Used Car price prediction using machine learning

1. ***Abstract*—There is a growing demand for used luxury cars in India. Research suggests that there is an increase of 30-40 percent of sales of old cars year by year.There are multiple reasons for this scenario most of the people are preferring to buy used cars:The age of the old cars are drastically reduced from averageused age of 5-6 to 2-3 years.Due to financial issues people are preferring to buy old cars.So many online platforms are offering transactions for used cars.The challenge in this system is there areno organized platforms to predict the accurate price.The object of this project is to implement multiple machine learning algorithms**
i.e. regression models to predict the prices of the used cars for different models. In this paper we implemented Ridge regressor, Linear regression and KNN regressor to predict the prices of the used cars.Various evaluation metrics are used tp measure the performance of the models i.e. Maximum Likelihood sequence estimation,Root Maximum Likelihood sequence estimation and R2 square.

2. ***Index Terms*—K-Nearest Neighbour regressor, Ridge regressor, Root Maximum Likelihood sequence estimation, Maximum Likelihood sequence estimation and R2 square.**

I. Introduction

In order to normalize the prices in the platforms we use machine learning techniques to predict the prices.For this experiment we use supervised machine learning algorithms Lasso regression , Linear regression and Random Forest regressor. Along with the machine learning techniques this paper tests the performance of the data on ANN(Artificial Neural Networks) using Keras regressor. In the end a comparative analysis is conducted to measure performance of the machine learning models with ANN. Amongst these algorithms RF regressor achieved MAE of 1.0970472 and R2 error value of 0.772584 which is highly accurate. In the year of 2016, nearly 70 million passenger cars were released, this raised demand for sales of used cars. The main reason for the surge in used cars is that most of the people are moving towards upgrading the existing models.With this the buyer as well as the seller are in search of information of prices trends.So machine learning models predict the prices based on the history. In this project we implement machine learning methods to predictthe car prices.The task is divided into two phases one is feature engineering and the other one is feature screening. In feature engineering we clean the data by removing the miss- ing values, removing outliers and selecting the best features and data normalization to improve the data quality. In the second phase we apply LightBGM(Light Gradient Boosting Machine) for correlation analysis and feature selection. After this process we apply machine learning algorithms XGBoost and RandomForest models.These models are mixed weighted to train and predict the data. In this paper Many business fields data analysis depends on probabilistic methods.The regression problem type case studies make use of MLR(Multi Linear Regression) models to predict the continuous value predictions. In this study several MLR types of algorithms are implemented on the feature set of mileage, fiscal power, mark, model etc.Amongst the all algorithms GBR (gradient Boosting Regressor) showed the highest R2 value.

Tools and technologies used for this project are: The projectis designed in Python programming language.Project is exe-cuted and evaluated in the Pycharm environment. For modelcreation and evaluation is performed using scikit-learn library.Mean Squared Error is the basic metric for measuring the performance of the regression model.Its loss function isoptimized using the mean of the squares of the predictionswhich is called least squares.MSE helps to punish the modelsfor larger error if the difference between the model predictedand expected value is high the squared value is also willbe high. Root Mean Squared Error is another parameter thatmeasures the performance of the regression models. RMSEis the extension of the MES metric. MES gives the squaredvalues of the parameters applying square root to the MESgives the same metrics and is easy to use. Another parameter tomeasure the performance of the model MAE is Mean AbsoluteError. The other metrics MSE and RMSE punish the largererrors whereas the MAE identifies the increase in the errors. In this project we are proposing a machine learning model to predict the price of the car based on the parameters likelocation, number of kilometers traveled, model released year,fuel type, mileage and type of the engine. This is a regressionproblem.In this paper we are proposing Linear and RandomForest machine learning regression models with feature engineering techniques.

The project workflow is divided into following steps; Step 1: collecting dataset.The dataset is collected from the publicly available repository. Step 2: Preprocessing the raw data. Removing/imputing the null values. Step 3:Exploratory data analysis to show the data patterns Step 4: Performing feature engineering to select the important features Step 5: Creating machine learning models Step 6: Evaluating the models

In this paper we are proposing multiple regression models SVM regressor, RandomForest regressor, Polynomial regression, Decision tree regression are used to predict the prices of the used cars. To predict the price different attributes are used including vehicle mileage, manufacturing year, fuel consumption, road tax and engine size.

In this paper used car price prediction experiments are conducted using Gradient Boosting Regression model and thismodel is evaluated using Mean Absolute Error(MAE) = 3D

Github Link: https://github.com/kranthisree009/ML-project.git

0.28 in addition to this we have used Random Forest Regressor to find the best price for used cars and the model achieved Mean Squared Error of 0.35 and multiple linear regression with Mean Squared Error of 0.55 .

With the increase of used car sales in this paper we have conducted a research of finding the trends of used car prices. So in the experimental analysis we have implemented RandomForest regressor and LightGBM regressor to find the prices. In the evaluation stage we have recorded error metrics of each algorithm Mean Squared Error, Mean Absolute Error and R2 square error.We concluded that LightGBM achieved a low error rate compared to the other algorithms. So we can use LightGBM in the future for other regression calculations

II. Motivation

New car price components are fixed price and the taxes imposed by the government. According to the study the average price of the new car is 48000 dollars which is expensive. Another study conducted to check the average ownership of the cars in years is declining year by year due to new features added to the new cars very frequently. On the other hand people are willing to buy used cars rather than new cars.

There are third party websites which help to sell or buy used cars on their platform. But the prices of the cars are fixed with the additional charges and sometimes these prices do not reflect the prices of market values. A system which includes multiple algorithms to analyze the prices is beneficial to both the seller and the buyer

III. Objectives

The object of this project is to implement multiple machine learning algorithms i.e. regression models to predict the prices of the used cars for different models. In this paper we imple- mented Ridge regressor, Linear regression and KNN regressor to predict the prices of the used cars.

Main features of the project are: To evaluate the performance of the machine learning models various evaluation metrics are used. Maximum Likelihood sequence estimation Root Maximum Likelihood sequence estimation R2 square are the metrics used to predict the car prices.

IV. Related work

Although the used car market has expanded in size in recent years, the pricing evaluation method of the used car market in my country has shown the issue that it does not meet market demand. With the aid of a precise used car price prediction, people may make informed decisions and keep as far away from the inflated costs of used automobiles on the market as they can. This study predicts the price of used automobiles using the random forest and LightGBM algorithms, and then compares and analyses the prediction outcomes. The trials' findings provide the following relevant evaluation criteria for the random forest and LightGBM models: The R value for MSE and MAE is 0.0373 and 0.0385, respectively [9].

Along with the increase in motor vehicle ownership, per capita ownership rates, and per capita ownership numbers, so too is the demand for second hand cars. The price of secondhand automobiles can be difficult to determine because it depends on a variety of factors, including the vehicle's condition and fundamental construction. Through data analysis and modelling, the used automobile trading platforms dis-cussed in this article will help resolve this problem. This is ac- complished by offering a thorough overview of the modelling procedure, including data pretreatment, feature engineering, and parameter adjusting. Three machine learning models, such as the gradient lifting decision, are effective in tackling regression problems [8].

Newly constructed automobiles are not able to reach buyers despite the large increase in car usage due to problems such as high prices, limited supply, financial difficulties, and other factors. As a result, the used car market is expanding quickly everywhere, although it is still quite young in India and is mostly controlled by the unorganised sector. When buying a used car, this raises the possibility of fraud. Therefore, it is necessary to develop a highly precise model that can calculate the price of a used car without favouring either the buyer or the merchandiser.This model creates a Random Forest Machine Learning model and a Supervised learning-based Artificial Neural Network model that can both learn from the provided automotive dataset [18].

China's used car trade business has expanded considerably. The scientific examination of car pricing, which reduces transaction risks in the used car market and promotes the market's healthy growth, makes it feasible to trade used cars at reasonable and fair prices. Even though there has been extensive research on used car price prediction, it is still challenging to provide an accurate and trustworthy estimate. This study proposes a more sophisticated machine learning method based on LightGBM to predict used car values reliably and imaginatively using actual transaction records from a used car selling website. By assessing the relative weights of each feature, the basic dataset is cleaned and the features are filtered[20].

In several business areas related to statistics and machine learning (ML), multiple linear regression (MLR) models are often used to estimate and fit a linear relationship between a continuous response variable and various explanatory factors. In our case study, we used a number of regression techniques based on supervised machine learning to anticipate the used car resale price given a variety of parameters including mileage, fuel type, financial power, mark, model, and the year of manufacturing. In every model that was looked at, the gradient boosting regressor showed a high R-squared score and a low root mean square error [4].

The market for used cars is one of the newest and fastest-growing industries. Online marketplaces have grown more crucial as the digital era has developed for satisfying the needs of both buyers and sellers in the used car market. However, choosing the worth of the used car on the buyer's end and its price on the seller's end offers a conundrum. Therefore, a pre-owned car price determination model is needed to project a used car's resale price for company finance and customer

purchase. Numerous ML models for determining the fair price of a used car have been proposed by researchers in order to make this study topic evergreen. This review article concentrates on a number of machine learning methods that have been proposed by researchers for calculating the price of used cars [13]

The major objective of this project is to develop a mathematical model that will predict the price of a used car basedon its current characteristics. It can be difficult to predict the cost of a used car because of the many factors that might affect the price, such as the car's current mileage, condition, model, and year. Furthermore, from the seller's perspective, it becomes difficult to anticipate a used car's price properly. Asa result, the goal of this project is to create tools and conduct research into models that can precisely estimate the price of a used car based on its specifications. A customer will be much better informed when making a purchase as a result [14].

With the use of several factors, including engine size, fuel type, gearbox, road tax, and mileage, this study aims to create a model that can estimate the realistic pricing of secondhand cars. The vendors, buyers, and automakers can all benefit from this strategy in the used car market. Eventually, it can produce a price prediction that is pretty accurate based on the infor- mation that users provide. The process of developing a model makes use of machine learning and data science. Scraping used car listings was utilised to gather the dataset. The study included a number of regression approaches, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression, to achieve the best level of accuracy [5].

Used cars are in demand today, and the used car market is growing. When buyers try to compare prices for used carsof the same type across numerous websites, they become perplexed and end up buying cars that are overpriced or overly expensive in comparison to what they should be. This study aims to calculate the price of a used car using factors such as brand, model, gearbox, and others. In order to avoid overpaying for an automobile, this study is expected to help prospective buyers determine a car's pricing based on its value[1] [17].

The usage of private vehicles, particularly cars, has increased during the Covid-19 pandemic due to the notion that they are the safest mode of transportation for preserving distance and avoiding the transmission of the virus. Based on data from two different Indonesian secondary car markets, a comparison of a Car X price sample in the city of Surabaya with the specifications for the car years 2015 to 2018 with car mileage under 1000 kilometres reveals that used cars come ina range of prices; as a result, a system for predicting used car prices is necessary so that people can learn the typical priceof used cars sold in the market [6] [10].

Due to the quick growth in the number of private vehicles and the expansion of the used vehicle market, used automobiles are currently the most popular choice among consumers wanting to buy cars. Both buyers and sellers have the option ofonline P2P exchange on the used automobile marketplace on the internet. In such systems, the reliability of the second-hand car price estimation mainly determines whether the vendor and the customer may have a more successful trading experience. In this study, a pricing evaluation methodology based on big data analysis is presented. The pricing data for each type of car is analysed using the BP neural network employing widely dispersed vehicle data as well as a substantial amount of vehicle transaction data [16] [15].

As motor vehicle technology develops, there is a rising market for motor vehicles that are now in circulation links, or "used cars." Given that they constitute a particular category of "e-commerce commodity," used cars are more sophisti- cated than typical "e-commerce commodities." As a result, estimating the value of secondhand cars can be difficult be- cause it depends on both their physical state and fundamental construction. There isn't yet a set criteria in place by the state for figuring out the worth of secondhand autos. The first step in this study is feature engineering, which involves feature screening and data preparation to address the relevant issue. Data cleaning involves eliminating outliers and filling in blanks, and data transformation involves transforming data [3] [2].

Every corporate organisation is aware that sometimes tough choices must be made. Making the wrong decisions can lead to large losses or even the failure of a business. This article is concerned with a retail operation that deals with the sale of used autos. The major goal is to develop a prediction model that can estimate the used car market value based on crucial factors. Extra Trees Regression, a feature engineering method, and Random Forest Regression, a machine learning method, are both utilised to achieve the goal because Extra Trees Regression matches the number of decision trees and Random Forest Regression is modelled for prediction analysis.The results of our plan are overwhelmingly favourable [12] [7].

We use machine learning approaches in this work to forecast the price of used cars with a minimum of human involve- ment and to deliver more objective results. The dataset is preprocessed using the Pycaret module of Python, and the algorithm comparison function compares the effectiveness of each technique. Both the Random Forest and Extra Trees Re- gressors perform admirably in this investigation. The algorithm was then optimised using the hyperparameter function. The results show that the performance with R2 = 0.9807, which was reached with incredibly random numbers, is the best. Following the algorithm's acquisition and testing on new data, the final algorithm model was created [19].

This research work primarily focuses on one of the retail industries, namely the used automobile sales company, in an effort to suggest a creative solution to this challenge. It is well recognised that making difficult but intelligent judgements is an essential duty for every organisation. Making poor decisionscan result in significant losses or possibly the closure of a corporation [11].

V. Data Description

The dataset is collected from open source repository Kaggle.The dataset contains 13 columns as the features.

- Name: The brand and model of the car.
- Location : The location in which the car is being sold oris available for purchase.
- Year : The year or edition of the model.
- Kilometers Driven : The total kilometers driven in the carby the previous owner(s) in KM.
- FuelType : The type of fuel used by the car. (Petrol,Diesel, Electric, CNG, LPG)
- Transmission : The type of transmission used by the car.(Automatic / Manual).
- Mileage :The standard mileage offered by the car com-pany in kmpl or km/kg
- Engine :The displacement volume of the engine in CC.

VI. Proposed Framework

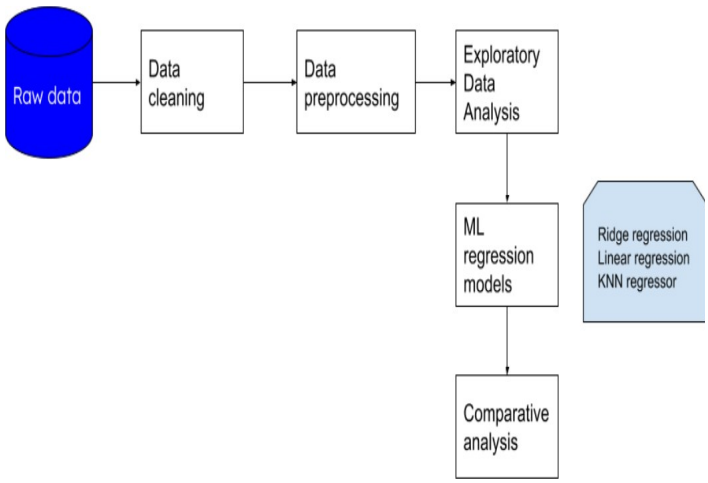

Figure 1. Workflow

A. Implementation:

- The implementation of the project is done in 4 phases:In the first step we have collected the dataset from the open source repository Kaggle. Collected raw data is processedby removing null values and duplicates.
- In the dataset half of the columns contains units in the columns power(bph) and mileage(km/kg) etc., in this scenario first we have converted the columns to strings and then removed the units from the columns. Again thesecolumns are converted to integer columns
- Exploratory data analysis is conducted to analyze the underlying structure of the data.
- Using pandas describe function we have understood the statistical features of the data to find the outliers and minimum and maximum values of the data.
- Data contains diverse columns in the data so data is normalized using scikit-learn standard scalar method.Normalized data is split into training and testing using scikit-learn train test split method.
- Training set is used to train the Linear regression, Ridge regression and KNN regressor models To compare the performance of these models various metrics are used(MLSE,RMLSE and R2 square errors)

B. Linear regressiion

The basic principle of linear regression is finding the linear relationship between the dependent and independent variables. If the dependent variable is single then the regression is called a single regression model and if the multiple dependent vari- ables are there then the regression is called multiple regression model. The relationship of the variables is represented usinga straight line.

To calculate the best fit line: Y = mx+b

Y : dependent variables X : independent variables b : is the intercept m:linear regression coefficient

Figure 2. Linear regression

C. Ridge regression

Ridge regression uses L2 regularization to add the penaltywhich is equivalent to the square of the magnitude of the coefficients.The minimization objective function of ridge re-gression is LS Obj+*(sum of square of coefficients)

D. KNN regressor:

K-Nearest Neighbor regressor is the parametric method to calculate the relationship between the dependent variables andthe continuous variables by using the average of the values in the neighborhood. The size of the neighborhood is decided by the analyst by performing the cross validation method.

E. Exploratory Data Analysis

Exploratory data analysis is an important step analysing the data. To perform this task we used the data visualization libraries matplotlib and seaborn.

- In the fig.3. we have visualized the box plots of owner type and price is visualized. Like natural cases the price

of first hand and second hand is high compared to the third and fourth above.

- In the second graph we have visualized the scatter plotof each feature in the dataset and observed the linear relationship with the target variable.
- Distribution of price is observed using histogram plot.It is observed that the distribution is between 0-20 and the maximum value is 40.The distribution of price is left skewed we can observe that long tail in the right hand side.

Figure 3. Owner type boxplot

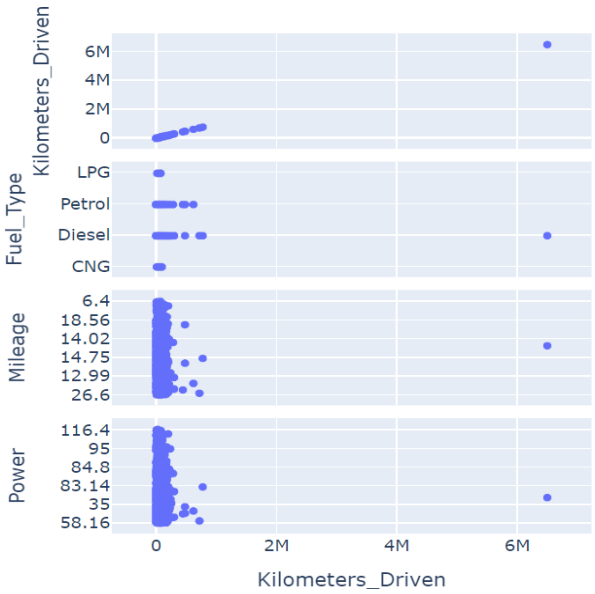

Figure 4. kilometers driven boxplot

VII. Results Analysis

In this paper we have implemented the 3 types of algorithms Linear regression, ridge regression and KNN regressor. For each algorithms we have recorded the various parameters to measure the regression capability MSLE, Root MSLEm R2 score and accuracy of the models. Among the three algorithms KNN regressor performed well. Linear regression model scored 73 percent accuracy, Ridge regression scored 73 percent accuracy and KNN regressor scored 86 percent accuracy. We have saved the best machine learning model KNN regressor in the pickle format and deployed onto the web application. Web application is designed using Flask frame work, HTML and CSS. Entire project is developed in Python programming language.

Figure 5. Fuel type boxplot

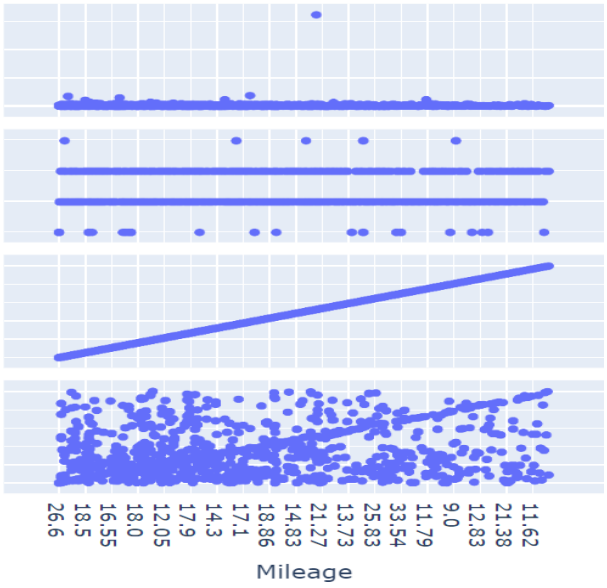

Figure 6. Mileage boxplot

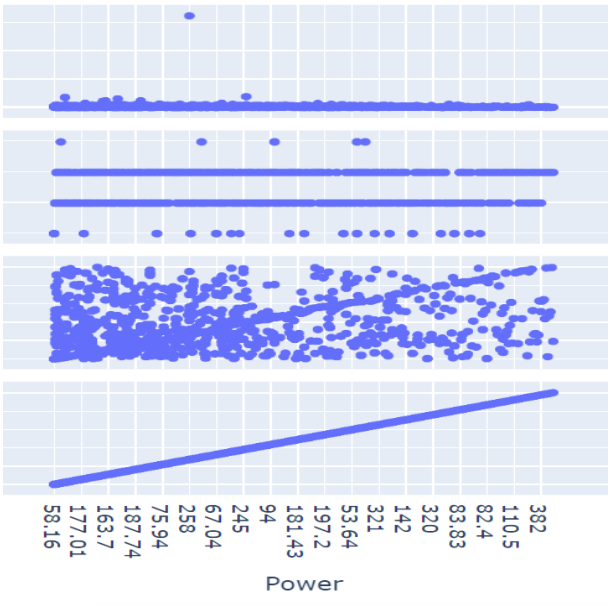

Figure 7. Power boxplot

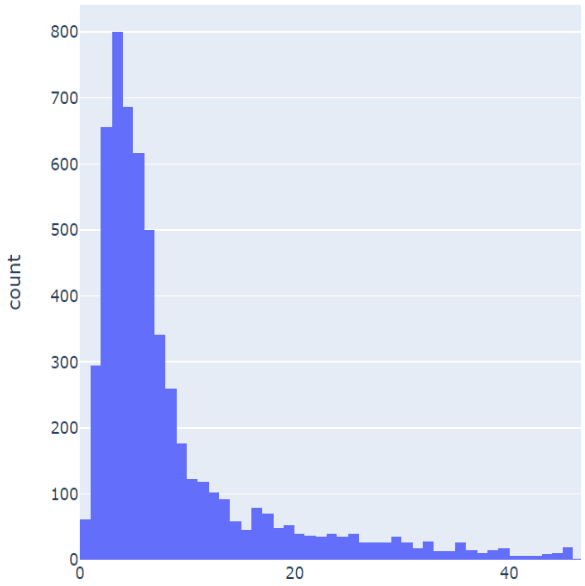

Figure 8. Price distribution

	Linear Regression	Ridge Regression	KNN
MSLE	31668.921464	32059.950187	15773.662179
Root MSLE	177.957640	179.052926	125.593241
R2 Score	0.738874	0.735650	0.869939
Accuracy(%)	73.887400	73.565000	86.993900

Figure 9. Comaparision of accuracies

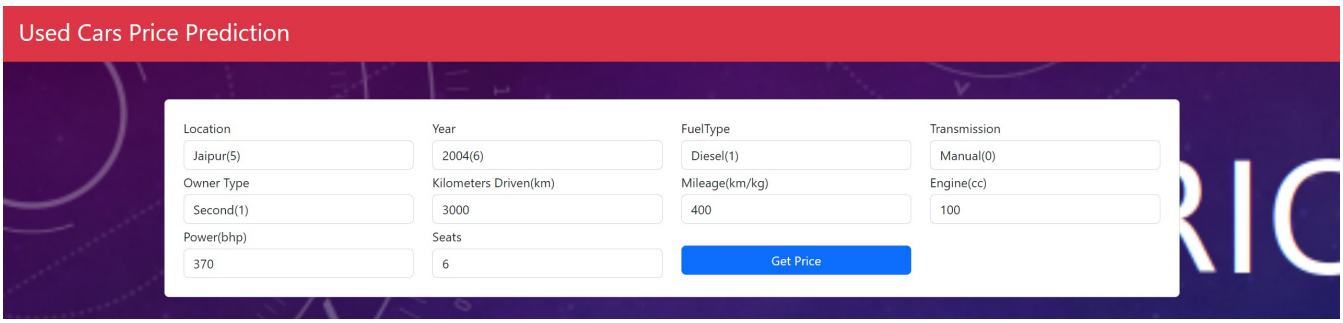

Figure 10. test case 1

Price : 310.2 K

Figure 11. test case result

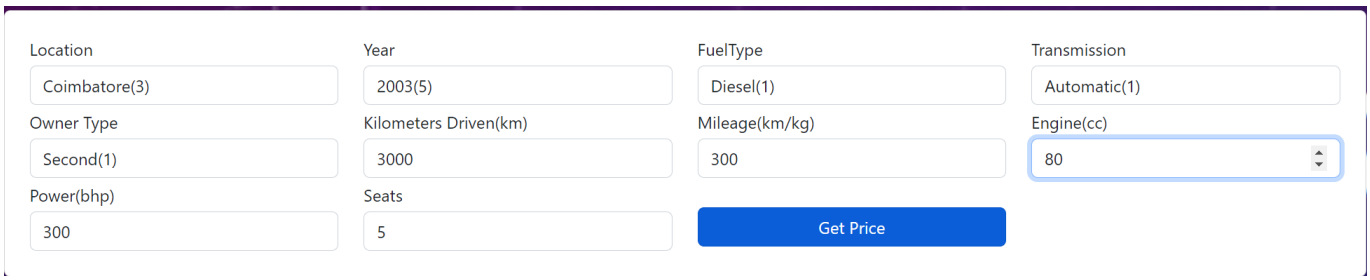

Figure 12. Test case2

Price : 332.4 K

Figure 13. test case 2 result

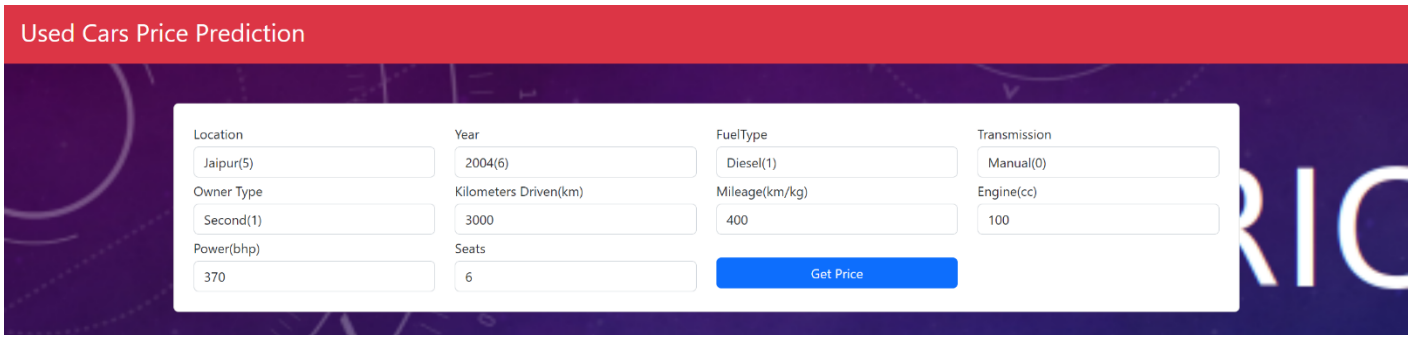

Figure 14. test case 3 result

REFERENCES

1. Fauzi Arifin Alghifari, Rachmadita Andreswari, and Edi Sutovo. Used cars price prediction in dki jakarta using extreme gradient boosting and bayesian optimization algorithm. In *2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, pages 01–05, 2022.
2. Rupesh Gupta, Avinash Sharma, Vatsala Anand, and Sheifali Gupta. Automobile price prediction using regression models. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 410–416, 2022.
3. Shengqiang Han, Jianhua Qu, Jinyi Song, and Zijing Liu. Second-hand car price prediction based on a mixed-weighted regression model. In *2022 7th International Conference on Big Data Analytics (ICBDA)*, pages 90–95, 2022.
4. Mustapha Hankar, Marouane Birjali, and Abderrahim Beni-Hssane. Used car price prediction using machine learning: A case study. In *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 1–4, 2022.
5. Chuyang Jin. Price prediction of used cars using machine learning. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, pages 223–230, 2021.
6. Annisaa Fauziyah Kinadi, Rachmadita Andreswari, Edi Sutoyo, Ramdhan Nugraha, and Anton Abdul Basah Kamil. Used car price prediction in surabaya using random forest regressor algorithms. In *2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, pages 1–4, 2022.
7. Sachin Kumar, Damandeep Kaur, and Anjum Parvez. Prediction of prices car price prediction with machne learning. In *2022 International Conference on Cyber Resilience (ICCR)*, pages 1–4, 2022.
8. Linyu Li and Zhihui Ye. Research on used car price prediction based on stacking model fusion. In *2022 International Conference on Informatics,Networking and Computing (ICINC)*, pages 86–90, 2022.
9. Yashi Li, Yuxuan Li, and Yuexi Liu. Research on used car price prediction based on random forest and lightgbm. In *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, pages 539–543, 2022.
10. Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. Prediction of pricesfor used car by using regression models. In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pages 115–119, 2018.
11. Chejarla Venkat Narayana, Chinta Lakshmi Likhitha, Syed Bademiya, and Karre Kusumanjali. Machine learning techniques to predict the priceof used cars: Predictive analytics in retail business. In *2021 Second In- ternational Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1680–1687, 2021.
12. Chejarla Venkata Narayana, Nukathoti Ooha Gnana Madhuri, Atmakuri NagaSindhu, Mulupuri Aksha, and Chalavadi Naveen. Second sale car price prediction using machine learning algorithm. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pages 1171–1177, 2022.
13. Punitha Ponmalar P and Angelin Christinal C. Review on the pre-owned car price determination using machine learning approaches. In *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pages 274–278, 2022.
14. Santosh Kumar Satapathy, Rutvikraj Vala, and Shiv Virpariya. An automated car price prediction system using effective machine learning techniques. In *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pages 402–408, 2022.
15. Samveg Shah, Mayur Telrandhe, Prathmesh Waghmode, and Sunil Ghane. Imputing missing values for dataset of used cars. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–5, 2022.
16. Ning Sun, Hongxi Bai, Yuxia Geng, and Huizhu 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/ DistributedComputing (SNPD)*, pages 431–436, 2017.
17. Radhika Swarnkar, Rhea Sawant, Harikrishnan R, and Srideviponmalar P. Multiple linear regression algorithm-based car price prediction. In *2023 Third International Conference on Artificial Intelligence and SmartEnergy (ICAIS)*, pages 675–681, 2023.
18. Janke Varshitha, K Jahnavi, and C. Lakshmi. Prediction of used car prices using artificial neural networks and machine learning. In *2022 International Conference on Computer Communication and Informatics(ICCCI)*, pages 1–4, 2022.
19. Feng Wang, Xusong Zhang, and Qiang Wang. Prediction of used car price based on supervised learning algorithm. In *2021 International Conference on Networking, Communications and Information Technology (NetCIT)*, pages 143–147, 2021.
20. Han Zhang. Prediction of used car price based on lightgbm. In *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, pages 327–332, 2022.