Machine Learning Report

Yongxing NIE
College of Engineering
Northeastern University
Toronto, ON
Nie.yo@northeastern.edu

Abstract

The objective of this project is to first visualize all the data, find hidden information through analysis, and then use the dataset to train three machine learning models according to the characteristics of the data, and determine the most suitable model for the data set according to the MSE of the model, and finally apply the model to predict the price of the house.

1. Introduction

In this project, I study the data of Boston house price in which I use different libraries like Numpy, Pandas, Matplotlib, and different machine learning algorithms. I study different columns of the table and try to correlate them with house prices and find a relation between them. I try to find and analyze those key factors like area, crime rate etc., which helps people in Boston area to enhance their decision and vision when buying a property.

2. Experiments

1.0 Datasets

The Boston Housing Dataset can be downloaded here: http://lib.stat.cmu.edu/datasets/boston.

The dataset is derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The dataset has in total 506 entries and 14 features. Features include:

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

Table 1: The first 6 rows of the dataset.

	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

1.1 Data cleaning

According to the original data description[1], the max value of MEDV seems to be censored at 50.00 (corresponding to a median price of $50,000). Based on that, values above 50.00 may not help to predict MEDV. Therefore, I removed cases with prices above $50,000.

From the distribution information, it is found that ZN are 0s in its minimum, 25%, and 50%. At the same time, CHAS are 0s in its minimum, 25%, 50%, and 75%. These demonstrate that ZN and CHAS, albeit conditional or categorical, are somehow not suitable to predict the house price MEDV. But I will not manually clean it, I will let the recursive features elimination and cross validation to decide whether to remove it or not.

Table 2: The distribution statistics of the dataset.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	506	506	506	506	506	506	506	506	506	506	506	506	506	506
mean	3.614	11.364	11.137	0.0692	0.5547	6.285	68.575	3.795	9.549	408.237	18.456	356.674	12.653	22.533
std	8.602	23.322	6.8604	0.2540	0.1159	0.703	28.149	2.106	8.707	168.537	2.1649	91.295	7.141	9.197
min	0.006	0	0.46	0	0.385	3.561	2.9	1.130	1	187	12.6	0.32	1.73	5
25%	0.082	0	5.19	0	0.449	5.886	45.025	2.100	4	279	17.4	375.378	6.95	17.025
50%	0.257	0	9.69	0	0.538	6.209	77.5	3.207	5	330	19.05	391.44	11.36	21.2
75%	3.677	12.5	18.1	0	0.624	6.624	94.075	5.188	24	666	20.2	396.225	16.955	25
max	88.97	100	27.74	1	0.871	8.78	100	12.127	24	711	22	396.9	37.97	50

The distribution histogram shows that CRIM, ZN, CHAS, and B are highly skewed. MEDV is normally distributed. Other features are either normally distributed or bimodal distributed.

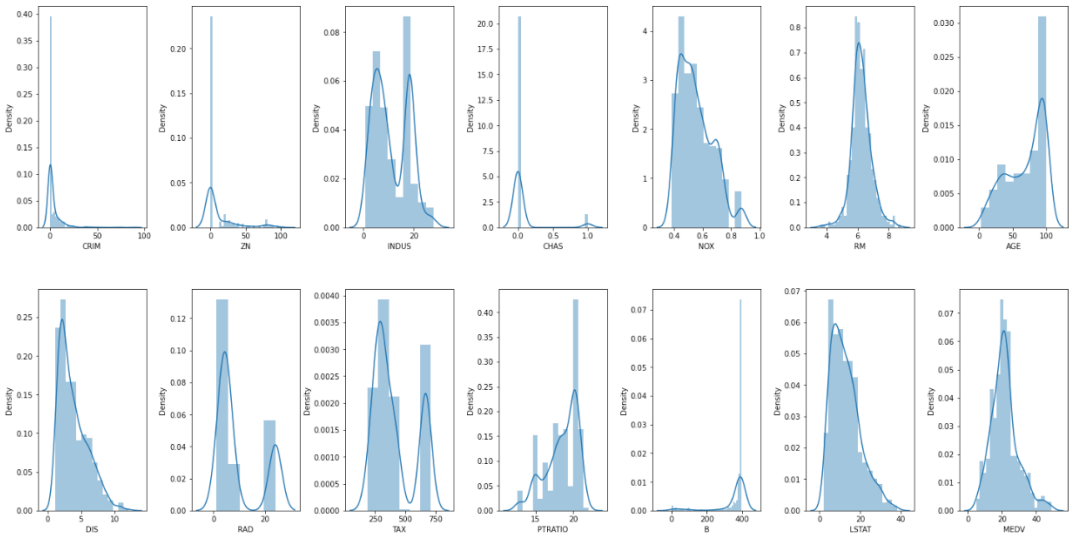

Figure1: The histogram distribution of each feature.

[1] http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

Outlier analysis shows that there are 8 features owning outliers. Since the 3 algorithms I used today are robust to outliers, I will not remove outliers this time.

```
Column CRIM outliers = 13.04%
Column ZN outliers = 13.44%
Column INDUS outliers = 0.00%
Column CHAS outliers = 100.00%
Column NOX outliers = 0.00%
Column RM outliers = 5.93%
Column AGE outliers = 0.00%
Column DIS outliers = 0.99%
Column RAD outliers = 0.00%
Column TAX outliers = 0.00%
Column PTRATIO outliers = 2.96%
Column B outliers = 15.22%
Column LSTAT outliers = 1.38%
Column MEDV outliers = 7.91%
```

1.2 Data characteristics

The attributes correlation graph tells that LSTAT is strongly negatively correlated with MEDV. It has a correlation score 0.76. And RM is highly positively correlated with MEDV with a score 0.69. The rest features are weakly correlated with MEDV.

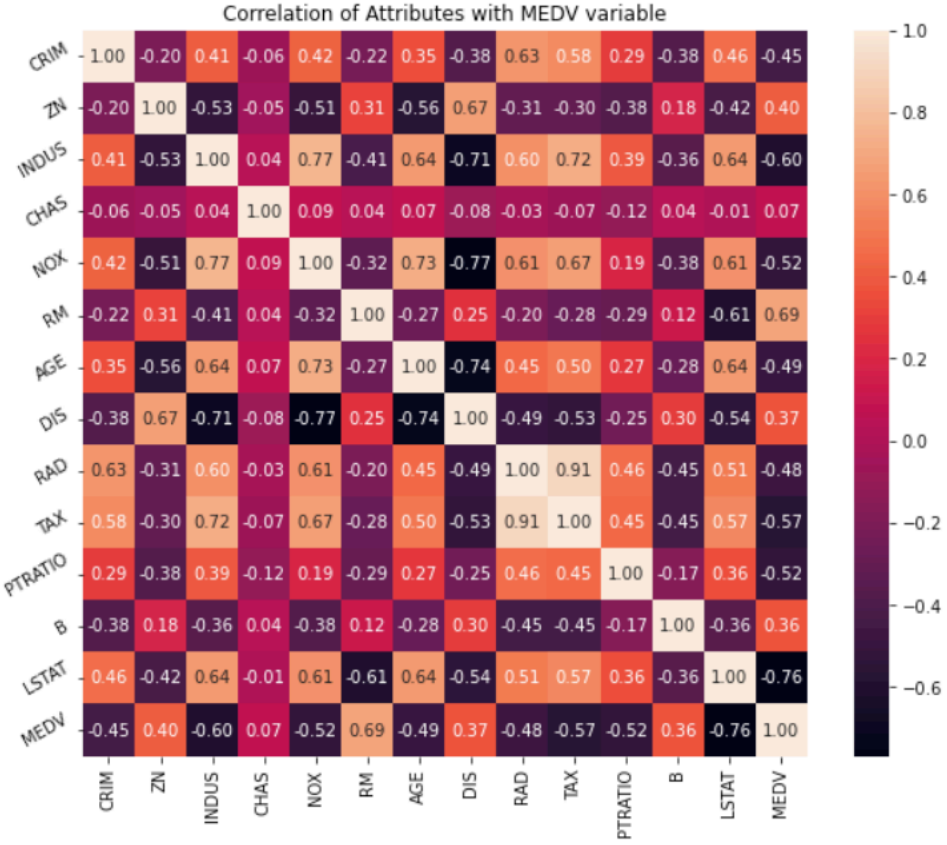

Figure 2: Correlation of Attributes with MEDV variable.

2.0 Methods

I applied three machine learning algorithms: Linear Regression, Decision Tree, and Random Forest, on the training data and then using the testing data to check the accuracy of each algorithm. I will choose my final prediction model according to their interpretability as well performance.

3. Results

In this report, I have analyzed the performance of 3 supervised learning algorithms on the Boston house price datasets.

Before training the model, I applied RFECV (recursing feature elimination cross validation) to eliminate the least important features to rebuild the model. I use Random Forest Classifier. The cross-validation repeats are 5. The table below are the elimination process. The iteration result shows that when the number of selected features is 8 or 4, the accuracy is above 0.94.

Table 2: Recursive Features Elimination Result.

Iterations	RandomForestClassifier cross validation score	Selected_features	Eliminated_features
1	0.852	10	3
2	0.948	8	5
3	0.884	6	7
4	0.940	4	9

Since when the number of selected features is 4 or 8, the cross-validation score is highest, I will train my model separately by 4 features and 8 features. The 4 features are CRIM, ZN, AGE, and LSTAT. The 8 features are CRIM, ZN, INDUS, NOX, AGE, TAX, PTRATIO, LATAT. The result that ZN is important in predicting the house price really surprises me.

After splitting the dataset into training and testing subsets, I trained 3 models by applying the 4 features and 8 features separately. The mean square error plot demonstrate that all 3 algorithms have small MSE.

Table 2: Scores of Different Algorithms.

	Linear regression	Decision Tree	Random Forest
n_selected features=4			
MSE	0.510	0.263	0.248
n_selected features=8			
MSE	0.582	0.263	0.262

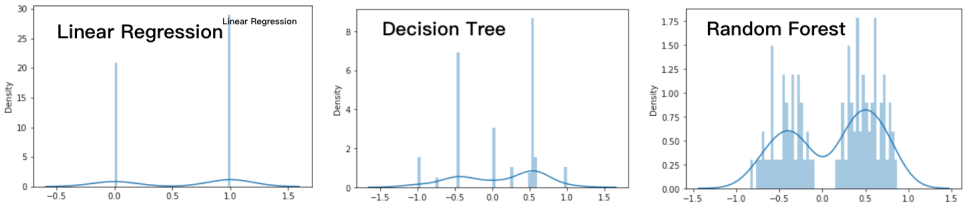

Figure 3: Bar Plot of Mean Square Error.

4. Conclusions

In this project, I analyzed and processed the dataset, trained it with 3 different algorithms. According to the above analysis, 3 algorithms all have small MSE, each of them can be used to predict the house price. And we can use 4 features or 8 features to build the prediction model.

Since before the training, I ignored the skewness of some features. In the future, I want to explore more that if log-transformed the skewed data, whether the performance will be optimized or not.

Acknowledgments

The learning code is adapted from: https://github.com/Unnati0104/Uber-Data-Analysis/blob/main/Uber_Data_Analysis.ipynb and references therein.

References

[1] Harrison Jr, David, and Daniel L. Rubinfeld. "Hedonic housing prices and the demand for clean air." Journal of environmental economics and management 5.1 (1978): 81-102.

[2] http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

[3] https://www.projectpro.io/article/uber-data-analysis-project-using-machine-learning-in-python/589.

[4] "HOUSING PRICE PROJECT REPORT" https://m2pi.ca/project/2020/bc-financial-services-authority/BCFSA-final.pdf.