Data Science and Business Analytics Intern at The Sparks Foundation

Bachelor of Technology

In

COMPUTER SCIENCE

Submitted by:

YASH BISHT

BTECH 3RD SEMESTER 2022

Under the Supervision of:

MR AJAY KUMAR

(Assistant Professor)

Department of Computer Science

Faculty of Engineering

GOEL INSTITUTE OF TECHNOLOGY AND MANAGEMENT , LUCKNOW, INDIA

FEBRUARY, 2022

CERTIFICATE

This is to certify that **Mr. Yash Bisht** (Roll. No. 2003600100111) ha**s** carried out the work presented in the synopsis titled **“**Data Science and Business Analytics Intern at The Sparks Foundation**”** submitted for partial fulfillment for the award of the **Bachelor of Technology In Computer Science** from **GITM, Lucknow** under my supervision.

It is also certified that:

(i) This synopsis embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/diploma/certificate.

(ii) The candidate has worked under my supervision for the prescribed period.

(iii) The synopsis fulfills the requirements of the norms and standards prescribed by the AKTU and GITM, Lucknow, India.

(iv) No published work (figure, data, table etc) has been reproduced in the synopsis without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.

Ajay Kumar
(GUIDE)
Goel Institute of Technology and Management , LUCKNOW

Date:28/02/2022
Place : Lucknow

DECLARATION

I hereby declare that the synopsis titled **“Data Science and Business Analytics Intern at The Sparks Foundation”** is an authentic record of the research work carried out by me under the supervision of Ajay Kumar, Department of Information technology, for the period of February, 2022 at GITM, Lucknow. No part of this synopsis has been presentedelsewhere for any other degree or diploma earlier.

I declare that I have faithfully acknowledged and referred to the works of other researchers wherever their published works have been cited in the synopsis. I further certify that I have not willfully taken other's work, para, text, data, results, tables, figures etc. reported in the journals, books, magazines, reports, synopsis, theses, etc., or available at web-sites without their permission, and have not included those in this B.Tech synopsis citing as my own work.

Date:28/02/2022

Byash

Yash Bisht

ACKNOWLEDGEMENTS

I am pleased to acknowledge mentor Ajay kumar sir for their invaluable guidance during the course of this Internship work.

I want to thank my advisers and everyone at the company for their patience and assistance during my on-site training. Thanks to their guidance, I was able to develop [mention specific skills and learn about the skills that will help me to expand my resume and advance my career.

We extend our sincere thanks to Mr. Brijesh Pandey who continuously helped me throughout the Internship tasks and without his guidance, this Internship would have been an uphill task.

We are also grateful to other members of the The Sparks Foundation team who co-operated with us regarding some issues.

We would also like to thank 'IBM-CLOUD' (https://cloud.ibm.com) Source banner which provide and helped me to deliver the data – analysis report in Jupyter notebook (I used IBM Cloud as PAAS)

Last but not the least, My Folks and my batchmates also co-operated with me nicely for the smooth report making and Data analysis tasks.

Yash Bisht
February
2022

INTRODUCTION

Overview of Company:

The sparks foundation is working to bring parity in education, making sure children have equal opportunity at success, irrespective of the financial background.

I have participated in the Graduate Rotational Internship Program (GRIP)organized by the company.

Website link: https://www.thesparksfoundationsingapore.org/

Company's Mission Statement:

"To inspire students, help them innovate and Let them integrate to build the next generation humankind."

Position:

Data Science and Business Analytics Intern at The Sparks Foundation

Duration:

The internship is of Two months started from 1 February 2022 and end at 03 March 2021.

Mentors:

My Supervisor during the Internship is Mr. Ajay kumar sir, works in The Sparks Foundation.

Overview of Tasks:

This report discusses the result of the work done in Analysis and prediction using "Prediction using Supervised ML and Python Libraries and Frameworks for Analysis and Visualization for the Prediction Systems on Jupyter Notebook Platform.

Background and Motivation:

Data analysis is important in **business to understand problems facing an organization**, and to explore data in meaningful ways. Data in itself is merely facts and figures. Data analysis organizes, interprets, structures and presents the data into useful information that provides context for the data.

It adds value to the organization, helping it to make informed business decisions and providing an edge over market competitors. Hence, a career in analytics is bound to guarantee you the role of one of the key decision-makers in the organization.

Methodology:

Prediction using Supervised ML

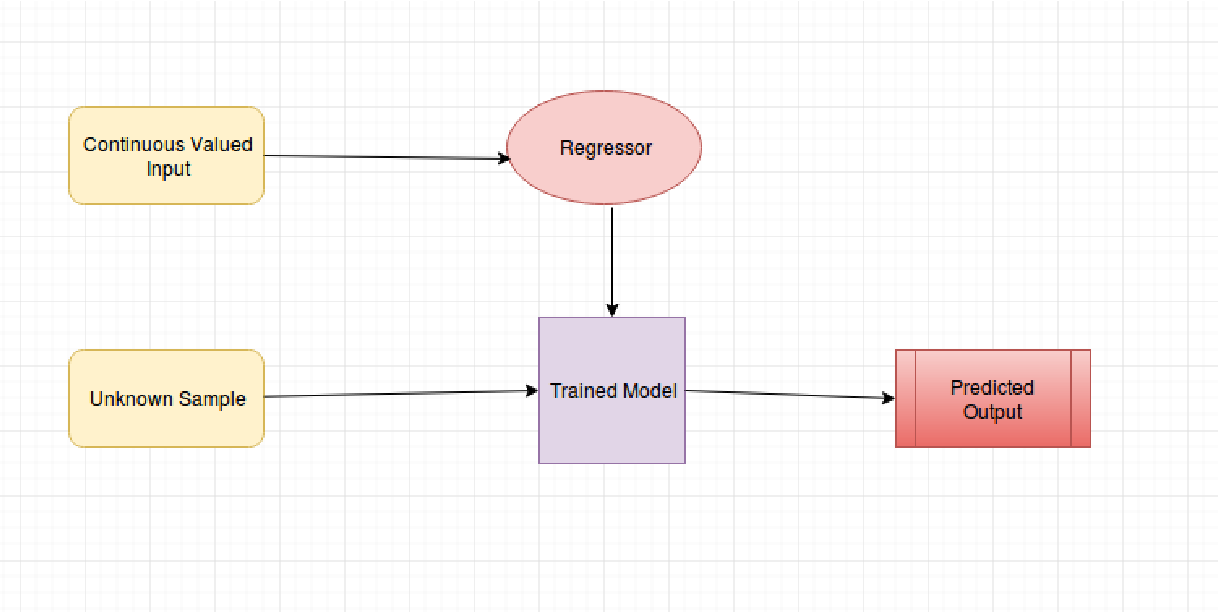
Problem Statement : create Linear Regression Model that predict the percentage of an student based on the no. of study hours.
Objective : The goal of this task is if we feed any new data(no. of study hours) in the model,
it would be able to predict the score of the student.

Format of the internship:
The internship will have a maximum duration of two months. asks are assigned and interns work independently. The completed tasks are submitted through Google forms. I can also ask any queries in the discussion forum, The Sparks Foundation Network on LinkedIn and our mentors and fellow interns will help me out.

Tasks List:

- LinkedIn Profile Improvement -I have to Improve my professional profile on LinkedIn.
- Technology (only Tech interns)- I have to Complete AT LEAST ONE TASK from the list of tasks given under your internship function. After that, I can do as many tasks as I want for learning & LoR.
- Peer-evaluation (mandatory for all): I have to Watch and comment on the at least 5 task videos on LinkedIn posted by fellow interns. Refer to FAQs for the steps of peer evaluation: https://lnkd.in/gnGiBbb
- Additional tasks for LoR (optional)- This will be shared via email. I have also take help in FAQs for this: https://lnkd.in/gnGiBbb

DATA FLOW DIAGRAM



Facilities required for proposed work (Tools Description)

I have used various tools during my internship which are:

LinkedIn:

LinkedIn is the **world's largest professional network on the internet**. You can use LinkedIn to find the right job or internship, connect and strengthen professional relationships, and learn the skills you need to succeed in your career. I have used this to apply for the Internship. And they contact me only through Gmail and LinkedIn.

Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows data scientists to create
And share documents that integrate live code, equations, computational output, visualizations, and
multimedia resources, along with explanatory text in a single document

IBM Watson Studio:

BM Watson® Studio empowers data scientists, developers and analysts to build, run and manage AI
models, and optimize decisions anywhere on IBM Cloud Pak® for Data. I have used to create
jupyter Notebook inside IBM Watson.

Python Machine Learning, Data Analysis and Data Visualization Libraries:

For Data Analysis and Reporting I have used many Libraries for example MatplotLib(for data visualization) Numpy and Pandas (for working on Series and DataFrames Data Structures) Sklearn (for Machine Learning).

REPORT

Data Science and Business Analytics Intern at The Sparks Foundation

Author : Yash Bisht

Task 1 : Prediction using Supervised ML

Problem Statement : create Linear Regression Model that predict the percentage of an student based on the no. of study hours.

Goal : The goal of this task is if we feed any new data(no. of study hours) in the model, it would be able to predict the score of the student

Step 1: Importing all libraries required in this notebook

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.linear_model import LinearRegression from
sklearn.model_selection import train_test_split from sklearn import
metrics
```

```
Waiting for a Spark session to start...
Spark Initialization Done! ApplicationId = app-20210407163902-0002
KERNEL_ID = 5a5ffc4f-e2ee-4054-b102-8eaf753208f6
```

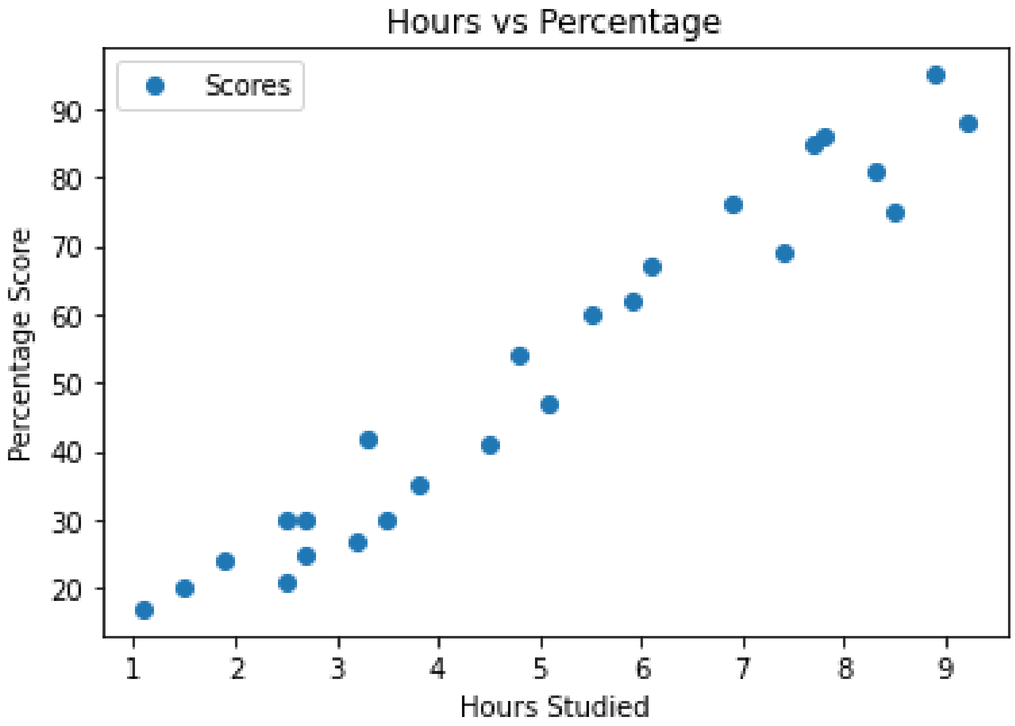
```
# Reading data from given remote link url =
"https://raw.githubusercontent.com/AdiPersonalWorks/Random/master/
student_scores%20-%20student_scores.csv" data = pd.read_csv(url) print("Data imported
successfully") data.head(10)
```

```
Data imported successfully

  Hours  Scores 0    2.5
21
1       5.1     47
2       3.2     27
3       8.5     75
4       3.5     30
5       1.5     20
6       9.2     88
7       5.5     60
8       8.3     81
9       2.7     25
```

plotting our data points on 2-D graph

```
data.plot(x='Hours', y='Scores', style='o') plt.title('Hours vs
Percentage') plt.xlabel('Hours Studied')
plt.ylabel('Percentage Score')  plt.show()
```



From the graph above, we can clearly see that there is a positive linear relation between the number of hours studied and percentage of score. **step 2: Preparing the data** dividing the data into "attributes" (inputs) and "labels" (outputs).

```
X = data.iloc[:, :-1].values Y = data.iloc[:,
1].values
```

Step 3: Splitting the data into training and test sets.

```
 X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size =0.2)

X_train.shape

(20, 1)

X_test.shape

(5, 1)
```

step 4: fitting linear regression model

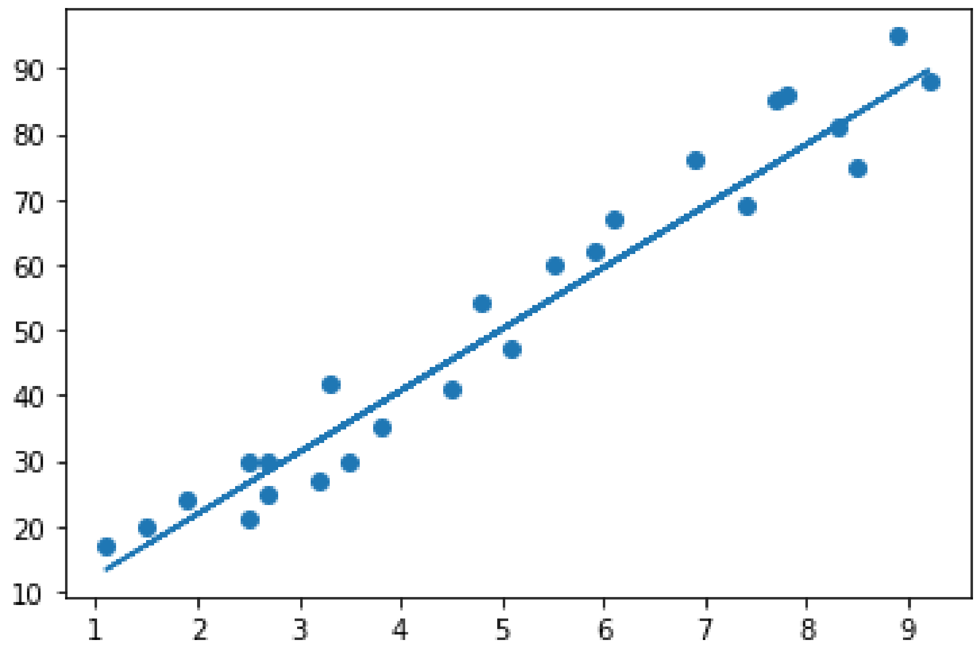
```
regressor = LinearRegression()

regressor.fit(X_train, Y_train) print("Training

complete.") Training complete.
```

step 5: Plotting the regression line

```
line = regressor.coef_*X+regressor.intercept_ plt.scatter(X, Y)
plt.plot(X, line); plt.show()
```



step 6 :Making Predictions

```
Y_predict = regressor.predict(X_test)
df = pd.DataFrame({"Actual": Y_test, "predicted": Y_predict}) df
```

```
  Actual  predicted 0      76
68.013668
1            95  86.868750
2            20  17.104948
3            24  20.875965
4            86  76.498455
```

Conclusion & Future Work

Conclusion

Regression analysis is a reliable method of identifying which variables have impact on a topic of interest.

The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.

What will be predicted score if a student studies for 9.25 hrs/ day?

```
predicted_score = regressor.predict([[9.25]])[0]
print(f"No. of Hours = {9.25}\n Predicted Score = {predicted_score}") No. of Hours = 9.25
 Predicted Score = 90.16838897911664
```

Evaluating the model

```
#by mean square error print('Mean Absolute
Error = ',
      metrics.mean_absolute_error(Y_test, Y_predict))

Mean Absolute Error =  6.327642755508299
```

Future Work

We can do overall exploratory data analysis and we can use Multiple Regression in the datasets.

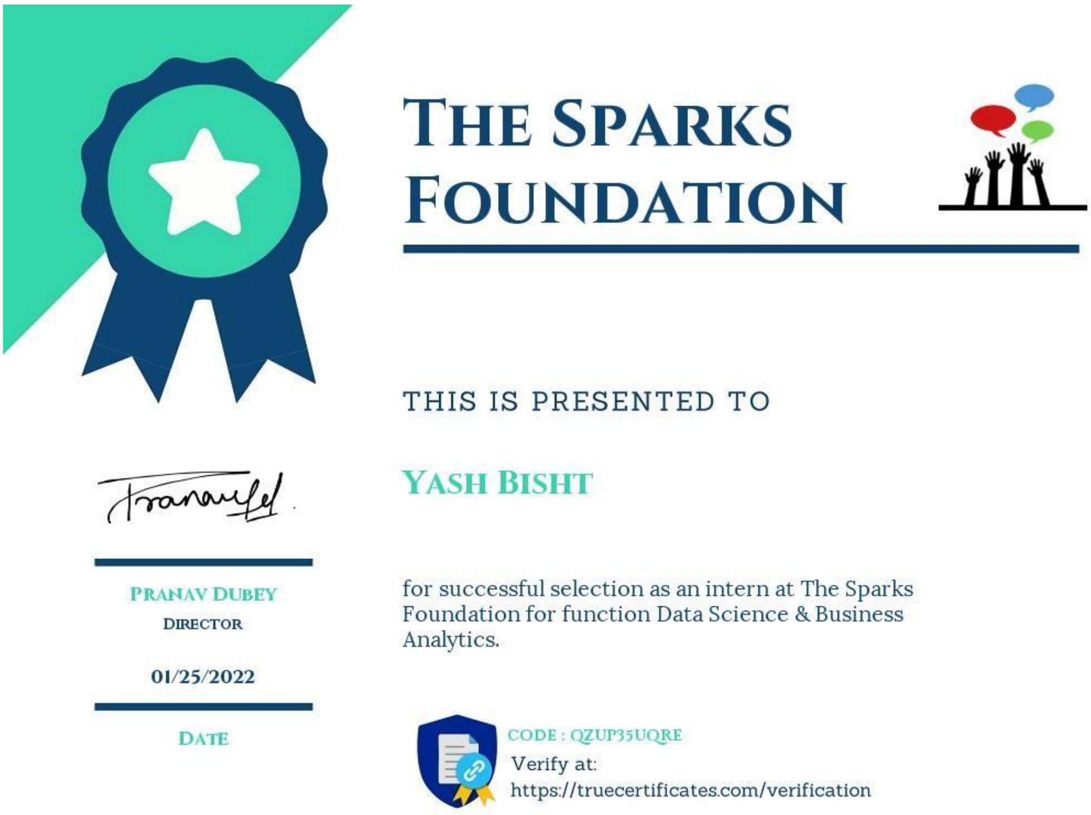
Multiple regression is an extension of linear regression models that allow predictions of systems with multiple independent variables

REFERENCES

List of references that are used are:

- https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/markd-jupyter.html
- https://scikit-learn.org/0.21/documentation.html
- https://pandas.pydata.org/docs/
- https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/8b6de2fb-b5f2-4c54-adf6-8ac9d5a51760/view?access_token=13e9f5c6146e1e0d0937b31faee92fdd95b3879835854ee8890c76fa085dfbd3
- https://jupyter-notebook.readthedocs.io/en/stable/

CERTIFCATE OF PARTICIPATION



THE SPARKS FOUNDATION

THIS IS PRESENTED TO

YASH BISHT

for successful selection as an intern at The Sparks Foundation for function Data Science & Business Analytics.

PRANAV DUBEY
DIRECTOR

01/25/2022

DATE

CODE : QZUP35UQRE

Verify at:
https://truecertificates.com/verification