International Conference on Emerging Trends in Technology, Science, and Education (ICETTSE-2022)

CAR AND HOUSE PRICE PREDICTION USING LINEAR, LASSO AND RIDGE REGRESSION

Pooruvi Singh
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
pooruvivirendrasingh2015@gmail.com

Omkar Tendolkar
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
omkartendolkar10@gmail.com

Aditya Kini
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
adityakini7686@gmail.com

Suraj Maurya
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
surajmaurya3118@gmail.com

Neelam Phadnis
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
kulkarnineelam@gmail.com

***Abstract*—In this research we used Linear, lasso and Ridge regression for generating models for car and house price prediction, for car price prediction we did comparison between Linear and Lasso regression and for house price prediction we did comparison between Linear and Ridge regression the dataset for both car and house price prediction is taken from website Kaggle.**

Keywords*—Linear regression, Lasso regression, Ridge regression, car price prediction, house price prediction)

I. Introduction

Car and house are the most important possession a person can have people value driving a car because it confers prestige and allows them to exercise personal control and autonomy while a house shields us from the whims of nature and hazards. A home provides a sense of security and well-being, which is more than just a physical structure but a symbol of power, authority, and a slew of other things but there exists a lack of transparency and knowledge in determination of prices of house or values of used cars because of which people are not assured of the price that they are paying for buying a house or a used car also there are involvement of intermediaries in determination of prices which in turn includes their charges hence there is a need of system that can determine the values of housing properties and value of used cars . In this paper we have made a comparative study of Linear and Lasso regression for car price prediction model and a comparative study of Linear and Ridge regression for house price prediction model. The dataset used here is taken from website Kaggle.

II. Literature review

Various researches regarding car and house price prediction are done previously

Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya and Pitchayaki Boonpou [1] did a comparative study on various regression models using the dataset collected from a German ecommerce website they observed that gradient boosted regression trees gave the best performance in comparison of random forest regression and multiple linear regression with a MAE of 0.28.

Pattabiraman Venkatasubbu and Mukkesh Ganesh [2] did a comparative study on Lasso regression, Multiple regression and regression trees for predicting the prices of used cars Multiple regression performed well in comparison of others with an error rate of 3.468%.

Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic and Jasmin Kevric [3] made use of three machine learning techniques Artificial Neural Network, Support Vector Machine and Random Forest for predicting prices of used cars in Bosnia and Herzegovina and obtained an accuracy of 87.38%.

CH.Raga Madhuri, Anuradha G, and M.Vani Pujitha [4] Predicted house price with the help of Regression techniques like Multiple linear, Ridge, LASSO, Elastic Net, Gradient Boosting and Ada Boost Regression where gradient boosting algorithm gave high accuracy as compared to other algorithms

Ayush Varma, Abhijit Sarma, Sagar Doshi and Rohini Nair [5] proposed a house price prediction system by making a optimal use of Linear regression, Forest regression and Boosted regression with a further increase in accuracy by making use of Neural networks.

P. Durganjali and M. Vani Pujitha [6] Used different classification algorithms like logistic regression, decision tree, Naïve Bayes, random forest and Ada boost for house price prediction where Ada Boost gave the highest accuracy of 96%

III. Methodology

The work flow for the model generation of car and house price prediction can be given as

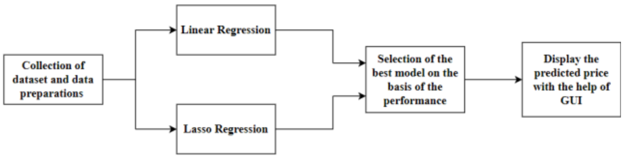

Collection of dataset and data preparations
Linear Regression
Lasso Regression
Selection of the best model on the basis of the performance
Display the predicted price with the help of GUI

Fig 1: Workflow for car price prediction

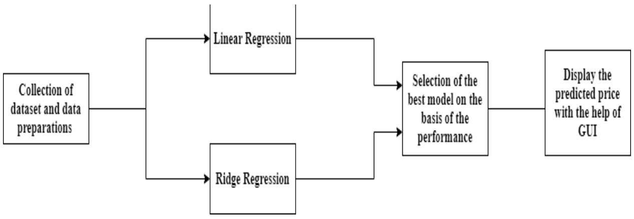

Collection of dataset and data preparations
Linear Regression
Ridge Regression
Selection of the best model on the basis of the performance
Display the predicted price with the help of GUI

Fig 2: Workflow for House price prediction

At the start of the process, we first collect the dataset and perform various data preprocessing steps to prepare our data after this stage the data is split up into training and test set where training data is used for model generation by using algorithms like linear and lasso regression for car price prediction and Linear and Ridge regression for House price prediction the algorithm that performs well is selected for prediction. Now in order to use this ML Model for prediction we develop a web application where users can predict the value of their used cars or housing property based on their choice, they would be redirected to the respective web page for prediction.

IV. IMPLEMENTATION

A. *Data understanding and preprocessing*

Dataset for both car and house price prediction are taken from website Kaggle the car dataset is uploaded by Neha Birla under open database license and contains information about the cars listed on www.cardekho.com and contains features like name, year, selling price, present price, Kms driven, fuel type, seller type, transmission, number of previous owner's car had. House dataset is of Mumbai area uploaded by Sameep Seth under public domain license it contains information of houses all over the Mumbai such as area, location, number of bedrooms and various amenities available but our research focuses on the prices of house in Kharghar, Mira Road East and Thane West and we have considered features like area, location, number of bedrooms and whether a house is new or resale.

Table I: Distribution of categorical data for car dataset

Attribute	**Count**
Fuel Type	
Petrol	239
Diesel	60
CNG	2
Seller Type	
Dealer	195
Individual	106
Transmission	
Automatic	261
Manual	40

Table II: Distribution of categorical data for house dataset

Attribute	**Count**
Number of bedrooms	
1 bedroom	439
2 bedrooms	593
3 bedrooms	271
Location	
Kharghar	512
Thane West	405
Mira Road East	386

The dataset may contain duplicate values or null values also we may drop some attributes that are not necessary in prediction process this data preparation is done with the help of python programming language for car dataset we drop the name feature as it is not important in prediction and for house dataset, we are considering area, location, and number of bedrooms.

Correlation Heat map is generated to know the statistical measure of linear relationship between variables or attributes of dataset, the value of correlation coefficient can range from -1 to 1 correlation heatmap can be drawn using python's seaborn library.

Seaborn is a Python module for creating statistical visualizations. It interacts well with panda's data structures and gives a high-level interface to matplotlib. The seaborn library's functions present a declarative, dataset-oriented API that makes it simple to convert data questions into images that can answer them [7].

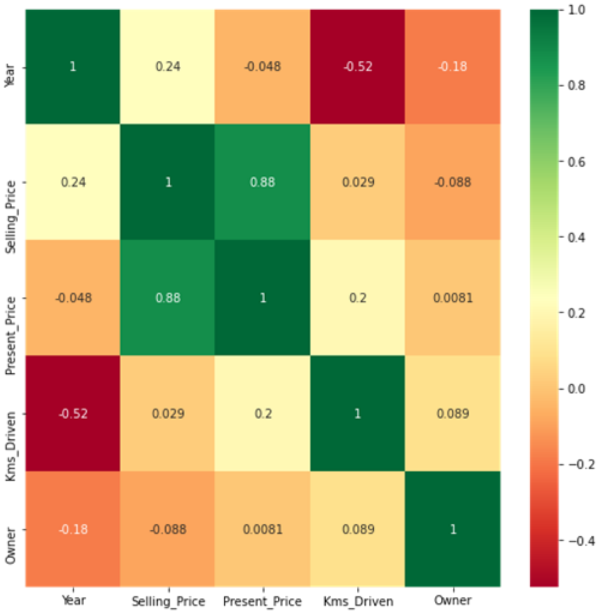


Fig 3: correlation heat map for car price prediction

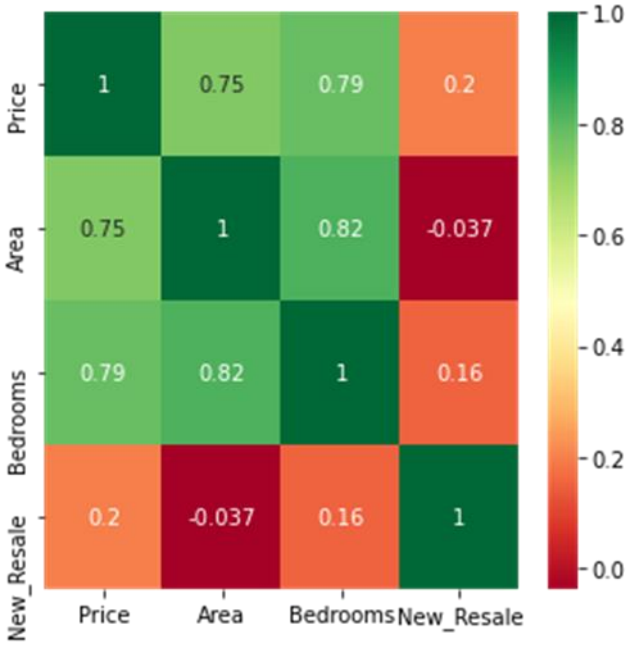


Fig 4: Correlation heat map for house price prediction

B. Study of Linaer and Lasso regression for Car price prediction

For the model generation of car price prediction, we have done a comparative study of linear regression and Lasso regression available in the SciKit learn a machine learning library.

Linear Regression

Linear regression is a type of supervised learning machine learning algorithm. It carries out a regression task. Based on independent variables, regression models a goal prediction value. It is mostly utilized in forecasting and determining the link between variables. Different regression models differ in terms of the type of relationship they evaluate between dependent and independent variables and the number of independent variables they employ.

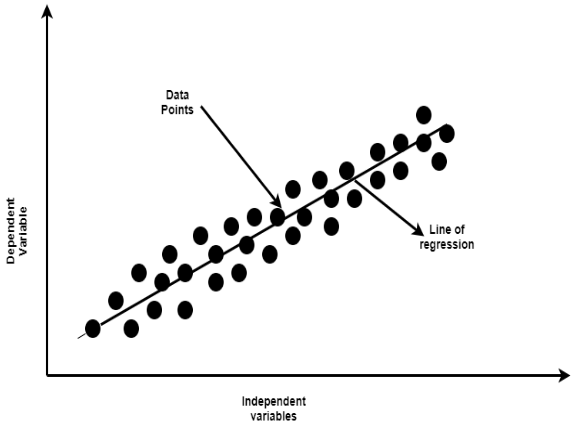


Fig 5: Linear regression

Lasso regression

Sometimes using standard regression methods to a set of candidate variables for generating a model tends to lead to overfitting in terms of the number of variables [8].

Least Absolute Shrinkage and Selection Operator is the abbreviation for LASSO It's a statistical formula for regularizing data models and selecting features.

Shrinkage is used in this model. Data values are shrunk towards a central point known as the mean in shrinkage. Simple and sparse models are provided by the lasso approach.

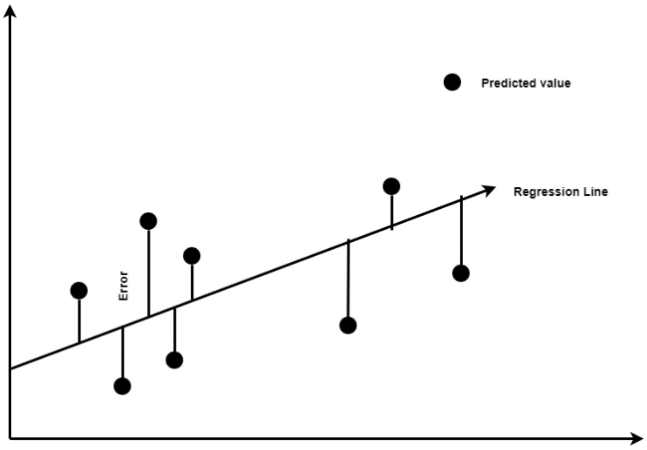


Fig 6: Lasso regression

For car price prediction the dataset was split up into 90 percent for training data and 10 percent for test data.

For Linear regression a R2 score of 0.879945 was obtained for training dataset and a R2 score of 0.836576 was obtained for test dataset. For Lasso regression a R2 score of 0.842785 was obtained for training dataset and a R2 score of 0.870916 was obtained for test dataset.

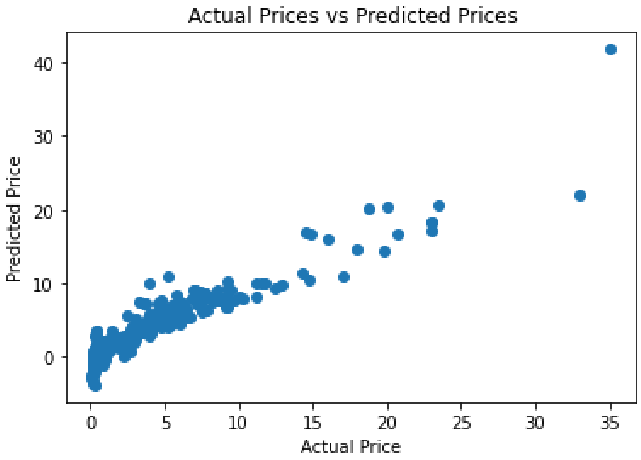


Fig 7: Actual vs Predicted price by linear regression for training dataset of car price prediction

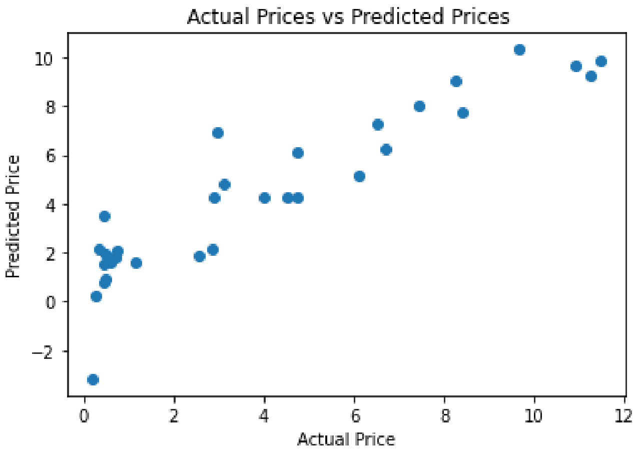


Fig 8: Actual vs Predicted price by Linear regression for test dataset of car price prediction

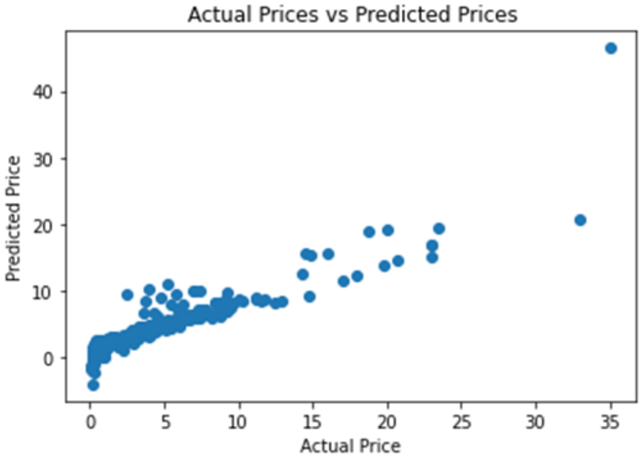


Fig 9: Actual vs Predicted price by Lasso regression for training dataset of car price prediction

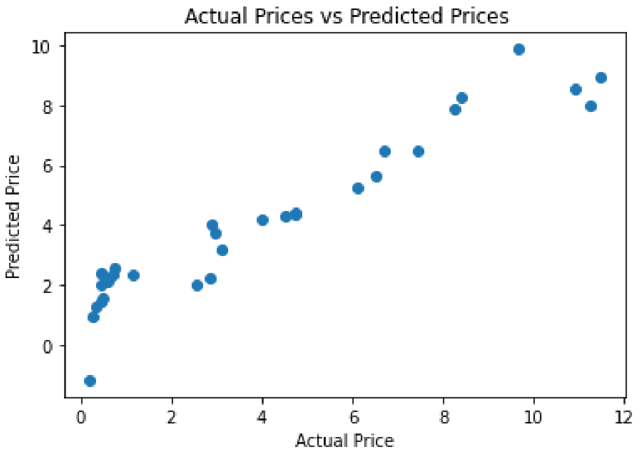


Fig 10: Actual Vs Predicted price by Lasso regression for test dataset of car price prediction

Table III: R2 score of linear and Lasso regression.

Algorithm	**R2 Score**
Linear regression	
Training dataset	0.879945
Test dataset	0.836576
Lasso regression	
Training dataset	0.842785
Test dataset	0.870916

C. *Study of Linaer and Ridge regression for house price prediction*

For the house price prediction model Linear and Ridge regression were studied here the dataset was divided into 70 percent for training data and 30 percent for test data

Ridge regression

Ridge regression is a model tuning technique that can be used to analyze data with multicollinearity. L2 regularization is achieved using this method. When there is a problem with multicollinearity, least-squares are unbiased, and variances are big, the projected values are far from the actual values. It reduces the size of the parameters. As a result, it's employed to avoid multicollinearity. By shrinking the coefficients, it minimizes the model's complexity.

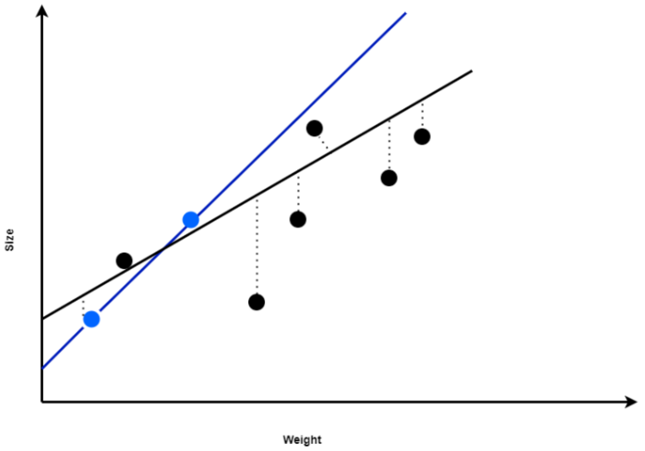


Fig 10: Ridge regression

For Linear regression a R2 score of 0.668408 was obtained for training dataset and a R2 score of 0.693642 was obtained for test dataset. For Ridge regression a R2 score of 0.668403 was obtained for training dataset and a R 2 score of 0.693933 was obtained for test dataset.

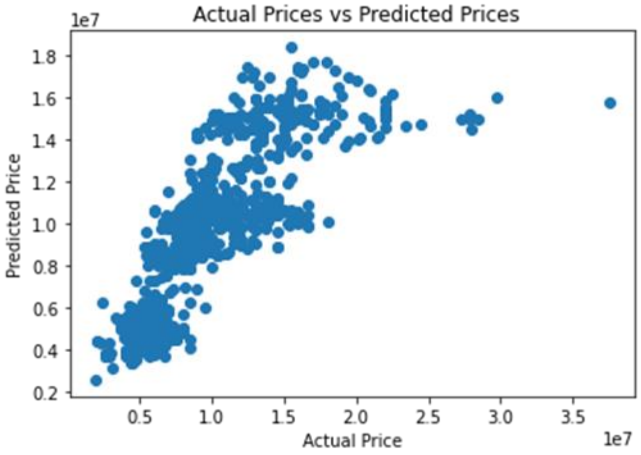


Fig 11: Actual Vs Predicted price by Linear regression for training dataset of house price prediction

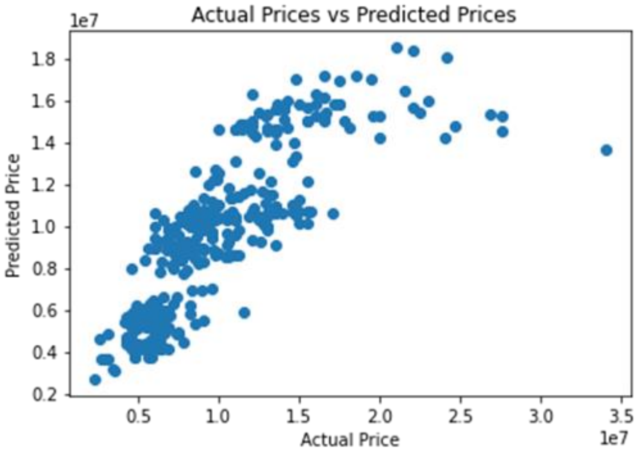


Fig 12: Actual Vs Predicted price by linear regression for test dataset of house price prediction

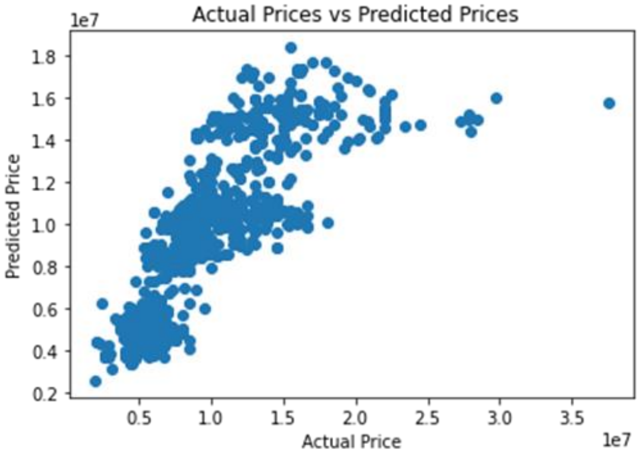


Fig 13: Actual Vs Predicted price by Ridge regression for training dataset of house price prediction

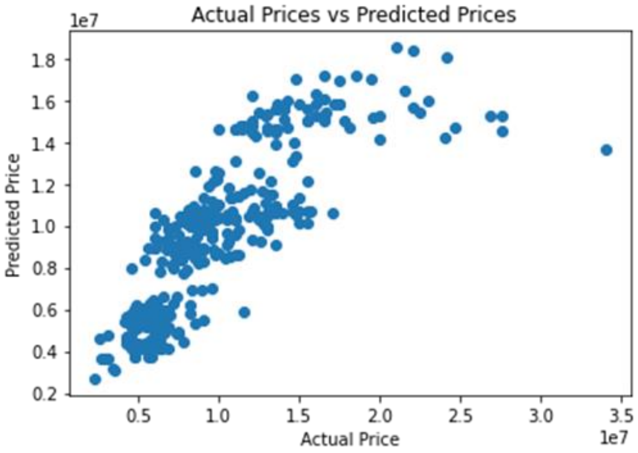


Fig 14: Actual Vs predicted price by Ridge regression for test dataset of house price prediction

Table IV: R2 score of Linear and Ridge regression.

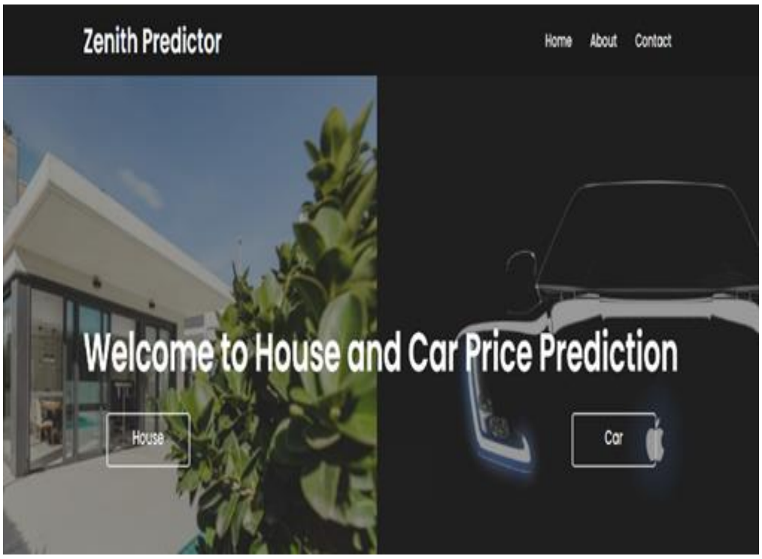
Algorithm	**R2 Score**
Linear regression	
Training dataset	0.668408
Test dataset	0.693642
Ridge regression	
Training dataset	0.668403
Test dataset	0.693933

D. Sample GUI for car and House price prediction

Both the models for car and house price prediction can be deployed with the help of a web application it consists of three web pages a home screen where user has choice to select whether they want to predict the car price or house price based on the user's choice user would be redirected to the desired web page for prediction

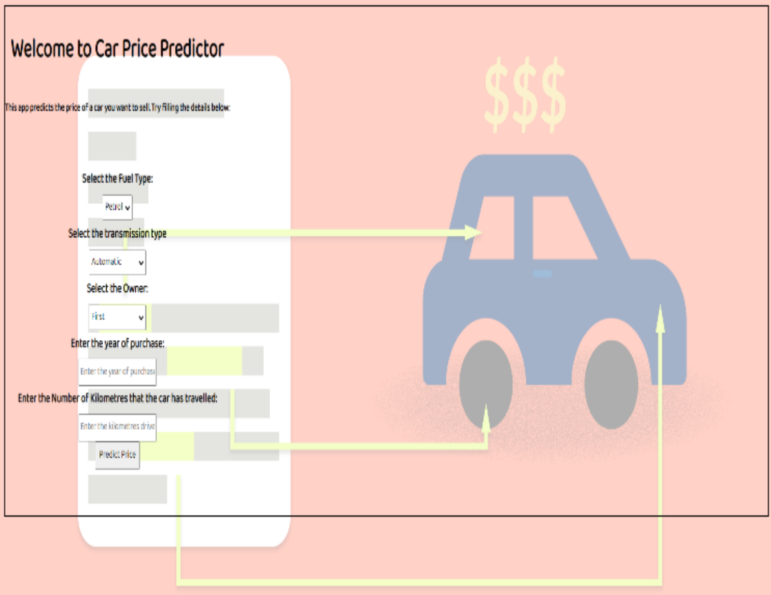
The web application is developed with the help of flask, JavaScript, HTML and CSS

Flask is a Python micro framework that provides the basic functionality of a web framework while also allowing more plug-ins to be added to expand the functionality and feature set. Flask is known as a Python micro framework because it keeps the basic functionality minimal while allowing for development flexibility [9].



Welcome to House and Car Price Prediction

Fig 13: Sample GUI of home page



Welcome to Car Price Predictor

This app predicts the price of a car you want to sell. Try filling the details below:

Select the Fuel Type: Petrol

Select the transmission type Automatic

Select the Owner: First

Enter the year of purchase: Enter the year of purchase

Enter the Number of Kilometres that the car has travelled: Enter the kilometres drive

Fig 14: sample GUI of car price predictor

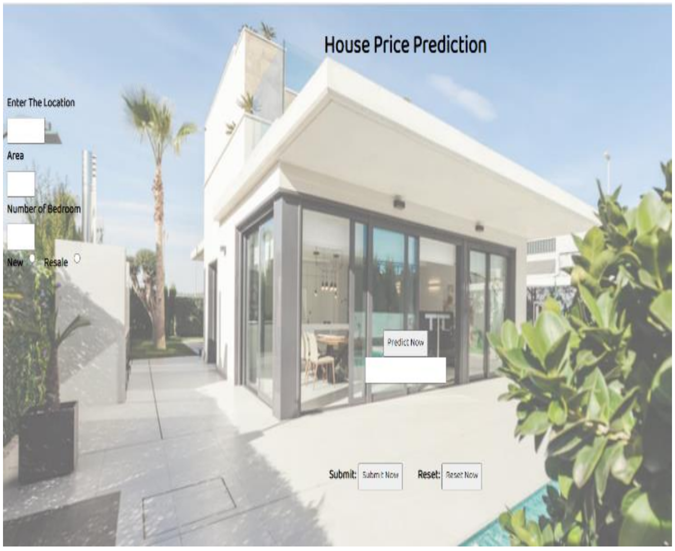


Fig 15: Sample GUI of house price prediction

V. CONCLUSION

The research focused on comparative study between Linear and Lasso regression for predicting prices of used cars and Linear and Ridge regression for predicting prices of Hosing property, the dataset for both the models were taken from website Kaggle the models can be deployed with help of a web application which makes it easier for users to predict the price of used cars or price of housing properties.

The R2 score of Linear regression on test dataset of car price prediction was 0.836576 and R2 score of Lasso regression on test dataset of car price prediction was 0.870916. The R2 score of Linear regression on test dataset of House price prediction was 0.693642 and R2 score of Ridge regression on test dataset of House price prediction was 0.693933.

VI. REFRENCES

[1] Nitis Monburinon, Prajak Chertchom, Suwat Rungpheung, Sabir Buya and Pitchayakit Boonpou, "Prediction of prices for used car by using regression models," In 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok Thailand, 2018, pp. 115-119.

[2] Pattabiraman Venkatasubbu and Mukkesh Ganesh, "Used cars price prediction using supervised learning techniques," *International Journal of Engineering and advanced technology(IJEAT),* vol.9, Issue 1S3, pp.216-223, Dec 2019.

[3] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic and Jasmin Kevric, "Car price prediction using machine learning techniques," TEM Journal, vol.8, Issue 1, pp.113-118, Feb 2019.

[4] CH.Raga Madhuri, Anuradha G and M.Vani Pujitha, "House price prediction using regression techniques: A comparitive study," In IEEE 6th International Conference on smart structures and systems(ICSSS), Chennai India, 2019.

[5] Ayush Varma, Abhijit Sarma, Sagar Doshi and Rohini Nair, "House price prediction using machine learning and neural networks," In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore India, 2018, pp.1936-1939.

[6] P.Durganjali, and M.Vani Pujitha, "House resale price prediction using classification algorithms," In IEEE 6th International Conference on smart structures and systems (ICSSS), Chennai India, 2019.

[7] Michael L. Waskom, "seaborn: statistical data visualization," *The Journal of Open Source Software",* April 2021. Available : https://joss.theoj.org/

[8] J.Rastam and J.A. Cook, "LASSO regression*," British Journal of Surgery (BJS),* vol.105, Issue 10, pp. 1338-1348, sept 2018.

[9] Fankar Armash Aslam, Hawa Nabeel Mohammed, Jumnal Musab Mohd. Munir, Murade Aaraf Gulamgaus and P.S. Lokhande, "Efficient way of web development using python and flask," *International Journal of Advanced Resaech in Computer Science*, vol.6, April 2015.