ALGONQUIN COLLEGE

Used Car Price Prediction

Yanhui Gao | # 041045034 | CST2208 Data Science Topic Final Project

Objective

Practice the machine learning workflow We have learned so far to predict a car's market price using its attributes. The purpose of this project is to understand and evaluate used car prices.

Study the Data

Checked the data

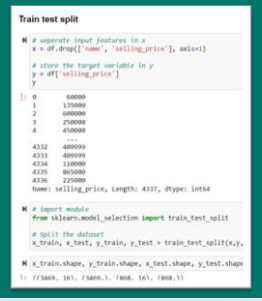
- 4340 observations and 8 features.
- 'selling_price' is technically the target.
- No missing value.
- No typos or inconsistent capitalization error.
- All data types are correct.

Data Cleaning

- Distribution of the numerical variables.
- Remove outliers with km_driven > 400,000
- Categorical Variables

Feature Engineering

- Group sparse classes
- Create dummy variables
- Visualize the correlation
- Splitting Training and Test data.



Model Training

Metric: Mean Absolute Error (MAE)

Linear Regression Model

Quick to train and test as a baseline algorithm.

Decision Tree Model

predicts the value of a target variable by learning simple decision rules inferred from the data features .

Random Forest Model

To account for the large number of features in the dataset and compare with the Decision Tree method.

Support Vector Machines

Try to use a typical competitive learning algorithm.

Results

Model	Train MAE	Test MAE
Linear Regression Model	227965	236407
Decision Tree Model	184773	197953
Random Forest Model	176124	177859
Support Vector Machines	307464	299928

Compared to Linear Regression, Decision tree, Support Vector Machines based methods, Random Forest did perform comparably well. However Support Vector Machines also be attributed to the difficulty in tuning most SVM methods.

Data Resource:
https://www.kaggle.com/code/mdejazulhassan/vehicle-dataset-from-cardekho/data