2019/2020

Professors: Dino Pedreschi, Anna Monreale

Data Mining Project, Data analysis on CarVana Dataset

Giacomo Ettore Rocco, Giuseppe Valentini, John Bianchi

Pisa 2019/20

University of Pisa
Department of Computer Science

1 Introduction

The project consists in exercises that require the use of data mining tools for analysis of data. Is been performed by using Python, especially the Pandas, Sklearn, Numpy libraries. The results of the different tasks is been reported in a unique paper.
Our analysis is based on a Dataset of the company "Carvana", an online dealer of used cars. Our aim is to predict whether a future purchase can be considered "good" or "bad". Initially there is a data understanding phase, in which the Dataset is manipulated, in terms of missing values management and understanding of the various attributes, which will then be used later for clustering analysis, using algorithms such as K-means, DBSCAN and Hierarchical. Then proceed with the association rules and classification.

2 Data Understanding

To perform the analysis of the models, in order to predict whether a car is a bad buy or a good buy, we need to understand how the data present in our training set are distributed. We will present a study of the records of the Carvana dataset, witch contains 58385 records with 34 different features. We will explain what every attributes stand for in Table 1.

2.1 Data Semantics

The columns "PRIMEUNIT" and "AUCGUART" are almost all composed by null elements, like it's shown in the table, for this reason we have chosen to completely cut away those columns, because we considered them just useless for the purposes of the project. "AcquisitionType" and "KickDate", that are described in the dictionary, are not present in the dataset. Transmission had two distinct values, "MANUAL" and "Manual", which referred to the same type of transmission, and we replaced them with the same value using "MANUAL".

Table 1: Feature Descriptions

Name	Type	Description	Domain	Missing values
RefId	Numeric, discrete	Unique (sequential) number assigned to vehicles	0 >N <73015	0,00%
IsBadBuy	Boolean	Identifies if the kicked vehicle was an avoidable purchase	{0 =positive ,1=negative}	0,00%
PurchDate	Numeric, data	The Date the vehicle was Purchased at	2009 to 2010	0,00%
Auction	Categorical	Auction provider at which the vehicle was purchased	{'ADESA', 'MANHEIM', 'OTHER'}	0,00%
VehYear	Numeric, discrete	The manufacturer's year of the vehicle	2000 >N <2011	0,00%
VehicleAge	Numeric, discrete	The Years elapsed since the manufacturer's year	0 >= N <10	0,00%
Make	Categorical	Vehicle Manufacturer	|Make| = 23	0,00%
Model	Categorical	Vehicle Model	|Model| = 779	0,00%
Trim	Categorical	Vehicle Trim Level	|Trim| = 117	3.2%

SubModel	Categorical	Vehicle Submodel	|SubModel| = 585	0.01%
Color	Categorical	Vehicle Color	|Color| = 17	0.01%
Transmission	Categorical	Transmission type	Auto, Manual	0.01%
WheelTypeID	Numeric, discrete	Id of the vehicle wheel	0 >= N <= 3	4,00%
WheelType	Categorical	The vehicle wheel type description	|WheelType| = 3	4,00%
VehOdo	Numeric	The vehicles odometer reading	4825<=N<=115717	0,00%
Nationality	Categorical	The Manufacturer's country	|Nationality| =4	0.006%
Size	Categorical	The vehicle's size category	|Size| = 12	0.006%
TopThreeAmericanName	Categorical	Identifies if the manufacturer is one of the top three American manufacturers	{'OTHER', 'FORD', 'CHRYSLER','GM'}	0.006%
MMRAcquisitionAuctionAveragePrice	Numeric, continuous	Acquisition price for this vehicle in average condition at time of purchase	0.0 <= N <= 35722.0	0.02%
MMRAcquisitionAuctionCleanPrice	Numeric, continuous	Acquisition price for this vehicle in the above Average condition at time of purchase	0.0 <= N <=36859.0	0.02%
MMRAcquisitionRetailAveragePrice	Numeric, continuous	Acquisition price for this vehicle in the retail market in average condition at time of purchase	0.0 <= N <=39080.0	0.02%
MMRAcquisitionRetail CleanPrice	Numeric, continuous	Acquisition price for this vehicle in the retail market in above average condition at time of purchase	0.0 <= N <=41482.0	0.02%
MMRCurrentAuctionAveragePrice	Numeric, continuous	Acquisition price for this vehicle in average condition as of current day	0.0 <= N <=35772.0	0.4%
MMRCurrentAuctionCleanPrice	Numeric, continuous	Acquisition price for this vehicle in the above condition as of current day	0.0 <= N <=36859.0	0.4%
MMRCurrentRetailAveragePrice	Numeric, continuous	Acquisition price for this vehicle in the retail market in average condition as of current day	0.0 <= N <=39080.0	0.4%
MMRCurrentRetailCleanPrice	Numeric, continuous	Acquisition price for this vehicle in the retail market in above average condition as of current day	0.0 <= N <=41462.0	0.4%
PRIMEUNIT	Boolean	Identifies if the vehicle would have a higher demand than a standard purchase	{'yes', 'no'}	95.4%
AcquisitionType	-	Identifies how the vehicle was acquired		100,00%
AUCGUART	Categorical	The level guaranteed provided by auction for the vehicle	{'Green','Yellow', 'Red'}	95.4%
KickDate	-	Date the vehicle was kicked back to the auction		100,00%
BYRNO	Numeric, discrete	Unique number assigned to the buyer that purchased the vehicle	|BYRNO| = 70	0,00%
VNZIP1	Numeric, discrete	Zip-code where the car was purchased	|VNZIP1| =142	0,00%
VNST	Categorical	State where the the car was purchased	|VNST| = 37	0,00%
VehBCost	Numeric, continuous	Acquisition cost paid for the vehicle at time of purchase	N >0	0,00%
IsOnlineSale	Boolean	Identifies if the vehicle was originally purchased online	{0,1}	0,00%
WarrantyCost	Numeric, discrete	Warranty price	N >0	0,00%

2.2 Distribution of the variables and statistics

A useful way to evaluate the frequency distribution of the features is plotting the density distributions. We made histograms using for the binning the Sturges rule:

$$Bins = 1 + 3.3 \log n \tag{1}$$

where n is the number of records. The red line represents the mean μ of the values, the yellow lines represent the standard deviation σ, and the green ones the outliers, considered in the range $[\mu \pm 3\sigma]$.

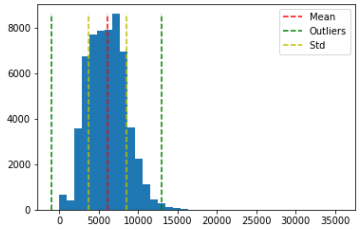

Figure 1: MMRAcquisitionAuctionAveragePrice

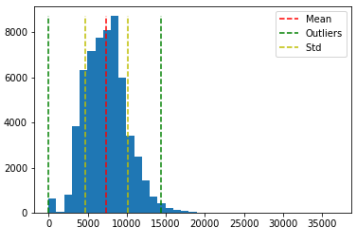

Figure 2: MMRAcquisitionAuctionCleanPrice

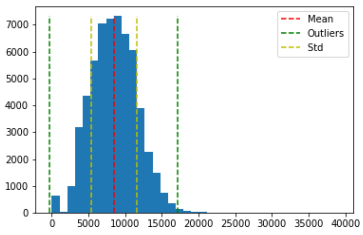

Figure 3: MMRAcquisitionRetailAveragePrice

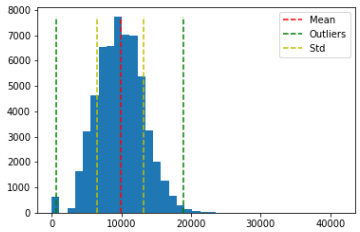

Figure 4: MMRAcquisitonRetailCleanPrice

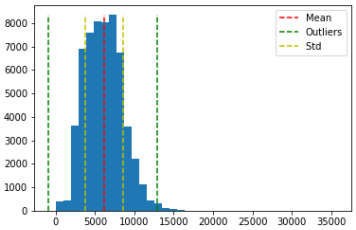

Figure 5: MMRCurrentAuctionAveragePrice

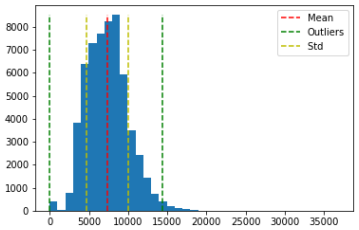

Figure 6: MMRCurrentAuctionCleanPrice

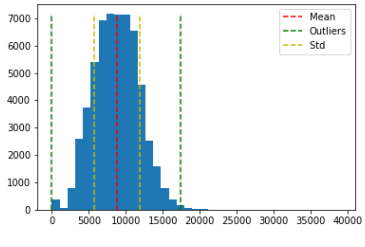

Figure 7: MMRCurrentRetailAveragePrice

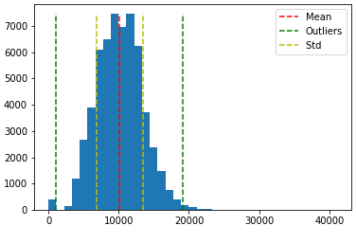

Figure 8: MMRCurrentRetailCleanPrice

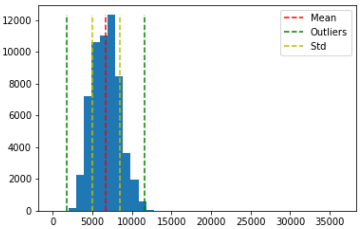

Figure 9: VehBCost

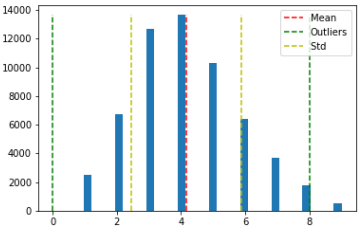

Figure 10: VehicleAge

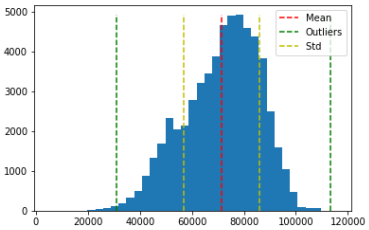

Figure 11: VehOdo

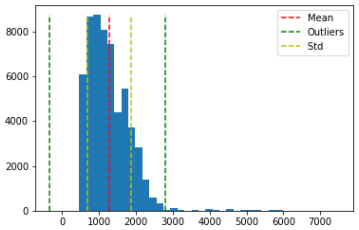

Figure 12: WarrantyCost

Table 2: Feature Distributions, shows a summary of the main statistics about numeric features.

Attributes	Mean	Std	Min	Max	25' Perc.	50' Perc.	75' Perc.
MMRAcquisitionAuction AveragePrice	6.128,1279	2456,6326	0.0	35722	4273	6097	7765
MMRAcquisitionAuction CleanPrice	7.372,9126	2715,5064	0.0	36859	5409	7308	9017
MMRAcquisition ReatailAveragePrice	8.497,2885	3151,1062	0.0	39080	6279	8448	10652
MMRAcquisitionRetail CleanPrice	9.851,7680	3378,8396	0.0	41482	7501	9798	12084
MMRCurrentAuction AveragePrice	6.131,6666	2432,1715	0.0	35722	4275	6063	7737
MMRCurrentAuction CleanPrice	7.389,9586	2682,3108	0.0	36859	5415	7311	9014
MMRCurrentRetailAve ragePrice	8.776,0651	3086,3737	0.0	39080	6538	8733	10910
MMRCurrentRetail CleanPrice	10.145,2270	3304,637	0.0	41062	7788	10103	12309
VehBCost	6.730,0083	1762,0752	1.0	36485	5430	6700	7900
VehOdo	71.478,0905	14591,2245	4825	115717	61785	73359	82427
VehYear	2.005,3446	1,7333	2001	2010	2004	2005	2007
VehicleAge	4,1749	1,7138	0	9	3	4	5
WarrantyCost	1.276,1050	598,8854	462	7498	5430	1155	1623

2.3 Correlation

After analyzed features individually, we proceeded, with the aim to reduce the dimensionality, analyzing how the feature are correlated each other. For performing this scope, we used Pearson correlation and Spearman correlation, the graphic about the correlation is similar, for that reason we showed just one of them with the graphic below.

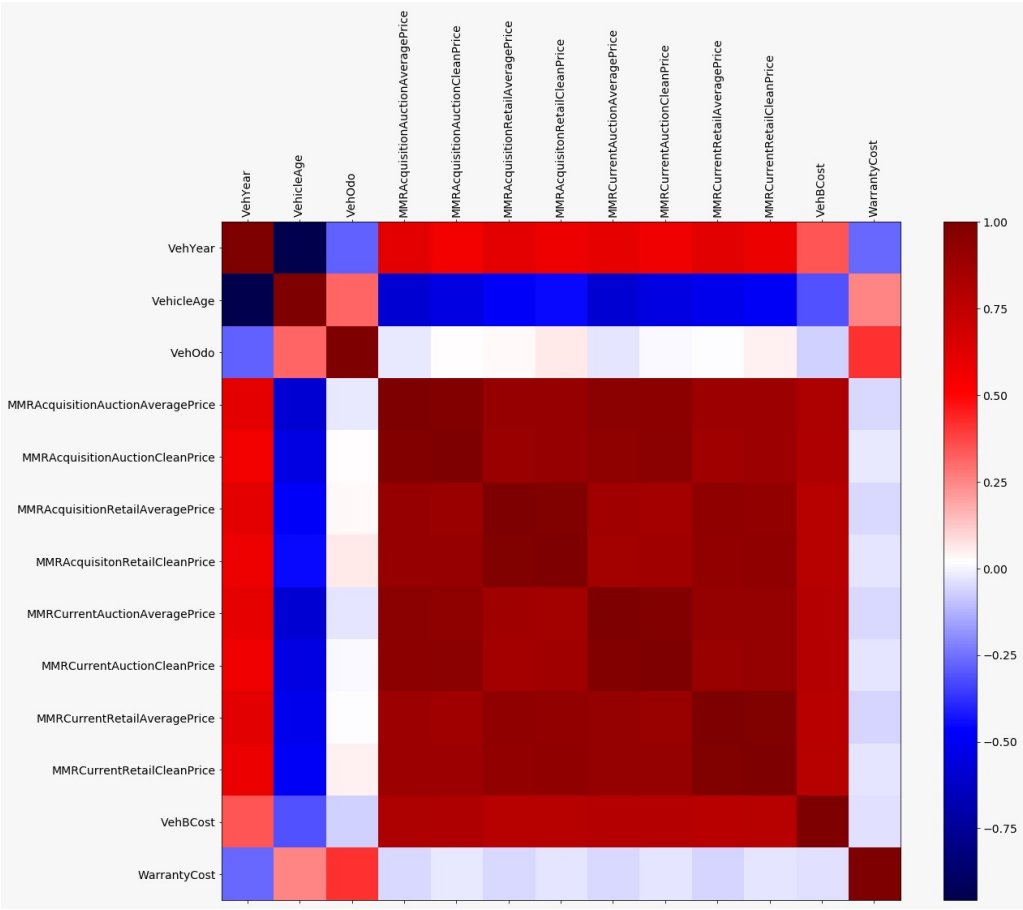

Figure 13 : Correlation matrices according to Pearson.

This correlation shows that all the MMR* cost, are strongly correlated. The strongest correlation is pair to pair, so that, we choose to use only 4 of the 8 values of MMR* without practically losing information. The two features "WheelType" and "WheelTypeID" is referring about the same information concern the type of pneumatic. We choosed to maintain "WheelType" without losing information.

"WarrantyCost" is indipendent from all the other MMR* cost as the graph below shows. For this reason we can assume that the MMR* cost are not influenced by the warranty cost.

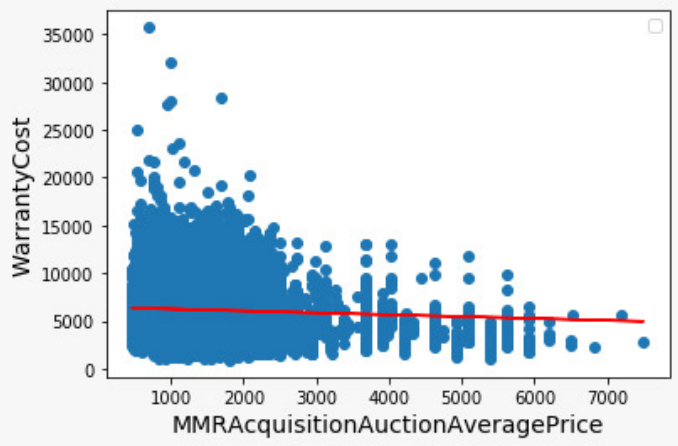

Figure 14: Linear regression between MMRAcquisitionActionnAveragePrice and WarrantyCost.

We can say, like for WarrantyCost, that "VehOdo" is independent from MMR* cost. This is a notable thing, because implies that cars's cost are independent from its distance made.

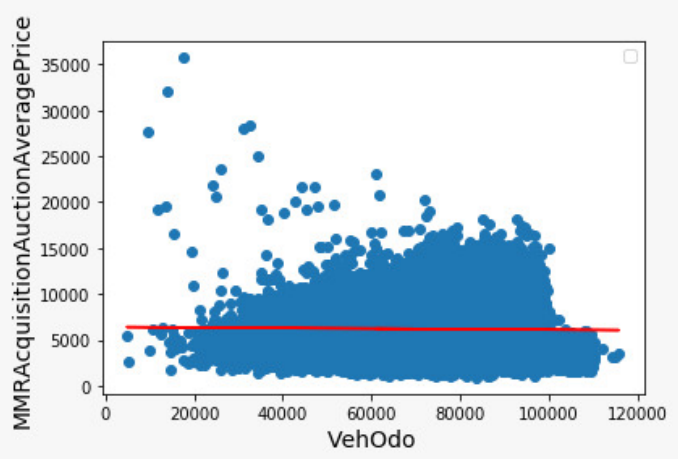

Figure 15: Linear regression between VehOdo and MMRAcquisitionActionnAveragePrice

"VehicleAge" and "VehYear" are negatively correlated. They represent in fact, both, the age of the vehicle. So one of the two is droppable without losing information, we choose to maintain "VehicleAge", that is a reusable data, independent from the current year, but is just how the vehicle is old in the moment of purchase.

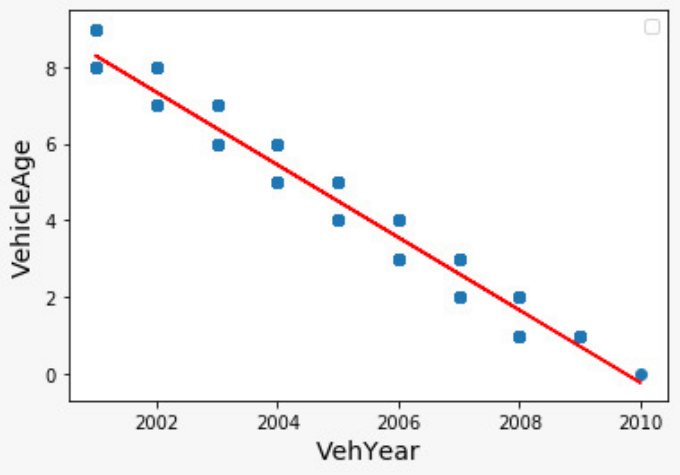

Figure 16: Linear regression between VehicleAge and VehYear. Notice the negative slope of the line: an increasing year implies a decreasing age.

2.4 Data transformations and missing values management

In Carvana dataset are present "Missing Values", there are multiple reasonable ways to address this problem, that are dependent by the nature of the feature and the number of missing values in that features. If the value of the feature is numerical, an approach can be filling missing values with the mean or the mode of the feature. Another way is retrieving the information analyzing records that are the same values/structure in the other features. We studied the different type of missing values and filling that in various ways:

- In "Transmission" we found 0.01% of missing values, we managed this values, seeing that for the 90% of the case one model have only one type of transmission, so that we could fill missing values using models: another record with the same model has the same type of transmission, the final result that every missing value is being filled with "AUTO".
- "Trim" feature has the 3.2% of missing values, is being setted with "bas", that means that is the base version of a car.
- "Color" has the 0.01% of missing values, filled with "NOT AVAIL".
- "SubModel" has the 0.01% of missing values, being filled with "Not Available".
- "WheelType" has the 4% of missing values, they are been managed using the mode of the WheelType's submodel of the record.
- In "Nationality" we found just 4 missing values and exploiting "Make". We filled the values.
- Just 4 missing values for "TopThreeAmericanName", so also in this case we exploited "Make".
- In MMR* values we found values with the value 0. We treated this values like missing values. We filled missing values using the mean, this for every MMR*, obviously mean values is being calculated excluding the zero values.

3 Clustering Analysis

In this section we analyze the dataset through clustering algorithms, such as K-Means, DB-SCAN and Hierarchical clustering. The attributes used are only the numerical ones.

3.1 K-means

The K-means algorithm is a clustering algorithm that allows you to divide a set of objects into K groups based on their attributes. The first step is the study of the value of K, that is the number of clusters in which you want to divide the dataset, which for this algorithm must be chosen initially. We analyzed the variation of the SSE value, that is the sum of the differences squared between each observation and the average of its group with the variation of K with it included in the range [0,50].

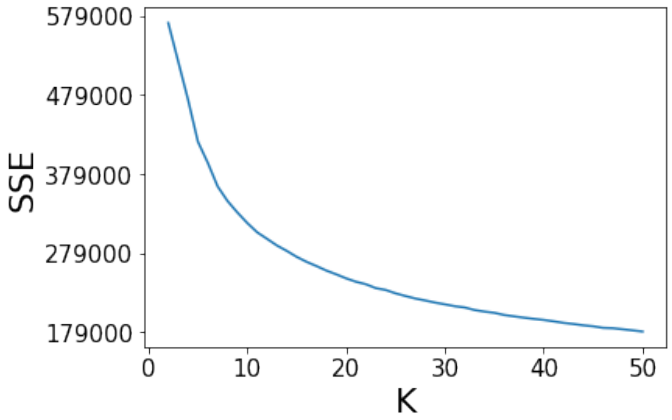

Figure 17: K variation

We can found the optimal value of K in the region where the curve starts to bend, witch in our case is approximately 10. K-Means is sensitive to the proportions of the characteristics, so we have scaled all the numerical values with z-score, i.e. the number of standard deviation from the mean of a data point, to make sure that no characteristics can dominate another.

The following table summarizes the observation made for each cluster found.

Table 3: K-means statistics

	Average VehicleAge	Average acquisition price	Average retail acquisition price	IsBadBuy percentage	number of elements of each cluster
Cluster 1	3.98	7230.52	9865.29	0.00 %	8253
Cluster 2	3.02	10625.55	13960.99	4.12 %	5020
Cluster 3	2.19	7449.25	9968.80	0.02 %	6572
Cluster 4	5.77	3729.50	5654.70	100.00 %	3787
Cluster 5	6.62	3269.04	5063.58	1.12 %	5443
Cluster 6	3.84	7299.79	10010.17	0.00 %	8540
Cluster 7	3.91	6744.92	10102.85	11.54 %	1499
Cluster 8	4.44	4641.25	6630.42	0.01 %	8046
Cluster 9	3.98	7267.89	10090.47	100.00 %	2978
Cluster 10	4.52	4423.50	6264.70	0.00 %	8246

From the table it can be seen that clusters 4 and 9 are composed only of values that have "IsBadBuy" = YES, in fact they are not composed of particularly young vehicles. However if you look at the age of the vehicle you will notice that the cluster with the highest average is 5, which has a low percentage of records with "IsBadBuy" = YES. This may demonstrate that the age of the vehicle does not affect the "IsBadBuy" attribute, but we are still not sure about this.

Below we show the progress of the various clusters in relation to the individual attributes considered.

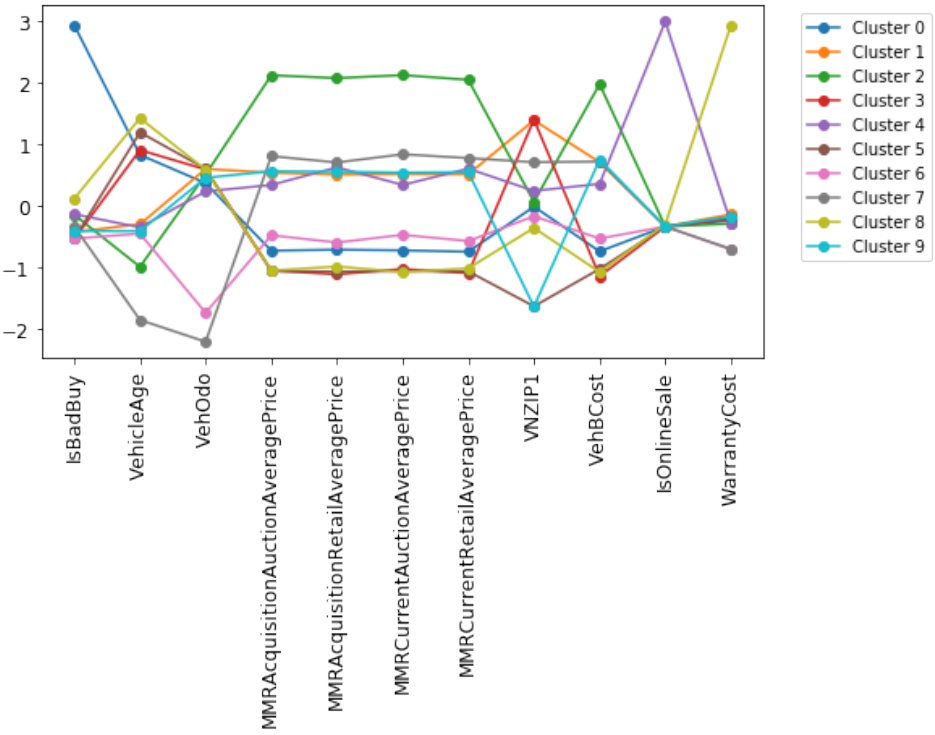

Figure 18: Distribution of the clusters in relation to the attributes.

3.2 Density-based scan clustering

The algorithm used for density-based clustering analysis is DBSCAN. The DBSCAN works by following two hyper parameters, the radius (eps), i.e. the measure within which two nodes must be found to be considered "close", and the minPts, the minimum number of points that must be close to another point in order to be able to consider it "core". We choose minPts = 10 following the observation "minPts = ln (n), where n is the number of records", and then we studied the variation of eps in relation to the ordered distances of the records and we obtained it according to the "Knee method".

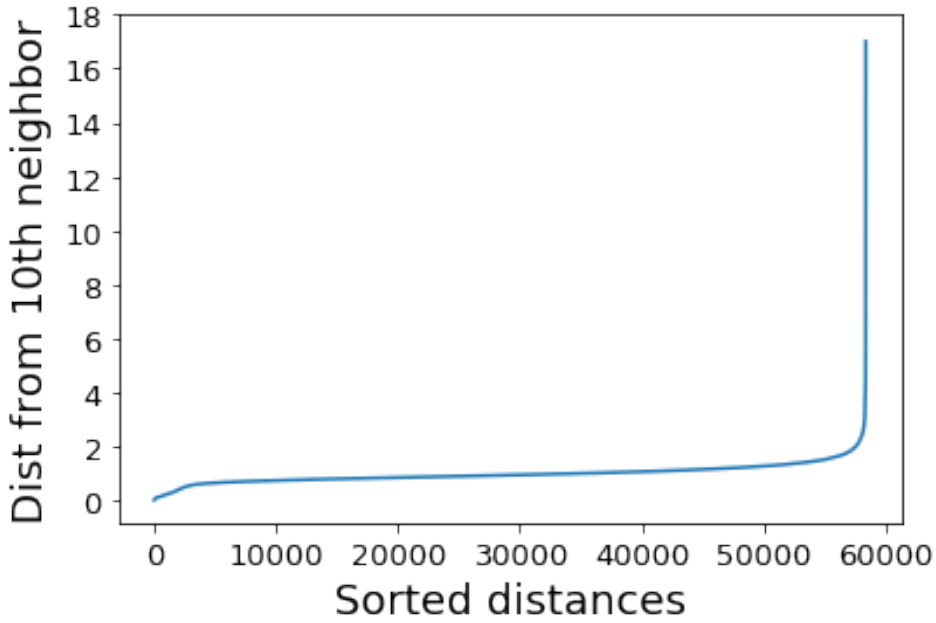

Figure 19: Eps variation.

From the figure it can be seen that the point where the straight line rises drastically is approximately before 2. After various attempts we have set eps = 1.8. The results obtained are summarized in Table 4.

Table 4: DBSCAN statistics

	Average VehicleAge	Average acquisition price	Average retail acquisition price	IsBadBuy percentuage	number of elements of each cluster
Outlier	4.42	10236.12	13453.39	67.08 %	161
Cluster 1	4.93	5449.75	7801.94	100.00 %	6898
Cluster 2	4.07	6275.70	8647.69	0.00 %	49811
Cluster 3	3.83	6833.07	10184.53	0.00 %	1312
Cluster 4	4.38	5766.92	9085.02	100.00 %	149
Cluster 5	6.92	2668.26	4347.06	100.00 %	53

The set formed by the outliers is given directly by the algorithm and includes all the noise points of the dataset, i.e. those points that are not part of any cluster. In addition to this, it can be seen that most of the elements are part of cluster 2, this is because the dataset is very dense. This composition of the clusters makes no further analysis useful.

3.3 Hierarchical clustering

Hierarchical clustering is a clustering approach that aims to build a cluster hierarchy. The matrix of the distances between the points was calculated with the Euclidean formula, then we analyzed the dendograms with the "Complete", "Ward" and "Average" methods. We then used the agglomerative strategy to study the result.

3.3.1 Average method

The hierarchical clustering method, average, calculates the distances between clusters using the Cartesian average between the intracluster points. This is the resulting dendogram:

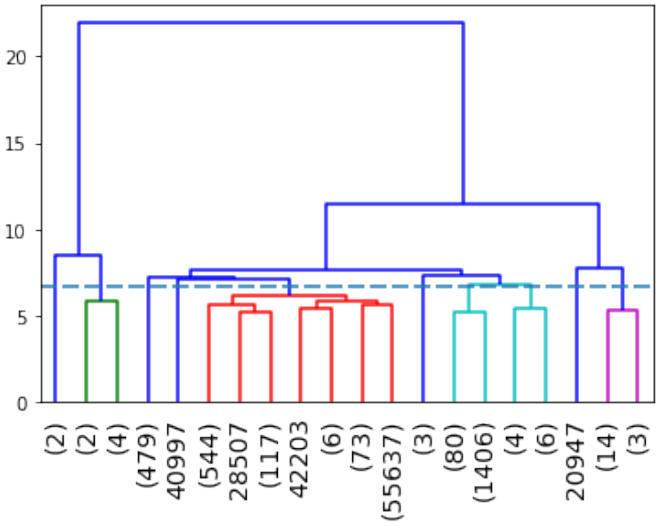

Figure 20: Hierarchical average.

We decided to cut the dendogram at height 7, to have 10 distinct clusters. Below is the table with the various statistics of the clusters found:

Table 5: Hierarchical Average statistics

	Average VehicleAge	Average acquisition price	Average retail acquisition price	IsBadBuy percentuage	number of elements of each cluster
Cluster 1	4.17	6196.36	8568.29	12.26 %	56379
Cluster 2	2.17	25959.33	30379.83	100.00 %	6
Cluster 3	5.40	5151.60	9000.20	10.00 %	10
Cluster 4	6.13	3721.64	5847.30	20.04 %	479
Cluster 5	3.91	6739.06	10091.75	11.57 %	1486
Cluster 6	2.59	18914.29	22362.12	94.12 %	17
Cluster 7	3.00	14960.67	19275.67	0.00 %	3
Cluster 8	1.50	33892.50	38482.50	100.00 %	2
Cluster 9	8.00	2411.00	3104.00	100.00 %	1
Cluster 10	2.00	19250.00	21290.00	100.00 %	1

As can be seen from the table above, a large cluster is created containing most of the records, and many smaller clusters. The analysis therefore does not lead to particular observations.

3.3.2 Complete method

The hierarchical clustering method, complete, calculates the distances between clusters using the maximum Cartesian distance between the most distant points of the respective clusters. This is the resulting dendogram:

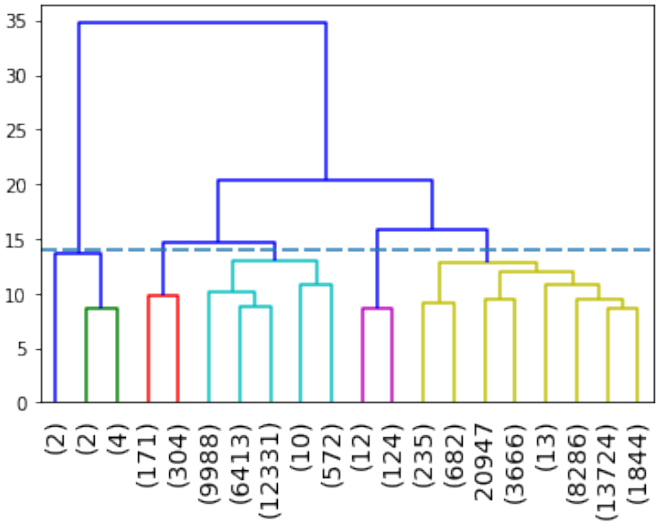

Figure 21: Hierarchical complete.

We decided to cut the dendogram at height 14, to have 5 distinct clusters. Below is the table with the various statistics of the clusters found:

Table 6: Hierarchical Complete statistics

	Average VehicleAge	Average acquisition price	Average retail acquisition price	IsBadBuy percentuage	number of elements of each cluster
Cluster 1	2.00	27942.62	32405.50	100.00 %	8
Cluster 2	4.86	4479.27	6452.50	13.48 %	29314
Cluster 3	3.44	7957.82	10786.33	11.00 %	28451
Cluster 4	2.80	15442.58	19214.05	22.06 %	136
Cluster 5	6.10	3721.82	5846.89	18.74 %	475

As shown in the table above, the results are better, cluster 2 contains the least expensive machines with just under half the values, and 13 % of these are purchases labeled as bad. The third has practically the same size but contains records representing more expensive cars with a lower percentage of bad purchases.

3.3.3 Ward method

The hierarchical clustering method, Ward, calculates the clustering pairs to be merged using an objective function, The initial cluster distances in Ward's minimum variance method are therefore defined to be the squared Euclidean distance between points:

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2$$

This is the resulting dendogram:

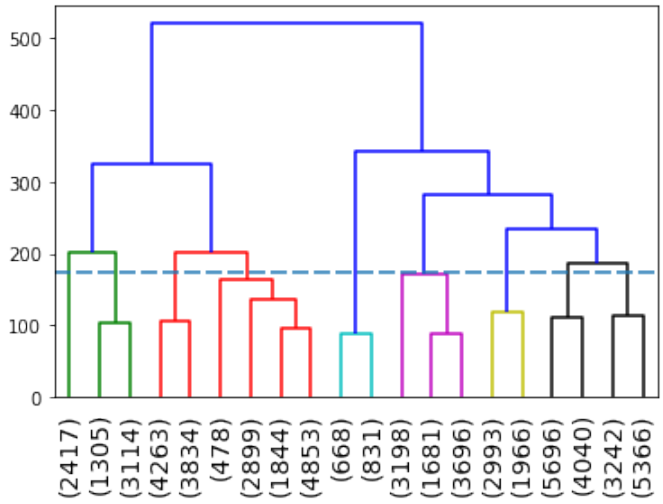

Figure 22: Caption for the image.

We decided to cut the dendogram at height 180, to have 9 distinct clusters. Below is the table with the various statistics of the clusters found:

Table 7: Hierarchical Single statistics

	Average VehicleAge	Average acquisition price	Average retail acquisition price	IsBadBuy percentuage	number of elements of each cluster
Cluster 1	2.98	6254.83	8511.28	0.00 %	8575
Cluster 2	5.71	3714.01	5566.63	0.95 %	10074
Cluster 3	2.99	10311.28	13586.79	2.04 %	4959
Cluster 4	3.61	7234.50	9804.12	0.02 %	8608
Cluster 5	5.48	4070.44	6076.70	100.00 %	4419
Cluster 6	3.91	6744.92	10102.85	11.54 %	1499
Cluster 7	4.63	4664.00	6720.07	0.00 %	8097
Cluster 8	3.98	7769.86	10726.98	100.00 %	2417
Cluster 9	3.86	7467.57	10118.17	0.00 %	9736

Compared to the previous methods, the cluster sizes are more uniform here. Particular importance should be given to cluster 5 which, containing particularly old cars, is composed only of purchases labeled as bad. On the contrary, in clusters 1 and 3, containing the younger cars, bad purchases are almost absent.

3.4 Evaluation of different clustering algorithms

Of all the clustering techniques we presented in this chapter, hierarchical clustering emerges victorious, thanks to its ability to split large clusters. Especially the Ward method is that what gives us more uniform clusters. On the contrary, DBSCAN does not produce useful results, given the large density of the dataset. Moreover, although we used two distance metrics to perform hierarchical clustering, like euclidean distance, similarity and cosine, the euclidean distance's results were always the best ones.

4 Association Rules

In this chapter we will define the association rules valid for the dataset used according to the paradigmatic principle of Market Basket Analysis and one-dimensional association rules using the "Apriori" algorithm.

The section is divided into four parts:
- Data preparation to use the Apriori algorithm.
- Extraction of frequent itemsets.
- Extraction of most interesting rules.
- Replacement of missing values with the rules obtained.

4.1 Data preparation

We have decided to delete the following attributes: "RefId", "PurchDate", "BYRNO", "VNZIP1", "IsOnlineSale" as we have found them to be superfluous for the purpose of our investigation.

Subsequently, using the Struges rule, we divided the values of the following column values into numeric intervals (bin): "VehOdo", "VehBCost", "WarrantyCost", "MMRAcquisitionAuctionAveragePrice", "MMRAcquisitionRetailAveragePrice", "MMRCurrentAuctionAveragePrice", "MMRCurrentRetailAveragePrice". This was done because the "Apriori" algorithm needs to operate with a limited number of possible values in order to generate a uniformly proportioned number of rules.

Finally we have transformed the types of values from numerical to literals for the attributes "IsBadBuy" and "VehicleAge" in order to not create possible conflicts with the "Apriori" algorithm.

4.2 Frequent itemsets extraction

Regarding the extraction of the most frequent itemsets, using the "Apriori" algorithm, we set the parameters as follows:

- MinSupport = 10%.
- zMin in Range [2,6], where zMin is the minimum number of items for each itemset.

We extracted five items with the highest support for each zMin value examined, summarized below in table 8.

Table 8: Frequent Itemsets

ZMin	itemsets	Support
2	IsBadbuy: 'No', Transmission: 'AUTO'	0.8455741298986024
2	Nationality: 'AMERICAN', Transmission: 'AUTO',	0.8139730063030968
2	Nationality:'AMERICAN', IsBadBuy:'No'	0.7344820498766785
2	Nationality:'AMERICAN', IsBadBuy:'No', Transmission: 'AUTO'	0.7152302000548095
2	Auction:'MANHEIM', Transmission: 'AUTO'	0.540165113729789
3	Nationality:'AMERICAN',IsBadBuy: 'No', Transmission: 'AUTO'	0.7152302000548095
3	Auction:'MANHEIM', IsBadBuy: 'No', Transmission: 'AUTO'	0.47855576870375444
3	Auction:MANHEIM', Nationality:'AMERICAN', Transmission: 'AUTO'	0.45365168539325845
3	WheelType:'Alloy', IsBadBuy: 'No', Transmission: 'AUTO'	0.42965538503699646
3	WheelType:'Alloy', Nationality: 'AMERICAN', Transmission: 'AUTO'	0.4281138668128254
4	Auction:'MANHEIM', Nationality: 'AMERICAN', IsBadBuy: 'No', Transmission: 'AUTO'	0.402593176212661
4	WheelType: 'Alloy', Nationality: 'AMERICAN', IsBadbuy: 'No', Transmission: 'AUTO'	0.3704782132090984
4	(WheelType:'Covers', Nationality: 'AMERICAN', IsBadbuy: 'No', Transmission: 'AUTO'	0.33803781858043297
4	TopThreeAmericanName: 'GM', IsBadbuy: 'No', Transmission: 'AUTO', Nationality: 'AMERICAN'	0.300561797752809
4	VehBCost: (Interval(6841.75, 9122.0, closed='left'), Nationality: 'AMERICAN', IsBadbuy: 'No', Transmission: 'AUTO'	0.29372773362565086
5	WheelType:'Alloy', Auction:'MANHEIM', Nationality:'AMERICAN', IsBadBuy: 'No', Transmission: 'AUTO'	0.21761098931214032
5	Make:'CHEVROLET', IsBadBuy: 'No', Transmission: 'AUTO', TopThreeAmericanName: 'GM', Nationality:'AMERICAN'	0.20971499040833105
5	WheelType:'Covers', Auction: 'MANHEIM', Nationality:'AMERICAN', IsBadBuy: 'No', Transmission: 'AUTO'	0.1809571115374075
5	MMRAcquisitionAuctionAveragePrice: Interval(5238.75, 7416.125), Interval(CurrenteAuctionAveragePrice: 4788.125, 6997.688), Nationality:'AMERICAN', IsBadBuy: 'No', Transmission: 'AUTO'	0.1772060838585914
5	TopThreeAmericanName: 'GM', Auction:'MANHEIM',IsBadBuy: 'No', Transmission: 'AUTO', Nationality:'AMERICAN'	0.17330090435735818
6	Make:'CHEVROLET', Auction:'MANHEIM', IsBadBuy: 'No', Transmission: 'AUTO', TopThreeAmericanName:'GM', Nationality:'AMERICAN'	0.12025554946560701
6	Trim:'LS', Make: 'CHEVROLET', IsBadbuy: 'No', Transmission: 'AUTO', TopThreeAmericanName: 'GM', Nationality:'AMERICAN'	0.11938202247191011
6	Make: 'CHEVROLET', WheelType: 'Covers', IsBadBuy: 'No', Transmission: 'AUTO', TopThreeAmericanName: 'GM', Nationality:'AMERICAN'	0.1160420663195396
6	MMRAcquisitionAuctionAveragePrice: Interval(7416.125, 9593.5), CurrentAuctionAverngePrice: Interval(6997.688, 9207.25), VehBCost: Interval(6841.75, 9122.0), Nationality:'AMERICAN', IsBadBuy: 'No', Transmission: 'AUTO'	0.11297615785146616
6	Make:'CHEVROLET', VehBCost: Interval(6841.75, 9122.0), IsBadBuy: 'No', Transmission: 'AUTO', TopThreeAmericanName:'GM', Nationality:'AMERICAN'	0.10264798574952042

The observations that can be made on the most frequent itemsets that have been found are as follows:

• The "IsBadBuy" attribute is mainly composed of "No", the "Transmission" attribute almost exclusively from "AUTO", therefore the most frequent item is the one that contains them both.
• As the zMin value increases, the higher value of support drops dramatically.
• A very interesting itemset is Auction: "MANHEIM", Nationality: "AMERICAN", IsBadBuy: "No", Transmission: "AUTO", calculated with zMin = 4.

4.3 Association Rule's extraction

We also used the "Apriori" algorithm to extract the most significant Association Rules for the dataset, setting the following parameters:

• MinSupport = 10%.
• MinConfidence in Range [60% , 69%], [70% , 79%], [80% , 89%], [90% , 99%].
• zmin = 2.

We therefore extracted the five most interesting rules, for each range of MinConfidence examined, but not taking into consideration those rules that found a confidence equal to 1, such as for example "CHRYSLER", ("AMERICAN"), considering them as obvious. (There can never be Chrysler machines that are not American, or "Chevrolet" that are not part of the General Motors group). We have summarized everything in table 9.

Table 9: Association rules

Conf	itemsets	Support	Confidence
90-99	Transmission:AUTO', WarrantyCost:(Interval(1969.714, 2472.286), Nationality:'AMERICAN',	0.10956768977	0.9971940763
90-99	Transmission:'AUTO',(TopThreeAmericanName:'CHRYSLER', VehBCost:(Interval(6841.75, 9122.0), Nationality:'AMERICAN',	0.11335297341	0.9966867469
90-99	Transmission:'AUTO',(TopThreeAmericanName:'CHRYSLER', VehBCost:(Interval(6841.75, 9122.0)	0.11335297341	0.9966867469
90-99	Transmission:'AUTO',(TopThreeAmericanName:'CHRYSLER', VehBCost:(Interval(6841.75, 9122.0), Nationality:'AMERICAN', IsBadbuy:'No'	0.10364140860	0.9965415019
90-99	Transmission:'AUTO',(TopThreeAmericanName:'CHRYSLER', VehBCost:(Interval(6841.75, 9122.0), IsBadbuy:'No'	0.10364140860	0.9965415019
80-89	IsBadbuy:'No',MMRCurrentretailaverageprice: (Interval(8057.938, 10444.25), Auction:'MANHEIM'	0.14505686489	0.8999043672
80-89	IsBadbuy:'No',MMRCurrentretailaverageprice: (Interval(8057.938, 10444.25), Size:'MEDIUM'	0.1071355165	0.8998705222
80-89	IsBadbuy:'No',MMRAcquisitionauctionaverageprice: (Interval(5238.75, 7416.125), Nationality:'AMERICAN',	0.24496437380	0.899723200
80-89	IsBadbuy:'No',MMRAcquisitionauctionaverageprice: (Interval(5238.75, 7416.125),MMRCurrentauctionaverageprice:(Interval(4788.125, 6997.688),Auction:'MANHEIM',Transmission:'AUTO'	0.12200260345	0.8997094859
80-89	IsBadbuy:'No',MMRAcquisitionauctionaverageprice: (Interval(5238.75, 7416.125), Nationality:'AMERICAN', Transmission:'AUTO'	0.240117155	0.8996342167

70-79	Nationality:'AMERICAN',VehBCost:(Interval(4561.5,6841.75), Size:'MEDIUM', IsBadbuy:'No'	0.14942449986	0.7997799779
70-79	Nationality: 'AMERICAN',MMRCurrentretailaverageprice: (Interval(5671.625, 8057.938),MMRCurrentauctionaverageprice: (Interval(2578.562, 4788.125))	0.11294190189	0.7988853889
70-79	Nationality:'AMERICAN',MMRCurrentretailaverageprice: (Interval(5671.625, 8057.938),MMRAcquisitionauctionaverageprice: Interval(3061.375, 5238.75),IsBadbuy:'No'),	0.10072965195	0.798723346462
70-79	Nationality:'AMERICAN',MMRAcquisitionretailaverageprice: (Interval(8509.688, 10861.25),MMRAcquisitionauctionaverageprice: (Interval(5238.75, 7416.125)),Transmission:'AUTO'	0.12253357083	0.7987049235
70-79	Nationality:'AMERICAN',VehBCost:(Interval(4561.5, 6841.75), Size:'MEDIUM'	0.1688133735	0.7986386840
60-69	VehBCost:(Interval(4561.5, 6841.75),MMRCurrentretailaverageprice: (Interval(5671.625, 8057.938), IsBadbuy:'No'	0.13245673922	0.6996221176
60-69	MMRAcquisitionretailaverageprice: Interval(6158.125, 8509.688),MMRCurrentretailaverageprice: (Interval(5671.625, 8057.938),VehBCost:(Interval(4561.5, 6841.75))	0.12460605645	0.6993847337
60-69	MMRAcquisitionretailaverageprice: Interval(6158.125, 8509.688),MMRCurrentretailaverageprice: (Interval(5671.625, 8057.938),VehBCost:(Interval(4561.5, 6841.75),Nationality:'AMERICAN',Transmission:'AUTO'	0.100010276	0.698862956
60-69	MMRAcquisitionretailaverageprice: Interval(6158.125, 8509.688),MMRCurrentretailaverageprice: (Interval(5671.625, 8057.938),VehBCost:(Interval(4561.5, 6841.75),Nationality:'AMERICAN',Transmission:'AUTO'	0.10001027	0.698862956
60-69	Nationality:'OTHER',WarrantyCost:(Interval(462.0,964.571), Size:'MEDIUM', WheelType:'Covers')	0.1009694436	0.698542481
60-69	WarrantyCost:(Interval(462.0,964.571),(Nationality:'OTHER', Size:'MEDIUM', IsBadbuy:'No', Transmission:'AUTO'	0.11780624828	0.6984868487

Therefore, the observations that can be deduced are the following:

• American cars have almost all the automatic transmission, almost always of the Chrysler group and, if paid at auction between 841.75 $ and 9122 $, they almost always prove to be a "good buy"
• Cars purchased at the Manheim auction, paid between $ 8057.938 and $ 10444.2, are not "bad buy".
• Medium-sized cars (Size = "MEDIUM"), paid between $ 8057.938 and $ 10444.2, are not "bad buy".

Listed below are the most interesting rules that provide for the "IsBadBuy" attribute, with the relative value of support and confidence.

The rules that provide for "IsBadBuy" = YES with MinSupport = 10% are only two. This is due to the fact that there are few records with attribute "IsBadBuy" = YES which makes them unusable.

Table 10: Rules for the 'IsBadBuy' attribute

IsBadBuy	itemsets	Support	Confidence
YES	IsBadbuy:'Yes', (Transmission:'AUTO')	0.1193477665113	0.1236864527122
YES	IsBadbuy:'Yes', (Nationality:'AMERICAN')	0.1013976431899	0.1213065038318
NO	IsBadbuy:'No',(MMRCurrentauctionaverageprice: Interval(6997.688, 9207.25), WheelType:'Covers', Nationality:'AMERICAN', Transmission:'AUTO')	0.1076664839682	0.939892344497
NO	IsBadbuy:'No',(MMRCurrentauctionaverageprice: Interval(6997.688, 9207.25), WheelType:'Covers', Nationality:'AMERICAN',)	0.1080775554946	0.939407473574
NO	IsBadbuy:'No',(VehBCost:Interval(6841.75, 9122.0), WheelType:'Covers', Nationality:'AMERICAN', Transmission:'AUTO')	0.127158125513	0.937965887555
NO	IsBadbuy:'No',(VehBCost:Interval(6841.75, 9122.0), WheelType:'Covers', Nationality:'AMERICAN'),	0.1277747328035	0.937539273595
NO	IsBadbuy:'No',(MMRCurrentauctionaverageprice:Interval(6997.688, 9207.25), WheelType:'Covers', Transmission:'AUTO')	0.1201870375445	0.936349079263

4.4 Replacement of missing values with the rules obtained

Exploiting the association rules extracted we wrote a method that, taken in input the original training set, replace, for each record with a null value, the value that has the highest confidence, among all the rules concerning the subsets of attributes of that record that deduces that particular attribute. This part would require a more detailed explanation, but for reasons of space we will not spent too much explanations on that. We report same examples of " filling null solutions" that our method gave, the number of records attribute used for searching in to the rules is two in the first two cases, and three in the last two.

- NATIONALITY: Record 18532 -> ("AMERICAN", ("DODGE", "No")).
- SIZE: Record 15769, 18532, 20016, 35157 -> ("MEDIUM", ("No", "AUTO")).
- TRANSMISSION: Record 23019 -> ("AUTO", ("LARGE", "AMERICAN", "No")).
- TOPTHREEAMERICANNAME: Record 18532-> ("CHRYSLER", ("DODGE", "No", "AUTO"))

5 Classification

The classification's task is the task that is supposed to learning from different decision trees/classification algorithms with different parameters and gain formulas with the object of maximizing the performances. In this phase we will explain out decision tree interpretation, and how we validate the tree with test and training test, and after we will discuss about the best prediction model.

Before starting we had to exclude some attributes, considered useless for classification purposes. To make the decision tree, we used the following attributes: "IsBadBuy", "VehicleAge", "VehOdo", "MMRAcquisitionAuctionAveragePrice", "MMRAcquisitionRetailAveragePrice", "MMRCurrentAuctionAveragePrice", "MMRCurrentRetailAveragePrice", "IsOnlineSale", "VehBCost", "Auction", "TopThreeAmericanName", "Transmission", "WheelType", "Nationality", "Color", "VNST", "Size". Subsequently we had to codify the categorical attributes among those that we kept. For the coding we used "One hot encoding", which creates a column for each unique value of each attribute, inserting 0 if the value is not present, 1 vice versa. The attributes encoded with this method are as follows: "Auction", "Transmission", "Nationality", "TopThreeAmericanName", "Color", "VNST", "Size", "WheelType".

For the choice of hyperparameters we used random search, the range of values analyzed is represented by table 11.

Table 11: Hyper-parameters

min samples split	max depth	min samples leaf	max leaf nodes	class weight	splitter
range(2,20)	range(1,20)	range(1,20)	range(2,20)	Balanced, None	Best, Random

The best combination of hyperparameters found, for both criteria (Gini, Entropy), is the following:

Criterion = Gini; Parameters: 'splitter': 'best', 'min samples split': 14, 'min samples leaf': 14, 'max leaf nodes': 9, 'max depth': 5, 'class weight': None

Criterion = Entropy; Parameters: 'splitter': 'best', 'min samples split': 5, 'min samples leaf': 13, 'max leaf nodes': 11, 'max depth': 17, 'class weight': None

Considering how unbalanced the target variable is ("IsBadBuy" = YES is numerically much lower than "IsBadBuy" = NO) we performed an oversampling procedure, using the "SMOTE" method. The name of the method stands for Synthetic Minority Oversampling Technique. This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input, the implementation of SMOTE does not change the number of majority cases. Another way to address this problem, was to unify the training set and the test set, in order to create a dataset more balanced, but we prefered to use the oversampling.

5.1 Decision Tree interpretation and validation

For the creation and validation of the decision trees we used two techniques. The first one uses the "Train to split" method, which divides the dataset into training (70%) and tests (30%) in a random way. Here are the tables with the results.

Table 12: TTS Training

Criterion	Accuracy	Precision 'YES'	Precision 'NO'	Recall 'YES'	Recall 'NO'
Gini	0.77	0.88	0.72	0.65	0.91
Entropy	0.78	0.95	0.71	0.61	0.97

		Predicted	
		T	F
Actual	T	32616	3207
	F	12657	23166

Figure 23: Confusion Matrix Gini

		Predicted	
		T	F
Actual	T	34663	1160
	F	14020	21803

Figure 24: Confusion Matrix Entropy

Table 13: TTS Test

Criterion	Accuracy	Precision 'YES'	Precision 'NO'	Recall 'YES'	Recall 'NO'
Gini	0.78	0.88	0.73	0.66	0.91
Entropy	0.79	0.95	0.72	0.62	0.97

		Predicted	
		T	F
Actual	T	13981	1372
	F	5294	10059

Figure 25: Confusion Matrix Gini

		Predicted	
		T	F
Actual	T	14858	495
	F	5886	9467

Figure 26: Confusion Matrix Entropy

It is interesting to note that the values, on training and on the test, are almost identical. As for the Recall "YES" value, this result was achieved thanks to the oversampling operation, otherwise it would have been much lower, given the low percentage of "IsBadBuy" = YES.

A second method used is cross-validation, which divides the dataset into X parts (five in our case), and creates multiple decision trees considering each time a different test set between the parts created. The decision tree with the highest cross-validation score is chosen. The results are shown in the table below.

Table 14: Decision tree CV based on Training

Criterion	Accuracy	Precision 'YES'	Precision 'NO'	Recall 'YES'	Recall 'NO'
Gini	0.75	0.79	0.72	0.77	0.74
Entropy	0.77	0.83	0.72	0.75	0.80

Table 15: Decision tree CV on Test

Criterion	Accuracy	Precision 'YES'	Precision 'NO'	Recall 'YES'	Recall 'NO'
Gini	0.75	0.45	0.86	0.65	0.73
Entropy	0.74	0.50	0.86	0.61	0.79

		Predicted	
		T	F
Actual	T	26532	9350
	F	10551	35449

Figure 27: Confusion Matrix Gini (Training) CV

		Predicted	
		T	F
Actual	T	28768	7114
	F	11463	34537

Figure 28: Confusion Matrix Entropy (Training) CV

		Predicted	
		T	F
Actual	T	11231	4063
	F	1793	3383

Figure 29: Confusion Matrix Gini (Test) CV

		Predicted	
		T	F
Actual	T	12157	3137
	F	2021	3155

Figure 30: Confusion Matrix Entropy (Test) CV

Of the two techniques used, the one using the cross-validation method, using Gini's criterion, gives us the best results. We then proceed with the application of the decision tree to the test set supplied to us.

5.2 Application to Test set

The test set consists of 14596 records, some of which have zero and null values. In order to run the model chosen on the test set we must manage the null values, to avoid inconsistencies the same choices were made, described in 2.4, used for the training set.

After having eliminated the attributes deemed useless, as for training, and having coded the categorical values in the same way, we proceed with the application of the model. The results that come out are the following:

Table 16: Application to test set

Accuracy	Precision 'YES'	Precision 'NO'	Recall 'YES'	Recall 'NO'
0.82	0.24	0.89	0.20	0.91

		Predicted	
		T	F
Actual	T	11702	1127
	F	1417	351

Figure 31: Confusion Matrix test set

As the dataset is structured, we are able to better predict when a car is a good buy, and not the other way around. The considerations that come out are the following:
- The higher the price of a car, the less chance it is of a bad purchase.
- Dodge, Chevrolet and Ford are almost never bad purchases.
- The older a car, the greater the chance of it being a bad purchase.

Contents