c67 fianl report_Model

Jinminli-1008264361

2024-12-03

Car Price in Serbia

Sicheng Huang, Jinmin Li, Zixiang Xiao, Yifan Wang
University of Toronto

Prof. So-hee Kang

Submitted on: December 3, 2024

Task distribution

Sicheng Huang: Cover page; Background and Significance; Impact on the Field; Limitations and Future Research;

Zixiang Xiao: Exploratory Data Analysis; Data Cleaning

Jinmin Li: Model building; Model Selection

Yifan Wang: Diagnostics; Model validation; Key Findings

Background and significance

Cars are among the most complex mass-produced products, representing the culmination of years of research and development. Which makes pricing of cars a complex process influenced by a wide range of factors that extend beyond simple production costs. It reflects a combination of economic, technological, and market dynamics. As a result, car price reflects a careful balance to ensure profitability, market competitiveness, and consumer accessibility.

The automotive industry in Serbia has a long tradition and plays a significant role in the country's economy. With factories like Fiat in Kragujevac, Serbia has become a regional hub for the production of cars and components. In addition to major manufacturers, the industry is supported by numerous companies producing parts and providing maintenance and repair services. The market is dynamic, with growing demand for electric and environmentally friendly vehicles, contributing to the modernization and global competitiveness of the sector. Generally speaking, automotive industry in Serbia is significant and makes up about 15% of industry output, which also make automotives aftermarket in Serbia significant.

Based on the data we are given; we have 17 variables (including quantitative variables and dummy variables) that influence after market price in Serbia. Model diagnostics, such as Cook's Distance, are used to identify and address outliers, improving the reliability of the models. Validation is conducted through a training-testing split, revealing that the Interaction Model demonstrates superior predictive performance. The purpose of this research project is to figure out what are the variables that determine car price in Serbia automotives aftermarket.

Data Cleaning

To prepare the car sales dataset for analysis, several data cleaning steps were undertaken to enhance data quality and ensure meaningful insights. Initially, the dataset was loaded, and non-numeric characters in the horsepower column were removed to convert the entries into numeric values. This was achieved by stripping any trailing text such as " HP" and converting the remaining string into a numeric format, facilitating accurate quantitative analysis involving horsepower.

The car_brand was extracted from the car_name by isolating the first word, assuming it represents the brand. This simplification allowed for easier grouping and categorization based on brand attributes. Outliers in the price column were then addressed using a custom function that calculated the 10th and 90th percentiles to establish a robust interquartile range. Data points outside 1.5 times this range were considered outliers and removed, reducing the impact of extreme values on the analysis.

Categorization played a significant role in the data cleaning process. The car_type variable was recoded into broader categories such as “Luxury,” “Family,” “Utility,” and others to simplify analysis and interpretation. Similarly, the emission_class was transformed into an emission_type variable, grouping different emission standards into “Large Displacement” and “Small Displacement” categories. Fuel types were consolidated into categories like “Oil,” “Electric,” “Mixture,” and “Unknown” to streamline fuel-based analyses.

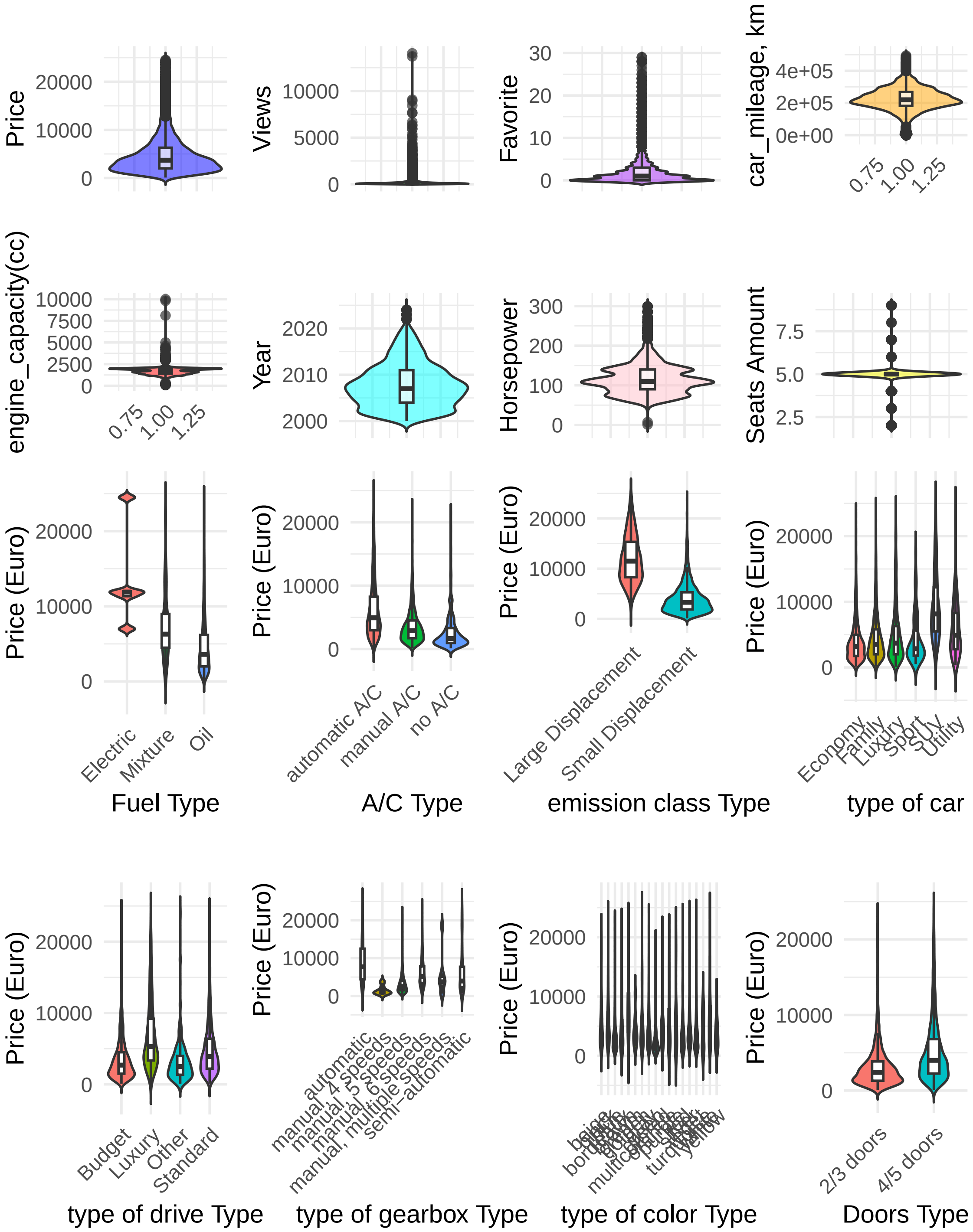
Brands were categorized into “Luxury,” “Standard,” “Budget,” “Other,” and “Unknown” segments to enable comparisons across different market tiers. Entries with missing emission classes were removed to maintain data consistency. Further filtering excluded cars with mileage exceeding 500,000 km, horsepower over 300, a “favorite” count over 30, and manufacturing years before 2000. Additionally, entries with “Unknown” fuel types or brand categories were omitted to focus on data with clear, usable information.

Finally, the horsepower column was reconverted to numeric to ensure consistency after the transformations and filtering. These comprehensive data cleaning steps resulted in a refined dataset, free from anomalies and inconsistencies, thereby facilitating a more accurate and insightful analysis of the car sales trends in Serbia for 2024.

Exploratory Data Analysis

To begin our investigation, we examined our dataset of car sales, which consists of 6,190 observations after cleaning. The continuous variables under consideration are price, views, favorite count, car mileage (in kilometers), engine capacity (in cubic centimeters), year of manufacture, horsepower, and number of seats. By plotting violin charts for each variable’s distribution, we observed that most variables are moderately distributed, with price and horsepower showing noticeable skewness toward the lower end. Specifically, car mileage tends to be skewed toward higher values, indicating that many cars have accumulated significant mileage.

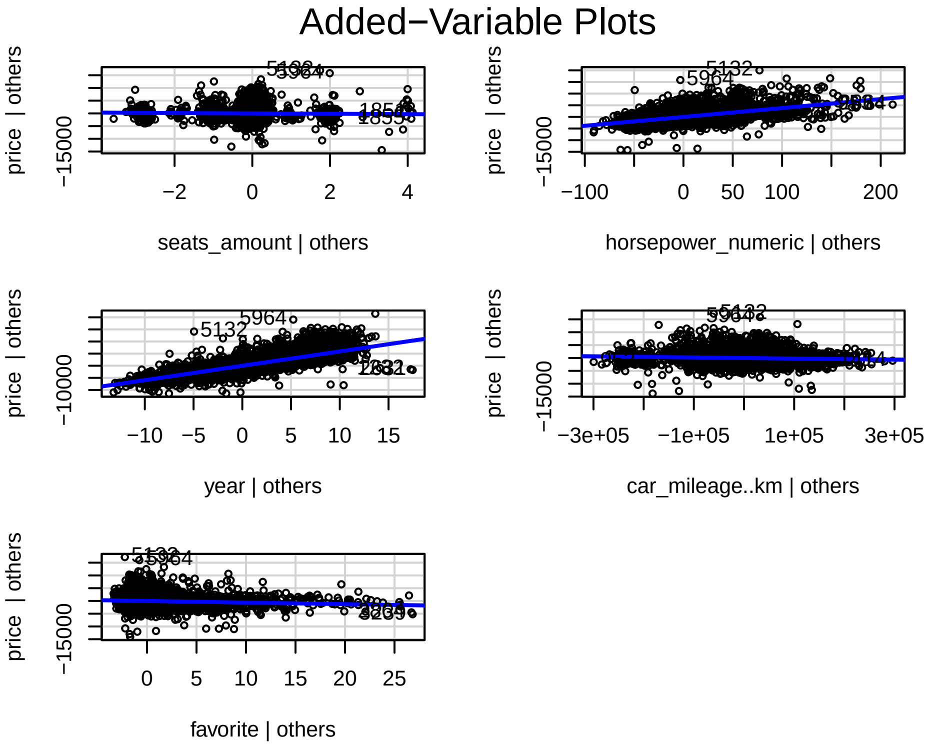
Further examination using violin plots of price against categorical variables such as fuel type, air conditioning type, emission type, car type, car brand category, gearbox type, color, and number of doors revealed distinct patterns. For instance, cars classified under the “Luxury” brand category tend to have higher prices compared to those in the “Standard” or “Budget” categories. Similarly, vehicles with automatic gearboxes generally command higher prices than those with manual transmissions.



We conducted a multiple linear regression analysis to model the car prices based on these variables. The summary of the regression model indicates that several predictors are statistically significant. Notably, year of manufacture and horsepower have strong positive relationships with price, suggesting that newer cars and those with more power are valued higher in the market.

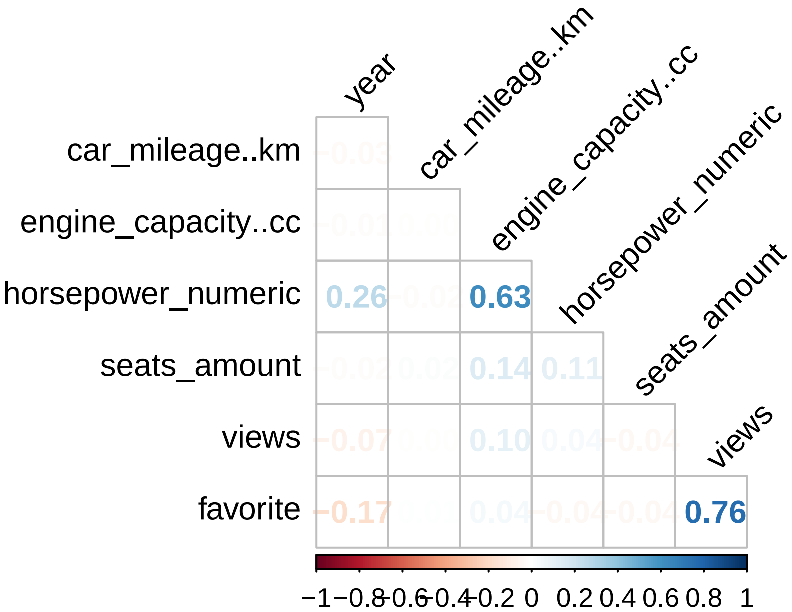
The emission type also plays a significant role; cars with “Small Displacement” engines tend to have lower prices, possibly due to perceptions about performance or efficiency.

Interestingly, the favorite count shows a significant negative relationship with price. This could imply that more affordable cars receive more interest or that users are favoriting cars they aspire to purchase. Car mileage negatively impacts price as expected, with higher mileage reducing the vehicle’s value. The type of drive (e.g., front-wheel drive) and certain fuel types like “Oil” and “Mixture” also negatively affect the price.



Variables such as seats amount, specific car types (like SUVs and Sports cars), and belonging to a “Luxury” or “Standard” brand category positively influence the price. The adjusted R-squared value of 0.7922 indicates that approximately 79% of the variability in car prices is explained by the model, signifying a strong fit.

In summary, the analysis reveals that factors such as year of manufacture, horsepower, emission type, car mileage, car type, and brand category are significant predictors of car price in our dataset. These findings can help buyers and sellers understand the key attributes that influence car valuation in the Serbian market for 2024.



After gaining insights on the relationships between car prices and individual factors, we can now examine the correlation among all the factors in the dataset using the correlation matrix shown in the heatmap above. The heatmap visualizes the relationships between the continuous variables, including year, car_mileage(km), engine_capacity(cc), horsepower_numeric, seats_amount, views, and favorite.

It is evident that views and favorite have a relatively strong correlation (0.76), indicating that cars with more views tend to have more favorites as well. This relationship suggests that these two variables are highly related, which may lead to issues like multicollinearity during model training. Therefore, it could be beneficial to consider dropping one of these factors when building a predictive model to avoid redundancy.

Another noticeable correlation is between engine_capacity..cc and horsepower_numeric (0.63), which makes intuitive sense, as larger engine capacities are typically associated with higher horsepower. However, retaining both these variables might result in moderate multicollinearity. Similarly, year shows a positive correlation with horsepower_numeric (0.26), indicating that newer car models tend to have more powerful engines.

```
##                year     car_mileage..km engine_capacity..cc  horsepower_numeric 
##            1.266095            1.235798            2.129853            2.090930 
##        seats_amount               views            favorite 
##            1.048701            1.649384            1.663423 
```

The VIF analysis for the variables indicates that none of the predictors exhibit problematic multicollinearity, as all VIF values are below 5. The highest VIF values are for favorite (1.663) and views (1.649), suggesting some moderate correlation but not enough to warrant immediate exclusion. Given that all values are relatively low, we can conclude that multicollinearity is not a major concern in this model, and all predictors can be retained for further analysis without significantly affecting model stability or interpretability.

Model

Overview of the Models

In this section, we developed three models to improve the accuracy of our predictions progressively: the main effect model, the interaction model, and the final model. Throughout this process, we added variables we deemed essential and removed those with minimal impact on the model to enhance its predictive performance.

Main Effect Model

We want to build a main effects model. After preliminary cleaning the data, we address multicollinearity by selecting one variable from each pair of variables highly correlated with one another; here we retain “favorite” and “horsepower.” After making a full model with all the variables, we have back-stepped out those whose t-value is small: “color”. This completes our main effects model.

- We used the step function to perform backward stepwise selection on the full model, iteratively removing variables with minimal impact to derive the initial Main Effect Model.

```
## Start:  AIC=93206.13
## price ~ favorite + year + factor(A.C) + factor(emission_type) + 
##     seats_amount + horsepower_numeric + car_mileage..km + factor(type_of_drive) + 
##     factor(doors) + factor(fuel_type) + factor(gearbox) + factor(car_type_1) + 
##     factor(car_brand_category)
## 
##                               Df  Sum of Sq        RSS   AIC
## <none>                                      2.1247e+10 93206
## - factor(fuel_type)           2 2.5508e+07 2.1273e+10 93210
## - factor(doors)               1 1.9673e+07 2.1267e+10 93210
## - seats_amount                1 3.1647e+07 2.1279e+10 93213
## - factor(type_of_drive)       2 4.6658e+07 2.1294e+10 93216
## - car_mileage..km             1 7.6011e+07 2.1323e+10 93226
## - factor(A.C)                 2 1.2866e+08 2.1376e+10 93239
## - favorite                    1 2.9394e+08 2.1541e+10 93289
## - factor(gearbox)             5 7.1315e+08 2.1961e+10 93400
## - horsepower_numeric          1 1.3485e+09 2.2596e+10 93585
## - factor(car_brand_category)  3 1.5961e+09 2.2843e+10 93648
## - factor(car_type_1)          5 1.7429e+09 2.2990e+10 93684
## - factor(emission_type)       1 2.4788e+09 2.3726e+10 93887
## - year                        1 1.5679e+10 3.6927e+10 96625
```

- Based on the results of the AIC method, we identified a simplified model with the following structure and summary, which we define as the Main Effect Model.We removed the variable “color” based on the AIC, and we excluded “post_info” due to its lack of statistical significance. Subsequently, we evaluated the final Main Effect Model by examining its coefficients, R^2, and adjusted R^2 , Adjusted R^2. We observed a reduction in Adjusted R^2, likely due to the decrease in the number of variables included in the model.This leads us to think of building the Interaction Model.
- In the Main Effect Model, the Multiple R^2 value is 0.7938, indicating that approximately 79.38% of the variability in the dependent variable is explained by the model. The

Adjusted R^2 is 0.7922, which accounts for the number of predictors in the model and suggests a strong overall fit.

- Formula of Main Effect Model:

$$\begin{aligned}\text{price} = \beta_0 \ &+ \beta_1 \cdot \text{factor(emission\_type)} + \beta_2 \cdot \text{seats\_amount} + \beta_3 \cdot \text{horsepower\_numeric} + \beta_4 \cdot \text{year} \\ &+ \beta_5 \cdot \text{factor(car\_type\_1)} + \beta_6 \cdot \text{factor(fuel\_type)} + \beta_7 \cdot \text{factor(gearbox)} + \beta_8 \cdot \text{factor(A.C)} \\ &+ \beta_9 \cdot \text{car\_mileage\_km} + \beta_{10} \cdot \text{favorite} + \beta_{11} \cdot \text{factor(car\_brand\_category)} + \epsilon\end{aligned}$$

Interaction model

From the car price data, we suspected that factors such as mileage, horsepower, and seating capacity may interact in a way that the combined effect on price is not simple. We noted that these continuous variables can interact with other predictors, such as the age or the number of seats, in order to improve the predictive power of our model.

- car_mileage..km * year : High mileage decreases the price of newer cars more significantly than older ones. When a newer car has high mileage, it indicates that it has been driven extensively in a short period, which may treat as a red flag for buyers.
- horsepower_numeric * year :Technological improvements in rencent years allow newer cars to have higher horsepower without drawbacks, altering its price impact.
- Backward stepwise selection systematically removes variables included that contribute minimally towards explanatory power using as criteria the AIC.

```
## Start:  AIC=91915.41
## price ~ factor(emission_type) + horsepower_numeric + year + factor(car_type_1) +
##     factor(fuel_type) + factor(gearbox) + factor(A.C) + car_mileage..km +
##     favorite + factor(car_brand_category) + car_mileage..km *
##     year + horsepower_numeric * year
##
##                              Df  Sum of Sq        RSS   AIC
## <none>                                     1.7260e+10 91915
## - factor(fuel_type)           2   51861852 1.7311e+10 91930
## - factor(A.C)                 2  101803605 1.7361e+10 91948
## - favorite                    1  232249317 1.7492e+10 91996
## - factor(gearbox)             5  330642489 1.7590e+10 92023
## - year:car_mileage..km        1  398678542 1.7658e+10 92055
## - factor(emission_type)       1 1270463987 1.8530e+10 92353
## - factor(car_brand_category)  3 1844925660 1.9104e+10 92538
## - factor(car_type_1)          5 2696117728 1.9956e+10 92804
## - horsepower_numeric:year     1 4040978450 2.1300e+10 93216
```

- Base on AIC, the step function gives us a more accurate variable list, by removing "seats_amount" and doors. We finalize the summing up of the model, retaining more impactful variables, and now present improved adjusted coefficients and newly achieved adjusted R-squared, reflecting our further enhanced predictive performance:

term	estimate	std.error	statistic	p.value
(Intercept)	-1.109004e+05	3.564364e+04	-3.11136676	1.870745e-03

factor(emission_type)Small Displacement	-1.962372e+03	9.211871e+01	-21.30264289	3.192232e-97
horsepower_numeric	-1.011806e+04	2.668645e+02	-37.91457875	6.111757e-283
year	5.900586e+01	1.772232e+01	3.32946659	8.752324e-04
factor(car_type_1)Family	-4.133372e+02	5.641363e+01	-7.32690311	2.656922e-13
factor(car_type_1)Luxury	2.485037e+02	6.255255e+01	3.97271973	7.186668e-05
factor(car_type_1)Sport	6.661925e+02	1.842194e+02	3.61629861	3.012340e-04
factor(car_type_1)SUV	2.115374e+03	8.453507e+01	25.02362778	1.049587e-131
factor(car_type_1)Utility	6.976672e+02	1.442002e+02	4.83818315	1.342180e-06
factor(fuel_type)Mixture	-3.063459e+03	7.710481e+02	-3.97311076	7.174906e-05
factor(fuel_type)Oil	-3.201372e+03	7.567248e+02	-4.23056361	2.364843e-05
factor(gearbox)manual, 4 speeds	-2.108384e+03	4.927560e+02	-4.27875858	1.907826e-05
factor(gearbox)manual, 5 speeds	-6.643978e+02	8.895319e+01	-7.46907191	9.196684e-14
factor(gearbox)manual, 6 speeds	-6.968143e+02	7.423301e+01	-9.38685232	8.504299e-21
factor(gearbox)manual, multiple speeds	-1.762112e+03	5.959018e+02	-2.95705007	3.117759e-03
factor(gearbox)semi-automatic	3.098911e+02	2.180681e+02	1.42107451	1.553457e-01
factor(A.C)manual A/C	-2.668047e+02	5.176804e+01	-5.15384952	2.631349e-07
factor(A.C)no A/C	-4.465543e+02	9.137327e+01	-4.88714332	1.048971e-06
car_mileage..km	1.374085e+00	1.152759e-01	11.91996318	2.110010e-32
favorite	-5.919053e+01	6.498644e+00	-9.10813649	1.112508e-19
factor(car_brand_category)Luxury	1.801719e+03	7.289595e+01	24.71630958	1.102879e-128
factor(car_brand_category)Other	-9.104098e+00	1.246150e+02	-0.07305781	9.417625e-01
factor(car_brand_category)Standard	6.808389e+02	5.117141e+01	13.30506388	7.634126e-40
year:car_mileage..km	-6.854964e-04	5.744357e-05	-11.93338737	1.802388e-32
horsepower_numeric:year	5.050606e+00	1.329376e-01	37.99229903	5.584915e-284

- We found that through the interaction terms we increased the Adjusted R^2 from 79.83% to 83.16%. Since that had been done, we could use it to our advantage and help create a better model. Now, we think such inclusion of these interaction terms tremendously enhanced the predictive power of our model, and had been the correct step in our analysis.
- year:car_mileage..km (-0.000679): As the interaction between year and car_mileage..km increases, the predicted price decreases slightly. A one-unit increase in their product reduces price by 0.000679.
- horsepower_numeric:year (5.03): A positive interaction—higher horsepower and later years jointly increase the price. Each unit increase in their product raises price by 5.03.
- We can conclude our interaction model with this formula:

$$\begin{aligned}\text{price} = \beta_0 \; & + \beta_1 \cdot \text{factor(emission\_type)} + \beta_2 \cdot \text{seats\_amount} + \beta_3 \cdot \text{horsepower\_numeric} + \beta_4 \cdot \text{year} \\ & + \beta_5 \cdot \text{factor(car\_type\_1)} + \beta_6 \cdot \text{factor(fuel\_type)} + \beta_7 \cdot \text{factor(gearbox)} + \beta_8 \cdot \text{factor(A.C)} \\ & + \beta_9 \cdot \text{car\_mileage\_km} + \beta_{10} \cdot \text{favorite} + \beta_{11} \cdot \text{factor(car\_brand\_category)} \\ & + \beta_{12} \cdot (\text{car\_mileage\_km} \times \text{year}) + \beta_{13} \cdot (\text{horsepower\_numeric} \times \text{year}) + \epsilon\end{aligned}$$

- After this step, we was on going into the testing of the assumption of linearity. When we delve a little deeper into the diagnostic tests, we find that residual analysis and investigation of multicollinearity between predictors. So we added the power and logarithmic transformation to the variable, to help solve the problem.

Final model (with transformation)

After the first two steps, we realized that we needed to perform some transformations on the variables to achieve a suitable linear model. So, we proceeded to the third step, where we applied logarithmic and square root transformations to certain variables.

- **favorite**: almost all the points were clustered on the far right side of the scatter plot. This indicated a high degree of skewness in its distribution. To address this, we applied a transformation to make its distribution more uniform, which improved the accuracy of our predictions.
- We also found that the **car_mileage** variable had values that varied greatly—some cars had only been driven fewer than one hundred kilometers, while others had been driven around 300,000 kilometers. Even though we removed some extreme values during data cleaning, the range was still quite large. To concentrate the data and reduce the impact of outliers, we applied a square root transformation to this variable.
- We can represent our regression model as:

$$
\begin{aligned}
\text{price} = \beta_0 \; & + \beta_1 \cdot \text{factor(emission\_type)} + \beta_2 \cdot \text{horsepower\_numeric} + \beta_3 \cdot \text{year} \\
& + \beta_4 \cdot \text{factor(car\_type\_1)} + \beta_5 \cdot \text{factor(fuel\_type)} + \beta_6 \cdot \text{factor(gearbox)} \\
& + \beta_7 \cdot \sqrt{\text{car\_mileage..km}} + \beta_8 \cdot \text{factor(A.C)} \\
& + \beta_9 \cdot (\text{car\_mileage..km} \times \text{year}) + \beta_{10} \cdot (\text{horsepower\_numeric} \times \text{year}) \\
& + \beta_{11} \cdot \log(\text{favorite} + 1) + \beta_{12} \cdot \text{car\_brand\_category} \\
& + \epsilon
\end{aligned}
$$

- Base on AIC, the step function gives us a more accurate variable list, we don't need to subtract variables at this stage.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.247622e+05	3.562498e+04	-3.5020985	4.648806e-04
factor(emission_type)Small Displacement	-1.964101e+03	9.172423e+01	-21.4131154	3.535555e-98
horsepower_numeric	-1.009349e+04	2.657117e+02	-37.9866274	6.720204e-284
year	6.573055e+01	1.770698e+01	3.7121261	2.073382e-04
car_type_1Family	-3.889749e+02	5.634797e+01	-6.9030854	5.598091e-12
car_type_1Luxury	2.733098e+02	6.251129e+01	4.3721661	1.250427e-05
car_type_1Sport	6.583379e+02	1.834746e+02	3.5881683	3.355943e-04
car_type_1SUV	2.129072e+03	8.418005e+01	25.2918783	2.290193e-134
car_type_1Utility	7.142246e+02	1.436038e+02	4.9735786	6.750823e-07
fuel_typeMixture	-3.169905e+03	7.678206e+02	-4.1284449	3.700460e-05
fuel_typeOil	-3.332421e+03	7.536273e+02	-4.4218418	9.954377e-06
gearboxmanual, 4 speeds	-2.121555e+03	4.906555e+02	-4.3239191	1.556969e-05
gearboxmanual, 5 speeds	-6.778624e+02	8.857116e+01	-7.6533083	2.259773e-14
gearboxmanual, 6 speeds	-7.180696e+02	7.395143e+01	-9.7100165	3.934702e-22
gearboxmanual, multiple speeds	-1.833237e+03	5.932783e+02	-3.0900119	2.010404e-03
gearboxsemi-automatic	2.885001e+02	2.171311e+02	1.3286909	1.839992e-01
I((car_mileage..km)^(1/2))	3.557703e+00	8.682002e-01	4.0977912	4.224628e-05
factor(A.C)manual A/C	-2.695346e+02	5.158418e+01	-5.2251407	1.797393e-07
factor(A.C)no A/C	-4.363378e+02	9.097721e+01	-4.7961220	1.655702e-06

car_mileage..km	1.473239e+00	1.166137e-01	12.6335035	3.859217e-36
log(favorite + 1)	-3.068412e+02	2.768831e+01	-11.0819745	2.834143e-28
car_brand_categoryLuxury	1.817695e+03	7.262970e+01	25.0268842	9.768493e-132
car_brand_categoryOther	-1.727669e+01	1.241530e+02	-0.1391565	8.893310e-01
car_brand_categoryStandard	6.845197e+02	5.094426e+01	13.4366394	1.368459e-40
year:car_mileage..km	-7.371893e-04	5.821014e-05	-12.6642755	2.634618e-36
horsepower_numeric:year	5.038327e+00	1.323635e-01	38.0643299	6.127489e-285

- Interpretation of transformation terms: For a 1-unit increase in log(favorite + 1), the price decreases by approximately 307 units. This suggests that higher favorite values are associated with lower prices.
- The coefficient of I((car_mileage..km) ^1/2 tells us 3.58 the car price is expected to change for a 1-unit increase in the square root of mileage, holding other variables constant. However, this is counterintuitive, as higher mileage often reduces value. This could be due to correlations with other variables in the model, and we need further investigations.

Table 3: Comparison of Models

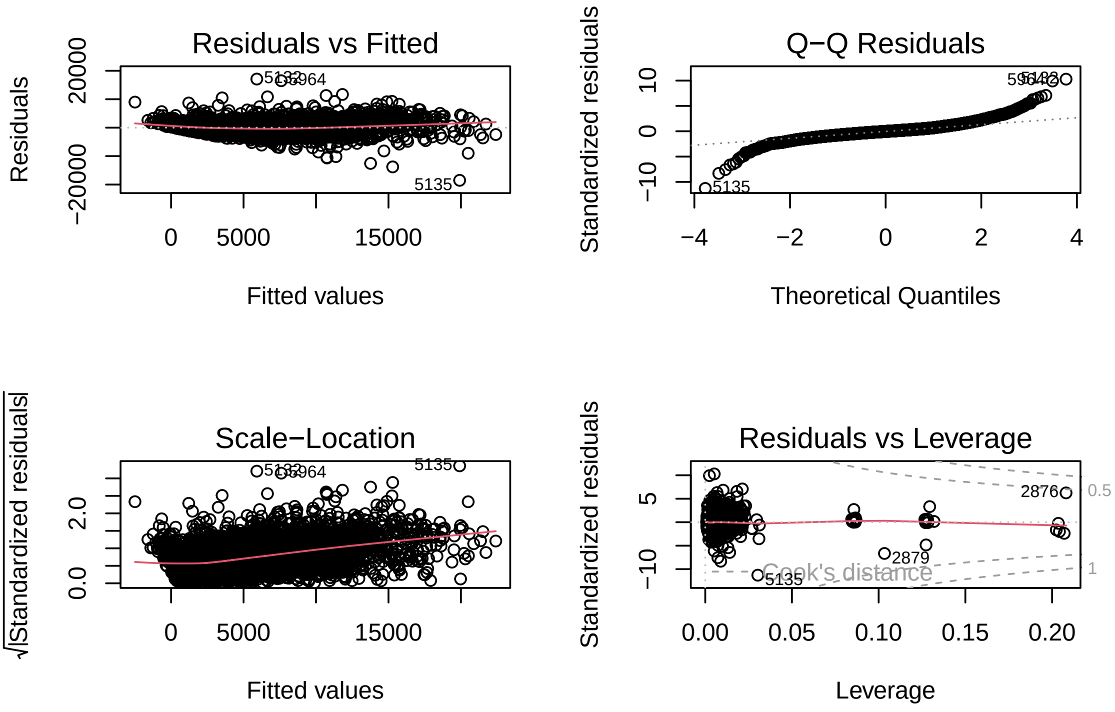
Model	R_Squared	Adjusted_R_Squared	AIC	BIC	Cp
Main Effect Model	0.7913994	0.7905193	93206.13	110963.0	84.5815
Interaction Model	0.8305512	0.8298915	91915.41	109658.9	115.6024
Final Model	0.8320581	0.8313769	91862.12	109612.3	143.2540

Model Selection

- Model Fit:
  - The R^2 and Adjusted R^2 values increase progressively from the Main Effect Model ($R^2 = 0.791$, Adjusted $R^2 = 0.790$) to the Final Model ($R^2 = 0.832$, $AdjustedR^2 = 0.831$). This indicates that each successive model explains more variation in the data.
- Model Complexity:
  - The C_p values increase across the models, from 84.58 in the Main Effect Model to 143.25 in the Final Model, suggesting that the Final Model is more complex, with additional parameters included.
- Model Selection Criteria:
  - The AIC and BIC values decrease slightly from the Main Effect Model (AIC = 93206.13, BIC = 110963.0) to the Final Model (AIC = 91862.12, BIC = 109612.3). The lowest values in the Final Model indicate that it is the most efficient model in balancing goodness of fit and complexity.
- Final Assessment: While the final model sustains its marginal improvement on both terms of R^2 and $AdjustedR^2$, along with improved but substantially greater complexity with ($C_p = 143.25$). Also, with somewhat superior AIC and BIC values for better performance to stand in the most of these above criteria, it is the final model with the superiority mark for the excellence in preforming it.

Diagnostics

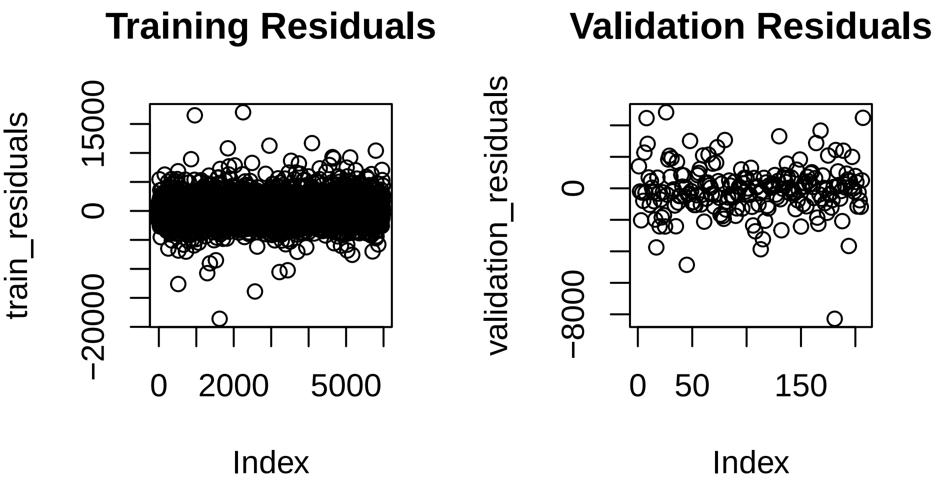
- After cleaning the data we run diagnostics on the model to check for model assumptions.
- The following graphs would give us the best feedback on the model assumptions:



- From the first graph we can know the residuals are roughly centered around zero across the fitted values, which supports normality in terms of mean zero.
- Points align closely with the diagonal line in the central portion of the plot, which supports normality for most residuals in the Q-Q plot.
- According to the third graph, residual variance is almost constant, but the spread does not seem extreme enough to strongly violate normality.
- From the last graph, the points may disproportionately impact the residuals but do not necessarily negate normality for the majority of data points.
- In conclusion, the residuals demonstrate reasonable normality across most of their distribution, especially in the central range. While there are deviations in the tails, the model generally satisfies the assumption of normality for residuals.

Model Validation

```
##                        Metric   Value
## 1    Average MSE (Training) 2762530
## 2 Average MSPE (Validation) 2822852
```



- During cross-validation, we used 70% of the data to train the model, and 30% of the data to keep aside for validation to ensure that the model is evaluated on data it has not seen during training. This approach helps to evaluate the model's performance more robustly and ensures that it is not overfitting to the training data. The Average MSE (Training) for the training dataset is 2762565 which tells us the average squared difference between the predicted and actual values on the training data. Average MSPE (Validation): The Average MSPE (Validation) for the validation dataset is 2820394. The slight difference between the MSE and MSPE shows that the model is performing similarly on both the training and validation datasets, which is a good sign of its generalizability.

Key Findings

- Our data analysis reveals a complex relationship between various car features and prices, with significant interactions between certain variables. It suggests that both direct factors like horsepower and indirect factors like car mileage influence the price and that these effects are moderated by other characteristics such as the car's age or size. We reduced some variables to decrease complexity. For example, we optimize the number of car_type categories from 8 to 3 and reduce fuel type categories. Price data was simplified to focus on the 10th to 90th quantiles, we also removed extreme outliers like cars with more than 500,000 km mileage or cars with more than 300 horsepower.
- According to the study, we found strong correlations between variables like views and favorites, and between engine_capacity and horsepower. This problem was solved by selecting the most significant variables and we selected horsepower numeric and removed engine capacity.
- The main effect model, interaction model, and transformation and power model were used to improve model performance which shows the complexity of the relationships affecting pricing. Firstly, the main Effect Model focused on the relationships between features and car price. Secondly, the interaction model interacted with terms such as horsepower numeric * year car_mileage * year and horsepower numeric * seat amount to find complex relationships. Thirdly, transformation and power model: logarithmic transformations of variables like favorite and square root transformations for car mileage.
- These analyses revealed key variables such as emission_type, seat amount horsepower

numeric, year car type 1, fuel type, gearbox, A.C, and car brand category are significantly associated with pricing so they can be key predictors. Additionally, the horsepower and the car mileage have relationships with the year by the interaction model. However, the year also emerges as a significant predictor of price when analyzed independently. In conclusion, if we want to buy a car with high horsepower, low car mileage, and fewer years, we need to pay more money.

Impact on the Field

In conclusion, we developed a model with a respectable adjusted R^2 value of 0.8331, indicating that the model explains 83.31% of the variation in the dataset. By refining the dataset, grouping similar categories, and incorporating interaction and polynomial terms, we significantly reduced the number of parameters from 95 to 66 during data cleaning and further to 28 in the final model. Our model provides a solution for real life scenarios for customers to evaluate car price using simplify and optimized model.

Our analysis revealed that the most influential positive factors on car prices are 'Car Type (Sport)', 'Car Brand Category (Luxury)', and 'Year', while the most influential negative factors are 'Horsepower Numeric', 'Fuel Type (Oil)', and 'Gearbox Manual (6 Speeds)'. These insights highlight the critical variables driving car pricing and underscore the model's efficiency in capturing key relationships within the data.

Limitations and Future Research

Our current model has significant limitations that require attention for improvement. First, the under representation of certain car types—electric cars (12 instances), hybrid cars (39), gas-powered cars (6), and diesel + gas cars (6)—introduces challenges such as data imbalance, model instability, and issues with statistical significance. This imbalance makes it difficult for the model to accurately generalize across all car types. Additionally, our QQ plot reveals deviations at both tails, indicating the presence of outliers that likely skew the results and reduce model accuracy.

Moreover, the high Mean Squared Error (MSE) difference suggests potential overfitting, which may stem from excessive model complexity or the influence of outliers. Another critical limitation lies in the categorization process we employed to reduce the number of parameters. While simplification was necessary, it inherently introduces subjectivity, potentially leading to inaccuracies in real-world predictions.

To improve the model, we need more data, especially for underrepresented car types like electric, hybrid, gas-powered, and diesel + gas cars, to address data imbalance and reduce bias. Techniques like oversampling rare types, weighting loss functions, and robust methods for handling outliers can further enhance performance. Simplifying the model with regularization can prevent overfitting, while objective categorization methods, such as clustering or expert input, can reduce subjectivity. Adding external variables like fuel availability or economic conditions could also improve predictions. These steps will make the model more balanced, accurate, and practical.

Reference

Jelenković, S., Brzaković, A., & Mihailović, B. (2020). The role and importance of dealers (sellers) for the automobile market in Serbia. Oditor, 6(3), 7–32. https://doi.org/10.5937/oditor2003007j

Domazet, I., & Stosic, I. (2017). Basic characteristics of competitive relations in the after-sales market of motor vehicles in Serbia. Ekonomika Preduzeca, 65(5–6), 413–426. https://doi.org/10.5937/ekopre1706413d