IBM MACHINE LEARNING

**Fatima, Sayeda**

**9/24/2022**

­­­-­­

SUPERVISED MACHINE LEARNING: REGRESSION MODELS FOR HOUSE PRICE PREDICTION

**Table of Contents**

[**1) Project Overview** 2](#_Toc114894107)

[**2) About the Dataset** 3](#_Toc114894108)

[**2a) Brief description of the data set you chose:** 3](#_Toc114894109)

[**2b) Summary of Data Attributes** 3](#_Toc114894110)

[**3) Main Objectives of Analysis** 4](#_Toc114894111)

[**4) Data Exploration, Data Cleansing and Features Engineering** 5](#_Toc114894112)

[**4a) Data Exploration** 5](#_Toc114894113)

[**4b) Data Cleansing & Features Engineering** 9](#_Toc114894114)

[**5) Summary of Training Different Regression Models** 16](#_Toc114894115)

[**5a) Machine Learning Regression Algorithm Development Approach** 16](#_Toc114894116)

[**5b) Summarizing Employed Models** 20](#_Toc114894117)

[**1) Ridge Regression (RR) Models** 20](#_Toc114894118)

[**2) Lasso Regression (LR) Models** 23](#_Toc114894119)

[**3) Elastic-Net (EN) Models** 26](#_Toc114894120)

[**4) Extreme Gradient Boosting (XGB) Regression Models** 29](#_Toc114894121)

[**6) Key Findings to the Main Objectives of Analysis** 32](#_Toc114894122)

[**6a) Result Summary** 32](#_Toc114894123)

[**6b) Recommended Model and Justification** 32](#_Toc114894124)

[**6c) Summarizing Model Drivers** 33](#_Toc114894125)

[**6d) Enlisting Top Contributory Factors** 33](#_Toc114894126)

[**6e) Visualizing Top Contributory Factors Driving Sale Value** 34](#_Toc114894127)

[**6f) Sale Price Prediction on Test/New Data** 35](#_Toc114894128)

[**7) Future Directions** 36](#_Toc114894129)

[**7a) Possible Flaws in Chosen Model** 36](#_Toc114894130)

[**7b) Recommendations** 36](#_Toc114894131)

[**8) Useful Links** 37](#_Toc114894132)

[**8a) Link to Other Useful Models** 37](#_Toc114894133)

[**8b) Github Link to Assignment Notebook and Other Files** 37](#_Toc114894134)

[**References** 38](#_Toc114894135)

# **1) Project Overview**

A fundamental issue real estate business face is assessing sale price based on house attributes. Grounded on hedonic price modelling theory, not only do neighbourhood-specific characteristics but also unit-specific attributes greatly drive house prices (Herath and Maier, 2010). Hence, identification of key factors driving sale prices of real estate is crucial to facilitate informed purchase decisions. It is here that Regression Machine Learning models can be very useful to gain deeper insight into underlying factors as well as their relationship in driving and estimating fair sale price of houses.

Hence, the main aim of the following regression modelling and analysis approach is to enable the business to:

\* Find best regression model for house price prediction

\* Identify different factors influencing house prices

\* Predict sale price based on contributory factors

# **2) About the Dataset**

## **2a) Brief description of the data set you chose:**

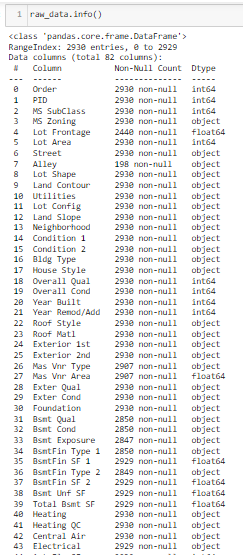
This project uses a hypothetical dataset 'Ames, Iowa Housing Dataset' which was downloaded from the following link:

<https://www.kaggle.com/datasets/prevek18/ames-housing-dataset>

## **2b) Summary of Data Attributes**

The dataset exhibits 2,930 data points (rows) and 82 features (columns) reflecting on housing characteristics.

The data also comes with ‘SalePrice’ Column which represents the Class requiring prediction.



# **3) Main Objectives of Analysis**

Real estate business performance is largely dependent on paying fair price of assets to prevent overpriced purchases and minimize loss. Hence, these businesses are continuously faced with the challenge to estimate realistic house prices and often rely on manual application of Hedonic Price Method (HPM) or hedonic regression analysis. Consequently, driven by HPM, this analysis is targeted towards answering the following queries

* What are the various contributory factors which drive house prices in a given area?
* Based on important factors, what will be projected price for different which housing units?

As a consequence, implementation of an automated machine learning (ML) HPM regression modelling process will enable the organization to:

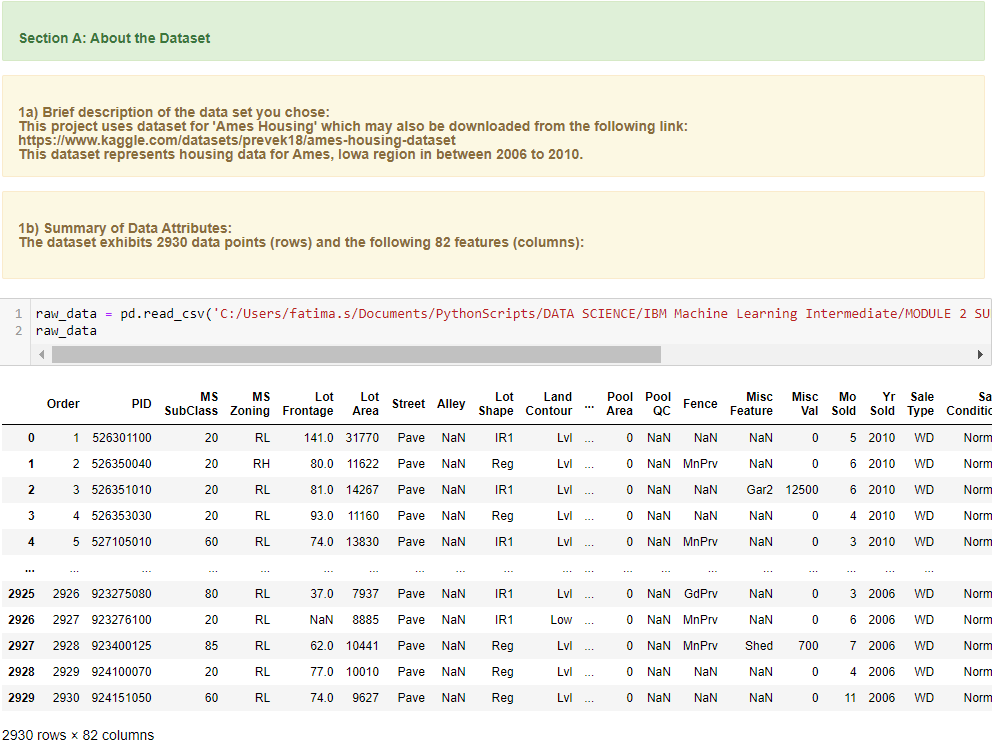
* identify key underlying factors which can appreciate or depreciate property values
* save valuable resources and funds in purchasing properties at right values
* effortlessly employ best ML HPM model and generate report with the click of a button

# **4) Data Exploration, Data Cleansing and Features Engineering**

Since the quality of any machine learning model highly depends on quality of data, hence, this stage is not only most important but is also time consuming. Hence, it was conducted in a step-by-step manner.

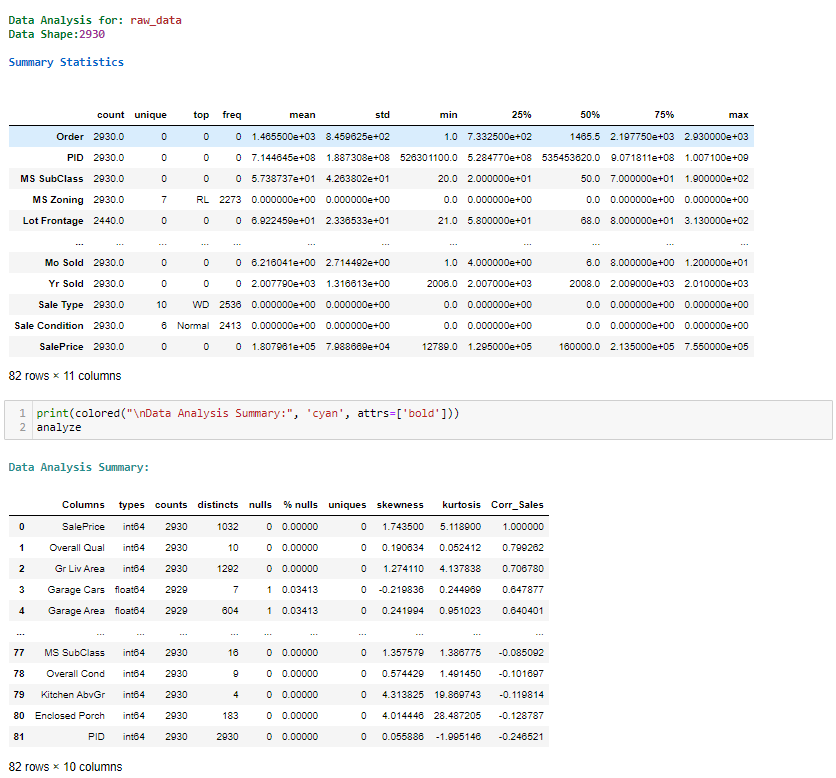
## **4a) Data Exploration**

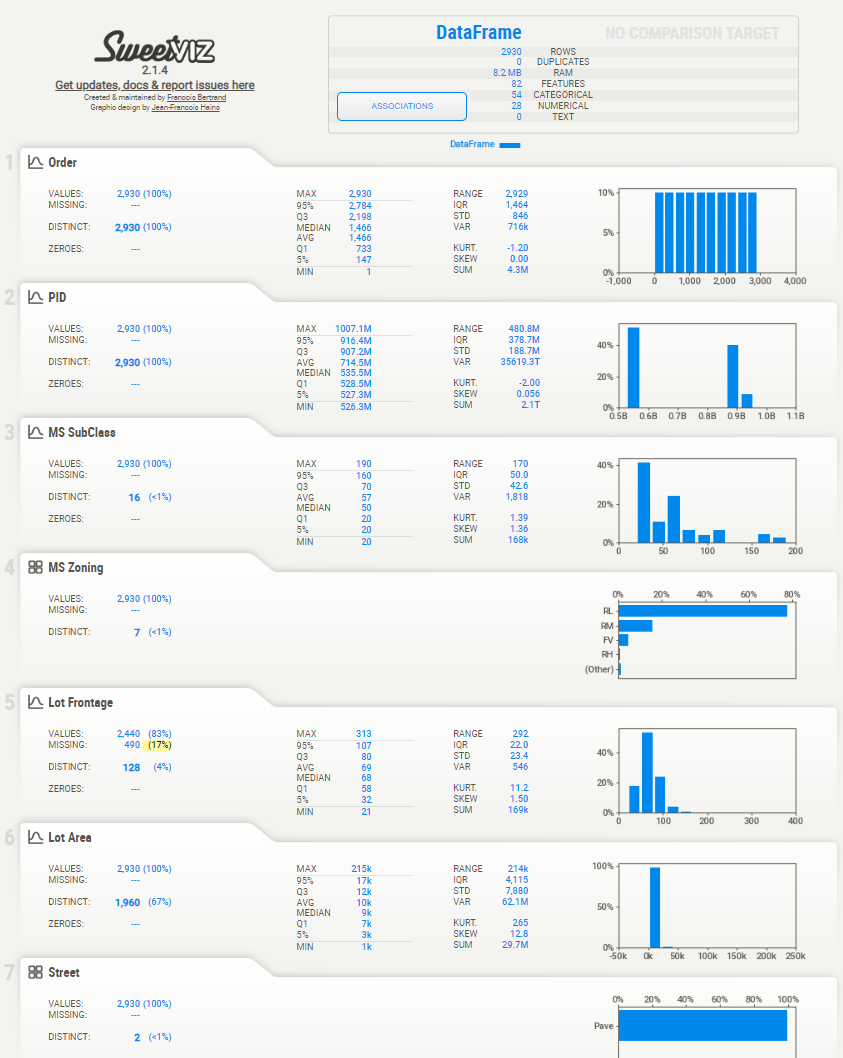
* Data was first loaded into pandas dataframe



* A method was created to conduct preliminary analysis including computation of:
* Descriptive statistics to summarize shape of a dataset’s distribution, its dispersion and central tendency
* Data analysis to depict data types, skewness, kurtosis, etc to facilitate subsequent data cleansing





* Additional Automated Exploratory Data Analysis was performed using Sweetviz to realize visual representation.

## **4b) Data Cleansing & Features Engineering**

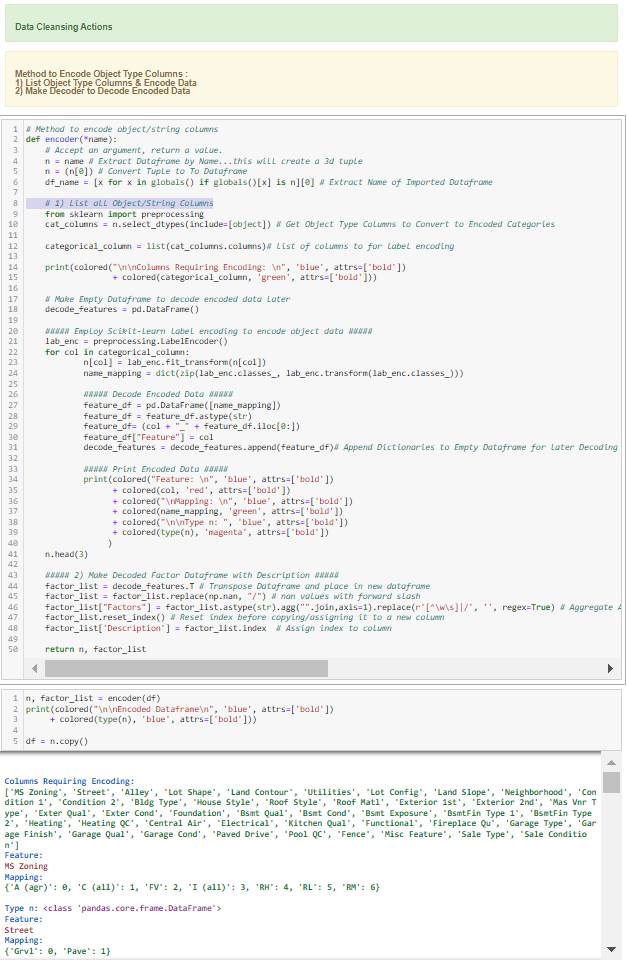
In machine learning, feature selection is the method to reduce the number of input variables during developing predictive modelling. This reduction in input variables is necessary not only to minimize computational cost of modelling but also to achieve improved performance of the model itself.

Among widely practices feature selection approaches include statistical-based feature selection methods which use statistical measures to evaluate relationship between each input variable and the target variable and then select those exhibiting strongest relationship with the latter. While these methods can be both speedy and effective, however, the ultimate choice of statistical measure is largely dependent on data types of both of these variables.

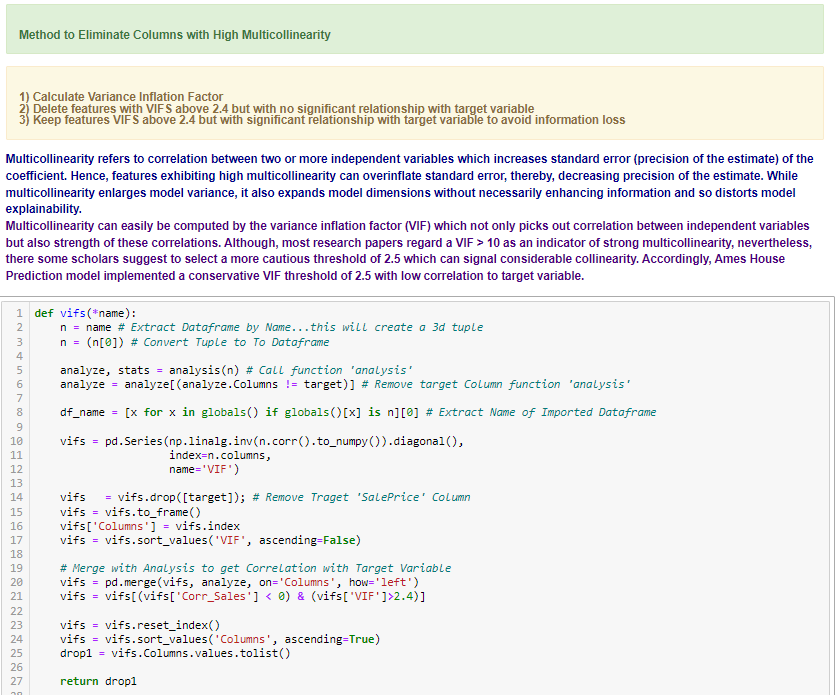
Irrespective of the statistical measure being employed, two dominant feature selection techniques, that is supervised and unsupervised, exist where the former can be further categorized into wrapper, filter and intrinsic techniques. Filter-based feature selection methods employs statistical measures to evaluate correlation between input and output variables so that those exhibiting highest correlations are selected. Statistical measures employed in filter-based feature selection are normally univariate in nature since they evaluate relationship of single input variables one by one with target variable, disregarding their interaction with each other.

Consequently, adopting filter-based feature selection methods, the housing price prediction model approached filter engineering in three steps.

1. Data Encoding
2. Managing Multicollinearity
3. Final Data Cleansing
4. Applying Outlier Treatment
5. Prior to final features selection, data encoding of object or string columns was carried out to facilitate any statistical computation during features selection process. Hence, after copying original dataset, an automated method was created and employed to encode object data using Scikit-learn label encoder.



1. Subsequently, multicollinearity was managed by another automated method, as follows:



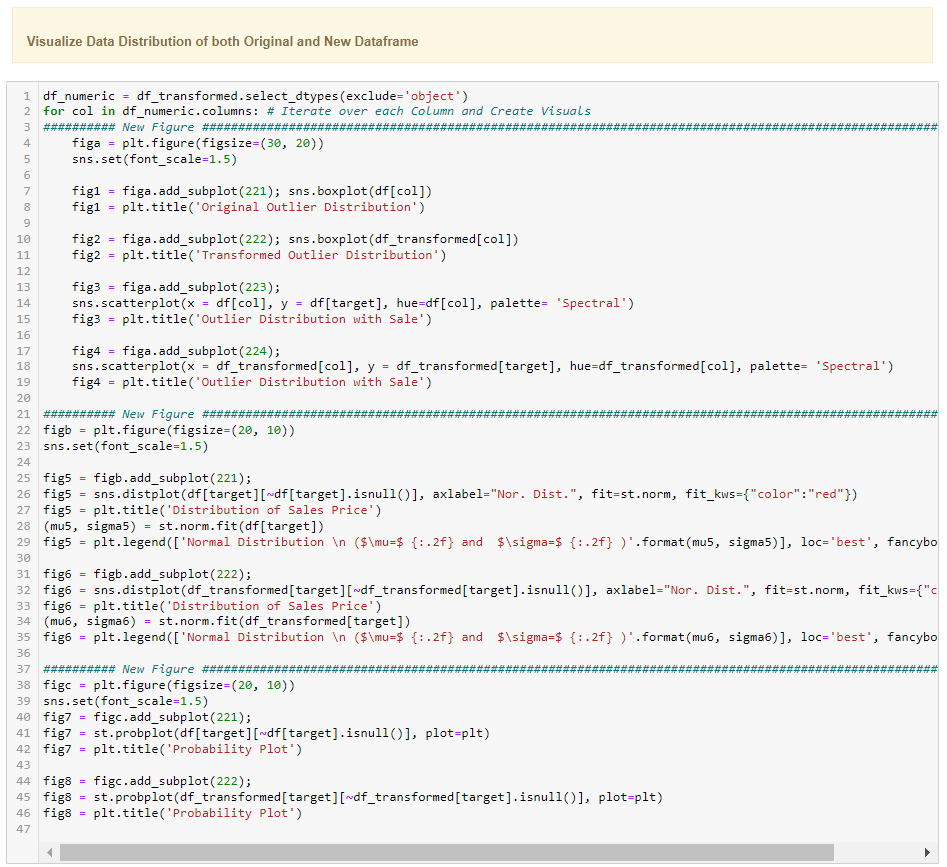
1. Using an additional automated method, statistical measures were then employed with supervised filter-based feature selection technique. Firstly, aforementioned “vif” method was called to identify columns exhibiting high multicollinearity but low correlation with sales. Then, columns with less unique features were marked. Consequently, highly skewed columns with low correlation to target were also enlisted. Lastly, Columns with extremely high null values were also keyed out. The lists were then combined to filter these columns out of the data-frame.

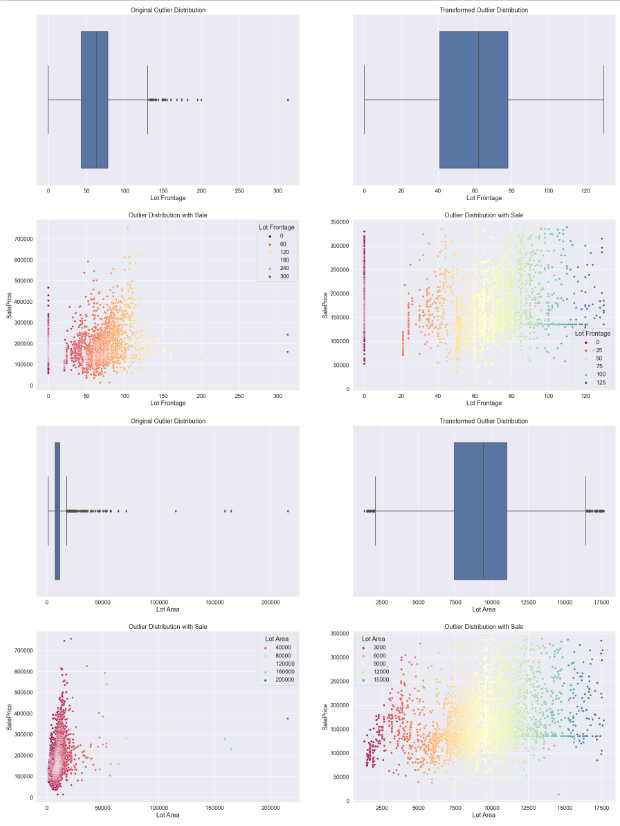


1. Lastly, an automated method was employed to replace outliers with mode, that is, most frequent value.



This was followed by another method to present graphical illustration of original and adjusted data side by side:



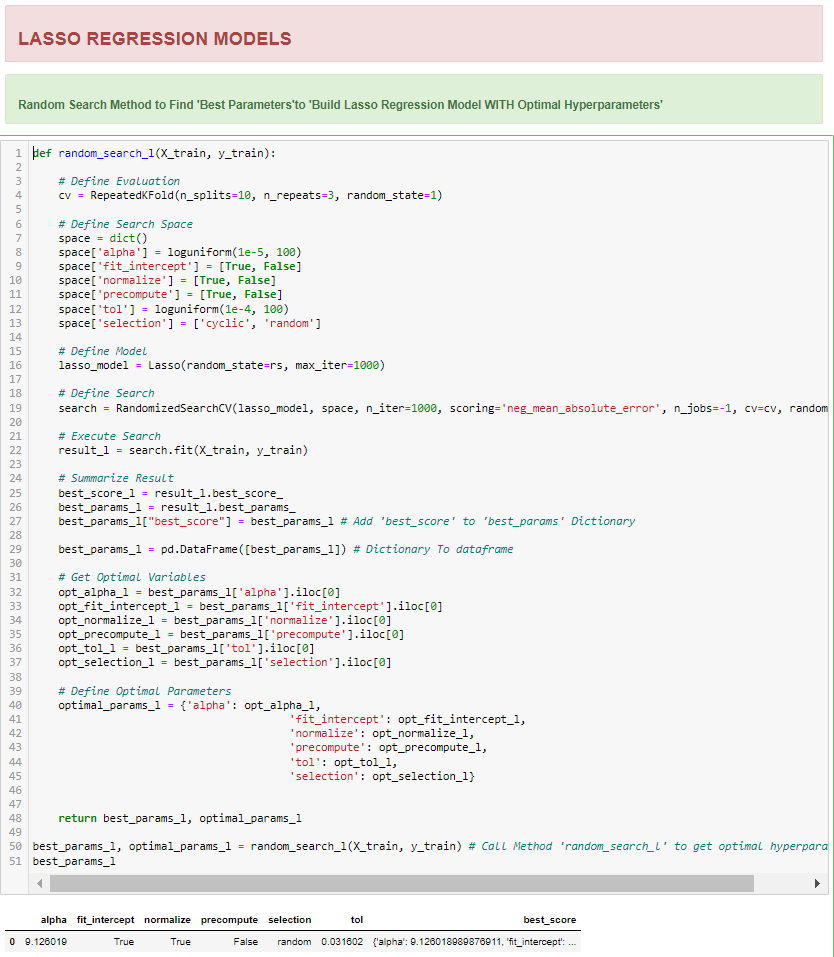


# **5) Summary of Training Different Regression Models**

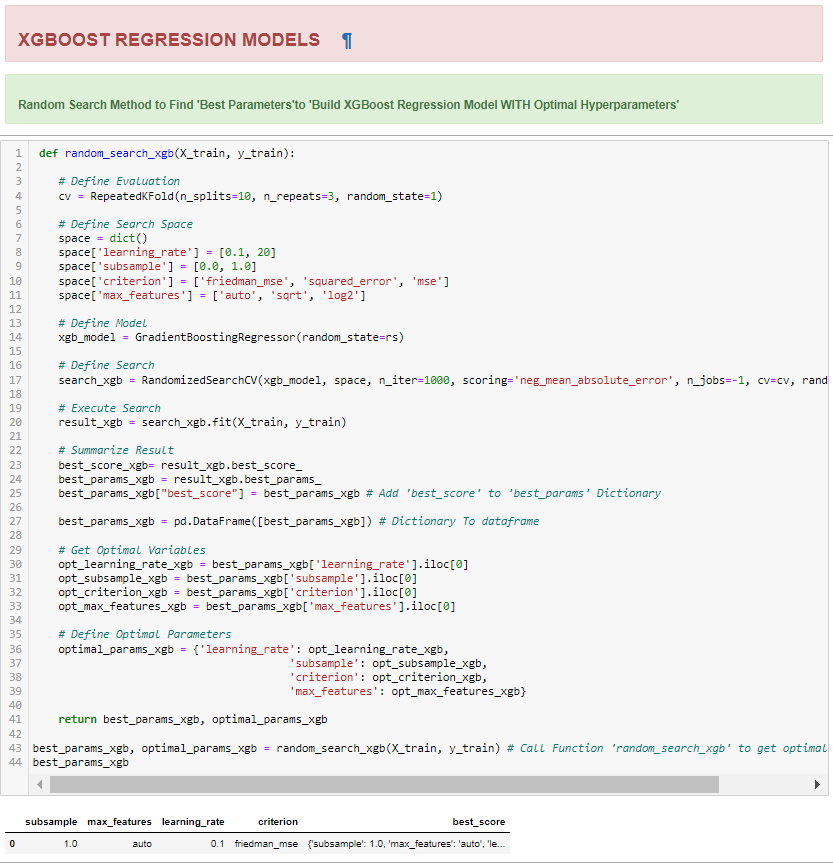
## **5a) Machine Learning Regression Algorithm Development Approach**

Using random search, an automated optimal hyper-parameter search method was created to find optimal model parameters. This approach was employed because best hyperparameters are not automatically learnt within estimators and its manual search not only slows down model development but may also lead to ineffective model construction. Hence, an exhaustive random search approach was used to pass parameter arguments to the constructor in order to find optimal hyperparameters for each model, as shown below:









## **5b) Summarizing Employed Models**

Following four main regression models have been used to predict house prices.

### **1) Ridge Regression (RR) Models**

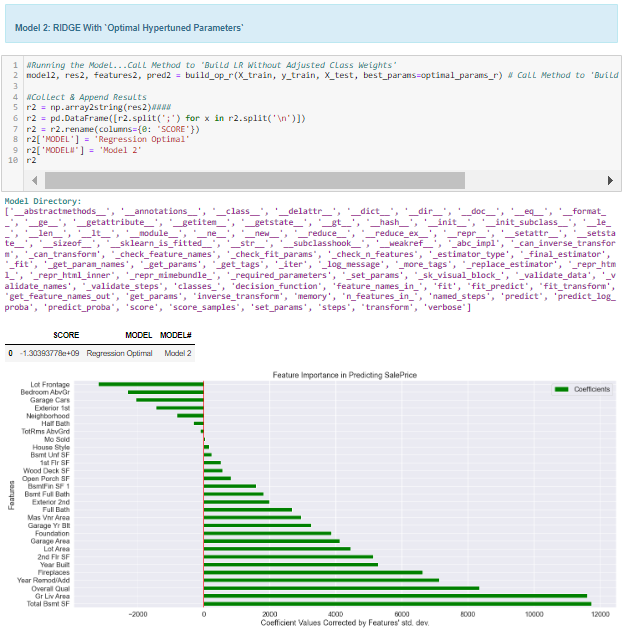
Due to the dependant nature of multiple variables in predicting variable 'y' where the output is influenced by multiple factors, ridge regression algorithm was employed. A single method was created to measure predictive capability of the following two RR models:



**1a) RR Model 1 Without Optimal Hyperparameter Tuning:** A simple algorithm was created without optimal hyperparameter tuning.



**1b) RR Model 2 WITH Optimal Hyperparameter Tuning:** A modified algorithm with auto-hyper-tuned parameters was created using random search method described above.



### **2) Lasso Regression (LR) Models**

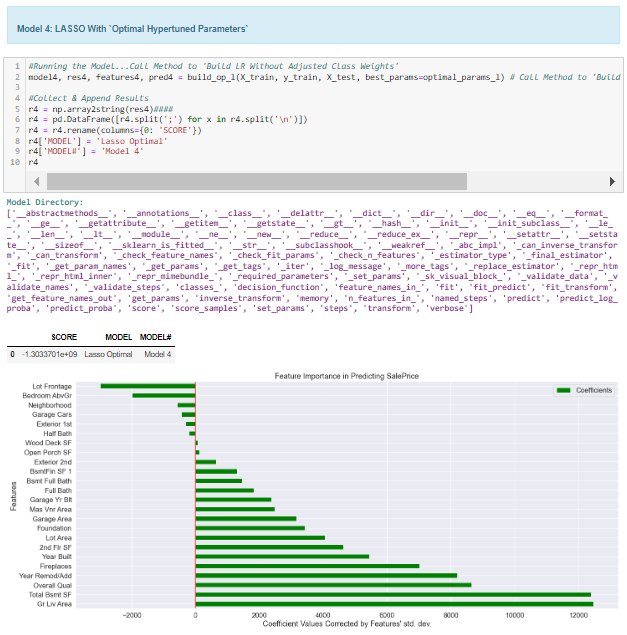
A single method was employed to run lasso regression models with and without optimal hyper-parameter tuning:



**2a) LR Model 1 Without Optimal Hyperparameter Tuning:** A simple algorithm was created without any hyperparameter tuning.



**2b) LR Model 2 WITH Optimal Hyperparameter Tuning:** A modified algorithm with auto hyper-tuned parameters was created using random search method described above.

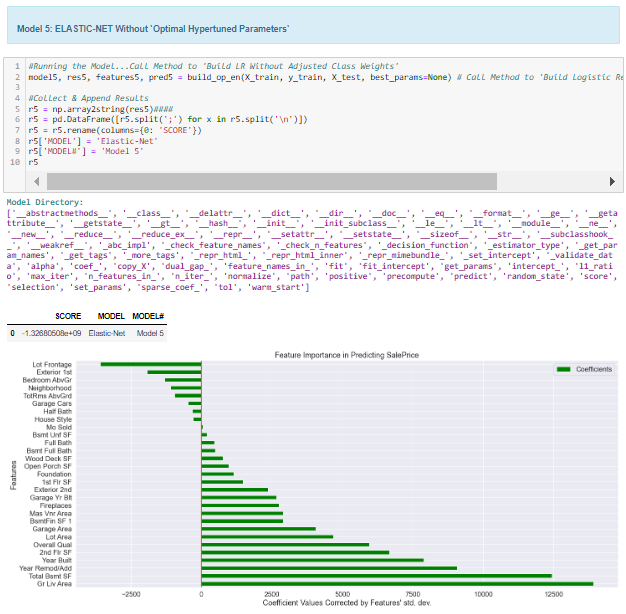


### **3) Elastic-Net (EN) Models**

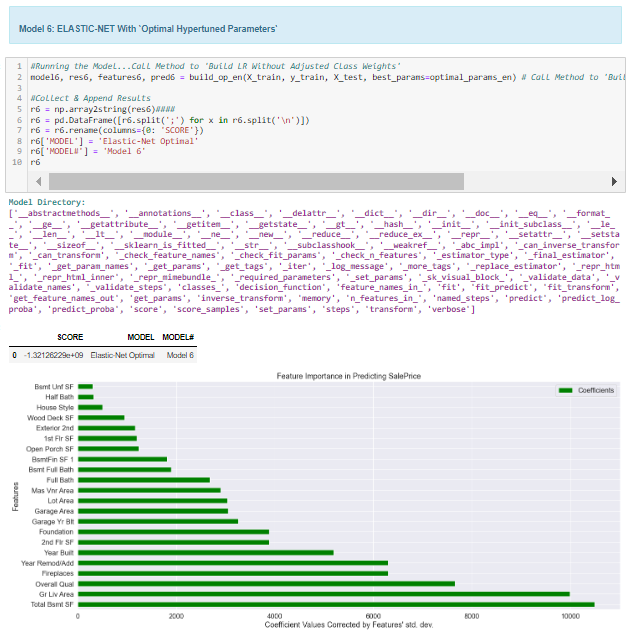
A single method was employed to run elastic-net models with and without optimal hyper-parameter tuning:



**3a) EN Model 1 Without Optimal Hyperparameter Tuning:** A simple algorithm was created without any hyperparameter tuning:



**3b) EN Model 2 WITH Optimal Hyperparameter Tuning:** A modified algorithm with auto hyper-tuned parameters was created using random search method described above:



### **4) Extreme Gradient Boosting (XGB) Regression Models**

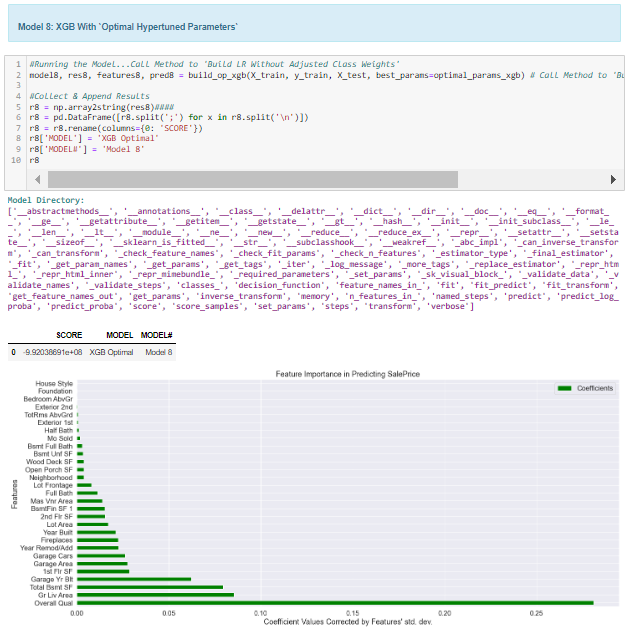
A single method was employed to run XGB regression models with and without optimal hyper-parameter tuning:



**4a) XGB Regression Model 1 Without Optimal Hyperparameter Tuning:** A simple algorithm was created without any hyperparameter tuning:

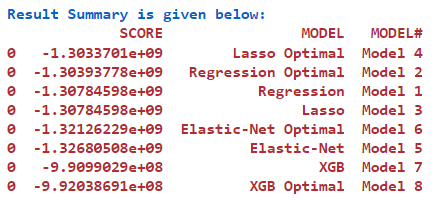


**4b) XGB Regression Model 2 WITH Optimal Hyperparameter Tuning:** A modified algorithm with auto hyper-tuned parameters was created using random search method described above:

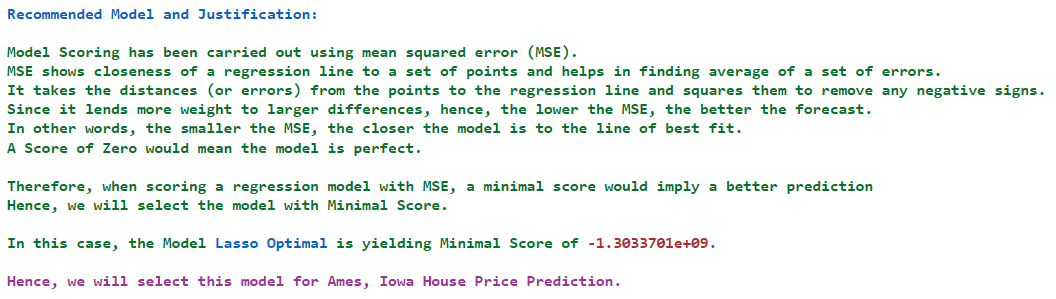


# **6) Key Findings to the Main Objectives of Analysis**

## **6a) Result Summary**



## **6b) Recommended Model and Justification**



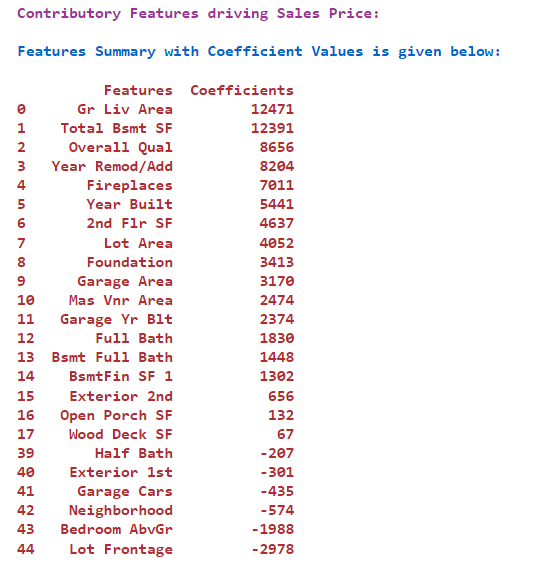
## **6c) Summarizing Model Drivers**

Main Drivers behind top performing Lasso Optimal model are as follows:

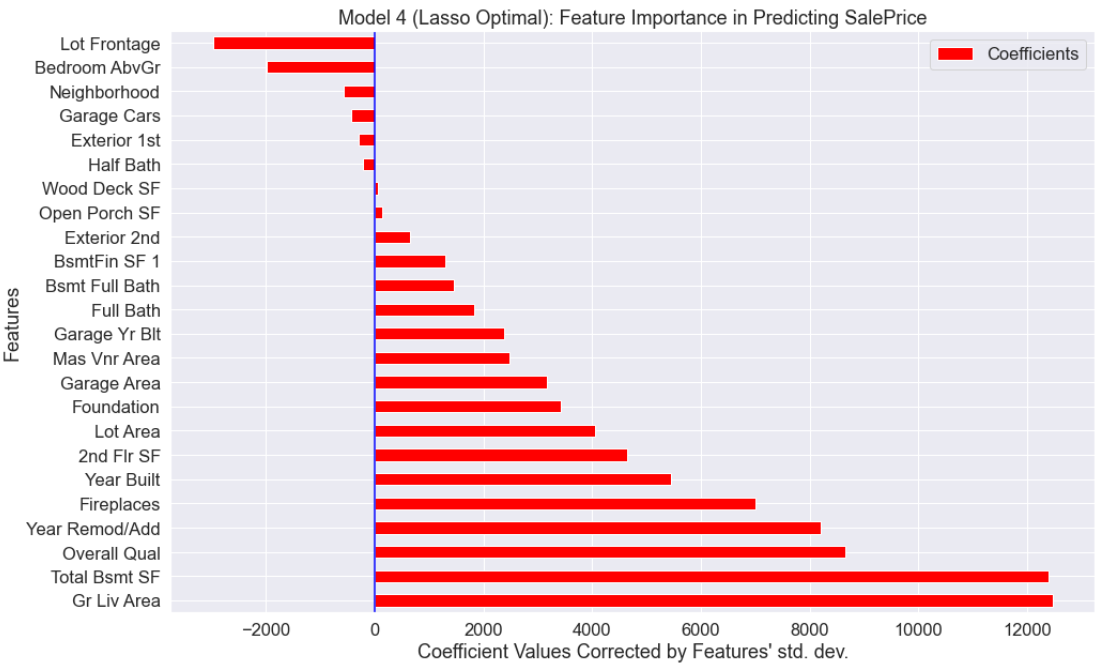
* Uses random search method to find optimal parameters to achieve best hyper-tuning
* Runs lasso regression model with above drivers to get top contributory features

## **6d) Enlisting Top Contributory Factors**

Top Factors Contributing to features driving sale value are cited under in ascending order of importance:

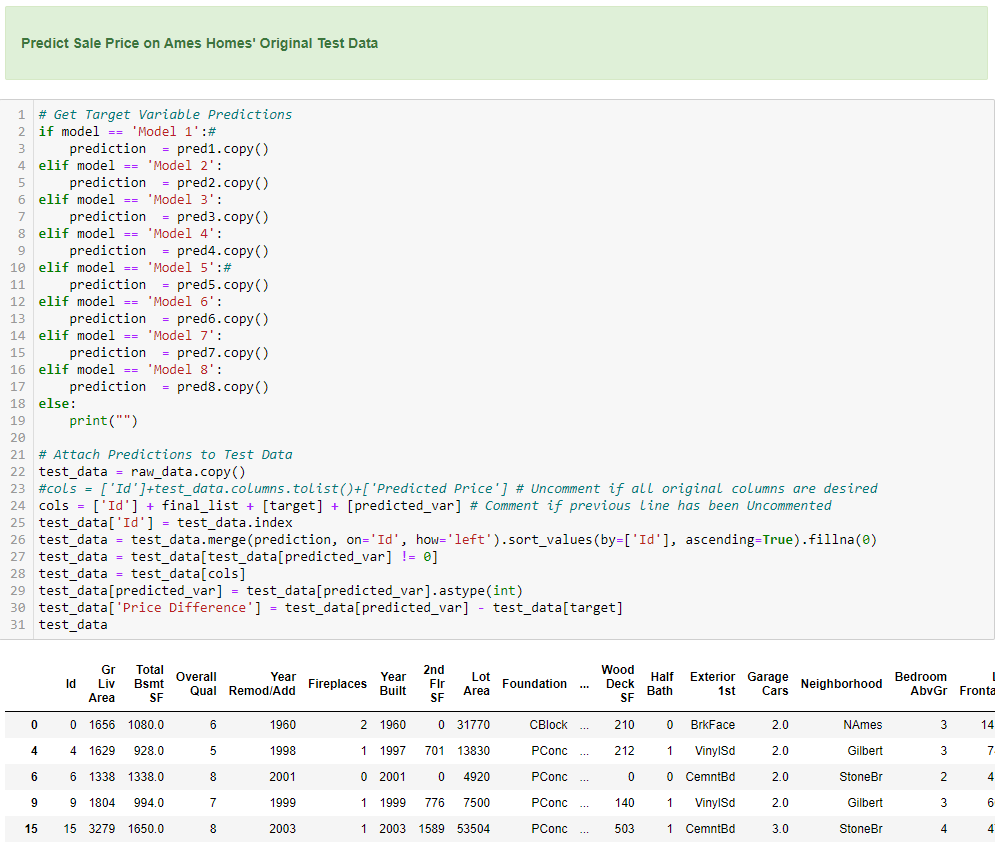


## **6e) Visualizing Top Contributory Factors Driving Sale Value**



## **6f) Sale Price Prediction on Test/New Data**

Predicted sale price for test data were extracted from top performing model as follows:

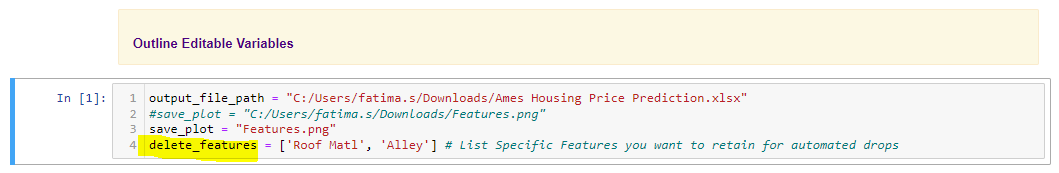


These findings were compiled and subsequently exported in the form of a well formatted coloured excel report.

# **7) Future Directions**

## **7a) Possible Flaws in Chosen Model**

* The model uses Mean Squared Error (MSE) as scoring method which may be highly biased for higher values.
* Just like “delete\_features” list, the model needs to incorporate user input list to keep features deemed necessary so that they are not automatically dropped



## **7b) Recommendations**

Following suggestions are likely to improve the model even further:

* Incorporate “keep\_features” list and use it to eliminate these from final “drop” list in row 16 of project notebook



* Introduce more models like Decision Trees and Random Forest
* Implement a method to combine best performing models to ensure enhanced performance and more effective generalization, (see, for example, 8) Useful Links, 8a-a)
* Apply other deep learning models like TensorFlow
* Because of biasness of MSE towards higher values, scoring may be substituted by Root Mean Squared Error (RMSE) which may reflect model performance whilst dealing with increased error values.
* Apply plot to depict both MSE and RMSE (see, Machines, 2022)

# **8) Useful Links**

## **8a) Link to Other Useful Models**

1. <https://www.kaggle.com/code/mgmarques/houses-prices-complete-solution>
2. <https://www.kaggle.com/code/marto24/beginners-prediction-top3>
3. <https://www.kaggle.com/code/mchatham/ames-housing-regression>
4. <https://www.kaggle.com/code/mkariithi/real-estate-sales-price-prediction/notebook>
5. <https://www.kaggle.com/code/bashkeel/eda-to-ensemble-model-lasso-ridge-xgboost>
6. <https://www.kaggle.com/code/gerlandore/advanced-house-regression-eda-model-comparison>
7. <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset/notebook>
8. <https://www.kaggle.com/search?q=ADVANCED+LINEAR+REGRESSION+BOSTON+HOUSE+PREDICTION>
9. <https://www.kaggle.com/code/koki25ando/nba-salary-prediction-using-multiple-regression>

## **8b) Github Link to Assignment Notebook and Other Files**

<https://github.com/FATIMASP/IBM-MACHINE-LEARNING-CERTIFICATION/tree/main/Supervised%20Machine%20Learning:%20Regression>

# **References**

Herath, S. & Maier, G., 2010. *The hedonic price method in real estate and housing market research: a review of the literature, (pp. 1-21),* Vienna, Austria: University of Economics and Business: Institute for Regional Development and Environment,

Machines, I. L., 2022. *Mean Squared Error.* [Online]   
Available at: https://insidelearningmachines.com/mean\_squared\_error/  
[Accessed 24 09 2022].