1. **Executive Summary**

Often when deciding on the value of a used car a dealership will need to determine the cars’ re-sale value. Frequently, we as consumers have perceptions about what cars of different ages, makes, models, etc. will maintain more of their value. Because of this, our team decided to analyze and evaluate the auction and retail prices of cars sold in auctions to determine which characteristics had the greatest impacts on the re-sale of a used car.

The dataset used in this analysis contains 10,061 cars sold in different auctions. This dataset was used in Kaggle competition but this version is a shortened adaptation of the dataset originally used. This adaptation contains 10,061 observations of cars sold at individual auction, age of the car, odometer reading, vehicle manufacturer, color of the car, type of wheel, size of vehicle, the price the car was purchased for as well as the retail price of the car. Additionally, this data set contains a variable categorizing whether this car was a good investment or not.

Our team was primarily interested in what factors have the greater impact on the re-sale value of used cars, so we decided to remove the variables for auction, color, and wheel type. This left us with the price we purchased the car at auction, the retail price, the age, the odometer reading, the make, and the size of the car for our analysis.

After running our initial analysis, we found that odometer reading and age had a lot of overlap, meaning there was high multicollinearity. After discovering this, we decided to combine the two variables by dividing our odometer reading by the age of the car to create a more accurate representation of the wear on the car by how old the car is (miles per year). We then added to our analysis to evaluate if the manufacturer or size of the car had a significant impact on the re-sale value.

Our team found that it is clear that miles per year has a very significant impact on the re-sale value of used cars. Interestingly, we found that the manufacturer of a car has little significant impact on the re-sale value compared to our baseline manufacturer of Acura. Additionally, we found that compact cars and vans hold less of their value in the re-sale market compared to our baseline car which was Large SUV.

Our team determined that about 3% of the variation in a used cars profitability can be explained by the miles put on the car each year, the size, and the manufacturer. This model also suggests that the manufacturer of a car has little influence on the re-sale market and compared to Large SUV’s, compact cars and vans will probably sell for less profit. This information may tell us that we as used car re-sellers will want to focus less on which manufacturer a car has or how many miles the car has and focus more on other factors unrepresented in this model, such as accident history, exterior damage, number of previous owners, etc. We would like to expand this model to include these types of variables to hopefully gain more insight into what other factors have greater influence on the re-sale profitability of used cars. We may also want to consider buying fewer Vans or compact cars and purchase more cars of other sizes.

1. Data Exploration

Table-1

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| *Auction* | Auction provider where vehicle was purchased |
| *Age* | The years elapsed since the manufacturer's year (how old is the vehicle) |
| *Make* | Vehicle manufacturer |
| *Color* | Vehicle color |
| *WheelType* | Vehicle wheel type description (Alloy, Covers) |
| *Odo* | Vehicle odometer reading |
| *Size* | Size category of the vehicle (Compact, SUV, etc.) |
| *MMRAauction* | Auction price for this vehicle (in average condition) at time of purchase |
| *MMRAretail* | Retail price for this vehicle (in average condition) at time of purchase |
| *BadBuy* | Identifies if the vehicle was a bad purchase (“YES”) or a good investment (“NO”) |

Table-1 above *explains each variable and what they represent. This dataset is a simplified version of a dataset used in a Caggle competition. We were unsure of which variables were omitted so our team decided to modify our analysis to gain different insights from the dataset’s original purpose. The dataset was originally used to predict whether a car purchased at auction was a bad investment or not. Our team decided to explore what factors impacted the potential profitability of the cars purchased at auction. It is our hope that this approach will give insight into which cars will sell for more profit than others so that we may focus our resources on more profitable used cars.*

The variables that appear to be important are odometer, age, MMRAauction and MMRAretail. The higher the odometer usually means the car is older and it has been used more. This would normally result in a lower price that the auction is able to sell it for. The age of the car can also affect the price of the car because older cars sell for less. MMRAauction and MMRAretail and both important because these two tell us how much they bought the car at the auction and how much they were able to sell the car for.

We think we should include additional variables like profit and the relationship between odometer and age. To calculate the new profit variable, we will calculate the difference between retail price of the car and the auction price of the car. This will essentially show us how much profit they made from selling the vehicle with the information that we have. We also made a new variable that showed the relationship between age and odometer which would show us how many miles per year the car accumulated. We did this by doing odometer divided by age.

We decided to remove some variables after looking at the data. We removed the variable auction because it was just the location where the auction was held. There were two names locations and then there was “other” and a majority of the cars were held at “other” auction locations. This didn’t give us much information so we removed it. We also removed the variable color because there was such a wide variety of colors that would not be significant. We removed wheel type because there were many NULLs instead of an option of alloy or covers.

We can visualize the data in unique ways if we run a histogram. Initially, we tried to run a box plot but ran into some issues. Running a histogram on age (Histogram 1 in our appendix) showed us that the age of the cars was pretty spread out but there weren’t as many one-year old cars and 9-year old cars as there were the rest of them. The majority age of cars was either 3, 4, and 5. After we made our new variable, profit, we decided to go back and run a histogram on that variable as well. (Histogram 2 in our appendix) This showed us that the majority of our profit was in the 800-1200 range and 3000-4000 range. We only made more than $6,000 of profit on 16 cars.

1. Main Analysis

We decided to start our analysis by deciding how to best measure the factors that affect re-sale value. Even though this dataset had a categorical variable for dictating whether or not the car was a bad investment, we didn’t have a clear picture of how much impact the observed variables in this dataset impacted profitability. We assumed that this was likely because this was a simplified version of the competition data. Taking this into consideration, we decided to create a new variable to attempt to measure the profitability of these used cars purchased at auction.

The profit variable was created by calculating the difference between the retail price of the car and the price the car was purchased for at auction. This new profit variable indicated that the average profit made from these cars was about $2539. We then decided to use a linear regression analysis as our observed dependent variable was continuous and not categorical in the way that the BadBuy variable was.

Our first model was a simple linear regression with our dependent variable as profit and our independent variable as age. As seen in the R code file in the appendix, we found that age was a very significant variable but did not explain much of the variation as expected. We then expanded our model to include the odometer rating, and then the make of the vehicle.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R^2 | Model P-value | Model RMSE | Adj R^2 |
| Profit~Age | 0.001 | 0.001 | 1356.07 | 0.001 |
| Profit~Age +Odo | 0.01 | 4.10e-16 | 1348.9 | 0.009 |
| Profit~Age +Odo +Make | 0.025 | 2.2e-16 | 1348.79 | 0.021 |

As expected, our model improved as we added additional variables. Notedly, we noticed that our models explained variation in profit climbed while our prediction measurement (RMSE) mostly remained the same or steadily improved by small increments.

We expected there to be some potential multicollinearity problems with our age and odometer variables. Logically one would expect the odometer to increase steadily along with age as the car is used over time. After confirming our suspicions with a partial R^2 analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Partial R^2 Model | R^2 | Model P-value | Adj R^2 |
|  | 0.1135 | 2.2e-16 | 0.1133 |

we decided to combine the variables together in the hopes of creating a more accurate measure of the wear on the used cars by dividing the odometer reading by the age of the cars. This had the double benefit of removing any potential multicollinearity issues. This new variable creates a miles per year or miles per age that we think provides more insight into the condition of the used car.

For our final model, we ended up using our new “odoage” (Odo/Age) variable, size, and make(manufacturer). We wanted to see if we should be purchasing less of certain sizes of cars and focusing more on other car sizes to maximize profitability

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R^2 | Model p-value | Model RMSE | Adj R^2 |
|  | 0.029 | 2.2e-16 | 1348.40 | 0.023 |

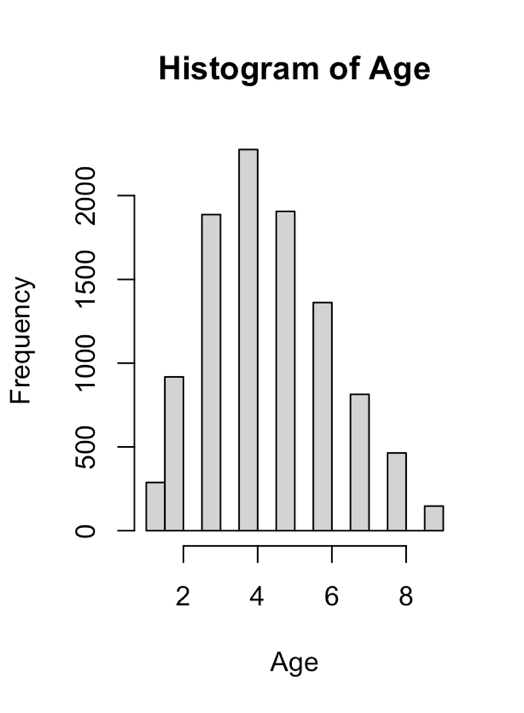
Despite our final model being the best for prediction and inference which can be determined by comparing our model’s RMSE, adjusted R^2, R^2, and model p-values. This model only explained about 3% of the variation in profit but we believe we gained some interesting insights. One of our results were that the make of a car had very little significance of this model. We also discovered that compared to our baseline car size which was Large SUV, vans and compact cars seem to have significantly less profitability. Even though that our model only explains a small amount of the variation in profit, we decided that there may be some actionable insights to be gained from this model.

As can be referenced in the appendix where our coefficients are found. We found that compared to large SUV’s vans and compact cars are approximately 5% less profitable on average and overall it seems that the age or odometer reading on a used car has very little to do with the cars’ potential profitability as the impact of these variables only slightly effect profitability.

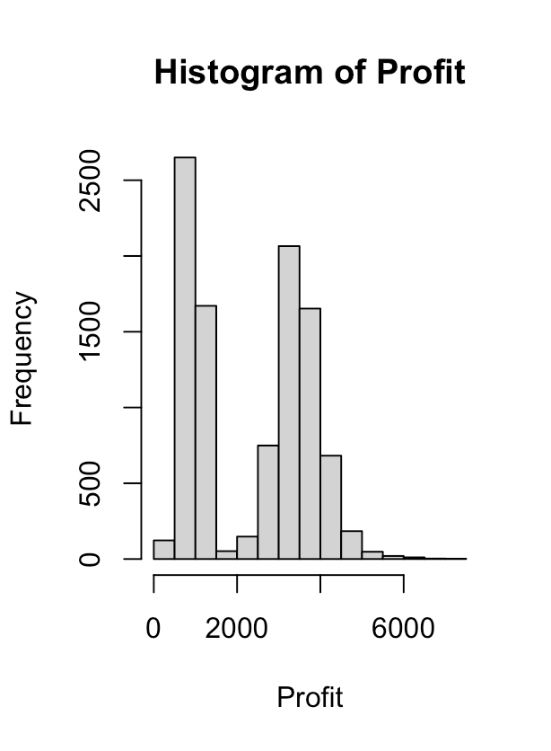
We recommend that dealerships or used car re-sellers should focus less attention on factors, such as manufacturer and mileage and focus perhaps on other factors that may be more impactful on profitability that are unrepresented in this model. Some of the unrepresented factors may be exterior damage, number of previous owners, accident history etc.

**APPENDIX**

**Histogram #1 -**

****

**Histogram #2**

****

**R-Code**

**Regression 1**

lm(formula = profit ~ Age, data = car\_train)

Residuals:

Min 1Q Median 3Q Max

-2432.9 -1415.0 520.2 1150.3 4832.9

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2493.201 44.573 55.935 < 2e-16 \*\*\*

Age -30.157 9.197 -3.279 0.00105 \*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1365 on 7040 degrees of freedom

Multiple R-squared: 0.001525, Adjusted R-squared: 0.001383

F-statistic: 10.75 on 1 and 7040 DF, p-value: 0.001046

**Regression 2**

lm(formula = profit ~ Age + Odo, data = car\_train)

Residuals:

Min 1Q Median 3Q Max

-2657.7 -1386.3 511.1 1158.8 4942.2

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.937e+03 8.425e+01 22.988 < 2e-16 \*\*\*

Age -5.561e+01 9.727e+00 -5.718 1.12e-08 \*\*\*

Odo 9.216e-03 1.186e-03 7.770 8.97e-15 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1359 on 7039 degrees of freedom

Multiple R-squared: 0.01002, Adjusted R-squared: 0.009735

F-statistic: 35.61 on 2 and 7039 DF, p-value: 4.108e-16

**Regression 3**

lm(formula = profit ~ Age + Odo + Make, data = car\_train)

Residuals:

Min 1Q Median 3Q Max

-3113.2 -1355.1 471.1 1159.4 4939.9

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.427e+03 5.206e+02 4.662 3.19e-06 \*\*\*

Age -5.736e+01 1.004e+01 -5.714 1.15e-08 \*\*\*

Odo 9.460e-03 1.221e-03 7.749 1.06e-14 \*\*\*

MakeBUICK -7.833e+02 5.337e+02 -1.468 0.1422

MakeCADILLAC -1.421e+03 9.324e+02 -1.524 0.1275

MakeCHEVROLET -5.678e+02 5.123e+02 -1.108 0.2677

MakeCHRYSLER -4.410e+02 5.133e+02 -0.859 0.3902

MakeDODGE -4.925e+02 5.128e+02 -0.961 0.3368

MakeFORD -5.613e+02 5.122e+02 -1.096 0.2732

MakeGMC -3.282e+02 5.356e+02 -0.613 0.5401

MakeHONDA -1.165e+02 5.392e+02 -0.216 0.8289

MakeHYUNDAI -6.475e+02 5.216e+02 -1.241 0.2145

MakeINFINITI -6.298e+02 7.226e+02 -0.872 0.3835

MakeISUZU -8.060e+02 6.533e+02 -1.234 0.2174

MakeJEEP -3.565e+02 5.215e+02 -0.684 0.4942

MakeKIA -5.002e+02 5.184e+02 -0.965 0.3346

MakeLEXUS 1.225e+03 7.925e+02 1.546 0.1222

MakeLINCOLN -1.634e+02 6.067e+02 -0.269 0.7877

MakeMAZDA -1.309e+02 5.261e+02 -0.249 0.8035

MakeMERCURY -5.369e+02 5.275e+02 -1.018 0.3088

MakeMINI -1.623e+03 9.322e+02 -1.741 0.0817 .

MakeMITSUBISHI -5.691e+02 5.277e+02 -1.078 0.2809

MakeNISSAN -1.460e+02 5.189e+02 -0.281 0.7784

MakeOLDSMOBILE -1.275e+03 5.607e+02 -2.274 0.0230 \*

MakePONTIAC -6.831e+02 5.154e+02 -1.325 0.1851

MakeSATURN -5.324e+02 5.193e+02 -1.025 0.3052

MakeSCION -3.092e+02 6.434e+02 -0.481 0.6308

MakeSUBARU 1.066e+03 1.083e+03 0.984 0.3250

MakeSUZUKI -4.410e+02 5.251e+02 -0.840 0.4010

MakeTOYOTA 1.875e+02 5.276e+02 0.355 0.7224

MakeVOLKSWAGEN 6.722e+01 6.427e+02 0.105 0.9167

MakeVOLVO 8.967e+02 1.083e+03 0.828 0.4079

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1351 on 7010 degrees of freedom

Multiple R-squared: 0.02569, Adjusted R-squared: 0.02138

F-statistic: 5.962 on 31 and 7010 DF, p-value: < 2.2e-16

**Regression 4**

lm(formula = profit ~ odoage + Make + Size, data = car\_train)

Residuals:

Min 1Q Median 3Q Max

-3010.0 -1369.0 504.6 1158.8 5083.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.646e+03 5.361e+02 4.936 8.14e-07

odoage 7.142e-03 1.815e-03 3.935 8.40e-05

MakeBUICK -4.109e+02 5.466e+02 -0.752 0.452236

MakeCADILLAC -1.435e+03 9.312e+02 -1.541 0.123313

MakeCHEVROLET -1.782e+02 5.254e+02 -0.339 0.734533

MakeCHRYSLER -1.125e+02 5.246e+02 -0.214 0.830185

MakeDODGE -1.923e+02 5.236e+02 -0.367 0.713369

MakeFORD -2.171e+02 5.254e+02 -0.413 0.679395

MakeGMC -1.470e+02 5.492e+02 -0.268 0.788967

MakeHONDA 2.739e+02 5.518e+02 0.496 0.619713

MakeHYUNDAI -3.224e+02 5.341e+02 -0.604 0.546046

MakeINFINITI -7.183e+02 7.215e+02 -0.996 0.319504

MakeISUZU -5.882e+02 6.647e+02 -0.885 0.376244

MakeJEEP -6.548e+01 5.384e+02 -0.122 0.903210

MakeKIA -1.895e+02 5.305e+02 -0.357 0.720891

MakeLEXUS 1.089e+03 7.905e+02 1.378 0.168301

MakeLINCOLN -4.922e+01 6.126e+02 -0.080 0.935963

MakeMAZDA 2.324e+02 5.389e+02 0.431 0.666299

MakeMERCURY -2.003e+02 5.403e+02 -0.371 0.710892

MakeMINI -1.103e+03 9.400e+02 -1.174 0.240495

MakeMITSUBISHI -2.535e+02 5.405e+02 -0.469 0.639086

MakeNISSAN 2.367e+02 5.317e+02 0.445 0.656184

MakeOLDSMOBILE -9.400e+02 5.725e+02 -1.642 0.100626

MakePONTIAC -2.609e+02 5.285e+02 -0.494 0.621547

MakeSATURN -2.153e+02 5.319e+02 -0.405 0.685706

MakeSCION -9.823e+00 6.538e+02 -0.015 0.988013

MakeSUBARU 1.289e+03 1.090e+03 1.183 0.236731

MakeSUZUKI -1.356e+02 5.369e+02 -0.253 0.800642

MakeTOYOTA 5.222e+02 5.402e+02 0.967 0.333722

MakeVOLKSWAGEN 2.917e+02 6.441e+02 0.453 0.650626

MakeVOLVO 9.101e+02 1.082e+03 0.841 0.400372

Size'LARGE TRUCK' 1.952e+02 1.376e+02 1.419 0.155936

Size'MEDIUM SUV' -1.738e+02 1.242e+02 -1.399 0.161901

Size'SMALL SUV' -2.157e+02 1.547e+02 -1.394 0.163429

Size'SMALL TRUCK' -3.258e+02 1.764e+02 -1.847 0.064820

SizeCOMPACT -4.230e+02 1.243e+02 -3.403 0.000669

SizeCROSSOVER -2.476e+02 1.591e+02 -1.556 0.119809

SizeLARGE -3.443e+02 1.268e+02 -2.715 0.006647

SizeMEDIUM -2.769e+02 1.197e+02 -2.313 0.020762

SizeSPECIALTY 7.321e+01 1.658e+02 0.442 0.658818

SizeSPORTS -7.835e+01 1.755e+02 -0.447 0.655201

SizeVAN -4.916e+02 1.287e+02 -3.820 0.000134

(Intercept) \*\*\*

odoage \*\*\*

MakeBUICK

MakeCADILLAC

MakeCHEVROLET

MakeCHRYSLER

MakeDODGE

MakeFORD

MakeGMC

MakeHONDA

MakeHYUNDAI

MakeINFINITI

MakeISUZU

MakeJEEP

MakeKIA

MakeLEXUS

MakeLINCOLN

MakeMAZDA

MakeMERCURY

MakeMINI

MakeMITSUBISHI

MakeNISSAN

MakeOLDSMOBILE

MakePONTIAC

MakeSATURN

MakeSCION

MakeSUBARU

MakeSUZUKI

MakeTOYOTA

MakeVOLKSWAGEN

MakeVOLVO

Size'LARGE TRUCK'

Size'MEDIUM SUV'

Size'SMALL SUV'

Size'SMALL TRUCK' .

SizeCOMPACT \*\*\*

SizeCROSSOVER

SizeLARGE \*\*

SizeMEDIUM \*

SizeSPECIALTY

SizeSPORTS

SizeVAN \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1349 on 7000 degrees of freedom

Multiple R-squared: 0.02905, Adjusted R-squared: 0.02336

F-statistic: 5.109 on 41 and 7000 DF, p-value: < 2.2e-16