

HOUSE PRICE PREDICTION USING REGULARIZATION METHODS

ABSTRACT:

In this study, we use the "house_price_dataset" to predict "SalePrice" using three different linear regression methods: Ridge, Lasso, and Elastic Net. We place special emphasis on certain characteristics, such as LotArea, MasVnrArea, YearBuilt, LotFrontage, BsmtFinSF1, BsmtUnfSF, GrLivArea, OverallQual, and TotalBsmtSF. The data is preprocessed, divided into training (70%) and testing (30%) sets, and regression models with gradient descent optimization are implemented from scratch. We thoroughly compare the performance of these models by measuring their accuracy with metrics like Mean Squared Error (MSE), Absolute Errors, Root Mean Squared Error (RMSE), R-squared (R^2), and Adjusted R-squared. This study helps in choosing the best model for practical applications by offering useful insights into how well certain regression algorithms predict house prices.

KEYWORDS: Ridge, Lasso, Elastic Net, SalePrice, regression, r^2 -squared

INTRODUCTION:

A) PROBLEM DEFINITION:

The current challenge is to develop precise predictive models employing the "house_price_dataset" for calculating the "SalePrice" of residential properties. In order to complete this work, you must overcome issues like handling missing data, encoding categorical variables, and spotting and dealing with outliers in the dataset. In addition, choosing the ideal feature sets and hyperparameters, as well as the most suitable regression technique—be it Ridge, Lasso, or Elastic Net—is essential for making correct predictions. The primary objective is to carefully assess and contrast the performance of these models using a variety of metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), and Adjusted R-squared, in order to identify the most efficient method for estimating property prices, which will aid in real estate pricing and investment decision-making processes.

Aim of this project:

To implement the three linear regression models named Lasso, Ridge, ElasticNet by using gradient descent (GD) optimization method from scratch and evaluate performance of models.

B) BACKGROUND STUDY:

Linear regression is a fundamental statistical approach for modeling relationships between one or more independent variables and a dependent variable. It is based on a linear connection and seeks coefficients that minimize the sum of squared errors. However, multicollinearity (correlations between independent variables) and overfitting can impair model performance in real-world datasets. Ridge, Lasso, and Elastic Net regression algorithms were created to overcome these issues.

Ridge regression includes a regularization element, the L2 penalty, that favors modest coefficient values, hence lowering multicollinearity. In contrast, Lasso regression employs the L1 penalty to promote sparse coefficients, thereby performing feature selection. Elastic Net regression combines L2 and L1 penalties in a way that strikes a balance between Ridge and Lasso, making it resistant to multicollinearity while also picking meaningful features. These techniques have found widespread use in a variety of fields, including finance, economics, and real estate, where predictive modeling is critical for decision-making.

RELATED WORKS:

1. HOUSE PRICE PREDICTION USING MACHINE LEARNING by Atharva Chouthai, Mohammed Athar Rangila, Sanved Amate, Prayag Adhikari, Vijay Kukre.

Numerous factors influence housing sales prices, including the size of the property, its location, the materials used in construction, the age of the property, the number of bedrooms and garages, and so on. The prediction model for dwellings in this paper is built using machine learning algorithms. To construct a predictive model, machine learning techniques such as logistic regression and support vector regression, as well as the Lasso Regression approach and Decision Tree, are used. They looked at housing data from 100 different properties.

2. HOUSE PRICE PREDICTION USING MACHINE LEARNING by Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, Dr. Vinayk A. Bharadi.

This paper surveyed to forecast the price of a house based on supplied features. To develop a prediction model, many Machine Learning models such as Linear Regression, Decision Tree, and Random Forest are employed. They took a step-by-step strategy to data collection, pre-processing data, data analysis, and model building. After that, Random forests produce the best results in terms of training data.

3. HOUSE PRICE PREDICTION USING MACHINE LEARNING by Ayush Kumar Tiwari, Akshita Sharma, Aman Goyal, Pragya Tewari.

In this work, the Decision tree AI estimate is used to create an assumption model for forecasting implicit selling costs for any land parcel. New characteristics, such as air quality and wrongdoing rate, were included to the dataset to help anticipate costs even more accurately. Because these elements aren't frequently associated with the datasets of other assumption structures, this system is fascinating.

C) OBJECTIVES AND CONTRIBUTION:

1. Data preprocessing:

Data preprocessing is used to convert data into clean data sets which are used for analysis. It consists of several steps like data cleansing, data reduction, data enrichment organization and transformation. So, it will be easy for machine learning models to read data and learn from data.

2. Descriptive statistics:

Descriptive statistics are used to describe data and visualize data. Understanding all the features of dataset is required to build model effectively and to evaluate the model. It is also used to know the mean, median and mode of all the features of dataset.

3. Model Training:

Model Training includes the creation of machine learning models. Many prebuilt models can be used but we are creating a linear regression model from scratch. We will train the model using train data which is done by splitting the data into train and test datasets.

4. Model Evaluation:

Model Evaluation is done by calculating different performance evaluation metrics like Mean Square Error (MSE), absolute error, Root means square error (RMSE), R2 values and Adjusted R2 values. This can be done using test dataset.

METHODOLOGY:

- ❖ Data Encoding and preprocessing: In data preprocessing we have analyzed all types of variables like categorical, numerical, temporal variables. There are many missing values in the dataset. We have changed all categorical features null values to “Missing” and changed all numerical features null values to median of that feature. Normalizing the numerical data was also done on the dataset by using standardScaler.
- ❖ Descriptive statistics: In this step we will describe some features of the dataset named LotArea, MasVnrAea, YearBuilt, LotFrontage, BsmtFinSF1, BsmtUnfSF, GrLivArea, VoerallQual, and TotalBsmtSF. We will get count, mean, standard deviation, min and max values of these features of the dataset to understand the data effectively.
- ❖ Visualization of data: All the dataset features are visualized, and analysis of the features are done using graphs.
- ❖ Splitting of data: In this step we will divide the entire dataset into two groups, training set and testing set. Training set is used to train machine learning model and learn from it. Testing data is used to evaluate model performance.
- ❖ Model creation: In this project we will create three linear regression models named Lasso, Ridge, ElasticNet using gradient descent (GD) optimization method from scratch. This is used to find salePrice values of houses.
- ❖ Model training: Model training is done by using training dataset. The model will get the relationship between the features and output, and it will give us the predicted values.
- ❖ Model Evaluation: Model evaluation is done using testing dataset. Using this dataset, we will calculate different performance evaluating features like MSE, Absolute error, RMSE, R2 value and Adjusted R2 values.
- ❖ Analysis of results: The performance evaluating features give us whether the model we created is effective or not. We will also compare the performance of different models.

MODEL DESCRIPTION:

- ❖ **StandardScaler:** A preprocessing method used frequently in machine learning and data analysis is called the StandardScaler. It is a phase in the preprocessing or data transformation process that datasets go through before machine learning models are trained. The StandardScaler job is to standardize or normalize a dataset's features (variables). Standard scaler is used to convert feature values to zero mean and unit standard deviation values.
- ❖ **Lasso regression model:** Lasso regression, also known as "Least Absolute Shrinkage and Selection Operator" regression, is a linear regression version that can be used for prediction as well as feature selection. It adds an L1 regularization factor to the linear regression objective function that penalizes absolute coefficient values, thereby encouraging some coefficients to become exactly zero. This distinguishing feature of Lasso makes it an invaluable tool for automatically picking the most relevant features from large datasets, simplifying models, and reducing multicollinearity. Lasso is a versatile technique for data analysis and predictive modeling because the regularization strength parameter, generally denoted as "alpha," allows for altering the trade-off between model accuracy and the degree of regularization.

Lasso regression model by using gradient descent optimization method from scratch: In this model we will create a model from scratch using optimization method.

- ❖ **Ridge regression model:** Ridge regression is a linear regression approach used for predictive modeling and dealing with dataset multicollinearity. It improves classic linear regression by incorporating an L2 regularization factor into the objective function, which penalizes coefficient squared values. This regularization encourages the coefficients to be small but not exactly zero, decreasing the influence of strongly correlated features while keeping them in the model. Ridge regression is useful for increasing the stability and generalization of linear regression models, particularly when working with datasets with several correlated independent variables. This results in more robust predictions and reduces the danger of overfitting.

Ridge regression model by using gradient descent optimization method from scratch: In this model we will create a model from scratch using optimization method.

- ❖ **ElasticNet regression model:** Elastic Net regression is a sophisticated linear regression technique that combines the characteristics of Ridge and Lasso regression. It integrates L1 (Lasso) and L2 (Ridge) regularization terms into the linear regression objective function, allowing it to handle multicollinearity while still performing feature selection. Elastic Net achieves a balance between the two by adjusting the relative strength of the L1 and L2 penalties. This balance makes it especially useful when dealing with datasets with many highly correlated independent variables, as it can shrink coefficients while also setting some coefficients exactly to zero, effectively selecting relevant features and improving model robustness while maintaining predictive accuracy.

ElasticNet regression model by using gradient descent optimization method from scratch: In this model we will create a model from scratch using optimization method.

EXPERIMENT AND RESULTS:

A) DATABASE:

Database house_price_prediction has train dataset and data description; train dataset has 1460 observations and 81 variables.

df

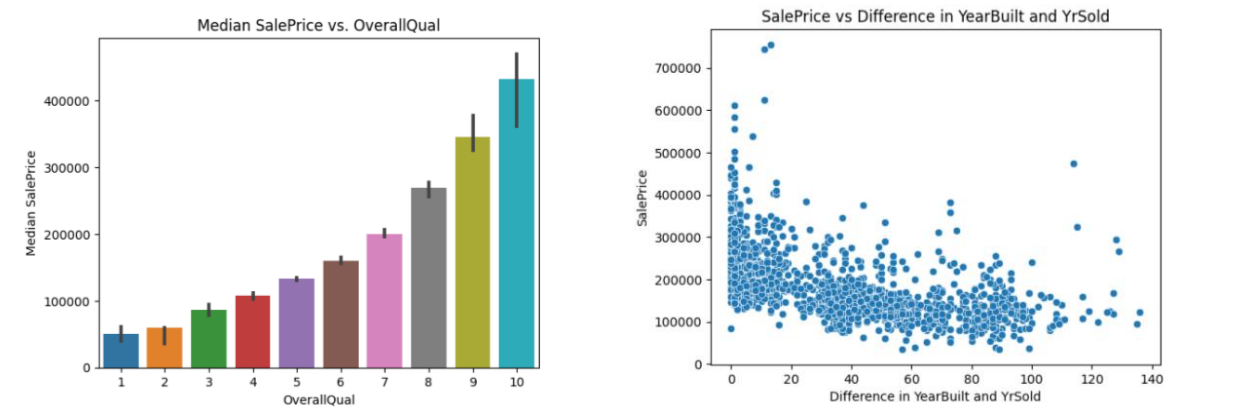
	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm
...
1455	1456	60	RL	62.0	7917	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
1456	1457	20	RL	85.0	13175	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	NWAmes	Norm
1457	1458	70	RL	66.0	9042	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Crawfor	Norm
1458	1459	20	RL	68.0	9717	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
1459	1460	20	RL	75.0	9937	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm

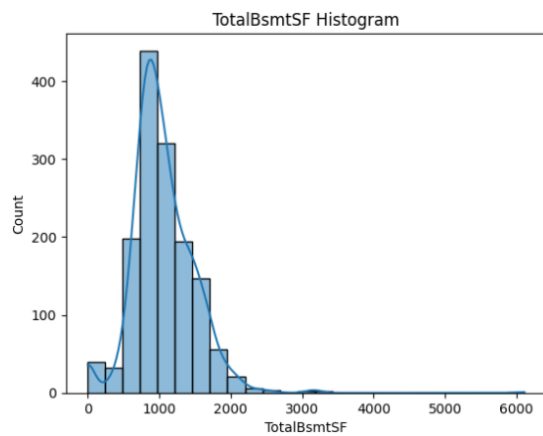
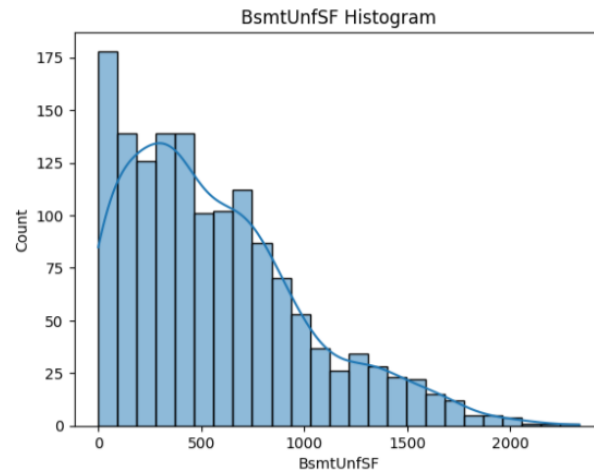
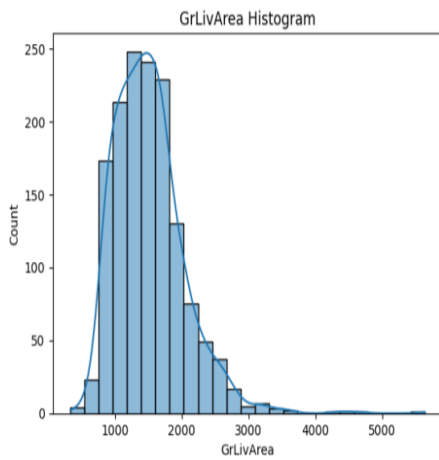
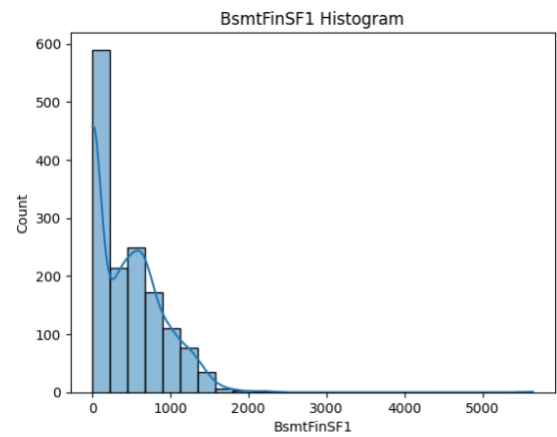
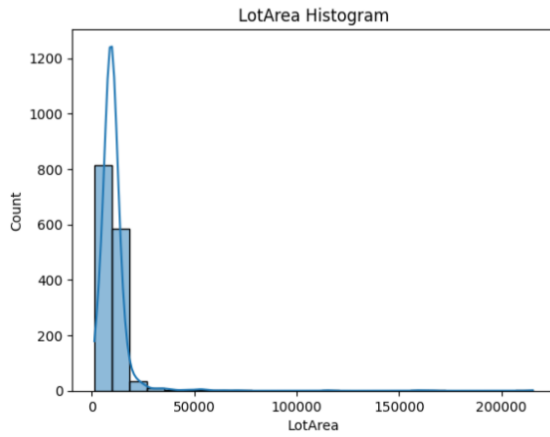
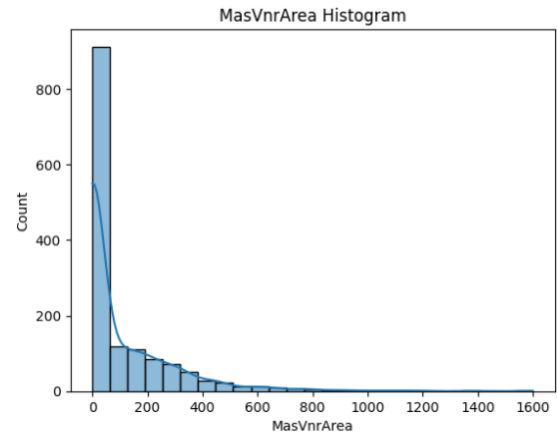
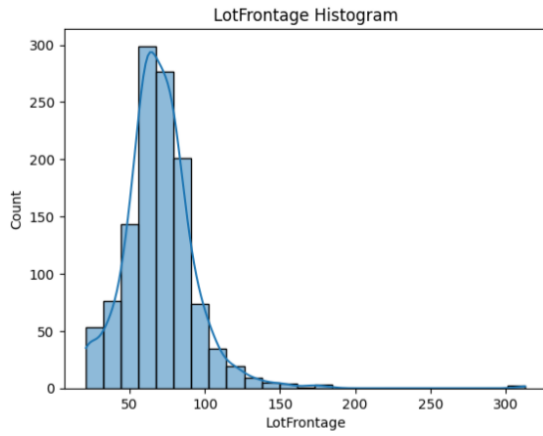
1460 rows × 81 columns

Here is descriptive statistics for some required variables:

	LotArea	MasVnrArea	YearBuilt	LotFrontage	BsmtFinSF1	BsmtUnfSF	GrLivArea	OverallQual	TotalBsmtSF
count	1460.000000	1452.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	10516.828082	103.685262	1971.267808	70.049958	443.639726	567.240411	1515.463699	6.099315	1057.429452
std	9981.264932	181.066207	30.202904	24.284752	456.098091	441.866955	525.480383	1.382997	438.705324
min	1300.000000	0.000000	1872.000000	21.000000	0.000000	0.000000	334.000000	1.000000	0.000000
25%	7553.500000	0.000000	1954.000000	59.000000	0.000000	223.000000	1129.500000	5.000000	795.750000
50%	9478.500000	0.000000	1973.000000	69.000000	383.500000	477.500000	1464.000000	6.000000	991.500000
75%	11601.500000	166.000000	2000.000000	80.000000	712.250000	808.000000	1776.750000	7.000000	1298.250000
max	215245.000000	1600.000000	2010.000000	313.000000	5644.000000	2336.000000	5642.000000	10.000000	6110.000000

Table: Descriptive Statistics of variables





B) TRAINING AND TESTING LOGS:

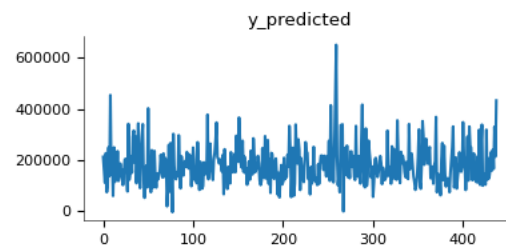
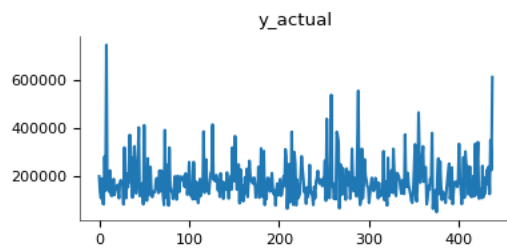
❖ Model 1: Lasso regression from sklearn

```
Lasso  
Lasso(alpha=0.1)
```

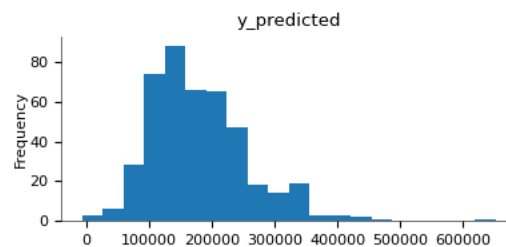
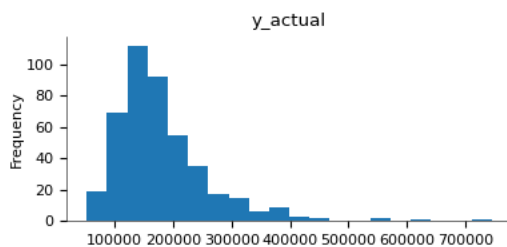
Performance metrics:

```
RMSE: 39664.98017356942  
MSE: 1573310652.169655  
MAE 21275.60100580882  
R2_score: 0.7682556141289717  
Adj_R2_score: 0.7149544053786352
```

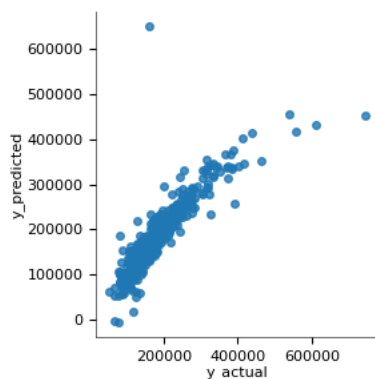
▶ Values



Distributions



2-d distributions



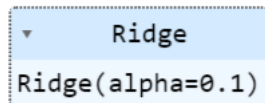
❖ Model 2: Lasso regression by using GD optimization method from scratch.

```
Predicted values [242621.32 165214.55 116505.93]
Real values      [200624 133000 110000]
Trained W        -7136.34
Trained b        13951.1
```

Performance metrics:

```
RMSE: 38753.90861124757
MSE: 1501865432.6489289
MAE 22713.022359345847
R2_score: 0.7787793009154425
Adj_R2_score: 0.7278985401259942
```

❖ Model 3: Ridge regression from sklearn

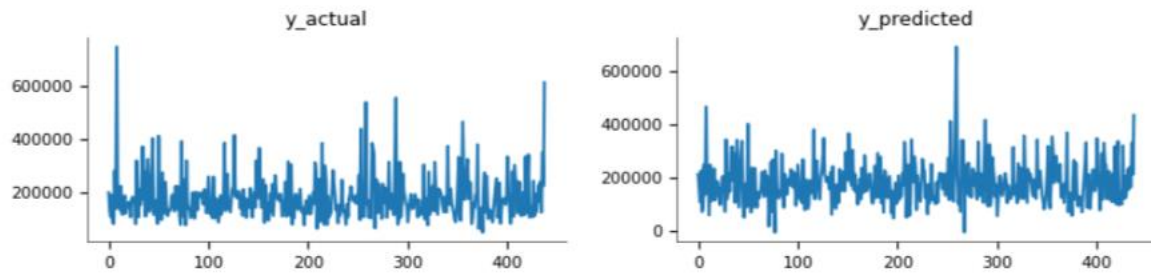


```
▼ Ridge
Ridge(alpha=0.1)
```

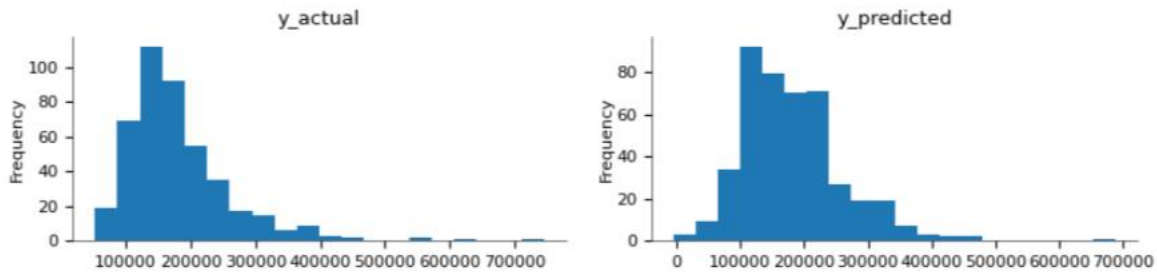
Performance metrics:

```
RMSE: 40545.00899618996
MSE: 1643897754.5011249
MAE 21375.130188648833
R2_score: 0.7578583256737874
Adj_R2_score: 0.7021657405787585
```

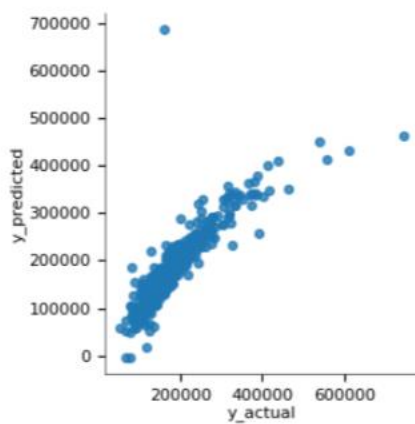
Values



Distributions



2-d distributions



❖ Model 4: Ridge regression by using GD optimization method from scratch.

```
Predicted values [242385.7 165005. 116568.24]
Real values      [200624 133000 110000]
Trained W        -7089.76
Trained b        14135.91
```

Performance metrics:

```
RMSE: 38770.826331612894
MSE: 1503176974.4360878
MAE 22694.02581557811
R2_score: 0.7785861143724097
Adj_R2_score: 0.7276609206780639
```

❖ Model 5: Elastic Net regression from sklearn

```
ElasticNet  
ElasticNet(alpha=0.1)
```

Performance metrics:

RMSE: 40213.89406674755

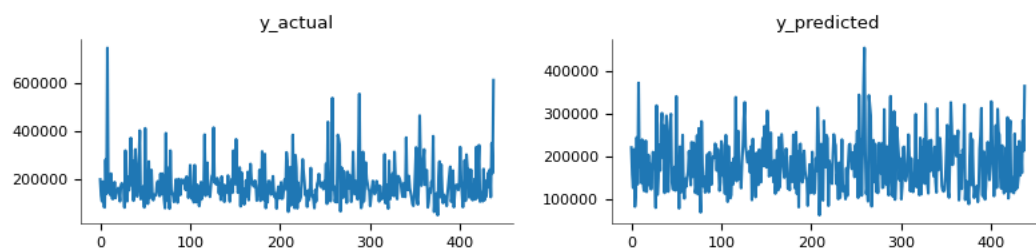
MSE: 1617157276.0115933

MAE 22837.715511988794

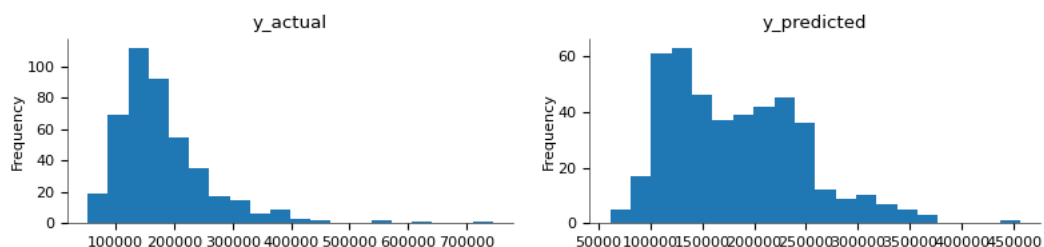
R2_score: 0.761797125526765

Adj_R2_score: 0.7070104643979209

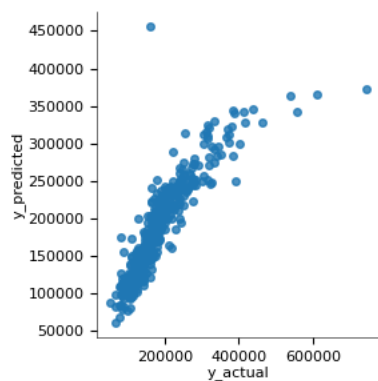
Values



Distributions



2-d distributions



❖ Model 6: ElasticNet regression by using GD optimization method from scratch.

```
Predicted values [242389.35 165004.69 116572.16]
Real values      [200624 133000 110000]
Trained W        -7087.53
Trained b        14139.92
```

Performance metrics:

```
RMSE: 40213.89406674755
MSE: 1617157276.0115933
MAE 22837.715511988794
R2_score: 0.761797125526765
Adj_R2_score: 0.7070104643979209
```

C) DISCUSSION AND COMPARISON

To predict the SalePrice of a house we implemented 6 models on the dataset. All the models worked well on the dataset. R2 value is a performance metrics we can use in comparing two models. R2 values give us the information about the proportion of variance in the dependent variable that can be explained by the independent variable. The higher the R-squared value, the better the model fits your data.

Model	R-squared value	Adjusted R-squared value
Lasso Regression from sklearn	0.7683	0.7149
Lasso Regression by using GD optimization method from scratch.	0.7788	0.7279
Ridge Regression from sklearn	0.7579	0.7022
Ridge Regression by using GD optimization method from scratch.	0.7786	0.7277
ElasticNet Regression from sklearn	0.7618	0.7070
ElasticNet Regression by using GD optimization method from scratch.	0.7618	0.7070

Table: R-squared and Adjusted R-squared values of different models

From the results table we can say that Lasso regression model using gradient descent optimization method from scratch worked well than other models like ridge, elasticnet. We can use this model in predicting saleprice of houses.

CONCLUSION:

The prediction of sale price of a house is done by using machine learning techniques. The house_price_prediction database was used for it. Many preprocessing steps like changing missing values to median value, normalizing the data using StandardScaler were used for it. We split data into two groups like train and test. We fitted 6 machine learning models named lasso, ridge, elasticnet regression model from Sklearn and these models by using gradient decent optimization method from scratch and evaluated them and we calculated performance metrics. All the models worked well on dataset but lasso regression model by using gradient decent optimization method from was better than all other models.

REFERENCES:

- ❖ Ayush Kumar Tiwari, Akshita Sharma, Aman Goyal, Pragya Tewari. “House Price Prediction Using Machine Learning” [Online]
- ❖ G.Naga Satish, Ch.V.Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu “House Price Prediction Using Machine Learning”. *IJITEE*, 2019.
- ❖ Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh - “A hybrid regression technique for house prices prediction” 2017, *IEEE*.
- ❖ Bharatiya, Dinesh, et al. “Stock market prediction using linear regression.” *Electronics, Communication, and Aerospace Technology (ICECA), 2017 International conference of. Vol. 2. IEEE, 2017.*
- ❖ CH. Raga Madhuri, G. Anuradha, M. Vani Pujitha -” House Price Prediction Using Regression Techniques: A Comparative Study” 2019 in (*ICSSS*), *IEEE*.
- ❖ <https://www.geeksforgeeks.org/implementation-of-elastic-net-regression-from-scratch/>
- ❖ Refcode1:
<https://memphis.instructure.com/courses/98483/files/folder/Ref.%20codes?preview=7774288>
- ❖ Refcode2:
<https://memphis.instructure.com/courses/98483/files/folder/Ref.%20codes?preview=7774291>