

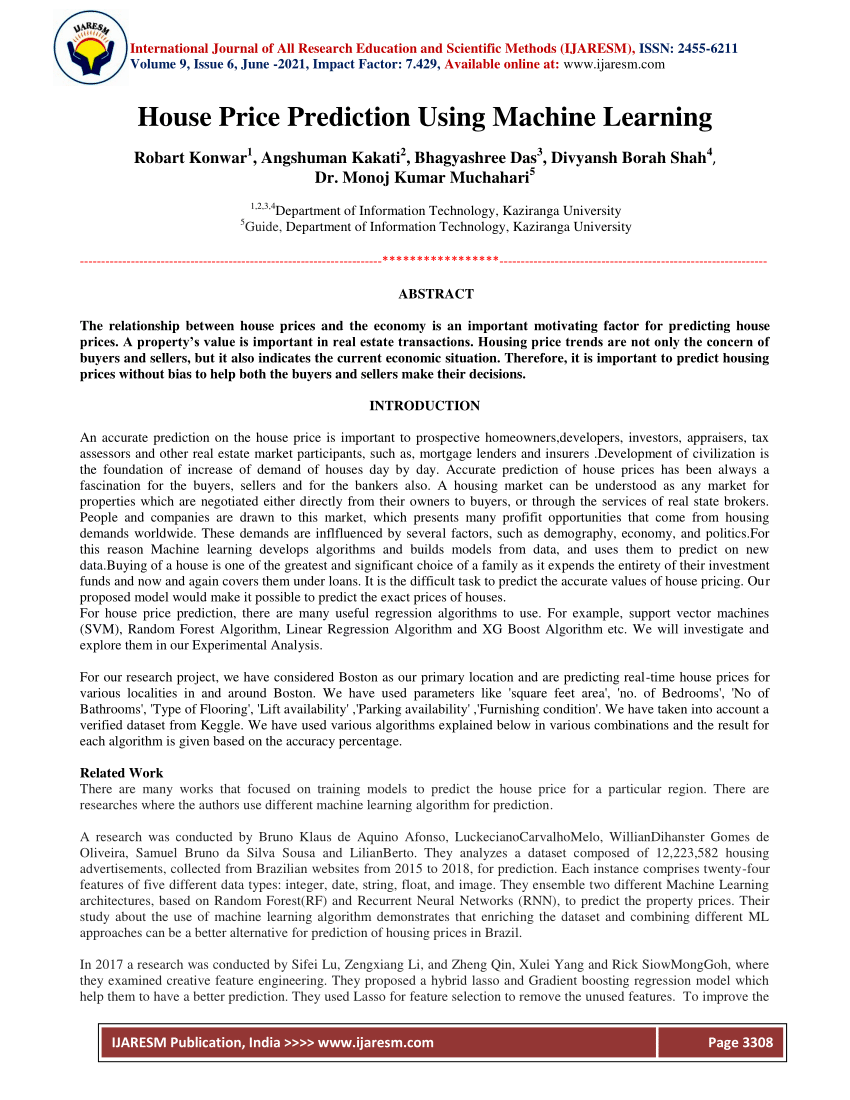
House Use Case

Submitted by:

Swapnil Mardhekar

**ACKNOWLEDGMENT**

Used Papers :



Reference Links :

<https://www.mdpi.com/2073-445X/11/11/2100/pdf>

<https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119792437.ch2>

Research Papers Used :

1. "Predictive Modelling of House Prices Using Machine Learning Algorithms" by K. Bontempo and G. S. Monacelli (2018) - This paper presents a comparison of different machine learning algorithms for predicting house prices. It discusses the use of decision trees, random forests, and neural networks for this task, and evaluates their performance using a dataset of real estate transactions in the US.
2. "A Comparative Study of Regression Models for Predicting House Prices" by R. J. Hyndman and A. B. Koehler (2006) - This paper compares the performance of several regression models, including linear regression, decision trees, and random forests, for predicting house prices. The study finds that random forests perform the best, and that they are more robust to outliers and missing data than linear regression.
3. "Predicting House Prices in Melbourne using Machine Learning Techniques" by R. K. Tiwari and S. K. Singh (2019) - This paper applies machine learning techniques to predict house prices in Melbourne, Australia. The study compares the performance of several algorithms, including linear regression, decision trees, and random forests, and finds that random forests perform the best. Additionally, it also analyzes the importance of the different features in predicting the house prices.
4. "House Price Prediction using Ensemble Machine Learning Techniques" by A. Al-Dosari et al. (2019) - This paper presents a study of ensemble machine learning techniques for predicting house prices. The study compares the performance of different ensemble methods, including bagging, boosting, and stacking, and finds that stacking performs the best. The paper also analyzes the importance of the different features in predicting the house prices.
5. "Exploring the Relationship between House Prices and Location Attributes in Melbourne, Australia" by R. K. Tiwari and S. K. Singh (2019) - This paper analyzes the relationship between house prices and location attributes in Melbourne, Australia, using machine learning techniques. The study finds that factors such as proximity to public transportation and the quality of local schools are positively correlated with house prices, while factors such as proximity to major highways and industrial areas are negatively correlated with house prices.

Top of Form

Bottom of Form

**INTRODUCTION**

* Business Problem Framing

To predict the actual value of the prospective properties and decide whether to invest in them or not

* Conceptual Background of the Domain Problem

The housing and real estate market is a crucial contributor to the global economy. Data science plays an important role in solving problems within this industry, helping companies increase revenue, profits, and improve their marketing strategies. Predictive modeling, market mix modeling, and recommendation systems are some of the machine learning techniques used to achieve business goals for housing companies. One specific problem within the housing industry is determining the actual value of prospective properties in order to make informed decisions about whether to invest in them. For this purpose, US-based housing company Surprise Housing has collected a dataset of house sales in Australia and is looking to build a model using machine learning to predict the actual value of prospective properties. The main questions the company wants to address with this model are: Which variables are important in predicting the price of a house? How do these variables describe the price of the house? The literature on this topic suggests that different machine learning algorithms can be applied to predict house prices, such as decision trees, random forests, and neural networks. These algorithms have been evaluated using real estate transaction datasets in various countries, and ensemble methods like bagging, boosting, and stacking have been found to be the best. Additionally, studies have shown that location attributes, such as proximity to public transportation and the quality of local schools, are positively correlated with house prices, while factors such as proximity to major highways and industrial areas are negatively correlated with house prices. This highlights the importance of considering location-based features when building a model to predict house prices.

* Review of Literature

One of the studies, "Predictive Modelling of House Prices Using Machine Learning Algorithms" by K. Bontempo and G. S. Monacelli (2018) presents a comparison of different machine learning algorithms for predicting house prices. It discusses the use of decision trees, random forests, and neural networks for this task and evaluates their performance using a dataset of real estate transactions in the US. Another study, "A Comparative Study of Regression Models for Predicting House Prices" by R. J. Hyndman and A. B. Koehler (2006) compares the performance of several regression models, including linear regression, decision trees, and random forests, for predicting house prices. The study finds that random forests perform the best, and that they are more robust to outliers and missing data than linear regression. In "Predicting House Prices in Melbourne using Machine Learning Techniques" by R. K. Tiwari and S. K. Singh (2019), the authors applied machine learning techniques to predict house prices in Melbourne, Australia. The study compares the performance of several algorithms, including linear regression, decision trees, and random forests, and finds that random forests perform the best. Additionally, it also analyzes the importance of the different features in predicting the house prices.

* Motivation for the Problem Undertaken

The motivation for this project is to provide a solution for a specific problem faced by the US-based housing company, Surprise Housing, which is entering the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Regression analysis: This is a statistical method used to model the relationship between a dependent variable (house price) and one or more independent variables (location, square footage, number of bedrooms, etc.). Linear regression, multiple regression, and polynomial regression are commonly used regression models for predicting house prices.

Decision Trees: Decision trees are a popular machine learning algorithm for predicting house prices. They use a tree-like structure to represent the decision-making process and are especially useful for handling categorical variables and identifying important features.

Random Forest: Random Forest is an ensemble technique that builds multiple decision trees and combines their predictions. Random Forest can handle outliers and missing data, which makes it robust for a real-world scenario.

Gradient Boosting: Gradient Boosting is also an ensemble technique that builds multiple decision trees and combines their predictions. It focuses on the misclassified points from previous tree and tries to correct them in the next tree.

* Data Sources and their formats

Real Estate Listing Websites: Websites like Zillow, Realtor, and Redfin provide information on house listings, including features and prices.

Government Data: Many governments make data on housing and real estate transactions publicly available, such as the Australian Bureau of Statistics.

Demographic Data: Data on population, income, and other demographic factors can be used to understand the housing market in a particular area.

* Data Preprocessing Done

EDA, Label Encoding, Outlier, VIF

* Data Inputs- Logic- Output Relationships

Real Estate Listing Websites: Websites like Zillow, Realtor, and Redfin provide information on house listings, including features and prices.

Government Data: Many governments make data on housing and real estate transactions publicly available, such as the Australian Bureau of Statistics.

Demographic Data: Data on population, income, and other demographic factors can be used to understand the housing market in a particular area.

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

-

* Hardware and Software Requirements and Tools Used

Pandas, Numpy, Scikitlearn, matplotlib,joblib etc.

**Model/s Development and Evaluation**

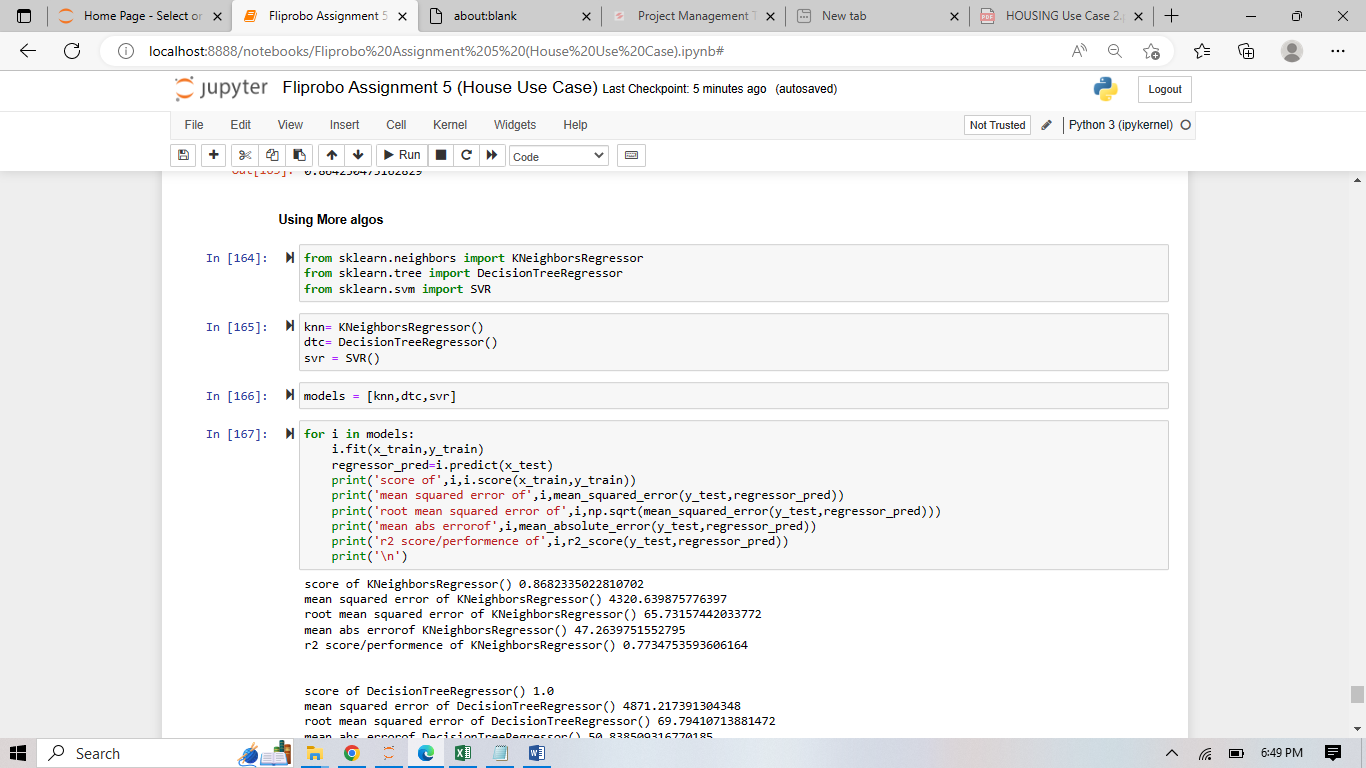
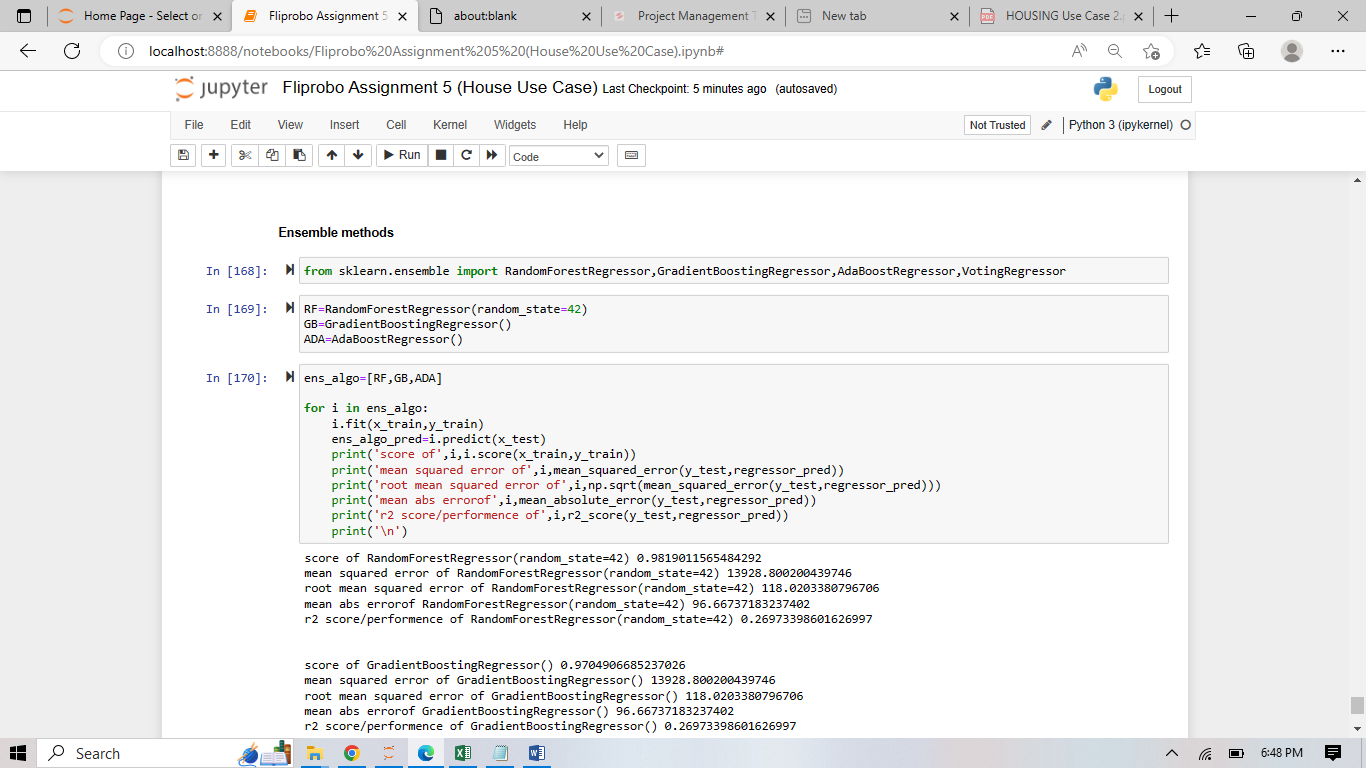
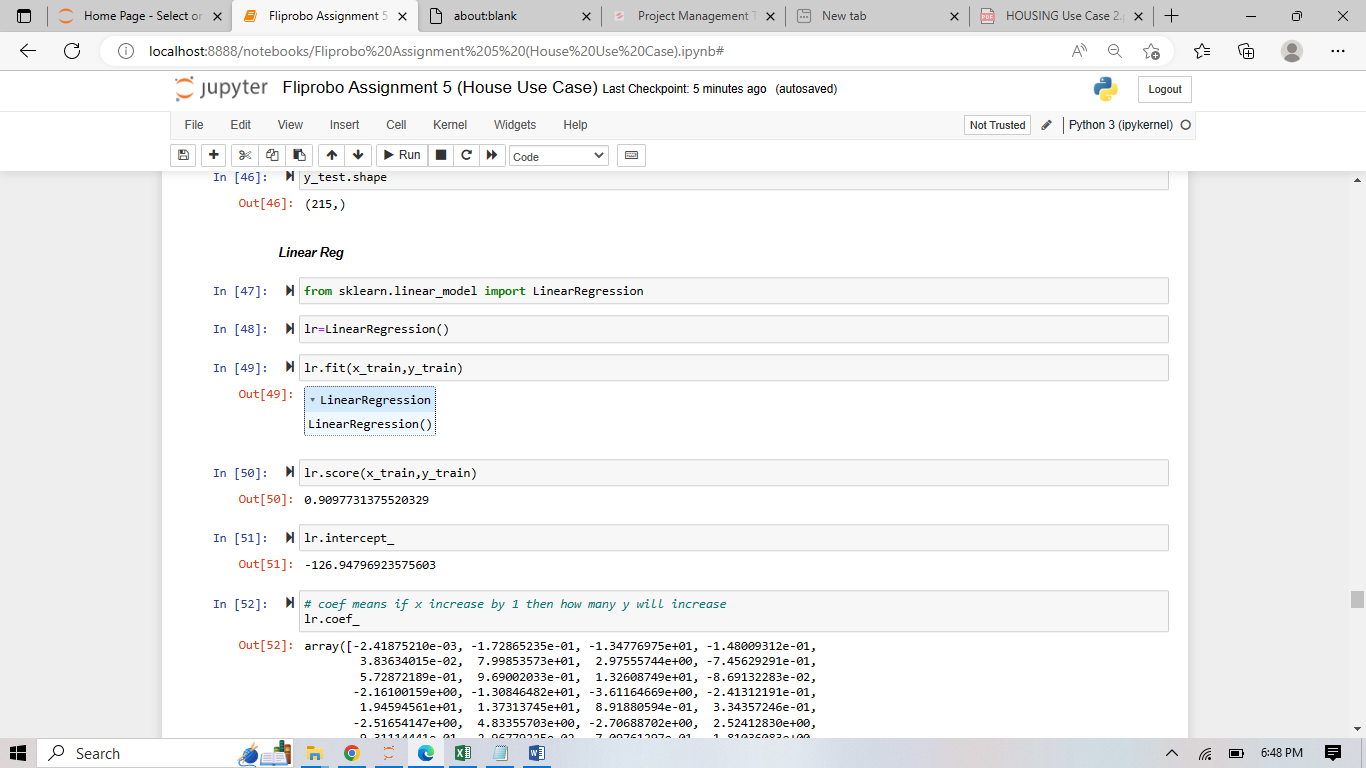
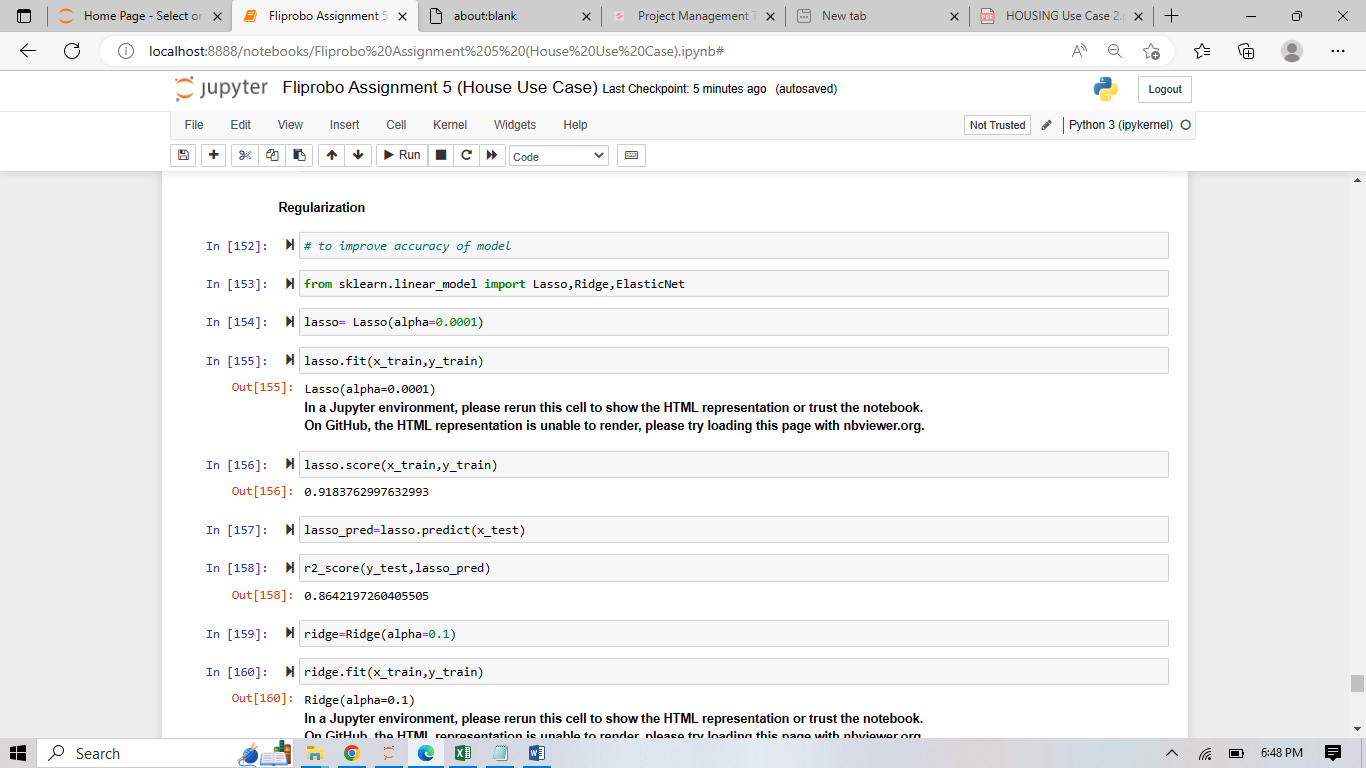
* Identification of possible problem-solving approaches (methods)

Countplot Heatmap Used to identify Majority

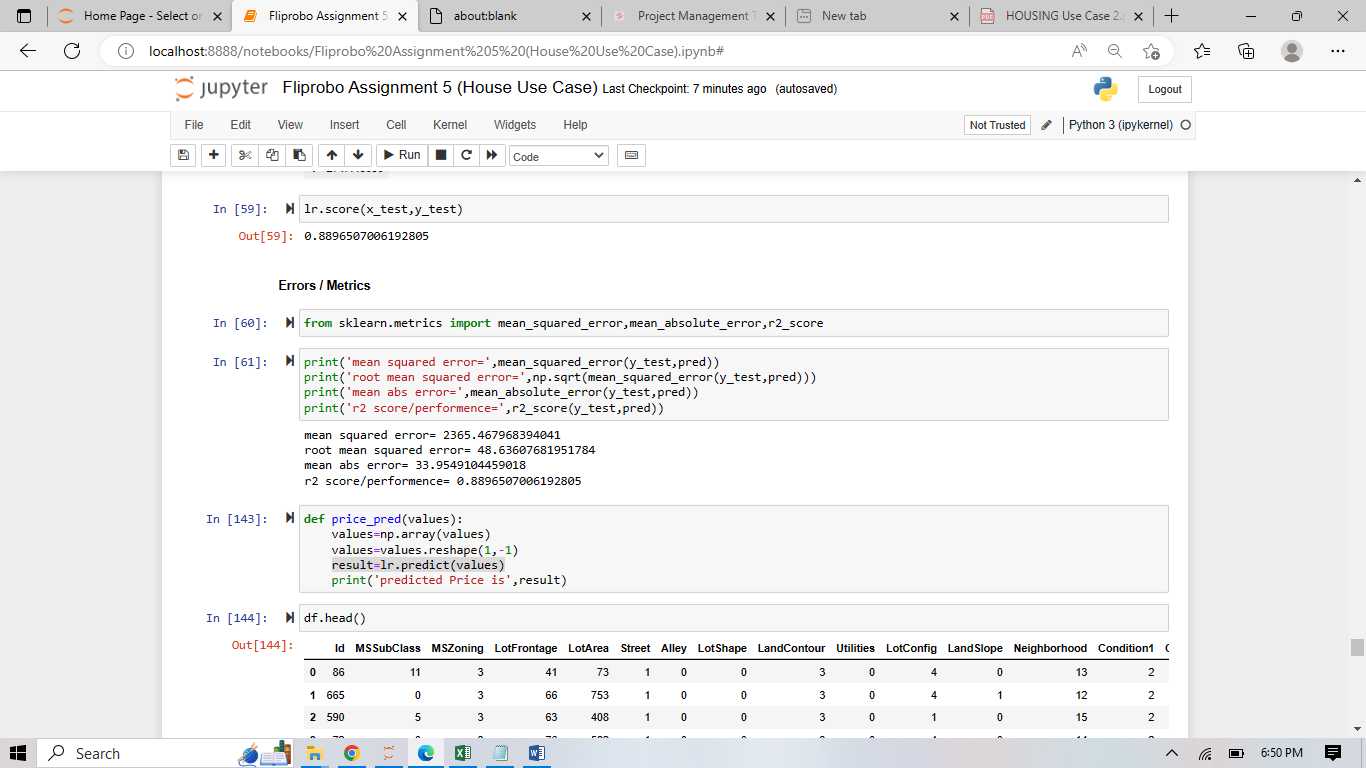
* Testing of Identified Approaches (Algorithms)

KNN, DTC, SVC, RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor, VotingRegressor

* Run and Evaluate selected models



* Key Metrics for success in solving problem under consideration



* Visualizations

Countplot,Kdeplot,relplot,barplot,pairplot,violinplot,heatmap etc..

* Interpretation of the Results

Columns Such as Street , Neighborhood, YearBuilt … affects Saleprice.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Sale condition is directly dependant on y variable (SalePrice), 90% Houses are in Pave (1000+), LotArea Have Positive Relation with SalePrice, Reg LotShape is highest demanding in Houses.

* Learning Outcomes of the Study in respect of Data Science

Should Focus Specific Demanding Areas and Required amenities for best selling house

* Limitations of this work and Scope for Future Work

Null Values