Business Analytics and Data Science

ASSIGNMENT 2

LEONARDO GONNELLI

Data Preprocessing & Feature Engineering

Feature Engineering

Derived features : `car_age`, `mileage_per_year`, `engine_efficiency`, `power_index`

Interaction features : `brand_model` (brand + model combination)

Tax-based features : 5 features capturing tax relationships (e.g. `tax_mileage` and `tax_per_mpg`)

Polynomial features : 8 polynomial terms (squared terms and interactions)

Encoding Strategy

Target encoding : applied to all categorical variables (including `brand_model`) with smoothing = 10 to reduce overfitting

One-hot encoding (for XGBoost) : replace categorical variables with binary columns

Frequency encoding : model and brand_model frequency added as additional features

Data Cleaning and Target Transformation

Dropped columns : `ID` (non-informative), year (we used `car_age` instead)

Outlier handling : IQR-based clipping to manage extreme values

Target transformation : price distribution being skewed, we applied `log1p(price)`, which improves models performance,

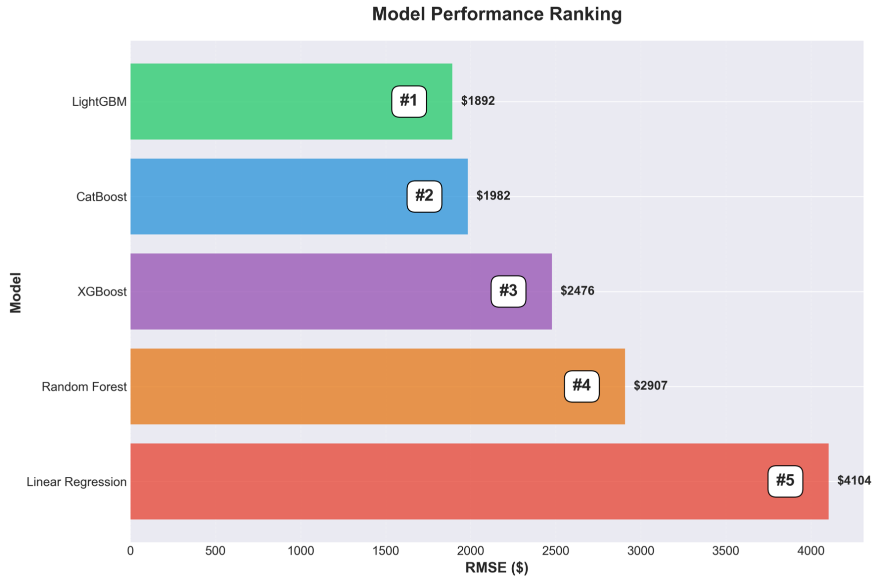
Model Comparison

Key Findings

LightGBM outperforms all other tested models, achieving a significant improvement over the baseline Linear Regression model (54% improvement).

Gradient boosting models outperform the tree-based model (random forest) and linear model.

LightGBM and CatBoost models outperform XGBoost as they handle categorical variables natively.



Conclusion and Further Improvements

Best model – LightGBM

Final test rmse (on the available half of the test set on Kaggle) : 1892.37

Achieves the best performance and good generalization.

Key success factors

Feature engineering, log transformation and target encoding significantly improved models performance (e.g. LightGBM's rmse : 1953 before target encoding -> 1892 after — ↓61).

Further Improvements Ideas

Advanced feature engineering (feature aggregation, combination target encoding...), experiment ensemble methods (model blending, stacking).