The Erdős Institute

Car sales price prediction

Mariana Khachatryan, Amogh Parab, Nasim Dehghan Hardoroudi, Adreja Mondol

Outline

- ❑ Motivation
- ❑ Modeling framework
- ❑ Results
- ❑ Conclusions

Motivation

- ❑ Buying and selling cars is common experience especially among people leaving in rural areas with little or no transportation

- ❑ Key Stakeholders

  - Individuals selling cars and car dealerships need price prediction model to set competitive and accurate prices for cars.

  - Dealerships want to maximize profit while ensuring quick car sales. Accurate price prediction results in competitive pricing and profitability.

  - Customers can use the model to estimate whether the set price is fare.

Modeling framework

Use car sales data from CarDekho online marketplace

Model the relationship between car sales price and different car features

Data processing involved:

- Data cleaning
- Feature engineering
- One hot encoding of categorical variables
- Removal of highly correlated features

Final data set:
6533 data points with 12 features

Split 80:20 and scale using Standard scaling

Test set

Training set

Baseline model

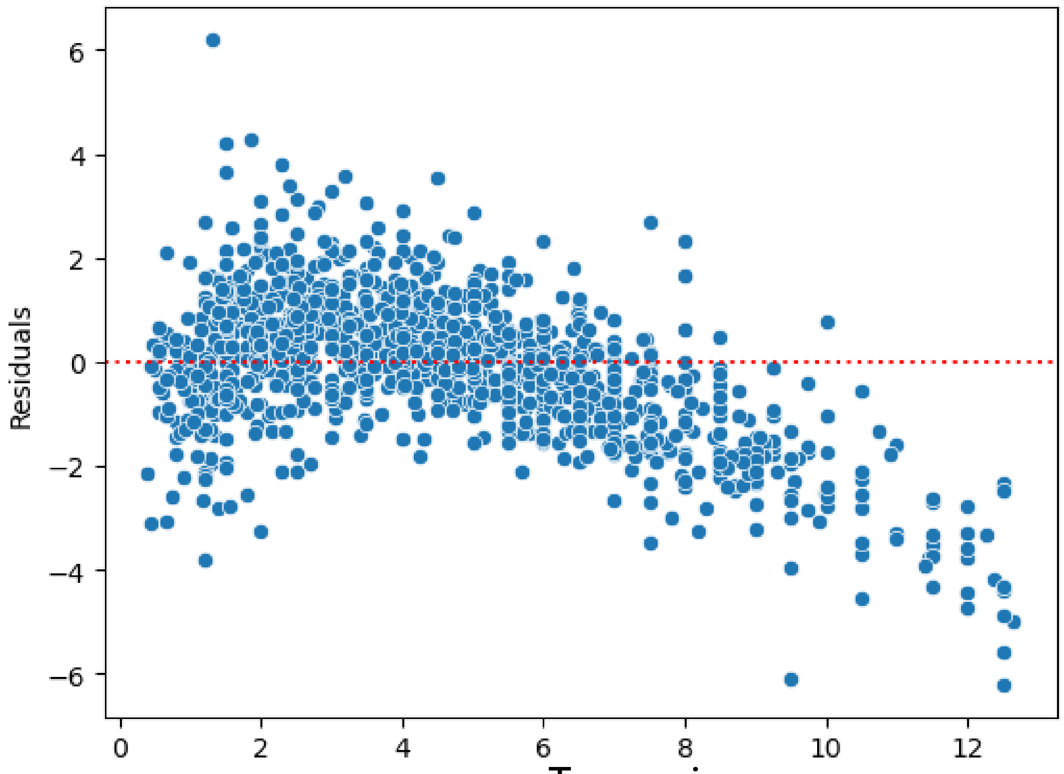
- Linear regression

Regression models with parameters from cross validation

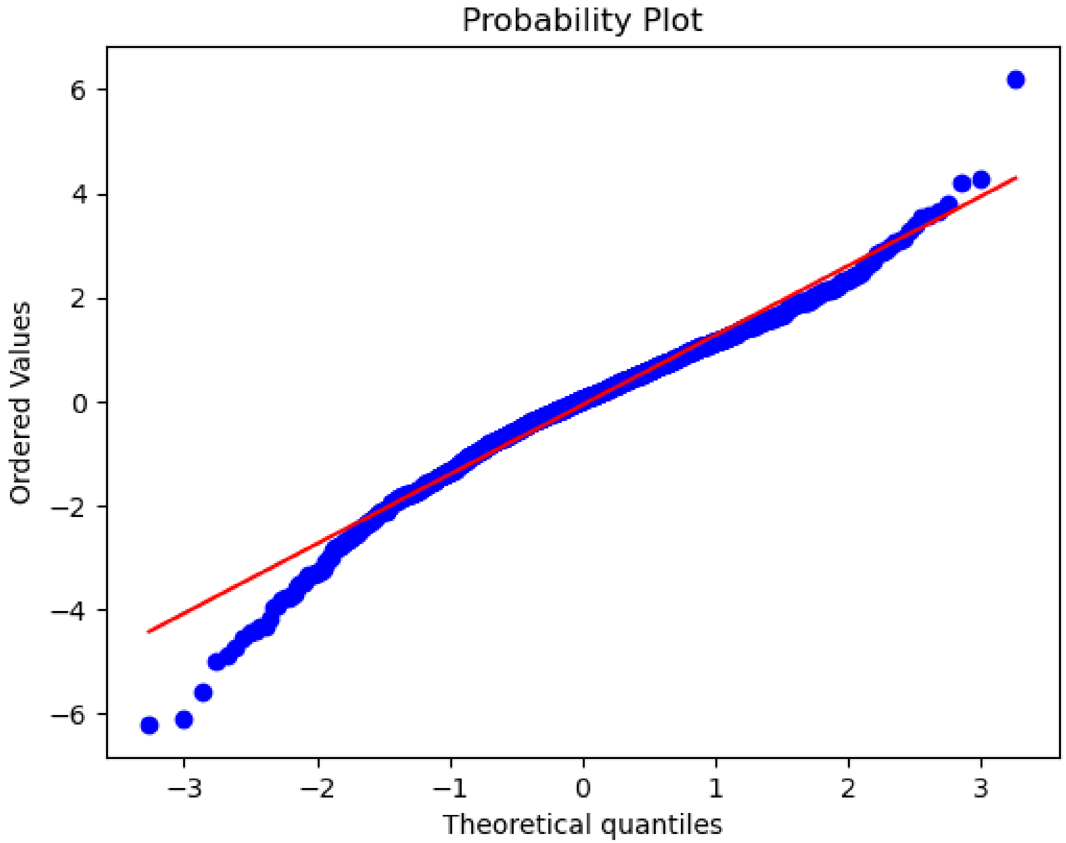
- Polynomial regression
- k-nearest neighbors
- Support Vector Machines
- Tree methods (best performance)

Results: Linear Regression (Base model)

- ❑ Root Mean Squared Error (RMSE) of 1.35 and R2=0.73.
- ❑ Calculate residuals (difference between predicted label values and true values) and check Linear Regression assumptions



The assumption of homoscedasticity is violated.



Normality is violated for lower for residual values below -3

Should consider other non-linear models.

Results from non-linear models

Used Grid Search Cross-Validation to tune model parameters.
Overall best model performance was obtained with XGBoost.
XGBoost outperforms SVMs and kNN because it is inherently nonlinear and is less sensitive to hyperparameter tuning.
XGBoost improves performance by combining multiple trees, which enhances it's ability to model complex patterns.
It also reduces overfitting by combining multiple trees and employing shrinkage/regularization.

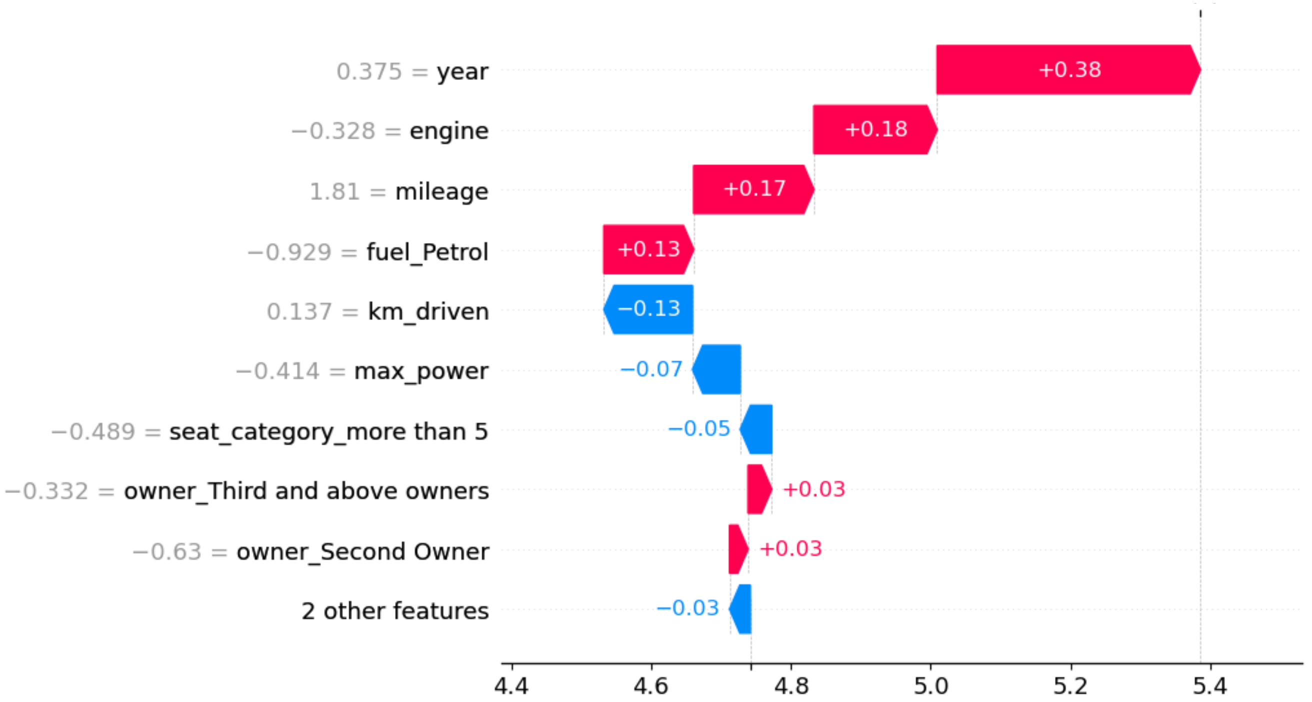
Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)	R^2
Linear Regression (Baseline)	1.02	1.35	33 %	0.73
2nd order Polynomial Regression	0.81	1.12	23%	0.82
K-Nearest Neighbours	0.78	1.13	22%	0.81
Support Vector Regressor	0.76	1.09	20%	0.82
XGBoost	0.60	0.87	16%	0.89

Results: SHapley Additive exPlanations (SHAP values) for describing feature importances

SHAP shows the contribution of each feature on the prediction of the model.

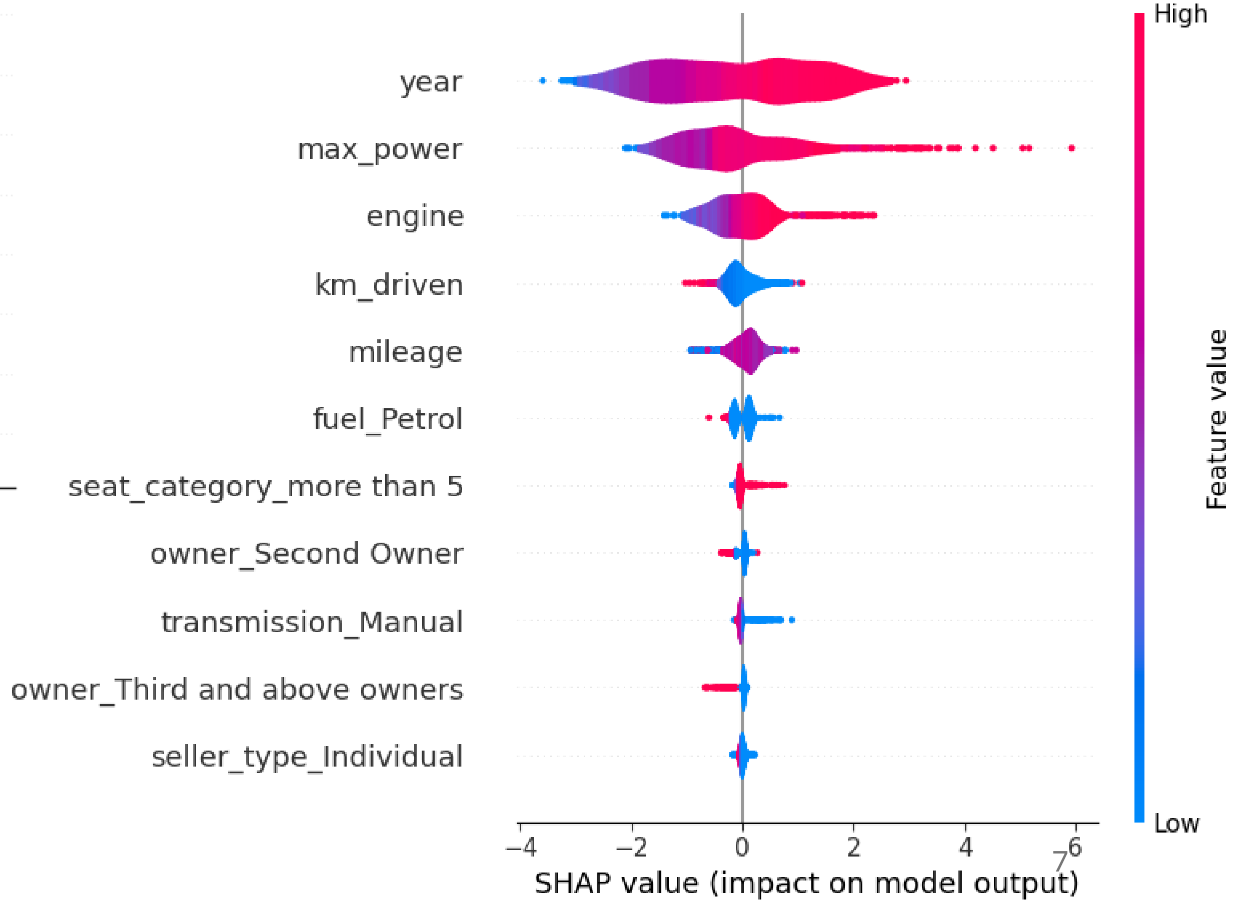
SHAP values for one single observation x are given by the length of the bar

Predicted value=5.386



The mean of all predictions=4.742

The global effect of the features on model prediction



Conclusions

- ❑ Base model has a poor performance as Linear Regression assumptions are violated

- ❑ Overall best model performance was obtained wit XGBoost with MAPE of 16% and R^2=0.89

- ❑ The four features that have the most influence on the predicted price are

    - year,
    - max power (measurement of the engine's power that accounts for frictional losses in the engine),
    - engine (the amount of air and fuel that can be pushed through the cylinders in the engine),
    - km driven