Evaluation of Used Car Value using Advanced Regression Techniques

Alex Leung
ac6leung@uwaterloo.ca

August 10, 2020

1 Introduction

How much would you pay for a used car? The answer to this question likely depends on the mileage on the car, the age of the car, and many other factors. Furthermore, the answer to this question is likely of great interest to both sellers and buyers in the used car market, in order to determine a fair price for any given car, as well as how to potentially increase the value of the car.

In this paper, we attempt to model the price of a used car based on a number of features using various advanced regression techniques. This paper was inspired by the work by Del Giudice et al. (2015), who modeled the market prices of a number of real estate properties within a region of Italy as a function of the properties' features using a semiparametric additive model.

The remainder of this paper is structured as follows: In section 2, we describe and explore the dataset on which we will perform the analysis. In section 3, we outline the techniques which will be used to model the data. We draw comparisons between several models and outline our statistical observations. In section 4, we discuss our observations in the context of the used car market. In section 5, we outline directions for future analysis in this area.

2 Data

The data that will be used is the "Used car catalog" by Kirill Lepchenkov on Kaggle. The dataset contains data on 38,531 used car listings from various web listings; according to the author, the data was collected from various web sources in Belarus on December 2nd, 2019. We have kept only 14 of the original 30 variates provided in the raw data before beginning our analysis. Reasons for removing the other 16 variates can be found in Appendix 1.

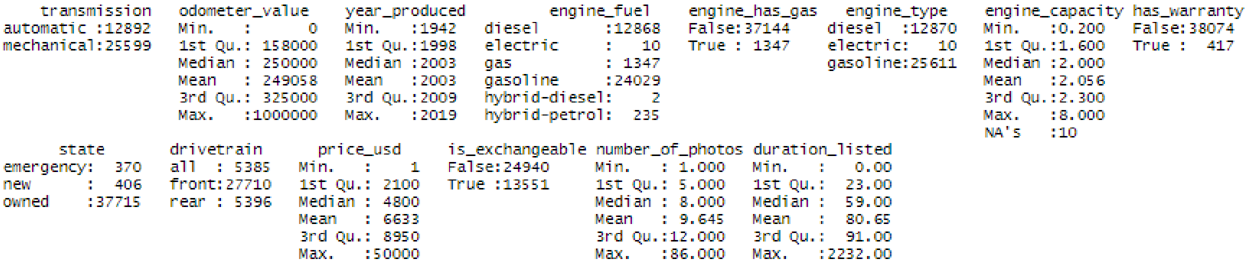
Name	Description	Name	Description
transmission	Transmission type.	has_warranty	If the car is covered under warranty.
odometer_value	Odometer reading, in km.	drivetrain	Front/rear/all wheel drive.
year_produced	The year the car was produced.	price_usd	Listed price, in USD.
engine_fuel	Fuel type.	is_exchangeable	If the seller is willing to exchange for another vehicle.
engine_type	Engine type.	number_of_photos	Number of photos posted with the ad.
engine_capacity	Engine capacity, in litres.	duration_listed	Number the days the posting has been up.
state	New/Owned/Emergency (severely damaged).	engine_has_gas	If the car uses propane.

Table 1: Names and brief descriptions of variates in the dataset.

2.1 Exploratory Data Analysis and Preprocessing

Before beginning our analysis, we consider whether the dataset might contain duplicates; since the data was scraped from different web sources, it might be the case that some sellers have posted their vehicles on different websites, thus creating duplicate observations for the same vehicle. As such, we can remedy this by removing any observed duplicates in our data. This reduces the dataset by only 40 observations (to 38,491 observations).

We begin our exploration of the data by calling a summary on the dataset using R, which yields the following output:



```
    transmission   odometer_value    year_produced          engine_fuel    engine_has_gas   engine_type    engine_capacity has_warranty
automatic :12892   Min.   :      0   Min.   :1942   diesel       :12868   False:37144    diesel  :12870   Min.   :0.200   False:38074
mechanical:25599   1st Qu.: 158000   1st Qu.:1998   electric     :   10   True : 1347    electric:   10   1st Qu.:1.600   True :  417
                   Median : 250000   Median :2003   gas          : 1347                  gasoline:25611   Median :2.000
                   Mean   : 249058   Mean   :2003   gasoline     :24029                                   Mean   :2.056
                   3rd Qu.: 325000   3rd Qu.:2009   hybrid-diesel:    2                                   3rd Qu.:2.300
                   Max.   :1000000   Max.   :2019   hybrid-petrol:  235                                   Max.   :8.000
                                                                                                          NA's   :10
      state         drivetrain       price_usd      is_exchangeable number_of_photos duration_listed
emergency:  370   all  : 5385   Min.   :    1   False:24940     Min.   : 1.000   Min.   :   0.00
new      :  406   front:27710   1st Qu.: 2100   True :13551     1st Qu.: 5.000   1st Qu.:  23.00
owned    :37715   rear : 5396   Median : 4800                   Median : 8.000   Median :  59.00
                                Mean   : 6633                   Mean   : 9.645   Mean   :  80.65
                                3rd Qu.: 8950                   3rd Qu.:12.000   3rd Qu.:  91.00
                                Max.   :50000                   Max.   :86.000   Max.   :2232.00
```

We believe that a model with the variate `age`, calculated as 2019 - `year_produced` would be more interpretable and more useful to our regression analysis; and as such, we perform this transformation.

Next, we make some concerning observations about the dataset. The minimum value for `odometer_value` is 0; since this is a used car dataset, this is quite dubious. It is likely that 0 is simply a placeholder for missing values. Additionally, there are many other suspicious values for `odometer_value` such as "1111111" or "999999" which lead us to believe that many reported values in this column are not reliable. Later analysis reveals that this column is highly correlated with `age` (correlation coefficient 0.48), which impedes the regression analysis as the necessary design matrices are not of full rank (so we drop the `odometer_value` altogether, avoiding these issues).

The variate `state` is a categorical variate which identifies a car as either new, owned (used), or emergency (severely damaged). Since we are only interested in used cars, we should remove all cars which `state` is not "owned", and then drop this variate entirely.

The variate `engine_fuel` appears to be a categorical variate with 6 possible values. A large majority (over 95%) of the observations fall under the "diesel" or "gasoline" categories. We remove all the other observations that do not fall into these two categories, because (1) reduction to a binary variate is convenient for analysis, and (2) this does not limit the contextual analysis of the problem, given the large majority of the market is made up by these two categories. In doing this, we also eliminate all the entries in which `engine_has_gas` is true (thus eliminating this variate from our analysis), and those where `engine_type` is "electric". This confounds the two variates `engine_fuel` and `engine_type`, so we choose to drop `engine_type`.

Thirdly, the minimum price (`price_usd`) is shown as $1. Observations with a nominal price are likely not providing the true asking price, but rather as a placeholder as a solicitation for offers. Further inspection reveals a handful of observations with nominal prices (less than $100). These observations are also removed in order to limit the extreme observations in the data.

After performing the above transformations, we are left with 36,129 observations of 9 explanatory variates; `transmission`, `engine_fuel`, `engine_capacity`, `has_warranty`, `drivetrain`, `is_exchangeable`, `number_of_photos`, `duration_listed` , `age`; and our response variate, `price_usd`. We have trimmed less than 10% of our original number of observations (based on extreme values), so our data should still be a good representation of the used car market.

We now explore the distributions of each of these variates. In Figure 1 below, we have plotted each of the variates in the dataset. We can see that most of the numerical variates follow a unimodal distribution, but have long right tails. We would like to transform this data so that the distributions of each of the variates are more symmetrical. We use simple power transformations on the variates `price_usd`, `number_of_photos`, `engine_capacity`, `duration_listed`, and `age`, all using a power of 0.25, to shift the mode of the distribution towards the right.

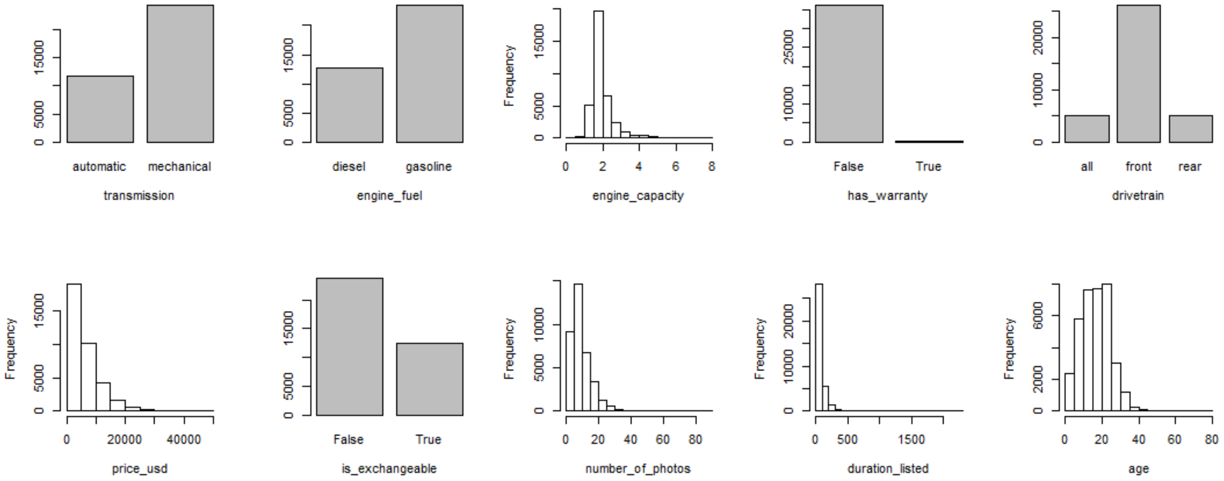


Figure 1: Distributions of each of the variates in the dataset.

After making the specified transformations, we find the distributions of the numerical variates to be much more symmetrical.

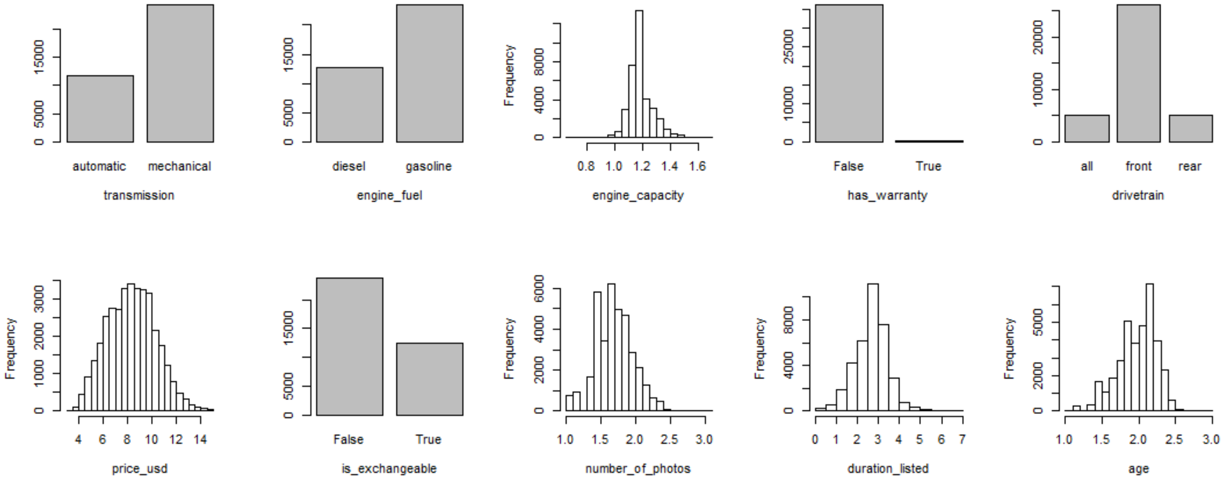


Figure 2: Distributions of each of the variates in the dataset, after applying appropriate power transformations to the numerical variates.

We can also explore the relationship between each of the explanatory variates with the response variate (price). In Figure 3 below, we have plotted the price (transformed) against each explanatory variate (transformed). Upon initial observation of these plots, we find that some explanatory variates appear to have quite strong correlations with the price of a car. Namely, there appear to be a clear positive correlation between price and `has_warranty`. There is a noticeable negative correlation between price and `age`. The variates `transmission` and `drivetrain` also appear to be correlated with price.

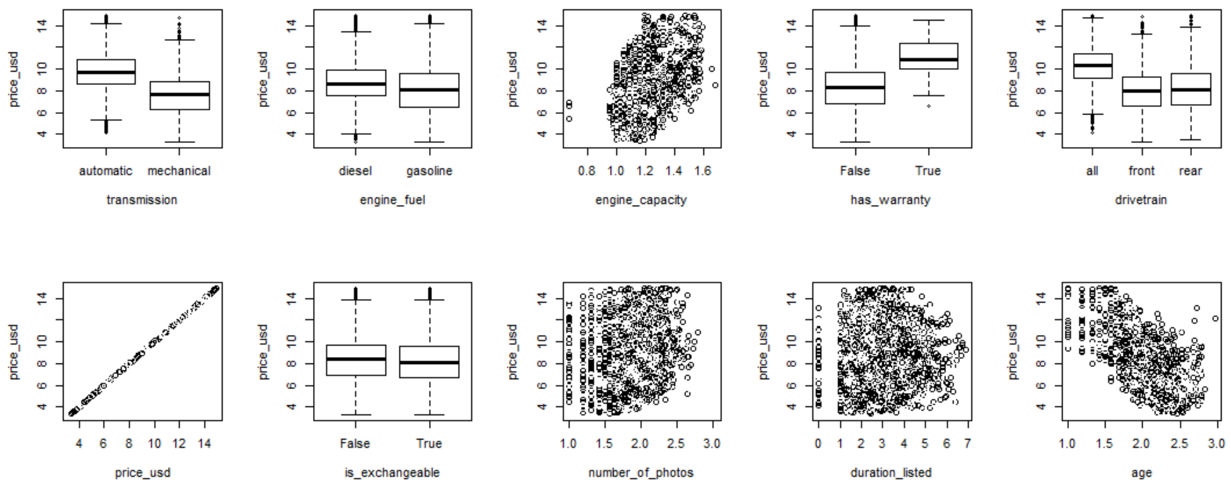


Figure 3: Plots for listing price (y) vs. each variate (x), after performing appropriate transformations.

3 Statistical Methods

In this section we will model price of a used car as a function of the explanatory variates. We will utilize two different models: (1) a mixed-effects additive model, emulating the work done by Del Giudice et al. (2015); and (2) traditional least squares linear regression. The results of these models are discussed and compared in section 3.3.

3.1 Penalized Spline as Semiparametric Mixed Model

As mentioned above, the fitting of this model will closely follow the work by Del Giudice et al. (2015). In the aforementioned paper, the price of a real estate property was modelled using a semiparametric mixed-effects model representation for penalized splines (Ruppert et al., 2003; Wand, 2003). The mixed model takes the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

In this model, $\mathbf{X}\boldsymbol{\beta}$ is the fixed effects (parametric) portion of the model, and $\mathbf{Z}\mathbf{u}$ is the random effects (non-parametric) portion of the model.

The parametric portion of the model is straight forward; this is analogous to simple linear regression where the design matrix $\mathbf{X}$ with n samples and p variates is multiplied by the coefficient vector $\boldsymbol{\beta}$.

In the non-parametric portion, the design matrix $\mathbf{Z}$ is multiplied by a random effects vector $\mathbf{u}$. If we define the design matrix $\mathbf{Z}$ as a $n \times k$ matrix of basis functions for a spline (where n is the sample size and k is the number of basis functions required); then the vector $\mathbf{u}$ becomes the coefficients for these basis functions at each knot. By placing assumptions on $\mathbf{u}$, namely, $E(\mathbf{u}) = 0$ and $\text{cov}(\mathbf{u}) = \sigma_u^2\mathbf{I}$, with $\sigma_u^2 = \frac{\sigma_\epsilon^2}{\lambda^2}$ for some $\lambda > 0$, the variance of $\mathbf{u}$ is finite and can no longer freely range from $-\infty$ to ∞. This reduces the effect of each basis function, thus smoothing the fit.

This formulation is equivalent to a penalized spline using ridge regression on the coefficients $\mathbf{u}$ to attain the fitted values given by (see Appendix 2 for derivation):

$$\hat{\mathbf{y}} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \hat{\lambda}^2\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} \mathbf{0}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \end{bmatrix},$$

and the smoothing parameter for a spline of degree d, $\hat{\lambda}$ is selected by the restricted maximum likelihood method (REML):

$$\hat{\lambda}^2 = \left(\frac{\hat{\sigma}^2_{\epsilon,REML}}{\hat{\sigma}^2_{u,REML}} \right)^d$$

Del Giudice et al. (2015) used the methodology as described above, using binary variates and discrete numerical variates with small cardinality (e.g., whether the property has had maintenance done, floor level of the property) as the linear fixed effects. Splines for the random effects part of the model, are fitted on the other numerical variates (e.g., square footage of the property) using standard power basis functions.

Our model of `price_usd` (transformed) will be linear in the transformed variates `transmission`, `engine_fuel`, `has_warranty`, `drivetrain`, and `is_exchangeable`. Splines will be fit to the transformed variates `age`, `engine_capacity`, `number_of_photos`, and `duration_listed`.

Considering the transformations we have perfomed on the data, the model is written mathematically as:

$$\begin{aligned} price\_usd^{0.25} &= \beta_0 + \beta_1 transmission + \beta_2 engine\_fuel + \beta_3 has\_warranty + \beta_4 drivetrain_{front} \\ &+ \beta_5 drivetrain_{rear} + \beta_6 is\_exchangeable + f_1(age^{0.25}) + f_2(engine\_capacity^{0.25}) \\ &+ f_3(number\_of\_photos^{0.25}) + f_4(duration\_listed^{0.25}) + \epsilon_i \end{aligned}$$

Fitting this model to the data yields the following coefficients and splines:

	coefficient	p-value
intercept	9.129	0.000
transmission	-0.195	0.000
engine_fuel	-0.390	0.000
has_warranty	-0.131	0.158
drivetrain (front)	-0.515	0.000
drivetrain (rear)	-0.188	0.000
is_exchangeable	-0.070	0.000

	effective df	lambda
f(age^0.25)	8.482	0.050
f(engine_capacity^0.25)	7.525	0.444
f(number_of_photos^0.25)	4.509	11.924
f(duration_listed^0.25)	6.583	1.363

Table 2: Results of fitting the mixed-effects model.

From the table above, we find that all the linear coefficients are highly significant, except that for `has_warranty`. In Figure 3, we observed from a bar plot that `has_warranty` was likely to have an effect on the price of the car. The reason for the low significance of this variate in our fit is likely due to the fact that there are very few cars where `has_warranty` is true, thus increasing the variance of our parameter estimate. For the fitted splines, an interesting result is that the smooth functions for `number_of_photos` and `duration_listed` are both quite linear in nature, and could possibly be modelled using a fixed effect.

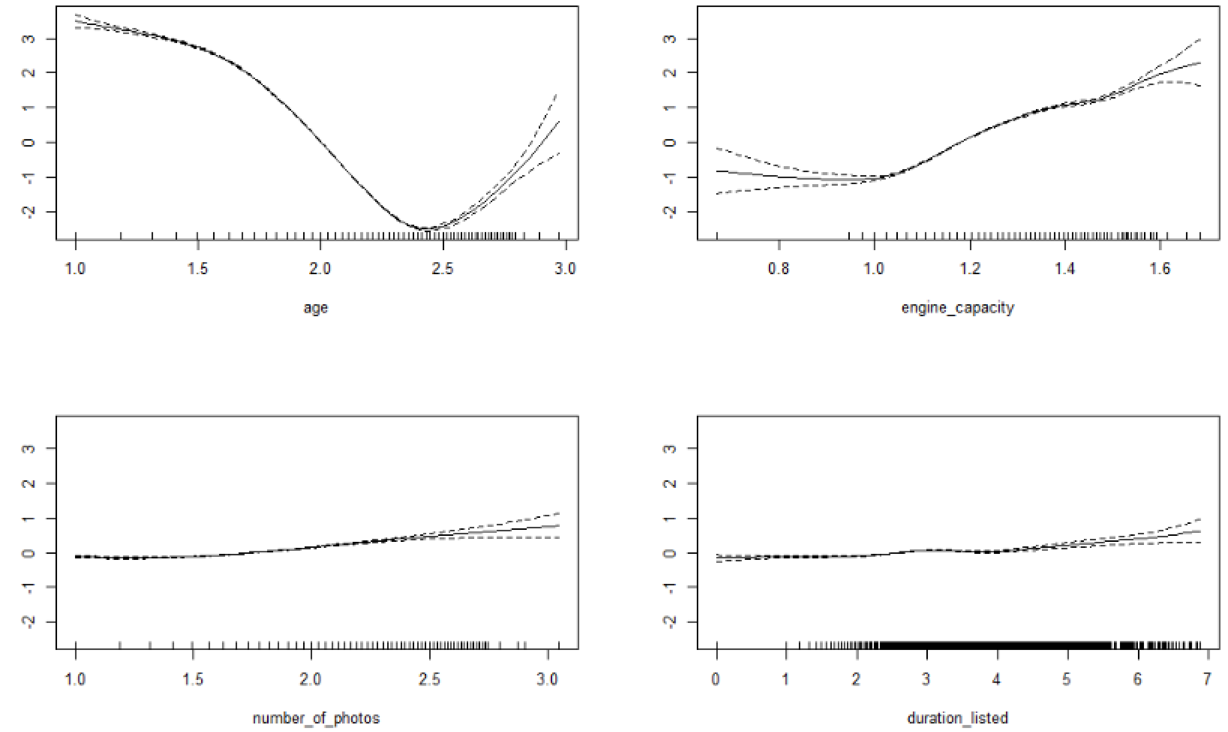


Figure 4: Non-linear effects in the mixed-effects model. Dashed lines represent standard error bands. Hash marks on the x-axes represent observations.

Next, we check this model fit by a residual plot (below). This residual plot does not raise any concerns; there are no discernable patterns, and appears to mostly satisfy the homoscedasticity assumption.

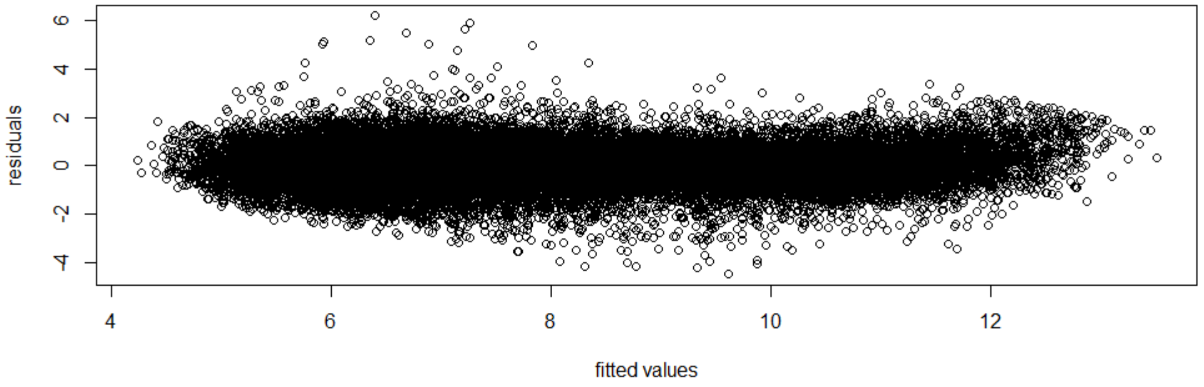


Figure 5: Residual plot under the mixed effects model.

Furthermore, this model appears to have a good fit, with an R-squared value of 0.849, and a mean absolute percentage error (MAPE) of 7.74%. However, we do note that these results cannot be directly translated over into the context of the problem, as we have performed transformations on the data. This is discussed further in section 3.3.

3.2 Traditional Least Squares Linear Regression

We fit a traditional linear model to the data, using ordinary least squares regression. This is a much less sophisticated model than that in section 3.1 above; we will compare the relative performance of each model on this dataset.

We assume a linear effect of all variates. Mathematically, we write the model as:

$$\begin{aligned} price\_usd^{0.25} &= \beta_0 + \beta_1 transmission + \beta_2 engine\_fuel + \beta_3 drivetrain_{front} \\ &+ \beta_4 drivetrain_{rear} + \beta_5 is\_exchangeable + \beta_6 (age^{0.25}) + \beta_7 (engine\_capacity^{0.25}) \\ &+ \beta_8 (number\_of\_photos^{0.25}) + \beta_9 (duration\_listed^{0.25}) + \epsilon_i \end{aligned}$$

Fitting this model to the data yields the following coefficients:

	coefficient	p-value
intercept	13.25	0
transmission	-0.32	0
engine_fuel	-0.48	0
has_warranty	-0.74	0
drivetrain (front)	-0.63	0
drivetrain (rear)	-0.27	0
is_exchangeable	-0.11	0
age^0.25	-5.79	0
engine_capacity^0.25	5.46	0
number_of_photos^0.25	0.50	0
duration_listed^0.25	0.11	0

Table 3: Coefficients of the simple linear regression model.

We check the residual plot of this model (below); most of the plot does not raise concern, except for the seemingly increased variance and skewness observed towards the left side of the plot. This is a relatively small concern, and preserving the simplicity of the model outweighs the need to correct this issue.

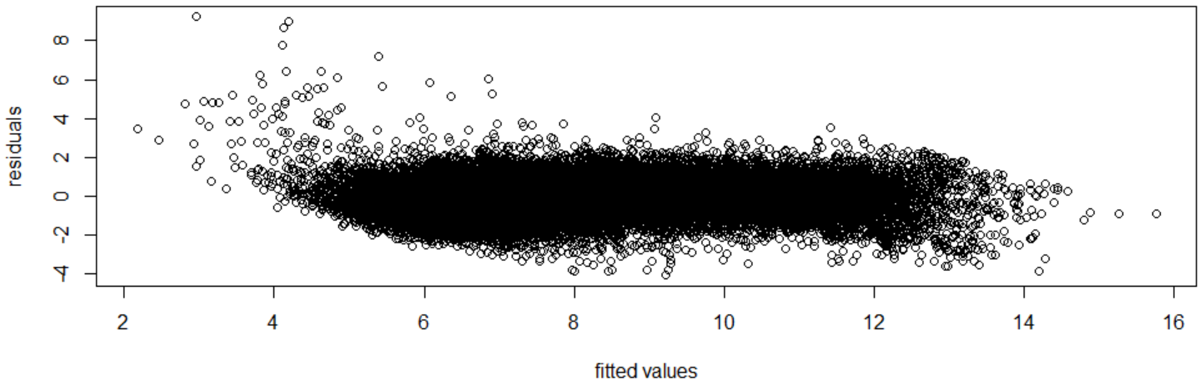


Figure 6: Residual plot under the simple linear regression model.

This model achieves an R-squared value of 0.819, and a MAPE of 8.41%.

3.3 Model Comparison and Statistical Conclusions

In this section we compare the two models specified in sections 3.1 and 3.2. We can begin by comparing the predictive results of the model, summarized in the below table.

Model	R squared	MAPE	Raw MAPE
Mixed Model	0.849	7.74%	34.74%
OLS	0.819	8.41%	38.13%

Table 4: Summary results of the two fitted models on the dataset.

If we use these two measures (R-squared and MAPE) as measures of goodness-of-fit, then we say that the mixed model provides a slightly better fit than the traditional linear model. There is very little difference between these two models, by these two measures.

Recall that the MAPE values are calculated on the transformed basis; where the response variate, price, was transformed using a power transform with a power of 0.25. If we transform these results back into absolute terms[1] (prediction error of actual price), we achieve the results in the third column (raw MAPE). Unfortunately, in the context of the actual original problem (estimation of the price of a used car), this model is not too accurate, as the estimated price will be off by over 30% on average. However, even in the context of the question (estimation of the price of a used car), the two models have a relatively small difference in raw MAPE of only 3.59%. However, the small disparity between the goodness-of-fit of the two models does not match up with the large difference in model complexity.

The main difference between these two models, put simply, is how the numerical variates are handled; in the mixed model, penalized smoothing splines are fit to the numerical variates, while the OLS model assumes simple linear relationships. An analysis of deviance performed between both models tells us that the non-linearity of the mixed model is statistically significant with a p-value of 0.

In spite of this, it appears, visually, that two of the four non-linear spline functions shown in Figure 4 (`number_of_photos` and `duration_listed`) are approximately linear, and the other two (`age` and `engine_capacity`) could be very well be approximated by linear functions for large portions of the plot. Take, for instance, the plot for `engine_capcity` in Figure 4. While there is some non-linearity present at the ends of the plot, most of the middle portion of the plot (between approximately 1.0 and 1.6 on the transformed scale) could be closely approximated by a line. At the extreme ends of the plot, we begin to see more curvature, but there are fewer observations with extreme values which result in wider standard error bands and also smaller contributions to residual errors. Another example is the plot for `age` (also in Figure 4), in which we observe an inflection point at 2.5 on the transformed scale, but we observe only very slight concavity to the left and to the right of this point. This could represent a point where the used cars are no longer sold for use, but rather have some value as an antique ($2.5^{(1/0.25)} \approx 39$ years).

As such, the reason that the OLS model can achieve most of the predictive accuracy as that of the mixed model is because the underlying relationships are linear (linear in transformed scale, exponential in untransformed scale), save for extreme values near the minimum or maximum of the distributions of each variate. This is also evidenced by the fact that the residual plot of the OLS model (Figure 6) exhibits heteroscedasticity only on the extreme ends of the plot.

Another consideration is the computational time[2] required to fit the two models. The OLS model can be fitted several orders of magnitude faster than the mixed model. The OLS model was fit in 0.05 seconds using the built in `lm()` function in R. The mixed model was fit in 7.77 seconds using the `gamm()` function in the `mgcv` package (Wood, 2011) in R.

[1] raw MAPE = $(1+\text{MAPE})^{(1/0.25)}$ - 1

[2] All computations were performed on the same machine. Processor: Intel(R) Core(TM) i5-4200U CPU @ 1.60GHz, 2301 Mhz, 2 Cores, 8GB RAM.

Another advantage of the OLS model is that it is much more interpretable than the mixed model. Given a change in one variate, one can immediately estimate the change in price using the coefficients of the OLS model. In the mixed model, it is difficult to calculate the change in price resulting from the change of a numerical variate, as one would have to calculate the new value using the coefficients for the basis functions of the splines, which are also not ncessarily readily available.

As such, when comparing the two models for the purpose of estimating the price of a used car, the OLS model is the better model, due to better interpretability, more suitability for the problem, and faster computational time.

4 Discussion and Conclusions

We have explored the relationship between the listing prices and various features of used cars. The features we explored are:

- Vehicle Transmission
- Fuel Type
- Warranty Coverage
- Drivetrain
- Whether the seller is willing to exchange for another car
- Vehicle Age
- Engine Capacity
- Number of photos in the ad
- Duration the ad has been listed

We explored these relationships using a dataset of over 30,000 used car listings, collected in Belarus in December 2019. Most of the cars in this dataset were priced in the range of $2,000 to $8,000 US dollars.

While our model is not accurate in estimating the price of a used car (average error of approximately 35%), our model can be used to estimate the marginal effects of a vehicles features' on its selling price, with high confidence. The marginal effect of each of these features on the price of an average listing (~$5,000), holding all other features constant, are summarized in the below table[3]:

Feature	Marginal effect on price ($)
Vehicle Transmission (Automatic over Mechanical)	+763 for Automatic transmission
Fuel Type (Gasoline over Diesel)	+1,148 for Gasoline
Warranty Coverage	-1,766 for Warranty Coverage
Drivetrain (All wheels over Front wheels)	+1,497 for AWD
Drivetrain (All wheels over Rear wheels)	+636 for AWD
Seller Willing to Exchange	-257 for willingness to exchange
Vehicle Age	-409 per year
Engine Capacity	+1,894 per litre
Number of Photos	+55 per photo
Duration Listed	+2 per day

Table 5: Marginal effects of used car features on listing price.

It should be noted that these relationships are simply averages and will be subject to greater variation if a car offers a very high or very low value of a given feature (e.g., if a car is very old).

Most of the results in the above table are not surprising. However, there are some peculiar results that can be discussed:

[3] More detailed mathematical derivation available in Appendix 3.

- We find that a car actually has a lower listing price if it is covered under warranty. This is possibly due to the fact that owners are generally only going to sell a car that is still under warranty under some unusual circumstances which would adversely lower the selling price.
- The number of photos in a listing obviously does not affect the value of the car itself, but rather it has been found that higher priced cars are sold with more photos in the listing. This is possibly due to the fact that purchasers of higher priced cars would like to perform more due diligence before purchasing, and as such, the sellers post more photos.
- Similarly, the duration for which an ad is listed does not affect the value of the car; but we find that higher priced cars have ads that are listed longer. This is possibly due to the fact that there is less demand for higher priced used cars, which means that the ads must be left up for longer periods before a buyer is found.

5 Future Work

Looking ahead, more comprehensive models could be fit for the purpose of the estimation of used car prices. There are many features of used cars which were not captured in the analysis of this paper. Namely, due to limited data quality, we were not able to capture features such as whether a car has air conditioning, power steering, or power windows. Furthermore, we have not considered factors such as the make and model of a car, which, pragmatically, have effects on the selling price of any car due to changes in perceived quality.

One might consider using other machine learning techniques and algorithms to perform regression analysis on this dataset. Algorithms such as decision trees and neural networks are better suited for performing regression analysis on non-linear data, such as this dataset (pre-transformation), and may yield better results.

References

Del Giudice, V., Manganelli, B., & De Paola, P. (2015). Spline smoothing for estimating hedonic housing price models. In O. Gervasi, B. Murgante, S. Misra, M. L. Gavrilova, A. M. A. C. Rocha, C. Torre, D. Taniar, & B. O. Apduhan (Eds.), *Computational science and its applications – iccsa 2015* (pp. 210–219). Springer International Publishing.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755453

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, *18*. https://doi.org/10.1007/s001800300142

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Data

Used-cars-catalog by Kirill Lepchenkov on Kaggle. Posted on December 2, 2019. https://www.kaggle.com/lepchenkov/usedcarscatalog

Appendix

Appendix 1 - Raw data and dropped features

Below are the names and short descriptions of each of the 30 variates in the original dataset.

Name	Description	Name	Description
manufacturer_name	Name of the car manufacturer.	is_exchangeable	If the seller is willing to exchange for another vehicle.
model_name	Name of the car model.	location_region	Location of seller.
transmission	Transmission type.	number_of_photos	Number of photos posted with the ad.
color	Car body colour.	up_counter	The meaning of this variable is unclear.
odometer_value	Odometer reading, in km.	feature_0	Denotes whether a feature is present in a car.
year_produced	The year the car was produced.	feature_1	Denotes whether a feature is present in a car.
engine_fuel	Fuel type.	feature_2	Denotes whether a feature is present in a car.
engine_has_gas	If the car uses propane.	feature_3	Denotes whether a feature is present in a car.
engine_type	Engine type.	feature_4	Denotes whether a feature is present in a car.
engine_capacity	Engine capacity, in litres.	feature_5	Denotes whether a feature is present in a car.
body_type	Body type.	feature_6	Denotes whether a feature is present in a car.
has_warranty	If the car is covered under warranty.	feature_7	Denotes whether a feature is present in a car.
state	New/Owned/Emergency (severely damaged).	feature_8	Denotes whether a feature is present in a car.
drivetrain	Front/rear/all wheel drive.	feature_9	Denotes whether a feature is present in a car.
price_usd	Listed price, in USD.	duration_listed	Number the days the posting has been up.

Since we are focusing on regression analysis, categorical variables with high-cardinality will limit the interpretability of our models (as one-hot encoding would result in an inundating number of dummy variables). As such, we exclude the following variables: `manufacturer_name` (55 unique values), `model_name` (1118 unique values), `color` (12 unique values), `body_type` (12 unique values), and `location_region` (6 unique values).

Furthermore, there are other variates in the dataset which are uninterpretable. Namely, the meaning of the variate `up_counter` is unclear. The author of the data has also stated that the variates `feature_0` through `feature_9` lack consistency[4]; that is, the presence of a given feature in one car does not match up with the presence of the same feature in another car. As such, these variables should be removed as well. It is assumed that these features were intended to represent features such as air conditioning, power windows, etc.

After removing the aforementioned features, we are left with 14 variates (13 explanatory plus 1 response).

[4] https://www.kaggle.com/lepchenkov/usedcarscatalog/discussion/157041

Appendix 2 - Derivation of the penalized spline estimator in the mixed model formulation

This appendix is derived from Ruppert et al. (2003), pp. 98-100.

The mixed model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \qquad \text{cov}\left(\begin{bmatrix}\mathbf{u} \\ \boldsymbol{\epsilon}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

where $\mathbf{G} = \sigma_u^2\mathbf{I}$ and $\mathbf{R} = \sigma_\epsilon^2\mathbf{I}$.

As mixed model regression:

By placing a normality assumption on $\mathbf{u}$, we can write

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \qquad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}).$$

Maximizing the joint density of $(\mathbf{y}, \mathbf{u})$ with respect to$\boldsymbol{\beta}$ and $\mathbf{u}$ is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}^\mathbf{T}\mathbf{G}^{-\mathbf{1}}\mathbf{u}$$

which leads to the solution

$$\begin{bmatrix}\beta \\ \mathbf{u}\end{bmatrix} = (\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{C} + \mathbf{B})^{-\mathbf{1}}\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{y}$$

with $\mathbf{C} = \begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-\mathbf{1}}\end{bmatrix}$.

When we set $\sigma_u^2 = \frac{\sigma_\epsilon^2}{\lambda^2}$, we get $\mathbf{G} = \frac{1}{\lambda^2}\mathbf{R}$.

$$\begin{aligned}\begin{bmatrix}\beta \\ \mathbf{u}\end{bmatrix} &= (\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{C} + \mathbf{B})^{-\mathbf{1}}\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{y} \\ &= (\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{C} + \mathbf{B})^{-\mathbf{1}}\mathbf{R}^{-\mathbf{1}}\mathbf{R}\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{y} \\ &= (\mathbf{C}^\mathbf{T}\mathbf{R}^{-\mathbf{1}}\mathbf{R}\mathbf{C} + \mathbf{B}\mathbf{R})^{-\mathbf{1}}\mathbf{C}^\mathbf{T}\mathbf{R}\mathbf{R}^{-\mathbf{1}}\mathbf{y} \\ &= (\mathbf{C}^\mathbf{T}\mathbf{C} + \lambda^\mathbf{2}\mathbf{D})^{-\mathbf{1}}\mathbf{C}^\mathbf{T}\mathbf{y}\end{aligned}$$

where $\mathbf{D} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\end{bmatrix}$.

And thus, the fitted values can be attained by

$$\hat{\mathbf{y}} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}$$

As ridge regression:

We regularize (penalize) the "roughness" of the spline by adding a penalty proportional to the coefficient of each basis function. More formally, we want

$$||\mathbf{u}||^2 \leq M$$

for some penalty M. Notice that this is a form of ridge regression.

The objective function for the penalized spline, for some value of λ becomes

$$arg\min_{\beta,u}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \lambda^2\mathbf{u}^T\mathbf{u}.$$

We can recast this using $\mathbf{C} = \begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix}$, $\quad \mathbf{D} = \begin{bmatrix}\mathbf{0}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k\end{bmatrix}$, and $\quad \mathbf{A} = \begin{bmatrix}\boldsymbol{\beta} \\ \mathbf{u}\end{bmatrix}$.

Then we get

$$\begin{aligned}
\mathbf{A}^* &:= arg \min_{\mathbf{A}} (\mathbf{y} - \mathbf{C}\mathbf{A})^T(\mathbf{y} - \mathbf{C}\mathbf{A}) + \lambda^2 \mathbf{A}^T\mathbf{D}\mathbf{A} \\
&= arg \min_{\mathbf{A}} \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{C}\mathbf{A} - \mathbf{A}^T\mathbf{C}^T\mathbf{y} + \mathbf{A}^T\mathbf{C}^T\mathbf{C}\mathbf{A} + \lambda^2\mathbf{A}^T\mathbf{D}\mathbf{A} \\
&= arg \min_{\mathbf{A}} \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{C}\mathbf{A} + \mathbf{A}^T(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^T\mathbf{A}
\end{aligned}$$

Taking the derivative with respect to **A** and applying Fermat's condition:

$$\begin{aligned}
&-2\mathbf{y}^T\mathbf{C} + 2\mathbf{A}^{*T}(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^T = 0 \\
&\implies \mathbf{y}^T\mathbf{C} = \mathbf{A}^{*T}(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^T \\
&\implies \mathbf{C}^T\mathbf{y} = (\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})\mathbf{A}^* \\
&\implies \mathbf{A}^* = (\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}
\end{aligned}$$

And thus, the fitted values can be attained by

$$\hat{\mathbf{y}} = \mathbf{C}\mathbf{A}^* = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}$$

Appendix 3 - Calculation of marginal effects

This appendix shows the derivation of the marginal effects in dollar terms as shown in Table 5 of section 4.

The regression analyses were performed on the transformed dataset, where all numerical variates, including the response variate, were transformed using a power transformation with power 0.25.

Let p denote the response variate (price) in the untransformed scale. Let p^* denote the response variate in the transformed scale. Thus, $p^* = p^{0.25}$.

Similarly, for an explanatory variate, let v be the variate in the untransformed scale, and $v^* = v^{0.25}$ be the variate in the transformed scale.

We want to take the coefficients β_i^* (in transformed scale) in Table 3, from the OLS model fitted on transformed data, so that they can be interpreted in the untransformed scale.

We define the marginal effects in the untransformed scale as β_i for some explanatory variate i.

Then, for numerical variates, we have

$$\beta_i = \frac{\partial p}{\partial v} = \frac{\partial p}{\partial p^*} \frac{\partial p^*}{\partial v*} \frac{\partial v^*}{\partial v}$$

where

$$\frac{\partial p}{\partial p^*} = 4p^{0.75}, \qquad \frac{\partial p^*}{\partial v*} = \beta_i^*, \qquad \frac{\partial v^*}{\partial v} = \frac{1}{4} v^{-0.75}$$

and therefore,

$$\beta_i = \left(\frac{p}{v}\right)^{0.75} \beta_i^*$$

The value for p was set as $5,000 (approximate average price), and the value for v was the average value for the respective variate.

For the categorical variates, the values of v were not transformed (they were simply one-hot encoded), so we have

$$\beta_i = \frac{\partial p}{\partial v} = \frac{\partial p}{\partial p^*} \frac{\partial p^*}{\partial v}$$

with

$$\frac{\partial p}{\partial p^*} = 4p^{0.75}, \qquad \frac{\partial p^*}{\partial v} = \beta_i^*$$

and therefore,

$$\beta_i = 4p^{0.75} \beta_i^*$$

Again, the value for p was set as $5,000 (approximate average price).