Chris Troop
CSC5610
2024-10-12

Project M2 - Linear Regression of Used Car Sale Prices

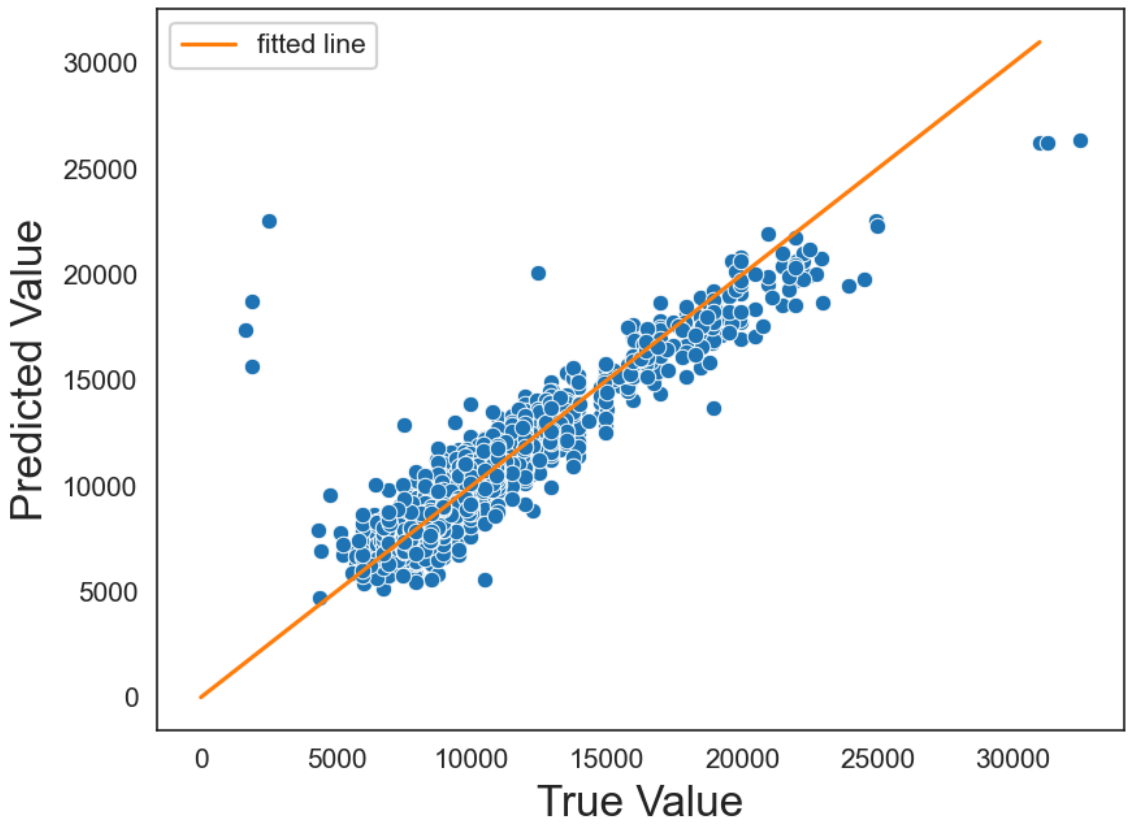
The data that was previously cleaned in the previous report “Used Car Data Report” was leveraged to produce a linear regression model to better understand what is driving the price of used cars in the data set. Each of the variables in the data set were evaluated against the “Price” variable for if there were any trends. For continuous variables (“Age_08_04”, “KM”, and “Weight”) linear regression was performed to assess if the variable contributed to “Price”. For categorical variables, Kruskal-Wallis tests were performed to identify if there was a difference in the mean price of cars within the categories of the variable. If the test failed to find a statistical difference with an alpha of 0.01, the variable was excluded from the linear regression model. The table below shows the results of all tests.

Variable Name	Associated visual/graphically	Associated statistically
Age_08_04	yes	yes
KM	yes	yes
Fuel_Type	no	no
HP	yes	yes
Color	yes	yes
Automatic	No	No
CC	yes	yes
Doors	yes	yes
Cylinders	No	No
Gears	No	No
Quarterly_Tax	yes	yes
Weight	yes	yes
Mfr_Guarantee	yes	yes
BOVAG_Guarantee	yes	yes
Guarantee_Period	yes	yes

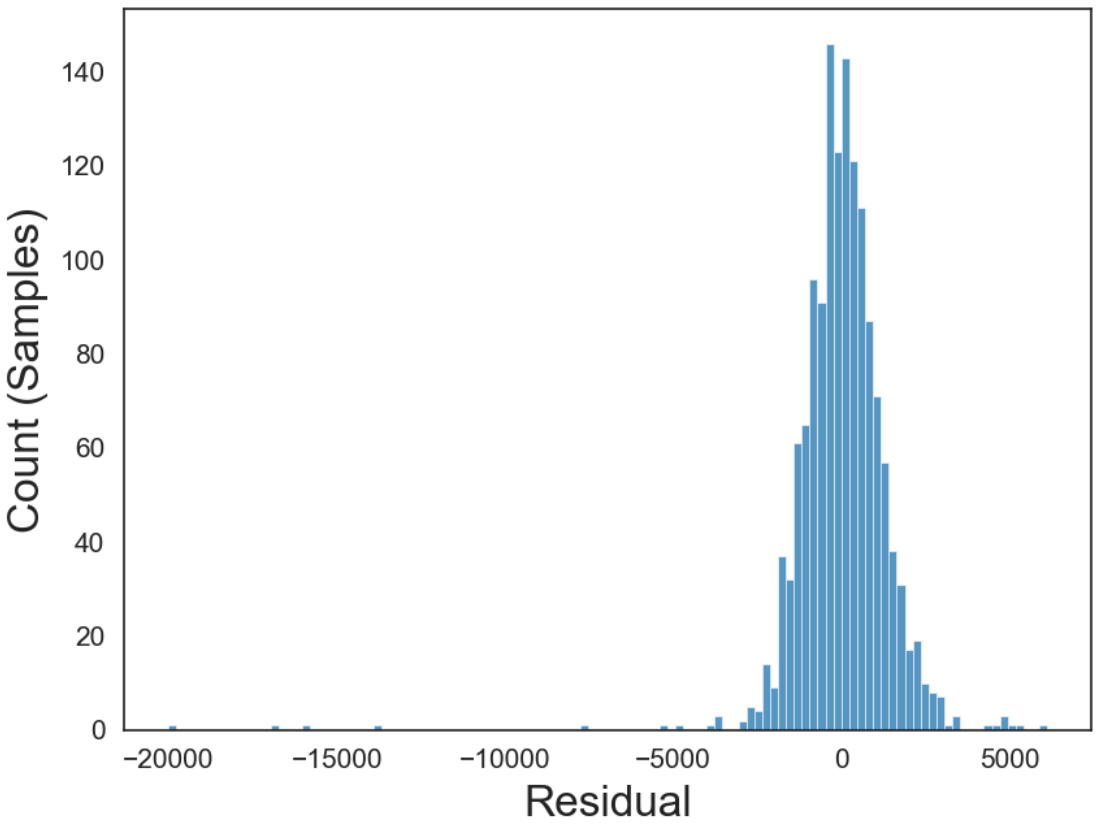
ABS	yes	yes
Airbag_1	yes	yes
Airbag_2	yes	yes
Airco	yes	yes
Automatic_airco	yes	yes
Boardcomputer	yes	yes
CD_Player	yes	yes
Central_Lock	yes	yes
Powered_Windows	yes	yes
Power_Steering	No	No
Radio	No	No
Mistlamps	yes	yes
Sport_Model	No	No
Backseat_Divider	yes	yes
Metallic_Rim	yes	yes
Radio_cassette	No	No
Parking_Assistant	No	No
Tow_Bar	yes	yes

Initially the model was showing an error for a large number of conditions at 7.29e+06. After removing all of the unassociated variables from the table above and scaling “KM”, “CC” and “Weight” by 0.0001, 0.1 and 0.1 respectively, the condition number was reduced to 2.23e+04. Still less than ideal but a 2 order of magnitude reduction nonetheless.

From the linear regression model depicted below, the RMSE (1452.20) and MAPE (10.8%) were fairly good in the context of used car prices. In my opinion, being able to predict the value of a vehicle within 10% is pretty good considering all the variables that are contributing to the price along with economic fluctuations that are not accounted for in this data set.

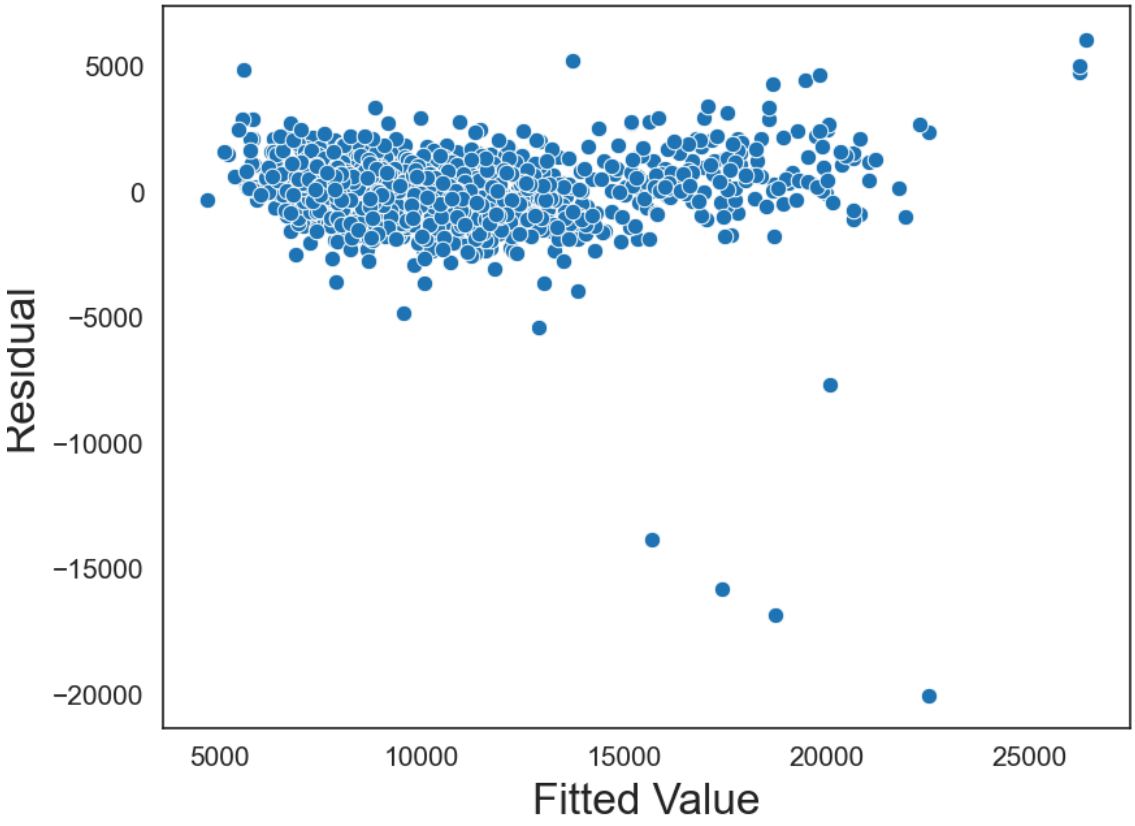


Upon analyzing the residuals, there are a few outliers that cause some issues but for the most part the residuals are normally distributed around 0 (see below). There are a few outliers past -5000 that create a tail.



A visual assessment of the residual dependency was conducted on the plot shown below. Aside from the few outliers mentioned earlier, most of the values are randomly distributed around the x = 0 line indicating there is not a dependency on the

predicting values.



Now that the model is produced we can look at each variable's significance on the predicted value.

Variable Name	Coefficient	unit
Age_08_04	-113.2467	month
KM	-167.0325	km
HP	27.1518	hp
Color (Black)	980.0459	category
Color (Black Metallic)	745.32151	category
Color (Blue)	711.1493	category
Color (Blue Metallic)	560.5724	category
Color (Green)	482.0471	category

Color (Green Metallic)	432.3569	category
Color (Grey)	917.2686	category
Color (Grey Metallic)	660.6544	category
Color (Red)	615.1845	category
Color (Red Metallic)	544.8029	category
Color (Silver)	1188.2934	category
Color (Silver Metallic)	704.4301	category
Color (Violet Metallic)	497.7693	category
Color (White)	-140.5631	category
Color (White Metallic)	1048.2185	category
Color (Yellow Metallic)	130.5588	category
CC	-16.2339	CC
Doors	57.6305	Num of doors
Quarterly_Tax	7.9251	money
Weight	121.5479	kgs
Mfr_Guarantee	196.9632	category
BOVAG_Guarantee	645.4539	category
Guarantee_Period	-18.9254	months
ABS	-209.1795	category
Airbag_1	198.1960	category
Airbag_2	100.6597	category
Airco	234.6329	category
Automatic_airco	2089.2859	category
Boardcomputer	-135.5092	category
CD_Player	112.2321	category
Central_Lock	-4.3456	category
Powered_Windows	258.2454	category

Mistlamps	12.3723	category
Backseat_Divider	1.7022	category
Metallic_Rim	245.5944	category
Tow_Bar	-119.5376	category

Based on this information we can conclude that the addition of an automatic aircon will be the number one feature to increase the sale price of the car followed by painting the car silver.

Used car prices are extremely variable. There are major contributing factors like inventory of new cars and inventory of used cars that drives prices up or down. Basic supply and demand still hold power in this market. The strength of the economy is also a driver in used car prices. Other data that would be helpful to bolster the model would be any sales promotions used on the purchase, sales date and location for normalizing out any inflation that could have impacted the price.