ReDI School of
Digital Integration

Predicting Fair Market Value for Used Cars

🚗💰

Tochukwu Ibechem

We use tech to connect human potential and opportunity with dignity & humility

Introduction

Problem Statement:

- Car buyers/sellers seek a reliable way to estimate prices of used cars.
- Used car pricing is complex due to varying depreciation, mileage, and regional factors.
- Goal: Build a predictive model to estimate used car prices using historical Craigslist listings.

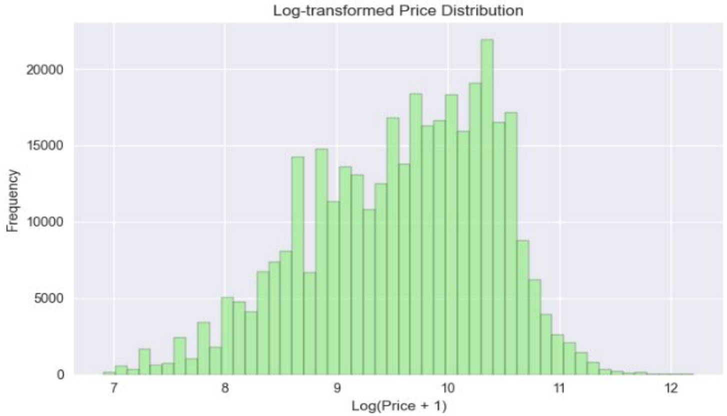
Dataset Used:

- Source: Kaggle - Craigslist Cars & Trucks
- Records: ~426,000 used car listings
- Features: 26 columns incl. make, model, year, odometer, fuel type, etc.

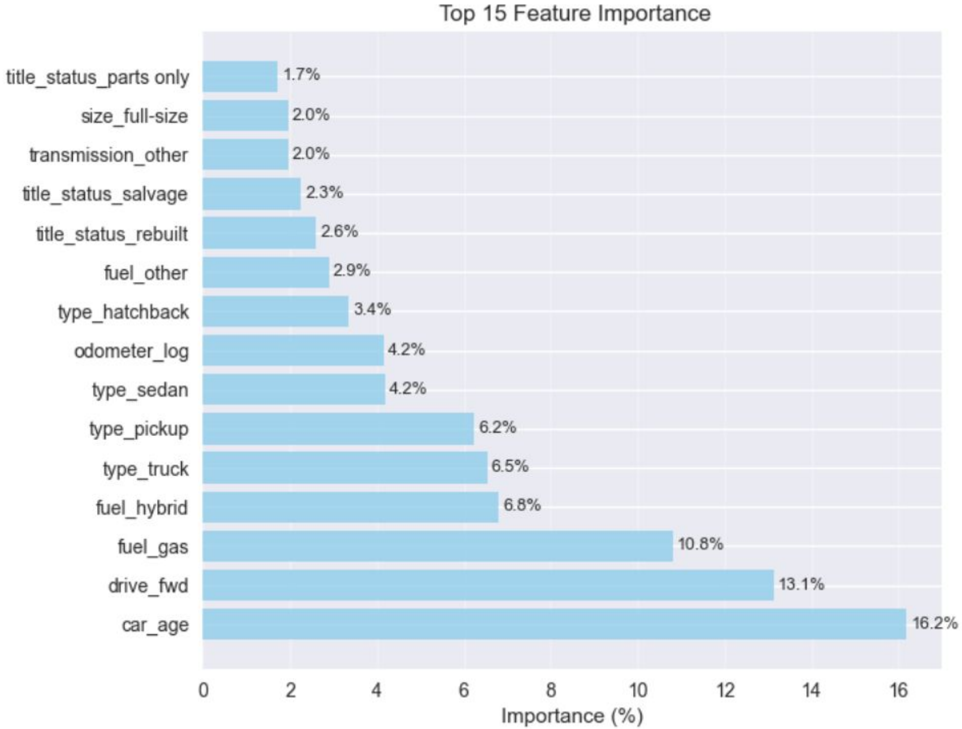
Approach:

- Use XGBoost Regressor (with log-transformed prices and odometer values) to predict car prices

Variables Analysis



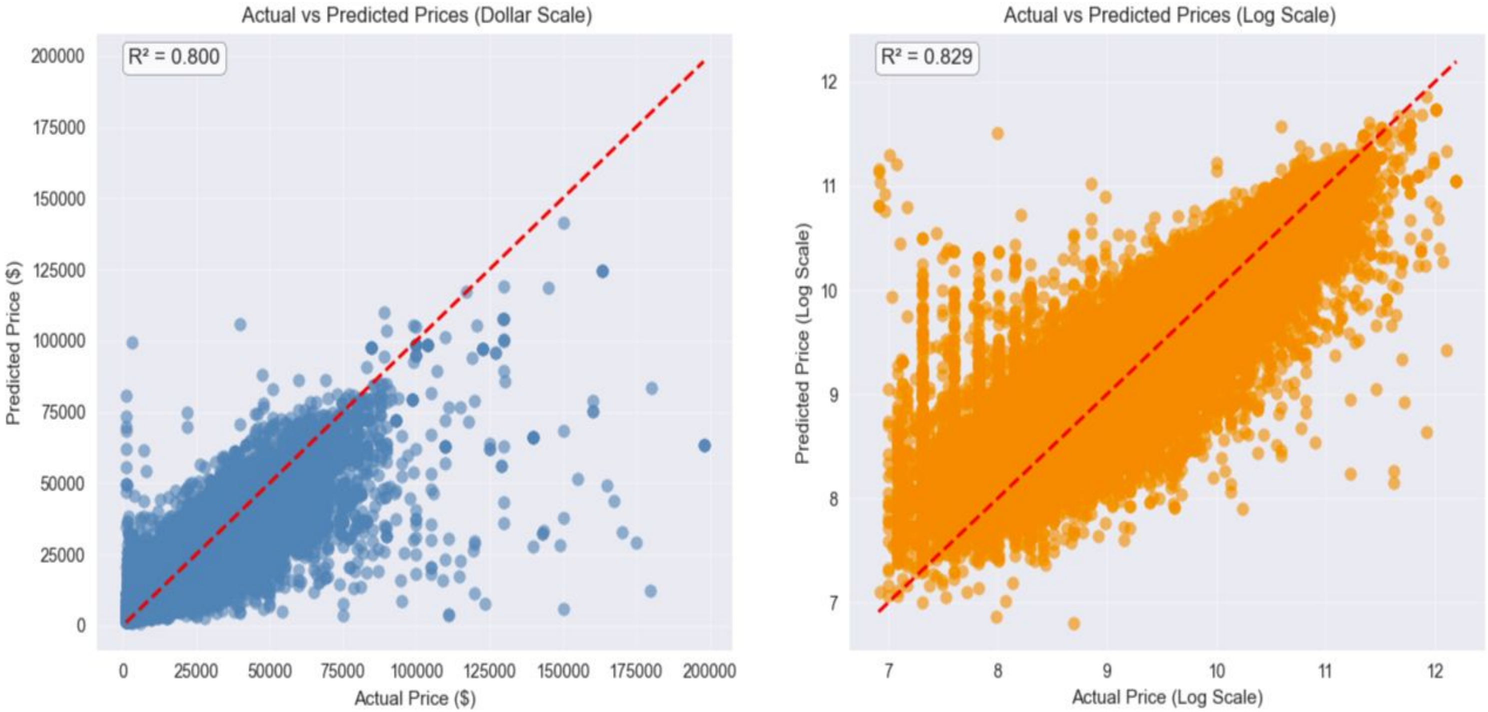
Log transformation is essential for modeling this skewed price data effectively.



The top 15 features contribute ~84.9% of total importance, showing that the model's decisions are concentrated in a few strong signals.

Vehicle age, drivetrain, fuel type, and body type are the strongest predictors of price.

Prediction Analysis



- Strong performance overall: High R^2, generalizes well, no major overfitting.
- Overpredicts slightly on average (~7%).
- Mean Relative Error: +6.6% (train), +7.7% (test) → Slight upward bias — model tends to overpredict.
- Predicts directionally well, but not precise for rare/exotic listings.

Insights

Challenges

- Price ranges are very broad ($0 to nearly $200K), which makes prediction inherently hard.
- Large memory requirement for Categorical features encoding (therefore used features with considerable unique values).
- Choice of effective new features

Learnings

- Data Quality is Critical
- Tree-based models (XGBoost) outperformed linear regression.
- Domain knowledge (e.g., "100K+ mileage matters")
- Feature engineering drives performance

What can be improved

- Improved feature engineering.
- Add Interaction Features: Interaction between age × mileage, fuel × type, or transmission × drive may reveal nonlinear effects.
- Build separate models for economy vs. luxury segments
- Data Quality Improvements: example: Better mileage normalization (e.g., age-adjusted mileage).

Thank you!