An analytical approach for predicting used BMW car price based on the presence of certain used car data

Nikita Agarwal

Under the guidance of

Prof: Dr. Balu Gokaraju

North Carolina Agricultural and Technical State University

1601 E Market St, Greensboro, NC 27411

nagarwal@aggies.ncat.edu

***Abstract*— Finding the price of a used car needs a fair system of evaluation that not only benefits the buyer, but also the car dealer. The most important question that needs to be answered concerns the identification of the car features that can significantly affect the price. The goal of this project is to create a machine learning model that can predict the price of a used car based on its features.**

A dataset containing 10782 used cars from the UK was used. The dataset contains significant used car features together with the price expressed in pounds (£) they have been sold for. The dataset has been analysed (data gathering, data preprocessing and data exploration) and then machine learning algorithms have been applied to it. Different machine learning prediction models were used to train the dataset. Based on analysis either the Extra tree regression Model or the Random forest regressor have turned out to be the best models for price prediction with an accuracy ranging from 90% to 97%.

***Index Terms*— Extra tree regression, Random forest regressor, prediction models, features**

I. Introduction

According to research, consumers in the UK are changing cars every 7.7 years. Since the invention of the car, the trading of used cars has been a substantial business for both private sellers and commercial car dealers [1]. The car pricing process is mostly based on comparisons to other advertisements on various websites or on the seller's gut feeling, not always leading to optimal pricing. In fact, price should be accurately matched to one's car, taking into account various features of the item that can influence its price. This is exactly the problem this project wants to address, trying to answer the following questions "What features do I need to take care of when determining the price of a used car?" and "how can I build a machine learning model which is able to predict prices of used cars?"- If this information was freely available, better pricing decisions could be taken, allowing for a fair and transparent price for all parties involved.

Machine learning can be an especially useful to explore these hidden features in the price of the used car [2].

A. Literature Review

Wu et al. [3] conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbor machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it [4].

Furthermore, Pudaruth [5] applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior color, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%.

Gonggie [6] proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models

Noor and Jan [7] build a model for car price prediction by using multiple linear regression. The dataset was created during the two-months period and included the following features: price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage in kilometer, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model

year and model as input features. With the given setup authors were able to achieve prediction accuracy of 98%

Dataset:

This data set is collected from used cars in UK. The data set contains information of price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size.

Dependent variable: price

Independent variables: model, year, transmission, mileage, fuel type, tax, miles per gallon, engine size.

Below is more information about the dataset:

- Dataset Title: 1000,000 UK used car dataset
- Number of Instances: 10782
- Number of Attributes: 9

Method:

Firstly, we have used the following regression models- Ridge, Lasso, Bayesian ridge, K neighbors regressor, Decision tree regressor, AdaBoost regressor, Bagging regressor, Extra trees regressor and Random Forest regressor and calculated r2-score. The Machine learning method for features selection such as Extra tree classifier, Correlation Feature importance and PCA Analysis were applied to select the best features. In addition, after PCA analysis we again used the above-mentioned regression techniques to calculate r2 score.
Secondly, we have compared the r2-score before and after PCA for all the regression models applied.
Based on the highest r2-score we have selected regression model and built the car price prediction model.

Process Block Diagram:

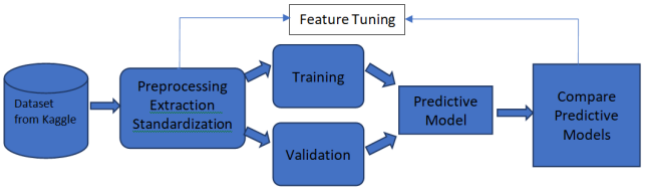

Figure 1: Process block diagram

Pre-Processing:

After having gathered the dataset, the data preprocessing on the car producers datasets has been performed. The first step was to investigate the structure of the car producers' datasets, looking at its shape and at the type of variables included. Functions like head(), shape, describe(), info, and, dtypes were used to investigate them. Most of the independent variables were string, integer and float data types.
There were 5 numerical columns, 3 categorical columns and 1 year column. Three categorical columns are as follows: Model, Transmission and Fuel type. Now we are converting these, 3 categorical columns into numerical columns by ordinal encoder in sklearn. First step for data cleaning was to remove duplicate rows from the data. In this dataset, number of duplicate rows are 179. So we removed these duplicate rows by drop_duplicates() and after removing these duplicate rows we have 10664 rows and 9 features.

The dataset does not contain any missing values or Zeros values that does not have a meaning. So, no removal of data was required. As a next step, we are checking for outliers in the dataset by histogram and box-whisker plots and extreme values were dropped because they inhibit prediction power of the model. There were only three data points greater then 178987 in the mileage column so they are outliers. In year column the data from 1995 to 2000 are outlier because our maximum data fall into the range of 2010 to 2020. In price column, price greater then 99950 are outliers. In mpg values greater then 400 and less then 8 are outlier. In addition, engine size has outliers as well. These values were also dropped from data-set because these are noise for the data. In addition, model transmission and fuel type have no outlier.

At the end of this process, we remove all duplicate rows, missing values and outliers and now the dataset is clean and we have 9799 rows and 9 features.

Feature Engineering and Training Data:

As a next step, the relationship between the dependent variable (price) and the independent variables has been investigated. Pearson correlation coefficient was used to select the relevant features from the dataset. The figure below shows the heatmap of the pearson correlation for datasets using corr() function which showed the correlation between data points as shown in figure 2.

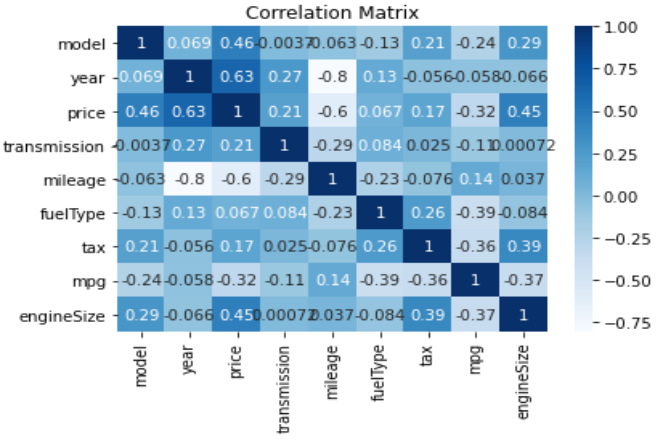

Figure 2: Correlation matrix plot

1. Price & Year have a direct connection -- maybe the older car has the less price.

2. Year & mileage are also correlated. – maybe more the year more will be mileage.
3. Price and mileage are also correlated—maybe the higher mileage of the car have lesser price.
4. miles per gallon (mpg) and tax seem not significant correlate to price.

Furthermore, we used Extra trees classifier to get the best features from the available features. We found that mpg, year, tax, model and mileage are the best features which plays vital role in building the price predication model.

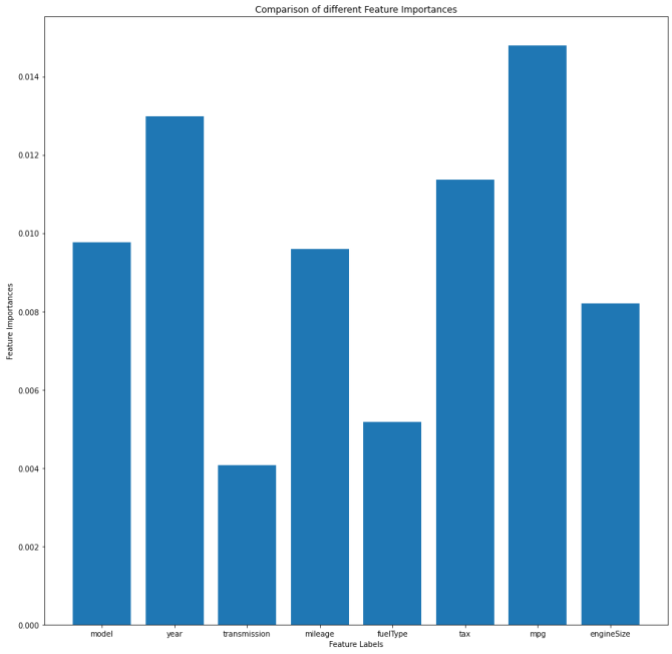

Figure 3: Bar plot for Extra trees classifier

Splitting the dataset into training and testing datasets. 80% of the dataset was used for training the model and 20% for testing the model for prediction. Random state of 42 was taken.

Machine Learning Model Development:

After having explored the dataset and the relationship between the variables for each of the car producers file, it has been decided to proceed with the creation of a predictive model that would have been able to identify the car features that mostly influence the price of a used car and then predict its price.

The following steps have been performed for each of the car producer's datasets

- Importing all the machine learning modules from sklearn machine learning library.
- Standardisation

The result of **standardization** (or **Z-score normalization**) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively. The equation is shown below:

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

This technique is to re-scale features value with the distribution value between 0 and 1.

- The price column was dropped to keep a realistic output (price) after scaling the input (features).

We have used the following regression models- Ridge, Lasso, Bayesian ridge, K neighbors regressor Decision tree regressor, AdaBoost regressor, Bagging regressor, Extra trees regressor and Random Forest regressor and calculated r2-score before and after PCA for all the regression models. Based on the highest r2-score we have selected regression model and built the car price prediction model.

Results:

Based on the regression model we have the following R2-score after the standardization of the original dataset as shown in the table 1.

Models	R2 Score
LR	0.761
R	0.761
L	0.761
BR	0.761
KNR	0.933
DTR	0.906
SVR	0.008
ABR	0.717
BR	0. 953
ETR	0.952
GBR	0.955
RFR	0.953

Table 1: R2 score of the original dataset

Figure 4 shows the comparison of R2 score after the standardization of the original dataset.

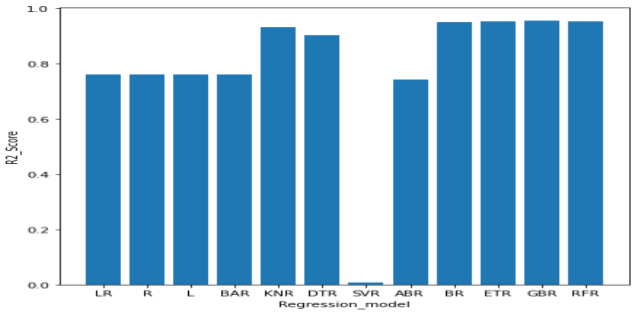

Figure 4: Bar plot for the comparison of R2 score

Root mean square error on standardize dataset:

Models	RMSE
LR	5566.262
R	5566.288
L	5566.216
BR	5566.365
KNR	2946.857
DTR	3541.547
SVR	11348.393
ABR	7456.760
BR	2464.032
ETR	2472.653
GBR	2408.301
RFR	2474.418

Table 2: RMSE of the original dataset

Based on the regression model we have the following R2 score after applying the PCA on standardize dataset.

Models	R2 Score
LR	0.684
R	0.684
L	0.684
BR	0.684
KNR	0.918
DTR	0.847
SVR	0.031
ABR	0.676
BR	0.920
ETR	0.924
GBR	0.867
RFR	0.921

Table 3: R2 score after PCA of the dataset

Below bar plot shows the comparison of R2 score after applying the PCA on standardize dataset.

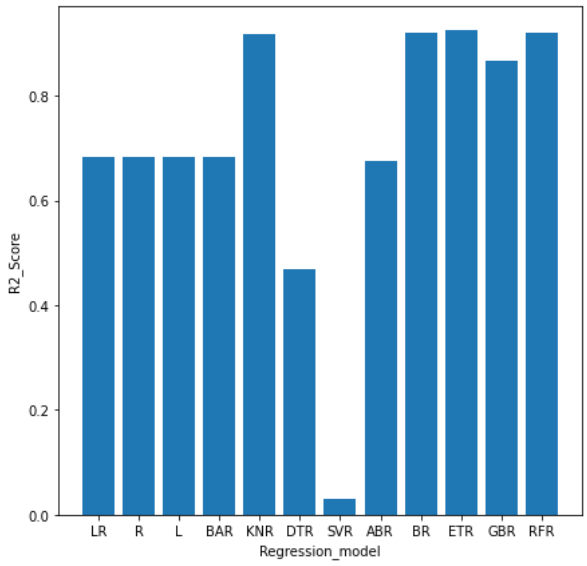

Figure 5: Bar plot for the comparison of R2 score after PCA

Root mean square error after PCA on standardize dataset:

Models	RMSE
LR	5566.282
R	5566.288
L	5566.216
BR	5566.365
KNR	2946.857
DTR	3453.756
SVR	11348.393
ABR	5971.693
BR	2476.385
ETR	2451.979
GBR	2411.199
RFR	2468.806

Table 4: RMSE after PCA on the dataset

Discussion:

Findings and Observation:

Our results from the regression model shows that the R2 score before PCA have good results compared to R2 score after PCA so we build our prediction model based on the R2 score after the standardization of the original dataset. We have used twelve regression models to calculate the R2 score and out of those we found that extra tree regressor and random forest regressor have the more then 90% accuracy and the smallest root mean square error that's why we build the predication model using extra tree regressor. We also found that predicated price of used car is somewhat similar to given price.

Our results from scatter matrix plot shows that the Price & Year have a direct connection -- maybe the older car has the lesser price. Year & mileage are also correlated– maybe more the year more will be mileage.

Our results from Extra trees classifier shows that mpg, year, tax, model and mileage are the best features which plays vital role in building the price predication model.

Below figure 6 shows comparison between car predicted price and given price:

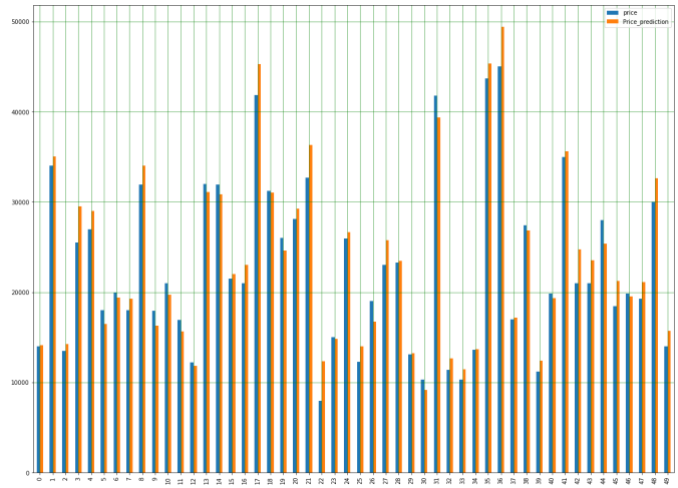

Figure 6: Comparison between car predicted price and given price:

Conclusion:

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data. In this research, dataset is standardized and cleaned to avoid unnecessary noise for machine learning algorithms. Data cleaning is one of the processes that increases prediction performance, yet insufficient for the cases of complex data sets. In this paper, twelve different machine learning techniques have been used to forecast the price of used cars in UK. Among all the regression model, extra tree regression gives us more than 92% accuracy before and after PCA that why we have chosen extra tree regression to build the predication model. From the prediction model we concluded that the predicated price is somewhat similar to the actual price.

The main limitation of this study is the low number of records that have been used. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

REFERENCES:

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from: http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January 2014].

[2]MOTORS MEGA. 2014. Available from: http://motors.mega.mu/news/2013/12/17/auto-market-8-decrease-sales-newcars/ [Accessed 17 January 2014].

[3] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817. [2] Du, J., [4]Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.

[5] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[6] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE

[7] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.