

Breast Cancer Classification Using Logistic Regression

Avril Luo

Problem & Data

Problem

1. Predict whether a breast tumor is malignant or benign
2. Binary classification task

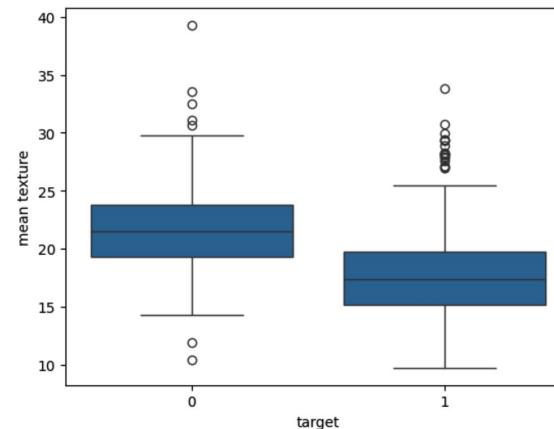
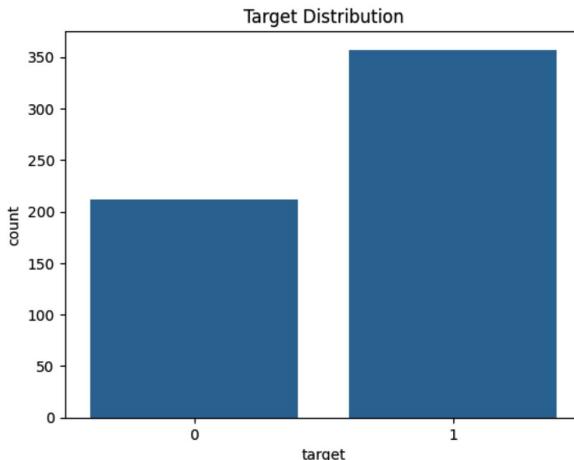
Data

1. Source: `sklearn.datasets.load_breast_cancer`
2. Samples: 569
3. Features: 30 numerical features
4. Target: Malignant vs Benign
5. All features are numerical and extracted from medical images

Exploratory Data Analysis (EDA)

Key EDA Findings

1. Target variable is relatively balanced
2. Several features (e.g. mean radius and mean texture) show clear separation between classes, suggesting strong predictive signals.
3. No missing values in the dataset



Model & Method

Model:

1. Logistic Regression (baseline)
2. Random Forest
3. Gradient Boosting

Method

1. Stratified 5-fold Cross-Validation
2. Final evaluation on held-out test set
3. Metrics: Accuracy, F1 score, Confusion Matrix

Why Logistic Regression?

1. Strong cross-validation performance
2. Simple and interpretable
3. Well-suited for structured numerical data

Results

Model Comparison (5-Fold Cross-Validation)

1. Logistic Regression (scaled):

- a. Accuracy ≈ 97.4%
- b. F1 ≈ 97.9%

2. Random Forest and Gradient Boosting performed slightly worse

	Model	CV Accuracy (mean)	CV F1 (mean)
0	Logistic Regression (scaled)	0.973669	0.979434
1	Random Forest	0.952569	0.962421
2	Gradient Boosting	0.949076	0.960242

Final Model Performance

- 1. Test Accuracy ≈ 95.6%
- 2. Low number of false negatives

Conclusion & Next Steps

Conclusion

1. On this dataset, a simple linear model outperformed more complex ensemble methods.
2. Simple models can achieve strong performance on structured data

Next Steps

1. Feature selection
2. Hyperparameter tuning
3. Try ensemble models (e.g. Random Forest)