**CAR DHEKO - USED CAR PRICE PREDICTION**

****

A machine learning-based solution to predict used car prices and deploys an interactive web application for users.

Submitted By,

MIRTHU BAASHINI B

Data Science Student

Batch – MDE95

GUVI GEEK NETWORKS

**ABSTRACT**

The Car Dheko - Used Car Price Prediction project aims to enhance customer experience and streamline the pricing process for used cars. By leveraging machine learning, we developed an accurate and user-friendly tool that predicts used car prices based on features such as make, model, year, fuel type, transmission type, and more. Historical data from multiple cities were consolidated, cleaned, and preprocessed to create a structured dataset.

Key steps included handling missing values, encoding categorical variables, scaling numerical features, and removing outliers. Various machine learning models, including Linear Regression, Decision Trees, Random Forests, and Gradient Boosting, were trained and evaluated using relevant performance metrics.

The Random Forest model was found to be the best-performing, with the following results:

* **Mean Absolute Error (MAE):** 72,268.15
* **Mean Squared Error (MSE):** 14,373,609,930.47
* **R-squared (R²):** 0.941

The final model was deployed as an interactive Streamlit application, allowing users to input car details and receive real-time price predictions. This tool empowers customers and sales representatives with an efficient solution for estimating used car prices.

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TABLE OF CONTENTS** | **PAGE.NO** |
| 1. | Introduction | 4 |
| 2. | Problem Statement and Objective | 5 |
| 3. | Data Overview | 6 |
| 4. | Methodology | 7 |
| 5. | Exploratory Data Analysis (EDA) | 8 |
| 6. | Model Development | 9 |
| 7. | Results and Discussion | 9 |
| 8. | Deployment | 10 |
| 9. | Challenges and Limitations | 10 |
| 10. | Conclusion | 11 |
| 11. | References | 11 |

**INTRODUCTION**

The used car market plays a pivotal role in the automotive industry, where pricing accuracy is crucial for both consumers and sellers. In traditional methods, pricing is often subjective, relying on the knowledge and judgment of the seller, which can lead to inconsistencies and inefficiencies. As the demand for used cars continues to rise, there is an increasing need for a more data-driven approach to determine fair and accurate pricing. This project aims to address these challenges by developing a machine learning model that can predict the prices of used cars based on various attributes, including the car’s make, model, year, fuel type, transmission type, and several other relevant factors. The goal is to build a system that provides accurate, real-time price predictions, simplifying the buying and selling experience for both customers and sales representatives.

In this project, we utilize historical data from Car Dekho, a popular platform for buying and selling used cars in India. The dataset contains comprehensive details about various used cars listed in multiple cities, including specifications, features, and price details. The raw data, however, is often unstructured, so one of the first tasks is to process and transform it into a structured format. We then apply machine learning techniques to create a predictive model that can estimate the price of a used car based on the input features. To make this model accessible and user-friendly, we integrate it into a Streamlit-based web application. This application will allow users to input the details of a car and receive an instant price prediction, making it a valuable tool for both prospective buyers and sales teams. Ultimately, the project aims to improve the efficiency of the pricing process, enhance customer satisfaction, and facilitate better decision-making within the used car market.

### ****PROBLEM STATEMENT AND OBJECTIVE****

**PROBLEM STATEMENT**

Determining the price of a used car is a complex task, influenced by several factors such as the car’s make, model, year, condition, kilometers driven, and more. Traditional pricing methods often rely on subjective judgment, leading to inconsistencies and inefficiencies in the used car market. This project aims to build a machine learning model that predicts the prices of used cars based on various attributes to offer a more consistent, accurate, and objective method of pricing.

**OBJECTIVE**

The goal of this project is to develop a machine learning-based system that predicts used car prices using historical data and various features like the car’s make, model, fuel type, and kilometers driven. By providing a real-time price prediction tool, the system aims to streamline the pricing process for customers and sales representatives, making it more data-driven and efficient. This system is further integrated into an interactive **Streamlit-based web application**, allowing users to input car details and get price predictions instantly.

**DATA OVERVIEW**

The dataset used in this project is collected from **Car Dekho**, a popular online platform for buying and selling used cars in India. It includes various attributes about used cars listed for sale across multiple cities. The dataset consists of both structured and unstructured data that requires preprocessing before being fed into machine learning models.

Key attributes in the dataset include:

* **Car Specifications:** Make, model, year of manufacture, fuel type, transmission type, and body type.
* **Condition Variables:** Number of previous owners, kilometers driven, and car’s condition.
* **Price Data:** Listed price, actual price, and price saving information (if available).
* **Other Features:** Includes details such as OEM, car variant, trending data, and more.

The raw data contains missing values, outliers, and categorical features that require encoding before being used in modeling. The dataset is cleaned and preprocessed to convert it into a structured format suitable for machine learning.

**METHODOLOGY**

This section describes the approach followed to solve the problem of predicting used car prices.

**Problem Definition**

The objective is to predict the price of a used car based on its attributes. This is a regression problem, where the goal is to estimate the price as a continuous value.

**Feature Engineering**

The features selected include attributes such as make, model, fuel type, kilometers driven, transmission type, and car condition. Categorical variables were encoded using **Label Encoding** and **One-Hot Encoding** techniques. Numerical variables were scaled using **StandardScaler**.

**Model Selection**

Various machine learning models were used to build the prediction model. These include:

* Linear Regression
* Decision Trees
* Random Forest
* Gradient Boosting

**Model Training and Evaluation**

The models were trained using a train-test split (80-20) and evaluated using **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared (R²)** as evaluation metrics.

**EXPLORATORY DATA ANALYSIS (EDA)**

EDA is an important step in understanding the dataset and the relationships between variables. During this phase:

**Visualizations**

Plots like histograms, scatter plots, and box plots were used to visualize the distribution of variables, correlations, and outliers.

**Key Insights**

Some key insights from the data included:

* Car prices show a strong correlation with the car’s age, kilometers driven, and brand.
* Certain features like body type and fuel type have clear distributions that influence the price prediction.
* Outliers were detected in the price and kilometers driven columns, which were handled during preprocessing.

**MODEL DEVELOPMENT**

In this section, the development of machine learning models is discussed.

**Model Training**

We trained several models, including Random Forest, Linear Regression, and Decision Trees. The **Random Forest** model performed the best, providing the most accurate predictions.

**Cross-validation**

Cross-validation was used to ensure that the model generalizes well on unseen data.

**Hyper parameter Tuning**

For models like Random Forest, hyper parameter such as the number of trees, max depth, and min samples split were optimized using Grid Search.

**RESULTS AND DISCUSSION**

**Model Performance**

The **Random Forest** model was the best performer with the following results:

* **MAE:** 72,268.15
* **MSE:** 14,373,609,930.47
* **R²:** 0.941

**Model Evaluation**

The performance of the Random Forest model indicates that it is highly accurate in predicting used car prices. The model is reliable and robust, which makes it suitable for real-time predictions in an interactive web application.

**DEPLOYMENT**

The final model was deployed as a **Streamlit** web application.

**Streamlit Application**

The application allows users to input car details (make, model, year, fuel type, transmission, etc.) and receive real-time price predictions.

**User Interface**

The user-friendly interface is designed to be intuitive, ensuring that users can easily input the required information and view the predicted price.

**CHALLENGES AND LIMITATIONS**

**Data Issues**

Some data preprocessing challenges included handling missing values, encoding categorical variables, and dealing with outliers.

**Model Limitations**

The model may not generalize well to extremely rare car models or unusual combinations of features.

**CONCLUSION**

The Car Dheko – Used Car Price Prediction project successfully developed a machine learning-based solution to predict the prices of used cars, utilizing a range of features such as make, model, year, fuel type, and transmission type. After data preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features, several machine learning models were trained and evaluated. Among the models tested, the Random Forest Regressor performed the best, achieving an R-squared value of 0.941, indicating a high level of accuracy in price predictions.

The model’s strong performance demonstrates the effectiveness of machine learning in automating the pricing process in the used car market. The integration of the model into an interactive **Streamlit-based web application** provides an accessible tool for users to input car details and receive real-time price predictions. This enhances the decision-making process for both prospective buyers and sellers. While the model performed well, future work could focus on incorporating additional features, exploring advanced algorithms like deep learning, and integrating real-time data updates for continuous model improvement.

Additionally, expanding the dataset to include more cities and rare car models would further improve the model’s generalizability. Overall, this project demonstrates the potential of data-driven approaches to streamline the pricing process, improve transparency, and increase customer satisfaction in the used car market.

**REFERENCE**

<https://docs.google.com/document/d/1zTFaw5oacPGUkVbxyhSsv5tYcr4U5fAhr2DoS7Ey9-Q/edit?tab=t.0>

**DECLARATION**

I declare that, this Documentation is prepared by Mirthu Baashini B, Data Science Student at Guvi Geek Network, Batch MDE95.