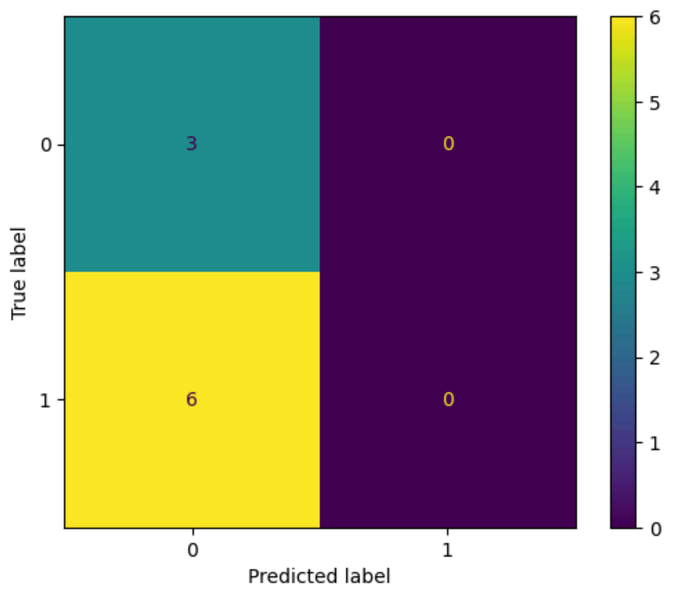
Report on 50 Startups

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.

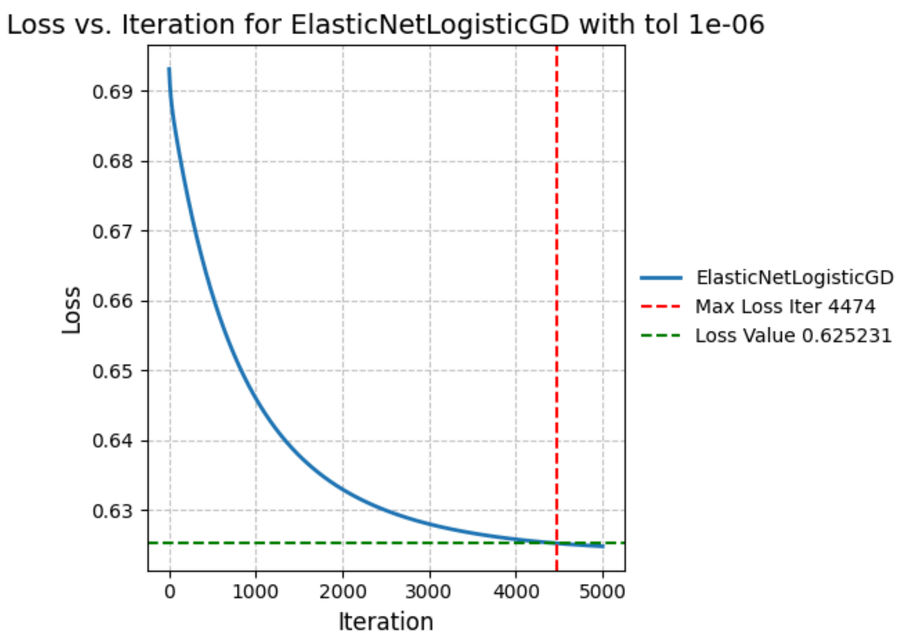
Preparation:

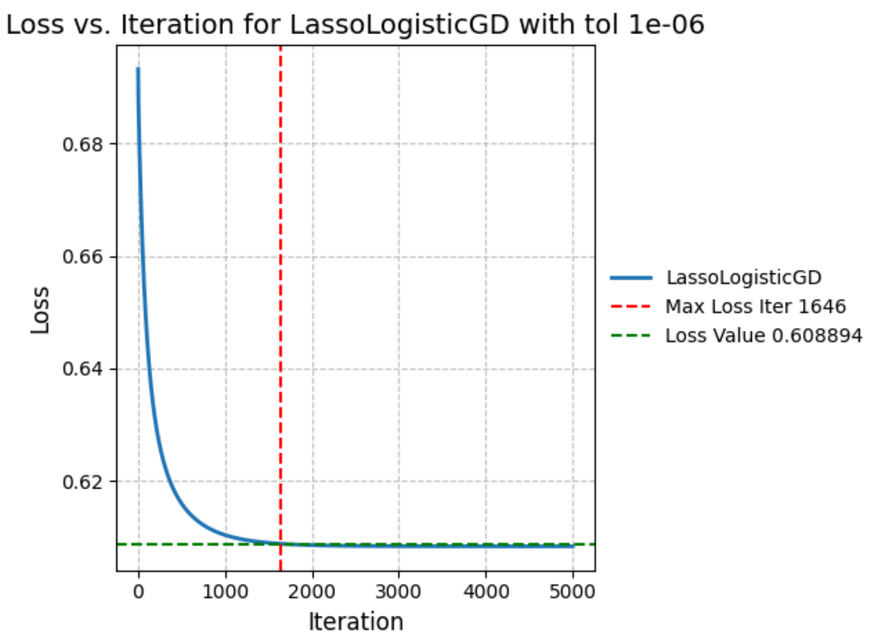
- Through initial analysis, the dataset reveals four predictor variables: R&D Spend, Administration, Marketing Spend, and Profit, each with 33 data points.
- The dataset is split into 75% for training and 25% for testing. An initial logistic regression model, termed *Initial_log*, is implemented using sklearn. This baseline model achieves an accuracy of 33.33%, specificity (Measure of True negatives out of all negatives) of 0% and and sensitivity (Measure of True Positives out of all positives) of 100%.

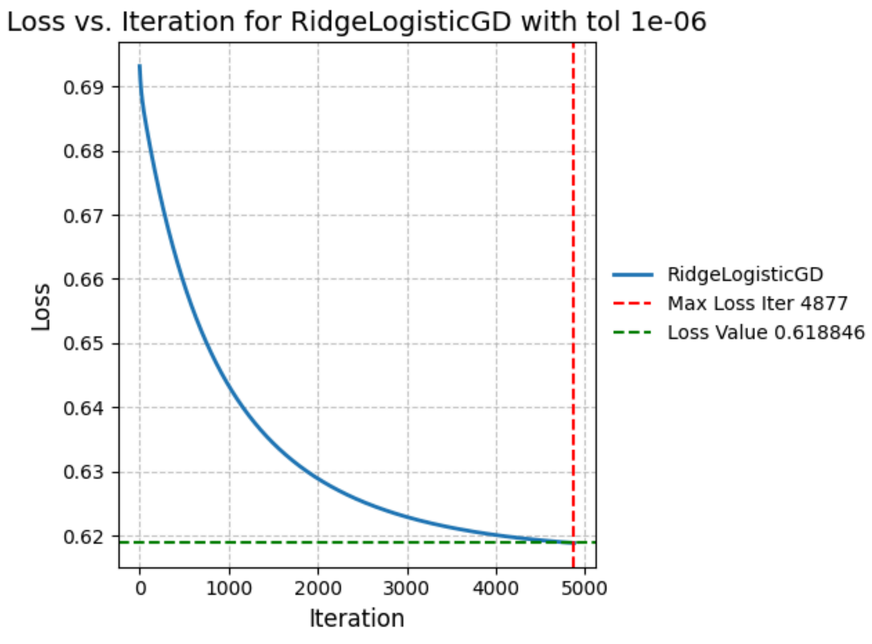


Use of Ridge, Lasso and Elastic Net:

- A confusion matrix reveals that *Initial_log* is good at only identifying those in California 3 out 9 times while Florida nothing.
- To address this, various regularization techniques are applied (trying to minimize the Loss):
  - 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 (Lasso) and L2 (Ridge) penalties).
  - (Where λ applies a penalty to introduce a small bias as to better fit the logistic curve to unknown data)
- After regularization, the accuracy remains 33.33%, but there are improvements in other metrics:
  - Precision (measure of True Positives out of all True Positives and False Positives): decreases to 20% from 33% for California and increases to 50% for Florida from 0.
  - Sensitivity: increases to 33% from 0 for Florida and decreases to 33% from 100% for California .

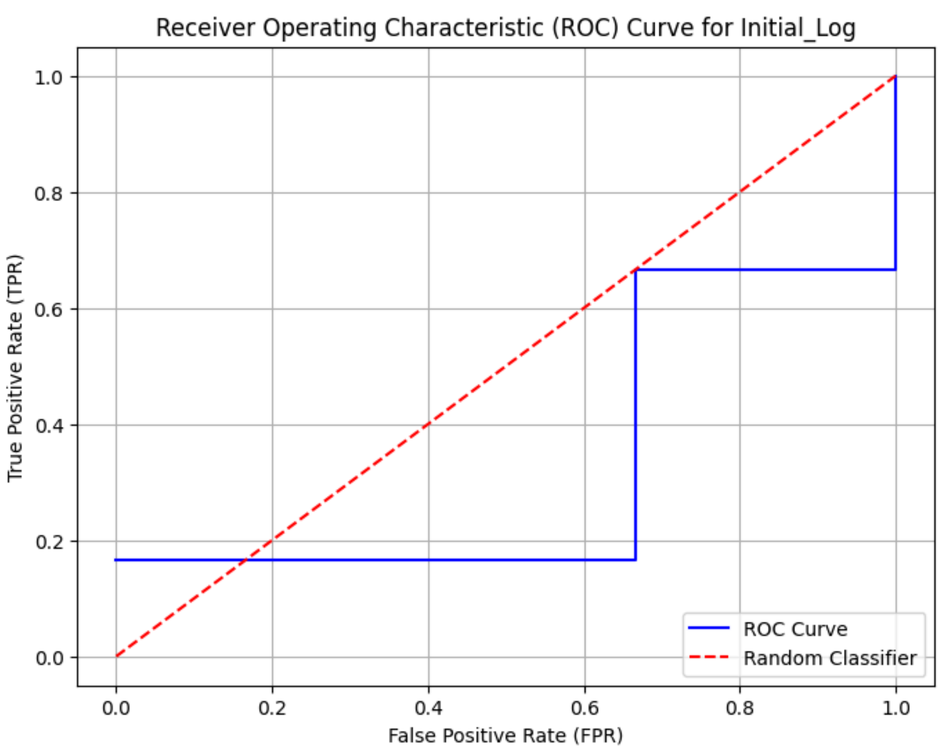


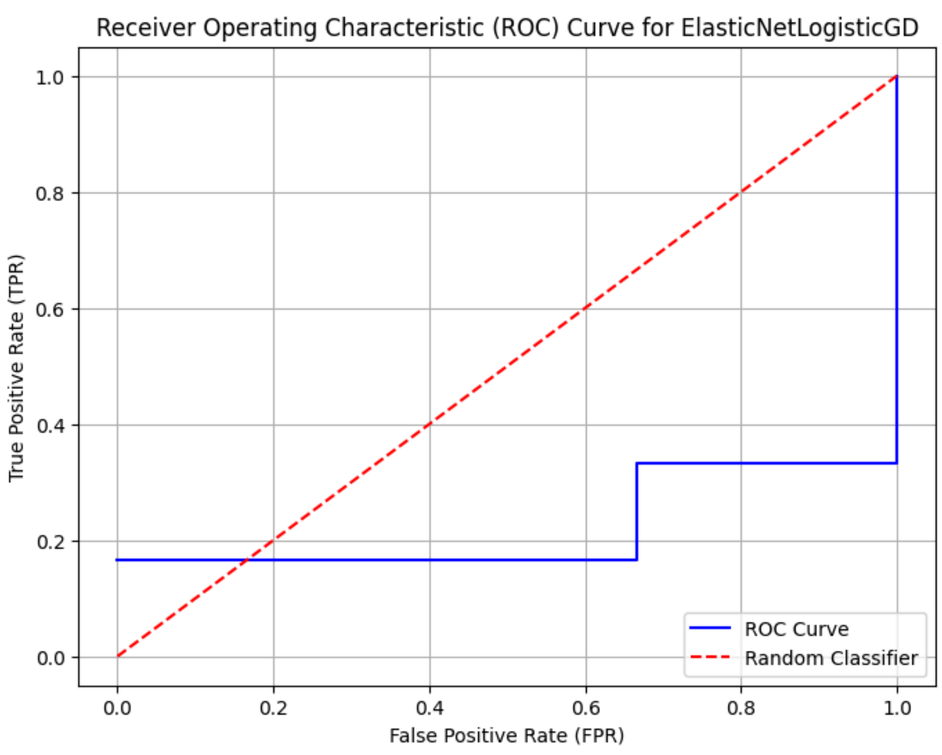


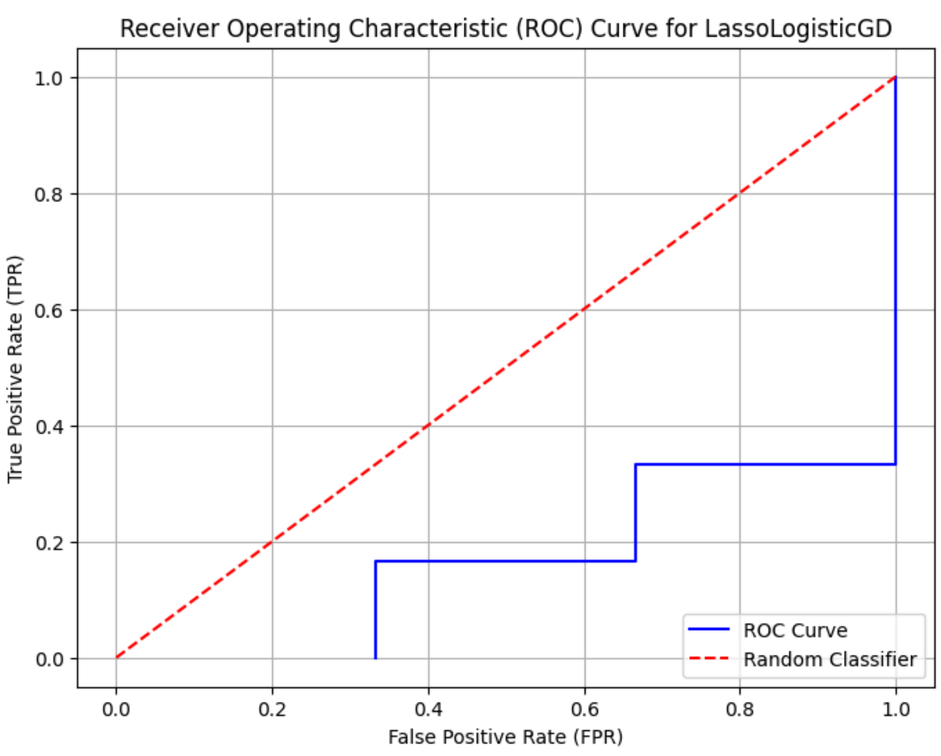


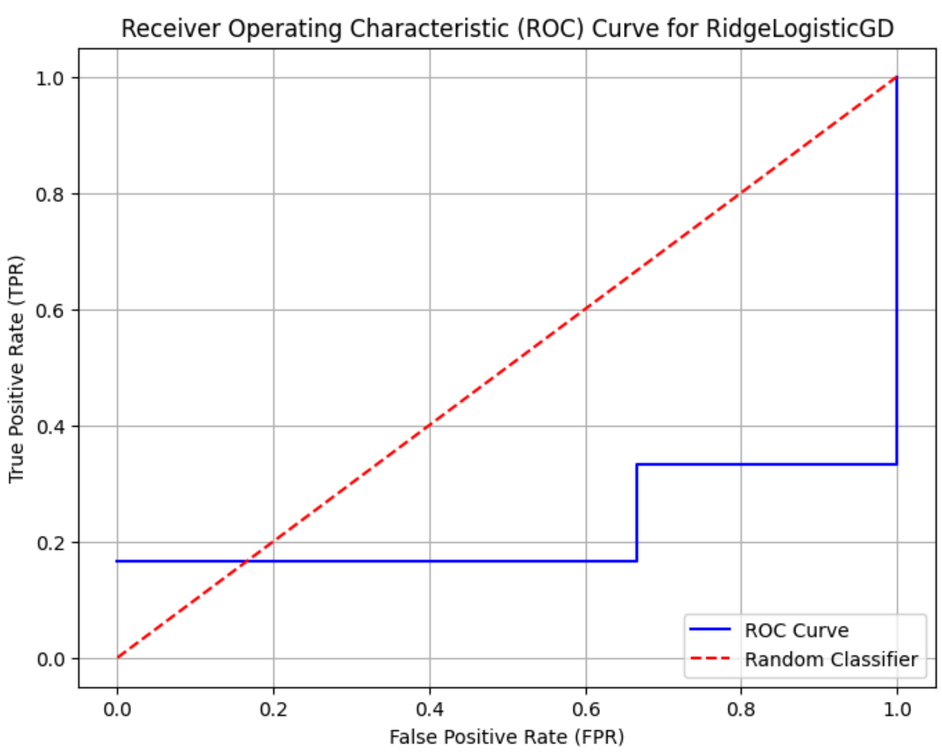
ROC Curve:

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









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. As such, one should consider:
  - Collecting more balanced data.
  - Exploring other models.