Predicting Car Prices Using Machine Learning

Done by: Asrar Sharaf

EPSILON AI ACADEMY

Project Proposal

1. Introduction / Background

The automotive industry has seen a significant rise in used and new car sales. Accurately predicting car prices helps buyers make informed decisions and enables sellers to optimize their pricing strategies. Machine Learning (ML) provides powerful tools to analyze market trends, understand factors affecting car prices, and make reliable price predictions. This project applies ML techniques to predict new and used car prices and extract actionable insights from car sales data.

2. Problem Statement

The goal of this project is to predict new and used car prices based on multiple features, such as car condition, mileage, manufacturer, fuel type, drivetrain, transmission, model year, and exterior/interior color. Additionally, the project aims to answer key business questions about market trends, price influencers, and customer preferences to guide decision-making.

3. Objectives

- Develop a machine learning model to accurately predict car prices.
- Analyze factors influencing car prices, such as condition, manufacturer, fuel type, and color.
- Provide actionable business insights and recommendations for car dealerships.
- Deploy an interactive application for price prediction and market analysis.

4. Dataset Description

The dataset was sourced from **Kaggle** and contains **9,246 car entries** initially, with 10 primary columns. After cleaning and preprocessing, the dataset was reduced to **5,235 entries** with **13 columns** including:

- condition – car condition (New, Used, Certified Pre-Owned)
- mileage_mi – mileage in miles
- price – car price
- state – car location
- model_year – year of manufacture
- manufacturer – car brand
- fuel_type – fuel type (Gasoline, Hybrid, Diesel, Flex Fuel)
- drivetrain – drivetrain type (All-wheel, Four-wheel, etc.)

- transmission – transmission type
- exterior_color / interior_color – car colors
- accidents_or_damage – accident history
- 1_owner_vehicle – ownership history

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9246 entries, 0 to 9245
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Car                   9071 non-null   object 
 1   Condition             9071 non-null   object 
 2   Mileage               9071 non-null   object 
 3   Price                 9071 non-null   object 
 4   Basics Info           9242 non-null   object 
 5   Vehicle History Info  9242 non-null   object 
 6   Vehicle Reviews Info  9242 non-null   object 
 7   Seller Rating         7716 non-null   float64
 8   Seller Rating Count   7716 non-null   object 
 9   Seller Address        8954 non-null   object 
dtypes: float64(1), object(9)
memory usage: 722.5+ KB
```

df.isnull().sum()

```
car                      175
condition                175
mileage                  175
price                    175
basics_info                4
vehicle_history_info       4
vehicle_reviews_info       4
seller_rating           1530
seller_rating_count     1530
seller_address           292
dtype: int64
```

The data cleaning process handled missing values, converted objective columns into numerical values where needed, creating new columns (feature engineering) and prepared the dataset for analysis, visualization ML modeling, and deployment.

5. Methodology

The project followed a structured approach:

Data Preprocessing & Feature Engineering:

- Handled missing values and removed duplicates.

- Converted mileage and price into numeric types.
- Creating new columns with valuable information
- Applied **One-Hot Encoding** for categorical features.
- Used **Robust Scaler** to normalize numerical columns.

Feature Selection:

- Employed **Wrapper Methods** and **Embedded Methods** to select the most influential features.

Machine Learning Models:

- Trained multiple regression models:
    - **XGB Regressor**
    - **Random Forest Regressor**
    - **Decision Tree Regressor**
    - **K-Nearest Neighbors (KNN) Regressor**
    - **Ridge Regressor**
- Created a **Voting Regressor** combining XGB and Ridge for improved predictions.

Evaluation Metrics:

- **R^2 Score**, **RMSE**, **MAE**, and cross-validation techniques were used to evaluate model performance.

Deployment:

- Developed a **Streamlit web application** for interactive analysis and ML-based car price prediction. Users can filter data, visualize market trends, and predict car prices based on selected features.

6. Business Questions and Insights

1. **Car Conditions:** New cars bring higher prices despite fewer units sold; used cars are more numerous but generate slightly lower revenue.
2. **Manufacturers:** Jeep, Ford, and Chevrolet dominate sales and revenue.
3. **Fuel Types:** Gasoline cars are most frequent and contribute the highest revenue; hybrid cars show growth potential.
4. **Exterior Colors:** Grey, black, and white are the most popular and generate the most revenue.
5. **Drivetrain Types:** All-wheel and four-wheel drive cars sell for higher prices and are more popular.
6. **Transmission:** Automatic cars dominate sales and revenue.
7. **Model Years:** Newer models, especially 2024, generate the highest sales revenue.

Business Recommendations:

- Focus on stocking more **new cars**.
- Collaborate with top manufacturers such as Jeep, Ford, and Chevrolet.
- Offer more **Gasoline and Hybrid vehicles**.
- Prioritize popular colors (**Grey, Black, White**).
- Highlight **all-wheel and four-wheel drive vehicles**.
- Ensure availability of **automatic transmission cars**.
- Keep **recent model years** in stock to meet customer demand.

7. Tools and Libraries

- **Python Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Plotly, Category Encoders
- **Machine Learning & Evaluation:** Scikit-Learn, XGBoost, imblearn, joblib
- **Web Deployment:** Streamlit

8. Expected Outcomes

- A robust machine learning model that predicts car prices accurately.
- Interactive visualizations of market trends and price influences.
- Actionable insights to guide inventory, marketing, and sales strategies.
- A user-friendly web application for real-time price prediction.