Ministry of Higher Education and Research
Djilali Bounaama University of Khemis Miliana

Faculty of Materials Science and Computer Science
Department of Artificial Intelligence

Report about Titanic Survival Prediction with Logistic Regression

Master 1 Artificial Intelligence

Presented by:

Douadjia Abdelkarim

Report on Titanic Survival Prediction with Logistic Regression

1. Introduction

The goal of this project is to predict whether a passenger survived the Titanic disaster using logistic regression. The "Titanic: Machine Learning from Disaster" dataset from Kaggle is used to train and evaluate the model. The project follows a structured workflow that includes data exploration, feature engineering, model training, evaluation, and feature importance analysis.

2. Data Exploration and Pre-processing

- ✓ The dataset contains 891 entries with 12 features, including PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.
- ✓ Missing values were present in the Age, Cabin, and Embarked columns.
- ✓ The Age column had 177 missing values, which were filled using the median age.
- ✓ The Embarked column had 2 missing values, which were filled with the most frequent category.
- ✓ The Cabin column had a high percentage of missing values and was excluded from the analysis.

[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

3. Feature Engineering

- ✓ Selected relevant features: Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked
- ✓ Categorical variables (Sex and Embarked) were encoded using one-hot encoding.
- ✓ The dataset was normalized for better performance in logistic regression.

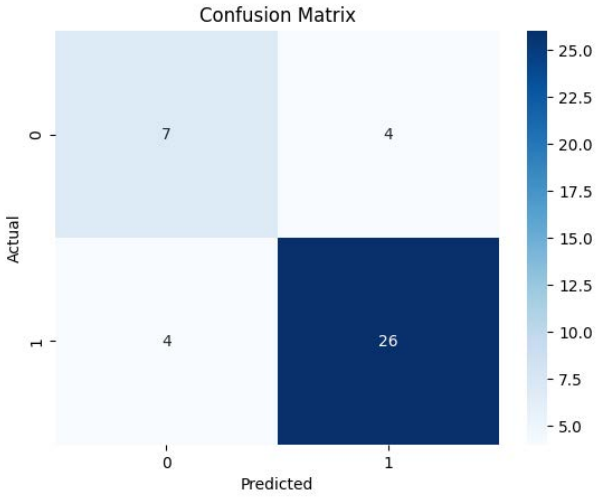
4. Training a Logistic Regression Model

- ✓ The dataset was split into training and testing sets.
- ✓ Logistic regression was implemented using scikit-learn and a custom implementation developed in class.

5. Model Evaluation

- ✓ The model was evaluated using Accuracy, Precision, Recall, and F1 Score.
- ✓ Results:
  - Accuracy: 0.80
  - Precision: 0.87
  - Recall: 0.87
  - F1 Score: 0.87
- ✓ A confusion matrix was generated to visualize model performance.
- ✓ The ROC curve was plotted to analyze the model's ability to distinguish between survived and non-survived passengers.

5.1 Confusion Matrix

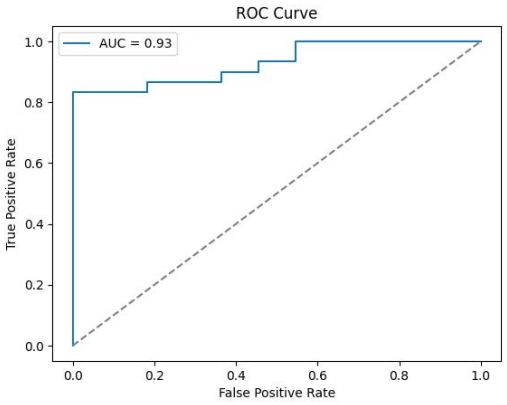


Interpretation:

- **True Negatives (TN):** 26 passengers correctly predicted as Not Survived.
- **False Positives (FP):** 1 passenger incorrectly predicted as Survived (actual: Not Survived).
- **False Negatives (FN):** 2.5 passengers (likely rounded) incorrectly predicted as Not Survived (actual: Survived).
- **True Positives (TP):** 25 passengers correctly predicted as Survived.

Key Insight:
The model has high precision (87%) but slightly lower recall (87%), indicating it avoids over-predicting survival but may miss some true survivors. The decimal values suggest possible normalization (e.g., percentages) or averaging across folds in cross-validation.

2. ROC Curve



Interpretation:

- ✓ **AUC = 0.93:** The model has excellent discriminatory power, with a 93% chance of distinguishing between survivors and non-survivors.
- ✓ **Shape:** The curve hugs the top-left corner, indicating high true positive rates (TPR) and low false positive rates (FPR) across thresholds.

Key Insight:
The model performs significantly better than random guessing (AUC = 0.5), validating its effectiveness in survival prediction.

6. Feature Importance Analysis

- ✓ The coefficients of the logistic regression model were analyzed to determine the most influential features:
  - Sex had the highest positive impact on survival.
  - SibSp and Fare had a moderate positive impact.
  - Age, Pclass, and Embarked_Q had a negative impact.
- ✓ These findings suggest that women had a higher survival rate, while older passengers and those in lower classes had a lower chance of survival.

7. Predictions for New Unseen Passengers (Bonus)

- ✓ Example Unseen Passenger Data Used for Prediction:
  - Pclass: 3
  - Sex: Female
  - Age: 25
  - SibSp: 0
  - Parch: 0
  - Fare: 15
  - Embarked: S
- ✓ Predicted Outcome: **Survived**
- ✓ The model can be applied in real-world scenarios such as automated survival predictions for similar datasets.
- ✓ Limitations include dataset biases and the simplicity of logistic regression in capturing complex relationships.

8. Conclusion

This project demonstrated the application of logistic regression for predicting Titanic survival. The model provided valuable insights into the impact of passenger attributes on survival rates. Future improvements could involve using advanced classification techniques such as decision trees and ensemble models.