Data Mining Project Analysis on Cadillac

ISM 6136 Data Mining

Dr. Kiran Garimella

University of South Florida

December 8, 2021

Lahari Goshika
Ms. BAIS

Background of the Problem

Today's used car market is increasing at a rapid pace. The problem is there needs to be a method to predict used car values and pricing. In the current market, there are various supply chain issues, with one being bottlenecks of vehicle supplies from international countries like Japan to the United States. Additionally, there are international supply issues with the U.S. importing microchips from foreign countries slowing down new car production and leaving new cars sitting in empty lots waiting for chips to be installed before sale.

Since we are amidst a pandemic and economic uncertainty, many citizens are having to face tough financial decisions to maintain necessities like housing and food. One of the tough decisions citizens are making are whether to sell their vehicle for supplemental income, or whether purchasing a used vehicle makes financial sense. The issue is many people do not know how much their used car is worth and face the potential of selling their car for a smaller price than they should have or also face the potential for purchasing a used vehicle for more than they should have.

This is when a full data mining analysis would be most helpful. A full data mining analysis can work with multiple variables such as mileage, year, make, model & driving history and can create a reliable estimate for the value of the car. There are many companies and sites that have data related to sold car histories etc. making the gathering of the requisite data easy. Using data mining tools to predict car prices will be able to provide assurance to the seller and buyer of the vehicle, that they are getting a fair deal.

This technique would not only be beneficial to individual people, but additionally, individual businesses. Companies like Carvana who purchase vehicles online, would be able to use the data mining techniques to determine what the fair offer price should be for the vehicle. Companies have made poor purchases of used cars in the past where they ended up with a vehicle that was worth less than what they paid for and end up sitting in a car lot until it is sold for a loss. This situation also occurs when a dealership takes in a used car as a trade in for a new vehicle.

Another problem is that automobile manufacturers are going to need to know how much they should price their new vehicles out of production. As the used car values increase, it will naturally increase the price of new vehicles, but there must be an equilibrium of what to price the new vehicles so that they're not too much more expensive than their used counterpart so they can still incentivize the sale of their new cars without eliminating the demand for their used vehicles as well. Being able to determine the used price of a vehicle would help the manufactures of new cars determine a price for their new vehicles.

All in all, the problem is finding a way to determine the price of used vehicles in a volatile economy. As we will showcase in this project, full data mining techniques can help solve this issue and be used by both individual people and businesses to find the most reliable price of used cars.

Motivation for solving the Problem: (Paul De Costa Jr)

Our motivation to solve the problem of getting the most accurate used car price is to prevent the buyers from overpaying or under selling for used vehicles especially during economic uncertainties. We are all consumers of motorized vehicles and have personally dealt with over and under evaluation of cars, both used and new. We feel that giving businesses and individuals a tool to derive a fair price for used vehicles would be a positive contribution to society.

This is when a full data mining analysis would be most helpful which we will describe in more detail in this project. A full data mining analysis can work with multiple variables such as mileage, year, make, model & driving history and can create a reliable estimate for the value of the car. There are many companies and sites that have data related to sold car histories etc. making the gathering of the requisite

data easy. Using data mining tools to predict car prices will be able to provide assurance to the seller and buyer of the vehicle, that they are getting a fair deal.

This technique would not only be beneficial to individual people, but additionally, individual businesses. Companies like Carvana who purchase vehicles online, would be able to use the data mining techniques to determine what the fair offer price should be for the vehicle. Companies have made poor purchases of used cars in the past where they ended up with a vehicle that was worth less than what they paid for and end up sitting in a car lot until it is sold for a loss. This situation also occurs when a dealership takes in a used car as a trade in for a new vehicle.

Solution Methodology

Start -> Data Collection -> Pre-processing of Data -> Applying algorithm -> Building prediction model -> Comparing the results

Data Description:

For this analysis, we are working on a dataset that consists of data of all the used cars which belongs to the year ranging from 1965 to 2019. The whole data set consists of 10 attributes or variables, and they are region, price, year, make, model, condition, cylinders, fuel, odometer, paint color.

The dataset consists of cars from different companies which further has different models, and we will be working on one specific model to understand the variation caused by the variables. And the model we will be working on is Cadillac.

Below is the sample snippet of the dataset.

REGION	PRICE	YEAR	MAKE	MODEL	CONDITION	CYLINDERS	FUEL	ODOMETER	PAINT COLOR
albuquerque	15500	1965	ford	mustang	excellent	8	gas	4800	blue
albuquerque	17995	2015	ford	transit	good	6	gas	71181	white
albuquerque	18995	2014	ram	promaster 2500	good	6	gas	80483	white
albuquerque	8998	2012	volkswagen	jetta tdi	excellent	4	diesel	89000	white
albuquerque	22500	2003	ford	mach1 mustang	excellent	8	gas	15700	white
albuquerque	18995	2014	chevrolet	express g4500	good	8	gas	93187	white
albuquerque	6995	2003	ford	f250	good	8	gas	145468	white
albuquerque	8495	2007	audi	a6 4.2 s-line	good	8	gas	87000	grey
albuquerque	5950	2007	nissan	titan	excellent	8	gas	172108	red
albuquerque	13995	2015	ford	f150	excellent	8	gas	117201	white
albuquerque	15995	2005	pontiac	gto	good	8	gas	61578	black
albuquerque	9995	2010	chevrolet	express g2500	good	8	gas	129753	white
albuquerque	17995	2017	ford	transit	good	6	gas	39803	white
albuquerque	5995	2008	dodge	grand caravan	good	6	gas	146131	red
albuquerque	22995	2017	chevrolet	equinox	good	6	gas	19126	red

As the predicted outcome for this data is price and it is a continuous variable, we will be using multiple linear regression for this model, and the prediction model is built using R-script.

Pre-processing of Data

The sourced dataset has been checked for missing values and null values, and it is found that the data is clean. We created a subset of the original dataset using the conditions such as make= 'Cadillac', the cars belonging to the year between 2012 and 2019. As most number of cars belong to this period, and to prevent outliers in the data, we have selected that period. We selected odometer (which denotes the number of miles the car has been driven), year, cylinders (number of cylinders the engine has) as the independent variables and the dependent variable is price. The newly created subset contains a sample size of 193, and we created a training and test data set in 80:20 proportion.

R-script

```
library(rio)
install.packages("writexl")
master_data=import("Used _Car_price_details.xlsx")
colnames(master_data)=tolower(make.names(colnames(master_data)))
attach(master_data)
cadillac=subset(master_data, make=="cadillac" &(year>2012) &(year<2019) )
set.seed(30011610)
cadillac$index=seq(1:nrow(cadillac))
cadillac_train=subset(cadillac, index>=1 & index<=153)
cadillac_test=subset(cadillac, index>153)
cadillac.out=lm(price~odometer+year+cylinders, data=cadillac_train)
summary(cadillac.out)
cadillac_train$cylinders=as.factor(cadillac_train$cylinders)
cadillac_train$year=as.factor(cadillac_train$year)
str(cadillac_train)
cadillac3.out=lm(price~odometer+year+cylinders,data=cadillac_train)
summary(cadillac3.out)
```

Output:

```
Call:
lm(formula = price ~ odometer + year + cylinders, data = cadillac_train)
Residuals:
    Min      1Q  Median          3Q   Max
-16266.1 -3727.6  329.1  2685.6 31496.1
```

Coefficients:

```
                 Estimate   Std. Error      t value Pr(>|t|)
(Intercept) 1.835e+04 2.194e+03     8.365 4.73e-14 ***
odometer      -1.141e-01 2.059e-02 -5.540 1.39e-07 ***
year2014    4.629e+03    1.581e+03       2.928 0.00397 **
year2015    9.158e+03    1.783e+03    5.137 8.92e-07 ***
year2016    1.176e+04    1.961e+03    5.999 1.53e-08 ***
year2017    1.677e+04    1.953e+03    8.588 1.31e-14 ***
year2018    1.602e+04    2.631e+03    6.088 9.85e-09 ***
cylinders6   3.009e+03 1.445e+03    2.083 0.03904 *
cylinders8   2.178e+04 1.635e+03 13.325 < 2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6411 on 144 degrees of freedom
Multiple R-squared: 0.7993,       Adjusted R-squared: 0.7881
F-statistic: 71.67 on 8 and 144 DF, p-value: < 2.2e-16
```

For the above model, Multiple R-squared value which explains the proportion of variance for a dependent variable explained by independent variables is 0.7993. As we see the p-values are below 0.05 and the beta coefficient values are significant.

Feature Engineering

As we see the variables year, cylinders are not continuous variables and in order to understand the variance properly we are converting these variables in to factor variables. And we have added another variable 'condition' to the regression model, to how it could affect the price variable. And it is found that the Multiple R-squared value is improved. And it also shows that condition good and new values are insignificant with a very high p-value. Below is the R -script and the results R-script

```
cadillac_train$cylinders=as.factor(cadillac_train$cylinders)
cadillac_train$year=as.factor(cadillac_train$year)
cadillac$condition=as.factor(cadillac$condition)
cadillac4.out=lm(price~odometer+year+cylinders+condition,data=cadillac_train)
summary(cadillac4.out)
```

Output :

```
Call:
lm(formula = price ~ odometer + year + cylinders + condition,
   data = cadillac_train)
Residuals:
    Min      1Q  Median      3Q   Max
-14142.2 -3566.6  -300.3  2735.8 25890.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.572e+04 2.141e+03    7.345 1.51e-11 ***
odometer         -9.664e-02 1.946e-02 -4.965 1.95e-06 ***
year2014          5.389e+03  1.484e+03    3.633 0.000392 ***
year2015          9.615e+03  1.658e+03    5.800 4.17e-08 ***
year2016          1.301e+04  1.872e+03    6.951 1.24e-10 ***
year2017          1.668e+04  1.886e+03    8.845 3.39e-15 ***
year2018          1.504e+04  2.487e+03    6.047 1.25e-08 ***
cylinders6        3.601e+03  1.354e+03     2.661 0.008706 **
cylinders8        2.196e+04  1.573e+03     13.958 < 2e-16 ***
conditiongood     1.227e+03 2.230e+03      0.550 0.583045
conditionlike new 1.850e+03 1.162e+03       1.592 0.113525
conditionnew      2.256e+04 4.455e+03     5.063 1.27e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5928 on 141 degrees of freedom
Multiple R-squared: 0.832,        Adjusted R-squared: 0.8189
F-statistic: 63.48 on 11 and 141 DF, p-value: < 2.2e-16
```

We checked for any probable outliers as part of the further processing, and we did so by checking at the indices whose variance is three times that of the mean. We discovered that there are two outliers, which are the 2017 and 2018 models with four and six cylinders, respectively.

R- script

```
leverage=hat(model.matrix(cadillac4.out))
```

plot(leverage, pch=19)
abline(3*mean(leverage), 0, col="red", lwd=3)
lev_index=cadillac_train[leverage>(3*mean(leverage)),]
lev_index
Output:

```
> lev_index
       region price year     make       model condition cylinders fuel odometer paint.color index
19158 hawaii 42998 2017 cadillac         ct6       new         4  gas      100       white    77
19161 hawaii 68870 2018 cadillac ats-v coupe       new         6  gas       20        grey    78
```

Once the indices of outliers are found, we trained the model by removing the outliers and ran a regression model.

R-script:
cadillac_train.final= subset(cadillac_train, index!=77 &index!=78)
cadillac5.out=lm(price~year+condition+cylinders+odometer, data=cadillac_train.final)
summary(cadillac5.out)
Output:

```
Residuals:
     Min       1Q   Median       3Q      Max
-15359.1  -3453.4   -177.9   2714.7  25350.7

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.643e+04  2.091e+03   7.859 9.29e-13 ***
year2014          5.224e+03  1.441e+03   3.624 0.000405 ***
year2015          9.470e+03  1.610e+03   5.882 2.85e-08 ***
year2016          1.289e+04  1.817e+03   7.093 5.95e-11 ***
year2017          1.732e+04  1.842e+03   9.401  < 2e-16 ***
year2018          1.359e+04  2.459e+03   5.524 1.57e-07 ***
conditiongood     1.336e+03  2.166e+03   0.617 0.538216
conditionlike new 1.959e+03  1.129e+03   1.736 0.084849 .
cylinders6        3.031e+03  1.327e+03   2.284 0.023870 *
cylinders8        2.149e+04  1.535e+03  13.998  < 2e-16 ***
odometer         -1.006e-01  1.894e-02  -5.311 4.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5755 on 140 degrees of freedom
Multiple R-squared:  0.8318,    Adjusted R-squared:  0.8198
F-statistic: 69.25 on 10 and 140 DF,  p-value: < 2.2e-16
```

So from the output we could see that the fit of the model did not improve significantly, but condition like new variable is significant with a better p-value.

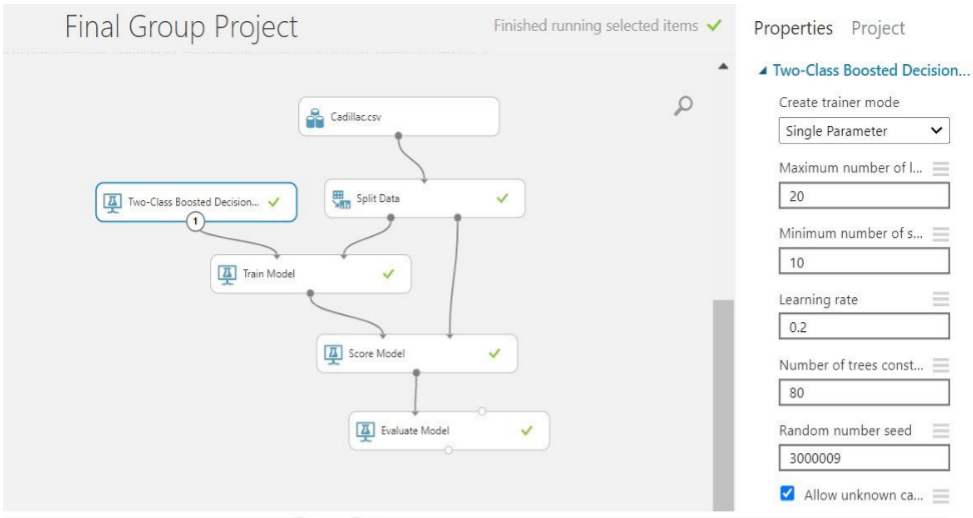
All the above models have taken year: 2013, condition: excellent, and cylinders: 4 as the base case and the models are evaluated against these conditions.

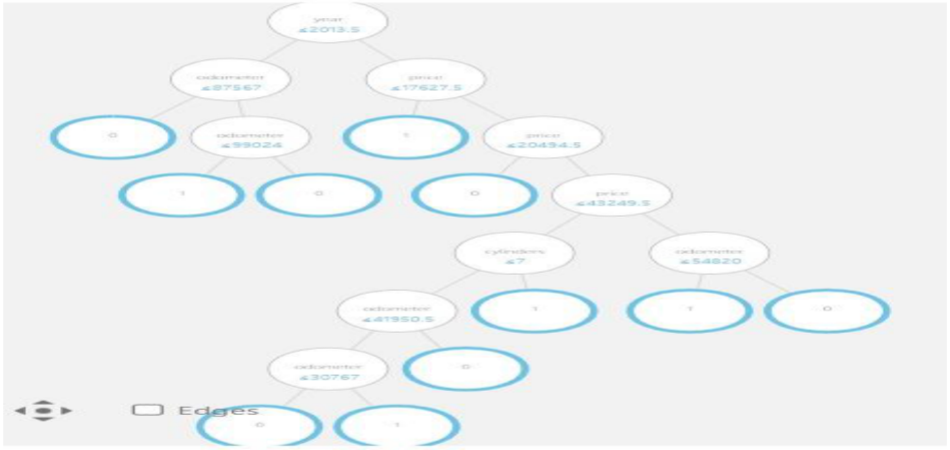
Running the prediction model

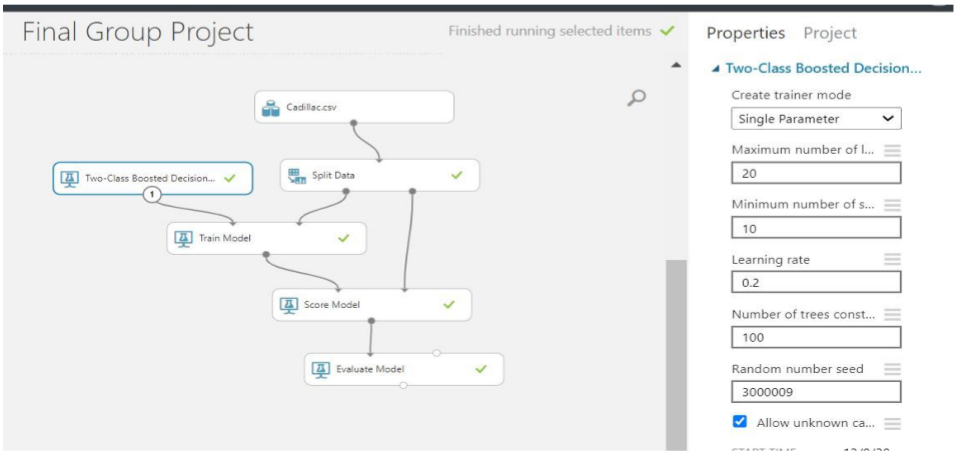
A prediction model has been run on the test data and also on the trained data, the data has been merged into single data file further analysis. A classification variable has been introduced in order to compare whether the predicted price is over priced or underpriced. Overpriced is categorized as 0 and Underpriced as 1.
predicted_model=predict(cadillac5.out, cadillac_test[,-11])
predicted_trainmodel=predict(cadillac5.out, cadillac_train.final[,-11])
cadillac_test$predicted_price= predicted_model
cadillac_train.final$predicted_price=predicted_trainmodel

Evaluation Metrics:

To understand the model better, we ran a two class boosted decision tree model in Azure ML with classification as prediction variable. We have split the data in 80:20 proportion and the number of trees is set to 80.

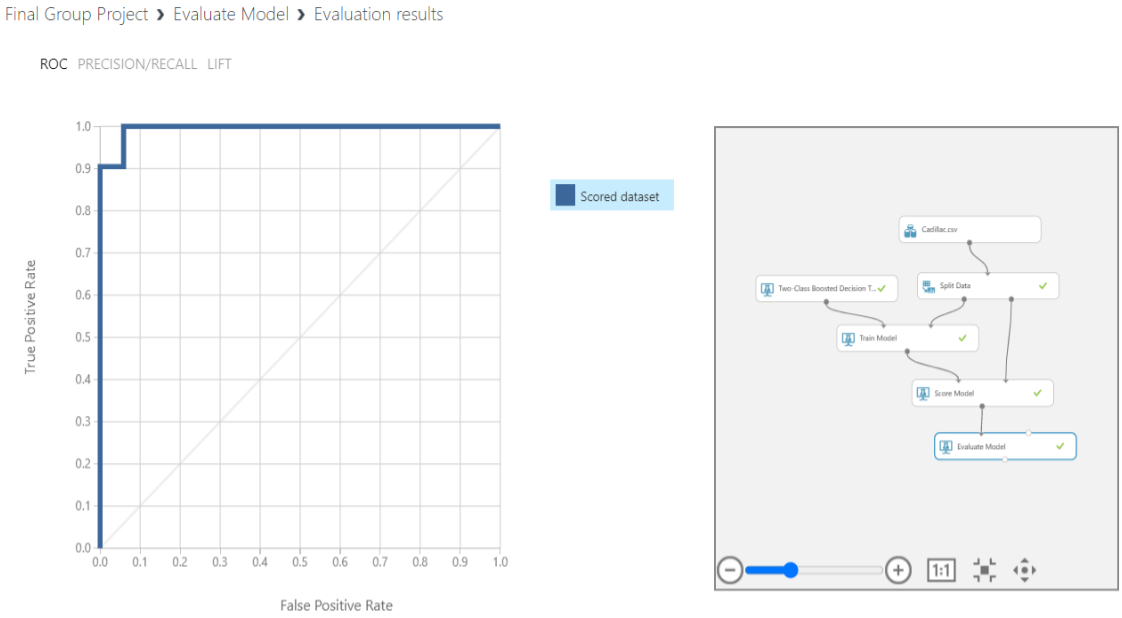






Summary sheet:

When two class boosted decision tree models have been run on the dataset with number of trees as 80 and number of trees as 100 by keeping the number of leaves as 20, we have got the below results. The results are same for both the experiments.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
21	0	0.974	0.955	0.5	0.994

False Positive	True Negative	Recall	F1 Score
1	16	1.000	0.977

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	18	0	0.474	0.921	0.923	1.000	0.857	0.850	1.000	0.000
(0.800,0.900]	1	1	0.526	0.921	0.927	0.950	0.905	0.889	0.941	0.053
(0.700,0.800]	2	0	0.579	0.974	0.977	0.955	1.000	1.000	0.941	0.053
(0.600,0.700]	0	0	0.579	0.974	0.977	0.955	1.000	1.000	0.941	0.053
(0.500,0.600]	0	0	0.579	0.974	0.977	0.955	1.000	1.000	0.941	0.053
(0.400,0.500]	0	2	0.632	0.921	0.933	0.875	1.000	1.000	0.824	0.171
(0.300,0.400]	0	0	0.632	0.921	0.933	0.875	1.000	1.000	0.824	0.171
(0.200,0.300]	0	0	0.632	0.921	0.933	0.875	1.000	1.000	0.824	0.171
(0.100,0.200]	0	1	0.658	0.895	0.913	0.840	1.000	1.000	0.765	0.230
(0.000,0.100]	0	13	1.000	0.553	0.712	0.553	1.000	1.000	0.000	0.994

Conclusion:

From the above results we could see that the accuracy is 0.974 and precision is 0.955, which denotes that the predicted model is significant.

The explanation for confusion matrix is as follows.

True Positive: It denotes that how many cases were identified as underpriced and truly underpriced.
True Negative: It denotes about how many cases were identified as overpriced and are actually overpriced.

False Positive: It denotes about how many cases were identified as underpriced but are actually overpriced.
False Negative: It denotes that how many cases were identified as overpriced but are actually underpriced.

Hence from the results of our model, we can suggest that the predicted outcomes from multiple linear regression and evaluation of results using two class boosted decision tree is significant, which further helps the user to price the used car according to the predicted model.

Recommendation:

For our model the accuracy obtained is 97.4% and precision is 97.5% the multiple R-squared value for the multiple linear regression model is 0.8318 and the adjusted R squared is 0.8198 which seems to be like there is no over fitting.

Because of the nature of this project, multiple algorithms can be merged as modules and their outputs can be mixed to improve the accuracy of the final result. These algorithms' output, however, must be in the same format as the others. The modules are simple to add once that criterion is met, as seen in the code. The project gains a lot of modularity and versatility as a result of this.

The dataset contains further room for development. When the size of the dataset is increased, the algorithms' precision improves. As a result, more data will almost certainly improve the model's accuracy in predicting the saleable price.