

Report on Used Cars Price Prediction

By

POOJITHA VUPPALAPATI – 1339322

SALOMI KRISHNAMURTHY - 1325463

SRI SAKTICHARAN NIRMAL KUMAR – 1337576

MEGHA VAMSHI KRISHNA VEMALA – 1325438

SHRUTHI MANCHAPPANAHALLI BASAVARAJ – 1330334

**ABSTRACT**

In the automobile market, estimating the cost of used cars is crucial for both buyers and sellers. This paper investigates the process of utilizing machine learning techniques to develop a forecast model for used automobile prices. Information on many features, including make, model, year, mileage, and condition, is gathered during data collecting. Encoding categorical variables and managing missing data are two important tasks that require feature engineering in order to prepare meaningful input variables for the model. For price prediction, a variety of machine learning methods can be used, such as neural networks, random forests, gradient boosting, and linear regression. The size and complexity of the dataset are among the criteria that influence the model choice. Evaluation measures are used to evaluate the correctness of the model using Root Mean Squared Error (RMSE). The model may be utilized in the real world to assist consumers in making knowledgeable decisions while shopping for secondhand cars after it has been trained and assessed.

**INTRODUCTION**

An important part of the automobile sector that affects both buyers and sellers is the prediction of used car pricing. It is now easier to predict a used car's pricing with accuracy because of advancements in machine learning algorithms and data availability. Offering insights into this intricate yet crucial market element, this research looks at the approaches employed in building a predictive model for used automobile prices.

Numerous factors, including make, model, age, mileage, condition, and location, have an impact on the used car market. Accurate price estimating requires an understanding of these factors and the incorporation of these variables into a predictive model. Furthermore, because the market is dynamic, strong models that can adjust to shifting circumstances and trends are required.

Predictive modeling is based on data collecting, which necessitates thorough information from sources including internet listings, car dealerships, and previous sales data. The gathered data contains both categorical and numerical features, therefore feature engineering and rigorous preprocessing are required to get it ready for model training.

Choosing a suitable machine learning algorithm for price prediction is crucial after data preparation. Different trade-offs in complexity and predictive accuracy are available for algorithms such as neural networks, random forests, gradient boosting, and linear regression. The choice of algorithm is influenced by various parameters, including processing resources, feature complexity, and dataset size.

From data collection and preprocessing to model selection and evaluation, we examine every step of the procedure in this study before deploying a predictive model for used car price prediction. With an eye on utilizing data-driven methods for efficient used car pricing, the goal is to offer insights and useful advice to stakeholders in the automotive sector.

**MOTIVATION:**

The necessity to handle several significant issues and satisfy the needs of both consumers and sellers in the automotive business is what drives the projection of used car prices.

Informed Decision-Making for Buyers: When purchasing a used car, buyers frequently lack clarity regarding the vehicle's fair market worth. Because they provide precise pricing projections based on a variety of variables, including car age, mileage, condition, and market trends, predictive models help consumers make well-informed selections. This guarantees that consumers receive the best value for their money and helps them avoid overpaying.

Optimized Pricing Strategy for Sellers: To draw in potential purchasers, sellers strive to maximize their profits while maintaining a competitive price. With the help of predictive algorithms, sellers can determine the best pricing to charge depending on the state of the market and the particulars of their cars. Sellers can shorten the time it takes to sell cars and increase profits by correctly pricing their inventory.

Transparency and Market Efficiency: By reducing the information asymmetry that exists between buyers and sellers, predictive models promote transparency in the market. Clear pricing promotes confidence and trust in the used automobile market, which facilitates easier transactions and increases market competition.

**BACKGROUND:**

The used car market is a complex and dynamic environment influenced by various factors:

*Vehicle Heterogeneity:* Used cars are available in a variety of types, models, ages, mileage, and conditions. The worth of the car is influenced by each of these elements, which makes it difficult to develop uniform pricing strategies.

*Market dynamics:* Several factors, including supply and demand, the state of the economy, and seasonal trends, affect used car prices. For predictive models to produce precise pricing projections, these dynamics must be taken into consideration.

*Data Availability:* There is a wealth of information about used automobile sales thanks to the growth of internet platforms, dealership databases, and market research publications. Robust predictive models can be created by utilizing machine learning and statistical modeling techniques to leverage this data.

**DATA DESCRIPTION:**

**A screenshot of a computer

Description automatically generated**

* The Dataset is taken from Kaggle.
* Link:https://www.kaggle.com/code/nitiruengcharoen/price-prediction-of-bmw-used-car
* The dataset used for this project consists of historical data on used car listings. It includes information such as the make, model, year, mileage, condition, location, and price of each car. The data was collected from various online sources, including classified ads websites and car dealership websites.
* Scale = 10,782 rows × 9 columns
* model (Categorical): The make and model of the car.
* year (Numerical): The year the car was manufactured.
* mileage (Numerical): The total distance the car has traveled (miles).
* mpg (Numerical): Miles per gallon (fuel efficiency).
* Fuel Type (Categorical): The type of fuel the car uses (e.g., Diesel, Petrol).
* transmission (Categorical):transmission type(Automatic, Manual).
* tax (Numerical): The tax amount for the car (in US dollars).
* engine Size (Numerical): The size of the car's engine (in liters).
* price (Numerical): The price of the car (in US dollars).

**TECHNOLOGY USED :**

* **Python**: In this project, Python is the main programming language used for data collecting, preprocessing, training, evaluating, and deploying models. NumPy, pandas, scikit-learn, Matplotlib, and Seaborn are some of the libraries that are utilized for different tasks.
* **NumPy**: NumPy is a Python library that offers numerous mathematical functions for effective numerical computing, as well as support for sizable, multi-dimensional arrays and matrices.
* **Pandas**: This is used in data manipulation activities like feature transformation, categorical variable encoding, and management of missing values.
* **Matplotlib**: It is used in this project to visualize model performance and investigate correlations between features and target variables.
* **Seaborn**: In this project, Seaborn is utilized for advanced data visualization to enhance the exploration of relationships between features and target variables, as well as to create visually appealing and informative plots to aid in model evaluation.

**METHODOLOGY:**

**Data** **Collection**: Consult internet listings, dealerships, and past sales data to learn about the brand, model, year, mileage, condition, and location of used automobiles, among other things.

**Data Preprocessing**: To get the data ready for modeling, clean it up and perform preprocessing operations such as addressing missing values, encoding category variables, and scaling numerical characteristics.

**Feature Extraction**: Generate new features, modify existing ones, and choose pertinent characteristics that are most predictive of the target variable (vehicle price) in order to extract useful insights from the data.

A screenshot of a computer program

Description automatically generated

**Random Forest Modelling**: By constructing an ensemble of decision trees and training each tree on a different subset of the data and features, random forest modeling produces a robust and accurate predictive model for used car price prediction. The final prediction is calculated as the average (regression) or majority vote (classification) of the predictions made by each individual tree.

**Training Set:** To reduce prediction mistakes and maximize its capacity to correctly estimate used car values, the machine learning model in this project's training phase learns the patterns and relationships in the data by modifying its parameters.

**Testing Set**: During the project's testing phase, the performance of the trained model is assessed using a different testing dataset to make sure it can generalize to predict used car prices and make correct predictions on data that hasn't been seen before.

**Model Evaluation**: To determine whether the trained model can generalize to new data, measure its performance using metrics like Root Mean Squared Error (RMSE) on a different testing dataset.

**Prediction**: Use the trained model in the real world by incorporating it into a web application that allows customers to enter the details of a used automobile and get the estimated price.

**DATA VISUALIZATION:**

**A group of different colored bars

Description automatically generated**

**Bar charts:** Show the frequency of categorical parameters such as manufacturer, model, and body type (SUV, sedan, etc.).

Overall, this graph provides a convenient way to quickly visualize the relationships between different features and the target variable in the dataset.

* **Price vs. Mileage:** A bar plot can show how a car's price tends to decrease as the mileage increases (negative correlation).
* **Price vs. Year:** This bar plot would likely show a negative correlation between price and car year, indicating depreciation.
* **Price vs. Features:** You can visualize the price against features like horsepower, number of doors, or safety rating to see if these features influence price.

By analyzing these scatter plots, you can identify trends that might be useful for predicting used car prices. For instance, a car with a higher mileage or older year might be predicted to have a lower price based on the trends in the scatter plot.

**CORRELATION MATRIX:**

**A screenshot of a graph

Description automatically generated**

A table that lists the direction and intensity of each association between pairs of attributes in your data collection is called a correlation matrix. Every cell displays a correlation coefficient, with 0 denoting no association and ranging from -1 (completely negative correlation) to +1 (absolutely positive correlation).

It is an excellent instrument for comprehending the ways in which various attributes (such as mileage, year, or engine size) affect a target variable (used car pricing) in your investigation.

It is possible to determine which characteristics have the strongest correlations with used car prices by examining a correlation matrix such as this one. Using this data, models for used automobile pricing prediction can be constructed. For example, since these have the strongest attributes, you may concentrate on features like age, mpg\_age\_ratio, eng\_efficiency, and mileage.

**EXPERIMENT AND RESULTS :**

A screenshot of a computer

Description automatically generated

It contains data on multiple features of used cars and their predicted prices. Here's a breakdown of the data:

* **Columns:**
* price: The actual selling price of the car
* price\_predict: The predicted price of the car using a model (possibly based on other features)
* model: The car model (e.g., 3 Series, X5, Z4)
* transmission: The car's transmission type (Automatic, Semi-Auto, Manual)
* mileage: The car's mileage
* fuelType: The car's fuel type (Diesel, Petrol)
* tax: The car tax
* engineSize: The car's engine size
* age: The car's age in years
* mpg: The car's miles per gallon
* eng\_efficiency: The car's engine efficiency
* mpg\_age\_ratio: The ratio of the car's miles per gallon to its age
* eng\_age\_ratio: The ratio of the car's engine size to its age
* **Rows:** Each row represents data for a specific used car. There are multiple cars listed (possibly 9 based on the partial data visible).

**CONCLUSION:**

To sum up, the creation of predictive models for used car pricing meets important demands in the automotive industry, which is advantageous to both buyers and sellers. These models enable sellers to maximize their pricing strategies and purchasers to make well-informed selections by offering precise price predictions based on a variety of parameters, including market trends and car features. The history of the used automobile market highlights its complexity, highlighting the wide range of vehicles it offers, its dynamic pricing dynamics, and the wealth of information that is at its disposal. Predictive models provide an answer to these problems by improving market efficiency and transparency using machine learning and statistical modeling approaches.

Predictive modeling approaches must be progressively developed and refined in the future to satisfy the changing demands of the automobile sector. This means adjusting to shifting market conditions, improving algorithms, and incorporating new data sources. Future used car pricing predictions could be more accurate and efficient as technology develops, which would be advantageous to all parties involved. Finally, used car pricing prediction models are essential for building confidence, enhancing judgment, and guaranteeing honest and open pricing procedures in the used automobile market.

**REFERENCES:**

www.datacamp.com

https://github.com/datacamp/careerhub-data

https://pandas.pydata.org/docs/reference/frame.html

https://scikit-learn.org/stable/index.html

https://towardsdatascience.com/random-forest-ca80e56224c1

**CONTRIBUTION**:

Poojitha Vuppalapati (ID: 1339322):

Data Collection & Preprocessing: Gathered data from online sources and handled preprocessing tasks such as missing value treatment and categorical variable encoding.

Sri SaktiCharan Nirmal Kumar (ID: 1337576):

Feature Extraction & Model Evaluation & Project idea inspiration : Engineered relevant features and assessed model performance using metrics like RMSE.

Salomi Krishna Murthy (ID: 1325463):

Model Building & Data Visualization: Constructed the random forest regression model and created visualizations (e.g., bar charts, scatter plots) to analyze feature-price relationships.

Megha Vamshi Krishna Vemala (ID: 1325438):

Correlation Analysis & Experiment Interpretation: Analyzed the correlation matrix to identify influential features and interpreted experiment results to draw insights.

Shruthi Manchappanahalli Basavaraj (ID: 1330334):

Project Inspiration & Technology Overview: Contributed to project idea , highlighting the importance of predicting used car prices, and provided insights into the technology stack used.