Group 4 Presentations
Predicting Used Car Prices Using Machine Learning

Yate Zhang, Yuchen Bi, Chris Wang, Jiancong Zhu

Agenda

- Executive Summary
- Problem Statement and Research Objects
- Data Collection & Preprocessing
- EDA
- Feature Engineering
- Modole Development & Evaluation
- Findings
- Q & A

Executive Summary

- The automotive resale market faces challenges in accurately pricing used cars.
- This project develops a machine learning model to predict used car prices based on Craigslist data.

Problem Statement and Research Objects

Problem: Accurate used car price prediction is crucial for fair market valuation because traditional methods fail to capture multiple influencing factors.

Objects: Develop a robust predictive model using machine learning.

- Identify key determinants (mileage, brand reputation, model year, fuel type, etc.)

- Engineer meaningful features to enhance model interpretability.

- Compare multiple ML models and select the most effective one.

- Provide scalable pricing recommendations for consumers and dealerships.

Data Collection & Preprocessing

•Dataset sourced come from Kaggle: Craigslist

•- 426,880 rows, 26 columns.

•- Key attributes: Price, year, manufacturer, model, odometer, fuel type, transmission.

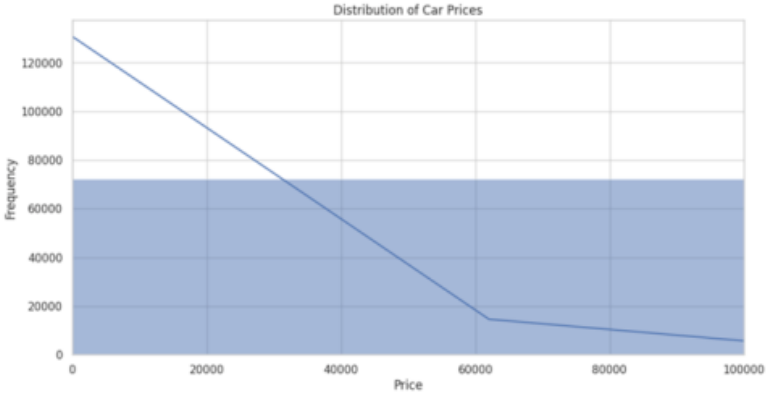
•- Removed columns with higher than 50% missing values.

•- Imputed missing values (median for numerics, mode for categoricals).

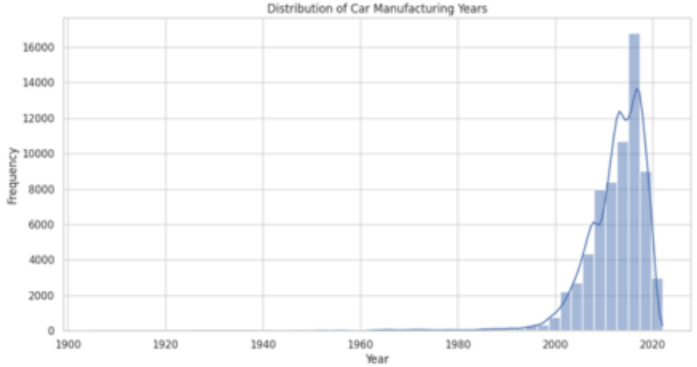
•- Removed unrealistic price values ($0 or >$100,000) and duplicate records.

EDA

- Car prices were highly skewed, with many listings having extremely low or unrealistically high values. To improve the reliability of the analysis, listings with prices below $100 and above $100,000 were removed to filter out unrealistic values and extreme outliers.



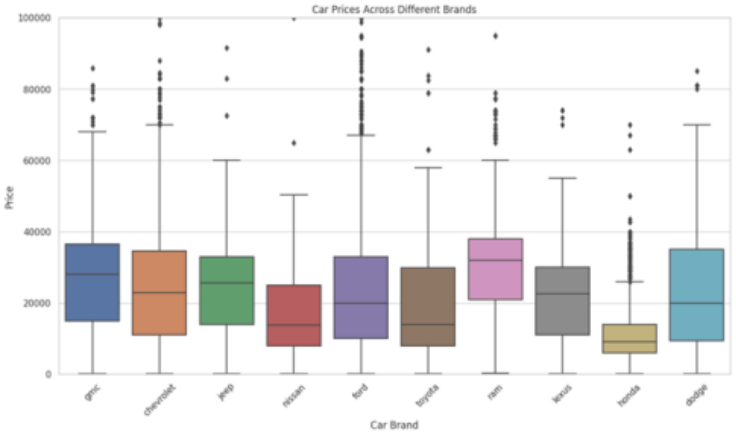
- This plot reveal that most listed vehicles were produced between 1995 and 2020, with a noticeable concentration in more recent years. This pattern aligns with market trends, where newer vehicles are more frequently listed for resale.



- Newer vehicles tended to have higher prices, while older vehicles showed a downward trend in value due to depreciation.

Similarly, cars with higher mileage generally had lower prices, confirming the expected negative correlation between mileage and resale value. However, a few anomalies were present, where some older vehicles maintained high prices, likely due to classic or luxury car categories.

- This variation suggests that luxury brands hold their value better or are priced at a premium due to perceived quality and demand. Furthermore, categorical variables such as fuel type and transmission type were explored using count plots. The majority of the listings were gasoline-powered vehicles, while electric and hybrid cars made up a smaller proportion of the dataset. Similarly, automatic transmission cars were more common than manual ones, reflecting market preferences.



Feature Engineering

•- Created new features:

• - Car age (2024 - year of manufacture).

• - Mileage per year (odometer / car age).

•- Encoded categorical variables using Label Encoding.

•- Standardized numerical variables (odometer, mileage per year) using StandardScaler.

•- Final dataset cleaned and stored for model training.

Model Development & Evaluation

Findings

Thank You and Q & A