**Prediction of Profit**

**With**

**Machine Learning Model**

***Name: Kartik Ningaraj Yanigar***

***E-mail:kartik.nyanigar.mec20@itbhu.ac.in***

***Ph No: 8867021358***

***Project for Internship at Exposys Data Labs***

ABSTRACT

A machine learning model is a computer program trained to recognize specific patterns. You teach a model on a set of data and give it an algorithm to use to reason about and learn from that data.

Once the model has been trained, you can use it to reason over data it hasn't seen before and make predictions about it. Here this concept has been used to predict the profits of a company based on some already given features.

Table of Contents

1. Introduction
2. Existing method
3. Proposed Method with Architecture
4. Methodology
5. Implementation
6. Conclusion

Introduction

A company should always create an attainable goal; otherwise, people will not be able to perform to their total capacity if they believe the company's goal is unattainable. Profit forecasting for a specific period is similar to goal-setting. If you know how much profit you can earn with the amount of R&D and marketing you undertake, you can generate more than the anticipated profit as long as the predicted value is met.

**What role does this model play in this situation?**

This machine learning model will be extremely useful in situations where we need to locate a profit based on how much we spend in and for the market. In a nutshell, this machine learning model will assist in determining profit based on the amount spent from the dataset of 50 businesses.

About 50 Companies dataset:

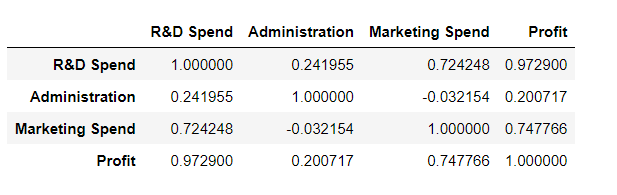
This particular dataset holds data from **50 Companies.**

The features in this dataset are**R&D spending, Administration Spending, Marketing Spending,**while the target variable is:**Profit.**

1. **1. R&D spending:**The amount which companies are spending on Research and development.
2. **2. Administration spending:**The amount which companies are spending on the Administration panel.
3. **3. Marketing spending:**The amount which companies are spending on marketing strategies.
4. **5. Profit:**How much profit that particular companies are making.

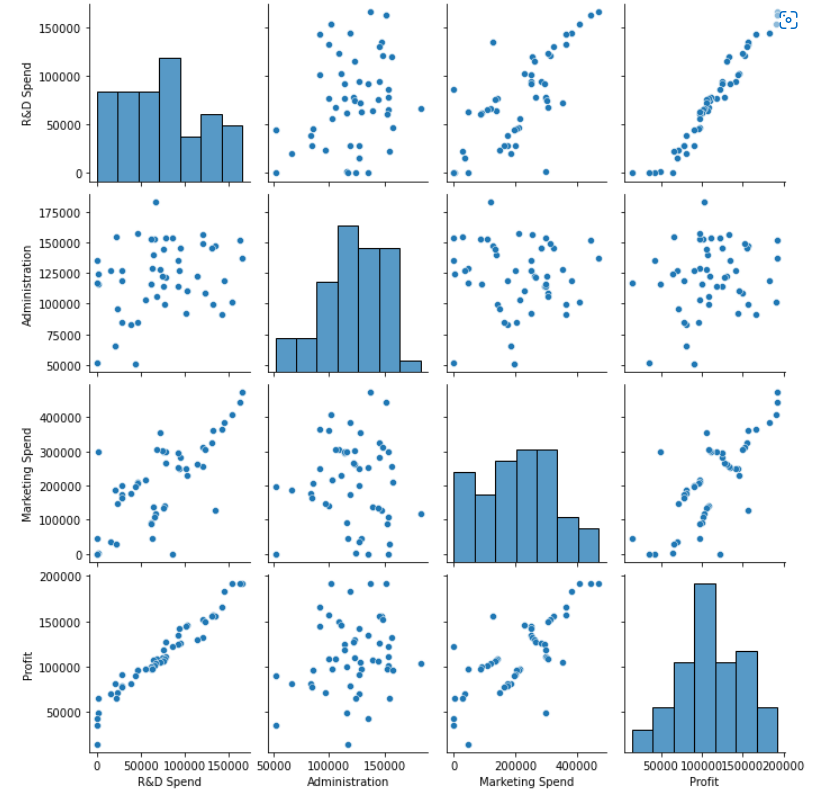
Existing Method

Exisitng dataset is a simple dataset with continuous variables. Thus simple regression algorithms are employed like Linear, Ridge, Lasso Rgression.



We can see that all three columns have a direct relationship with the profit, which is our target variable. Administration does not have good correlation with profit which proves that it vaguely determines Profit.

The Pairplot of Data gives:

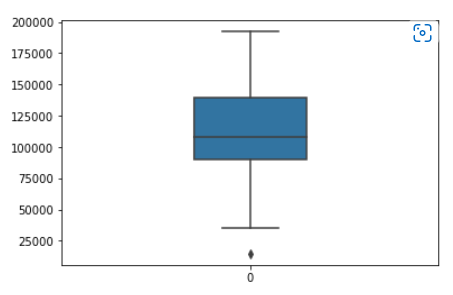


1. Research and development are directly proportional to the profit.

2. The marketing spend seems to be directly proportional with the profit.

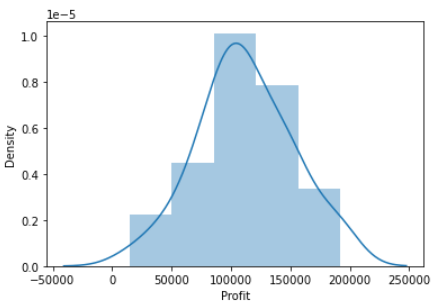
3. There is no relationship between the second column and profit i.e. our target column

Checking for outliers in profit:



Since the dataset is small the single outlier wont have a significant effect.

Profit Distribution:

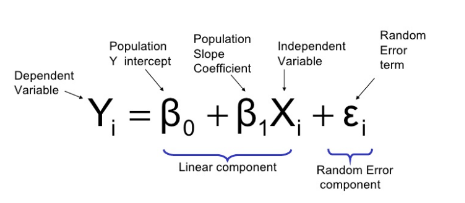


Proposed method with Architecture

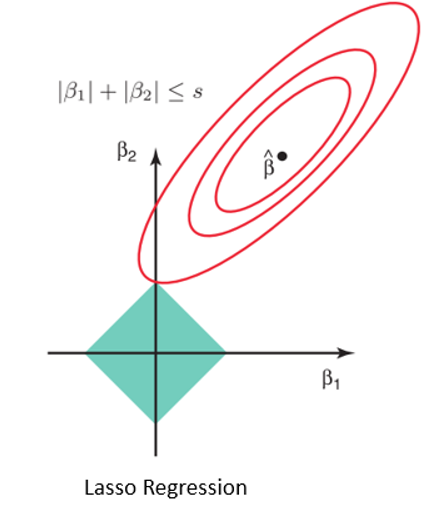
Thus according to industrial experts Multiple Linear Regression, Support Vector Regression etc. is recommended. But the type of algorithms used depends on the dataset.

In this project Lasso Regression is proposed in which the model is penalized for the sum of absolute values of the weights.

Basic equation:



And under Lasso Regression:



Methodology

Simple Linear Regression is a linear regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.

Example : Salary = a0 + a1\*Experience ( y = a0 + a1x form ).

Ridge Regression or L2 Regularisation avoids problem with simple linear regression by regularisation of coefficients.

Lasso Regression enhances the performance of linear regression by regularisation and shrinkage of the coefficients.

Support Vector Regression is a regression model in which we try to fit the error in a certain threshold (unlike minimizing the error rate we were doing in the previous cases). SVR can work for linear as well as non-linear problems depending on the kernel we choose. There is an implicit relationship between the variables, unlike the previous models, where the relationship was defined explicitly by an equation (coefficients are sufficient to balance the scale of variables). Therefore, feature scaling is required here.

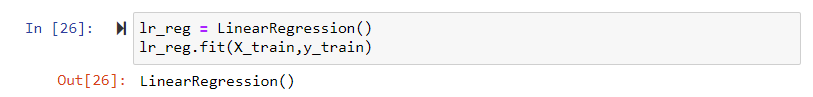
Decision Tree Regression builds a regression model in the form of a tree structure. As the dataset is broken down into smaller subsets, an associated decision tree is built incrementally. For a point in the test set, we predict the value using the decision tree constructed

Random Forest Regression – In this, we take k data points out of the training set and build a decision tree. We repeat this for different sets of k points. We have to decide the number of decision trees to be built in the above manner. Let the number of trees constructed be n. We predict the value using all n trees and take their average to get the final predicted value for a point in the test set.

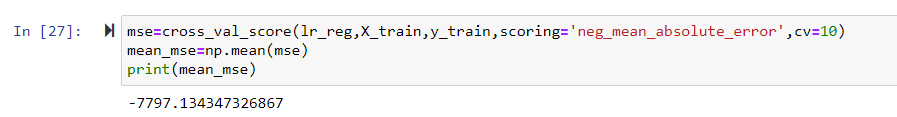
Implementation

1. **Linear Regression-**

Linear Regression simply fits a line in data with least squared error.



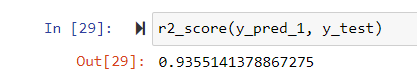
And cross validating:



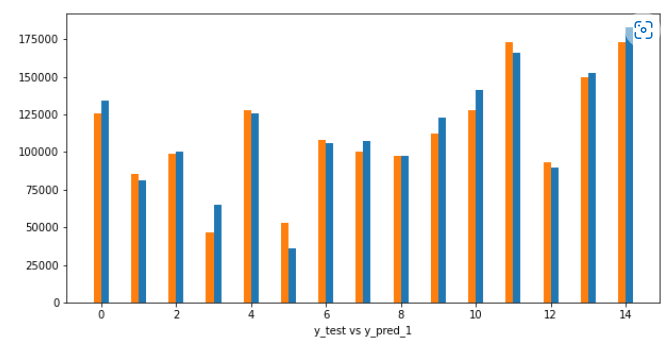
Predicting the test target variables:



And then R squared value:



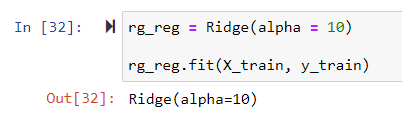
Graph of target variable and predicted values:



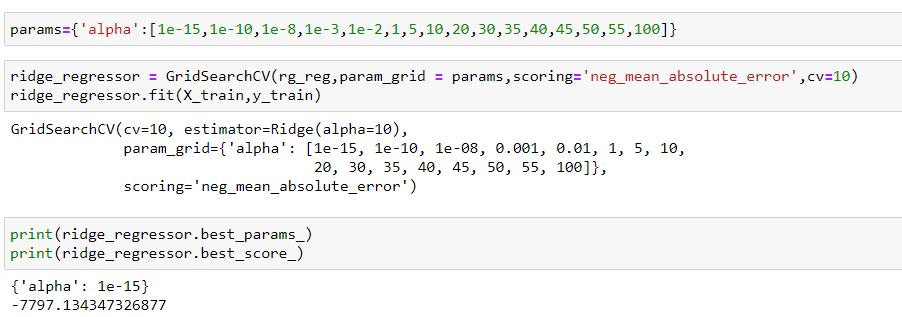
But the coefficients can be regularised and shrunk to perform better on the test data.

1. **Ridge Regression-**

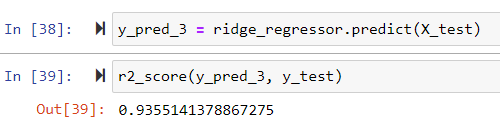
In Ridge Regression the coefficients are regularized to perform better on test data.



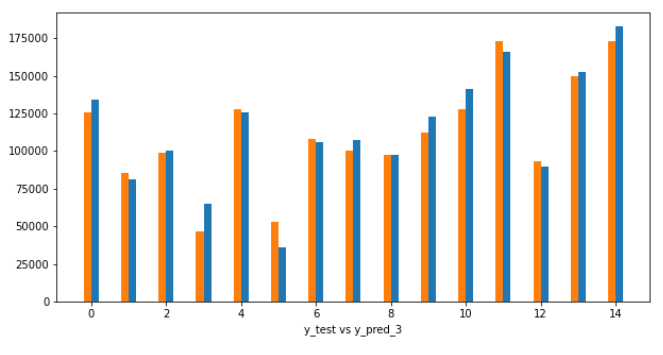
Hyperparameter tuning to achieve to get best results:



Predicting and R squared value:



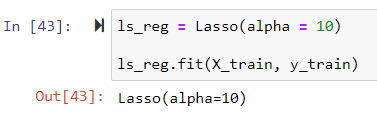
Graph of Test variable and predicted values:



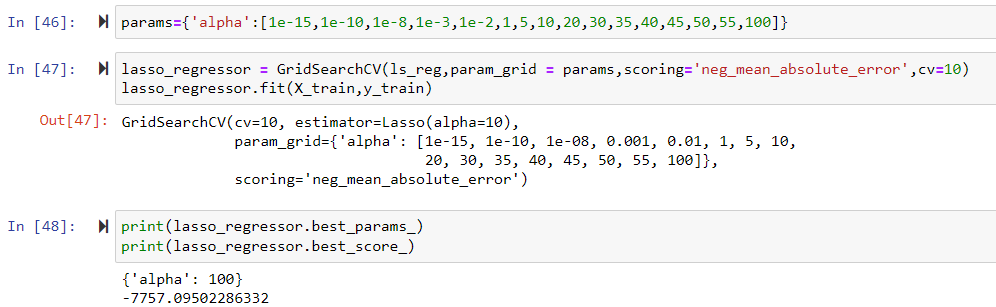
The performance can be bettered by shrinking the coefficients.

1. **Lasso Regression-**

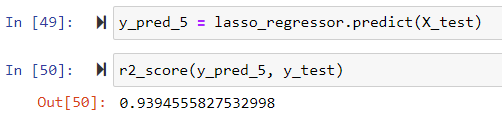
In Lasso Regression the coefficients are regularised and shrunk to achieve better performance with test data.



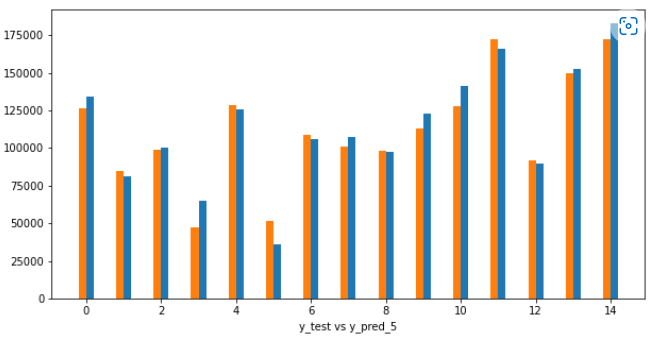
Hyper parameter Tuning:



Predicting the target values and R squared value:

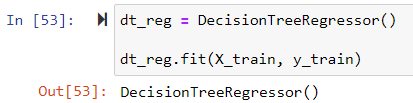


Graph of Test Variable and predicted values:

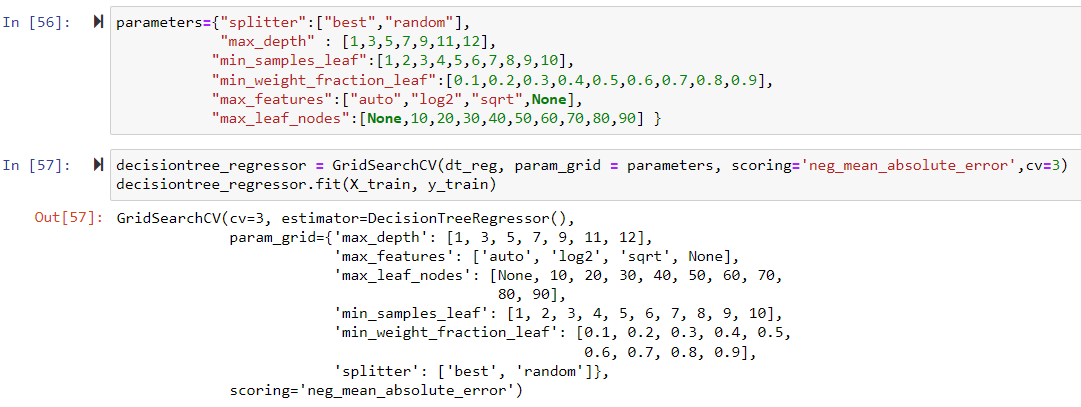


1. **Decision Tree Regression-**

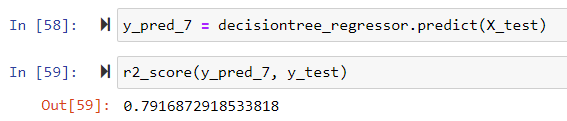
In this algorithm tree structured classifiers are used to achieve regression.



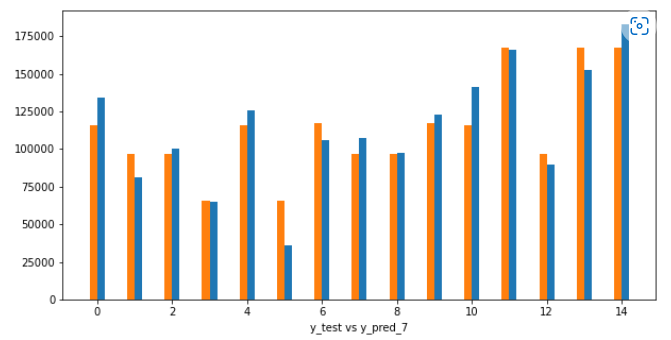
Hyper Parameter Tuning:



Predicting and R squared values:

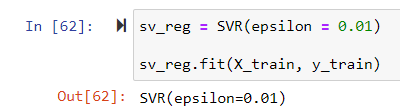


Graph of Test variables and predicted values:

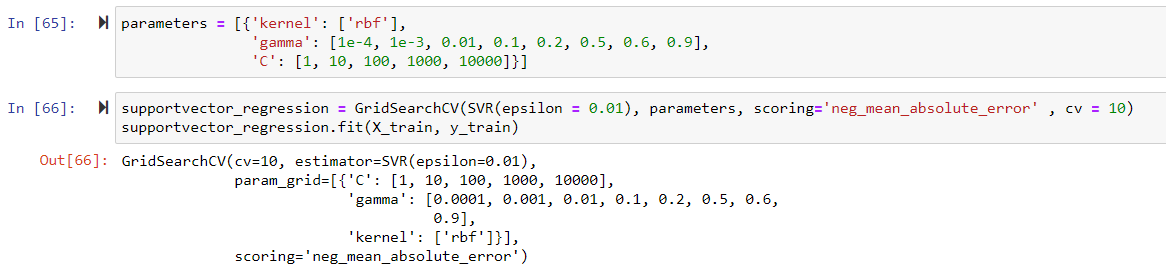


1. **Support Vector Regression-**

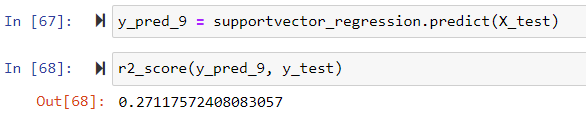
In this algorithm we try to fit a best fit hyperplane in the feature space.



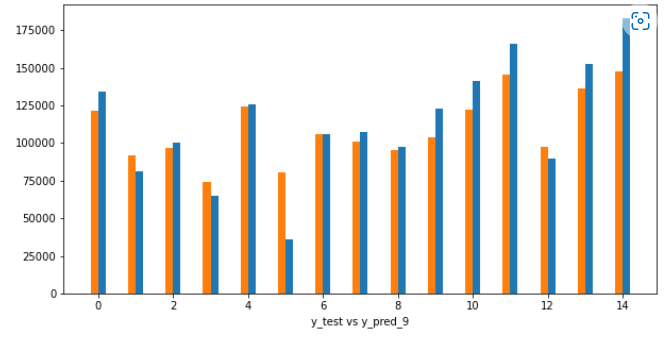
Hyper Parameter Tuning:



Predicting and R squared values:

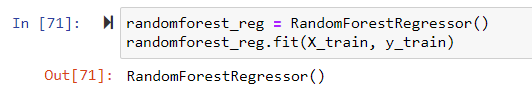


Graph of Test variables and predicted values:

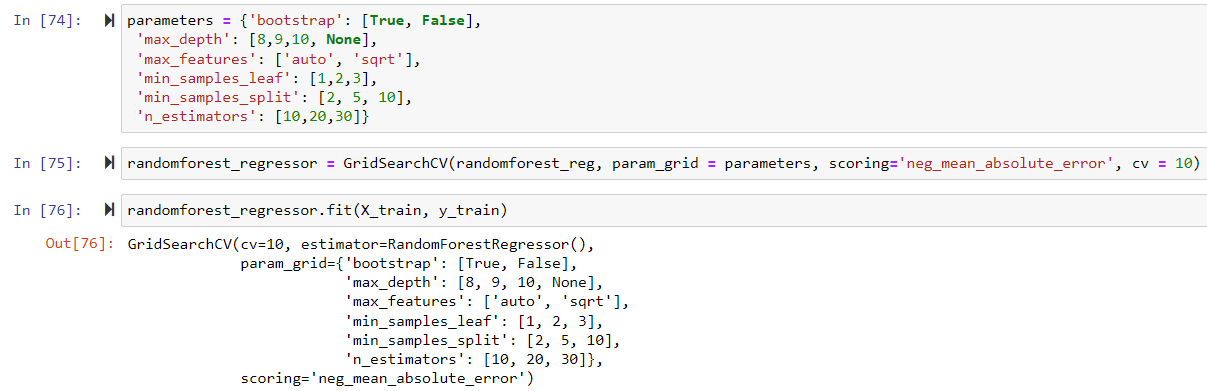


1. **Random Forest Regressor-**

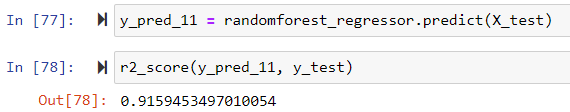
In this algorithm we ensemble learning method for regression.



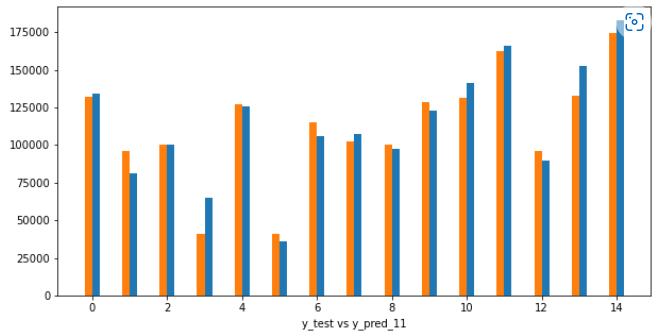
Hyper Parameter Tuning:



Predicting and R squared values:

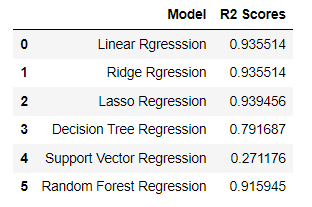


Graph of Test variables and predicted values:



**Conclusion**

Thus by enlisting each model considered and the corresponding R squared scored.



Thus Lasso Regression is the best model to be chosen for Profit Prediction according to data by given 50 Companies.

-Thank You-