Data Science Project Training Report

on

Machine Learning Domain Projects for Regression, Classification and Clustering using Various Datasets

BACHELOR OF TECHNOLOGY

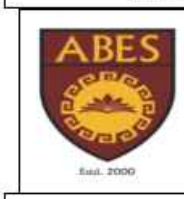
Session 2021-22
in

Information Technology

By
CHANDAN SHUKLA
2000321540023

AATIF JAMSHED
ASSISTANT PROFESSOR

DEPARTMENT OF INFORMATION TECHNOLOGY
ABES ENGINEERING COLLEGE, GHAZIABAD



AFFILIATED TO
DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW
(Formerly UPTU)

Student's Declaration

I / We hereby declare that the work being presented in this report entitled **"BOSTON HOUSE PRICE PREDICTION"** is an authentic record of my / our own work carried out under the supervision of Mr. **AATIF JAMSHED, Assistant Professor, Information Technology.**

Date:

Signature of student
(Name: CHANDAN SHUKLA)
(Roll No. 2000321540023)
Department: Information Technology

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

Signature of HOD
Dr. Amit Sinha

Information Technology

Signature of Teacher
Aatif Jamshed

Assistant Professor
Information Technology

Date:.....

Table of Contents

INTRODUCTION

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses in Boston's suburbs.

Once we get a good fit, we will use this model to predict the monetary value of a house located at the Boston's area.

A model like this would be very valuable for a real state agent who could make use of the information provided in a daily basis.

For the prediction Model we are going to use Regression (A Supervised Learning method).

REGRESSION

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.

The two basic types of regression are simple linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome.

Regression can help finance and investment professionals as well as professionals in other businesses. Regression can also help predict sales for a company based on weather, previous sales, GDP growth, or other types of conditions.

ABOUT THE DATASET

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive Data, and has been used extensively throughout the literature to benchmark algorithms. However, these comparisons were primarily done outside of Delve and are thus somewhat suspect. The dataset is small in size with only 506 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Dataset Naming

The name for this dataset is simply boston. It has two prototasks: nox, in which the nitrous oxide level is to be predicted; and price, in which the median value of a home is to be predicted.

Miscellaneous Details

- **Origin**
    1. The origin of the boston housing data is Natural.
- **Usage**
    1. This dataset may be used for Assessment.
- **Number of Cases**
    1. The dataset contains a total of 506 cases.
- **Order**
    1. The order of the cases is mysterious.

Variables:

There are 14 attributes in each case of the dataset. They are:

1. CRIM - per capita crime rate by town

2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per 10,000 dollars.
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in 1000's dollars

PROJECT EXPLAINATION

Data Collection

- It was obtained from the StatLib archive Data, and has been used extensively throughout the literature to benchmark algorithms. However, these comparisons were primarily done outside of Delve and are thus somewhat suspect. The dataset is small in size with only 506 cases.
- The dataset is collected from Kaggle.

Loading the collected data

- The CSV data is loaded with the help of read_csv method in pandas library.
- The dataset consists of 506 samples and 14 features with 1 prediction feature

```
# Initializing column names
columns = ['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS',
'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV']


# Loading Boston Housing Dataset
boston = pd.read_csv('../data/housing.csv', delimiter=r"\s+", names =
columns)
```

Let's see the number of samples and factors

```
# TODO : Let's know how many factors of an individual and Number of
Samples
print("The Boston housing Price Prediction Dataset has")
print("\t\tNumber of Factors : \t", boston.shape[1] - 1)
print("\t\tNumber of Samples : \t", boston.shape[0])



OUTPUT :
The Boston housing Price Prediction Dataset has
              Number of Factors :   13
              Number of Samples :   506
```

Feature Engineering

Let's check for null values.

```
# TODO : Check for null values and visualizing it using heatmap
boston.isnull().sum()
OUTPUT :
CRIM        0
ZN          0
INDUS       0
CHAS        0
NOX         0
RM          0
AGE         0
DIS         0
RAD         0
TAX         0
PTRATIO     0
B           0
LSTAT       0
MEDV        0
dtype: int64
```

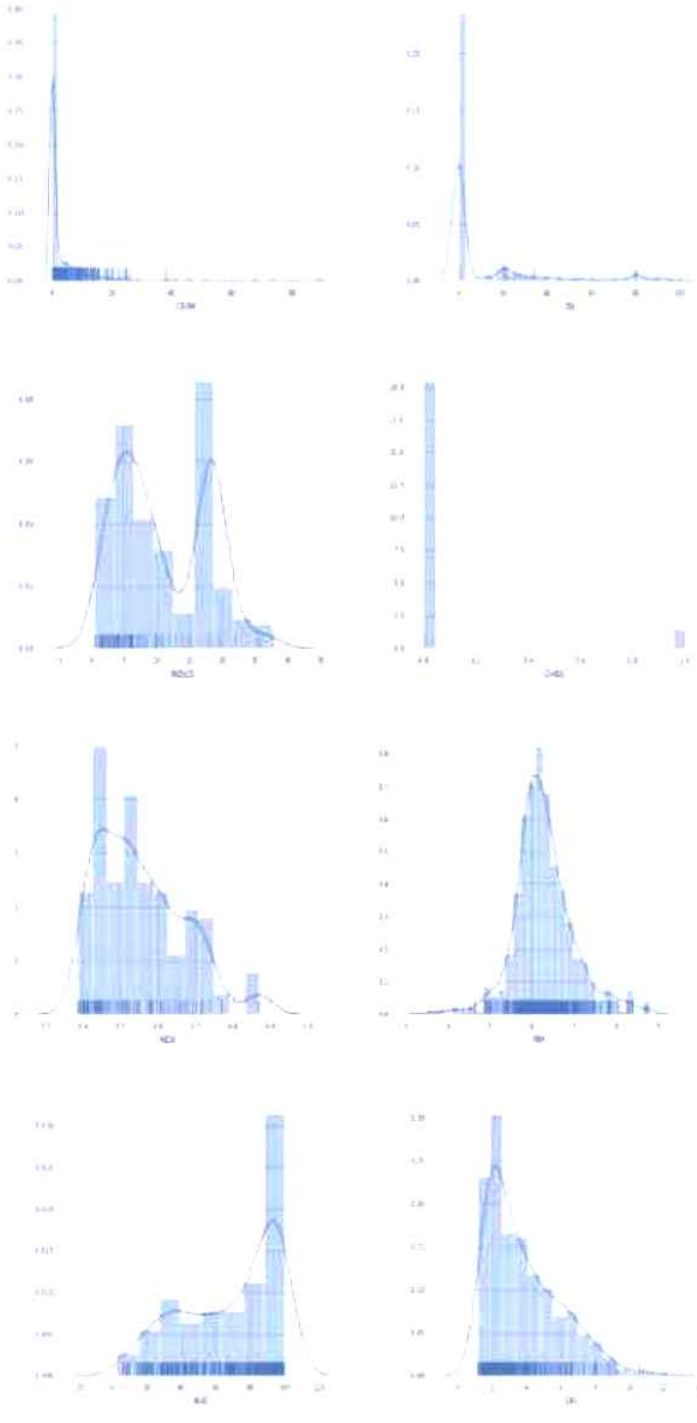
Let's check for Datatypes of each features.

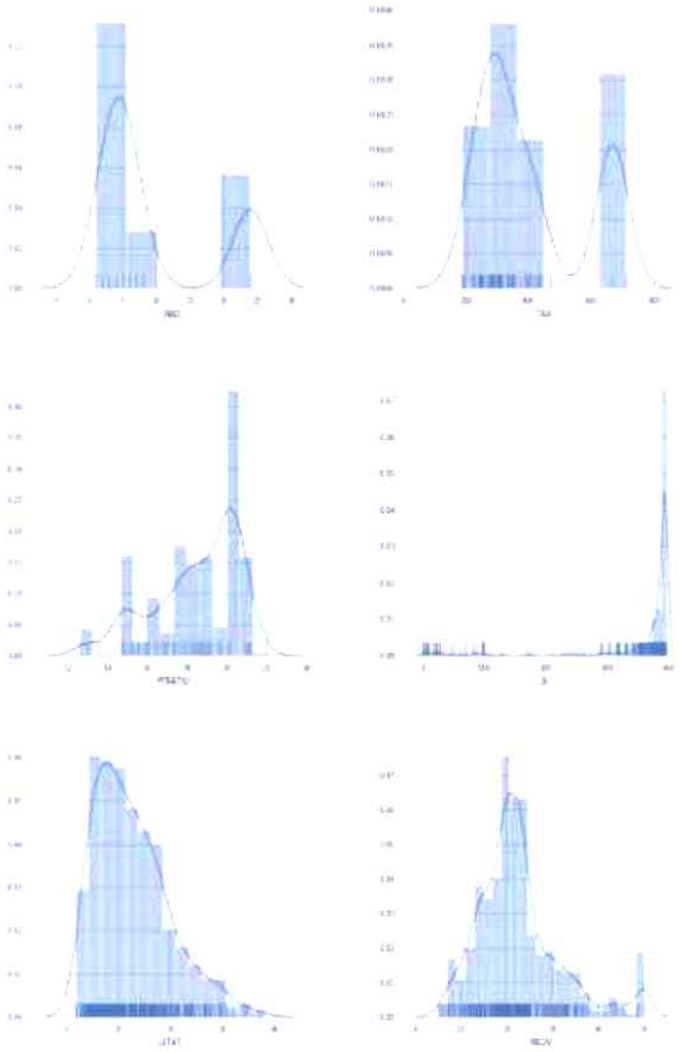
```
# TODO : Let's check for data types of all the columns
boston.dtypes

OUTPUT :

CRIM        float64
ZN          float64
INDUS       float64
CHAS          int64
NOX         float64
RM          float64
AGE         float64
DIS         float64
RAD           int64
TAX         float64
PTRATIO     float64
B           float64
LSTAT       float64
MEDV        float64
dtype: object
```

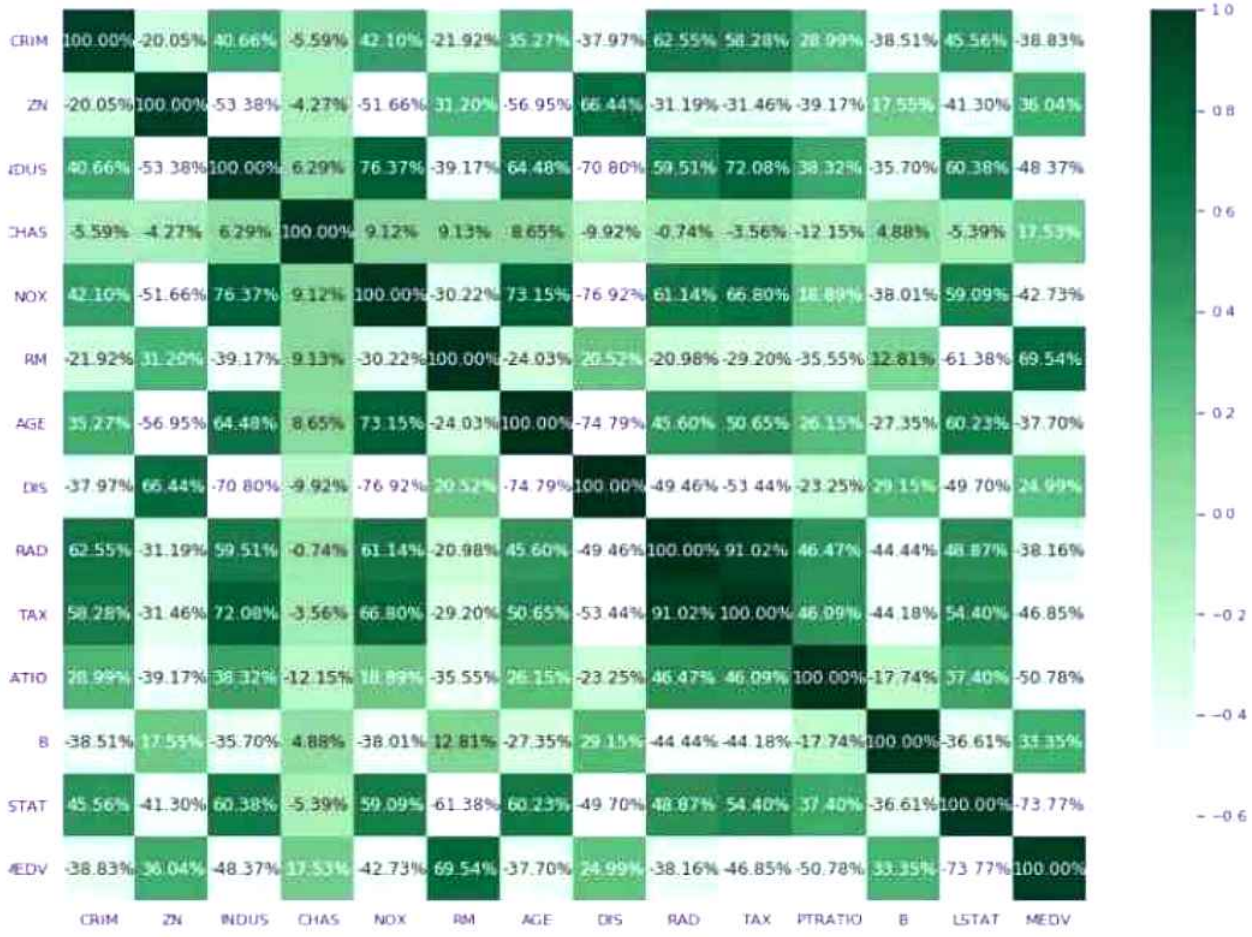
Though all the features are numerical values, we are going to see the distribution of data and some are categorical also.





FEATURE OBSERVATION AND SELECTION

-->Correlation Matrix



-->Important Features with their Importance

LSTAT 0.737663
RM 0.695360
PTRATIO 0.507787
INDUS 0.483725
TAX 0.468536
NOX 0.427321
CRIM 0.388305
RAD 0.381626
AGE 0.376955
ZN 0.360445
B 0.333461
DIS 0.249929
CHAS 0.175260

Steps Performed

- Data Exploration
- Exploratory Data Analysis
- Feature Observation
- Building Different Models
- Evluating and Selecting Optimal Model
- Cross Validation
- Making Predictions

CONCLUSION

By analysing historical data for house prices in Calgary along with various relevant features, we established some interesting patterns and trends. Using machine learning techniques, we were then able to identify a subset of the original features that are in a sense sufficient to describe our data. Having selected the most important features, we then trained a model for house price prediction. This model can therefore be used to predict the price of houses basis on various features.