2024S-T2 BDM 2203

Project Report: Car Price Prediction

**Submission By:**

Govind Dogra

Abdul Rahim Mohammed

Diksha Gori

Jashanpreet Kaur

Rajkaran Yasodhadevi Ponnuvelusamy

|  |  |  |
| --- | --- | --- |
| Members | Tasks | |
| Govind Dogra | Power BI Dashboard | * Data Preparation * KPIs * Visualizations |
| Abdul Rahim Mohammed | Data Preparation | * Data Collection * Data Preprocessing * GUI Development * Feature engineering |
| Diksha Gori | Deployment | * Deployment * EDA |
| Jashanpreet Kaur | Project Management | * Report Writing * Presentation Preparation |
| Rajkaran Yasodhadevi Ponnuvelusamy | Model Implementation  Model Evaluation | * Linear Regression * Decision Tree Regressor * XGBoost Regressor * Evaluation Metrics (MSE, R² Score) |

**Introduction:** The primary objective of this project is to develop a machine learning model that can predict car prices based on features such as make, model, year, and condition. By understanding the factors that most influence pricing, the model can provide actionable insights for pricing strategies, helping to align listed prices with market expectations. This project encompasses data collection, preprocessing, model training, and evaluation. It focuses on regression models, as the target variable car price is continuous. The project is limited to the available dataset and the selected machine learning models, with scope for further enhancement in future iterations.

**Data Understanding and Preparation**

**Data Collection**  
**Dataset Source:** Vehicle listings, auction outcomes, and dealership records are possible sources of the information. Making sure the data includes a wide variety of automobiles and is reflective of the market is crucial.

**Feature Overview**: The dataset typically includes features such as:

* **Make and Model**: The brand and specific type of vehicle, which often affect perceived value.
* **Year**: Indicates the age of the car, with newer cars generally being more expensive.
* **Odometer Reading**: Reflects how much the car has been used, with higher mileage often leading to lower prices.
* **Condition**: A qualitative measure of the car's wear and tear.
* **Other Features**: May include location (e.g., region), fuel type, and transmission, each contributing differently to the car's price.

**Data Cleaning**

**Handling Missing Values:** In real-world datasets, missing data can frequently be a problem. Missing values can be filled in using approaches like mean/mode imputation or more complex ones like regression imputation. Rows that have a large amount of missing data can also be eliminated.

**Finding outliers:** Extreme values that dramatically deviate from other data are known as outliers. They may skew the model's learning process and produce predictions that are not accurate. These numbers can be recognized and handled with the aid of outlier detection techniques like Z-score or IQR approaches.

**Data Transformation**

**Encoding Categorical Variables:** Numerical input is necessary for many machine learning models. The conversion of categorical variables, such as "Make" or "Model," into numerical representations is accomplished through the use of one-hot encoding, label encoding, or binary encoding techniques. Every approach offers advantages based on the type of data.

**Feature Scaling:** There are wide ranges in the scale of features such as pricing and odometer readings. Normalization or standardization of characteristics to have comparable ranges can improve the performance of models like neural networks and linear regression. Z-score normalization and Min-Max scaling are two popular scaling techniques.

**Data Exploration**

EDA facilitates comprehension of the underlying trends, feature relationships, and any problems in the data.

**Statistical Summary:** Standard deviation, mean, and median are examples of descriptive statistics that provide light on the distribution and central tendencies of the data.

**Correlation Analysis:** Determining whether features may be good predictors of the target variable involves examining the correlation between the features. For instance, a strong association between "Year" and "Price" would imply that newer vehicles are typically costlier.

**Methods of Visualization:** Histograms, scatter plots, and heatmaps are examples of visualization techniques that are used to analyze data distributions, find abnormalities, and reveal linkages.

**Modeling**

**Linear Regression:** A basic model based on the assumption that independent variables (features) and dependent variables (price) have a linear relationship. Simple to understand, it might not fully convey intricate relationships.

**Decision Tree Regressor:** A nonlinear model called a decision tree regression divides data into subgroups according to feature values and then arranges the pieces into a structure like a tree. In comparison to linear regression, it is more flexible, but if not adjusted properly, it may become overfit.

**The XGBoost Regressor:** A complex ensemble method that enhances accuracy by aggregating predictions from several weak learners (typically decision trees). Performance and versatility in handling different kinds of data and complexity are two of XGBoost's best-known attributes.

**Model Implementation**

**Training and Validation:** Training and validation sets (such as an 80-20 split) are usually created from the dataset. Cross-validation methods, such k-fold cross-validation, are employed to make sure the model functions effectively on various data subsets.

**Hyperparameter tuning:** Performance can be enhanced by adjusting the parameters of models such as XGBoost. The best combination of parameters is found using methods like random or grid search.

**Model Evaluation**

**Average Squared Error (MSE):** Calculates the average squared deviation between the values that were expected and those that were observed. MSE values below.001 suggest higher accuracy.

**R² Score:** The ratio of predictability in the dependent variable to that in the independent variables is indicated by the R2 score. Excellent prediction accuracy is indicated by an R2 value that is near to 1.

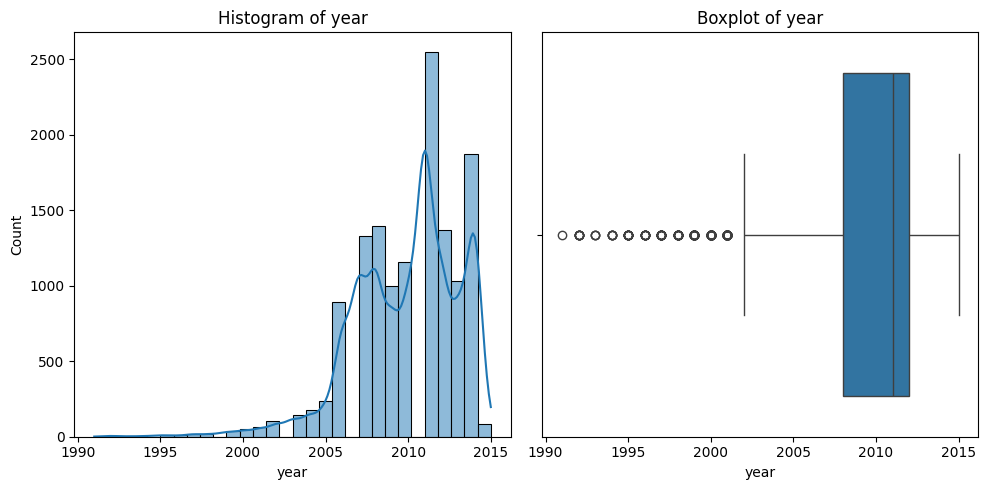
**Key Visualizations**

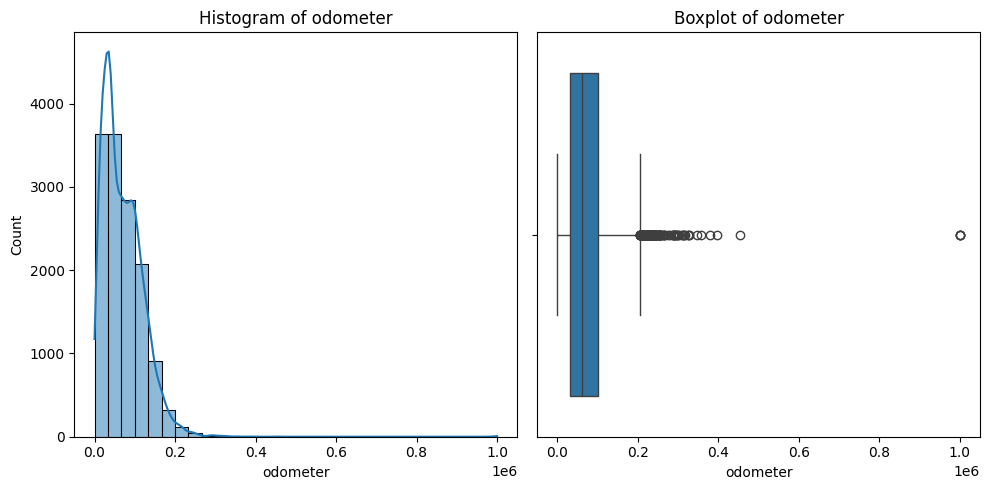
**Price Distribution Visualization**

**Visualization Type**: Histogram or Box Plot

**Histogram:** A histogram shows the frequency of automobile prices at various intervals (bins). This aids in determining the most prevalent price ranges, the presence of several modes (peaks), and the degree of data skewness (left or right). The histogram may show that the majority of cars fall into a particular price range, which represents market trends. For example, a peak in the histogram at about $15,000 would indicate that used cars in the sample typically sell for this amount.

**Box Plot:** A box plot shows the median, quartiles, and any possible outliers in an overview of the price distribution. This is especially helpful in determining how dispersed the prices are and whether any extreme levels call for additional research. If major outliers are present, such as very expensive or very luxurious cars, the box plot can highlight them and highlight how they may distort the results of the analysis as a whole if left unchecked.

****

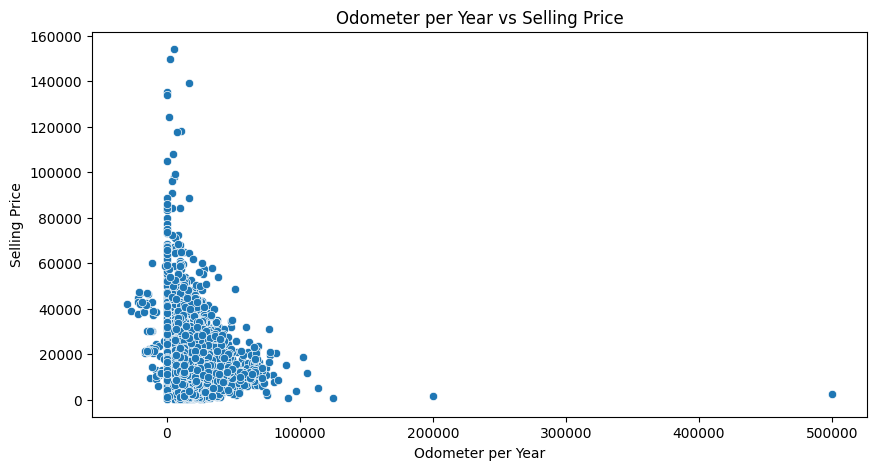
****

#### Feature Relationship Visualization

**Visualization Type**: Scatter Plot, Pair Plot

**Scatter Plot:** To show the relationship between two continuous variables, scatter plots are utilized. Plotting "Odometer Reading" against "Price," for instance, can show how mileage influences a car's worth. A scatter plot with a decreasing slope would show that costs usually decrease with increasing miles. The scatter plot might reveal a linear or non-linear relationship between "Year" and "Price," with newer cars generally having higher prices.

**Pair Plot:** A pair plot is a scatter plot matrix for more than one feature. This enables a thorough understanding of the relationships between each feature and the cost. It comes in very handy when handling several variables.



#### Categorical Feature Impact Visualization

**Visualization Type**: Box Plot for Categorical Features

**Box Plot for Categorical Features:** A box plot can compare the distribution of prices across different categories within a categorical feature. For example, comparing prices across different "Fuel Types" can reveal which type generally leads to higher or lower car prices. A box plot for "Make" might reveal that certain brands, like "BMW" or "Mercedes," generally have higher median prices compared to "Ford" or "Toyota." This insight could inform brand-specific pricing strategies.



**Power BI Dashboard**

1. Changed datatype of column ‘mmr’ and ‘sellingprice’ to currency

2. Changed some Column names for clarity and better understandability.

3. Converted ‘saledate’ column with values such as (Thu Dec 18 2014 12:30:00 GMT-0800 (PST)) into 3 new columns as Day\_sold, Date\_sold and Time\_sold.

4. KPI’s:

* Total Sales revenue: total revenue generated by selling cars.

Query:

Total Sales Revenue = SUM(car\_prices[Selling\_Price])

* Number of cars sold

Query:

Number of Cars Sold = DISTINCTCOUNT(car\_prices[vin])

* Total Variants: total variants of cars.

Query:

Number of Cars Sold = DISTINCTCOUNT(car\_prices[vin])

* Monthly Target: defined the monthly sales target with defined threshold.

5. **Visualisations:**

* **Total Sale by brand:**

**Insight:** This bar chart displays the overall sales income for each automobile brand.

**Key observations**: Ford has the most sales, followed by Chevrolet and Nissan.

Other brands, such as Toyota, BMW, and Mercedes-Benz, make substantial contributions, although their sales are smaller than Ford and Chevrolet.

* **Top 5 models by revenue**:

**Insight**: This chart shows the top five automobile models by revenue.

**Key observations**:

The Ford F-150 is the best-selling vehicle, generating the greatest income.

Other models, like as the Altima, Escape, Fusion, and G Sedan, follow, each generating significant sales.

* **Geographical Sales distribution:**

**Insight**: The map depicts sales geographically across the United States.

**Key Observations:** Sales are concentrated in certain places, with larger bubbles suggesting higher quantities.

Sales are mostly concentrated in the United States.

* **Dealership with Highest Sales:**

**Insight:** This table ranks the top dealerships and sellers based on total sales income.

**Key observations:**

Ford Motor Credit Company LLC leads with revenues of more than $314 million.

Other major vendors are Hertz Corporation, Nissan-Infiniti, and Axis Corporation.

The list includes a full analysis of the top-grossing dealerships.

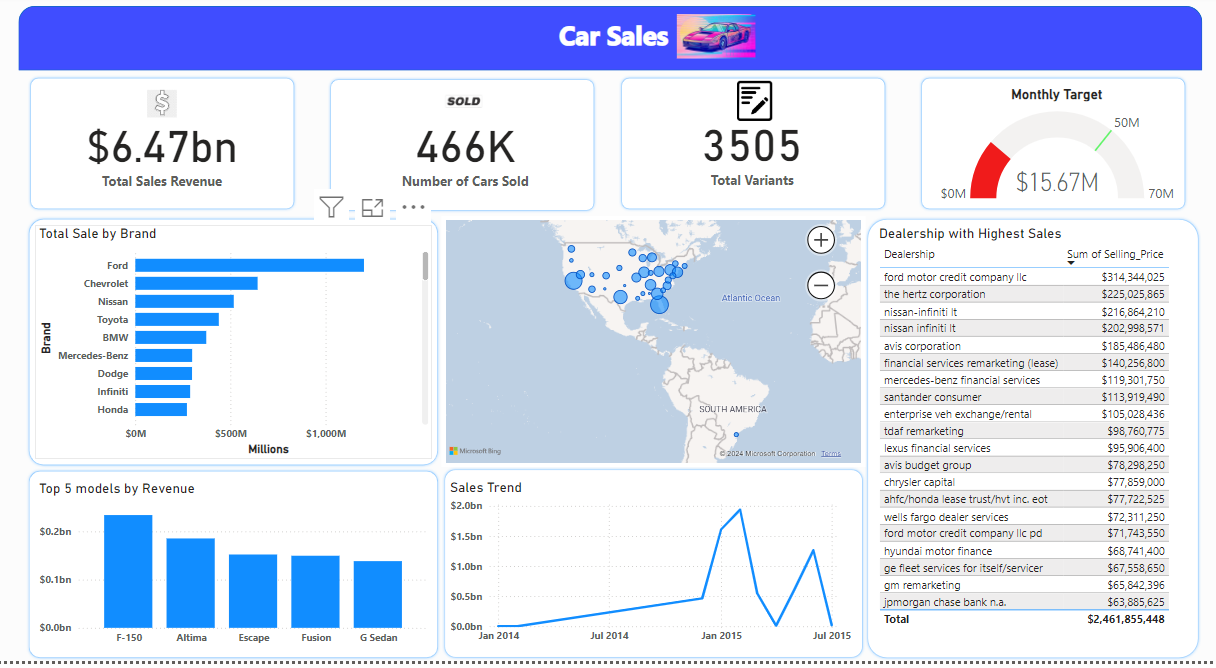
* **Sales Trend over time**

**Insight**: A line chart shows the total sales trend over time.

**Key observations**: Sales increase significantly in early 2015.

There is also a significant drop following the peak, with sales returning later, showing variations in demand or sales activity.

**Dashboard:**

****

**Conclusion:** A machine learning model for automobile price prediction and a dashboard for visualizing automotive sales data are the two main components of the project. The dashboard helps to detect important patterns and performance indicators in the automobile market by offering insights into sales trends, top-selling brands and models, and geographic sales distribution. In terms of predicting automobile pricing based on attributes like make, model, year, and condition, the XGBoost Regressor has proven to be the most successful regression technique used in the machine learning component. When combined, these elements seek to improve comprehension and judgement regarding vehicle pricing and sales tactics.