CAPSTONE PROJECT

Sri Lankan Vehicle Price prediction Machine Learning Model

Abstract

Vehicle price could be predicted with 0.99 R2 score with decision tree machine learning model.

Tharindu Dissanayake

Tharindu.Dissanayake@dialog.lk

1.0 Introduction

Vehicles are essential for transportation. However, the price of vehicles varies in Sri Lanka as a non-vehicle manufacturing county. In this project, the machine learning model was created to predict Sri Lankan vehicle prices given the details of the vehicle. The machine learning models were created using the supervised learning method and they were regression-type machine learning models. The vehicle price is determined by several features such as brand, model, manufactured year, condition, transmission, body, fuel type, capacity, and mileage. These features were used in the model to predict the prices of vehicles.

Several machine learning algorithms were used to create the models in this project. The R2 score was used to determine the best model from the models. The R2 score varies between 0 and 1. If the R2 value is close to 1 then the predicted value is close to the actual value in a regression model. Therefore, the highest R2 score was used to determine the best model out of several models.

This project report contains six chapters including Chapter 1 which is the introduction to this project. Chapter 2 describes the data used in the machine learning model. Chapter 3 presents the design of the machine learning method to predict vehicle prices. Chapter 4 includes the results observed from the proposed machine learning method. Chapter 5 share the conclusion of the project and Chapter 6 share the discussion regarding the project.

2.0 Data

The data needed for the model building was obtained from the Kaggle website [1]. Dataset was the Sri Lanka vehicle prices dataset on the Kaggle website. The data contain details related to Sri Lankan vehicles that were listed for sale. To build this dataset, data was taken for ikman.lk website which is Sri Lankan online vehicle buying and selling platform. The data is updated monthly from ikman.lk website using an automated script. Figure 1 shows a sample of the dataset.

	Title	Sub_title	Price	Brand	Model	Edition	Year	Condition	Transmission	Body	Fuel	Capacity	Mileage	Location	Seller_type	published_date
35502	Toyota Premio 2018 for sale	Posted on 16 Nov 10:21 am, Akkarepattu, Ampara	Rs 16,500,000	Toyota	Premio	Toyota	2018	Used	Automatic	Saloon	Petrol	1,499 cc	34,000 km	Akkarepattu, Ampara	Member	11/16/2022 10:21
113285	Toyota Vitz 2017	Posted by Ajith on 2021-09-25 8:26 pm, Anurada...	Rs. 8,500,000	Toyota	Vitz	NaN	2017	Used	Automatic	Car	Petrol	1000	32000	Anuradapura	Member	9/25/2021 20:26
35505	Honda Vezel Z grade 2014 for sale	Posted on 27 Oct 11:33 pm, Maharagama, Colombo	Rs 8,395,000	Honda	Vezel	Z grade	2014	Used	Automatic	NaN	Petrol	1,500 cc	91,000 km	Maharagama, Colombo	Premium-Member	10/27/2022 23:33
107144	Bajaj CT-100 2005	Posted by Udara on 2021-10-02 9:38 pm, Homagama	Rs. 77,000	Bajaj	CT-100	NaN	2005	Used	Manual	Motorbike	Petrol	100	50000	Homagama	Member	10/2/2021 21:38
1251	Toyota Aqua S Limiterd 2012 for sale	Posted on 28 Sep 9:37 pm, Ambalangoda, Galle	Rs 5,575,000	Toyota	Aqua	S Limiterd	2012	Used	Automatic	Hatchback	Petrol	1,500 cc	88,000 km	Ambalangoda, Galle	Member	9/28/2022 21:37

Figure 2.1: Sample of the Sri Lanka vehicle prices dataset [1].

3.0 Methodology

The supervised regression machine learning model was created for the project. However, the dataset could not be directly fed to regression machine learning algorithms to generate models. Therefore, the dataset was pre-processed before feeding it to algorithms to generate models.

To pre-process datasets to use in regression algorithms, the dataset was analyzed to identify issues to remove them from the dataset, and then the dataset was converted to numerical values. Figure 3.1 shows the data information of the dataset. As in figure 3.1, the Edition column does not have more than half of the data points. Therefore, this column was removed from the dataset. Then all the rows with missing values were removed to remove null data points.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 123971 entries, 0 to 123970
Data columns (total 16 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Title           123971 non-null  object
 1   Sub_title       123971 non-null  object
 2   Price           123971 non-null  object
 3   Brand           123971 non-null  object
 4   Model           123970 non-null  object
 5   Edition         47538 non-null   object
 6   Year            123971 non-null  int64
 7   Condition       123971 non-null  object
 8   Transmission    123970 non-null  object
 9   Body            118876 non-null  object
 10  Fuel            123971 non-null  object
 11  Capacity        123766 non-null  object
 12  Mileage         123971 non-null  object
 13  Location        123971 non-null  object
 14  Seller_type     123971 non-null  object
 15  published_date  123971 non-null  object
dtypes: int64(1), object(15)
memory usage: 15.1+ MB
```

Figure 3.2: Data information of the dataset.

Categorical columns such as Brand, Model, Condition, Transmission, Body, and Fuel were processed then. Incorrect categories were removed, and data was changed to numerical values using predefined values unique to the category. Mileage, Capacity, and Price columns were also processed to change their values to numerical-only values, and then outliers were removed. Finally, brand type, vehicle manufacture year, condition type, transmission type, body type, fuel type, mileage (km), capacity (cc), published year, published month, and published day were used as feature variables, and price (Rs) was used as the target variable. The correlation matrix shown in figure 3.2 were observed to check the correlation between the feature variables.

	Brand_type	Model_type	Year	Condition_type	Transmission_type	Body_type	Fuel_type	Mileage_km	Capacity_cc	published_year	published_month	published_day
Brand_type	1.000000	0.252560	0.020206	-0.019509	0.139888	0.368803	0.195846	-0.256732	-0.473164	-0.391104	0.198899	0.001860
Model_type	0.252560	1.000000	-0.126948	-0.015022	0.043570	0.477476	0.057281	-0.047414	-0.133160	-0.498710	0.247847	0.013330
Year	0.020206	-0.126948	1.000000	0.075175	-0.336634	0.001798	-0.107602	-0.460994	-0.261757	-0.020432	0.023403	-0.005125
Condition_type	-0.019509	-0.015022	0.075175	1.000000	-0.005384	-0.033929	0.029696	-0.136481	-0.018466	0.077925	-0.060270	-0.000719
Transmission_type	0.139888	0.043570	-0.336634	-0.005384	1.000000	0.009811	0.092604	0.102549	0.088338	0.012562	-0.019360	-0.002545
Body_type	0.368803	0.477476	0.001798	-0.033929	0.009811	1.000000	0.050784	-0.161485	-0.326230	-0.936736	0.481363	-0.008141
Fuel_type	0.195846	0.057281	-0.107602	0.029696	0.092604	0.050784	1.000000	-0.056139	-0.302537	-0.074597	0.019351	-0.000320
Mileage_km	-0.256732	-0.047414	-0.460994	-0.136481	0.102549	-0.161485	-0.056139	1.000000	0.492845	0.165180	-0.087987	0.000395
Capacity_cc	-0.473164	-0.133160	-0.261757	-0.018466	0.088338	-0.326230	-0.302537	0.492845	1.000000	0.369350	-0.199370	0.002186
published_year	-0.391104	-0.498710	-0.020432	0.077925	0.012562	-0.936736	-0.074597	0.165180	0.369350	1.000000	-0.516222	0.010117
published_month	0.198899	0.247847	0.023403	-0.060270	-0.019360	0.481363	0.019351	-0.087987	-0.199370	-0.516222	1.000000	0.040693
published_day	0.001860	0.013330	-0.005125	-0.000719	-0.002545	-0.008141	-0.000320	0.000395	0.002186	0.010117	0.040693	1.000000

Figure 3.2: Correlation Matrix.

Linear regression, decision tree regression, lasso regression, and random forest algorithms were used in the project to create models using the processed dataset. 70% of the dataset, which was 61863 data rows, was used to train the models and 30% of the dataset, which was 26514 data rows, was used to test the models. The R2 score values were used to get the best model.

4.0 Results

The R2 score of the machine learning models was used to select the best model. Table 4.1 and figure 4.1 show the R2 score values of models. According to R2 scores, the decision tree regression model with a 0.999819 R2 score value is the best model to predict vehicle prices.

Table 4.1: R2 score values of machine learning models.

ML Algorithm	R2 Score
Linear regression	0.616834
Decision tree regression	0.999819
Lasso regression	0.616834
Random forest	0.991348

Figure 4.1: R2 score values of machine learning models.

Table 4.1 and figure 4.1 show feature importance scores for different features of the decision tree regression model. According to figure 4.1, capacity followed by manufactured year is the two most important parameters when predicting vehicle prices using the decision tree regression model.

Table 4.2: Feature importance scores for different features.

Features	Feature Importance
Brand_type	0.061604
Year	0.286298
Condition_type	0.001704
Transmission_type	0.080054
Body_type	0.128455
Fuel_type	0.006954
Mileage_km	0.026924
Capacity_cc	0.370727
Published_year	0.00071
Published_month	0.022561
Published_day	0.01401

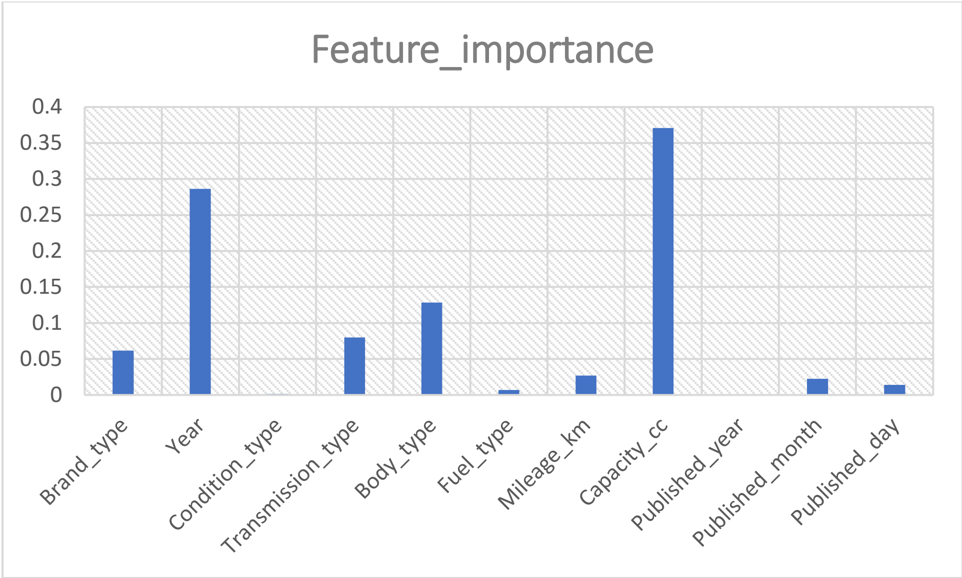

Figure 4.2: Feature importance scores for different features.

5.0 Conclusion

The decision tree regression model is the best model to predict vehicle prices. It shows an R2 score of 0. 999819. Vehicle capacity and vehicle manufactured year are the most important features in detecting the vehicle price using the decision tree regression model.

6.0 Discussion

This machine learning model could mostly be used to predict used vehicle prices as most of the available data is from used vehicles.

Due to the high amount of data points and limited CPU available, most of the machine learning algorithms could not be used to create models in this project.

Reference

[1] L. Jayawardena, “Sri Lanka Vehicle Prices Dataset” in *Kaggle*, March. 2021. [Online]. Available: https://www.kaggle.com/datasets/lasaljaywardena/sri-lanka-vehicle-prices-dataset