Logistic Regression Report on Titanic Survival Prediction

Name: Priyank, Roll no. 2521CS07

Abstract

This report presents a logistic regression analysis on the Titanic dataset. The goal is to predict passenger survival (**Survived** variable) using demographic and travel features. We preprocess the data, train a logistic regression model, and evaluate performance using accuracy, classification metrics, and a confusion matrix. Only the counts-based confusion matrix is presented for clarity.

1 Theory of Logistic Regression

Logistic Regression is a widely used statistical method for binary classification tasks, where the goal is to predict the probability of an outcome belonging to one of two classes (e.g., survived or not survived). Unlike linear regression, which produces continuous values, logistic regression applies the **sigmoid function** to a linear combination of input features, mapping predictions into the range $(0, 1)$. The model parameters are estimated using **maximum likelihood estimation**, ensuring the predicted probabilities best fit the observed data. A threshold, typically 0.5, is then applied to classify observations into their respective categories. Logistic regression is computationally efficient, interpretable, and provides a probabilistic framework for decision-making, making it highly suitable for problems such as predicting survival on the Titanic dataset.

2 Dataset and Features

We used the Titanic dataset containing 891 passenger records with the following features:

- Passenger class (`Pclass`)
- Sex (`Sex`)
- Age (`Age`)
- Number of siblings/spouses aboard (`SibSp`)
- Number of parents/children aboard (`Parch`)
- Ticket fare (`Fare`)
- Port of embarkation (`Embarked`)

The target variable is **Survived** (0 = Not Survived, 1 = Survived).

3 Preprocessing

- Missing numeric values (*Age, SibSp, Parch, Fare*) were filled using the median.
- Missing categorical values (*Sex, Embarked*) were filled using the mode.

- One-hot encoding was applied to categorical features.
- Features were standardized using z-score normalization.
- Dataset split: 80% training, 20% testing.

4 Model

A Logistic Regression classifier (`lbfgs` solver, max_iter = 1000) was trained.

5 Results

The model achieved an accuracy of about **80%** on the test set.

5.1 Confusion Matrix

The confusion matrix (counts) is shown below:

$$\text{CM} = \begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix} = \begin{bmatrix} 98 & 12 \\ 23 & 46 \end{bmatrix}$$

Where:

- TN (True Negatives): 98 (correctly predicted Not Survived)
- FP (False Positives): 12 (incorrectly predicted Survived)
- FN (False Negatives): 23 (missed Survived cases)
- TP (True Positives): 46 (correctly predicted Survived)

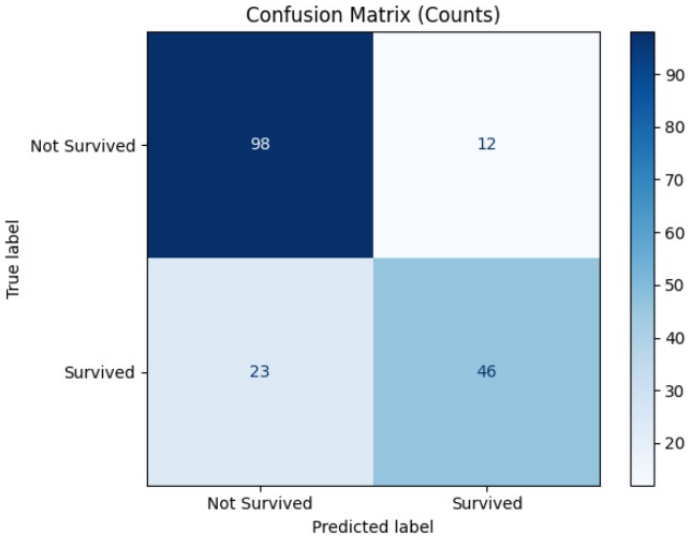

Figure 1: Confusion Matrix of survival case

6 Conclusion

The logistic regression model performs well with an overall accuracy of $\approx 80\%$. However, recall for the *Survived* class is lower than for *Not Survived*, indicating the model is more conservative in predicting survival. Future improvements could involve feature engineering (e.g., family size, passenger titles) or handling class imbalance.