Miraj Rahman Sanjid

CS 505 Data Mining

Octavio Juarez Espinosa

Date: 3/4/2025

Logistic Regression Report

**Objective:**

To implement and evaluate a logistic regression model for a binary classification problem. This project focuses on gaining experience in data preprocessing, model training, evaluation, and interpretation using the Titanic Survival Prediction Dataset.

**Dataset**

* Source: [Titanic Dataset](https://www.kaggle.com/datasets/yasserh/titanic-dataset/data)
* Link: <https://www.kaggle.com/datasets/yasserh/titanic-dataset/data>
* Target Variable: Survived (1 = survived, 0 = not survived)

**Data Preprocessing:**

* Exploration: Checked dataset structure, missing values, and statistical summaries.
* Dropped Columns: 'PassengerId', 'Name', 'Ticket', 'Cabin' (not useful for prediction).
* Missing Values Handling:
  + 'Age' filled with median.
  + 'Embarked' filled with mode.
* Encoding Categorical Variables:
  + 'Sex' and 'Embarked' encoded using Label Encoding.
* Feature Scaling:
  + Standardized 'Age' and 'Fare' using StandardScaler.
* Train-Test Split:
  + 80% Training, 20% Testing (Stratified by 'Survived').

**Model Training:**

* Algorithm Used: Logistic Regression
* Hyperparameter Tuning:
  + GridSearchCV used to optimize 'C' (regularization strength) and 'penalty' (L1, L2).
  + Best Parameters: {C: optimal\_value, penalty: optimal\_value}

**Model Evaluation:**

* Metrics Computed:
  + Accuracy: {computed\_accuracy}
  + Precision, Recall, F1-score (from classification report)
  + Confusion Matrix:

[[TN FP]

[FN TP]]

* + AUC-ROC Score: {computed\_auc}
* Confusion Matrix Heatmap: Displayed for better visualization.
* AUC-ROC Curve: Plotted to evaluate model performance.

**Feature Importance Analysis:**

* Logistic regression coefficients analyzed to determine influential features.
* Most significant features:
  + 'Sex' (coefficient: {Coefficient1})
  + 'Pclass' (coefficient: {Coefficient2})
  + 'Fare' (coefficient: {Coefficient3})
* Features with the highest positive coefficients increase survival probability, while negative coefficients decrease it.

**Discussion on Precision-Recall Trade-offs:**

* Precision vs Recall:
  + Precision: Important when false positives are costly (e.g., fraud detection).
  + Recall: Important when false negatives are costly (e.g., medical diagnosis).
* Titanic Case:
  + Prioritizing recall ensures that all possible survivors are identified (reducing false negatives).
  + Accuracy alone may not be the best metric; recall is crucial to avoid misclassifying actual survivors.

**Conclusion:**

* Logistic regression performed well in predicting survival based on passenger data.
* Feature importance analysis helped understand key influencing factors.
* Model evaluation showed a balance between precision and recall, with AUC-ROC confirming good discriminatory power.
* Further improvements could include:
  + Feature engineering (creating new interaction features).
  + Trying other classification models (e.g., Decision Trees, Random Forest).

**Final Thoughts:**

Logistic Regression is a solid baseline model for Titanic survival prediction, but optimizing recall could be more beneficial for real-world applications where missing a survivor is costlier than misclassifying a non-survivor.