Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics Program

Price predictor for used cars

Econometrics Final Project Report

Authors:

Nazarii Peniaha

Vasyl Burak

Kyrylo Shamanskyi

17 May 2023

Contents

1 Introduction

In recent years, the used car market has expanded quickly. Many people are turning to the used car market as a more economical choice due to the rising expense of buying a new automobile and the increased popularity of ride-sharing services. However, purchasing a used automobile can be intimidating as it can be challenging to determine whether you are receiving a reasonable price. We are motivated to learn more about used automobile prices in order to inform prospective purchasers. In general, we want to offer useful information that will aid consumers in detecting a price for used cars more wisely if, for example, they want to sell it.

Our business concept is to offer an analysis of the used cars market with an emphasis on the costs of various models, brands, and conditions.

Prospective used automobile buyers and sellers seeking a dependable and reasonably priced vehicle would be our target audience for this analysis. The study might also be helpful for auto dealers and insurance providers that want to more correctly price their products or plans.

2 Data description and analysis

It was quite easy to find a dataset on the Internet that provided all the necessary information for our analysis. The dataset is sourced from Kaggle and contains used car prices from some popular car brands: Mercedes, Audi, BMW, Hyundai....
Variables:

1. model: The model of the certain brand of the car.
2. year: The year the car was manufactured.
3. price: The selling price of the car in GBP(British pound sterling).

4. transmission: The type of transmission (automatic or manual).

5. mileage: The distance the car has traveled in miles.

6. fuelType: The type of fuel used by the car (diesel, petrol, other).

7. tax: road tax paid every year by the owner of every motor vehicle which is being used on the roads.

8. mpg: The miles per gallon fuel efficiency of the car.

9. engineSize: The size of the car's engine in liters.

For further analysis, we will use the variable **price** itself as dependent, and all other variables as independent. After analyzing the csv files for each of the brands, we came to the conclusion that the creation of a model for predicting the price of the automobile will be the same for each of them. Accordingly, we chose only one model from 3 popular brands, namely Audi A6, Hyundai Tucson, Merc C Class and created price prediction models for them.

3 Methodology explanation

Creating a price prediction model consists of several stages:

1. Selection of data from the general data set only about the model we are interested in, for example, Audi A6.

2. Conversion of data in columns transmission and fuelType:

    - Automatic or Semi-auto will be replaced by 1 and Manual with 0 respectively.
    - Petrol will be replaced by 1 and Diesel with 0 respectively.

3. Defining the statistically significant variables.

4. For the statistically significant variables, we will create a linear regression model that will predict the price of the machine based on the input data provided.

5. Make a tests for the received model.

Audi:

After completing the first two steps, we will receive data about the Audi A6 with changed data according to the above rules from the initial csv file. We will select a linear regression model as follows, first including all independent variables in the model, and then gradually discarding statistically insignificant ones, guided by the results of the t-test. For defining the statistically significant variables in the model for Audi A6 we will use hypothesis testing:

$$H_0 : \beta_i = 0 \quad vs \quad H_1 : \beta_i \neq 0$$

If the p-value is less or equal than the significance level ($\alpha = 0.05$), then we can reject the null hypothesis. So the variable is statistically significant for our model and we select it in our model. if the p-value is greater then alpha I will not select this variable in the model. As a result, we will leave only variables whose p-value is less than 0.05. The independent variables to be included in the model and their coefficients are listed in Figure 1.

	coef_value	p-value
intercept	-3925675.37	0.0
transmission	1160.17	0.0
mileage	-0.06	0.0
fuelType	-3767.09	0.0
tax	-24.06	0.0
mpg	-576.22	0.0
engineSize	2182.14	0.0
year	1973.01	0.0

Figure 1. *Coefficients and p-values of the model.*

So, we got a model for predicting the price of a car, which will be given the corresponding statistically significant parameters(Figure 2).

$$price = \beta_0 + \beta_1 * transsmision + \beta_2 * mileage + \beta_3 * fuelType + \beta_4 * tax + \beta_5 * mpg + \beta_6 * engineSize + \beta_7 * year + u$$

Figure 2. *The equation of the model.*

We will also consider how our model selects prices compared to real ones — judging not only by the R-squared value but also by plotting the prices for the first 50 cars(Figure 3).

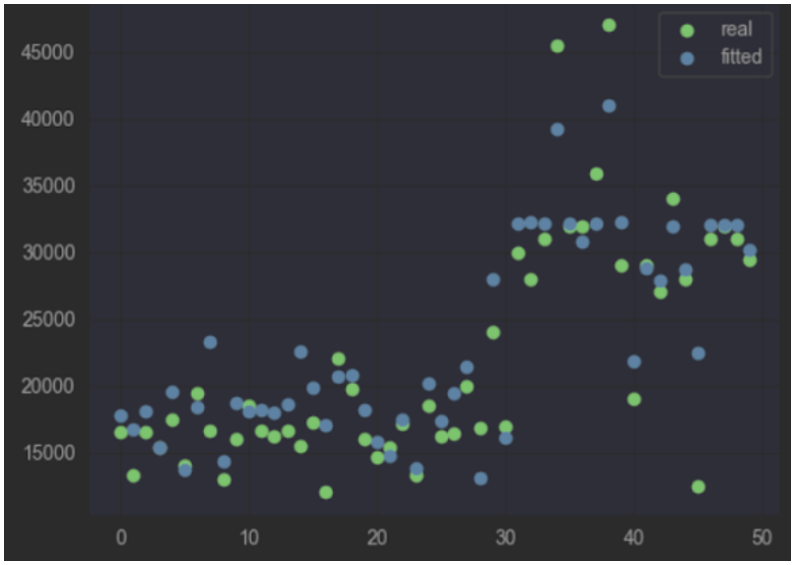

Figure 3. *Comparing graphic.*

Since many young buyers will be interested in the volume of the car's engine, we decided also to make a logistic regression model to predict whether the engine volume is larger than the average value. For all cars with a volume smaller than the average, 0 was assigned, and for all with a larger volume - 1. After discarding all statistically insignificant variables, we obtained a model with the following coefficients as in Figure 4.

	coef_value	p-value
intercept	2139.93	0.00
year	-1.06	0.00
price	0.00	0.00
transmission	10.91	0.00
mileage	0.00	0.02
fuelType	-5.12	0.00
mpg	-0.25	0.00

Figure 4. *Logit model.*

LLR p-value is below 0.05, then we can conclude that the model overall is useful and is well at predicting the values of the response variable.

We performed similar actions for Mercedes C Class and Hyundai Tucson(because we decided also to analyze cheaper brand. One of the cheapest is Hyundai) and obtained the following coefficients and models.

Mercedes C Class model coefficients:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.499e+06	1.06e+05	-42.508	0.000	-4.71e+06	-4.29e+06
year	2235.0887	52.410	42.646	0.000	2132.334	2337.843
transmission	2675.6560	306.472	8.731	0.000	2074.788	3276.524
mileage	-0.1081	0.005	-22.038	0.000	-0.118	-0.099
fuelType	1502.8673	146.999	10.224	0.000	1214.661	1791.073
engineSize	8495.2703	147.305	57.671	0.000	8206.464	8784.077
tax	-7.4786	1.347	-5.554	0.000	-10.119	-4.839
mpg	-66.8890	5.112	-13.085	0.000	-76.911	-56.867

Mercedes C Class model:

$$price = \beta_0 + \beta_1 * year + \beta_2 * transsmision + \beta_3 * mileage + \beta_4 * fuelType + \beta_5 * engineSize + \beta_6 * tax + \beta_7 * mpg + u$$

Hyundai Tucson model coefficients:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.712e+06	8e+04	-21.418	0.000	-1.87e+06	-1.56e+06
year	1705.8869	79.099	21.566	0.000	1550.710	1861.064
transmission	-1.712e+06	8e+04	-21.418	0.000	-1.87e+06	-1.56e+06
mileage	-0.0659	0.006	-11.689	0.000	-0.077	-0.055
fuelType	-1850.1597	286.987	-6.447	0.000	-2413.170	-1287.149
engineSize	1562.0157	390.071	4.004	0.000	796.774	2327.257
tax	9.7538	1.636	5.962	0.000	6.544	12.963
mpg	-49.5536	14.970	-3.310	0.001	-78.921	-20.186

Hyundai Tucson model:

$$price = \beta_0 + \beta_1 * year + \beta_2 * transsmision + \beta_3 * mileage + \beta_4 * fuelType + \beta_5 * engineSize + \beta_6 * tax + \beta_7 * mpg + u$$

4 Results

We obtained models to predict the price of three different models, accordingly we can obtain similar models for any car model. Looking at each of our models, we can give thorough answers to the **questions**:

1. What are the most important features that impact car prices?

   After considering the example of 3 different models of cars, we found that the important features of the influence on the price are: transmission, mileage, fuelType, tax, mpg, engineSize, year.

2. Can we accurately predict car prices based on the available features?

   Comparing the results of the model prediction with real prices, you can even look at the scatter plot in the first part, we will get that 35 percent of the selected prices are with an error of more than 10 percent. In fact, this is unacceptable with a maximum car price of 60k, so more than the available features is needed to accurately predict the price.

3. What are the best strategies for buying or selling used cars based on the available data?

   The smartest thing to do is to operate with the mileage, since it has a negative coefficient in our model, then the sellers will be able to increase the price of the car by reducing the mileage. Thus, buyers are interested in finding out the real mileage of the car

Hypothesis:

1. **Cars with lower mileage have higher prices than those with higher mileage.**

   All our models have a negative mileage factor. This indicates that with higher values of the mileage variable, the price of the car decreases.

   We can also test whether mileage is a statistically significant variable. And in the alternative hypothesis, test whether the sign has a negative coefficient:

$$H_0 : \beta_{mileage} = 0 \quad vs \quad H_1 : \beta_{mileage} < 0$$

The p-value for our test is lower than 0.05. Therefore, we reject the null hypothesis. And conclude that cars with lower mileage have prices higher than those with lower mileage.

2. **Cars with newer model years have higher prices than those with older model years.**

   That hypothesis we test with z-test and by analyzing sign of the coefficient of the variable. And also we use correlation analysis. We firstly determine the relationships between different variables. We explore correlation between the price and other variables(Figure 4).

	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
year	1.000000	0.694233	0.190899	-0.773525	0.153435	0.231409	-0.187740	-0.148636
price	0.694233	1.000000	0.254510	-0.678710	0.251708	0.368708	-0.395439	0.378114
transmission	0.190899	0.254510	1.000000	-0.238973	0.023754	0.172898	-0.069091	0.044991
mileage	-0.773525	-0.678710	-0.238973	1.000000	-0.188849	-0.330324	0.282765	0.093611
fuelType	0.153435	0.251708	0.023754	-0.188849	1.000000	0.311885	-0.351741	0.017782
tax	0.231409	0.368708	0.172898	-0.330324	0.311885	1.000000	-0.564355	0.225370
mpg	-0.187740	-0.395439	-0.069091	0.282765	-0.351741	-0.564355	1.000000	-0.199724
engineSize	-0.148636	0.378114	0.044991	0.093611	0.017782	0.225370	-0.199724	1.000000

Figure 5. *Correlation table.*

We see that variable price is most strongly correlated with mileage (negative correlation) and year (positive correlation). This correlation is another confirmation of hypotheses 1 and 2.

3. **Cars with automatic transmissions have higher prices than those with manual transmissions.**

   In our model, the default for the car is a manual transmission. The transmission coefficient reflects the price difference for the same cars, but with different gearboxes. This coefficient is positive, which means that with an automatic transmission, the price increases.

5 Conclusion/limitations and next steps

During the entire project, aspects that affect the price of used cars on the market were investigated by analyzing data selected from the dataset. Possible approaches were explored and at each stage the most useful approach was chosen, both for analysis and for model creation.

Since we did not manage to achieve the proper accuracy of our model, in the future we would like to find a dataset that has a larger number of features that can influence the price of a used car.

If we can achieve high prediction accuracy through models, then we will create a general-purpose application that will predict the price of used cars based on input data.

References

[1] Link for the Google Colab of the project: https://colab.research.google.com/drive/1Fpjj8IX8M2MkcbiGXQey7TYI4KH9zB0j?usp=sharing

[2] Link to the data set https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes?select=merc.csv

[3] *Introductory Econometrics: A Modern Approach* Jeffrey M. Wooldridge, South-Western College Publishers