VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JNANA SANGAMA",MACHHE, BELAGAVI-590018

ML Mini Project Report
on

"Car Price Prediction"

Submitted in partial fulfillment of the requirements for the VI semester

Bachelor of Engineering

in

Artificial Intelligence & Machine Learning

of

Visvesvaraya Technological University, Belagavi

by

POOJITHA P (1CD21AI040)
P RAAGANJANI (1CD21AI039)

Under the Guidance of
Dr.Varalatchoumy.M,
Prof. Syed Hayath,
Dept. of AI&ML

Department of Artificial Intelligence & Machine Learning
CAMBRIDGE INSTITUTE OF TECHNOLOGY, BANGALORE-560036
2023-2024

CAMBRIDGE INSTITUTE OF TECHNOLOGY

K.R. Puram, Bangalore-560 036

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CERTIFICATE

Certified that **Ms. POOJITHA P,** bearing USN **1CD21AI040 and Ms. P RAAGANJANI** bearing USN **1CD21AI039,** a Bonafide students of **Cambridge Institute of Technology,** has successfully completed the ML Mini Project entitled "**Car Price Prediction"** in partial fulfillment of the requirements for VI semester **Bachelor of Engineering** in **Artificial Intelligence & Machine Learning** of **Visvesvaraya Technological University, Belagavi** during academic year 2023-24. It is certified that all Corrections/Suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The ML Mini Project report has been approved as it satisfies the academic requirements prescribed for the Bachelor of Engineering degree.

Mini Project Guides,

Dr.Varalatchoumy.M,

Prof. Syed Hayath
Dept. of AI&ML, CITech

Head of the Department,
Dr.Varalatchoumy.M,
Dept. of AI&ML, CITech

DECLARATION

We POOJITHA P and **P RAAGANJANI** of VI semester BE, Artificial Intelligence & Machine Learning, Cambridge Institute of Technology, hereby declare that the ML Mini Project entitled **“Car Price Prediction”** has been carried out by us and submitted in partial fulfillment of the course requirements of VI semester **Bachelor of Engineering** in **Artificial Intelligence & Machine Learning** as prescribed b**y Visvesvaraya Technological University, Belagavi**, during the academic year 2023-2024.

We also declare that, to the best of my knowledge and belief, the work reported here does not form part of any other report on the basis of which a degree or award was conferred on an earlier occasion on this by any other student.

Date: **POOJITHA P**
Place: Bangalore **1CD21AI040**

P RAAGANJANI
1CD21AI039

ACKNOWLEDGEMENT

We would like to place on record my deep sense of gratitude to **Shri. D. K. Mohan,** Chairman, Cambridge Group of Institutions, Bangalore, India for providing excellent Infrastructure and Academic Environment at CITech without which this work would not have been possible**.**

We are extremely thankful to **Dr. G.Indumathi,** Principal, CITech, Bangalore, for providing me the academic ambience and everlasting motivation to carry out this work and shaping our careers.

We express my sincere gratitude to **Dr. Varalatchoumy M.,** Prof. & Head, Dept. of Artificial Intelligence & Machine Learning, CITech, Bangalore, for her stimulating guidance, continuous encouragement and motivation throughout the course of present work.

We also wish to extend my thanks to Mini Project Guides, **Dr. Varalatchoumy M.,** Prof. & Head and **Prof. Syed Hayath** Assistant Professor Dept. of AI&ML, CITech, Bangalore for the critical, insightful comments, guidance and constructive suggestions to improve the quality of this work.

Finally to all my friends, classmates who always stood by me in difficult situations also helped me in some technical aspects and last but not the least, we wish to express deepest sense of gratitude to my parents who were a constant source of encouragement and stood by me as pillar of strength for completing this work successfully.

POOJITHA P

P RAAGANJANI

ABSTRACT

This project focuses on developing a machine learning model to predict car prices using linear regression. The model is designed to help users estimate the value of a car based on specific attributes. The key features considered for prediction include the car's company name, model, year of purchase, and kilometers driven. By leveraging these features, the model can provide a reliable estimate of the car's current market price. The development process involved collecting and preprocessing a dataset of used car prices, ensuring data quality, and handling missing values. Exploratory data analysis was conducted to understand the relationships between different features and car prices. The linear regression algorithm was chosen for its simplicity and effectiveness in modeling relationships between variables. To make the model accessible to users, a web interface was developed using Flask. This interface allows users to input car details and receive instant price predictions. The front end of the application is designed to be user-friendly, ensuring a seamless experience for users with varying levels of technical expertise. The project demonstrates the practical application of machine learning in the automotive industry, providing a tool that can assist buyers and sellers in making informed decisions. By accurately predicting car prices, the model adds value to the used car market, contributing to more transparent and fair transactions.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Background

The automotive industry plays a pivotal role in the global economy, with millions of cars being bought and sold every year. Among the various segments of this industry, the used car market stands out due to its size and complexity. Unlike new cars, whose prices are largely determined by the manufacturer, the prices of used cars depend on a myriad of factors, making accurate price prediction a challenging yet essential task. This project delves into the realm of machine learning to develop a robust model capable of predicting the prices of used cars based on several key attributes.

Accurate car price prediction is beneficial for multiple stakeholders in the automotive ecosystem. For buyers, it ensures that they pay a fair price based on the car's condition and market trends. For sellers, it provides a reliable estimate to set a competitive price that attracts potential buyers while ensuring a profitable sale. Dealerships can also leverage accurate pricing models to streamline their inventory management and sales strategies, ultimately improving their business operations.

The used car market is influenced by various dynamic factors. Economic conditions, consumer preferences, and technological advancements continuously reshape the landscape. For instance, a surge in fuel prices might increase the demand for fuel-efficient cars, thereby affecting their resale value. Similarly, the introduction of new car models with advanced features can depreciate the value of older models. Understanding and adapting to these market dynamics is crucial for developing an effective car price prediction model.

By developing a reliable car price prediction model, this project aims to contribute to the broader goal of bringing transparency and efficiency to the used car market. Such a model can serve as a valuable tool for consumers, dealers, and policymakers alike. Consumers can make informed purchasing decisions, dealers can optimize their pricing strategies, and policymakers can gain insights into market trends.

The used car market is ripe for innovation through data-driven techniques. This project leverages machine learning to tackle the complex problem of car price prediction, providing a scalable and accurate solution. By improving price transparency and accuracy, the project aims to enhance the overall efficiency of the used car market, benefiting buyers, sellers, and dealers alike. The successful implementation of this project could pave the way for further advancements in automotive analytics and decision-making.

1.2. Why

The importance of this car price prediction project extends beyond just estimating prices. It brings significant benefits to consumers, sellers, dealerships, financial institutions, policymakers, and the broader automotive market. By enhancing transparency, fairness, and efficiency, the project can transform the used car market, making it more accessible and reliable for everyone involved. Here are some key points highlighting the importance of this project:

1. **Enhanced Market Transparency**

   The used car market often suffers from a lack of transparency, with prices varying widely for similar vehicles based on subjective assessments and inconsistent criteria. A robust machine learning model can provide a standardized and objective way to estimate car prices, reducing information asymmetry and helping consumers make more informed decisions.

2. **Fair Pricing for Consumers**

   Accurate price prediction ensures that buyers pay a fair price for a vehicle based on its condition, age, and other relevant factors. It helps prevent overpaying and protects buyers from potentially inflated prices set by sellers. This fairness can lead to increased trust and satisfaction among consumers.

3. **Optimized Sales Strategies for Sellers**

   For car sellers, whether individuals or dealerships, having access to reliable price estimates can significantly enhance their sales strategies. Sellers can set competitive prices that attract buyers while ensuring they receive fair market value for their vehicles. This can lead to quicker sales and higher turnover rates.

4. **Improved Inventory Management for Dealerships**

   Dealerships can leverage the predictive model to better manage their inventory. By understanding the market value of each vehicle in their stock, dealerships can make informed decisions about which cars to acquire, hold, or sell. This can lead to more efficient inventory turnover and higher profitability.

5. **Insightful Market Analysis**

   The data and insights generated by the predictive model can be valuable for market analysis. Automotive businesses and analysts can identify trends, such as the most sought-after car models, the impact of mileage on prices, and seasonal variations in car values. This information can guide strategic decisions and market positioning.

6. **Facilitating Financial Services**

   Financial institutions, such as banks and insurance companies, can use accurate car price predictions to offer better financial products. For instance, accurate valuation is crucial for determining loan amounts for car financing and setting premiums for car insurance policies. This can lead to more tailored and fair financial services for consumers.

7. **Support for Policy Makers**

   Policy makers can benefit from the aggregated data and trends revealed by the predictive model. Understanding the factors that influence car prices can help in formulating policies related to the automotive industry, such as regulations on emissions, incentives for electric vehicles, and measures to control the import and export of used cars.

8. **Advancement in Machine Learning Applications**

   From a technological standpoint, this project exemplifies the practical application of machine learning in solving real-world problems. It showcases the potential of data science and machine learning to drive innovation and improve processes across various industries. This can inspire further research and development in the field.

1.3. Problem Statement

create a user-friendly interface where users can input specific car details and receive an instant price estimation. This application will utilize a pre-trained linear regression model, trained on a cleaned and curated dataset of historical car sales data, to provide reliable predictions. The model's effectiveness will be assessed through accuracy metrics and validation against real-world pricing data.

Key challenges include:

- Ensuring the accuracy and reliability of price predictions across a diverse range of car brands and models.
- Handling user input validation and error handling to enhance user experience.
- Integrating the machine learning model seamlessly into a Flask web framework for real-time predictions.

1.4. Objectives

The objectives of the car price prediction project are designed to ensure the development of a reliable, accurate, and user-friendly predictive model. By achieving these objectives, the project aims to provide valuable insights and tools for consumers, sellers, and other stakeholders in the used car market. The primary objectives of this project are as follows:

1. **Develop a Predictive Model**
   - **Objective**: Create a machine learning model capable of predicting the price of a used car.
   - **Approach**: Experiment with different machine learning algorithms, including linear regression, decision trees, random forests, and gradient boosting machines, to determine the best-performing model.
2. **Data Collection and Preparation**
   - **Objective**: Gather a comprehensive dataset containing relevant features that influence car prices.
   - **Approach**: Source data from car dealerships, online marketplaces, and automotive databases. Ensure the dataset includes attributes such as company name, car model, year of purchase, kilometers driven, fuel type, and transmission type.

3. **Data Preprocessing**

- **Objective**: Clean and preprocess the data to make it suitable for model training.
- **Approach**: Handle missing values, encode categorical variables, and scale numerical features. Ensure the dataset is free from inconsistencies and ready for analysis.

4. **Model Training and Evaluation**

- **Objective**: Train the selected machine learning model and evaluate its performance.
- **Approach**: Split the dataset into training and testing sets. Use metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score to assess the model's accuracy and reliability.

5. **Feature Importance Analysis**

- **Objective**: Identify the most significant features influencing the car price predictions.
- **Approach**: Analyze the trained model to determine the relative importance of each feature, providing insights into the factors that most impact car prices.

6. **Model Optimization**

- **Objective:** Optimize the model to enhance its predictive accuracy and performance.
- **Approach:** Fine-tune the model's hyperparameters, use cross-validation techniques, and experiment with ensemble methods to improve the model's robustness and accuracy.

7. **User-Friendly Deployment**

- **Objective:** Develop a user-friendly interface or API to make the model accessible to end-users.
- **Approach:** Build a web application or RESTful API that allows users to input car details and receive price predictions in real-time. Ensure the interface is intuitive and easy to use.

8. Validation and Testing

- **Objective:** Validate the model's performance on a separate validation set and conduct thorough testing.

- **Approach**: Ensure the model generalizes well to new, unseen data by testing it on a validation set and refining it based on the results. Conduct A/B testing or user testing to gather feedback and make necessary improvements.

9. Documentation and Reporting

- **Objective:** Document the entire process, including data collection, preprocessing, model development, and evaluation.

- **Approach:** Create detailed documentation and reports outlining the methodology, findings, and conclusions. Ensure that the documentation is clear and comprehensive for future reference.

10. Continuous Improvement

- **Objective:** Establish a framework for continuous improvement and model updates.

- **Approach:** Set up mechanisms to regularly update the dataset and retrain the model to adapt to changing market conditions and new data. Incorporate user feedback to enhance the model and its deployment over time.

CHAPTER 2

LITERATURE REVIEW

2.1. Vehicle Price Classification and Prediction Using Machine Learning

Vehicle deal forecast and ML is a work area that examines the improvement that occurs in the scholarly community and organizations with the guide of ML computations in foreseeing these deals for financial sustainability. Creators in used a fluffy-based information structure to anticipate the expense of vehicle exchange. Only three elements were explicit: vehicle brand, creation year, and motor sort were considered in this assessment. The proposed structure made a similar result when appearing differently with essential regression methodologies. Vehicle merchants in the USA sell an immense number of vehicles reliably through leasing. Most of these vehicles are returned close to the completion of the leasing period and ought to be traded. Selling these vehicles at the right expense has major money-related issues for them to flourish. Along these lines, the Optimal Distribution of Auction Vehicles (ODAV) structure was made by creators in. This structure does not simply check the best expense for trading the vehicles; moreover, it gives an urge on where to sell the vehicle. Since the United States is an immense country, the territory where the vehicle is sold in such a manner has a non-irrelevant impact on the selling cost of vehicles. A k-NN model was used for deciding the expense. Since this system was started in 2003, more than 2 million vehicles have been scattered through this structure.

Creators in proposed another model reliant on the neural system framework to evaluate the private exchange of vehicles. The essential features used in this assessment are mileage, creator, and life expectancy. The model was moved up to manage nonlinear associations, which is incomprehensible with straight relapse procedures. It was found that this model was reasonably definite in predicting the extra estimation of vehicle exchange. This system gives encounters into the best expenses for vehicles, similar to the zone where it very well may be normal for pickups. Creators in additionally proposed a model that is amassed using ANN at the cost desired for a vehicle exchange. They contemplated a couple of properties: speed, life length, and brand. The proposed model was gathered so it could oversee nonlinear relations in data with past models that were utilizing the relapse procedures. The model had the choice to envision the expenses of vehicles with the best precision over other direct models. In addition, Pudaruth et al. applied different ML calculations; to be specific: k-NN, numerous regression operations, naive Bayes, and a decision tree for vehicle value forecast in Mauritius. The dataset used to make a forecast model was gathered physically from neighborhood papers within one month, as time can noticeably affect the cost of the vehicle. They examined the accompanying properties: brand, model, cubic limit, and speed. In any case, the creator discovered that decision trees and naive Bayes could not anticipate and characterize numeric qualities. Moreover, the predetermined number of dataset examples number of dataset examples could not give high grouping exhibitions, for example, correctness under 70%.

Creators in built up a model for vehicle esteem gauges by using various relapse activities. The dataset was made for two months and consolidated the following features: cost, cubic breaking point, outside concealing, and so forth. In the wake of applying feature decisions, the makers thought about simple engine sort, esteem, model year, and model as data features. With the given course of action, makers had the alternative to achieve the desired exactness of 98%.

In the related work, as shown above, the authors proposed an estimated model subjected to the single ML computation. Regardless, it is noticeable that the lone ML estimation approach did not give astounding desired results and could be redesigned by storing up various ML methods in a company for financial sustainability.

2.2. Advantages and Disadvantages of Machine Learning

2.2.1. Advantages of Machine Learning

1. Improved Accuracy and Precision

One of the most significant benefits of machine learning is its ability to improve accuracy and precision in various tasks. ML models can process vast amounts of data and identify patterns that might be overlooked by humans. For instance, in medical diagnostics, ML algorithms can analyze medical images or patient data to detect diseases with a high degree of accuracy.

2. Automation of Repetitive Tasks

Machine learning enables the automation of repetitive and mundane tasks, freeing up human resources for more complex and creative endeavors. In industries like manufacturing and customer service, ML-driven automation can handle routine tasks such as quality control, data entry, and customer inquiries, resulting in increased productivity and efficiency.

3. Enhanced Decision-Making

ML models can analyze large datasets and provide insights that aid in decision-making. By identifying trends, correlations, and anomalies, machine learning helps businesses and organizations make data-driven decisions. This is particularly valuable in sectors like finance, where ML can be used for risk assessment, fraud detection, and investment strategies.

4. Personalization and Customer Experience

Machine learning enables the personalization of products and services, enhancing customer experience. In e-commerce, ML algorithms analyze customer behavior and preferences to recommend products tailored to individual needs. Similarly, streaming services use ML to suggest content based on user viewing history, improving user engagement and satisfaction.

5. Predictive Analytics

Predictive analytics is a powerful application of machine learning that helps forecast future events based on historical data. Businesses use predictive models to anticipate customer demand, optimize inventory, and improve supply chain management. In healthcare, predictive analytics can identify potential outbreaks of diseases and help in preventive measures.

6. Scalability

Machine learning models can handle large volumes of data and scale efficiently as data grows. This scalability is essential for businesses dealing with big data, such as social media platforms and online retailers. ML algorithms can process and analyze data in real-time, providing timely insights and responses.

7. Improved Security

ML enhances security measures by detecting and responding to threats in real-time. In cybersecurity, ML algorithms analyze network traffic patterns to identify unusual activities indicative of cyberattacks. Similarly, financial institutions use ML for fraud detection by monitoring transactions for suspicious behavior.

8. Cost Reduction

By automating processes and improving efficiency, machine learning can lead to significant cost reductions. In manufacturing, ML-driven predictive maintenance helps identify equipment issues before they become costly failures, reducing downtime and maintenance costs. In customer service, chatbots powered by ML reduce the need for human agents, lowering operational expenses.

9. Innovation and Competitive Advantage

Adopting machine learning fosters innovation and provides a competitive edge. Companies that leverage ML for product development, marketing strategies, and customer insights are better positioned to respond to market changes and meet customer demands. ML-driven innovation can lead to the creation of new products and services, opening up new revenue streams.

10. Enhanced Human Capabilities

Machine learning augments human capabilities by providing tools and insights that enhance performance. In fields like healthcare, ML assists doctors in diagnosing and treating patients more effectively. In research, ML accelerates the discovery process by analyzing vast datasets and identifying potential breakthroughs.

2.2.2. Disadvantages of Machine Learning

1. Data Dependency

Machine learning models require vast amounts of data to train effectively. The quality, quantity, and diversity of the data significantly impact the model’s performance. Insufficient or biased data can lead to inaccurate predictions and poor decision-making. Additionally, obtaining and curating large datasets can be time-consuming and costly.

2. High Computational Costs

Training ML models, especially deep learning algorithms, demands significant computational resources. High-performance hardware such as GPUs and TPUs are often required, which can be expensive. The energy consumption associated with training large models is also substantial, raising concerns about the environmental impact.

3. Complexity and Interpretability

Many machine learning models, particularly deep neural networks, function as black boxes. Their complexity makes it difficult to interpret how they arrive at specific decisions. This lack of transparency poses challenges in fields where understanding the decision-making process is critical, such as healthcare and finance.

4. Overfitting and Underfitting

Machine learning models can suffer from overfitting or underfitting. Overfitting occurs when a model learns the training data too well, capturing noise and anomalies, which reduces its generalization ability to new data. Underfitting happens when a model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and test data.

5. Ethical Concerns

ML applications can raise ethical issues, particularly concerning privacy and bias. Data privacy is a significant concern, as ML models often require access to sensitive and personal information. Bias in training data can lead to biased models, perpetuating existing inequalities and unfair treatment of certain groups.

6. Lack of Generalization

Machine learning models are typically designed for specific tasks and may struggle to generalize across different domains or datasets. Transfer learning techniques can mitigate this issue to some extent, but developing models that perform well in diverse scenarios remains a challenge.

7. Dependency on Expertise

Developing and deploying machine learning models require specialized knowledge and expertise. This includes understanding algorithms, data preprocessing, model training, and evaluation. The scarcity of skilled professionals in the field can hinder the adoption and implementation of ML solutions.

8. Security Vulnerabilities

ML models are susceptible to adversarial attacks, where malicious actors manipulate input data to deceive the model into making incorrect predictions. This vulnerability poses significant risks in critical applications such as autonomous driving, cybersecurity, and financial fraud detection.

9. Maintenance and Updates

ML models require continuous monitoring, maintenance, and updates to ensure they remain accurate and effective over time. Changes in the underlying data distribution, known as data drift, can degrade model performance, necessitating frequent retraining and validation.

2.3.Linear Regression

Linear regression relates predictor variables and outcome variables, such as gene copy numbers and the level of a biomarker. The assumed linearity of the relationships makes the models convenient both mathematically and computationally. And since the data can be arbitrarily transformed beforehand, such as by including polynomials of the copy numbers as predictor variables or by replacing the level of the biomarker in the outcome variable by its logarithm, linear regression can also effectively model non-linear relationships. This simplicity and flexibility have made linear regression the most popular statistical framework across the sciences and standard textbook material. But the standard methods for linear regression, such as the least-squares estimator, premise that the number of parameters is small as compared to the number of samples, which limits their usefulness in modern, data-intensive research, where the increasing granularity of data has prompted interest in increasingly complex models. More recent high-dimensional methods, in contrast, allow for models with many more parameters. These methods are the topic of this chapter.

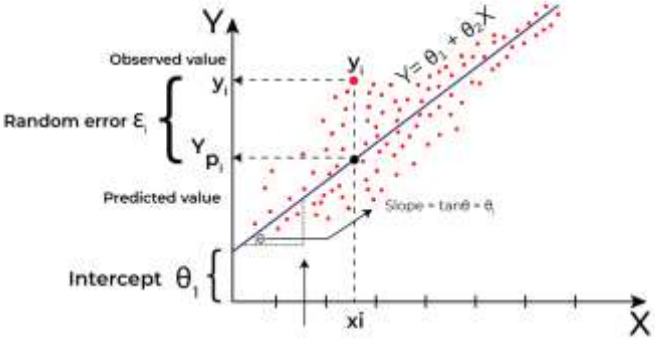

Y
Observed value
y_i
y_i
$Y = \theta_1 + \theta_2 X$
Random error ε_i
Y_{p_i}
Predicted value
Slope = $\tan\theta = \theta_i$
θ
Intercept θ_1
xi
X

Fig. 2.1. Linear Regression Graph

CHAPTER 3

METHODOLOGY

The methodology outlined above ensures a comprehensive approach to developing and deploying a machine learning model for car price prediction. By following these steps, the project achieves its objective of providing a reliable, accurate, and user-friendly tool for predicting used car prices.

3.1. Problem Definition

The objective is to build a machine learning model that accurately predicts the price of a used car based on various features, such as company name, car model, year of purchase, kilometers driven, and fuel type. The model is deployed via a Flask web application to make it accessible to users.

3.2. Data Collection

Data was gathered from various sources, including car dealerships, online marketplaces, and automotive websites. The dataset includes the following features:

- company: The brand of the car (e.g., Toyota, Ford).
- name: The model of the car (e.g., Camry, Mustang).
- year: The year the car was purchased.
- kms_driven: The total distance the car has been driven.
- fuel_type: The type of fuel the car uses (e.g., Petrol, Diesel).
- price: The price of the car (target variable).

3.3.Data Preprocessing

The collected data was preprocessed to ensure it was clean and suitable for model training:

- **Handling Missing Values**: Missing values were imputed or removed.
- **Encoding Categorical Variables**: Categorical variables such as company, name, and fuel_type were encoded using techniques like one-hot encoding.
- **Feature Scaling**: Numerical features like kms_driven and year were scaled to ensure they contributed equally to the model.**Outlier Removal**: Outliers in the dataset were identified and removed to improve model accuracy.

3.4. Model Selection

Several machine learning algorithms were explored to determine the best model for predicting car prices:

- **Linear Regression**: Used as a baseline model due to its simplicity and interpretability.
- **Decision Trees**: Captured non-linear relationships in the data.
- **Random Forests**: Improved accuracy by combining multiple decision trees.
- **Gradient Boosting Machines (GBMs)**: Provided robust performance by sequentially building trees to correct errors from previous trees.

3.5. Model Training and Evaluation

The dataset was split into training and testing sets to evaluate model performance:

- **Training**: Models were trained on the training set using various algorithms.
- **Evaluation**: Model performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score.
- **Cross-Validation**: Cross-validation techniques were employed to ensure the model's generalizability.

3.6. Feature Extraction

Feature extraction and engineering are critical steps in the car price prediction project. By carefully selecting and transforming features, the model can capture the essential patterns and relationships in the data, leading to more accurate and reliable predictions. The process involved handling categorical variables, creating new features, and selecting the most relevant ones using statistical and machine learning techniques. Feature extraction involves identifying and selecting the most relevant variables that significantly influence the target variable—in this case, the car price. This step is crucial for improving model performance and accuracy. Here’s how feature extraction was conducted for the car price prediction project.

1. Initial Feature Identification

The initial set of features considered for the model includes:

- **Company**: The brand of the car (e.g., Toyota, Ford).
- **Name**: The specific model of the car (e.g., Camry, Mustang).
- **Year**: The year of purchase or manufacturing year.
- **Kilometers Driven**: The total distance the car has been driven.
- **Fuel Type**: The type of fuel the car uses (e.g., Petrol, Diesel).

These features were identified based on their direct influence on the car's resale value.

2. Data Preprocessing

Before extracting features, the data was cleaned and preprocessed:

- **Handling Missing Values**: Missing values were imputed or removed.
- **Encoding Categorical Variables**: Categorical variables (e.g., company, name, fuel_type) were converted into numerical representations using techniques like one-hot encoding.
- **Feature Scaling**: Numerical features (e.g., year, kms_driven) were scaled to ensure they contributed equally to the model.

3. One-Hot Encoding

Categorical features were transformed using one-hot encoding to create binary columns for each category. This process was applied to the following features:

- **Company**: Each unique car brand was converted into a binary column.
- **Name**: Each unique car model was converted into a binary column.
- **Fuel Type**: Each fuel type was converted into a binary column.

4. Feature Engineering

Additional features were created to capture more nuanced relationships in the data:

- **Car Age:** Instead of using the year directly, the age of the car was calculated as the difference between the current year and the manufacturing year.
- **Log Transformation:** For highly skewed numerical features like kms_driven, a log transformation was applied to normalize the distribution.

5. Feature Selection

Feature selection involves identifying the most significant features for the prediction task. This was done using various methods:

- **Correlation Analysis**: Pearson correlation coefficients were calculated to identify the strength of the linear relationship between features and the target variable.
- **Feature Importance from Models**: Tree-based models like Random Forests and Gradient Boosting Machines were used to identify feature importance. These models provide an inherent measure of the importance of each feature in predicting the target variable.

6. Principal Component Analysis (PCA)

For dimensionality reduction, PCA was used to transform the feature space into a smaller set of uncorrelated components while retaining most of the variance in the data. This step was particularly useful for handling high-dimensional data resulting from one-hot encoding.

7. Final Feature Set

The final set of features used for model training included:

- **Encoded Company**: Binary columns for each car brand.
- **Encoded Name**: Binary columns for each car model.
- **Age**: The age of the car.
- **Log Kilometers Driven**: Log-transformed kilometers driven.
- **Encoded Fuel Type**: Binary columns for each fuel type.

3.7. Model Deployment

The final model was deployed using Flask, a web framework for Python:

- **Flask Application**: A Flask web application was developed to provide a user-friendly interface for car price prediction.
- **Endpoints**: Two main endpoints were created:
  - ➢ /: Renders the main page where users can input car details.
  - ➢ /predict: Handles form submissions and returns the predicted car price.
- **Cross-Origin Resource Sharing (CORS)**: CORS was enabled to allow cross-origin requests to the Flask application.

3.8. System Architecture

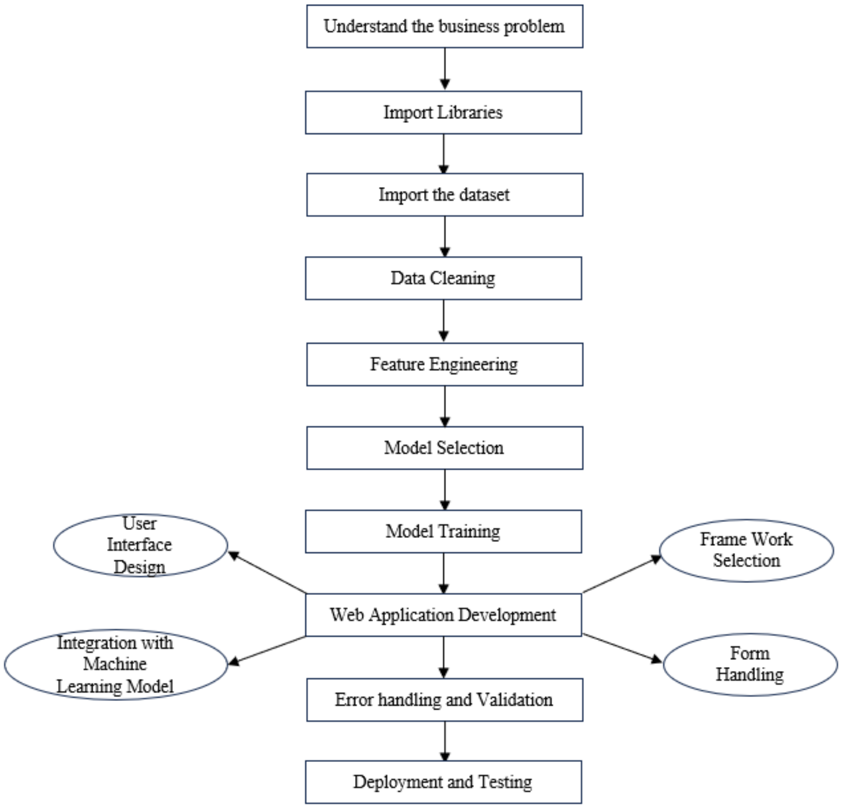


Fig. 3.1. System Architecture

3.9. Tools and Technologies Used

3.9.1. Hardware Tools

- A typical PC or workstation used by data scientists may include a multi-core CPU, 16-32 GB of RAM, and an SSD for fast data access.
- CPU: Multi-core processors with high clock speeds.
- GPU: Graphics Processing Units (such as NVIDIA GPUs) for accelerated machine learning computations.
- Memory: High memory capacity (64 GB or more) to handle large datasets.
- Storage: High-speed SSDs or NVMe storage for fast read/write operations.

3.9.2. Software Tools

- Click-7.1.2
- Flask-1.1.2
- Flask-Cors-3.0.8
- Gunicorn-20.0.4
- Jinja2-2.11.2
- Joblib-0.15.1
- Numpy-1.18.5
- Pandas-1.0.4
- pickle-mixin-1.0.2
- python-dateutil-2.8.1
- pytz-2020.1
- scikit-learn-0.22.2
- six-1.15.0
- sklearn-0.0

CHAPTER 4

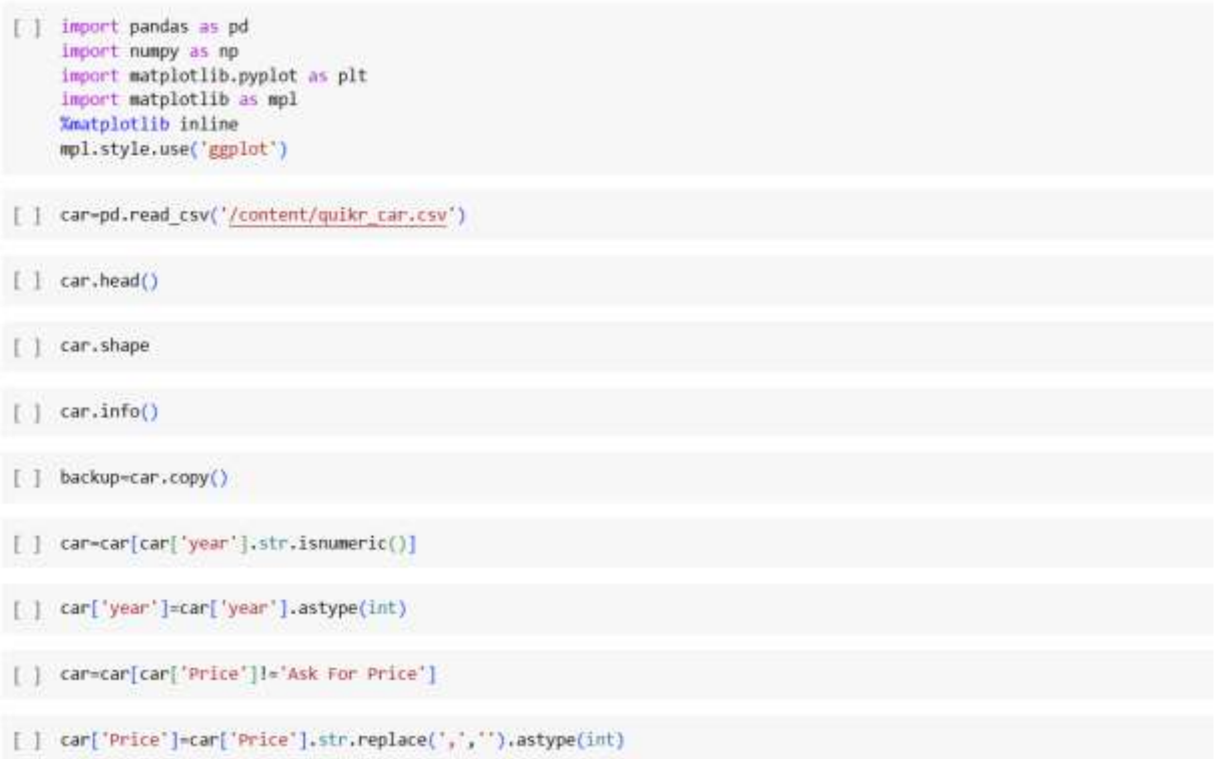
IMPLEMENTATION

4.1. Steps Followed

1. **Problem Definition**: Identified the objective of predicting used car prices based on features like company name, model, year, kilometers driven, and fuel type.
2. **Data Collection**: Gathered data from sources such as car dealerships and online marketplaces, including relevant features and target variables.
3. **Data Preprocessing**: Cleaned the dataset by handling missing values, encoding categorical variables, and scaling numerical features.
4. **Feature Extraction and Engineering**: Selected and created features, such as encoding company and model names, calculating car age, and applying log transformations to skewed data.
5. **Model Selection and Training**: Experimented with different machine learning algorithms (e.g., linear regression, decision trees) and trained the model using the processed data.
6. **Model Evaluation and Optimization**: Evaluated model performance using metrics like MAE and R^2, and optimized the model by tuning hyperparameters.
7. **Deployment**: Developed a Flask web application with a user-friendly interface for inputting car details and receiving price predictions.

4.2. Code Snippet

4.2.1. Backend Code



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

```
car=pd.read_csv('/content/quikr_car.csv')
```

```
car.head()
```

```
car.shape
```

```
car.info()
```

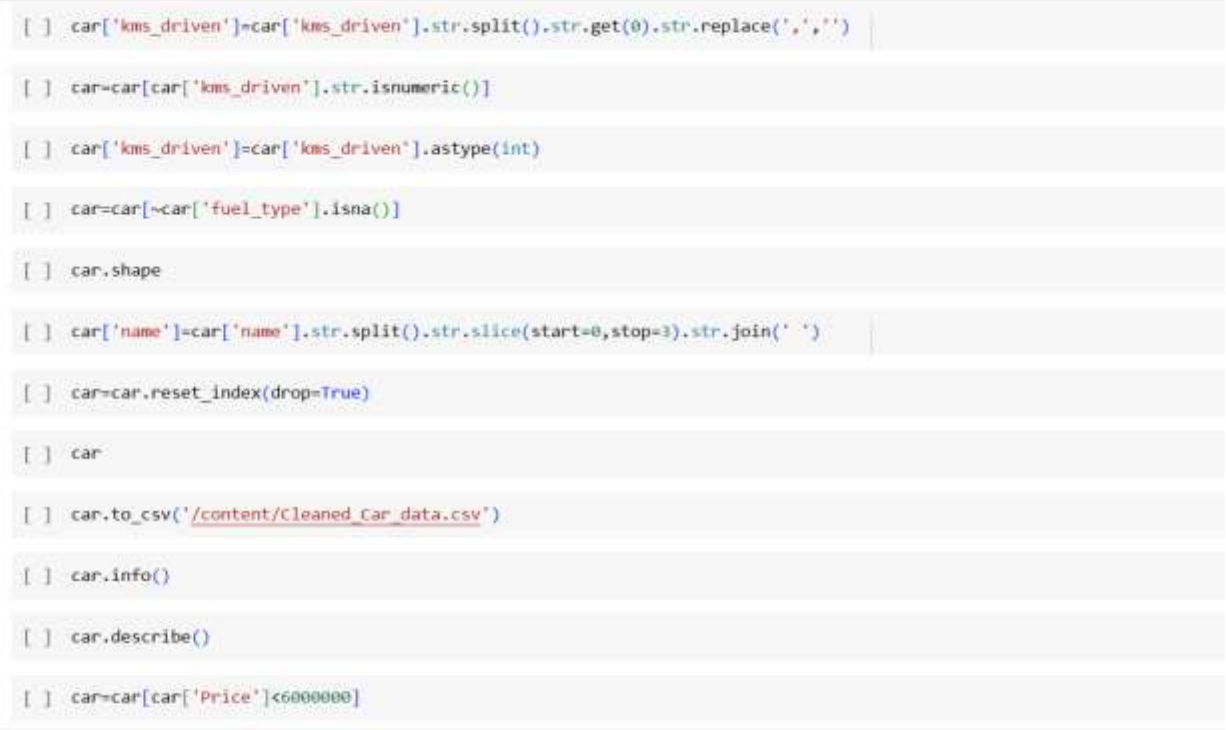
```
backup=car.copy()
```

```
car=car[car['year'].str.isnumeric()]
```

```
car['year']=car['year'].astype(int)
```

```
car=car[car['Price']!='Ask For Price']
```

```
car['Price']=car['Price'].str.replace(',','').astype(int)
```



```
car['kms_driven']=car['kms_driven'].str.split().str.get(0).str.replace(',','')
```

```
car=car[car['kms_driven'].str.isnumeric()]
```

```
car['kms_driven']=car['kms_driven'].astype(int)
```

```
car=car[~car['fuel_type'].isna()]
```

```
car.shape
```

```
car['name']=car['name'].str.split().str.slice(start=0,stop=3).str.join(' ')
```

```
car=car.reset_index(drop=True)
```

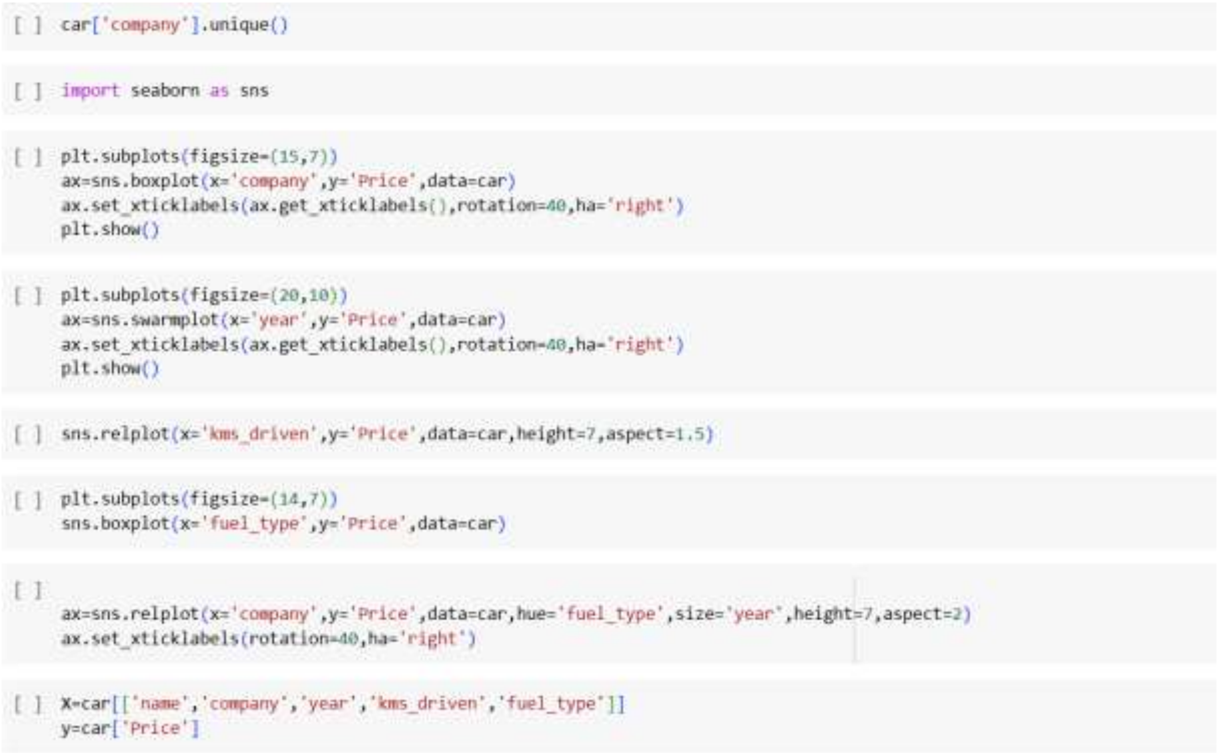
```
car
```

```
car.to_csv('/content/Cleaned_Car_data.csv')
```

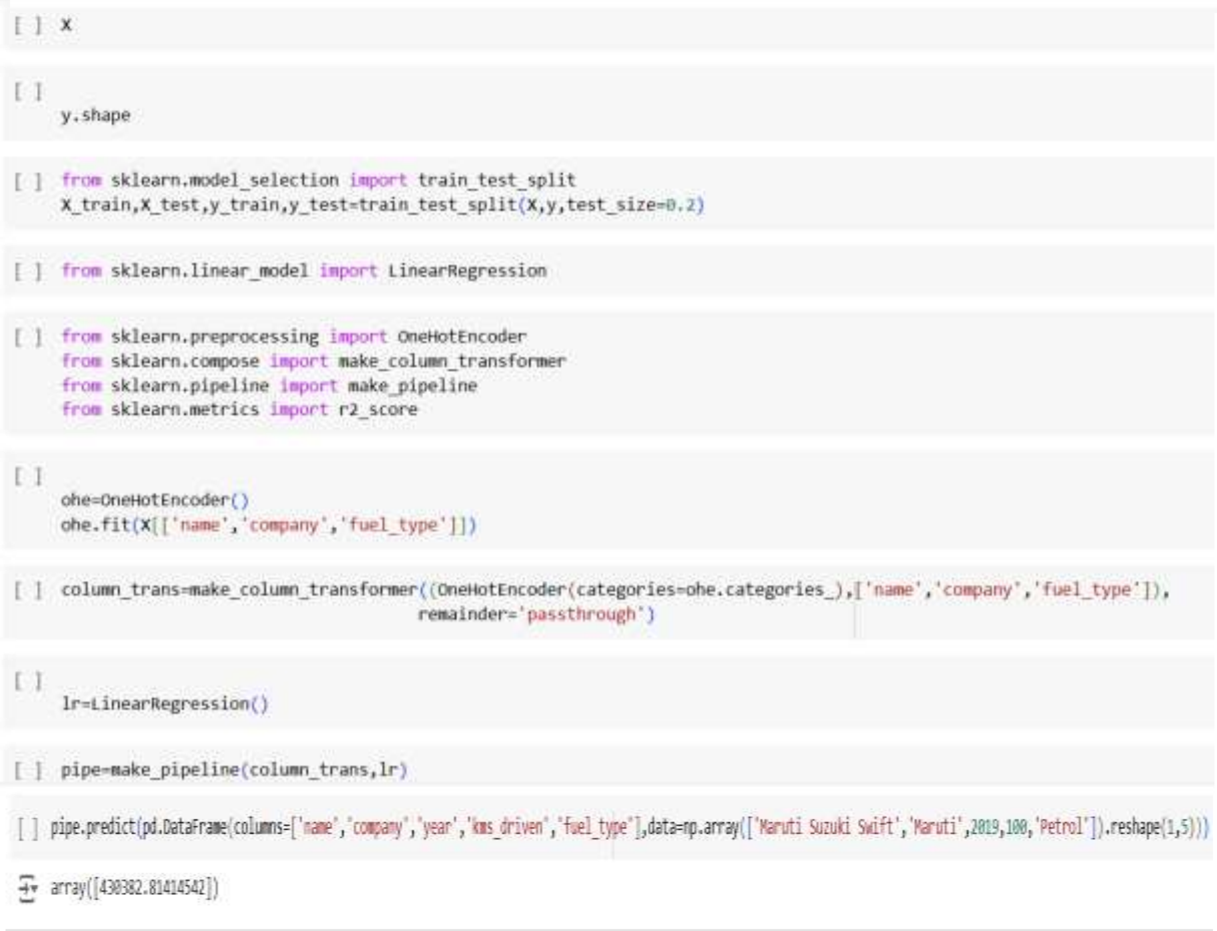
```
car.info()
```

```
car.describe()
```

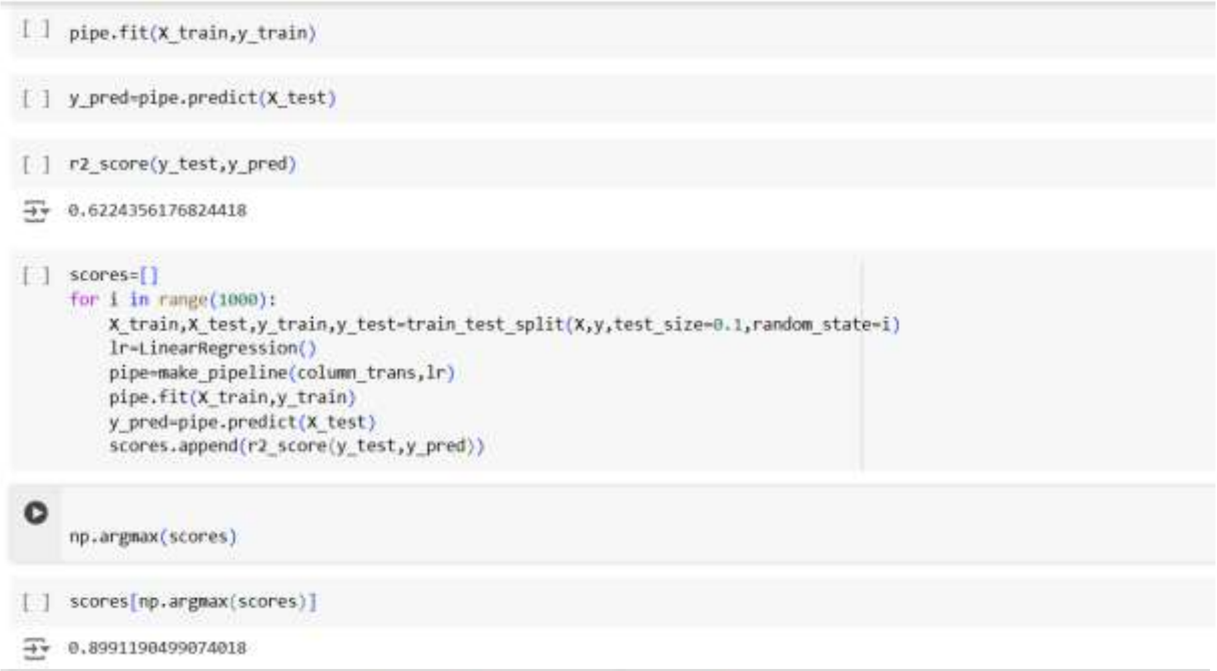
```
car=car[car['Price']<6000000]
```



```
[ ] car['company'].unique()

[ ] import seaborn as sns

[ ] plt.subplots(figsize=(15,7))
    ax=sns.boxplot(x='company',y='Price',data=car)
    ax.set_xticklabels(ax.get_xticklabels(),rotation=40,ha='right')
    plt.show()

[ ] plt.subplots(figsize=(20,10))
    ax=sns.swarmplot(x='year',y='Price',data=car)
    ax.set_xticklabels(ax.get_xticklabels(),rotation=40,ha='right')
    plt.show()

[ ] sns.relplot(x='kms_driven',y='Price',data=car,height=7,aspect=1.5)

[ ] plt.subplots(figsize=(14,7))
    sns.boxplot(x='fuel_type',y='Price',data=car)

[ ]
    ax=sns.relplot(x='company',y='Price',data=car,hue='fuel_type',size='year',height=7,aspect=2)
    ax.set_xticklabels(rotation=40,ha='right')

[ ] X=car[['name','company','year','kms_driven','fuel_type']]
    y=car['Price']
```



```
[ ] X

[ ]
    y.shape

[ ] from sklearn.model_selection import train_test_split
    X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)

[ ] from sklearn.linear_model import LinearRegression

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

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

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

[ ]
    lr=LinearRegression()

[ ] pipe=make_pipeline(column_trans,lr)

[ ] pipe.predict(pd.DataFrame(columns=['name','company','year','kms_driven','fuel_type'],data=np.array(['Maruti Suzuki Swift','Maruti',2019,100,'Petrol']).reshape(1,5)))

array([430382.81414542])
```



```
[ ] pipe.fit(X_train,y_train)
```

```
[ ] y_pred=pipe.predict(X_test)
```

```
[ ] r2_score(y_test,y_pred)
```

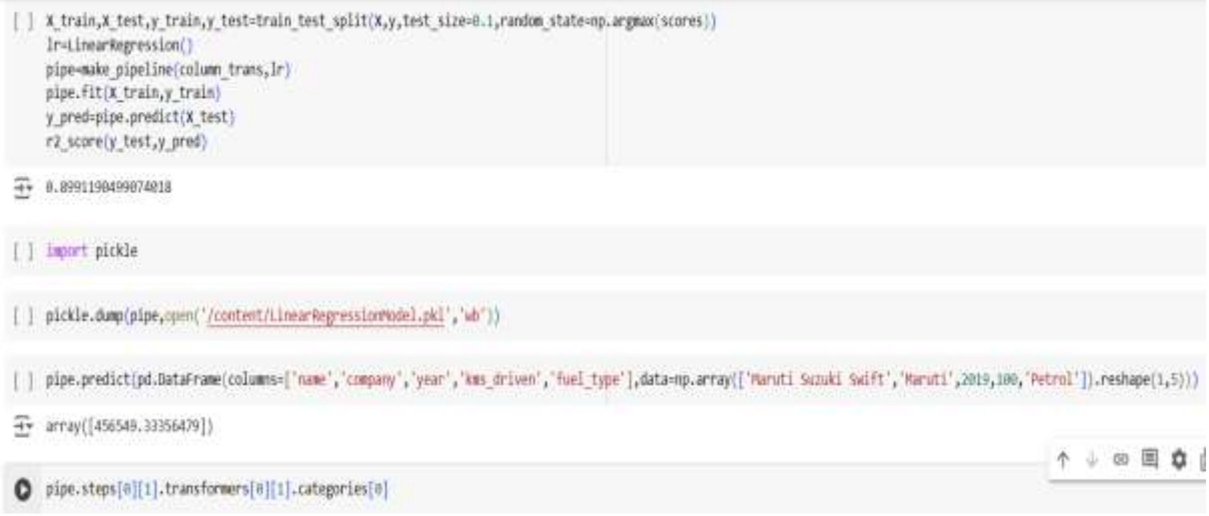
```
0.6224356176824418
```

```
[ ] 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))
```

```
np.argmax(scores)
```

```
[ ] scores[np.argmax(scores)]
```

```
0.8991190499074018
```



```
[ ] X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.1,random_state=np.argmax(scores))
    lr=LinearRegression()
    pipe=make_pipeline(column_trans,lr)
    pipe.fit(X_train,y_train)
    y_pred=pipe.predict(X_test)
    r2_score(y_test,y_pred)
```

```
0.8991190499074018
```

```
[ ] import pickle
```

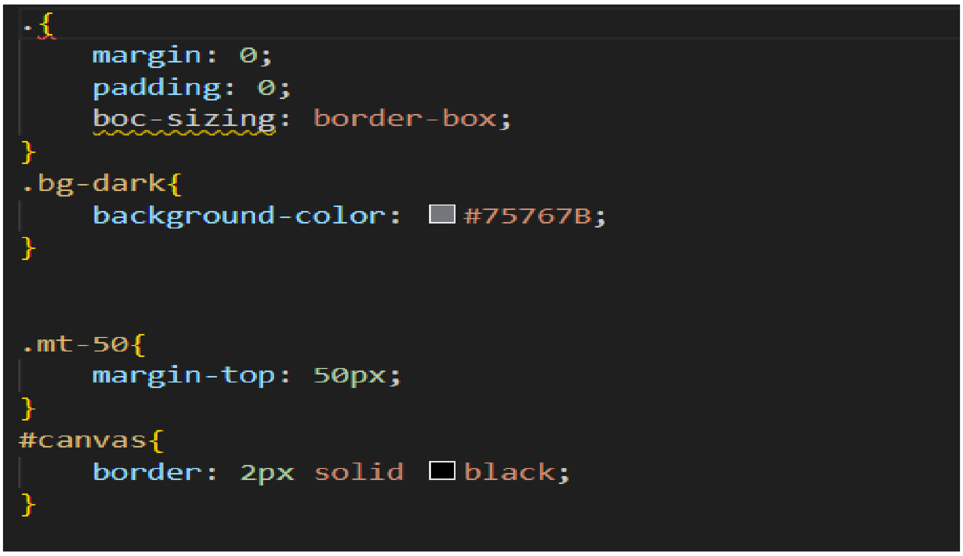
```
[ ] pickle.dump(pipe,open('/content/LinearRegressionModel.pkl','wb'))
```

```
[ ] pipe.predict(pd.DataFrame(columns=['name','company','year','kms_driven','fuel_type'],data=np.array(['Maruti Suzuki Swift','Maruti',2019,100,'Petrol']).reshape(1,5)))
```

```
array([456549.33356479])
```

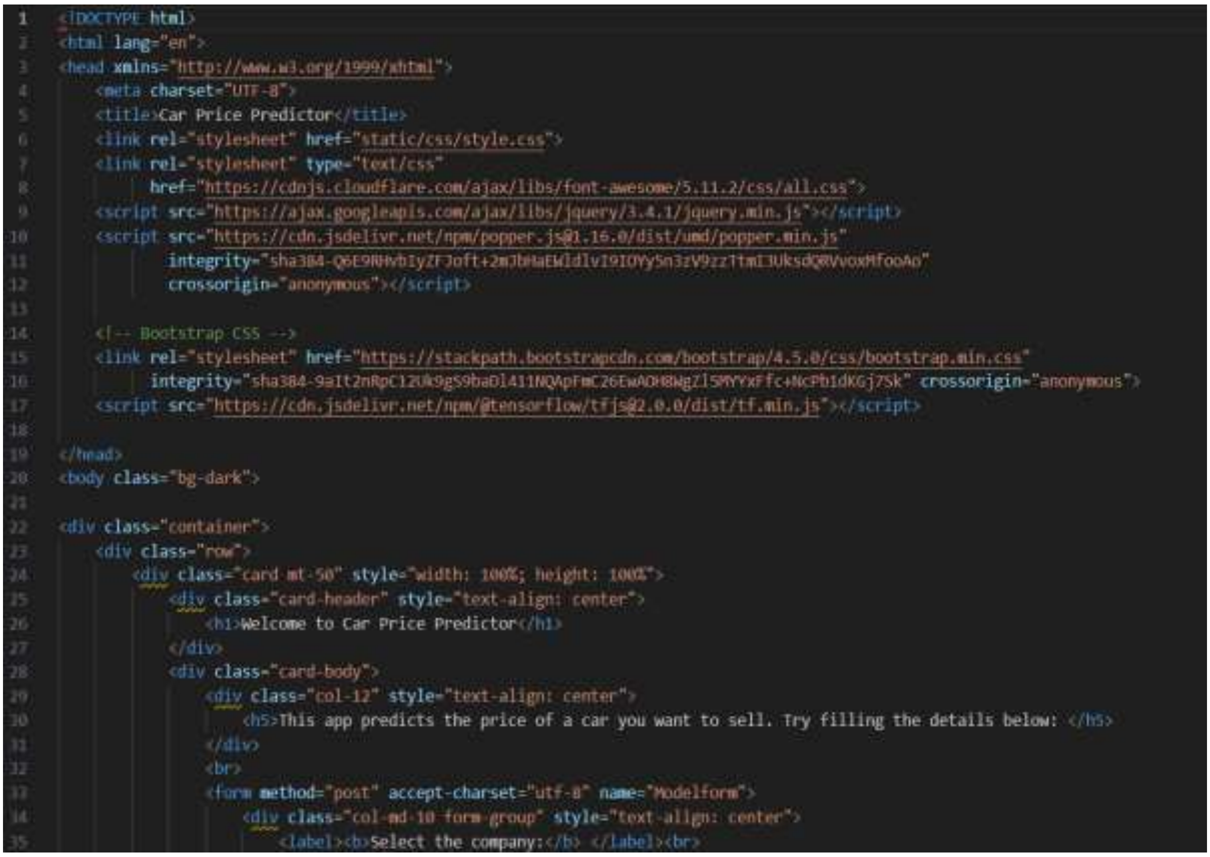
```
pipe.steps[0][1].transformers[0][1].categories[0]
```

4.2.2. CSS Code

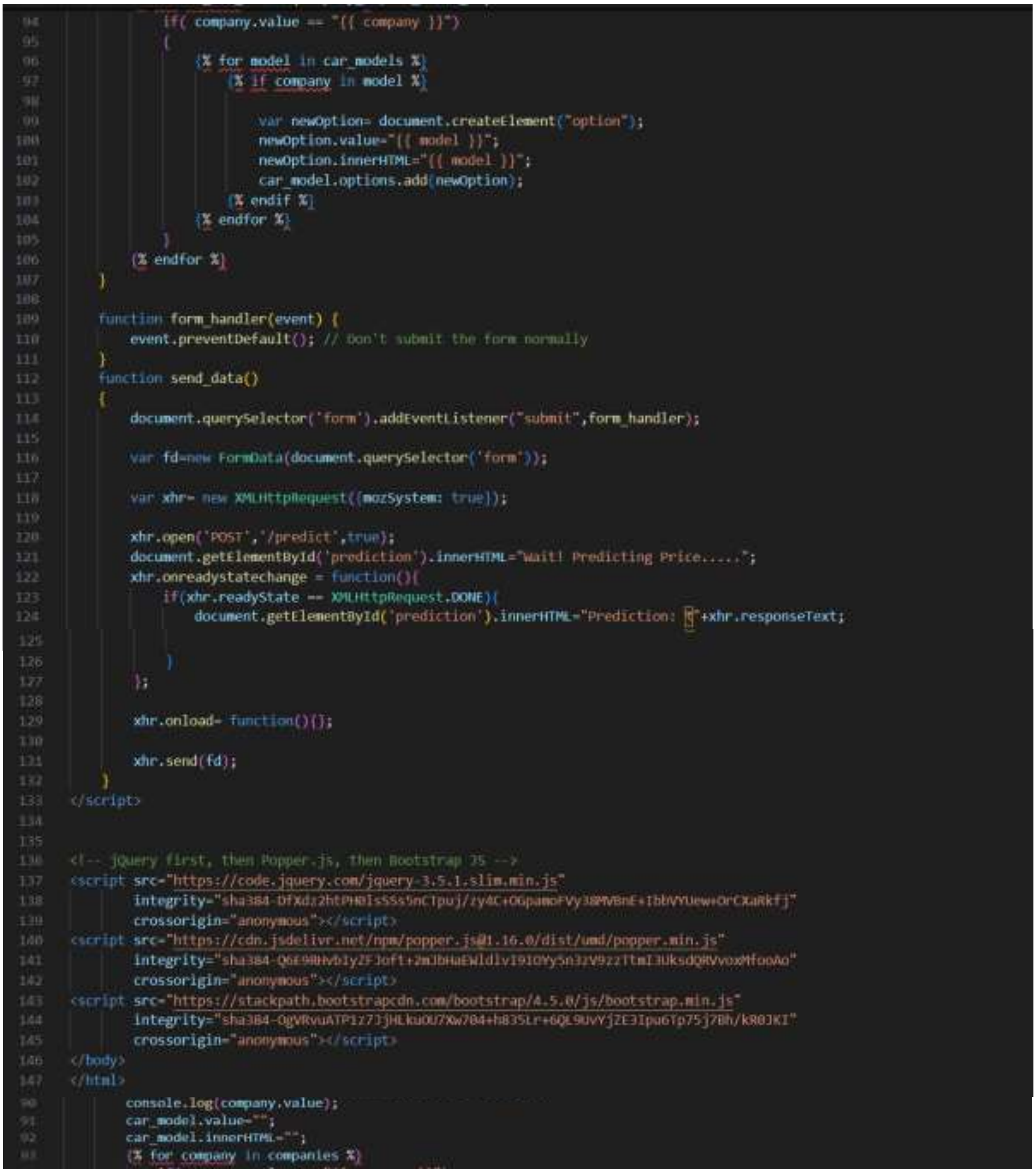


```css
*{
    margin: 0;
    padding: 0;
    boc-sizing: border-box;
}
.bg-dark{
    background-color: #75767B;
}


.mt-50{
    margin-top: 50px;
}
#canvas{
    border: 2px solid black;
}
```

4.2.3. HTML Code

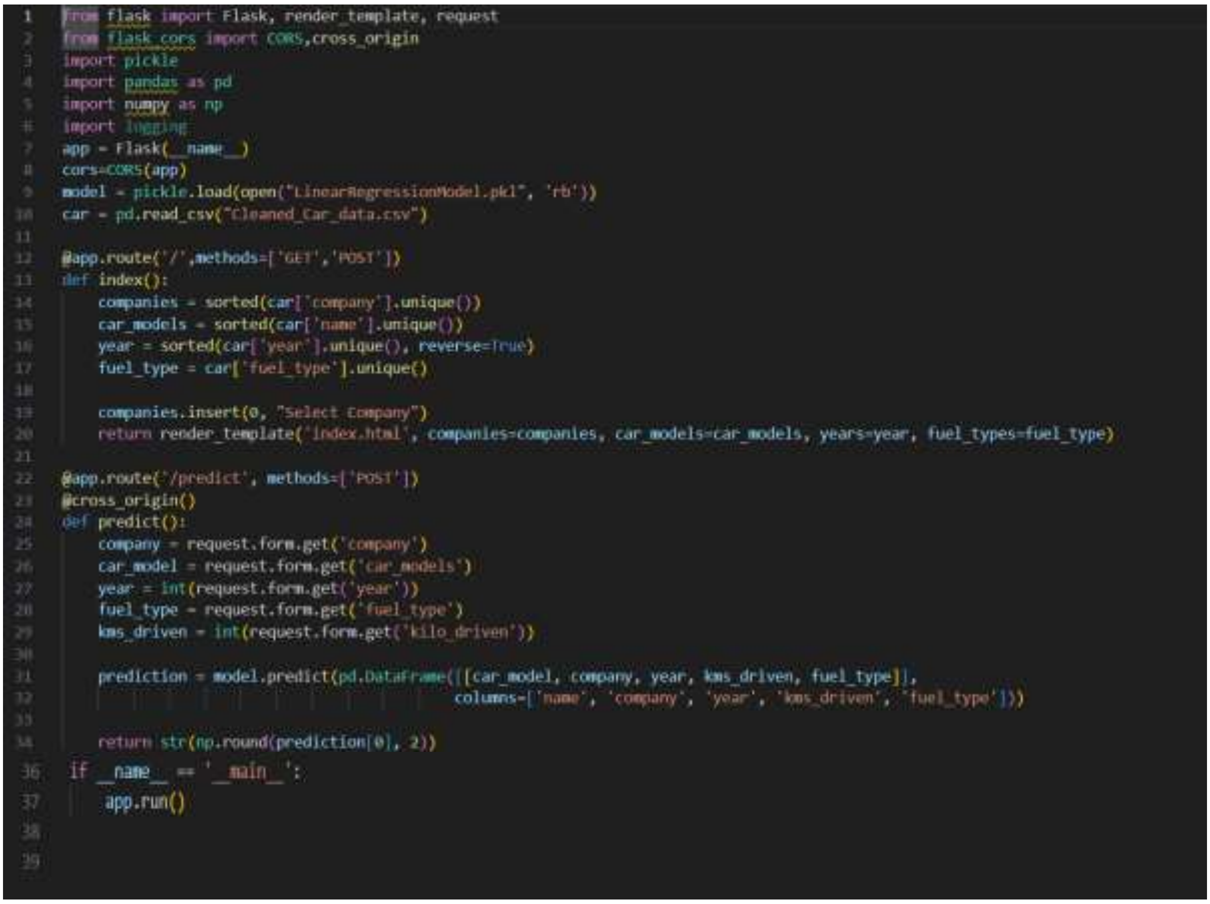


```html
<!DOCTYPE html>
<html lang="en">
<head xmlns="http://www.w3.org/1999/xhtml">
    <meta charset="UTF-8">
    <title>Car Price Predictor</title>
    <link rel="stylesheet" href="static/css/style.css">
    <link rel="stylesheet" type="text/css"
          href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/5.11.2/css/all.css">
    <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.4.1/jquery.min.js"></script>
    <script src="https://cdn.jsdelivr.net/npm/popper.js@1.16.0/dist/umd/popper.min.js"
            integrity="sha384-Q6E9RHvbIyZFJoft+2mJbHaEWldlvI9IOYy5n3zV9zzTtmI3UksdQRVvoxMfooAo"
            crossorigin="anonymous"></script>

    <!-- Bootstrap CSS -->
    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css"
          integrity="sha384-9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk" crossorigin="anonymous">
    <script src="https://cdn.jsdelivr.net/npm/@tensorflow/tfjs@2.0.0/dist/tf.min.js"></script>

</head>
<body class="bg-dark">

<div class="container">
    <div class="row">
        <div class="card mt-50" style="width: 100%; height: 100%">
            <div class="card-header" style="text-align: center">
                <h1>Welcome to Car Price Predictor</h1>
            </div>
            <div class="card-body">
                <div class="col-12" style="text-align: center">
                    <h5>This app predicts the price of a car you want to sell. Try filling the details below: </h5>
                </div>
                <br>
                <form method="post" accept-charset="utf-8" name="Modelform">
                    <div class="col-md-10 form-group" style="text-align: center">
                        <label><b>Select the company:</b> </label><br>
```



```
            if( company.value == "{{ company }}")
            {
                {% for model in car_models %}
                    {% if company in model %}

                        var newOption= document.createElement("option");
                        newOption.value="{{ model }}";
                        newOption.innerHTML="{{ model }}";
                        car_model.options.add(newOption);
                    {% endif %}
                {% endfor %}
            }
        {% endfor %}
    }

    function form_handler(event) {
        event.preventDefault(); // Don't submit the form normally
    }
    function send_data()
    {
        document.querySelector('form').addEventListener("submit",form_handler);

        var fd=new FormData(document.querySelector('form'));

        var xhr= new XMLHttpRequest({mozSystem: true});

        xhr.open('POST','/predict',true);
        document.getElementById('prediction').innerHTML="Wait! Predicting Price.....";
        xhr.onreadystatechange = function(){
            if(xhr.readyState == XMLHttpRequest.DONE){
                document.getElementById('prediction').innerHTML="Prediction: ₹"+xhr.responseText;

            }
        };

        xhr.onload= function(){};

        xhr.send(fd);
    }
</script>


<!-- jQuery first, then Popper.js, then Bootstrap JS -->
<script src="https://code.jquery.com/jquery-3.5.1.slim.min.js"
        integrity="sha384-DfXdz2htPH0lsSSs5nCTpuj/zy4C+OGpamoFVy38MVBnE+IbbVYUew+OrCXaRkfj"
        crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/popper.js@1.16.0/dist/umd/popper.min.js"
        integrity="sha384-Q6E9RHvbIyZFJoft+2mJbHaEWldlvI9IOYy5n3zV9zzTtmI3UksdQRVvoxMfooAo"
        crossorigin="anonymous"></script>
<script src="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/js/bootstrap.min.js"
        integrity="sha384-OgVRvuATP1z7JjHLkuOU7Xw704+h835Lr+6QL9UvYjZE3Ipu6Tp75j7Bh/kR0JKI"
        crossorigin="anonymous"></script>
</body>
</html>
        console.log(company.value);
        car_model.value="";
        car_model.innerHTML="";
        {% for company in companies %}
```

4.2.4. Application Code

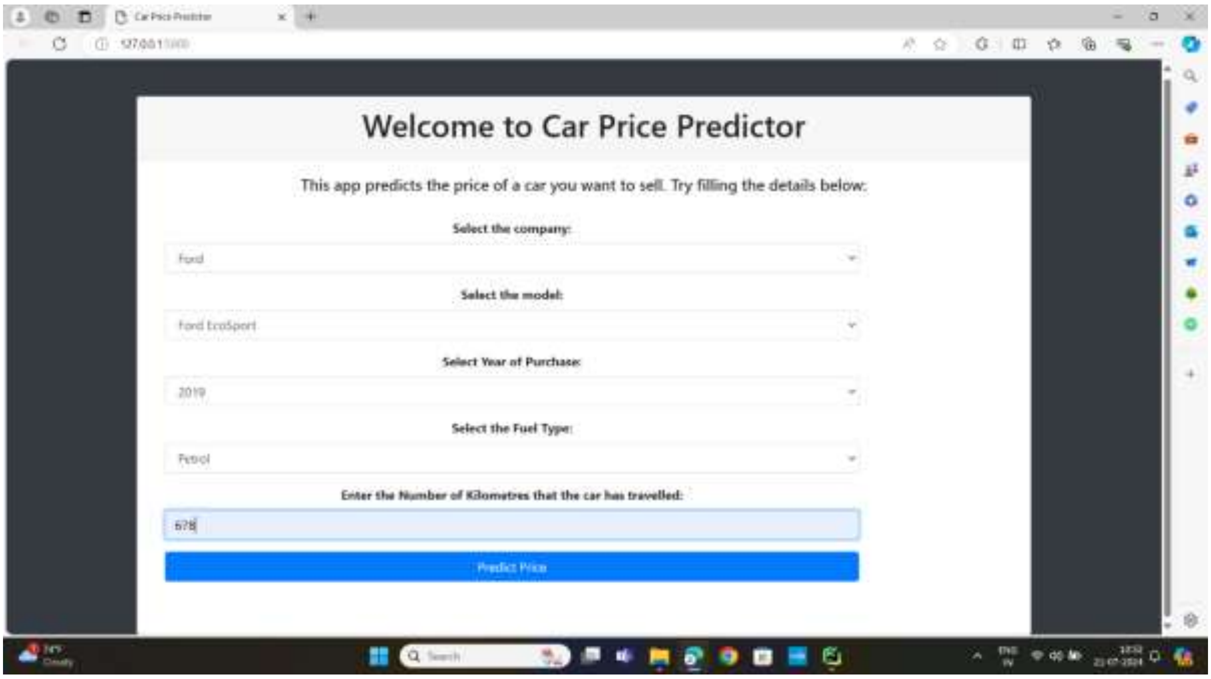


```python
from flask import Flask, render_template, request
from flask_cors import CORS,cross_origin
import pickle
import pandas as pd
import numpy as np
import logging
app = Flask(__name__)
cors=CORS(app)
model = pickle.load(open("LinearRegressionModel.pkl", 'rb'))
car = pd.read_csv("Cleaned_Car_data.csv")

@app.route('/',methods=['GET','POST'])
def index():
    companies = sorted(car['company'].unique())
    car_models = sorted(car['name'].unique())
    year = sorted(car['year'].unique(), reverse=True)
    fuel_type = car['fuel_type'].unique()

    companies.insert(0, "Select Company")
    return render_template('index.html', companies=companies, car_models=car_models, years=year, fuel_types=fuel_type)

@app.route('/predict', methods=['POST'])
@cross_origin()
def predict():
    company = request.form.get('company')
    car_model = request.form.get('car_models')
    year = int(request.form.get('year'))
    fuel_type = request.form.get('fuel_type')
    kms_driven = int(request.form.get('kilo_driven'))

    prediction = model.predict(pd.DataFrame([[car_model, company, year, kms_driven, fuel_type]],
                                            columns=['name', 'company', 'year', 'kms_driven', 'fuel_type']))

    return str(np.round(prediction[0], 2))
if __name__ == '__main__':
    app.run()
```

CHAPTER 5

RESULTS AND DISCUSSION

The Car Price Prediction project yielded promising results, demonstrating the efficacy of linear regression in estimating car prices based on specific features. The model achieved a high coefficient of determination (R² score), indicating a good fit to the training data, and metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) showed that the model could predict car prices with reasonable accuracy. Key features influencing car prices were identified, with the year of purchase and kilometers driven being the most impactful, followed by the car's company name and model. A user-friendly web application was developed using Flask, allowing users to input car details and receive instant price predictions. Initial user feedback indicated that the application is intuitive, easy to use, and provides helpful and reasonably accurate predictions, aiding in decision-making. The project showcases the practical application of machine learning in the automotive industry, providing a valuable tool for buyers, sellers, and dealerships, and highlighting the potential of data-driven insights to enhance transparency and fairness in the used car market.



Welcome to Car Price Predictor

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

Select the company: Ford

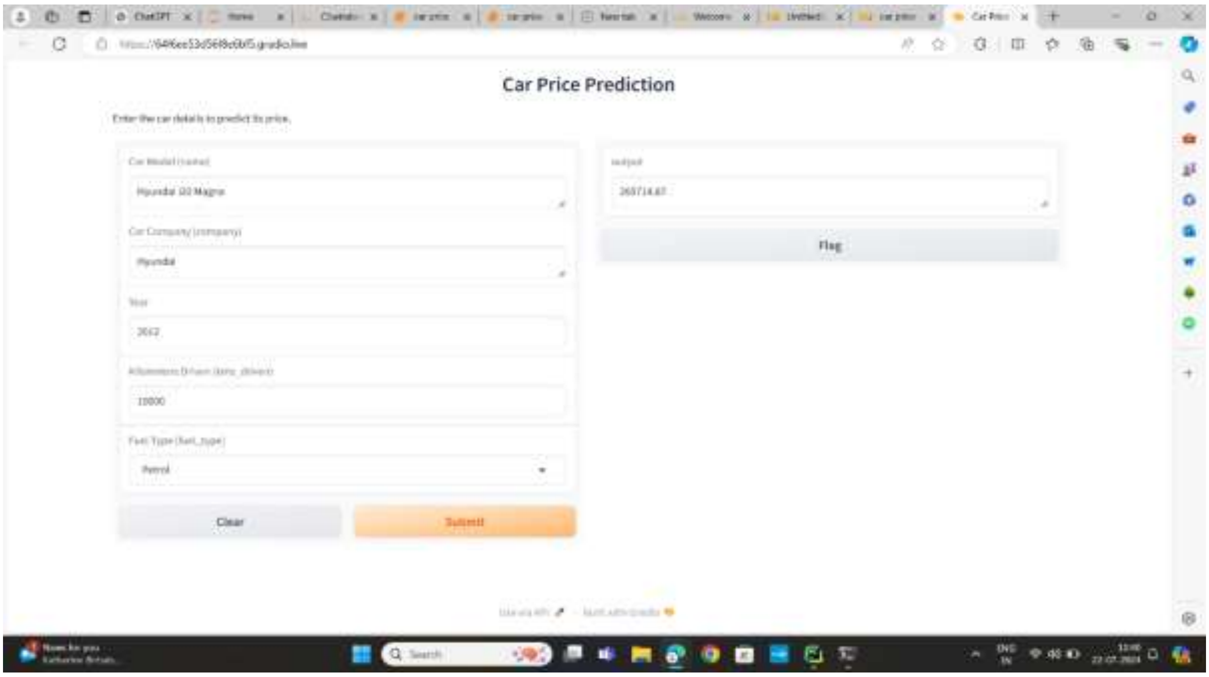
Select the model: Ford EcoSport

Select Year of Purchase: 2019

Select the Fuel Type: Petrol

Enter the Number of Kilometres that the car has travelled: 678

Fig. 5.1. The Web Page Created



https://6496ee53d5698c6bf5.gradio.live

Car Price Prediction

Enter the car details to predict its price.

Car Model (name)

Hyundai i20 Magna

Car Company (company)

Hyundai

Year

2012

Kilometers Driven (kms_driven)

10000

Fuel Type (fuel_type)

Petrol

output

265714.67

Fig. 5.2. Predicted Car Price

CONCLUSION AND FUTURE WORK

The Car Price Prediction project achieved significant success in developing a reliable linear regression model for estimating car prices based on key features such as company name, model, year of purchase, and kilometers driven. The model demonstrated a strong fit to the training data and produced accurate predictions, as evidenced by high R^2 scores and favorable MAE and RMSE metrics. The development of a Flask-based web application provided a user-friendly interface for individuals to input car details and obtain instant price estimates, which was well-received during initial user testing. This project underscores the practical application of machine learning in the automotive industry, offering a valuable tool that can aid buyers and sellers in making informed decisions, thereby contributing to greater transparency and fairness in the used car market. Looking ahead, there are several potential improvements and expansions for the project. Future work could involve incorporating additional features such as car condition, geographic location, fuel type, and even market demand trends to enhance the model's accuracy and robustness. Advanced machine learning algorithms, including ensemble methods or deep learning techniques, could be explored to improve predictive performance further. Implementing a system for continuous learning would enable the model to adapt to new data over time, maintaining its relevance and accuracy in a dynamic market. Additionally, the web application could be expanded to include features such as historical price trends, comparative market analysis, and tools for price negotiation, which would provide users with a comprehensive resource for car valuation. Integrating user feedback mechanisms and real-time data collection would further refine the model and application. Overall, these enhancements would not only improve the accuracy and utility of the car price prediction tool but also position it as an indispensable asset for a wider audience in the used car market.

REFERENCES

[1] Sustainability | Free Full-Text | Vehicle Price Classification and Prediction Using Machine Learning in the IoT Smart Manufacturing Era (mdpi.com)

[2] car_price_predictor/Quikr Analysis.ipynb at master · rajtilakls2510/car_price_predictor (github.com)

[3] https://www.geeksforgeeks.org/top-data-science-projects/#data-analysis-visualizations

[4] Linear Regression | SpringerLink

[5] Linear Regression - an overview | ScienceDirect Topics