

**HOUSING PROJECT**

Submitted by:

SANAA KHAN

**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**

**Problem Statement:**

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 there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter 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. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

**Business Goal:**

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:**

For **each project, a** literature review **provides a** clear **idea, which serves** as **a basis. Most** of the authors **here conclude** that artificial neural networks have **a big impact on prediction,** but in the real **world, so should** other **algorithms. Considered**  
**under** consideration. **Conducting** this **research helped me** know both the **benefits** and **his**  
**corn, which** helped me successfully implement the project. Motivation for the Problem Undertaken

Accurately estimating the value of **a home is of great interest to many stakeholders, including homeowners, homebuyers,** real estate agents, creditors, and investors. It is also **difficult. It's well known** that factors such as size, number of **rooms,** and location affect price, **but** many **others do too. In addition, Price**  
**is** sensitive to changes in market demand and the **details** of each **situation. B. When** a property needs to be **sold** urgently.

The sales price of a property can be predicted in **a number of** ways, often based on regression techniques. **Essentially, all** regression **procedures have** one or more predictor variables as input and a single target variable as output.  
**This post compares the performance of various**machine learning **techniques** in predicting the selling price of **a home** based on **many characteristics** such as  
**square feet,** number of **bedrooms** and **bathrooms,** and **geographic location.**

**Analytical Problem Framing:**

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?   
  
**Data Sources and their formats:**

Data contains 1460 entries each having 81 variables

Data contains null values. You need to treat them using the domain knowledge and your own understanding.

Extensive EDA has to be performed to gain relationships of important variable and price.

Data contains numerical as well as categorical variable. You need to handle them accordingly.

You have to build machine learning models, apply regularisation and determine the optimal values of hyper parameters.

You need to find important features which affect the price positively or negatively.

Two datasets are being provided to you (test.csv, train.csv). You will train on train.csv dataset and predict on test.csv file

**Data Pre-processing Done:**

Basic Data cleanup was performed by doing missing values check and imputation using business logic.

Above 19 columns have missing values.

It can be seen that-

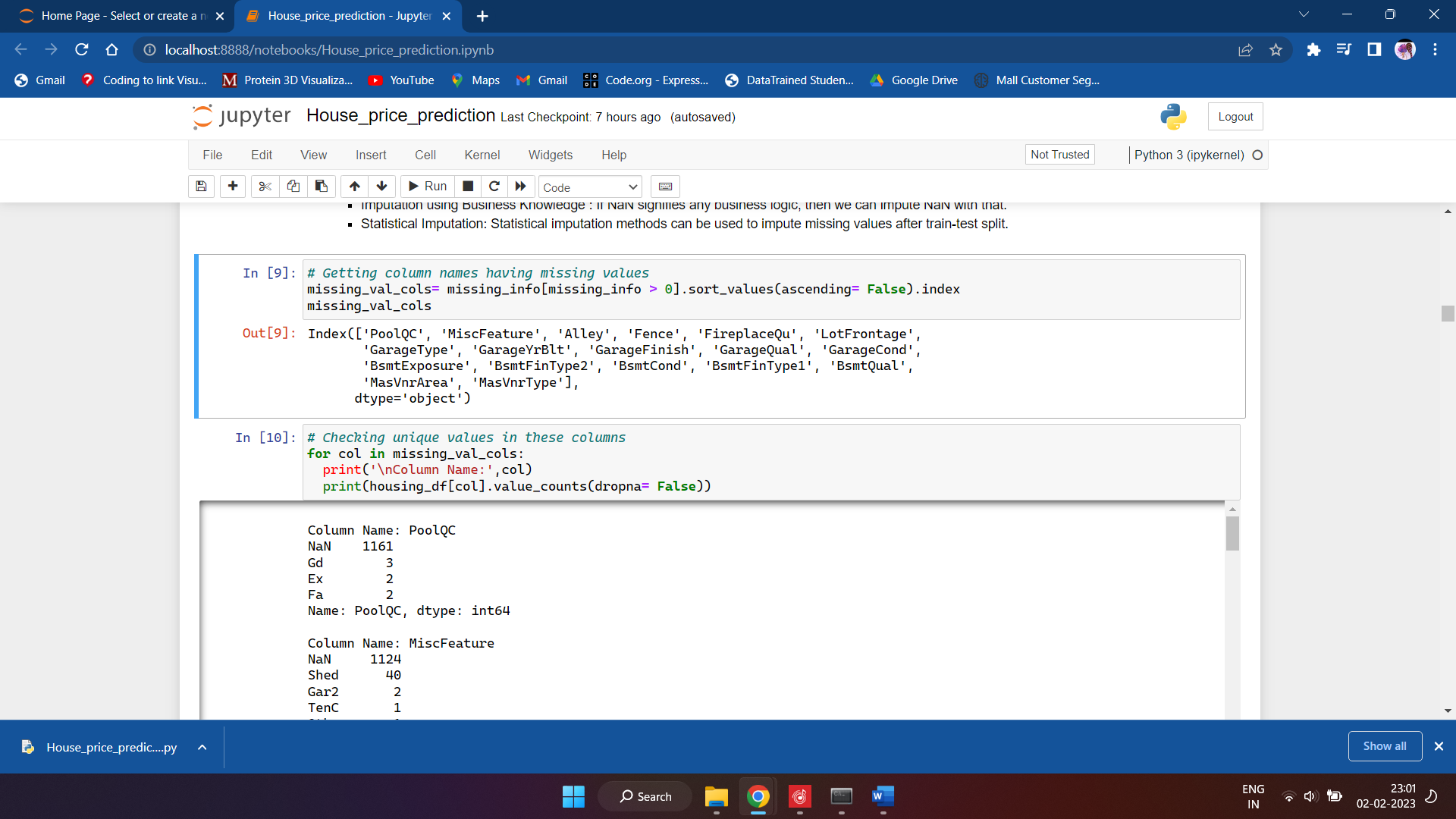
PoolQC, MiscFeature, Alley, Fence and FireplaceQu have very high percentage of missing value.

I will check all the columns to understand if these are actually missing or have some meaning. Once it is identified, imputation can be performed:

Imputation using Business Knowledge: If NaN signifies any business logic, then we can impute NaN with that.

Statistical Imputation: Statistical imputation methods can be used to impute missing values after train-test split.

**The following data had missing values:**



**Data Inputs- Logic- Output Relationships:**

we will replace NaN values for above attributes withh 'Not Present'.

For rest of the columns, we'll check if they have any relation with other columns and if we can use that relation in observed daat to impute these columns:

**LotFrontage, GarageYrBlt, MasVnrArea, MasVnrType, Electrical**

Initially GarageYrBlt and GarageType both had 5.55% missing value. After imputing NaN values of GarageType with 'Not Available', we can see that GarageYrBlt value is NaN for only those observations where GarageType is 'Not Available'. We can conclude that if garage is not available then there will be no 'GarageYrBlt' value for that. So we can safely **impute GarageYrBlt NaN values with 0.**

Performed statistical imputation for rest of the columns after train-test split: **LotFrontage, MasVnrArea, MasVnrType, Electrical**

**Hardware and Software Requirements and Tools Used:**

*The following libraries were used to perform analysis:*

* Pandas
* Numpy
* Matplotlib.pyplot
* Seaborn for data visualisation
* Sklearn for statistical analysis
* Sweetviz
* Sklearn.model selection
* I performed autoEDA using SweetViz

**Model/s Development and Evaluation:**

**Observations from AutoEDA**

**Numerical Associations with SalePrice:**

* GrLivArea: 0.71
* GarageArea: 0.62
* TotalBsmtSF: 0.61
* 1stFlrSF: 0.61
* TotRmsAbvGrd: 0.53
* YearBuilt: 0.52
* YearRemodAdd: 0.51
* MasVnrArea: 0.48
* BsmtFinSF1: 0.39
* LotFrontage: 0.35
* WoodDeckSF: 0.32
* 2ndFlrSF: 0.32
* OpenPorchSF: 0.32
* LotArea: 0.26

**Categorical Associations with SalePrice:**

* OverallQual: 0.83
* Neighborhood: 0.74
* GarageCars: 0.70
* ExterQual: 0.69
* BsmtQual: 0.68
* KitchenQual: 0.68
* FullBath: 0.58
* GarageFinish: 0.55
* FireplaceQu: 0.54
* Foundation: 0.51
* GarageType: 0.50
* Fireplaces: 0.48
* BsmtFinType1: 0.46
* HeatingQC: 0.44

**Testing of Identified Approaches (Algorithms):**

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

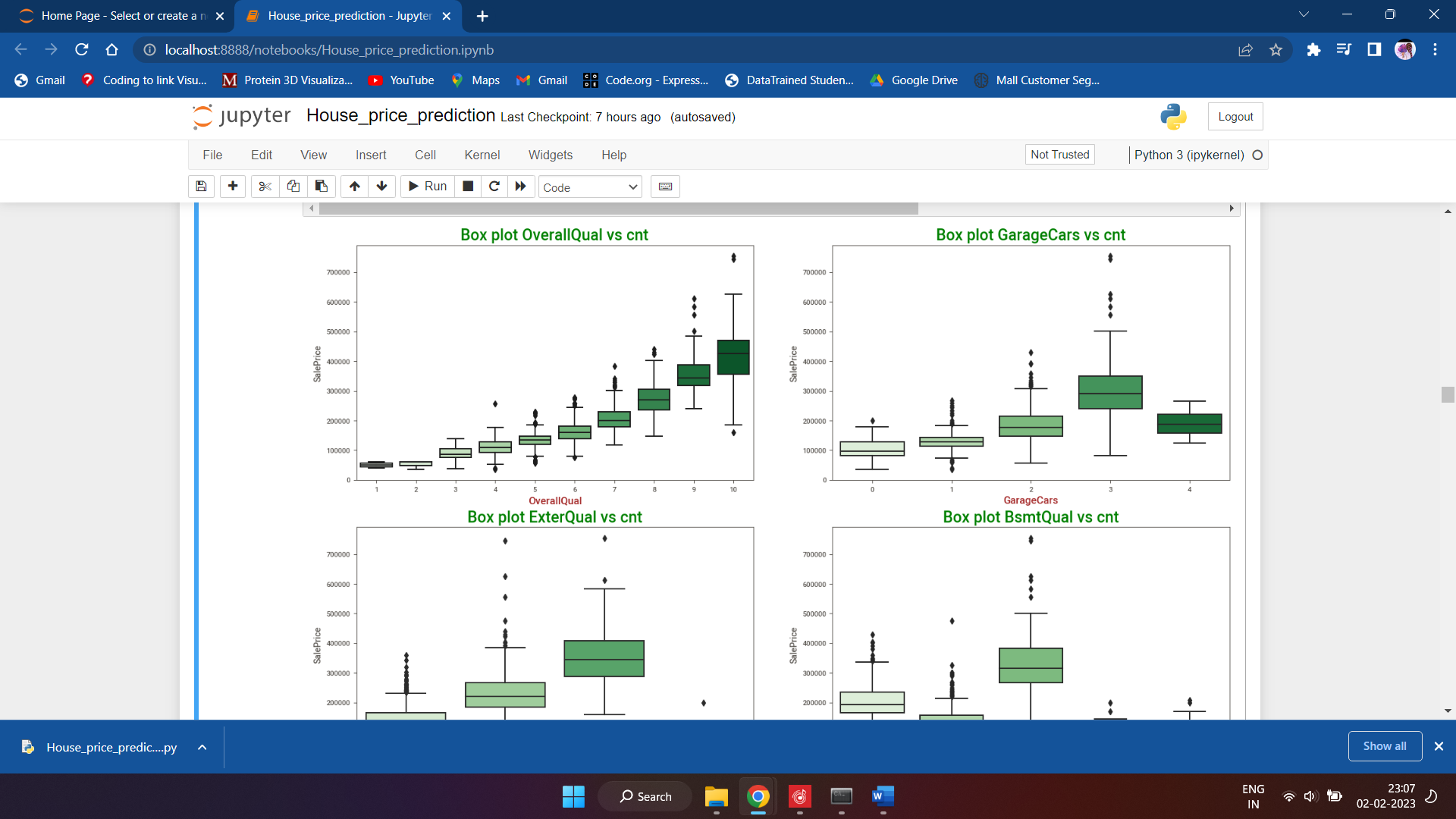
Run and evaluate selected models

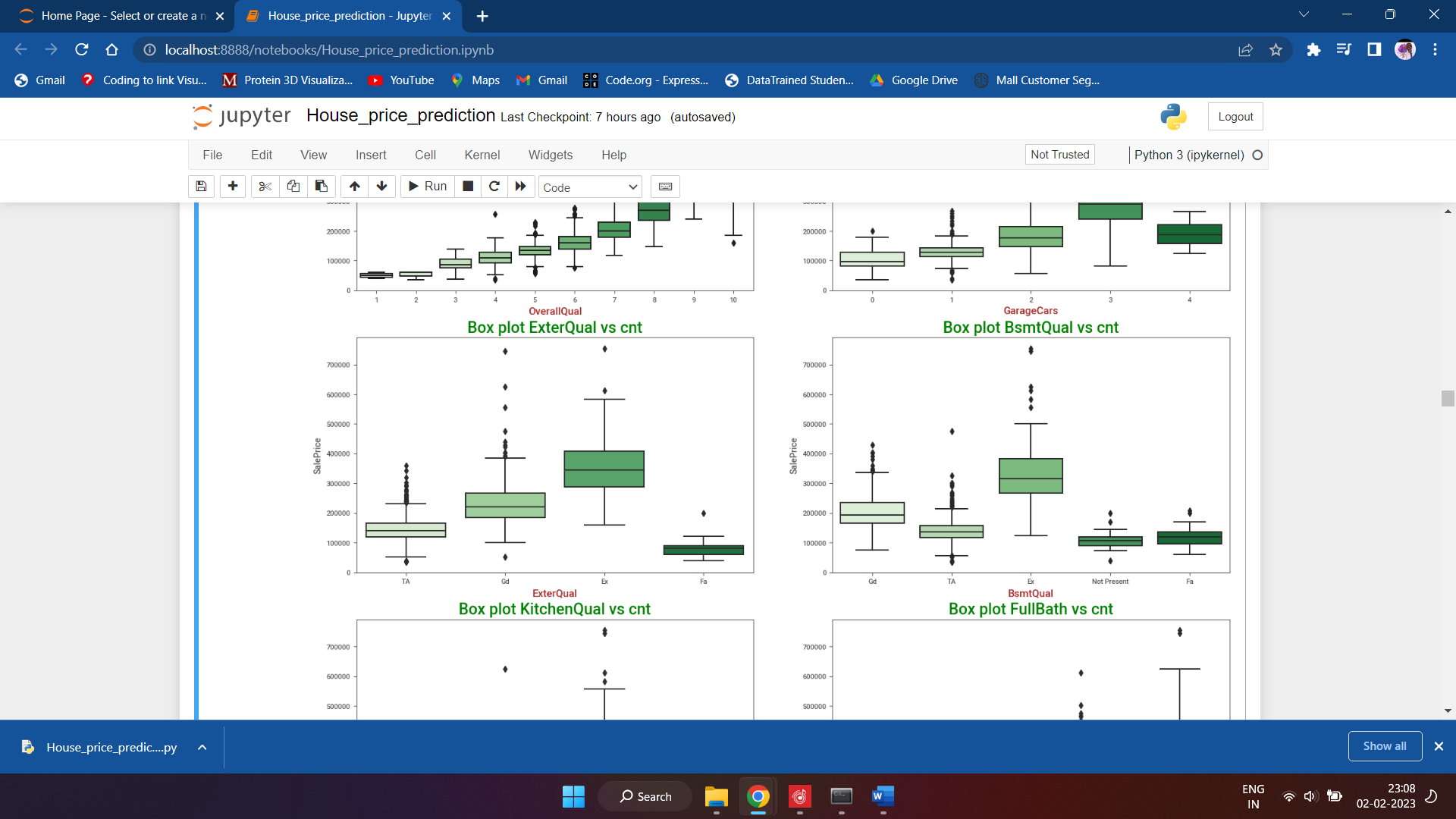
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

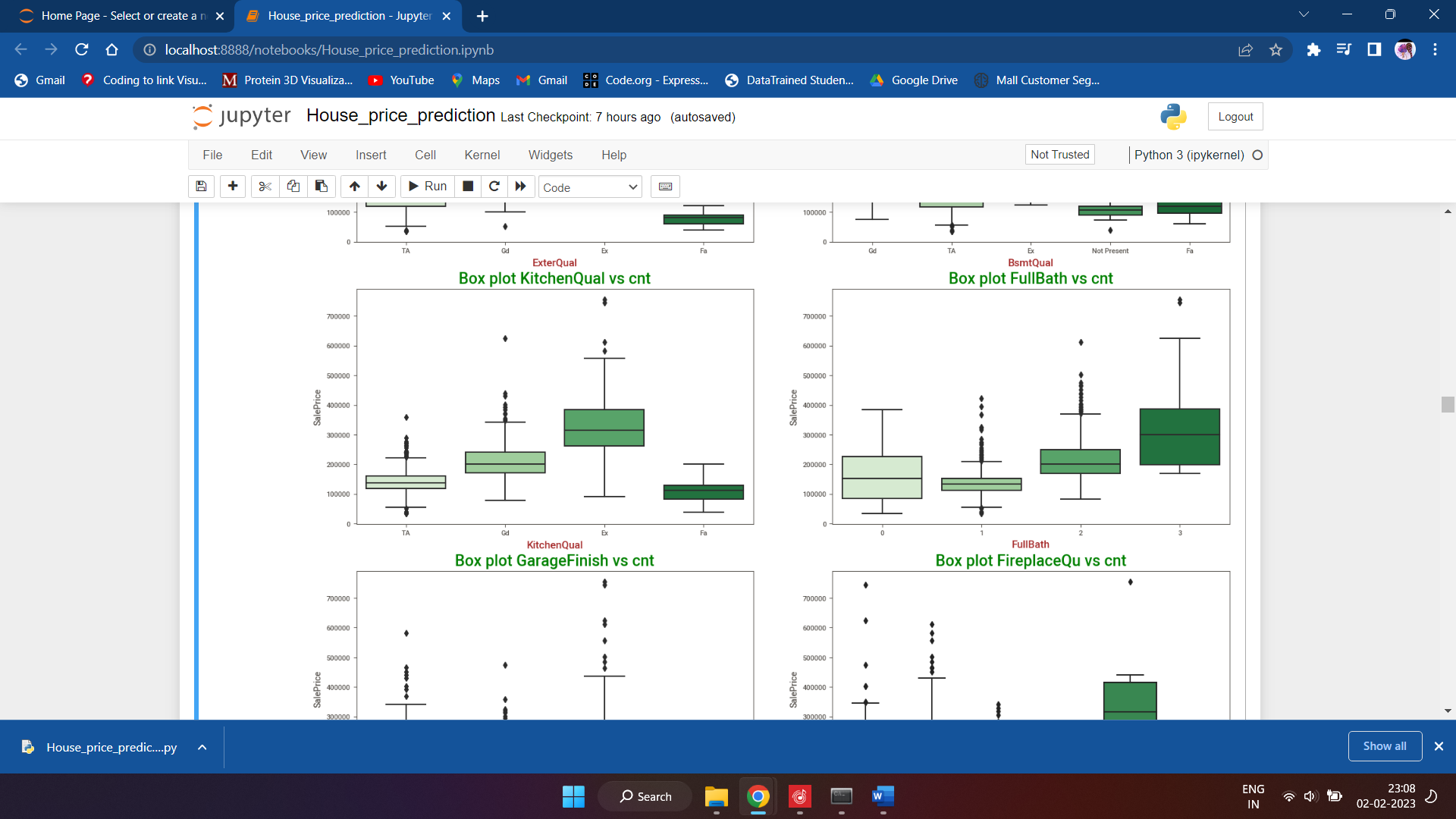
Key Metrics for success in solving problem under consideration

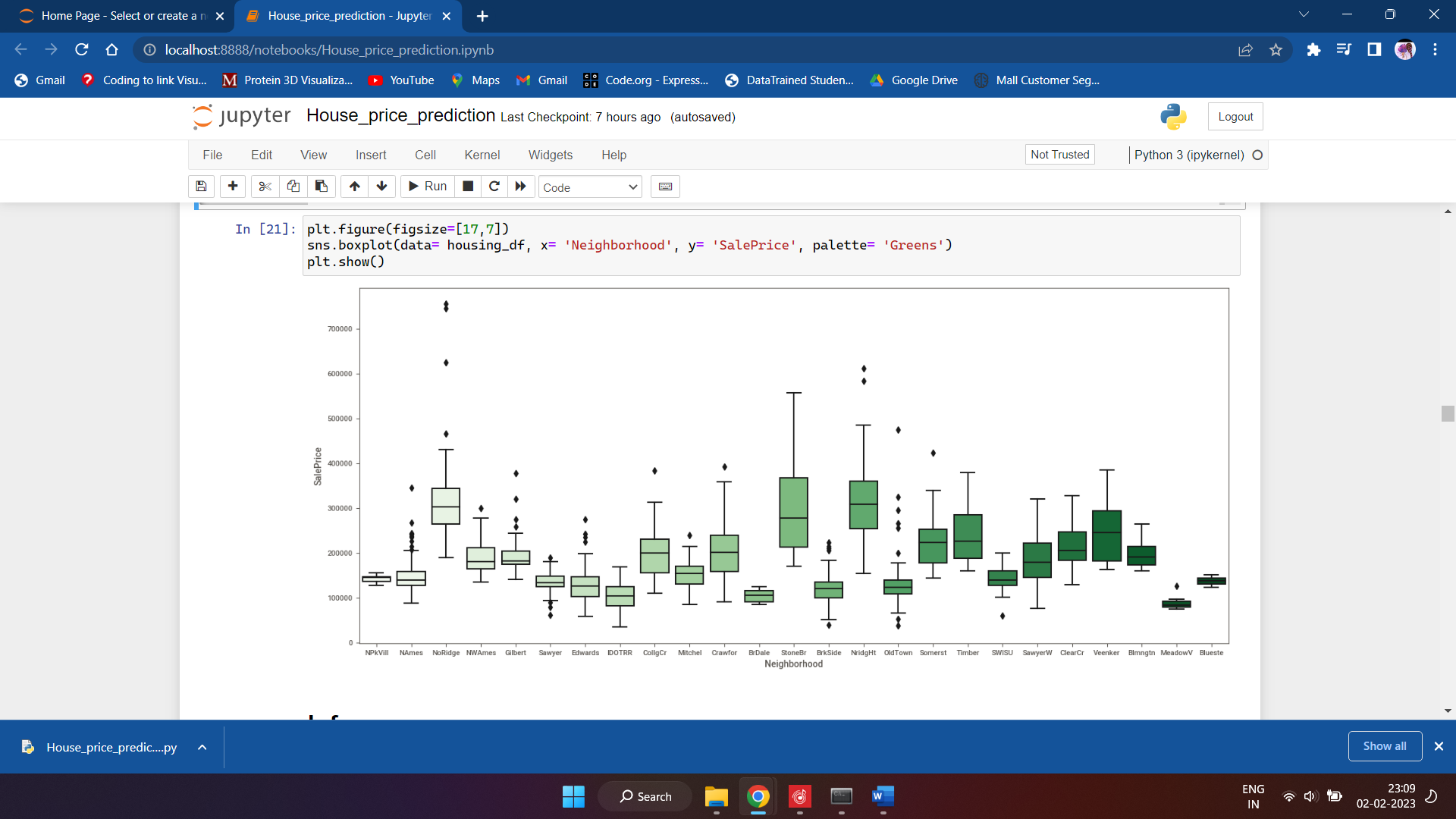
What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

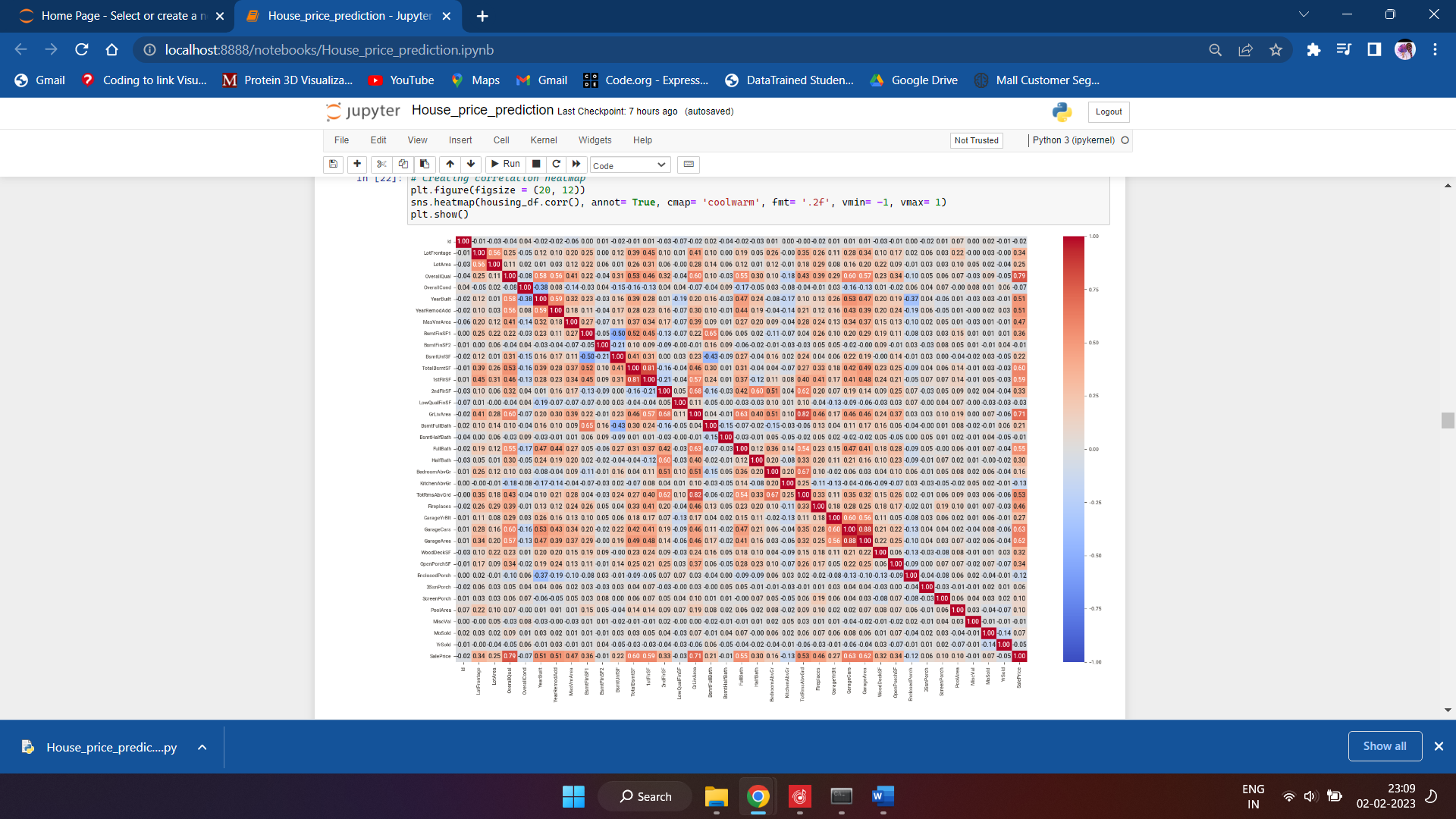
**Visualizations:**

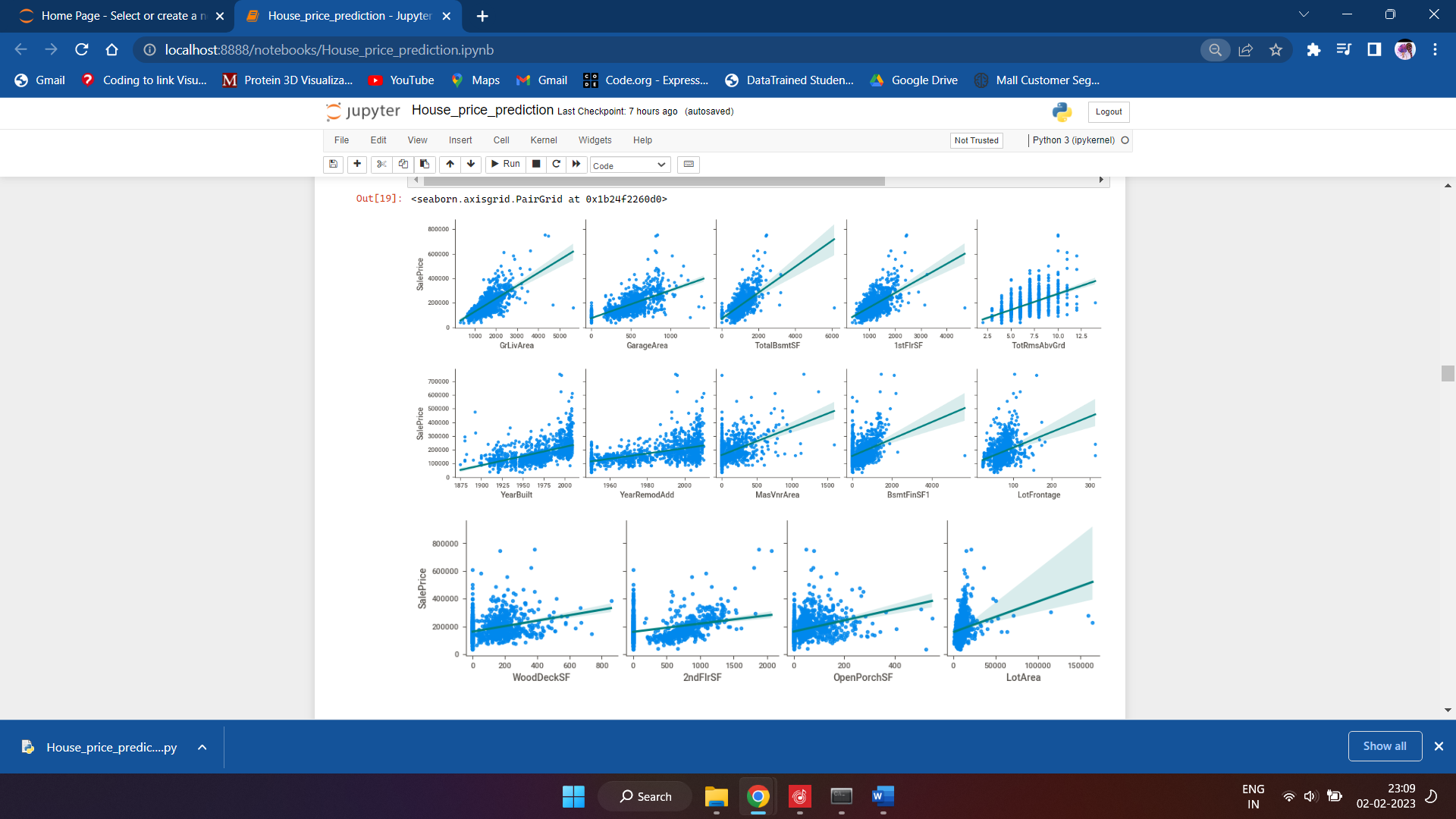












**CONCLUSION:**

Ridge and Lasso both the models have almost same test and train accuracy. So it can be said that there is no overfitting.

Lasso and Ridge both have similar r2 score and MAE on test dataset. But Lasso has eliminated 110 features and final no. of features in Lasso Regression model is 116. Where Ridge has all 226 features. So, our Lasso model is simpler than Ridge with having similar r2 score and MAE.

\* Ridge Regression model on test dataset: r2 score= 0.8912, MAE= 0.0934, RMSE= 0.1357

\* Lasso Regression model on test dataset: r2 score= 0.8947, MAE= 0.0914, RMSE= 0.1335

Considering above points we can choose our Lasso Regression model as our final model.

Below are the top 25 features in the Lasso regression model.