Prediction of Used Car Prices using Regression Models

Data science approach

[Abdulwahid] - [30/10/2022]

Agenda

1. Problem Definition
2. CRISP Methodology
3. Findings and Insights
4. Recommendations

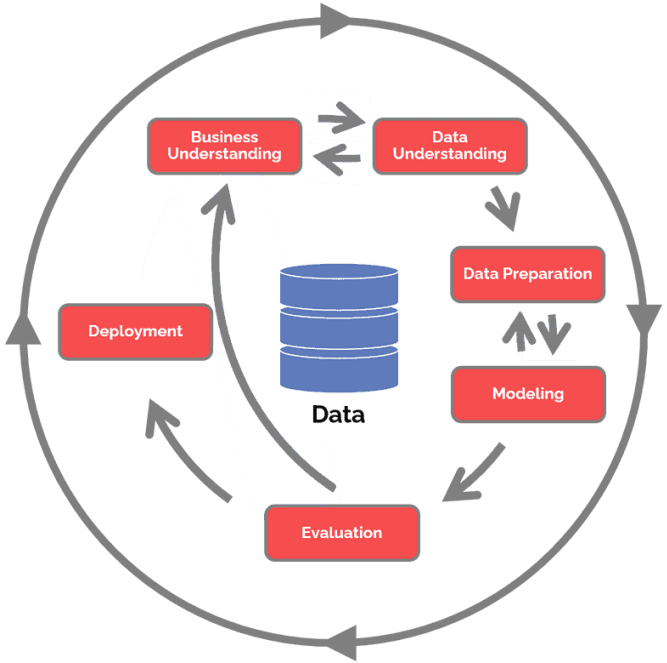
Problem Definition

Determining the worthiness of cars based on their characteristics

- What features impact the price?
- What are the best prices for used cars?

CRISP Methodology

The main stages for a successful data science project



(Data Science Process Alliance, 2022)

Business Understanding

Business Tasks

- Improving a company's advertising approach (used car selling company)

- Predicting the best prices to sell their recent used cars

Data Understanding

- Data was obtained from GitHub as a part of a data science challenge
- The dataset has 16 features with target variable price

- Conducted a descriptive and exploratory data analysis
- To find trends and insights
- To determine the best price predictors

Data Preparation

Data cleaning

- ✓ Handling missing values for training and testing set
- ✓ Checking feature data types
- ✓ Checking for inconsistent data

Data Transformation

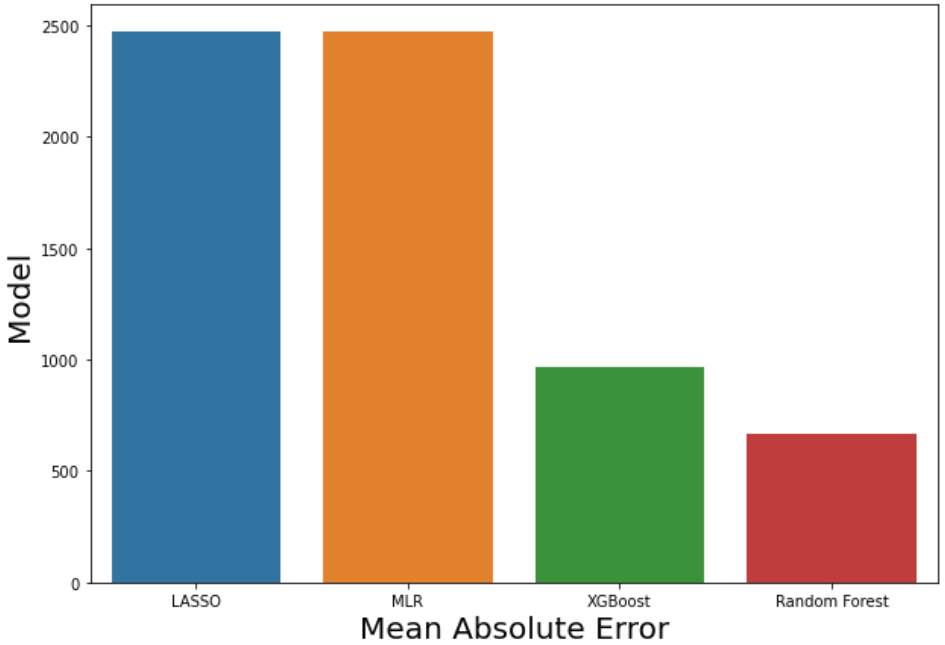
- ✓ Label encoding (from categorical to numerical)
- ✓ Feature scaling (range of 0 to 1)
- ✓ Dropping unwanted features

Model Development

- Training and testing are provided
- Split data features from target variable
- Models are built by fitting training set
- Model testing using testing set

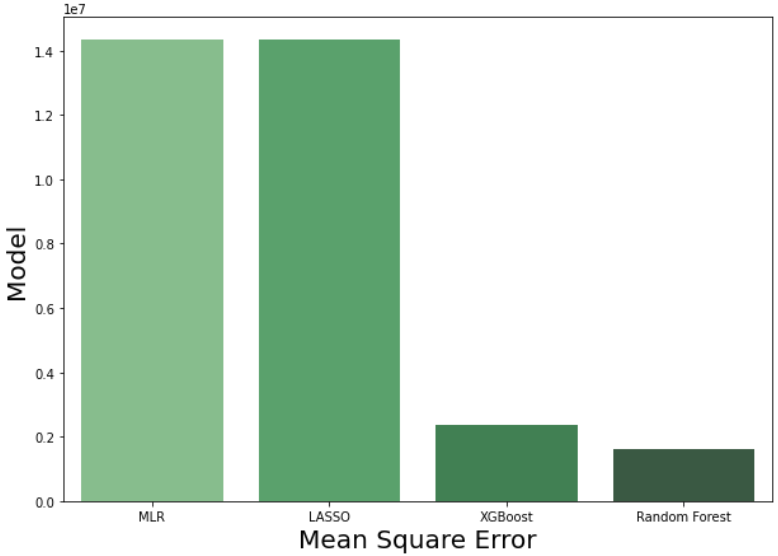
Model Evaluation

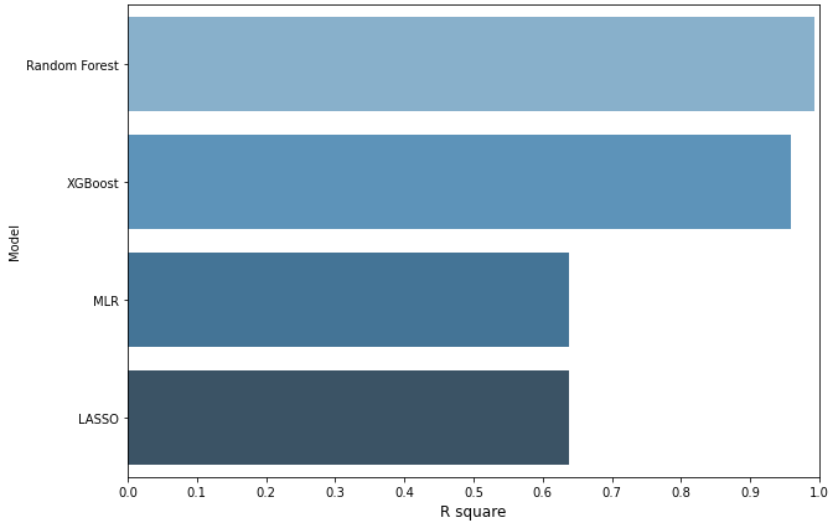
Model evaluation matrix (Mean Absolute Error)



Model Evaluation

Model evaluation matrix (Mean Square Error, R square)

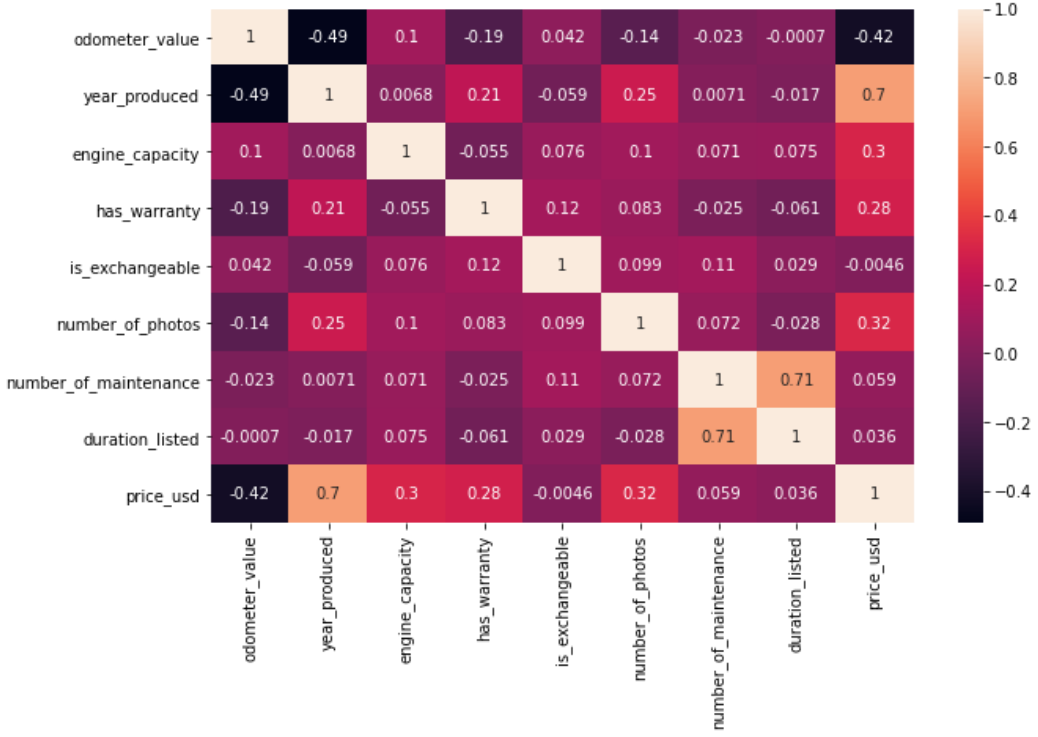




Deployment (Insights and Findings)

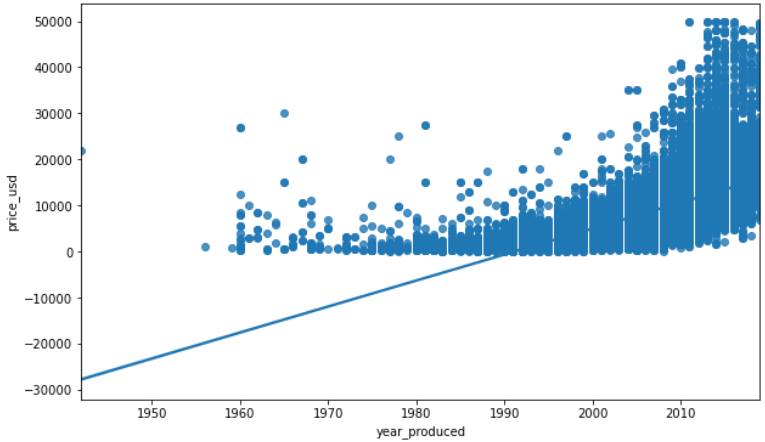
Pearson Correlation (Heatmap)

- Year produced has the highest correlation with price
- Exchangeable has the lowest correlation to the price

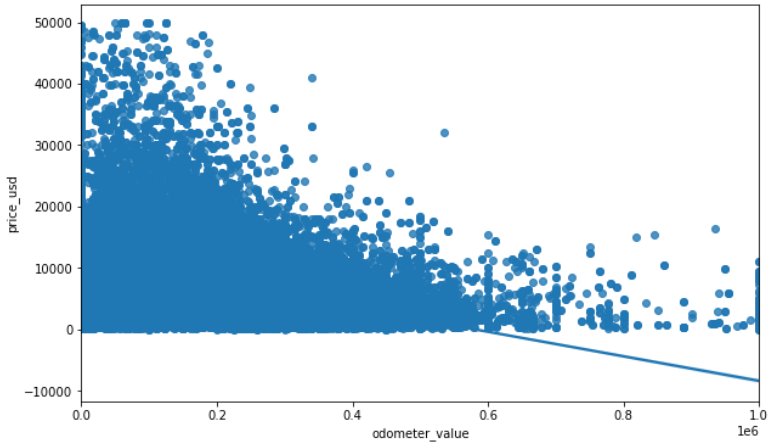


Deployment (Insights and Findings)

Regression Plots



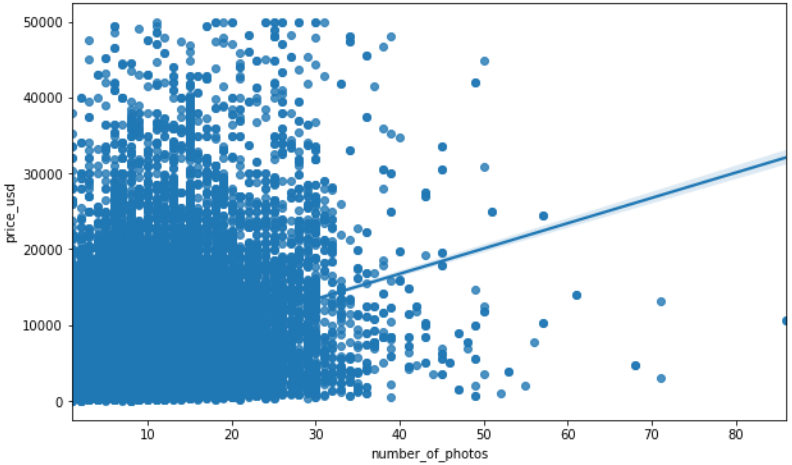
P-value = 0.0



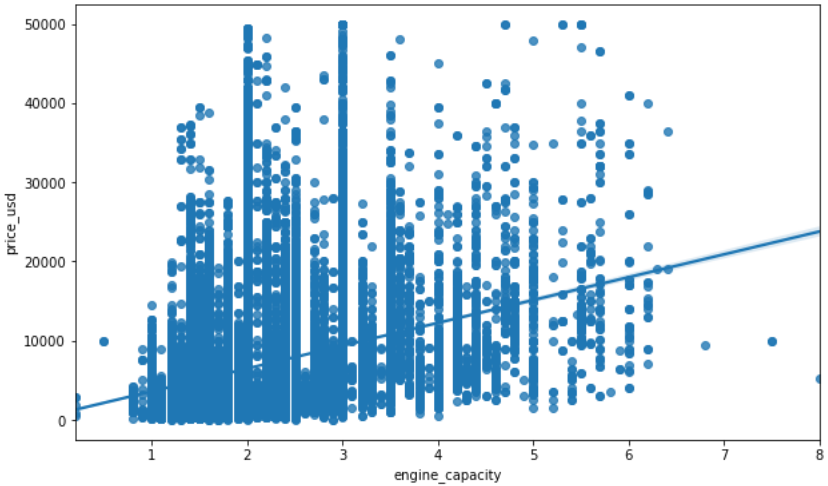
P-value = 0.0

Deployment (Insights and Findings)

Regression Plots



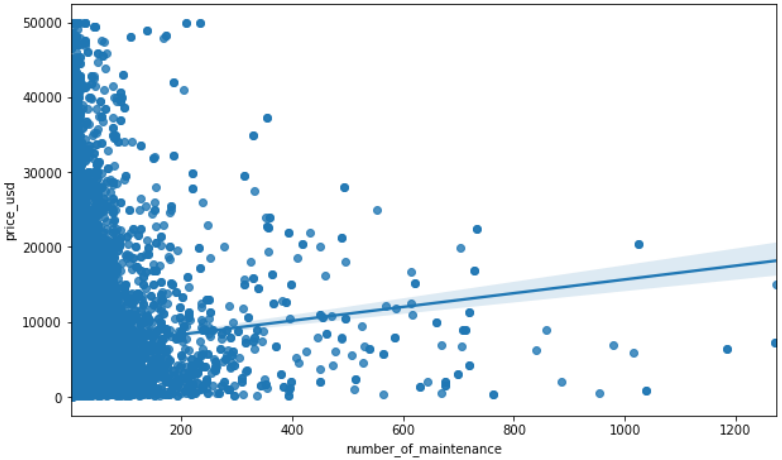
P-value = 0.0



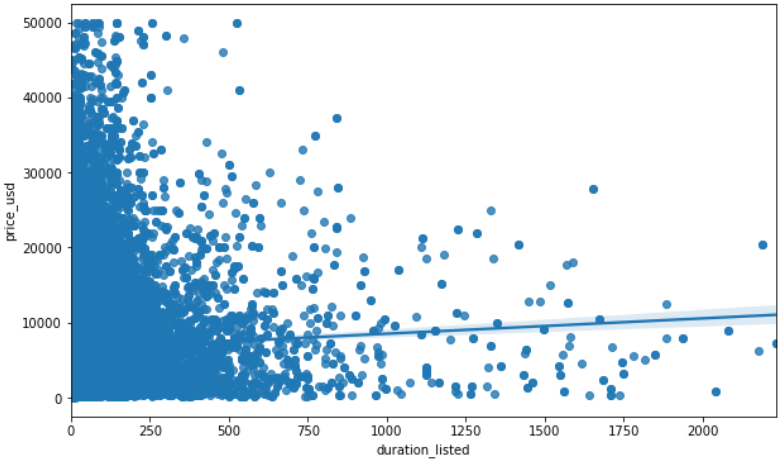
P-value = 0.0

Deployment (Insights and Findings)

Regression Plots

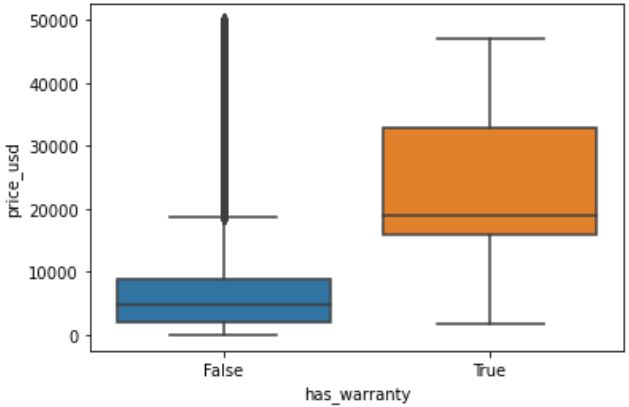


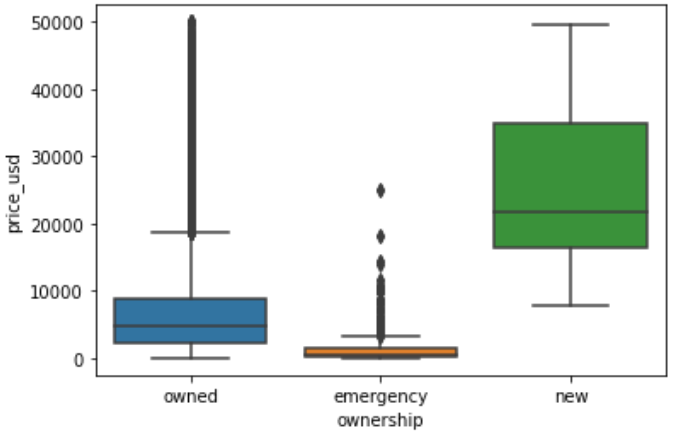
P-value = 1.0242198585534666e-39

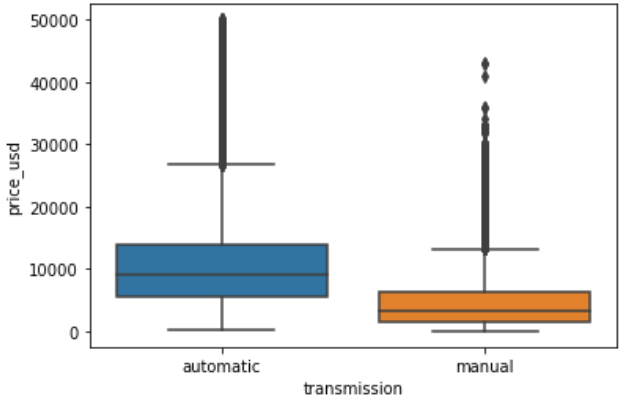


P-value = 9.7320020809580 92e-16

Deployment (Insights and Findings)

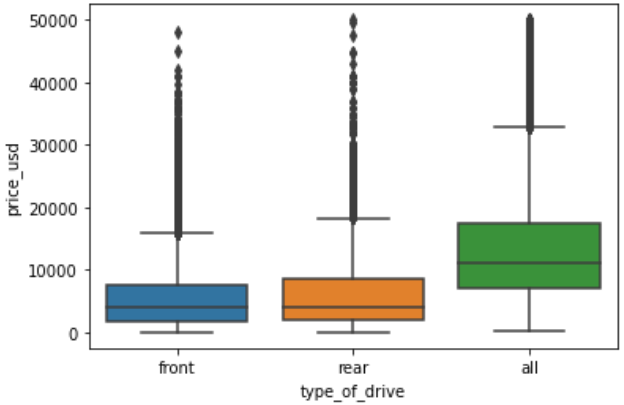


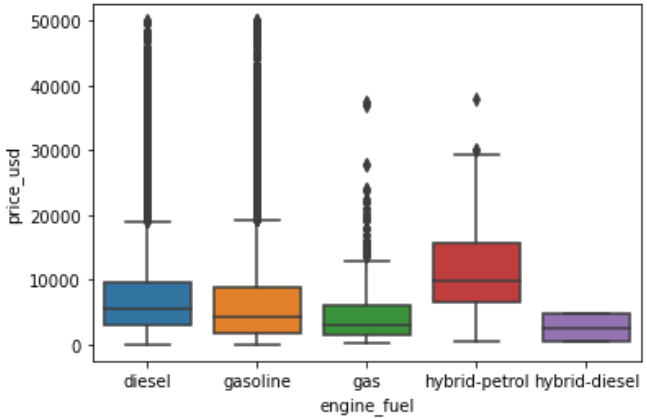


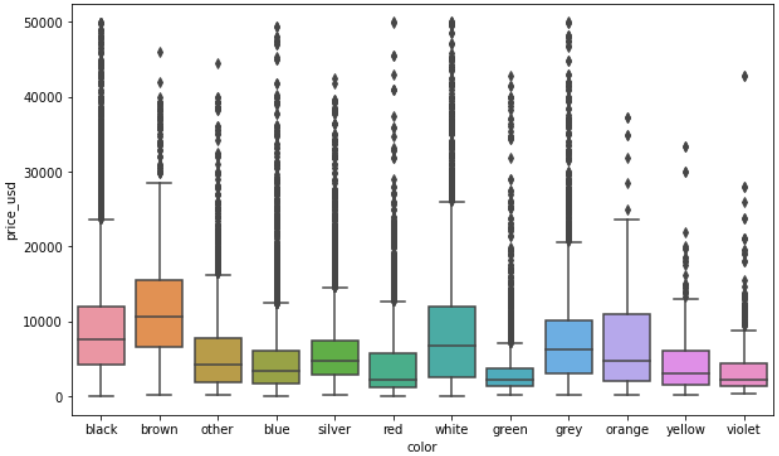


Box Plots

Deployment (Insights and Findings)

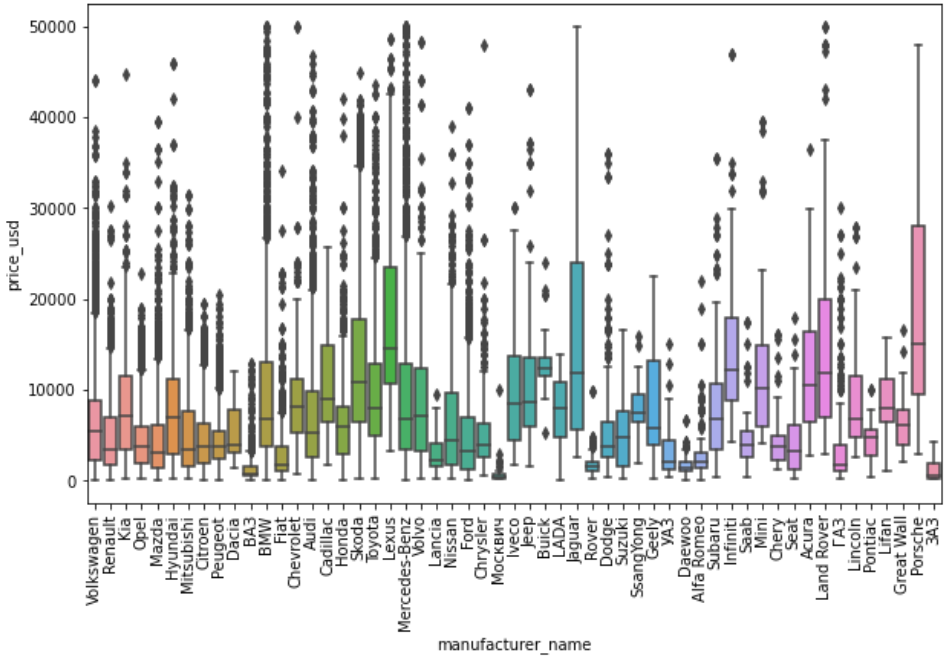


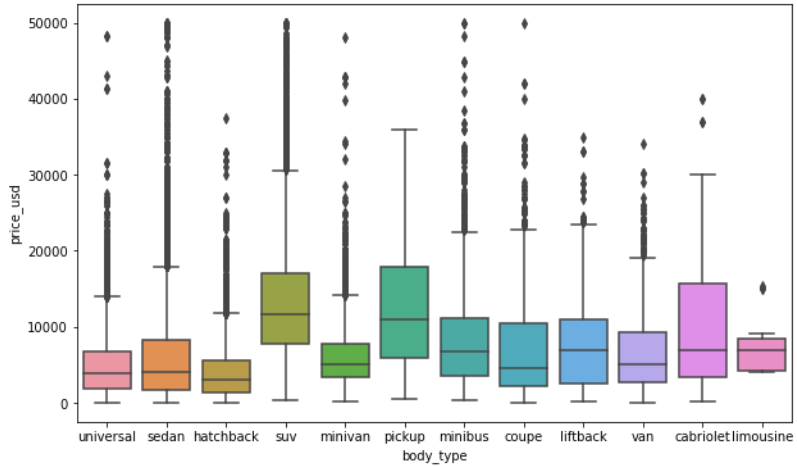




Box Plots

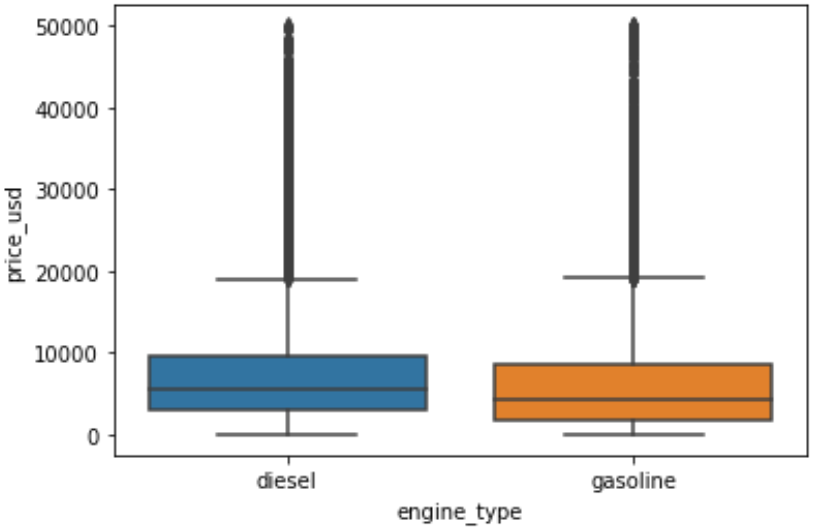
Deployment (Insights and Findings)

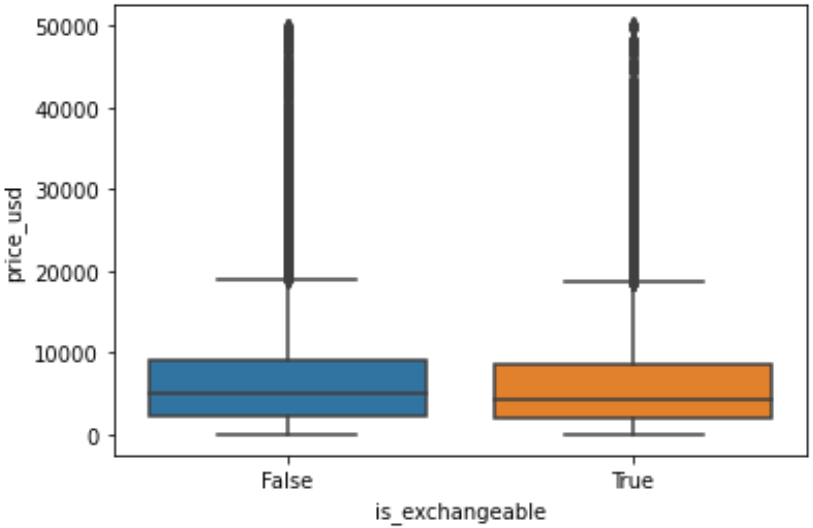




Box Plots

Deployment (Insights and Findings)





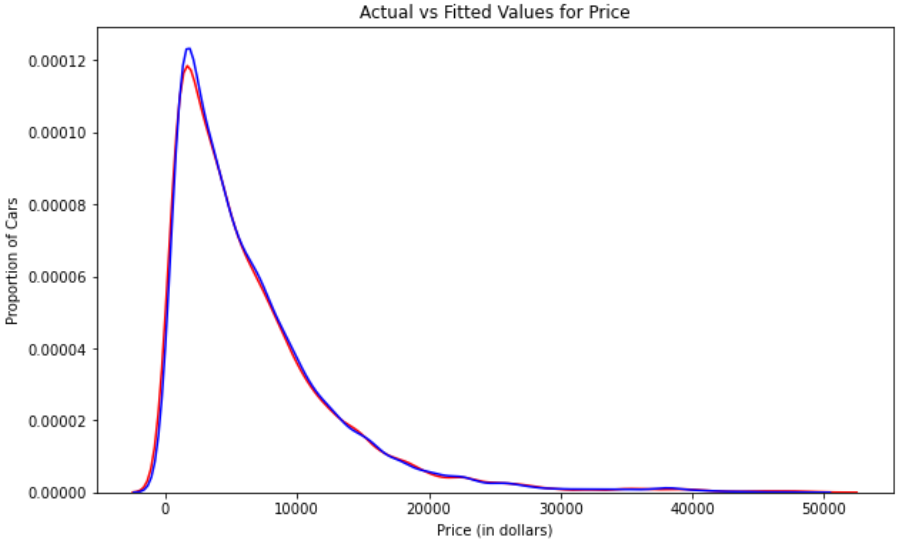
Box Plots

Recommendations

Features that can be utilized to improve the advertising approach.

- ✓ Manufacturer name
- ✓ transmission
- ✓ color
- ✓ Odometer value
- ✓ Year produced
- ✓ Engine fuel
- ✓ Engine capacity
- ✓ Body type
- ✓ Has warranty
- ✓ ownership
- ✓ Type of drive
- ✓ Number of photos

Random Forest's Plot



Reference

1. CRISP-DM - Data Science Process Alliance. (2022). Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Thank You