Building and Training a Supervised Machine Learning Model using Scikit- Learn for Elevating Business Sales

Jayesh Tembhekar[1], Jayesh Katkar [2], Darshan Dahekar[3], Prasad Daf[4]

[1-4]Student, Dept. of Computer Technology, K.D.K College of Engineering, Maharashtra, India

---***---

***Abstract -** This Machine Learning model will help to predict the real estate prices in the coming days. We will understand the problem of a real estate company from its CEO and then apply ML to solve it. Prices are predicted manually and often leads to faulty price tags, This ML model will take the real-world data and throw the well-tested results by going through various ML algorithm like Decision Tree, Linear Regression, Random Forest Regression. This project will walk you through analyzing the data, importing it to Jupyter notebook, will look at promising attributes from a cluster of data, finding out correlation, plotting graphs, also creating the pipeline, dealing with missing values, etc. In the end, this ML model will be used to predict the prices by giving the set of features at near to 0% error rate by going through RMSE. To make the ML model more robust, cross validation, train-testing splitting, stratified shuffle split, and sampling work in action.*

Key words: **Predictive Analysis, Real Estate ML model, Data Visualization, Batch Learning, Regression task, RMSE, Supervised Learning, Unsupervised Learning.**

1. INTRODUCTION

The world has picked up a pace and we have to set ourselves aligned to it. AI & ML in the future in every sector we can think of. It provides an unbiased testimony for humans. That is why Real Estate can fully depend on AI & ML operations.

Our project will start by collecting the raw datasets of Houses, Hotels, and other different real estates from Kaggle.com. Now we have to perform data cleaning to clean or correcting out the inaccurate records.

Like in the corporate business environment, the cleaned data has to be visualized by making graphs, histograms of housing data by library Matplotlib to make sense out of it.

Training and Testing split is dividing the dataset into two for testing the models to get the idea of the accuracy of the model.

To eliminate the repetitive pre-processing steps, we will build the pipeline that remembers steps and helps us to automate the continuous flow of the dataset.

At last, we will decide the desired model by implementing regression algorithms using Scikit-learn

We will use the RMSE as an evaluation metric to check the model performance.

Cross-checking the models build on regression algorithm and it will throw out the most trustable output by Random Forest Algorithm.

2. SCIKIT-LEARN: MACHINE LEARNING IN PYTHON.

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language. Emphasis is put on ease of use, performance, documentation, and API consistency.

Scikit-learn exposes a wide variety of machine learning algorithms, both supervised and unsupervised. Importantly, the algorithms, implemented in a high-level language.

It is an open-source Python library that implements a range of machine learning, pre-processing, cross validation and visualization algorithms using a unified interface.

2.1 Algorithm Used:

a) Decision Tree: Splitting the dataset based on different conditions.

b) Linear Regression: To find the relationship between the dependent value and independent value.

c) Random Forest Regression: To perform both regression and classification tasks from multiple ML models to make a more accurate prediction than a single model.

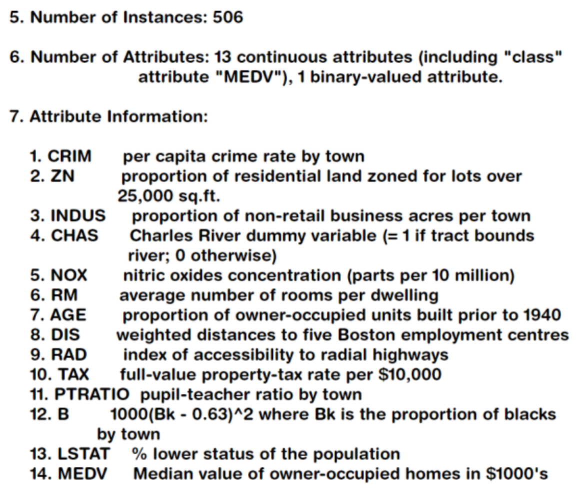
3. LIFECYCLE.

3.1 Obtain Raw Data.

We have obtained the data that we need from available data sources. We have gathered our housing data from UCI Machine Learning Repository. Data was divided into two separate files, we have made it organize by giving labels/features to their respective values.

a. Housing.names

b. Housing.data



5. Number of Instances: 506

6. Number of Attributes: 13 continuous attributes (including "class"
attribute "MEDV"), 1 binary-valued attribute.

7. Attribute Information:

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
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

Fig -1: Labels for each column of raw data.

3.2 Data Cleaning / Scrub.

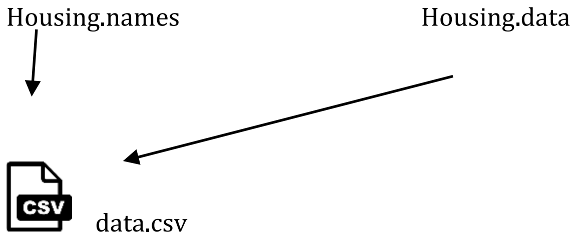


Fig – 2: Formation of final data for ML model.

3.3 Exploration Data Analysis (EDA).

EDA is an important initial step for any the knowledge discovery process, in which data scientists interactively explore unfamiliar datasets by issuing a sequence of analysis operations (e.g. filter, aggregation, and visualization).

3.4 Data Modelling.

a. Data Slicing:

This stage is all about building a model that best solves your problem. A model can be a Machine Learning Algorithm that is trained and tested using the data.

4. FUTURE SCOPE

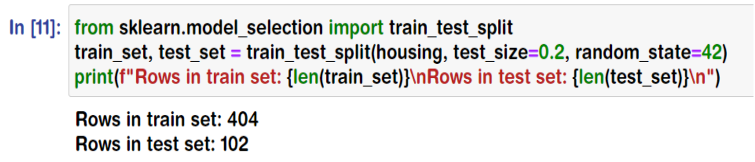
This predictive ML model can be applied for future housing datasets. As datasets available in a variety of labels, we have to make minute changes and pass the data as required. ML model will tend to work as effectively as the first time. So, it has a great future scope of application.

This stage always begins with a process called Data Splicing, where you split your entire data set into two proportions.

One for training the model (training data set) and the other for testing the efficiency of the model (testing data set).

This is followed by building the model by using the training data set and finally evaluating the model by using the test data set.

Train Set	Test Set
80%	20%



```
In [11]: from sklearn.model_selection import train_test_split
         train_set, test_set = train_test_split(housing, test_size=0.2, random_state=42)
         print(f"Rows in train set: {len(train_set)}\nRows in test set: {len(test_set)}\n")

         Rows in train set: 404
         Rows in test set: 102
```

Fig-3: sklearn code snippet for data slicing.

5. CODING ENVIRONMENT.

a) Python 3.7.6: High level, interpreted general-purpose programming language.

b) Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, visualizations, machine learning, data cleaning, narrative text, and much more.

5.1 Packages / Libraries:

a) Matplotlib: It is a comprehensive library for creating static, animated, and interactive visualizations in python.
b) Numpy: To add support for large, multi-dimensional arrays and matrices along with mathematical functions.

6. CONCLUSION

As the world is hit by Covid-19, it will be a great challenge for the business sector to come back in the mainstream. Real Estate is all time in-demand business thus our ML model will be a great help for elevating sales by giving well-deserved capital to real estate. The inclusion of ML and AI in business is a 21st-century need. Consumers will thrive for a fair deal in purchasing, to create an ease of doing business mindset, AI & ML is the future.

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of my guides, and teachers. My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

REFERENCES

[1] Chenchen Fan, Zechen Cui, Xiaofeng Zhong., "House Prices Prediction with Machine Learning Algorithms", ICMLC 2018: Proceedings of the 2018 10th International Conference on Machine Learning and Computing February 2018.

[2] T. Milo, Amit Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview", SIGMOD '20: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, June 2020.

[3] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel "Scikit-learn: Machine Learning in Python", November 2011, The Journal of Machine Learning Research.