

# Customer Purchase Decision Prediction for Pre- Owned Cars Using Logistic Regression

Jayasree S, P VedhaShree, Sivani D, Swetha M, S Uma Devi

Department of Computer Science and Engineering

Avinashilingam Institute for Home Science and Higher Education for Women

## ABSTRACT

The rapid growth of the pre-owned car market has created a need for accurate prediction of customer purchase decisions to support effective sales and marketing strategies. This paper presents a machine learning-based approach for predicting whether a customer is likely to purchase a pre-owned car using the Logistic Regression algorithm. A Kaggle dataset containing customer attributes such as age, gender, and estimated salary is utilized for model development. The data undergoes preprocessing techniques including cleaning, feature selection, and normalization to improve model performance. Logistic Regression is trained on historical customer data to identify patterns associated with purchasing behaviour and classify customers as potential buyers or non-buyers. The system is implemented using Python and data analysis libraries including Pandas, NumPy, Matplotlib, and Scikit-learn. Experimental results demonstrate that the proposed model achieves an accuracy of approximately 88.8%, indicating its effectiveness in predicting customer purchase decisions. The proposed system provides a reliable and efficient solution for customer behaviour analysis, enabling dealers to identify potential customers, optimize marketing efforts, and support data-driven decision-making in the pre-owned car market.

**Key words:** Machine Learning, Logistic Regression, Customer Purchase Prediction, Pre-Owned Cars, Predictive Analytics, Data Analysis.

## 1. INTRODUCTION

The pre-owned car market has experienced significant growth in recent years due to

increasing vehicle costs and the demand for affordable transportation. As the market expands, understanding customer purchasing behavior has become essential for automobile dealers and businesses. Predicting whether a customer is likely to purchase a pre-owned car can help improve sales performance, customer targeting, and marketing strategies. Traditional methods of identifying potential buyers rely on manual analysis and personal experience, which may not always provide accurate results. Machine learning techniques offer a more efficient and data-driven approach by analyzing historical customer data and identifying patterns that influence purchasing decisions. This study proposes a Logistic Regression-based prediction system that analyzes customer attributes and predicts whether a customer is likely to purchase a pre-owned car. The system aims to support decision-making, reduce manual effort, and improve overall business performance.

## 2. LITERATURE REVIEW

Several researchers have applied machine learning techniques to predict customer behaviour and analyse factors influencing purchase decisions in the automotive industry. Logistic Regression has been widely used for classification problems due to its simplicity, interpretability, and effectiveness in predicting binary outcomes. Studies have shown that machine learning models can successfully identify potential buyers by analysing customer demographics, preferences, and vehicle-related attributes.

Research on pre-owned car analysis has also focused on predicting vehicle prices using

algorithms such as Linear Regression, Random Forest Regression, Support Vector Regression, and Multiple Linear Regression. These models utilize vehicle characteristics such as age, mileage, fuel type, engine capacity, and vehicle condition to improve prediction accuracy. Comparative studies have demonstrated that machine learning approaches provide more reliable results than traditional manual estimation methods.

Recent advancements in machine learning and data analytics have enabled businesses to process large volumes of customer and vehicle data efficiently. Various studies have highlighted the importance of data preprocessing, feature selection, and exploratory data analysis in improving model performance. These

research contributions provide a strong foundation for developing an intelligent system that predicts customer purchase decisions in the pre-owned car market. The findings from existing literature support the use of Logistic Regression as an effective and efficient technique for customer purchase prediction

### 3. PROBLEM STATEMENT

Identifying potential customers in the pre-owned car market is a challenging task for dealers and businesses. Traditional approaches depend largely on assumptions, past experience, and manual evaluation of customer information, which can lead to inaccurate predictions and missed sales opportunities. Moreover, the increasing volume of customer data makes manual analysis time-consuming and inefficient. Therefore, there is a need for an intelligent system that can accurately analyze customer characteristics and predict purchasing behaviour. The proposed solution utilizes Logistic Regression to classify customers as potential buyers or non-buyers based on demographic and behavioral factors, thereby improving decision-making and marketing effectiveness.

### 4. OBJECTIVES

The primary objective of this study is to develop a machine learning-based system for predicting customer purchase decisions in the pre-owned car

market. The system aims to collect and preprocess customer-related data, analyze factors influencing purchasing behavior, and apply Logistic Regression for classification. It also seeks to evaluate model performance using appropriate metrics and provide accurate predictions that can assist dealers in identifying potential customers and improving business strategies.

## 5. PROPOSED METHODOLOGY

The proposed system utilizes the Logistic Regression algorithm to predict whether a customer is likely to purchase a pre-owned car. The methodology consists of data collection, preprocessing, exploratory data analysis, model training, testing, and performance evaluation. The overall workflow is designed to transform raw customer data into meaningful insights that support accurate purchase decision prediction.

### 5.1 Dataset Description

The dataset used in this study was obtained from Kaggle and contains information related to Toyota vehicles in the pre-owned car market. The dataset includes attributes such as price, age, kilometers driven (KM), fuel type, horsepower, transmission type, engine capacity, number of doors, and vehicle weight. These features are used to analyze vehicle characteristics and predict customer purchase decisions. The dataset serves as the foundation for training and testing the machine learning model.

| Sl.no | Price | Age | KM    | Fuel type |
|-------|-------|-----|-------|-----------|
| 1.    | 13500 | 23  | 46986 | Diesel    |
| 2.    | 13750 | 23  | 72937 | Diesel    |
| 3.    | 13950 | 24  | 41711 | Diesel    |
| 4.    | 20950 | 25  | 31461 | Petrol    |
| 5.    | 19950 | 22  | 43610 | Petrol    |

**Table-1:** Sample Dataset (Toyota vehicle dataset)

### 5.2 Data Preprocessing

Data preprocessing is performed to improve data quality and ensure accurate analysis. Initially, the

dataset is loaded using the Pandas library and converted into a structured format. Missing and invalid values are identified and handled appropriately. Numerical attributes are filled using statistical measures such as mean values, while categorical attributes are handled using suitable techniques. Duplicate records and unnecessary columns are removed to eliminate redundancy and reduce noise. Data type conversions are also performed to ensure consistency across all attributes.

### 5.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to understand the distribution and characteristics of the dataset. Various visualization techniques such as histograms, scatter plots, box plots, and count plots are used to identify trends, patterns, and relationships among variables. Distribution analysis helps understand the spread of data, while relationship analysis reveals correlations between different features. These visualizations provide valuable insights that support effective model development.

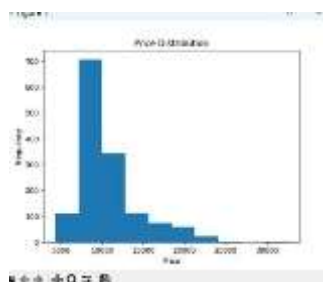


**Fig -1:** Price Distribution Histogram The

histogram shows how car prices are distributed across different ranges Where,  
Most cars are concentrated in the lower price range (around 5000–10000).  
As the price increases, the number of cars decreases.  
Very few cars are in the higher price range (above 20000).

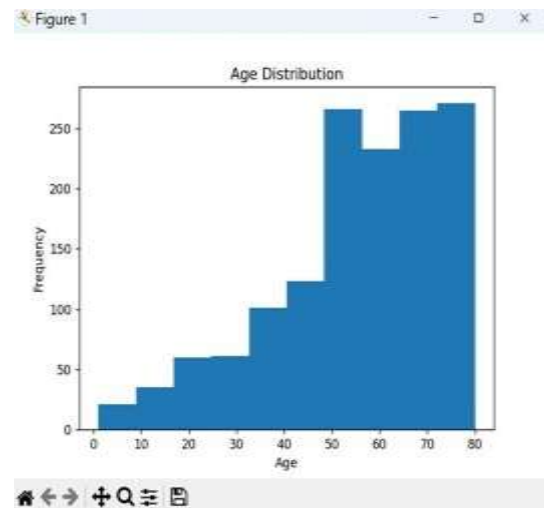


**Fig -2:** Age Distribution Histogram This

histogram represents the age of the cars and the frequency of each cars where,

X-axis = age groups

Y-axis = frequency (number of cars in each age group)

It shows how the data is spread and which age range has more cars.

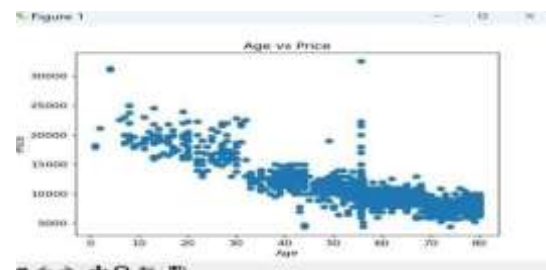


**Fig -3:** Scatter plot - Age Vs Price

The scatter plot represents the age of the cars and price of the cars based on the age where, Each dot represents one car.

It shows a negative trend → as the age of the car increases, the price generally decreases.

Helps identify patterns, correlations, and outliers.

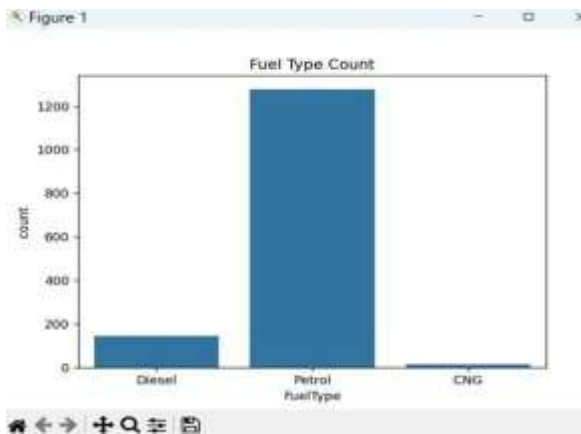


**Fig -4:** Fuel Type Distribution

The bar diagram represents the fuel type and the count of various cars where, The bar chart shows the distribution of cars based on fuel type. Petrol cars have the highest count, indicating they are the most commonly used. Diesel cars are fewer compared to petrol and CNG cars have the lowest count, showing minimal usage.

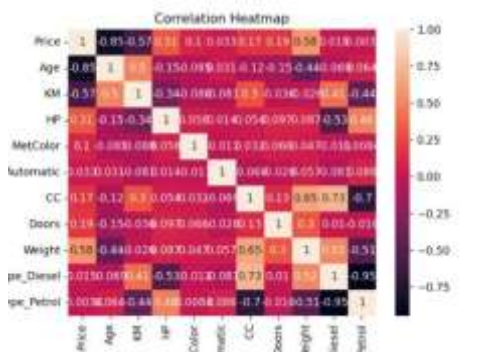


**Fig -5:** Correlation Heatmap

The correlation heatmap illustrates the relationships among the numerical features in the Toyota vehicle dataset. The analysis shows that **Price** has a strong negative correlation with **Age** (-0.85), indicating that older vehicles generally have lower prices. Similarly, **Price** exhibits a moderate negative correlation with **KM** (-0.57), suggesting that vehicles with higher mileage tend to have lower market values. On the other hand, **Weight** (0.58) and **HP** (0.31) show positive correlations with price, indicating that heavier and more powerful vehicles are often priced higher. The heatmap also reveals strong relationships between certain vehicle characteristics, such as **CC** and **Weight** (0.65). These insights help identify influential features and support the development of

an effective Logistic Regression model for customer purchase prediction.

## 5.4 Logistic Regression Model

Logistic Regression is employed as the primary classification algorithm for predicting customer purchase decisions. It is widely used for binary classification problems due to its simplicity and effectiveness. The model analyzes the relationship between customer attributes and the target variable to estimate the probability of purchase. Based on this probability, the model classifies customers as either potential buyers or non-buyers. The algorithm learns from historical customer data and identifies patterns that influence purchasing behavior.

## 5.5 Model Training and Testing

The prepared dataset is divided into training and testing datasets. The training dataset is used to train the Logistic Regression model, allowing it to learn relationships between input features and purchase decisions. The testing dataset is then used to evaluate the model's ability to make predictions on unseen data. This process ensures that the model generalizes well and performs effectively in real-world scenarios.

## 5.6 Performance Evaluation

The performance of the proposed model is evaluated using standard classification metrics such as accuracy, precision, recall, and confusion matrix. Accuracy measures the overall correctness of predictions, while precision and recall provide insights into the model's classification capability. The confusion matrix is used to compare actual and predicted values, helping to assess model effectiveness. These evaluation metrics validate the reliability and predictive performance of the proposed system.

## 5.7 System Workflow

The overall workflow of the proposed system consists of data collection, data preprocessing, exploratory data analysis, model training, testing, prediction, and performance evaluation. The

system processes customer information and generates a binary output indicating whether a customer is likely to purchase a pre-owned car. This prediction assists dealers in identifying potential customers, improving marketing strategies, and supporting data-driven business decisions.

## 6. RESULTS AND DISCUSSION

The developed model successfully predicts whether a customer is likely to purchase a pre-owned car. Various visualizations such as Age distribution, Price distribution, scatter plots are generated to understand customer behavior. The results indicate that customer attributes significantly influence purchasing decisions and can be effectively utilized for prediction.

### 6.1 Model Accuracy

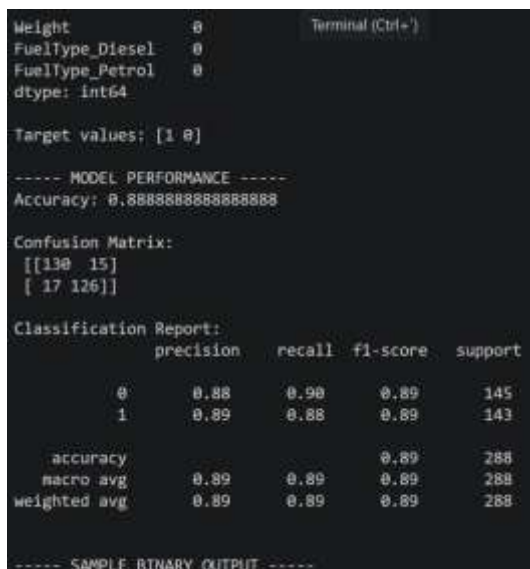


**Fig 6:** Accuracy Result of Logistic Regression Model

The performance of the Logistic Regression model was evaluated using standard classification

metrics. The model achieved an accuracy of **88.89%** on the testing dataset, indicating its effectiveness in predicting customer purchase decisions. The precision, recall, and F1-score values were approximately **0.89**, demonstrating

balanced classification performance across both classes. These results confirm that the proposed model can reliably identify potential customers in the pre-owned car market.

### 6.2 Confusion Matrix Analysis

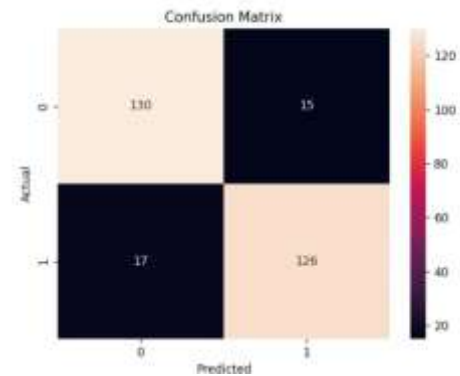


**Fig 7:** Confusion Matrix

The confusion matrix evaluates the performance of the Logistic Regression model by comparing actual and predicted outcomes. The model correctly classified **130 non-purchasing customers** and **126 purchasing customers**, with only a few misclassifications. The results demonstrate the effectiveness of the proposed model in predicting customer purchase decisions and indicate good overall classification performance.

## 7. CONCLUSION

This study presented a machine learning-based approach for predicting customer purchase decisions in the pre-owned car market using Logistic Regression. The proposed model effectively analyzes customer attributes and identifies patterns that influence purchasing behavior. Experimental results demonstrate that the model provides reliable predictions and achieves satisfactory performance in classifying potential buyers and non-buyers. By utilizing data-driven techniques, the system helps dealers improve customer targeting, optimize marketing strategies, and enhance decision-making processes. The study highlights the effectiveness of Logistic Regression as a simple and efficient classification technique for customer purchase prediction.

## 8. ACKNOWLEDGEMENT

The authors express their sincere gratitude to Mrs. S. Uma Devi, M.E., Assistant Professor, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, for her valuable guidance, continuous support, and encouragement throughout the development of this project. The authors also thank the department and institution for providing the necessary facilities to carry out this work.

## 9. FUTURE SCOPE

Future enhancements may include the use of advanced machine learning algorithms, larger real-time datasets, and additional customer and vehicle attributes to improve prediction accuracy. The system can also be deployed as a web or mobile application and integrated with recommendation features to provide a better user experience and support more effective decision-making.

## REFERENCES

- [1] Sukanta Maji and Gautam Bandyopadhyay, "The Effect of Risk Perception on Pre-Owned Car Purchase Decision: A Logistic Regression Approach," *Journal of Management*, 2018.
- [2] G. Kalpana, A. Kanaka Durga, Anoop Reddy, and G. Karuna, "Predicting the Price of Pre-Owned Cars Using Machine Learning and Data Science," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 2022.
- [3] "Customer Purchase Behavior Prediction Using Machine Learning Algorithms: A Comprehensive Review," *International Journal of Computer Trends and Technology (IJCTT)*, 2025.
- [4] Rakesh Sambyal, Rachit Kaushik, Tushar Tyagi, and Rudransh Sharma, "Used Car Price Prediction Using One Hot Encoding," *International Journal For Multidisciplinary Research (IJFMR)*, 2025.
- [5] Andrés Martínez, Claudia Schmuck,

Sergiy Pereverzyev, Clemens Pirker, and Markus Haltmeier, "A Machine Learning Framework for Customer Purchase Prediction in the Non-Contractual Setting," *European Journal of Operational Research*, vol. 281, no. 3, pp. 588–596, 2020.

[6] Radhika Swarnkar, Rhea Sawant, Harikrishnan Ramachandran, and Srideviponmalar P., "Multiple Linear Regression Algorithm-Based Car Price Prediction," *Proceedings of the Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, IEEE, 2023.

[7] "An Empirical Analysis of Machine Learning Algorithms for Used Car Price Prediction System," *IEEE Conference Publication*, 2024.