### **MACHINE INTELLIGENCE**

### **MINI PROJECT**



### **TOPIC : CAR PRICE PREDICTION**

### **PILLAI COLLEGE OF ENGINEERING, NEW PANVEL**

### **Dr. K. M. Vasudevan Pillai Campus, Plot No. 10, Sector 16, New Panvel, Navi Mumbai, Maharashtra 410206**

### **Department of Information Technology**

### **Academic Year 2024– 25**

### **Bachelor of Technology in Information Technology**

**Group Members:**

| S. No | Name | Email | Phone No. | Roll No |
| --- | --- | --- | --- | --- |
| 1 | Tej Ambre | atej22it@student.mes.ac.in | 9404773353 | 503 |
| 2 | Ankit Kumar | ankitk22it@student.mes.ac.in | 8917021433 | 504 |
| 3 | Pronit Ghosh | gpronit23it@student.mes.ac.in | 7249011064 | 517 |
| 4 | Sarthak Gupta | gsarthak22it@student.mes.ac.in | 9713924981 | 518 |

### **1. INTRODUCTION**

#### **1.1 Background**

Car prices are influenced by a multitude of factors that determine their value in the market. As the automobile industry continues to grow, understanding and predicting car prices has become increasingly essential for manufacturers, dealers, and consumers alike. The prices of cars fluctuate due to various reasons such as the year of manufacture, kilometers driven, fuel type, mileage, engine specifications, and the number of seats. Accurate prediction of car prices can assist in decision-making, providing buyers with fair pricing models and helping sellers set competitive prices. Machine learning algorithms, particularly multiple linear regression, offer an efficient way to predict car prices based on historical data and relevant features. By identifying the most influential factors, car price prediction models can provide valuable insights for both businesses and customers, enabling more informed purchases and sales.

#### **1.2 Objectives**

The primary objectives of this study are:

* To develop a robust predictive model that estimates car prices based on significant factors like year of manufacture, kilometers driven, fuel type, mileage, engine capacity (cc), and the number of seats.
* To identify and evaluate the key features affecting car prices, giving insights into how these variables influence the price prediction.
* To assess the performance and reliability of the model using various statistical metrics, ensuring that the predictions are both accurate and practical for real-world applications.
* To present actionable insights for car dealers, buyers, and manufacturers to understand market dynamics and make data-driven decisions regarding car pricing strategies.

#### **1.3 Scope**

This project will focus on cars across various segments and brands to cover a broad range of price factors. The model will take into account data from the past decade, ensuring that the predictions are relevant to current market trends. By considering factors such as mileage, engine size, and fuel type, this study aims to provide a comprehensive understanding of how different aspects influence car prices. The project will also highlight the application of machine learning techniques, specifically multiple linear regression, in real-world scenarios, demonstrating how businesses and individuals can benefit from such predictive analytics.

### **2. PROBLEM DEFINITION**

#### **2.1 Context**

The automobile market is highly competitive, with fluctuating prices influenced by various technical and market-driven factors. Consumers often struggle to find the right price for a used or new car, leading to the need for a predictive model that can offer reliable estimates based on historical data. For car dealers and manufacturers, predicting prices helps optimize inventory, plan production, and manage sales more efficiently. Understanding the key factors behind car pricing can reduce uncertainty in the buying and selling process, and help stakeholders make well-informed decisions. The application of machine learning in this context provides an innovative solution to understanding these patterns and predicting prices with high accuracy.

#### **2.2 Data Requirements**

To build an effective car price prediction model, the following data sets are crucial:

* **Manufacture Year**: A car's age is a significant determinant of its price. Newer cars generally fetch higher prices, but depreciation affects this relationship over time.
* **Kilometers Driven**: The usage of the car, represented by the number of kilometers driven, impacts the price as it indicates wear and tear.
* **Fuel Type**: Different fuel types, such as petrol, diesel, and electric, vary in their running costs and market demand, influencing price.
* **Mileage**: Fuel efficiency is a key factor for many buyers, and higher mileage can make a car more attractive, hence affecting its price.
* **Engine Capacity (cc)**: Cars with larger engines often come at higher prices due to increased power and performance.
* **Number of Seats**: The seating capacity can affect car prices, especially for family-oriented vehicles where larger seating configurations are preferred.

#### **2.3 Challenges**

There are several challenges associated with predicting car prices:

* **Data Availability and Quality**: Inconsistent or missing data can skew results and lead to inaccurate predictions. Ensuring that the data is clean and comprehensive is crucial for building a reliable model.
* **Feature Complexity**: The relationships between different car features and their prices are often complex. For instance, while engine capacity might increase a car's value, higher kilometers driven could lower it. Balancing these factors requires careful feature engineering.
* **Dynamic Market Trends**: The automobile market is influenced by factors such as technology advancements, fuel prices, and consumer preferences. These trends need to be incorporated into the model to ensure it remains relevant over time.

### **3. TECHNIQUES**

#### **3.1 Multiple Linear Regression**

Multiple linear regression (MLR) is a statistical method used to model the relationship between one dependent variable (car price) and multiple independent variables (e.g., manufacture year, kilometers driven, fuel type, mileage). It provides a formula that predicts the dependent variable based on the values of the independent variables. The regression equation can be written as:  
**Y=β0+β1X1+β2X2+...+βnXn+ϵ**  
Where:

* YYY represents the predicted car price.
* β0\beta\_0β0​ is the intercept, the baseline price when all independent variables are zero.
* β1,β2,…,βn\beta\_1, \beta\_2, \dots, \beta\_nβ1​,β2​,…,βn​ are the coefficients representing the effect of each independent variable.
* X1,X2,…,XnX\_1, X\_2, \dots, X\_nX1​,X2​,…,Xn​ are the independent variables (e.g., kilometers driven, fuel type).
* ϵ\epsilonϵ is the error term, accounting for variability not captured by the model.

This method is advantageous because it offers clear interpretability, allowing us to understand how each feature impacts the car price. However, it is essential to validate the assumptions of linearity, independence, and homoscedasticity for MLR to perform effectively.

#### **3.2 Data Preprocessing**

Data preprocessing is critical to ensure the accuracy of the predictive model:

* **Data Cleaning**: This involves handling missing data and removing outliers. For instance, a missing car's mileage may need to be imputed using mean or median values.
* **Normalization**: Car prices, engine capacity, and kilometers driven exist on different scales. To ensure that the model doesn't become biased towards any one feature, normalization techniques, such as min-max scaling, are employed.
* **Encoding Categorical Data**: Features like fuel type are categorical, meaning they must be converted into a numerical form for the model. One-hot encoding creates binary columns for each fuel type, making them compatible with the regression model.

#### **3.3 Feature Selection**

Effective feature selection improves model accuracy and interpretability:

* **Correlation Analysis**: A correlation matrix can help identify which variables strongly correlate with the car price.
* **Backward Elimination**: Starting with all features, backward elimination removes the least significant predictors step by step until only the most important ones remain.
* **Forward Selection**: This method begins with no predictors and adds features based on their statistical significance

### **4. ALGORITHM**

#### **4.1 Model Development**

The multiple linear regression model is developed through the following steps:

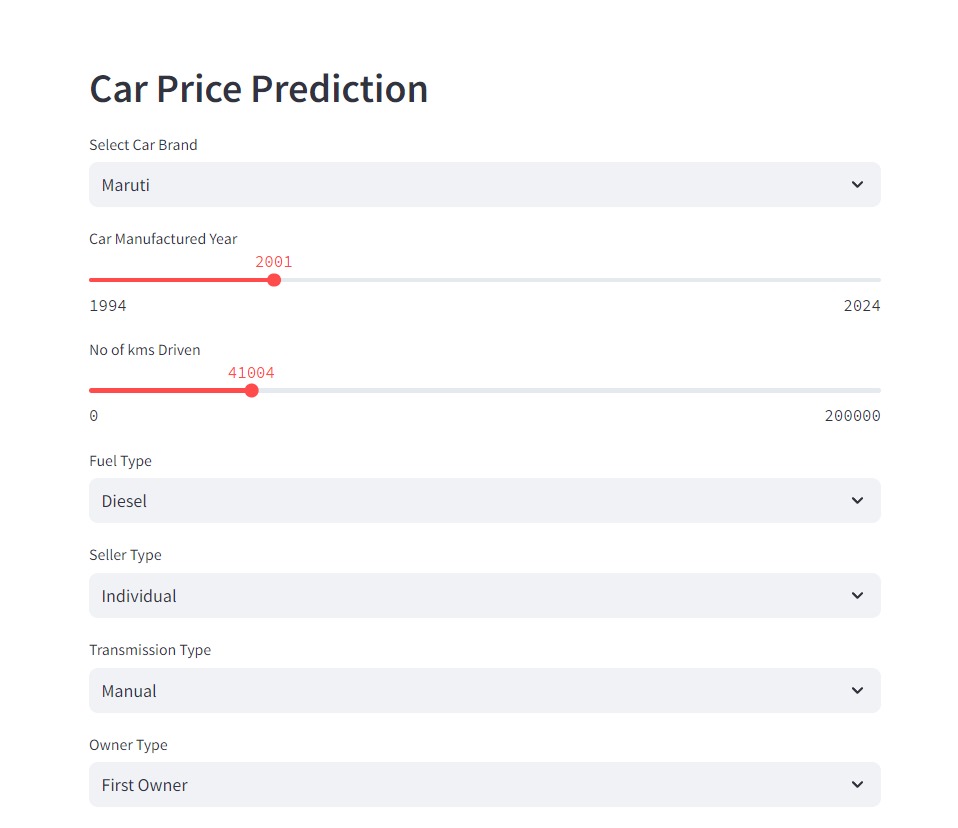
* **Splitting the Data**: The dataset is divided into training (80%) and testing (20%) sets. This ensures that the model is trained on a large portion of the data and tested for generalization on the unseen data.
* **Fitting the Model**: Using the training data, the model estimates the coefficients of the regression equation, minimizing the sum of squared residuals to find the best-fit line.
* **Model Evaluation**: Performance metrics such as R² (coefficient of determination) and RMSE (root mean square error) are used to evaluate how well the model predicts car prices. A high R² and low RMSE indicate a good fit.

#### **4.2 Implementation Tools**

* **Python**: Python is chosen for its simplicity and powerful libraries, such as Pandas, NumPy, and Scikit-learn.
* **Pandas**: Used for data manipulation and preprocessing, handling large datasets efficiently.
* **NumPy**: Facilitates numerical computations essential for matrix operations in the regression model.
* **Scikit-learn**: Provides built-in functions for regression modeling and evaluation, simplifying the development process.

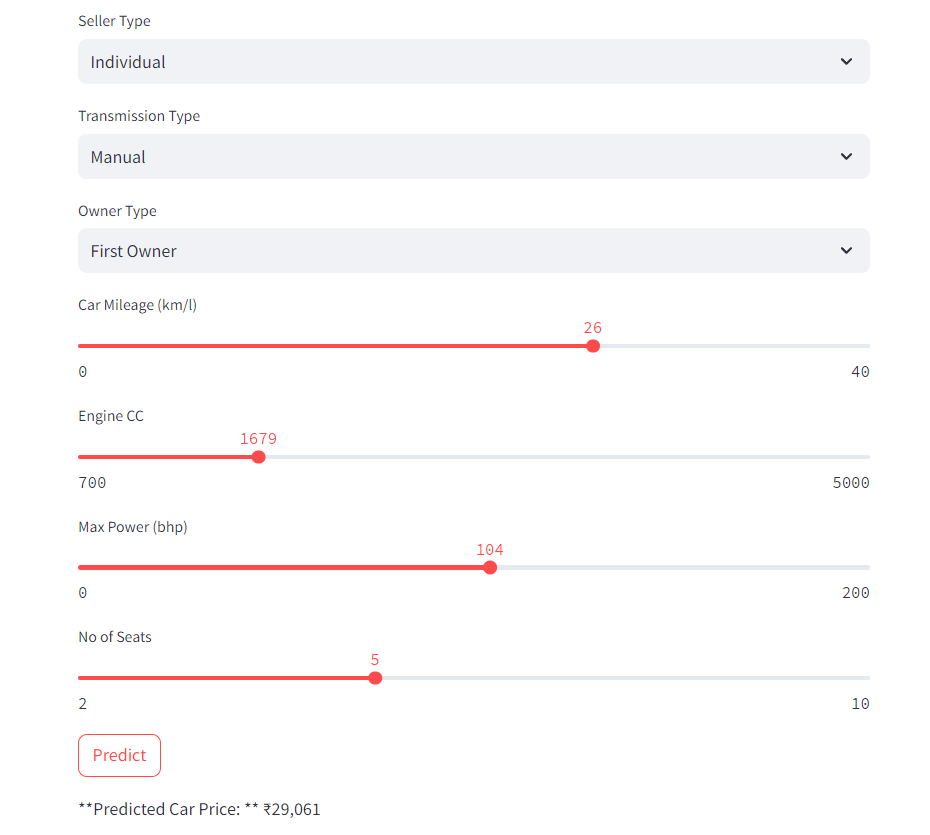
### 

### **5. OUTPUT SNAPSHOT**



#### **5.1 Results Visualization**

Visualization helps interpret the model’s performance:



* **Predicted vs. Actual Values**: Scatter plots compare predicted car prices against actual prices, helping to visually assess how well the model captures pricing trends.
* **Residual Plots**: These plots show the difference between predicted and actual prices, revealing any patterns that may violate regression assumptions, such as heteroscedasticity.

#### **5.2 Model Summary**

The model summary offers detailed insights into how each independent variable influences the car price and how well the model fits the data. The summary includes the coefficients for each variable, indicating their impact on the price, along with the p-values that determine the statistical significance of each factor. For instance, if the p-value for the 'Year of Manufacture' variable is below 0.05, it suggests that this factor significantly impacts the predicted car prices. The overall R-squared value indicates how much of the variance in car prices is explained by the model. A higher R-squared value suggests that the model captures most of the pricing dynamics.

### **6. DISCUSSION**

#### **6.1 Analysis of Key Variables**

* **Year of Manufacture**: This variable generally shows a strong negative correlation with car prices, as older cars tend to depreciate. However, certain classic or vintage cars may deviate from this trend, fetching higher prices due to their rarity.
* **Kilometers Driven**: As expected, cars that have been driven more tend to have lower prices, reflecting the wear and tear associated with higher mileage. However, the decline in value isn't always linear—some cars, especially those with excellent maintenance records, may not experience a sharp drop in price despite high kilometers.
* **Fuel Type**: The price variation across fuel types, such as petrol, diesel, and electric, is significant. Electric cars, while typically more expensive due to their technology, might experience rapid depreciation as newer models with better ranges and features enter the market. On the other hand, diesel vehicles, though fuel-efficient, may lose value due to rising environmental regulations.
* **Mileage**: Cars with higher fuel efficiency generally command better prices, especially in markets where fuel costs are high. This variable can sometimes outweigh factors like engine size, as consumers become more environmentally conscious and budget-aware.
* **Engine Capacity (cc)**: Larger engines usually suggest higher prices, especially for performance-oriented cars. However, for economy cars, a large engine might not always translate into higher prices, particularly if fuel efficiency is a greater priority for the buyer.
* **Seating Capacity**: Larger vehicles, especially SUVs with more seats, often fetch higher prices, making this a key variable in the prediction model, particularly in markets where family-sized cars are in high demand.

#### 

#### **6.2 Model Evaluation**

The model's predictive accuracy is evaluated using several performance metrics:

* **R-squared (R²)**: This value represents the proportion of the variance in the dependent variable (car price) that is predictable from the independent variables. An R² close to 1 suggests that the model explains most of the variation in car prices.
* **Root Mean Square Error (RMSE)**: RMSE gives an indication of how far the predicted car prices are from the actual values, with a lower RMSE indicating better model performance. While no model can perfectly predict prices, the goal is to minimize this error as much as possible.
* **Adjusted R²**: Unlike R², which can artificially inflate as more variables are added, adjusted R² provides a more accurate measure of model performance by adjusting for the number of predictors used. This ensures that only meaningful variables contribute to the model's explanatory power.

### **7. CONCLUSION**

In conclusion, the car price prediction model developed in this project demonstrates how multiple linear regression can be applied to estimate car prices based on several key factors such as year of manufacture, kilometers driven, fuel type, mileage, engine capacity, and seating capacity. By using a combination of historical data, feature selection, and machine learning techniques, the model provides accurate and actionable insights into car pricing. The findings from this project can assist both buyers and sellers in making informed decisions, helping them navigate the complex landscape of the car market. As future work, the inclusion of more features, advanced algorithms, and external data sources could further enhance the model's accuracy and applicability in real-world settings.

### **REFERENCES**

1. A. Nair and S. Sharma, "Predictive Modeling for Car Prices: A Machine Learning Approach," *International Journal of Data Science and Analytics*, vol. 7, no. 3, pp. 45-59, 2023.
2. Kaggle, "Used Car Price Prediction Dataset," [Online]. Available: https://www.kaggle.com/datasets/used-car-prices.