Challenge 0
Jovana Radinovic

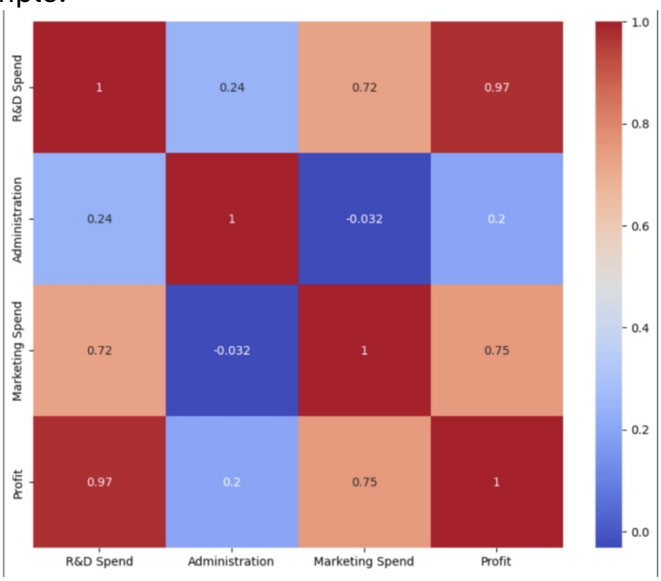
Overview
In the first challenge, we need to analyze the data we gathered from dataset "50_Startups.csv" We need to clean data, handle missing data, normalize categorical variables, build a classification model to distinguish between the two states using standard and regularized logic regression and evaluate model performance through accuracy metrics and visualizations like ROC curve.

Setup
Essential libraries used for this task included:

- **Matplotlib**: For creating 2D visualizations.
- **NumPy**: For handling multidimensional arrays and numerical operations.
- **Pandas**: For data manipulation and analysis.
- **Scikit-learn**: For machine learning models and utilities.

Data Preparation and Initial Insights

- **Dataset Inspection**:
  - The dataset contains 50 rows and 5 columns.
  - A correlation matrix was calculated to identify relationships between variables. For example:

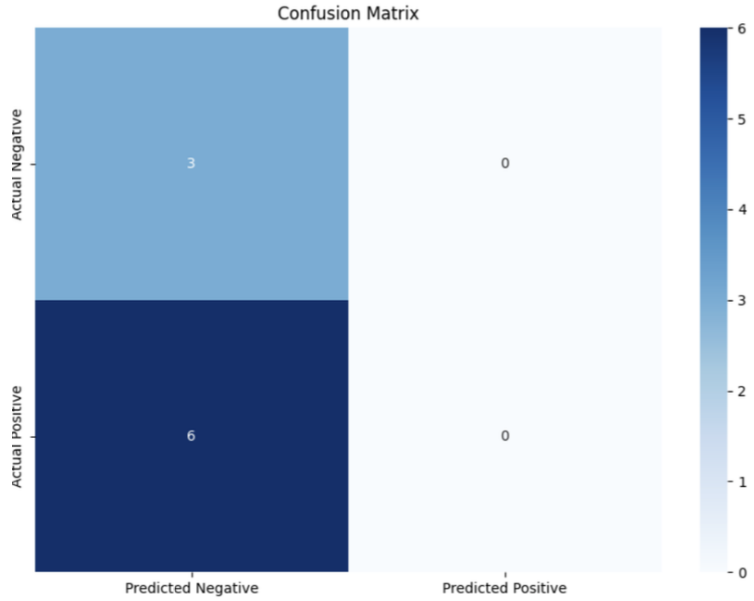


"R&D Spend" showed a strong positive correlation with "Profit" (0.9729) and "Marketing Spend" (0.7242).
  - No significant inverse correlations were found.
- **Preprocessing Steps**:
  - Missing values were replaced with column means to maintain data distribution.
  - Data was normalized to a range of [-1, 1].

Model Training and Evaluation

1. Logistic Regression:

- **Data Splitting**:
  - The dataset was split into training (75%) and testing (25%) sets.
- **Initial Results**:
  - Model accuracy was low at 33.3%.

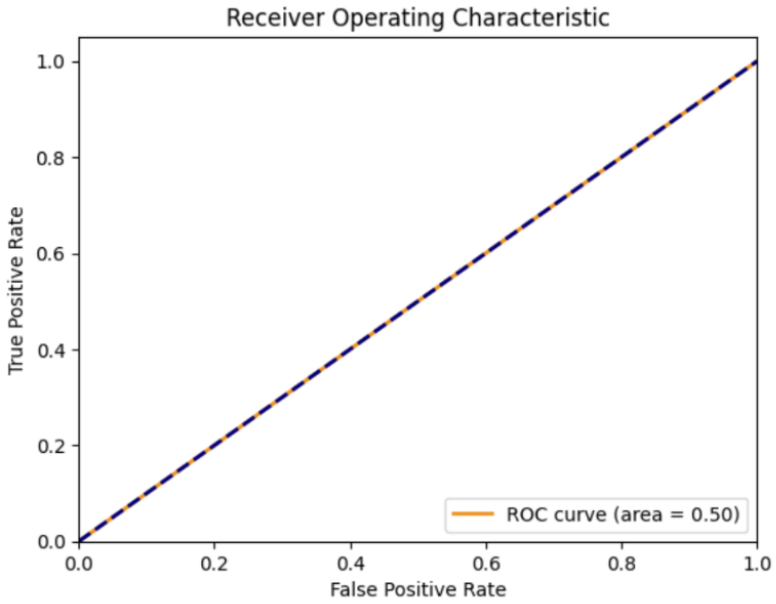


- The model is precise when predicting negative cases but still has problems predicting positive cases.
- The possible motives could be: -The parameter is not optimized
  -Dataset is not balanced
  -The model is wrong
- We can rule out that the dataset is not balanced and that the model is wrong, so the only possible explanation is that the parameter is not optimized.

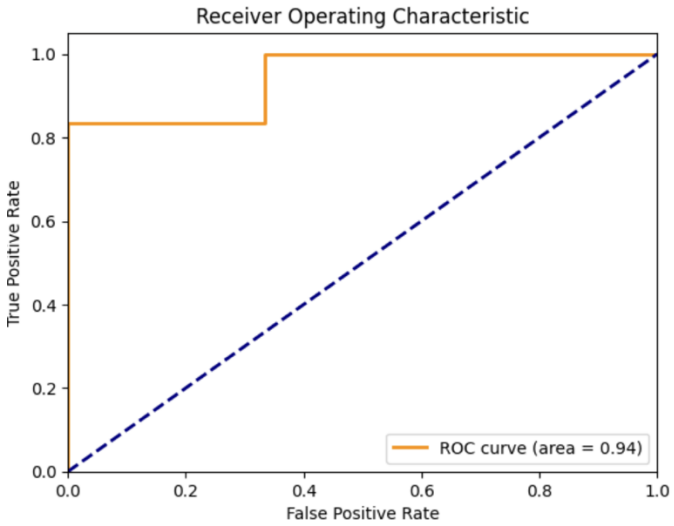
2. Regularization Techniques:

Three regularization methods were applied to improve the model:

- **L2 Regularization (Ridge)**:
  - Adjusted hyperparameters using grid search.
  - Result: ROC-AUC improved from 0.33 to 0.50.



- **L1 Regularization (Lasso)**:
  - Enhanced feature selection by penalizing irrelevant features.
  - Result: Accuracy increased to 66.7%.



- **Elastic Net**:
  - Combined L1 and L2 regularization for balanced feature selection and weight penalty.
  - Result: Performance slightly worsened compared to L1 regularization.