Profit Prediction of 50 Startup Companies

Name: Brinta Debnath

ABSTRACT

In this project we tried to predict the profit of Startup companies depending on some past data. So here at first, we got a raw data, and we cleaned the data. The data contains 50 rows and 4 columns. We checked if there are any null values in the data then we renamed the columns like R&D Spend as RDS, Administration as ADMS, Marketing spent as MKTS. Then we checked if there are any duplicate values in the data. We checked the correlation of the features and plotted a pair plot. The we used the OLS model for our prediction. For testing the goodness of the model, we calculated the t-value and p-value of every feature and also for checking the accuracy we get the r-square and adjusted r-square value and found the model accuracy is 94.75%. Then we fitted simple linear regression and multiple linear regression. Then checked multicollinearity and the normality of the data. After that we again used the multiple linear regression model again and we got better accuracy which is 96.13%.

INTRODUCTION

It is an era of startup. Young people are coming forward with innovative ideas which can make a big change in the society. Some of the startups became so big that they make huge amount of profit. And to make a business they need investment. we can say that startups pipeline operates on the same principles which are like other MNCs the major difference between both is that on the one hand startups work to make products that are beneficial for the customers on a small scale while other established companies do that work on a large scale by re-doing something which is already being done.

The competition goal is to predict the profit of startup profit on the bases of data provided which are on the bases of Research and Development Spend (R&D Spend), Administration Spend and Marketing Spend. We use multiple regression in this model because we must predict profit (dependent variable) on bases of multiple field (independent variables) rather than one field just like we done in Simple Linear Regression. This model can help those people who want to invest in startup company by analysing profit of the company.

Methodology

A dataset was collected containing information on startups such as R&D Spend, Administration Cost, Marketing Spend, and Profit. The data was analysed using Multiple Linear Regression. The linear regression model was trained on the training data to fit the model to the data. The model was then used to predict the profit of startups given their R&D Spend, Administration Cost, and Marketing Spend.

DATA COLLECTION:

Data is collected from the given link.

LINK: https://drive.google.com/file/d/1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF/view

DATA DESCRIPTION:

This dataset holds data from **50 startups.** The features in this dataset are **R&D spending, Administration Spending and Marketing Spending** while the target variable is: **Profit.**

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

OLS

Estimating the coefficient in a linear regression equation is a common statistical task, and Ordinary Least Squares regression (OLS) is a popular method for doing so. One popular method for doing this is the Ordinary Least Squares (OLS) method of linear regression. The strategy is dependent on reducing the squared difference between the observed and predicted values. This OLS Estimator Works Well. The ideal estimator would be objective and have the lowest possible variance (efficient). Given that OLS estimators have the smallest variance of any linear and unbiased estimator, they are the most effective estimators.

R-square

The R-squared statistic indicates the degree to which observed values are consistent with the regression line. In the context of multiple regression, it is referred to as the coefficient of multiple determination. R-squared is a simple measure of how well a linear model can account for variance in a response variable.

But if you want to know if the coefficient estimates and predictions are biased, you cannot just look at the R-squared value; you must analyse the residual plots. The R-squared value does not reveal whether a regression model is suitable. A high R-squared value indicates that the model does not adequately fit the data, while a low value indicates that the model is accurate.

Adjusted R-square

Regression models with varying numbers of predictors can be compared in terms of their explanatory power using the adjusted R-squared statistic.

Let's say we want to see how well a five-predictor model does in comparison to a single-predictor model, and so we calculate their R squared. Is it indicative of how much better the five-factor model is that it has a larger R-squared value? Or is there a correlation between the number of variables and the R-squared value? All you have to do is check the adjusted R-squared values. In order to account for the fact that more predictors means a more complex model, R-squared has been tweaked to produce the adjusted R-squared. Only if the addition of the new term improves the model beyond what would be expected by chance does the adjusted R-squared value rise. It goes down when a predictor boosts the model by less than would be expected by chance. While typically not the case, the adjusted R-squared can be negative. Consistently it is less than the R-squared.

VIF

To quantify the extent to which one independent variable's behaviour (variance) is influenced by its interaction/correlation with the other independent variables, the variance inflation factor (VIF) was developed. Using variance inflation factors, one can rapidly quantify a variable's impact on the overall regression standard error.

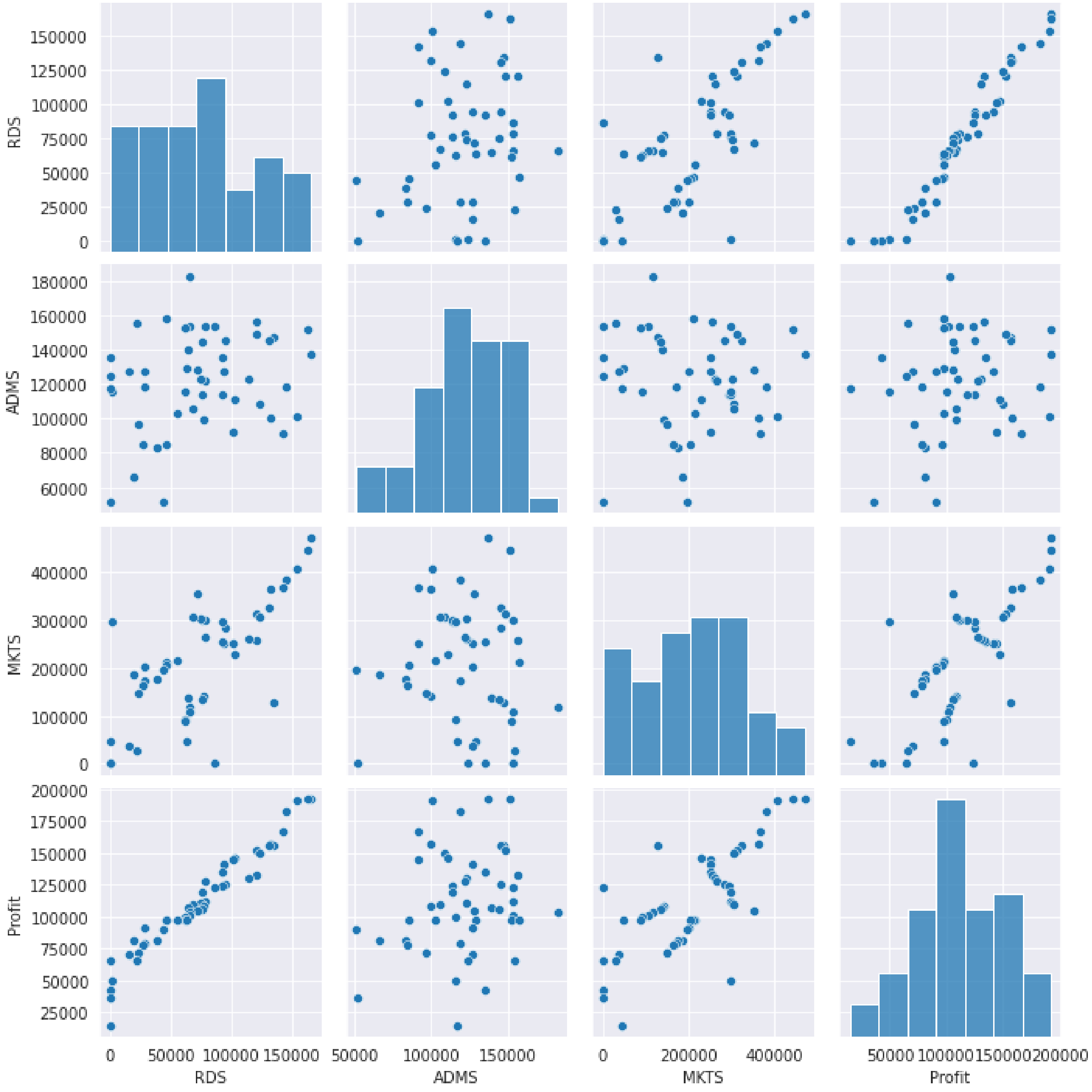
When conducting an ordinary least square (OLS) regression analysis, the variance inflation factor (VIF) can be used to gauge the degree of multicollinearity present in the data. Type II error and variance both increase due to multicollinearity. To put it another way, it makes a variable's coefficient consistent but unreliable. By default, the VIF threshold is set to 5, so only variables with a VIF of 5 or lower will be considered for model inclusion. If the VIF is less than 5, then the predictor has a low correlation with other predictors. VIF values above 10 indicate an excessively high correlation between model predictors, while values between 5 and 10 indicate a moderate correlation.

Q-Q Plot

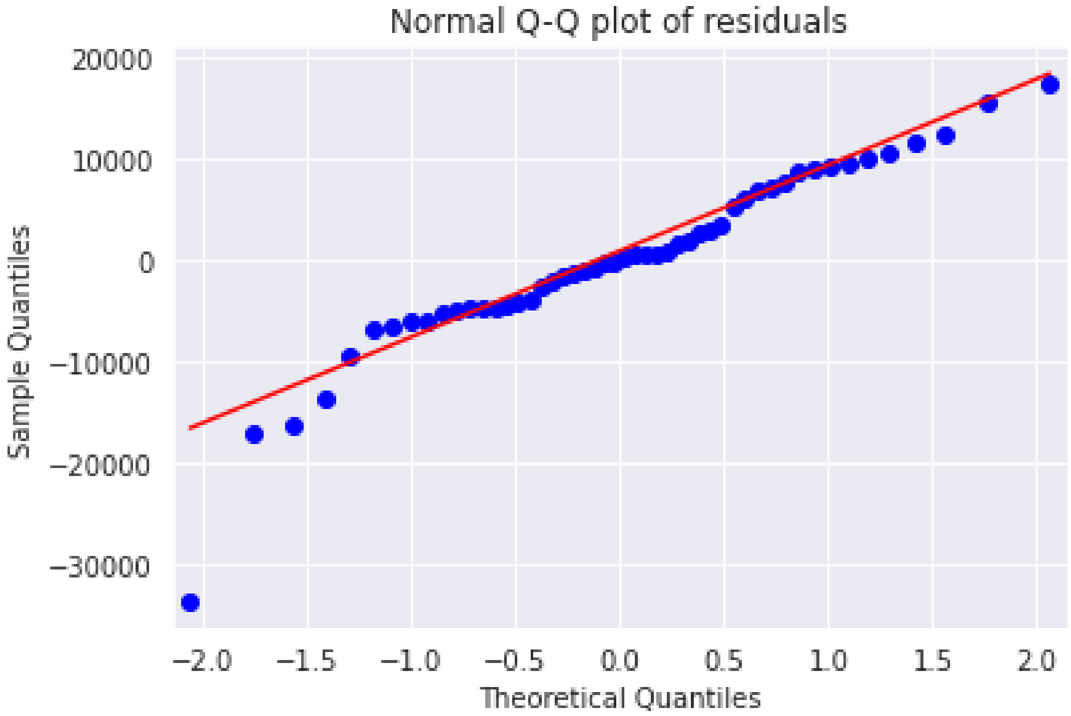
Plotting two quantiles against one another creates what are known as Q Q Plots. A quantile is a percentage below which a set of values falls. The median, for instance, is a quantile where half of the data lies below it and half lies above it. Q-Q plots are used to determine whether two data sets are drawn from the same distribution. If the two data sets share a common distribution, the points will cluster around a line drawn at a 45-degree angle on the Q-Q plot.

❖ Data Visualization

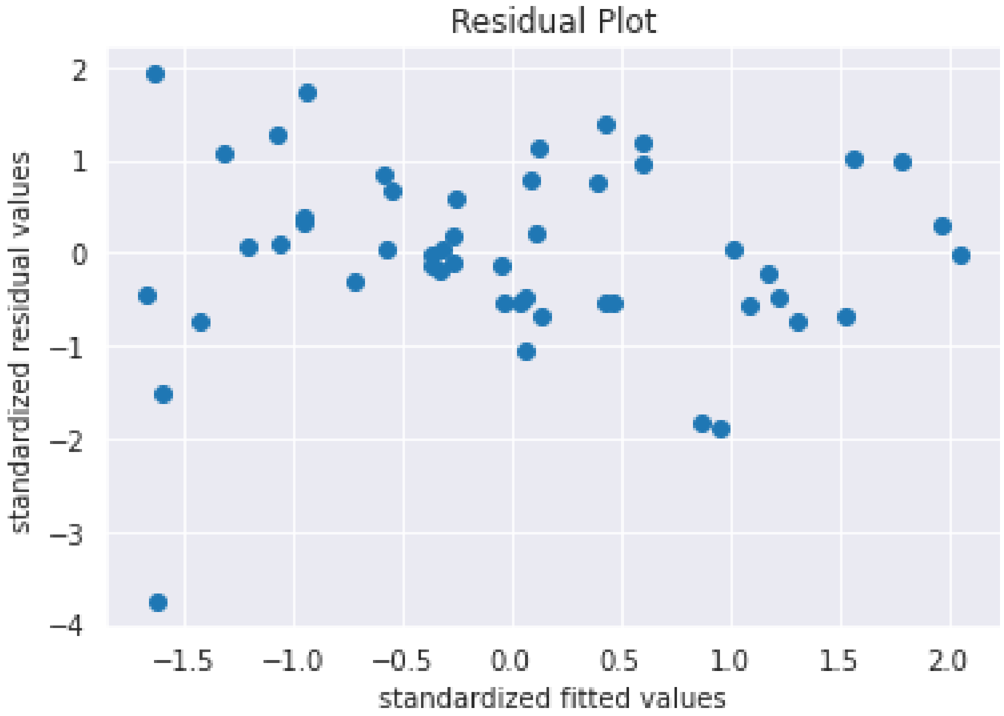
Pair Plot



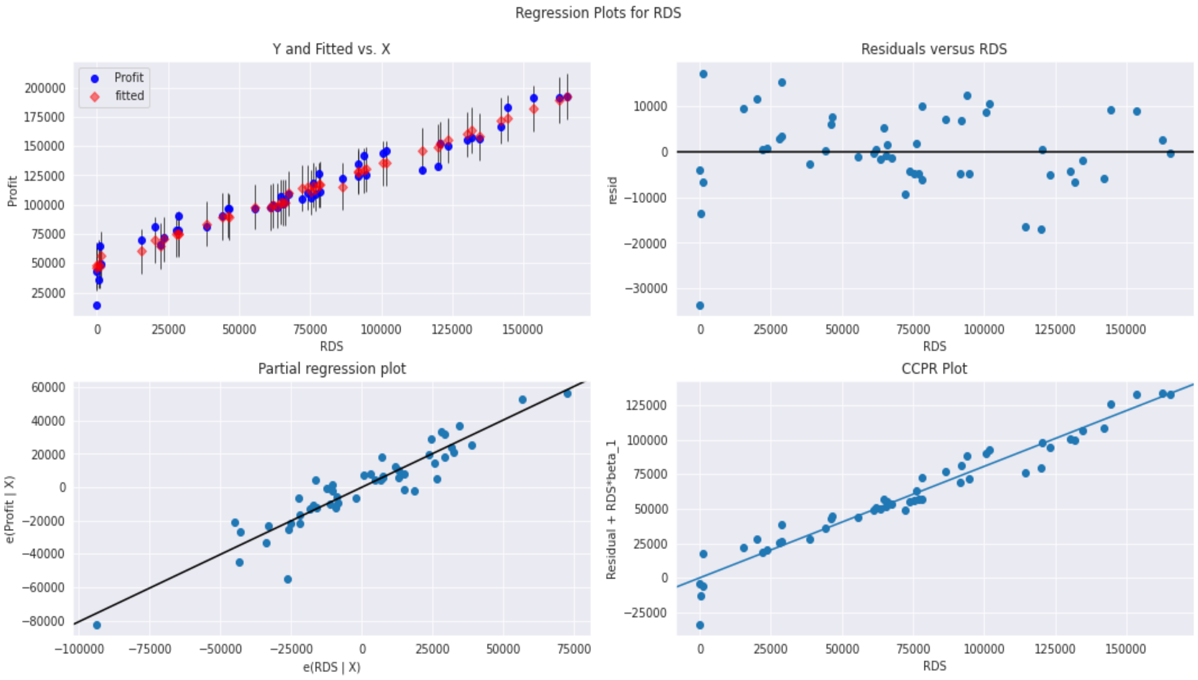
Q-Q Plot



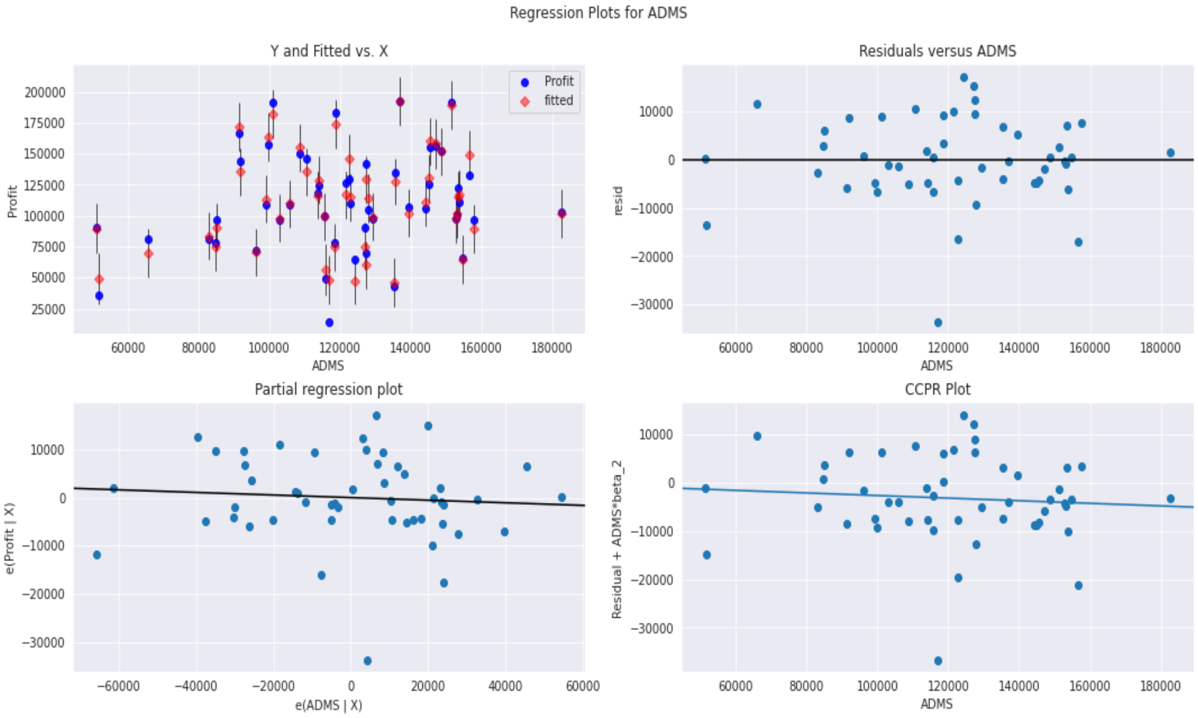
Residual Plot



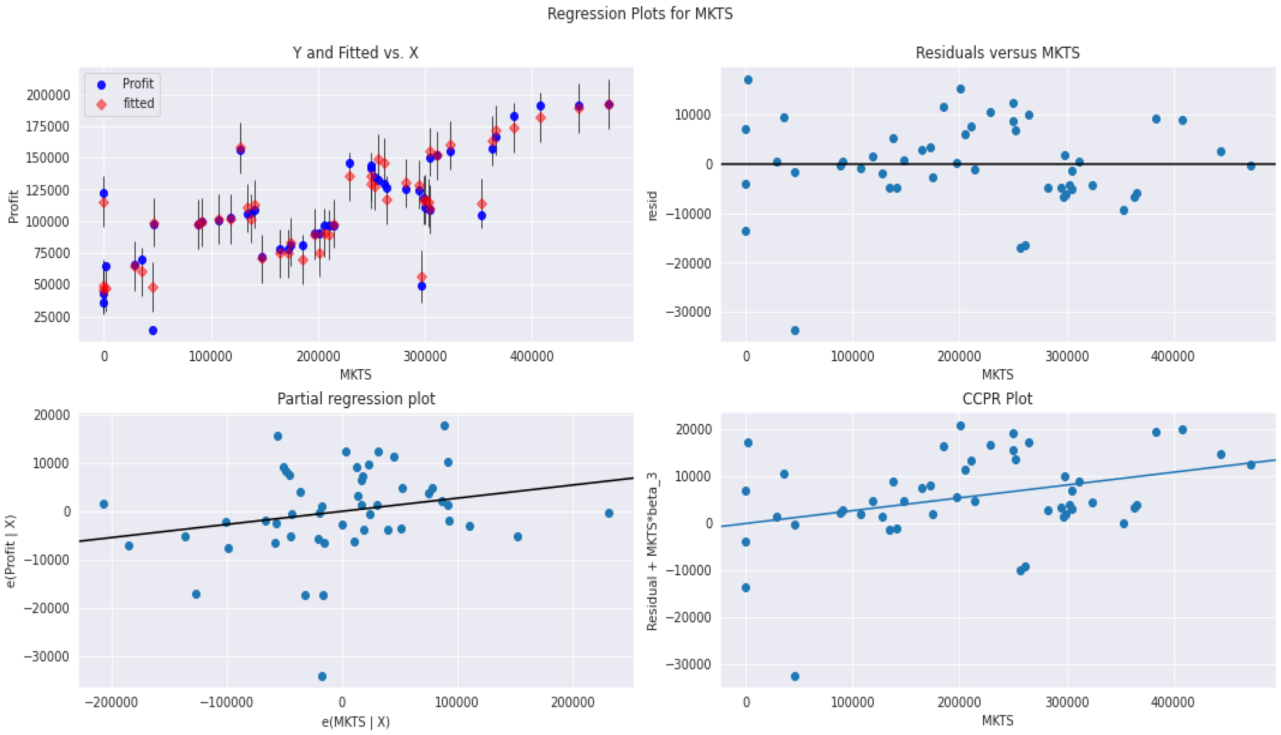
Regression Plots for RDS



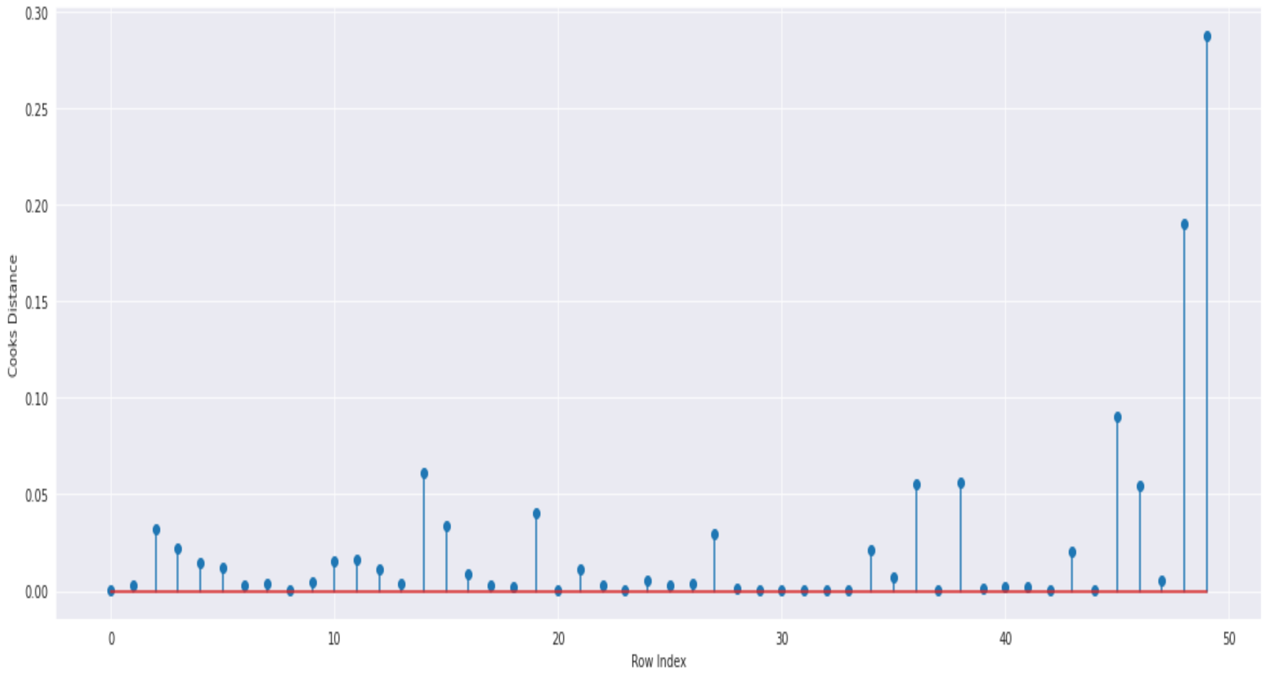
Regression Plots for ADMS



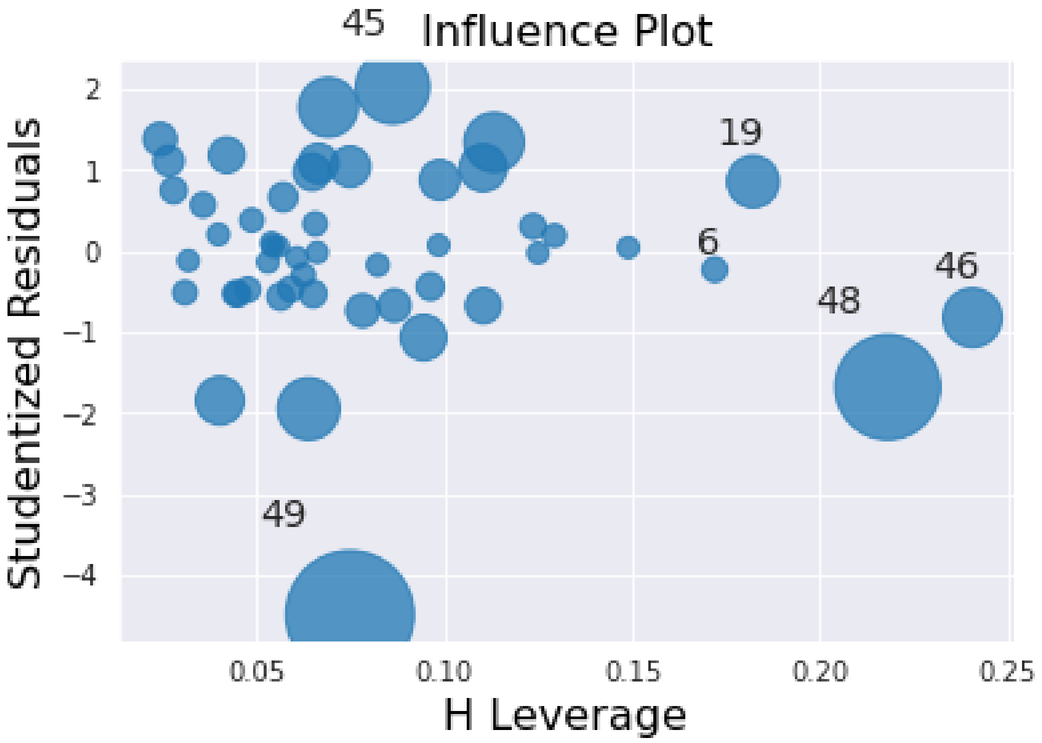
Regression Plots for MKTS



Stem Plot



Influence Plot



Conclusion

In this project we wanted to create a model to predict the profit of 50 startup companies. At first model the model accuracy was 94%. Then we checked the normality, multicollinearity and the high leverage values form the data and built a new model. And this time the accuracy is 96%. So we got our best model to predict the profit of this 50 companies.