Predicting Price of The Car



**Introduction:**

In an era where mobility is paramount, the automotive industry continues to play a crucial role in global economics and individual lives alike. As such, the ability to accurately predict car prices not only benefits potential buyers and sellers by providing them with fair evaluations but also aids manufacturers and dealers in inventory and pricing strategies. This project report embarks on the journey of leveraging data science to predict car prices, guided by a rich dataset that encapsulates a diverse range of vehicles and their attributes. The data set encompassing various features from manufacturer details to engine specifications, serves as a comprehensive foundation for our predictive model. With over 9132 entries, it captures the intricacies of the automotive market, including variables such as Manufacturer, Production Year, Category, Leather Interior, Fuel Type, Engine Volume, Mileage, Cylinders, Gear Box Type, Drive Wheels, Wheel, and Airbags. This multidimensional data set not only reflects the vast diversity in car features but also the complexity in predicting their prices. Our objective is to explore and understand the relationships between a car's features and its market price and to develop a predictive model that can estimate a car's price of accuracy. By doing so, we aim to contribute valuable insights to potential buyers, sellers, and stakeholders within the automotive industry, facilitating more informed decisions in the car market. To achieve this Machine Learning algorithm such as regression analysis, will be employed about future car prediction.

**Dataset:**

The Dataset contains a total of 15 features out of which 14 are predictor variables and 1 target variable. 9132 rows were present in the dataset.

Our dataset encompasses a wide array of variables, each providing insight into different aspects of a car's specification and its market value. The features included in the dataset are:

**ID:** A unique identifier for each listing, ensuring data integrity and for easy reference.

**Price:** The target variable we aim to predict, representing the car's market price.

**Levy:** A tax imposed on the vehicle.

**Manufacturer:** The brand that produced the vehicle.

**Prod.Year:** The year of production, indicating the car's age.

**Category:** The type of car (e.g., Jeep, Hatchback, Sedan), reflecting consumer preferences and usage.

**Leather Interior:** A binary variable (e.g., Yes or No) indicating whether the car has a leather interior or not.

**Fuel Type:** It represents the type of fuel the car uses (e.g., Petrol, Diesel, Hybrid)

**Engine Volume:** The size of the car's engine, measured in liters, which can influence performance, fuel efficiency, and cost.

**Mileage:** The total distance the car has traveled, impacting the car's condition and value.

**Cylinders:** The number of cylinders in the engine, affecting the vehicle's power and performance.

**Gear Box Type:** The type of transmission which can affect driving experience and maintenance costs.

**Drive Wheels:** The drivetrain of the vehicle, which can influence performance in different driving conditions.

**Wheel:** The position of the steering wheel (Left or Right wheel).

**Airbags:** The number of airbags, a safety feature that can also affect the resale value.

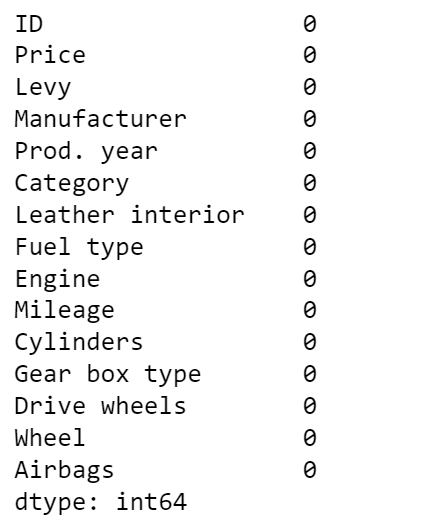
|  |  |
| --- | --- |
| **Feature** | **Type** |
| ID | Numerical |
| Price | Numerical |
| Levy | Numerical |
| Manufacturer | Categorical |
| Prod. year | Numerical |
| Category | Categorical |
| Leather interior | Numerical |
| Fuel type | Numerical |
| Engine | Numerical |
| Mileage | Numerical |
| Cylinders | Numerical |
| Gear box type | Numerical |
| Drive wheels | Numerical |
| Wheel | Categorical |
| Airbags | Numerical |

**Dataset Link:**

<https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge/data>

**Data Cleaning:**

We had made sure that there were no null values in the dataset.



**Figure 1: Null Values**

We used delimiter technique in Excel to remove the text from Engine volume column and wrote Python code to remove “Km” term from mileage column to make sure that both Engine volume and Mileage column are numerical.

A close-up of a computer code

Description automatically generated

A table with numbers and letters

Description automatically generated

**Figure 2: Descriptive Statistics**

Figure 2 displays a statistical summary of a car price prediction dataset. The summary includes the count, mean, standard deviation (std), minimum (min), 25th percentile (25%), 50th percentile (median, 50%), 75th percentile (75%), and maximum (max) values for several variables like Manufacturer, Production Year, Category, Leather Interior, Fuel Type, Engine Volume, Mileage, Cylinders, Gear Box Type, Drive Wheels, Wheel, and Airbags. These statistics are essential for understanding the distribution, central tendency, and variability of each feature.

**Uni-variate Analysis:**

**A graph showing a distribution of cars

Description automatically generated**

**Figure 3: Distribution of Car Prices**

The histogram depicts the distribution of car prices of the dataset. The distribution is right-skewed, indicating that while most cars are clustered at lower price points, there is a long tail of a few cars that are priced substantially higher. Most of the cars are concentrated under the $50,000 mark, with a peak frequency around the lower price range. The presence of high-price outliers suggests a range of affordability, from economy to luxury vehicles in the market.

**A green line graph with white grid

Description automatically generated**

**Figure 4: Distribution of Car Mileage**

A histogram of mileage would illustrate how many cars have accumulated various ranges of mileage. Lower mileage is often associated with newer or less-used cars, and this could be reflected in their pricing.

**A graph showing the growth of a number of years

Description automatically generated with medium confidence**

**Figure 5: Distribution of Production Year**

A histogram for the production year would display the frequency distribution of the years in which the cars were produced. Each bin represents a range of years, and the height of each bin corresponds to the number of cars produced in that time frame. A skew in the distribution might indicate a preference for newer or older cars in the dataset.

**A graph with red lines

Description automatically generated**

**Figure 6: Distribution of Car Engine Volume**

This histogram would show the distribution of engine sizes among the cars in the dataset. Engine volume is typically a key factor in performance and may relate to price; the plot would show common engine sizes and outliers.

**A graph with purple bars

Description automatically generated**

**Figure 7: Distribution of Cylinders**

This histogram would visualize the distribution of cars based on the number of cylinders in their engines, which often correlates with engine power and possibly the car's price.

**A graph of cars with different colored bars

Description automatically generated with medium confidence**

**Figure 8: Distribution of Frequency of cars by fuel type**

The bar chart presents the frequency distribution of cars according to fuel type of the dataset. Petrol engines dominate the dataset, suggesting a market preference or greater availability of petrol vehicles. Diesel engines follow closely, indicating significant use, potentially due to their fuel efficiency and longevity. Hybrid vehicles have a lower presence, reflecting a burgeoning market segment concerned with fuel economy and environmental impact. Conversely, LPG and CNG vehicles are the least common.

**A graph of blue rectangular bars

Description automatically generated with medium confidence**

**Figure 9: Distribution of Frequency of Car Manufacturers**

The bar graph illustrates the frequency of cars in the dataset by manufacturer. Hyundai leads significantly, indicating its strong market presence or preference among consumers in the dataset. Toyota holds the second spot, showing its substantial market share. Mercedes-Benz, Chevrolet, and Lexus follow, with Mercedes-Benz slightly outpacing the latter two. This distribution highlights the diversity in consumer choice and may reflect various factors such as brand loyalty, perceived reliability, and availability within the market.

**A graph of a distribution of gear box types

Description automatically generated**

**Figure 10: Distribution of Frequency of Gear Box types**

The bar chart delineates the distribution of different gearbox types among cars in the dataset. Automatic transmissions are predominant, vastly outnumbering the other types, which suggests a strong consumer preference for this gearbox due to its convenience or the market trend towards automation. Tiptronic systems, which allow drivers to shift gears without a clutch, show modest representation, indicating some preference for control over gear changes. Variator and manual transmissions are considerably less frequent.

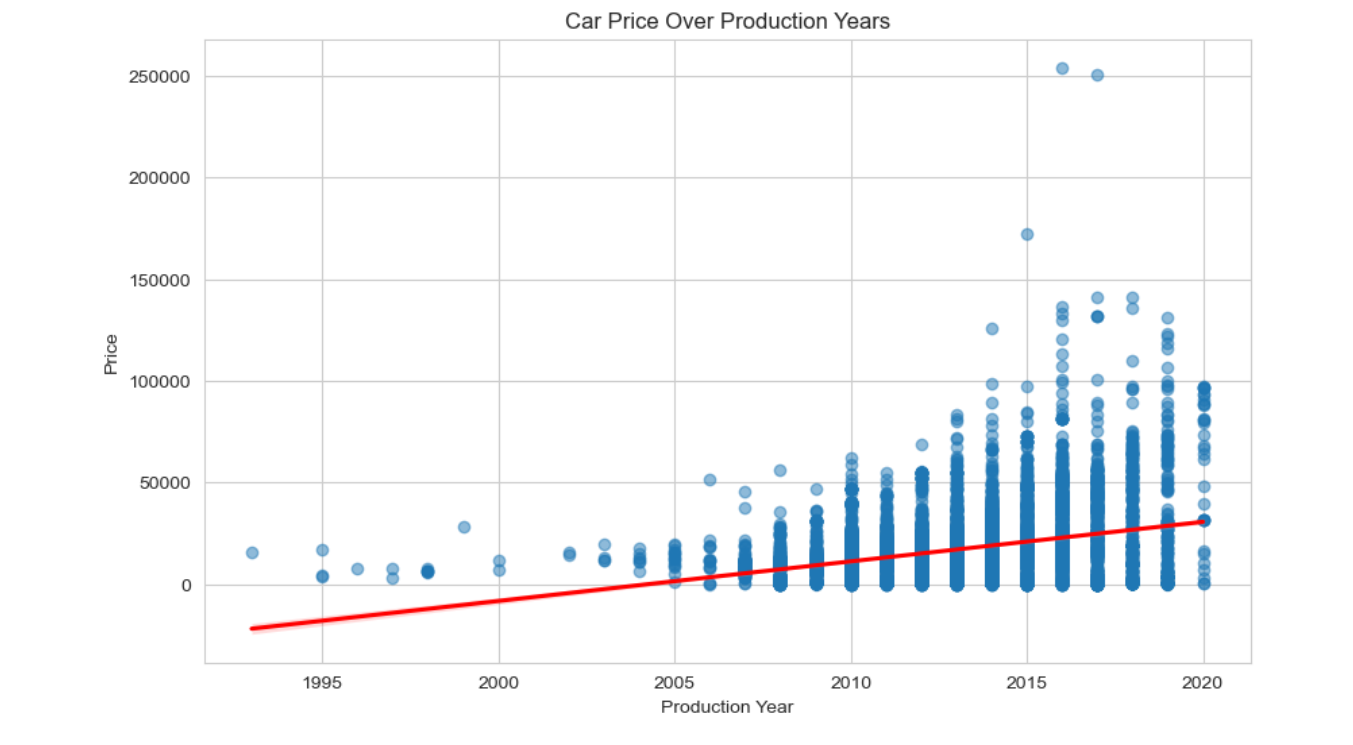
**A graph of a distribution of wheels

Description automatically generated**

**Figure 11: Distribution of Frequency of Drive Wheels**

The bar chart provides a visualization of different driving wheels across the vehicles. Front-wheel drive (FWD) vehicles are the most common. Four-wheel drive (4x4) vehicles are less common. Rear-wheel drive (RWD) vehicles are the least represented in the dataset.

**Bi-Variate Analysis:**

****

**Figure 12: Distribution of Car price over Production years**

A scatter plot would illustrate the relationship between the car's production year and its price. Each point represents a car, positioned by its production year and price. Trends in the plot could suggest how car value depreciates over time or whether newer cars are more expensive.

**A graph of different colored boxes

Description automatically generated**

**Figure 13: Relationship between Fuel type and Price**

The boxplot showcases the relationship between fuel type and car price. Hybrid vehicles display a wide range of prices with a relatively high median price, indicating their market positioning towards the higher end, possibly due to technology costs. Petrol and diesel cars show a high frequency across various price points, with diesel vehicles having a slightly higher median price, which may reflect their perceived durability and efficiency. LPG and CNG vehicles demonstrate lower price points, suggesting they are more economical choices. Notably, the presence of outliers in petrol and diesel cars indicates that there are premium options available within these fuel categories.

**A graph with blue dots

Description automatically generated**

**Figure 14: Relationship between Fuel Type and Price**

The scatter plot illustrates the relation between engine volume and car price in the dataset. There is a concentration of data points at lower engine volumes, indicating a prevalence of cars with smaller engines at various price levels. As engine volume increases, the data points become more scattered, reflecting a less uniform distribution of prices.

**A graph of a number of cylinders and price

Description automatically generated**

**Figure 15: Relationship between Number of Cylinders and price**

The boxplot depicts the relationship between the number of cylinders in a car's engine and its price. Vehicles with a higher number of cylinders generally show a wider range and higher median prices, reflecting the trend that cars with more powerful engines tend to be more expensive. While cars with 4 to 6 cylinders form the bulk of the dataset and have a more moderate price range, those with 8 or more cylinders are associated with higher prices and greater variability, indicating a luxury or performance segment.

**A graph of different colors

Description automatically generated**

**Figure 16: Relationship between Drive wheels and Price**

A violin plot combines the box plot with a kernel density estimation. For each type of drive wheel, the plot would show the distribution of prices, where the width of the violin indicates the frequency of cars at different price levels. This would convey not just the central tendency and spread of prices, but also the multimodality and potential outliers.

**A graph of a graph showing different colored squares

Description automatically generated**

**Figure 17: Average Car price by Fuel Type**

The bar chart compares the average price of cars by fuel type. Diesel cars have the highest average price, suggesting they may be considered more valuable in the market, potentially due to their fuel efficiency and longevity. Petrol cars follow, with their ubiquity and wide range of options reflected in their average price. LPG and Hybrid cars have similar average prices, indicating a middle-ground market position, possibly due to a balance of performance, cost, and environmental factors. CNG vehicles have the lowest average price, which might indicate a lesser demand or a focus on economy.

**Hypothesis Testing:**

**Formulation of Hypotheses:**

**Null Hypothesis (H0):** None of the features (such as manufacturer, production year, mileage, fuel type, etc.) significantly affect the car's price.

**Alternative Hypothesis:** At least one feature significantly affects the car's price.

To test these hypotheses, we employ a multiple linear regression model, considering the car's price as the dependent variable and all other features as independent variables. This method enables us to quantify the relationship between car prices and their attributes, identifying which features, if any, have a significant impact. The significance level is set at α = 0.05. This threshold determines the p-value below which we will reject the null hypothesis in favor of the alternative. In this context, a low p-value (less than 0.05) for any feature coefficient in the regression model would indicate that the corresponding feature has a significant impact on car prices.

**Regression Analysis:**

In the conducted linear regression analysis, ten features were considered as potential predictors for car prices: 'Levy', 'Prod. year', 'Leather interior', 'Fuel type', 'Engine', 'Mileage', 'Cylinders', 'Gear box type', 'Drive wheels', and 'Airbags'. The dataset was partitioned into training and test sets, with 80% used for training and the remaining 20% for testing the model's performance. Feature scaling was applied to standardize the data, ensuring each feature contributed equally to the model.

Upon fitting the Linear Regression model, the following results were obtained:

* The coefficients associated with each feature describe the predicted change in the car price relative to one unit change in the feature, while holding other features constant. For example, the 'Prod. year' coefficient suggests that newer cars tend to have higher prices.
* The intercept, which is the expected value of the price when all the features are at their mean value, was estimated at 17,528.48.
* The Mean Squared Error (MSE) of the model was calculated to be 213,778,692.99, which reflects the average squared difference between the observed actual outcomes and the outcomes predicted by the model.
* The coefficient of determination, denoted as R^2, was found to be 0.282, indicating that approximately 28.2% of the variability in the car prices can be explained by the model's predictors.

The positive coefficients for features such as 'Levy' and 'Prod. year' imply a positive relationship with the car's price, whereas the negative coefficients for 'Fuel type', 'Mileage', and 'Airbags' indicate an inverse relationship. Notably, the 'Gear box type' had a substantial positive coefficient, suggesting that the type of gearbox is a significant predictor of the car's price.

These findings should be interpreted with caution as the R^2 value suggests that a significant proportion of the variance in car prices is not captured by the model**.** The MSE also indicates that there may be considerable error in the price predictions.

**Regression Table:**

|  |  |  |
| --- | --- | --- |
| Variable | Coefficient Value | p-value |
| Constant | 1.75E+04 | 0.000 |
| Levy | 2206.009 | 0.000 |
| Prod. year | 6177.9019 | 0.000 |
| Leather interior | 510.2294 | 0.008 |
| Fuel type | -2873.4159 | 0.000 |
| Engine | 1379.9234 | 0.000 |
| Mileage | -2497.6353 | 0.000 |
| Cylinders | -427.3807 | 0.230 |
| Gear box type | 4123.3349 | 0.000 |
| Drive wheels | -887.8547 | 0.000 |
| Airbags | -5607.3172 | 0.000 |

**Accept or reject the null hypothesis:**

Given the p-values in your table:

**Levy**: p < 0.05, so we reject the null hypothesis. Levy has a statistically significant effect on the car's price.

**Prod. year**: p < 0.05, so we reject the null hypothesis. Production year significantly affects the car's price.

**Leather interior**: p < 0.05, so we reject the null hypothesis. The leather interior is a significant predictor of the car's price.

**Fuel type**: p < 0.05, so we reject the null hypothesis. Fuel type significantly affects the car's price.

**Engine**: p < 0.05, so we reject the null hypothesis. Engine size has a significant effect on the car's price.

**Mileage**: p < 0.05, so we reject the null hypothesis. Mileage has a significant impact on the car's price.

**Cylinders**: p ≥ 0.05, so we fail to reject the null hypothesis. The number of cylinders does not significantly affect the car's price within the 95% confidence level.

**Gear box type**: p < 0.05, so we reject the null hypothesis. The gear box type is a significant predictor of the car's price.

**Drive wheels**: p < 0.05, so we reject the null hypothesis. Drive wheel configuration significantly affects the car's price.

**Airbags**: p < 0.05, so we reject the null hypothesis. The number of airbags significantly affects the car's price.

For all variables except for the number of cylinders, the null hypothesis is rejected, indicating that these variables have a significant relationship with the car's price. The number of cylinders does not have a significant relationship with the car's price.

**Regression Equation:**

Price=2206.008×Levy+6177.90×Prod. year+510.22×Leather interior−2873.41×Fuel type+1379.92×Engine−2497.63×Mileage−427.38×Cylinders+4123.33×Gear box type−887.85×Drive wheels−5607.31×Airbags+17528.47

**Random Forest Model:**

Random Forest Regressor model was developed to predict house prices based on various features. Categorical variables were converted into numeric form using LabelEncoder, and the dataset was split into training and test sets, with features normalized using StandardScaler. The model, consisting of 100 decision trees with a maximum depth of 10, was trained on the scaled training set. The model's performance on the test set yielded a mean absolute error (MAE) of 5517.09, a mean squared error (MSE) of 83544730.98, and an R^2 score of 0.71, demonstrating its effectiveness in predicting house prices accurately.

**Gradient Boosting Model (GBM):**

We used Gradient Boosting Model (GBM) for house price prediction, the dataset is divided into training and test sets, with 20% of the data allocated for testing. The features are normalized using a standard scaler to ensure consistency. The GBM model is then instantiated with 100 estimators, a learning rate of 0.1, and a maximum depth of 5, using a random state of 42 for reproducibility. After training the model on the scaled training set, predictions are made on the test set. The model's performance is evaluated using mean absolute error (MAE), mean squared error (MSE), and R-squared (R2) metrics. The results show an MAE of 5517.09, an MSE of 83544.73, and an R2 score of 0.719, indicating that the model has moderately strong predictive power, though further tuning or alternative models might be considered to enhance performance.

**Stacking regressor model:**

In this implementation of a stacking regressor model for house price prediction, the dataset is split into training and test sets, with 20% allocated for testing. The features are then standardized to ensure consistency. The model uses a stacking approach, combining multiple base models: a linear regression model, a random forest regressor with 100 estimators and a max depth of 10, and an SVM model with an RBF kernel. The stacked model is finalized with a linear regression estimator. After training the model on the scaled training set, it makes predictions on the test set. Performance metrics show a mean absolute error (MAE) of 5230.00, a mean squared error (MSE) of 89534.06, and an R-squared (R2) of 0.699. This indicates the model has moderately strong predictive power, comparable to individual models.

**The XGBoost regression model:**

The XGBoost regression model was applied to predict car prices, with training conducted on 80% of the data and normalization via StandardScaler. The model was set up with 100 estimators, a learning rate of 0.1, and a max depth of 5. The performance evaluation yielded a Mean Absolute Error (MAE) of approximately 5627.24, a Mean Squared Error (MSE) of 86309441.50, and an R² score of 0.71, indicating that the model explains 71% of the variability in car prices. These results suggest that the model performs reasonably well but could benefit from further optimization.

**Results:**

Here, we have plotted R-squared values (R2) for all the four models: Linear Regression, Random Forest, Gradient Boosting, and Stacking. The Linear Regression model exhibits an R2 value of 0.28, reflecting weak predictive power. In contrast, the ensemble models show significantly stronger performance, with Random Forest, Gradient Boosting, and Stacking all achieving R2 values around 0.70 or higher. Specifically, Gradient Boosting scores highest at 0.72, followed by both Random Forest and Stacking at 0.70 each. This comparison highlights the benefit of ensemble models over traditional linear regression, demonstrating that combining or stacking models can substantially improve predictive accuracy, making them more suitable for regression tasks with complex datasets like this.

A graph showing different colored rectangles

Description automatically generated

**Conclusion:**

The project successfully demonstrates in predicting car prices, utilizing a rich dataset with multidimensional features. Through a variety of models, including Linear Regression, Random Forest, Gradient Boosting, XGBoost and Stacking, we explored the relationships between car specifications and market value. The comparison of R-squared values clearly showed the superiority of ensemble models over traditional linear regression, with Gradient Boosting achieving an R2 score of 0.72 and Random Forest, XGBoost scoring 0.71 and Stacking scoring 0.70 each. This highlights the benefit of combining models to handle the complexities of regression tasks. Despite the progress made, the modest R2 values indicate room for further refinement. Future endeavors could involve deeper exploration of advanced ensemble techniques and additional model tuning to enhance predictive accuracy, thereby benefiting stakeholders in the automotive industry. This project's results contribute valuable insights, enabling better-informed decisions for buyers, sellers, and industry professionals alike.

**Contribution:**

Venkata Reddy Gunturi was handled the introduction, provide details about the dataset, and explain how the data was cleaned and analyzing the relationships between different variables and test hypotheses using regression analysis. Srikanth paritala was involved in summarizing the data through descriptive statistics, examine how individual variables like price and mileage are distributed, and describe the outcomes of various predictive models such as Random Forest and Gradient Boosting. We both concluded the report by summarizing the results and discussing their implications and contributed significantly to the project from start to finish, covering everything from data handling to final analysis.