Logistic Regression to Predict Survivors of the Titanic Shipwreck

Swathi V
Dept. of Engineering Design
Indian Institute of Technology Madras
Chennai, India
ed18b034@smail.iitm.ac.in

***Abstract*—When the widely considered "unsinkable" Titanic sank, unfortunately only about 32% of the passengers onboard were able to survive. In this analysis, we try to analyse if any subset of passengers had a greater chance of survival using logistic regression. We try to identify factors that played a major role in the survival of passengers.**

***Index Terms*—Logistic regression, maximum likelihood estimation, confusion matrix, accuracy, precision, recall, F1 score, Titanic shipwreck**

I. Introduction

Logistic regression is a statistical tool used to model a binary classification problem, and to subsequently predict a qualitative response (such as Yes/No). Logistic regression models are used to make decisions by modelling the probability of the first class.

The mathematical function at the core of logistic regression is the **sigmoid function** $\sigma(x)$, which is an S-shaped curve that can map any real value to a number between 0 and 1. The decision boundary on x for the classification problem is then learned by setting a threshold on the value of $\sigma(x)$.

In this article, we use logistic regression to try and model the factors that influenced the survival of passengers aboard the Titanic. We hypothesize that factors such as age, sex, and socio-economic status influenced to some extent who had access to lifeboats, and were able to survive as a result.

We first explore the mathematics behind logistic regression, after which we apply the same to the Titanic problem. We start by exploring and analysing the data, and then building models and comparing them based on their performance on a subset of the data. Finally, we use the best model to make predictions on the remaining data.

II. Logistic Regression

In this section, we analyse how logistic regression is used to model a binary classification problem, the two classes being represented by 0 and 1.

The probability $P(Y = 1|X)$, denoted henceforth as $P(X)$, is modelled using a logistic function, given by

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

From 1, we can see that a low value of $\beta_0 + \beta_1 x$ gives a value of $P(X)$ that is close to 0, and a high value of $\beta_0 + \beta_1 x$ gives a value of $P(X)$ close to 1.

A. Estimating the Regression Coefficients

The optimal regression coefficients can be estimated by **maximum likelihood estimation**. The **likelihood function** is given by

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_i'=0} (1 - p(x_{i'})) \tag{2}$$

The coefficients β_0^* and β_1^* are chosen to maximize this likelihood function, and numerical methods like **gradient descent** can be used for the same.

B. Making Predictions

For every new data point X, a prediction can be made by estimating $P(X)$ as

$$P(X) = \frac{e^{\beta_0^*+\beta_1^* X}}{1 + e^{\beta_0^*+\beta_1^* X}} \tag{3}$$

If this value exceeds a particular threshold (0.5 for a symmetric loss matrix), the class can be predicted as 1 and vice-versa.

C. Multiple Logistic Regression

Multiple logistic regression is used when, instead of a single input feature X, we have a vector of input features $X_1, X_2, \cdots, X_p$. In this case, the model is taken as

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\beta_p X_p}} \tag{4}$$

To estimate the regression coefficients and the make predictions for a new data point, methods described above can be used.

D. Quantifying the Performance of the Model

The performance of a classification model can be quantified using a **confusion matrix** as shown in Fig. 1 and associated scores such as **accuracy**, **precision**, **recall**, and **F1 score**.

Accuracy: The accuracy of a model, which represents the total ratio of correct classifications, is given by

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{5}$$

Precision: The precision of a model, which represents the ratio of the correct classifications out of all positively predicted values, is given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

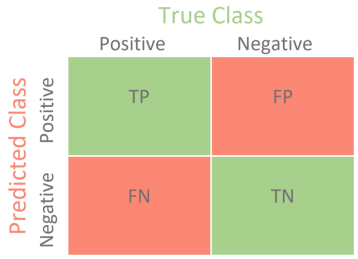

Fig. 1. Confusion matrix

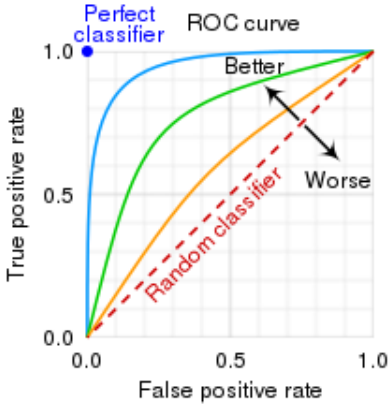

Fig. 2. Receiver operating characteristic curve

Recall: The recall of a model, also known as the sensitivity, is the ratio of correct predictions out of all the true positives. It is given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

F1 Score: The F1 score combines both precision and recall into a single metric, and is given by

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Area under ROC curve: The *receiver operating characteristics* curve plots the sensitivity and specificity for every every decision threshold between 0 and 1. The area under this curve (AUC) provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0, one whose predictions are 100% correct has an AUC of 1. The benefit of using AUC over other classification metrics is that AUC is invariant to the scale chosen as well as the classification threshold chosen. Fig. 2 shows the ROC curves of a random as well as an ideal classifier, and models between the two that can be evaluated using the AUC.

Using just one of the above metrics is not sufficient to quantify the performance of a model, as it depends on the domain of application. For example, in a medical image classification for diagnosis of a disease, false negatives can prove to be much costlier than false positives. In such a case, a model with a high recall would be preferred, even if it means a low precision.

III. Data

Two different datasets, train and test, contain a list of passengers and information like age, gender, socio-economic class, etc. The features present in each dataset are given in I.

TABLE I
VARIOUS FEATURES THAT ARE PRESENT IN THE DATASETS

Feature	Description
PassengerId	A unique key representing the passenger id
Survival	0 for No and 1 for Yes
PClass	Ticket class, where 1 represents 1st class, 2 represents 2nd class, and 3 represents 3rd class
Name	Passenger name
Sex	
Age	
SibSp	Number of siblings / spouses onboard the Titanic
Parch	Number of parents/ children onboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of embarkation, where S = Southampton, Q = Queenstown, and C = Cherbourg

A. Preliminary Data Analysis

The given training dataset contains a list of **889** passengers. The distribution of these passengers based on the target variable (survival) is given in Fig. 3. Since the dataset is not extremely skewed, we do not perform data augmentation.

```
Total number of observations: 889
Distribution of observations based on the target variable:
0    549
1    340
Name: Survived, dtype: int64
```

Fig. 3. Distribution of training data

IV. The Problem

The objective of this article is to understand if there exists any relationship between passenger details like age, sex, and socio-economic status and the survival rates using the train dataset. The subsequent step would be to try and predict survival of the passengers in the test data set. The steps followed to explore the same are:

- Data cleaning and imputation
- Data visualization and exploratory analysis
- Building a model to fit the data
- Quantifying the performance of the model
- Using the model to predict survival

A. Data Cleaning and Imputation

Features like *Name*, *PassengerId*, and *Ticket*, which are not related to passenger survival, were removed. Rows where *Embarked* was Null were also removed, as these were very few.

Categorical variables *Embarked* and *Sex* are converted to one-hot encodings using the *get_dummies* method in pandas.

For rows where *Age* is missing, data imputation needs to be done, and the distribution of age with *PClass* and *Sex* is considered for the same, which is shown in Fig.4. Since the distribution of age varies mainly with *Pclass* and not *Sex*, imputation is done based only on the former.

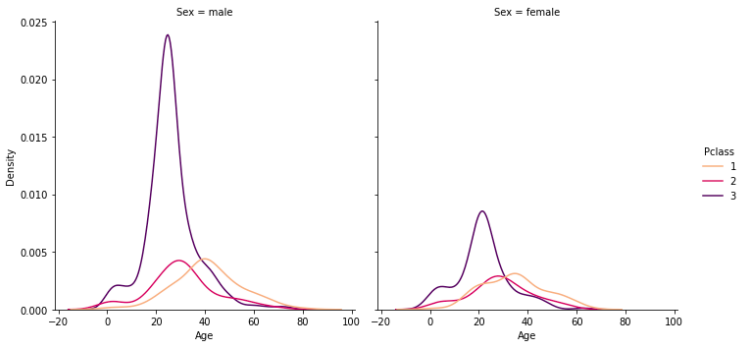

Fig. 4. Distribution of age with class and sex

Additionally, continuous variables *Fare* and *Age* are also binned and copies are created. This can be used to create an alternate model. The bin edges were created by visualizing the spread of the data, shown in Fig. 5 and Fig. 6. The distribution of data in each bin is shown in Fig. 7 and Fig. 8.

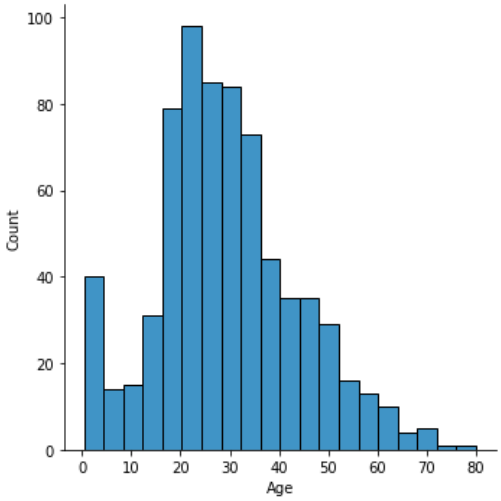

Fig. 5. Spread of the age variable

Age and *Fare* values are now scaled to a range 0-1 to enable faster convergence during training. The data is now ready to be visualized and analysed.

B. Data Visualization and Exploratory Analysis

The class ratio of the training dataset is (340:549), which is not very skewed. The effect of various independent variables on survival are analysed in Fig.9 to Fig. 15.

Notably, a much larger proportion of females have survived the shipwreck (Fig. 9). Since the cabin feature is very sparsely

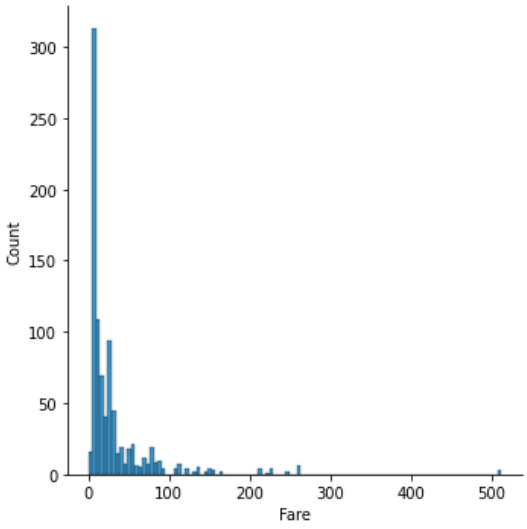

Fig. 6. Spread of the fare variable

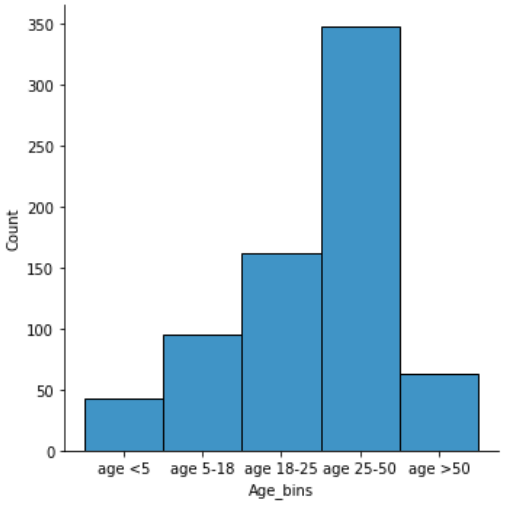

Fig. 7. Distribution of training data in each age bin

available (more than 650 rows have missing cabin details), and the survival rates do not seem to follow a clear pattern with respect to the variable *Cabin* (Fig, 12), this column is dropped. All other factors are considered while building the model.

C. Building Models to Fit the Data

1) Model 1: Using Continuous Variables: The variables 'Pclass', 'male', 'Age', 'SibSp', 'Parch', 'Fare', 'C', and 'Q' are used as the input variables.

2) Model 1: Using Binned Variables: The variables 'Pclass', 'male', 'age <5', 'age 5-18', 'age 18-25', 'age 25-50', 'fare <10', 'fare 10-20', 'fare 20-30', 'fare 30-100', 'SibSp', 'Parch', 'C', 'Q' are used as the input variables.

In both the models, variables 'female' and 'S' are removed because they are correlated to the variables 'male' and ('S', 'Q') respectively. In addition to this, in model 2, the varibles

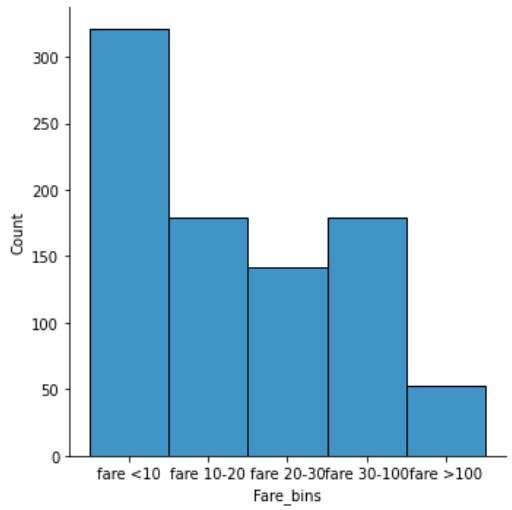

Fig. 8. Distribution of training data in each fare bin

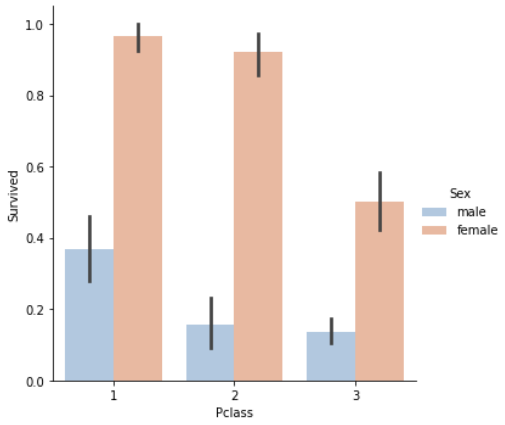

Fig. 9. Variation of survival rates with *Pclass* and *Sex*

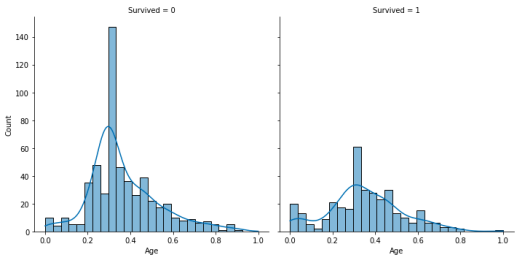

Fig. 10. Variation of survival rates with *Age*

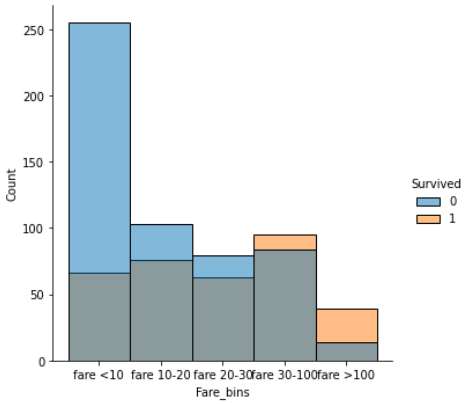

Fig. 11. Variation of survival rates with *Fare*

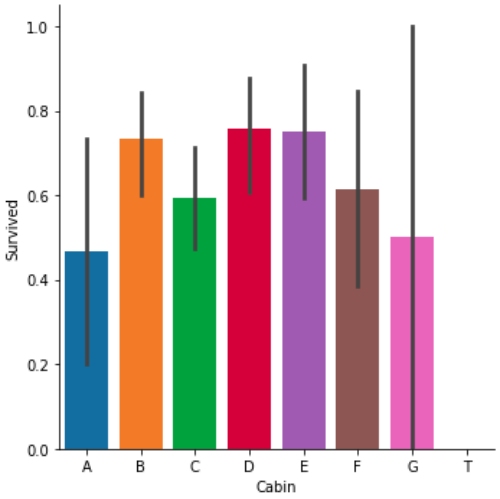

Fig. 12. Variation of survival rates with *Cabin*

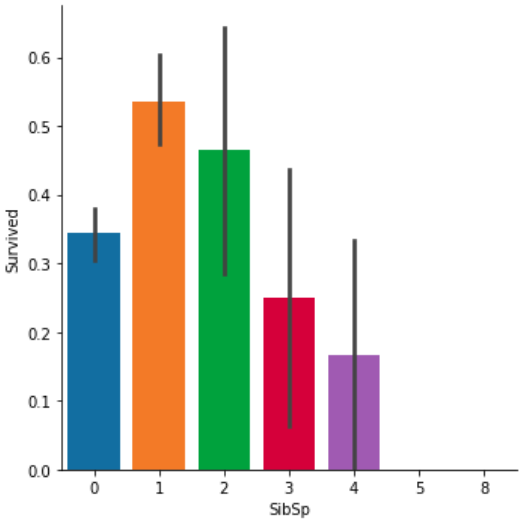

Fig. 13. Variation of survival rates with *SibSp*

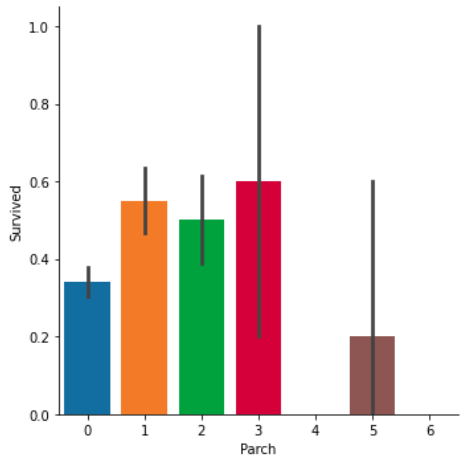

Fig. 14. Variation of survival rates with *Parch*

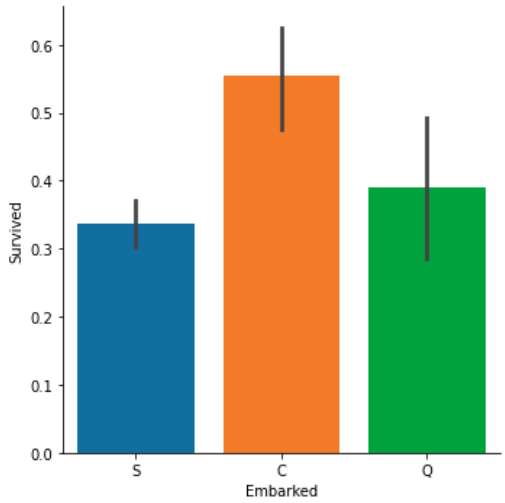

Fig. 15. Variation of survival rates with *Embarked*

‘age >50’ and ‘fare >100’ are removed for the same reason. A validation dataset of 30% is used to ensure that the models do not overfit.

D. Comparing Model Performances

The performance metrics for both models are given in Fig. 16 and Fig. 17, for both training and validation datasets. In this case, since the loss matrix is symmetric, i.e., both false positives and false negatives are equally costly, we look for a high F1 score and accuracy. The ROC curves for both models are shown in Fig. 18 and Fig. 19 respectively.

We see that both models perform equivalently. Further, both models do not overfit significantly. For further analysis, we go ahead with model 1 as it has a lower number of input variables.

The coefficients of each feature is shown in Fig. 20.

The observations from the coefficients are:

- As expected from Fig. 9, the highest (negative) weight is given to the feature ‘male’, which is an alias for the feature ‘Sex’.

```
Model 1


Training data:
Confusion matrix:
[[331  67]
 [ 49 175]]
Accuracy:  0.8135048231511254
Precision:  0.7231404958677686
Recall:  0.78125
F1_score:  0.7510729613733906


Validation data:
Confusion matrix:
[[149  32]
 [ 20  66]]
Accuracy:  0.8052434456928839
Precision:  0.673469387755102
Recall:  0.7674418604651163
F1_score:  0.7173913043478259
```

Fig. 16. Performance of Model 1

```
Model 2


Training data:
Confusion matrix:
[[331  66]
 [ 49 176]]
Accuracy:  0.815112540192926
Precision:  0.7272727272727273
Recall:  0.7822222222222223
F1_score:  0.7537473233404711


Validation data:
Confusion matrix:
[[148  32]
 [ 21  66]]
Accuracy:  0.8014981273408239
Precision:  0.673469387755102
Recall:  0.7586206896551724
F1_score:  0.7135135135135136
```

Fig. 17. Performance of Model 2

- The weight given to the feature ‘Age’ is very low, even though age seems to be an important factor in deciding survival from Fig. 10. This is because a logistic regression model cannot fit this distribution accurately.
- The weights given to ‘Fare’, ‘SibSp’, ‘Parch’, and ‘Embarked’ are also pretty low. The evidence for this can be seen in Fig. 11 to Fig. 15.

Inspired by the above observations, we build a new model which does not use the features ‘Fare’, ‘SibSp’, ‘Parch’, and ‘Embarked’. The performance of this model on the train and test dataset is shown in Fig. 21. However, it is outperformed by the above mentioned models.

E. Using the Model to Predict

Model 1 was used to predict survival of passengers in the test dataset. As done in the training dataset, imputation was done for Age, and additionally for Fare as well. The test dataset contained 418 passengers, and the outcomes were as follows:

- Survived - 157

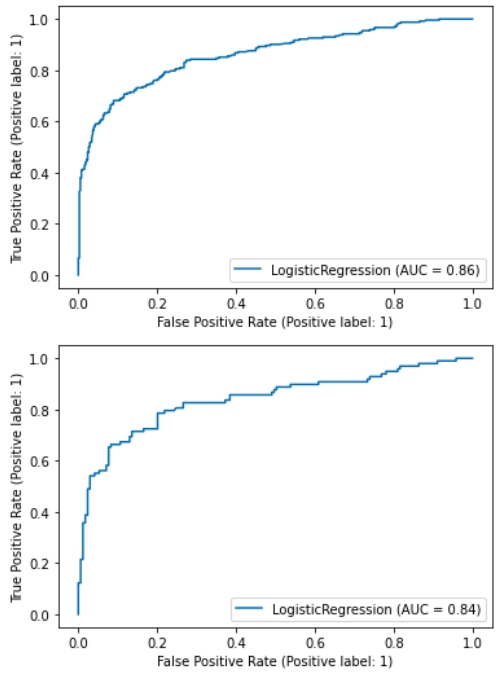

Fig. 18. ROC curve of Model 1

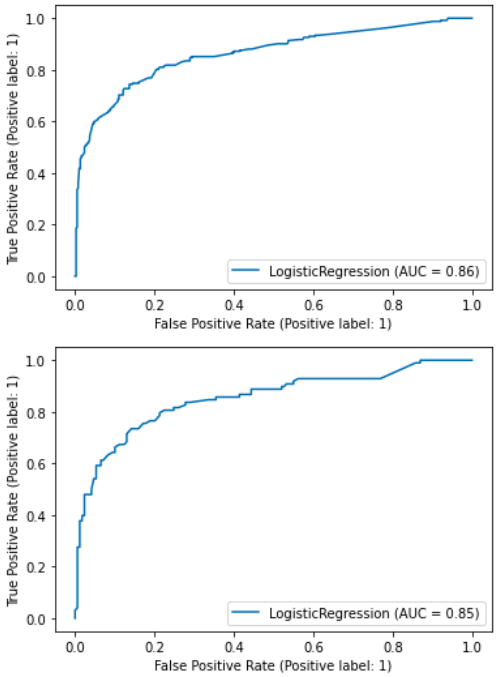

Fig. 19. ROC curve of Model 2

```
Model1 :

Pclass  : -1.1774659585883178
male  : -2.6261432136243363
Age  : -0.04374144174187662
SibSp  : -0.4038383263203895
Parch  : -0.02327934699125502
Fare  : 2.4619860250596126e-05
C  : 0.570156134878336
Q  : 0.2468132587887023
```

Fig. 20. Coefficients of each feature for Model 1

```
Model 3


Training data:
Confusion matrix:
[[324  69]
 [ 56 173]]
Accuracy:  0.7990353697749196
Precision:  0.7148760330578512
Recall:  0.7554585152838428
F1_score:  0.7346072186836518


Validation data:
Confusion matrix:
[[149  34]
 [ 20  64]]
Accuracy:  0.797752808988764
Precision:  0.6530612244897959
Recall:  0.7619047619047619
F1_score:  0.7032967032967032
```

Fig. 21. Results of using feature selection

- Did not survive - 261

These results have been recorded in the file 'predictions.csv'.

V. Conclusions

With a logistic regression model, the survivors of the titanic shipwreck can be predicted with an accuracy of about 80%. Thus, more than just an element of luck, there is an obvious influence of age, sex, and other socio-economic factors on whether a person survived the shipwreck or not.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, "Classification" in *An Introduction to Statistical Learning,* New York, Springer, 2013.