Titanic Survival Prediction using Logistic Regression (scikit-learn)

Vaibhav Chourasia

October 2, 2025

1 Introduction

This report demonstrates the application of logistic regression using the scikit-learn library to predict survival on the Titanic dataset. Logistic regression is a widely used statistical method for binary classification, modeling the probability of an event (here, survival) based on input features.

2 Libraries and Tools Used

The following Python libraries were used:

- **NumPy**: For numerical operations and array manipulation.
- **Pandas**: For loading and preprocessing tabular data.
- **Matplotlib**: For visualizing data and model predictions.
- **scikit-learn**: For logistic regression modeling and feature scaling.

3 Dataset and Preprocessing

The Titanic dataset was obtained from `https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv`. Only the following features were selected:
`Pclass, Sex, Age, SibSp, Parch, Fare`, with `Survived` as the target variable.

3.1 Preprocessing Steps

1. Missing values in `Age` were filled with the median value.
2. The `Sex` column was mapped to numeric values: male=1, female=0.
3. Features were standardized using scikit-learn's `StandardScaler`:

   $$x' = \frac{x - \mu}{\sigma}$$

   where μ and σ are the mean and standard deviation of each feature.
4. Data was split into training (80%) and testing (20%) sets using a random permutation.

4 Logistic Regression Model (scikit-learn)

Note: In scikit-learn, the cross-entropy loss (log-loss) and the gradients are computed internally by the library. The user does not need to implement gradient descent manually; the solver optimizes the loss function automatically to learn the weights.

The logistic regression model predicts the probability of survival as:

$$p(y = 1|x) = \frac{1}{1 + e^{-z}}, \quad z = w^\top x + b$$

where w is the vector of learned coefficients, b is the intercept, and x is the input feature vector. scikit-learn internally optimizes the log-likelihood function to find the best w and b.

5 Model Training and Results

The model was trained using scikit-learn's `LogisticRegression` with `max_iter=3000` to ensure convergence.

The learned weights (intercept + coefficients) obtained from the code are:

$$[w_0, w_1, w_2, w_3, w_4, w_5, w_6] = [-0.724, -0.813, -1.233, -0.474, -0.397, -0.069, 0.181]$$

Here, w_0 is the intercept and the remaining coefficients correspond to the features:

$$w = [w_{\text{Pclass}}, w_{\text{Sex}}, w_{\text{Age}}, w_{\text{SibSp}}, w_{\text{Parch}}, w_{\text{Fare}}]$$

5.1 Final Logistic Regression Equation

The logistic regression model can be expressed as:

$$p(x) = \frac{1}{1 + \exp\Big(-\big(w_0 + w_1 \cdot \text{Pclass} + w_2 \cdot \text{Sex} + w_3 \cdot \text{Age} + w_4 \cdot \text{SibSp} + w_5 \cdot \text{Parch} + w_6 \cdot \text{Fare}\big)\Big)}$$

Substituting the learned weights, the equation from the code output is:

```
p(x) = 1 / (1 + exp(-(-0.724 - 0.813*Pclass - 1.233*Sex - 0.474*Age - 0.397*SibSp - 0
```

5.2 Test Accuracy

The model achieved a test accuracy of approximately:

$$\text{Accuracy} = 0.78 \quad \text{(may vary depending on random seed)}$$

This demonstrates that scikit-learn's logistic regression effectively captures the relationships between input features and survival probability on the Titanic dataset.

6 Results and Graphs

6.1 Logistic Curve: Age vs Survival

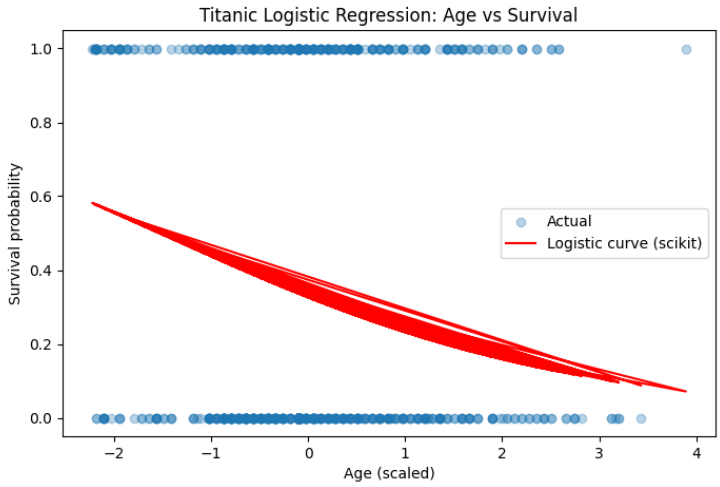

Figure 1: Logistic regression curve for the Age feature with actual survival data points.

6.2 Predicted vs Actual Survival

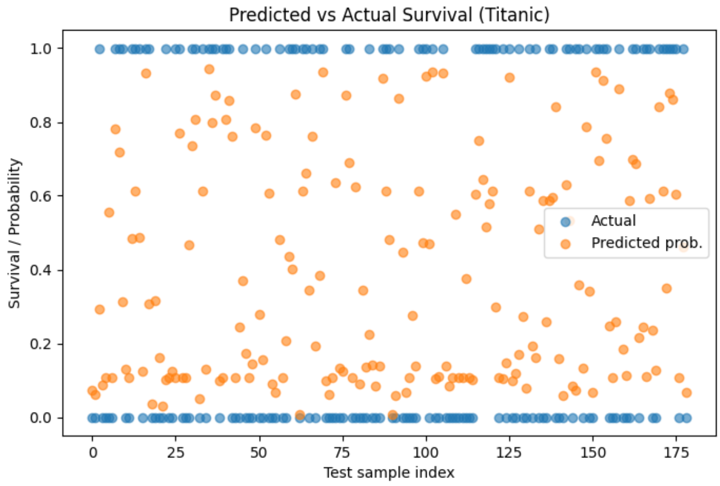

Figure 2: Scatter plot comparing predicted survival probabilities and actual survival labels on the test set.

7 Conclusion

This report demonstrates how scikit-learn's logistic regression can be applied to predict survival on the Titanic dataset. The model successfully captures the relationship between features such as Age, Sex, and Pclass with survival probability. Visualizations show how the logistic curve fits the data, and the predicted probabilities align reasonably with actual outcomes.