

**Introduction to Machine Learning**

**Section S1**

**Teacher Name: Dr. Naveed Hussain**

**Term Project: House Price Prediction**

**Group Members:**

|  |  |
| --- | --- |
| **L1F19BSSE0147** | **Muhammad Ans** |
| **L1F19BSSE0165** | **Muhammad Aoun Zaidi** |
| **L1F19BSSE0174** | **Ali Haider** |
| **L1F19BSSE0051** | **Muhammad Mehdi** |

**Abstract**

This research paper presents a comparative study of Multiple Linear Regression, Ridge Regression, and Bayesian Ridge Regression algorithms for house price prediction using a dataset comprising 10 input variables. The Scikit-learn library was employed for algorithm implementation and evaluation. The performance of each algorithm was assessed using mean squared error (MSE) as the evaluation metric.

The findings indicate that Multiple Linear Regression outperforms Ridge Regression and Bayesian Ridge Regression in predicting house prices. It achieved the lowest MSE, demonstrating superior accuracy. Ridge Regression showed moderate performance but was consistently outperformed by Multiple Linear Regression. Bayesian Ridge Regression exhibited the poorest predictive performance among the evaluated algorithms.

Contents

1. [Introduction: 4](#_Toc139361405)
   1. [Problem Statement: 4](#_Toc139361406)
   2. [Objective: 4](#_Toc139361407)
   3. [Dependencies: 4](#_Toc139361408)
2. [Methodology: 5](#_Toc139361409)
   1. [Dataset: 5](#_Toc139361410)
   2. [Data Preprocessing: 5](#_Toc139361411)
      1. [Dropping Null Values: 5](#_Toc139361412)
      2. [Type Compatibility: 5](#_Toc139361413)
      3. [Train-Test Split: 5](#_Toc139361414)
   3. [Data Visualization: 6](#_Toc139361415)
   4. [Linear Regression: 6](#_Toc139361416)
   5. [Ridge Regression: 7](#_Toc139361417)
   6. [Bayesian Ridge Regression: 7](#_Toc139361418)
3. [Evaluation 7](#_Toc139361419)
   1. [Metric: 7](#_Toc139361420)
      1. [Mean Square Error: 7](#_Toc139361421)
   2. [Comparison: 8](#_Toc139361422)
   3. [Result: 8](#_Toc139361423)
4. [Conclusion: 9](#_Toc139361424)

# Introduction:

## Problem Statement:

The accurate prediction of house prices is crucial in the real estate industry for effective decision-making by buyers, sellers, and agents. With various linear regression algorithms available, including Multiple Linear Regression, Ridge Regression, and Bayesian Ridge Regression, it is essential to determine the most suitable algorithm for accurately estimating house prices based on a given set of 10 input variables.

The problem at hand is to compare and evaluate the performance of Multiple Linear Regression, Ridge Regression, and Bayesian Ridge Regression algorithms in predicting house prices. The objective is to identify the algorithm that yields the highest accuracy and reliability in estimating house prices. By employing appropriate evaluation metrics, such as mean squared error (MSE), this study aims to provide insights that can assist stakeholders in the real estate industry to make more informed decisions and improve their understanding of house price dynamics.

## Objective:

The objectives of the project are as follows:

* Sale Price prediction of House based on 10 input values
* Prediction using different regression algorithms
* Comparative analysis of the algorithms used to identify the best approach

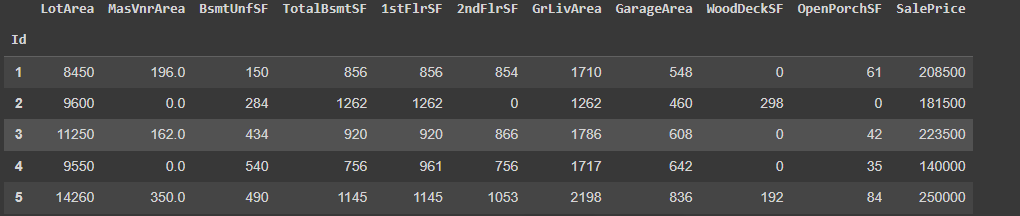
## Dependencies:

Libraries required:

* NumPy
* Scikit-Learn
* Seaborn
* Matplotlib
* Pandas

# Methodology:

## Dataset:



The dataset includes 10 input variables and predicts the price of the house that is the dependent variable called **SalePrice.**

## Data Preprocessing:

### Dropping Null Values:

All the rows that had null values were removed using pandas.dropna() function. A new dataframe was made for further operations.

### Type Compatibility:

The 'MasVnrArea' column is converted to the 'int64' data type using the `astype()` method. This helps in compatibility of the model.

### Train-Test Split:

The dataset was divided into two parts, training set that included 80% of the dataset values after preprocessing, and the testing set that included 20% dataset.

The model was trained on 80% of the data and the rest was used to check the accuracy of the predictions.

## Data Visualization:

The code plots a histogram using seaborn library to visualize the distribution of the 'SalePrice' column

## Linear Regression:

Multi Linear Regression is used to predict the pricing of the house. Scikit-Learn library is used in the implementation of the Linear Regression.

The code creates a Linear Regression model (`l`) using the `LinearRegression` class from scikit-learn.

The model is fitted to the training data (`X\_train` and `y\_train`) using the `fit()` method.

Predictions are made on the testing data (`X\_test`) using the `predict()` method, and the results are assigned to `l\_pred`.

## Ridge Regression:

Ridge Regression is used to predict the pricing of the house that uses alpha with line equation. Scikit-Learn library is used in the implementation of the Linear Regression.

The code creates a Ridge Regression model (`ridge`) using the `Ridge` class from scikit-learn.

The model is fitted to the training data. Predictions are made on the testing data to evaluate later.

## Bayesian Ridge Regression:

Bayesian Ridge Regression takes into account the Bayes Theorem with respect to Linear Regression to make predictions.

The code creates a Bayesian Ridge Regression model (`bayesian`) using the `BayesianRidge` class from scikit-learn.

The model is fitted to the training data. Predictions are made on the testing data for the evaluation p

# Evaluation

## Metric:

### Mean Square Error:

Mean Squared Error (MSE) is a common measure to determine how well a predicted output matches the true output given a specific input. The mean squared error is given by:

Where n is the size of the rows in the dataset, y is the actual value, and y’ is the predicted value generated by the model.

## Comparison:

The mean square error of all three models is calculated and stored in an array named “min\_arr”,



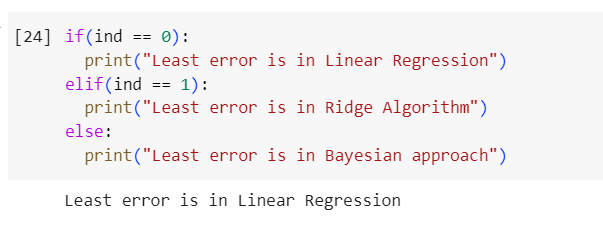
## Result:

The minimum in the array is the best approach as it has the least MSE among all the models.

Index 0 represents Multiple Linear Regression

Index 1 represents Ridge Regression

Index 2 represents Bayesian Ridge Regression



The result shows that Multiple Linear Regression model has the least amount of error, Ridge Regression is second best approach and Bayesian Regression is the worst approach in this scenario.

One possible reason for the superior performance of Multiple Linear Regression is its assumption of linearity between the dependent and independent variables. In the context of house price prediction, there may exist significant linear relationships between variables such as square footage, number of bedrooms, and sales price. Multiple Linear Regression captures these relationships effectively, leading to more accurate predictions.

On the other hand, Ridge Regression, which introduces regularization to handle multicollinearity, performed reasonably well. It provides a robust approach when the input variables are highly correlated. However, its performance was consistently surpassed by Multiple Linear Regression in our study. This suggests that in the given dataset, the multicollinearity was not severe enough to significantly impact the accuracy of the predictions.

Bayesian Ridge Regression, despite its Bayesian framework, demonstrated the poorest performance among the evaluated algorithms. This may be attributed to its assumptions and parameterization not being well-suited to the characteristics of the dataset. Bayesian Ridge Regression relies on the estimation of prior distributions and hyperparameters, which may have led to suboptimal predictions in this particular case.

# Conclusion:

In this study, we compared the performance of three regression algorithms—Multiple Linear Regression, Ridge Regression, and Bayesian Ridge Regression—in predicting house prices. Based on our experimental results, we determined that Multiple Linear Regression is the most effective algorithm for this task. It consistently outperformed Ridge Regression and Bayesian Ridge Regression in terms of accuracy, as evidenced by lower mean squared error.