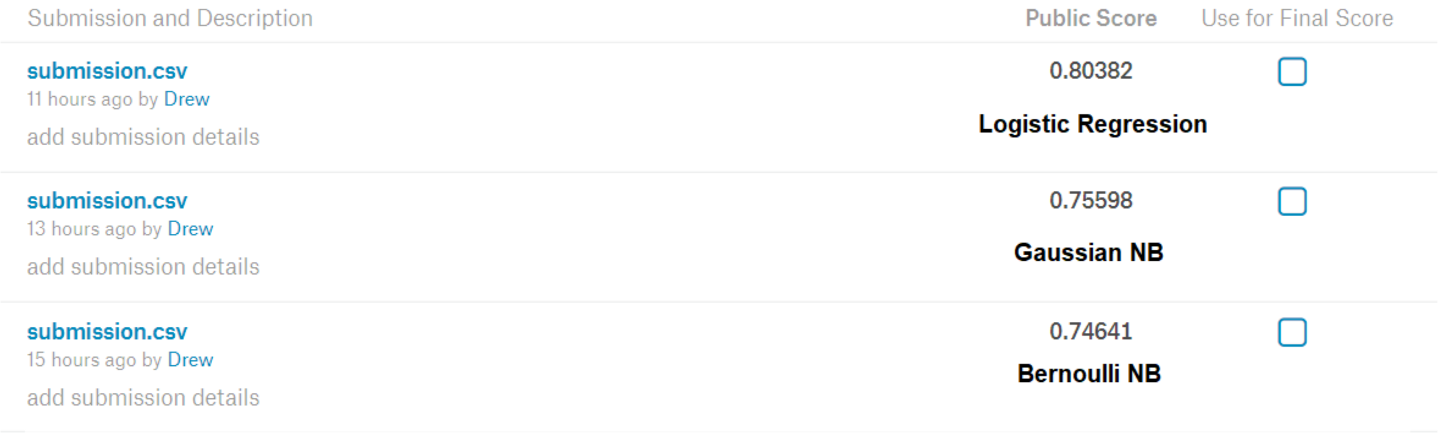


- The ROC curve comparing the three classifiers clearly shows that using Logistic Regression is the better of the three. It stays furthest away from the random classifier line (toward the top-left corner) showing that the model was the best measure of separability, which is also shown in the confusion matrix.
- The PR curves also confirms this as the PR curve for Logistic Regression is closer to top right corner.
- It is therefore recommended that the Logistic Regression model is used.

Conclusion:

- The main objective of this study was to test data using three binary classifiers: Logistic Regression Gaussian Naïve Bayes, and Bernoulli Naïve Bayes on the Titanic data set.
- Cross-validation used was K-fold cross-validation, which randomly splits a training set into 10 distinct subsets called folds, then trains and evaluates the model n (i.e. 10) times, picking a different fold for evaluation every time and training on the other n (i.e. 9) folds.
- Results showed that using a Logistic Regression performed the best as it scored with an average ROC AUC of 0.871 in a 10-fold cross-validation design.
- Results also showed that the model with the best ROC curve, PR curve, PR versus threshold curve, and learning curve was Logistic Regression.
- Accuracy score on the test set submitted to Kaggle showed a score of 0.80382.

Figure 35 – Kaggle Score for All Three Models:



Submission and Description	Public Score	Use for Final Score
submission.csv 11 hours ago by Drew add submission details	0.80382 **Logistic Regression**	□
submission.csv 13 hours ago by Drew add submission details	0.75598 **Gaussian NB**	□
submission.csv 15 hours ago by Drew add submission details	0.74641 **Bernoulli NB**	□