

Report

Assignment 3 – MLG382



May 21, 2023

Belgium campus

Table of Contents

[Intro 2](#_Toc135586334)

[Process 2](#_Toc135586335)

[1. Data Preparation: 2](#_Toc135586336)

[2. Feature Engineering: 2](#_Toc135586337)

[3. Model Training and Evaluation: 2](#_Toc135586338)

[Results: 2](#_Toc135586339)

[Recommendations and Further Analysis: 2](#_Toc135586340)

**Group Members:**

* Kagiso Kgobane
* Gregory Avvakoumides
* Acseivire Mboto

# Intro

The goal of this analysis was to build a multiple linear regression model to predict the profitability (Profit) of startups based on their investment in Research and Development (R&D Spend), Marketing Spend, and an interaction term between R&D Spend and Marketing Spend. The model was evaluated using the R-squared (r2) score, which measures the proportion of the variance in the target variable explained by the model.

# Process

## Data Preparation:

The "50\_Startups" dataset was loaded, consisting of information about 50 startups, including their R&D Spend, Marketing Spend, and Profit.

## Feature Engineering:

An interaction term between R&D Spend and Marketing Spend was created to capture any potential combined effect of these variables on profitability. This additional feature aimed to improve the model's performance.

## Model Training and Evaluation:

The dataset was split into training and testing sets using an 80-20% ratio. A multiple linear regression model was trained on the training set, using the features (R&D Spend, Marketing Spend, and Interaction) and the corresponding target variable (Profit). The trained model was used to predict the profitability of the test set. The model's performance was evaluated using the R-squared score, which quantifies the proportion of the variance in the test set's Profit that can be explained by the model.

# Results:

The multiple linear regression model achieved an R-squared score of 0.9149371871179751. A high R-squared score suggests that the model captures a large portion of the variability and provides a better fit to the data.

# Recommendations and Further Analysis:

Based on the current model's performance, there are possible actions for improvement, such as exploring additional features, trying different algorithms, optimizing hyperparameters, or increasing the dataset size.