CAR AND HOUSE PRICE PREDICTION WEB APPLICATION

(AT21TECSM50115)

A **Mini-Project Report** Submitted in partial fulfilment of the requirements of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER ENGINEERING

BY

Pooruvi Singh (Roll No 56) (Leader)

Aditya Kini (Roll No 27)

Suraj Maurya (Roll No 32)

Omkar Tendolkar (Roll N0 59)

Supervisor

Mrs.Neelam Phadnis

DEPARTMENT OF COMPUTER ENGINEERING

SHREE L. R. TIWARI COLLEGE OF ENGINEERING

KANAKIA PARK, MIRA ROAD (E), THANE -401 107, MAHARASHTRA.

University of Mumbai

(AY 2021-22)

Declaration by the Candidate

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date:

(Pooruvi Singh)
Roll No.: 56 Exam. Seat No.: 22MS16057

(Aditya Kini)
Roll No.:27 Exam. Seat No.: 22MS16028

(Suraj Maurya)
Roll No.: 32 Exam. Seat No.: 22MS16033

(Omkar Tendolkar)
Roll No.: 59 Exam. Seat No.: 22MS16060

Shree Rahul Education Society's (Regd.)
SHREE L. R. TIWARI COLLEGE OF ENGINEERING
Kanakia Park, Near Commissioner's Bungalow, Mira Road (East), Thane 401107, Maharashtra
(Approved by AICTE, Govt. of Maharashtra & Affiliated to University of Mumbai)
NAAC Accredited | ISO 9001:2015 Certified
Tel. No.: 022-28120144 / 022-28120145 | Email: slrtce@rahuleducation.com | Website: www.slrtce.in
DEPARTMENT OF COMPUTER ENGINEERING

CSM501 Mini-Project – 2A

Fifth Semester, 2021-2022 (Odd Semester)

CERTIFICATE

This is to certify that the **Mini-Project** entitled **"CAR AND HOUSE PRICE PREDICTION WEB APPLICATION "** is a bonafide work of

Pooruvi Singh (Roll No. 56)

Aditya Kini (Roll No. 27)

Suraj Maurya(Roll No. 32)

Omkar Tendolkar (Roll No. 59)

submitted to the University of Mumbai in partial fulfilment of the requirement of course name "**Mini-Project 2A**" having course code **CSM501** for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering"**.

Signature of Supervisor/Guide

Name: Mrs. Neelam Phadnis

Date: _____

Signature of the H.O.D.

Name: Mrs. Neelam Phadnis

Date: _____

Signature of the Principal

Name: Dr. Deven Shah

Date: _____

Shree Rahul Education Society's (Regd.)
SHREE L. R. TIWARI COLLEGE OF ENGINEERING
Kanakia Park, Near Commissioner's Bungalow, Mira Road (East), Thane 401107, Maharashtra
(Approved by AICTE, Govt. of Maharashtra & Affiliated to University of Mumbai)
NAAC Accredited | ISO 9001:2015 Certified
Tel. No.: 022-28120144 / 022-28120145 | Email: slrtce@rahuleducation.com | Website: www.slrtce.in
DEPARTMENT OF COMPUTER ENGINEERING

CSM501 Mini-Project – 2A

Fifth Semester, 2021-2022 (Odd Semester)

Mini-Project Report Approval

This Mini-project report entitled "**CAR AND HOUSE PRICE PREDICTION WEB APPLICATION ”** by

Pooruvi Singh (Roll No. 56)

Aditya Kini (Roll No. 27)

Suraj Maurya (Roll No. 32)

Omkar Tendolkar (Roll No. 59)

is belonging to the course name "**Mini-Project – 2A”** having course code **CSM501** submitted as a Term work and approved for the degree of Batchelor of Engineering in Computer Engineering.

Examiners

1. Name: _____(Internal)

   Signature: _____

2. Name: _____(External)

   Signature: _____

Date:

Place:

Acknowledgement

I Miss Pooruvi Virendra Singh, Leader of team Zenith along with my other team members would like to express our gratitude to our project guide Mrs.Neelam Phadnis for continuously guiding us through the course of the project . We are really thankful for your patient guidance,enthusiastic encouragement that has motivated us throughout the process.

Pooruvi Singh
Roll No.: 56 Exam. Seat No.: 22MS16057
Aditya Kini
Roll No.: 27 Exam. Seat No.: 22MS16028
Suraj Maurya
Roll No.: 32 Exam. Seat No.: 22MS16033
Omkar Tendolkar
Roll No.: 59 Exam. Seat No.: 22MS16060

Abstract

The used car market is an ever-rising industry with the emergence of the online portals such as cars24, Quikr and many others has facilitated the need for both the customer and seller to be informed about the trends and patterns that determines the value of the used car in the market. The price of a new car is fixed by the manufacturer so the customers are assured of the money they invest but for used cars there is a need for a system that predicts the worthiness of a car using a variety of features. Similarly Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation hence to even tackle this there is a requirement of a predictive model which predicts the value of housing property based on various factors. In this project we propose a web application that integrates two machine learning models that will predict the resale value of used cars and value of housing properties and uses three regression algorithms: linear regression, ridge regression, lasso regression.

Organization of the Report

1) Introduction

An overview of the system, problem statement, objectives, importance, scope.

2) Literature review

A glimpse over existing system, identifying problems with previous system, discussing limitation of previous system.

3) Proposed System

An introduction to the proposed system, architecture and framework, Details of algorithms used.

4) Requirement Analysis

Different software and hardware requirements and UML diagrams

5) Implementation and Results

Implementation of code and final output

6) Testing

Various test cases of project

7) Conclusion and Future work

Final conclusion of project regrading accuracy of models and future scope for project

8) Outcomes

Attainment of each member in various course outcomes

9) References

The various research papers and scholarly articles referenced.

Table of Contents

CAR AND HOUSE PRICE PREDICTION WEB APPLICATION	i
Declaration by the Candidate	ii
Mini Project Report Approval	iii
Acknowledgement	iv
Abstract	v
Table of contents	vi
List of figures	vii
List of tables	viii
List of Abbreviations	ix
1 Introduction	11
1.1 Introduction	11
1.2 Background and Motivation	11
1.3 Problem statement	12
1.4 Project Objectives	12
1.5 Project importance	12
1.6 Scope of project work	13
2 Literature Review	14
2.1 Survey of Existing system	14
2.2 Problems with present system	15
3 Proposed system	16
3.1 Introduction	16
3.2 Architecture	16
3.3 Algorithm	18
4 Requirement analysis	20
4.1 Hardware and software requirements	20
4.2 Design details	23
5 Implementation and results	26
6 Testing	37
7 Conclusion and future work	39

8 Outcomes	39
9 References	40

LIST OF FIGURES

1.6 Scope	13
3.2 a Architecture of house price prediction	16
3.2 b Architecture of car price prediction	17
3.3.a Linear	18
3.3.b Ridge	19
3.3.c Lasso	19
4.2.1 Architecture diagram	20
4.2.2 Data flow diagram level 0	23
4.2.2 Data flow diagram level 1	23
4.2.3 Use case diagram	24
4.2.4 Sequence diagram	24
4.2.5 Activity diagram	25
5.1 Home Page	25
5.2 Car price page	26
5.3 House price page	27

List of Tables

Details of hardware and software	20
Outcomes	40

List of Abbreviations

SK-learn	SciKit Learn
AWS	Amazon Web Services
ML	Machine learning
JS	Java Script
HTML	Hyper Text Markup language
CSS	Cascading Style Sheets
OS	Operating Systems
AI	Artificial Intelligence

1 Introduction

1.1 Introduction

Car and House price prediction web application is a system that integrates two machine learning models that will predict the value of used cars on the basis of factors like kilometers driven, fuel type, company, model name etc. and the value of housing property on the basis of factors like area, number of bedrooms, amenities available, location etc. A machine learning model is a program which is trained to recognize certain types of pattern, which is trained over a training data set which contains number of records or rows and an algorithm to generate model and accuracy of the model is determined by a test dataset.

Here we will implement and evaluate the performance of various machine learning algorithms like Linear Regression, Ridge Regression, Lasso Regression.

1.2 Background and Motivation

Considering the demand for private cars all around the world, the demand of the second-hand car market has been rising and creating a chance in business for both buyer and seller. In several countries, buying a used car is the best choice for a customer because its price is reasonable and affordable for the buyer. After a few years of using them, it may get a profit from reselling again. However, various factors influence the price of a used car such as how old those vehicles are and the condition in the current scenario of them. Normally, the price of used cars in the market is not constant Hence many a times intermediaries are involved in the buying and selling process which may determine prices that are not worthy paying

Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation.

Hence our motivation is

a) To Cut the intermediary cost

b) Bring transparency to consumers

c) To regulate the reselling Item system

d) To determine whether the car is worth the posted price.

1.3 Problem statement

The process of Resale of Real estate properties and used automobiles are the least transparent with the involvement of an intermediary and the resale cost is not evaluated properly by considering all the factors which leads to ambiguity. Hence we propose a Machine learning model that will help to determine the resale value of car on the basis of factors like Kms driven, manufacturer, engine type etc. and the value of housing properties on the basis of factors like area, number of bedrooms, location etc.

1.4 Project Objectives

a) To develop a web application that helps users to predict the value of the used cars and housing properties.

b)To help car dealers better understand what makes a car desirable , the important feature in order to provide better services.

c) To predict the efficient house pricing for real estate customers by considering a number of factors.

d) To develop the frontend an interactive client-side interface with the help of technologies like HTML, CSS and JavaScript so that user can see and interact directly

e)To develop the backend it will help to deploy and integrate our machine learning model with the help of technology like Flask.

1.5 Project Importance

The project holds importance in many fields

Social:

The Model will Eliminate the need of intermediaries determining the price and people will be assured of the price they are paying or getting for their property or cars.

Commercial:

The model can be used by various organizations and ventures that provide services like buying and selling to customers in a whole new revolutionized way.

Industrial:

This will help automobile industries to manipulate the design of the cars, the business strategy etc. to meet certain price levels. It will help budding automobile industries to enter into the market and understand various factors affecting the automobile price.

1.6 Scope of Project Work

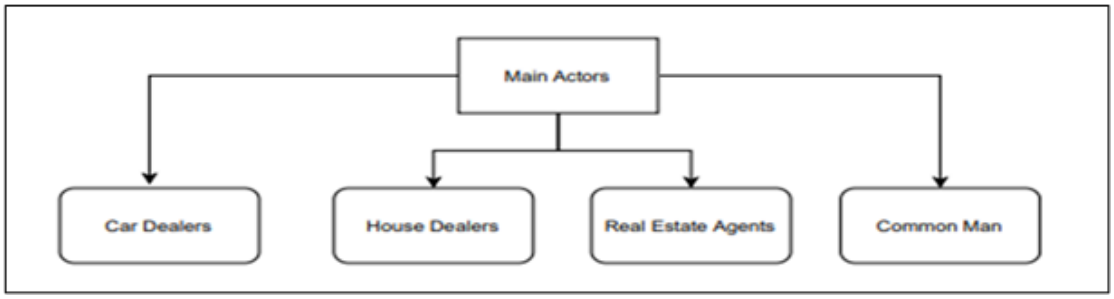


Fig 1.6 Scope

Main cases associated

Real estate agents: Evaluation of their properties

Car dealers: Evaluation of their cars

Common man: Evaluation of their car and properties

The Client

The proposed system will be able to predict used cars and house market value can help both buyers and sellers.

Used car sellers (dealers)

They are one of the biggest target groups that can be interested in results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and offer a better service.

Individuals

There are lots of individuals who are interested in the used car market at some point in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less than its market value.

2 Literature Review

2.1 Survey of Existing System

The existing system involves the intervention of intermediaries in buying and selling process moreover there is no transparency as to what factors were considered in determining the resale value of a car or value of housing property

Several related works have been done previously on the subject of used car price prediction and House price prediction .

Pudaruth predicted the price of used cars in Mauritius using multiple linear regression, k-nearest neighbors, naive Bayes and decision trees. Although their results were not good for prediction due to a less number of car observations. Pudaruth concluded in his paper that the decision tree and naive Bayes are unable to use for variables with a continuous value [1].

Noor and Jan used multiple linear regression to predict vehicle car price. They performed variable selection technique to find the most influencing variables then eliminate the rest. The data contain only selected variables that used to form the linear regression model. The result was impressive with R-square = 98% [2].

Peerun et al did research to evaluate the performance of the neural network in used car price prediction. The predicted value, however, is not very close to the actual price, especially on cars with a higher price.They concluded that support vector machine regression slightly outperform neural network and linear regression in predicting used car prices[3].

Sun et al proposed the application of an online used car price evaluation model using the optimized BP neural network algorithm. They introduced a new optimization method called Like Block-Monte Carlo Method (LB-MCM) to optimize hidden neurons. The result showed that the optimized model yielded higher accuracy when it compared to the non-optimized model. Based on the previous related works, we realized that none of them had implemented gradient boosting

techniques in the prediction of used car price yet. Thus, we decided to build a used car price evaluation model using gradient boosted regression trees [4].

Sifei Lu et.al, introduced a hybrid model for the regression of Lasso and Gradient to predict the price of the individual home. This approach has recently been used as the key kernel for the Kaggle Challenge "House prices: Advanced techniques for regression" [5].

Muhammad Fahmi Mukhlishin et.al, uses several methods to predict the value of land and house. This paper compares Fuzzy logic, Artificial Neural Network, and K-Nearest Neighbor to find the most appropriate method to determine the sellers ' price [6].

Atharva choogle et.al, House price forecasting has been introduced using data mining techniques. It provides a description of the prediction markets and also the current markets that help to make useful predictions in understanding the market. It is therefore necessary to predict the efficient pricing of real estate customers for their budgets and priorities [7].

2.2 Problems with Present System

Most of the existing prediction systems makes use of KNN and Naive bayes but following are some issues related to them

a) Accuracy depends on the quality of the data

b) With large data, the prediction stage might be slow

c) Sensitive to the scale of the data and irrelevant features

d) Require high memory – need to store all of the training data

e) Given that it stores all of the training, it can be computationally expensive

f) Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases.

g) Naive Bayes algorithm faces the 'zero-frequency problem' where it assigns zero probability to a categorical variable whose category in the test data set wasn't available in the training dataset. It would be best if you used a smoothing technique to overcome this issue.

h) Its estimations can be wrong in some cases, so you shouldn't take its probability outputs very seriously.

3 Proposed System

3.1 Introduction

The web application integrates two machine learning models one for car price prediction and another for house price prediction the algorithms that will be used to generate thee models are linear regression, lasso regression and ridge regression the dataset for house and car price prediction are taken from website called Kaagle that contains more than 3000 tuples scikit learn library is used that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology like NumPy, pandas, and Matplotlib.It provides functionality for Regression, including Linear and Logistic Regression. Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning.

3.2 Architecture

Proposed system for house price prediction

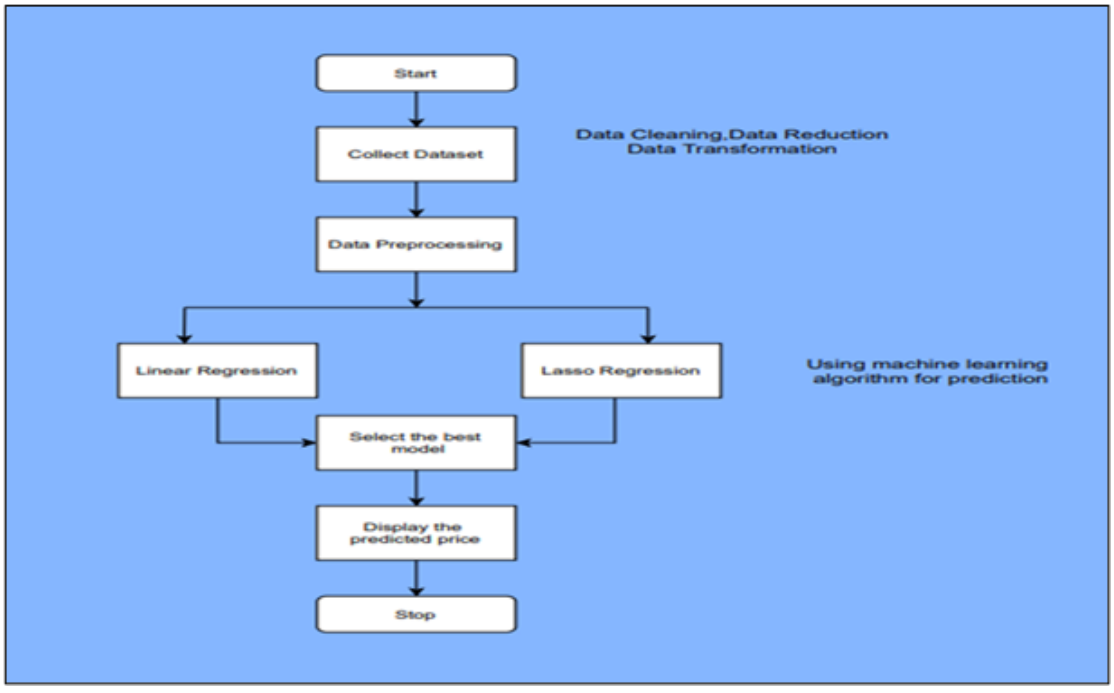


Fig 3.2 a Architecture of House price prediction system

Proposed system for car price prediction

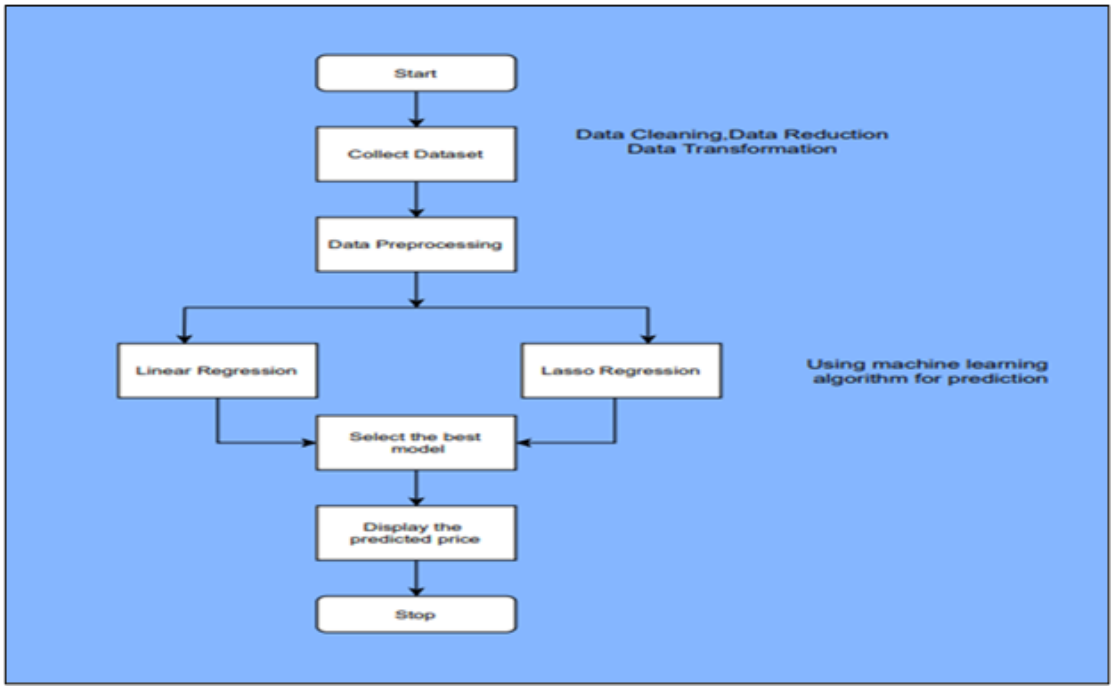


Fig 3.2 b Architecture of Car price prediction system

The process starts by collecting the dataset. The next step is to do Data Preprocessing which includes Data cleaning, Data reduction, Data Transformation. Then, using various machine learning algorithms we will predict the price. The algorithms involve Linear Regression, Ridge Regression and lasso regression. The best model which predicts the most accurate price is selected. After selection of the best model the predicted price is displayed to the user according to the user's inputs.

Data cleaning: Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.

Data reduction: Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form.

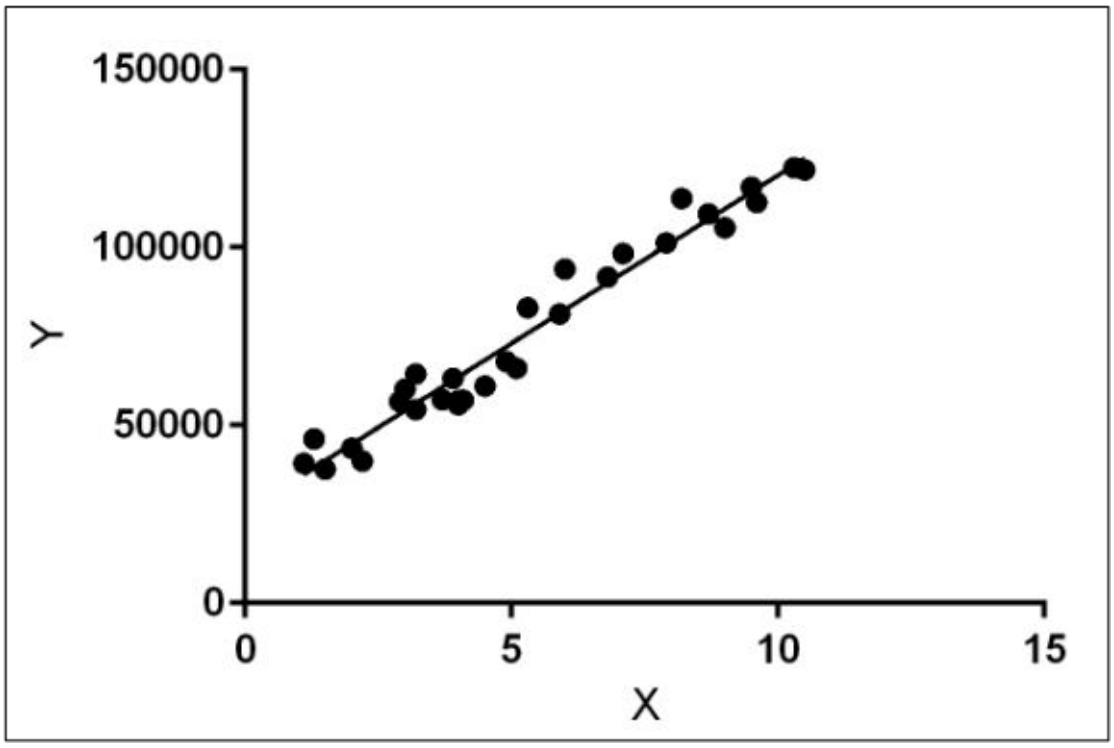
The purpose of data reduction can be two-fold: reduce the number of data records by eliminating invalid data or produce summary data and statistics at different aggregation levels for various applications.

Data Transformation: Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system.

3.3 Algorithm

Linear regression

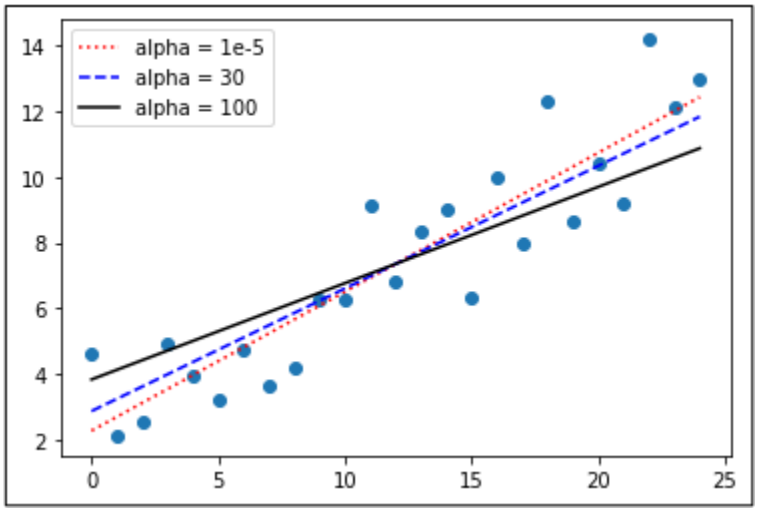
Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering and the number of independent variables being used.



3.3 a Linear regression

Ridge regression

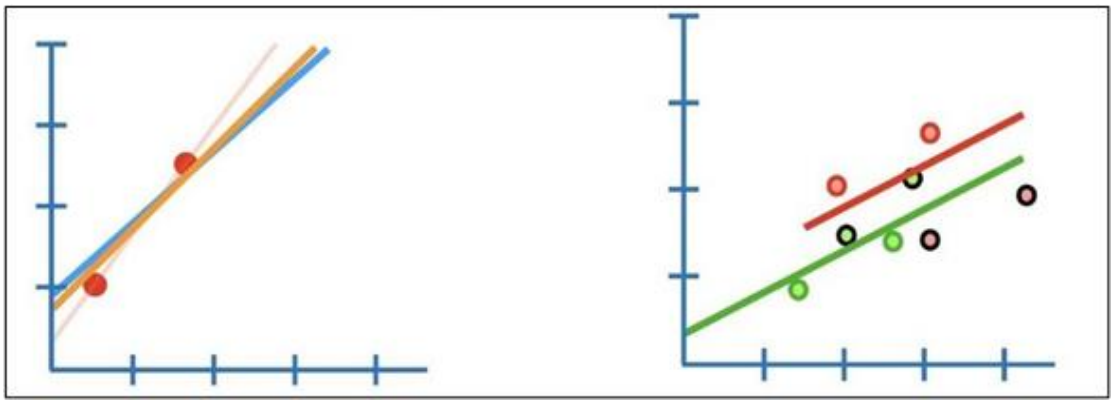
Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.



3.3 b Ridge regression

Lasso regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.



3.3 c Lasso regression

4 Requirement Analysis

4.1 Hardware and Software Details

HTML	The HyperText Markup Language, or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets and scripting languages such as JavaScript.
CSS	Cascading Style Sheets is a style sheet language used for describing the presentation of a document written in a markup language such as HTML. CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript.

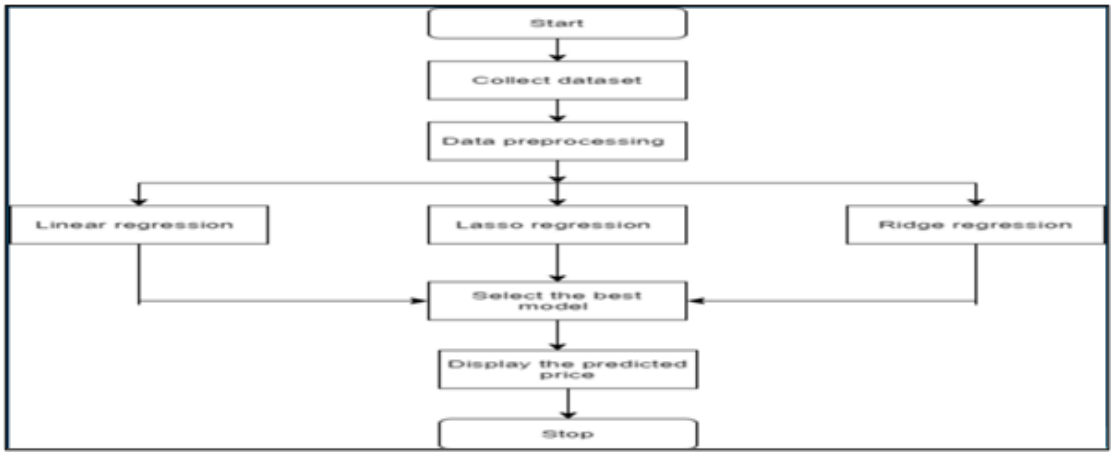
	Bootstrap is a free and open-source CSS framework directed at responsive, mobile-first front-end web development. It contains CSS- and JavaScript-based design templates for typography, forms, buttons, navigation, and other interface components.
	JavaScript, often abbreviated as JS, is a programming language that conforms to the ECMAScript specification. JavaScript is high-level, often just-in-time compiled, and multi-paradigm. It has curly-bracket syntax, dynamic typing, prototype-based object-orientation, and first-class functions.
	Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

	The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.
	Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.
	Heroku is a cloud platform as a service supporting several programming languages. One of the first cloud platforms, Heroku has been in development since June 2007, when it supported only the Ruby programming language, but now supports Java, Node.js, Scala, Clojure, Python, PHP, and Go.

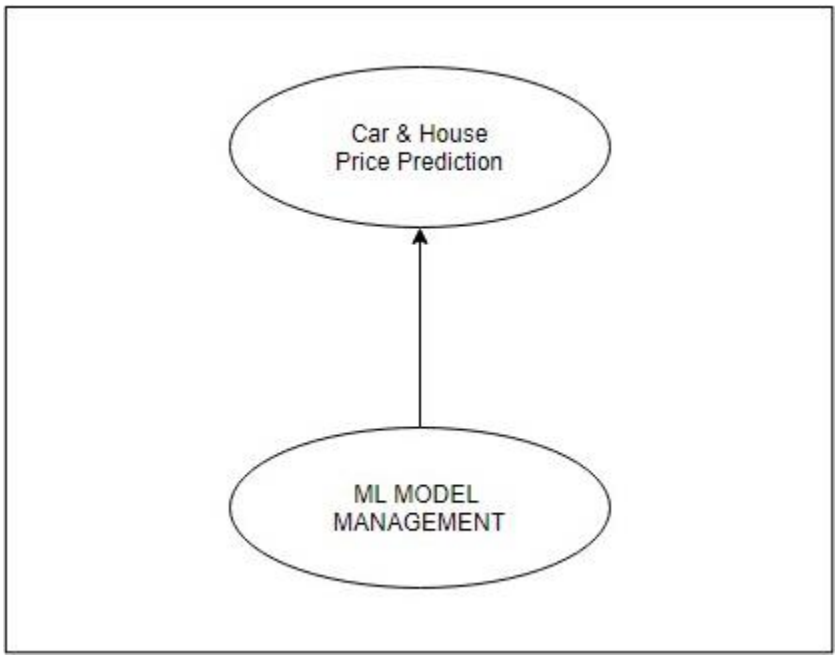
4.2 Design Details

Following are the various diagrams related to design details

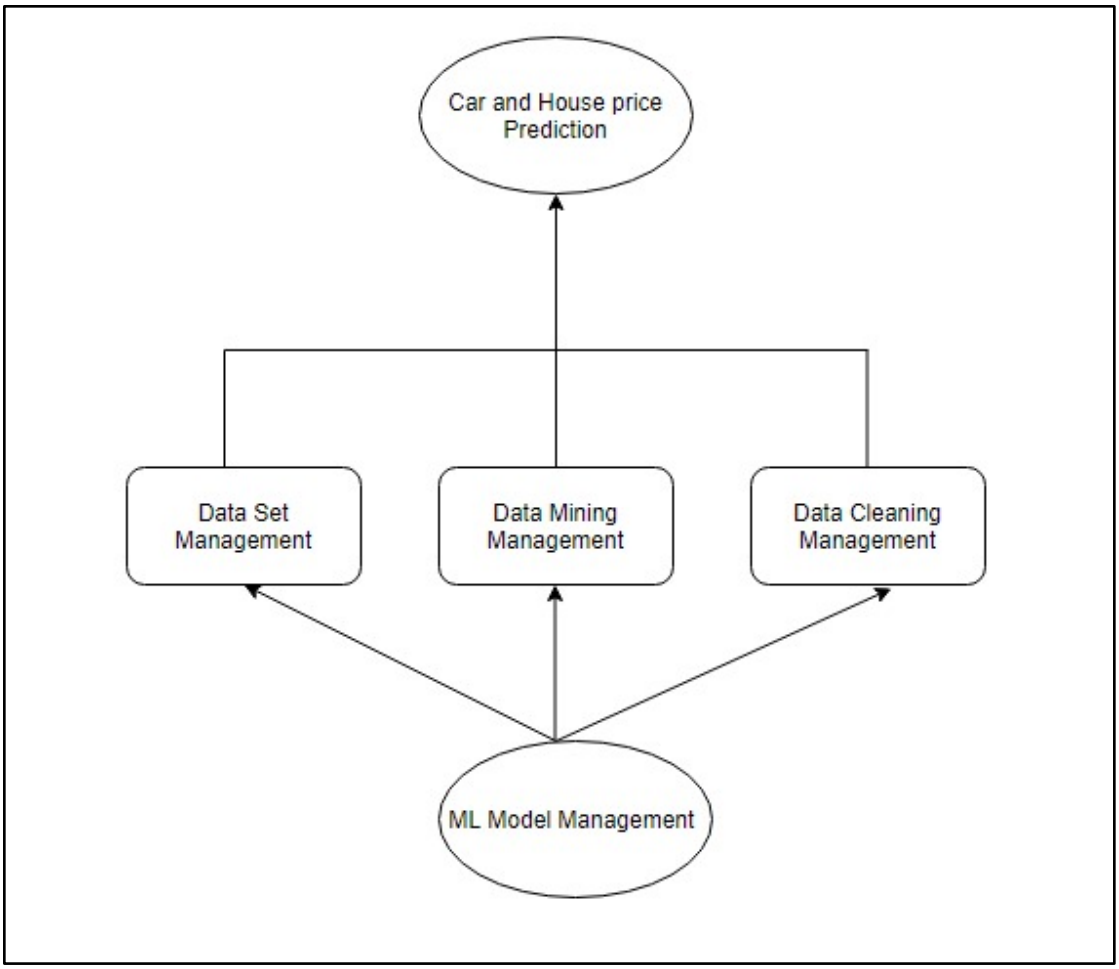
4.2.1 Architecture diagram



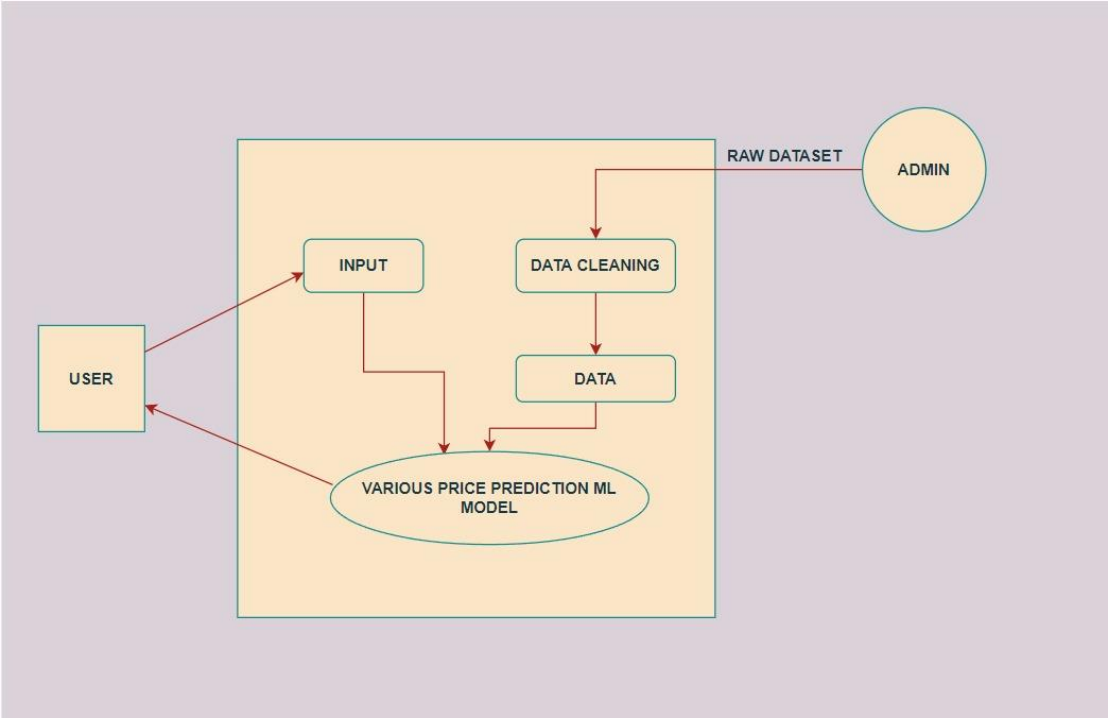
4.2.2 Data flow diagram level 0



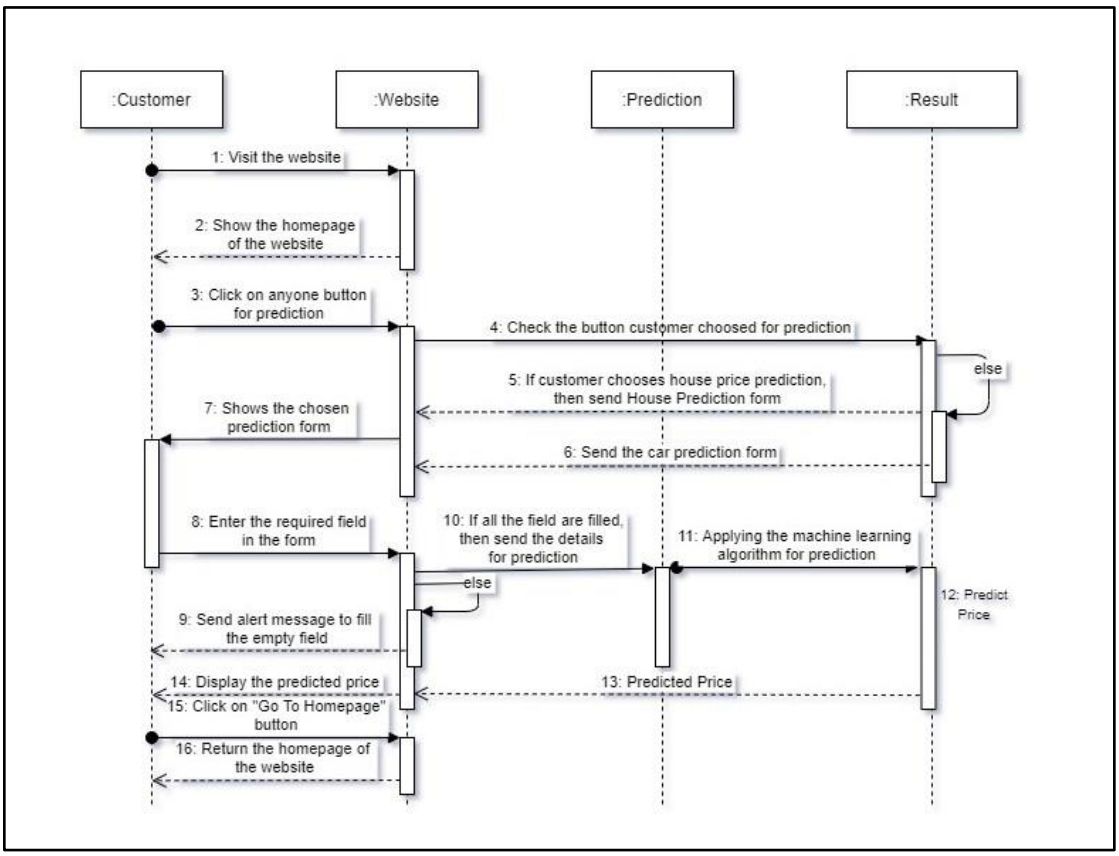
4.2.3 Data flow diagram level 1



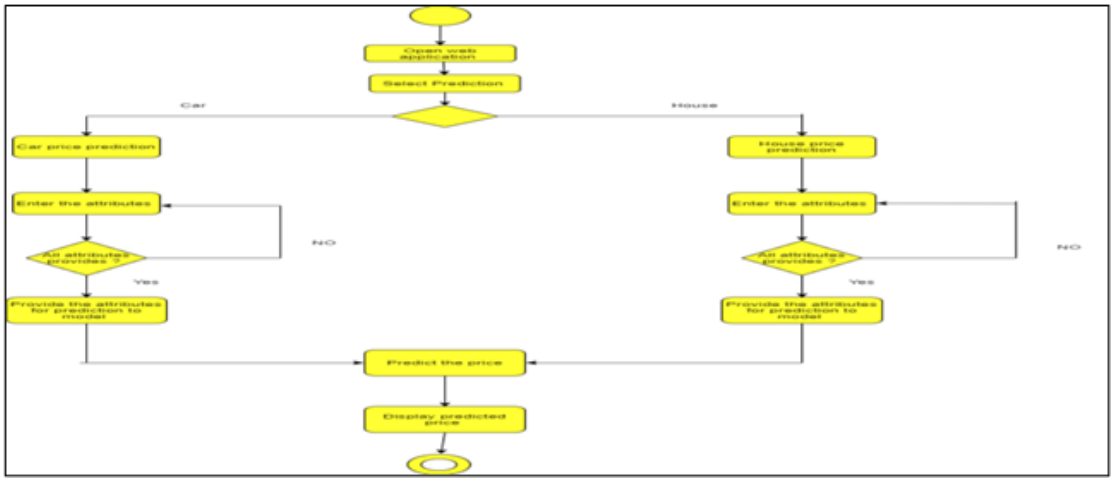
4.2.4 Use case diagram



4.2.5 Sequence Diagram

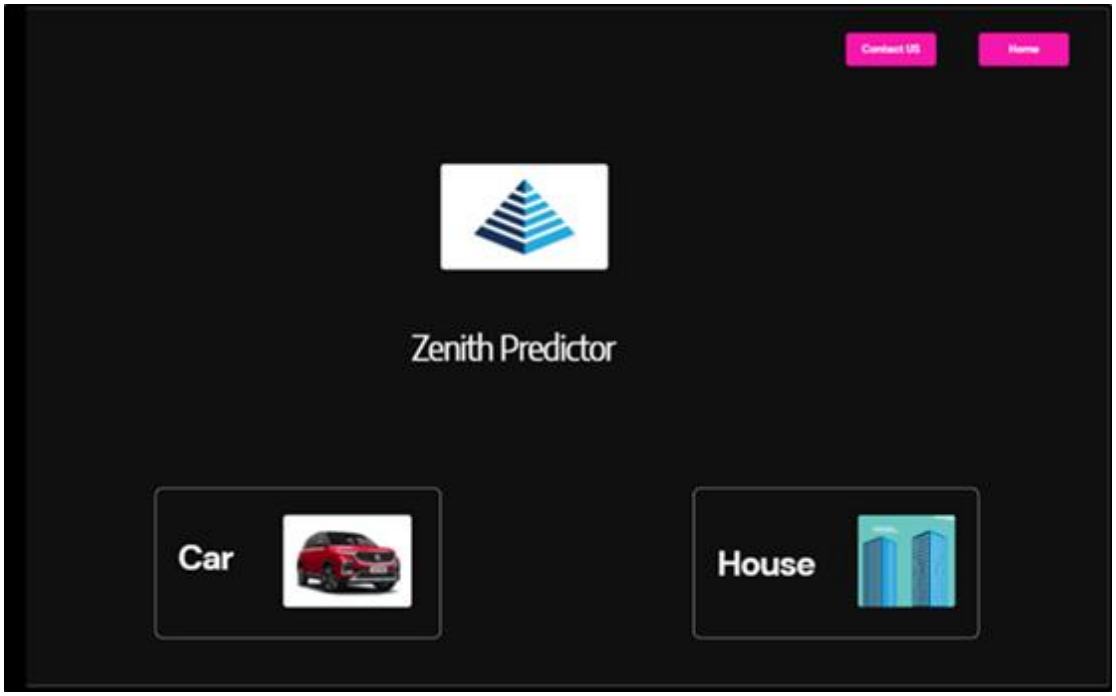


4.2.6 Activity diagram



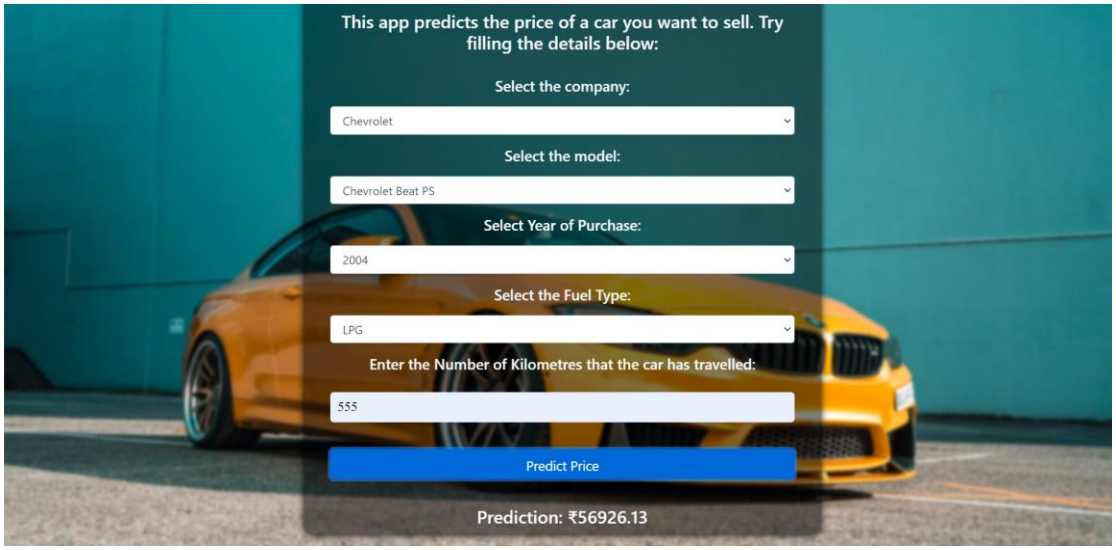
5 Implementation and Results

Home page



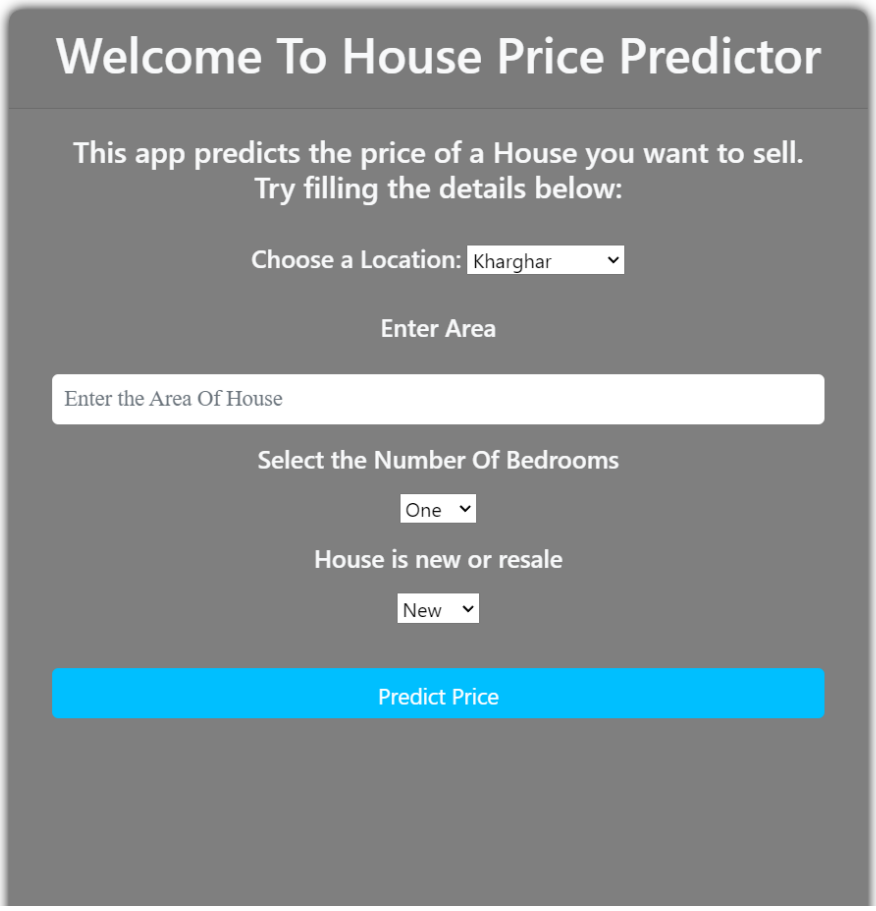
5.1 Home page

Car Price prediction



5.2 car price page

House Price Prediction



5.3 House price page

Implementation of car price prediction model

a) Importing dependencies and checking quality of data

[3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
%matplotlib inline
mpl.style.use('ggplot')
```

[4]:
```
car=pd.read_csv('OneDrive/Desktop/car_price_predictor-master/quikr_car.csv')
```

[5]:
```
car.head()
```

[5]:

	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing XO eRLX Euro III	Hyundai	2007	80,000	45,000 kms	Petrol
1	Mahindra Jeep CL550 MDI	Mahindra	2006	4,25,000	40 kms	Diesel
2	Maruti Suzuki Alto 800 Vxi	Maruti	2018	Ask For Price	22,000 kms	Petrol
3	Hyundai Grand i10 Magna 1.2 Kappa VTVT	Hyundai	2014	3,25,000	28,000 kms	Petrol
4	Ford EcoSport Titanium 1.5L TDCi	Ford	2014	5,75,000	36,000 kms	Diesel

[6]:
```
car.shape
```

[6]:
```
(892, 6)
```

Anomalies in data

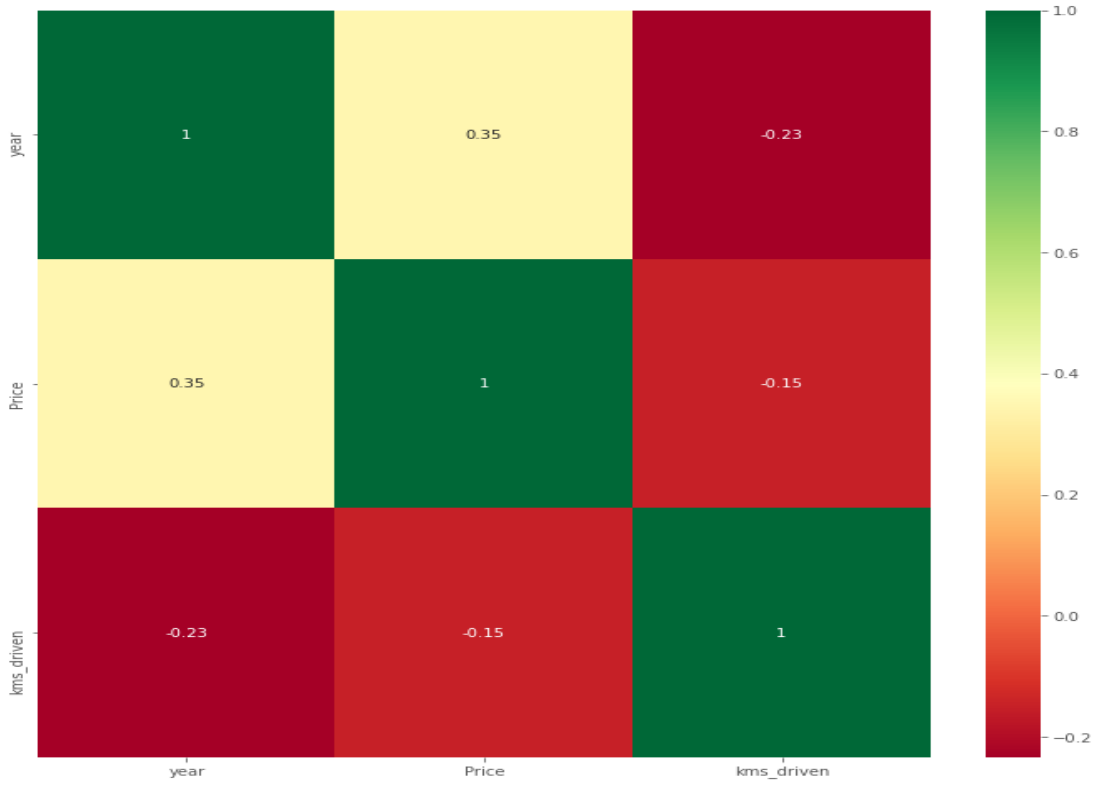
- names are pretty inconsistent
- names have company names attached to it
- some names are spam like 'Maruti Ertiga showroom condition with' and 'Well mentained Tata Sumo'
- company: many of the names are not of any company like 'Used', 'URJENT', and so on.
- year has many non-year values
- year is in object. Change to integer
- Price has Ask for Price
- Price has commas in its prices and is in object
- kms_driven has object values with kms at last.
- It has nan values and two rows have 'Petrol' in them
- fuel_type has nan values

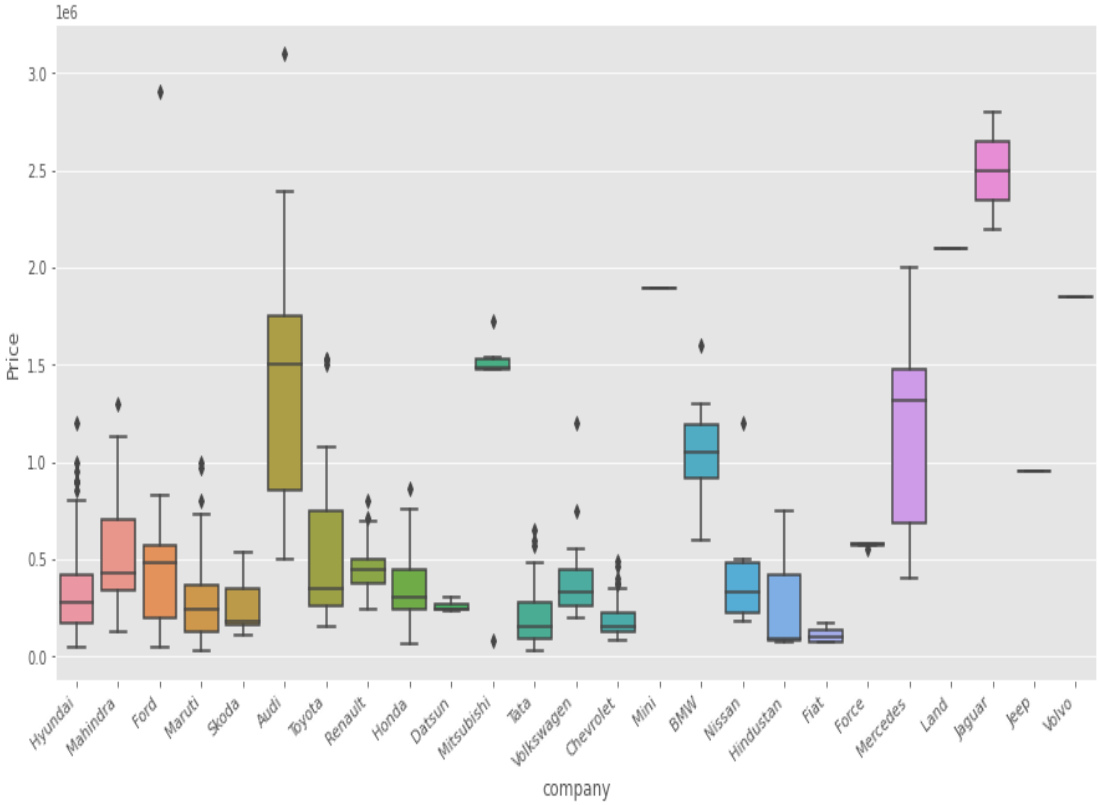
b) Cleaning of data

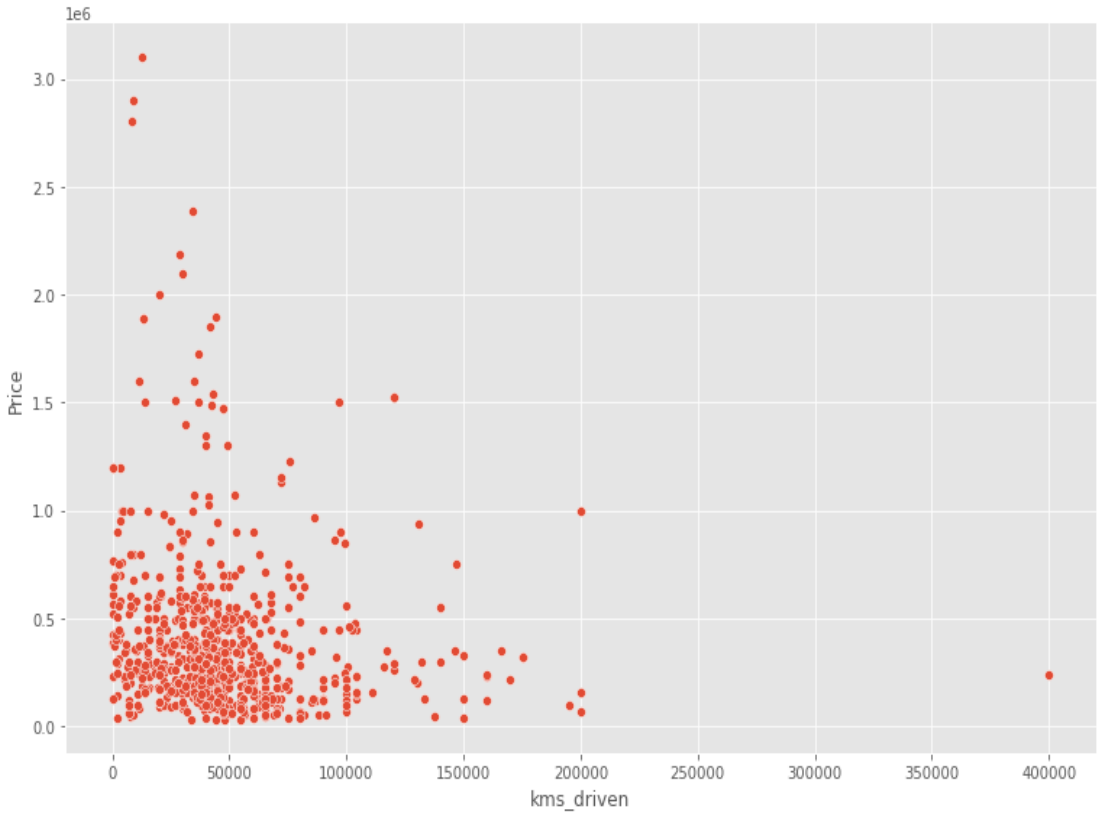
Cleaning Data

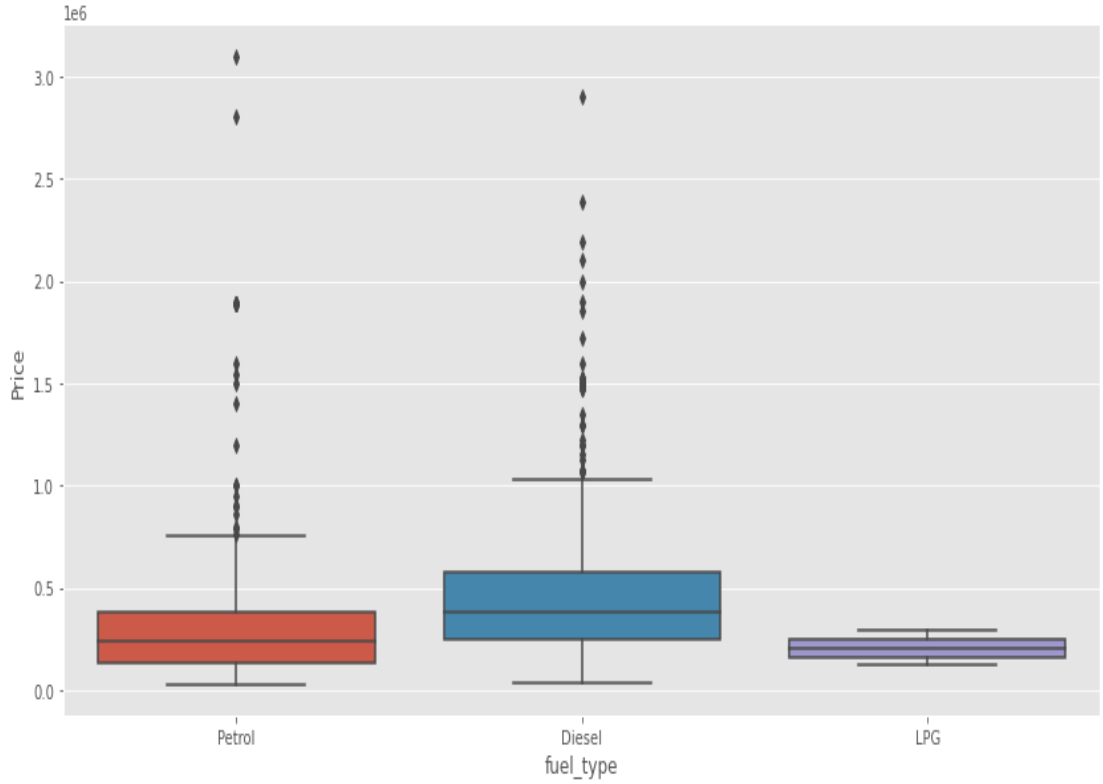
```
[22]: # year has many non-year values
car=car[car['year'].str.isnumeric()]

[23]: # year is in object. Change to integer
car['year']=car['year'].astype(int)

[24]: # Price has Ask for Price
car=car[car['Price']!='Ask For Price']

[25]: # Price has commas in its prices and is in object
car['Price']=car['Price'].str.replace(',','').astype(int)

[26]: # kms_driven has object values with kms at last.
car['kms_driven']=car['kms_driven'].str.split().str.get(0).str.replace(',','')

[27]: # It has nan values and two rows have 'Petrol' in them
car=car[car['kms_driven'].str.isnumeric()]

[28]: car['kms_driven']=car['kms_driven'].astype(int)

[29]: # fuel_type has nan values
car=car[~car['fuel_type'].isna()]

[30]: car.shape

[30]: (816, 6)

[31]: car['company'].unique()

[31]: array(['Hyundai', 'Mahindra', 'Ford', 'Maruti', 'Skoda', 'Audi', 'Toyota',
       'Renault', 'Honda', 'Datsun', 'Mitsubishi', 'Tata', 'Volkswagen',
       'Chevrolet', 'Mini', 'BMW', 'Nissan', 'Hindustan', 'Fiat', 'Force'
```

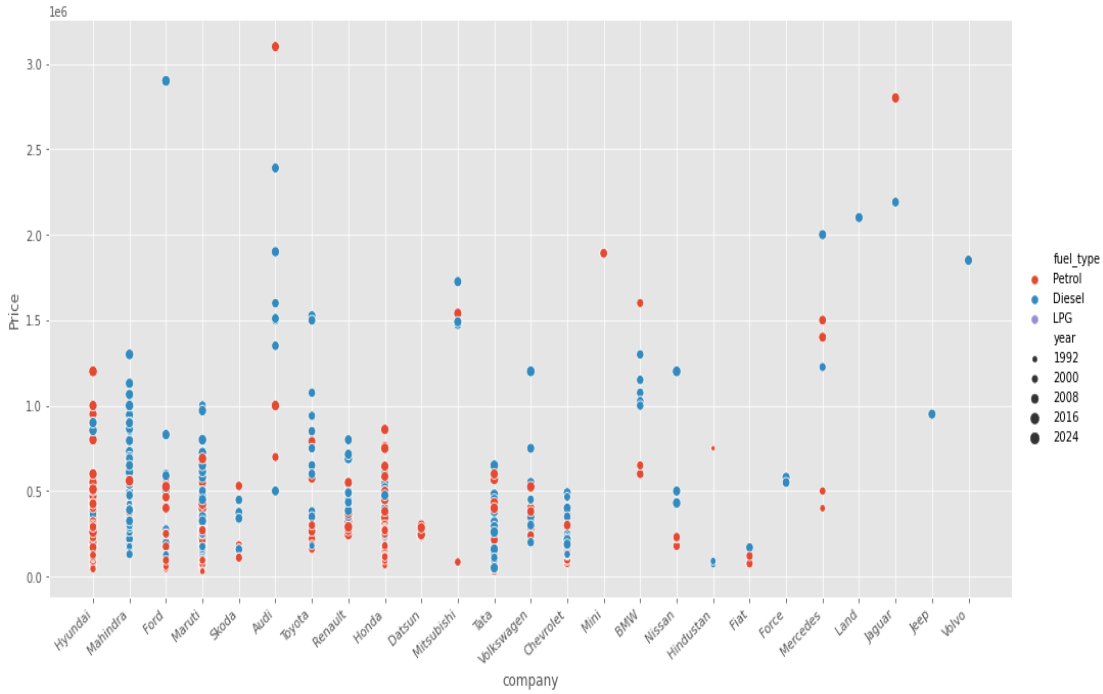
c) Data visualization




1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
Hyundai
Mahindra
Ford
Maruti
Skoda
Audi
Toyota
Renault
Honda
Datsun
Mitsubishi
Tata
Volkswagen
Chevrolet
Mini
BMW
Nissan
Hindustan
Fiat
Force
Mercedes
Land
Jaguar
Jeep
Volvo
company


1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
0
50000
100000
150000
200000
250000
300000
350000
400000
kms_driven





d) Implementing Linear regression

Applying Train Test Split

```
[71]: from sklearn.model_selection import train_test_split
      X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.1)

[72]: from sklearn.linear_model import LinearRegression

[74]: from sklearn.preprocessing import OneHotEncoder
      from sklearn.compose import make_column_transformer
      from sklearn.pipeline import make_pipeline
      from sklearn.metrics import r2_score

[75]: # Creating an OneHotEncoder object to contain all the possible categories

[76]: ohe=OneHotEncoder()
      ohe.fit(X[['name','company','fuel_type']])

[76]: OneHotEncoder()

[77]: # Creating a column transformer to transform categorical columns
      column_trans=make_column_transformer((OneHotEncoder(categories=ohe.categories_),['name','company','fuel_type']),
                                           remainder='passthrough')
```

Linear Regression Model

```
[78]: lr=LinearRegression()

[79]: # Making a pipeline
      pipe=make_pipeline(column_trans,lr)

[80]: # Fitting the model
      pipe.fit(X_train,y_train)
```

```
[80]: # Fitting the model
      pipe.fit(X_train,y_train)

[80]: Pipeline(steps=[('columntransformer',
                       ColumnTransformer(remainder='passthrough',
                                         transformers=[('onehotencoder',
                                                        OneHotEncoder(categories=[array(['Audi A3 Cabriolet', 'Audi A4 1.8', 'Audi A4 2.0', 'Audi A6 2.0',
             'Audi A8', 'Audi Q3 2.0', 'Audi Q5 2.0', 'Audi Q7', 'BMW 3 Series',
             'BMW 5 Series', 'BMW 7 Series', 'BMW X1', 'BMW X1 sDrive20d',
             'BMW X1 xDrive20d', 'Chevrolet Beat', 'Chevrolet Beat...
                                                                                  array(['Audi', 'BMW', 'Chevrolet', 'Datsun', 'Fiat', 'Force', 'Ford',
             'Hindustan', 'Honda', 'Hyundai', 'Jaguar', 'Jeep', 'Land',
             'Mahindra', 'Maruti', 'Mercedes', 'Mini', 'Mitsubishi', 'Nissan',
             'Renault', 'Skoda', 'Tata', 'Toyota', 'Volkswagen', 'Volvo'],
            dtype=object),
                                                                                  array(['Diesel', 'LPG', 'Petrol'], dtype=object)]),
                                                                 ['name', 'company',
                                                                  'fuel_type'])])),
                       ('linearregression', LinearRegression())])

[81]: y_pred=pipe.predict(X_test)

[82]: # Checking R2 Score
      r2_score(y_test,y_pred)

[82]: 0.8430114544991612

[83]: y_pred_train=pipe.predict(X_train)

[84]: r2_score(y_train,y_pred_train) # may be OF!

[84]: 0.9512932474239895

[85]: # Finding the model with a random state of TrainTestSplit where the model was found to give almost 0.92 as r2 score
```

e) R2 score obtained is 0.920084

```
[85]: # Finding the model with a random state of TrainTestSplit where the model was found to give almost 0.92 as r2_score
scores=[]
for i in range(1000):
    X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.1,random_state=i)
    lr=LinearRegression()
    pipe=make_pipeline(column_trans,lr)
    pipe.fit(X_train,y_train)
    y_pred=pipe.predict(X_test)
    scores.append(r2_score(y_test,y_pred))

[86]: np.argmax(scores)

[86]: 655

[87]: scores[np.argmax(scores)]

[87]: 0.9200894544056878
```

f) Implementing Lasso regression

R2 score 0.72500

```
[14]: column_trans=make_column_transformer((OneHotEncoder(categories=ohe.categories_),['name','company','fuel_type']),
                                       remainder='passthrough')

[18]: la=Lasso(max_iter=10000)

[19]: pipe=make_pipeline(column_trans,la)

[20]: pipe.fit(X_train,y_train)

[20]: Pipeline(steps=[('columntransformer',
                       ColumnTransformer(remainder='passthrough',
                                         transformers=[('onehotencoder',
                                                        OneHotEncoder(categories=[array(['Audi A3 Cabriolet', 'Audi A4 1.8', 'Audi A4 2.0', 'Audi A6 2.0',
       'Audi A8', 'Audi Q3 2.0', 'Audi Q5 2.0', 'Audi Q7', 'BMW 3 Series',
       'BMW 5 Series', 'BMW 7 Series', 'BMW X1', 'BMW X1 sDrive20d',
       'BMW X1 xDrive20d', 'Chevrolet Beat', 'Chevrolet Beat...
                                                                              array(['Audi', 'BMW', 'Chevrolet', 'Datsun', 'Fiat', 'Force', 'Ford',
       'Hindustan', 'Honda', 'Hyundai', 'Jaguar', 'Jeep', 'Land',
       'Mahindra', 'Maruti', 'Mercedes', 'Mini', 'Mitsubishi', 'Nissan',
       'Renault', 'Skoda', 'Tata', 'Toyota', 'Volkswagen', 'Volvo'],
      dtype=object),
                                                                              array(['Diesel', 'LPG', 'Petrol'], dtype=object)]),
                                                                    ['name', 'company',
                                                                     'fuel_type'])])),
                       ('lasso', Lasso(max_iter=10000))])

[21]: y_pred=pipe.predict(X_test)

[22]: r2_score(y_test,y_pred)

[22]: 0.7250045789706804
```

Implementation of house price prediction model

a) Data understanding and pre-processing

```
[1]: import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     from sklearn.linear_model import Ridge
     from sklearn import metrics

[2]: # loading the data from csv file to pandas dataframe
     House_dataset = pd.read_csv('OneDrive/Desktop/Miss/56 Pooruvi Singh/Mumbai1.csv')

[3]: # inspecting the first 5 rows of the dataframe
     House_dataset.head()
```

[3]:

	Price	Area	Location	Bedrooms	New_Resale
0	4500000	600	Kharghar	1	0
1	6700000	650	Kharghar	1	0
2	4500000	650	Kharghar	1	0
3	5000000	665	Kharghar	1	0
4	12500000	1550	Kharghar	3	0

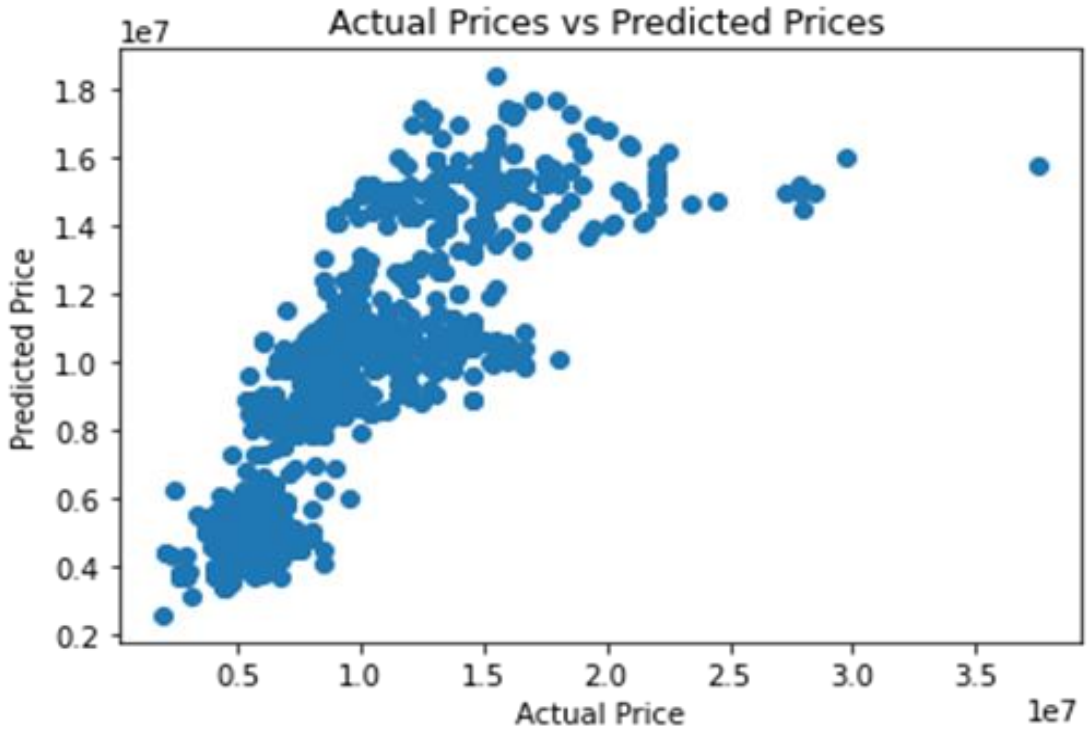
```
[4]: # checking the number of rows and columns
     House_dataset.shape

[4]: (1303, 5)

[5]: # checking the number of rows and columns
     House_dataset.shape

[5]: (1303, 5)
```

b) Data pre-processing

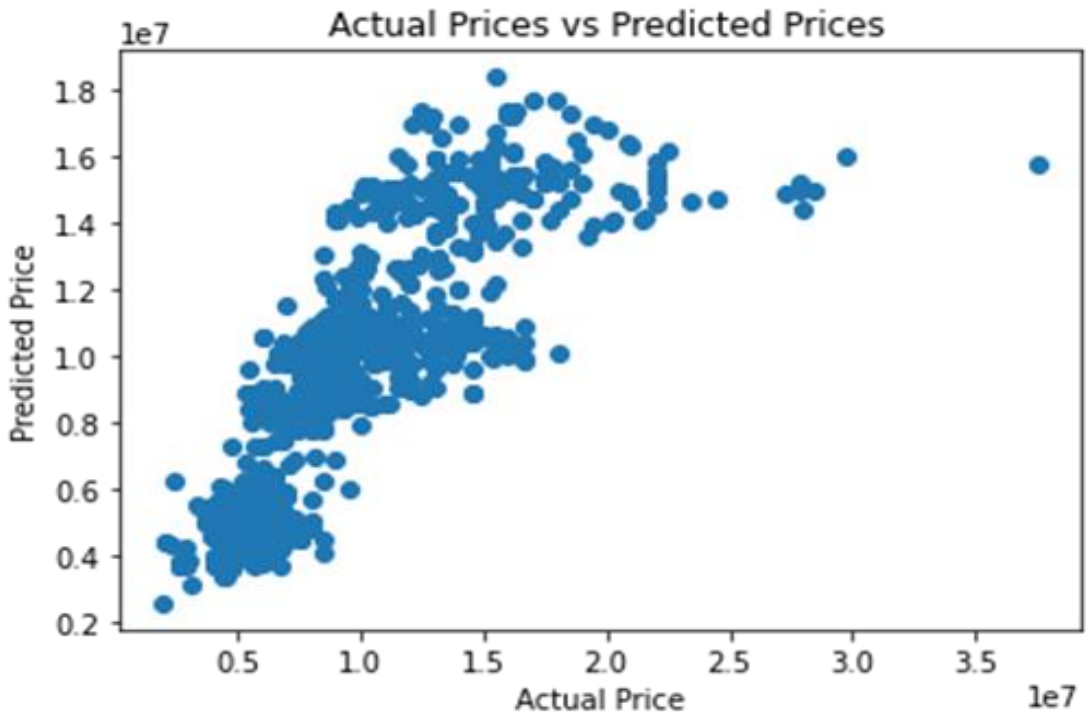
```
[6]: # getting some information about the dataset
     House_dataset.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1303 entries, 0 to 1302
     Data columns (total 5 columns):
      #   Column      Non-Null Count  Dtype
     ---  ------      --------------  -----
      0   Price       1303 non-null   int64
      1   Area        1303 non-null   int64
      2   Location    1303 non-null   object
      3   Bedrooms    1303 non-null   int64
      4   New_Resale  1303 non-null   int64
     dtypes: int64(4), object(1)
     memory usage: 51.0+ KB

[ ]:

[7]: # checking the number of missing values
     House_dataset.isnull().sum()

[7]: Price        0
     Area         0
     Location     0
     Bedrooms     0
     New_Resale   0
     dtype: int64

[8]: import seaborn as sns
     #get correlations of each features in dataset
     corrmat = House_dataset.corr()
     top_corr_features = corrmat.index
     plt.figure(figsize=(5,5))
     #plot heat map
     g=sns.heatmap(House_dataset[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

c) Data visualization

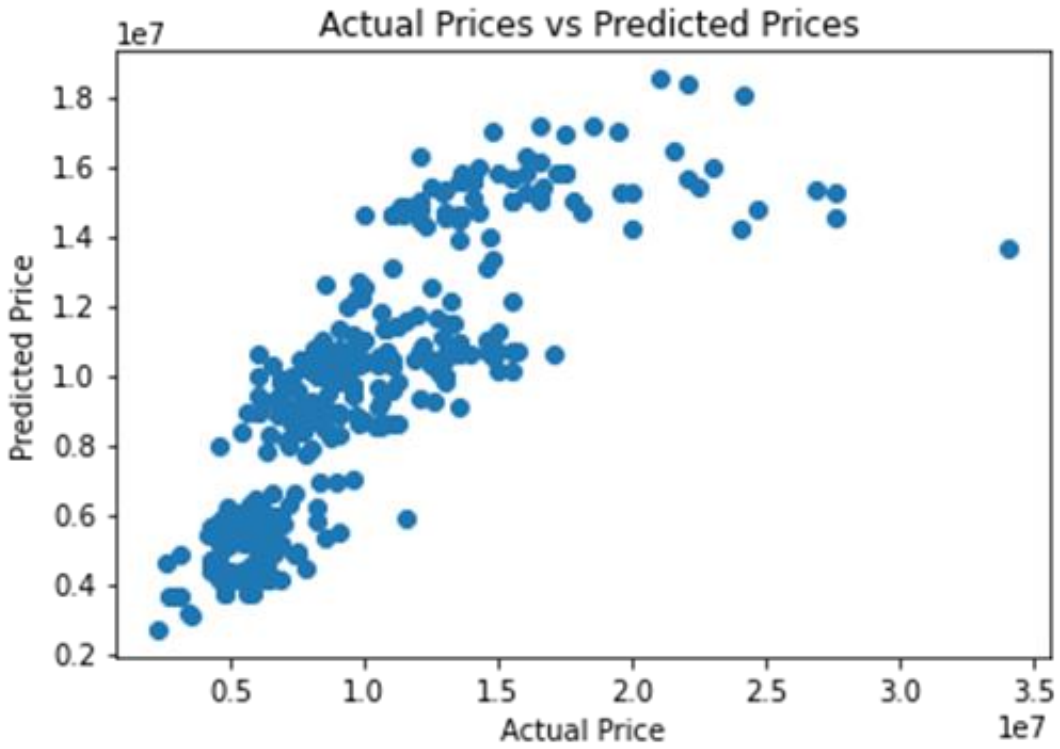
d) Actual vs Predicted price by linear regression for training dataset of House price prediction



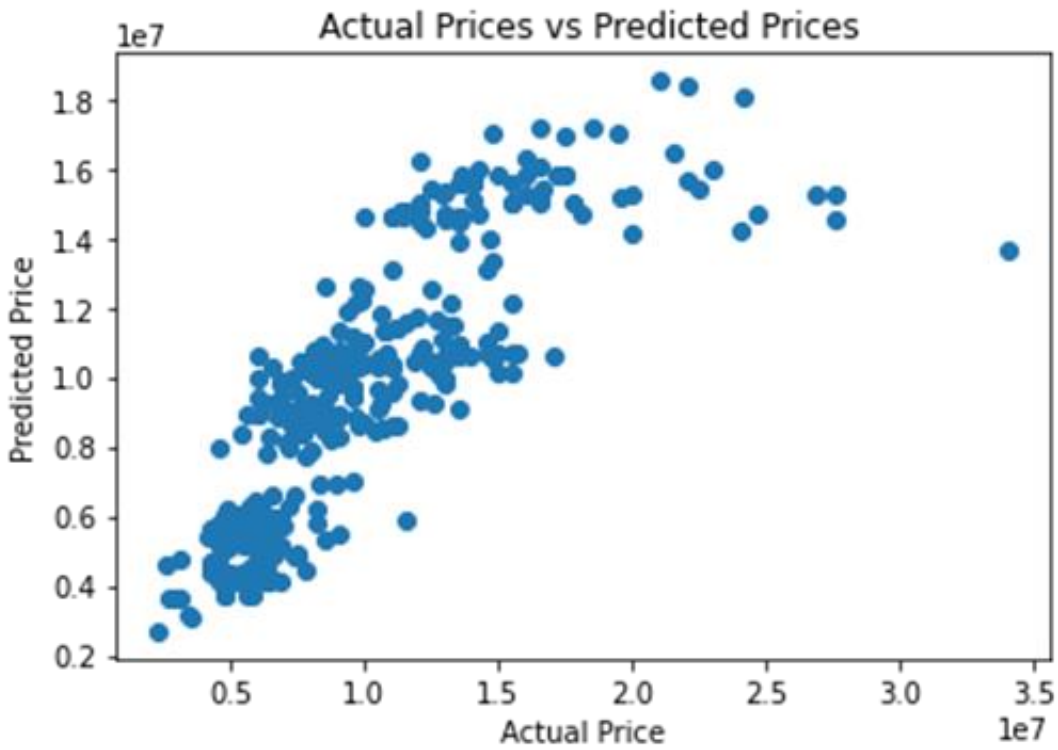
e) Actual vs Predicted price by Ridge regression for training dataset of House price prediction



f) Actual vs Predicted price by linear regression for test dataset of House price prediction



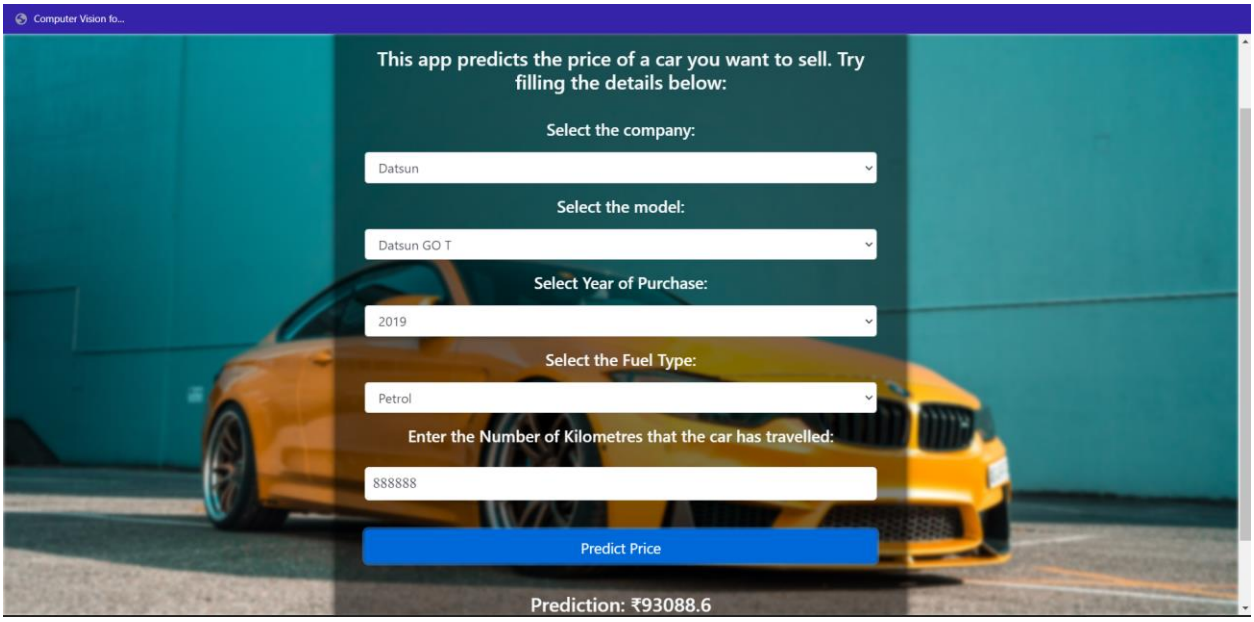
g) Actual vs Predicted price by Ridge regression for test dataset of House price prediction



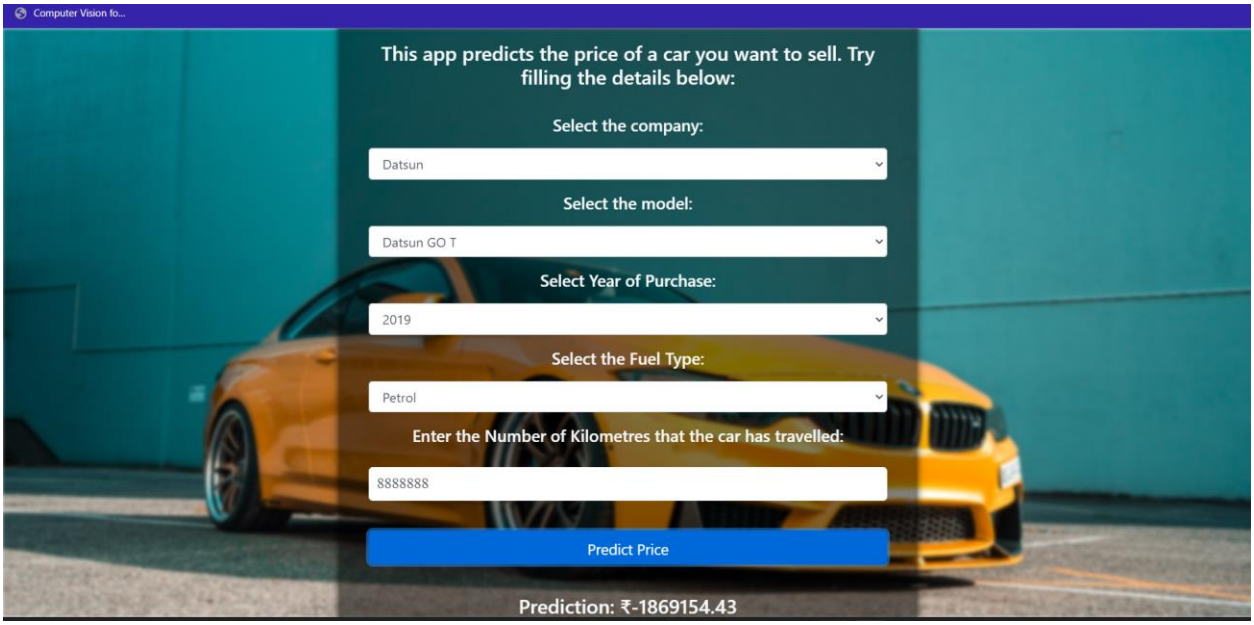
- For the house price prediction model Linear and Ridge regression were studied here the dataset was divided into 70 percent for training data and 30 percent for test data.
- For Linear regression a R2 score of 0.668408 was obtained for training dataset and a R2 score of 0.693642 was obtained for test 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.

6 Testing

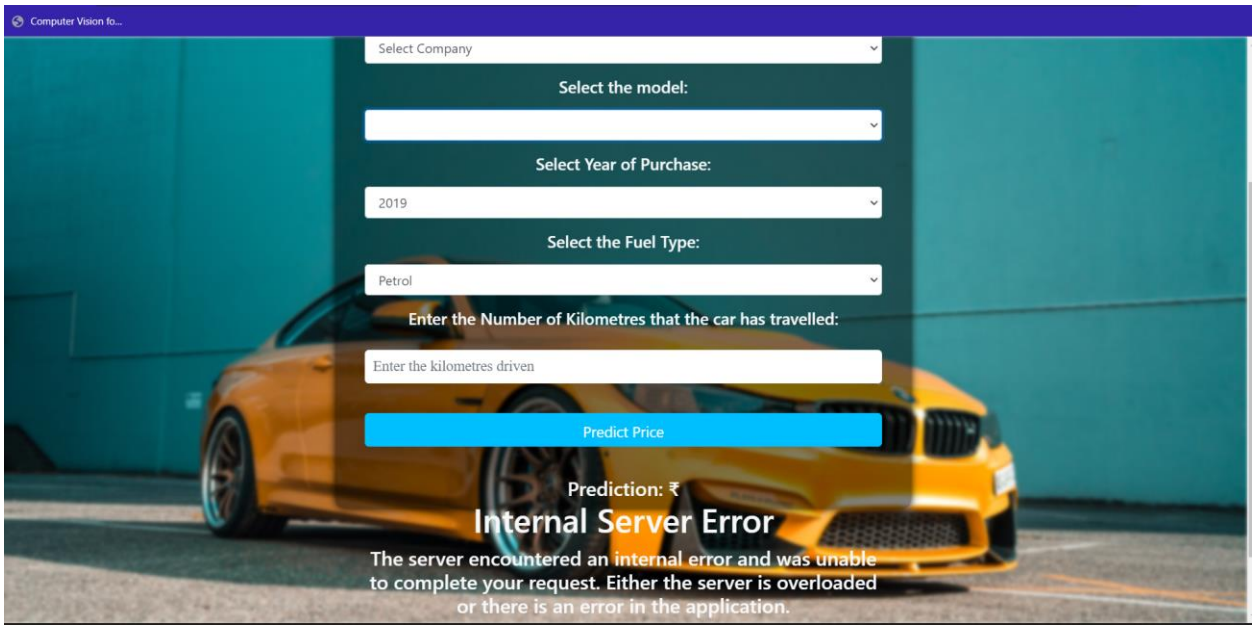
a) Entering value of kilometres driven in lakhs and below the prediction value is positive and acceptable



b) Entering value of kilometres driven in ten lakh and above leads to negative prediction values



c) Missing any value in prediction leads to production server error



d) If the area of house entered is very low then also the prediction value is in lakhs

```
[28]: lin_reg_model.predict(pd.DataFrame([[1,2,1,0]],columns=['Area','Location','Bedrooms','New_Resale']))

[28]: array([1792949.71106833])
```

7 Conclusion and Future work

Hence, we have implemented a machine leaning model that helps user predicting prices of used cars on various features like KMS Driven, Fuel type, purchase year, company name and model name. Value of Housing properties on the basis of location, area, no of bedrooms and new or resale We have obtained a R2 score of 0.92 for car price prediction and a R2 score of 0.66 for house price prediction

As a part of future work we can implement the same model with help of other algorithms like KNN, Naïve Bayes and by using neural networks also we can have a larger dataset for prediction by storing the new data as entered by users also for the better prediction we can consider various other features like for car price we can consider the color, cosmetic condition etc. and for house price various amenities that are available in the locality and in the campus. Also, along with the prediction model we can implement a platform where user can buy or sell their house or car.

8 Outcomes

SrNo	Course Outcomes	Pooruvi(GL)	Aditya	Omkar	Suraj
1	Learner will be able to Identify societal/research/innovation/entrepreneurship problems through appropriate literature surveys	5	4	5	5
2	Identify Methodology for solving above problem and apply engineering knowledge and skills to solve it	4	5	5	5
3	Validate, Verify the results using test cases/benchmark data/theoretical/ inferences/experiments/simulations	5	5	5	4
4	Analyze and evaluate the impact of solution/product/research/innovation /entrepreneurship towards societal/environmental/sustainable development	4	4	5	5
5	Use standard norms of engineering practices and project management principles during project	5	5	4	4
6	communicate through technical report writing and oral presentation.	5	5	5	5
7	Gain technical competency towards participation in Competitions, Hackathons, etc.	5	5	5	5
8	Demonstrate capabilities of self-learning, leading to lifelong learning.	5	5	4	5
9	Develop interpersonal skills to work as a member of a group or as leader	5	5	5	5
		5	5	5	5

9 References

[1] S. Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques,” International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.

[2] N. Kanwal and J. Sadaqat, “Vehicle Price Prediction System using Machine Learning Techniques,” International Jounal of Computer Applications, vol. 167, no. 9, pp. 27–31, 2017.

[3] S. Peerun, N. H. Chummun, and S. Pudaruth, “Predicting the Price of Second-hand Cars using Artificial Neural Networks,” The Second International Conference on Data Mining, Internet Computing, and Big Data, no. August, pp. 17–21, 2015.

[4] N.Sun, H. Bai, Y. Geng, and H. Shi, “Price evaluation model in second-hand car system based on BP neural network theory,” in 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), jun 2017, pp. 431–436.

[5] Aminah Md Yusof and Syuhaida Ismail ,Multiple Regressions in Analysing House Price Variations. IBIMA Publishing Communications of the IBIMA Vol. 2012 (2012), Article ID 383101, 9 pages DOI: 10.5171/2012.383101.

[6] Babyak, M. A. What you see may not be what you get: A brief, nontechnical introduction to over fitting regression-type models. Psychosomatic Medicine, 66(3), 411-421.
[7] Atharva chogle, priyanka khaire,Akshata gaud, Jinal Jain .House Price Forecasting using Data Mining Techniques International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 12, December 2017.