ITEC4433 - Data Warehousing and Data Mining Term Project Report

"HOW TO PREDICT CAR SELLING"

Student Name & Surname: Gizem Başaran
Student Number: 20SOFT1041

The Definition of the Problem:

To analyze Vehicle dataset using Weka. And estimate the selling price of the vehicles with Linear Regression and Random Forest Algorithms.

Data Acquisition and Dataset Properties

I got my dataset from kaggle. The subject is to be able to make a price estimation with used vehicle data. It has 8 features. these; name, year, selling_price, km_driven, fuel, seller_type, transmission, owner. It contains 4340 sample data in total.

Attributes:
Name: Name of the cars.
Year: Year of the car when it was bought.
Selling_price: Price at which the car is being sold.
Km_driven: Number of Kilometres the car is driven.
Fuel: petrol / diesel / CNG / LPG / electric.
Seller_type: Individual or a Dealer.
Transmission: Automatic/Manual.
Owner: Number of previous owners of the car.

Selecting the Data Mining Method

I checked if I have empty data by preprocessing.

```
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   name           4340 non-null   object
 1   year           4340 non-null   int64
 2   selling_price  4340 non-null   int64
 3   km_driven      4340 non-null   int64
 4   fuel           4340 non-null   object
 5   seller_type    4340 non-null   object
 6   transmission   4340 non-null   object
 7   owner          4340 non-null   object
dtypes: int64(3), object(5)
```

Missing values control with isna() in Python Pandas Library.

```
name             0
year             0
selling_price    0
km_driven        0
```

```
fuel             0
seller_type      0
transmission     0
owner            0
dtype: int64
```

Afterwards, I tried the classify methods from Weka and chose the Random Forest machine learning algorithm and reached 70% accuracy.
I selected the factors that I thought were related to the vendor's pricing and used them to fit a linear model to the training data to create a linear regression model. Using an optimization approach, I have estimated the coefficients of this linear model. Linear Regression machine learning algorithm reached 85% accuracy.

Executing the Data Mining Function

I applied the algorithm I chose by choosing Classify and Random Forest from Weka.
I applied Linear Regression using pandas, numpy, LinearRegression libraries using python. I removed these object features to feature selection part; seller_type, owner, fuel, name, transmission.

Random Forest:

```
Correctly Classified Instances        3054               70.3687 %
Incorrectly Classified Instances      1286               29.6313 %
Kappa statistic                          0.3536
Mean absolute error                      0.1468
Root mean squared error                  0.2866
Relative absolute error                 72.7831 %
Root relative squared error             90.2699 %
Total Number of Instances             4340
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.881	0.499	0.768	0.881	0.821	0.419	0.821	0.893	First Owner
	0.442	0.112	0.574	0.442	0.499	0.362	0.768	0.556	Second Owner
	0.160	0.008	0.289	0.160	0.206	0.204	0.724	0.132	Fourth & Above Owner
	0.181	0.033	0.291	0.181	0.223	0.185	0.704	0.199	Third Owner
	0.118	0.001	0.333	0.118	0.174	0.196	0.945	0.409	Test Drive Car
Weighted Avg.	0.704	0.357	0.675	0.704	0.683	0.383	0.798	0.743	

Obtaining the Results and Interpretation

Linear Regression:

```
MAE: 0.9937753386731192
MSE:  2.162871527787962
RMSE: 1.4706704347976682
```

Random Forest:

```
MAE:  0.1468
```

```
MSE:  0.2866
RMSE: 0.5354
```

Examples for Customers:

```
prediction: 366599.48, target value: [350000]
prediction: 502965.82, target value: [450000]
prediction: 1157524.28, target value: [930000]
prediction: 823426.73, target value: [685000]
prediction: 332507.89, target value: [325000]
prediction: 502965.82, target value: [450000]
```