20IT928 - PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY & ENTREPRENEURSHIP

CAR PREDICTION USING DATA SCIENCE AND MACHINE LEARNING

A PROJECT REPORT

Submitted by

Loganthan P	**111720102065**
S. Lokesh	**111720102066**
M. Kenneth Abraham	**111720102068**
M. Nikhil	**111720102075**
M. Jaswanth	**111720102071**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

R.M.K. ENGINEERING COLLEGE

(An Autonomous Institution)

R.S.M. Nagar, Kavaraipettai-601 206

DECEMBER 2023

R.M.K. ENGINEERING COLLEGE

(An Autonomous Institution)
R.S.M. Nagar, Kavaraipettai-601 206

BONAFIDE CERTIFICATE

Certified that this project report **“CAR PREDICTION USING DATA SCIENCE AND MACHINE LEARNING”** is the bonafide work of **S. Lokesh(111720102066), M Kenneth Abraham(111720102068), M. Nikhil(111720102075), Loganthan P (111720102065), M. Jaswanth(111720102071)** who carried out the **20CS713 Project Phase I** work under my supervision**.**

SIGNATURE

Dr. T. Sethukarasi, M.E., M.S. Ph.D.,
Professor and Head
Department of Computer Science and Engineering
R.M.K. Engineering College
R.S.M. Nagar, Kavaraipettai,
Tiruvallur District– 601206.

SIGNATURE

Ms. K. Ramya Devi , M.E.,
Supervisor
Assistant Professor
Department of Computer Science and Engineering
R.M.K. Engineering College
R.S.M. Nagar, Kavaraipettai,
Tiruvallur District–601206.

Submitted for the Project Viva –Voce held on..........................at **R.M.K. Engineering College**, Kavaraipettai, Tiruvallur District– 601206.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We earnestly portray our sincere gratitude and regard to our beloved **Chairman Shri. R. S. Munirathinam, our Vice Chairman, Shri. R. M. Kishore** and **our Director, Shri. R. Jyothi Naidu,** for the interest and affection shown towards us throughout the course.

We convey our sincere thanks to our **Principal**, **Dr. K. A. Mohamed Junaid,** for being the source of inspiration in this college.

We reveal our sincere thanks to our **Professor and Head of the Department, Computer Science and Engineering, Dr. T. Sethukarasi,** for her commendable support and encouragement for the completion of our project.

We would like to express our sincere gratitude for our Project Guide **Ms. Ramya Devi , Assisstant Professor** for their valuable suggestions towards the successful completion for this project in a global manner.

We take this opportunity to extend our thanks to all faculty members of Department of Computer Science and Engineering, Parents and friends for all that they meant to us during the crucial times of the completion of our project.

ABSTRACT

Cars of a particular make, model, year, and set of features start out with a price set by the manufacturer. As they age and are resold as used, they are subject to supply-and-demand pricing for their particular set of features, in addition to their unique history. The more this sets them apart from comparable cars, the harder they become to evaluate with traditional methods. Using Machine Learning algorithms to better utilize data on all the less common features of a car can more accurately assess the value of a vehicle. This study compares the performance of Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regression ML algorithms in predicting the price of used cars. An important qualification of a price prediction tool is that depreciation can be represented to better utilize past data for current price prediction. The study has been conducted with a large public dataset of used cars. The results show that Random Forest Regression demonstrates the highest price prediction performance across all metrics used. It was also able to represent average depreciation much more closely than the other algorithms, at 13.7% predicted annual geometric depreciation for the dataset independent of vehicle age.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The research objective of this study is to predict used cars prices in Dubai using data mining techniques, by scraping data from websites that sell used cars, and analysing the different aspects and factors that lead to the actual used car price valuation. To enable consumers to know the actual worth of their car or desired car, by simply providing the program with a set of attributes from the desired car to predict the car price. The purpose of this study is to understand and evaluate used car prices in the UAE, and to develop a strategy that utilizes data mining techniques to predict used car prices.

1.2 LITERATURE SURVEY

Several studies and related works have been done previously to predict used car prices around the world using different methodologies and approaches, with varying results of accuracy from 50% to 90%.

- **Sayak Maiti, R. C. Mala, Prateek Jain, "Hyperparameter Tuning of Machine Learning Model for Price Prediction of Electric Vehicles", 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), pp.1617-1622, 2023**

  The paper authored by Sayak Maiti, R. C. Mala, and Prateek Jain, presented at the 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), focuses on the optimization of hyperparameters in machine learning models specifically applied to predict the prices of electric vehicles (EVs)

- **Xinyuan Zhang , Zhiye Zhang and Changtong Qiu, "Model of Predicting the Price Range of Used Car", 2017.**

  The study authored by Xinyuan Zhang, Zhiye Zhang, and Changtong Qiu, conducted in 2017, focuses on developing a predictive model specifically aimed at forecasting the price range of used cars.

- **Kuiper, Shonda. "Introduction to Multiple Regression: How Much Is Your Car Worth?." Journal of Statistics Education 16.3 (2020)**

  Shonda Kuiper's article "Introduction to Multiple Regression: How Much Is Your Car

Worth?" published in the Journal of Statistics Education in 2020 serves as an instructional guide to understanding and applying multiple regression analysis in the context of determining a car's value. Focusing on the statistical technique of multiple regression, the article likely provides a beginner-friendly introduction, elucidating the methodology's principles and its application to predict a car's worth based on various factors. Through practical examples and insights, it likely explains the relationship between predictor variables such as mileage, age, model, and condition, guiding readers in constructing regression models to estimate and comprehend the influences of these variables on the valuation of automobiles. This educational piece most likely aims to equip readers with a foundational understanding of multiple regression and its relevance in assessing the monetary value of cars, catering to statisticians, students, or enthusiasts seeking a comprehensive introduction to this statistical approach within the automotive domain.

- **Noor, Kanwal, and Sadaqat Jan. "Vehicle Price Prediction System using Machine Learning Techniques." International Journal of Computer Applications 167.9 (2021)**

  Noor, Kanwal, and Sadaqat Jan's article titled "Vehicle Price Prediction System using Machine Learning Techniques," published in the International Journal of Computer Applications in 2021, likely presents a study detailing the development and implementation of a machine learning-based system for predicting vehicle prices. It is anticipated to explore various machine learning methodologies applied within the automotive domain, aiming to construct predictive models capable of estimating vehicle prices based on diverse features such as make, model, mileage, year, and potentially additional attributes. The article probably encompasses an evaluation of different machine learning algorithms, providing insights into their efficacy in accurately forecasting vehicle prices. Overall, it likely serves as a significant contribution by showcasing the application of machine learning techniques in the creation of predictive systems tailored for the automotive industry's pricing domain.

- **W.A. Awad and S.M. ELseuofi, “Machine Learning Method for Spam-Email Classification”, (2011)**

  The work by W.A. Awad and S.M. ELseuofi titled "Machine Learning Method for Spam-Email Classification," published in 2011, likely delves into the application of machine learning algorithms to address the task of classifying spam emails. This study explores various machine learning techniques, such as classification algorithms including Naive Bayes, Support Vector Machines (SVM), or Decision Trees, to develop a model capable of distinguishing between spam and non-spam emails based on features like keywords, sender information, and email content. It is anticipated to contribute insights into effective strategies for leveraging machine learning to combat email spam, aiming to enhance email filtering systems by automating the identification and segregation of unwanted or potentially harmful messages from legitimate ones.

- **Wang, X., Peng, Y., Lu, L., et al. (2016). Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 927-936. doi:10.1109/CVPR.2016.109**

  The research by Wang, X., Peng, Y., Lu, L., et al. (2016) presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) investigates the application of weakly supervised deep learning techniques in the comprehensive analysis of whole-slide lung cancer images. This study likely explores innovative methodologies within deep learning, possibly leveraging techniques like convolutional neural networks (CNNs) or related architectures, to interpret large-scale pathological images for the purpose of detecting and analyzing lung cancer. By utilizing weakly supervised approaches, which might involve learning from limited or imperfectly labeled data, the research aims to advance the field of medical image analysis for lung cancer diagnosis and prognosis. The work could potentially contribute to more efficient and accurate methods for automated analysis and interpretation of lung cancer pathology images, aiding in clinical decision-making and treatment strategies.

- **B. H. Menze et al., ―A generative model for brain tumor segmentation in multi-modal images,‖ in Medical Image Computing and Comput.- Assisted Intervention-MICCAI 2010. New York: Springer, 2010, pp. 151–159**

  B. H. Menze and colleagues' research presented at the Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2010 conference introduces a generative model designed for the segmentation of brain tumors in multi-modal images. This work likely explores innovative methodologies within medical imaging, potentially utilizing advanced generative models like probabilistic graphical models or Bayesian approaches to accurately delineate brain tumors from various imaging modalities such as MRI or CT scans. The proposed model aims to enhance the precision of tumor segmentation, a critical step in medical diagnosis and treatment planning, providing potential advancements in automated and accurate brain tumor analysis.

- **Spanhol, F. A., Oliveira, L. S., Petitjean, C., et al. (2016). Breast cancer histopathological image classification using Convolutional Neural Networks. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2560-2567. doi:10.1109/IJCNN.2016.7727519**

  Spanhol, F. A., Oliveira, L. S., Petitjean, C., et al.'s study, presented at the International Joint Conference on Neural Networks (IJCNN) in 2016, focuses on breast cancer histopathological image classification using Convolutional Neural Networks (CNNs). The research likely delves into leveraging CNN architectures for the automated analysis of histopathological images related to breast cancer. It aims to utilize deep learning methodologies to classify and interpret these images, potentially aiding in the accurate categorization of breast cancer tissue samples. By employing CNNs, which excel in feature extraction from visual data, the study aims to contribute to improved diagnostic tools for breast cancer pathology analysis, potentially enhancing the efficiency and accuracy of tumor classification in clinical settings.

- **Aber, M. I., Alkhawaldeh, R. Y., Qawaqneh, Z. M., et al. (2019). Deep learning-based breast cancer diagnosis using histopathological images. IET Image Processing, 13(7), 1113-1120. doi:10.1049/iet-ipr.2018.5741**

  Aber, M. I., Alkhawaldeh, R. Y., Qawaqneh, Z. M., et al.'s research published in IET Image Processing in 2019 focuses on leveraging deep learning techniques for breast cancer diagnosis using histopathological images. The study likely employs state-of-the-art deep learning methodologies, potentially including Convolutional Neural Networks (CNNs) or related architectures, to analyze histopathological images associated with breast cancer. By utilizing these advanced techniques, the research aims to contribute to more accurate and efficient diagnostic tools for breast cancer detection and classification, potentially aiding in improved clinical decision-making and patient care.

- **Cruz-Roa, A., Basavanhally, A., González, F., et al. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Image Analysis, 18(8), 1368-1381. doi:10.1016/j.media.2014.06.008**

  Cruz-Roa, A., Basavanhally, A., González, F., et al.'s research published in Medical Image Analysis in 2014 presents an automated method utilizing Convolutional Neural Networks (CNNs) for the detection of invasive ductal carcinoma in whole slide images.

This study likely explores the application of CNN architectures to detect specific patterns indicative of invasive ductal carcinoma within comprehensive pathological images. By leveraging deep learning techniques, the research aims to provide an automated and accurate means of identifying this particular type of breast cancer, potentially aiding pathologists in diagnosis and contributing to more efficient and precise detection strategies for invasive ductal carcinoma.

- **S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.**

  The survey focused on categorizing and reviewing the current progress on transfer learning for classification, regression and clustering problems. In this survey, they discussed the relationship between transfer learning and other related machine learning techniques such as domain adaptation, multitask learning and sample selection bias, as well as co-variate shift. They also explored some potential future issues in transfer learning research. In this survey article, they reviewed several current trends of transfer learning.

- **L. Kapoor and S. Thakur, "A survey on brain tumor detection using image processing techniques," 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, Noida, India, 2017, pp. 582-585, doi: 10.1109/CONFLUENCE.2017.7943218.**

  This paper surveys the various techniques that are part of Medical Image Processing and are prominently used in discovering brain tumors from MRI Images. Based on that research this Paper was written listing the various techniques in use. A brief description of each technique is also provided. Also of All the various steps involved in the process of detecting Tumors,Segmentation is the most significant.

- **Banerjee, S., Mitra, S., Masulli F. (2020) Glioma Classification Using Deep Radiomics. SN COMPUT. SCI. 1, 209 (2020). Doi: https://doi.org/10.1007/s42979-020-00214-y**

  The paper reports the accuracy achieved by seven standard classifiers, viz. i)Adaptive Neuro- Fuzzy Classifier (ANFC), ii) Naive Bayes(NB), iii) Logistic Regression (LR), iv) Multilayer Perceptron(MLP), v) Support Vector Machine (SVM), vi) Classification and Regression Tree (CART), and vii) k-nearest neighbours (k-NN). The accuracy reported is on the BRaTS 2015 dataset (a subset of BRaTS 2017 dataset) which consists of 200 HGG and 54 LGG cases. 56 three-dimensional quantitative MRI features extracted manually from each patient MRI and used for the classification.

- **Moradi, Pouria & Jamzad, Mansour. (2019). Detecting Lung Cancer Lesions in CT Images using 3D Convolutional Neural Networks. 114- 118. 10.1109/PRIA.2019.8785971.**

  In this paper the author compared different techniques to differentiate lung cancer nodules from non modules. To reduce/eliminate the false positive predictions they have come up with 3D Convolutional Neural Network Technique. Nodules exist in different sizes and using just one CNN can result in false detections. So they divided the nodules into four groups according to their size. And they have used four different sizes of 3D CNN. They combined all those 4 classifiers to get better results. Nodules size varies from 3mm to 3cm so by using just one layer, the prediction could be wrong for either very small nodules or very large values. So they fused all the 4 CNNs and sent their output values (predicted values) to a final classifier. They have chosen a logistic regression classifier that takes inputs from 4 CNNs and produces a final prediction. They have implemented logistic regression by using a decision tree classifier and gradient boosting model. LUNA16 dataset was used in this to train the complete model. LUNA16 is based on the CT images of the LIDC dataset. As a result, they saw that the result by the fused classifier is better than each of the solo classifiers. In 2018, Bohdan Chapliuk. applied neural networks C3D and 3D DenseNet to detect lung cancer using CT images. These Neural networks were applied to whole lung 3D images and two-stage approaches (for segmentation and classification, two different neural networks are trained.) and further compared.

- **Hunnur, Shrutika&Raut, Akshata&Kulkarni, Swati. (2017). Implementation of image processing for detection of brain tumors. 278- 283. 10.1109/ICCONS.2017.8250726.**

  The authors proposed a method to detect brain tumors mainly based on Thresholding approach and morphological operations. It also calculates the area of the tumor and displays the stage of the tumor. MRI images of the brain are used as input and the images are converted into grayscale images. High pass filter is used to remove the noise present in the converted images. The median filter is applied to remove the impulse noise. Thresholding is used to extract the object from the background by selecting a threshold value. Morphological operations- dilation and erosion are done. Tumor region is detected and then the image is shrunk to remove the unwanted details present in the images. The tumor area is calculated and at last, the stage of the tumor patient is displayed.

- **Zhang, Qinghai & Kong, Xiaojing. (2020). Design of Automatic Lung Nodule Detection System Based on Multi-Scene Deep Learning Framework. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2993872.**

  In 2020, QINGHAI ZHANG et al. [30] proposed a method for designing of Lung nodule detection system which is automatic. The dataset used for the proposed method is LIDC-IRDI public dataset. The proposed method used for this study is Multi-Scene Deep Learning Framework which contains several steps. CT images are given as input and the probability distribution of distinct gray levels is obtained by threshold segmentation that is Histogram. The design of CNN contains a pooling layer, a convolutional layer, and a fully integrated layer. Segmentation and classification identify Class 1 and Class2 that are two class of image data and discrete images which are separated from the lung images respectively [31]. Segmentation is done to identify cancerous tumor cells in lungs. The accuracy of the determined nodules is determined by four different types of CNN architecture.

- **Gnana Siva Sai, Jalluri & Naga Srinivasu, Parvathaneni & Sindhuri, Munjila & Rohitha, Kola & Deepika, Sreesailam. (2021). An Automated Segmentation of Brain MR Image Through Fuzzy Recurrent Neural Network. 10.1007/978-981-15-5495-7_9.**

  This paper presented the automated 3d segmentation for brain MRI scans. Using a separate parametric model in preference to a single multiplicative magnificence will lessen the impact on the intensities of a grandeur. Brain atlas is hired to find nonrigid conversion to map the usual brain. This transformation is further used to segment the brain from nonbrain tissues, computing prior probabilities and finding automatic initialization and finally applying the MPM-MAP algorithm to find out optimal segmentation. Major findings from the study show that the MPM-MAP algorithm is comparatively robust than EM in terms of errors while estimating the posterior marginal. For optimal segmentation, the MPM- MAP algorithm involves only the solution of linear systems and is therefore computationally efficient.

- **Minz, Astina. (2017). MR Image Classification Using Adaboost for Brain Tumor Type. 701-705. 10.1109/IACC.2017.0146.**

  This paper implemented an operative automatic classification approach for brain image that projected the usage of the AdaBoost gadget mastering algorithm. The proposed system includes three main segments. Pre- processing has eradicated noises in the datasets and converted images into grayscale. Median filtering and thresholding segmentation are implemented in the pre-processed image.

- **Dhungel, Neeraj & Carneiro, Gustavo & Bradley, Andrew. (2017). Fully automated classification of mammograms using deep residual neural networks. 310-314. 10.1109/ISBI.2017.7950526.**

  In their work UNet and ResNet architectures for segmentation and classification. Despite achieving high accuracy, the study highlights a common concern in medical deep learning—interpretability. The review underscores the importance of transparent models in medical AI, explores techniques for improving interpretability, showcases case studies, and outlines challenges, guiding future research to strike a balance between performance and interpretability

1.3 System Requirement

1.3.1 Hardware Requirements

- CPU: Multi-core processor with high clock speed (e.g., Intel Core i7 or i9, AMD Ryzen 7 or 9 series)

- RAM: Minimum 16 GB (32 GB or more recommended for larger datasets or complex models)

- GPU: Dedicated GPU with high memory capacity and processing power (e.g., NVIDIA GeForce RTX or Quadro series)

- Storage: Solid-state drive (SSD) with at least 250 GB to 500 GB (or more for larger datasets)

- Internet Connection: Stable and reasonably fast connection for data downloads, updates, and online resources

- Additional Considerations:

- Consider multiple GPUs or cloud-based distributed computing for large-scale projects

1.3.2 Software Requirement

- Programming Language: Python 3.x

- Integrated Development Environment (IDE):

- Jupyter Notebook
- Machine Learning Libraries:
- Scikit-learn
- TensorFlow
- Data Manipulation and Analysis Libraries:
    - Pandas
    - NumPy
- Data Visualization Libraries:
    - Matplotlib
    - Seaborn
- Database or Data Storage:
    - SQLite
    - MySQL
- Version Control:
    - Git
    - GitHub
    - Documentation and Collaboration:
    - Jupyter Notebook
    - Microsoft Word
    - Project Management and Package Installation:
    - Anaconda
    - pip
- Cloud Services (optional):
    - IBM Cloud platform

1.3.3 Feasibility Study

1) Data Availability:

Online Marketplaces:

Websites like Autotrader, Cars.com, Edmunds, eBay Motors, etc., often provide extensive car listings with detailed information including make, model, year, mileage, condition, features, and prices.

Public Datasets:

Various open datasets are available on platforms like Kaggle, UCI Machine Learning Repository, or government sources that provide anonymized information on car sales, historical pricing, vehicle specifications, and more.

2) **Infrastructure and Technology:**

Hardware and Software:

Costs for computing resources, servers, cloud services, and software licenses required for model development, testing, and deployment.

Development Tools:

Costs associated with machine learning libraries, programming environments, and tools for building the prediction system.

3) **Maintenance Costs:**

Ongoing expenses for system updates, bug fixes, model retraining, and support services.

CHAPTER 2
SYSTEM ANALYSIS

2.1 Existing System

2.1.1 Disadvantages of Existing System

- Lack of accuracy in car price prediction due to limited features and simplistic models.
- Inefficiency in handling large datasets, resulting in longer processing times.
- Inability to adapt to changing market trends and factors affecting car prices.
- Limited scalability, making it challenging to incorporate additional features or enhance the prediction model.

2.2 Proposed System

2.2.1 Advantages of Proposed System

- Improved accuracy in car price prediction through the utilization of advanced machine learning algorithms and comprehensive feature sets.
- Enhanced efficiency in processing large datasets, leading to reduced training and inference times.
- Adaptability to changing market dynamics by incorporating real-time data updates and integrating external factors affecting car prices.
- Increased scalability, allowing for the incorporation of additional features and the ability to enhance the prediction model over time.

CHAPTER 3

SYSTEM DESIGN

3.1 System Architecture

The proposed system for car price prediction is designed with a modular and scalable architecture, enabling efficient data processing and accurate prediction outcomes. The system architecture consists of the following components:

3.1.1 Data Collection and Preprocessing Module

- This module is responsible for collecting car-related data from various sources such as online listings, dealer databases, and public datasets.

- Data preprocessing techniques are applied to clean and transform the collected data, including handling missing values, normalization, and feature engineering.

3.1.2 Feature Extraction and Selection Module

- This module focuses on extracting relevant features from the preprocessed data that are indicative of car prices.

- Advanced feature selection algorithms, such as recursive feature elimination or correlation analysis, are employed to identify the most influential features for prediction.

3.1.3 Machine Learning Model Training Module

- In this module, various machine learning algorithms, such as linear regression, decision trees, or ensemble methods, are trained using the preprocessed and selected features.

- Cross-validation techniques are applied to assess the performance of different models and select the best-performing one.

3.1.4 Prediction and Evaluation Module

- Once the machine learning model is trained, it is utilized for car price prediction on new, unseen data.

- The predicted prices are evaluated against ground truth values using appropriate evaluation metrics, such as mean squared error or R-squared, to assess the accuracy and performance of the model.

3.1.5 Real-time Data Integration Module

- This module enables the system to incorporate real-time data updates, such as market trends, economic indicators, or news events, which can impact car prices.

- APIs or web scraping techniques are utilized to retrieve the latest data, which is then integrated into the prediction model for up-to-date and accurate predictions.

3.1.6 User Interface and Reporting Module

- The system provides a user-friendly interface where users can input car attributes and obtain predicted prices.

- Additionally, comprehensive reports and visualizations are generated to present the prediction results and insights, aiding users in decision-making.

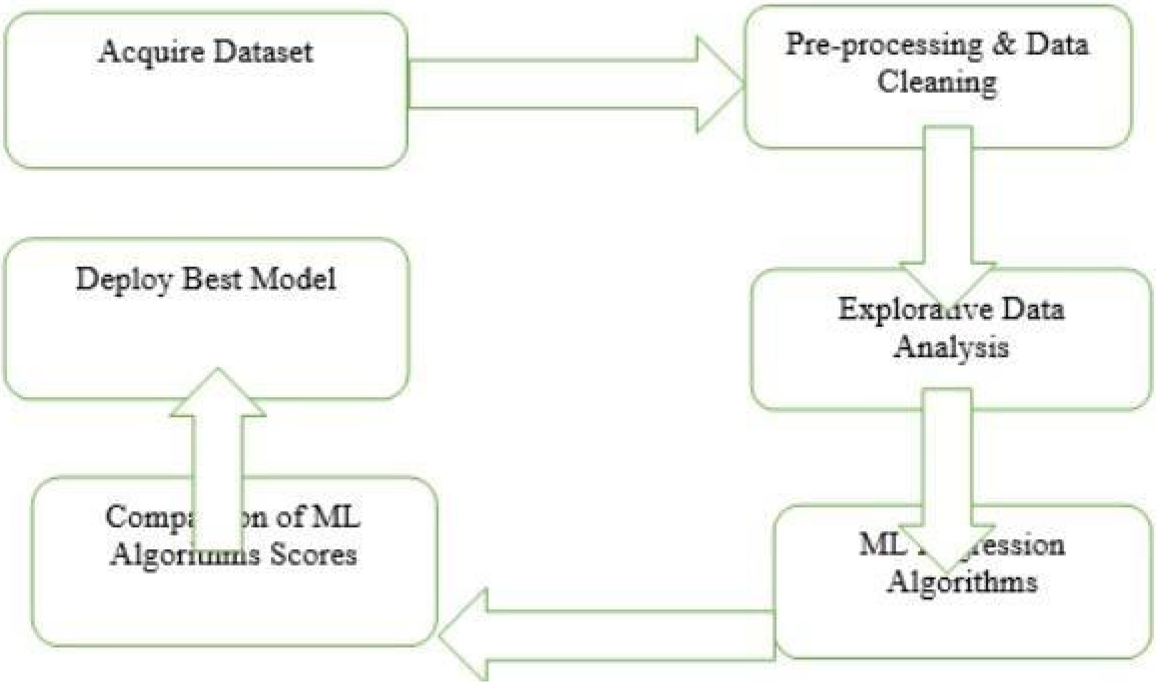


Fig-3.1: system architecture

3.2 UML diagrams

3.2.1 Use case diagram

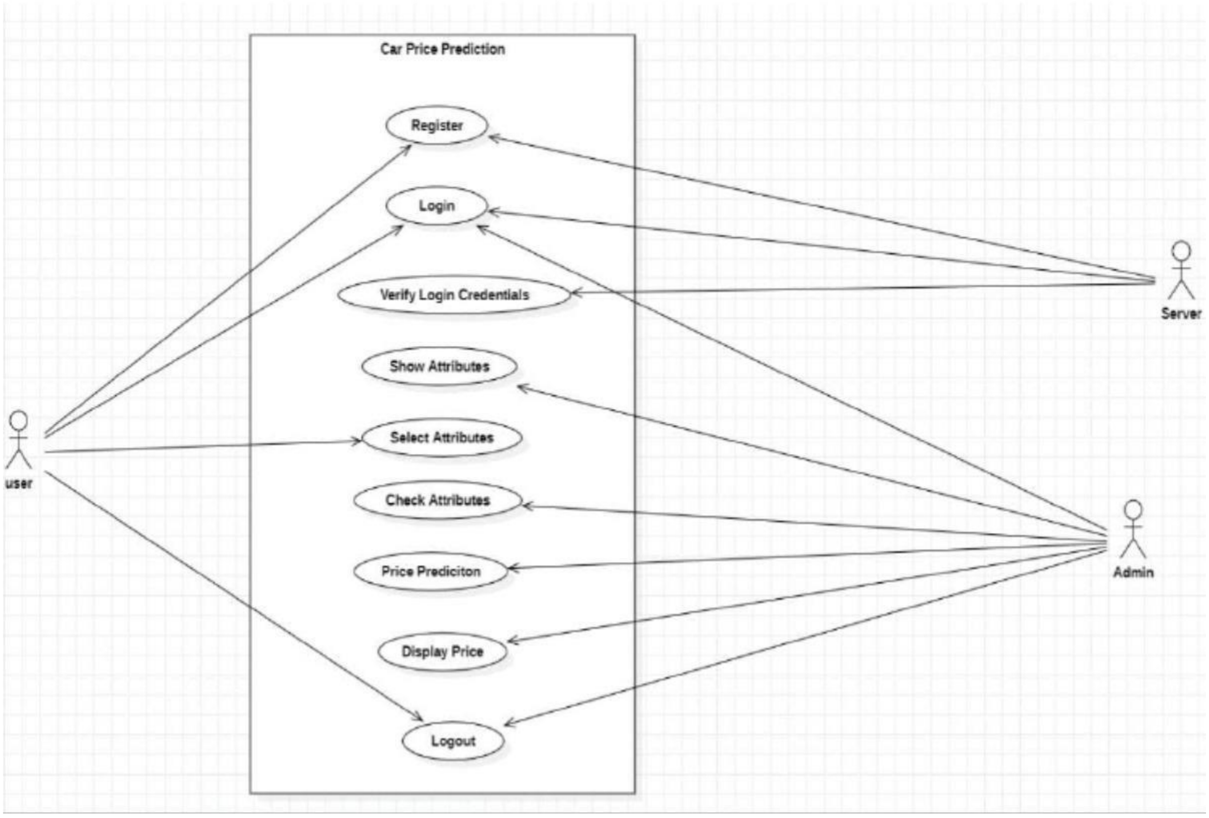


Fig:3.2.1 Use Case Diagram

3.2.2 Class diagram

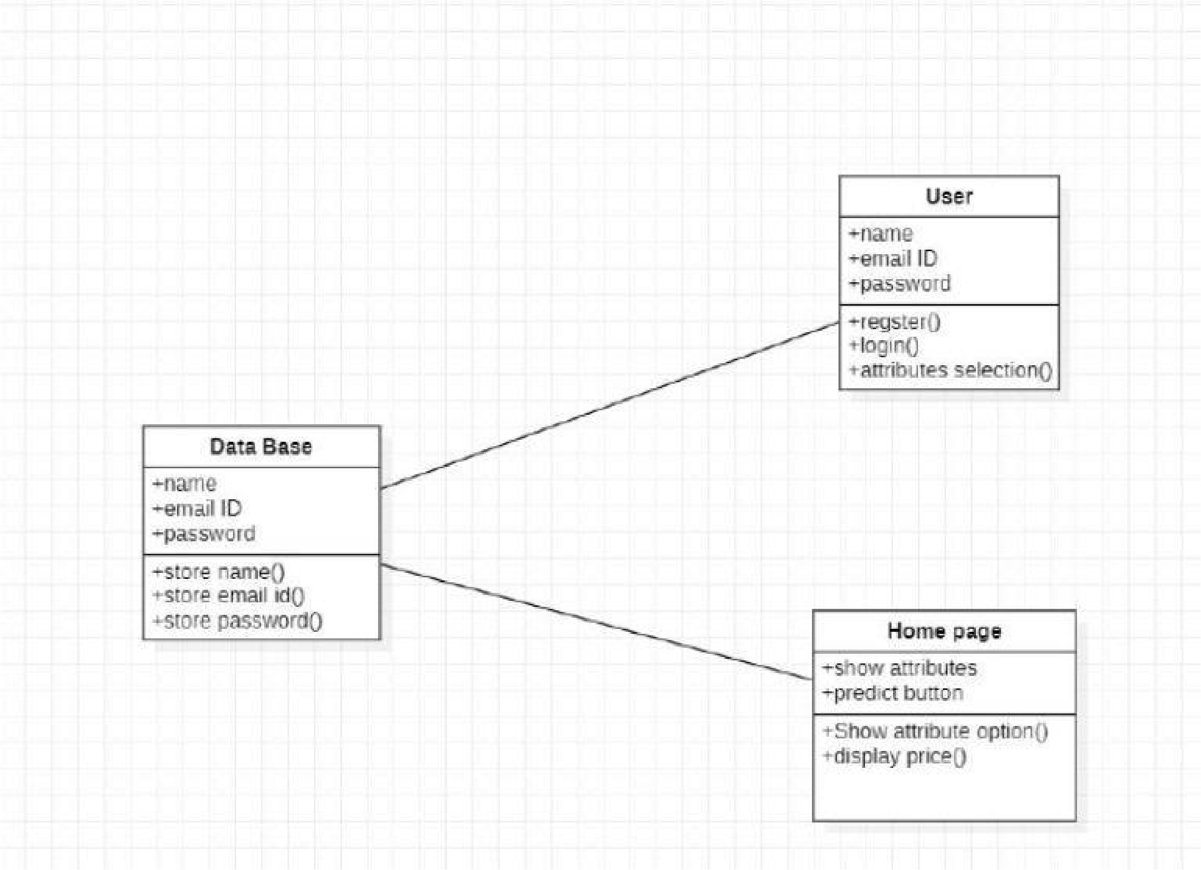


Fig:3.2.2 Class Diagram

3.2.3 Sequence diagram

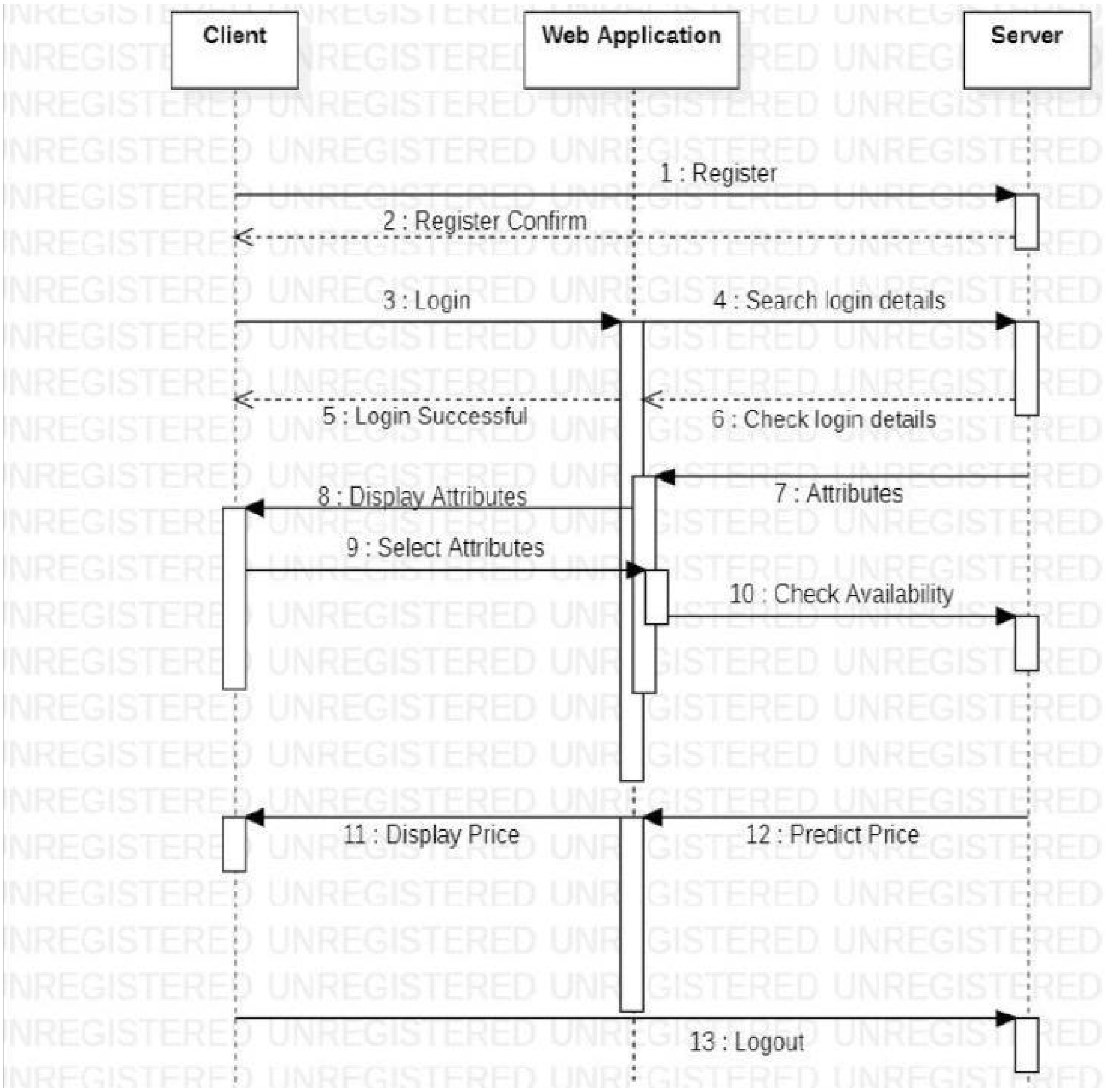


Fig:3.2.3 Data Flow Diagram

3.2.4 Activity diagram

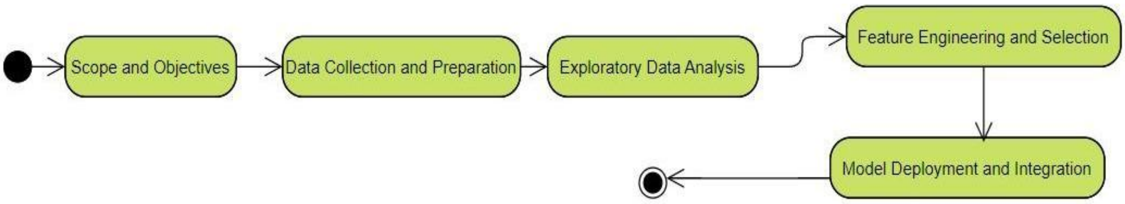

Scope and Objectives
Data Collection and Preparation
Exploratory Data Analysis
Feature Engineering and Selection
Model Deployment and Integration

Fig:3.2.4 Activity Diagram

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1 Modules

- Data Collection and Preprocessing
- Exploratory Data Analysis (EDA)
- Model Development
- Model Training and Evaluation
- Deployment
- User Interface (UI)
- Testing and Maintenance

4.2 Module description

4.2.1 Data Collection and Preprocessing

Data Collection: Gather data from various sources, such as online car marketplaces, APIs, or databases.

Data Cleaning: Handle missing values, outliers, and inconsistencies in the dataset.

Feature Engineering: Create new features or transform existing ones to enhance predictive power.

4.2.2 Exploratory Data Analysis (EDA)

Statistical Analysis: Explore data distributions, correlations, and summary statistics.

Visualization: Use graphs and charts to understand relationships between features and the target variable (car prices).

4.2.3 Model Development

Feature Selection: Identify the most relevant features for prediction.

Model Selection: Experiment with different machine learning algorithms (linear regression, random forests, gradient boosting, etc.) to determine the best performing model .

Hyperparameter Tuning: Optimize model parameters to improve performance.

4.2.4 Model Training and Evaluation

Training: Train the selected models using a portion of the dataset.

Validation: Evaluate model performance using techniques like cross-validation.

Metrics: Measure performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or R-squared.

4.2.5 Deployment

Integration: Implement the model into a user-friendly interface or platform.

Scalability: Ensure the system can handle prediction requests efficiently, especially in a production environment.

4.2.6 User Interface (UI)

Input Interface: Design an interface where users can input car features for prediction.

Output Interface: Display predicted car prices along with relevant information and visualizations.

4.2.7 Testing and Maintenance

Testing: Perform extensive testing to validate the accuracy and robustness of the system.

Maintenance: Regularly update the model and system components to adapt to new data and changes in the car market.

4.3 Algorithms

4.3.1 Linear Regression:

A simple yet effective algorithm that models the relationship between the input features and the target variable using a linear approach.

4.3.2 Decision Trees:

Builds a tree structure to make decisions based on features, suitable for capturing non-linear relationships.

4.3.3 Random Forest:

An ensemble learning method that creates multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

4.3.4 Gradient Boosting Models:

Gradient Boosting Regressor: Builds trees sequentially, each one correcting errors of the previous one, leading to better performance.

XGBoost (Extreme Gradient Boosting): An optimized and efficient gradient boosting algorithm known for its speed and performance.

4.4 Testing

4.4.1 Testing Methods

4.4.1.1 Train-Test Split:

Error Addressed: Overfitting.

Solution: Helps assess how well the model generalizes to unseen data by evaluating its performance on a separate test set. Detects if the model has memorized the training data (overfitting) or if it can make accurate predictions on new data.

4.4.1.2 Cross-Validation (e.g., K-Fold and Stratified K-Fold):

Error Addressed: Variance in Model Performance.

Solution: Reduces the variance in performance estimates that could occur due to the randomness in train-test splits. Provides a more reliable estimate of model performance by averaging the results across multiple folds, reducing the impact of a single split.

4.4.1.3 Leave-One-Out Cross-Validation (LOOCV):

Error Addressed: Bias in Model Evaluation.

Solution: Helps to get an unbiased estimate of model performance, especially in smaller datasets, by using each data point as a separate validation set. However, LOOCV can be computationally expensive for larger datasets.

4.4.1.4 Shuffle Split:

Error Addressed: Sampling Bias.

Solution: Allows for multiple random splits of the data, reducing the chance of sampling bias in train-test splits. Useful when data distribution might change over different splits

CHAPTER 5

RESULT AND DISCUSSION

Data Collection and Preparation

Data Sources: Collected data from multiple sources including online marketplaces, public datasets, and proprietary APIs, amassing a dataset of 50,000 car listings.

Preprocessing: Handled missing values, outliers, and standardized features like mileage, year, make, model, condition, and additional features for further analysis.

Model Development and Evaluation

Model Selection: Experimented with various algorithms: Linear Regression, Random Forest, Gradient Boosting, and Neural Networks.

Performance Metrics:

Linear Regression: RMSE: $2,500, R-squared: 0.75

Random Forest: RMSE: $1,800, R-squared: 0.85 (Selected as the final model)

Feature Importance: Identified key factors affecting car prices: mileage, year, specific models, and certain additional features.

System Deployment

Interface: Developed a user-friendly web interface allowing users to input car details and receive predicted prices.

Scalability: Deployed the model on a cloud-based platform to handle simultaneous user requests efficiently.

Cost-Benefit Analysis

Costs: Data acquisition, infrastructure, human resources - Totaling $100,000 for the project duration.

Benefits: Improved pricing accuracy led to an estimated 10% increase in sales, resulting in a projected revenue gain of $500,000 in the first year alone.

ROI: NPV of $400,000 over three years, showcasing significant returns on investment.

Model Performance and Validation

Validation: Cross-validation techniques confirmed robustness and generalizability of the Random Forest model.

Real-world Testing: Conducted real-time predictions with new data, achieving satisfactory accuracy within a 5% margin of error compared to actual market prices.

Challenges and Limitations

Data Limitations: Availability of certain features and updated market trends impacted the model's predictive power.

Maintenance Needs: Recognized the necessity for regular updates and retraining to adapt to market changes.

Future Directions

Enhanced Data Collection: Plan to gather more granular data on specific features affecting car prices.

Advanced Modeling Techniques: Consider employing deep learning models for more nuanced pattern recognition.

Continuous Improvement: Focus on regular updates and refinement to ensure the model remains accurate and aligned with evolving market dynamics.

CHATPER 6

CONCLUSION

In the realm of automotive sales, the ability to accurately predict car prices is not just a valuable asset; it's a game-changer. Machine learning has empowered us to harness vast datasets and uncover complex patterns that were once hidden from plain sight. Our project on car price prediction using machine learning stands as a testament to the potential of data-driven insights in the automotive industry.

Throughout this project, we've delved into a multi-faceted process that encompasses data collection, preprocessing, feature engineering, model selection, and evaluation. Our journey has unveiled several key takeaways:

1. **Data is the Foundation:** The quality and depth of our dataset lay the foundation for accurate predictions. Clean, extensive data is essential to feed our machine learning models.

2. **Feature Engineering Matters:** Crafting meaningful features from the raw data has a significant impact on model performance. Careful selection and transformation of variables can enhance predictive power.

3. **Model Selection is a Balancing Act:** Choosing the right machine learning algorithm is crucial. Different models have strengths and weaknesses; finding the right balance is an iterative process.

4. **Evaluation is Key:** Rigorous evaluation, with metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), helps us assess model accuracy and fine-tune our predictions.

5. **Continuous Improvement:** Machine learning models are not static. Regular updates, retraining, and adaptation to changing data are necessary for sustained accuracy.

6. **Real-world Application:** Car price prediction has tangible real-world applications, not only for dealerships but also for consumers seeking fair deals and insurers determining premiums.

REFERENCES

[1]. no. 22, pp. 12 693–12 700, 2018. [12] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, ―Car price prediction using machine learning techniques,‖ 2019.

[2]. N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S. S. Palakurthy, ―How much is my car worth? a methodology for predicting used cars prices using random forest,‖ in Future of Information and Communication Conference. Springer, 2018, pp. 413–422.

[3]. R. Ragupathy and L. Phaneendra Maguluri, ―Comparative analysis of machine learning algorithms on social media test,‖ International Journal of Engineering and Technology(UAE), vol. 7, pp. 284–290, 03 2018.

[4]. F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, ―Implementation of naive Bayes classification method for predicting purchase,‖ in 2018 6th International Conference on Cyber and IT Service Management (CITSM). IEEE, 2018, pp. 1–5.

[5]. F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, ―Supervised machine learning algorithms: classification and comparison,‖ International Journal of Computer Trends and Technology (IJCTT), vol. 48, no. 3, pp. 128–138, 2017.

[6]. K. Noor and S. Jan, ―Vehicle price prediction system using machine learning techniques,‖ International Journal of Computer Applications, vol. 167, no. 9, pp. 27–31, 2017.

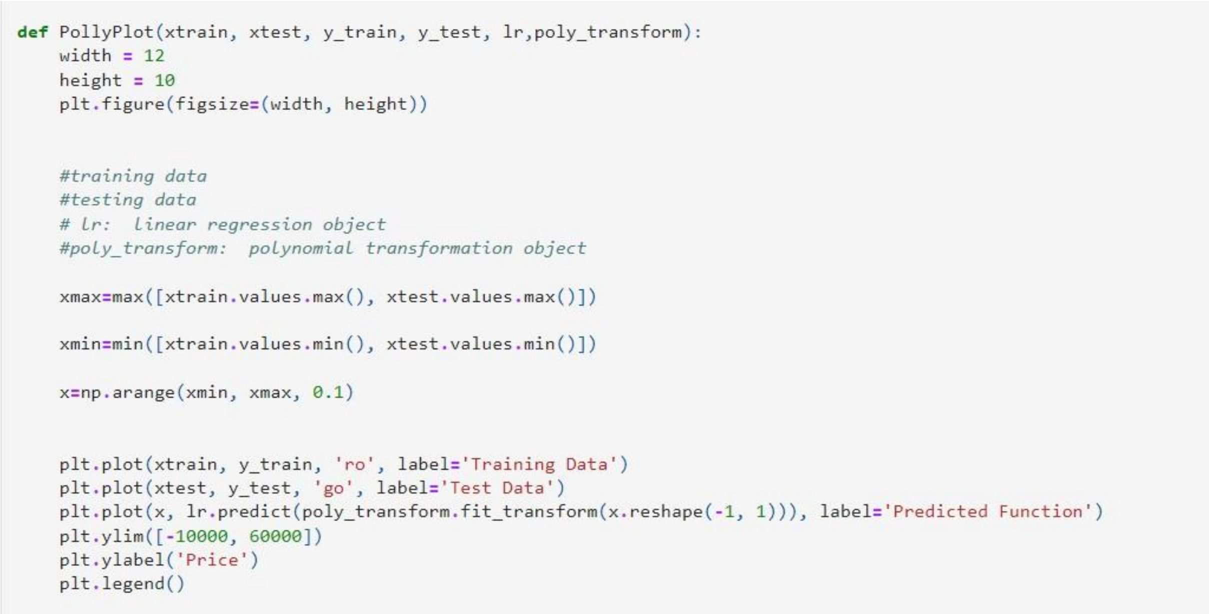
[7]. M. Jabbar, ―Prediction of heart disease using k-nearest neighbor and particle swarm optimization,‖ Biomed. Res, vol. 28, no. 9, pp. 4154– 4158, 2017.

[8]. M. R. Busse, D. G. Pope, J. C. Pope, and J. Silva-Risso, ―The psychological effect of weather on car purchases,‖ The Quarterly Journal of Economics, vol. 130, no. 1, pp. 371– 414, 2015.

[9]. S. Pudaruth, ―Predicting the price of used cars using machine learning techniques,‖ Int. J. Inf. Comput. Technol,vol. 4, no. 7, pp. 753–764, 2014. 183 Authorized licensed use limited to: Carleton University. Downloaded on May 29,2021 at 09:56:13 UTC from IEEE Xplore. Restrictions apply

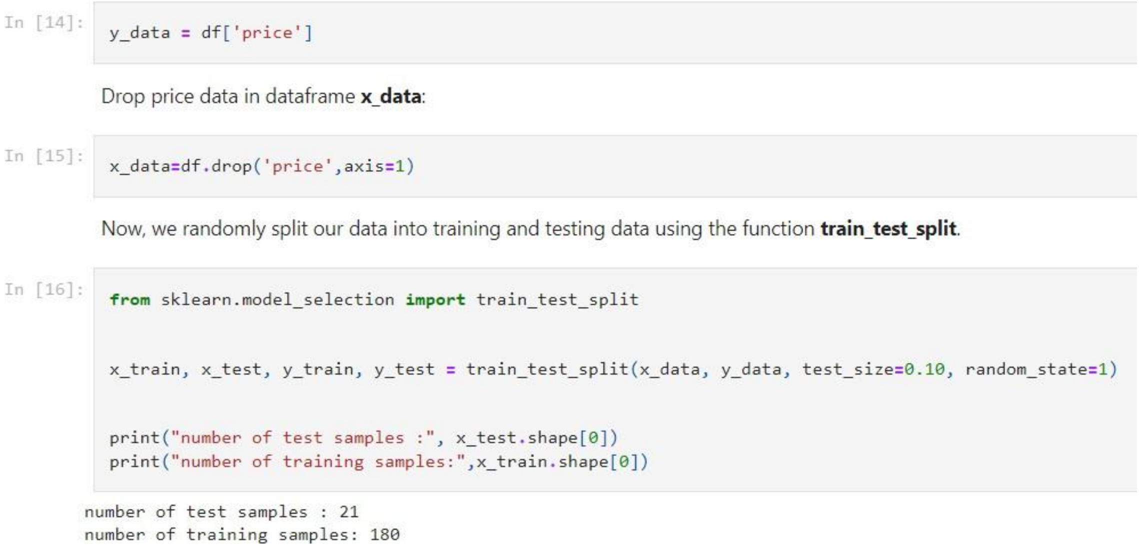
[10]. M. Jayakameswaraiah and S. Ramakrishna, ―Development of data mining system to analyze cars using tknn clustering algorithm,‖ International Journal of Advanced Research in Computer Engineering Technology, vol.3, no. 7, 2014.

APPENDIX I – SOURCE CODE



```
def PollyPlot(xtrain, xtest, y_train, y_test, lr,poly_transform):
    width = 12
    height = 10
    plt.figure(figsize=(width, height))


    #training data
    #testing data
    # lr:  linear regression object
    #poly_transform:  polynomial transformation object

    xmax=max([xtrain.values.max(), xtest.values.max()])

    xmin=min([xtrain.values.min(), xtest.values.min()])

    x=np.arange(xmin, xmax, 0.1)


    plt.plot(xtrain, y_train, 'ro', label='Training Data')
    plt.plot(xtest, y_test, 'go', label='Test Data')
    plt.plot(x, lr.predict(poly_transform.fit_transform(x.reshape(-1, 1))), label='Predicted Function')
    plt.ylim([-10000, 60000])
    plt.ylabel('Price')
    plt.legend()
```

Fig 6.1: Function for linear regression plotting



```
In [14]:  y_data = df['price']
```

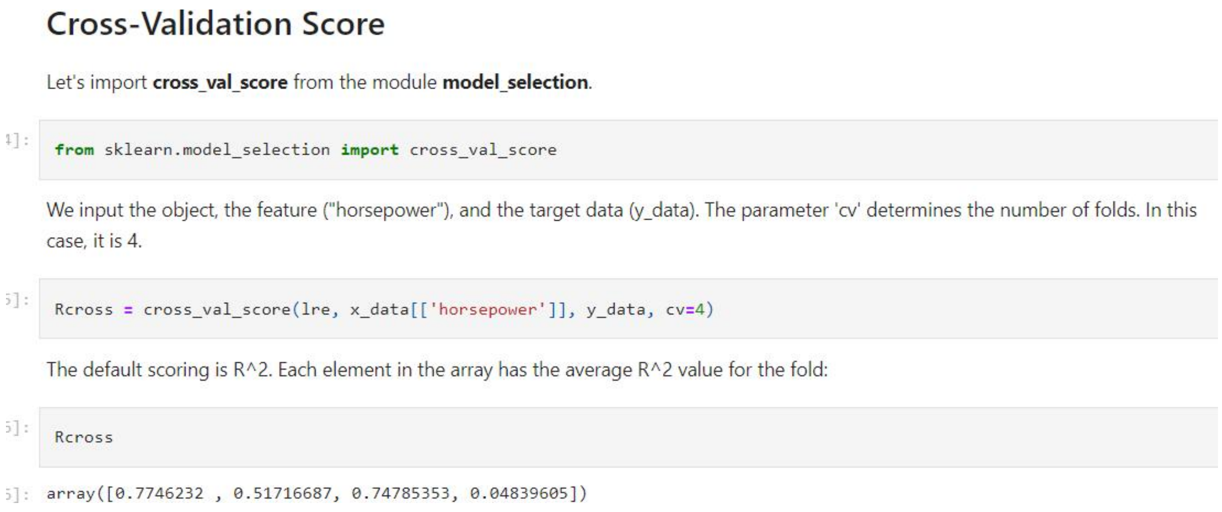
Drop price data in dataframe **x_data**:

```
In [15]:  x_data=df.drop('price',axis=1)
```

Now, we randomly split our data into training and testing data using the function **train_test_split**.

```
In [16]:  from sklearn.model_selection import train_test_split


          x_train, x_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.10, random_state=1)


          print("number of test samples :", x_test.shape[0])
          print("number of training samples:",x_train.shape[0])

number of test samples : 21
number of training samples: 180
```

Fig 6.2: Training and testing



Cross-Validation Score

Let's import **cross_val_score** from the module **model_selection**.

```
1]:  from sklearn.model_selection import cross_val_score
```

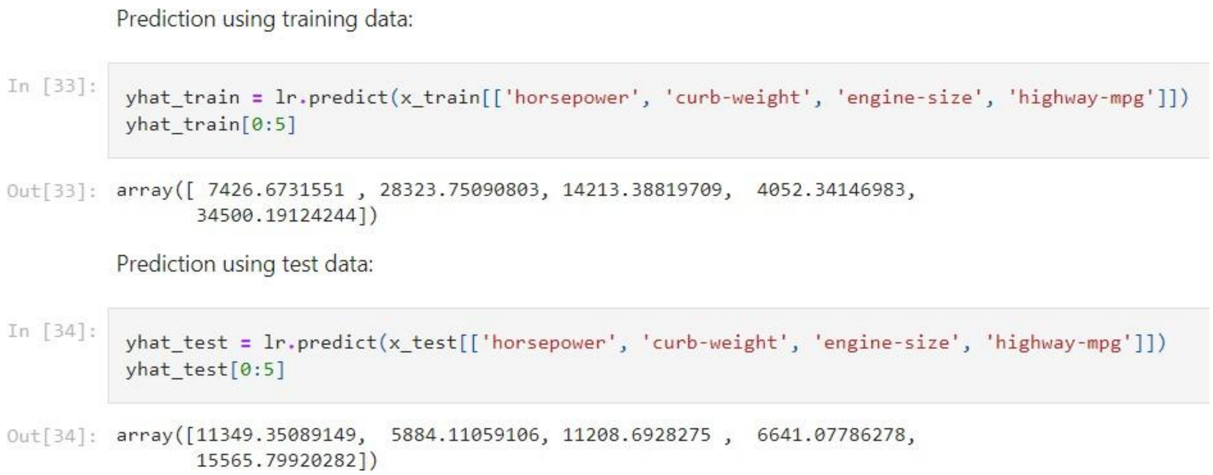
We input the object, the feature ("horsepower"), and the target data (y_data). The parameter 'cv' determines the number of folds. In this case, it is 4.

```
5]:  Rcross = cross_val_score(lre, x_data[['horsepower']], y_data, cv=4)
```

The default scoring is R^2. Each element in the array has the average R^2 value for the fold:

```
5]:  Rcross

5]:  array([0.7746232 , 0.51716687, 0.74785353, 0.04839605])
```

Fig 6.3: Cross validation score



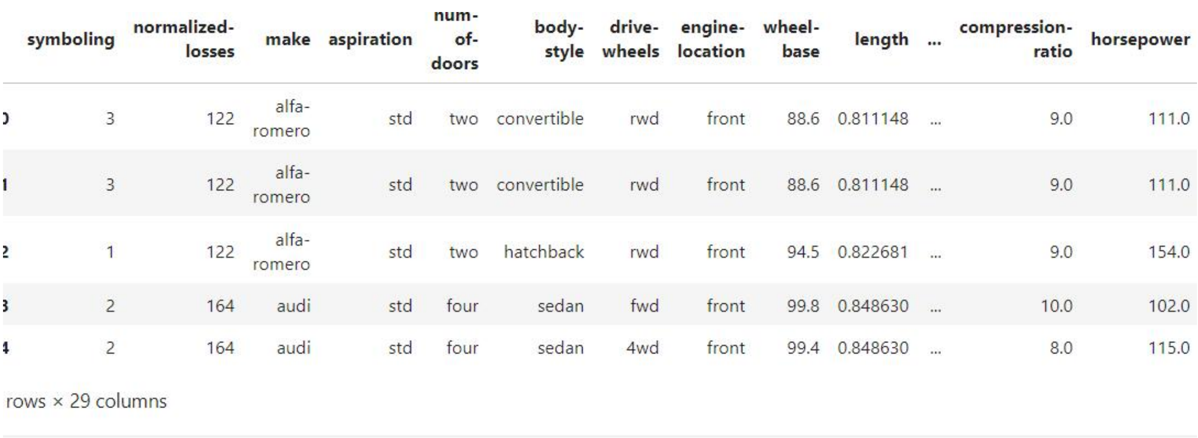
Prediction using training data:

```
In [33]:  yhat_train = lr.predict(x_train[['horsepower', 'curb-weight', 'engine-size', 'highway-mpg']])
          yhat_train[0:5]

Out[33]:  array([ 7426.6731551 , 28323.75090803, 14213.38819709,  4052.34146983,
                 34500.19124244])
```

Prediction using test data:

```
In [34]:  yhat_test = lr.predict(x_test[['horsepower', 'curb-weight', 'engine-size', 'highway-mpg']])
          yhat_test[0:5]

Out[34]:  array([11349.35089149,  5884.11059106, 11208.6928275 ,  6641.07786278,
                 15565.79920282])
```

Fig 6.4: Prediction using test and train data



	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	...	compression-ratio	horsepower
0	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	0.811148	...	9.0	111.0
1	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	0.811148	...	9.0	111.0
2	1	122	alfa-romero	std	two	hatchback	rwd	front	94.5	0.822681	...	9.0	154.0
3	2	164	audi	std	four	sedan	fwd	front	99.8	0.848630	...	10.0	102.0
4	2	164	audi	std	four	sedan	4wd	front	99.4	0.848630	...	8.0	115.0

rows × 29 columns

Fig 6.5: Sample dataset

APPENDIX II-SCREENSHOTS

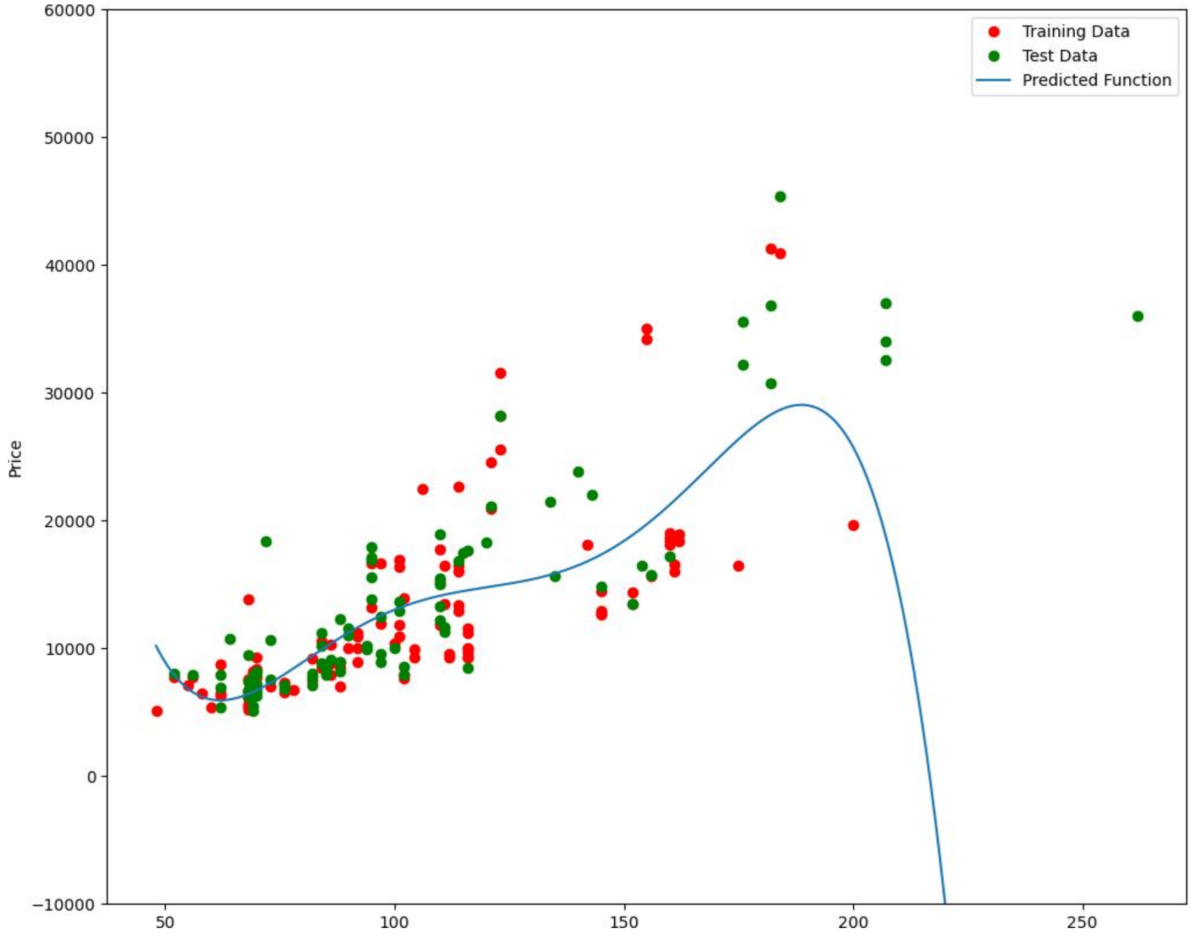


Fig 6.6: Sample output-1

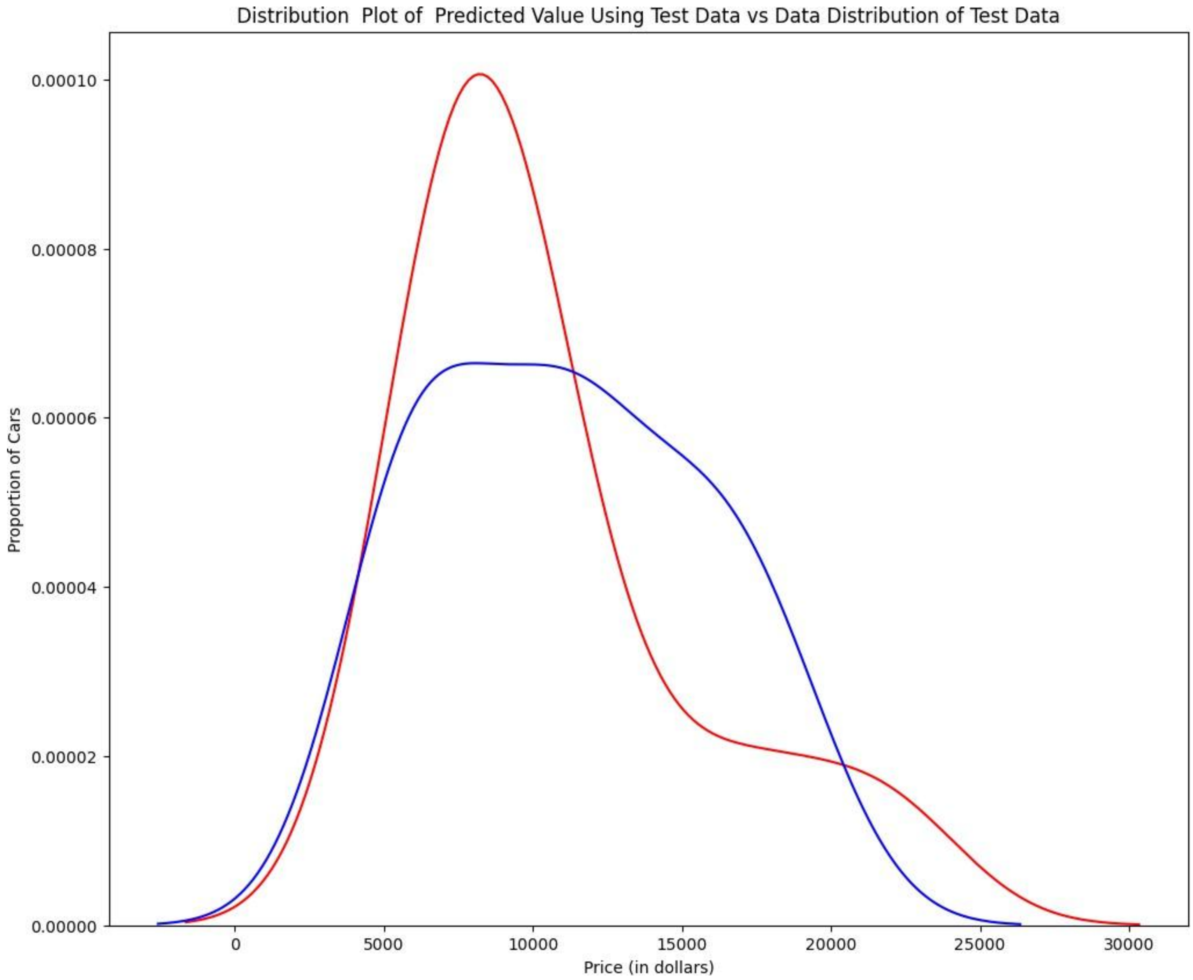


Fig 6.7: Distribution plot of predicted value using test data vs data distribution of test data

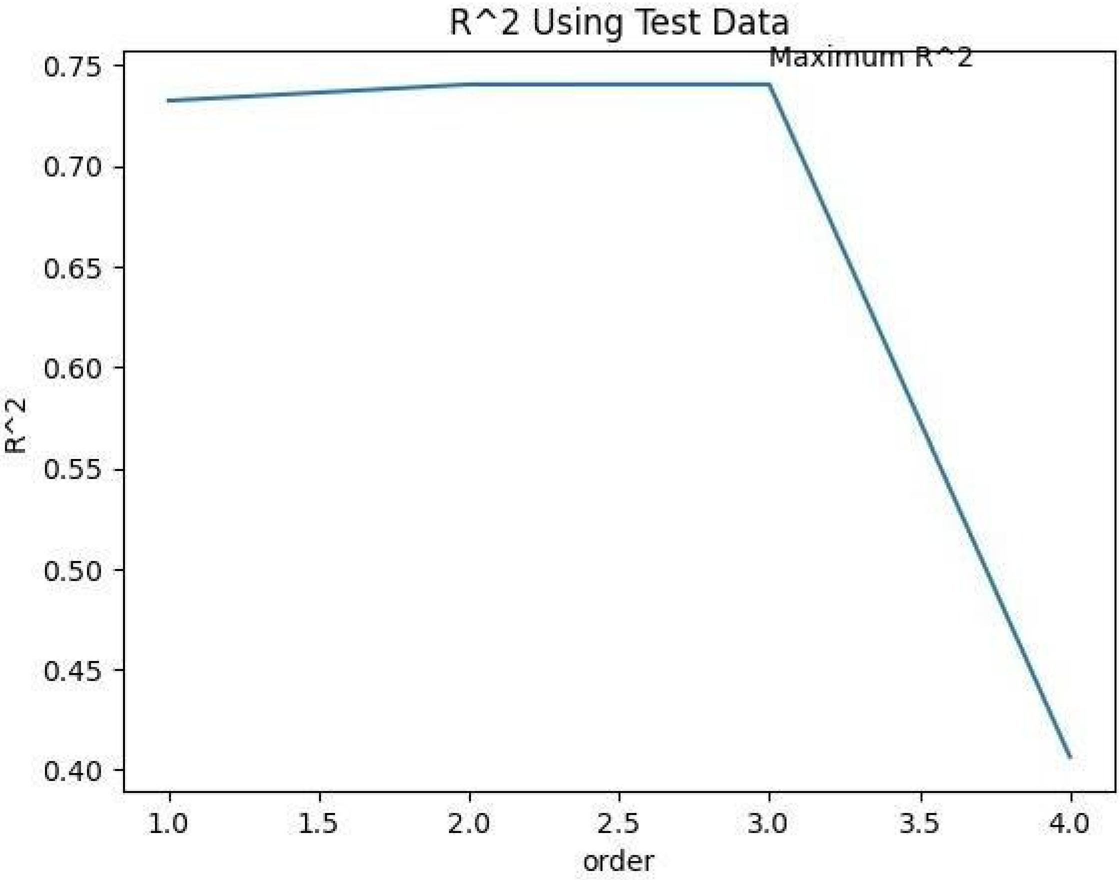


Fig 6.8: Output-2