HOUSE PRICE PREDICTION BASED ON RANDOM FOREST AND LINEAR REGRESSION

Abstract—This Paper contains the procedure of finding of predictions which is made to predict the price of a house. Linear Regression and Random Forest are used as a Machine Learning Methodologies. The Dataset used in this Procedure consists of the attributes that defines the price of a house. The dataset contains about 14 Dimensions and the dataset his features that is directly related to the price of the house. I assess the models and explain why they were specifically used and presents the final results. The Dataset used in this Problem was clean enough, but have need some data cleaning and data management, which I will discuss later in this Paper. This paper also contains the process how I reach to the conclusion based on inference and data analysis. The Literature Review of the related works is also presented in this report to expand the scope of discussion for the reader. An Evaluation of Both Random Forest and Linear Regression is also conducted to identify pros and cons of both models and more specifically the completion time of each models. Random Forests and the Linear Regression are both the example of the supervised Learning.

Keywords- *Random Forests, Linear Regression, House Price Prediction, Real Estate*

I. Introduction

Is from the name, this project is related to a real world business, Real Estate. In this Project I had used two most popular machine learning models, Linear Regression Model and the Random Forests Model. These model will help the real estate businesses to find the suitable price of the house based on their previous history. The Random Forest Model is a supervised Machine Learning Model. The Random Forest model will predict a suitable price for the house by looking to the different parameters of the House that are present in the Dataset that is previously collected from buying houses manually. This Project is the solution to the Real world problem, which many Real Estate businesses are facing. There are many factors that impact the price of a house, the most important factor is the Location of the House which we have included as different parameters. Other Parameters are also very effective like the population around the house.

Objectives

- ~ Finding the Effective Attributes that contribute to the price of a house.
- ~ To use two machine learning models to predict the price of the house.
- ~ Finding the correlation between the different features in the dataset by using the Data Analysis Techniques.
- ~ Visualizing the data for better understanding.
- ~ Highlighting the Possible Improvements for future data analysis.

II. Literature Review

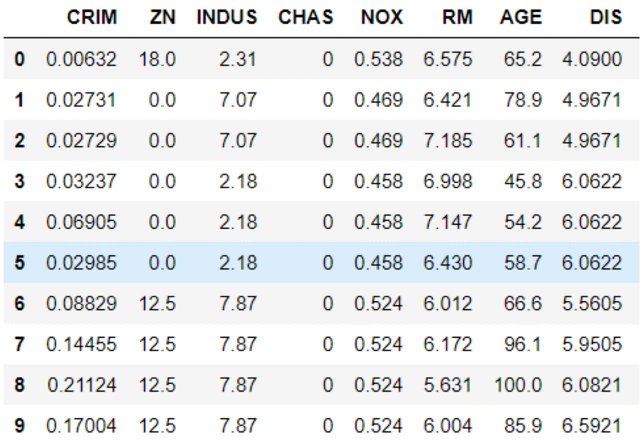
This section is dedicated to the literature review and a related discussion. The work discussed will be evaluated in relation to the research being investigated.

The report that I want to discuss is a journal [Journal of Environmental Economics and Management], Written by *David Harrison Jr.* and *Daniel L Rubin Feld* at the Science Direct. According to the authors, “This paper investigates the methodological problems associated with the use of housing market data to measure the willingness to pay for clean air. With the use of a hedonic housing price model and data for the Boston metropolitan area, quantitative estimates of the willingness to pay for air quality improvements are generated. Marginal air pollution damages (as revealed in the housing market) are found to increase with the level of air pollution and with household income. The results are relatively sensitive to the specification of the hedonic housing price equation, but insensitive to the specification of the air quality demand equation. This research was supported by the National Bureau of Economic Research. All statistical analyses were performed on the NBER Center for Computational Research's TROLL System. This Report is in support of our Project in the way of predicting the right price of the houses under the study of the Environmental Economics studies.

III. Data Management

Data Source: - This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. (b) Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

Data Description: - The Dataset have 14 attributes including the target attribute which is Price of the House. The shape of the dataset is 506x14. The attribute CRIM means per capita crime, ZN is the Proportion of residential land, INDUS means non-retail business acres per town, CHAS means Charles River Dummy variables (=1 if tract bounds river; 0 otherwise), NOX for nitric oxides concentration, RM for average number of rooms for dwelling, AGE for proportion of owner occupied units built Prior to 1940, DIS for weighted Distances, RAD means index of accessibilities to radial highways, TAX attribute is for full-value property-tax rate per $10,000, PTEATIO is for pupil-teacher ratio by town, LSTAT is the percentage of lower status of the population, MEDV is our target attribute which means Median value of owner-occupied homes in $1000's. Number of instances are 506. Missing attributes values are None.



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622
6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505
8	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921

External Libraries

Numpy:- A library that includes support for large multidimensional arrays and matrices, extensively used as it provides access to very large library of high-level math functions for operations on these arrays and matrices.

Pandas:- A popular data framework library in Python, one can manipulate and present data.

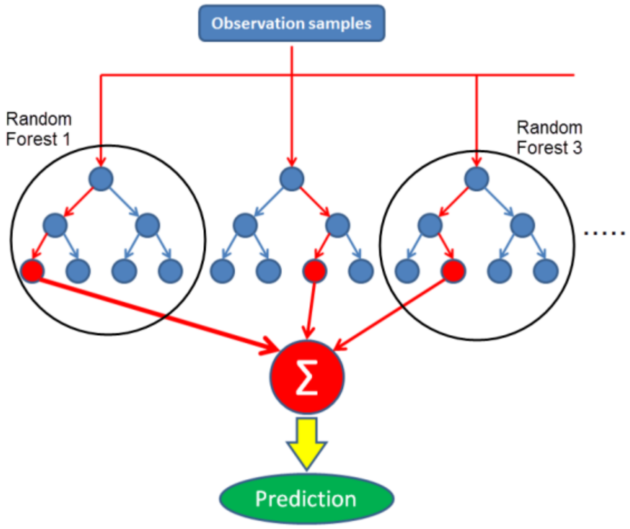
Scikit-learn:- A machine learning library in Python that is extensively used within Data Science and Computer Science. It was used in this project for the implementation of the algorithms as it is easy and straightforward to implement.

Seaborn and Matplotlib:- These libraries are for visualization of data, there are different types of graphs that can be developed using this library. The graph's axis and scales can be changed accordingly and some elements of customization in color are possible.

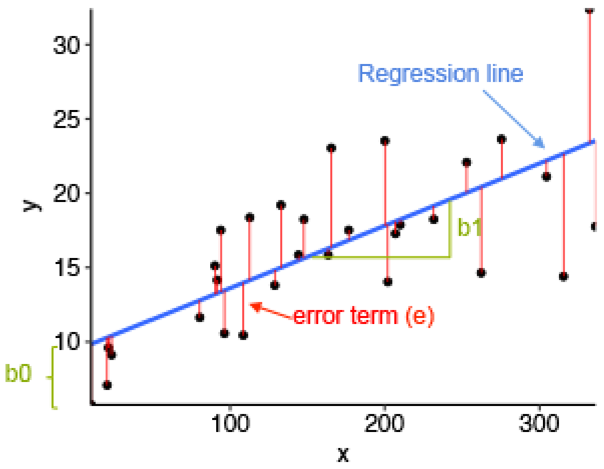
IV. Methodology

To achieve the results of our project I have used two models, The first one is Random Forest and the second one is Linear Regression. But along with the Linear regression I also used BinarytreeRegressor to check out the results. Below is the how and why of the models.

Random Forest: - For us one aims of the project was the prediction of the price of a house base on the attributes in the data. Random Forest is a supervised machine learning algorithm. A big part of machine learning is classification — we want to know what class an observation belongs to. The ability to precisely classify observations is extremely valuable for various business applications like predicting whether a particular user will buy a product or forecasting whether a given loan will default or not. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction



Linear Regression: - The second model that we have used in our project is Linear Regression. Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares.



The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

Random Forest Justification: - One of the Aim of this project is to predict the price of the house with high accuracy. Is I mentioned previously, Random Forests provide higher accuracy through cross validation. Upon till now our data set is perfect and there is not a single instance with missing values, but if there were any case Random forest can handle this without effecting the accuracy. I choose random forest because of its ability of scalability. Right now our data is limited and contains only 506 records, but in future if we have more records, for us it will not be problem and will help us improve our model because it can take care of large datasets as well. The scalability and flexibility of the Random forest is what it makes fit for our dataset.

Linear Regression Justification: - I choose the Linear Regression Model, because in our data set we have a lot of dependent and independent variable and this model will help us have the Relationship between these dependent and independent variables. Though this model does not help us the way random forest does but still it was good then other machine learning models. This model helped us discover the relationship between variables and it needed a statistical analysis to understand what the relationship is.

V. EXPLORATORY DATA ANALYSIS

We have clear understanding of our dataset we can now perform some experiments and conduct data analysis with our dataset so we can have a general idea of interrelationships.

By looking into the Number of rooms greatly impacts the price of the data. When there are more rooms in the house the price is respectively high. It means that the price is directly proportional to the number or rooms in house. The column that denotes the number of room is named is RM. To grab the RM column we us the index of the "RM" with our DataFrame. To return the series that contains the different values which our dataset contains for the "RM" column. By calling the .sum function on the series it will return the sum of the repeating values in our series.

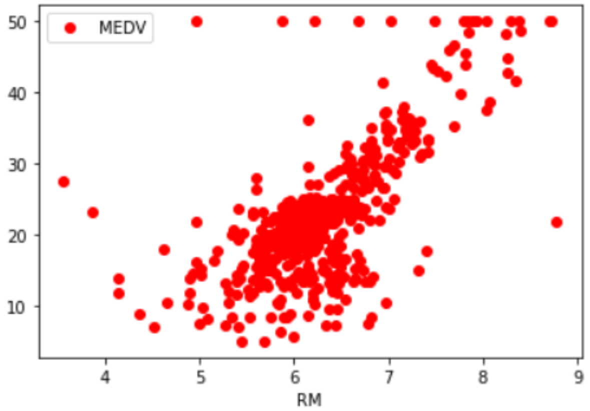


Figure 1Plot between Price and RM

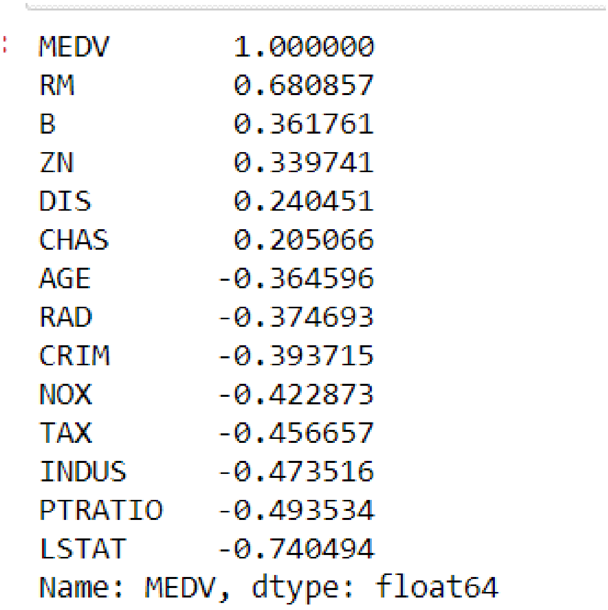
To plot the result of the Number of Rooms in a houses versus the price of the house we will use the method plot with the data frame to plot.

From the scatter plot above we can clearly see that the RM value highly effects the MEDV value. Although we have some values that are not matching to the pattern. These values are called outliers. Those outlier values act is the noise in our dataset and our machine learning model should be clever enough not to learn from these noise. If we have enough amount of noise in our data, we can go to the data source providers and can manually do queries about these outlier's data. But in our data these values are very less in amount so we just ignore them.

So far we have looked it only one relationship between our data attributes which is not enough for us. In order to see other relationships in our data set we can again look into the correlation in our data. The strong correlation in our data set is with number of Rooms which is represented by the "RM" column. This is the strong ever correlation.

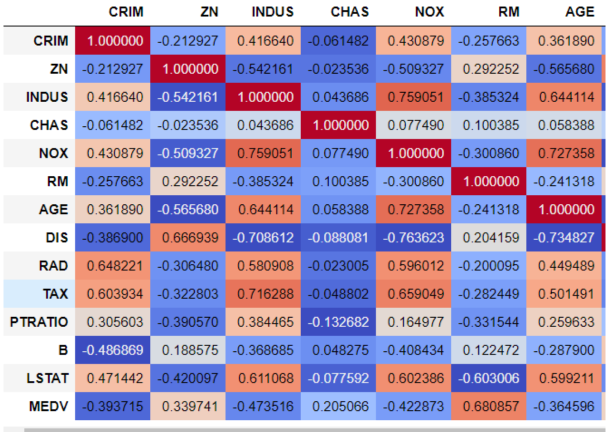
In the list of the correlation table between the different attributes other factorS that effect our Price is the crime rate in the area which is represented by the column named is "CRIM".

The correlation table which is obtained with the help of correlation matrix.



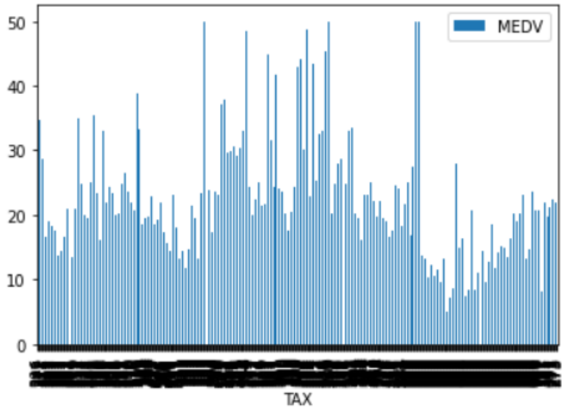
```
MEDV       1.000000
RM         0.680857
B          0.361761
ZN         0.339741
DIS        0.240451
CHAS       0.205066
AGE       -0.364596
RAD       -0.374693
CRIM      -0.393715
NOX       -0.422873
TAX       -0.456657
INDUS     -0.473516
PTRATIO   -0.493534
LSTAT     -0.740494
Name: MEDV, dtype: float64
```

In order to show the better understanding we can visualize the result with the seaborn library.



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
CRIM	1.000000	-0.212927	0.416640	-0.061482	0.430879	-0.257663	0.361890
ZN	-0.212927	1.000000	-0.542161	-0.023536	-0.509327	0.292252	-0.565680
INDUS	0.416640	-0.542161	1.000000	0.043686	0.759051	-0.385324	0.644114
CHAS	-0.061482	-0.023536	0.043686	1.000000	0.077490	0.100385	0.058388
NOX	0.430879	-0.509327	0.759051	0.077490	1.000000	-0.300860	0.727358
RM	-0.257663	0.292252	-0.385324	0.100385	-0.300860	1.000000	-0.241318
AGE	0.361890	-0.565680	0.644114	0.058388	0.727358	-0.241318	1.000000
DIS	-0.386900	0.666939	-0.708612	-0.088081	-0.763623	0.204159	-0.734827
RAD	0.648221	-0.306480	0.580908	-0.023005	0.596012	-0.200095	0.449489
TAX	0.603934	-0.322803	0.716288	-0.048802	0.659049	-0.282449	0.501491
PTRATIO	0.305603	-0.390570	0.384465	-0.132682	0.164977	-0.331544	0.259633
B	-0.486869	0.188575	-0.368685	0.048275	-0.408434	0.122472	-0.287900
LSTAT	0.471442	-0.420097	0.611068	-0.077592	0.602386	-0.603006	0.599211
MEDV	-0.393715	0.339741	-0.473516	0.205066	-0.422873	0.680857	-0.364596

The relationship between the price of the house and the TAX. When there is increase in the TAX value the price significantly decreases. We can have the graph of the TAX and the MEDV to show the relationship.



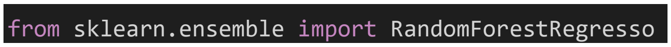
This is the relationship of the TAX and the MEDV. We can see that there are some irregularities but still it shows us quite good pattern between the price of the house and the tax.

VI. TESTING AND RESULTS

This section is dedicated for the test and result of our models. Random Forests and Linear Regression showed good results to us. But still there is some difference in the accuracy of the two models. Both the model predicts us the price of the house though the accuracy of the Linear Regression was not quite impressive but still was considerable. For this problem, looking into the dataset the random forest showed us very great result in predicting the price of the house. The models successfully returned us the suitable price of the property based on the previous records and this is how a supervised learning model works.

Random Forest Results

By using the Pipeline with the dataset we have created an easy to use scenario where every model can be fit on the data in a very simple way. Python is a rich language with different models, I used scikit learn library to import the Random Forest. By using the following command, I got accessed to the Random Forest algorithm.

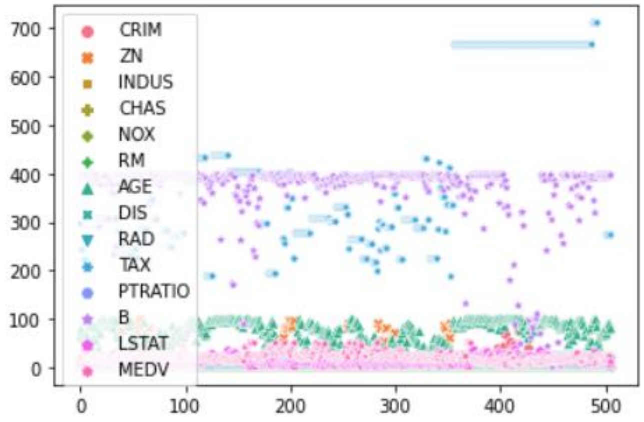


```
from sklearn.ensemble import RandomForestRegresso
```

Random forest algorithm gives us very good result as expected the RMSE score of the Random Forest was 1.207 which is more then expected. RMSE is the square root of the variance of the residuals. And this value indicated the absolute fit of the model to the data-how close the observed data points are to the model's predicted values.

From the test data and the predicted data there is a very close relationship.

From the data of the all features we have the following graph.



Linear Regression Results

Although Linear Regression model is very popular for modelling a target value based on independent predictors. But in our case it didn't give us the expected results is by the Random Forest model. The MSE score of the model was 23.38 while the RMSE score was 4.38 which is very high is compared to the Random Forest.

I have tried the DecisiontreeRegressor as well which overfit on the data and followd the whole graph. While the Linear Regression Model was good in predicting but the RMSE score was very demotivating in this case. The Mean value of the Linear Regression was 5.03 and the Standard deviation of the Linear Regression was 1.0594.

To use the Linear Regression model as like the Random Forest Resgression I just have had to import it from the scikit learn library.

from sklearn.ensemble import RandomForestRegressor

d_model = LinearRegression()

With the help of scalling the values in the dataset every model can now easily be fit very easily like the Linear Regression and the Random Forest.

VIII. Conclusions

All the Objectives of the project have been achieved, By looking into the objective section all of our aims are achieved successfully. We were looking to predict the price of a house and our model can now predict the price of the house once someone pass the features of the house. Another objective was to find the attributes that effect the price of the house and we have find it and visualize it and find it was RM and CRIM which was effecting the price of the house more then any other attributes. We have used two models as per requirements and we also find the correlation between the data as this was one of the objective of our Project.

We visualize the data by using the Python Libraries which was quite helpful in achieving our results.

The Size of the dataset will affect the model accuracy. If we increase the size of the dataset by having more data related to the real estate our model is scalable and flexible and will predict the price more accurately which will help the real world real estate business and will help the real estate business to find the good price for the house instead of manually calculating and thinking about the price.

VIII. REFERENCES

[1]. Journal of Environmental Economics and Management.
URL: https://www.sciencedirect.com/science/article/abs/pii/0095069678900062

[2]. Python Machine Learning and Deep Learning with Python, scikit-learn.
URL:https://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow/dp/1789955750

[3]. Real Estate Dataset
URL: http://lib.stat.cmu.edu/datasets/

[5]. Real Estate Market Analysis

URL: https://www.amazon.com/Real-Estate-Market-Analysis-Information

[4]. Python Data Analysis Library

URL: https://www.pandas.pydata.org

[6]. Python for Data Analysis
URL: https://www.oreilly.com/library/view/python-for-data/9781491957653/

[7]. Artificial Intelligence in Real estate

URL: https://www.amazon.com/Artificial-Intelligence-Real-Estate-Investing-ebook/dp/B07N89FWPQ

[8]. The Machine Learning workshop

URL: https://www.amazon.com/Machine-Learning-Workshop

[9]. Data Driven Decisions

URL: https://www.amazon.com/Marketing-Analytics-Essential-

[10]. Data visualization with Python.

URL:https://www.amazon.com/Data-Visualization-Python

IX. APPENDICIES

I have used Jupyter notebook is a code editor for this project and all the notebooks used in this project can be accessed form the following links.

I have used Python 3 along with Numpy, pandas, Scikit learn, Matplotlib and Seaborn.

Click here to access the Notebook that contain the whole code for achieving the results of the project.

URL: https://drive.google.com/file/d/1ntLqtHm3qfIh0LLB_XUCWdqfgmTr-qsz/view?usp=sharing