Name: Rohit Paradkar

Institution:

K.J. Somaiya Institute of Engineering and Information Technology

Project Name:

Profit Prediction

Submitted To:

EXPOSYS DATA LAB

1. Abstract

The profit earned by a company for a particular period depends on several factors like how much R&D Spend, Administration Cost and Marketing Spend. So, for predicting the profit of a company for a particular period we need to train a machine learning model with a dataset that contains historical data about the profit generated by the company of 50 Companies. So, we have used five different machine learning algorithms to predict the profit. Out of five model Linear Regression model is best fit from the models such as Linear Regression, Random Forest, Decision Tree, Polynomial Regression, Artificial Neural Network

Table of Contents

2. Introduction

In today's world, data is generated everywhere, much like moving to another location (GPS data) or surfing. Internet (Internet history), image storage, etc. This information is a personalized environment for users. However, the challenge here is that this data is very large, this data grows so much that it cannot be processed by one person or even a team Production source (if mobile data is enabled, the data will be generated by that user). Well, there Machine learning will emerge that will capture all that data and provide what the user wants. Core concept of Machine learning is used to predict a company's profits. This is because the decision is very difficult or because there are many factors that affect a company's revenue and its scope, predict the company's revenue from many sources such as R & D costs, management, marketing, company standards and more. This High factor that affects a company's bottom line make things unpredictable to the average person. That's why, looking at the history of the company, that is, past revenue records and management costs, this is a model that recognize patterns of factors that influence profits and help you predict profits better.

2.1 Objective:

There are many ways to predict the profits of an individual company from marketing costs to R & D costs, there are many algorithms that do this multiple regression. Get the right values using all the different techniques, such as Linear Regression, Random Forest, Decision Tree, Polynomial Regression, Artificial Neural Network some of these values are closer to a particular input data and the goals in it than other algorithms. The purpose here is to develop the best algorithms for forecasting business profits.

3. Existing Method

By using a single independent variable, such as the cost of capital of a business project Dependent variable, i.e. The company's profits from this project are roughly predicted. Linear regression uses a single independent variable to predict the value of the dependent variable creates a regression line along with the given data and uses it to predict dependent variables Regression line. There are several other techniques such as the classification trees and random forests used of many dependent variables for predicting the value of the dependent variable.

3.1 Disadvantages of existing system:

- Most of people uses single algorithm to predict target variable which is not enough.
- The data is not completely consumed by the most of the model. The main purpose is to predict the value of the dependent variable, the value of the company's profits.

4. Proposed method with Architecture

The main purpose is to predict the value of the dependent variable, the value of the company's profits. It is based on company data from the last few years. So, from all the techniques used before Profit Prediction is created by calculating the average of all these predictions of the dependent variable. As a predicted dependent variable.

4.1 Benefits of the proposed system:

- Use all the given data to predict the value of the independent variable.
- Using 5 different models to predict the result will defiantly give the edge over using single model.
- Use of Artificial Neural Network is also giving very accurate result.

4.2 Algorithms

4.2.1 Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a task of regression. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

4.2.2 Decision Tree:

Decision trees are one of the most widely used and practical approaches to supervised learning. It can be used to solve both regression and classification problems, making the latter more practical.

This is a tree-structured classifier with three types of nodes. The root node is the initial node that represents the entire sample and can be further divided into other nodes. Internal nodes represent the characteristics of the dataset, and branches represent decision rules. Finally, the leaf node represents the result. This algorithm is very helpful in solving decision-related problems.

4.2.3 Random Forest:

Random forest regression is an effective way to use the decision tree algorithm [4]. First random n data Points are selected from the training set and these points are used to build a set of decision tree models. Another set of n random data points is selected and all data points from Includes training set. Constructed from given training data after a series of decision trees [5] Random Forest regression uses all these models to predict values.

4.2.4 Polynomial Regression:

Polynomial regression is a regression algorithm that models the relationship between the dependent variable (y) and the independent variable (x) as an nth degree polynomial. The polynomial regression equation is shown below.

$y= b_0+b_1x + b_2x^2+ b_3x^3+....$

1n

This is also known as the special case of multiple regression in ML. This is to add some polynomial terms to the multiple regression equation and convert it to polynomial regression. This is a linear model with some changes to improve accuracy. The dataset used in training polynomial regression is non-linear in nature. Use a linear regression model to fit complex, non-linear functions and datasets.

4.2.5 Artificial Neural Network:

The term "artificial neural network" refers to a biologically inspired subarea of artificial intelligence modeled on the brain. Artificial neural networks are usually computer networks based on biological neural networks that build the structure of the human brain. Just as the human brain has neurons that are connected to each other, artificial neural networks have neurons that are connected to each other at different layers of the network. These neurons are called nodes.

By combining all these algorithms, we can find the range of accuracy for model.

5. Methodology:

1) **Data visualization and Pattern evaluation:**
   In this step data is load into pandas data frame and visualize with respect to the given target attribute. Data visualization is a graphical representation of information and data. By using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in your data. This blog on data visualization techniques will help you understand the detailed techniques and benefits. In the world of big data, data visualization tools and technologies are essential to analyze vast amounts of information and make data-based decisions.
2) **Feature extraction:**
   In this project we are basically using correlation for feature extraction. Correlation is a statistical measure of the amount by which two or more variables change together. Simply put, it shows how much one variable changes with respect to small changes in another. It can have positive, negative, and zero values, depending on the direction of change. A high correlation value between the dependent variable and the independent variable indicates that the independent variable is very important in determining the output.

3) **Feature scaling:**
   This is a data preprocessing step that applies to the independent variable or characteristic of the data. Basically, it helps to normalize the data within a certain range. It can also help speed up the calculation of the algorithm.

4) **Training and Cross validation:**
   Cross-validation is a better model evaluation method than residuals. The problem with residual assessment is that it doesn't show how well the learner works when asked to make new predictions for data that he hasn't seen yet. One way to solve this problem is to not use the entire dataset when training learners some data will be deleted before training begins. Once the training is complete, you can use the deleted data to test the performance of the model trained with the "new" data. This is the basic idea of the whole class of model validation methods called cross-validation.

5) **Model evaluation and selection:**
   Evaluation is constantly excellent in any subject right. In the case of device learning, it's far high-quality the practice. In this post, I will nearly cowl all of the famous in addition to not unusual place metrics used for device learning.

- Confusion Matrix
- Classification Accuracy.
- Logarithmic loss.
- Area under Curve.
- F1 score.
- Mean Absolute Error.
- Mean Squared Error.

6) Selection of model:

Base on the performance matrix and comparing the model accuracy. We will get the most accurate model accurate model for Machine Learning.

6. Implementation:

Python

Python is a commonly used high-level programming language developed by. Guido van Rossum [30], easy to interpret and read. Python has something specific. It is functional and suitable for both quantitative and analytical purposes. Data science Python is universally used, not just one. Dynamic and open-source languages are the best choice. Its rich library is also used however, manipulating the data is very easy for novice data analysts. Let's take a brief look at the Python libraries used in this task.

NumPy:

NumPy is a library of multidimensional array objects and sets of arrays. Processing routine. NumPy is used in combination with the SciPy and Matplotlib packages. this the combination is used for engineering calculations. Math and logical operations. Run using NumPy

Panda:

Pandas is a software library designed for manipulating and analyzing data. In the Python programming language. This is open-source BSD Three Clause license. It's based on the NumPy package and Data Frame Its main data structure.

Matplotlib:

Matplotlib is a Python module used to draw attractive graphs. Visual representation in data science is an important step. You will soon understand how to do it the data is divided by visual representation. There are many libraries to express. As for data, matplotlib is very popular and easy to visualize.

Scikit-learn:

Scikit-learn is a free Python library. There are multiple clustering classifications Regression algorithms such as Random Forest, DBSCAN, k-Means, Gradient, etc. Boosting, support vector machines, and programmed gradient boosting interface with NumPy and SciPy libraries.

Seaborn:

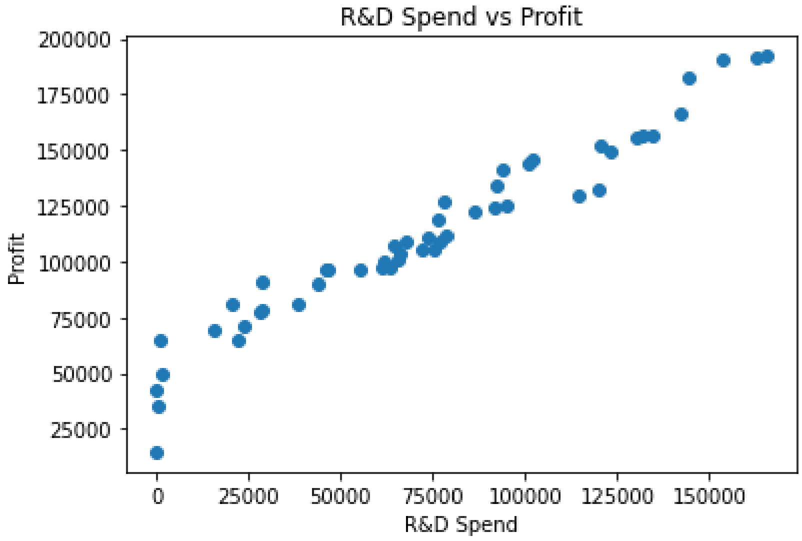
Seaborn is an open-source Python library used for statistical graphics. It offers with a dataset-oriented API for analyzing relationships between different variables a resource for selecting the color palette that is actually in the data.

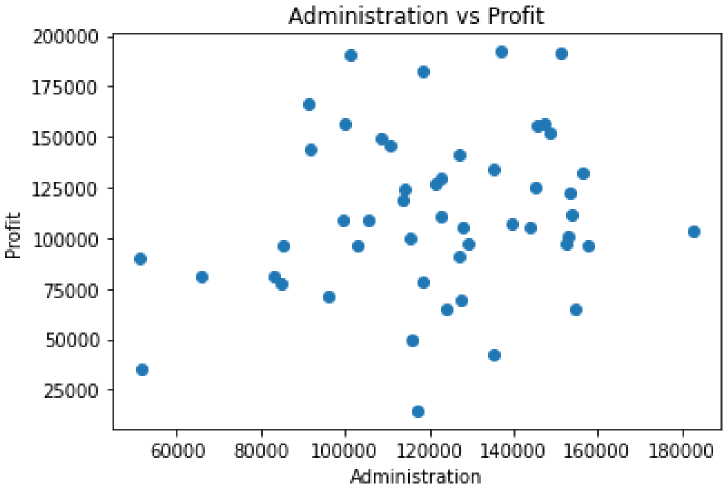
Keras:

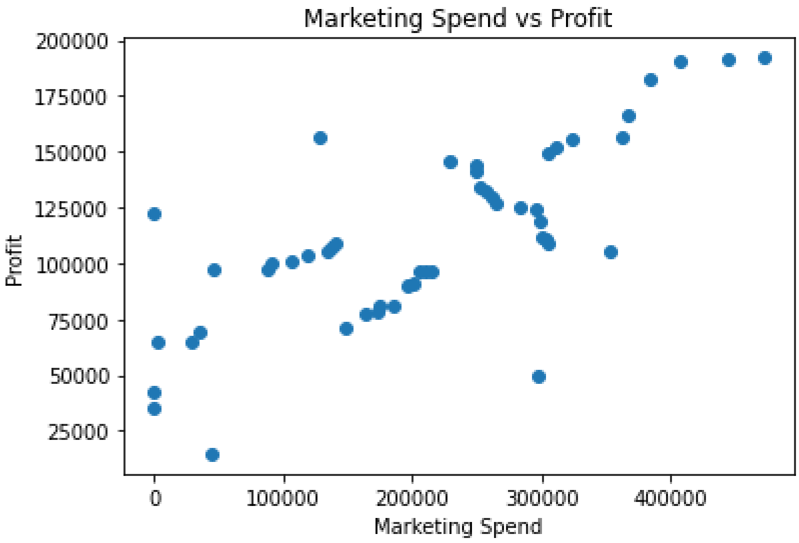
Keras is built on minimal structure and provides a clean and easy way to build deep learning models based on TensorFlow or Theano. Keras is designed to quickly define deep learning models. Well, Keras is the perfect choice for deep learning applications. Keras is built on minimal structure and provides a clean and easy way to build deep learning models based on TensorFlow or Theano. Keras is designed to quickly define deep learning models. Well, Keras is the perfect choice for deep learning applications.

6.1 Data Visualization:

Data visualization is very important technique for data science. In the given data R&D Spend, Administration Cost and Marketing Spend given attribute which is visualize against the profit. The graph of that results are as follows:







6.2 Feature Selection:

There are many different factors that can improve a machine learning model. Effective for any task. One of the feature selection methods is data correlation.

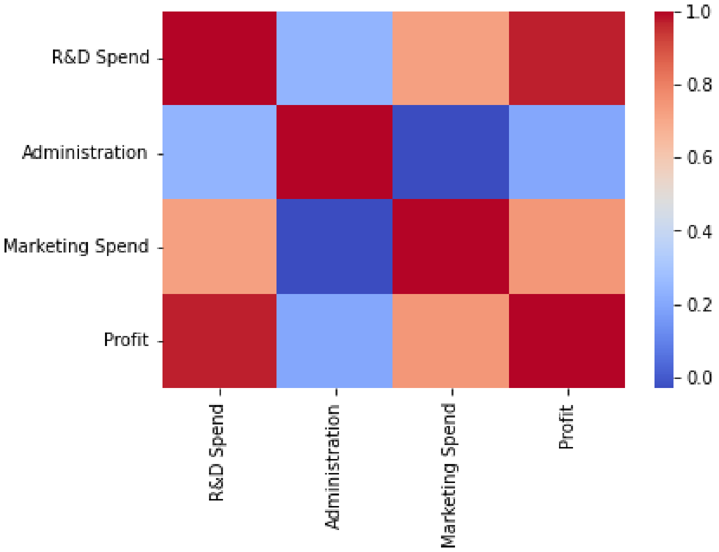
6.2.1Data Correlation Method:

Data correlation is a useful technique for predicting one attribute from another and is used as a basic quantity in many modelling techniques. For functions. As it increases, the correlation becomes positive, so it is different from other characteristics. If negative, one characteristic increase and another decreases. If not for relationships between any two attributes, it can be said that there is no correlation. If they're a linear relationship between constant variables, Pearson correlation Coefficients are used. When there is a non-linear relationship between constant variables Next, Spearman's correlation coefficient is used. Since the dataset under consideration is linear, the Pearson correlation coefficient is also linear. Used to select features in this survey. This correlation of all attributes. To improve the efficiency of machine learning models. Attributes with a negative correlation have been removed. It's a statistical measurement Linear correlation of two variables X and Y. There are values between +1 and 1. Where 1 is linear positive correlation, 0 is no linear correlation, and 1 is linear correlation. Negative correlation. The motivation for considering correlation is that people know the score One indicator can predict another that is more relevant More accurately. The more accurate the prediction, the stronger the relationship between variables.

Order in which features are correlated are as follows:

Features	Correlation
R&D Spend	0.972900
Administration	0.747766
Marketing Spend	0.200717
Profit	1.000000

The heat map for correlation between non-numerical attributes is plotted as follows:



6.3 Cross-validation:

Cross-validation (CV) is a statistical analysis technique used to evaluate the effectiveness of machine learning techniques and is a resampling technique used to do so. Algorithm validation when data is inadequate. Hierarchy is the process of rearranging data to ensure that all convolutions are in place. Representative of the whole data divided into folds can be controlled by criteria. To ensure that each fold has the same percentage of results with a particular category value. The result value of the class. This process is called stratified k-fold mutual verification. Common cross-validation techniques include K-validation, stratified K-validation, and non-excluded cross-validation. The motivation behind the 10-direction stratified cross-validation is the estimator smaller variance than single holdout set estimation. Very important when the amount of data is limited. There is a wide variety Estimating the results of different data samples or specific data partitions created Training and test set. A four-direction stratified cross-validation eliminates this variance performance is degraded by comparing more than four individual partitions estimated to be less susceptible to data splitting by averaging.

Result of Cross Validation of models are as follows:

Model	Cross Validation Accuracy
Linear Regression	0.895185
Random Forest	0.867456
Decision Tree	0.895185
Polynomial Regression	0.790378

6.4 Training the model:

For model training data is split between following order:

Training – 80%

Testing - 20%

The dependent variables and this training process depend on the algorithm: SVR, Random Forest, and more. Regression and each of these models have their own training method for a given pre-processed dataset. Because they all use a different class in Sklearn.

6.5 Testing the model:

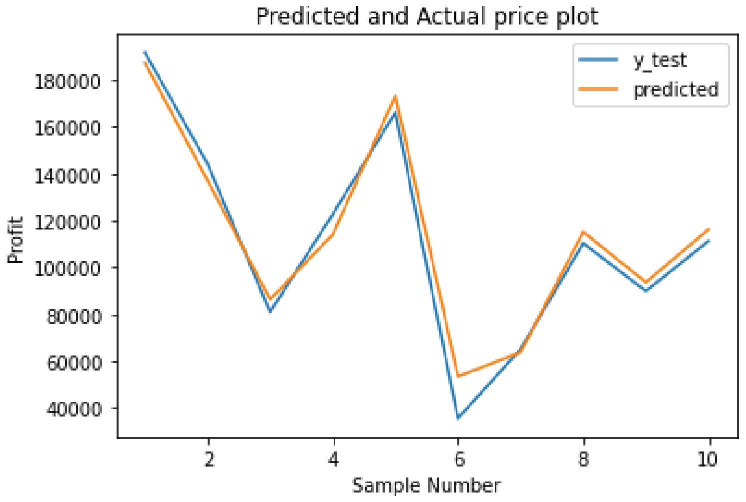
Performance Metrics:

You can use several metrics when assessing the performance of your model. that is Needed to understand how each metric measures to select an evaluation metric. Estimate the model better. The main purpose of this paper was to compare performance evaluation of machine learning techniques by evaluating all these performance metrics. Such as Accuracy score, Mean Absolute Error, and Max error.

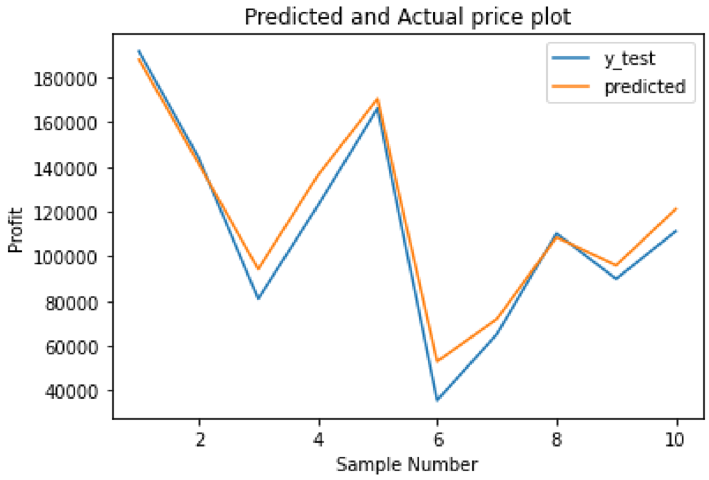
Accuracy of models are as follows:

Model	Accuracy
Linear Regression	89.5185
Random Forest	86.7456
Decision Tree	79.0378
Polynomial Regression	89.5185
ANN (Artificial Neural Network)	90.5484

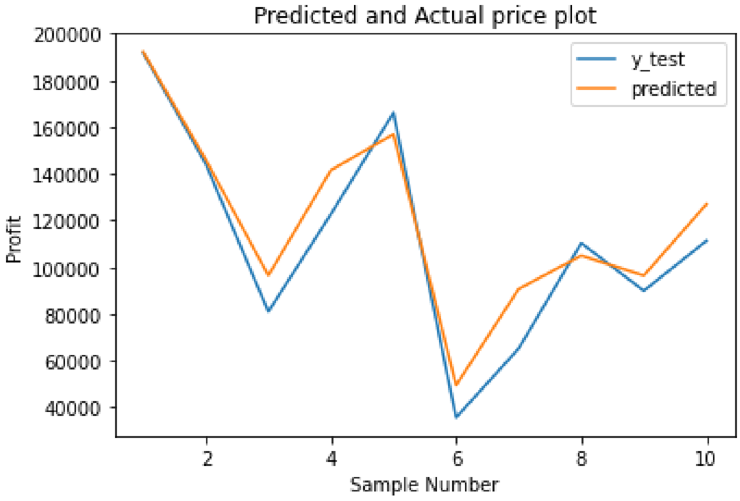
Linear Regression:



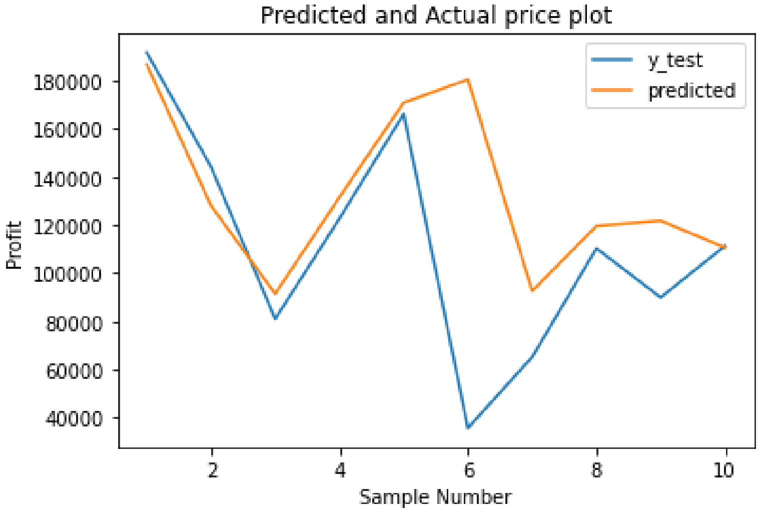
Random Forest:



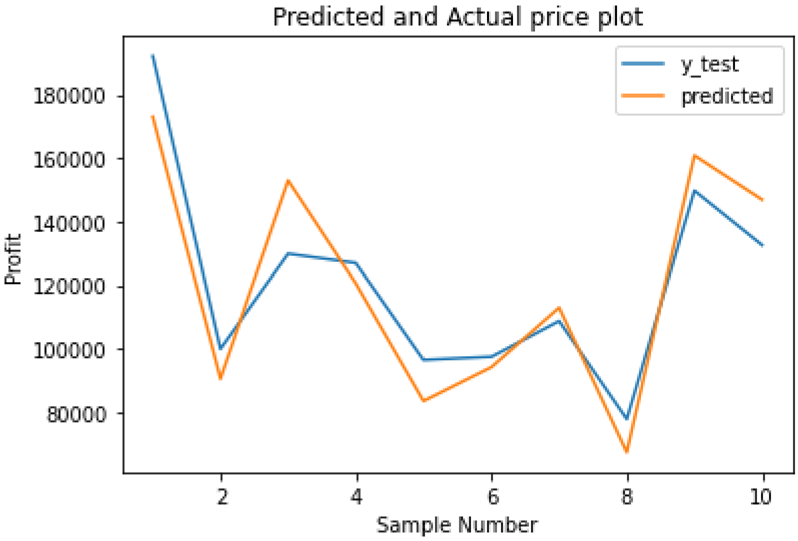
Decision Tree:



Polynomial Regression:



Artificial Neural Network (ANN):



7. Conclusion

Profit forecasts play an important role in all areas of the business. With the help of R&D, Administration, Marketing Spend helps you get the details you need about Profit. Different types of machine learning methods such as simple Linear regression, random forest regression, ANN, Polynomial Regression, Decision Tree etc. were evaluated for profit data to find key factors influencing profit. After running metrics such as accuracy, mean absolute error, maximum error. Linear regression, Polynomial Regression and ANN shows more accurate results than other algorithms. But in case of Polynomial Regression there is always chance of overfitting. ANN is good algorithm but we need more processing power and it is time consuming.

Therefore, Linear Regression proves to be a good algorithm. According to the collected data and thus fulfilling the aim of this project.