Analysis of Used Honda Cars

Summary

Data collection is done by scraping the website. 4,000 out of 17,000 car items are collected. The analysis is done to get the statistics from the data. The analysis of specific car model is done to see better understanding of each car model. Machine learning models are trained to predict price of car based on available data. Unfortunately, the evaluation results are not good. Several ideas seems viable to improve the models.

Data Collection

Data collection is done in 2 steps:

1. Visit the website and save the HTML files.
   Used Honda cars can be found on this web address https://www.mudah.my/malaysia/cars-for-sale/honda.
   The web address is iteratively visited by defining the page number in the argument. Every visited page will be saved as HTML file. All HTML files can be found in "html_pages" folder.
2. Read HTML files and retrieve car's information
   Each HTML page contains 40 items of car listings. There are around 17,000 items in Honda car listings. Only 4000 items were retrieved due to time limitation. The car's attributes are year, listing name/title, price, mileage, location, detailed page URL. The data is saved in csv file.

Potential improvements:

- Visit each car detailed page and retrieve more information

Furthermore, data cleaning and preprocessing are done to make it ready for analysis. This step includes:

- Clean nan rows
- Extract car model from the listing title
- Generate the middle value of mileage as real number for easier future usage. The mileage can also be separated into 2 components such as mileage minimum and mileage maximum.

Analysis

The data contains 3982 rows after preprocessing. The followings are several key findings from the data:

- The most popular model in the car listings is Honda Civic, followed by Honda City and Honda HR-V.
- The most common listing locations are Selangor, Kuala Lumpur, and Johor.
- Top 3 of car production year: 2018, 2019, 2017.

- Top 3 of car mileage: 85000-89999, 80000-84999, and 90000-94999.
- Concerning on middle of mileage, the mean of mileage is ~80,000

Analysis of specific car model

The following is the analysis of model Honda City only. The model can be replaced to get each analysis separately.
Figure 1 shows the price and years. When the cars are grouped by year and calculated the average price, you can see the chart on figure 2. As we expected, recent yeary have higher price.

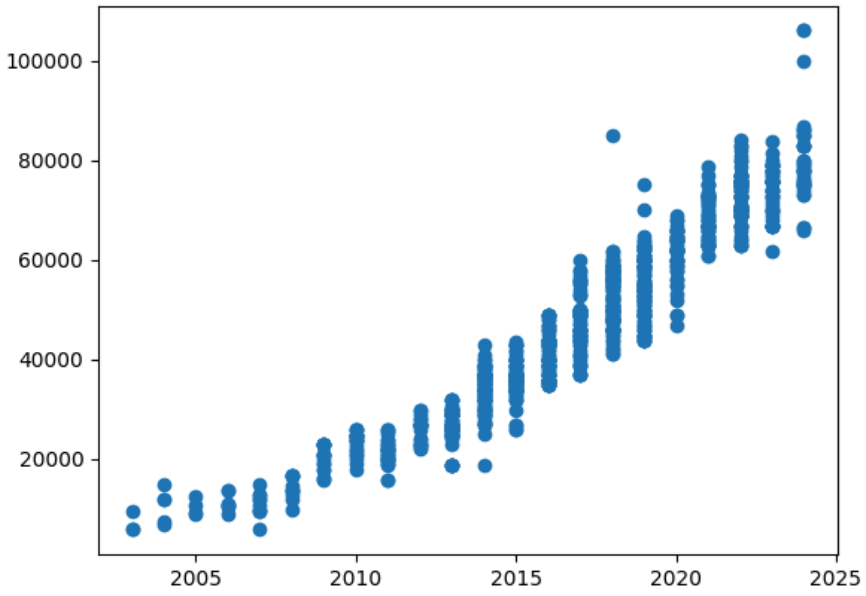

Figure 1. Honda Civic price and year distribution

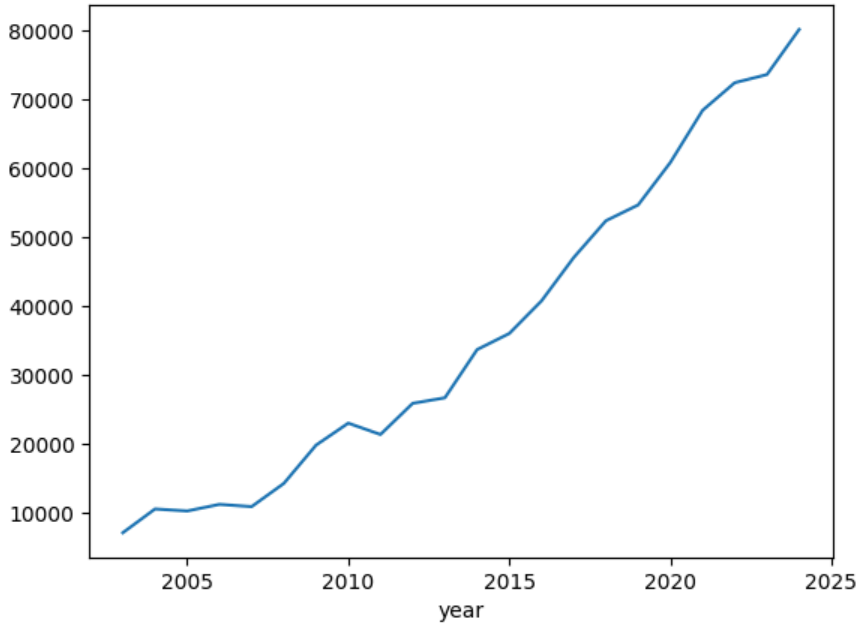

Figure 2. Average price of Honda Civic

Furthermore, price and year have very strong correlation. Pearson r value is 0.95 (p-value < 0.05). Additionally, price and mileage have quite good correlation. Peason r value is 0.73 (p-value < 0.05).

Predictive Modelling

This section describes the processes to train machine learning model. The objective is to predict car's price based on available data. Prior to training the model, feature engineering is done to convert categorical data such as car model and location into numerical values. The features include year, location, car model and mileage.

The data is split into two parts, 80% of data is training set and 20% is testing set. The models include Linear Regression, AdaBoost, and Neural Network (Multi Layer Perceptron). Mean Absolute Error (MAE) is used as the evaluation metric. Evaluation results shows bad performance from all models (shown in table 1). Possible reason is the features are too little. We can include information from car's detail.

Table 1. Evaluation results

Model	MAE
Linear Regression	21466.16
AdaBoost	18247.30
Neural Network	25437.25