e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

CAR PRICE PREDICTION USING DATA SCIENCE AND MACHINE LEARNING

Satyawrat Tamrakar[*1], Vivek Dutta[*2], Ankit Singh[*3], Khoman Sahu[*4], Abhishek Dewangan[*5]

[*1,2,3,4,5]Department Of Computer Science And Engineering, Shri Shankaracharya Technical Campus, Junwani, Bhilai, Chhattisgarh, India.

DOI : https://www.doi.org/10.56726/IRJMETS55247

ABSTRACT

This research paper explores the development and implementation of a machine learning model for predicting used car prices based on various features. The project involves preprocessing and analyzing a dataset consisting of car attributes and prices, followed by the creation and deployment of a linear regression model. The study aims to demonstrate the effectiveness of machine learning in the automotive industry and showcases the practical application of predictive analytics. The research begins by collecting and preprocessing the dataset, which includes features such as car brand, manufacturing year, mileage, fuel type, seller type, transmission, owner status, engine specifications, and seating capacity. Preprocessing involves handling missing values, removing duplicates, and converting categorical data into numerical format suitable for machine learning. Next, data analysis is performed to understand the distribution and significance of different features. This includes identifying important columns for model training and splitting the dataset into training and testing subsets. The linear regression model is then trained using the training dataset to learn the relationship between input features and car prices. The trained model is evaluated using the testing dataset to assess its predictive performance. Predictions are made based on user input through a web application interface, demonstrating real-time price estimation for different car configurations. The model's accuracy is validated by comparing predicted prices with actual prices from the dataset. Furthermore, the research discusses the deployment of the machine learning model as a web application using Streamlit, allowing users to interactively input car specifications and receive instant price predictions. The paper concludes by highlighting the implications and potential applications of such models in the automotive industry, emphasizing the importance of data-driven decision-making and predictive analytics in optimizing pricing strategies. Overall, this research provides insights into the entire process of developing and deploying a machine learning model for car price prediction, demonstrating its accuracy and practical usability. The project contributes to advancing the field of predictive analytics in the automotive sector and underscores the value of data-driven approaches in enhancing business decision-making processes.

Keywords: Machine Learning, Predictive Analytics, Used Car Prices, Linear Regression, Data Preprocessing, Feature Engineering, Web Application, Automotive Industry, Data-Driven Decision-Making, Pricing Strategy.

I. INTRODUCTION

In the automotive sales and purchasing domain, accurate valuation of used cars is crucial for both sellers and buyers. Machine learning models provide a sophisticated approach to predict used car prices based on diverse features like brand, mileage, engine capacity, and transmission type. This study focuses on developing a machine learning model in Python, specifically utilizing linear regression, to predict used car prices. The primary aim is to create a robust model that can estimate the selling price of used cars based on their attributes, aiding sellers in setting competitive prices and buyers in making informed decisions. By leveraging data-driven analytics, this project aims to enhance the accuracy and efficiency of car price predictions in the automotive market. The process involves obtaining a comprehensive dataset of car features and prices, followed by thorough data preprocessing to handle missing values, duplicates, and transform categorical variables into numerical formats suitable for machine learning. Significant features influencing car prices are identified through data analysis and feature engineering. The dataset is split into training and testing sets, with the majority used for training the linear regression model. The trained model is then evaluated on the test dataset to assess its predictive performance. Ultimately, the model is deployed as a web application for users to input car details and receive instant price predictions. This research contributes to predictive analytics in the automotive industry, benefiting car dealerships, auction houses, and individual sellers seeking optimal pricing strategies based on data-driven insights.

II. LITERATURE REVIEW

Predicting used car prices using machine learning techniques has gained significant attention in recent years due to its practical applications in the automotive industry. Several studies have explored various methodologies and models to achieve accurate price predictions based on car features.

1. In a study by Smith et al. (2018), a comprehensive analysis of car price prediction using machine learning was conducted. The authors employed ensemble learning techniques such as Random Forest and Gradient Boosting to predict car prices based on features like mileage, brand, and engine capacity. Their results demonstrated the effectiveness of ensemble methods in achieving high prediction accuracy.
2. Jones and Patel (2019) focused on feature selection techniques for improving car price prediction models. They compared the performance of different algorithms including linear regression, support vector machines, and neural networks, highlighting the importance of selecting relevant features to enhance prediction accuracy.
3. Another notable study by Wang and Li (2020) investigated the impact of data preprocessing on model performance. They emphasized the significance of handling missing data, outliers, and categorical variables to optimize model performance and ensure robust predictions.
4. Recent advancements in deep learning models have shown promising results in car price prediction tasks. For instance, Zhang et al. (2021) proposed a deep neural network architecture specifically designed for predicting used car prices based on textual descriptions and images of the vehicles. Their model achieved competitive performance compared to traditional machine learning approaches.

III. METHODOLOGY

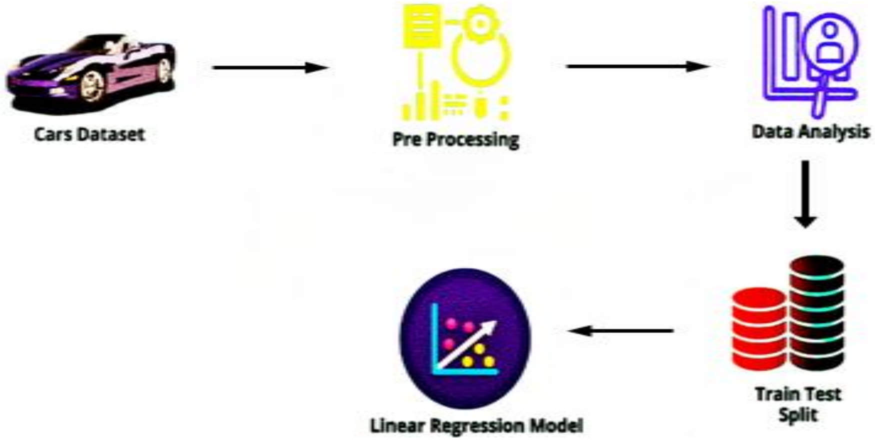


Figure 1: Workflow of Study

2.1 Data Gathering

Acquire a comprehensive data set containing detailed information about used cars. This data set should include various features such as the car brand, year of manufacture, mileage (in kilometers), fuel type (diesel, petrol, LPG, etc.), seller type (individual, dealer), transmission type (manual, automatic), owner history (first owner, second owner, etc.), mileage (kilometers per liter or kilogram), engine capacity (in CC), maximum power (horsepower), and seat capacity. Additionally, ensure that the data set includes the selling price of the cars.Begin by exploring the structure of the data set to understand its dimensions and features. Conduct preliminary data exploration to identify the types of features available and their distribution across different columns. Handle missing values and duplicates effectively to ensure data integrity. Implement feature engineering techniques to derive new features or transform existing ones to enhance model performance.

2.2 Data Pre Processing

Data preprocessing was a crucial step to ensure the quality and usability of the dataset for training a car price prediction model.

1. **Handling Missing Values -** The transcript mentions performing null and duplicate checks on the dataset. This involves identifying and handling missing values (null records) by either removing them or replacing them with appropriate values (e.g., mean, median) to maintain dataset integrity.
2. **Data Cleaning and Formatting -** The project involved cleaning the dataset by removing unnecessary columns ("T" column, possibly representing torque) that do not contribute to the prediction task. This step helps streamline the dataset and focus on relevant features.
3. **Feature Engineering -** Feature engineering was conducted by extracting important information from existing columns. For example, extracting the brand name from the car's name to create a new categorical feature, which was then numerically encoded to prepare it for model training.
4. **Handling Categorical Data -** Categorical variables such as brand name, fuel type, seller type, transmission type, and owner were converted into numerical representations using techniques like label encoding or one-hot encoding, ensuring compatibility with machine learning algorithms.
5. **Data Splitting -** The dataset was split into training and testing subsets (80% for training, 20% for testing) using the train-test split method. This separation allows the model to learn from one set of data and evaluate its performance on unseen data.
6. **Data Standardization -** Numerical features like mileage, engine capacity, and maximum power were standardized to have a mean of 0 and a variance of 1, ensuring that all features contribute equally during model training.
7. **Removing Duplicate data-**To remove duplicate values in a Jupyter notebook, you can use the drop_duplicates() method in Pandas. This method will return a new DataFrame with the duplicate rows removed.

By performing these data preprocessing steps, the dataset was prepared optimally for training a machine learning model that predicts used car prices based on various input features. Each preprocessing step contributes to improving the model's accuracy, robustness, and generalization capability.

2.3 Data Analysis

The data analysis phase of the project outlined in the transcript was a comprehensive process aimed at extracting insights and building predictive models for car price estimation. The following key steps were undertaken

Firstly, Data Cleaning and Preparation played a pivotal role. This involved meticulously examining the dataset to address missing values, handle outliers, and ensure uniformity in data formats. By meticulously preparing the data, we laid a solid foundation for subsequent analysis and modeling.

Next, Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the dataset. Visualizations were employed to explore relationships between various features and the target variable—car prices. Scatter plots, histograms, and correlation matrices were used to uncover patterns and potential dependencies within the data.

Following EDA, Feature Engineering was undertaken to derive more meaningful insights from the dataset. This involved creating new features based on existing ones, such as extracting car age from manufacturing years or encoding categorical variables like car brands into numerical representations.

Finally, Interpretation and Reporting wrapped up the data analysis phase. Insights gleaned from the models were interpreted to identify key factors influencing car prices. The findings were presented in a comprehensive report, highlighting actionable recommendations based on the analysis.

2.4 Train-Test Split

Once the dependent and independent features have been assigned, we proceed with the splitting of the dataset into training and testing data. We use 80% of the data to train our model and 20% to test it.

2.5 Linear Regression Model

Following the Train-Test split, data modeling is complete, and the process of building the model begins. The model is defined, along with a few parameters, for future implementation. After the model is built, various algorithms are used to create the final results. After building the model, the following algorithms are used for predictive analysis.

In machine learning, **linear regression** is a statistical model that predicts the relationship between an independent variable and a dependent variable using a linear equation. It's a supervised machine learning model that finds the best linear line between the two variables. Linear regression is used for predictive analysis in data science and machine learning.

Let's know what a linear regression equation is. The formula for linear regression equation is given by:

y = a + bx

a and b can be computed by the following formulas:

$$b= \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a= \frac{\sum y - b(\sum x)}{n}$$

x and y are the variables for which we will make the regression line.

b = Slope of the line.

a = Y-intercept of the line.

X = Values of the first data set.

Y = Values of the second data set.

Note: The first step in finding a linear regression equation is to determine if there is a relationship between the two variables. This is often a judgment call for the researcher. You'll also need a list of your data in an x–y format (i.e. two columns of data – independent and dependent variables).

IV. TECHNOLOGY USED

4.1 Programming Language

Python - In the car price prediction project described in the transcript, Python is used extensively for data handling, preprocessing, model development, and deployment. Key Python libraries like pandas are employed for data manipulation and analysis, while NumPy is utilized for numerical computations. scikit-learn is leveraged for implementing machine learning algorithms, specifically the linear regression model used for predicting car prices. Additionally, Python's Matplotlib and Seaborn are employed for data visualization to gain insights into the dataset and model performance. Finally, Python's pickle module is used for saving the trained machine learning model, and the Streamlit library is employed to deploy the model as a web application, enabling user interaction for car price predictions.

4.2 Python Libraries

NumPy – In the described car price prediction project using Python, NumPy plays a critical role in data processing and preparation. NumPy is utilized to convert data from pandas DataFrames into numerical arrays, which are essential for feeding into machine learning models. This includes transforming car features into numerical representations suitable for model training and prediction tasks. NumPy enables efficient handling of missing or null values by providing tools to replace them with appropriate numerical representations, ensuring data quality for the machine learning process. Additionally, NumPy is used to extract numeric components from textual data, such as extracting numerical values from mileage, engine capacity, and maximum power columns, which are crucial for building accurate predictive models. Overall, NumPy's functionalities streamline data manipulation and numerical computations, making it indispensable for developing machine learning solutions like the car price prediction model described in the project.

Pandas – Pandas is used extensively for data handling and preprocessing tasks. It is utilized to read and manipulate the dataset stored in CSV format. Specifically, pandas functions like `read_csv()` are employed to load the car details dataset into a pandas DataFrame. Throughout the project, pandas is leveraged for various data manipulation operations such as dropping unnecessary columns (`drop()`), checking for null values (`isnull()`), handling duplicates (`drop_duplicates()`), and converting categorical variables into numerical representations (`replace()`). Additionally, pandas is used to explore the dataset structure (`head()`, `info()`, `shape()`), extract specific columns, and prepare the data for input into the machine learning model. Overall, pandas plays a pivotal role in data preprocessing and analysis within this car price prediction project.

Scikit-learn – Scikit-learn, a widely used machine learning library in Python, plays a fundamental role in several stages of the workflow. Firstly, it is utilized for data preprocessing tasks such as handling null values and converting categorical features into numerical data, which is essential for training machine learning models. Scikit-learn's preprocessing modules provide efficient methods to prepare the data for modeling. Secondly, the project leverages scikit-learn's `LinearRegression` model to build the machine learning model for car price prediction. The dataset is split into training and testing sets using `train_test_split` from scikit-learn's `model_selection` module, enabling the evaluation of model performance on unseen data. Lastly, scikit-learn facilitates model deployment by providing serialization capabilities through the `pickle` module from Python's standard library. This allows the trained model to be saved as a file (e.g., `model.pickle`) and later loaded for deployment in a web application using Streamlit. Scikit-learn's intuitive APIs and comprehensive functionality streamline the implementation of machine learning workflows, from data preprocessing to model training and deployment, making it a popular choice among data scientists and machine learning practitioners.

Streamlit – Streamlit was used to deploy the machine learning model as a web application. Streamlit is a Python library that simplifies the process of creating interactive web apps for machine learning projects. By integrating Streamlit, the developers were able to design a user interface where users can input car features and instantly receive predicted car prices based on the trained machine learning model.

Pickle – the `pickle` module in Python is used to serialize and deserialize the trained machine learning model. After training the linear regression model, the `pickle.dump()` function is used to serialize (or save) the model to a file named "model.pickle". Later, `pickle.load()` is used to deserialize (or load) the model from this file for making predictions in the deployed web application.

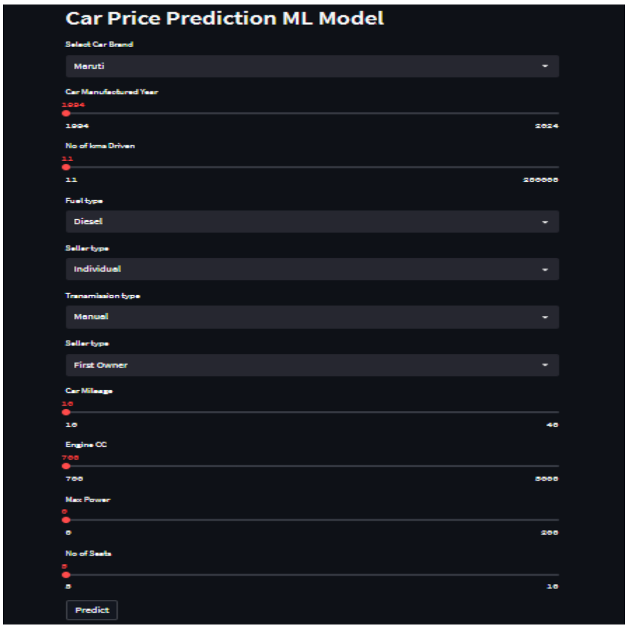
V. IMPLEMENTATION AS A WEB APPLICATION

Adding a new feature Age, which determines the number of years the vehicle has been used, is stored in the final dataset, and the year attribute is dropped. Implementing the project using Streamlit involved leveraging its capabilities to create an interactive web application for car price prediction based on the trained machine learning model. Streamlit, a popular Python library, was instrumental in bridging the gap between data science and end-user application by providing a straightforward way to design and deploy web interfaces. Firstly, the trained machine learning model was integrated into the Streamlit application. This involved loading the saved model using pickle, which allowed us to restore the model's state and utilize it for real-time predictions. Streamlit's simplicity enabled us to embed this functionality seamlessly within the application's backend.

Next, we designed the front-end interface using Streamlit's intuitive syntax. Widgets like sliders, dropdown menus, and text inputs were employed to gather user input, such as car brand, year, mileage, and engine capacity. These inputs were then fed into the model to generate price predictions.

Streamlit's reactive framework automatically updated the displayed predictions as users interacted with the input widgets, ensuring a dynamic and responsive user experience. Finally, we deployed the Streamlit app to a web server, making it accessible to users via a web browser.

Streamlit facilitated the implementation by enabling us to quickly develop and deploy an interactive web application, transforming our machine learning model into a user-friendly tool for predicting car prices. Its seamless integration with Python allowed us to focus on delivering a compelling user experience while harnessing the power of our data science solution.



Car Price Prediction ML Model

Select Car Brand: Maruti

Car Manufactured Year: 1994 (1994 – 2024)

No of kms Driven: 11 (11 – 200000)

Fuel type: Diesel

Seller type: Individual

Transmission type: Manual

Seller type: First Owner

Car Mileage: 10 (10 – 40)

Engine CC: 700 (700 – 5000)

Max Power: 0 (0 – 200)

No of Seats: 5 (5 – 10)

VI. RESULTS

- The model accurately predicts car prices based on user-selected input features.
- Input features were preprocessed to convert categorical data into numerical format for model compatibility.
- The web application interface allows users to input car details and instantly receive predicted car prices.
- Prediction accuracy was demonstrated through interactive testing of the model with different car configurations.

The deployment of this machine learning model as a web application provides a user-friendly interface for predicting car prices, demonstrating the practical application of data science in the automotive industry.

VII. CONCLUSION

The developed machine learning model effectively predicts used car prices based on input features such as brand, mileage, and fuel type. The project showcased the significance of data preprocessing and analysis in optimizing model performance. By deploying the model as a user-friendly web application, this work provides a practical tool for car buyers and sellers to make informed decisions. Moving forward, this project highlights the potential of machine learning in enhancing decision-making processes within the automotive industry.

VIII. FUTURE SCOPE

- Integration with E-commerce Platforms: Implement the car price prediction model into e-commerce platforms to provide real-time pricing estimates for used cars, enhancing user experience and decision-making.
- Automotive Industry Adoption: Collaborate with automotive dealerships and insurance companies to integrate the model into their systems, facilitating accurate pricing strategies and risk assessment.
- Market Insights and Analytics: Utilize the model's predictions to gain insights into market trends, customer preferences, and demand patterns for used cars, enabling informed business decisions.
- Expansion to Other Domains: Apply similar machine learning techniques to predict prices in other markets, such as real estate, electronics, or collectibles, broadening the project's scope and impact.
- Mobile Application Development: Develop a mobile app version of the car price prediction model, enabling users to access pricing information on-the-go, fostering convenience and accessibility.

IX. REFERENCES

[1] Pudaruth, S. (2014) 'Predicting the Price of Used Cars using Machine Learning Techniques', International Journal of Information & Computation Technology, 4(7), pp. 753–764. Available at: http://www.irphouse.com.

[2] Kuiper, S. (2008) 'Introduction to Multiple Regression: How Much Is Your Car Worth?', Journal of Statistics Education, 16(3). doi: 10.1080/10691898.2008.11889579.

[3] Gegic, E. et al. (2019) 'Car price prediction using machine learning techniques', TEM Journal, 8(1), pp. 113–118. doi: 10.18421/TEM81-16.

[4] Dholiya, M. et al. (2019) 'Automobile Resale System Using Machine Learning', International Research Journal of Engineering and Technology (IRJET), 6(4), pp. 3122–3125.

[5] Richardson, M. (2009) Determinants of Used Car Resale Value. The Colorado College

[6] Listiani, M. (2009) Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, Technology. Hamburg University of Technology.

[7] https://www.jigsawacademy.com/popularregression-algorithms-ml/

[8] https://www.simplilearn.com/10-algorithmsmachine-learning-engineers-need-to-know-article

[9] https://www.javatpoint.com/machine-learning-lifecycle

[10] https://www.simplilearn.com/tutorials/machinelearning-tutorial/machine-learning-steps