CognoRise INFOTECH

Car Price Prediction

Table Of Content

1. Objective
2. Dataset Overview
3. Data Preprocessing
4. Modelling
5. Evaluation
6. Conclusion

Objective

The objective of this project is to build a machine learning model that accurately predicts car prices based on key features such as engine size, horsepower, mileage, fuel type and others. The goal is to provide a reliable tool for buyers and sellers to make informed decisions, uncover the most influential factors affecting car pricing, and deliver a scalable solution for the automotive market

Dataset Overview

Dataset Source : https://www.kaggle.com/datasets/goyalshalini93/car-data

Size: 205 rows, 26 features.

Dataset Dictionary

- **Car_ID**: Unique ID for each observation (Integer).
- **Symboling**: Insurance risk rating; +3 (risky) to -3 (safe) (Categorical).
- **Car Name**: Name of the car (Categorical).
- **Fuel Type**: Type of fuel used (gas/diesel) (Categorical).
- **Aspiration**: Type of aspiration in the car (Categorical).
- **Door Number**: Number of doors in the car (Categorical).
- **Car Body**: Type of car body (Categorical).
- **Drive Wheel**: Type of drive wheel (Categorical).
- **Engine Location**: Location of the car engine (Categorical).
- **Wheelbase**: Distance between front and rear wheels (Numeric).
- **Car Length**: Length of the car (Numeric).
- **Car Width**: Width of the car (Numeric).
- **Car Height**: Height of the car (Numeric).
- **Curb Weight**: Weight of the car without passengers/baggage (Numeric).
- **Engine Type**: Type of engine (Categorical).
- **Cylinder Number**: Number of cylinders in the car (Categorical).
- **Engine Size**: Size of the engine (Numeric).
- **Fuel System**: Type of fuel system (Categorical).
- **Bore Ratio**: Bore-to-stroke ratio (Numeric).
- **Stroke**: Stroke volume inside the engine (Numeric).
- **Compression Ratio**: Compression ratio of the engine (Numeric).
- **Horsepower**: Engine power in horsepower (Numeric).
- **Peak RPM**: Maximum revolutions per minute (Numeric).
- **City MPG**: Mileage in the city (Numeric).
- **Highway MPG**: Mileage on the highway (Numeric).
- **Price**: Dependent variable, price of the car (Numeric).

Data Preprocessing

Data Cleaning

Missing Values

Missing data was identified and addressed by filling with median values to ensure that no null values would disrupt the modeling process.

Duplicated Data

Duplicated rows in the dataset were checked and removed to avoid bias and redundancy during analysis.

Feature Transformation

New features were engineered to enhance the dataset. For example company name is obtained from car name

Feature Scalling

StandartScaller

Scaling was performed on numeric variables (e.g., Horsepower, Curb Weight, Wheelbase) to ensure all features have a comparable range, which is critical for machine learning models sensitive to the magnitude of input data (e.g., gradient-based methods).

Data Splitting

Split Data

Data splitting refers to dividing the dataset into training and testing sets. Typically, an 80:20 split was used to ensure the model learns from the training set and is validated on unseen test data.

Modelling

Baseline Model

A simple model used to set a performance benchmark. This model helps evaluate the improvement brought by more complex algorithms.

Algorithms Used:

- Linear Regression
- Decision Tree

Ensemble Model

Combines predictions from multiple models to improve performance and robustness.

Algorithms Used:

- Random Forest
- Gradient Boosting (XGBoost)

Benefits:

- Handles non-linear relationships better.
- Reduces overfitting.

Key Metrics:

- R^2 Score
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

Evaluation

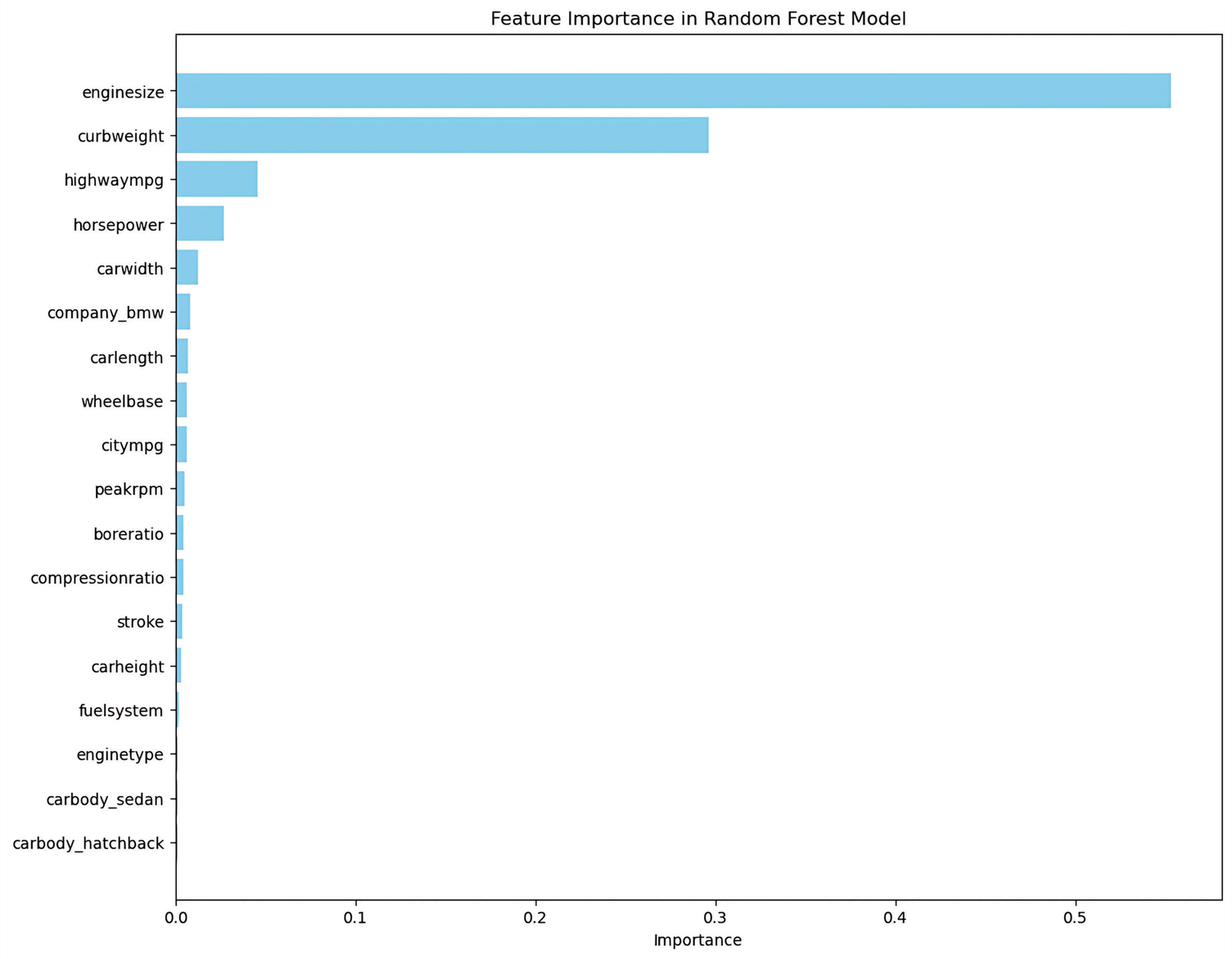
Model Performance Comparison Table

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R^2)	Root Mean Squared Error (RMSE)
Random Forest	3,275,513.57	1,284.36	0.9586	1,888.46
Gradient Boosting	5,688,260.73	1,684.46	0.9279	2,385.01
Linear Regression	8,635,315.60	1,924.64	0.8906	2,938.59
Decision Tree	8,920,432.77	1,955.96	0.8870	2,986.71

Key Insight :

- Best Performance: The Random Forest model has the lowest MSE and RMSE, indicating it is the most accurate model for predicting car prices.
- Comparison of Ensemble Models: Both ensemble models (Random Forest and Gradient Boosting) outperform the baseline models (Linear Regression and Decision Tree) in all metrics.
- Baseline Model Limitations: The baseline models show significantly higher error metrics, suggesting they are less effective in capturing the underlying patterns in the data.

Feature Important



Key Insight :

- Engine Size is the most significant factor affecting car price predictions, indicating that engine capacity plays a primary role in determining car value.
- Curb Weight is the second most important feature, suggesting that the vehicle's weight (likely linked to material quality and performance) is crucial.
- Highway MPG and Horsepower also stand out as significant contributors, highlighting that fuel efficiency and vehicle power are key determinants of car pricing.

Conlusion

- **Model Accuracy:** The Random Forest model demonstrated superior performance with an R^2 score of 0.9586, making it a highly accurate tool for predicting car prices.
- **Key Influential Factors:** Engine size and curb weight emerged as the most important predictors, followed by highway MPG and horsepower, highlighting the critical role of performance and efficiency in determining car prices.
- **Scalability and Practicality:** The model is scalable and adaptable to various automotive markets, providing a reliable framework for similar predictions. Its insights can guide manufacturers, buyers, and sellers in making informed decisions.

CognoRise
INFOTECH

Thank you