Università di Pisa

Corsi di laurea in Computer Science,
Data Science and Business Informatics
Anno Accademico 2017/2018
Corso di Data Mining

Studenti:

- Antonio Sisbarra (518552)
- Francesco Spinnato (572125)

CARVANA

Relazione di analisi su dataset: CarVana - Don’t get kicked!

1. Introduzione

Uno dei più grandi rischi per una concessionaria nell'acquistare un'auto usata all'asta, è che il veicolo appena comprato possa avere problemi seri che impediscano la successiva vendita ai clienti. In gergo automobilistico questi acquisti sfortunati vengono chiamati "kicks".
Questi pessimi affari si possono verificare per esempio quando ci sono odometri manomessi, guasti meccanici non facilmente risolvibili, problemi con il rilascio del libretto del veicolo da parte venditore oppure altre complicazioni impreviste. Errori di questo tipo possono essere molto costosi per la concessionaria contando i costi di trasporto, gli eventuali lavori di riparazione a distanza e le inevitabili perdite nella rivendita.
Da questo problema nasce la sfida proposta dalla start-up statunitense Carvana, pubblicata su Kaggle[1], il cui obiettivo è quello di elaborare un modello che possa individuare e prevenire il rischio di comprare auto problematiche.
Il nostro elaborato si divide in quattro sezioni:

1. **Data Understanding**: analisi del dataset rispetto a: data quality, data semantic. Preprocessing e eliminazione di attributi rindondanti.

2. **Clustering**: K-Means, DBScan e hierarchical.

3. **Association rule mining:** ricerca di pattern frequenti e regole di associazione.

4. **Classification**: costruzione di alberi decisionali con diversi modelli e valutazione della performance.

2. Data Understanding

2.1 Analisi qualitativa e quantitativa dei dati

Il training set necessario all'analisi è fornito da Carvana e contiene dati relativi alle vetture prodotte tra il 2001 e il 2010. Esso contiene 72983 record e 33 attributi.
Tra gli attributi del training set se ne contano 18 di tipo numerico e 14 di tipo nominale. Nella tabella sono riportati tutti gli attributi ordinati per tipologia.

[1] https://www.kaggle.com/c/DontGetKicked

Tipologia		Attributo
Numerico	Discreto	VehYear, VehicleAge, WheelTypeID, BYRNO, VNZIP
	Binario	IsBadBuy,
	Continuo	VehBCost, VehOdo, MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitonRetailCleanPrice, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice, WarrantyCost
Categorico	Nominale	Auction, Make, Model, Trim, SubModel, Color, Nationality, TopThreeAmericanName, AcquisitionType, AUCGUART, VNST
	Binario	IsOnlineSale
	Ordinale	Size, WheelType, PurchDate

L'attributo IsBadBuy indica se il veicolo è un cattivo acquisto con il valore "1", altrimenti il valore è "0".

Le informazioni sull'età e la data d'acquisto del veicolo sono contenute negli attributi PurchDate, VehicleAge e VehYear, mentre Make, Model, Trim e SubModel indicano il la casa di produzione dell'automobile, il modello e diverse altre caratteristiche tecniche.

Ulteriori specifiche riguardo colore, trasmissione, tipo di ruota, dimensione, chilometraggio, sono contenute negli attributi WheelTypeID, WheelType, Color, Transmission, VehOdo, Size.

TopThreeAmericanName, Nationality, VNZIP1 (stato di acquisto), VNST (codice postale), BYRNO (codice assegnato all'acquirente), VehCost", "IsOnlineSale", "WarrantyCost" riguardano nazionalità e provenienza, costo e modalità d'acquisto.

La sigla MMR sta per "Manheim Market Report" ed è uno degli indicatori principali per il prezzo all'ingrosso dei delle automobili. In particolare "Acquisition" indica il prezzo a cui il veicolo è stato venduto all'asta, "Clean" e "Average" si riferiscono alla condizione del veicolo (rispettivamente buona o nella media), "Auction" contiene il prezzo atteso del veicolo nell'asta e "Retail" è una stima del prezzo che il cliente sarebbe disposto a pagare in concessionaria.

Infine “PRIMEUNIT” indica livello di richiesta sul mercato rispetto a un acquisto standard, mentre “AUCGUART” descrive il rischio del veicolo, ossia il livello di garanzie che il venditore concede.

Dall’**analisi qualitativa** dei dati sono emersi alcuni errori o inesattezze nei valori degli attributi:

- il valore “Toyota Scion” dell’attributo Make. Non esiste sul mercato una casa produttrice con entrambi i nomi. Il valore è stato corretto in “Scion” dal momento che questi è la casa produttrice appartenente alla Toyota.
- I valori di Transmission, Trim e Model sono stati tutti convertiti in maiuscolo, visto che vi era la presenza minoritaria di alcuni valori in minuscolo.
- Per i valori di Model e SubModel sono stati identificati numerosi errori di incompletezza e di battitura. Questi attributi contengono moltissime informazioni potenzialmente utili riguardo diverse caratteristiche tecniche del veicolo, che tuttavia sono probabilmente state inserite in fase di compilazione senza un pattern specifico e senza ricerca di completezza. Pertanto è stata eseguita un'estrazione di tali specifiche in fase di preprocessing.

Dell’**analisi quantitativa** dei dati vengono riportate le osservazioni più significative:

- Gli attributi AUCGUART e PRIMEUNIT presentano una forte scarsità di valori (circa 90% NULL). Per questa ragione i due attributi, nonostante siano dei criteri interessanti per la valutazione di un’automobile, non sono stati presi in considerazione per le analisi future.
- 2360 automobili non hanno indicato il Trim (assetto interno), e 3174 non hanno indicazioni riguardo il tipo di cerchione.
- 846 automobili sono sfornite del dato riguardante il prezzo del veicolo in condizioni medie al momento dell’acquisto (MMRAcquisitionAuctionAveragePrice). Nel dataset sono ben 8 gli attributi MMR: siamo in presenza di un caso di ridondanza di valori la cui gestione verrà illustrata nella sezione seguente.

Il training set si compone di una grande quantità di dati, molti dei quali superflui o ridondanti. Nella sezione relativa al *preprocessing* si affronterà una riflessione sull’eliminazione di questi.

2.2 Distribuzione delle variabili

Dopo aver osservato le distribuzioni dei valori di tutti gli attributi, è stata fatta una disamina delle più interessanti.

IsBadBuy: i buoni acquisti sono più frequenti

Nel dataset si nota una distribuzione evidentemente sbilanciata per valori di IsBadBuy: l'87,7% delle auto risulta buon acquisto, mentre solo il 12,3% non lo è.
Le auto scadenti rappresentano eccezioni alle aste. Durante la costruzione e la valutazione dei modelli di classificazione, si dovranno bilanciare i valori di questo attributo così da poterne valutare l'efficienza con le diverse distribuzioni dei valori della classe.

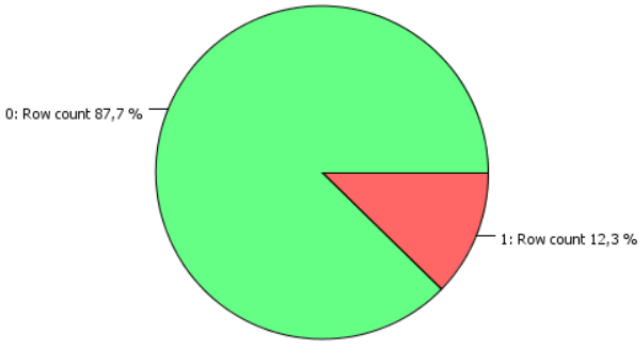


Figura 1 - Distribuzione IsBadBuy

Make: la casa automobilistica più frequente è Chevrolet

Le auto americane rappresentano l'83,77% dell'intero dataset. Il marchio più frequente risulta Chevrolet (23,81%), seguito da Dodge, Ford e Chrysler.

VehicleAge: l'età più frequente è 4 anni, le auto a rischio sono le più anziane

La distribuzione dei valori dell'età dei presenta un picco di occorrenze per i 4 anni: esse rappresentano il 23,21% del training set. Seguono poi quelli di 3 e 5 anni. Inoltre, si è notato che la percentuale di cattivi acquisti aumenta con l'età del veicolo: un 10% per i veicoli fino a 2 anni d'età, mentre si arriva al 20% per i veicoli di 6 anni e 31% per i veicoli di 9 anni.

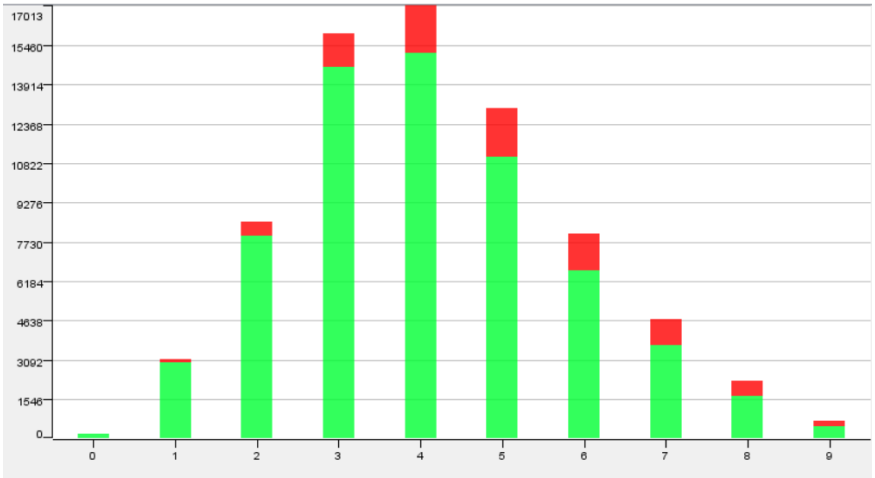


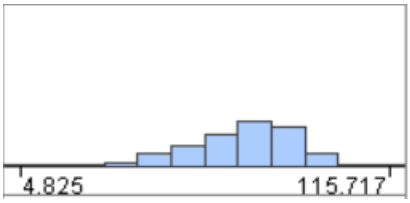
Figura 2 - Distribuzione BadBuy in base all'età del veicolo

Transmission: quasi tutti i veicoli sono a cambio automatico

Nel dataset si nota un notevole sbilanciamento per quanto riguarda il tipo di cambio: automatico nel 96,49% dei casi e manuale nel 3,5%.

Altre statistiche

- La media del chilometraggio delle automobili è di 71500 km, con una distribuzione Poissoniana, asimmetrica, negativa a sinistra,



quindi le macchine a chilometraggio alto sono le più frequenti.

- La media del prezzo dei veicoli è di 6730$ circa, mentre per quello della garanzia di 1276$.
- Le auto con attributo Nationality settato “OTHER” sono tutte europee, di marca Volkswagen, Volvo e Mini, 195 in totale**.**

2.3 Preprocessing

2.3.1 Analisi di Model e SubModel ed estrazione di subfeatures

Come già accennato gli attributi Model e SubModel contengono numerose informazioni tecniche. Prendendo ad esempio il seguente veicolo:

- Make: TOYOTA;
- Model: CAMRY 4C 2.4L I4 SFI;
- SubModel: 4D SEDAN CE;

possiamo ricavare il numero di cilindri (4C), la cilindrata (2.4L), il tipo di motore (I4), il sistema di iniezione (SFI) e il numero di porte (4D). Per estrarre queste informazioni da tutti i valori di Model e Submodel sono state utilizzate le espressioni regolari (regex). In totale sono stati quindi creati sei nuovi attributi: CilinderNumber, EngineType, EnigineLiters, InjectionType, DriveWheels e DoorsNumber. La maggioranza di questi nuovi attributi presenta numerosi missing values, pertanto non è stato possibile utilizzarli estensivamente. Tuttavia questa estrazione di valori è stata utile per aggiungere al dataset un nuovo attributo "ModelCleaned", corretto degli errori di battitura, che contiene i modelli delle automobili senza le specifiche tecniche sopracitate. In questo modo è stato possibile farsi un'idea dei modelli di automobili con una percentuale particolarmente bassa o alta di "kick", come si può vedere nella seguente tabella in cui ne sono stati elencati alcuni.

Model	Count	%IsBadBuy	Model	Count	%IsBadBuy
AVIATOR	17	47,1	*ACCORD*	315	8,3
PROTEGE	51	39,2	*G6*	1192	8,2
WINDSTAR	254	35,0	*CAMRY*	560	8,0
MAXIMA	167	33,5	*COLORADO PICKUP*	137	8,0
COOPER	24	33,3	*CHARGER*	683	7,9
MONTERO SPORT	53	32,1	*300*	469	7,9
G35	19	31,6	*GALANT*	461	7,4
MOUNTAINEER	87	31,0	*HHR*	745	7,1
INTREPID	35	28,6	*UPLANDER*	975	7,1
GRAND MARQUIS	46	28,3	*IMPALA*	4784	7,0
TRACKER	50	28,0	*TITAN PICKUP*	165	6,7
XL-7	38	26,3	*MAGNUM*	507	6,5
EXPLORER	812	26,2	*MILAN*	120	5,8
GRAND VITARA	39	25,6	*NITRO*	113	5,3
TIBURON	47	25,5	*LACROSSE*	143	4,9
B PICKUP	24	25,0	*AVENGER*	1142	4,3
G5	36	25,0	*ASCENDER*	119	4,2
MONTEREY	40	25,0	*AURA*	154	3,2

PATHFINDER	52	25,0	*YARIS*	38	0,0

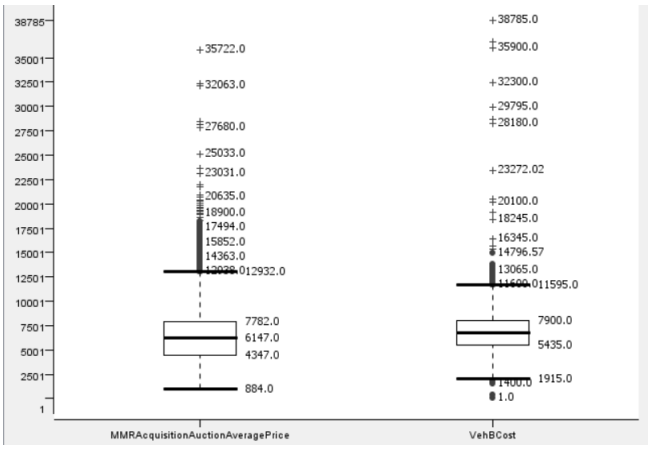
Questo nuovo attributo verrà inoltre utilizzato in seguito per la gestione dei missing values.

2.3.2 Gestione Missing Values

- **Size:** sono state trovate 5 righe con Size a NULL; si è deciso di eliminarle perché non inficiano sulla qualità delle analisi.
- **Transmission:** 9 righe a NULL, tutte di nazionalità americana; sostituite con cambio automatico vista la provenienza.
- **WheelTypeId:** Sono stati sostituiti i 3169 missing value con la moda dei rispettivi modelli, per capire quali fossero i cerchioni più popolari per quel determinato modello di veicolo.
- **Nationality:** 5 veicoli NULL, gestiti a mano guardando la casa produttrice
- **Color:** 102 missing values sostituiti con il valore più frequente. OTHER mantenuti.
- **MMRAcquisitionAveragePrice:** 846 missing values sono stati sostituiti con la media dei prezzi dei rispettivi modelli.

2.3.3 Outliers

Gli outliers nel dataset sono stati rilevati tramite l'utilizzo di un BoxPlot, che evidenzia i valori che non rientrano nel range interquartile. In particolare, sono stati rilevati valori identificati come outlier per gli attributi WarrantyCost, VehicleAge, MMRAcquisitionAuctionAveragePrice e VehBCost, ma solo per gli ultimi due si è deciso di eliminare alcuni valori sospetti:



- **VehBCost:** 1$ e 225$ sono evidenti errori di inserimento dati.
- **MMRAcquisitionAuctionAveragePrice:** Ci sono 53 valori che vanno da 12000$ a 35000$ circa; disturbavano le analisi e quindi rimossi nella fase di clustering.

2.3.4 Trasformazione dei valori

Al fine di compattare il dataset e rendere le analisi più semplici, sono state svolte le seguenti operazioni di pre-processing:

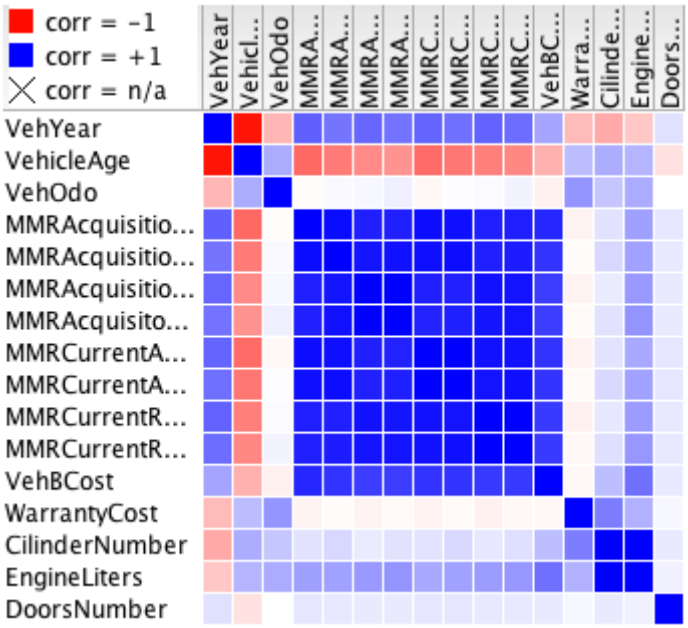
- **Nationality:** unione di TOP LINE ASIAN e OTHER ASIAN in ASIAN. Il primo valore contiene un'informazione aggiuntiva sulla qualità, superflua rispetto alla nazionalità, perciò è stato inglobato nella seconda. Inoltre, è stata operata

una trasformazione da stringa a intero: AMERICAN -> 0, ASIAN -> 1, OTHER -> 2.

- **Transmission:** da stringa a valore numerico binario: AUTO -> 0, MANUAL -> 1
- **Size:** è stato eseguito il binning per compattare i valori in tre categorie: (Compact, Sport) -> SMALL -> 0; (Crossover, Medium e Small Suv e Specialty) -> MEDIUM -> 1; (Large Suv, Truck e Van) -> LARGE -> 2.

2.4 Correlazione tra attributi ed eliminazione delle variabili ridondanti

Per trovare il livello di relazione lineare tra gli attributi numerici del dataset si è utilizzato il coefficiente di correlazione di Pearson. I valori risultanti sono stati disposti in una matrice di correlazione in cui il colore blu indica una correlazione positiva, mentre il colore rosso indica una correlazione negativa. Seguono delle riflessioni sulla correlazione tra attributi e sull'eliminazione di attributi ridondanti.

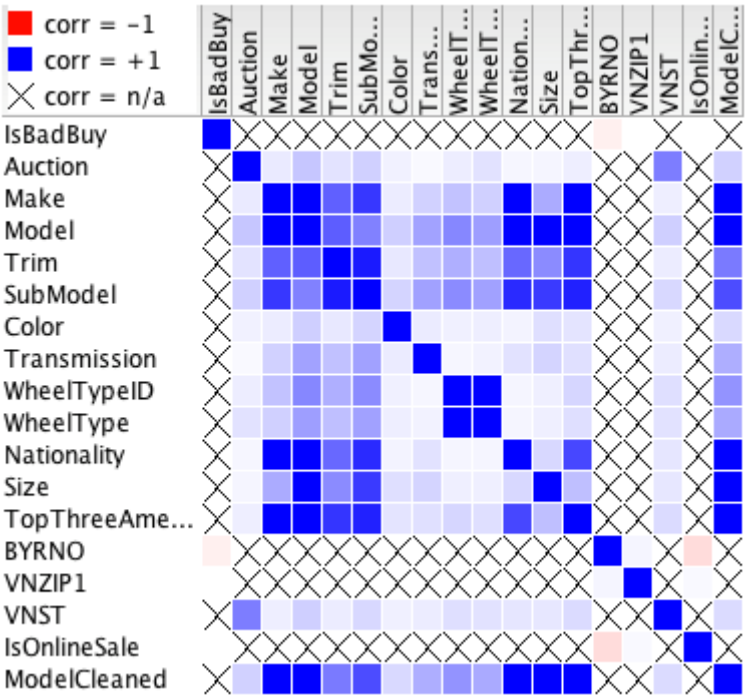


a. Tra il prezzo pagato all'asta (VehBCost) e le valutazioni di mercato MMR esiste una correlazione positiva, tuttavia questo valore è stato mantenuto poiché ritenuto significativo per la valutazione della vettura.
b. Gli MMR hanno tra loro una correlazione molto elevata (tra 0.99 e 0.89). Questi attributi si dividono in due categorie: MMRCurrent e MMRAcquisition. I primi valori rappresentano la base di prezzo dei veicoli nel giorno corrente pertanto, essendo strettamente legati ad un fattore temporale, non sono stati ritenuti utili ai fini dell'analisi. Gli MMRAcquisition segnano invece la base di prezzo dei veicoli al momento dell'acquisto. Visto che l'analisi riguarda le aste, e vista in ogni caso l'elevata correlazione tra questi attributi, si è deciso di

scartare gli MMR riguardanti i prezzi retail, e di mantenere esclusivamente l'attributo MMRAcquisitionAuctionAveragePrice.

c. VehYear è stato eliminato in quanto ridondante, visto che si dispone di già di VehicleAge.
d. Gli attributi numerici estratti attraverso le espressioni regolari, ossia CilinderNumber, EngineLiters, DoorsNumber, pur essendo di per sé interessanti, sono stati scartati in quanto composti da un eccessivo numero di missing values.

Per gli attributi nominali Knime dà la possibilità di eseguire il test chi quadrato di Pearson, il quale permette di confrontare la distribuzione del numero di valori per due o più gruppi utilizzando la stessa variabile categorica. Il valore ottenuto viene quindi normalizzato nel range [-1,1], dove 0 indica che non c'è correlazione, mentre –1 o 1 indica una forte correlazione (diretta o inversa). In questo test bisogna prestare particolare attenzione ai missing values, i quali vengono considerati come possibili valori.



Si è quindi scelto di eliminare i seguenti attributi:

- Trim: superfluo e troppo dettagliato.
- SubModel: troppo dettagliato, sostituito da ModelCleaned.
- Model: sostituito da ModelCleaned.
- WheelType: ridondante, nel dataset è già presente WheelTypeId.
- TopThreeAmericanName: ridondante dato che la nazionalità è deducibile dall'attributo Make.
- BRYNO: codici che identificano il compratore, ritenuti superflui per l'analisi.
- Auction: superfluo per le analisi, che presenta inoltre il 40% dei valori settati ad OTHER.
- VNZIP1: I codici postali sono stati ritenuti irrilevanti ai fini del progetto

- VNST: La locazione geografica dell'asta è ritenuta irrilevante ai fini dell'analisi
- IsOnlineSale, ritenuta inutile in quanto troppo sbilanciata (solo il 2% di vendite online)

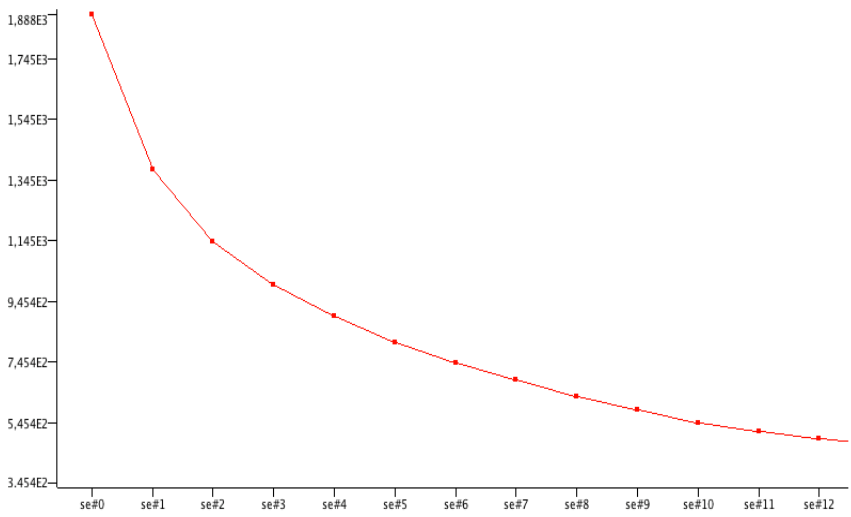
Il Training Set, alla fine delle operazioni di data understanding, presenta le seguenti caratteristiche: 72976 righe e 14 attributi (IsBadBuy, VehicleAge, Make, ModelCleaned, Color, Transmission, VehOdo, Nationality, Size, VehBCost, WarrantyCost, WheelTypeId, MMRAcquisitionAuctionAveragePrice).

3. Clustering

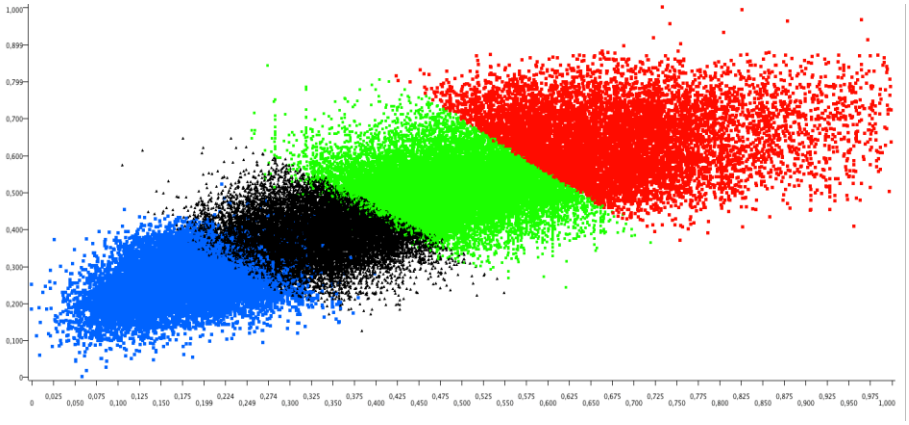
Per le operazioni di clustering sono stati scelti gli attributi VehBCost, MMRAcquisitionAveragePrice e WarrantyCost poiché rappresentano un aspetto economico del veicolo misurabile in dollari. Lo scopo di questa sezione è di mostrare la non rindondanza di tali attributi. Il dataset è stato normalizzato (min-max normalization), per poi applicare i seguenti algoritmi: K-Means, DBScan e hierarchical.

3.1 K-Means

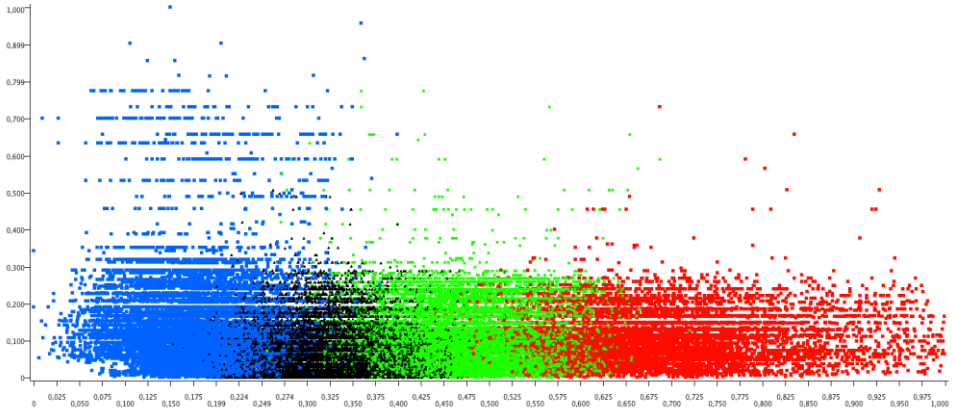
K-Means è un algoritmo di clustering partizionale, con l'obiettivo minimizzare la somma delle distanze di ciascun oggetto dal centroide del cluster cui è assegnato.
Come funzione di distanza, è stata scelta quella euclidea. Per identificare il numero ottimale di cluster (k) si è osservata la distribuzione di SSE generata iterativamente per ogni valore di k da 2 a 25. Dal grafico degli SSE ottenuti per i diversi k si è scelto k = 4 in quanto con più di quattro cluster la curva si stabilizza producendo rendimenti decrescenti. K-Means si è rivelato particolarmente suscettibile agli outliers con valori particolarmente elevati dell'attributo MMR, e solo eliminandoli è stato possibile ottenere una buona divisione in cluster. Dallo scatter plot (in figura) risulta evidente che i valori di MMRAcquisitionAuctionAveragePrice e VehBCost si raggruppano in quattro cluster vicini ma ben distinti. Questi attributi sono tra loro direttamente proporzionali e in



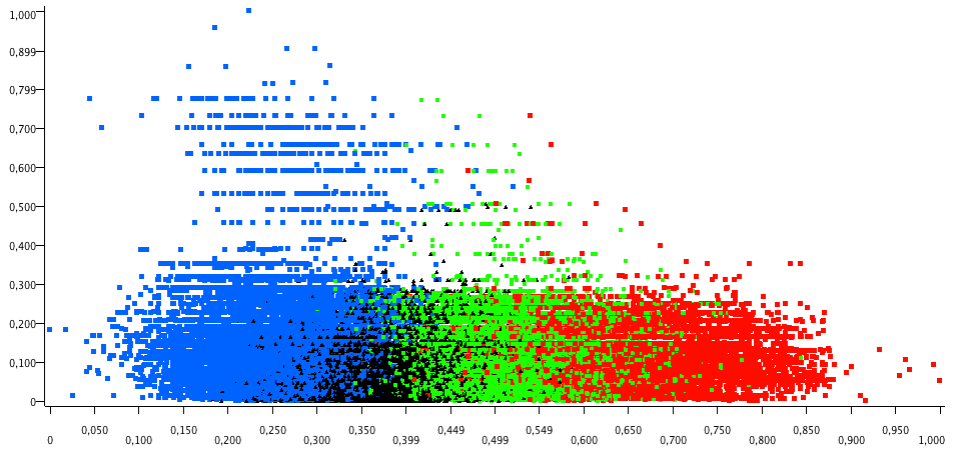
particolare la loro correlazione è pari a 0,835.



I valori MMRAcquisitionAuctionAveragePrice e WarrantyCost (in figura) si distribuiscono in un unico cluster ben diviso in quattro zone. Da notare l'esistenza di outliers in ogni cluster aventi prezzi bassi e costi di garanzia molto alti. Dal grafico è inoltre evidente la scarsa correlazione tra i due attributi (0,05).

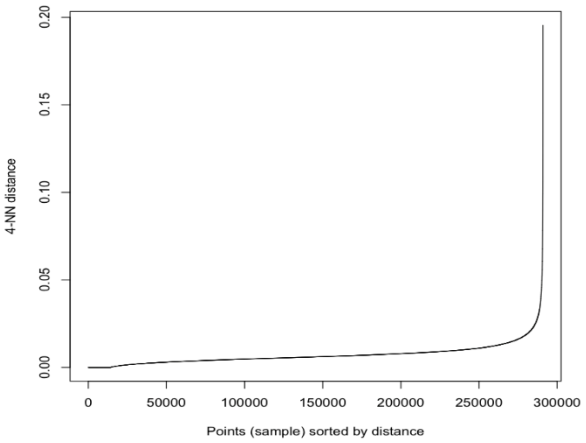


Molto simile alla precedente si presenta la distribuzione per gli attributi VehBCost e WarrantyCost la cui correlazione è di 0,033. Anche in questo caso si nota un unico grande cluster suddiviso in quattro zone e la presenza di numerosi outliers.

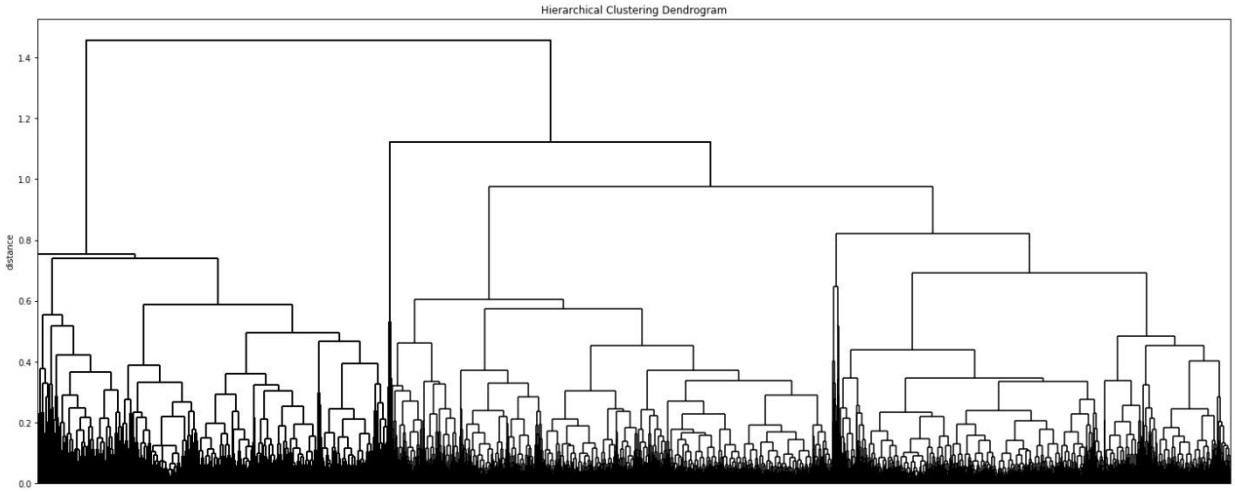


3.2 DBScan

DBScan è un algoritmo basato sulla densità che necessita due parametri: epsilon (eps), che indica la distanza massima tra punti che sono considerati vicini, e MinPts, che rappresenta il numero minimo di punti vicini richiesti per formare un cluster. Per la scelta di epsilon è stata calcolata la distanza tra i punti con l'algoritmo k-nearest neighbor, che computa la media delle distanze di ogni punto con i kappesimo nearest neighbor. Il valore di k in questo caso corrisponde a MinPts. Successivamente le distanze k sono state esposte graficamente in ordine ascendente. Dal grafico, iterando per valori di k compresi tra 2 e 25, il valore ottimale di epsilon varia in misura estremamente ridotta, rimanendo stabile a circa 0,03. L'algoritmo si è rivelato efficace solo per la ricerca degli outliers che identifica come noise, ma in generale inefficace nell'analisi poiché, anche variando il valore di MinPts, esso restituisce sempre un unico grande cluster.

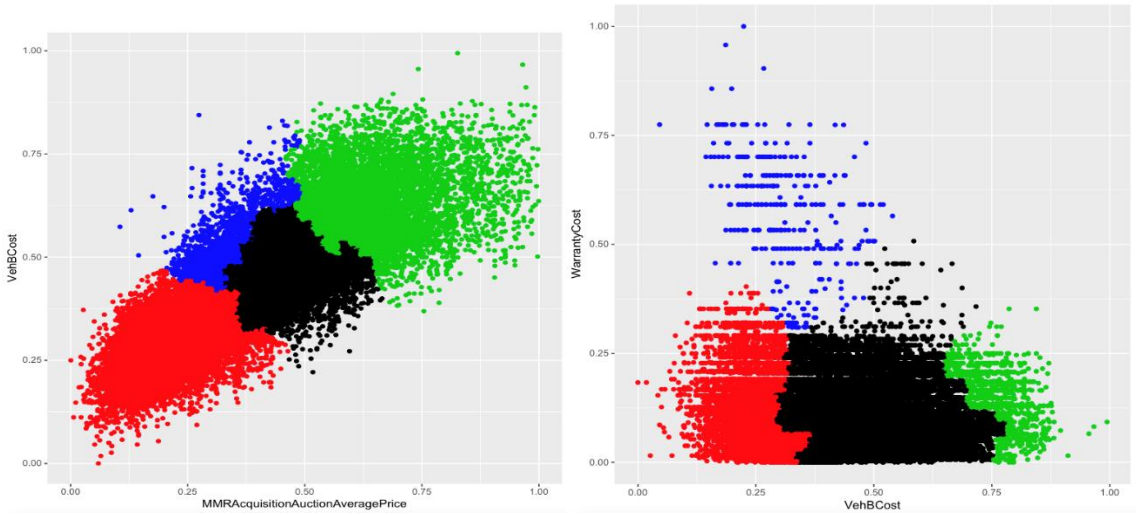


3.3 Hierarchical



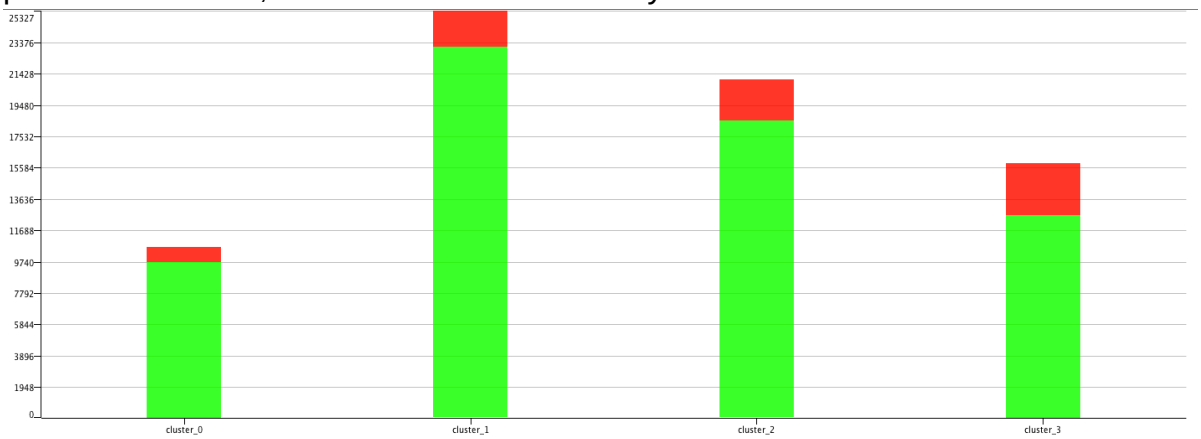
L'algoritmo gerarchico, per motivi legati ai tempi di calcolo, è stato applicato a un campione formato dal 50% del dataset. Il metodo single linkage identifica un unico grande cluster per tutti e tre gli attributi di costo, riuscendo a riconoscere solo qualche outlier. Questo risultato non stupisce in quanto questo algoritmo evidenzia in maniera netta tutte le similitudini e somiglianze tra gli elementi, privilegiando l'omogeneità tra gli elementi del gruppo a scapito della differenziazione tra gruppi. Il

metodo complete linkage offre invece un risultato che rispecchia meglio la natura dei dati. Infatti questo algoritmo di aggregazione evidenzia in maniera netta le dissomiglianze tra elementi, privilegiando la differenza tra i gruppi piuttosto che l'omogeneità degli elementi di ogni gruppo. Dall'analisi del dendrogramma ottenuto, si è scelto di tagliarlo in modo da formare quattro cluster, i quali dividono il dataset in zone ben definite. Da notare, nel grafico tra i valori VehBCost e WarrantyCost, un cluster formato da soli outliers aventi prezzi bassi e costi di garanzia molto alti.

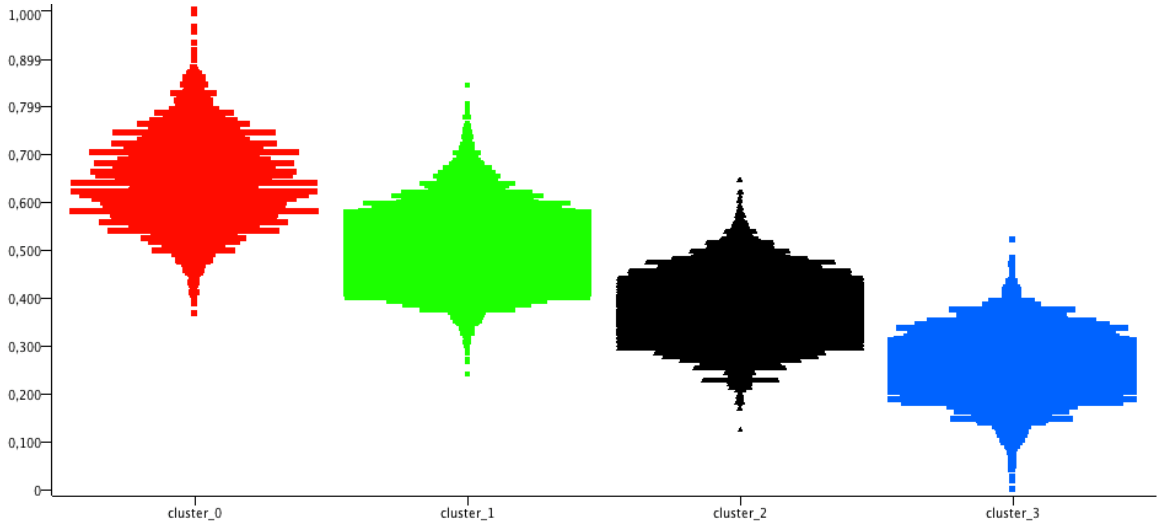


3.4 Conclusioni

Analizzando la distribuzione di tutti e tre gli attributi a confronto con IsBadBuy non si nota nulla di fuori dalla norma. Prendendo tuttavia in considerazione solamente MMR e VehBCost si può notare come nel cluster avente prezzi e MMR bassi (ossia quello che nello scatter plot si trova in basso a sinistra) ci sia una percentuale di BadBuy più elevata. Dal clustering pertanto parrebbe esistere una relazione tra prezzo del veicolo, MMR e l'attributo IsBadBuy.



Distribuzione BadBuy in base ai cluster di K-Means



Cluster di VehBCost con K-Means

4. Association Rule Mining

Il training set sottoposto alla ricerca delle regole presenta le seguenti caratteristiche: 72976 righe e 14 attributi: IsBadBuy, VehicleAge, Make, ModelCleaned, Color, Transmission, VehOdo, Nationality, Size, VehBCost, WarrantyCost, WheelTypeId, MMRAcquisitionAuctionAveragePrice.
Gli attributi numerici VehOdo, VehBCost, WarrantyCost e MMRAcquisitionAuctionAveragePrice sono stati organizzati in sette intervalli di uguale frequenza, contenenti circa 10400 valori ciascuno.
Anche per l'attributo VehicleAge è stato effettuato il binning per frequenza, suddividendo i valori in quattro intervalli.

	VehOdo	VehBCost	WarrantyCost	MMRAcquisition AuctionAveragePrice	VehicleAge
1	[5368,53943]	[1400,4775]	[462,702]	[884,3534]	[0,2]
2	(53943,63791]	(4775,5610]	(702,905]	(3534,4582]	(2,3]
3	(63791,70702]	(5610,6310]	(905,1,065]	(4582,5578]	(3,4]
4	(70702,75951]	(6310,7095]	(1065,1270]	(5578,6611]	(4,9]
5	(75951,81027]	(7095,7705]	(1270,1503]	(6611,7529]	
6	(81027,86832]	(7705,8500]	(1503,1918]	(7529,8628]	
7	(86832,115717]	(8500,38785]	(1918,7498]	(8628,35722]	

Essendo il dataset composto da molteplici valori, alcuni dei quali piuttosto infrequenti, si è scelto un supporto minimo relativamente basso (min_sup = 0.05), in modo da poter trovare anche regole più "rare". La confidenza minima è stata fissata invece a 0.7. In ogni caso le regole ottenute sono state interpretate con molta attenzione in quanto il dataset è estremamente sbilanciato. Sono infatti molto più frequenti automobili di nazionalità americana, con cambio automatico e che sono dei buoni acquisti. Nella tabella di seguito sono indicate le regole più interessanti.

	Premessa		Conseguenza	Sup	Conf	Lift
1	[Size = 1, Nationality = 0, ModelCleaned = PT CRUISER]	==>	Make = CHRYSLER	0.051	1.0	8.25
2	[VehBCost = [1400,4775], VehicleAge = (4,9]]	==>	MMRAcquisitionAuction AveragePrice = [884,3534]	0.085	0.81	5.7
3	[Nationality = 1, WheelTypeID = 2]	==>	WarrantyCost = [462,702]	0.063	0.73	5.29
4	[Size = 1, WarrantyCost = [462,702]]	==>	Nationality = 1	0.068	0.73	4.56
5	[WheelTypeID = 2, WarrantyCost = (1918,7498]]	==>	Size = 2	0.053	0.86	3.13
6	[MMRAcquisitionAuctionAveragePrice = [884,3534], VehBCost = [1400,4775]]	==>	VehicleAge = (4,9]	0.085	0.92	2.36
7	[VehicleAge = [0,2]]	==>	WheelTypeID = 2	0.111	0.70	1.47
8	[VehBCost = [1400,4775], VehicleAge = (4,9]]	==>	IsBadBuy = 0	0.079	0.75	0.86
9	[[VehicleAge = (4,9], VehOdo = (86832,115717]]	==>	IsBadBuy = 0	0.067	0.78	0.89
10	[Size = 0]	==>	Nationality = 0	0.064	0.75	0.85

Le regole più importanti per lo scopo di questo progetto sono quelle che implicano l'attributo IsBadBuy. Tuttavia non sono state ottenute regole aventi come conseguenza "IsBadBuy" con un lift significativamente superiore a 1. Se invece si osservano eventi con lift minore di 1, ossia in cui premessa e conseguenza sono correlate negativamente, si possono trovare regole interessanti:

- (8) Veicoli di età elevata, compresa tra i 4 e 9 anni, con prezzi bassi, tra 1400 e 4775 dollari, tendono a non essere buoni acquisti.
- (9) Veicoli di età elevata, compresa tra i 4 e 9 anni, e con molti chilometri, tra 86832 e 115717, tendono a non essere buoni acquisti.

Sono state invece trovate numerose regole con lift elevati che potrebbero prestarsi a diversi utilizzi non pertinenti al fine di questo progetto.

- (2) Veicoli con costo d'acquisto compreso tra 1400 e 4775 dollari, di età compresa tra 4 e 9 hanno MMR compresi tra 884 e 3534 dollari.
- (3) Veicoli asiatici con copricerchi hanno costo di garanzia tra 462 e 702 dollari.
- (5) Veicoli con cerchi in lega e costo di garanzia tra 1918 e 7498 dollari sono di grandi dimensioni.
- (7) Veicoli con età tra 0 e 2 anni tendono ad avere cerchi in lega.
- (10) Veicoli piccoli tendono a non essere americani.

Essendo il training set estremamente sbilanciato a favore dei buoni acquisti, nelle regole non appare mai IsBadBuy = 1. Per ovviare a questo problema è stato effettuato un downsampling del dataset in modo da avere il 50% di veicoli con IsBadBuy = 0 e 50% con IsBadBuy = 1. Si è quindi effettuato nuovamente il binning per frequenza dei valori numerici e si sono ricalcolate le regole con il medesimo supporto e confidenza minimi.

	Premessa		Conseguenza	Sup	Conf	Lift
1	[VehicleAge = (5,9], MMRAcquisitionAuctionAveragePrice = [884,3,203], VehBCost = [1,620,4,440]]	==>	IsBadBuy = 1	0.05	0.7	1.41
2	[Size = 2, VehicleAge = [1,3]]	==>	IsBadBuy = 0	0.05	0.7	1.4

I risultati, come si può vedere in tabella, sono coerenti con le regole viste in precendenza.

In sintesi da questa analisi risulta che i veicoli più a rischio hanno chilometraggio ed età elevati, prezzi e MMR bassi. Veicoli invece fino ai tre anni di età, di grandi dimensioni, tendono ad essere dei buoni acquisti.

5. Classification

5.1 Preparazione dei dati e *feature selection*

Utilizzando il nodo Feature Elimination di Knime è stato possibile verificare quali fossero gli attributi in grado di minimizzare l'errore di classificazione: VehicleAge, VehOdo, Nationality, Size, MMRAcquisitionAuctionAvgPrice, VehBCost e WarrantyCost. Sugli attributi numerici è stato mantenuto il binning utilizzato per le regole associative. Per quanto riguarda il test set, utilizzando lo stratified sampling e effettuando varie prove, si è deciso di utilizzare il 30% del dataset.

5.2 Creazione di alberi decisionali e migliorie performance

Al fine di trovare il miglior modello decisionale di seguito viene riportata una tabella con i confronti tra tre alberi di decisione (applicati sul test set) con le relative performance. (Per True Positive si intende un NotBadBuy classificato come NotBadBuy)

Alg.	Accuracy	Precision	Recall	F-measure	TP	FP	TN	FN
Decision Tree Learner (Gini)	86,918%	0.878	0.988	0.928	18864	2616	51	231
Decision Tree Learner (Gain)	86,787%	0.878	0.986	0.929	18810	2603	60	270
Random Forest	86,368%	**0.878**	0.980	0.927	18767	2599	**75**	375

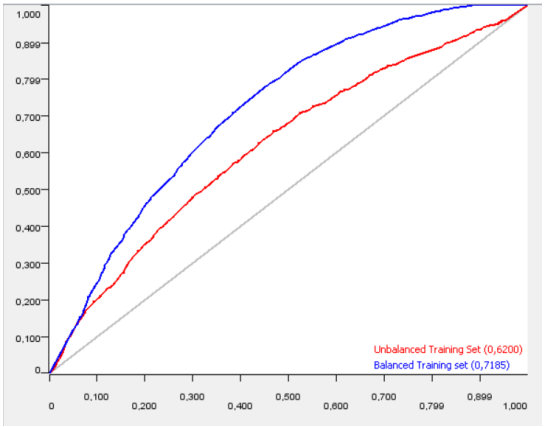
Per l'algoritmo Decision Tree Learner è stato applicato un post pruning con il metodo Reduced Error Pruning, per ridurre il rischio di overfitting. Per la Random Forest è stato utilizzato il nodo di Weka, visto che ha offerto migliori performance. Effettuando un primo confronto tra gli output degli algoritmi, quello che sembra performare meglio è la Random Forest, con un numero più alto sui veri negativi (BadBuy predetti bene).

5.2.1 Confronto tra alberi con dataset bilanciato e sbilanciato

Fino ad ora si è mantenuto lo sbilanciamento tra il numero di cattivi e buoni acquisti. Ai fini di una completa valutazione degli alberi si è deciso di confrontare due alberi su test set diversi: uno realistico (classe sbilanciata) e uno ideale (classe bilanciata). Si è deciso di utilizzare l'Equal Size Sampling per bilanciare le classi. Anche se l'albero con il dataset realistico ha un'accuratezza maggiore rispetto al dataset ideale, quest'ultimo predice (su test set sbilanciato come prima) più cattivi acquisti rispetto al primo. Questo non stupisce, poiché i valori 1 per IsBadBuy sono in netta minoranza (12% del dataset) rispetto agli altri. I risultati del confronto sono mostrati di seguito.

Alg.	Accuracy	Precision	Recall	F-measure	TP	FP	TN	FN
Random Forest (classi sbilanciate)	86,368%	0.878	0.980	0.927	18767	2599	75	375
Random Forest (classi bilanciate)	**61,693%**	0.928	0.611	0.939	11691	906	**1768**	7451

Effettuando i test su dataset sbilanciato, il modello costruito con dataset bilanciato risulta essere migliore per quanto riguarda la predizione dei BadBuy. I modelli costruiti con l'algoritmo Decision Tree Learner (Gini) con i due diversi dataset sono stati confrontati con una curva ROC.

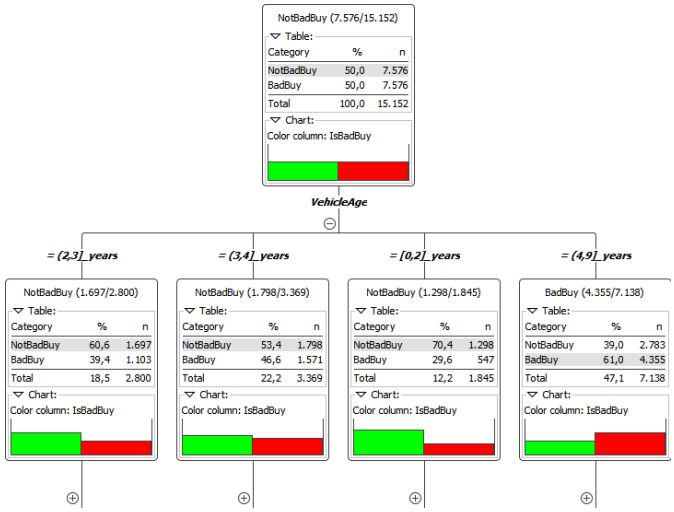


5.3 Interpretazione degli alberi decisionali

Dalla lettura dell'albero decisionale (vedi immagini in fondo alla sezione), si giunge alle seguenti conclusioni:

- Le auto più rischiose da acquistare sul mercato hanno un'età compresa tra i 4 e i 9 anni.
- Tra queste quelle a costo basso (tra 1400 $ e 4765$) hanno una maggiore incidenza di BadBuy.

- Il chilometraggio degli autoveicoli ha un'incidenza negativa sulla bontà di un acquisto, nella categoria delle "datate" quelle con un chilometraggio tra gli 86000 e 115000 km sono tra le più rischiose.
- Le auto meno rischiose da acquistare hanno un'età compresa tra i 0 e i due anni, con un costo di acquisto tra 6300$ e 8500$.
- Tra le automobili giovani quelle più a rischio sono con chilometraggio superiore ai 70000 km.
- Esaminando le automobili ad età media (intervallo 2-3 anni) risulta una maggioranza di NotBadBuy tra le automobili a dimensione Large e di nazionalità Americana.
- Negli autoveicoli ad età media un costo di garanzia basso (tra 462$ e 723$) è indice di BadBuy
- L'MMRAcquisitionAuctionAveragePrice alto (tra 6000$ e 7000$) negli autoveicoli ad età medio-alta (3-4 anni) è sintomo di buon acquisto.



Primo livello del Decison Tree

Focus sui NotBadBuy

Focus sui BadBuy delle automobili datate

In conclusione, secondo le nostre analisi un rivenditore di autoveicoli che vuole minimizzare le probabilità di acquisto di un cosiddetto “BadBuy” dovrebbe tenere conto di alcuni fattori di un’auto: il chilometraggio, l’età del veicolo, il costo del veicolo, il costo della garanzia, la dimensione e la provenienza.
Tutti gli alberi sono stati testati con la cross validation. Per ognuno di essi lo splitting del training set con la migliore accuracy (senza rischio di overfitting è il seguente): N* record per training: 72720 (70%); N* di record per test set: 21816 (30%).

Nelle figure è riportato l'albero ottenuto dal nodo Knime Decision Tree Learner, visto che il nodo Random Forest di Weka non riporta un albero grafico come output.

5.4 Il miglior modello predittivo

Durante le analisi si è potuta constatare una situazione contrastante sul significato di Accuracy nel caso in questione, infatti se è vero che il modello originale sbilanciato usato come training ha riportato un Accuracy alta (88% circa), è vero anche che il modello bilanciato, utilizzando un Equal Sampling sui BadBuy, ha classificato correttamente un maggior numero di acquisti "rischiosi".
Questo comportamento dei modelli di prediction è dovuto al fatto che il dataset iniziale disponeva di un numero esiguo di BadBuy (12%) rispetto al totale e, di conseguenza, si è deciso di utilizzare per la classification un modello bilanciato.
Tra gli algoritmi presi in considerazione il migliore è risultato il Random Forest di Weka.
Secondo il nostro parere, il classificatore con training bilanciato è il migliore, perché consente di identificare meglio le automobili a rischio e quindi ideale per l'obiettivo della ricerca