AutoML: A Machine Learning Approach for Car Price Prediction

Mihir V Panchal

A Student of Computer Science Engineering with Specialization in Artificial Intelligence & Machine Learning

Project Name: AutoML

Project Date: September, 2024

Aim of the Report

The aim of this report is to present a comprehensive analysis and implementation of a machine learning model designed for predicting the prices of second-hand cars. By utilizing a synthetic dataset, the report will explore the relationships between various car features—such as manufacturer, model, year of purchase, fuel type, and mileage—and their impact on pricing. The report seeks to demonstrate the effectiveness of the developed model in providing accurate price predictions, thereby aiding potential buyers in making informed purchasing decisions. Additionally, it will discuss the methodologies employed in data preparation, model training, and integration with a user-friendly web interface.

Abstract

This report details the development of a machine learning model for predicting the prices of second-hand cars, utilizing a synthetic dataset generated from key vehicle features. The dataset includes variables such as manufacturer, model, year of purchase, fuel type, and mileage. Through exploratory data analysis (EDA), we examined the relationships among these features, identifying significant patterns and trends that influence car pricing. A linear regression model was chosen for its simplicity and effectiveness in predicting continuous outcomes. The model was evaluated using the R^2 score, achieving a commendable accuracy of 0.84. Furthermore, the report describes the integration of the machine learning model with a static HTML website, developed using Flask, which allows users to input car details and receive price predictions in real-time. This project aims to provide potential buyers with valuable insights, facilitating informed decision-making in the second-hand car market.

1. Introduction

The automotive market has seen a significant shift towards online platforms where users can buy and sell second-hand cars. As potential buyers often face challenges in estimating fair prices for vehicles, this project aims to leverage machine learning to provide accurate price predictions for second-hand cars. Using a dataset from Quickr, this project seeks to assist users by creating a model that predicts car prices based on key features such as the manufacturer, model, year of purchase, fuel type, and the total kilometers driven. The goal is to empower consumers with data-driven insights, enhancing their purchasing decisions.

2. Data Description

The dataset utilized in this project is sourced from Quickr, a well-known online marketplace for buying and selling various products, including automobiles. The dataset comprises several critical features that are instrumental in determining the price of second-hand cars:

- **Company**: This feature identifies the manufacturer of the car, which can significantly influence its resale value due to brand reputation and reliability.

- **Model**: The specific model of the car plays a crucial role in price differentiation, as certain models may be more sought after than others.

- **Year of Purchase**: The age of the car is a key determinant of its value; typically, newer cars fetch higher prices compared to older models.

- **Fuel Type: Different fuel types** (e.g., petrol, diesel, electric) can impact pricing due to varying demand and operational costs associated with each type.

- **Number of Kilometers:** This feature reflects the usage of the vehicle, with higher kilometers generally indicating greater wear and tear, which can lower the car's value.

Together, these features provide a comprehensive view of the factors affecting second-hand car prices, making them essential for effective price prediction.

3. Exploratory Data Analysis (EDA)

Summary of EDA

The Exploratory Data Analysis (EDA) phase was crucial in understanding the dataset and preparing it for model training. Initially, the data was examined for completeness and accuracy, revealing some issues that required attention. This involved data cleaning, where duplicate entries and irrelevant columns were removed to streamline the dataset.

Missing value imputation was another vital step in the EDA process. Various strategies were employed to fill in gaps in the data, ensuring that the model would not be hindered by incomplete information. Techniques such as mean, median, or mode imputation were considered based on the nature of the missing data.

Insights from Data Visualization

Data visualization played an essential role in EDA, enabling the identification of trends and relationships within the dataset. Visual tools such as scatter plots, histograms, and correlation matrices were used to depict how features interacted with one another and their impact on car prices. For instance, visualizing the relationship between the year of purchase and price helped illustrate the depreciation curve typical in the automotive market, providing insights into how car age correlates with its resale value.

4. Methodology

Machine Learning Tools and Techniques

The methodology employed in this project involved several key machine learning tools and techniques:

- **Train-Test Split**: The dataset was divided into training and testing subsets to ensure the model could be evaluated effectively. The training set was used to train the model, while the testing set provided an unbiased assessment of its performance.

- **Linear Regression Model:** A linear regression model was selected due to its straightforward nature and effectiveness in predicting continuous outcomes. This model assumes a linear relationship between the independent variables (features) and the dependent variable (price).

- **One-Hot Encoding:** Given that some features were categorical (such as Company and Fuel Type), one-hot encoding was utilized to convert these categories into a numerical format that the model could interpret. This process allows the model to understand the significance of different categories without assuming any ordinal relationships.

- **Pipeline:** To streamline the modeling process, a pipeline was established. This pipeline integrated the various preprocessing steps and the model training into a single cohesive workflow, improving efficiency and ensuring consistency.

- **Pickle:** After training the model, the `pickle` library was employed to save the trained model to a file. This allows for easy loading and reuse of the model without the need for retraining, which is especially useful for deployment.

Preprocessing Steps

In addition to one-hot encoding, various preprocessing steps were undertaken to prepare the data for model training. This included normalization of numerical features to bring them onto a similar scale, enhancing the model's performance by improving convergence during the training phase. By carefully processing the data, the integrity and quality of the input to the machine learning model were ensured.

5. Implementation

Static HTML Website

To provide a user-friendly interface for users, a static website was designed using HTML and CSS. The website allows potential buyers to enter relevant information about the car they are interested in, such as the manufacturer, model, year of purchase, fuel type, and kilometers driven. The clean and simple design ensures that users can navigate easily and input their data without confusion.

Integration with Flask

The static website is seamlessly integrated with the machine learning model through the Flask framework, a lightweight web application framework for Python. By utilizing Flask, the model can respond to user inputs in real-time. The `render_template` function is employed to dynamically display the prediction results based on the information provided by the user. This integration not only enhances the interactivity of the application but also provides a practical way to showcase the machine learning model’s capabilities.

User Input Processing

Once users submit their details through the web form, the input is processed by the backend Python code. This includes converting the user inputs into the appropriate format, applying the same preprocessing steps as used during model training, and then passing the processed data to the trained model for prediction. The predicted price is then returned and displayed to the user, providing instant feedback on their input.

6. Results

Model Performance

The model's performance was evaluated using the R^2 score, which measures how well the model's predictions match the actual values. The achieved R^2 score of 0.84 indicates that the model explains 84% of the variance in the car prices, reflecting a strong predictive capability. This level of accuracy is significant in practical applications, where precise price predictions can aid buyers in making informed decisions.

Example Predictions

To illustrate the model's functionality, an example prediction was conducted. For a user inputting details for a 2018 Toyota petrol car with 20,000 kilometers, the model predicted a price of approximately $25,000. This prediction showcases the model’s ability to account for various features and provide a reasonable estimate, facilitating more confident purchasing decisions.

7. Conclusion

This project successfully demonstrates the application of machine learning in the context of car price prediction, providing a valuable tool for potential buyers in the second-hand car market. By utilizing a robust dataset and implementing effective machine learning techniques, the model achieves a commendable level of accuracy. Future work may involve exploring additional features, experimenting with more advanced algorithms, or expanding the model’s capabilities to include real-time market data. Such enhancements could further refine the model's predictions, ultimately leading to a more comprehensive solution for car buyers.

8. References

- Quickr dataset documentation
- Relevant machine learning literature and resources