Advanced Car Price Modelling and Prediction

Michail Tsagris and Stefanos Fafalios

Abstract The scope of the paper is modelling and prediction of brand new car prices in the Greek market. At first the most important car characteristics are detected via a state-of-the-art machine learning variable selection algorithm. Statistical (log-normal regression) and machine learning algorithms (random forest and support vector regression) operating on the selected characteristics evaluate the predictive performance in multiple predictive aspects. The overall analysis is mainly beneficiary for consumers as it reveals the important car characteristics associated with car prices. Further, the optimal predictive model achieves high predictability levels and provides evidence for a car being over or under-priced.

Keywords Car market · Price prediction · Variable selection · Nonlinear models

1 Introduction

Car price modelling and prediction has not attracted significant research interest, especially in the field of economics, the most suitable environment for research in this area. Some examples include Eckard (1985) who showed, empirically, that U.S. state regulation of new car dealer entry produces higher new car prices, as predicted by economic theory. Verboven (1996) explained the presence of price discrimination across European countries and attributed this phenomenon to cross-country differences in price elasticities, differences in quota regimes and differences in the degree of collusive behaviour. Matas (2009) constructed a quality-adjusted price index for the Spanish car market over the period 1981–2005. However, those papers cannot be exploited for the present analysis, due to the rapid technological advances and the adoption of sophisticated functionalities such as ESP, parking and weather sensors not available at that time.

M. Tsagris (✉)
Department of Economics, University of Crete, Gallos Campus, Rethymnon, Crete, Greece
e-mail: mtsagris@uoc.gr

S. Fafalios
Gnosis Data Analysis, Herakleion, Crete, Greece

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,
https://doi.org/10.1007/978-3-030-85254-2_29

More recently, Busse et al. (2013) compared the relationship between prices of gasoline to prices of used and new cars, and Gegic etal. (2019) predicted the discretized, into mutually exclusive classes, car prices. Xia et al. (2020) predicted the car sales using a highly versatile machine learning algorithm and Alberini et al. (2016) conducted, in the Swiss market, an analysis that resembles to some extent the current analysis. Aiming at examining whether fuel economy is capitalized in the car price, they linked price to car characteristics, collecting panel data but without considering key features, such as brands.

Wu et al. (2009) proposed an expert system to forecast the price of used cars using an adaptive neuro-fuzzy inference system. Lessmann and Voss (2017) empirically investigated numerous linear and nonlinear statistical models for forecasting the resale prices of used cars. Andrews and Benzing (2007) analysed the influence of auction, seller and product on the price premium in an eBay used car auction market. On a different route, Raviv (2006) examined the sequence of winning bids in the public auction of used cars in New Jersey providing evidence of an order-dependent increase in the price.

The current paper combines the concepts of the more recent work and attempts to extend it by considering the Greek car retail market in December 2020. Its scope is oriented towards the prediction of the prices of brand new cars. Specifically, using information from the characteristics of new cars the goal of the paper is to accurately model and predict the car prices. This information is mainly of importance for consumers who want to know the impact of each car characteristic on its price and whether additional characteristics and associated costs translate to true consumer value. To this end, a selection of the important car characteristics affecting its pricing is initially performed. Using the selected characteristics, statistical and machine learning algorithms yield interesting conclusions regarding the effect of those characteristics on the prices. Machine learning algorithms proved useful in terms of predictive performance while further examination of the final model's predictability pointed out the consumer benefits by providing strong evidence as to which cars are estimated to be over or under-priced.

The rest of the paper is organised as follows. The data analysed are described followed by a delineation of the models and algorithms whose predictive capability is assessed in multiple directions. Finally, conclusions close the paper.

2 Data Description

Cross-sectional data on brand new cars were accessed from the popular Greek car site autotriti in December 2020 covering a total of 1,600 brand new cars on 39 characteristics (price, horsepower, engine displacement, fuel type, time to 100 km/h, fuel consumption, etc.). Missing data information mandated a pre-processing prior to the analysis.

2.1 Data Cleaning Process

The car prices ranged from €9,100 up to €283,760 with mean and median values equal to €36,796 and €27,680 respectively. To safeguard against the high variability and in order to make safer predictions, only cars priced under €50,000 were selected. According to the World Bank the estimated Greek GDP per capita for 2019 was $19,582 justifying the choice of our selection.

A further investigation emerged the necessity to remove more cars. Three additional cars in the current dataset operating with LPG were excluded. This initial "cleaning" process divulged that Alfa Romeo should participate with only 9 car models, Mitsubishi with 9 models, and Land Rover and Lexus with 1 model each. Since these brands had less than 10 models and hence carry little information they were removed from further analysis. The reason being is the tenfold cross-validation protocol, where the split of the data took place in a stratified manner ensuring that each fold contained all brands. These actions along with the removal of cars with missing information on their characteristics ensured that 909 cars (models) would participate in the analysis, a number high enough to perform valid statistical inference and draw reliable conclusions. The 39 car characteristics of these 909 cars appear in Table 1, while the type of fuel and category appear on the contingency Table 2.

2.2 Brands and Car Prices

The ranges of the car prices grouped by the brand are visualised in Fig. 1. BMW, Subaru, Mercedes and Volvo sell the most expensive cars, all above €20,000, whereas Dacia, Fiat and Suzuki are the lowest priced cars, with no car over €30,000. The two most expensive cars belong to BMW, both valued more than €49,000, whereas the lowest-priced car is manufactured by Seat, valued €9,100. It is worth highlighting that Dacia and Fiat trade cars are also priced below €10,000. Nissan, Mazda, Audi and Honda produce cars at a wide range of prices, whereas the range of prices of Dacia, Fiat and Citroen is relatively small, compared to the other brands. Evidently, the car prices differ significantly across the brands and no statistics are necessary to validate this.

According to the overall number of sales in the Greek market,[1] Toyota is the most popular brand, whereas Subaru is the least popular brand. Figure 1 manifested that Toyota and Peugeot sell medium-priced cars, yet these two brands hold the highest number of sales. Fiat and Dacia on the other hand sell the most economic cars, yet they acquired the 15th and 16th position in the number of sales, respectively. The Spearman correlation between the brands ranked according to their median car prices and ranked according to their number of sales is equal to -0.291. This manifests a

[1] Information accesses through autotriti.

Table 1 Car characteristics

Characteristic	Mean	Median	Minimum	Maximum	Characteristic	Yes	No
Engine (cc)	1,454	1,496	898.00	2,494	Air-condition	293	616
Time to 100 km/h (s)	10.41	10.30	5.30	17.10	Clima	649	260
Consumption (L/100 km)	5.00	5.00	3.00	8.10	Rear electric windows	784	125
CO_2 emissions (g/km)	118.20	116.00	76.00	189.00	Fog lights	795	114
Reservoir autonomy (km)	1,020	980.00	593.00	1,842	Parking sensors	645	264
Taxation (€)	138.40	114.00	0.00	525.00	Rain sensors	572	337
Length (mm)	4,311	4,363	3,466	4,871	Xenon lights	225	684
Width (mm)	1,798	1,800	1,595	1,969	Cruise control	757	152
Height (mm)	1,529	1,495	1,353	1,801	Leather lining	207	702
Distance between wheels (mm)	2,634	2,649	2,300	2,920	Light alloy wheels	768	141
Port baggage size (L)	424.60	400.00	170.00	780.00	Sunroof	217	692
Fuel tank size (L)	49.71	50.00	32.00	70.00	Navigator	383	526
Weight (kg)	1,314	1,325	835.00	1,836	Manual gear box	684	225
Horsepower	132.70	122.00	60.00	1,603			
Rounds/m at maximum hp	4,872	5,000	400.00	6,600			
Torque (Nm)	230.50	240.00	91.00	445.00			
Rounds/m at max (Nm)	2,099	1,750	0.00	5,000			
Cylinders	3.67	4.00	3.00	4.00			
Maximum speed	194.10	193.00	150.00	250.00			
Guarantee in mechanics (years)	4.34	5.00	2.00	8.00			
Guarantee in rust (years)	12.18	12.00	6.00	30.00			
Guarantee in colour (years)	2.97	3.00	2.00	5.00			

Table 2 Car category and type of fuel

	Fuel			
Category	Diesel	Gasoline	Hybrid	Row totals
Big	14	16	1	31
Small-medium	63	96	17	176
Medium	31	37	0	68
Mini	2	44	8	54
Off-road	136	231	36	403
Polymorph	16	17	0	33
Small	24	114	6	144
Column totals	286	555	68	909

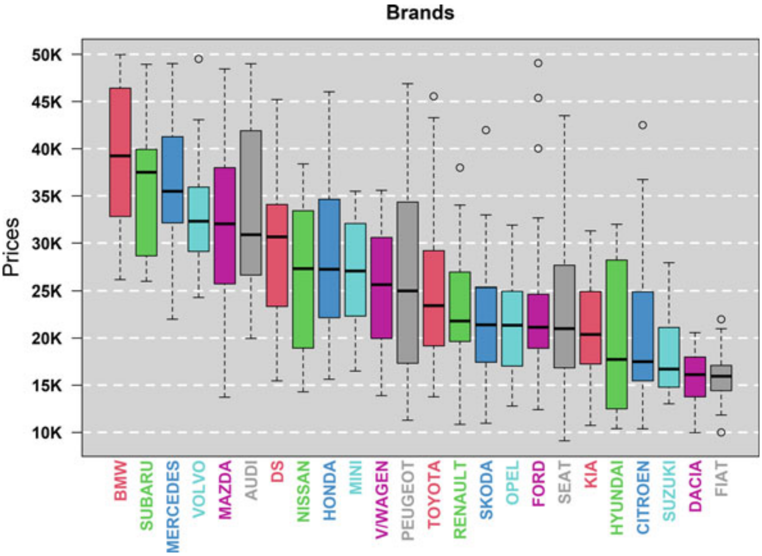


Fig. 1 Box plot of the car prices across brands ordered according to their median values

negative correlation, which is however non-statistically significant at the 5% level (p-value = 0.260) and implies that brand car prices and the number of sales seem not to be statistically significantly associated.

3 Data Analysis

The available 39 car characteristics serve as candidate predictor variables for the logarithm of car prices (y). The logarithmic transformation was chosen due to the right skewness of the price distribution.

3.1 Selection of the Important Car Characteristics

The task of selecting the important predictor variables (car characteristics) operated under the Forward Backward with Early Dropping (FBED) algorithm[2] (Borboudakis and Tsamardinos 2019). In brief, the algorithm proceeds in a forward manner, while dropping the non-significant predictor variables at each step, attempting to identify the predictor variables that are statistically significantly associated (at the 5% significance level) with the response variable. Upon completion of the forward search, a backward search is applied to remove any falsely selected variables. The FBED algorithm detected 18 car characteristics that are statistically significantly associated with the logarithm of the price.

3.2 The Log-Normal Regression Model

The logarithm of the car price implies that a log-normal regression was fitted with the relevant density of the log-normal distribution given by

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right), \tag{1}$$

where μ and σ^2 refer to the mean and variance parameters of the underlying normal distribution, respectively. Hence, the regression model is of the form $E\left(\ln(y)|\mathbf{X}\right) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ and $\boldsymbol{\beta}$ define the design matrix of the predictor variables and the vector of coefficients, respectively. This model signifies that fitted and predicted car prices based upon the log-normal regression model, when back-transformed to Euros, are equal to $\exp\left(\hat{y} + 0.5\hat{\sigma}^2_{y|\mathbf{x}}\right)$, where $\hat{\sigma}^2_{y|\mathbf{x}}$ is the estimated regression variance, since the mean of the log-normal distribution is $E(y) = \exp\left(\mu + 0.5\sigma^2\right)$. The logarithmic transformation was not applied to the continuous car characteristics despite their units of measurement being positive so as to keep their effects more interpretable. No interactions among the predictor variables were added either, as this would surge the number of estimated parameters.

According to the coefficients of the log-normal regression (not shown here), there is a mixture of car characteristics affecting its pricing, both mechanical and image related. The interpretation of those coefficients is straightforward; each of them refers to the expected percentage-wise price change for a given unit change in the values of each car characteristic, ceteris paribus. Overall, the coefficients possess the correct sign and their magnitude was also justified by univariate analyses. In terms of model fit, the log-normal model explains the 94.00%[3] of the logarithm of the price variability, providing good evidence of a highly acceptable model fit.

[2] FBED is publicly available in the *R* package *MXM* (Tsagris and Tsamardinos 2019).

[3] The adjusted coefficient of determination is equal to 93.68%.

However, there are three issues that should be considered. The reported (adjusted) coefficient of determination is not a valid prediction evaluation criterion. Secondly, the influence of the characteristics on the car's price might be far from linear and subsequently the prediction error of the log-normal regression model is not the lowest that can be achieved.

3.3 Car Price Prediction

The model's coefficient of determination is substantially high and despite revealing a very satisfactory fit, it cannot provide information on the model's predictability as it was computed on the same data the model was fitted, and hence it overestimates the model's true predictive performance.

A better strategy is to apply the tenfold cross-validation (CV) pipeline. This commences with splitting the dataset into ten mutually exclusive folds or sets in a stratified manner. The cars are randomly assigned to each fold in a stratified manner so that the distribution of the brands is nearly the same in all folds and hence each brand will be represented in each fold. One fold is left aside playing the role of the test set, while the other nine folds are collected in what is termed the training set. In the training set, the FBED algorithm selects the most important variables utilising the log-normal regression model. A predictive model is subsequently built and validated on the test set, i.e. using the values of the selected car characteristics in the test set, the car prices are predicted. Three metrics evaluating the predictive performance were estimated during the tenfold CV procedure. The percentage of variance explained (PVE), the mean absolute error[4] (MAE) and the mean error (ME) per brand are defined as

$$\text{PVE} = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} \tag{2a}$$

$$\text{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{2b}$$

$$\text{ME} = \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)}{N} \tag{2c}$$

respectively, where $\hat{y}_i$ refers to the predicted values of the test set whose sample size is equal to N. The PVE can be interpreted as the out of sample coefficient of determination. MAE, on the other hand, is easier to interpret and states that, on average, the predicted car price may deviate by plus or minus a value. Lastly, the ME shows the average direction (or sign) of the errors.

This process is repeated ten times so that all folds have played the role of the test set. Due to inherent variability in the results, the pre-described tenfold CV pipeline

[4] MAE serves the purpose of practical interpretation.

is repeated ten times with different generated folds each time[5] and the capability of the predictive model stems from the aggregation of all folds across all repetitions.

3.3.1 Machine Learning Predictive Algorithms

Highly versatile machine learning algorithms, such as random forest (RF) and support vector regression (SVR) were employed to assist in obtaining more accurate predictions. The algorithms' perk is the exploitation of the nonlinear functional relationship between the car characteristics and its price which can result in more accurate predictions. The RF algorithm is built upon creating numerous regression trees, justifying its name. RF relies on "bagging" (bootstrap aggregation) (Breiman 2001) and random selection of features (Ho 1995). The algorithm randomly draws a subset of variables with a bootstrap sample,[6] termed $\mathbf{X}_b$ and Y_b, and builds a tree using this subset. The tree discretizes the continuous variables into classes seeking for the optimal split, with the number of splits being a hyper-parameter that requires tuning. The process of randomly selecting variables and bootstrap samples is repeated B times with the predictions being computed over aggregation of all tree-based predictions $\hat{y} = \frac{\sum_{b=1}^{B} f(\mathbf{x}_b)}{B}$.

SVR is more complex and relies upon the following constrained minimization as described in Meyer et al. (2020)

$$\begin{array}{ll} \min_{\mathbf{a},\mathbf{a}^*} & \frac{1}{2}\left(\mathbf{a}-\mathbf{a}^*\right)^T \mathbf{Q}\left(\mathbf{a}-\mathbf{a}^*\right) + \epsilon \sum_{i=1}^{n}\left(a_i + a_i^*\right) + \sum_{i=1}^{n} y_i \left(a_i + a_i^*\right) \\ \text{s.t.} & 0 \leq a_i, a_i^* \leq C, \quad i = 1, \ldots, n \\ & \sum_{i=1}^{n}\left(a_i - a_i^*\right) = 0, \end{array}$$

where $\mathbf{a}$, $\mathbf{a}^*$ are the vector of parameters to be estimated, ϵ is a very small quantity, C is the cost, a tunable hyper-parameter and $\mathbf{Q}$ is an $n \times n$ positive semidefinite matrix with elements $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, where $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi\left(\mathbf{x}_j\right)$ denotes the kernel matrix. The radial basis function $\exp\left(-\gamma |\mathbf{u}-\mathbf{v}|^2\right)$ was selected as the kernel with γ being a tunable hyper-parameter.

The number of splits of the variables examined in the RF was $(1, 3, 5, 10)$. The cost hyper-parameter in SVR laid hold of ten equidistant values spanning from 0.2 to 2, while the γ parameter also took ten values, equally spread between $1/d^2$ and $1/d^{0.5}$, where d denotes the number of variables. Within the CV protocol, the predicted car prices were now based upon the RF with 4 hyper-parameter values (four splits), and with SVR using all combinations of the cost and γ, in the car characteristics that were selected by FBED. This results in four sets of predicted car prices for the RF and 100 sets for the SVR.

[5] This avoids "lucky" splits that could yield a high predictive performance.

[6] Sample with replacement, of the same size.

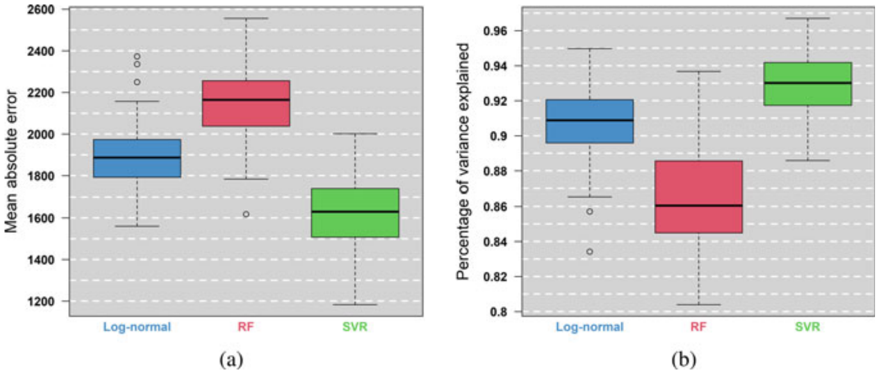


Fig. 2 Performance metrics of the CV protocol: **a** MAE expressed in € and **b** PVE

3.3.2 Predictive Performance Evaluation Results

The results of the ten times repeated tenfold CV appear in Fig. 2. The RF algorithm yielded the least accurate results, followed by the log-normal regression, while SVR produced the optimal predictions. The estimated PVE (2a) and MAE (2b) values of SVR equal 92.86% and €1,622.19 respectively. MAE, on the other hand, is easier to interpret and states that, on average, the model's predictions of a car price may deviate by plus or minus €1,622.19. This error is satisfactorily small relative to the range of car prices as it corresponds to 3.97% of the observed range of prices (€9,100–€49,983).

Figure 3 visualises the predictive performance of the SVR. In Fig. 3a MAE is classified according to the brand. MAE differs significantly across the brands and surprisingly enough, the error is high in the cheaper cars, Suzuki and Fiat. However, the highest error was observed for Mazda cars which does come by surprise as this brand produces the fifth most expensive cars (see Fig. 1). What is not surprising though is that the eight most expensive cars are the ones with an MAE more than €2000.

The error increases as the prices increase,[7] as observed in Fig. 3b, where MAE is classified according to intervals of €10,000. This is also evident in the scatter plot of Fig. 3c that visually contrasts the predicted prices against the observed prices, with the blue line corresponding to the 45° line, or perfect agreement. The higher the prices, the higher the spread of their predictions around the blue line, yet the correlation between these two is really high and equal to 0.968.

[7] The majority of the cars (71%) are priced less than €30,000 explaining the overall MAE of €1,622.19.

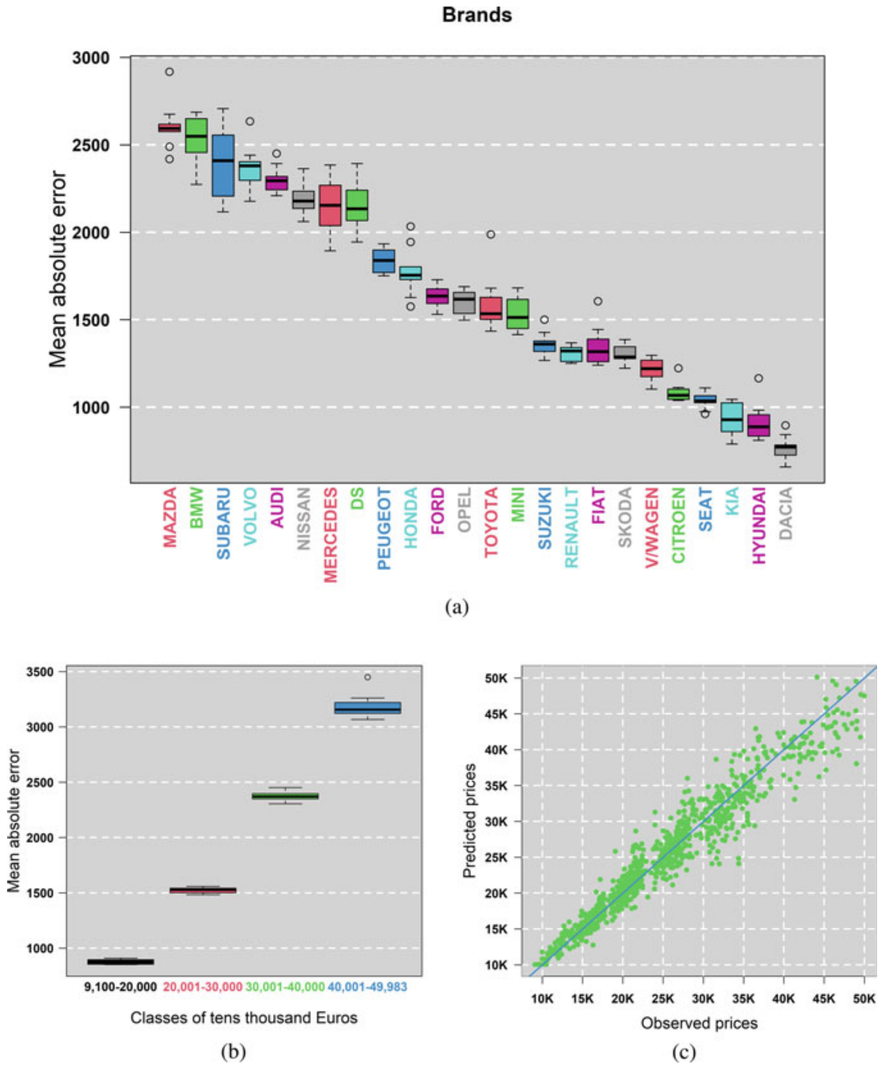


Fig. 3 Performance metrics of the CV protocol. **a** MAE according to brand, **b** MAE in classes of ten thousand € and **c** observed versus predicted values

3.3.3 Estimated over and Under-Priced Cars

From the consumer's point of view it is also worthy to characterize a car as being over or under-priced, based on the predictions of the cross-validation.[8] Figure 4 displays

[8] To be fair when assessing the over/under-pricing of a car, we had to use the predicted prices and not the fitted prices.

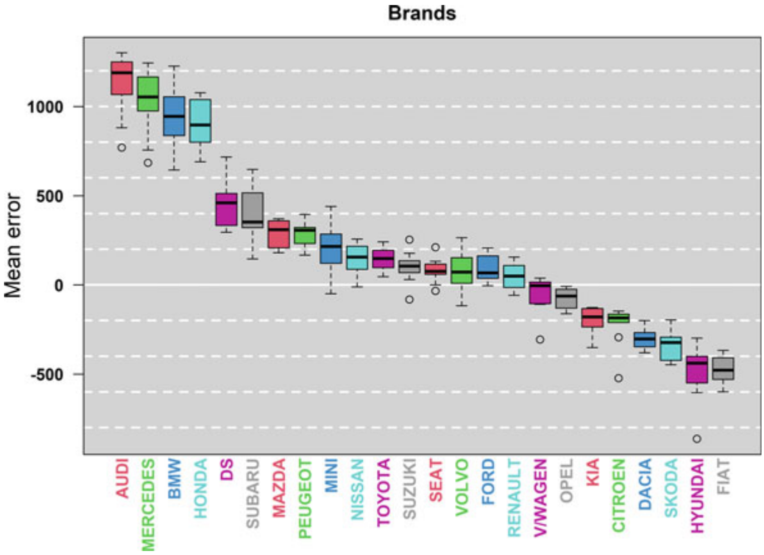


Fig. 4 Performance metrics of the CV protocol II. ME per brand

ME, an informative measure that sheds light into the problem of detecting over and under-priced cars. The boxes of brands located above the zero vertical line indicate brands estimated to be over-priced, whereas boxes of brands located below indicate brands classified as under-priced. Seat is estimated to produce the most over-priced cars, whereas Fiat is estimated to produce the most under-priced cars. Reputation, that is linked to reliability, could be the causal factor attributing to this phenomenon. A suitable proxy for this variable could be the reliability rating index. Even though this index exists, it ought to be country-specific and unfortunately, no values exist for the Greek market.

The largest negative difference between the actual and estimated prices was €11,342.89 observed at an Audi brand. and the largest positive difference was equal to €7,499.14, observed at a BMW branded car. The predicted price of the Audi brand car priced at €49,010 was €37,667.11. By examining the characteristics of cars whose true prices range within a €500 range it is obvious that this is an over priced car. There are 8 cars at a similar range of prices offering the same or better characteristics with 5 of them being produced by the most expensive brands (Volvo, Mazda, Subaru). This implies that a consumer who desires to purchase an Audi car has more affordable options from similar level brands, with similar characteristics. On the other hand, a BMW brand car priced at €31,594 was predicted to be worth €39,093.14. When considering cars priced within a window of €500 far from the predicted value, it is apparent that the characteristics of three cars (Subaru, Toyota and BMW) are similar to that of this BMW car, but at different brands. The referred car, given its high-class brand, can be seen as a value-for-money car. Thus, in this instance, comparing prestigious cars alone, a BMW and a Subaru cars are more expensive than this BMW car by more than €7,000, despite all three cars having similar characteristics.

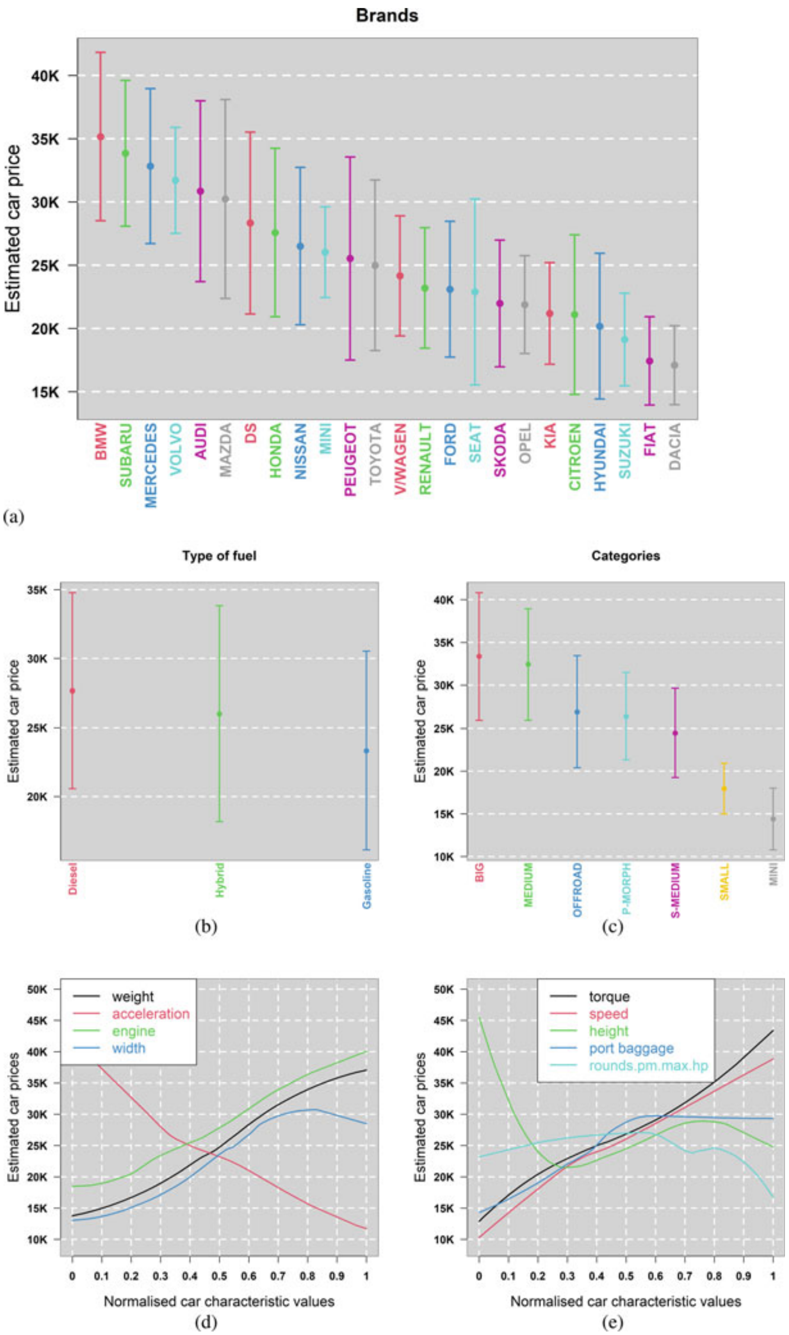
3.4 Estimated Individual Effect of the Car Characteristics

Since SVR does not return coefficients demonstrating the effect of each car characteristic on the price, the individual conditional expectation (ICE) plots (Goldstein et al. 2015) will portray these effects visually. The advantage of these plots is the visualisation of the nonlinear effect of the independent variables on the response variable. In this case study though a bootstrap variant of ICE plots has been implemented, within this bootstrap variant frame a car characteristic is chosen and its values are sampled with replacement. The optimal SVR, corresponding to the hyperparameters that yielded the optimal predictive performance, is fitted. This process is repeated 100 times and the average estimated prices are computed. For the continuous car characteristics the ICE is estimated using a locally-weighted polynomial regression.[9]

Figure 5 shows the effect of each characteristic on the estimated car price. Specifically Fig. 5a–c demonstrate the car brand, fuel and category effect on the estimated prices. The effect of the brand is sorted in descending order. This order is in an almost perfect agreement with the true order of the brands sorted according to their average prices. The only discrepancy is that SVR estimates Ford to be more expensive than Seat, whereas the opposite is true in this sample. As for the fuel the order is as expected, with diesel cars being the most expensive followed by hybrid and gasoline cars. Finally, for the car category the estimated order is distorted only between off-road and polymorphic cars. These plots evidently signify that SVR has managed to detect the correct effect of these categorical valued car characteristics. It must be highlighted that the estimated effects based on the log-normal regression coefficients were not as accurate as the SVR estimated effects.

Figure 5d, e visualises the effect of the continuous car characteristics. Since these characteristics are measured in different units, they were first normalised to be mapped on the same scale using $\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$, where $i = 1, \ldots, n$, with $n = 909$ cars. The car weight (in kg), acceleration (number of seconds the car requires to reach a speed of 100 km/h), engine displacement (in cc^3) and width (in mm) are plotted in Fig. 5d. As expected, the acceleration has a negative impact on the car price as the longer the time required for a car to reach the 100 km/h speed, the less expensive it is. The heavier and the bigger (in terms of engine) the car is, the more expensive it is. Car width is positively associated with its price up to a certain point, above which the car width influences price in a negative manner. The car torque (in Nm) and its maximum speed are two characteristics positively associated with the price as depicted in Fig. 5e. On the same graph, it is observed that the size of the port baggage

[9] For the categorical car characteristics, this kernel regression step is omitted.



(a)

(b)

(c)

(d)

(e)

Fig. 5 Effect of each car characteristic on the estimated car price

(in litres) has a partial positive effect up to a certain threshold, after which it becomes flat and hence does not affect the price. The effect of the height is highly nonlinear and a possible explanation could be that shorter cars are perhaps convertibles or sportive. As the height increases the car becomes sedan, hatchback or coupe style, so usually less expensive. A higher car indicates an SUV type or Jeep type which are more expensive. Rounds per minute at the maximum horsepower does not seem to be associated with the price. Note that these are individual effects, so plotting these in higher dimensions could reveal more interesting patterns with regards to their combined effects.

4 Conclusions

A variable selection algorithm identified significant associations between car characteristics and pricing in the Greek market. Among the 18 identified characteristics, the most important were the car's weight, brand, time to reach 100 km/h, category, torque, type of fuel and maximum speed. It is natural to assert that the effect of the identified car characteristics differs across brands. Allowing for interactions between car brands and the rest of the variables would relax the rather restrictive assumption imposed on the slopes of the continuous variables. However, the addition of such interaction terms would increase dramatically the number of estimated parameters and negatively affect the validity of the estimates. Such an approach would require either larger car samples or the missing information for all cars to be available on the autotriti's website.

Although the model fit with the use of a log-normal distribution was excellent, the predictive ability of the model was sub-optimal. This discrepancy could be due to nonlinear underlining associations between car characteristics and price, an association better captured by using machine learning techniques. Such methods produced models with greater predictive ability than the regression models at the expense of interpretability. The SVR predicted the car prices with a deviation of nearly €1,622 which can serve as a guideline for consumers.

Price prediction that is higher than the actual price indicates evidence of a value-for-money car or evidence of an interesting purchase. On the other hand, a lower than the actual predicted price signals a rather over-priced car whose purchase might not be for the consumer's best of interest. Two extremely over and under-priced cars exhibited this phenomenon. Note, however, that the characterization as over or under-priced relies upon the available data and it could also be attributed to chance as these are estimates. Moreover, there are unobserved factors contributing to the car price, such as research and development costs and marketing/advertisement expenses that are not publicly available. Brand reputation and reliability were not measured either and the records of each brand regarding mechanical faults observed after the purchase cannot be undisclosed to the public.

Collectively, all the aforementioned results are beneficial for consumers and can act as a guide for choosing a value-for-money car. The results further signified that not all machine learning algorithms do necessarily outperform statistical models, yet, some of them can.

If documentation of the production cost of each car was available, application of a stochastic frontier model (Battese and Coelli 1995) would further evaluate the profit efficiency of the car companies. This would enable companies to better orient their strategies and make the whole vehicle market more efficient.

Acknowledgements The authors are grateful to Nikolaos Pandis, Eleftherios Pavlos and George Filis who read earlier versions of this manuscript and provided constructive feedback.

References

Alberini A, Bareit M, Filippini M (2016) What is the effect of fuel efficiency information on car prices? Evidence from Switzerland. Energy J 37

Andrews T, Benzing C (2007) The determinants of price in internet auctions of used cars. Atl Econ J 35:43–57

Battese GE, Coelli TJ (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empir Econ 20(2):325–332

Borboudakis G, Tsamardinos I (2019) Forward-backward selection with early dropping. J Mach Learn Res 20:1–39

Breiman L (2001) Random forests. Mach Learn 45:5–32

Busse MR, Knittel CR, Zettelmeyer F (2013) Are consumers myopic? Evidence from new and used car purchases. Am Econ Rev 103:220–56

Eckard EW Jr (1985) The effects of state automobile dealer entry regulation on new car prices. Econ Inq 23:223–242

Gegic E, Isakovic B, Keco D, Masetic Z, Kevric J (2019) Car price prediction using machine learning techniques. TEM J 8:113–118

Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24:44–65

Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol 1, pp 278–282

Lessmann S, Voss S (2017) Car resale price forecasting: the impact of regression method, private information, and heterogeneity on forecast accuracy. Int J Forecast 33:864–877

Matas A, Raymond JL (2009) Hedonic prices for cars: an application to the Spanish car market, 1981–2005. Appl Econ 41:2887–2904

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2020) e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-4

Raviv Y (2006) New evidence on price anomalies in sequential auctions: used cars in New Jersey. J Bus Econ Stat 24:301–312

Tsagris M, Tsamardinos I (2019) Feature selection with the R package MXM. F1000Research 7:1505

Verboven F (1996) International price discrimination in the European car market. RAND J Econ 27:240–268

Wu JD, Hsu CC, Chen HC (2009) An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Syst Appl 36:7809–7817

Xia Z, Xue S, Wu L, Sun J, Chen Y, Zhang R (2020) ForeXGBoost: passenger car sales prediction based on XGBoost. Distrib Parallel Databases 38:713–738