

NAME OF THE PROJECT

HOUSING : PRICE PREDICTION

Submitted by:

V TARAK RAM SAI

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

This is regarding a client which used data analytics to purchase houses at a price below their actual values and flip them at a higher price. The focus of this project is to build a machine learning model that can accurately predict the price of a house.

* Conceptual Background of the Domain Problem

You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

* Review of Literature

In this report, we investigate the application of supervised machine learning techniques to predict the price of houses. The predictions are based on historical data that has collected by the company. In this paper, the price evaluation model based on big data analysis is proposed, which takes advantage of widely circulated house data and a large number of housing features data to analyse the price data by using various Machine Learning models.

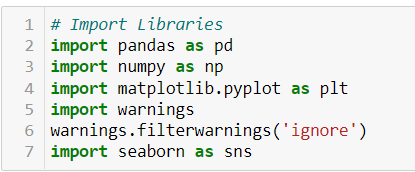
* Motivation for the Problem Undertaken

The purpose of this project is to determine the features of houses that decide the cost of the house. It aims to establish a house price evaluation model to predict the price that best matches the house by observing the data collected by the client.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Any machine learning model should follow the below steps while dealing a business problem. They are:
* **i.)** **Business Understanding:** The first step is to comprehend the research’s background, the problem description, and how the proposed project will achieve the goals.
* **ii.**) **Data Understanding:** The second stage requires collection of data listed in the project resources. This involves in determining the data requirements and exploring key data attributes.
* **iii.**) **Data Preparation:** The third stage involves data cleaning and should the handle the missing values in the data.
* **iv.**) **Modelling:** This involves determining the modelling technique and testing the design.
* **v.**) **Evaluation:** Here, we should evaluate the achieved results and should determine the performance of the model with best accuracy.
* **vi.**) **Deployment:** The last stage is implementation of the model.
* Data Preprocessing

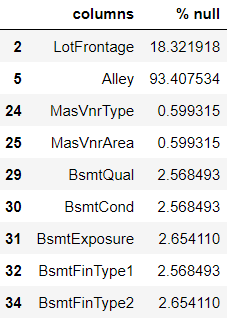
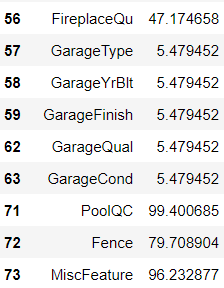
It entails converting raw data into comprehensible format that a machine learning model can understand. The data pre-processing involves data cleaning which involves handling missing values, transformation of data i.e. normalizing the data and data reduction which involves only required features.





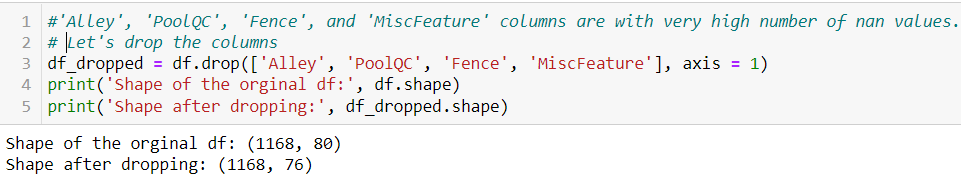
**Data Cleaning**

The first step of pre-processing is data cleaning by checking and eliminating any missing values because they affect the accuracy of the model. This is achieved by either filling the missing values with a mean or mode function or by dropping all the missing values. In this case, there are plenty of missing values in Alley, FireplaceQu, PoolQC, Fence and MiscFeature columns.

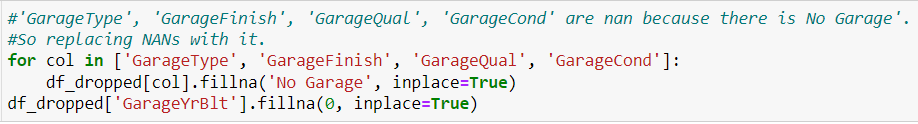
 

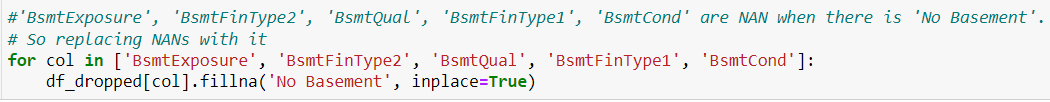
**Data Reduction**

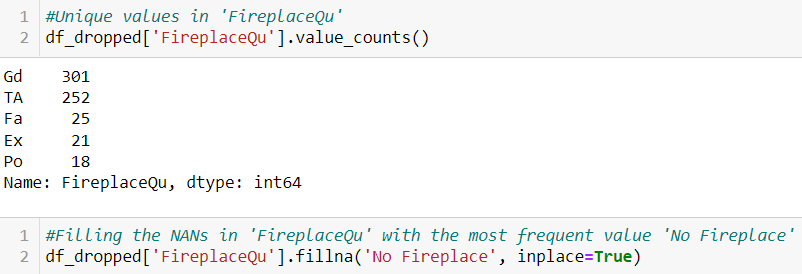
The next step of data processing is data reduction. This is used to remove duplicate features present in the data i.e. unwanted features for prediction.



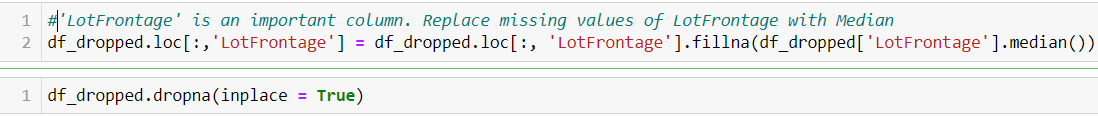
**Cleaning of Data**



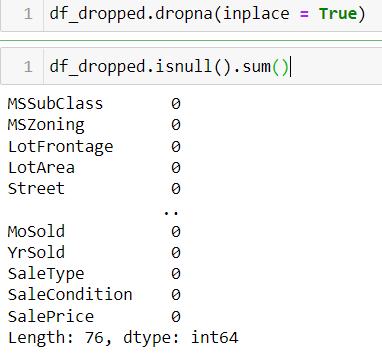




### Handling Missing Values

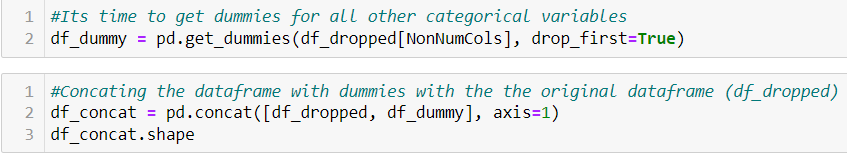


We have to drop the rows with missing values.



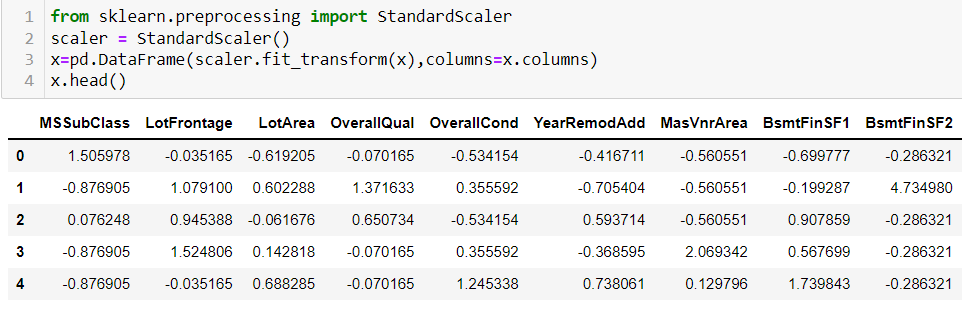
**Encoding**

Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models.



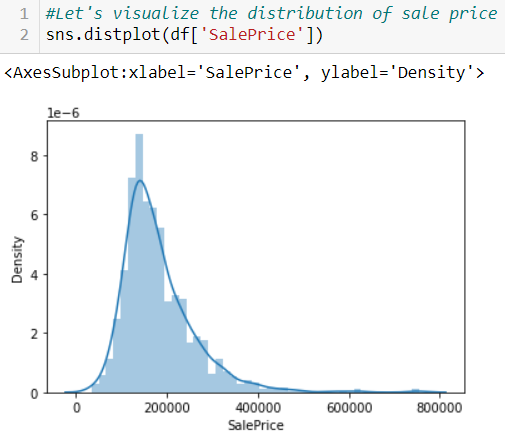
**Scaling**

Feature scaling is **a method used to normalize the range of independent variables or features of data**. To convert data into a distribution with a mean of 0 and standard deviation of 1, we will use a standard scalar.

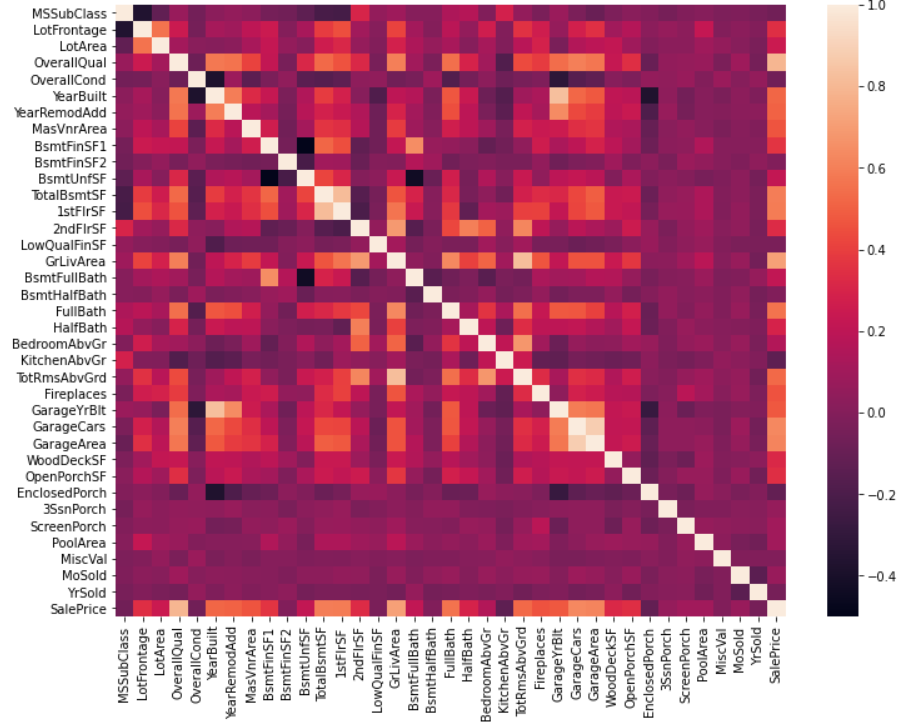


* Data Inputs- Logic- Output Relationships

The Relation between all the features in the dataset are determined in the below graphs as follows



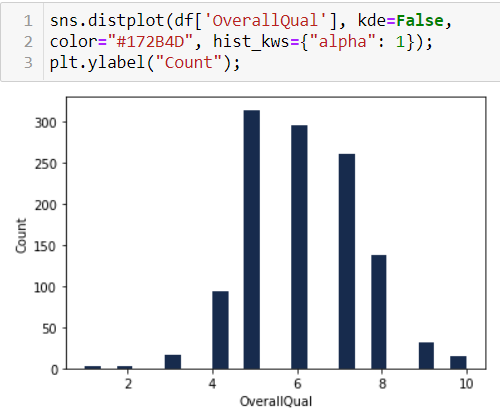
Our dataset contains a lot of variables, but the most important one for us to explore is the target variable. We need to understand its distribution.



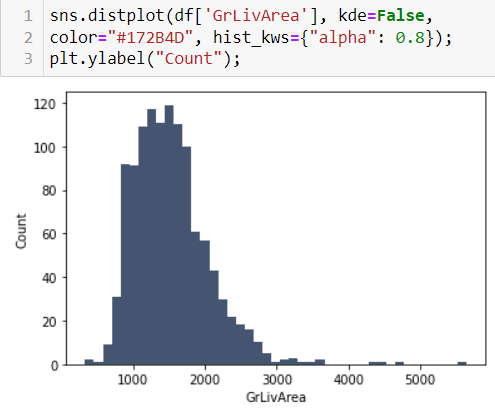
We can see that there are many correlated variables in our dataset. Wwe notice that Garage Cars and Garage Area have high positive correlation which is reasonable because when the garage area increases, its car capacity increases too. We see also that Gr Liv Area and TotRms AbvGrd are highly positively correlated which also makes sense because when living area above ground increases, it is expected for the rooms above ground to increase too.

Regarding negative correlation, we can see that Bsmt Unf SF is negatively correlated with BsmtFin SF 1, and that makes sense because when we have more unfinished area, this means that we have less finished area. We note also that Bsmt Unf SF is negatively correlated with Bsmt Full Bath which is reasonable too.

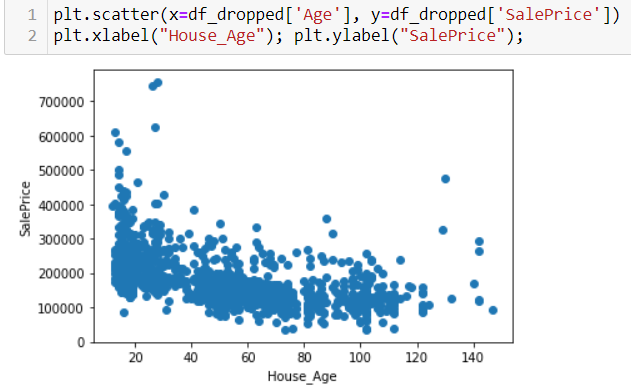
Most importantly, we want to look at the predictor variables that are correlated with the target variable (SalePrice). By looking at the last row of the heatmap, we see that the target variable is highly positively correlated with Overall Qual and Gr Liv Area. We see also that the target variable is positively correlated with Year Built, Year Remod/Add, Mas Vnr Area, Total Bsmt SF, 1st Flr SF, Full Bath, Garage Cars, and Garage Area.



We see that Overall Qual takes an integer value between 1 and 10, and that most houses have an overall quality between 5 and 7.



We can see that the above-ground living area falls approximately between 800 and 1800 ft2.



If the House is newly built, then the cost will be more.

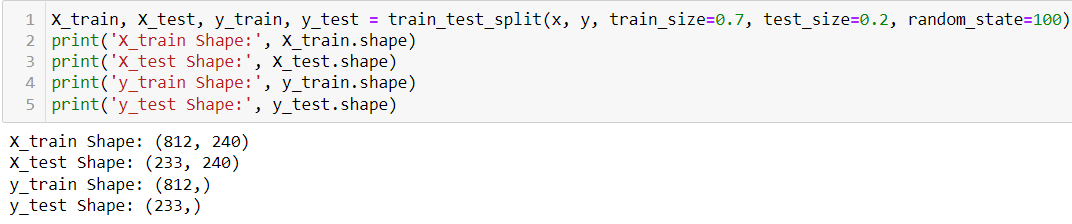
* Hardware and Software Requirements and Tools Used

The hardware requirements for the project includes a laptop with at least 4GB RAM. This project uses a Jupyter Notebook as a code editor. The Machine Learning models are implemented using python version 3.7 with libraries like numpy, pandas, matplotlib, seaborn and sklearn.

**Model/s Development and Evaluation**

The independent variables are declared in x and the dependent variable i.e. ‘SalePrice’ is declared in y as follows-



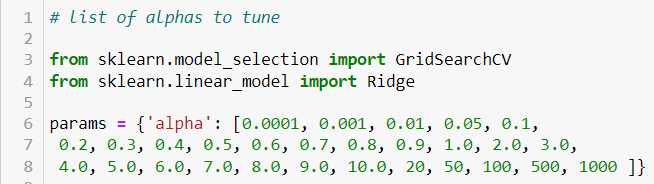


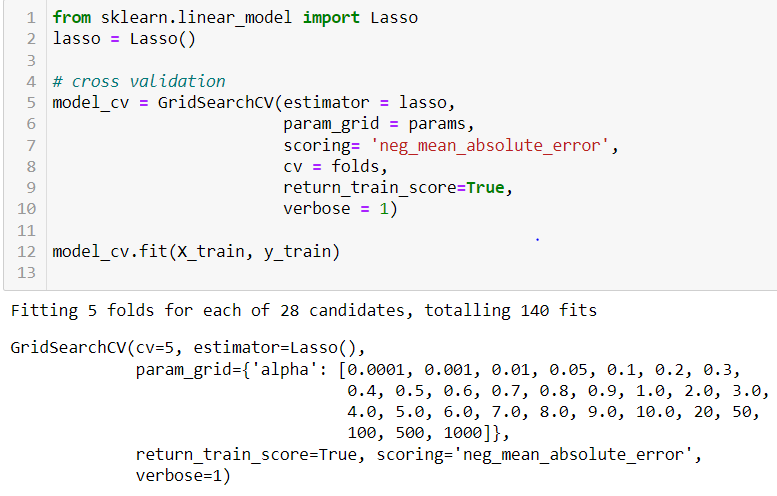
**Hyper Parameter Tuning**

In machine learning, hyper parameter optimization or tuning is **the problem of choosing a set of optimal hyper parameters for a learning algorithm**. A hyper parameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

**Lasso Regression**

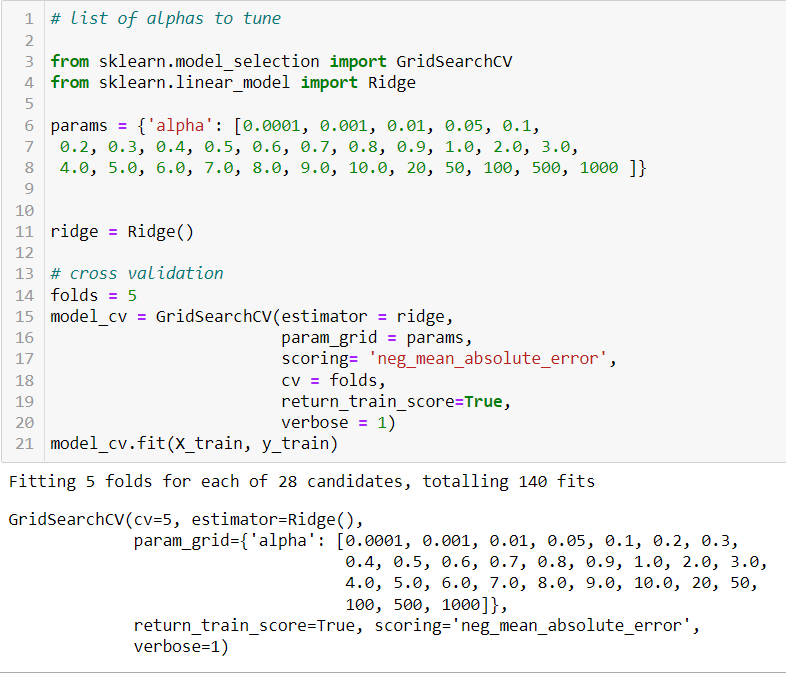
The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi-collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.





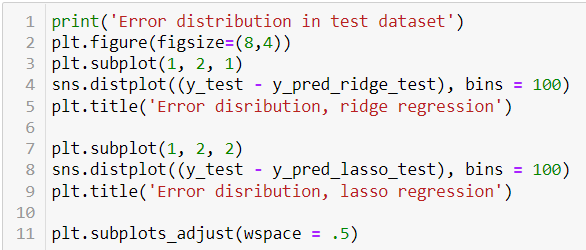
**Ridge Regression**

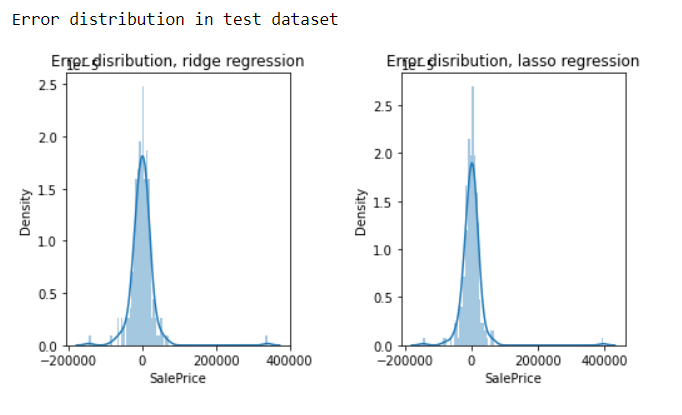
A Ridge regressor is basically a regularized version of Linear Regressor. The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e. helps in reducing the variance of the estimates.



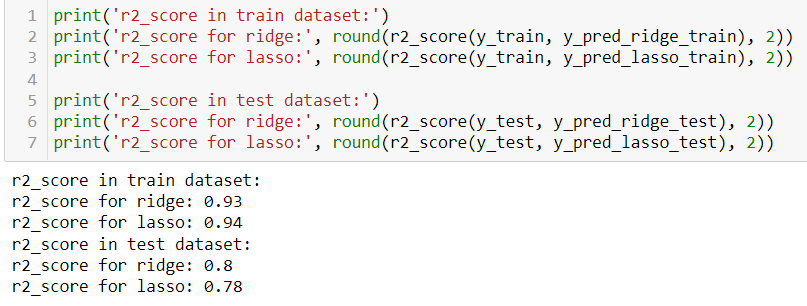
**Predicting Test Data**







**Evaluation Metrics**



The columns with positive correlation with the house price are as follows - ['MSZoning', 'LotFrontage', 'LotArea', 'Street', 'LotShape', 'LandContour', 'LotConfig', 'Neighborhood', 'Condition1', 'Condition2', 'HouseStyle', 'OverallQual', 'OverallCond', 'YearRemodAdd', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'MasVnrArea', 'ExterCond', 'Foundation', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinSF1', 'BsmtFinSF2', 'TotalBsmtSF', 'Heating', 'HeatingQC', '2ndFlrSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath', 'Functional', 'Fireplaces', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageCars', 'PavedDrive', 'WoodDeckSF', 'OpenPorchSF', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'SaleType', 'SaleCondition']

The columns with negative correlation with the house price are as follows - 'MSSubClass', 'LotShape', 'LandContour', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'Heating', 'HeatingQC', 'Electrical', 'LowQualFinSF', 'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual', 'Functional', 'FireplaceQu', 'GarageType', 'GarageQual', 'GarageCond', 'MiscVal', 'MoSold', 'YrSold', 'SaleType', 'SaleCondition'

**CONCLUSION**

* Key Findings and Conclusions of the Study

The company can purchase houses below the market value by highlighting the negative parameters. Afterwards, the company can work little bit on the negative parameters to decrease the magnitude of negative weight and sell the houses by mainly highlighting the positive features.

In this paper, we built several regression models to predict the price of some house given some of the house features. We evaluated and compared each model to determine the one with highest performance.

* Learning Outcomes of the Study in respect of Data Science

We also looked at how some models rank the features according to their importance. In this paper, we followed the data science process starting with getting the data, then cleaning and pre-processing the data, followed by exploring the data and building models, then evaluating the results and communicating them with visualizations.

As a recommendation, we advise to use this model (or a version of it trained with more recent data) by people who want to buy a house in the area covered by the dataset to have an idea about the actual price.

The model can be used also with datasets that cover different cities and areas provided that they contain the same features. We also suggest that people take into consideration the features that were deemed as most important as seen in the previous section; this might help them estimate the house price better.