

HOUSING PROJECT

Submitted by:

KALPANA KANWARIA

**ACKNOWLEDGMENT**

I would like to show my thankful gratitude to my flip robo mentor Ms. Khushboo garg and Mohammed Kashif for his assistance to solve the project and my data trained mentors. Along with that, I took reference from Kaggle, GitHub, stack overflow website and google.

**INTRODUCTION**

* Business Problem Framing

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy.It is a very large market and every company wants to increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases.

* Conceptual Background of the Domain Problem

**Describe the domain related concepts that you think will be useful for better understanding of the project.**

To understand the project and to work on that a Data Scientist should have a better knowledge of python programming for coding. Python libraires like( Pandas, Seaborn, NumPy and matplotlib, scikit learn etc), Machine Learning algorithms.

* Review of Literature

**This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.**

* Motivation for the Problem Undertaken

**Describe your objective behind to make this project, this domain and what is the motivation behind.**

Our main objective for making this project is to help companies to achieve their desired goal for house sales and purchase with the help of machine learning.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

**Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.**

Definitely, In making the project we have used mathematical calculations ,statistical calculations and analytical modelling. Some of the major mathematical concepts have been used in ML like linear algebra, calculus, probability etc. statistics is used here to analyse and visualize data to find unseen patterns.

* Data Sources and their formats

**What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.**

Data sources are the platforms from where we get the dataset to work on it. For ex- GitHub, Kaggle, skearn and many more. . The origin of obtaining data is unlimited we are generating data in every single nanosecond. A data can be in any format(csv, json, excel, tsv etc).Datasets contains the data points having information. They can be in any format (object, integer, float).It can a labelled dataset (supervised machine learning) or without label which comes under unsupervised machine learning

* Data Preprocessing Done

**What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?**

For cleaning of data we have used so many steps like to remove skewness we have used power transform function, to remove outliers we have use zscore and to remove the multicollinear features we used drop method.

* Data Inputs- Logic- Output Relationships

**Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.**

Data contains the input or independent variables and as well as output or target or label variables. Relationship between input and output variables can be positive or negative. Output variable depend on the data of the independent variables. linear relationship between input and output is a positive relation or vice versa.

* State the set of assumptions (if any) related to the problem under consideration

**Here, you can describe any presumptions taken by you**.

* Hardware and Software Requirements and Tools Used

**Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.**

Hardware- Mac OS high sierra laptop, Logitech Bluetooth keyboard, Logitech Bluetooth mouse.

Software-Anaconda Navigator (python 3.6)- used for programming

Libraries-pandas(to make data frames),seaborn(visualization of data),matplotlib(visualization of data),NumPy( to work with arrays),sklearn(to build ML models).

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

**Describe the approaches you followed, both statistical and analytical, for solving of this problem**.

We have used a supervised machine learning approach to solve this problem. In Supervised machine learning the model learns the relationship between the labelled input and output.

* Testing of Identified Approaches (Algorithms)

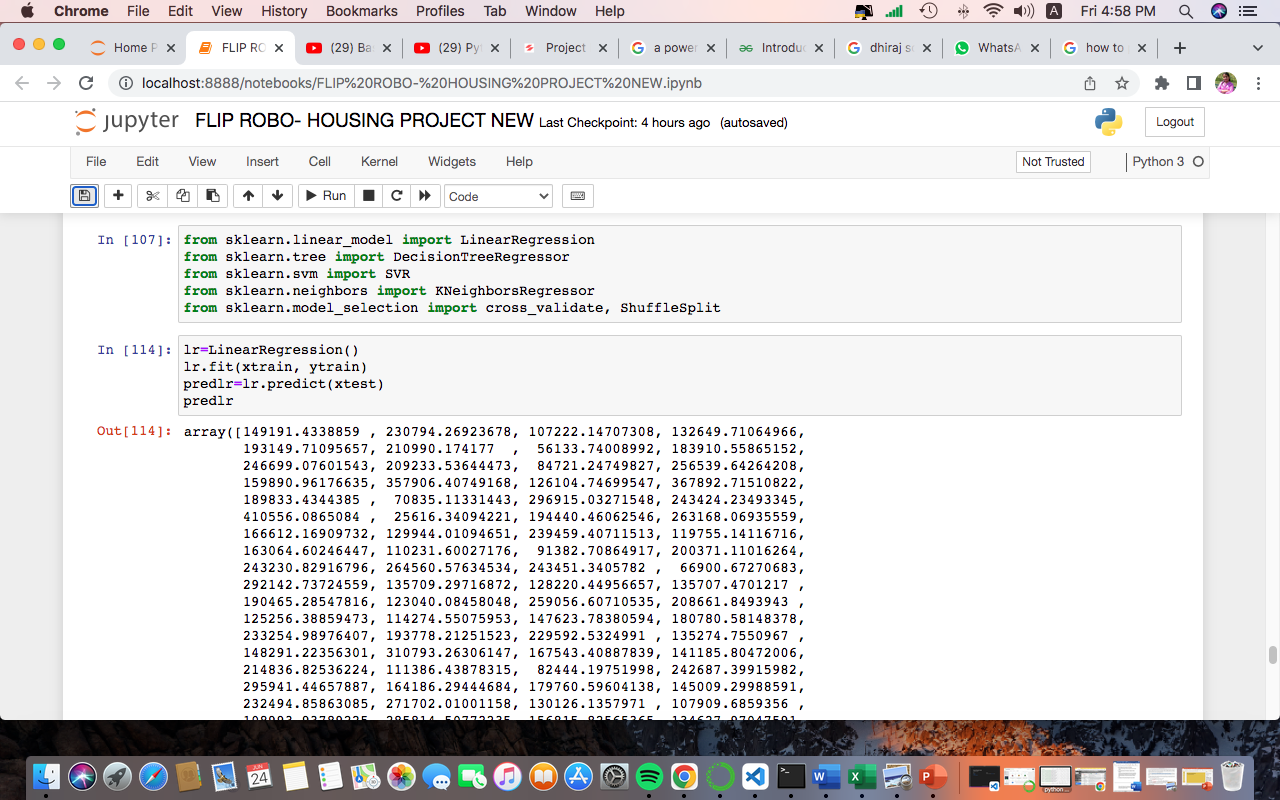
**Listing down all the algorithms used for the training and testing.**

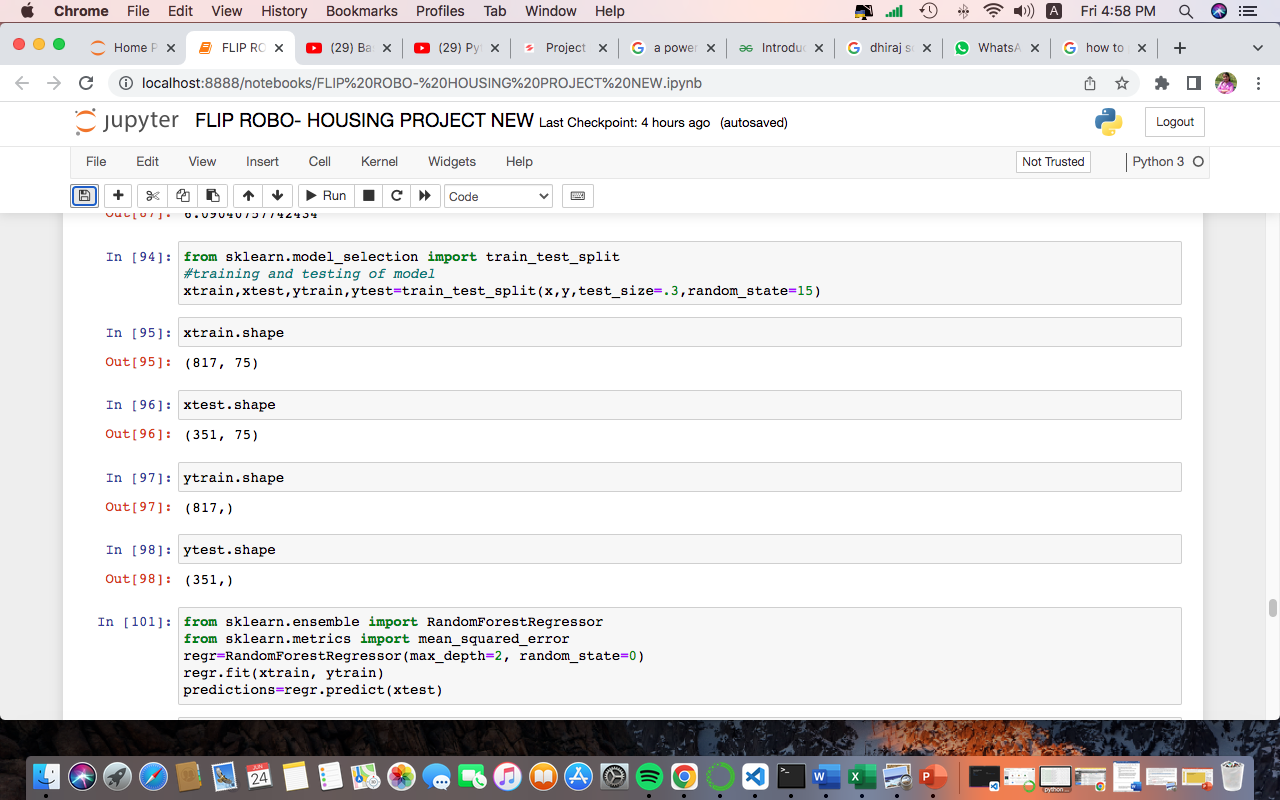
We have used so many algorithms in training and testing of data some of them are-Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support vector Machine etc.

* Run and Evaluate selected models

**Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.**

We have used machine learning algorithms such as Random Forest Regressor, Support Vector Machine, Linear Regression etc.





* Key Metrics for success in solving problem under consideration

**What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.**

* Visualizations

**Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.**

**If different platforms were used, mention that as well**.

We have used seaborn library to visualize the data and matplotlib to plot the different graphs and to analyse the data. We have use heatmap from seaborn library to detect the missing or null values. Corelation matrix to know the relation among the features. We have used matplotlib boxplot to see the outliers, plot to check the skewness etc.

* Interpretation of the Results

**Give a summary of what results were interpreted from the visualizations, pre-processing and modelling.**

**CONCLUSION**

* Key Findings and Conclusions of the Study

**Describe the key findings, inferences, observations from the whole problem.**

* Learning Outcomes of the Study in respect of Data Science

**List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.**

Data visualization as the name suggests is used to analyse the data in form of graphs, plots and maps. As we know with visualization technique things are easy to understand and are memorable for a long time which is not preferable in case of theory.

Data cleaning is a approach used to clean the raw data. By data cleaning we can get the cleaned and informative data by working on that we can extract the useful information from it.

There are many machine learning algorithms for model building and get predictions but the most used and easy to understand is the Random Forest Machine learning algorithm and K nearest Neighbour which is most widely used ML algorithm by data scientist in supervised machine learning.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.