Breast Cancer Analysis

Predicting Survival Time and Classifying Cancer Aggressiveness

Presented by : Alvin Asingo

PROJECT OVERVIEW

Objective

To predict survival time and classify breast cancer aggressiveness.

Business Problem

- Breast cancer is a leading cause of death among women. Early detection and personalized treatment are crucial.
- The goal is to help a hospital classify patients' cancer status and predict survival time to improve treatment strategies.

Data Understanding

Dataset Source:

- SEER Program, NCI via Kaggle (2006-2010, 4024 female patients).

Key features:

- Numerical:Age, Tumor Size, Regional Nodes Examined/Positive, Survival Months
- Categorical: Race, Marital Status, Stages, Grade, Hormone Receptor Status, etc

Predictive Modeling:- Regression

Objective:

- Estimate survival time using numerical factors.

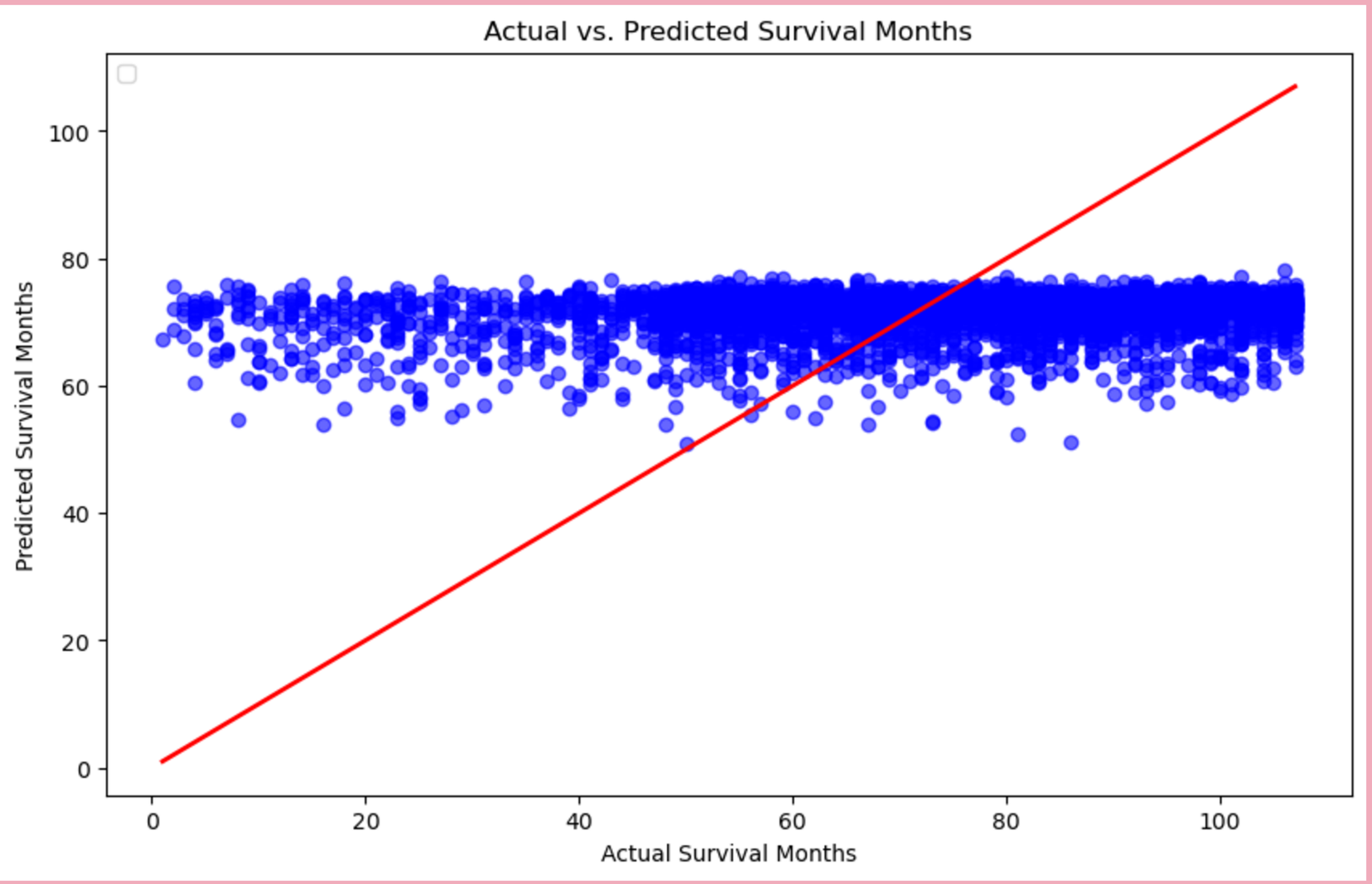
Selected Features:

- Age, Tumor Size, Regional Node Examined, Regional Node Positive, Survival Months.

Key Results:

- Low R-squared values; some significant predictors.
- Possible model refinements: explore non-linear relationships or add more variables.

View the graph (next slide)for an idea of the regression

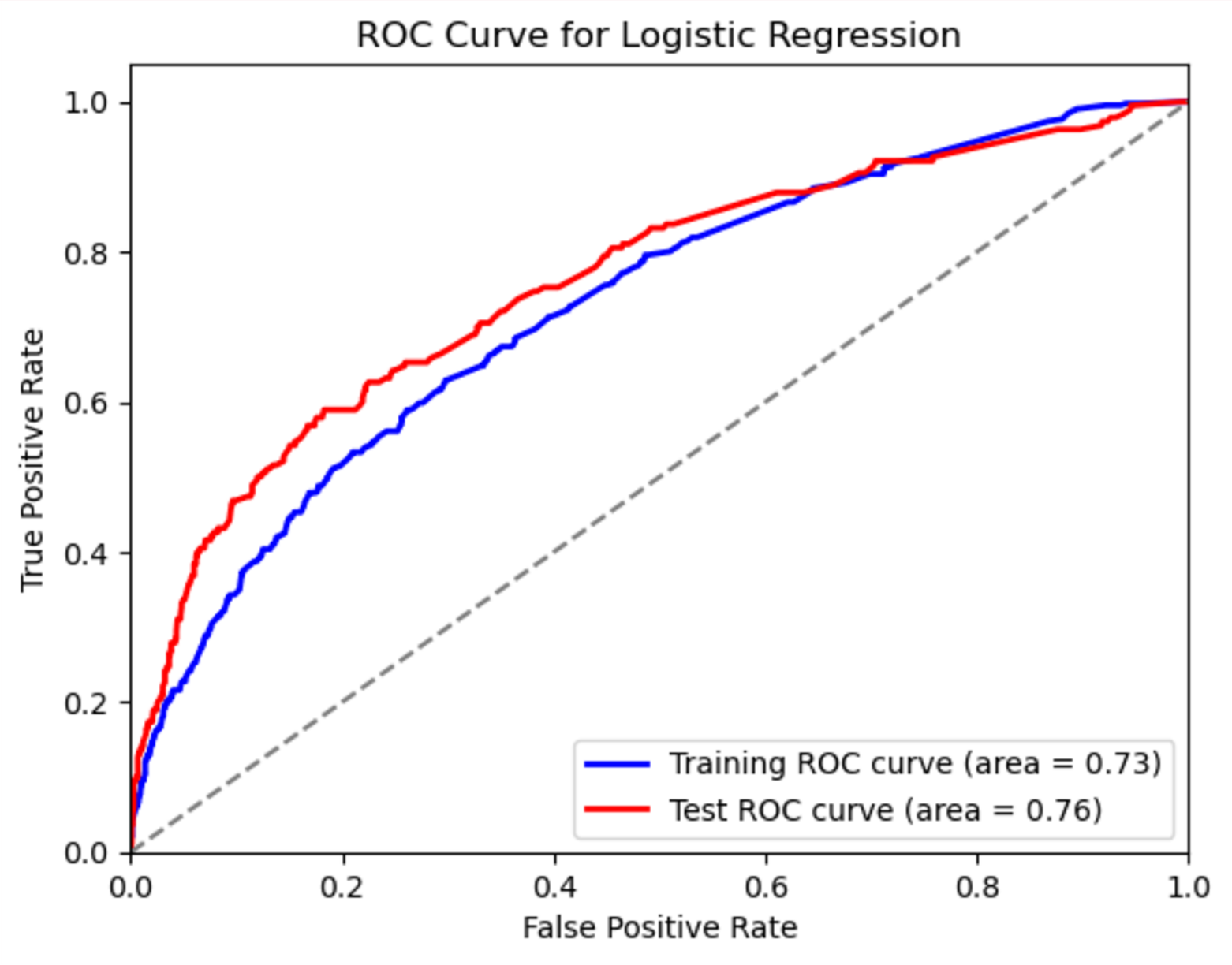


The graph shows that while the model predicts survival months, the predictions deviate quite a bit from the actual values

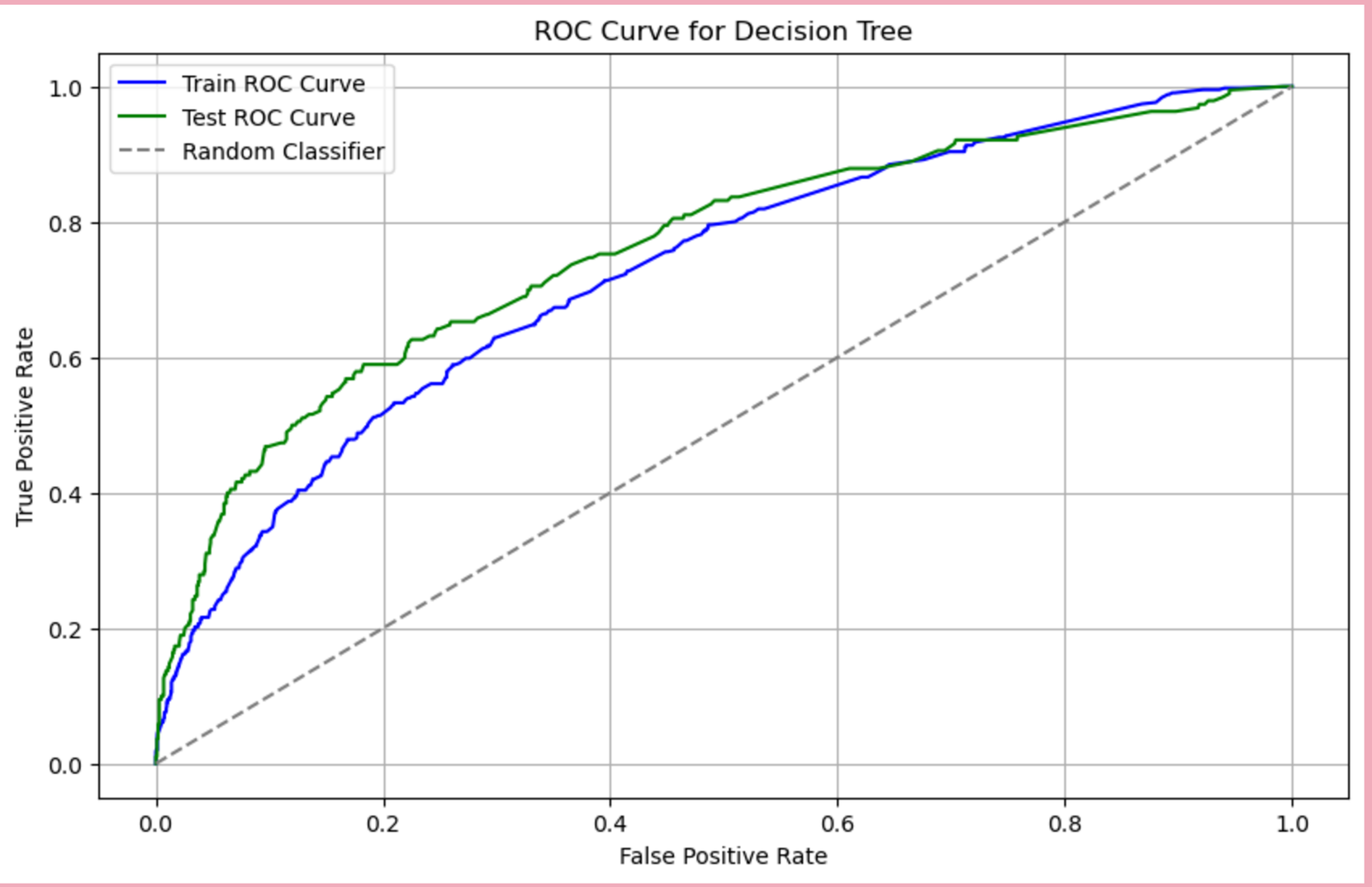
The red line is refference for ideal predictions but most points do not lie on this line. This can indicate that the model is not perfect in predicting survival months.

Classification - Model Comparison

- Objective:
  - Classify patients into aggressive vs. non-aggressive cancer.
- Models Used:
  - Logistic Regression
  - Decision Tree (with SMOTE and hyperparameter tuning)
- Key Findings:
  - Logistic Regression: High accuracy but struggles with minority class (Class 1).
  - Decision Tree: Improved recall for minority class but lower overall accuracy.



The model is performing reasonably well on both training and testing data, and is significantly better than a random classifier. The slight underfitting indicated by the difference between the training and testing curves



The curve suggests that the model is better than a random classifier. However,there is possible overfitting as evident in the gap between the train and test curves

Insights and Recommendations

- Model Suitability:
  - Logistic Regression for overall accuracy.
  - Decision Tree with SMOTE for minority class sensitivity.
- Next Steps:
  - Explore advanced models like Random Forest or Gradient Boosting.
  - Consider ensemble methods and further hyperparameter tuning.
  - Balance between precision and recall depending on hospital needs.

Thank You