VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

An Internship Project Report
On

"Boston House Price Prediction"

Submitted in partial fulfillment of the requirements for the VII Semester of degree of **Bachelor of Engineering in Information Science and Engineering** of Visvesvaraya Technological University, Belagavi

by

Suprith Satish
1RN19IS161

Under the Guidance of

Ms. Priyanka M R
Asst. Professor
Department of ISE

ESTD:2001
An Institute with a Difference

Department of Information Science and Engineering

RNS Institute of Technology

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru-560098

2022-2023

RNS Institute of Technology

Channasandra, Dr.Vishnuvardan Road, Bengaluru-560098

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

Certified that the Internship work entitled ***“Boston House Price Prediction”*** has been successfully completed by **Suprith Satish** bearing USN **1RN19IS161,** bonafide student of **RNS Institute of Technology** in partial fulfillment of the requirements for the Final year degree in **Bachelor of Engineering in Information Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during academic year 2022-2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The Internship report has been approved as it satisfies the academic requirements in respect of Internship work for the said degree.

Ms. Priyanka M R
Internship Guide
Asst. Professor
Dept. Of ISE

Dr. R Rajkumar/ Mr. Pramod R
Internship Coordinator
Asso./Asst. Professor
Dept. Of ISE

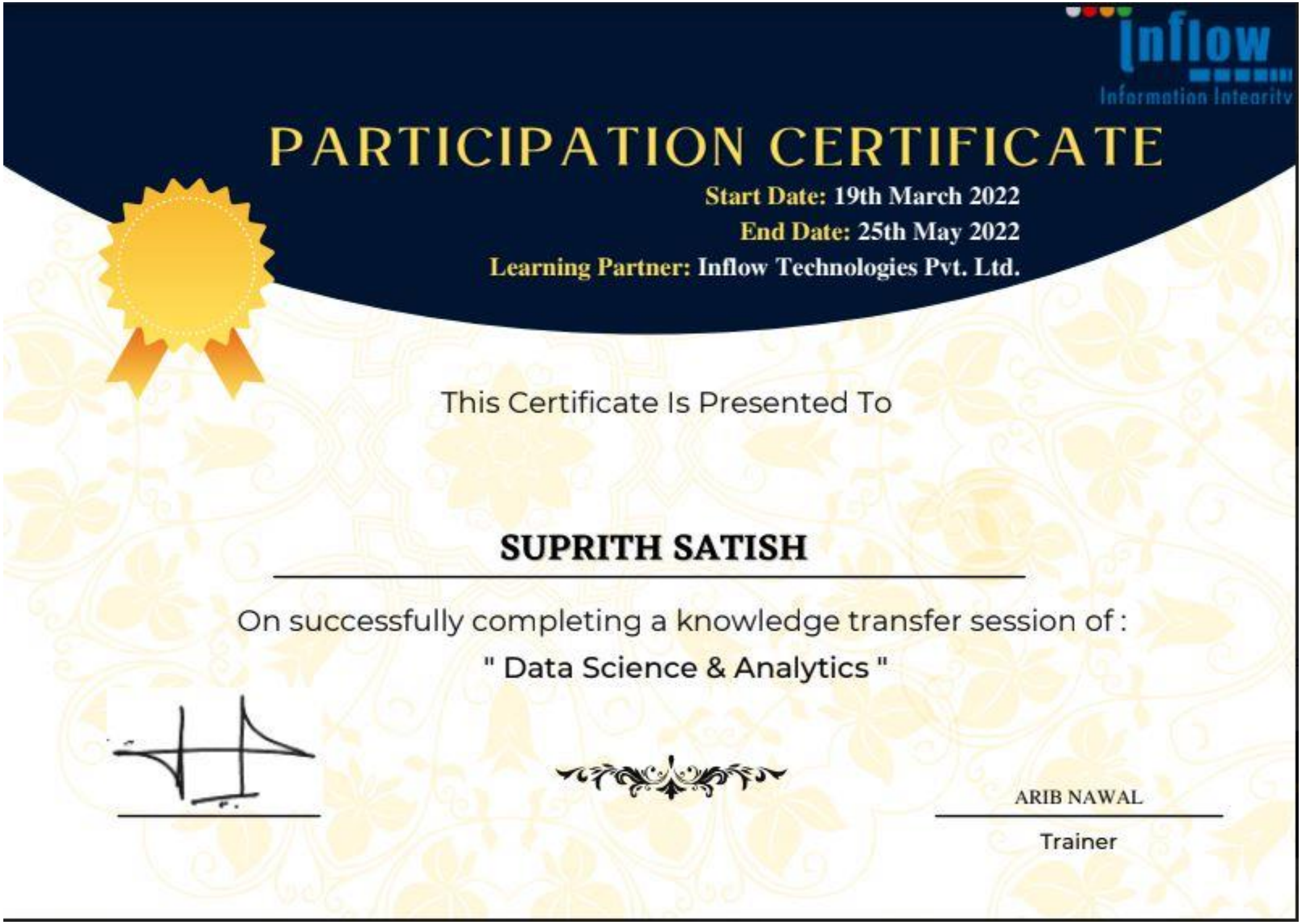
Dr. Suresh L
Professor, HoD
Dept. Of ISE
RNSIT

External Viva

Name of the Examiners **Signature with Date**

1.

2.



Inflow Information Integrity

PARTICIPATION CERTIFICATE

Start Date: 19th March 2022

End Date: 25th May 2022

Learning Partner: Inflow Technologies Pvt. Ltd.

This Certificate Is Presented To

SUPRITH SATISH

On successfully completing a knowledge transfer session of :

" Data Science & Analytics "

ARIB NAWAL

Trainer

ABSTRACT

"Boston House Prediction" is a Machine Learning based project which aims at predicting the Boston house prices based on set of features.This model undergoes training with dataset values in order to predict the price values.It uses Linear Regression to implement the same. The dataset for this project originates from the UCI Machine Learning Repository. The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts. We evaluate the performance and predictive power of a model that has been trained and tested on data collected from homes in suburbs of Boston, Massachusetts. A model trained on this data that is seen as a good fit could then be used to make certain predictions about a home's monetary value. A model like this would be very valuable for a real estate agent who could make use of the information provided in a daily basis.

ACKNOWLEDGMENT

At the very onset, I would like to place on record our gratefulness to all those people who have helped us in making this internship work a reality. Coming up with this Internship work to be a success was not easy. Our Institution has played a paramount role in guiding us in the right direction.

I would like to profoundly thank the **Management** of **RNS Institute of Technology** for providing such a healthy environment for the successful completion of this Internship / Professional Practice work.

I would also like to thank our beloved Principal, **Dr. M K Venkatesha**, for providing the necessary facilities to carry out this work.

I am extremely grateful to our beloved HoD, **Dr. Suresh L,** Dept. Of ISE, for having accepted to patronize me in the right direction with all his wisdom.

I am extremely grateful to our internship coordinators **Dr. R Rajkumar**, Associate Professor and **Mr. Pramod R,** Assistant Professor, Dept. Of ISE, for their continuous guidance and efforts to help students carry out their internships and the project successfully.

I would like to thank the internal guide **Ms. Priyanka M R,** Assistant Professor, Dept. Of ISE, for her continuous guidance and constructive suggestions for this work.

It’s my pleasure to thank Inflow Technologies for providing me the best platform to complete the internship work and glad to thank the external guide **Mr. Arib Nawal,** for corporate exposure, training and guidance.

Last but not the least, I am thankful to all the staff members of Information Science and Engineering Department for their encouragement and support throughout this work.

Suprith Satish 1RN19IS161

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

1.1 ORGANIZATION / INDUSTRY

1.1.1 COMPANY PROFILE

Inflow Technologies is IT Services Company registered under Govt. of India, Ministry of Micro, Small and Medium Enterprises. It also offers customized training programs, workshops and internship programs across engineering colleges in India.

Mission: To empower the students with necessary knowledge and industry skills for them to succeed in their future endeavors.

The Core Services they offer:

- Software Development and Testing Services on cutting edge technologies for small,mid-size and large enterprise clients.
- Permanent and Temporary staffing solutions to IT companies.
- Career Coaching, Mentoring, and Consultation to IT professionals.
- Technology trainings, Training on Leadership skills and Training on soft skills

1.1.2 DOMAIN / TECHNOLOGY

Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given. Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.

Machine learning is closely related to and often overlaps with computational statistics, a discipline that also specializes in prediction-making. It has strong ties to mathematical optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR), search engines and computer vision. Machine learning is sometimes conflated with data mining, although that focuses more on exploratory data analysis. Machine learning and pattern recognition can be viewed as two facets of the same field.

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning “signal” or “feedback” available to a learning system. These are:

1. Supervised learning: The computer is presented with example inputs and their desired outputs, given by a “teacher”, and the goal is to learn a general rule that maps inputs to outputs.
2. Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
3. Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

The most widely used Machine learning algorithms are:

- Support Vector Machines
- Linear Regression
- Logistic Regression
- Polynomial Regression
- Decision Trees
- K-Nearest Neighbor algorithm

- Neural Networks
- Clustering analysis

The applications of Data Science are:

- Digital Advertisements (Targeted Advertising and re-targeting)

  If you thought Search would have been the biggest application of data science and machine learning, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital bill boards at the airports – almost all of them are decided by using data science algorithms.

  This is the reason why digital ads have been able to get a lot higher CTR than traditional advertisements. They can be targeted based on user's past behavior. This is the reason why I see ads of analytics trainings while my friend sees ad of apparels in the same place at the same time.

- Recommender Systems

  Who can forget the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them, but also add a lot to the user experience.

  A lot of companies have fervidly used this engine / system to promote their products / suggestions in accordance with user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, LinkedIn, IMDB and many more uses this system to improve user experience. The recommendations are made based on previous search results for a user.

- Price Comparison Websites

  At a basic level, these websites are being driven by lots and lots of data which is fetched using APIs and RSS Feeds. If you have ever used these websites, you would know the convenience of comparing the price of a product from multiple vendors at one place. PriceGrabber, PriceRunner, Junglee, Shopzilla, DealTime are some examples of price comparison websites. Nowadays, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.

- Fraud and Risk Detection

  One of the first applications of data science originated from Finance discipline.

Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paper work while sanctioning loans. They decided to bring in data science practices in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

- Image Recognition

  You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. Similarly, while using Whatsapp web, you scan a barcode in your web browser using your mobile phone. In addition, Google provides you the option to search for images by uploading them. It uses image recognition and provides related search results.

- Airline Route Planning

  Airline Industry across the world is known to bear heavy losses. Except a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements.

  Now using data science, the airline companies can:

  1. Predict flight delay.
  2. Decide which class of airplanes to buy.
  3. Whether to directly land at the destination, or take a halt in between.

  Southwest Airlines, Alaska Airlines are among the top companies who've embraced data science to bring changes in their way of working.

1.2 PROBLEM STATEMENT

Given a list of attributes of the household, the goal is to predict the monetary value of any house located in Boston city. The features can be summarized as follows

- LSTAT: This is the percentage lower status of the population
- MEDV: This is the median value of owner-occupied homes in $1000s (Target Variable)

- CRIM: This is the per capita crime rate by town
- ZN: This is the proportion of residential land zoned for lots larger than 25,000 sq.ft.
- INDUS: This is the proportion of non-retail business acres per town.
- CHAS: This is the Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX: This is the nitric oxides concentration (parts per 10 million)
- RM: This is the average number of rooms per dwelling
- AGE: This is the proportion of owner-occupied units built prior to 1940
- DIS: This is the weighted distances to five Boston employment centers
- RAD: This is the index of accessibility to radial highways
- TAX: This is the full-value property-tax rate per $10,000
- PTRATIO: This is the pupil-teacher ratio by town
- B: B =1000(BK — 0.63)², where BK is African American people proportion descent by town

1.3 PROPOSED SYSTEM

In this project, we will evaluate the performance and predictive power of a model that has been trained and tested on data collected from homes in suburbs of Boston, Massachusetts. A model trained on this data that is seen as a *good fit* could then be used to make certain predictions about a home's monetary value. The dataset for this project originates from the UCI Machine Learning Repository. The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts. The model can be also be used to set the price of any house in the area based on attributes of the house.

CHAPTER 2

REQUIREMENT ANALYSIS, TOOLS & TECHNOLOGIES

2.1 Hardware & Software Requirements

- Windows / Unix Os
- RAM – 4 GB
- Jupyter Notebook
- Anaconda Navigator
- Google Colab
- Tableau
- External Libraries – Numpy, Pandas, Scikit-learn, Matplotlib

2.2 Functional Requirements

- The system should be able to predict the MEDV (median value of owner-occupied homes in $1000s) i.e. the Target variable for any new house based of the household parameters.
- The system should predict results with accuracy greater than 70 percent.
- The system should be able to visualize the difference between predicted and actual data.

2.3 Non Functional Requirements

- The system should provide good performance. Performance defines how fast a software system or a particular piece of it responds to certain users' actions under a certain workload.
- The system should be able to provide scalability. Scalability assesses the highest workloads under which the system will still meet the performance requirements.
- The system should provide good reliability. Reliability specifies how likely the system or its element would run without a failure for a given period of time under predefined conditions.

CHAPTER 3

DESIGN AND IMPLEMENTATION

3.1 FLOW CHART

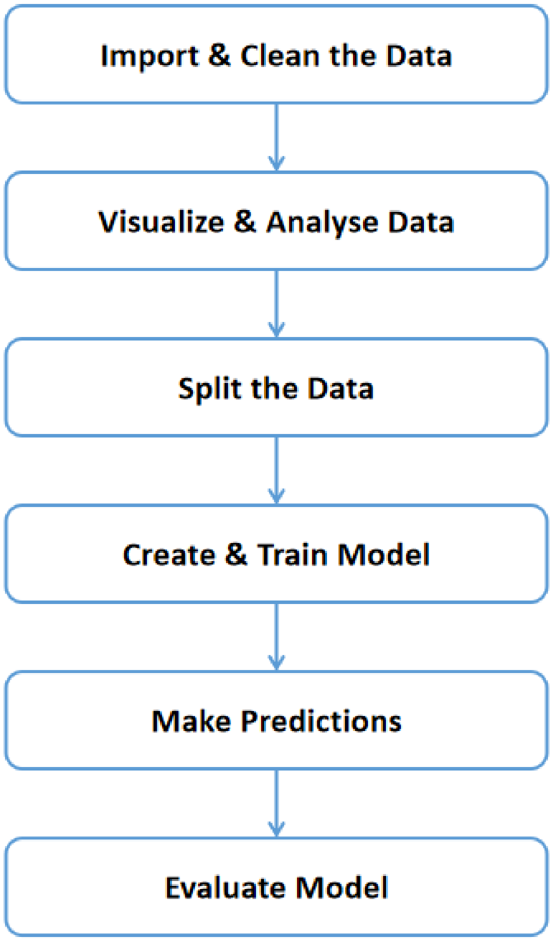


Fig 3.1 Flowchart representing Prediction process

3.2 ALGORITHM

Linear Regression is a machine learning algorithm based on supervised learning. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

$$observed\ data \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

$$predicted\ data \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

$$error \rightarrow \quad \varepsilon = y - y'$$

3.3 LIBRARIES / APIs

- NumPy: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations and much more.
- Pandas: Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or Python.
- Scikit-learn: Scikit-learn is an open source Python library that implements a range of machine learning, pre-processing, cross-validation and visualization algorithms using a unified interface.
- Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib consists of several plots like line, bar, scatter, histogram etc. Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is a low level graph plotting library written in python. It is open source and we can use it freely.

CHAPTER 4

OBSERVATIONS & RESULTS

4.1 TESTING

Software Testing is the process used to help identify the correctness, completeness, security and quality of the developed computer software. Testing is the process of technical investigation and includes the process of executing a program or application with the intent of finding errors.

4.1.1 UNIT TESTING

Unit testing is a level of software testing where individual units/ components of software are tested. The purpose is to validate that each unit of the software performs as designed. A unit is the smallest testable part of any software. It usually has one or a few inputs and usually a single output.

Table 4.1.1 Unit test case for Dataset Loading

SI No. of test case:	1
Name of test:	Check test
Item / Feature being tested:	Input Dataset
Sample Input:	Syntax to display top 5 element from dataset specified in a particular directory
Expected output:	Displays top 5 elements
Expected output:	Displays top 5 elements
Remarks	Test Succeeded

4.1.2 INTEGRATION TESTING

Integration testing is done to test the modules/components when integrated to verify that they work as expected i.e. to test the modules which are working fine individually does not have issues when integrated. The goal of integration testing is to detect any irregularity between the units that are integrated together.

Table 4.1.2 Integration test to check for train-test-split

SI No. of test case:	3
Name of test:	Check test
Item / Feature being tested:	Train-test-split function
Sample Input:	Input features and ratio and check for the correct splitting of the dataframe
Expected output:	Split according to the given ratio
Expected output:	Splitted according to the given ratio
Remarks	Test Succeeded

4.1.3 SYSTEM TESTING

System Testing (ST) is a black box testing technique performed to evaluate the system's compliance against specified requirements. System Testing is carried out on the whole system in the context of either system requirement specifications or functional requirement specifications or in the context of both. System testing tests the design and behavior of the system and also the expectations of the customer. It is performed to test the system beyond the bounds mentioned in the software requirement specifications (SRS).

Table 4.1.3 System test to check the accuracy using LR

SI No. of test case:	4
Name of test:	Check test
Item / Feature being tested:	Learning algorithm’s efficiency
Sample Input:	Training feature and output feature
Expected output:	Improved efficiency Accuracy:71%
Remarks	Test Succeeded

4.2 GRAPHS

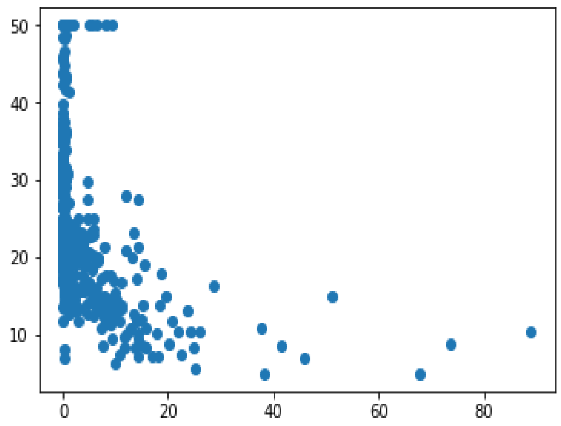


Fig 4.2.1: CRIM VS MEDV

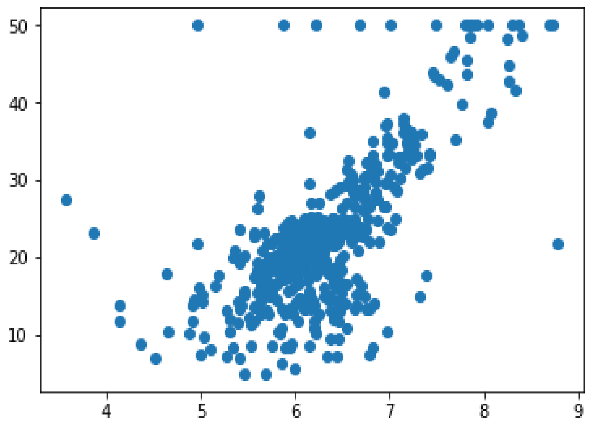


Fig 4.2.2: RM VS MEDV

- CRIM VS MEDV - This graphs depicts the relation between the per capita crime rate along the x axis and the median value of owner-occupied(target variable) homes along the y axis.
- RM VS MEDV - This graphs depicts the relation between the average number of rooms per dwelling along the x axis and the median value of owner-occupied(target variable) homes along the y axis.

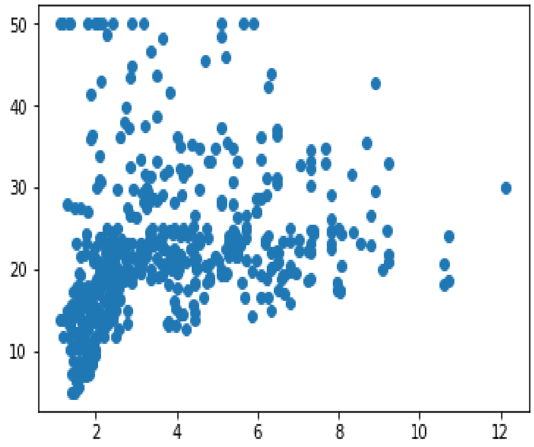


Fig 4.2.3: DIS VS MEDV

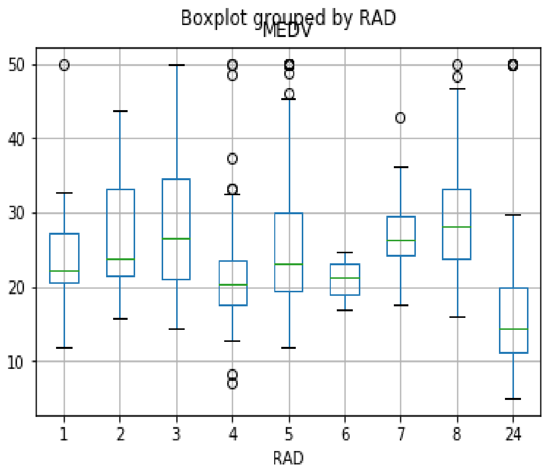


Fig 4.2.4: RAD VS MEDV

- DIS VS MEDV - This graphs depicts the relation between the weighted distance to five Boston employment centers along the x axis and the median value of owner-occupied(target variable) homes along the y axis.

- RAD VS MEDV - This graphs depicts the relation between the index of accessibility to radial highways along the x axis and the median value of owner-occupied(target variable) homes along the y axis.

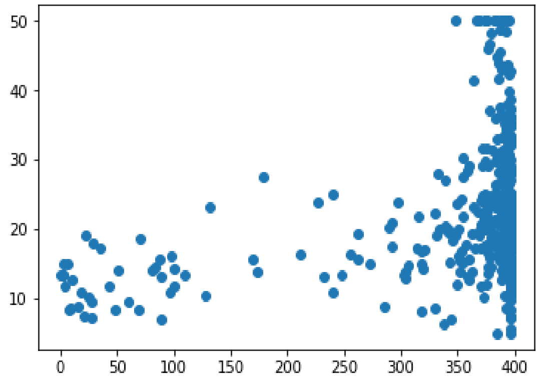


Fig 4.2.5: B VS MEDV

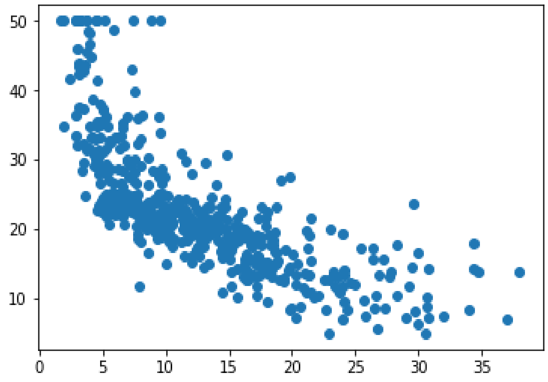


Fig 4.2.6: LSTAT VS MEDV

- B VS MEDEV - This graphs depicts the relation between 'B' along the x axis and the median value of owner-occupied(target variable) homes along the y axis. where B =1000(BK — 0.63)², where BK is African American people proportion descent by
- LSTAT VS MEDEV - This graphs depicts the relation between the percentage lower status of the population along the x axis and the median value of owner-occupied(target variable) homes along the y axis.

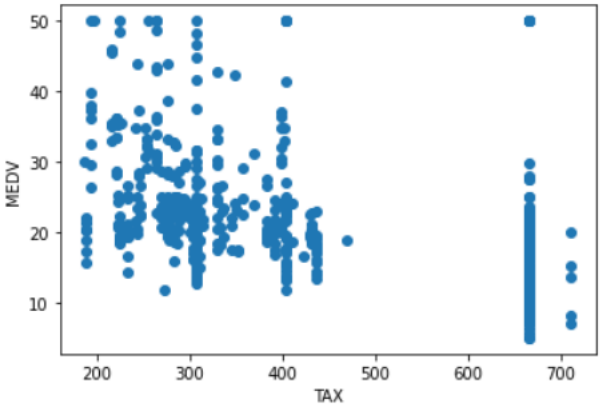
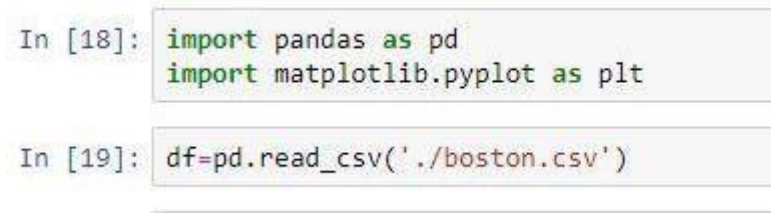


Fig 4.2.7: TAX VS MEDV

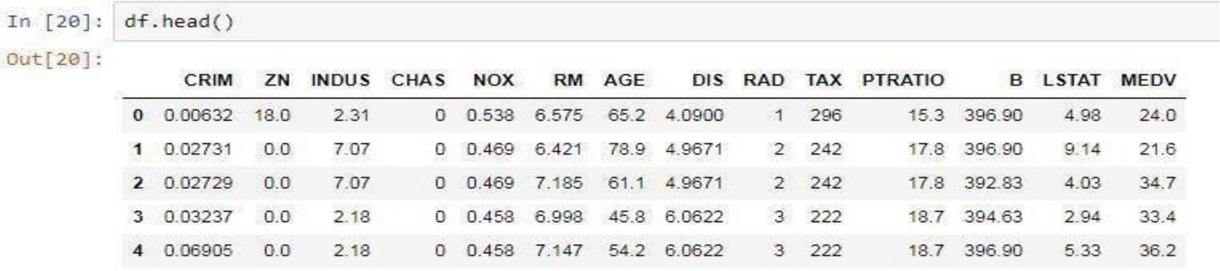
- TAX VS MEDEV - This graphs depicts the relation between the full-valued property tax per $10,000 along the x axis and the median value of owner-occupied(target variable) homes along the y axis.

4.3 SNAPSHOTS



```
In [18]:  import pandas as pd
          import matplotlib.pyplot as plt

In [19]:  df=pd.read_csv('./boston.csv')
```

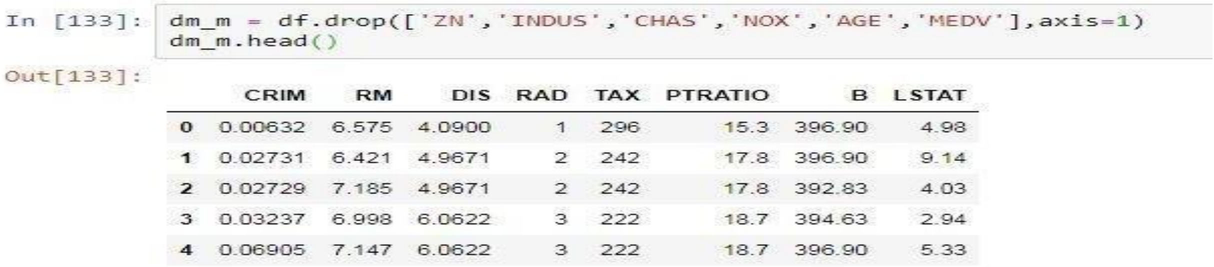
Fig 4.3.1: Importing Dataset



```
In [20]:  df.head()
Out[20]:
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

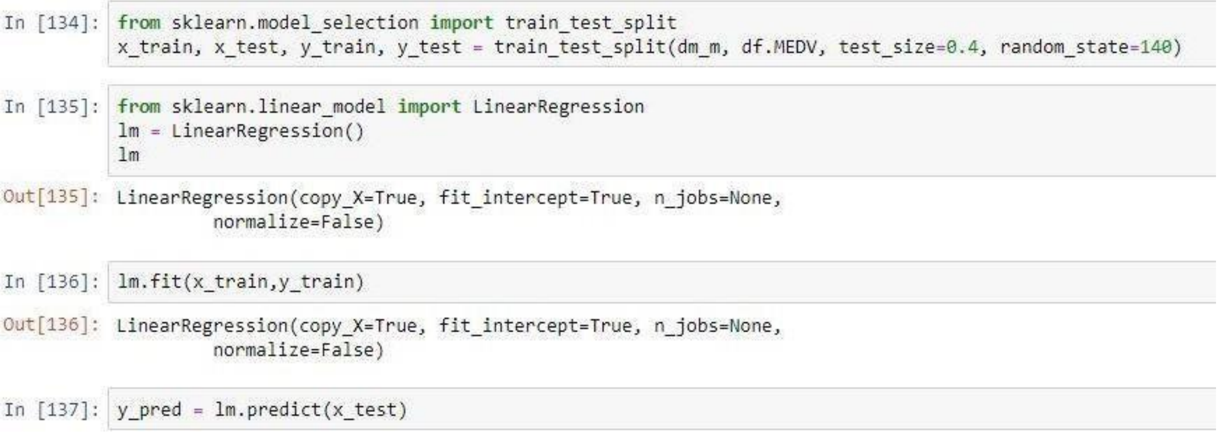
Fig 4.3.2: Sample Dataset



```
In [133]:  dm_m = df.drop(['ZN','INDUS','CHAS','NOX','AGE','MEDV'],axis=1)
           dm_m.head()
Out[133]:
```

	CRIM	RM	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	6.575	4.0900	1	296	15.3	396.90	4.98
1	0.02731	6.421	4.9671	2	242	17.8	396.90	9.14
2	0.02729	7.185	4.9671	2	242	17.8	392.83	4.03
3	0.03237	6.998	6.0622	3	222	18.7	394.63	2.94
4	0.06905	7.147	6.0622	3	222	18.7	396.90	5.33

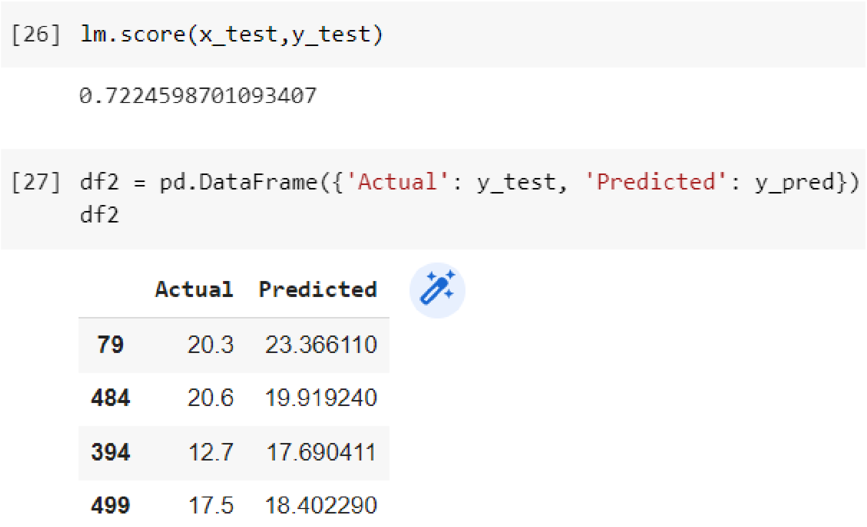
Fig 4.3.3: Sample Dataset after cleaning



```
In [134]:  from sklearn.model_selection import train_test_split
           x_train, x_test, y_train, y_test = train_test_split(dm_m, df.MEDV, test_size=0.4, random_state=140)

In [135]:  from sklearn.linear_model import LinearRegression
           lm = LinearRegression()
           lm

Out[135]:  LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                    normalize=False)

In [136]:  lm.fit(x_train,y_train)

Out[136]:  LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                    normalize=False)

In [137]:  y_pred = lm.predict(x_test)
```

Fig 4.3.4: Regression model

4.4 RESULTS

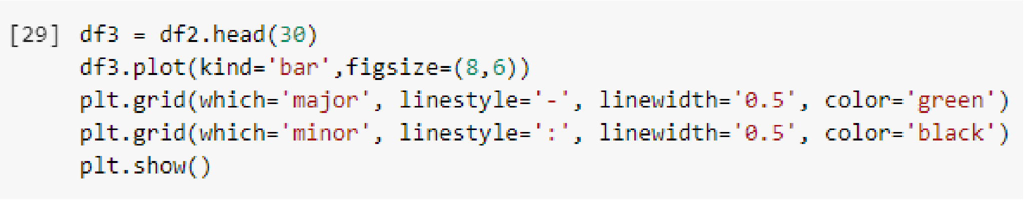


```
[26] lm.score(x_test,y_test)

     0.7224598701093407
```

```
[27] df2 = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
     df2
```

	Actual	Predicted
79	20.3	23.366110
484	20.6	19.919240
394	12.7	17.690411
499	17.5	18.402290

Fig 4.4.1: Coefficients and Accuracy



```
[29] df3 = df2.head(30)
     df3.plot(kind='bar',figsize=(8,6))
     plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
     plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
     plt.show()
```

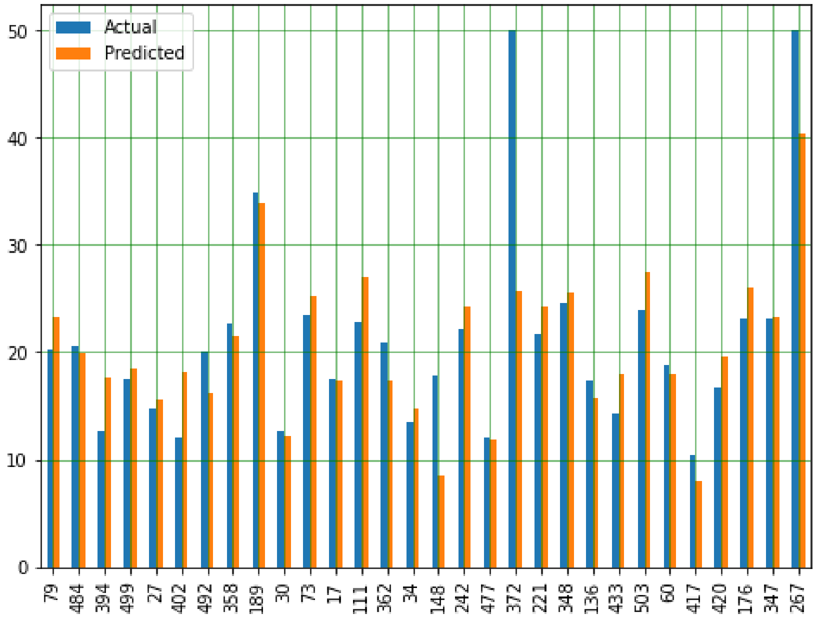


Fig 4.4.2: Variation in Actual and Predicted Values

CHAPTER 5

CONCLUSION & FUTURE WORK

We achieved the goals that we had set for this project. We have gained the complete information about the linear regression algorithm. We initially explored the dataset given, observe the features and get relative relationships between them, Plotted charts and graphs to perform data analysis. And finally developed an optimal data model, which satisfies all the requirements and predicts the price of the house accurately.

In this project, we analyzed the Boston House pricing dataset. We found that predicting the price of a house is highly dependent on the median value of the owner-occupied homes, the pupil-teacher ratio by town, the percentage of people with lower status, and the average number of rooms per dwelling. Linear Regression algorithm is used to find the result. We can use another kind of machine learning algorithm to implement this project. Implement GUI for this project. To develop improved algorithms and data capturing sensors to reduce the level of failure, to enroll and failure to acquire rate. To make it easily accessible to those in need of the information. Understand and implement the same for prediction of prices in general and not specific to Boston.

REFERENCES

- Machine Learning by Tom Mitchell
- Machine Learning algorithms at www.medium.com
- Matplotlib Pyplot at www.w3schools.com
- Linear Regression Analysis at www.ncbi.nlp.nih.gov.in
- https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in- python with-scikit-learn-83a8f7ae2b4f
- Dr. Patrick MD, PhD; Dr. Thomas R. MD, MPH; Linear Regression in Medical Research.https://journals.lww.com/anesthesiaanalgesia/fulltext/2021/01000/linear_regression_in_medical_research.18.aspx
- Dr.. Khushbu Kumari**,** Dr. Suniti Yadav, Linear Regression Analysis Study https://www.j-pcs.org/article.asp?issn=2395-5414;year=2018;volume=4;issue=1;spage=33;epage=36;aulast=Kumari