Predicting Car Prices Using Machine Learning

Done by: Asrar Sharaf

EPSILON AI ACADEMY

Final Report:

1. Title & Abstract

Title: Predicting Used Car Prices Using Machine Learning

Abstract:

This project focuses on predicting new and used car prices using machine learning models. A dataset of 5,235 cleaned car entries from Kaggle was analyzed to understand factors influencing car prices, such as condition, manufacturer, fuel type, color, drivetrain, transmission, and model year. Multiple regression models were applied, including XGB, Random Forest, Decision Tree, KNN, and Ridge, with a Voting Regressor combining Ridge and XGB achieving the best performance. The project also provides actionable business insights and an interactive Streamlit dashboard for data exploration and price prediction.

2. Introduction & Background

The automotive market has grown significantly, with used car sales becoming a major segment. Predicting accurate car prices helps sellers set competitive prices and buyers make informed decisions. Machine learning offers powerful methods to uncover patterns in historical data, analyze factors affecting car prices, and make precise predictions. This project aims to leverage ML models to predict car prices and provide insights for strategic business decisions.

3. Dataset Description & Cleaning

Initial Dataset:

- Source: Kaggle
- Entries: 9,246
- Columns: 10 (including Car, Condition, Mileage, Price, Basics Info, Vehicle History Info, Vehicle Reviews Info, Seller Rating, Seller Rating Count, Seller Address)
- Missing values: Between 0.04% and 16.5% per column

Data Cleaning:

- Removed duplicates and rows with excessive missing values
- Converted mileage and price to numeric types

- Extracted and created new features: model_year, manufacturer, fuel_type, drivetrain, transmission, exterior_color, interior_color, accidents_or_damage, 1_owner_vehicle, and state
- Final Dataset: 5,235 entries, 13 columns

---

4. Exploratory Data Analysis (EDA)

Car Conditions:

- Most vehicles sold are **Used (2,905)**, followed by **New (1,997)** and **Certified Pre-Owned (333)**.
- New cars generate the highest total sales value (~86.9 million), indicating higher prices despite lower numbers.

Manufacturers:

- Top by sales volume: **Jeep (542)**, **Ford (509)**, **Chevrolet (489)**
- Top by total price: **Jeep (21.3M)**, **Ford (16.6M)**, **Chevrolet (13.6M)**

Fuel Types:

- Gasoline dominates (4,755 cars), followed by Hybrid (273), Diesel (134), Flex Fuel (73)
- Total sales value: Gasoline (153M), Hybrid (11.3M), Diesel (7.2M), Flex Fuel (1.9M)

Exterior Colors:

- Most popular: **Grey (2,030)**, **Black (1,253)**, **White (1,101)**
- Highest total price: Grey (65.9M), Black (43M), White (38.8M)

Drivetrain:

- All-wheel drive (2,273) most common, followed by Front-wheel (1,308), Four-wheel (1,170), Rear-wheel (484)
- Total price: All-wheel (78.7M), Four-wheel (48.9M), Front-wheel (29M), Rear-wheel (16.8M)

Transmission:

- Fully Automatic dominates (4,807), followed by manual, CVT, Semi-Automatic
- Total sales value: Fully Automatic (160M), Manual (~5M), others lower

Model Years:

- Most cars are from 2024 (1,913), 2021 (512), 2023 (399), 2020 (382)
- Total sales value: 2024 (82.6M), 2023 (16.8M), 2021 (16.1M)

5. Feature Engineering & Transformation

- Categorical features were transformed using **One-Hot Encoding**
- Numerical features scaled with **RobustScaler**
- Feature selection applied: **Wrapper Method** and **Embedded Method**
- Key features: condition, mileage_mi, model_year, manufacturer, fuel_type, drivetrain, transmission, exterior_color, interior_color, accidents_or_damage, 1_owner_vehicle

6. Machine Learning Models

Models Used:

- **XGB Regressor**
- **Random Forest Regressor**
- **Decision Tree Regressor**
- **K-Nearest Neighbors (KNN) Regressor**
- **Ridge Regressor**
- **Voting Regressor** (combination of XGB + Ridge, best performing)

Evaluation Metrics:

- R^2 Score, RMSE, MAE, Cross-Validation

Best Model:

- **Voting Regressor (XGB + Ridge)** achieved the highest R^2 score on test data with the lowest RMSE and MAE, indicating strong generalization and accurate predictions.

7. Insights & Business Recommendations

Insights:

1. New cars bring higher revenue despite fewer units.
2. Jeep, Ford, and Chevrolet dominate both sales and revenue.
3. Gasoline cars dominate the market; Hybrid cars are growing.
4. Popular colors (Grey, Black, White) correspond to higher total prices.
5. All-wheel and Four-wheel drive cars generate more revenue.
6. Fully Automatic transmissions are preferred and more profitable.
7. Newer models, especially 2024, bring the highest revenue.

Recommendations:

1. Focus on stocking **new cars** and promoting them with deals/trade-ins.
2. Collaborate with **top manufacturers** for joint promotions.
3. Expand offerings of **Gasoline and Hybrid cars**.
4. Prioritize **popular colors** in inventory.
5. Highlight **all-wheel and four-wheel drive vehicles** in marketing.
6. Ensure availability of **automatic transmission cars**.
7. Keep **recent model years** in stock to attract buyers.

---

8. Deployment

A **Streamlit web application** was developed to allow:

- Interactive filtering of car data by condition, fuel type, transmission, drivetrain, color, and manufacturer.
- Visualization of market trends and total car prices by key features.
- Real-time **price prediction** using the trained Voting Regressor model.

This provides a user-friendly interface for both analysis and price estimation.

---

9. Conclusion

This project successfully demonstrates how machine learning can predict car prices and provide actionable business insights. By leveraging historical data, EDA, and multiple ML models, the project delivers both accurate predictions and strategic recommendations. The Streamlit deployment makes the analysis and prediction accessible and practical for real-world applications.