Report On

Predict Profit Value of the Company

By,

Samarth Pandey

EXPOSYS DATA LABS

Data Science Internship

Table of Contents

1.) Abstract: -

A company should always set a goal that should be achievable, otherwise, employees will not be able to work to their best potential if they find that the goal set by the company is unachievable. The task of profit prediction for a particular period is the same as setting goals. If you know how much profit you can make with the amount of R&D and marketing you do, then a business can make more than the predicted profit provided the predicted value is achievable.

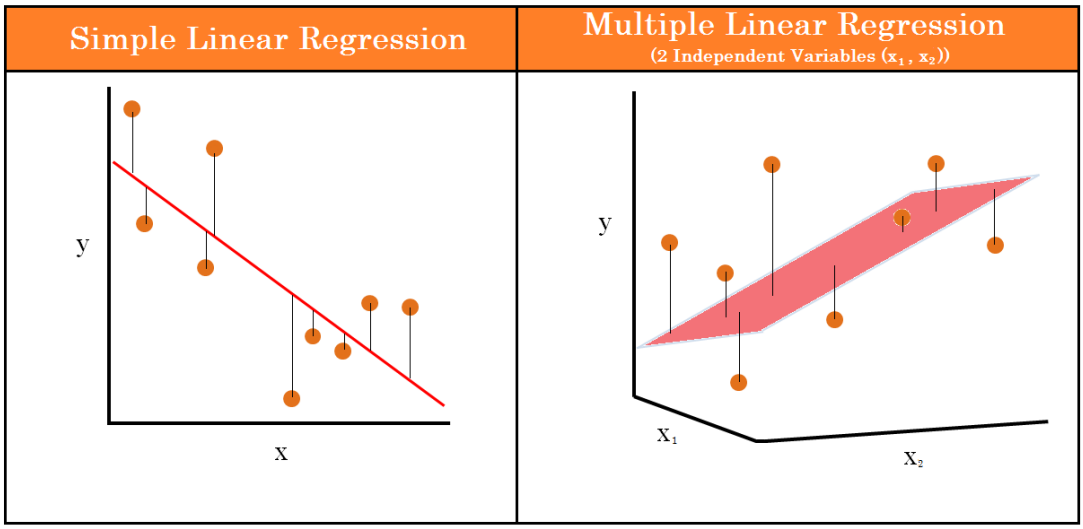
2.) Introduction: -

The profit earned by a company for a particular period depends on several factors like how much time and money a company spends on R&D, marketing and many more. 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. The task of predicting profit is an important task for every business to set an achievable goal. For example, if the business spends $500 on marketing, it can't expect a profit of $20,000. Likewise, there are many other factors on which the profit of a business depends.

3.) Comparison between Linear and Multiple Regression: -

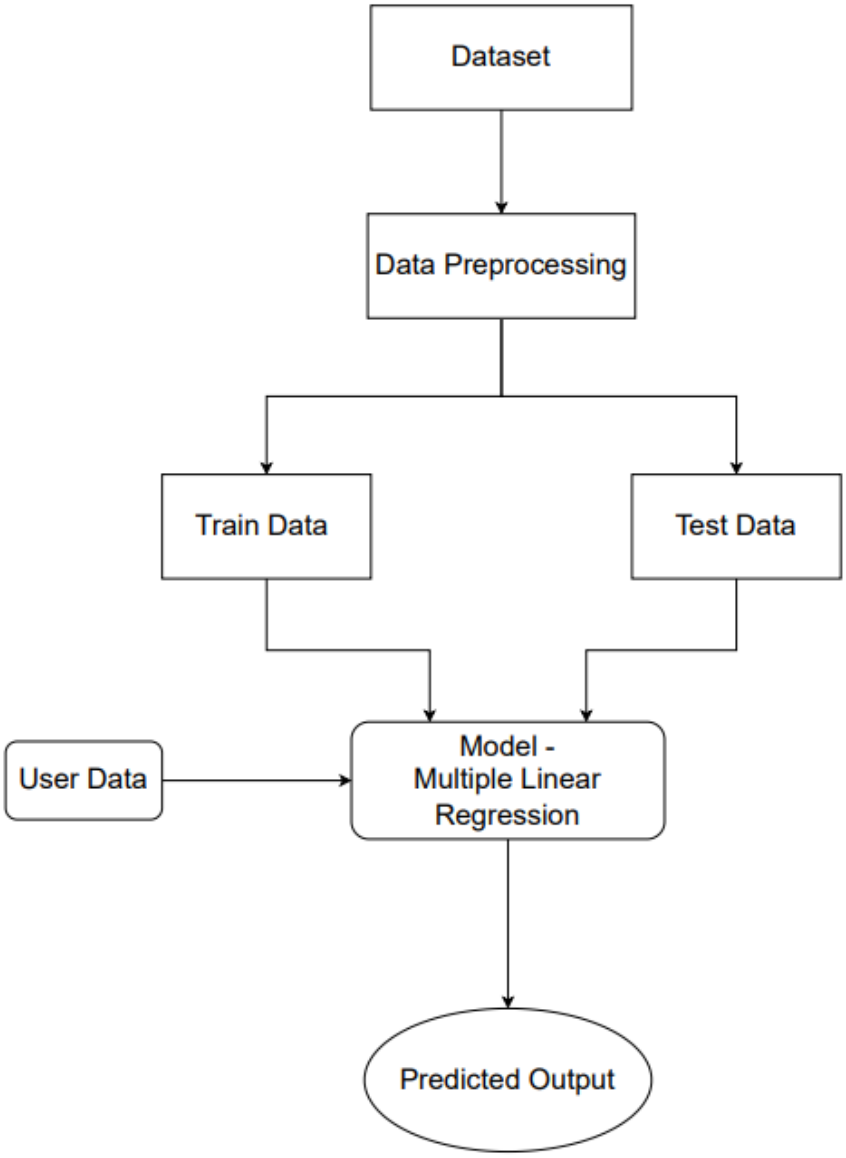
The major difference between linear regression and multiple linear regression is that in linear regression there is only one independent variable while when we check out, Multiple linear regression there is more than one independent variable. Let us take an example of both the scenarios

→Linear regression: When we want to predict the height of one person just from the weight of that person.
→Multiple Linear regression: If we alter the above problem statement just a little bit like, if we have the features like height, age, and gender of the person and we have to predict the weight of the person then we have to use the concept of multiple linear regression.

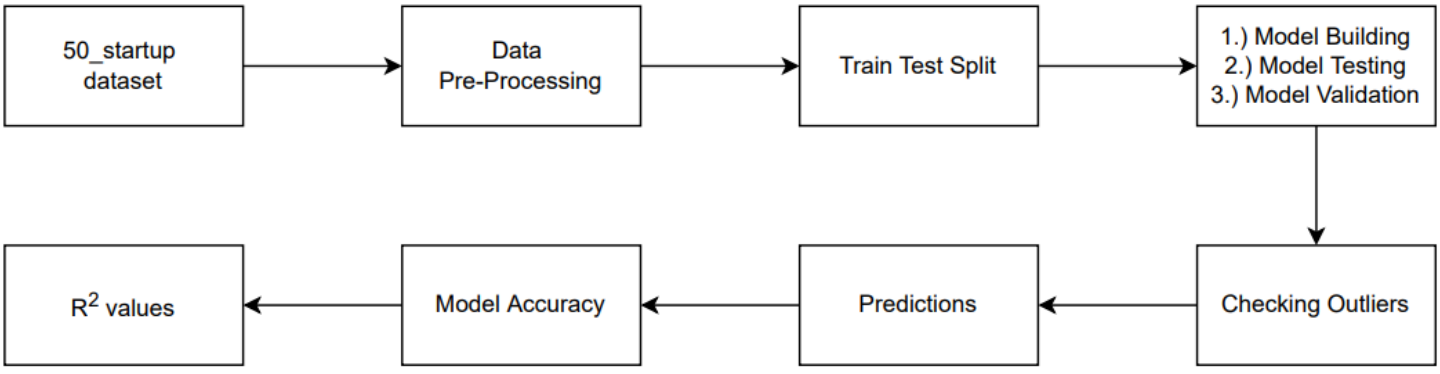


4.) Proposed Method with Architecture: -

The competition goal is to predict the profit of start-up profit on the bases of data provided which are on the bases of Research and Development Spend (R&D Spend), Administration Spend, Marketing Spend and State. We use multiple regression in this model because we have to 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 start-up company by analysing profit of the company.

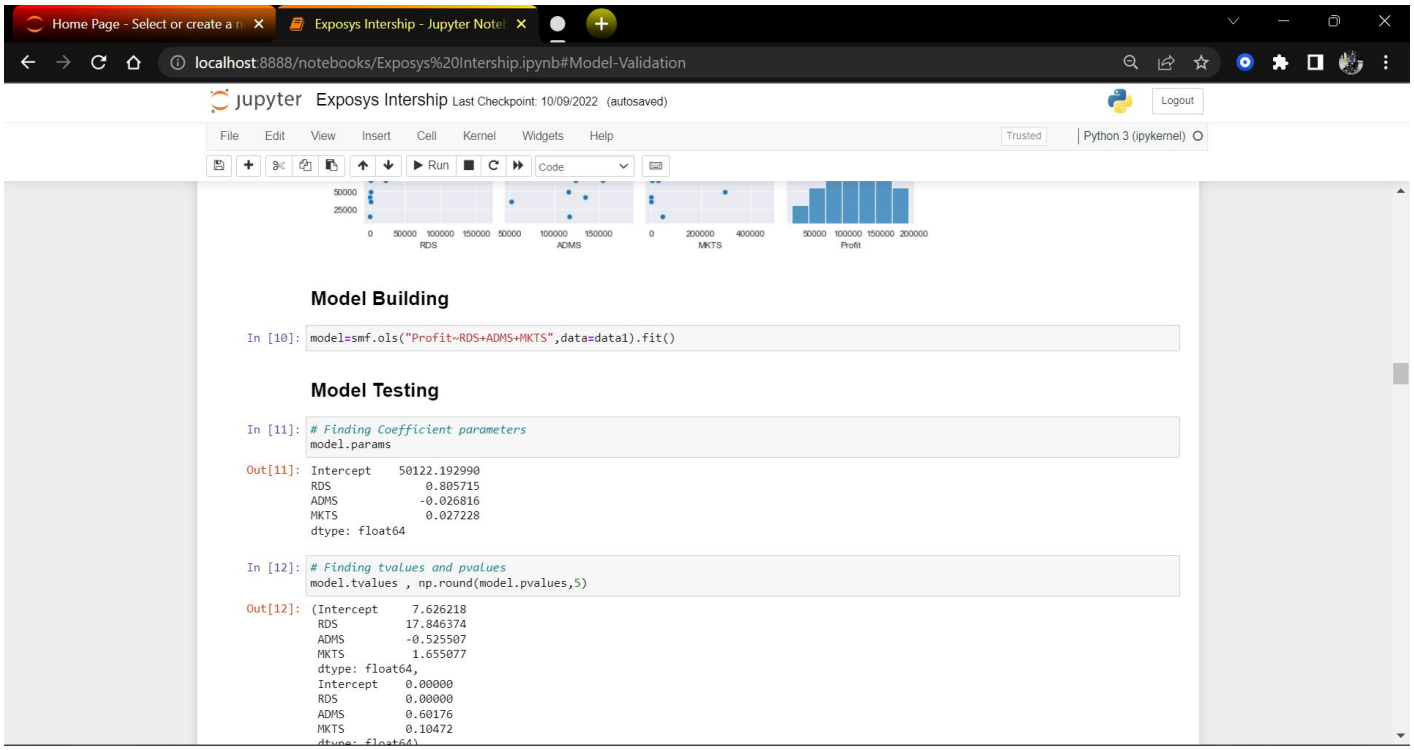


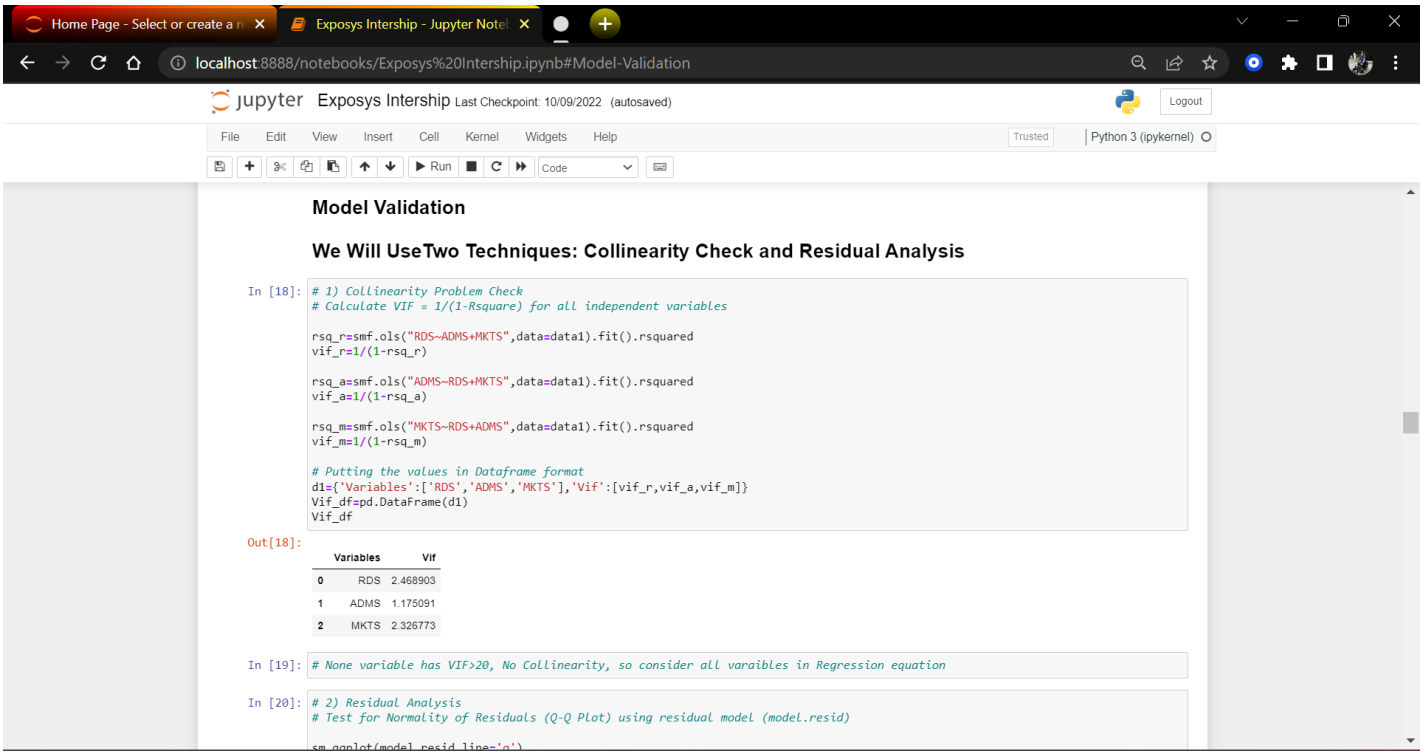
5.) Methodology



We first read the dataset and perform data exploration and data cleaning techniques, and make sure there are no null values or missing values in the dataset. And, then we spit the dataset into train and test and the build the model, test it and then validate it. Checking of outliers is done and then the predictions are made, after that model accuracy is checked and the R^2 values are generated.

7.) Implementation: -





Home Page - Select or create a n | Exposys Intership - Jupyter Notel

localhost:8888/notebooks/Exposys%20Intership.ipynb#Model-Validation

jupyter Exposys Intership Last Checkpoint: 10/09/2022 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

Checking Outliers or Influencers

Two Techniques : 1. Cook's Distance & 2. Leverage value

In [28]:
```
# 1. Cook's Distance: If Cook's distance > 1, then it's an outlier
# Get influencers using cook's distance
(c,_)=model.get_influence().cooks_distance
c
```

Out[28]:
```
array([3.21825244e-05, 3.27591036e-03, 3.23842699e-02, 2.17206555e-02,
       1.44833032e-02, 1.17158463e-02, 2.91766303e-03, 3.56513444e-03,
       4.04303948e-05, 4.86758017e-03, 1.51064757e-02, 1.63564959e-02,
       1.15516625e-02, 4.01422811e-03, 6.12934253e-02, 3.40013448e-02,
       8.33556413e-03, 3.30534399e-03, 2.16819303e-03, 4.07440577e-02,
       4.25137222e-04, 1.09844352e-02, 2.91768000e-03, 2.76030254e-04,
       5.04643588e-03, 3.00074623e-03, 3.41957068e-03, 2.98396413e-02,
       1.31590664e-03, 1.25992620e-04, 4.18505125e-05, 9.27434786e-06,
       7.08656521e-04, 1.28122674e-04, 2.09815032e-02, 6.69508674e-03,
       5.55314705e-02, 6.55050578e-05, 5.61547311e-02, 1.54279607e-03,
       1.84850929e-03, 1.97578066e-03, 1.36089280e-04, 2.05553171e-02,
       1.23156041e-04, 9.03234206e-02, 5.45303387e-02, 5.33885616e-03,
       1.90527441e-01, 2.88082293e-01])
```

In [29]:
```
# Plot the influencers using the stem plot
fig=plt.figure(figsize=(20,7))
plt.stem(np.arange(len(data1)),np.round(c,5))
plt.xlabel('Row Index')
plt.ylabel('Cooks Distance')
plt.show()
```

Home Page - Select or create a n | Exposys Intership - Jupyter Notel

localhost:8888/notebooks/Exposys%20Intership.ipynb#Model-Validation

jupyter Exposys Intership Last Checkpoint: 10/09/2022 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

Model Predictions

In [39]:
```
# say New data for prediction is
new_data=pd.DataFrame({'RDS':70000,"ADMS":90000,"MKTS":140000},index=[0])
new_data
```

Out[39]:

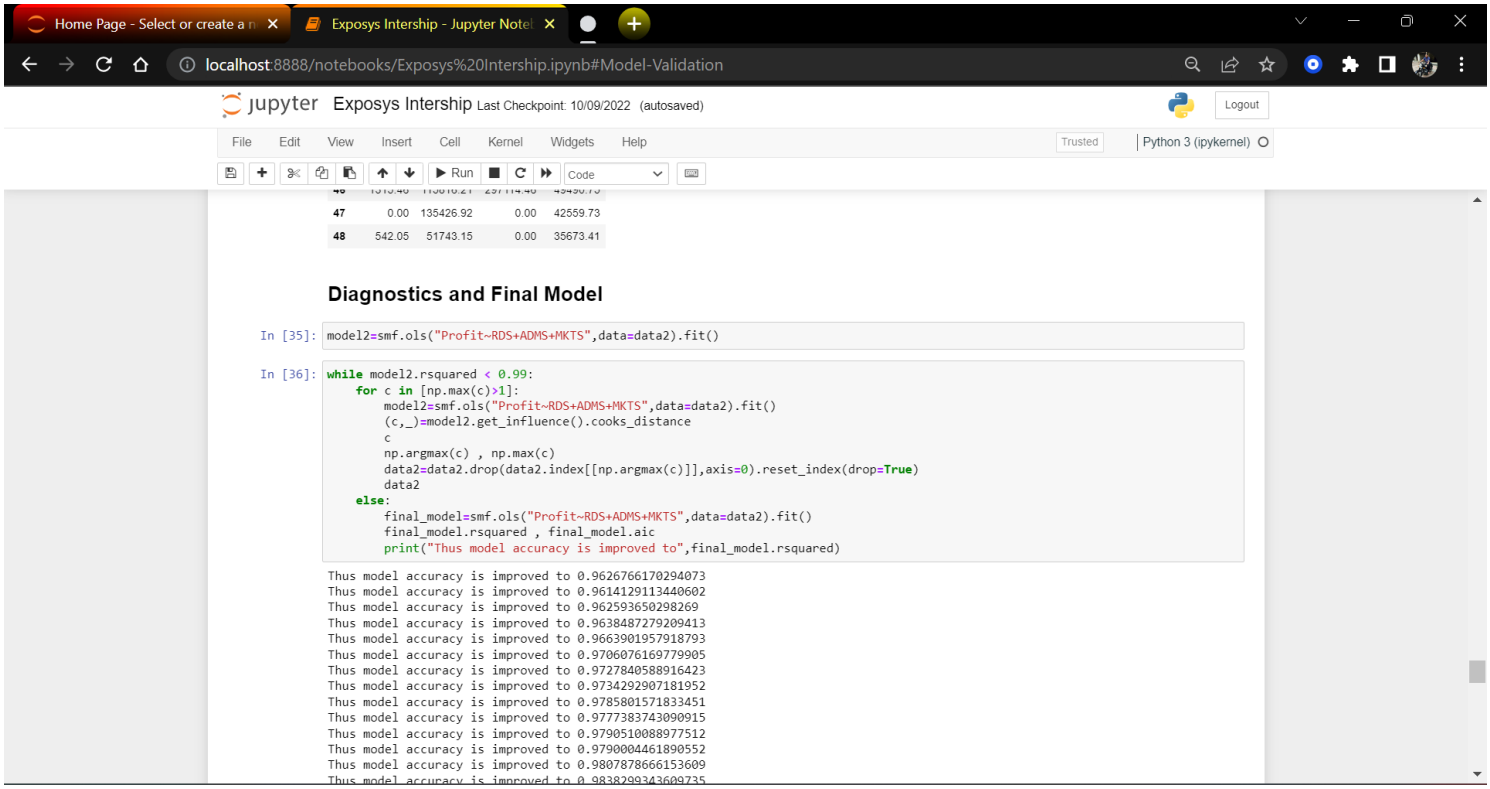
	RDS	ADMS	MKTS
0	70000	90000	140000

In [40]:
```
# Manual Prediction of Price
final_model.predict(new_data)
```

Out[40]:
```
0    104858.729408
dtype: float64
```

In [41]:
```
# Automatic Prediction of Price with 90.02% accurcy
pred_y=final_model.predict(data2)
pred_y
```

Out[41]:
```
0     165589.539700
1     158552.826483
2     156789.000710
3     149524.698853
4     150122.356712
5     126598.769555
6     130104.785747
7     127878.387928
8     117298.757074
9     111329.242429
10    110009.916133
11    102331.717613
12    109661.804131
```



7.) Conclusions: -

The project's primary goal was 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.

Multiple Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of three stages: 1) analysing the correlation and directionality of the data, 2) estimating the model, i.e., fitting the line, and 3) evaluating the validity and usefulness of the model.