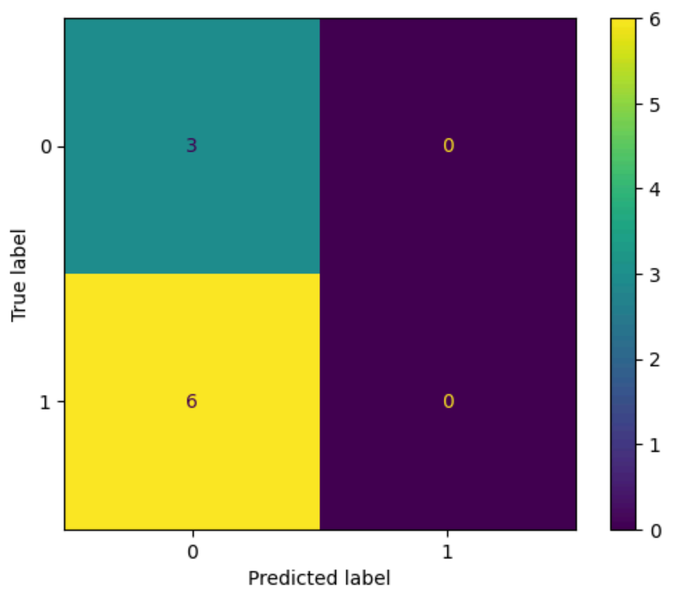
Report on 50 Startups

Introduction:

- This report analyzes 50 startups using a machine learning model, specifically Logistic Regression. Logistic Regression is chosen due to its capability for binary classification, making it suitable for this analysis. The target variables are categorical and indicate the location of the startups: California and Florida. The objective is to understand the impact of various factors on predicting startup success in these locations.

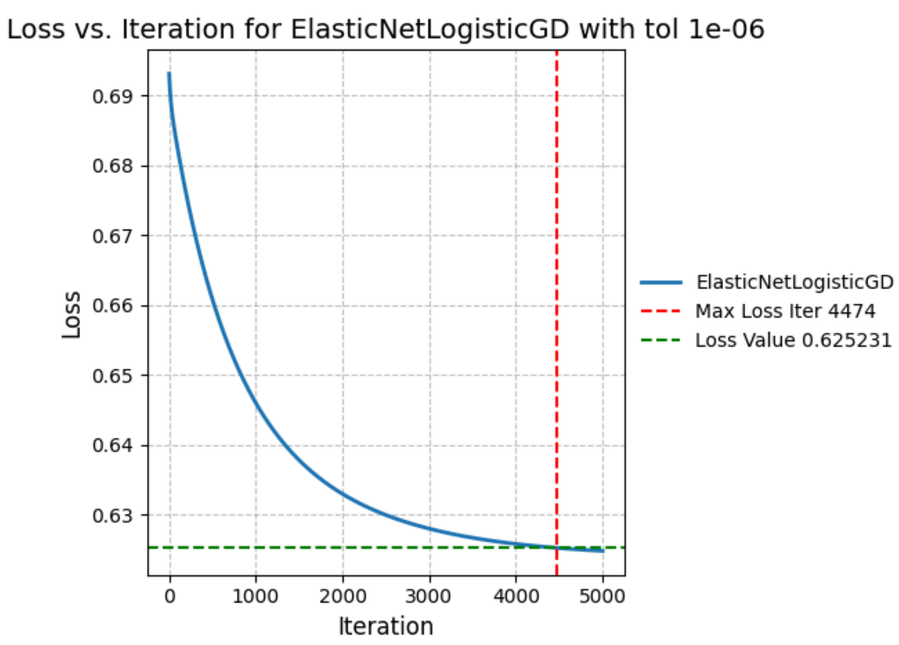
Preparation:

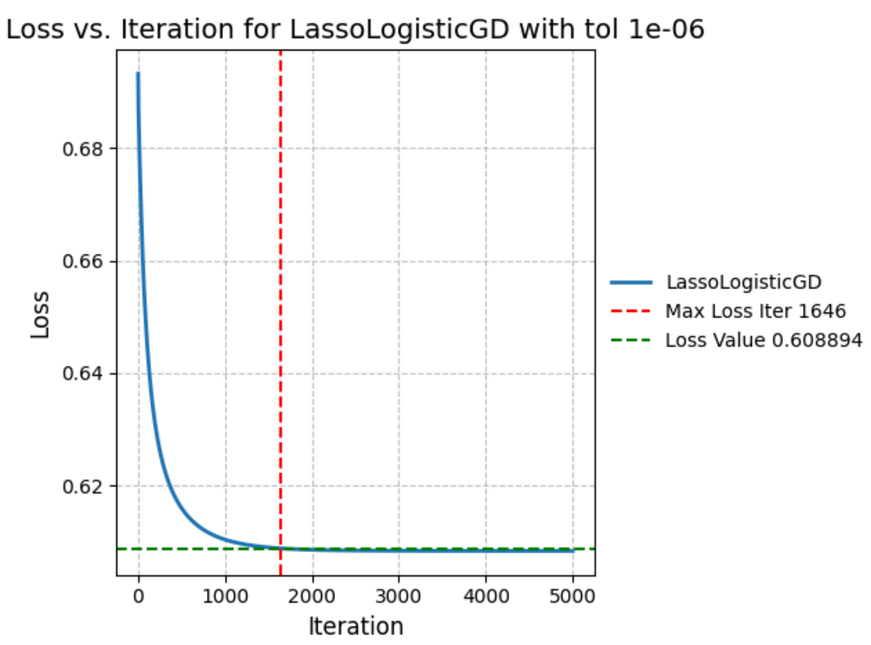
- Through initial analysis, the dataset reveals four predictor variables: R&D Spend, Administration, Marketing Spend, and Profit, each with 33 data points.
- The dataset is split into 75% for training and 25% for testing. An initial logistic regression model, termed *Initial_log*, is implemented using sklearn. This baseline model achieves an accuracy of 33.33%.

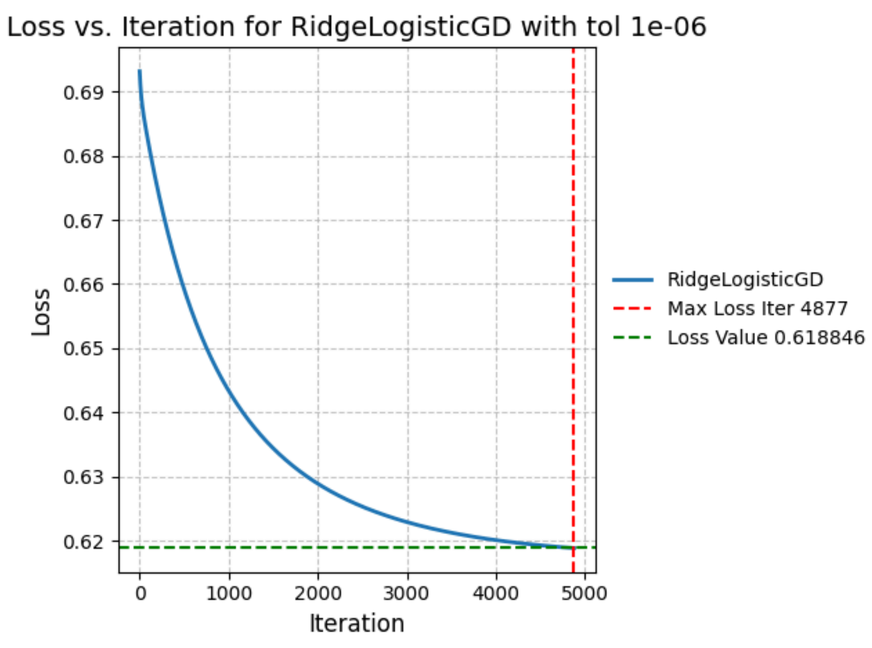


Use of Ridge, Lasso and Elastic Net:

- A confusion matrix reveals that Initial_log correctly classifies only 3 out of the 9 test samples. This is consistent with its low accuracy.
- To address this, various regularization techniques are applied:
  - Ridge Regression: $\lambda = 0.001$, learning rate = 0.1.
  - Lasso Regression: $\lambda = 0.001$, learning rate = 1.
  - Elastic Net: $\lambda = 0.001$, learning rate = 1, $\alpha = 0.5$ (equal balance of L1 and L2 penalties).
- After regularization, the accuracy remains 33.33%, but there are improvements in other metrics:
  - Precision: 20% for California and 50% for Florida.
  - Sensitivity: 33% for both locations.

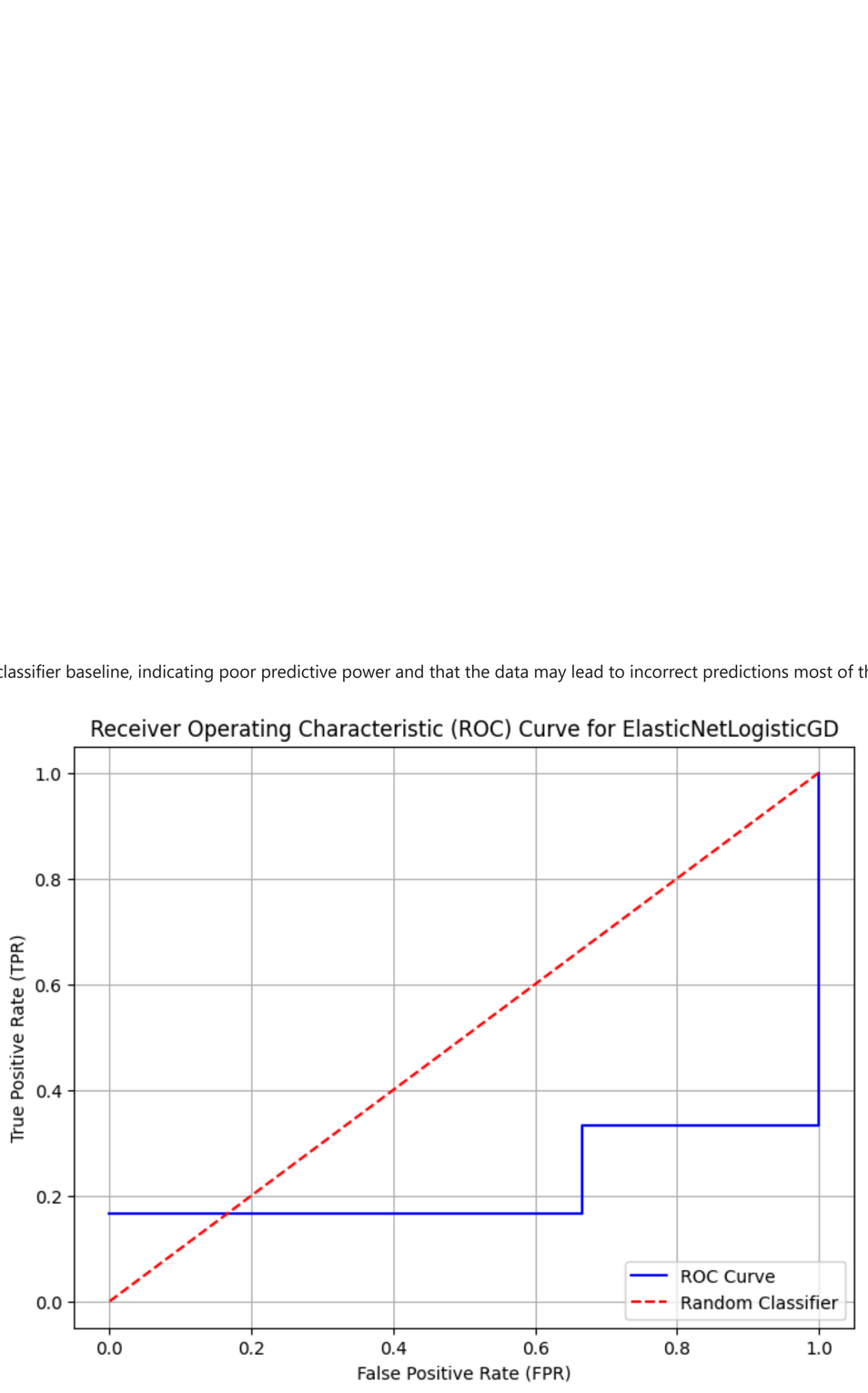






ROC Curve:

- The ROC curves for both *Initial_log* and the regularized models are plotted. Both models perform below the classifier baseline, indicating poor predictive power and that the data may lead to incorrect predictions most of the time.



Conclusion:

- The analysis demonstrates that while logistic regression is suitable for binary classification, the dataset's characteristics significantly limit model performance. Regularization improves some metrics, but the overall accuracy remains unchanged. Future work should consider:
  - Collecting more balanced and comprehensive data.
  - Exploring advanced models like decision trees.