Coursework Report on Machine Learning project based on Profit Prediction of companies

Poorva Suresh Araj - Student Id: 000997514 - as0000x

Abstract

This report describes the use of machine learning model to predict future profit of companies based on the expenditures spent on their Marketing, Research and Development, and Administration. It also explains the implementation of multi-linear regression algorithm which was used to develop this model. This machine learning model uses a powerful technique which can be used to understand the factors that influence profitability of company.

1. Introduction

Machine learning helps to predict future values based on historical data by using various algorithms. It allows software applications to work on big data collected from various sources to gather insights and predictions without explicitly programmed. In this project Multi-linear regression algorithm is used to understand the relation between revenue spent for marketing, research and development of company and administration costs with the profits earned by companies. As a result it predicts the profits expected to earn by a company. Marketing plays a crucial role in company's future as it helps to expand the business to customers. While research and development department helps a company to come up with various innovative ideas which could help company in long run. The data set involved in this project includes the data of thousand companies and their expenditures and profits.

2. Multi-Linear regression Model

Linear Regression Algorithm is based on Supervised learning. Supervised learning uses a existing training data set to teach models to receive desired outputs. Linear regression can have single or multiple continuous independent variables(x) based on which they are named as simple linear regression and multiple linear regression respectively. It predicts a dependent variable (y) based on given independent variables(x).

- Following equation represents multi-variate Linear regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_p x_{ip} + \epsilon$$

where, y_i = Dependent Variable,
$x_{i1,2..n}$ = Independent Variables 1 to n
$\beta_0 = y - intercept$
$\beta_p = SlopeCoefficient$
$\epsilon = Residuals$

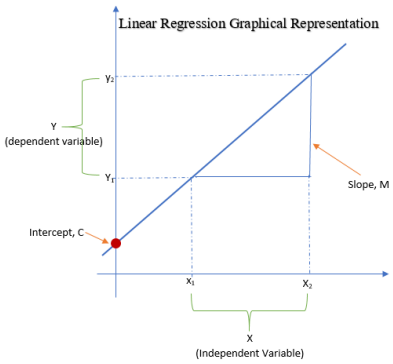

Figure 1. Linear Regression

- This Machine learning model tries to learn by getting the most accurate value of y-intercept and slope coefficient so that it can fit the data which has least square error.
- cost function or loss function allows to interpret how well the linear regression line fits the data. Linear regression model first identifies y-intercept and slope based on the data provided to the model. The predicted values of y-intercept and slope are used to plot a straight continuous line. If all sample values are on line then the cost function will have value 0 and model will have the most accuracy. The distance between the sample data point and predicted data points is called as error.The square of this error is calculated to avoid negative distance differences. Lower the value of cost function, higher is the accuracy of data prediction model.
- Gradient Descent algorithm is used to reduce the value of cost function.This algorithms optimizes the values of y-intercept β 0 and slop coefficient β P in order to find the line which could efficiently fit around the data.

3. Experiments

In order to implement the linear regression algorithm, Python version 3.10.0. Jupyter-notebook is used as an IDE.

3.1. Implementation and Experimental settings

- Libraries and Dataset Loading: : Numpy, Pandas, matplotlib, Seaborn and Sklearn are the essential libraries which are used in order to support model prediction. With the help of pandas libraries a CSV(Comma separated Values) file of sample dataset of 1000 companies is loaded and saved as the data frame. Then the dataset is divided into two main parts: Dependent(y) and independent columns(X1..Xn).

- Correlation Heatmap: Linear relationship between variables is represented by correlation matrix. It helps to understand dependence between two variables. If there is strong relationship between independent variable(y) with dependent variable(x) then it is considered as correct features to train the model. corr() method is applied on data frame to find the correlation between dependent and independent variables. Seaborn Library allows to graphically represent heatmap of correlation matrix. values of correlation are between -1 to 1. If the value is around 1 then there is strong co-relation between two variables and if the value is around -1 then the variables are inversely correlated. If the value is 0 then the variables shows no evidence of correlation and can be dropped from features list.

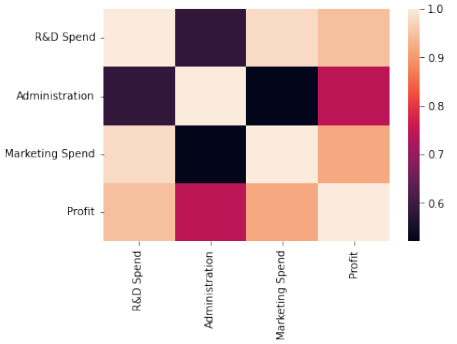

Figure 2. Correlation Heatmap

As per observations based on Figure 2 correlation between dependent variable 'profit' with independent variable 'RD spend' is around 0.9, which suggests it has strong correlation. Similarly correlation between 'profit'- 'Marketing spend' and 'profit' - 'administration' is between 0.8 to 1., which is also good enough to consider as features for prediction.

Figure 2 also proves that explanatory variables(x) are independent from each other as they show no sign of correlation and yields values under 0.7.

- Categorical Data Encoding: companies.info() command provides an output total number of columns, column names and its data types. As per the output column name 'State' has data type object which stores categorical values. Machine learning algorithm expects only numerical values as an output and categorical values cannot be interpreted by it. Hence,'state' column is converted into numerical column data type.

- Categorical Encoding involves following two techniques:

  1. Label Encoding
  2. and One-hot Encoding

- In Label Encoding integers are assigned to unique values of the column. From the dataset 'New York', 'California', and 'Florida' are unique values. Label Encoding has assigned integers to each state by alphabetical order. So California as 0, Florida as 1 and New York is assigned 2 as numerical integer.

- One hot encoding creates additional features(columns) based on the unique row values of column 'state'.So three columns named 0,1, and 2 are added in dataset. Rows will have values in the form of 0 and 1.

- But One hot encoding results in dummy variables trap. Dummy variables explains the scenario where variables correlates with each other. To avoid this scenario out of three new features, one feature is dropped out of data frame.

3.2. Linear Regression Model Fitting

After the data Pre-processing, further steps involves fitting the processed data into the linear regression model. Scikit-learn library provides support to develop a machine learning model and to fit our dataset in various algorithms simply by using various packages.

1. sklearn.model_selection(train_test_split):
   This package divides the dataset into two parts viz. training(X_train, y_train) and testing (X_test, y_test) dataset. The training dataset is feed to machine learning model to learn through the data based on algorithm logic. Testing dataset is used as an input for making predictions and provide output. This output(y-pred) is afterwards compared with y-test(testing data of dependent variable) in order to understand the accuracy of model. For this report, 20% data is kept as testing data, with 'shuffle =True' and 'random_ state=0'.
2. Further sklearn.linear_model (LinearRegression):
   This package fit the training dataset which is

X_train and y_train into the linear regression model. As linear regression is a supervised learning algorithm, X_train and y_train data is provided so that machine can learn with the help of it.

3. regressor_model.predict() function predicts the dependent values (y_pred) of input data(X_test). This command provides the output of predicted values which is 'Profit' based on testing dataset containing features RD Spend, Administration, Marketing Spend, and States.

3.3. Evaluation Criteria

- The regression coefficients of each feature columns which are calculated by the model can be obtained by command: regressor_model.coef_ and y-intercept can be obtained by command regressor_model.intercept_.
- Mean square error is the sum square of differences between actual values and predicted values. It is given by following equation:

$$\frac{1}{2} \cdot \frac{1}{m} \sum_{i=0}^{m-1} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2. \quad (1)$$

- R^2 or R-square also called as coefficient of Determination.It is used to evaluate the performance of linear regression model.It takes values only between 0 to 1.

$$R^2 = \frac{sumofsquaresoftheresidualerrors}{totalsumoferrors}$$

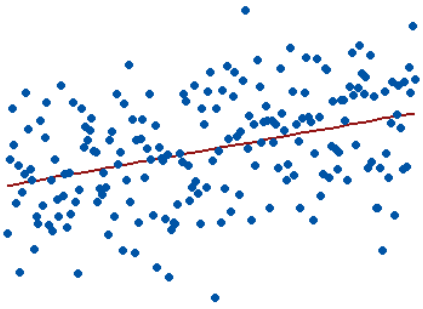

Figure 3. smaller R-Squared value

As per observations based on Figure 3 R-squared value is smaller if the data points are spread away from regression line. while R-squared value is larger if the data points are closer to regression line as per figure 4

3.4. Results

The values of regression coefficients are: -8.80536598e+02, -6.98169073e+02, 5.25845857e-01, 8.44390881e-01, 1.07574255e-01 of respective features. The value of y-intercept is -51035.2297240225.

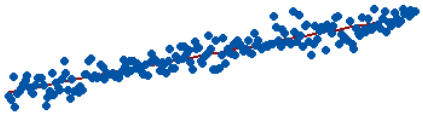

Figure 4. larger R-squared value

Based on the values of regressor coefficient and intercept the multi linear regression equation becomes

$$y_i = -51035.2297 + (-8.805e + 02x_{i1})$$

$$+(-6.98e{+}02x_{i2}){+}(5.25e{-}01x_{i3}){+}(8.44e{-}01x_{ip}){+}\epsilon$$

The performance metrics of the model are given by following commands: For r square score: r2_score(y_test, y_pred) For mean squared error: mean_squared_error(y_pred, y_test and For mean squared log error is mean_squared_log_error(y_test, y_pred).

R2 score is normalized by the variance of the target values. High values of R-squared explains there is less difference between the observed data and fitted values. As the model implemented of this report has given 0.9112 value, it refers that 91% data fit the regression model

The value of mean square error of our model is 192148061.81, which is quite large.But as the mean square error is affected by the scale of the target values(profit). If the value is compared with the mean of profit which is 119546.164656 then the output is justified.

3.5. Discussion

As the dataset has continuous numerical values Linear regression machine learning algorithm expected to be suitable for modelling. Dataset included multiple features hence it is considered multivariate-linear regression and its corresponding equation was used to analyze the model throughout.Other than evaluation metrics such as R squared and mean square errors, Various manual methods are used to understand the dataset for evaluation. It Includes .describe() which obtains mean, standard deviation, len()length and .info() for data types, and Plotting the values for visual representations of data in order to analyze it.

4. Conclusion

As per the results of evaluation metrics R-squared and mean-squared errors, we can conclude that the model was able to predict the future values of data efficiently with maximum accuracy score.