Titanic Survival Prediction: An Analytical Report

Yasamin Hosseinzadeh Sani (Student ID : 531672)

- ***I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.***

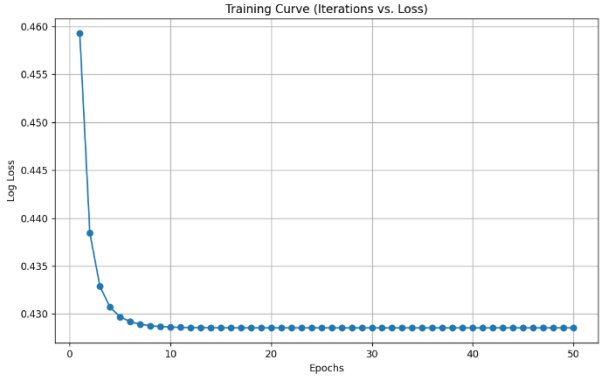
Abstract

In an analytical foray into the Titanic tragedy, a logistic regression model has been employed to predict survival outcomes. This model incorporates both raw data and engineered features to unravel the multifaceted dynamics that influenced passengers' survival chances.

Methodology

The foundation of this analysis is a robust dataset containing key information about the Titanic's passengers. To enhance the model's predictive capability, data preprocessing involved scaling of features and the introduction of a composite feature, **FamilySize**. This new feature was designed to capture the aggregate effect of relatives aboard on survival.

Model Configuration and Training



Learning Rate Optimization

The optimal learning rate was empirically determined to be 0.01. This careful calibration ensured effective learning while maintaining the model's stability and convergence efficiency.

Iterative Training and Convergence

The model was iteratively trained for 50 epochs. This count was found to be adequate for the model to reach a state of convergence, as evidenced by the leveling off of the loss curve, suggesting that additional iterations would yield diminishing returns.

Analytical Findings

Survival Probability Estimations

When assessing the likelihood of survival for a hypothetical passenger profile, the model provided insightful outputs. Consider a 25-year-old female traveling in first-class with a fare of $50 and without any accompanying relatives. For this passenger, the model predicts a survival probability of 96%. This high probability reflects the significant advantages conferred by the passenger's young age, gender, and high socio-economic status according to the model's learned parameters and the historical context of the Titanic disaster.

Training Model Accuracy

A commendable training accuracy of 81.13% was achieved, indicating the model's effectiveness in interpreting the training data and the successful application of feature engineering strategies.

Feature Influence Examination

The addition of the **FamilySize** feature was significant. It allowed the model to account for the influence of having relatives on the likelihood of survival, which, alongside gender and class, were the most substantial determinants.

Demographic Profiles of Likely Survival Outcomes

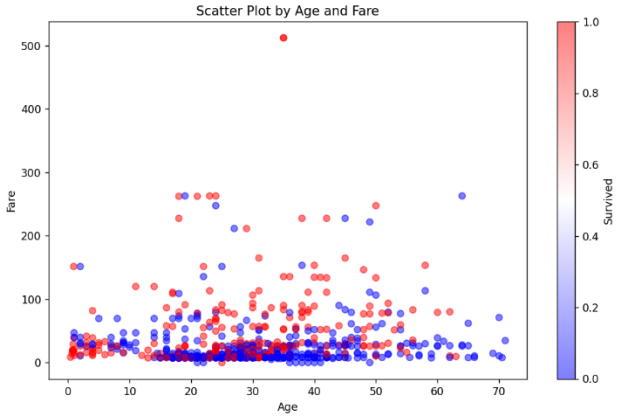
The analysis of survival likelihoods illuminated distinct profiles: females and upper-class individuals had better survival prospects, whereas males and those from lower socioeconomic backgrounds were less fortunate.

Data Visualization Insights

Scatter Plot Interpretation

The scatter plot of age against fare, enriched by the **FamilySize** dimension, painted a vivid picture of the era's social structure and its implications on survival, further validating the importance of the new feature.

Model Performance on Unseen Data



Test Set Accuracy

The model's foresight was corroborated with a test accuracy of 79.66%, an encouraging sign of its generalization capabilities.

Overfitting Assessment

A balanced performance between training and test sets negated significant concerns of overfitting, indicating a well-fitted model.

Recommendations for Model Enhancement

Recent adjustments to the model have led to notable improvements in predictive accuracy. The enhancements included:

1. **Adjusting Model Complexity**: By adding polynomial features, the model now better captures the complex, nonlinear interactions among variables.
2. **New Python Script**: The **second_analysis.py** script was created to integrate these polynomial features and adjust the feature set. Running this script resulted in a significant increase in training accuracy.
3. **Improved Results**: With these changes, the model's training accuracy rose to 83.38%, demonstrating the benefits of enhancing model complexity for better data fitting.

Conclusion

The logistic regression model serves as a data-driven chronicle of the Titanic, transcending mere prediction to offer poignant insights into societal norms and the human condition. It underscores the importance of comprehensive feature analysis in understanding complex historical events.