

End-to-End Car Price Prediction – Project Report

Author: Abdelrahman Solii

Date: 2025-12-07

Project: End-to-End Car Price Prediction

Abstract

This project aims to predict the resale price of cars using machine learning techniques. Using a dataset containing car features such as manufacturer, model, mileage, engine size, and age, we develop a model that accurately estimates car prices. The system also includes a web interface for real-time price prediction.

1. Introduction

The used car market often suffers from inconsistent pricing due to subjective appraisals and lack of transparency. This project seeks to automate car price estimation, providing accurate and reliable predictions for buyers and sellers.

Objectives:

- Analyze car datasets to understand features impacting price.
 - Build and train machine learning models to predict car prices.
 - Deploy the trained model through a user-friendly web interface.
-

2. Dataset Overview

The dataset contains **[19237_ROWS]** samples and **[18_FEATURES]** features. Key columns include:

Feature	Description
manufacturer	Car brand (e.g., Toyota, Hyundai)
model	Car model name
category	Vehicle type (e.g., SUV, sedan)
leather_interior	Boolean: leather interior presence
fuel_type	Type of fuel (petrol, diesel, electric)

Feature	Description
mileage	Distance driven in km
gear_box_type	Manual / Automatic
drive_wheels	FWD / AWD / RWD
wheel	Number of wheels
color	Exterior color
price	Selling price (target variable)
levy	Tax/levy amount
engine_volume	Engine size in liters
cylinders	Number of cylinders
airbags	Number of airbags
car_age	Age in years
price_per_mileage	Derived feature: price/mileage
engine_power	Engine horsepower
brand_mean_price	Average price per manufacturer
model_mean_price	Average price per model

Dataset Source: Collected from online car listings and historical sales data.

3. Data Preprocessing

Data preprocessing steps include:

1. **Missing Values:** Checked and handled using mean/mode imputation.
2. **Duplicate Records:** Removed duplicates to avoid bias.
3. **Feature Engineering:**
 - o car_age calculated from manufacturing year.
 - o price_per_mileage derived for better modeling.

4. Encoding Categorical Variables:

- Used LabelEncoder for manufacturer, model, fuel_type, gear_box_type, etc.

5. Train/Test Split:

- 80% training, 20% testing set.
-

4. Feature Engineering & (EDA)

- **The Feature engineering added a lot of value to the training by creating new important features that would help the model predict the price.**
- **The Evaluation went from(77%) to (99%).**

Key observations from data visualization:

- **Price Distribution:** Average car price: **18559\$**, median: **13172\$**
 - **Manufacturer Impact:** Brands like BMW, Mercedes have higher mean prices.
 - **Mileage vs Price:** Negative correlation; higher mileage generally leads to lower price.
 - **Engine Volume & Power:** Positive correlation with price.
 - **Fuel Type & Category:** SUVs and diesel cars tend to have higher prices.
-

5. Model Selection & Training

Several regression models were tested but the best was:

Model	R² Score	RMSE
CatBoost Regressor	0.9998	166.75
XGBoost Regressor	0.9982	469.53

Selected Model: CatBoost Regressor (best combination of accuracy and stability).

Training Details:

- Used default parameters with early stopping for CatBoost.
-

6. Model Evaluation

Evaluation on test set:

- **R² Score:** 0.9998
- **RMSE:** 166.75
- **RMSE :** 941.575

Feature Importance:

- Top features influencing price: mileage, engine_volume, car_age, manufacturer_mean_price, model_mean_price.

(Insert bar plot for feature importance.)

7. Deployment

The trained model is deployed using **streamlit** and **Gunicorn**.

Instructions:

1. Install dependencies: pip install -r requirements.txt
2. Run the app locally: python car_pred_price.py or via Gunicorn:
 1. gunicorn car_pred_price:app --bind 0.0.0.0:5000

Deployment Files:

- car_pred_price.py: stramlit application
 - cars_prediction.sav: Trained model
 - Procfile: For PaaS deployment
 - requirements.txt: Dependencies
-

8. Conclusion & Future Work

Conclusion:

- The CatBoost model provides highly accurate car price predictions.

- Key factors affecting car price include mileage, engine volume, car age, and manufacturer/model averages.

Future Work:

- Incorporate more features (e.g., location, demand trends, accident history).
 - Deploy as a full web platform with user accounts and batch predictions.
 - Explore ensemble models for improved accuracy.
-

9. References

- [Scikit-learn Documentation](#)

[CatBoost Documentation](#)

[Used Car Price Prediction Examples](#)
