A. Background

Titanic was the largest passenger ship ever made, collided with an iceberg on april 15,1912. It almost killed 1502 out of 2224 passengers. The major reason for loss of life was insufficient life boats but few people survived.

Here, we are analysing with a sample of 1309 observations to understand more about the survivability of the passengers travelled.

B. Objective

Build a predictive model that answers the question: “What sorts of people were more likely to survive?”

The project explores s 4 methods (Naïve Bayes, Logistic Regression, kNN and Decision Trees), and identifies the best method.

C. Data Exploration

-->The Dataset (Meta Data)

Variable	Description	Variable Info
survived	Survival of a person	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st class, 2 = 2nd class, 3 = 3rd class
sex	Gender	M= Male, F= Female
Age	Age in years	Decimal (1 to 80 years)
sibsp	Number of siblings / spouses aboard the Titanic	integer
parch	Number of parents / children aboard the Titanic	integer

ticket	Ticket number	Varchar
fare	Passenger fare	Float
cabin	Cabin number	Varchar
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S =Southampton
body	Recovered or not	Provided in numbers ; Not recovered is null
dest	port of destination	string

Below is the number of observations and variables from the provided data set.

```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   survived  1309 non-null   int64
 1   pclass    1309 non-null   int64
 2   name      1309 non-null   object
 3   sex       1309 non-null   object
 4   age       1046 non-null   float64
 5   sibsp     1309 non-null   int64
 6   parch     1309 non-null   int64
 7   ticket    1309 non-null   object
 8   fare      1308 non-null   float64
 9   cabin     295 non-null    object
 10  embarked  1307 non-null   object
 11  boat      486 non-null    object
 12  body      121 non-null    float64
 13  dest      745 non-null    object
dtypes: float64(3), int64(4), object(7)
```

The above table is derived from the python code , refer to appendix 1.1

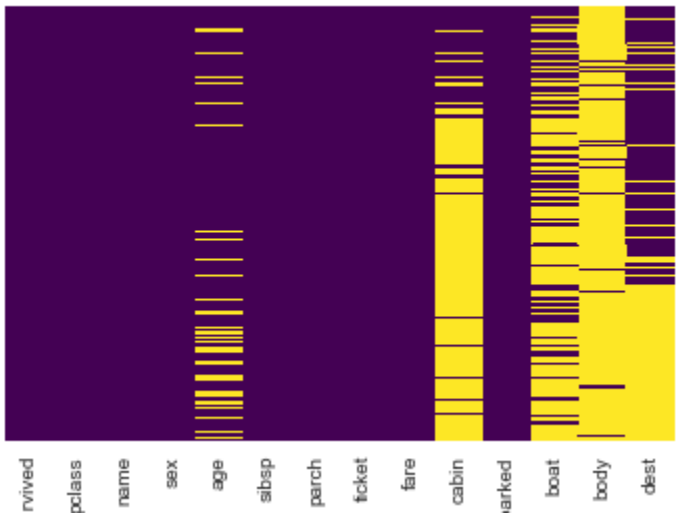
- Descriptive Statistics:-

5-Number Summary:

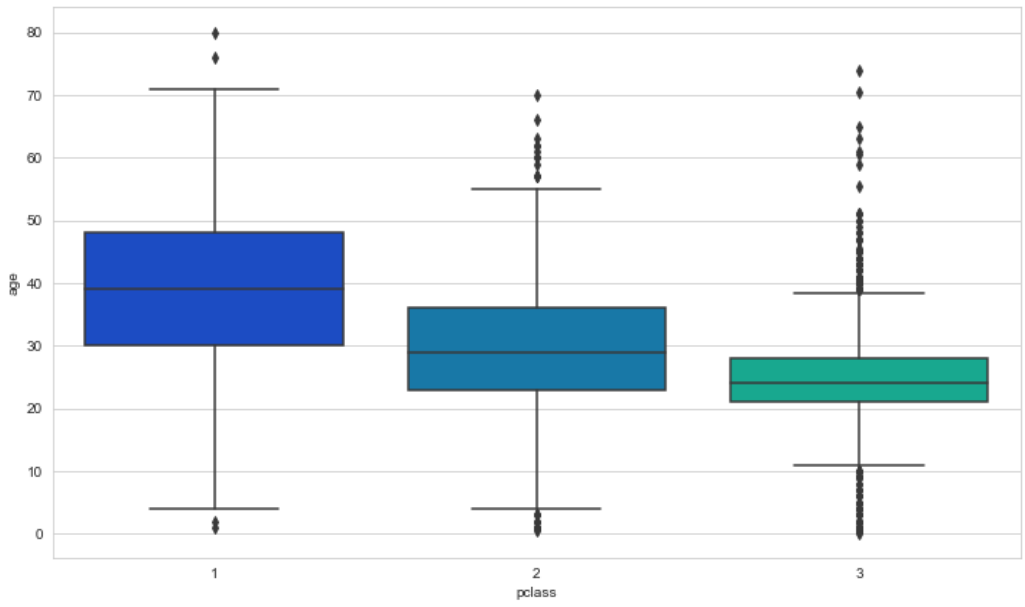
	survived	pclass	age	sibsp	parch	fare	body
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	121.000000
mean	0.381971	2.294882	29.881135	0.498854	0.385027	33.295479	160.809917
std	0.486055	0.837836	14.413500	1.041658	0.865560	51.758668	97.696922
min	0.000000	1.000000	0.166700	0.000000	0.000000	0.000000	1.000000
25%	0.000000	2.000000	21.000000	0.000000	0.000000	7.895800	72.000000
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200	155.000000
75%	1.000000	3.000000	39.000000	1.000000	0.000000	31.275000	256.000000
max	1.000000	3.000000	80.000000	8.000000	9.000000	512.329200	328.000000

Conclusions through graphs:-

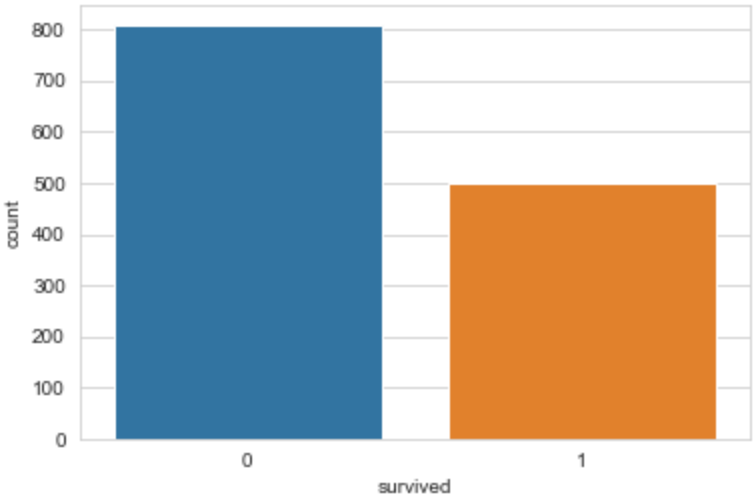
1. We verified the number of null values are present for the fields- age, cabin, boat, body and dest through heatmap.



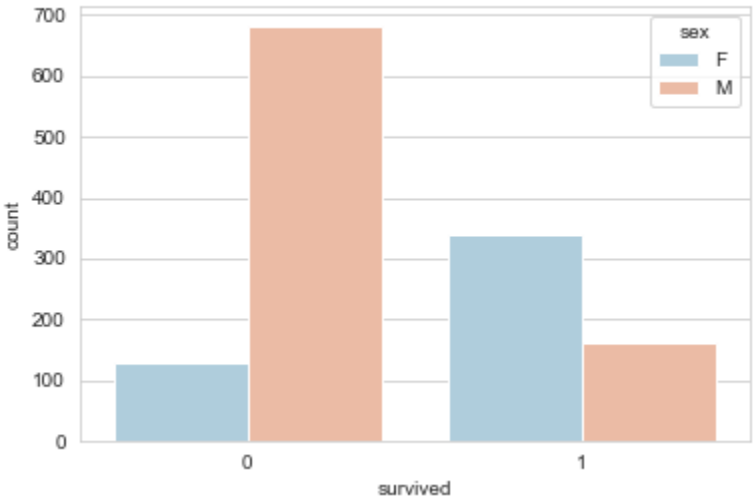
2. We classified the pclass and observed the age groups w.r.t pclass. For example,the people with age between 28 to 50 years are mostly in first class and people with age



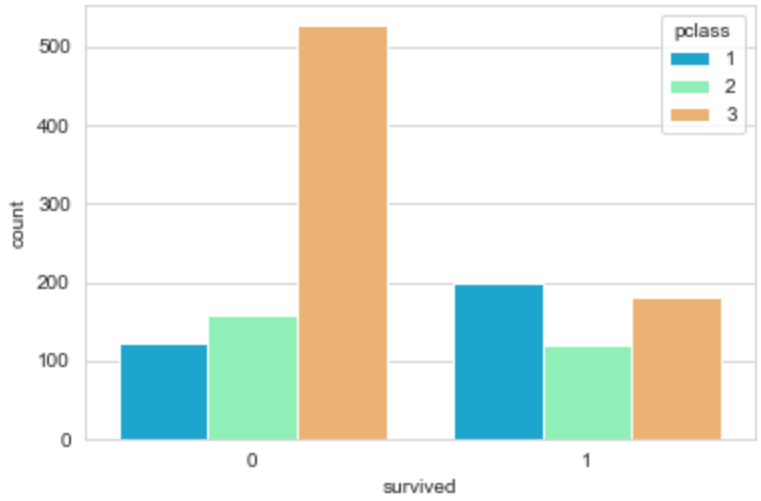
3. The number of people who are not survived are 808 and the people who survived are 500.



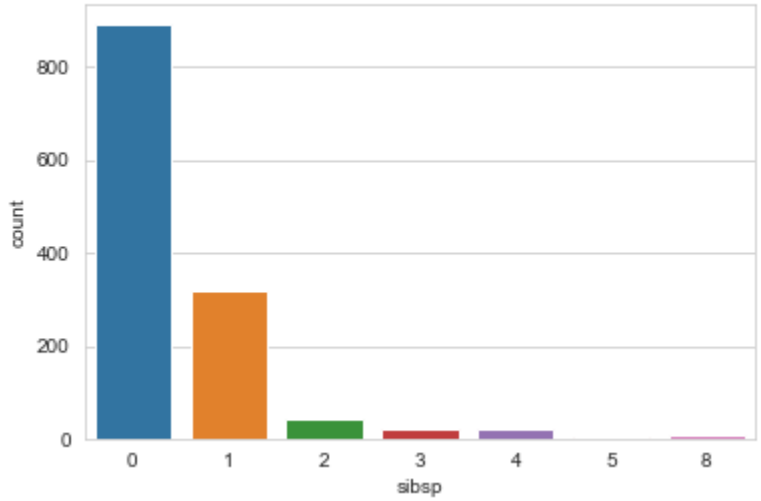
4. The survival rate of female is 339 and it's more than male, which is 161.



5. The people in pclass 3 have less chance of survival rate compared to other classes (pclass 1 and 2). From the below graph, we can conclude that 528 people from pclass 3 could not survive.



6. The below is the analysis of the number of the people in the ship who have spouses/siblings.



- Data Wrangling:-

Modified Variables:

- We modified the “age” variable based on the median of age on pclass. We did this as we identified there were 263 nulls in the data and we considered the age as an important factor influencing the data. The median values were 39,29 and 24 and the corresponding “pclass” was 1, 2 and 3 respectively. *Refer to appendix 3.1.*
- We converted the variable “Sex” into Binary as 0’s and 1’s as it was in object and it must be in float inorder to include that into the x variable. M =1 and F= 0. *Refer to appendix 3.2.*

- Modified the variable “Boat” into another variable “boat_mod” replacing the null values with 0’s and the nonnull values with 1’s in an idea whether the person got into a boat or not. *Refer to Appendix 3.3.*
- Modified the variable “Body” into a variable “body_mod” replacing the null values with 0’s and the nonnull values with 1’s in an idea, if the person is dead whether the person's body is recovered or not. It will make us classify in a better way for non-survived patients. *Refer to Appendix 3.4.*

Dropped Variables:

- Dropped a variable “cabin” as there were 1014 null values in it and was not useful in tuning the model.
- Dropped a variable “dest” as the ship itself did not reach destination and also there were many nulls to it. Intuitively thought to remove the variable as it doesn’t help the classifiers significantly.
- Dropped a variable “Ticket” as it doesn’t signify anything in decision making and it is an alphanumeric variable, Hence intuitively thought to remove the variable.*Refer to appendix 3.5.*

Retained variables:

- · The retention was based on the variables when split into x and Y. since we modified a few variables into different variables we have removed those variables.
- · The retained variables are “pclass” , “sibsp”, “parch”, “fare”, “Gender_Binary”, “embarked_Binary”, “boat_mod”, “body_mod”. These are the retained x variables and “survived” is y – target variable.

Additional wrangling:

- We modified the nulls in fare with median of fare and nulls in embarked_binary with 1 or “s” as per original data , as that’s the mode for the embarked_Binary variable. *Refer to appendix 3.6.*

D. Optimal Method

The data which is wrangled then was split into X_train, X_test, y_train, y_test with 30% of the data into test and 70% for training. The code random state 100 was used inorder to keep the 100th combination of splitting throughout the modeling.

The methods we tried were

1. Logistic regression:

- We used a plain linear regression without using solver or multiclass and ran the linear regression model to fit with the train data and then we tried with the score to understand the fit of the model. The score turned out to be 0.9847. The output of the predict was in an array and did some manual verification inorder to find the accuracy.
- The 10-fold cross validation error was `[0.95652174, 0.95652174, 0.9673913 , 0.9673913 , 0.97826087,1. , 0.94505495, 0.96703297, 0.98901099, 0.98901099]` and the mean is 0.9716. The test error is 1-0.9716.
- The confusion matrix accuracy is 0.9847, *Refer to appendix 4.1.*
- The significance of each and every variable was found and except for age and fare, all others were significant, *Refer to appendix 4.2.*

2. Naïve Bayes:

- Since the random state is mentioned as 100 the same test and train data is used in order to do naïve bayes method and the score of the model is 0.9746.
- The 10 fold cross validation error was `[0.95652174, 0.95652174, 0.93478261, 0.95652174, 0.95652174, 1. , 0.94505495, 0.96703297, 0.97802198, 0.97802198]` and the mean is 0.9629. The test error is 1- 0.9629.
- The feature significance suggests that “boat_mod”, “body_mod” and “fare” are top 3 important features in classifying with 3 NN- boat_mod has the highest significance with 0.3557 *Refer to appendix 4.4.*
- The confusion matrix accuracy is 0.9746, *Refer to appendix 4.3.*

3. KNN:

- Knn has a Mean accuracy score of 0.9160 and the distance used is by default the minkowski distance, and the number of n’s that is taken in the model is 3, as it is taught the same in the class.
- The 10 fold cross validation error was`[0.88043478, 0.80434783, 0.84782609, 0.93478261, 0.85869565,0.89130435, 0.91208791, 0.91208791, 0.9010989 , 0.84615385])` and the mean is 0.8789. The test error is 1- 0.8789.

- The feature significance suggests that “boat_mod”, “fare” and “sibsp” are top 3 important features in classifying with 3 NN- boat_mod has the highest significance with 0.261.
- The confusion matrix accuracy is 0.9160 , *Refer to appendix 4.5.*

4. Decision Tree:

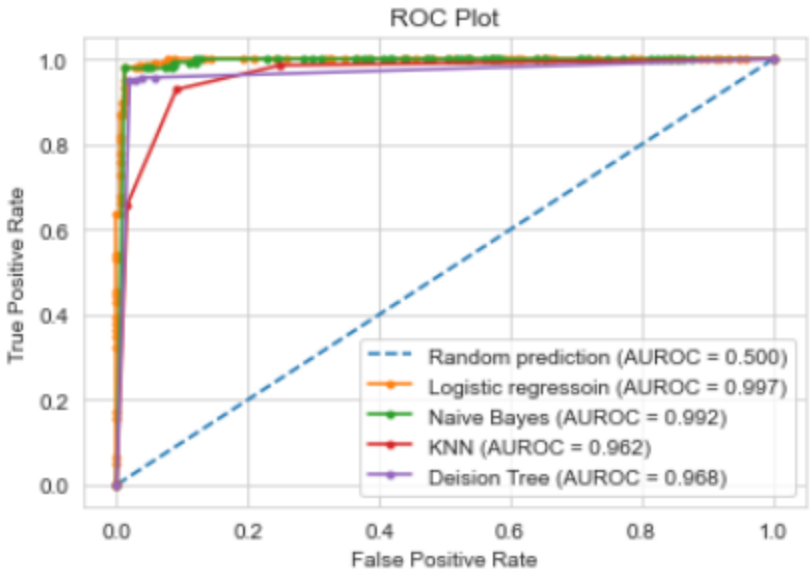
- Decision tree mean accuracy is 0.9695.
- The 10 fold cross validation error was `[0.93478261, 0.93478261, 0.9673913 , 0.97826087, 0.97826087,0.97826087, 0.93406593, 0.96703297, 0.97802198, 0.96703297]` and the mean is 0.9618. The test error is 1- 0.9618.
- The Decision tree feature importance signifies that the tree is divided from the first classifier as “boat_mod” *Refer to appendix 4.6.*
- The confusion matrix is attached. *Refer to appendix 4.8.*

The optimal method with the above insights and scores is Logistic Regression. The scores are mentioned below

	Model	Score
1	LogisticRegression	0.984733
3	Naive Bayes	0.974555
0	Decision Tree	0.969466
2	K-Nearest Neighbours	0.916031

By drafting the ROC curve

- The x-axis showing 1 – specificity (= false positive fraction = FP/(FP+TN)).
- The y-axis showing sensitivity (= true positive fraction = TP/(TP+FN))



The Logistic regression has the best Area Under receiver operating characteristic curve, and the scores are mentioned in the above chart.

Conclusion:

Based on the Test error method, as recommended by the template the Logistic regression has the least test error with just 0.0284. Hence, Logistic Regression is the optimal method for titanic analysis.

E. Appendix

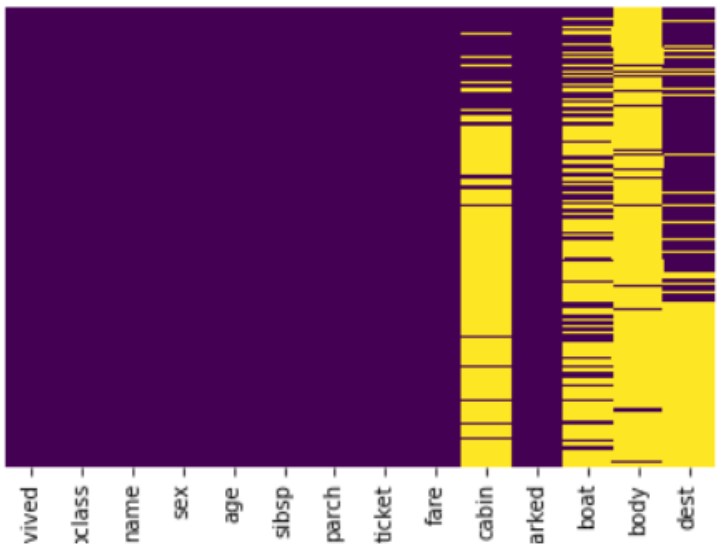
1.1

Exploring data Types

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   survived  1309 non-null   int64
 1   pclass    1309 non-null   int64
 2   name      1309 non-null   object
 3   sex       1309 non-null   object
 4   age       1046 non-null   float64
 5   sibsp     1309 non-null   int64
 6   parch     1309 non-null   int64
 7   ticket    1309 non-null   object
 8   fare      1308 non-null   float64
 9   cabin     295 non-null    object
 10  embarked  1307 non-null   object
 11  boat      486 non-null    object
 12  body      121 non-null    float64
 13  dest      745 non-null    object
```

3.1



3.2, 3.3, 3.4

name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	dest	Gender_Binary	embarked_Binary	boat_mod	body_mod
Allen, Miss. Elisabeth Walton	F	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO	0.0	1.0	1.0	0.0
Allison, Master. Hudson Trevor	M	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON	1.0	1.0	1.0	0.0
Allison, Miss. Helen Loraine	F	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON	0.0	1.0	0.0	0.0
Allison, Mr. Hudson Joshua Creighton	M	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON	1.0	1.0	0.0	1.0
Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	F	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON	0.0	1.0	0.0	0.0

3.5

	survived	pclass	name	sex	age	sibsp	parch	fare	embarked	boat	body	Gender_Binary	embarked_Binary	boat_mod	body_mod
0	1	1	Allen, Miss. Elisabeth Walton	F	29.0000	0	0	211.3375	S	2	NaN	0.0	1.0	1.0	0.0
1	1	1	Allison, Master. Hudson Trevor	M	0.9167	1	2	151.5500	S	11	NaN	1.0	1.0	1.0	0.0
2	0	1	Allison, Miss. Helen Loraine	F	2.0000	1	2	151.5500	S	NaN	NaN	0.0	1.0	0.0	0.0
3	0	1	Allison, Mr. Hudson Joshua Creighton	M	30.0000	1	2	151.5500	S	NaN	135.0	1.0	1.0	0.0	1.0
4	0	1	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	F	25.0000	1	2	151.5500	S	NaN	NaN	0.0	1.0	0.0	0.0

3.6

```
x.embarked_Binary.mode()
```

```
0    1.0
dtype: float64
```

4.1

Logistic regression Confusion Matrix

```
confusion_matrix(y_test,y_pred)
```

```
array([[250,   3],
       [  3, 137]])
```

4.2

	feature	feature_importance
5	boat_mod	6.586166
3	Gender_Binary	2.452210
6	body_mod	1.033202
1	parch	0.278077
4	embarked_Binary	0.221869
0	sibsp	0.206931
7	age	0.008997
2	fare	0.006481

4.3 Naïve bayes

```
confusion_matrix(y_test,y_pred1)
```

```
array([[246,   7],
       [  3, 137]])
```

4.4

	Variable	importance
0	boat_mod	0.355725
1	body_mod	0.032061
2	fare	0.013232
3	Gender_Binary	0.003562
4	pclass	0.000000
5	sibsp	0.000000
6	parch	0.000000
7	embarked_Binary	0.000000

4.5

KNN Confusion Matrix

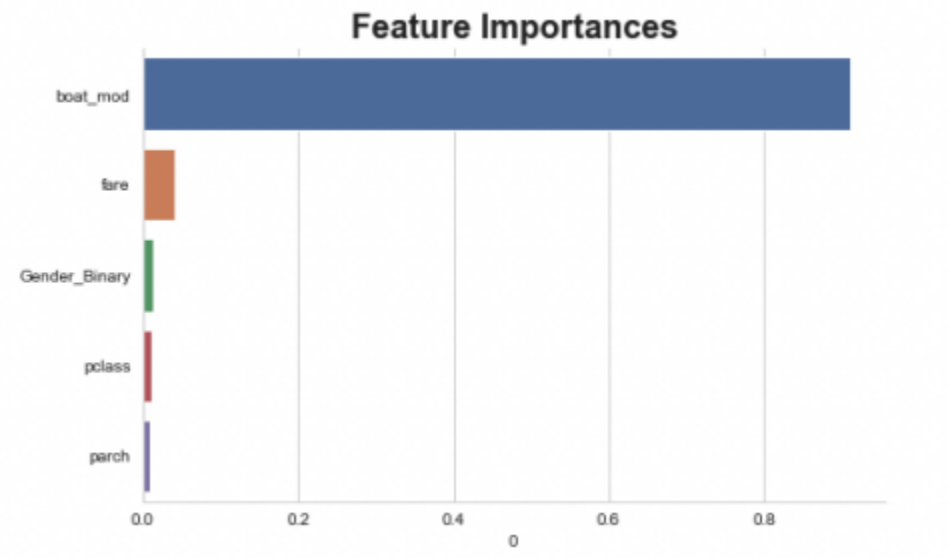
```
confusion_matrix(y_test,y_pred2)
```

```
array([[230,  23],
       [ 10, 130]])
```

4.7

	Variable	importance
0	boat_mod	0.263104
1	fare	0.066158
2	sibsp	0.039186
3	Gender_Binary	0.030534
4	pclass	0.016285
5	parch	0.012214
6	body_mod	-0.000509
7	embarked_Binary	-0.005089

4.6 Decision tree



4.8

Decision Tree Confusion Matrix

```
confusion_matrix(y_test,y_pred3)
```

```
array([[248,   5],
       [  7, 133]])
```