I.J. Information Technology and Computer Science, **2026, 1, 124-139**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2026.01.07

Car Price Analysis Using Data Collected from an Online Sales Platform

Bui Quang Phu
Faculty of Computer Science, University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
E-mail: 20520273@gm.uit.edu.vn
ORCID iD: https://orcid.org/0009-0002-9724-2921

Pham Hoang Phuc
Faculty of Computer Science, University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
E-mail: 20520278@gm.uit.edu.vn
ORCID iD: https://orcid.org/0009-0000-4335-4501

Pham The Son
Faculty of Information Science and Engineering, University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
E-mail: sonpt@uit.edu.vn
ORCID iD: https://orcid.org/0000-0003-1959-2838
*Corresponding author

Received: 19 June 2025; Revised: 08 October 2025; Accepted: 11 November 2025; Published: 08 February 2026

Abstract: In this paper, we aim to develop a car price prediction model using data collected from an online sales platform. To accomplish the proposed objective, we applied the following approaches and techniques: (1) Collecting sales data from the online sales platform; (2) Exploratory analysis of data before and after data preprocessing; (3) Experimenting to find a suitable prediction model for the collected dataset. The novelty of this study lies in constructing a real-world dataset of pre-owned car prices collected directly from an online sales platform and in building a car price prediction model using an empirical approach combined with machine learning models. Unlike previous studies based on existing structured datasets, this study emphasizes the discovery of data-driven insights through exploratory analysis and the identification of key variables affecting car prices. At the same time, essential insights regarding car prices were obtained from the dataset. Experimental results show that the model using the XGBoost algorithm achieved an R^2 of 0.776 for the default parameter case and an R^2 of 0.779 for the optimized parameter case. These findings provide a practical solution for real-world car price prediction systems, allowing buyers and sellers to make more informed pricing decisions.

Index Terms: Data Collection, Predictive Analytics, Price Prediction Analysis, Exploratory Data Analysis, EDA, Extreme Gradient Boosting, XGBoost, MSE.

1. Introduction

In recent years, data analytics has become an integral part of data-driven decision-making. In particular, it plays a vital role in big data and artificial intelligence. Among the various analytical methods, predictive analytics has received considerable attention due to its ability to forecast future outcomes and support strategic business decisions. Predictive analysis helps us achieve many essential benefits, especially in the e-commerce industry, by predicting to support business decision making, assisting managers to make critical decisions based on past data[1,2]. In addition, predictive analysis can help us identify customer behaviors and create predictive models of customer purchasing behavior[3].

However, despite the growing research interest in predictive analytics, studies focusing on real-world datasets collected directly from local online marketplaces remain limited. Most previous research has used pre-structured or open datasets, which may not fully capture the diverse and unstructured nature of real-world market data. Therefore, there is still a research gap in developing and validating predictive models on authentic user-generated datasets that reflect real-world market behavior.

This research aims to address that gap by focusing on the problem of used car price prediction in the Vietnamese online market. Among the many benefits that predictive analysis brings, we are particularly interested in the problem of

price prediction analysis in the e-commerce industry, enabling sellers and buyers to propose reasonable prices based on product characteristics. In this paper, we solve the problem of car price prediction to determine a suitable and well-founded price, avoiding the situation where buyers or sellers offer a price that is too low or too high[4,5]. Based on the analyzed data, sellers can optimize their sales strategy and determine the best time to sell their cars.

To achieve this proposed objective, we collected raw data on used car prices at ChoTot Car Market (available at: https://xe.chotot.com) using the raw collection method. Then, we processed the data and explored it multiple times to gain a deeper understanding of the dataset. Based on the results of the exploratory data analysis, we developed a suitable model to predict used car prices using new data samples.

During the experiment, we utilized supporting libraries such as scikit-learn, NumPy, Pandas, and many other libraries to facilitate preprocessing, exploratory data analysis, and model development. Through numerous experiments, we found that a suitable model for the collected dataset is XGBoost, an ensemble model that utilizes multiple decision trees to make the most accurate predictions. This paper has contributed to the field of predictive analytics by (1) building a dataset of used car prices collected from an online sales platform; (2) finding a model that can predict car prices based on new data. However, the prediction results of the model are still different, but not significant; (3) finding meaningful insights that show the relationships between car prices and car information.

For ease of follow-up, this paper is organized as follows: Section I introduces the study's main objectives and the problem to be addressed. Section II provides an overview of previous related research. Section III presents the data collection method and describes the dataset that was collected. Section IV proposes the data processing and analysis methods. Section V empirically proposes the most suitable predictive model. Finally, Section VI concludes and summarizes the contributions of the paper.

2. Related Works

Predictive analysis is one of the key problems in data science, applied in many fields. In recent years, there have been many scientific works using machine learning algorithms to solve problems related to predictive analysis. Especially in the field of car price prediction, there are many related studies applying machine learning to build car price prediction models to help buyers and sellers estimate car prices. In this section, we present an overview of previous studies related to building car price prediction models.

Ahmad et al. [6], in the paper *"Car Price Prediction Using Machine Learning"*, the authors studied and built a model to predict car prices based on vehicle specifications. The method focuses on machine learning models, including Linear Regression, AdaBoostRegressor, Lasso Regression, and Ridge Regression. The study is based on a dataset collected from online car trading floors and auction dealers. In addition, the study integrates additional information on customer reviews, market trends, and regional factors. The results of the Lasso Regression and Ridge Regression models achieved an accuracy of nearly 90% in predicting car prices. The limitation of the study does not mention the possibility of economic fluctuations. The impact of factors such as tax policies, fuel prices, or geographical factors has not been clarified. However, this study did not conduct an in-depth feature analysis to identify variables affecting the car price prediction results. Still, it only focused on the process of comparing the algorithm and model performance. The study did not clearly evaluate the model's generalization ability when applied to the market or real-world data sets. This gap suggests that further research is needed to develop adaptive models that are suitable for regional car price fluctuations.

Gupta et al. [7], in the study *"A Machine Learning Approach for Predicting Price of Used Cars and Power Demand Forecasting to Conserve Non-Renewable Energy Sources"*, proposed a model to automatically predict used car prices to help buyers avoid being unreasonably priced by sellers. The method uses machine learning techniques to predict car prices based on previous user data and car characteristics based on a collected dataset, and preprocesses the data to create 4000 car models. The result is a model with an accuracy of 92.38%. The study provides valuable insights into car price forecasting and power demand forecasting. Two separate, unrelated problems will reduce the urgency of the car price forecasting problem. The research objectives are quite broad, which may affect the generalizability of the analysis. These issues highlight a research gap that can assess how power demand influences car prices over time.

Budiono et al. [8], in the study *"Used Car Price Prediction Model: A Machine Learning Approach"*, developed a used car price prediction model, helping buyers and sellers easily price cars based on characteristics such as year of manufacture, car type, car condition, etc. In the study, web scraping techniques were applied to collect 504 used car data samples from car sales websites as input data for the model. Using K-Nearest Neighbors (KNN), Regression Model to build the model. The results showed that the KNN model achieved a low error rate of 8.3% and an R^2 value of 98.8%. Limitations, the study was based on only 504 data samples, which is a relatively small data set, not to mention the ability to expand the model with information such as geographic area, market fluctuations, and seasonal factors. Based on the presented results, the study has high prediction accuracy, but it depends on the dataset. Therefore, it limits the model's generalizability to other markets. Additionally, the study has not been compared with other advanced models, such as XGBoost or Random Forest. The study did not analyze other important influencing variables, such as fuel prices, that may affect vehicle values over time. Therefore, future work should address these limitations by using more advanced ensemble models, such as XGBoost or Random Forest, and incorporating larger and more diverse datasets.

Jin et al. [9], in their paper titled *"Price Prediction of Used Cars Using Machine Learning"*, aimed to build a model to predict the reasonable price of used cars based on characteristics such as mileage, year of manufacture, fuel

consumption, transmission, road tax, fuel type, etc. This model aims to support sellers, buyers, and manufacturers in the used car market. The method used in the study includes machine learning and analytical techniques in data science. The data set used in the study was collected from used cars. Before building the model, the data was visualized and segmented to optimize the performance of the regression models. The result was that using random forest regression, the highest R-square was 0.90416, indicating high predictive ability. However, the study mainly focused on traditional regression models without applying new methods such as gradient boosting that can help improve the accuracy of prediction. Although the survey removed outliers and applied preprocessing, it did not evaluate the reliability of the model across vehicle brand variables. The paper did not analyze the importance of variables in the dataset, leading to difficulties in interpreting which variables have the most decisive influence on car price prediction. These gaps highlight opportunities for further research, more general and interpretable models in the future.

Das Adhikary et al. [10], in the paper *"Prediction of Used Car Prices Using Machine Learning"*, studied and built a model to predict used car prices in the Indian market, helping consumers better understand price trends. The method used in the article employs machine learning algorithms to predict used car prices, including K-Nearest Neighbor, Random Forest Regression, Decision Tree, and Light Gradient Boosting Machine (LightGBM). The results found the best model and compared it with other models. The model is capable of predicting used car prices with high accuracy. In this study, mainly focusing on model comparison, not explaining the results in detail, limits the model to real data. In addition, the model may not generalize well to new markets due to the region-specific data set. In the future, it can be extended by integrating automatic data collection techniques and combining cross-market data sets to enhance the data connectivity and usability of the model.

Nandan et al. [11], in the study *"Pre-Owned Car Price Prediction by Employing Machine Learning Techniques"* developed a price prediction system for used cars, which helps buyers predict the value of used cars based on the characteristics of the car. The method used in this work is the regression technique in machine learning, specifically Linear Regression, LASSO Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting. The result is a car price prediction model. In which the error values of MAE, MSE, and RMSE are compared. The study has shown that the optimal model helps predict used car prices more accurately. However, the study mainly focused on the algorithm performance without analyzing the potential influence of variables on car prices. The paper did not have a strategy for handling missing values, nor did it evaluate the impact of data imbalance that could affect the model performance. Therefore, there is still a research gap in developing explainable car price prediction models.

Longani et al. [12], in the paper *"Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques"* developed a fair price prediction system for used cars in the Mumbai region to ensure fairness between sellers and buyers. The paper's approach is based on ensemble techniques in machine learning using the Random Forest Algorithm and eXtreme Gradient Boost algorithms to develop a prediction model. The results of both methods achieved similar performance in predicting the prices of used cars. However, eXtreme Gradient Boost gave better results as measured by Root Mean Squared Error (RMSE) of 0.53, which is much lower than 3.44 of the Random Forest Algorithm. However, the study did not address the generalizability of the model to regions other than Mumbai. Although the approach demonstrates good predictive performance, the study focuses on comparing algorithms, but the model lacks the ability to interpret or analyze the importance of variables. Moreover, the dataset is limited in terms of region, making it challenging to explore the impact on price trends. Future studies can extend this line of research by incorporating multi-regional datasets to assess different price trends from region to region.

Benabbou et al. [13], in the study *"Machine Learning for Used Cars Price Prediction: Moroccan Use Case"* built a used car price prediction system in Morocco to support users in making reasonable pricing decisions. The approach used in the study is supervised learning algorithms; collecting data using a web crawler, then applying preprocessing techniques such as feature selection and listwise deletion. Finally, using XGBoost to train the model and deploying it using the Flask web framework. The XGBoost model achieved high performance with $R^2 = 90.7\%$, showing good predictive ability. However, the model is only applied to predict car prices in Morocco, and its performance has not been evaluated when applied to other markets. The study presents the integration of the process from data collection to model deployment, but it has not evaluated the suitability of each algorithm for the dataset. The results may not be generalizable because the dataset may not represent the diversity of the automobile market. Future studies should incorporate market characteristics to improve the reliability and adaptability of the model.

Wu et al. [14], in the paper *"An Expert System of Price Forecasting for Used Cars Using Adaptive Neuro-Fuzzy Inference"* built an expert system to predict used car prices based on factors such as car brand, year of manufacture, engine type, and car equipment. The approach in the paper focuses on using an adaptive neuro-fuzzy inference system (ANFIS) to predict used car prices. To evaluate the effectiveness of ANFIS, this model is compared with an artificial neural network (ANN) with a back propagation (BP) algorithm. Experimental results show that ANFIS has better car price prediction ability than the ANN model. The study did not mention factors such as location, car condition, or market fluctuations. The relatively small dataset in the study is limited to a single online platform, which may limit the generalizability of the model to a larger market. The study focuses on the accuracy of the model and analyzes its performance. The lack of integration of economic or spatial data may affect the practical applicability of the model. This also leaves room for further studies to evaluate ANFIS-based models using larger datasets.

Synthesizing related studies shows that most works focus on comparing the accuracy between machine learning algorithms, but have not analyzed in depth the importance of input variables affecting car prices. Many models are built on data sets limited in scale or geographical scope, leading to low generalizability when applied to other markets. Other

factors such as socio-economic indicators, tax policies, fuel prices, and regional factors are hardly integrated into the prediction model. In addition, few studies pay attention to model interpretability and handling missing values, which significantly affect the reliability of prediction results. Therefore, the next research direction needs to focus on developing a prediction model with good adaptability, combined with feature analysis.

3. Dataset Construction

ChoTot (available at: https://www.chotot.com/) is a prominent and established e-commerce platform in Vietnam. The web-based system allows users to post products for sale in many different fields. One of them is a field specializing in cars (ChoTot Car Market), which allows users to post vehicles for sale, such as motorbikes, electric vehicles, used cars, etc. (available at: https://xe.chotot.com). We focus on collecting data on used cars, and this is the primary source of data that we collect and analyze. In the ChoTot e-commerce web system, there are many types of users selling products. Among the many types of users, we only collect data from users who have been identified by the system as partners. We select the news for sale labeled "Partner" for the following reasons: ChoTot Car Market is an e-commerce website system. On the system, there will be many users registered to post for sale in many product categories. Registering a user to sell is very easy and free. These users are at risk of posting fake sales information. Therefore, the system allows users to verify the reliability of users through personal documents. Authenticated users will post true sales information. Sales information from authenticated users will be labeled "Partner". In short, the information of "Partner" posted for sale will be more trustworthy than the remaining types of users. In this section, we present the data collection process and describe in detail the collected raw dataset.

3.1. Data Collection Process

We have collected used car sales data from ChoTot Car Market. The raw dataset includes information about the seller, detailed information about the car, car images, and especially the car price. From the collected raw dataset, we conducted exploratory analysis experiments many times to better understand the dataset, detect errors, and abnormalities in the data. The data collection process is carried out according to the following procedure:

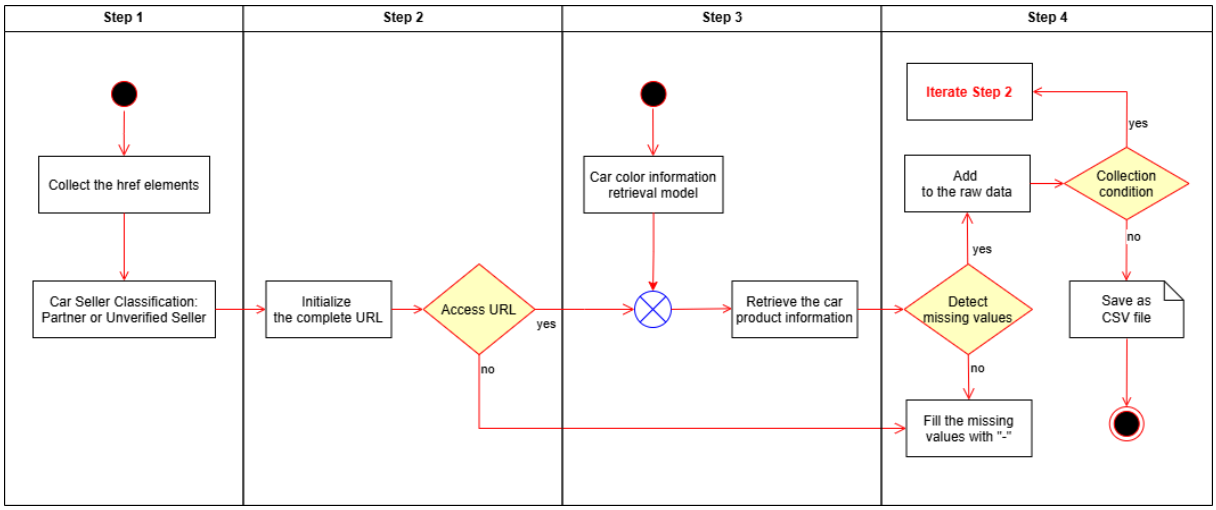

Fig.1. Data collection process

The process of collecting data on buying and selling used cars at ChoTot Car Market is designed and presented through the following detailed steps Fig.1:

- Step 1: The objective of this step is to get the href information of each car sale post and classify the car seller. First, access ChoTot Car Market (available at: https://xe.chotot.com/mua-ban-oto) to get a list of cars for sale. Next, focus on getting: (1) The href tag information used to access the detailed information page of each car; (2) whether the car seller is a partner or not. We filter out partner sales posts to obtain accurate information. Because the ChoTot Car Market platform has authenticated users to make the sales information more trustworthy, it avoids fake sales posts. Therefore, authenticated users will post more trustworthy car sales information than unauthenticated users. In this step, we will filter out the sales posts of authenticated users.
- Step 2: The task of this step is to initialize the complete URL of the car product detail page and access that URL to enter the car product detail page. First, after collecting enough hrefs, connect to the root URL after "https://xe.chotot.com/" to create a complete URL to access each car product. Next, access each complete URL to connect to the product detail page of each car and retrieve the necessary information. If it can be accessed, we will go to the next step. If not, we will fill in the symbol "-" for all the attributes that need to

be collected in the dataset.

- Step 3: When accessing the product detail page, we will get all the information about the car product. In the car product detail page, there is an information section about the car images. We use an additional color recognition model with data of car images to create additional car color information. Data is collected from the car detail information page, such as: title, seller name, color, seller classification (this information is important because we know whether the seller is a partner or not), car price, address, car brand, car model... If any value is missing, fill in the symbol "-". Regarding the car color data, we perform the following additional testing method to consistently represent the data: (Method 1) Access the car description section of the car listing information page. Next, get the description information, remove special characters, and non-letter symbols. Then, extract color information from the product description; (Method 2) If method 1 fails, proceed to method 2. Use the car object prediction model in the image to identify the color of the car.
- Step 4: To collect more data, return to Step 1. Then, the next page URL is dynamically generated using the following URL pattern:

$$\text{https://xe.chotot.com/mua-ban-oto?page=<page\_number>} \quad (1)$$

where <page_number> represents the index of the next page to be collected.
Repeat the URL pattern in (1) until the required amount of data is reached or all the listing pages have been collected. Finally, save all raw data as a .csv file for further data analysis.

3.2. Dataset Description

The initial raw dataset consists of 21 columns (attributes) and 2688 rows, including 4 numerical attributes and 17 categorical attributes. The descriptions and meanings of these attributes are presented in the following table:

Table 1. Attribute descriptions

No.	Attribute	Data Type	Description
1	Color	String	Car color
2	Title	String	Car sale listing title
3	Sales user	String	The name of the car seller or the name of the car dealership.
4	Seller classification	String	Seller Type (Individual, Semi-professional)
5	Address	String	Contact address of the seller
6	Make	String	Car make (Toyota, Ford, …)
7	Model	String	Car model or series (Corolla, …)
8	Year	Integer	Car production year
9	Odometer	Integer	Number of kilometers traveled based on the odometer
10	Condition	String	Car condition (new, pre-owned)
11	Transmission	String	Transmission type of the car (automatic, manual, …)
12	Fuel type	String	Fuel type of the car (gasoline, diesel, …)
13	Body style	String	Car body style or design (sedan, …)
14	Seat capacity	Integer	The seat capacity of the car
15	Country	String	Car manufacturing country
16	Warranty	String	Information on the car warranty policy
17	Weight	String	Weight of the car
18	Load capacity	String	Maximum load capacity of the car
19	Partner	String	Whether the car seller is a partner or not
20	Href	String	URL of the car sale post
21	Price	Integer	The sale price of the car

In the raw dataset we collected, there are 5 attributes with missing values, as follows:

- "Seller" attribute: 1007 missing values.
- "Odometer" attribute: 284 missing values.
- "Body style" attribute: 366 missing values.
- "Seat capacity" attribute: 311 missing values.
- "Price" attribute: 3 missing values.

The dataset used in this study consists of 2,688 samples, which is smaller than some existing public datasets. However, this data is collected directly from an online car trading platform, reflecting the actual and updated market situation instead of synthetic data. Although the data size is limited, we have applied many measures to ensure the reliability of the results, including thorough data preprocessing and cleaning; feature normalization and noise removal; cross-validation and hyperparameter tuning to avoid overfitting. In addition, the dataset includes many different brands, car models, and price levels, ensuring diversity and representation of the car market within the scope of the study. This study mainly aims to propose and verify the feasibility of a car price prediction model based on real data.

4. Data Analysis

After completing the data collection process, we proceed to process and analyze the data, with the stages performed according to the process Fig.2. For each stage in the process, we present details about the implementation process as well as the results achieved in each stage.

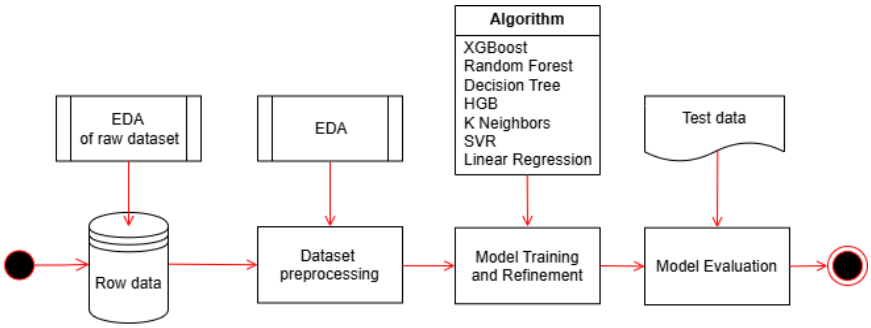

Fig.2. Data analysis and predictive modeling process

In the process of analyzing and building a predictive model Fig.2, we proceed in two main stages. First, conduct exploratory data analysis (EDA) and data cleaning. Specifically, we will conduct an EDA before and after data preprocessing. Because after preprocessing, some attributes will be removed because they do not have predictive value in the model. Therefore, when conducting EDA before preprocessing to consider the attributes that can be removed, and exploratory analysis after preprocessing, explore the relationship between the attributes with each other. Additionally, the exploratory analysis process helps determine the method for handling missing and invalid values. Second, after having a clean dataset, we conduct experiments to find the most reasonable price prediction model for the dataset.

4.1. Exploratory Analysis of the Raw Dataset

For the raw dataset, the values of the data samples are mostly categorical variables, so we will initially apply visualization techniques to explore and find meaningful insights. Specifically, we focus on three types of visualizations, including pie charts, bar charts, and combined line and bar charts. The result of this EDA is a data dashboard Fig.3, Fig.4.

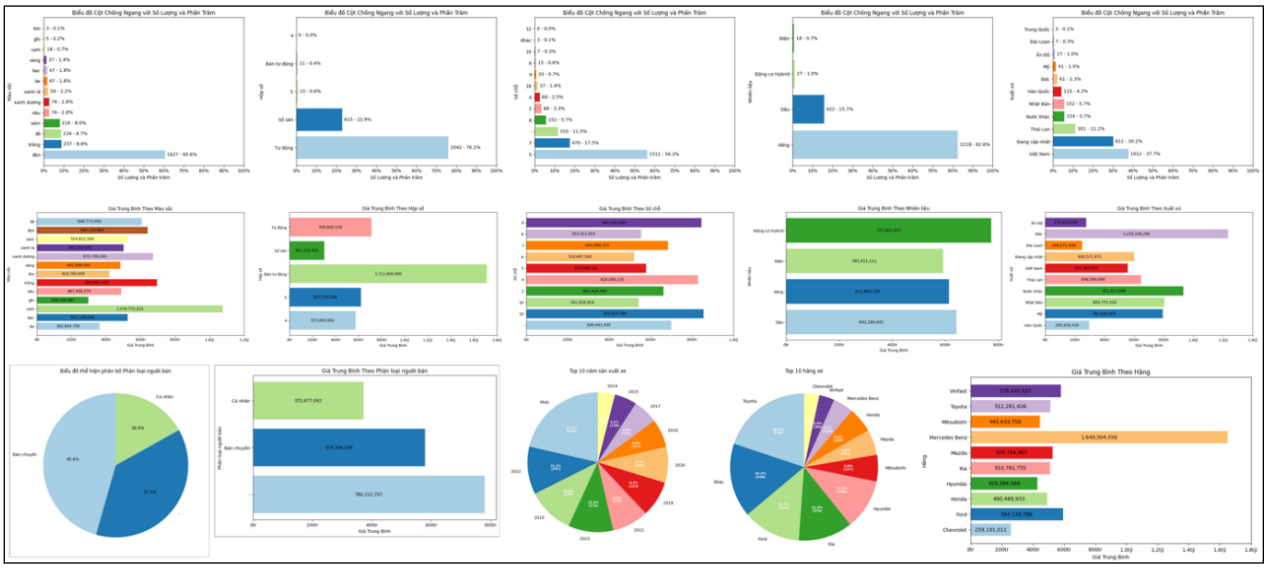

Fig.3. Exploratory analysis dashboard for the raw dataset

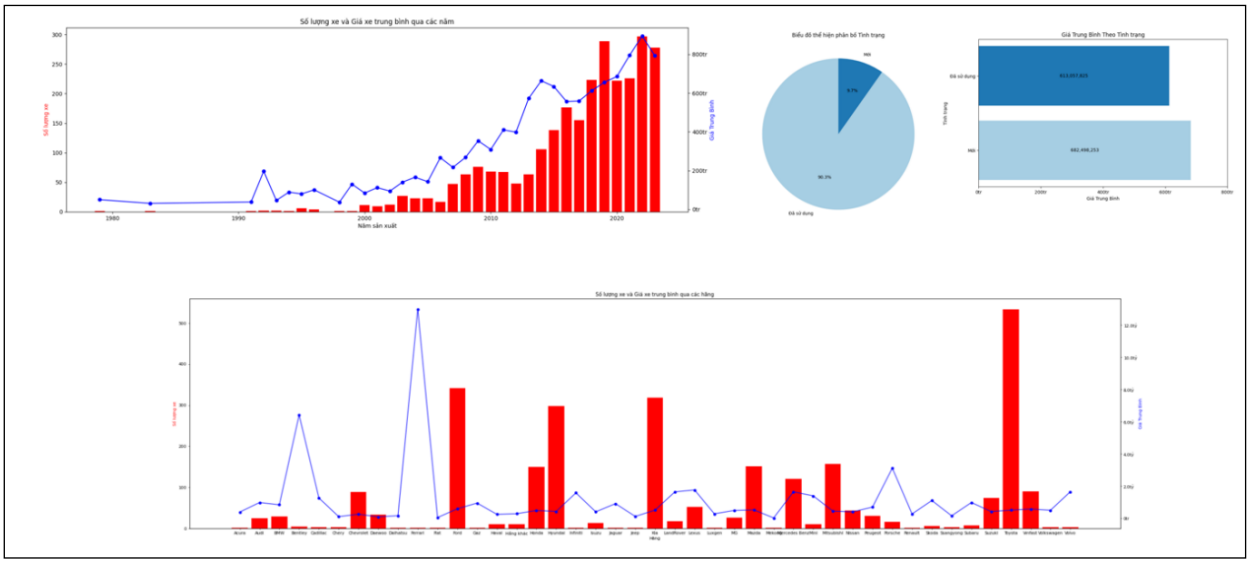

Fig.4. Exploratory analysis dashboard for the raw dataset with the attributes “Year”, “Condition”, and “Make”

From the data dashboard Fig.3, Fig.4, we can draw some observations as follows:

- The “Color” attribute: The popular color of cars sold on ChoTot Car Market is mainly black, accounting for 60.6% of the total. However, black is a fairly popular color, and when combining prices, we see that the prices of black cars are concentrated at an average price. Among the car colors in the dataset, orange cars stand out, with the highest average price (approximately 1 billion 76 million VND). In fact, we see that orange is the color most commonly used for luxury cars, so the price is high.
- The “Seller classification” attribute: Based on the visual results, we see that the number of shops and car dealers is larger than that of individual car dealers. Car dealers account for 45.6% of the total number of car dealers. Combined with the car price, car dealers have a higher average price than individual car dealers. From this insight, we can predict that car dealers will often sell many types of cars, and the cost of maintaining and managing additional cars will result in higher prices.
- The “Make” attribute: Based on the visual results of car manufacturers, Toyota cars account for the majority, about 19.9%. In addition to Toyota, there are other popular brands such as Ford 12.7%, Kia 11.9%, Hyundai 11.1%, Mitsubishi 5.8%, Mazda 5.6%, Honda 5.5%, Mercedes-Benz 4.5%, Vinfast 3.4%, Chevrolet 3.3%... Through the data just listed, it shows that Toyota cars are being traded more than other cars. However, when combined with prices and the top 10 car manufacturers, Toyota has an average price of about 512 million VND. In our opinion, Toyota is a popular and famous car manufacturer, so it is used by consumers. Therefore, there is a price segment from popular to high-end. Next, the top 10 cars with a high average price are Mercedes-Benz at approximately 1 billion 650 million VND. In particular, the collected dataset shows that Ferrari has the lowest sales rate but the highest price, approximately 12 billion VND. This can be considered an exceptional data sample. We will consider removing or retaining it in the following preprocessing process.
- The “Condition” attribute: The majority of cars sold on ChoTot Car Market are pre-owned cars. Pre-owned cars account for 90.3%, which is much higher than new cars.
- The “Year” attribute: In the collected dataset, the year of manufacture of cars ranges from 1990 to 2023. However, there are two special years, 1979 and 1983. This is considered an exception because the cars were manufactured for a long time. The year of manufacture of cars in the data set is mainly concentrated after 2014. In addition, it was also discovered that the year 2022 has the highest number of cars and prices; the number of cars accounts for 11.1%, and the price is about 850 million VND.
- The “Transmission” attribute: In the dataset, information about the car's gearbox is also of interest to sellers, divided into three main types of gearboxes: automatic, semi-automatic, and manual. Currently, most of the cars in ChoTot Car Market are cars with automatic gearboxes because automatic gearboxes are quite convenient when moving in the city, so many people choose to use them. The highest average price is a car with a semi-automatic gearbox, approximately 1 billion 712 million VND, because this is an improved gearbox that combines an automatic and manual gearbox. In more detail, the price of a car with a semi-automatic gearbox is about 710 million VND higher than an automatic one and about 300 million VND higher than a manual one.
- The “Fuel type” attribute: Car fuels focus on four types of fuel, such as gasoline, oil, electricity, and hybrid (a combination of gasoline and electricity). In Vietnam, gasoline cars are the most popular. Therefore, at ChoTot Car Market, gasoline cars also account for the majority, about 82.6%, and the lowest is electric cars,

about 0.7%. Because electric cars have only recently become widely popular. In addition, the price of hybrid cars is the highest, at about 773 million VND. Because this is an advanced engine that combines two fuels together. The lowest price is for electric cars, about 591 million VND.

- The "Body style" attribute: Cars at ChoTot Car Market have many types of models, including Sedan, SUV, Minivan, Hatchback, Pickup... The most popular models are the Sedan, at approximately 30.2%, and the SUV, at approximately 26.7%. The average price of Sedan and SUV cars is about 574 million VND and about 787 million VND respectively. For Couple cars (2-door cars), about 1 billion 110 million VND is much higher than other car models.
- The "Seat capacity" attribute: Based on the analysis results, the number of seats in cars is mainly 4-seat and 7-seat. In addition, the data set also contains 2-seat and 9-seat cars, with a negligible number.
- The "Country" attribute: The origin of cars at ChoTot Car Market is commonly cars assembled in Vietnam, about 37.7%. The remaining cars are from other countries such as China, Thailand, Japan, Korea, Germany, USA, India... The average price is low, focusing on cars from Taiwan at about 248 million VND, India at about 277 million VND, and Korea at about 295 million VND. In particular, Germany is one of the countries that manufactures and assembles luxury cars such as Audi, BMW, Maybach, Mercedes Benz, Opel, Porsche,... so German cars have the highest price of about 1 billion 233 million VND.

Next, we conduct a detailed analysis of the values of the "Sales user" attribute to find more meaningful insights. Specifically, we visualized the number of cars associated with the seller as follows:

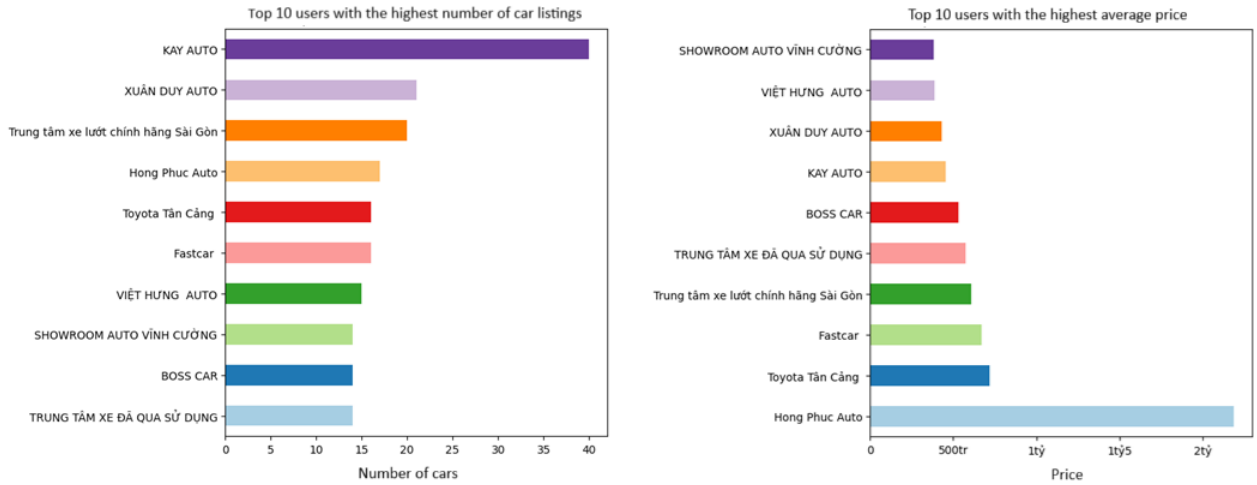

Fig.5. Visualization of the top 10 car sellers

The analysis results show that among the top 10 users posting for sale at ChoTot Car Market include Fig.5: "KAY AUTO", "XUAN DUY AUTO", "Saigon Genuine Used Car Center". The users posting for sale focus on popular and cheap cars. Among them, the user "Hong Phuc Auto" has a higher average price than the other users posting for sale. However, the number of sales is only average. But in more detail, the user "Hong Phuc Auto" focuses on selling cars with an average price of about 500 million VND, and the most expensive car model of "Hong Phuc Auto" is the Bentley Mulsanne Speed Model 2015, about 8 billion 300 million VND. All users posting for sale in the top 10 are partners of ChoTot Car Market. This is reliable information and has good reference value.

The remaining attributes, such as "Title", "Address", and "Odometer", require further processing so that they can be presented after the data preprocessing and cleaning section. In addition, attributes such as "Warranty", "Weight", and "Load capacity" have only one value so that these attributes will be removed in the data preprocessing section.

4.2. Dataset Preprocessing and Cleaning

The original raw dataset has inconsistent attributes in terms of data format and many missing values. Therefore, we need to preprocess the data to make the dataset more consistent and more suitable for inclusion in the prediction model. The data preprocessing process is carried out through a flowchart that details the steps in Fig.6.

The data processing and cleaning procedure was carried out in detail through the following main steps:

- Step 1: Replace the "-" symbol values with NaN (Not a Number) to denote those values as missing values that need to be processed.
- Step 2: Remove and replace unnecessary columns.

  - Delete the columns "Title", "Sales user", "Href" because they do not have predictive meaning, and the columns "Warranty", "Weight", "Load capacity" because they only have one value.

- Replace the column “Address” with the column “Province” because this column only includes values related to the province. This replacement makes data processing and analysis easier.

- Step 3: Handle missing values.
  - Categorical variables: There are 3 columns that need to be replaced: “Seller classification”, “Body style”, and “Make”. Replace the NaN value with the most common value for each group found when grouping. The “Make” column needs to be replaced with a local mode. If the local mode is NaN, replace the NaN value with a global mode.
  - Numerical variables: Proceed similarly to the categorical variable. Here, there are two columns to replace, “Seat capacity” and “Odometer”. For “Odometer”, there is one difference: find the maximum and minimum values for each group. Then, replace the values outside the minimum and maximum similarly as above.
  - Target variable: For “Price”, we proceed to remove three rows with missing values.

- Step 4: Convert the data types of the properties to match each other. Specifically, convert “Year”, “Odometer”, “Seat capacity”, and “Price” to integers.

After performing preprocessing steps on the dataset, we proceed to store the processed dataset. With this version, we can perform post-preprocessing exploratory analysis to find relationships between variables and find valuable insights in the dataset.

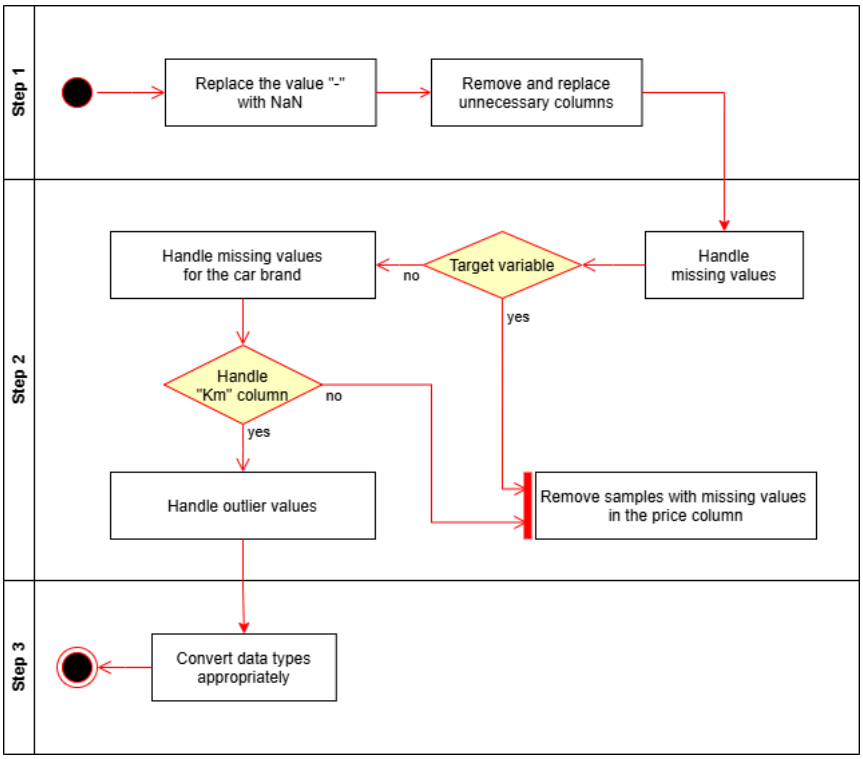

Fig.6. Steps for data processing and cleaning

4.3. Exploratory Analysis of the Preprocessed Dataset

After preprocessing the data, we conduct exploratory data analysis again for the attributes/columns that have full values. In this section, we will call the attribute/column a variable and divide the attributes/columns into the following variables: numeric value distortion and categorical value distortion. And, we choose the variable “Price” as the target variable. The remaining variables are independent variables.

In this section, we used visualization techniques. The result is a dashboard consisting of correlation charts and frequency distributions of categorical variables shown as follows:

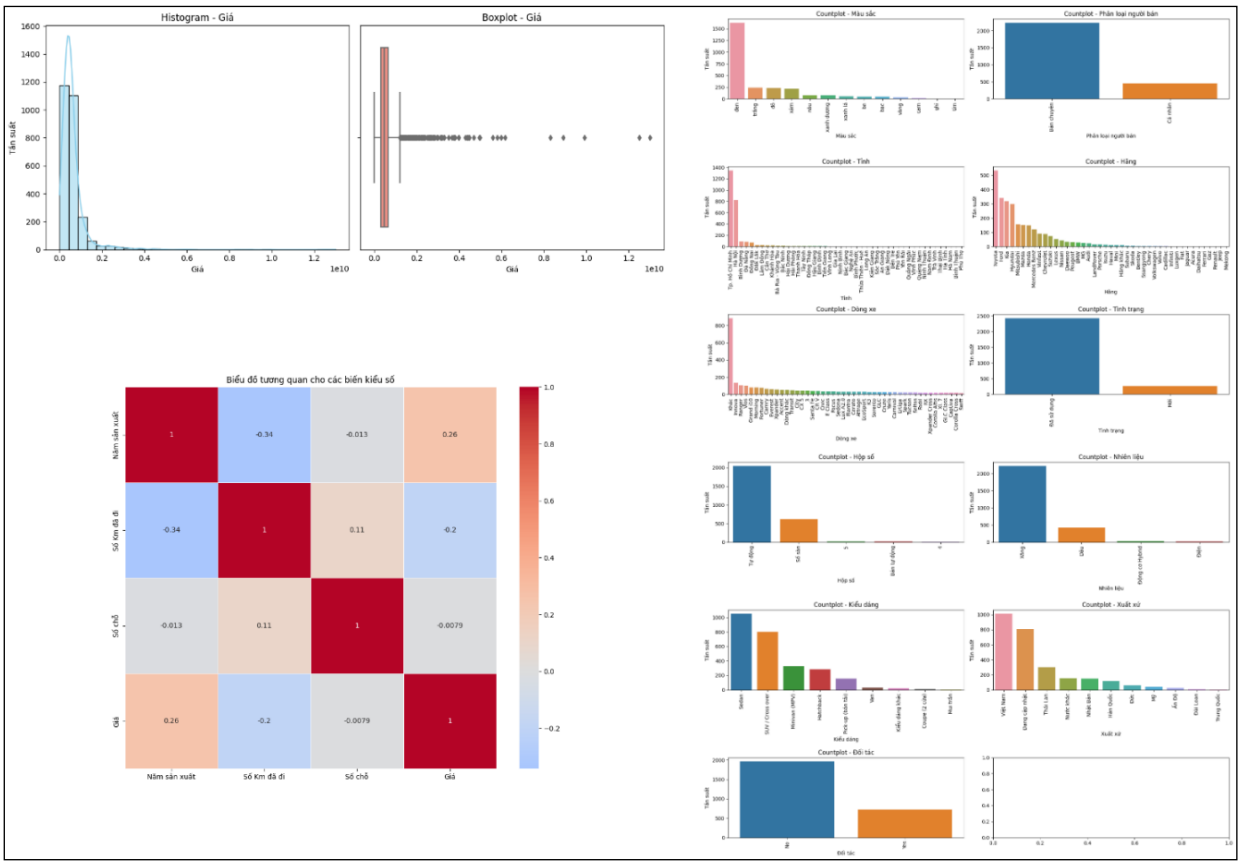

Fig.7. Exploratory results of the processed dataset

Based on the exploratory results of the processed dataset Fig.7, we draw the following conclusions:

- Target Variable ("Price" attribute): The "Price" variable has a right-skewed distribution with skewness = 7.39, indicating that most of the cars are concentrated in the low average price range. At the same time, there are a few cars with high values. The difference between the mean (VND 619 million) and the median (VND 476 million) reflects the significant impact of outliers. In addition, kurtosis = 89.78 shows that the distribution has a very sharp peak, with many outlier data points located far from the center.
- Numerical Variables: We analyzed each variable and examined the correlation between the variables with each other for car prices Fig.7. From the analysis results, we have the following conclusions:

  - The variable "Year" (Car production year) is likely to directly impact car prices as it reflects the car's newness. The cars in the dataset were highly concentrated in the period 2014-2021. In addition, the distribution is skewed to the left with a skewness of -1.39. The distribution is slightly kurtotic with a kurtosis of 2.59, indicating a high concentration of cars in recent years. This suggests that the majority of the cars are new, while some older cars appear as 1979 outliers.
  - The variables "Year" and "Seat capacity" are symmetrical variables with an error of 0.1
  - The variable "Odometer" (Number of kilometers traveled based on the odometer) has a mean of about 53,393 km and a median of 50,000 km. The distribution is slightly less symmetrical with a skewness of 0.12 and a kurtosis of -1.01. Therefore, the data on vehicles with kilometers traveled is relatively uniform, with few obvious outliers.
  - Next, analyzing the correlation between the variables, only "Year" and "Odometer" are correlated with each other, and the correlation is weak.
  - Finally, analyzing the correlation between the variables and the target variable "Price", no variable shows a significant correlation with the car price.

- Categorical Variables: For categorical variables, analyze their distribution based on frequency as well as the distribution among labels in the categorical variable with respect to price. The analysis helps to find variables that influence price.

  - Using visual techniques to sketch a boxplot frequency chart for the classification variables for the price. At the same time, analyze to find the most common classification values that affect the price Fig.7.

- Next, apply the ANOVA analysis technique to find the variables that affect the car price. The result is that the variable "Seller classification" has the most influence on the car price (with F-Test = 65.92 and p-value close to 0). And, the variable "Province" has the least influence on the price (with F-Test = 1.82 and p-value = 0.00215).
- ANOVA analysis of the variable "Make" shows an evident influence on car price (with F-Test = 55.37 and p-value close to 0). This result confirms that the average prices between different brands differ.
- Variables such as "Condition" and "Fuel type" have P-value and F-Test significance coefficients exceeding the scale threshold, so these variables are not a sufficient basis to evaluate the influence on the price.

Analyzing the data further, we set out to find out which "Make" car models are being sold in which "Province" with the highest average "Price". Specifically, the steps to solve the problem are as follows:

- Step 1: Group the data by the variable "Make" and the variable "Province", then calculate the average value by "Price" for each group found.
- Step 2: Find the "Province" with the highest average "Price" from the grouped data.
- Step 3: Select the value of the variable "Make" and the variable "Province" with the highest average price found. Then, sort in descending order by the variable "Price".
- Step 4: Finally, select the value of the variable "Province" and the variable "Make" with the highest average price.

After following the above steps, the "Province" with the highest average price is "Hanoi". And the "Hanoi" combined with the variable "Make" has the highest average price, which is "Panamera", and has the lowest average price, which is "CD5".

Next, we build a visualization to show the ratio of new and used car condition (the variable "Condition") for each car brand (the variable "Make"). The resulting visualization is as follows:

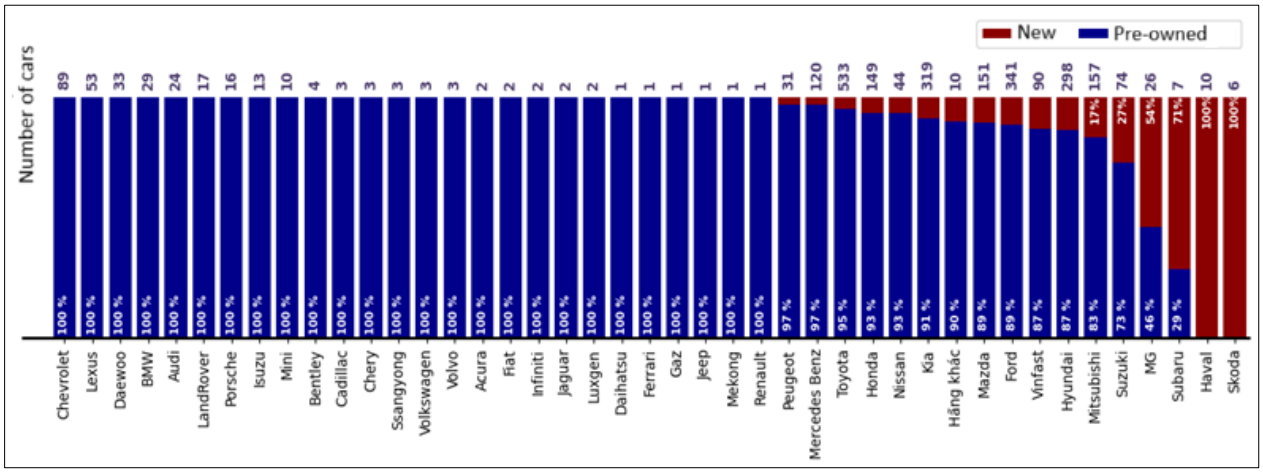

Fig.8. Chart illustrating the status of cars by manufacturer (The percentage of new and used cars by brand)

Based on the visualization results Fig.8, we conclude that the data of the cars collected at ChoTot Car Market are mostly used cars, with only a few new cars from manufacturers such as Mercedes-Benz, Toyota, Ford, etc.

In addition, we observed the values of the variables "Year", "Odometer", and "Seat capacity", which can be considered as classification distortions. Next, we conducted an ANOVA analysis to examine the possibility of affecting the car price variable. The results showed that "Year" and "Odometer" have a fairly high impact on the price. On the contrary, the variable "Seat capacity" has no impact on the price.

After completing the exploratory analysis, we removed the variables that do not affect the car price, which are "Condition", "Fuel type", and "Seat capacity". The reasons for the removal have been presented above.

Next, we apply two more data encoding techniques to convert categorical values into numeric values. Specifically, we use the Label Encoder and One Hot Encoder techniques. The end of the data preprocessing and exploration of the post-analysis dataset are two versions of the dataset "Label" and "OneHot", which are called label-encoded dataset and one-hot encoded dataset respectively. With the two versions, "Label" and "OneHot", we use them to train the model and predict car prices for new data. Then, analyze and evaluate which version will give better prediction results.

5. Model Building and Refinement

5.1. Experimental Study

In this section, we conduct two experiments: the first experiment uses a label-encoded dataset; the second experiment uses a one-hot encoded dataset to find the most suitable model.

To evaluate the performance of the model, we used three scales: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-Squared (R^2). The meaning of the measures is explained as follows:

- MSE: measures the difference between the predicted and actual values, calculated as the average of the squared error between the expected and actual values; the smaller the better.
- MAE: measures the difference between the predicted and actual values, calculated as the average of the absolute value of the error between the predicted and actual values; the smaller the better.
- R^2: evaluates the explanatory power of a regression model, measuring the percentage of variation in the dependent variable explained by the model; the closer to 1, the better.

The experimental process to find the suitable model is as follows: First, use XGBoost, Random Forest, Decision Tree, Histogram Gradient Boosting, K-Neighbors, SVR, and Linear Regression to train the model using both label-encoded datasets and one-hot encoded datasets, respectively. Second, apply MSE, MAE, and R^2 to select the most suitable model and dataset (label-encoded or one-hot encoded).

In the experiment, we omitted the data normalization process to reduce preprocessing costs; however, this did not affect the model's accuracy. Based on the overall research results [15]. There was no performance difference or statistical difference between normalized data and non-normalized data when applying models such as XGBoost, LightGBM, Random Forest, and CatBoost. At the same time, it demonstrated high stability, as confirmed by the Wilcoxon and Friedman tests. Therefore, the decision tree-based algorithms are not affected by the feature ratio because the node splitting process depends on the relative order and not on the absolute value of the data.

Similarly, in the study of artificial intelligence in geosciences [16] when using XGBoost, LightGBM, and NGBoost to predict geophysical logs, it was also found that the decision tree is not sensitive to the expansion of feature vectors; there is no need to normalize or expand the well logs. And, in the study [17] XGBoost also proves that this model works stably without normalizing the input data.

Therefore, normalization is not required when applying XGBoost, because the nature of the tree-based algorithm depends on the branching rule based on relative values (threshold-based splits) rather than on absolute magnitude. Therefore, when evaluating the performance of XGBoost, normalizing or not normalizing the data has almost no effect on the results. Normalization can be omitted to save preprocessing costs without reducing the model accuracy. However, denormalization of data may result in high MSE and MAE values.

Firstly, to be able to choose a model that fits the dataset, we have built pipelines[18] for seven models, with each pipeline consisting only of the default pre-installed model. The reason we did not standardize this dataset is that the dataset only has two variables after removing the variables with no values, and the remaining variables have been encoded as one-hot.

After running the built pipelines, we have created a ranking table of the model results based on the metrics MSE, MAE, and R^2 Table 2. This table is arranged in descending order of R^2.

Table 2. Experimental results on multiple models

No.	Model	MSE	MAE	R^2
1	XGBoost	6.943×10^{16}	1.057×10^{8}	0.776
2	Random Forest	9.384×10^{16}	1.061×10^{8}	0.698
3	Decision Tree	1.26×10^{17}	1.241×10^{8}	0.594
4	Histogram Gradient Boosting	2.052×10^{17}	2.092×10^{8}	0.339
5	K Neighbors	3.072×10^{17}	2.971×10^{8}	0.011
6	SVR	3.233×10^{17}	2.909×10^{8}	-0.041
7	Linear Regression	5.337×10^{25}	7.257×10^{11}	-1.718×10^{8}

Secondly, based on the results presented in Table 2, the models were tested on the same dataset, and the results clearly indicate that the XGBoost model outperforms other models. More specifically, the Random Forest model (R^2 = 0.698), Decision Tree (R^2 = 0.594), and XGBoost showed the most stable and accurate prediction ability, with the XGBoost results reaching MSE = 6.943×10^{16}, MAE = 1.057×10^{8}, and R^2 = 0.776. The XGBoost model showed better generalization ability thanks to the gradient boosting mechanism combined with regularization (L1, L2) to help reduce overfitting. In addition, the remaining models are not optimal enough to handle data with complex distributions and large variable values, such as Histogram Gradient Boosting, which achieved an R^2 of only 0.339. Mainly, when the data exhibits a significant difference between numerical attributes, K-Neighbors (R^2 = 0.011) and SVR (R^2 = -0.041) are not suitable,

as they are sensitive to data ratios and feature distances. Next, the Linear Regression model, with an extremely low R^2 value of -1.718×10^8, suggests that the relationship between the independent variables and the dependent variable of car price is nonlinear and is not well-suited for linear models. It can be concluded that thanks to the sequential boosting mechanism, XGBoost can learn from the errors of previous trees and optimize globally through a custom loss function. To ensure the model operates stably even when the data has a skewed distribution, integrating tree pruning and shrinkage helps reduce noise and increase generalization ability. From the analysis just presented, it can be concluded that XGBoost is the most suitable model for the car price prediction problem because it is stable, and the ability to model nonlinearly balances well between accuracy. The remaining models exhibit poor performance due to the complex structure of the data or overfitting.

One more reason we selected the model with the first rank as the model that best fits the dataset is Extreme Gradient Boosting (XGBoost). This is an ensemble-type machine learning algorithm for solving prediction and classification problems[17]. It has some characteristics as follows:

- Ensemble: Combine multiple weak models into a stronger model, improving the accuracy and generalizability of the model.
- Boosted decision tree: The algorithm builds multiple decision trees, each of which is built sequentially and learns from the errors of the previous tree.
- Optimizing the loss function: the loss function specifies the error between the predicted value and the actual value, helping the model adjust the prediction to minimize the error.
- Regularization: Use algorithms such as L1 (lasso) and L2 (ridge) to control the complexity of the model and avoid over-fitting.

Based on the above characteristics, we selected the XGBoost algorithm to build a model for the preprocessed dataset. Here, we will use the default parameters of the model to predict car prices. Then, we conduct model tuning, which means searching for parameter values to produce the best prediction results. The statistical data evaluating the model will be presented in the following section.

The performance of the XGBoost model with an R^2 of 0.776 is the highest among the tested models, but it is still only average compared to many similar studies. The obtained results may come from the following reasons: (1) the limited size of the dataset may lead to the model not having enough information to learn complex rules in the relationship between variables; (2) the actual data collected from the car trading floor will contain many noisy data samples and outliers, which may reduce the accuracy of the model; (3) the data distribution of the target variable (car price) is skewed, and the model does not learn the extreme values well; (4) the complex nonlinear relationship between the features with each other, beyond the model's representation capacity, without creating new features (feature engineering).

After selecting a valid XGBoost model, the next stage is hyperparameter tuning to find the optimal parameters that will optimize the model. In the process of fine-tuning hyperparameters, apply the Randomized Search Cross-Validation (RandomizedSearchCV) method to optimize the XGBoost model by finding the best set of parameters within a sufficiently large space, while ensuring computational efficiency. The search range includes hyperparameters that have a significant impact on the learning ability and generalization of the model: n_estimators $\in$ [100, 1000], max_depth $\in$ [3, 10], learning_rate $\in$ [0.01, 0.3], min_child_weight $\in$ [1, 10], gamma $\in$ [0, 0.5], subsample and colsample_bytree $\in$ [0.5, 1.0], reg_alpha (L1) $\in$ [0, 1] and reg_lambda (L2) $\in$ [0, 10]. The search process is limited to 50 iterations (n_iter = 50) to balance the coverage of the parameter space and the training cost. Each implementation is evaluated using cross-validation, and random_state = 42 to ensure reproducibility. Once completed, the model with the best parameter set is fully retrained on the training set and used for formal evaluation on the testing set.

During the prediction process, we found that the “OneHot” dataset gave better results than “Label”, so we will present the results of “OneHot”. The parameters chosen during the model tuning process include:

Table 3. The best parameter values for XGBoost

No.	Parameter	Meaning	Default Value	Best Value
1	reg_lambda	Ridge Regularization (L2)	None	1
2	reg_alpha	Lasso Regularization (L1)	None	1
3	n_estimators	The number of boosting iterations, equivalent to the number of decision trees trained	100	450
4	max_depth	The maximum depth of the tree	None	12
5	learning_rate	Determines the extent to which each new tree contributes to the overall model.	None	0.1
6	gamma	Minimum loss function reduction to make one split in the tree.	None	0.3
7	colsample_bytree	Percentage of columns selected to build the tree.	None	0.7

The best values of the above parameters have been found using Random Search. We will proceed to use the XGBoost model with the best set of parameters found to evaluate the model's performance on the preprocessed dataset.

5.2. Model Evaluation

Along with the above three metrics, we have combined two more methods: Train-Test Split and Cross-Validation (K-fold) to evaluate the model more effectively.

For the Train-Test Split, we have divided the dataset into a training set and a testing set with a ratio of 8:2. Then, we trained the XGBoost model with the training set and predicted the car price on the testing set. We have obtained the results of the three evaluation metrics based on the default parameter set and the best parameter set as follows:

Table 4. The evaluation metrics results based on the train-test split

Evaluation Metric	Results	
	Default Value	Best Value
MSE	6.943×10^{16}	6.682×10^{16}
MAE	1.057×10^{8}	0.957×10^{8}
R^2	0.776	0.779

From the results table, it can be seen that in the best parameter set, the values of MSE, MAE, and R2 are all improved compared to the default parameter set. This demonstrates that the optimal parameter set will yield better performance than the default parameter set. We continue to evaluate this best parameter set using the K-Fold method.

For K-Fold, we divided it into 10 folds to build the model and predict car prices 10 times based on those folds. For each experiment, nine folds will be selected as the training set, and the remaining fold will be the testing set. In each evaluation, we will save the results of the three metrics mentioned above in Table 5.

Table 5. The evaluation metrics results are based on the cross-validation (K-Fold)

Metric	Results with the best value parameter									
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
MSE	1.036×10^{17}	5.115×10^{16}	5.937×10^{16}	2.402×10^{17}	6.391×10^{17}	4.332×10^{16}	9.639×10^{16}	1.440×10^{17}	1.374×10^{17}	1.181×10^{17}
MAE	1.066×10^{8}	0.972×10^{8}	1.058×10^{8}	1.054×10^{8}	1.276×10^{8}	0.894×10^{8}	0.973×10^{8}	1.310×10^{8}	1.133×10^{8}	1.041×10^{8}
R^2	0.727	0.788	0.847	0.690	0.420	0.842	0.809	0.768	0.734	0.673

Based on Table 5, the results clearly show that the model's oscillation level is relatively stable. Specifically, the mean and standard deviation of the scales are presented in Table 6.

Table 6. Statistical summary of cross-validation results with 10 folds

Metric	Mean	Variance	Standard Deviation
MSE	1.632×10^{17}	4.03×10^{33}	6.35×10^{16}
MAE	1.078×10^{8}	1.56×10^{14}	1.25×10^{7}
R^2	0.73	0.016	0.127

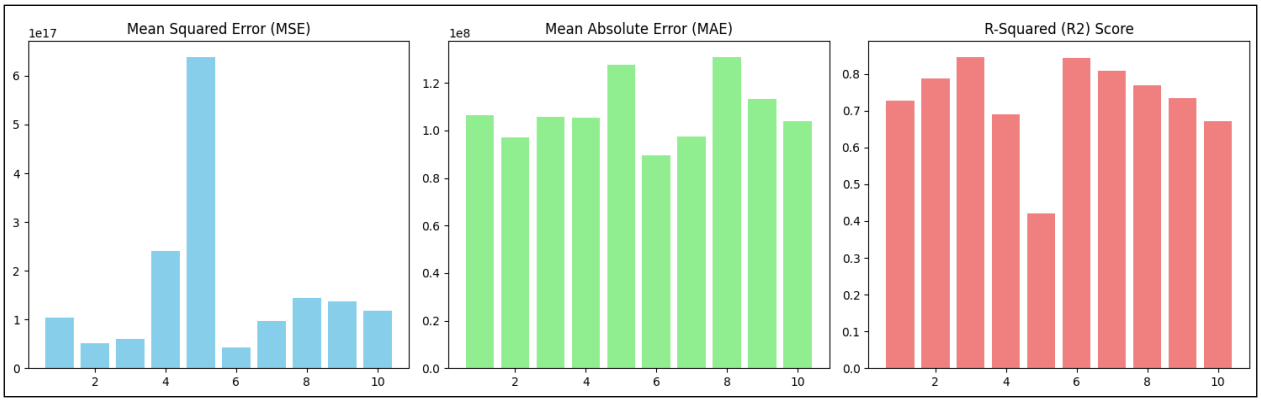

Fig.9. The evaluation metrics results for each fold

Comment on the results in Table 6:

- The MSE has an average of 1.06×1017, standard deviation of 6.35×1016. The results show that the average square error between the folds varies, but it remains within the allowable range.
- The MAE has an average of 1.07×108, standard deviation of 1.25×107. This value in VND currency is entirely acceptable, indicating that the average absolute error remains quite stable across the data splits.
- The R^2 has an average of 0.730, variance of 0.016, and standard deviation of 0.127, clearly showing the relatively good consistency of the model between folds.

In general, the variance and standard deviation values between the 10 folds are all at low levels on both scales, R^2 and MAE, indicating that the XGBoost model maintains a good generalization ability and stability, and does not depend strongly on the way the dataset is divided. This confirms that the XGBoost model in the price prediction problem exhibits good reliability and robustness.

From the results, we can observe that the model's performance does not differ significantly at each evaluation on the folds. This demonstrates that the model is neither overfitting nor underfitting when given any fold as a test set.

We visualize the results of the above three scales across each fold through the bar charts above Fig.9.

6. Conclusions

In this research, we developed a model for predicting the prices of pre-owned cars using data collected from an online marketplace. The dataset was preprocessed and analyzed in detail to identify essential insights, uncover meaningful relationships between variables, and develop a price prediction model. The experimental results demonstrated that the XGBoost model achieved the best performance, with $R^2 = 0.776$ for default parameters and $R^2 = 0.779$ after optimization, confirming the model's robustness to this data type.

In addition to model construction, this study presents a practical data workflow for car price analysis, encompassing the data collection process, preprocessing, and exploratory analysis, which can be applied in similar online marketplace scenarios. These findings also provide valuable insights into key factors that influence car prices, such as brand, year, mileage, and car condition, which can help sellers and buyers make more informed car pricing decisions.

However, this study is limited by the scope and representativeness of the self-collected dataset, which may not fully represent market diversity, regional differences, or temporal fluctuations such as holidays, seasons, or promotions. Furthermore, the study only analyzed numeric and structured categorical data; it did not integrate text processing of car descriptions.

In the future, we plan to implement automatic data collection from other online platforms to ensure data timeliness and integrate multimodal features such as text and image data. Additionally, deploying the predictive model as a real-time web service will enable users to estimate car prices based on the latest market trends dynamically.

Acknowledgment

This research is funded by University of Information Technology - Vietnam National University HoChiMinh City under grant number D1-2025-58.

References

[1] Bokonda PL, Ouazzani-Touhami K, Souissi N. Predictive analysis using machine learning: Review of trends and methods. 2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2020, p. 1–6. https://doi.org/10.1109/ISAECT50560.2020.9523703.

[2] Punia S, Shankar S. Predictive analytics for demand forecasting: A deep learning-based decision support system. Know-Based Syst 2022;258. https://doi.org/10.1016/j.knosys.2022.109956.

[3] Jamarani A, Haddadi S, Sarvizadeh R, Haghi Kashani M, Akbari M, Moradi S. Big data and predictive analytics: A systematic review of applications. Artif Intell Rev 2024; 57:176. https://doi.org/10.1007/s10462-024-10811-5.

[4] Ranjeeth S, Latchoumi TP, Paul PV. A Survey on Predictive Models of Learning Analytics. Procedia Comput Sci 2020; 167:37–46. https://doi.org/https://doi.org/10.1016/j.procs.2020.03.180.

[5] Habel J, Alavi S, Heinitz N. A theory of predictive sales analytics adoption. AMS Review 2023; 13:34–54. https://doi.org/10.1007/s13162-022-00252-0.

[6] Ahmad M, Farooq MA, Hussain MZ, Hasan MZ, Mustafa M, Khalid A, et al. Car Price Prediction using Machine Learning. 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), IEEE; 2024, p. 1–5.

[7] Gupta S, Vijarania M, Udbhav M. A machine learning approach for predicting price of used cars and power demand forecasting to conserve non-renewable energy sources. Renewable Energy Optimization, Planning and Control: Proceedings of ICRTE 2022, Springer; 2023, p. 301–10.

[8] Budiono DA, Utomo KS, Wibowo KJ, Wiradinata MJ. Used car price prediction model: a machine learning approach. Int J Comput Inf Syst(IJCIS) 2024; 5:59–66.

[9] Jin C. Price prediction of used cars using machine learning. 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), IEEE; 2021, p. 223–30.

[10] Das Adhikary DR, Sahu R, Pragyna Panda S. Prediction of used car prices using machine learning. Biologically Inspired Techniques in Many Criteria Decision Making: Proceedings of BITMDM 2021, Springer; 2022, p. 131–40.

[11] Nandan M, Ghosh D. Pre-owned car price prediction by employing machine learning techniques. Journal of Decision Analytics and Intelligent Computing 2023; 3:167–84.
[12] Longani C, Prasad Potharaju S, Deore S. Price prediction for pre-owned cars using ensemble machine learning techniques. Recent Trends in Intensive Computing, IOS Press; 2021, p. 178–87.
[13] Benabbou F, Sael N, Herchy I. Machine Learning for Used Cars Price Prediction: Moroccan Use Case. In: Lazaar M, Duvallet C, Touhafi A, Al Achhab M, editors. Proceedings of the 5th International Conference on Big Data and Internet of Things, Cham: Springer International Publishing; 2022, p. 332–46.
[14] Wu J Da, Hsu CC, Chen HC. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Syst Appl 2009; 36:7809–17. https://doi.org/10.1016/J.ESWA.2008.11.019.
[15] Pinheiro JMH, de Oliveira SVB, Silva THS, Saraiva PAR, de Souza EF, Godoy R V, et al. The impact of feature scaling in machine learning: Effects on regression and classification tasks. ArXiv Preprint ArXiv:250608274 2025.
[16] Wang H, Wu Y, Zhang Y, Lai F, Feng Z, Xie B, et al. Uncertainty and explainable analysis of machine learning model for reconstruction of sonic slowness logs. Artificial Intelligence in Geosciences 2023; 4:182–98. https://doi.org/https://doi.org/10.1016/j.aiig.2023.11.002.
[17] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: Association for Computing Machinery; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.
[18] Ru-tao Z, Jing W, Gao-jian C, Qian-wen L, Yun-jing Y. A Machine Learning Pipeline Generation Approach for Data Analysis. 2020 IEEE 6th International Conference on Computer and Communications (ICCC), 2020, p. 1488–93. https://doi.org/10.1109/ICCC51575.2020.9345123.

Authors' Profiles

Bui Quang Phu holds a Bachelor's degree in Computer Science from the University of Information Technology, Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam. Passionate about Machine Learning and Data Science.

Pham Hoang Phuc received a Bachelor's degree in Computer Science from the University of Information Technology, Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam. His research interests include Machine Learning and Software.

Pham The Son is a lecturer at the Faculty of Information Science and Engineering, University of Information Technology, Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam. He holds Bachelor's and Master's degrees in Computer Science from the University of Information Technology, VNU-HCM. His research interests include Explainable Artificial Intelligence, Machine Learning, Data Analytics, and Intelligent Systems Applications.

How to cite this paper: Bui Quang Phu, Pham Hoang Phuc, Pham The Son, "Car Price Analysis Using Data Collected from an Online Sales Platform", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.1, pp.124-139, 2026. DOI:10.5815/ijitcs.2026.01.07