EXPOSYS DATA LABS

Bengaluru, Karnataka, 560064

“Profit Prediction of 50 Companies using Data Science”.

A Dissertation work submitted in partial fulfilment of requirement for the award of degree of

Internship

By

Name- **Sasanapuri Tarun**

College- **Aditya Institution of Technology and Management (AITAM)**

Under the guidance of

Exposys Data Labs

ABSTRACT

Start-up companies are newly founded companies or entrepreneurial ventures that are in the initial phase of development. They are most commonly with high-tech projects, development and production, distribution of new products, processes or services.

To overcome this problem the techniques of Machine Learning can easily be utilized in order extract useful pieces of information of the specified data in startupcompanies and the profits. The accurate analysis of company database benefits in early prediction, profit and R&D Spend.

The project provides the R&D Spend, Administration Cost and Marketing Spend of 50 companies are given along with profit earned. The target is to prepare an MLmodel which can predict the profit value of a company if the value of R&D Spend,Administration Cost and Marketing Spend. Here we apply the ML Model like Regression algorithms. In the project we are applying the python model.

Table of Contents

Chapter Name	Page no
Abstract	2
1 Introduction	4-5
1.1 Data Science	4
1.2 Machine Learning	5
2 Existing Methods	6
2.1 Issues in existing Systems	6
3 Proposed method	7-9
3.1 Regression Algorithm	7
3.2 Algorithm	8
4 Methodology	9
4.1 Prediction Model	9
4.2 Data Preprocessing	9
4.3 Training and Testing	9
4.4 Data Model	9
5 Implementation	10-11
5.1 Source Code	10-11
6 Conclusion	12
7 References	13

INTRODUCTION

1. Data Science

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis can come from many different sources in various formats.

The Data Science Lifecycle

Now that we know what data science is, next up let us focus on the data science lifecycle. Data science's lifecycle consists of five distinct stages, each with its owntasks:

1. **Capture**: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.

2. **Maintain**: Data Warehousing, Data Cleansing, Data Staging, Data Processing,Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

3. **Process**: Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data Scientists take the prepared data and examine its patterns,ranges, and biases to determine how useful it will be in predictive analysis.

4. **Analyze**: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the real meat of the lifecycle. This stageinvolves performing the various analyses on the data.

5. **Communicate**: Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readableforms such as charts, graphs, and reports.

1.2. Machine Learning

Machine learning (ML) is a field of inquiry devoted to understanding and buildingmethods that 'learn', that is, methods that leverage data to improve performance onsome set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order tomake predictions or decisions without being explicitly programmed to do so.

Machine learning algorithms are used in a wide variety of applications, such as inmedicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

2.Existing Methods

Existing research handled for diabetes detection. Data mining approaches like clustering and classifications were studied in existing systems. Profit Prediction using algorithms such as Regression algorithms. The Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future prediction such as weather condition, sales, profit and administration, etc. For such case we need some technology which can make predictions more accurately.

2.1.Issues in existing systems

- Using machine learning the accuracy of detection is less.
- High false positives.
- There is no interactive tool for users to profit.

3.Proposed Method

3.1.Regression Algorithm

Linear Regression

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

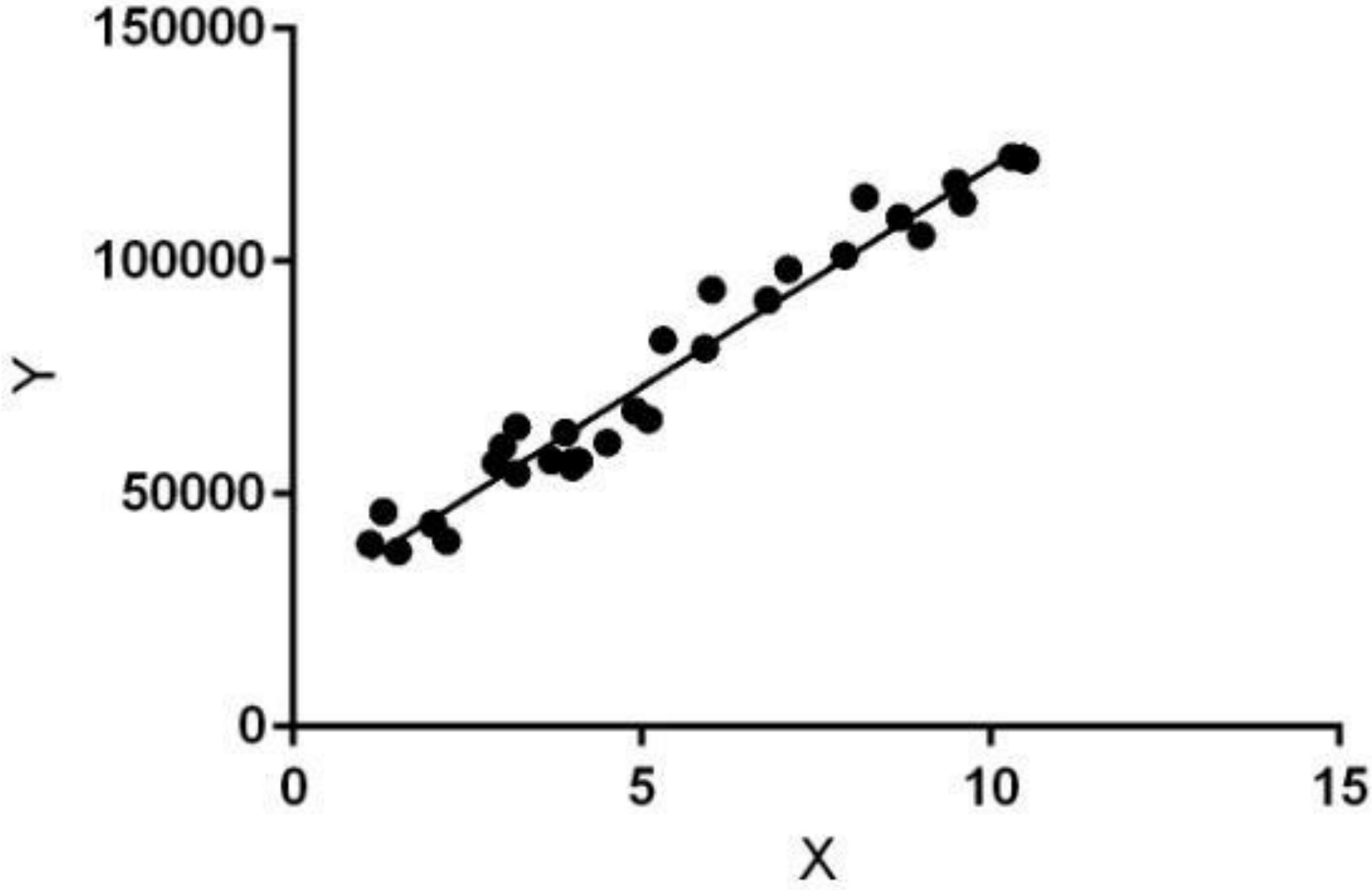
The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables.

The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.

One of the most important supervised learning tasks is regression. In regression setof records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.



3.2.Algorithm

Steps involved in linear regression algorithm:

Step 1: Import the libraries.

Step 2: Encoding the categorical data using the package sklearn module.

Step 3: Avoid dummy variable trap

Step 4: Splitting the data into Train and Test set.

Step 5: Fitting Multiple Linear Regression Model to Training set.

Step 6: Predicting the test results.

Step 7: Calculate the regression metrices.

4. Methodology

We’ll start with importing Pandas and NumPy into our python environment and loading a .csv

dataset into a pandas data frame named as df. To see the first five records from the dataset we use

pandas df.head() function. We’ll also use seaborn and matplotlib for visualization.

2. Data Preprocessing

The pre-processing techniques used in the presented work are:

- **Data Cleaning**: Data is cleansed through processes such as filling in missingvalue, thus resolving the inconsistencies in the data.
- **Data Reduction**: The analysis becomes hard when dealing with huge databases. Hence, we eliminate those independent variables(symptoms) which might have nonull values.

3. Training and Testing

We'll be using a simple learning machine model called Random Forest Classifier. We train the model with standard parameters using the training dataset. We evaluate the performance of our model using test dataset. Our model has a classification accuracy of 80.5%.

4. Data Model

The trained model is written in a pklfile called prediction-model which is used topredict the output of the user input. Flask is a micro web framework written in Python.

5. Implementation

5.1 Source Code

Import the necessary libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.pipeline import Pipeline,make_pipeline
```

Importing statsmodels

```
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

Importing metrics

```
from sklearn.metrics import
    accuracy_score,classification_report,confusion_matrix,auc,roc_curve
```

Importing sklearn-dataprocessing

```
from sklearn.model_selection import train_test_split,cross_val_score,GridSearchCV
from sklearn.feature_selection import f_regression
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import
    PolynomialFeatures,StandardScaler,OneHotEncoder
from sklearn.metrics import r2_score,mean_squared_error
```

Importing sklearn-models

```
from sklearn.linear_model import LinearRegression,Ridge
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

import warnings
warnings.filterwarnings(“ignore")

sns.set_theme(style='darkgrid',palette='Accent')
pd.options.display.float_format=“{:,.2f}".format
```

Load the dataset

The dataset is available at Kaggle

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

```
df=pd.read_csv(‘50Stratups.csv’)
x=df.iloc[:,0:2].values
y=df.iloc[:,1].values
x=x[:,1:]
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=1/2.5,random_state=0)
y=y_test.reshape(-1,1)
from sklearn.linear_model import LinearRegression
r=LinearRegression()
r.fit(x_train,y_train)
y_pred=r.predict(x_test)
x_pred=r.predict(y)
plt.scatter(x_train,y_train,color='red')
plt.plot(y_pred,x_pred,color='blue')
plt.title('LinearRegression')
plt.xlabel('R&D speed')
plt.ylabel('Administration')
plt.show()
#claculating the regression metrices
from sklearn.metrics import

 r2_score
```

6. Conclusion

In conclusion, the linear regression model developed in this project can accurately predict the profit value of a company based on R&D Spend, Administration Cost and Marketing Spend. In this, project we have applied the ML model like Regression Algorithms such as, Linear algorithm to check the relationship betweenthe independent and dependent variables and evaluating the regression metrices like MSE, RMSE and R-Squared value

7. REFERENCES

1. https://drive.google.com/file/d/1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF/view
2. https://www.javatpoint.com/machine-learning-random-forest-algorithm
3. https://en.wikipedia.org/wiki/Ensemble_learning
4. https://www.w3schools.com/css/css_form.asp
5. https://www.w3schools.com/howto/howto_css_modals.asp
6. https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science