iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34543

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Data Science upon Machine Learning - Apartment Price Prediction

Nusaiba K K[1], Silja Varghese[2]
[1]*M.Tech CSE,* [2]*Assistant Professor, Nehru College of Engineering and Research Centre, Pampady, Thrissur, Kerala, India*

Abstract: ***Nowadays problems and solutions in technology is mostly related to data. From a single individual to the whole world there comes an importance for data. Data mining, Data visualization, Data extraction etc. came into existence from this importance. Now we are in a world where everything is going to be automated. Artificial Intelligence, machine learning and deep learning are becoming more and more advance in these days. All these technologies paved the way to Data science. Data science is the study of data. The goal of Data Science is to gain insights and knowledge from any type of data. This paper proposes a step by step procedure of handling of data. The first step in data preparation is data collection. The data should be much large in size. After data collection the data should be loaded .When the data is been loaded successfully, it should be carefully analyzed in-order to find what all cleaning procedures are required to apply on it. Then the data should be cleaned by various methods. This step will include handling of missing values, removal of unnecessary features and after cleaning the data, the next step is feature engineering and dimensionality reduction. After that, the data is analysed to detect the outliers and they are removed efficiently. After Outlier removal unnecessary data's are also removed, because this is the last step of our data pre-processing. The next step after pre-processing the data is the machine learning model building, where we are going to predict the price of apartments. We are using k-fold-cross-validation and GridSearchCV to analyse the best algorithm as well as the best parameters. The main step of the proposed system is that we are going to analyse all the possible machine learning algorithms on this data set and thereby cross validate them to extract the best score to analyse the best algorithm of all the algorithms applied. The best algorithm that we found will be used to build the model.***

Keywords: Machine learning, Data-visualization, Data-cleaning, Feature Engineering, Outlier detection, Dimensionality reduction, K-Fold-Cross-Validation.

I. INTRODUCTION

"Data plays an important role in the current world. Whichever the situation or field we take, there comes an inevitable role of data. Today, in the current pandemic world too, we know that data is extremely useful. Data Science is the emerging technology which deals with the data to showcase and extract the real time usefulness of data. This paper proposes the methods of applying the concepts of data science upon a machine learning model. In Business, Health, Science, Education, and Finance and in every such field there exists an extraordinary presence of data. The raw data which is available to us is of no use. Because it is huge in size and difficult to interpret. In-order to find the inner meaning or the relevant information contained inside data, we have to work on data. This highlights the importance of data science.

II. METHODOLOGY

Raw data consists of many missing values, unnecessary values, and many other problems. In order to get relevant information from the data, have to study the data and extract the needed information from the data. Data is abundant in the present scenario but without data science the data is not useful as needed. Various steps involved in the proposed system are data collection, data analysis, data cleaning, feature engineering, dimensionality reduction, outlier detection and removal, model building, K-Fold-Cross-Validation and Grid-search-CV.

A. Data Collection.

Data is the resource we need to visualize the role of data science. Data is huge in size and found everywhere. But collecting useful and relevant data is a tedious process. There are various methods available to extract the data. Certain data are available for free and certain data can be bought for money.Kaggle.com is a platform where we can find datasets available for free. Data collected should be relevant to our problem set. Different forms of data are available today. This paper focuses on structured data. Mainly numerical data. Numerical data is the data which contains numbers as its data points.

B. Data Analysis.

The data is analysed carefully in-order to find the problems of the data. The data contains missing values, unnecessary values and Outliers. Sometimes it requires assumptions and sometimes it requires business logic and sometimes it requires technical knowledge to deal with the data.

C. Data Cleaning

The first step in data cleaning is handling the missing values. The python function “is null ()” can be used to find the null values present in the dataset. The rows with null values can be removed if they are negligible in size .Otherwise the missing data points can be substituted by the average of similar data points. Unnecessary columns can be removed in this stage. There will be so many numbers of columns in the dataset. Some data columns will be crucial to the prediction process. Some will be extremely irrelevant. Some will be relevant but can be removed if they are assumed to be irrelevant according to the problem statement. Each and every column of the dataset should be cleaned to make the column values consistent. Considering one column will lead us to various diversities and we have to deal with it accordingly. Sometimes one column will have values of different data types or of different forms. Some can be neglected and the rest have to be handled wisely. If the data is cleaned efficiently the model will be proportionally efficient.

D. Feature Engineering.

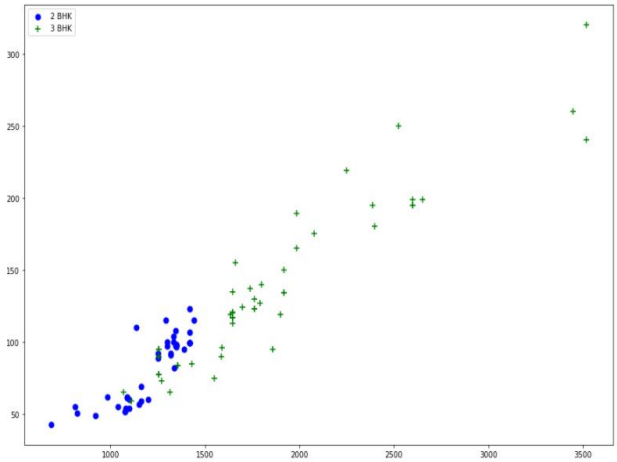
Feature Engineering is the process of extracting relevant feature from the dataset. The feature will be useful for model building or for any data pre-processing steps. The Feature we build for the purpose of data pre-processing can be dropped if it is no more in use. These features play an important role in data pre-processing.

E. Dimensionality Reduction.

Upon exploring the dataset, sometimes we can find columns of data with high dimensionality curse or high dimensionality problem. It will be hard to handle such data points. So in-order to avoid such an issue in the future we should analyse the data in the pre-processing stage itself. Rows will lesser number of data points can be grouped together to “Other” category. This is the most effective way to handle high dimensionality problem. In our proposed system there is a categorical data column which faces high dimensionality problem or dimensionality curse. The data points with lesser number of data points have been found by using python group-by function applied upon location column. The rows with lesser number of data points can be categorized into “Other” category.

F. Outlier Detection and Removal.

Outliers are those data points which deviate from the normal data points. Various data visualization techniques and various business logics can be used in finding outliers. Scatter plot is widely used for outlier detection. Various common business logic can also be used to find the outlier values from the data. After finding the outliers we can remove them from our dataset to make our dataset efficient as much as possible. Efficient dataset leads to efficient prediction model.



Fugure1.Scatter plot

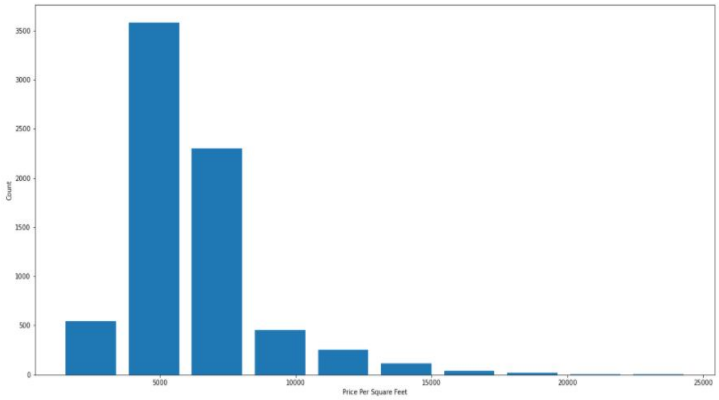


Figure 2.Histogram

G. Model Building

The model for predicting the price of the apartment can be built by using regression algorithms. The dataset is divided into training and testing dataset. Training data is used to train the model and testing data is used to test the model. In the proposed system 20 percentage is the testing data and 80 percentage is the training data. Linear regression, Lasso Regression, Decision Tree etc. can be used to build the model. But here comes the problem of finding the best algorithm for building the model. First we can build the model using Linear Regression and then move forward to the later data science concepts. The proposed model provides 85 percentage of accuracy on applying linear regression.

H. GridSearchCV

Python provides various tools and technologies for the data scientists, only we need is to use them efficiently. GridSearchCV is a python library used for finding the scores of the algorithms we provided. We can create a dictionary with the algorithms and their parameters and can provide it for finding the best algorithm and best parameters. Linear Regression, Lasso Regression and Decision Tree used for this proposed system and the output obtained clearly indicates that Linear Regression is the best algorithm to build the model. Even after building the model there comes a problem which is upon each execution of the algorithm it provides different results. So here comes another scenario to deal with the data.

I. K-Fold-Cross-Validation

In K-Fold-Cross-Validation the training and testing data is divided into k number of folds and each fold is a random selection of the other. This is a build-in python library which helps us to make the result optimized. Here in this proposed system, the dataset is divided into 5 folds.--

III. CONCLUSION

In this paper, we have analysed various concepts of data science which is applied on a machine learning model for predicting the price of houses. From data collection to model building there arises various data in-consistencies and data problems .Data science and data visualization helps us to overcome the various abnormalities efficiently. Data cleaning is the most crucial step we have encountered and if it is 100 percentage efficient then the model can bring us a score more than 90 percentage. Each and every step leads us to a problem and with the help of various python libraries and data science the problems are solved efficiently.

IV. FUTURE SCOPE

Data-science is a trending and upcoming technology and hence it has a broader scope than any other fields. Also the need and amount of data nowadays is also tremendous .Here in the proposed system we are dealing with structured data. This idea can be extended to unstructured data as well. Nowadays there comes wide variety of dataset. Applying the data-science concepts upon such dataset can bring tremendous model building or advantageous results. Also Hybrid dataset and hybrid algorithms can be applied for increasing the efficiency of the model. Scope of data and therefore data science is unpredictable because such a huge requirements are present for the both.

REFERENCES

[1] K. Biron, W. Mansoor, S. Miniaoui, S. Atalla, H. Mukhtar, and K. F. B. Hashim, "Data science tools for crime investigation, archival, and analysis," in 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

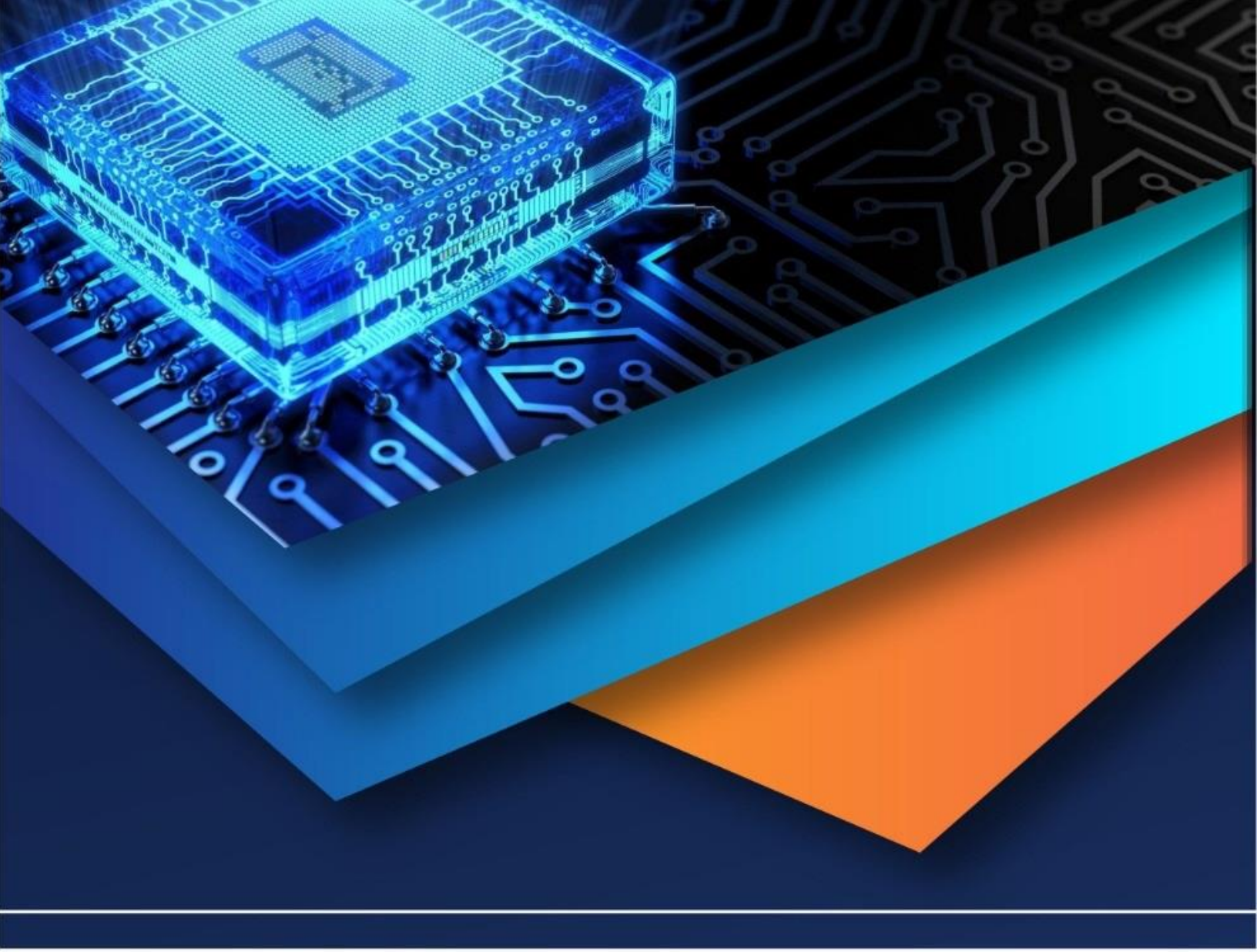
[2] R. Al-Zaidy, B. C. Fung, A. M. Youssef, and F. Fortin, "Mining criminal networks from unstructured text documents," Digital Investigation, vol. 8, no. 3-4, pp. 147– 160, 2012. [4]Prediction of Crop Yield using Machine Learning", *International Research Journal of Engineering & Technology*, vol. 5, no. 2, Feb 2018.

[3] M. R. Keyvanpour, M. Javideh, and M. R. Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework," Procedia Computer Science, vol. 3, pp. 872–880, 2011.

[4] F. Bex, S. Van den Braak, H. Van Oostendorp, H. Prakken, B. Verheij, and G. Vreeswijk, "Sense-making software for crime investigation: how to combine stories and arguments?" Law, Probability & Risk, vol. 6, no. 1-4, pp. 145–168, 2007.

[5] J. Xu and H. Chen, "Criminal network analysis and visualization," Communications of the ACM, vol. 48, no. 6, pp. 100–107, 2005.

[6] D. E. Brown, "The regional crime analysis program (recap): a framework for mining data to catch criminals," in SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218), vol. 3. IEEE, 1998, pp. 2848–2853.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY