Modeling the relationship between car sales price and different car features

Buying and selling cars is common experience especially among people leaving in rural areas with little or no transportation. It is thus interesting to study what factors do influence car sales price significantly and what can be improved to have better sales and fair car sales prices. Using different Machine Learning Regression methods we will develop models to predict car sales price using CarDekho Dataset .

Authors

- Mariana Khachatryan
- Adreja Mondol
- Amogh Parab
- Nasim Dehghan

KPI

Technical performance Key Performance Indicators (KPIs)

1. Mean Absolute Error (MAE)
2. Root Mean Squared Error (RMSE)
3. R-squared- proportion of variance in target variable that is predictable from input features

Business impact KPIs

These KPIs measure how effectively the model adds value to the business, such as improving pricing strategies or increasing customer satisfaction.

1. Revenue Increase: Accurate price predictions could help optimize the selling price of cars, leading to better margins.
2. Cost Reduction: If the model is used to automate pricing, it can reduce the need for manual evaluation and pricing, saving labor costs.
3. Inventory Turnover Rate: How quickly cars are being sold after their prices are set by the model.

Key Stakeholders

1. 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.
2. Customers can use the model to estimate whether the set price is fare.

Modeling approach

We compare predictions of different Machine Learning models:

- Linear Regression
- Tree Methods
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)

We start with linear regression model as a baseline model. We also use Elastic Net linear model from python sklearn library. Elastic Net uses combination of Lasso and Ridge regularizations. We then compare results from different regression models. We use grid search to find optimal model parameters for different regression methods. We want to learn:

- Which features have the most influence on sales price prediction?
- Which regression model will provide the best performance?
- What is the prediction error? The model framework is displayed in the following diagram:

As we are trying different regression models including one's that use distance metric, we use one-hot encoding for categorical features and

Results

Overall best model performance was obtained with Extreme Gradient Boosting (XGBoost)**.** XGBoost with RMSE of 0.87 and $R^2 = 0.89$. The performance of different models is summarized in the following table.

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

XGBoost outperforms SVMs and KNN because it is inherently nonlinear and robust to scale 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.

Conclusions

We have used SHapley Additive exPlanations (SHAP values) for describing importances of each of the features on the prediction of the model. We have learned that the four features that have the most influence on the predicted price are year, max power, engine and km driven. Here engine refers to the amount of air and fuel that can be pushed through the cylinders in the engine. Max power is a measurement of the engine's power that accounts for frictional losses in the engine