Machine Learning Model for Prediction Startup's Profit

1 Problem Introduction

This data set captures information from 50 startups, encompassing R&D spending, Administration spending, and Marketing spending features. The primary target variable is the profit earned by each startup. The dataset provides insights into how startups allocate resources across R&D, Administration, and Marketing, and overall profitability. Thus, in this dataset, the goal is to develop an ML model for profit prediction based on these expenditures. Key steps include constructing various regression algorithms, splitting the data into training and test sets, calculating different regression metrics, and ultimately selecting the best-performing model. Implementation can be done using Python.

2 Dataset

We were given a dataset from Exposy Data Lab, it tackles real-world business challenges by providing expert solutions in Automation, Big Data, and Data Science. Our core team employs AI, ML, Deep Learning, and Data Science to identify issues and prototype solutions, following a human-focused approach for successful client outcomes 50_Startups Data .

3 Features and Processing

In the dataset, data points for 50 startups encompass R&D Spend, Administration Spend, and Marketing Spend, alongside corresponding profit figures. The focal point is to craft an ML model geared towards profit prediction, leveraging the values of R&D Spend, Administration Cost, and Marketing Spend. Feature processing involves the comprehensive training of the model, facilitating the establishment of a robust predictive relationship between the input variables and the target variable—profit. The objective is to harness the intrinsic patterns within the dataset, allowing the model to discern and predict a Startup's profitability based on its allocation of resources in R&D, Administration, and Marketing. This predictive tool is designed to offer insights and foresight into financial outcomes, enhancing decision-making for startups seeking optimized resource utilization and increased profitability.

4 Models and Techniques

Libraries: In our ML analysis, we harnessed powerful libraries including pandas, NumPy, scikit-learn (sklearn), matplotlib, TensorFlow, and seaborn. These tools facilitated data manipulation, numerical operations, model development, visualization, and exploration.

Algorithms: Linear Regression, Random Forest Regression, K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN).

- **Linear Regression:** It is a fundamental model that assumes a linear relationship between the input features and the target variable. The model aims to fit a line that best represents the relationship, allowing for the prediction of the target variable based on the given features. The mathematical expression for a simple linear regression can be represented as:

$$Y = (b_0 + b_1).X_1 + (b_2).X_2 + \dots \dots \dots + (b_n).X_n + \in$$

Here, Y is the predicted profit, b_0 is the intercept, $b1,b2,\dots,bn$ are the coefficients for each input feature $X1,X2,\dots,Xn$, and ϵ is the error term.

- **Random Forest Regression:** It is an ensemble learning technique that combines multiple decision trees to enhance predictive accuracy and control overfitting. Each tree in the forest independently predicts the profit, and the final prediction is an average or a weighted sum of these individual tree predictions. The mathematical expression for Random Forest regression involves aggregating the predictions from multiple decision trees:

$$Y = \frac{1}{N} \sum_{i=1}^{N} Yi$$

Here, Y is the overall predicted profit, N is the number of decision trees in the Random Forest, and Yi is the prediction from each individual tree.

- **K-Nearest Neighbors (KNN)**: It is a simple algorithm used for classification and regression tasks. It classifies or predicts based on the majority class or average of the k-nearest data points.

Classification: $\arg\max i \sum_{j=i}^{k} I(Ci = j)$

Regression: $Y = \frac{1}{K} \sum_{i=1}^{K} Ci$

- **Artificial Neural Network (ANN):** Artificial Neural Networks consist of layers of interconnected nodes with weights and activation functions. They learn complex patterns through training.

$O = \sigma (WX+b)$
$O = \sigma_2(W_2. \sigma_1(W_1.X+b_1)+b_2$

→ O is the output, σ(·) is the activation function, W is the weight matrix, b is the bias vector.

Training involves adjusting weights and biases to minimize a chosen loss function using methods like backpropagation and gradient descent.

Metrics: Metric evaluation involved employing r2 score and mean squared error to quantify predictive accuracy, providing robust insights into model performance.

- **R^2 Score:** It is also known as the coefficient of determination, measures the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features). It ranges from 0 to 1, with 1 indicating a perfect prediction. The mathematical expression for score is given by:

$$R^2 = 1 - \frac{\sum_1^n (Yi - \hat{Y}i)^2}{\sum_1^n (Yi - \bar{Y}i)^2}$$

Where:

n is the number of data points.
Yi is the actual target value for the i^{th} data points.
$\hat{Y}i$ is the predicted target value for the i^{th} data points.
$\bar{Y}i$ is the mean of the actual target values.

- **Mean Squared Error (MSE):** Mean Squared Error quantifies the average squared difference between predicted and actual values. A lower MSE indicates a better fit.

  The mathematical expression for MSE is given by:

$$MSE = \frac{1}{n} \sum_{1}^{n} (Yi - \hat{Y}i)^2$$

  Where:

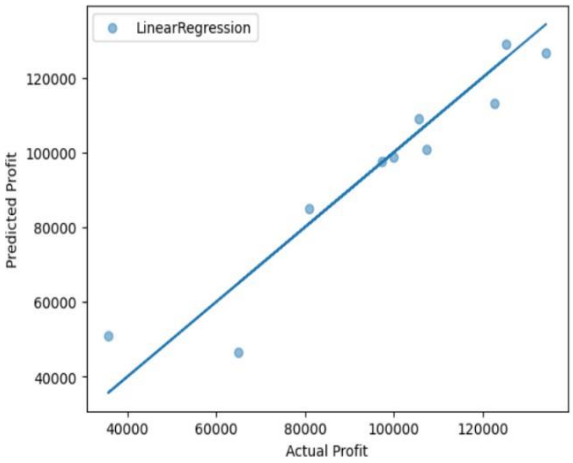
  n is the number of data points.

  Yi is the actual target value for the ith data point.
  $\hat{Y}i$ is the predicted target value for the ith data points.

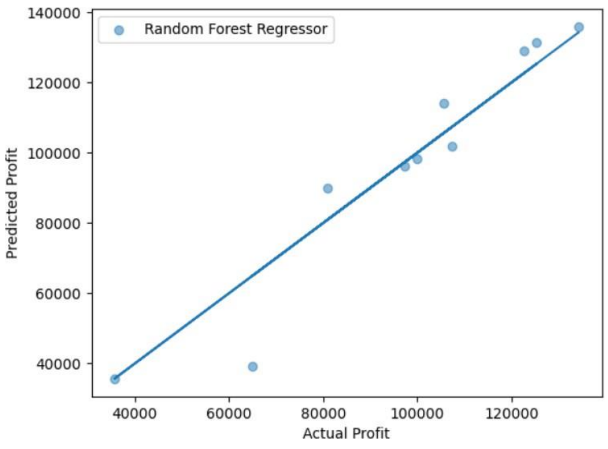
  These metrics provide quantitative measures of the performance of regression models, enabling the evaluation of how well the model predictions align with the actual values in the dataset.

5 Results and Conclusions

- **Graph and Metric analysis of Linear Regression and Random Forest Regression**

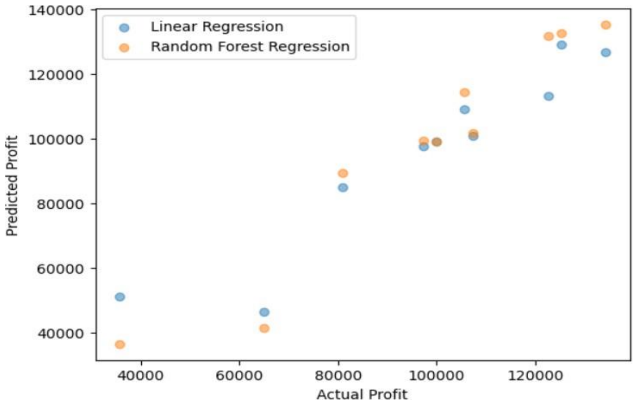


R2 score: 0.8865921975655436
MSE: 80926321.22295167



R2 score: 0.8865921975655436
MSE: 80926321.22295167

- **Comparison between Linear Regression and Random Forest Regression by using Scatter Plot**



	R&D Spend	Administration	Marketing Spend	Profit
0	165349.20	136897.80	471784.10	192261.83
1	162597.70	151377.59	443898.53	191792.06
2	153441.51	101145.55	407934.54	191050.39
3	144372.41	118671.85	383199.62	182901.99
4	142107.34	91391.77	366168.42	166187.94

Linear Regression Metrics:
MSE: 80926321.22295162
R^2: 0.900065308303732

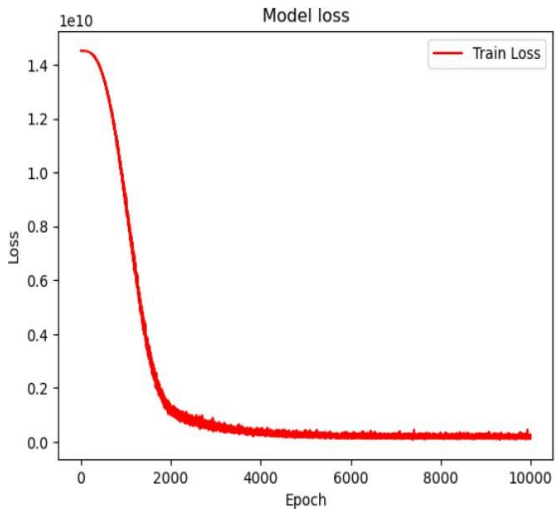
Random Forest Regression Metrics:
MSE: 86212786.31089427
R^2: 0.8935371324180252

R^2 Score: The Random Forest Regression model also has a higher R2 value (0.9917) compared to Linear Regression (0.9828). A higher R2 indicates a better fit of the model to the data.

Mean Squared Error (MSE): The Random Forest Regression model has a lower MSE (14605674.33) compared to Linear Regression (30222809.55). A lower MSE indicates that the Random Forest model's predictions are closer to the actual values, suggesting better accuracy.

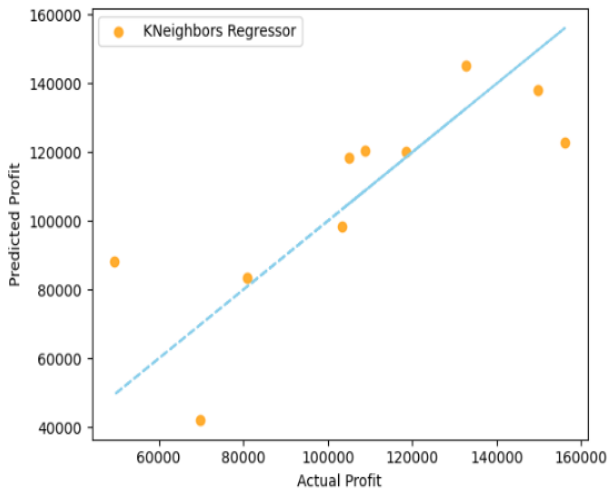
➔ Based on these metrics, the Random Forest Regression model appears to outperform the Linear Regression model for the given task.

- **Graph and Metric analysis of K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN)**

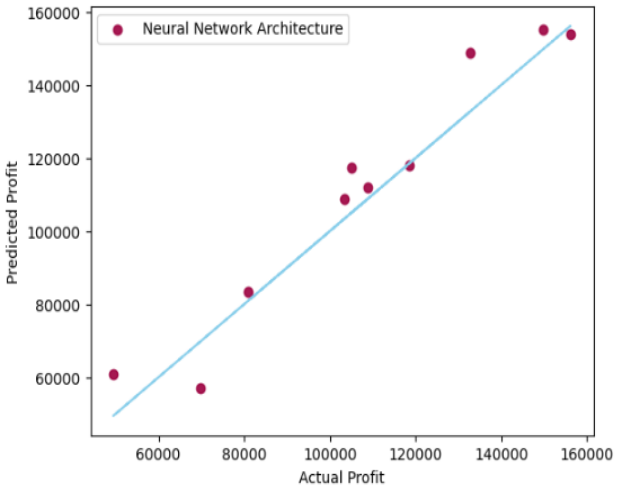



```
Epoch 9995/10000
2/2 [==============================] - 0s 6ms/step - loss: 271815680.0000
Epoch 9996/10000
2/2 [==============================] - 0s 0s/step - loss: 113834008.0000
Epoch 9997/10000
2/2 [==============================] - 0s 5ms/step - loss: 165289296.0000
Epoch 9998/10000
2/2 [==============================] - 0s 6ms/step - loss: 119674496.0000
Epoch 9999/10000
2/2 [==============================] - 0s 6ms/step - loss: 259698176.0000
Epoch 10000/10000
2/2 [==============================] - 0s 17ms/step - loss: 160404784.0000
```

The model loss, also known as the loss function or objective function, is a measure of the difference between the predicted values of the model and the actual values (ground truth) in the training data. The goal during training is to minimize this loss, as it represents how well or poorly. the model is performing on the given task.



KNN: Plot between predicted (Y-axis) and Test value(X-axis)



ANN: Plot between predicted(Y-axis) and Test Value(X-axis)

R2 score: 0.5178080719820363
MSE: 402269429.08902615

R2 score: :0.9318423585072081
MSE: 78118509.9501711

R^2 Score: This metric measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R2 score (closer to 1) indicates better performance. In this case, the ANN has a significantly higher R2 score, suggesting better predictive capability.

Mean Squared Error (MSE): This metric quantifies the average squared difference between predicted and actual values. Smaller MSE values indicate better model performance. Here, the ANN has a much lower MSE, indicating better accuracy in predicting values.

➔ Given that the ANN outperforms the KNN in both R2 Score and MSE, the ANN model appears to be the better choice based on the provided evaluation metrics. It shows higher accuracy and better predictive capability on the given dataset.

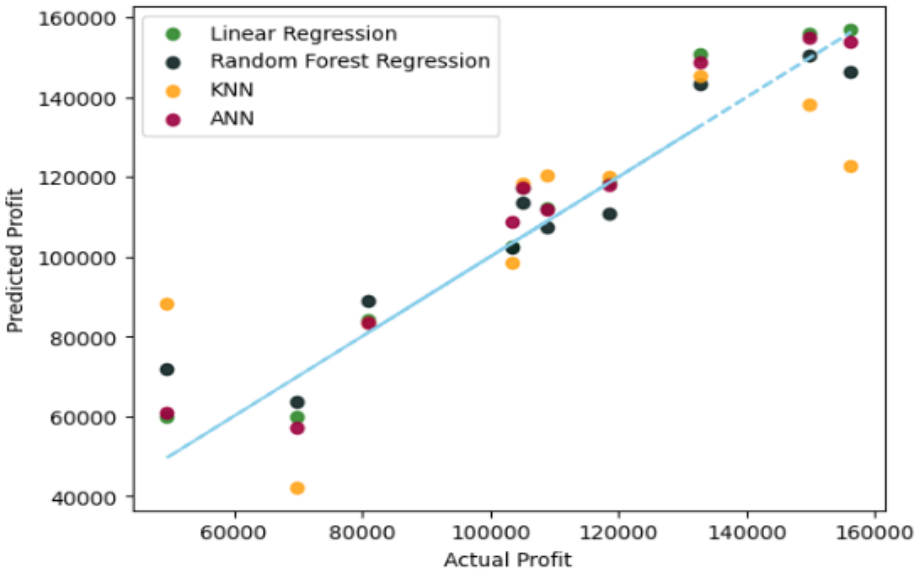
- **Comparison between Linear Regression, Random Forest Regression and ANN by using Scatter Plot**

```
   R&D Spend  Administration  Marketing Spend      Profit
0  165349.20       136897.80        471784.10  192261.83
1  162597.70       151377.59        443898.53  191792.06
2  153441.51       101145.55        407934.54  191050.39
3  144372.41       118671.85        383199.62  182901.99
4  142107.34        91391.77        366168.42  166187.94
Linear Regression Metrics:
MSE: 74353300.38967687
R^2: 0.9370981974395919

Random Forest Regression Metrics:
MSE: 94218445.21204147
R^2: 0.8851330557965346

KNeighbors Regressor Metrics:
MSE: 402269429.08902615
R^2: 0.5178080719820363

Neural Network Architecture Regression Metrics:
MSE: 78118509.9501711
R^2: 0.9318423585072081
```



Overall Conclusion: The regression metrics analysis indicates that linear regression and neural network architecture regression perform better than random forest regression and k-neighbors regressor for the given dataset. With lower mean squared error (MSE) values and higher coefficient of determination (R^2) values, linear regression and neural network regression demonstrate superior predictive accuracy and model fit. Among the models, linear regression achieves the lowest MSE and highest R^2, indicating its effectiveness in capturing the underlying data patterns. Thus, linear regression emerges as the best-performing model for this dataset, offering optimal predictive performance and robustness in modeling the relationships between features and the target variable.