STARTUP PROFIT PREDICTION

SRINIVASAN K DATA SCIENCE INTERN

EXPOSYS DATA LABS

ABSTRACT:

Start-ups play a major role in economic growth. They bring new ideas, spur innovation, create employment thereby moving the economy. There has been an exponential growth in start-ups over the past few years.

Predicting the success of a start-up allows investors to find companies that have the potential for rapid growth, thereby allowing them to be one step ahead of the competition.

The use of mathematical and statistical concepts such as Regression Analysis to determine the profit of a start-up depending on their expenditure will help us invest more strictly and intelligently.

TABLE OF CONTENTS:

INTRODUCTION:

The objective of this project is to predict whether a start-up company has amassed profit depending on the expenditure the company spend in the respective fields of Research and Development, Marketing and Administration.

The profit of the company should be inferred from the given details, and their connection should be exploited to do so.

PROBLEM STATEMENT:

In the given dataset, R&D Spend, Administration Cost and Marketing Spend of 50 Companies are given along with the profit earned. The target is to prepare an ML model which can predict the profit value of a company if the value of its R&D Spend, Administration Cost and Marketing Spend are given.

- Construct Different Regression algorithms
- Divide the data into train set and test set
- Calculate different regression metrics
- Choose the best model

EXISTING METHODS:

For these kinds of problems, there are several existing mechanisms to predict the profit under different schemes. A few of them are:

- **Clustering Algorithms:**

  Falling under the family of unsupervised ML algorithms, clustering is used to analyse unlabelled data, segregate it into groups with similar traits, and assign it into clusters. This is a subjective task, so you can use different algorithms to solve it.

- **Association Rules:**

  Another ML method that every data scientist should learn to be in high demand is association rules. A popular technique for uncovering interesting relationships between different variables in huge databases, association rules are actively harnessed to build recommendation engines

- **Markov Chains:**

  Markov chains are a common way to statistically model random processes. This method is used to describe a possible sequence of events (transitions) based solely on the process' present state, independently from its full history.

The need-to-know idea is that, the existing methodologies for using these products does not appease the given dataset. The given dataset is purely mathematical and thus need a statistical and mathematical approach to provide the ideal details.

PROPOSED METHODS:

The proposed method for this project is **Regression Analysis.**

Regression Analysis is a predictive modelling technique (supervised Machine Learning Method) that analyses the relation between the target or dependent variable and independent variable in a dataset.

The different types of regression analysis techniques get used when the target and independent variables show a linear or non-linear relationship between each other, and the target variable contains continuous values.

The regression technique gets used mainly to determine the predictor strength, forecast trend, time series, and in case of cause & effect relation.

There are five kinds of basic regression models:

- **Linear Regression (or) Multi-Linear Regression:**

  Linear regression is one of the most basic types of regression in machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other.

  In case the data involves more than one independent variable, then linear regression is called multiple linear regression models.

  The linear regression can be explained mathematically as follows: $y = mx + c + e$

  The slope (m), the intercept (c) and the possible errors (e) determine the outcome of our prediction.

  The values of m and c get selected in such a way that it gives the minimum predictor error. It is important to note that a simple linear regression model is susceptible to outliers.

- **Logistic Regression:**

  Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc. This means the target variable can have only two values, and a sigmoid curve denotes the relation between the target variable and the independent variable.

  Binary classification or separation of discreet dependent values with the help of independent variables can be done with Logistic Regression.

  We cannot perform this case study analysis using Logistic Regression.

- **Ridge Regression (or) L2 Regression:**

  This is another one of the types of regression in machine learning which is usually used when there is a high correlation between the independent variables.

  This is because, in the case of multi collinear data, the least square estimates give unbiased values. But, in case the collinearity is very high, there can be some bias value.

  Therefore, a bias matrix is introduced in the equation of Ridge Regression. This is a powerful regression method where the model is less susceptible to overfitting.

  Ridge Regression is a popular type of regularized linear regression that includes an L2 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction task.

- **LASSO Regression (or) Penalized Regression:**

  The full form of LASSO is the Least Absolute Shrinkage and Selection Operation. As the name suggests, LASSO uses the "shrinkage" technique in which coefficients are determined, which get shrunk towards the central point as the mean.

  The LASSO regression in regularization is based on simple models that possess fewer parameters. We get a better interpretation of the models due to the shrinkage process. The shrinkage process also enables the identification of variables strongly associated with variables corresponding to the target.

- **Polynomial Regression:**

  Polynomial Regression is another one of the types of regression analysis techniques in machine learning, which is the same as Multiple Linear Regression with a little modification. In Polynomial Regression, the relationship between independent and dependent variables, that is X and Y, is denoted by the n-the degree.

  It is a linear model as an estimator. Least Mean Squared Method is used in Polynomial Regression also. The best fit line in Polynomial Regression that passes through all the data points is not a straight line, but a curved line, which depends upon the power of X or the value of n.

  Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables do we add some polynomial terms to linear regression to convert it into Polynomial regression.

METHODOLOGY:

The proposed methods are implemented and are compared to pick the model to apply.

The Data Summary:

There are four columns in the given data set.

- The Research and Development Cost
- Marketing Spend
- Administration Spend
- Profit

All the values are quantitative.

Data Pre-Processing:

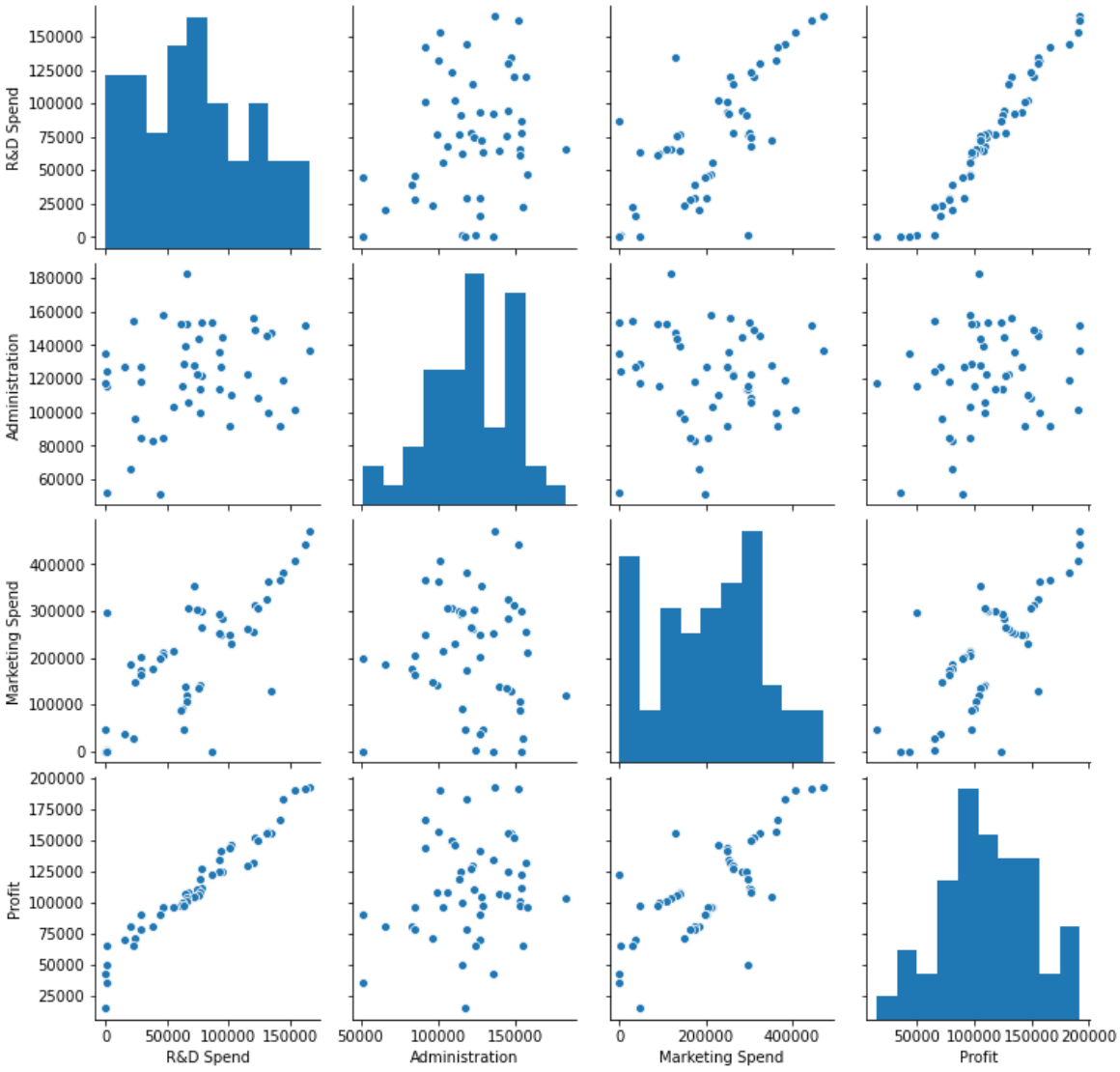
There are several tasks in data pre-processing. They are:

1. Data Visualization:

The data is accumulated and included in the file and used to check.

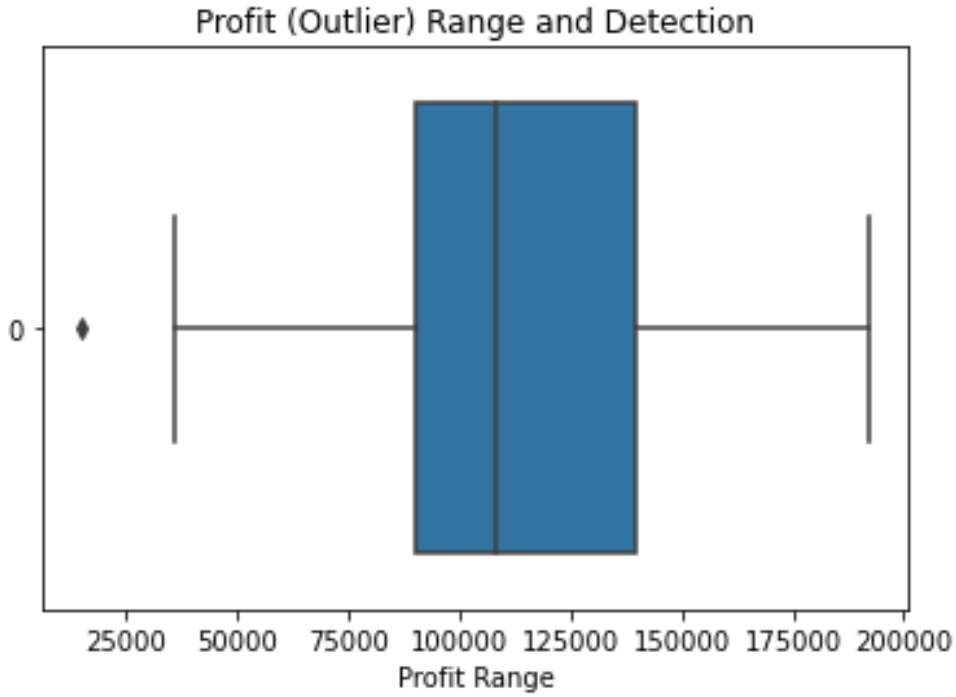
Then it is visualized both statistically and mathematically.

The visualization using the Pair Plot symbolizes the rate of dependency between each of the variables available.



2. **Picking the Outliers in the Dependent Variable:**

The outliers in the profit are found using Box Plot.



Model Description:

1. **Splitting the Dataset into Dependent and Independent Variables:**

The dataset is split into two data frames: X (independent) and Y (dependent) variables.

They are converted into data frames for easier access.

2. **Splitting the Dataset Testing and Training Sets:**

The dataset is further split into train and test splits.

Thus, we have x_train and y_train, used for training the model and x_test and y_test for testing the model.

3. **Applying the Models:**

The Regression Models are applied to be fit the Training Dataset and then are used to predict the Testing Dataset.

Then, their metrics are found and analysed.

4. **Picking the Models:**

Finally, the model is picked using the metrics of regression.

REGRESSION METRICS:

It is necessary to obtain the accuracy of training data, But it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.

So to build and deploy a generalized model we require to Evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result.

The four basic evaluation parameters that are used in Regression Analysis are:

R2 Score (or) the Co-efficient of Determination:

It works by measuring the amount of variance in the predictions explained by the dataset. Simply put, it is the difference between the samples in the dataset and the predictions made by the model.

Mean Squared Error:

MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.

Hence, **MSE** = $\frac{\sum(\boldsymbol{True} - \boldsymbol{Prediction})^2}{\boldsymbol{N}}$**,** where, N is the number of values.

Root Mean Squared Error:

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

Hence, **RMSE =** $\sqrt{\frac{\sum(True - Prediction)^2}{N}}$, where, N is the number of values.

Mean Absolute Error:

The absolute difference means that if the result has a negative sign, it is ignored. Hence, MAE = True values – Predicted values. MAE takes the average of this error from every sample in a dataset and gives the output.

IMPLEMENTATION:

The implementation of the project is done using Python and the required libraries are imported.

Then, the data is checked for discrepancies and is then visualized graphically.

Then the models are applied and the methods are checked. Finally, the model is picked.

It tries to determine how strongly related one dependent variable is to a series of other changing variables referred to as independent variables.

The dependent variable is the one that we focus on.

Independent variables are the factors that may or may not affect the dependent variable.

Dependent receives the impact, while Independent provides (or not) the impact.

Regression analysis helps us determine which factors matter and their relationships. It also helps us find out what their effects are on sales figures. Here, we use the available independent variable to predict the dependent profit value.

The code for the implementation is provided in the annexure.

RESULT:

After the implementation of all models, the data is then obtained based on their evaluation metrics.

Regression	R2 SCORE	MSE	RMSE	MAE
Linear	0.9341560653448712	62240269.84291537	7889.250271281509	6489.660170486654
Ridge	0.9341560648946424	62240270.076091304	7889.250286059589	6489.660170486654
Lasso	0.934156066059954	62240269.2969394	7889.250236678984	6489.660143898275
Polynomial	0.934156066059954	89144211.20924094	9441.62121720846	8209.075128317101

From this, it is inferred that Linear Regression provides a perfectly appropriate idea for us to depend upon.

CONCLUSION:

The machine learning model that predicts the profit of a start-up was built successfully using the Regression Analysis. For further improvement, we could be able to perform an extraction of a detailed feature to evaluate it in a much smoother manner.