Don’t Get Kicked

Ananya Suresh Kumar

Ivy Tien

Siddharth Singh

Kaushal Desai

Adityadev Singh

New York University – Stern School of Business, Data Mining for Business Analytics

1. INTRODUCTION

When you visit an auto dealership to purchase a used vehicle, the most important thought on your mind is whether you are being offered a good deal or not. A good deal is probably one where the car appears to be in reliable working condition and its price tag seems reasonable. The auto dealership itself goes through a similar thought process when it acquires the used vehicles that are sold to its customers. It serves as an intermediary between first time/used vehicle owners and people interested in purchasing these vehicles on the secondary market.

The vehicles are acquired through a variety of sources. An important segment of these acquisitions are auto auctions that contribute to roughly 26% of the total number. Since car dealerships are active players on the auction market, one may think that they have a fair idea of what they are buying. However, they often end up with a raw deal. This can happen due to a variety of reasons – mechanical or electrical issues with the cars that are not detected at the time of purchase, tampered components such as odometers, issues with the paperwork associated with the vehicle. Such vehicles are referred to as 'kicked' vehicles. These vehicles either have to be scrapped altogether by the dealership or significant amounts have to be spent on repairing them before they are put up for re-sale. Either way, the car dealership ends up losing a lot of money. Also, if these issues go undetected and the kicked cars are sold to their customers, it opens up the dealership to a loss of reputation for putting out a bad product in the market.

The motive of this study was to analyze whether the car dealerships could nip this problem in the bud by leveraging the power of data analytics to make informed decisions at the time of car purchase. 8.6 million cars were sold in the used car market through auto auctions last year out of which roughly 12% ended up being kicked cars. Therefore, the monetary impact of this study to the used vehicle industry could be in the range of hundreds of million dollars and potentially even in billions.

We split the report into three major sections: Methodology, Analysis, and Insights/Conclusion. The section on Methodology deals with the sources of our data and our data dictionary. As part of our analysis, we first tried to understand the impact of our predictions on the car dealership. We then proceeded to perform a visual analysis of the data. This was followed by modeling in Azure using 6 different machine learning algorithms. We selected the best two models - Decision Forest and Decision Jungle and ran further optimizations. Once this was complete, we performed modeling in Rattle to see if any insight was missed by our Azure models. Based on our analysis, we drew up the top insights and proceeded to visually verify if they were meaningful. We concluded by stating our findings and suggesting the parameters that car dealers should keep in mind when buying cars at an auto auction.

2. METHODOLOGY

2.1. Data Source

The dataset we used is hosted by Kaggle in a competition called 'Don't Get Kicked! The data was collected by Carvana, a technology start-up based in Phoenix, whose mission is to make car buying better by bringing technology, transparency, and exceptional customer service to the car buying process.

2.2. Data Dictionary

The target variable is 'IsBadBuy', which has two possible values: Yes and No. We consider 'Yes' as the positive case (1) meaning the car is kicked and 'No' as the negative case (0) meaning car is a good buy. The dataset includes 72983 records and 32 predictor variables (for the full data dictionary, see Appendix A), but the variables we used to build our model include the following:

Vehicle Attributes (Vehicle Age, Make, Model, Submodel, Trim, Color, Size, Transmission, Wheel Type, VehOdo)

Price Attributes (VehBCost, Warranty Cost, MMR Acquisition Auction Average/Clean Price, MMR Acquisition Retail Average/Clean Price, MMR Current Auction Average/Clean Price, MMR Current Retail Average/Clean Price)

Special Features (PRIMEUNIT, AUCGUART)

Other Attributes (VNST, Auction, IsOnlineSale, Nationality, Top3, VNZIP1)

After exploring the data, we assessed several needs for transformation, imputation, and exclusion. For example: Replacing NULL values present in the data with blanks. Also, some cars had NULL values in Nationality, Top3 and Make fields, which were duly replaced by, correct data by comparing cars with matching model. We transformed IsBadBuy, IsOnlineSale and VNZIP1 from Numerical to Categorical fields. RefID was retained as an identity field using Clear feature setting of Metadata Editor in Azure. The dataset was split 80/20.

Following predictors were dropped from our analysis –
PurchDate – Narrow range indicates no significant value to our prediction
VehYear, WheelTypeID – Correlated to VehicleAge and WheelType respectively
BYRNO – It gives information about car buyer and hence not useful in our analysis

3. ANALYSIS

The analysis was split into four sequential phases – Monetary Impact, Data Exploration, Data Modeling, and Prediction Analysis.

3.1. Monetary Impact

The first phase in our analysis was aimed at understanding the monetary impact of the different predictions our models would make. This was necessary to ascertain the decisive metrics for evaluating model performance. We calculated the average cost of vehicle purchase from our data to be approximately $7000. The re-sale price was estimated at $8700. In case the purchased car was kicked, we estimated the repair costs to be $3000. Since there was no information available on the percentage of kicked cars that were scrapped completely, we assumed all cars could be repaired. Based on the above information, we estimated a profit of $1700 for a good purchase by the dealership. In case of a kicked car, the dealer has to pay for the purchase price plus repair costs that worked out to $10000 and then sold it for $8700 resulting in a loss of $1300. The impact of our predictions were tabulated as follows –

Actual \ Predicted	Good Buy	Bad Buy
Good Buy	$1,700	($1,700)
Bad Buy	($1,300)	$1,300

Note: Figures in red indicate losses

If our models correctly predicted good buys or bad buys, it would have a positive monetary impact for the dealership. But if the models incorrectly predicted a car as a bad buy when it was actually good (false negative), the dealership would lose out on the opportunity of earning $1700. Also, if our models incorrectly predicted a car purchase as a good buy when it was actually bad (false positive), the dealership would lose $1300. Since both the incorrect prediction scenarios resulted in significant losses, we needed to select a metric that gave weightage to both - a false positive (FP) and a false negative (FN).

The metrics of choice to satisfy this condition were –

a. F1 score - 2TP/(2TP+FP+FN)
b. AUC – Sensitivity/(1 – Specificity)

3.2. Data Exploration

Before building our models, we explored the data visually using Excel and Tableau.

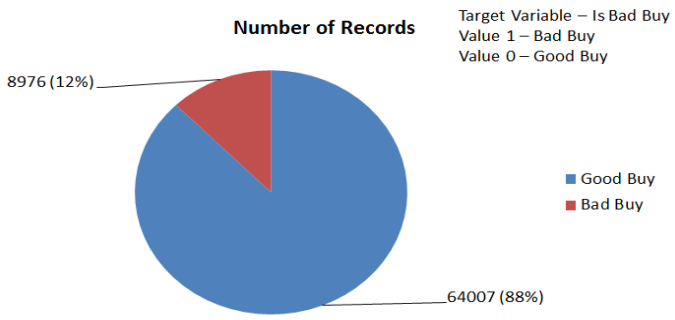

Figure 1: Distribution of Good buys and Bad Buys

88% of all records were classified as Good purchases while 12% were classified as bad purchases or kicked cars as visualized in Figure1.

We identified the sources of the data procured by Carvana in Figure2. We came to know that over half the data was being sourced from Manheim auction house, about a fifth of it was from Adesa auction house and the remaining data was clubbed from other sources.

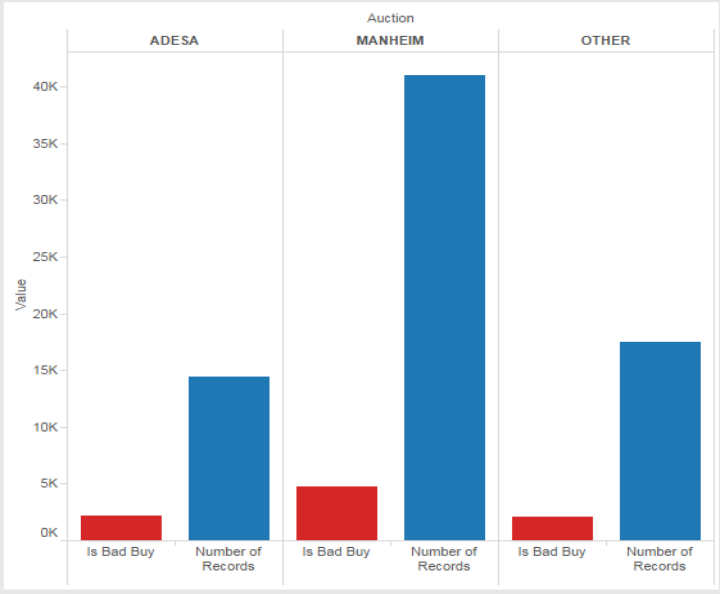

Figure 2: Records and Bad Buys by Auction house

Finally, we identified Texas, Florida, California and North Carolina as some of the states with the biggest auction markets for vehicles.

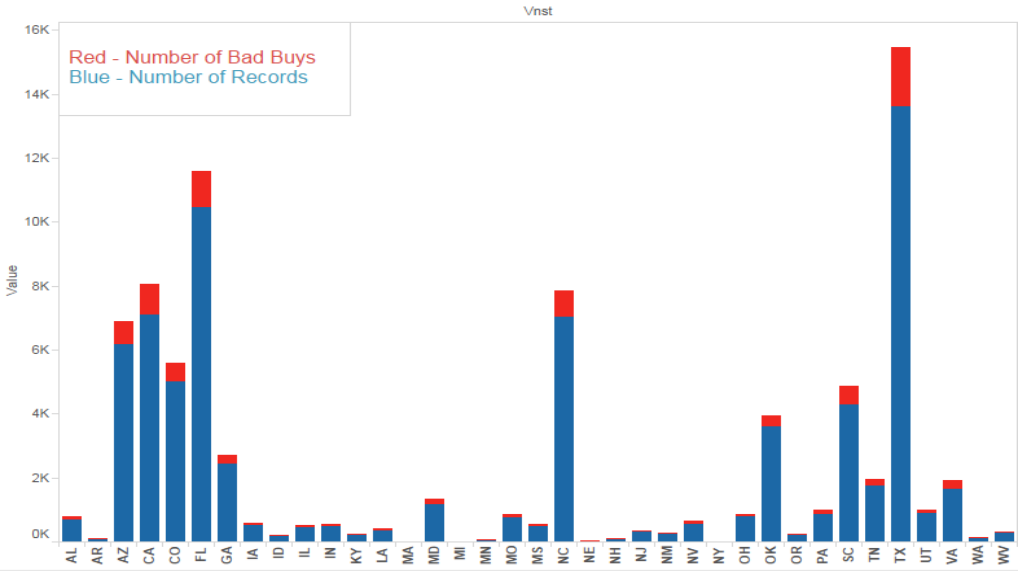

Figure 3: State-wise distribution of records and bad buys

3.3. Data Modeling

Now that we had an idea of the data we were dealing with, we decided to use Azure as our primary modeling tool since we had a number of categorical predictors with more than 32 levels. Filter Based Feature selection with Chi squared scoring method was used to select the top 20 features.

Since our data was heavily skewed towards good buys, we used SMOTE to oversample Bad buys with a SMOTE percentage of 300.

In the initial iterations, we ran six different models –Logistic Regression, Decision Forest, Boosted Decision Tree, Decision Jungle, Support Vector Machine, and Neural Network.

The results are listed as follows –

Model	TP	FP	TN	FN	Precision	Recall	Accuracy	Specificity	F1 Score	AUC
Logistic Regression	4	4	12212	1286	0.500	0.003	0.904	1.000	0.006	0.695
Decision Forest	517	474	12352	1254	0.522	0.292	0.882	0.963	0.374	0.741
Boosted Decision Trees	220	676	11552	1070	0.246	0.171	0.871	0.945	0.201	0.677
Decision Jungle	528	424	12402	1243	0.555	0.298	0.886	0.967	0.388	0.736
Support Vector Machines	20	21	12195	1270	0.488	0.016	0.904	0.998	0.030	0.602
Neural Networks	1287	12184	44	3	0.096	0.998	0.098	0.004	0.174	0.256

Table 1: Model Performance Evaluation

We noticed that two of the models – Decision Forest and Decision Jungle were faring much better than the others based on AUC and F1 score. Therefore, we decided to optimize the results for these two models.

Predictors selected in our final analysis were – Auction, Vehicle Age, Make, Model, Color, Transmission, WheelType, VehicleOdo, PRIMEUNIT, AUCGUART, VNST, VehBCost, IsOnlineSale, WarrantyCost

Filter Based Feature Selection was dropped since we had eliminated the highly correlated features leaving only 15 predictors.

The optimized model results were –

Model	TP	FP	TN	FN	Precision	Recall	Accuracy	Specificity	F1 Score	AUC
Decision Forest	548	498	12328	1223	0.524	0.309	0.882	0.961	0.389	0.739
Decision Jungle	462	245	12581	1309	0.653	0.261	0.894	0.981	0.373	0.756

Table 2: Optimized modeling

We used Permutation Feature Importance to observe the most important predictors. The top two features came out to be WheelType and Make.

We decided to run the Decision Tree model in Rattle to see if we were missing any important feature. The top node came out to be Vehicle Age as shown in Figure 4.

```
1) root 58386 7126 [0,0] (0.877950194 0.122049806)
  2) VehicleAge< 4.5 35586 3060 [0,0] (0.914011128 0.085988872)
```

Figure 4: Top nodes of Decision Tree in Rattle

3.4. Prediction Analysis

We performed a visual analysis of the three important predictors to confirm if they could be used by car dealerships.

First, we analyzed the distribution of kicked cars by Age. As seen in Figure 5, vehicles older than 4 years exceeded the average percentage of Bad buys which was 12%. The dealers would therefore be advised to exercise caution when purchasing cars older than 4 years.

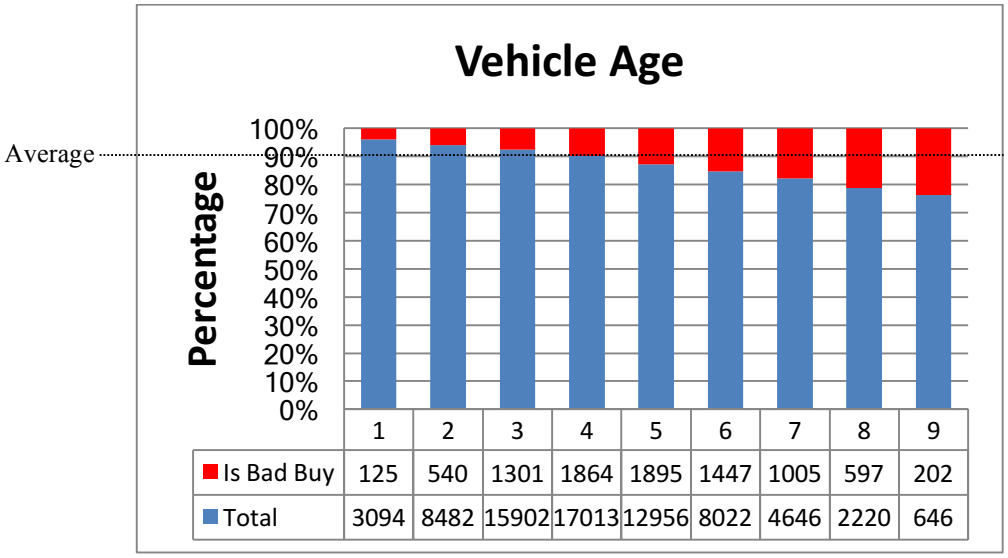

Figure 5: Impact of Vehicle Age on Bad buys

The next feature that we visualized was Wheel Type. Wheels in our data were classified into three categories – Alloy, Covers, and Special. There were 3174 records in which data was missing for this field. An astonishing 70% of these records corresponded to bad buys as is visible in Figure 6. A possible explanation for the same has been mentioned in our conclusion.

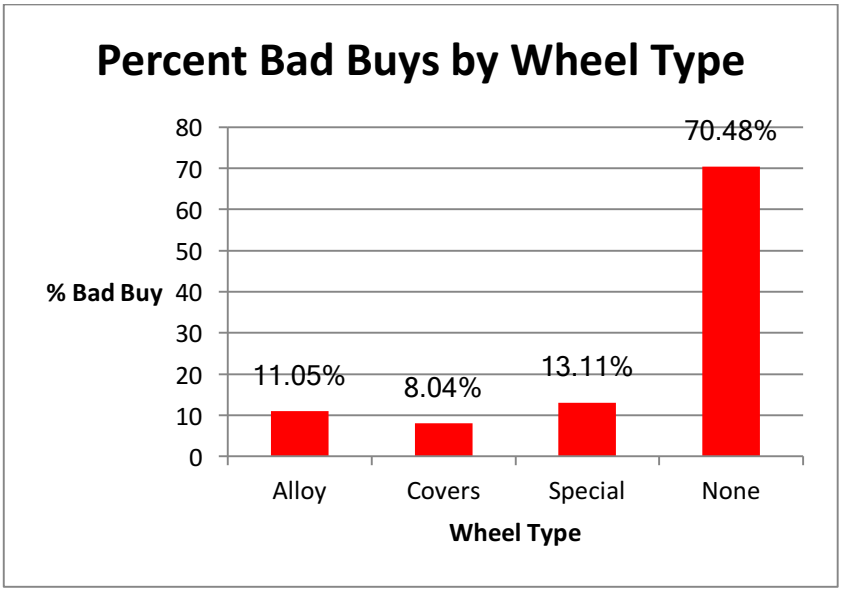

Figure 6: Percentage Bad buys by Wheel Type

On analyzing the impact of Vehicle Make on bad buys, we observed that there were differing trends on the brands that contributed the largest absolute number of bad buys and the brands that had the highest percentage of bad buys of the total cars sold for that specific brand. Ford, Chevrolet, Dodge, and Chrysler had the highest absolute number of kicked cars. Plymouth, Lexus, Infiniti and Mini had the highest percentage of kicked cars, but a very small representation in absolute numbers as visualized in Figure 7.

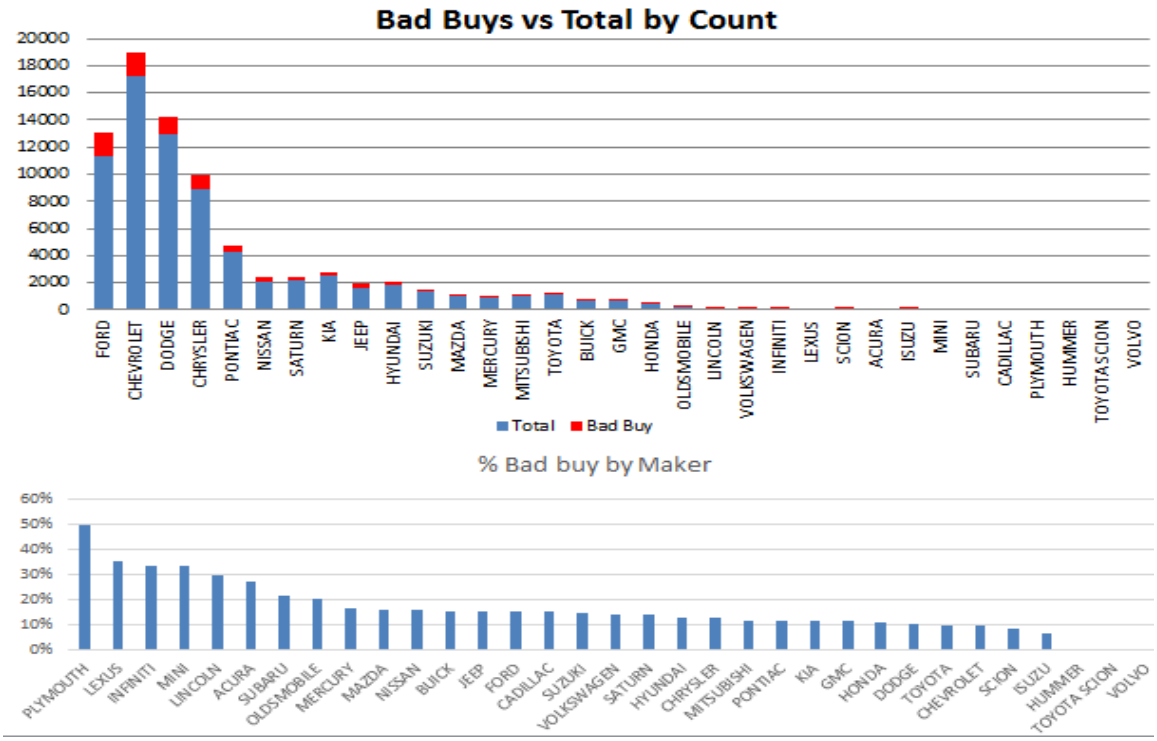

Figure 7: Bad Buys by Brand

4. CONCLUSION

Based on both Rattle and Azure modeling results, we learned a few business insights about vehicle auction for used car dealers. First of all, the age of the used vehicle plays an important role in predicting whether the vehicle might be a kicked car or not. Since older cars most likely also have higher mileage, older engines, and more outdated technical designs, this finding seems to be reasonable. We can also tell from the dataset that older used vehicles have lower average acquisition cost (cars less than 1 year old have an average acquisition cost of $8,890 whereas cars more than 7 years old have an average acquisition cost of less than $6,000). When we analyzed the number of used cars sold, the result shows a much lower number of cars sold if the vehicle is over 7 years old. Considering these two results together, it seems that dealers typically are not willing to pay too much for older vehicles and that fewer sellers are interested in selling since the prices will be low. This finding also explains why in the retail used car market, older cars typically have lower retail value.

The second important business insight we learned from our models is that brand is also a significant predictor for kicked cars. This result is logical and expected because certain brands tend to produce higher quality cars. In addition, owners of cars with higher quality (and probably higher price tag) are more likely to take care of their vehicles. The dataset shows that certain makers, such as Chevrolet and Dodge, sold a lot more used cars than other brands while they both had around average percentage of bad buys. It seems that these two brands are better choices when making purchases than others. On the other hand, Nissan and Jeep both had slightly above average bad buy percentages and much fewer numbers of used cars sold at these auctions. This result could mean that dealer didn't want to over-pay for these brands because they expected more problems from these brands. Finally, our models show correlation between brand, model and sub-model. Therefore, model and sub-model are also important factors to consider when buying used cars for similar reasons.

One interesting finding from our model is the number of bad buys when the information of wheel type is missing from our dataset. We think the missing information could either mean that wheels were missing from the cars or that those car wheels were badly damaged. In either case, those cars were probably in bad condition. Even if car dealers could purchase those cars at very cheap prices, they could end up in losses because of the repair costs they had to incur to resell those cars.

Overall, we think this dataset provides very interesting and important information for used car dealers as well as any car buyer. By knowing the important predictors for kicked cars, car dealers can make better purchase choices and avoid more bad buys. Kicked cars cost car dealers significant amount of money not only because of investment wasted but also potential profits lost had they purchased good cars. For used car dealers, a quick turn-around at good retail price is the only way they can make good profits considering other expenses they have to incur to keep their business going. Kicked cars require both time and money to be in re-sale condition and the retail prices could fall while kicked cars are being repaired. With the insights from the dataset and our model, we can suggest that dealers should pay special attention to older vehicles and certain brands if other conditions of the used vehicles are similar. Dealers can also avoid over-paying for vehicles that are more likely to be bad buys and perhaps reduce the losses if these cars turn out to be kicked cars.

APPENDIX A: DATA DICTIONARY

Field Name	Definition
RefID	Unique (sequential) number assigned to vehicles
IsBadBuy	Identifies if the kicked vehicle was an avoidable purchase
PurchDate	The Date the vehicle was Purchased at Auction
Auction	Auction provider at which the vehicle was purchased
VehYear	The manufacturer's year of the vehicle
VehicleAge	The Years elapsed since the manufacturer's year
Make	Vehicle Manufacturer
Model	Vehicle Model
Trim	Vehicle Trim Level
SubModel	Vehicle Submodel
Color	Vehicle Color
Transmission	Vehicles transmission type (Automatic, Manual)
WheelTypeID	The type id of the vehicle wheel
WheelType	The vehicle wheel type description (Alloy, Covers)
VehOdo	The vehicles odometer reading
Nationality	The Manufacturer's country
Size	The size category of the vehicle (Compact, SUV, etc.)
TopThreeAmericanName	Identifies if the manufacturer is one of the top three American manufacturers
MMRAcquisitionAuctionAveragePrice	Acquisition price for this vehicle in average condition at time of purchase
MMRAcquisitionAuctionCleanPrice	Acquisition price for this vehicle in the above Average condition at time of purchase
MMRAcquisitionRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition at time of purchase
MMRAcquisitonRetailCleanPrice	Acquisition price for this vehicle in the retail market in above average condition at time of purchase
MMRCurrentAuctionAveragePrice	Acquisition price for this vehicle in average condition as of current day
MMRCurrentAuctionCleanPrice	Acquisition price for this vehicle in the above condition as of current day
MMRCurrentRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition as of current day
MMRCurrentRetailCleanPrice	Acquisition price for this vehicle in the retail market in above average condition of current day
PRIMEUNIT	Identifies if the vehicle would have a higher demand than a standard purchase
AcquisitionType	Identifies how the vehicle was acquired (Auction buy, trade in, etc.)
AUCGUART	The level guarantee provided by auction for the vehicle (Green light - Guaranteed/arbitratable, Yellow Light -

	caution/issue, red light - sold as is)
BYRNO	Unique number assigned to the buyer that purchased the vehicle
VNZIP	Zip code where the car was purchased
VNST	State where the car was purchased
VehBCost	Acquisition cost paid for the vehicle at time of purchase
IsOnlineSale	Identifies if the vehicle was originally purchased online
WarrantyCost	Warranty price (term=36month and millage=36K)