University of Central Lancashire
UCLan

School of Engineering

Year 4: 2020-21

Programming with Data

EL4013

Used Car Price Prediction

Saad Bin Munir

Abstract

This project predicts the price of a car by building a model using previous data. The used cars dataset is obtained from Kaggle website. Different regressor modelling techniques (i.e. Linear Regressor, Decision Tree, Multilayer Perceptrons and Random Forest) are applied and their variances and accuracies are computed. The Random Forest Regressor gave the best results showing the variance of 96% and an accuracy of 89%.

University of Central Lancashire
UCLan

Motivation

- There is always a need of car price prediction in order to find out the worth of vehicle before spending money towards it.
- Using the model in this project, a user will be able to predict price of a car while buying or selling it.

University of Central Lancashire
UCLan

Datasets

- The datasets consists of 100,000 used cars sold in the UK.
- The datasets are obtained from Kaggle website.
- The datasets contain columns listing price, brand, model, year, mileage, mpg, transmission type, fuel type, and engine size.

University of Central Lancashire
UCLan

Data Preparation and Cleaning

- The data was provided in separate spreadsheets for each brand of vehicles. An extra column containing brand name is added to each data frame and the data is combined into a single data frame.
- Each column in dataset is closely analysed and minimum and maximum values are found out.
- The data had some values which were unrealistic, for example engine size of 0, mileage of 0 for some old cars, mpg less than 10, and year of 2060 etc. These values are discarded.
- Some cheap cars had very high price, and some expensive cars had very low price listed in data frame, these values are analysed and discarded if needed.

University of Central Lancashire
UCLan

Research Question

Predict the price of a used car by building a model using data of used cars previously sold in the market.

University of Central Lancashire
UCLan

Methods

We used quantitative research methods to analyse the data as our aim was to predict the price which is a numeric value.

Correlation is initially used to analyse the effect of each variable on price of a car. This method is important in our case as it gives the relationship of different attributes with price.

Regression methods are used to train the model in order to predict the price of car. These methods are chosen as they best describe the impact of independent variables on dependent variables and their relationship.

Methods (Cont'd)

Linear Regressor is initially used as it best describes the linear relationship between different numeric variables.

Decision Tree Regressor is chosen because it is a powerful algorithm for fitting a complex dataset.

Multilayer Perceptron (MLP) is used because it is capable of handling the data which is non linearly separable.

Random Forest Regressor is chosen because of it's ability to contain more than one decision tree.

University of Central Lancashire
UCLan

Findings

The relationship of each attribute, found out using correlation, is as follows:

University of Central Lancashire
UCLan

Findings (Cont'd)

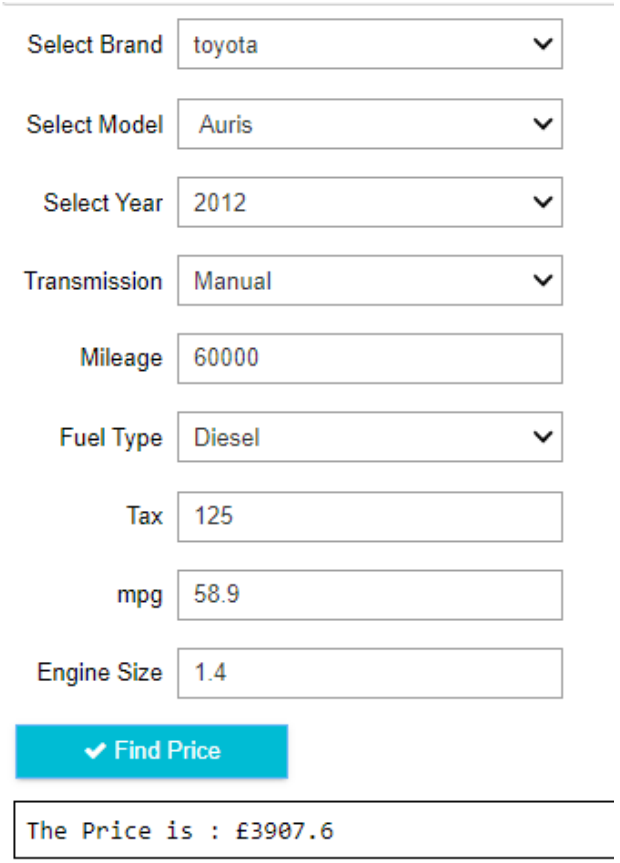
The variance and accuracy of dataset from each model is as follows:

	Model	R2 Score	RMSE	% Accuracy
0	Linear Regressor	0.874771	3473.584531	79
1	Decision Tree Regressor	0.943045	2342.554850	86
2	Multilayer Perceptrons Regressor	0.896446	3158.693904	81
3	Random Forest Regressor	0.963185	1883.367924	89

University of Central Lancashire
UCLan

User Interface

- The user interface takes car attributes from the user and displays the worth of car accordingly
- The appearance of user interface and price prediction display can be seen from the image.



University of Central Lancashire
UCLan

Conclusions

- It is concluded that the engine size and mileage greatly affect the price of car while the tax has least effect on price.
- The Random Forest Regressor is most accurate model for our dataset with R^2 score of 96% and accuracy of 89%.
- The variance and accuracy can be further improved if each brand of car is separately trained instead of training them altogether.

University of Central Lancashire
UCLan

Limitations

- The results are only applicable to 9 brands of cars because of limitation of dataset.
- This model does not take location, colour and condition of car into account which may affect the predicted price.

University of Central Lancashire
UCLan

References

ADITYA, 100,000 UK Used Car Data set. Available: https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes.

KARACA, E., Predicting Audi Car Prices. Available: https://www.kaggle.com/emrekaraca/predicting-audi-car-prices.