

PROGETTO DI DATA MINING

2019/2020

"IS IT A BADBUY?"

WILLIAM GUGLIELMO, 516326

ALDO IANNARELLI, 602009

VITALBA MACALUSO, 603160

DANIELE TRIBUZIO, 602020

Progetto di Data Mining

7 gennaio 2020

Indice

1 Comprensione dei dati	2
1.1 Semantica dei dati	2
1.2 Distribuzione delle variabili e statistiche	3
1.3 Valutazione della qualità dei dati	6
1.3.1 Missing Values	6
1.3.2 Outliers	8
1.4 Trasformazione delle variabili	8
1.5 Correlazioni a coppie ed eliminazione di eventuali variabili ridondanti	9
1.5.1 Correlazioni a coppie	9
1.5.2 Eliminazione di eventuali variabili ridondanti	9
2 Clustering	10
2.1 Clusterizzazione con tecnica K-Means	10
2.2 Clusterizzazione con tecnica DBscan	11
2.3 Clusterizzazione con tecnica gerarchica agglomerativa	12
2.4 Valutazione finale del migliore algoritmo di clusterizzazione	13
3 Association Rules	14
3.1 Estrazione dei pattern frequenti e analisi del numero di pattern rispetto al parametro MinSup	14
3.1.1 Itemsets frequenti	14
3.1.2 Itemsets chiusi	15
3.1.3 Itemsets massimali	15
3.2 Estrazione delle regole associative per differenti valori di MinConf	16
3.3 Regola associative utili per il trattamento dei valori mancanti	17
3.4 Regola associative utili per predire la variabile target	17
4 Classification	17
4.1 Learning	17
4.2 Decision Tree	18
4.2.1 Interpretazione dell'albero	19
4.2.2 Confronto accuracy senza il filtraggio del Dataset	19
4.3 Random Forest	19
5 Conclusioni	20

1 Comprensione dei dati

Il DataSet fornito appartiene a Carvana, startUp americana il cui obiettivo è quello di rivoluzionare il modo in cui acquistare e rivendere veicoli usati, facendo della qualità una prerogativa fondamentale. Carvana si discosta dalle più tradizionali modalità di rivendita dei veicoli, facendo proprie tecnologie e modelli scientifici che sono top of the line.

I veicoli che Carvana rivende sono provenienti da vendite all'asta. Ogni acquisto viene gestito da un team di *buyers* interni con competenze specifiche sulla categoria di veicoli. Il modello di business di Carvana utilizza un algoritmo che consente di valutare se un determinato veicolo in vendita, proveniente dal wholesale, rispetti gli standard di qualità necessari a superare l'ispezione condotta da meccanici con certificazione ASE.

L'obiettivo del lavoro è quello di implementare un modello in grado di individuare se il veicolo che verrà acquistato all'asta risulti un Bad-Buy o meno.

1.1 Semantica dei dati

Il DataSet consiste di 58386 records relativi a veicoli su cui sono state rilevate 34 variabili di cui 1 sola variabile target (IsBadBuy).

- **RefId** (numerica): ID assegnato al veicolo;
- **IsBadBuy** (categoriale dicotomica): se il veicolo è da considerarsi un Bad-Buy (valore 1) o no (valore 0);
- **PurchDate** (categoriale): data in cui il veicolo è stato acquistato all'asta;
- **Auction** (categoriale): fornitore dell'asta presso il quale il veicolo è stato acquistato (ADESA, MANHEIM, OTHER);
- **VehYear** (numerica): anno di produzione del veicolo;
- **VehicleAge** (numerica): età in anni del veicolo;

In questa sezione possiamo trovare tutte le variabili relative ai Brand:

- **Make** (categoriale): produttore del veicolo;
- **Model** (categoriale): modello del veicolo;
- **Trim** (categoriale): assetto del veicolo;
- **SubModel** (categoriale): sotto-modello del veicolo;

Di seguito vengono elencate le caratteristiche specifiche dei veicoli:

- **Color** (categoriale): colore del veicolo;
- **Transmission** (categoriale dicotomica): se il cambio del veicolo è automatico o manuale;
- **WheelTypeID** (numerica): valore assegnato al tipo di ruote del veicolo;

- **WheelType** (categoriale): tipo di ruote del veicolo;
- **VehOdo** (numerica): contatore chilometrico del veicolo;
- **Nationality** (categoriale): paese del produttore del veicolo;
- **Size** (categoriale): categoria di appartenenza del veicolo rispetto alle dimensioni;
- **TopThreeAmericanName** (categoriale): se il produttore è uno dei tre maggiori produttori americani.

Nel Database troviamo indicatori come il Manheim Market Report (MMR), il principale indicatore dei prezzi nel wholesale. I prezzi calcolati si basano su oltre 10 milioni di transazioni di vendita negli ultimi 13 mesi. Questi sono:

- **MMRAcquisitionAuctionAveragePrice**: (numerica): il prezzo dei veicoli in medie condizioni al momento dell'acquisto;
- **MMRAcquisitionAuctionCleanPrice**: (numerica): il prezzo del veicolo in condizioni superiori alla media al momento dell'acquisto;
- **MMRAcquisitionRetailAveragePrice**: (numerica): il prezzo d'acquisto del veicolo ,in medie condizioni, presso il concessionario al momento dell'acquisto;
- **MMRAcquisitionRetailCleanPrice**: (numerica): il prezzo di acquisizione dell'auto in condizioni superiori alla media in concessionario al momento dell'acquisto;
- **MMRCurrentAuctionAveragePrice**: (numerica): il prezzo dell'automobile in medie condizioni nell'anno corrente del DataSet;
- **MMRCurrentAuctionCleanPrice**: (numerica): il prezzo dell'auto in condizioni superiori alla media nell'anno corrente del DataSet;
- **MMRCurrentRetailAveragePrice**: (numerica): il prezzo d'acquisto del veicolo ,in condizioni medie, presso il rivenditore d'auto nell'anno corrente del DataSet;
- **MMRCurrentRetailCleanPrice**: (numerica): il prezzo d'acquisto in condizioni sopra la media presso concessionari nell'anno corrente del DataSet.

Le altre variabili rimanenti sono:

- **PRIMEUNIT** (categoriale dicotomica): se il veicolo presenta una domanda al di sopra dello standard (YES) o no (NO);
- **AUCGUART** (categoriale ordinale): livello di garanzia fornito dall'asta per il veicolo;
- **BYRNO** (numerica): ID assegnato all'acquirente del veicolo;

- **VNZIP1** (numerica): codice postale del luogo in cui il veicolo è stato acquistato;
- **VNST** (categoriale): Paese in cui il veicolo è stato acquistato;
- **VehBCost** (numerica): costo di acquisto del veicolo al momento dell'asta;
- **IsOnlineSale** (categoriale dicotomica): se il veicolo è stato originariamente acquistato online (valore 1) o no (valore 0);
- **WarrantyCost** (numerica): prezzo della garanzia per 36 mesi e 36k miglia.

1.2 Distribuzione delle variabili e statistiche

La distribuzione della variabile target è fortemente sbilanciata perché l'87.7% dei veicoli risulta essere un Good-Buy. La Figura 1.1 rappresenta la distribuzione di questa variabile.

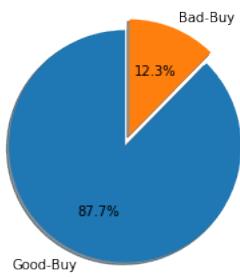


Figura 1.1: Percentuale di veicoli Bad-Buy e Good-Buy

Ora ci si focalizzerà sulla distribuzione di alcune variabili presenti nel DataSet.

Auction. Il dominio di Auction è rappresentato da tre categorie riferite al fornitore d'asta: 'ADESA', 'MANHEIM' e 'OTHER'. La distribuzione di questa variabile è rappresentata nella Figura 1.2, in cui nell'asse delle ascisse sono inserite le tre categorie e nell'asse delle ordinate le frequenze relative. Il 56% circa dei veicoli presenti nel DataSet è stato formato da 'MANHEIM', mentre il 20% circa da 'ADESA'.

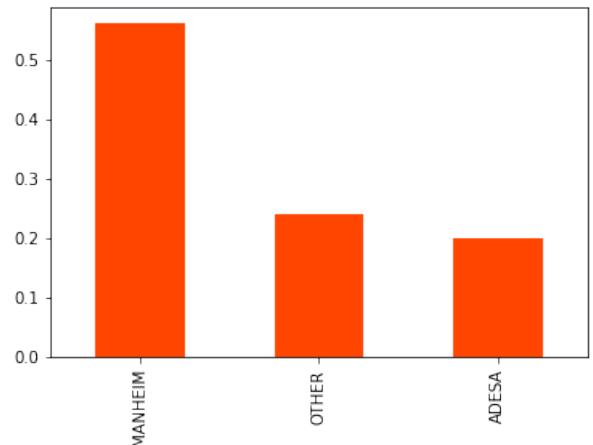


Figura 1.2: Bar plot dei fornitori d'asta

VehYear. Di tutti i veicoli, soltanto uno presenta età pari a 0 (produzione nell'anno 2010) e nessuna auto risulta essere stata prodotta più di 9 anni prima della costruzione del DataSet. L'anno in cui si sono prodotti veicoli con maggiore frequenza risulta essere il 2006, con 13668 unità. Il 50% dei veicoli è stato prodotto prima del 2006 (la media infatti si trova in corrispondenza dell'anno 2005), mentre l'anno medio di produzione è tra il 2005 e il 2006. Da ciò possiamo dedurre che la distribuzione di tale variabile non è perfettamente simmetrica, come si può vedere dalla Figura 1.3, ma leggermente asimmetrica positivamente.

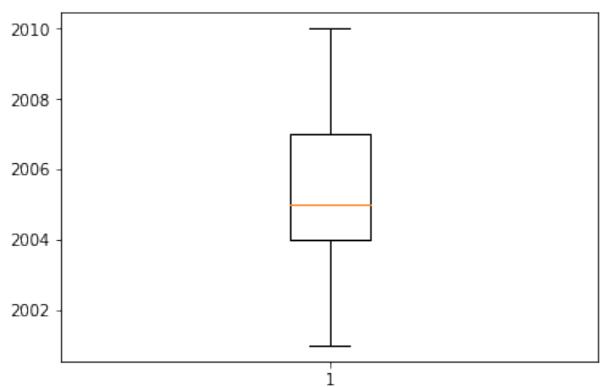


Figura 1.3: Box plot dell'anno di produzione dei veicoli

Make. Dalla Figura 1.4 si può notare come il maggiore produttore di veicoli acquistati da Carvana sia 'CHEVROLET' con il 23.7%, seguito da 'DODGE' con il 17.7% e da Ford con il 15.4%. Tutti i veicoli degli altri 30 produttori sono in percentuale inferiori al 13%.

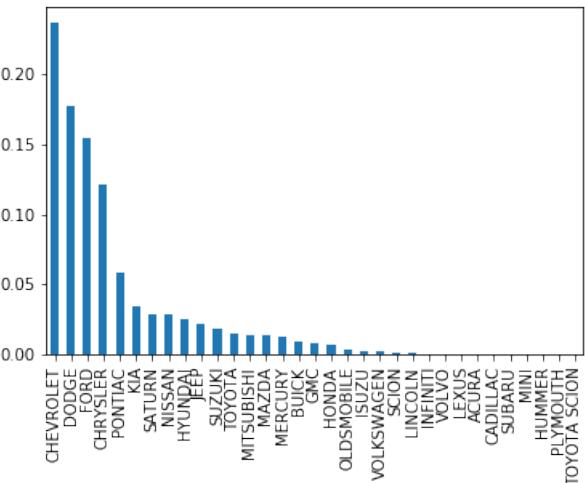


Figura 1.4: Bar plot dei produttori di veicoli acquistati da Carvana

Transmission. Il dominio di questa variabile è rappresentato da due categorie: ‘AUTO’ e ‘MANUAL’. Il 96.5% dei veicoli presenti nel DataSet ha il cambio automatico, quindi la distribuzione risulta fortemente sbilanciata rispetto alla variabile presa in esame. Si è voluto rappresentare la distribuzione del tipo di cambio rispetto ai produttori di veicoli, mediante bar plot impilato. Il risultato si può visualizzare nella Figura 1.5. Dalla rappresentazione grafica si evince che il produttore con la maggiore percentuale di auto con cambio manuale è Mini (41.2%).

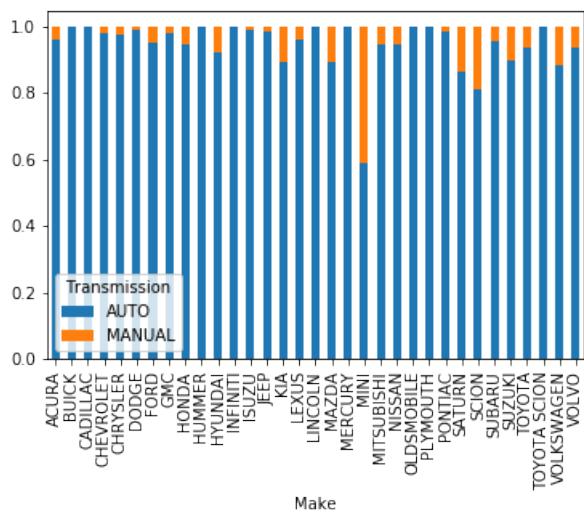


Figura 1.5: Bar plot del tipo di cambio del veicolo rispetto ai produttori

WheelType. Relativamente al tipo di ruote, la variabile presenta tre categorie: ‘Alloy’, ‘Covers’ e ‘Special’. Nella Figura 1.6 è mostrata la distribuzione della variabile. Soltanto l’1.1% dei veicoli possiede come tipologia di ruote ‘Special’. La variabile, inoltre, presenta 2577 missing values.

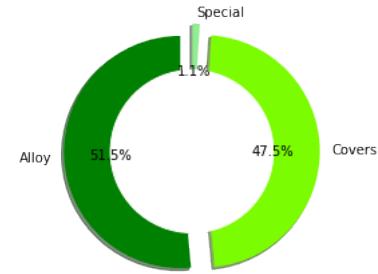


Figura 1.6: Donut chart del tipo di ruote dei veicoli

VehOdo. Questa variabile non presenta missing values e, mediante la regola di Sturges, abbiamo stabilito che il numero ottimale di classi necessario per la rappresentazione grafica è pari a 17. Il density plot e l’istogramma ottenuti (Figura 1.7) mostrano che nel DataSet si ha maggiore frequenza di unità il cui contachilometri segna tra 70.000 e 80.000 km.

Si è scelto inoltre di verificare mediante il test di Anderson-Darling se la variabile può considerarsi distribuita normalmente. Il valore osservato della statistica test è risultato pari a 333.327. Dal momento il valore critico al 5% è uguale a 0.787, l’ipotesi di distribuzione normale della variabile presa in esame è stata scartata.

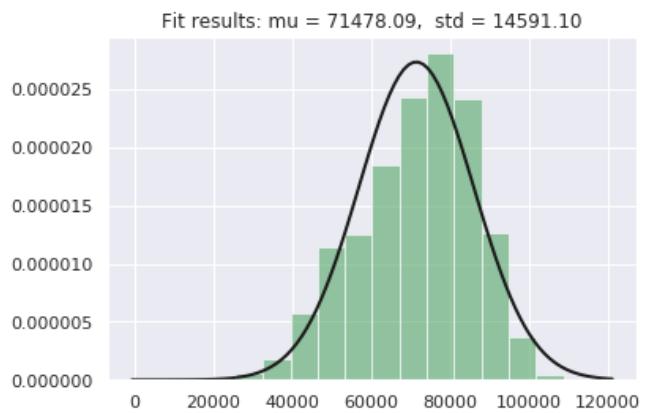


Figura 1.7: Density plot dei chilometri percorsi dai veicoli

Nationality & TopThreeAmericanName. La variabile TopThreeAmericanName può assumere i valori: ‘CHRYSLER’, ‘FORD’, ‘GM’ e ‘OTHER’. Un veicolo è stato prodotto da uno dei tre maggiori produttori americani se il valore di questa variabile appartiene all’insieme {‘CHRYSLER’, ‘FORD’, ‘GM’}.

Nel DataSet non sono presenti veicoli americani che non siano stati prodotti da uno dei tre maggiori produttori americani. Le percentuali di distribuzione sono mostrate in Figura 1.8.

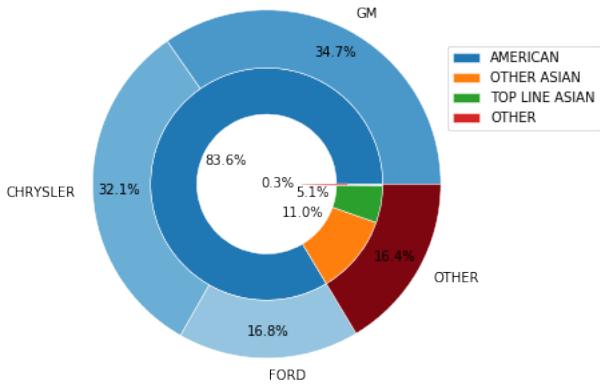


Figura 1.8: Il donut chart esterno rappresenta la distribuzione di TopThreeAmericanName, mentre il donut chart interno rappresenta la distribuzione di Nationality

VehBCost. In media, il costo d'acquisto del veicolo all'asta è intorno ai 6730 dollari. Per avere un andamento dei costi dei veicoli, si è optato per un istogramma con numero di classi pari a 17, applicando la regola di Sturges, unito ad un density plot. Come si deduce dalla Figura 1.9, la frequenza maggiore è compresa tra i 7000 e i 8000 dollari.

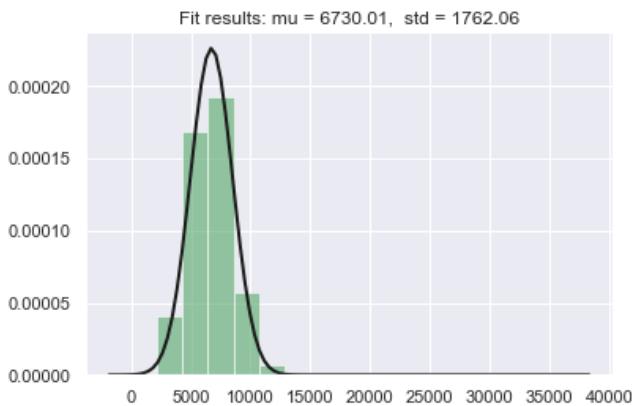


Figura 1.9: Costo di acquisizione del veicolo al tempo di vendita

WarrantyCost. Questa variabile non presenta missing values e, quindi, scegliamo di rappresentare anche questa variabile con 17 classi.

I valori minimo e massimo della variabile sono rispettivamente 462 e 7498 dollari. Per rappresentare la distribuzione è stato utilizzato l'istogramma in Figura 1.10. Dal grafico si può osservare che la maggior parte dei valori è compresa fra 500 e 1500 dollari; allontanandosi da questi valori, il numero di veicoli tende via via ad abbassarsi.

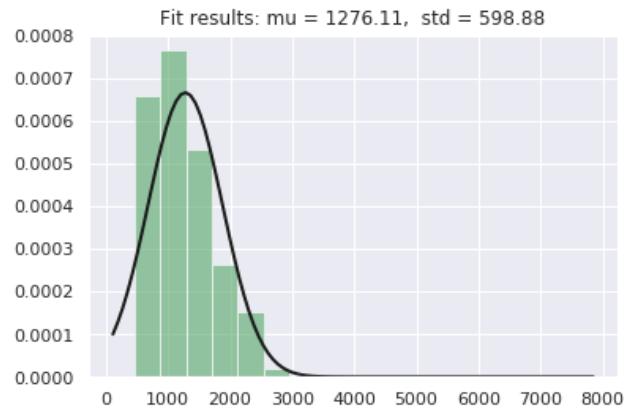


Figura 1.10: Density plot del prezzo di garanzia dei veicoli

MMRPrices. Affinché i prezzi del Report Manheim Market possano essere compresi al meglio, in questa fase è stato deciso di rappresentare un bar chart multiplo in cui viene considerata la media dei prezzi delle auto suddivise per Size, confrontando gli indicatori MMR al tempo di acquisto (che nelle figure riportano un colorazione più scura) con quelli al tempo corrente (con colorazione più chiara). La suddivisione per Size consente di avere un'idea di come si distribuiscono i prezzi a seconda della dimensione del veicolo. Emerge chiaramente come i veicoli di piccole dimensioni, che hanno in genere basse prestazioni, abbiano un costo inferiore rispetto a quelli di grandi dimensioni, che hanno prestazioni maggiori.

Dalle Figure 1.11, 1.12, 1.13 e 1.14 si può notare che:

- per i prezzi relativi all'asta non si verificano significativi scostamenti tra il tempo di acquisto e il tempo corrente;
- per i prezzi relativi al mercato del retail vi è un aumento al tempo corrente rispetto al tempo di acquisto.

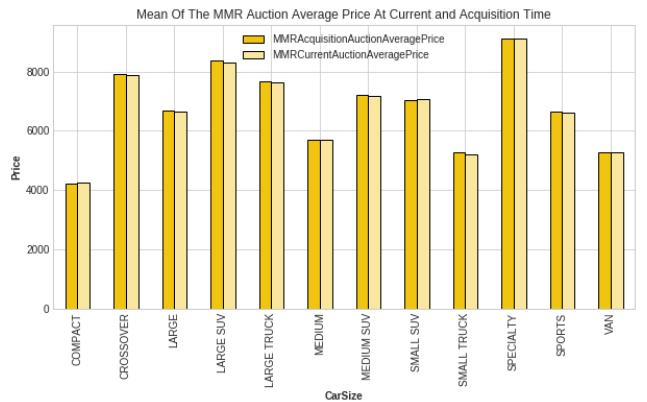


Figura 1.11: Media dei prezzi MMR Auction Average suddivisi per CarSize

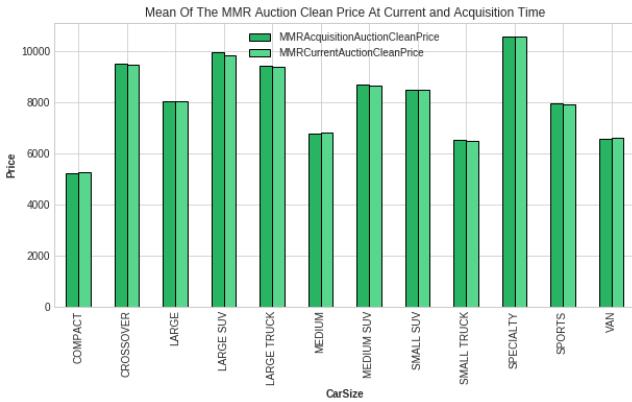


Figura 1.12: Media dei prezzi MMR Auction Clean suddivisi per CarSize

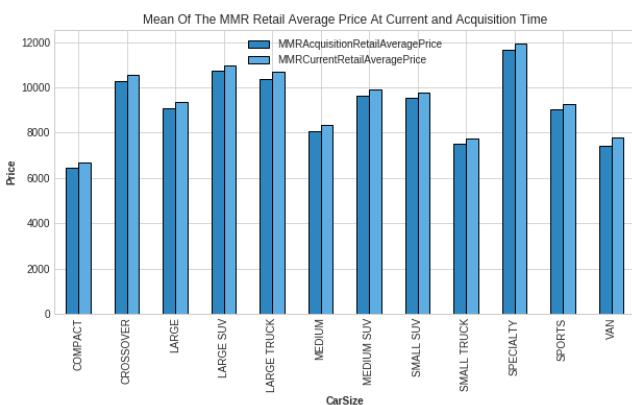


Figura 1.13: Media dei prezzi MMR Retail Average suddivisi per CarSize

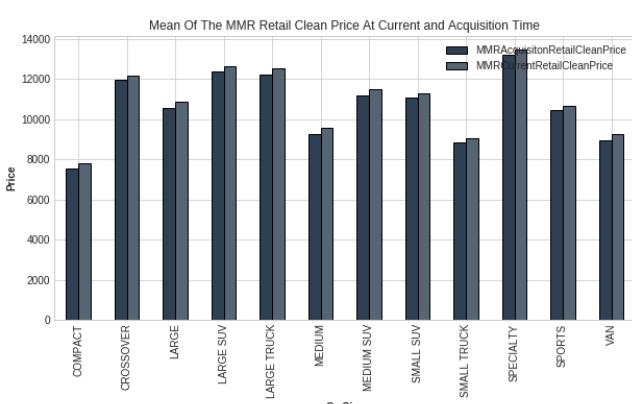


Figura 1.14: Media dei prezzi MMR Retail Clean suddivisi per CarSize

1.3 Valutazione della qualità dei dati

Per primo, si è verificata l'accuratezza sintattica. La variabile Transmission presenta un valore ‘Manual’ che appartiene al dominio ma non è scritto correttamente: tale valore viene

sostituito con ‘MANUAL’.

Inoltre, alcune delle variabili quantitative relative al prezzo e la variabile ‘WheelTypeID’, presentano dei valori pari a 0 che, per un prezzo non sembra essere un valore ragionevole così come per un ID del tipo di ruota.

Si nota inoltre come, considerando Make (CADILLAC), Model (SRX AWD V6 3.6L V6 S) e SubModel (4D SUV), si ha un veicolo che possiede due valori differenti in riferimento a Size (‘MEDIUM SUV’ e ‘SPECIALTY’). Si provvede a modificare questo attributo utilizzando l’attributo ‘MEDIUM SUV’ nella variabile Size in quanto è la moda e il SubModel corrispondente riporta la dicitura SUV.

ATTRIBUTO	NULL VALUES
WheelTypeID	4
MMRAcquisitionAuctionAveragePrice	648
MMRAcquisitionAuctionCleanPrice	552
MMRAcquisitionRetailAveragePrice	648
MMRAcquisitionRetailCleanPrice	648
MMRCurrentAuctionAveragePrice	393
MMRCurrentAuctionCleanPrice	300
MMRCurrentRetailAveragePrice	393
MMRCurrentRetailCleanPrice	393
VehBCost	0
WarrantyCost	0

Tabella 1.1: Valori nulli per variabile

Successivamente, sono stati trattati i valori nulli come missing values. La Tabella 1.1 fornisce il conteggio dei valori nulli per ciascuna delle variabili numeriche.

1.3.1 Missing Values

La Tabella 1.2 mostra gli attributi in cui sono presenti missing values, ordinati in senso decrescente. Le variabili PRI-MEUNIT e AUCGUART presentano più del 95% di missing values. Prendendo in esame il loro significato, si è deciso di considerare queste variabili soltanto nell’ambito della classificazione dei veicoli, con opportuni accorgimenti. Le altre variabili presenti nella Tabella hanno un numero di missing values decisamente inferiore e verranno trattate successivamente.

ATTRIBUTES	MISSING VALUES	MISSING VALUES %
PRIMEUNIT	55703	95,40%
AUCGUART	55703	95,40%
WheelType	2577	4,41%
WheelTypeID	2577	4,41%
Trim	1911	3,27%
MMRAcquisitionAuctionAveragePrice	661	1,13%
MMRAcquisitionRetailAveragePrice	661	1,13%
MMRAcquisitionRetailCleanPrice	661	1,13%
MMRCurrentAuctionAveragePrice	638	1,09%
MMRCurrentRetailAveragePrice	638	1,09%
MMRCurrentRetailCleanPrice	638	1,09%
MMRAcquisitionAuctionCleanPrice	565	0,97%
MMRCurrentAuctionCleanPrice	545	0,93%
Transmission	8	0,01%
SubModel	7	0,01%
Color	7	0,01%
Nationality	4	0,01%
Size	4	0,01%
TopThreeAmericanName	4	0,01%

Tabella 1.2: Missing values per attributo

La Heatmap in Figura 1.15 misura la correlazione di nullità, cioè con quale misura la presenza o l'assenza di una variabile influisce sulla presenza di un'altra. La correlazione di nullità va da -1 (se la presenza di una variabile è correlata alla mancanza di un'altra), a 0 (se le variabili che appaiono o non appaiono non hanno alcun effetto l'una sull'altra) e infine a 1 (se la presenza di una variabile è correlata alla presenza di un'altra).

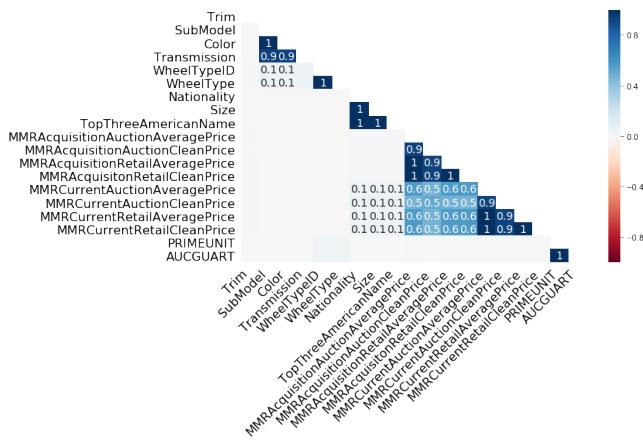


Figura 1.15: Heatmap dei missing values

Nel DataSet fornito non appaiono correlazioni di nullità negativa, ma ci sono diverse correlazioni positive pari o prossime a 1. La presenza di Color è correlata alla presenza di SubModel che, a loro volta, sono correlate con la presenza di Transmission.

Come ci si può aspettare, la presenza della variabile WheelType è totalmente correlata con quella di WheelTypeID (ogni ID del tipo di ruote, infatti, fa riferimento ad un tipo di ruote: 1.0 indica Alloy, 2.0 indica Covers e 3.0 indica Special).

La variabile TopThreeAmericanName è correlata con la presenza di Nationality e Size. Infine, l'assenza della variabile PRIMEUNIT è correlata totalmente con l'assenza di AUC-

GUART.

In particolare, tra queste ultime due variabili sembra esserci associazione: nonostante i missing values, per i veicoli ad alto rischio non c'è domanda, mentre per i veicoli a basso rischio la domanda è molto alta (come mostrato nella Tabella 1.3).

AUCGUART	GREEN	RED
NO	2565	62
YES	56	0

Tabella 1.3: Tabella di frequenza di PRIMEUNIT e AUCGUART

Ritornando alle variabili in cui sono presenti missing values, si è optato per:

- sostituire i valori mancanti con i rispettivi valori più frequenti, considerando il modello del veicolo;
- sostituire i valori mancanti di Trim facendo riferimento al modello, al sotto-modello e alle dimensioni del veicolo, in quanto si è visto che determinati modelli e sotto-modelli (con certe dimensioni) presentano soltanto alcune tipologie di assetto. In caso di più tipologie di assetto, si prenderà la moda;
- sostituire i valori mancanti delle variabili relative agli indici dei prezzi con la rispettiva mediana. Questa scelta è stata dettata dal fatto che, in tali variabili, sono presenti prezzi molto elevati relativi ad auto molto più costose rispetto alle altre (si ipotizza in questa fase che siano auto di lusso). Pertanto la mediana, essendo uno stimatore robusto, risente meno di questi valori rispetto alla media. Per maggiore chiarezza, nella Tabella 1.4 sono mostrate le mediane relative agli indici dei prezzi;

ATTRIBUTO	MEDIANA
MMRAcquisitionAuctionAveragePrice	6133.0
MMRAcquisitionAuctionCleanPrice	7350.0
MMRAcquisitionRetailAveragePrice	8484.0
MMRAcquisitionRetailCleanPrice	9844.0
MMRCurrentAuctionAveragePrice	6086.0
MMRCurrentAuctionCleanPrice	7330.0
MMRCurrentRetailAveragePrice	8756.0
MMRCurrentRetailCleanPrice	10128.0

Tabella 1.4: Mediane degli indici dei prezzi MMR

- sostituire i valori mancanti di Transmission con la sua moda, considerando ciascun modello del veicolo;
- sostituire i valori mancanti della variabile SubModel prendendo la moda rispetto al modello e alle dimensioni del veicolo;

- sostituire i valori mancanti di Color con la categoria ‘NOT AVAIL’;
- sostituire i valori mancanti della variabile Nationality, con ‘AMERICAN’ se questo è un TopThreeAmericanName; sostituire i valori mancanti rimanenti prendendo la moda rispetto al produttore del veicolo, in quanto si ipotizza che un produttore produca i propri veicoli, in genere, sempre nello stesso Paese;
- sostituire i valori mancanti della variabili TopThreeAmericanName con la moda rispetto alla nazionalità dell’auto.
- sostituire i valori mancanti della variabili Size con la moda rispetto al modello del veicolo;
- sostituire i valori mancanti delle variabili AUCGUART e PRIMEUNIT con la modalità ‘NOT AVAIL’. Tale sostituzione si è mostrata molto utile nell’ambito della classificazione dei veicoli in Bad-Buy e non Bad-Buy. Come verrà mostrato in seguito, infatti, AUCGUART è risultata essere la terza migliore variabile di classificazione.

1.3.2 Outliers

Per la ricerca degli outliers è stato utilizzato uno scatterplot, in cui nell’asse delle ascisse poniamo la variabile MMRAcquisitionAuctionAveragePrice e nell’asse delle ordinate la variabile VehBCost.

La scelta di queste due variabili è stata dettata dal fatto che, essendo gli indici MMR fortemente correlati tra loro (come si vedrà in seguito, la correlazione più bassa è pari a 0.88), le due variabili risultano essere le maggiormente correlate, con un coefficiente di correlazione pari a 0.82. Al di sopra del valore di 16345 dollari della variabile VehBCost sono stati individuati 10 potenziali outliers. Dalla Figura 1.16 è possibile vedere anche che questi ultimi sono tutti Bad-Buy e appartengono alle categorie ‘MEDIUM SUV’, ‘SPECIALTY’ e ‘LARGE TRUCK’ della variabile Size.

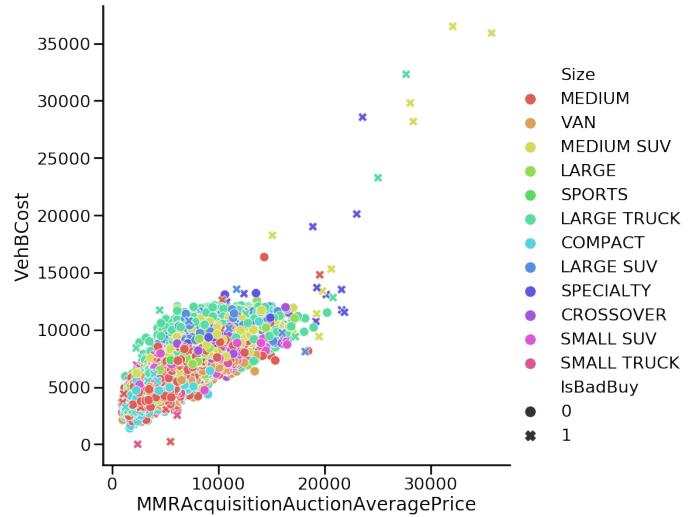


Figura 1.16: Scatterplot di MMRAcquisitionAuctionAveragePrice e VehBCost con punti differenziati per Size e IsBadBuy

1.4 Trasformazione delle variabili

Per prima cosa, si è optato per binarizzare la variabile Transmission con 0, per la categoria ‘MANUAL’ e con 1 per la categoria ‘AUTO’.

Inoltre, tenendo conto del fatto che ‘CHRYSLER’, ‘DODGE’, ‘JEEP’, ‘PLYMOUTH’ appartengono al produttore ‘CHRYSLER’, ‘FORD’, ‘LINCOLN’, ‘MERCURY’ appartengono al produttore ‘FORD’ e ‘BUICK’, ‘CADILLAC’, ‘CHEVROLET’, ‘GMC’, ‘OLDSMOBILE’, ‘PONTIAC’, ‘SATURN’ appartengono al produttore ‘GM’, si è scelto di aggregare le categorie ‘CHRYSLER’, ‘FORD’, ‘GM’ della variabile TopThreeAmericanName come ‘American Top3’ e successivamente di binarizzarla con valore 1 se i produttori appartengono alle Top3 americana e con valore 0 se non vi appartengono. La Figura 1.17 mostra la distribuzione della ‘nuova’ variabile.

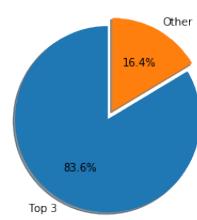


Figura 1.17: Distribuzione della nuova variabile TopThreeAmericanName trasformata

I valori della variabile WheelType sono stati trasformati da float a interi.

La variabile PurchDate è stata trasformata da una stringa nel formato ‘mm/gg/aaaa’ a un intero contenente solo l’anno.

Infine, sono state adottate delle trasformazioni su tutti gli indici MMR, i quali risultano altamente correlati tra loro e con la variabile VehBCost (come vedremo in seguito). Al fine di migliorare la qualità dell'analisi è stato preferito:

- trasformare gli indici MMR relativi al tempo d'acquisto, in termini di "margine di guadagno massimo possibile":

- 1) Difference=MMR_Acquisition-VehBCost , per ognuno dei 4 MMR;
- 2) Markup_max=Difference_Auction+Difference_Retail.

- trasformare anche gli indici MMR relativi al tempo corrente in termini di "margine di guadagno massimo possibile", costruendo un VehBCost corrente:

- 1) Rate_of_Change=MMR_Current/MMR_Acquisition , per ognuno dei 4 MMR;
- 2) ROC_average=Media tra i Rate_of_Change;
- 3) VehBCost_current=VehBCost*ROC_average;
- 4) Difference=MMR_Current-VehBCost , per ognuno dei 4 MMR;
- 5) Markup_max=Difference_Auction+Difference_Retail.

Queste trasformazioni hanno aiutato ad eliminare l'elevata correlazione tra le variabili suddette e ridurre la dimensionalità del DataSet, senza però perdita di informazioni.

1.5 Correlazioni a coppie ed eliminazione di eventuali variabili ridondanti

1.5.1 Correlazioni a coppie

La Figura 1.18 mostra la correlazione esistente tra le variabili quantitative originarie del DataSet.

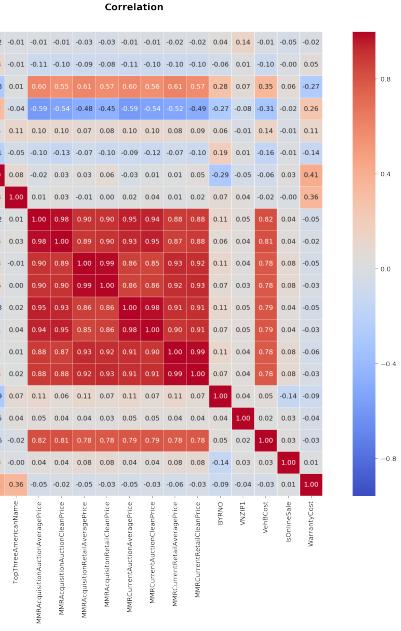


Figura 1.18: Correlazione di Pearson tra le variabili

Le variabili VehYear e VehicleAge risultano correlate al 96%. Inoltre, tali variabili risultano mediamente correlate negativamente con gli indici MMR: i coefficienti di correlazione variano dal -49% al -61%.

La variabile VehBCost aveva un'alta correlazione con gli indici MMR prima di effettuare le trasformazioni. Infine, come precedentemente accennato, tutti gli indici dei prezzi MMR risultavano altamente correlati tra loro.

Nella Figura 1.19 si può vedere come, sia per i Bad-Buy che per i Good-Buy, l'indice dei prezzi considerato diminuisce all'aumentare degli anni del veicolo. Fino al terzo anno di età del veicolo, i veicoli che non sono Bad-Buy presentano un indice dei prezzi inferiore rispetto a quelli Bad-Buy. Oltre i 3 anni di età, l'indice dei prezzi dei veicoli Bad-Buy risulta inferiore rispetto ai Good-Buy. Questo fenomeno potrebbe essere legato ad una maggiore svalutazione dei veicoli Bad-Buy.

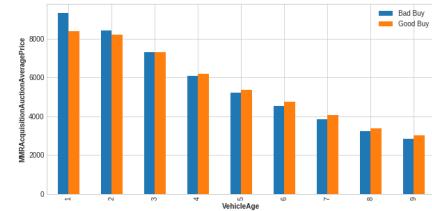


Figura 1.19: Correlazione di Pearson tra le variabili

1.5.2 Eliminazione di eventuali variabili ridondanti

All'interno del DataSet, si riscontra la presenza di variabili ridondanti (che restituiscono lo stesso significato di altre già presenti). Nella scelta tra due variabili ridondanti, viene

prediletta quella che ipoteticamente potrebbe essere più funzionale per le analisi successive. In quest'ottica, si è stabilito di eliminare la variabile VNZIP1 e tenere in considerazione la variabile VNST perché, sebbene abbia meno valori (152 contro 37), restituisce la medesima informazione della variabile eliminata. Inoltre, si è deciso di tenere in considerazione la variabile WheelType e di rimuovere WheelTypeID perché, pur essendo WheelTypeID numerica, il suo dominio reale fa riferimento ad una variabile categoriale. La variabile VehYear è stata eliminata ed è stata considerata soltanto la variabile VehicleAge per entrambi i motivi citati precedentemente.

Infine, anche se non si possono considerare ridondanti, sono state eliminate anche le variabili RefId e BYRNO, in quanto l'ID rispettivamente dei veicoli e dei *buyers* risultano essere irrilevanti ai fini dell'analisi.

2 Clustering

In questa sezione ci si occuperà di clusterizzare le unità presenti nel DataSet, al fine di ottenere gruppi di veicoli che presentano caratteristiche comuni. Ciò che ci si aspetta è di trovare eventuali correlazioni che, nella fase precedente di Data Understanding, non erano emerse. Gli attributi scelti per effettuare l'analisi dei cluster saranno tutti quantitativi e verranno specificati di volta in volta per ogni tecnica utilizzata. Per prima cosa, verrà applicata la tecnica di clusterizzazione K-means. Successivamente, verranno mostrati anche i risultati relativi alle tecniche DBSCAN e Hierarchical.

Si è deciso, inoltre, di non considerare la variabile VehicleAge tra le numeriche scelte per la clusterizzazione, in quanto non è di interesse trovare clusters di veicoli che presentano come caratteristica comune l'età. Considerando VehicleAge nella clusterizzazione, i risultati ottenuti si basano quasi esclusivamente sui valori assunti da tale variabile. Il nostro obiettivo è stato quello di identificare eventuali correlazioni con le variabili considerate.

2.1 Clusterizzazione con tecnica K-Means

Per l'applicazione del K-means sono stati scelti 7 attributi: VehOdo, VehBCost, Markup_Acquisition_AP_max, Markup_Acquisition_CP_max, Markup_Current_AP_max, Markup_Current_CP_max e WarrantyCost. La funzione di distanza utilizzata dall'algoritmo sarà la distanza Euclidea.

Il valore iniziale dei centroidi è stato definito utilizzando il metodo 'k-means++'. Questo metodo inizializza i centroidi in modo che siano distanti tra loro, permettendo così risultati migliori rispetto al metodo di inizializzazione casuale. Quest'ultimo, infatti, potrebbe condurre ad una soluzione non ottimale ma locale. Con il metodo scelto, invece, i centroidi saranno più precisi e pertanto saranno necessarie meno iterazioni. Per identificare il numero di cluster ottimale (k), si è scelto di calcolare il valore dell'SSE e del coefficiente Silhouette per un numero di clusters che va da 2 a

15. Successivamente, sono stati rappresentati graficamente gli andamenti di queste due misure per i diversi valori di k considerati, come mostrano la Figura 2.1 e 2.2.

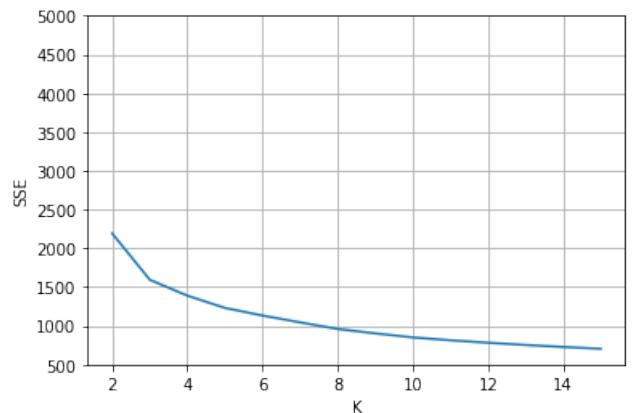


Figura 2.1: Andamento dell'SSE per valori di k crescenti

Il valore dell'SSE diminuisce al crescere di k e, a partire da $k=3$, si può notare che questo decresce in misura sempre minore.

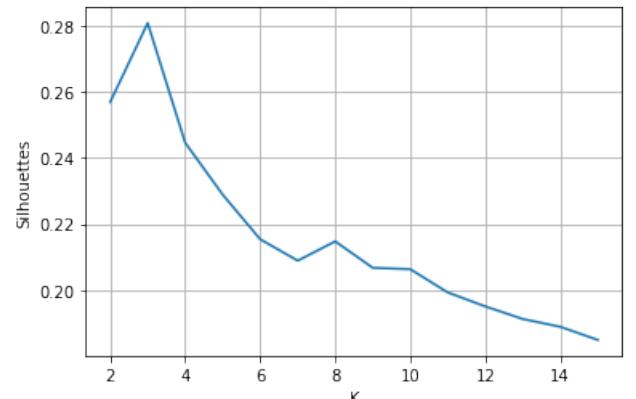


Figura 2.2: Andamento del coefficiente Silhouette per valori di k crescenti

Il valore del coefficiente Silhouette nel migliore dei casi è prossimo a 1, pertanto si ricerca il valore di k tale per cui questo risulta massimo.

Il valore ottimale di k è stato ottenuto, quindi, tenendo in considerazione il 'gomito' della curva SSE e il 'picco' della curva Silhouette. Il numero di clusters ottimale trovato è pari a 3. Per questo valore di k , l'SSE risulta pari a 1594.66, mentre il Silhouette score è 0.28.

La Tabella 2.1 e la Figura 2.3 mostrano le caratteristiche dei centroidi ottenuti.

	VehOdo	VehBCost	Markup_AAP	Markup_ACP	Markup_CAP	Markup_CCP	WarrantyCost
Cluster 0	54443,99	6722,74	951,17	3223,57	1253,51	3561,20	910,11
Cluster 1	79408,35	6257,52	-771,41	1744,44	-547,24	1969,46	1481,78
Cluster 2	78829,29	7401,44	4683,97	7812,75	4807,76	7958,34	1384,50

Tabella 2.1: Tabella di descrizione dei centroidi

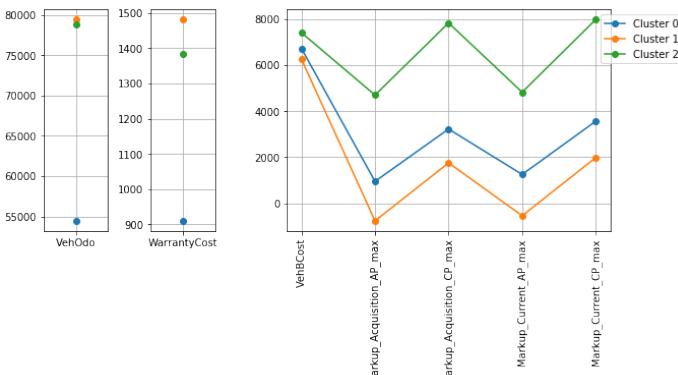


Figura 2.3: Plot dei valori assunti dai centroidi

Dalla tabella e dalla figura mostrate, si può notare come i clusters risultino molto più influenzati dalle variabili relative ai margini di guadagno piuttosto che da VehBCost e VehOdo. I centroidi dei Cluster 1 e 2, infatti, presentano valore simile per l'attributo VehOdo, mentre i centroidi dei Cluster 0 e 1 presentano valore simile per l'attributo VehBCost. Di seguito verranno descritti nel dettaglio i tre centroidi formatisi.

- **Centroide 0** suggerisce che il Cluster 0 è composto da veicoli con un basso valore di km percorsi e di costo assicurativo. Il prezzo medio pagato al tempo dell'acquisto per veicolo è pari a circa 6722 dollari e il margine di guadagno massimo al tempo d'acquisto risulta in media pari 951 dollari.
- **Centroide 1** suggerisce che il Cluster 1 è composto dai veicoli con il più alto valore di km percorsi e di costo assicurativo. Il prezzo medio pagato al tempo dell'acquisto è più basso rispetto a quello rilevato negli altri clusters e pari a circa 6257 dollari. Per questi veicoli non risulta esserci in media un margine di guadagno al tempo d'acquisto. Questo valore, infatti, è negativo sia per i veicoli in condizioni nella media (vedi Markup_AAP) sia per i veicoli in condizioni al di sopra della media (vedi Markup_CAP). Ad ogni modo, come ci si aspettava, per i veicoli in condizioni superiori alla media si riscontra una perdita potenziale inferiore rispetto agli stessi in condizioni standard.
- **Centroide 2** suggerisce che il Cluster 2 è composto da veicoli con il più alto valore di prezzo medio pagato al tempo dell'acquisto, seppur abbiano un valore relativo

ai km percorsi abbastanza alto. Questi veicoli, essendo stati pagati ad un prezzo medio alto, si presuppongono siano in condizioni migliori rispetto agli altri. Infatti, il margine di guadagno massimo risulta il più alto, sia per i veicoli in condizioni standard che per i veicoli in condizioni superiori alla media.

Nella Tabella 2.2 è mostrata la distribuzione dei centroidi rispetto alla variabile target IsBadBuy.

Bad-Buy		
	0	1
Cluster 0	19746	3786
Cluster 1	16528	1674
Cluster 2	14904	1748

Tabella 2.2: Distribuzione della variabile target nei tre clusters

Il 16% circa dei veicoli presenti nel Cluster 0 risulta Bad-Buy. Nel Cluster 1, il 9% circa dei veicoli risulta Bad-Buy, e nel Cluster 2 circa il 10%. Nella Figura 2.4 è riportato lo scatter plot dei veicoli rispetto alle variabili VehOdo (asse x) e Markup_AAP (asse y), facendo distinzione tra Bad-Buy e Good-Buy. Si precisa che il Cluster 0 è indicato con il colore turchese, il Cluster 1 con il colore viola e il Cluster 2 con il colore giallo. Si può notare che molti veicoli classificati come Bad-Buy si distribuiscono nella parte superiore del grafico, ovvero per valori alti del margine di guadagno, ma sembrano essere outliers. Inoltre, sembra che per un maggior numero di veicoli Bad-Buy sia stato registrato un valore alto dei km percorsi.

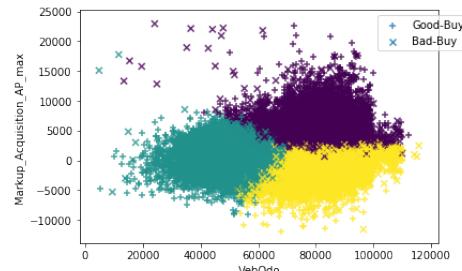


Figura 2.4: Scatter plot della variabile target nei tre clusters

2.2 Clusterizzazione con tecnica DBscan

Per l'applicazione del DBscan sono state considerate le stesse 7 variabili utilizzate per il K-means. La funzione di distanza utilizzata sarà la distanza Euclidea.

Per identificare il numero di clusters ottimale (k), è necessario definire il valore dei parametri Epsilon e MinPoints. Il valore di Epsilon dipende dal valore considerato per il parametro MinPoints. Tale valore dipende generalmente dal numero di attributi considerati per la clusterizzazione. Le regole maggiormente utilizzate sono $\text{MinPoints} = |D| + 1$ e $\text{MinPoints} = 2 * |D|$. Pertanto, si è optato per rappresentare grafi-

camente la distanza (Epsilon) per valori di MinPoints pari a 8 e 14. I risultati sono mostrati nella Figura 2.5.

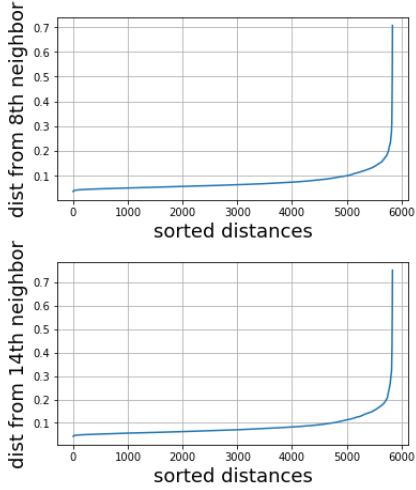


Figura 2.5: Grafico k-distance per la scelta di Epsilon e MinPoints

Per problemi legati alla complessità computazionale, è stato effettuato una clusterizzazione con tecnica K-Means, ottenendo un numero di clusters pari a 1/10 dell'intero DataSet, ovvero 5838. Successivamente, il grafico k-distance è stato ottenuto su questi punti. Il migliore valore per il parametro Epsilon si trova in prossimità del 'gomito' della funzione rappresentata. Questo perché, al di sopra della suddetta parte della curva, all'aumentare di Epsilon non aumenta ulteriormente la distanza massima tra i punti. Determinati i parametri, è stato applicata la tecnica DBscan su tutto il dataset.

Si ottengono dei buoni risultati considerando MinPoints=8 e Epsilon=0.18. E' stato individuato un unico cluster composto da 58349 veicoli e 37 noise points. Il coefficiente Silhouette calcolato risulta pari a 0.648. La Tabella 2.3 mostra la distribuzione del cluster e dei noise points rispetto alla variabile target.

	0	1	% di Bad-Buy
Cluster 0	51165	7184	14,0%
Noise	13	24	68,9%

Tabella 2.3: Distribuzione della variabile target nel cluster

Il 14% dei veicoli presenti nel Cluster 0 e il 68% dei noise points risulta Bad-Buy. Tale tecnica pertanto è risultata utile all'individuazione degli outliers. Nella Figura 2.6 si nota che effettivamente i noise points sono distribuiti nella parte superiore del grafico, ovvero per valori elevati della variabile relativa al margine di guadagno possibile, e possono considerarsi outliers.

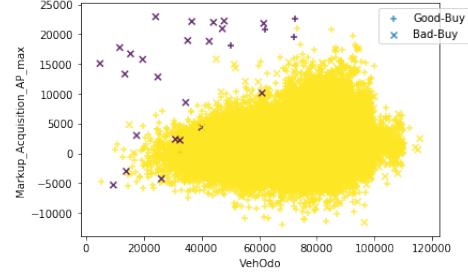


Figura 2.6: Scatter plot della variabile target nel clusters e per i noise points

2.3 Clusterizzazione con tecnica gerarchica agglomerativa

Per l'applicazione della tecnica gerarchica agglomerativa abbiamo considerato le stesse 7 variabili utilizzate per il K-means e il DBscan. La funzione di distanza utilizzata sarà la distanza Euclidea.

Per problemi legati alla complessità computazionale, è stata effettuata dapprima una clusterizzazione con tecnica K-Means, ottenendo un numero di clusters pari a 1/10 dell'intero DataSet, ovvero 5838. Su questo SubSet è sono stati considerati 4 differenti metodi di aggregazione: legame completo, legame singolo, legame medio e metodo di Ward. I risultati sono mostrati mediante l'utilizzo del dendogramma nella Figura 2.7.

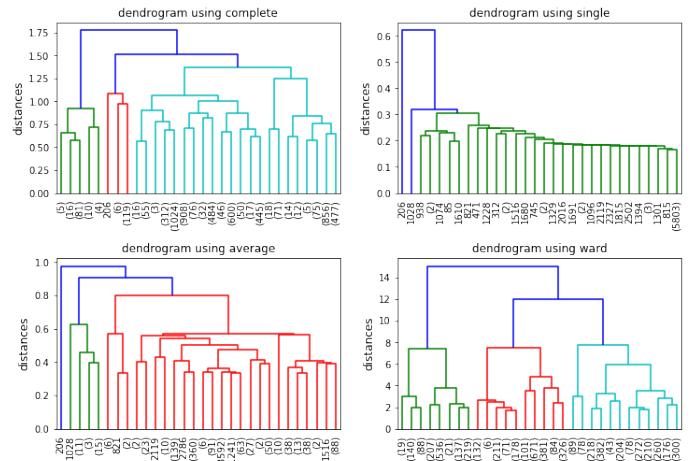


Figura 2.7: Dendrogramma per differenti metodi di aggregazione

Per ognuno dei metodi considerati, l'altezza a cui è stato scelto di effettuare il taglio si è basata principalmente sulla possibilità di ottenere 3 clusters. La Tabella 2.4 mostra i risultati ottenuti. A prima vista, si può notare come, utilizzando il metodo del legame singolo, la distanza di merging è molto bassa fino alla penultima iterazione.

	Altezza taglio	Composizione
Complete	1.40	[275, 903, 57208]
Single	0.31	[58384, 1, 1]
Group Average	0.85	[33, 58352, 1]
Ward's Method	8	[11457, 23803, 23126]

Tabella 2.4: Composizione dei clusters ottenuti

La tabella mostra che, utilizzando il metodo del legame singolo, tutti i veicoli tranne due risultano appartenenti ad un unico cluster. Questo può essere spiegato dal fatto che il legame singolo unisce i punti basandosi sulla minima distanza tra tutte le coppie. Il metodo del legame medio permette di ottenere clusters fortemente sbilanciati. I clusters 1 e 3 infatti hanno dimensione molto piccola rispetto al cluster 2. Utilizzando il metodo del legame completo e il metodo di Ward, invece, i clusters risultano più bilanciati.

Il Silhouette score maggiore si ha utilizzando il metodo del legame singolo, pari a 0.675. La Tabella 2.5 mostra la distribuzione dei clusters ottenuti con questo metodo rispetto alla variabile target.

Bad-Buy		
	0	1
Cluster 1	51178	7206
Cluster 2	0	1
Cluster 3	0	1

Tabella 2.5: Distribuzione della variabile target nei clusters ottenuti con il metodo del legame singolo

Il Cluster 1 è composto al 12.3% circa da veicoli Bad-Buy, il Cluster 2 e il Cluster 3 contengono un solo veicolo ciascuno che risulta Bad-Buy. La Figura 2.8 mostra la distribuzione dei veicoli con riferimento ai clusters ottenuti e alla variabile target IsBadBuy. I risultati risultano molto simili a quelli ottenuti mediante la tecnica di clusterizzazione DBscan.

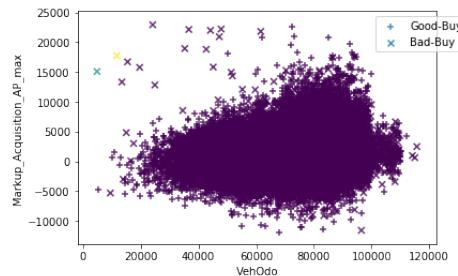


Figura 2.8: Scatter plot della variabile target nei tre clusters ottenuti con il metodo del legame singolo

Nonostante il Silhouette score sia molto elevato utilizzando questo metodo, la scelta si orienta verso una clusterizzazione più bilanciata seppur con un Silhouette score inferiore. Per questo motivo, è stato considerare il metodo di Ward,

in quanto i clusters risultano bilanciati e il Silhouette score il più alto tra le clusterizzazioni bilanciate ottenute. La Tabella 2.6 mostra la distribuzione della variabile target nei tre clusters identificati.

Bad-Buy		
	0	1
Cluster 1	10243	1214
Cluster 2	21619	2184
Cluster 3	19316	3810

Tabella 2.6: Distribuzione della variabile target nei clusters ottenuti con il metodo di Ward

Nel Cluster 1, il 10,6% dei veicoli risulta Bad-Buy, mentre nel Cluster 2 circa il 9%. La maggiore presenza di veicoli Bad-Buy si ha nel Cluster 3, in cui questi risultano il 16,5%. La Figura 2.9 mostra dei risultati molto simili a quelli ottenuti mediante la tecnica K-Means.

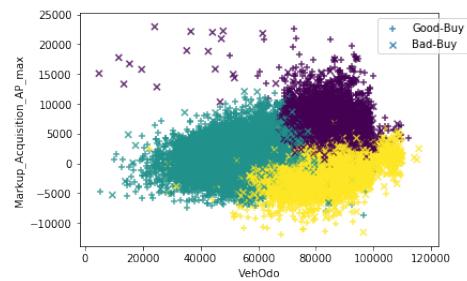


Figura 2.9: Scatter plot della variabile target nei tre clusters ottenuti con il metodo di Ward

2.4 Valutazione finale del migliore algoritmo di clusterizzazione

La Tabella 2.7 mostra il Silhouette score per le tecniche considerate. La tecnica K-Means crea 3 clusters di dimensioni bilanciate ma con un basso Silhouette score, così come i metodi del legame completo e di Ward. Le altre tecniche considerate creano clusters fortemente sbilanciati con un Silhouette score elevato. Nonostante il Silhouette score faccia pensare che il metodo migliore sia quello del legame singolo, riteniamo che clusters fortemente sbilanciati come nel nostro caso non siano un buon risultato. La clusterizzazione per cui si ottiene un unico cluster o più di un cluster contenente un solo veicolo non è efficiente. Per questo motivo, la nostra scelta si è orientata verso la tecnica K-Means e il metodo di Ward.

	N. di clusters	Silhouette score
K-Means	3	0.28
DBscan	1	0.65
Legame completo	3	0.23
Legame singolo	3	0.67
Legame medio	3	0.66
Metodo di Ward	3	0.23

Tabella 2.7: Silhouette score per le differenti tecniche considerate

Il metodo di Ward permette ottenere dei buoni risultati ma dei cluster meno definiti rispetto alla tecnica K-Means. Il Silhouette score per le due tecniche considerate conferma la nostra ipotesi.

In conclusione, la tecnica DBscan, così come il metodo del legame singolo e medio, siano da ritenere delle buone tecniche per l'individuazione degli outliers. La migliore tecnica di clusterizzazione è il K-Means.

3 Association Rules

In questa sezione verranno estratti i pattern frequenti e le regole associative. A tale scopo, il DataSet è stato trasformato in un DataSet di tipo transazionale e gli attributi quantitativi sono stati discretizzati sulla base dei centroidi ottenuti precedentemente mediante l'utilizzo della tecnica di clusterizzazione K-Means. Il risultato di questa discretizzazione è sintetizzato nella Tabella 3.1

Attributo	Intervallo		
	1	2	3
VehBCost	[-inf, 6490.13)	[6490.13, 7062.09)	[7062.09, inf)
VehOdo	[-inf, 66636.64)	[66636.64, 79118.82)	[79118.82, inf)
WarrantyCost	[-inf, 1147.305)	[1147.305, 1433.14)	[1433.14, inf)
Markup_Acquisition_AP_max	[-inf, 89.88)	[89.88, 2817.57)	[2817.57, inf)
Markup_Acquisition_CP_max	[-inf, 2484.005)	[2484.005, 5518.16)	[5518.16, inf)
Markup_Current_AP_max	[-inf, 353.135)	[353.135, 353.135)	[353.135, inf)
Markup_Current_CP_max	[-inf, 2765.33)	[2765.33, 5759.77)	[5759.77, inf)

Tabella 3.1: Intervalli definiti per le variabili quantitative

Si è cercato di estrarre pattern frequenti con diversi valori di support e diverse tipologie (frequenti, chiusi e massimali).

3.1 Estrazione dei pattern frequenti e analisi del numero di pattern rispetto al parametro MinSup

3.1.1 Itemsets frequenti

Un itemset è frequente se il suo support è maggiore o al più uguale alla soglia fissata come MinimumSupport (minsup).

- **minsup=0.10** produce 9702 itemsets, di cui 2011 con support inferiore a 0.11. Risultano, inoltre, 358 itemsets di dimensione pari a 2 e support < 0.15. Il 65% degli

itemsets ha dimensione ≤ 5 e support < 0.20. Il support massimo si ha per l'itemsets composto da Transmission = 1, IsOnlineSale = 0, pari a 0.94. Il support minimo si ha per l'itemset composto da VehBCost = [-inf, 6490.13), WheelTypeID = 1.0, Auction = 'MANHEIM', TopThreeAmericanName = 1, IsOnlineSale = 0, maggiore comunque di 0.10.

- **minsup=0.20** produce 1429 itemsets, di cui 165 con support inferiore a 0.21. Sono presenti 146 itemsets di dimensione pari a 2 e support < 0.25. Il 75.6% degli itemsets ha dimensione ≤ 5 e support < 0.30. Il support minimo si ha per l'itemset composto da Markup_Current_CP_max = [5759.77, inf), Markup_Current_AP_max = [3030.635, inf), TopThreeAmericanName = 1, IsBadBuy = 0, Transmission = 1, pari a 0.20.
- **minsup=0.30** produce 305 itemsets, di cui 39 con support inferiore a 0.31. Risultano, inoltre, 42 itemsets di dimensione pari a 2 e support < 0.35. Il 68.9% degli itemsets ha dimensione ≤ 5 e support < 0.40. Il support minimo si ha per l'itemset composto da Markup_Acquisition_CP_max = [2484.005, 5518.16), TopThreeAmericanName = 1, Transmission = 1.
- **minsup=0.40** produce 92 itemsets, di cui 8 con support inferiore a 0.41. Sono presenti 20 itemsets di dimensione pari a 2 con support < 0.45. Inoltre, il 67.4% degli itemsets ha dimensione ≤ 5 e support < 0.50. Il support minimo si ha per l'itemset composto da Auction = 'MANHEIM', Nationality = 'AMERICAN', IsBadBuy = 0, Transmission = 1.

Support	Itemsets
0.9406	{Transmission = 1, IsOnlineSale = 0}
0.8538	{IsBadBuy = 0, IsOnlineSale = 0}
0.8456	{IsBadBuy = 0, Transmission = 1}
0.8359	{TopThreeAmericanName = 1, Nationality = 'AMERICAN'}
0.8240	{IsBadBuy = 0, Transmission=1, IsOnlineSale = 0}
0.8146	{TopThreeAmericanName = 1, IsOnlineSale = 0, Nationality = 'AMERICAN'}
0.8146	{TopThreeAmericanName = 1, IsOnlineSale = 0}
0.8146	{Nationality = 'AMERICAN', IsOnlineSale = 0}
0.8140	{TopThreeAmericanName = 1, Transmission = 1, Nationality = 'AMERICAN'}
0.8140	{TopThreeAmericanName = 1, Transmission = 1}

Tabella 3.2: Primi 10 Frequent Itemsets

Al diminuire del valore soglia per il minsup, il numero di pattern frequenti aumenta. I 10 pattern con support maggiore sono gli stessi a prescindere dalle 4 soglie fissate per il minsup, in quanto il decimo itemsets ha support pari a 0.8140. Inoltre, è stato riscontrato che i veicoli più frequenti sono quelli di nazionalità americana, cambio automatico, considerati Good-Buy e non acquistati originariamente online.

3.1.2 Itemsets chiusi

Un itemset è definito chiuso se nessuno dei suoi adiacenti supersets ha lo stesso suo valore di support.

- **minsup=0.10** produce 5434 itemsets, di cui 1180 con support inferiore a 0.11. Risultano, inoltre, 328 itemsets di dimensione pari a 2 e support < 0.15 . Il 63.6% degli itemsets ha dimensione ≤ 5 e support < 0.20 . Il support massimo e minimo si hanno per gli itemsets frequenti trovati precedentemente.
- **minsup=0.20** produce 812 itemsets, di cui 100 con support inferiore a 0.21. Sono presenti 93 itemsets di dimensione pari a 2 e support < 0.25 . Il 72.8% degli itemsets ha dimensione ≤ 5 e support < 0.30 . Il support minimo si ha per l'itemset composto da Markup_Current_CP_max = [5759.77, inf), Markup_Current_AP_max = [3030.635, inf), Nationality = 'AMERICAN', TopThreeAmericanName = 1, IsBadBuy = 0, Transmission = 1.
- **minsup=0.30** produce 179 itemsets, di cui 23 con support inferiore a 0.31. Risultano, inoltre, 26 itemsets di dimensione pari a 2 e support < 0.35 . Il 67.0% degli itemsets ha dimensione ≤ 5 e support < 0.40 . Il support minimo si ha per l'itemset composto da Markup_Acquisition_CP_max = [2484.005, 5518.16), Nationality = 'AMERICAN', TopThreeAmericanName = 1, Transmission = 1.
- **minsup=0.40** produce 56 itemsets, di cui 4 con support inferiore a 0.41. Sono presenti 12 itemsets di dimensione pari a 2 con support < 0.45 . Inoltre, il 71.4% degli itemsets ha dimensione ≤ 5 e support < 0.50 . Il support minimo si ha per l'itemset composto da Auction = 'MANHEIM', TopThreeAmericanName = 1, Nationality = 'AMERICAN', IsBadBuy = 0, Transmission = 1.

Dalla Tabella 3.3 si evince che i primi 6 itemsets frequenti sono anche chiusi.

Support	Itemsets
0.9406	{Transmission = 1, IsOnlineSale = 0}
0.8538	{IsBadBuy = 0, IsOnlineSale = 0}
0.8456	{IsBadBuy = 0, Transmission = 1}
0.8359	{TopThreeAmericanName = 1, Nationality = 'AMERICAN'}
0.8240	{IsBadBuy = 0, Transmission=1, IsOnlineSale = 0}
0.8146	{TopThreeAmericanName = 1, IsOnlineSale = 0, Nationality = 'AMERICAN'}
0.8140	{TopThreeAmericanName = 1, Transmission = 1, Nationality = 'AMERICAN'}
0.7935	{TopThreeAmericanName = 1, Nationality = 'AMERICAN', Transmission = 1, IsOnlineSale = 0}
0.7345	{TopThreeAmericanName = 1, Nationality = 'AMERICAN', IsBadBuy = 0}
0.7155	{TopThreeAmericanName = 1, Nationality = 'AMERICAN', IsBadBuy = 0, IsOnlineSale = 0}

Tabella 3.3: Primi 10 Closed Itemsets

Al diminuire del valore soglia per il minsup, il numero di pattern frequenti aumenta. I 10 pattern con support maggiore sono gli stessi a prescindere dalle 4 soglie fissate

per il minsup, in quanto il decimo itemsets ha support pari a 0.7155. Anche in questo caso, è stato riscontrato che i veicoli più frequenti sono quelli di nazionalità americana, cambio automatico, considerati Good-Buy e non acquistati originariamente online.

3.1.3 Itemsets massimali

Un itemset frequente è massimale se nessuno dei suoi adiacenti supersets è frequente.

- **minsup=0.10** produce 584 itemsets, di cui 489 con support inferiore a 0.11. Risultano, inoltre, 44 itemsets di dimensione ≤ 3 e support < 0.15 . Il 53.6% degli itemsets ha dimensione ≤ 5 e support < 0.20 . Il support massimo si ha per l'itemset composto da TopThreeAmericanName = 1, Nationality = 'AMERICAN', IsBadBuy = 0, Transmission = 1, IsOnlineSale = 0 ed è pari a 0.697. Il support minimo si ha per l'itemset composto da VehBCost = [-inf, 6490.13), WheelTypeID = 1.0, Auction = 'MANHEIM', Nationality = 'AMERICAN', TopThreeAmericanName' = 1, IsOnlineSale = 0, pari a 0.10.
- **minsup=0.20** produce 106 itemsets, di cui 67 con support inferiore a 0.21. Sono presenti 16 itemsets di dimensione ≤ 3 e support < 0.25 . Il 62.3% degli itemsets ha dimensione ≤ 5 e support < 0.30 . Il support minimo si ha per lo stesso itemset chiuso trovato precedentemente con il minsup considerato.
- **minsup=0.30** produce 30 itemsets, di cui 17 con support inferiore a 0.31. Risultano, inoltre, 13 itemsets di dimensione ≤ 3 e support < 0.35 . Il 70.0% degli itemsets ha dimensione ≤ 5 e support < 0.40 . Il support minimo si ha per lo stesso itemset chiuso trovato precedentemente con il minsup considerato.
- **minsup=0.40** produce 16 itemsets, di cui 4 con support inferiore a 0.41. Sono presenti 9 itemsets di dimensione ≤ 3 con support < 0.45 . Inoltre, il 93.8% degli itemsets ha dimensione ≤ 5 e support < 0.50 . Il support minimo si ha per lo stesso itemset chiuso trovato precedentemente con il minsup considerato.

Dalla Tabella 3.4 si evince che nessuno dei primi 10 itemsets chiusi è massimale.

Support	Itemsets
0.6969	{TopThreeAmericanName = 1, Nationality = 'AMERICAN', IsBadBuy = 0, Transmission = 1, IsOnlineSale = 0}
0.4671	{Auction = 'MANHEIM', IsBadBuy = 0, Transmission=1, IsOnlineSale = 0}
0.4451	{WheelTypeID = 2.0, Transmission = 1, IsOnlineSale = 0}
0.4429	{Auction = 'MANHEIM', TopThreeAmericanName = 1, Nationality = 'AMERICAN', Transmission=1, IsOnlineSale = 0}
0.4362	{WarrantyCost = [inf, 1147.305), Transmission=1, IsOnlineSale = 0}
0.4211	{VehBCost = [-inf, 6490.13), Transmission=1, IsOnlineSale = 0}
0.4182	{WheelTypeID = 1.0, IsBadBuy = 0, Transmission=1, IsOnlineSale = 0}
0.4181	{WheelTypeID = 1.0, TopThreeAmericanName = 1, Nationality = 'AMERICAN', Transmission = 1, IsOnlineSale = 0}
0.4175	{VehBCost = [7062.09, inf), Transmission=1, IsOnlineSale = 0}
0.4133	{WheelTypeID = 2.0, IsBadBuy = 0, IsOnlineSale = 0}

Tabella 3.4: Primi 10 Maximal Itemsets con minsup=0.40

Al diminuire del valore soglia per il minsup, il numero di pattern frequenti aumenta. In questo caso, i 10 pattern con support maggiore sono diversi dai 10 pattern identificati come chiusi. L'itemset massimale con support maggiore presenta support inferiore al decimo itemset chiuso. Anche in questo caso, è stato riscontrato che i veicoli più frequenti sono quelli di nazionalità americana, cambio automatico, considerati Good-Buy e non acquistati originariamente online. A seguire, troviamo i veicoli che sono stati forniti da Manheim e i veicoli con cerchioni non in lega.

3.2 Estrazione delle regole associative per differenti valori di MinConf

Abbiamo fissato inizialmente il valore di minsup a 0.10.

- **minconf=0.10** produce 380320 regole, di cui il 39.1% con Lift < 1. Inoltre, 9699 regole implicano IsBadBuy.
- **minconf=0.20** produce 228165 regole, di cui il 32.6% con Lift < 1. Il 2.5% delle regole implica la variabile target.
- **minconf=0.30** produce 150412 regole, di cui il 30.5% con Lift < 1. Le regole che implicano la variabile IsBadBuy sono 5638. Da questa soglia di minconf a minconf = 0.70, il numero di regole che implicano la variabile target rimane invariato.
- **minconf=0.40** produce 93508 regole, di cui il 22.1% con Lift < 1.
- **minconf=0.50** produce 66836 regole, di cui il 17.6% con Lift < 1.
- **minconf=0.60** produce 49316 regole, di cui il 19.0% con Lift < 1.
- **minconf=0.70** produce 40280 regole, di cui il 22.4% con Lift < 1.
- **minconf=0.80** produce 34487 regole, di cui il 22.0% con Lift < 1. Il 16.3% delle regole implica IsBadBuy.
- **minconf=0.90** produce 20185 regole, di cui il 17.3% con Lift < 1. Per tale soglia di minconf, il numero di regole che implicano la variabile target si dimezza rispetto a quello della soglia di minconf = 0.80.
- **minconf=1.00** produce 4359 regole, di cui nessuno con Lift < 1. Per questo valore di minconf, nessuna regola implica IsBadBuy.

La Tabella 3.5 mostra le regole associative più interessanti. Nella tabella sono state riportate le 6 regole con valore migliore per il parametro confidence. Considerando tale parametro, le migliori regole risultano essere quelle che implicano gli attributi TopThreeAmericanName e Nationality.

Consequent	Antecedent	Support	Confidence
{TopThreeAmericanName = 1}	{ Nationality = 'AMERICAN'}	0.85	1.00
{ Nationality = 'AMERICAN'}	{TopThreeAmericanName = 1, IsBadBuy = 0, Transmission= 1}	0.72	1.00
{IsOnlineSale = 0}	{Markup_Acquisition_CP_max=(-inf, 2484.005), Nationality='AMERICAN', IsBadBuy=0, Transmission=1}	0.22	0.99
{Transmission=1}	{Auction = 'MANHEIM'}	0.54	0.96
{ IsBadBuy = 0}	{WarrantyCost = [-inf, 1147.305), Transmission=1, IsOnlineSale = 0}		
{Markup_Current_CP_max=(-inf, 2765.33)}	{Markup_Acquisition_CP_max=(-inf, 2484.005), VehBCost= (-inf, 6490.13), IsOnlineSale = 0}	0.17	0.98

Tabella 3.5: Regole associative più interessanti

La Figura 3.1 mostra gli istogrammi del parametro di confidence e dell'indice Lift per il valore di support pari a 0.40, ovvero il valore per cui si riesce a predire la variabile target con la maggiore accuratezza. Come ci aspettavamo, il numero di regole associative trovate diminuisce all'aumentare del valore attribuito al minconf. Notiamo che il maggior numero di regole si ha per il valore di minconf fissato a 0.10. Inoltre, il 50% delle regole presenta un valore di confidence < 0.29. Per quanto concerne il Lift, abbiamo dapprima verificato l'ipotesi di normalità della distribuzione mediante il test di Anderson-Darling per un livello di significatività del 5%: si rifiuta tale ipotesi in quanto il test è risultato pari a 159.05 e il valore critico uguale a 0.786. La distribuzione del Lift risulta leggermente asimmetrica positivamente, con media pari a 1.034 e mediana pari a 1.009.

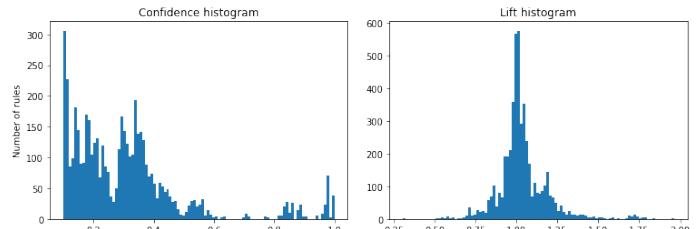


Figura 3.1: Istogramma della confidence e del Lift delle regole

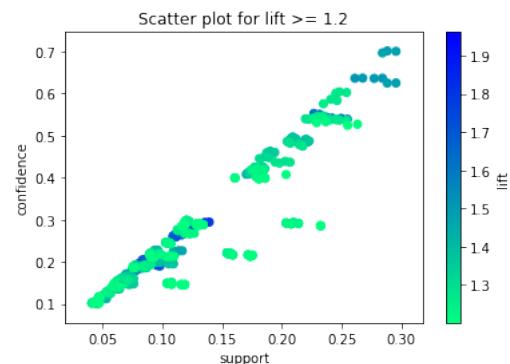


Figura 3.2: Scatter plot del Lift per valori ≥ 1.2

La Figura 3.2 mostra, infine, lo scatter plot dell'indice Lift per valori ≥ 1.2 . Si evince che si ottengono valori di Lift ≥ 1.2 per valori di support al massimo pari a 0.30 e valori di confidence al più pari a 0.70.

3.3 Regole associative utili per il trattamento dei valori mancanti

Il DataSet, come mostra la Tabella 2, contiene dei valori mancanti. Per la loro sostituzione sono state considerate le regole associative più interessanti che presentavano l'attributo di interesse nella consequence. Tali regole sono mostrate nella Tabella 3.6. L'accuracy con cui si predicono Transmission=1 e Nationality='AMERICAN' è risultata molto elevata.

Antecedent	Consequent	Support	Accuracy	Lift
{VehBCost=[7062.09,inf), IsOnlineSale=0}	{WheelTypeID=1.0}	0.26	0.61	1.19
{Auction='MANHEIM'}	{Transmission=1}	0.45	0.97	1.01
{VehBCost=[-inf,6490.13), Transmission=1, IsOnlineSale=0}	{Markup_Acquisition_AP_max=[-inf, 89.88]}	0.21	0.49	1.34
{VehBCost=[7062.09,inf)}	{Markup_Current_AP_max=[3030.6 35, inf]}	0.18	0.42	1.37
{WheelTypeID=1.0, Transmission=1, IsOnlineSale=0}	{Nationality='AMERICAN'}	0.42	1.0	1.20
{WarrantyCost=[-inf, 1147.305), IsBadBuy=0}	{Size='MEDIUM'}	0.27	0.64	1.51

Tabella 3.6: Regole associative più interessanti per la sostituzione dei missing values

3.4 Regole associative utili per predire la variabile target

Dal momento che la variabile target è IsBadBuy, la scelta del valore da attribuire al minsup e al MinimumConfidence (minconf) è stata influenzata dalla presenza di tale variabile nella consequent delle regole associative trovate. Le regole che implicano IsBadBuy=0 sono mostrate nella Tabella 3.7. Per ottenere queste regole sono stati utilizzati valori come minsup=0.4 e accuracy=0.85.

Antecedent	Consequent	Support	Accuracy
{Transmission=1, IsOnlineSale=0}	{IsBadBuy=0}	0.82	0.88
{Nationality='AMERICAN', TopThreeAmericanName=1}	{IsBadBuy=0}	0.73	0.88
{Auction='MANHEIM'}	{IsBadBuy=0}	0.49	0.89
{WheelTypeID=1.0}	{IsBadBuy=0}	0.44	0.86
{WarrantyCost=[-inf, 1147.305)}	{IsBadBuy=0}	0.42	0.88

Tabella 3.7: Regole più interessanti che implicano IsBadBuy=0

Per ottenere regole associative che implicano IsBadBuy=1, invece, è stato necessario diminuire il valore di minsup al 2% per avere accuracy maggiore possibile. Nella Ta-

bella 3.8 sono mostrate le regole per cui si è ottenuta accuracy migliore. Purtroppo l'accuracy migliore è il 20% per questa categoria di veicoli.

Antecedent	Consequent	Support	Accuracy
{VehOdo=[79118.82, inf), Markup_Current_CP_max=[-inf, 2765)}	{IsBadBuy=1}	0.022	0.20
{WarrantyCost=[1433.14, inf), VehBCost=[-inf, 6490.13)}	{IsBadBuy=1}	0.027	0.20
{Markup_Current_AP_max=[-inf, 353.135)}	{IsBadBuy=1}	0.057	0.15

Tabella 3.8: Regole più interessanti che implicano IsBadBuy=1

I risultati migliori sono stati ottenuti per IsBadBuy=0.

4 Classification

In questa sezione verranno applicati due differenti algoritmi di classificazione (Decision Tree e Random Forest) utilizzando parametri differenti al fine di massimizzare le performance del modello. La costruzione di questi modelli supervisionati ci permetterà di individuare quale veicolo può essere considerato Bad-Buy o meno.

4.1 Learning

Per entrambi i modelli non sono state prese in considerazione variabili ridondanti o considerate irrilevanti. In particolare sono state considerate le variabili categoriali 'Make', 'Model', 'Trim', 'SubModel', 'Color', 'WheelType', 'Nationality', 'Size', 'TopThreeAmericanName', 'AUGUART', 'VNST', 'IsOnlineSale' e le variabili numeriche 'PurchYear', 'VehYear', 'VehicleAge', 'VehOdo', 'VehBCost', 'WarrantyCost', 'Markup_Current_AP_max', 'Markup_Current_CP_max', 'Markup_Acquisition_AP_max' e 'Markup_Acquisition_CP_max'.

Prima di eseguire il learning, dal DataSet sono stati filtrati i 37 veicoli risultati come rumore nel DBScan.

Si ricorda, inoltre, che nella fase di comprensione dei dati la variabile target IsBadBuy risulta fortemente sbilanciata. Dal momento che c'è un forte sbilanciamento verso il valore 0, questo potrebbe condurre ad un modello in grado di prevedere, per qualsiasi veicolo, sempre tale valore (cioè non è un Bad-Buy). Pertanto, sono stati applicati tre diversi approcci per cercare di ovviare a questo problema:

- Utilizzare la seguente funzione per bilanciare il peso delle classi durante la costruzione dell'albero:

$$\frac{n_samples}{(n_classes * np.bincount(y))}$$

e considerato come accuracy la media della recall, in modo da avere anche un accuracy bilanciata;

- Ricerca del modello con undersampling dei dati, scegliendo casualmente un numero di veicoli non Bad-Buy

in modo da avere un training set bilanciato 50:50 di veicoli Bad-Buy e non Bad-Buy;

3. Ricerca del modello con oversampling dei dati, replicando casualmente i veicoli Bad-Buy per avere anche in questo caso un training set bilanciato 50:50.

I risultati migliori sono stati ottenuti utilizzando il primo approccio, probabilmente perché l'algoritmo prende in considerazione l'intero training set. Il secondo restituisce risultati molto simili al primo ma con un accuracy leggermente inferiore. L'ultimo, invece, ha restituito i risultati peggiori, probabilmente perché l'aumento del numero di sample richiede una diversa ricerca dei parametri.

Chiaramente, bilanciare il training set con il secondo approccio riduce di poco i tempi di learning, mentre il terzo li fa aumentare di molto. Per evitare di introdurre altri fattori casuali durante il training, si è utilizzato il primo approccio, avendo così una migliore accuracy (bilanciata), seppur aumentando leggermente i tempi di learning.

Per la ricerca del modello migliore è stata utilizzata una randomized grid search con cross validation. Nello specifico è stata usata la classe `RandomizedSearchCV` della libreria sklearn. In generale, la randomized grid search produce risultati molto simili ad una grid search completa.

La ricerca del modello è stata eseguita sui seguenti parametri:

- **criterion:** Gini ed Entropy;
- **max_depth:** infinita e i valori compresi fra 2 e 50;
- **min_sample_split:** i valori sono compresi tra 2 e 300;
- **min_sample_leaf:** i valori sono compresi tra 1 e 200.

Per la classe `RandomizedSearchCV` sono stati utilizzati i valori di default per i parametri `n_iter` e `cv`, che sono rispettivamente 100 e 5. Non è stato necessario dividere manualmente il dataset di training in training set e validation set, in quanto questa divisione è effettuata dalla classe `RandomizedSearchCV` durante la scelta del modello.

Si ricorda che nei risultati seguenti, il mean validation score è calcolato usando come funzione di scoring l'accuracy bilanciata.

4.2 Decision Tree

Con i parametri sopra elencati, la ricerca del modello per il miglior Decision Tree ha restituito i seguenti risultati:

- **Modello con rank 1:**

- **mean validation score:** 0.619 (std: 0.005);
- **criterion:** Gini;
- **max_depth:** 6;
- **min_sample_split:** 5;

- **min_sample_leaf:** 100.

- **Modello con rank 2:**

- **mean validation score:** 0.619 (std: 0.006);
- **criterion:** Gini;
- **max_depth:** 6;
- **min_sample_split:** 200;
- **min_sample_leaf:** 10.

- **Modello con rank 3:**

- **mean validation score:** 0.618 (std: 0.006);
- **criterion:** Entropy;
- **max_depth:** 8;
- **min_sample_split:** 2;
- **min_sample_leaf:** 200.

Per verificare le reali performance del Decision Tree generato è necessario utilizzare un DataSet che non è stato coinvolto nelle fasi di training e validazione. A questo scopo, per lo svolgimento del progetto, è stato predisposto il DataSet di test, separato dal DataSet di training. Per poter predire i veicoli presenti nel validation set è, prima di tutto, necessario applicare al test set tutte le trasformazioni applicate al training set. Una volta fatto, è possibile utilizzare il modello per predire se i veicoli del test set sono Bad-Buy oppure no.

Dal valore del cross validation nel grid search risulta che il miglior modello per il Decision Tree è quello con i parametri `criterion=Gini`, `max_depth=6`, `min_sample_split=5` e `min_sample_leaf=100`. Una volta scelto il modello, è possibile eseguire il fit sull'intero training set e testarlo successivamente sul test set. I risultati ottenuti per il training e il test set sono riportati rispettivamente nelle Tabelle 4.1 e 4.2.

Target	Precision	Recall	F1-Score
IsNotBadBuy	0.918	0.697	0.792
BadBuy	0.205	0.555	0.299

Tabella 4.1: Performance del Decision Tree sul training set

Target	Precision	Recall	F1-Score
IsNotBadBuy	0.916	0.692	0.788
BadBuy	0.195	0.542	0.287

Tabella 4.2: Performance del Decision Tree sul test set

Per quanto riguarda il training set, il modello produce un'accuracy pari a 0.697 per IsNotBadBuy; per quanto riguarda, invece, il test set si produce comunque una buona accuracy pari a 0.692.

Dai risultati possiamo dire che il modello si adatta discretamente ai dati ed inoltre possiamo assumere di non trovarci

in una situazione di overfitting, proprio perché le performance del test set e del training set, sono molto simili. È stata inoltre individuata la Confusion Matrix prodotta dal modello in Figura 4.1 e l'area definita dalla ROC Curve pari a 0.62 in Figura 4.2.

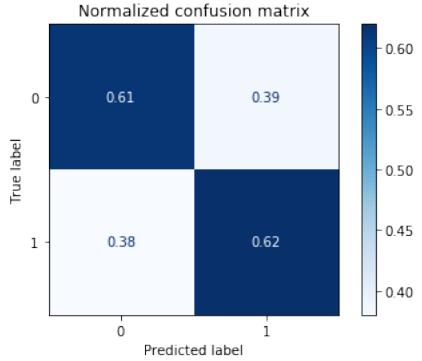


Figura 4.1: Confusion Matrix normalizzata

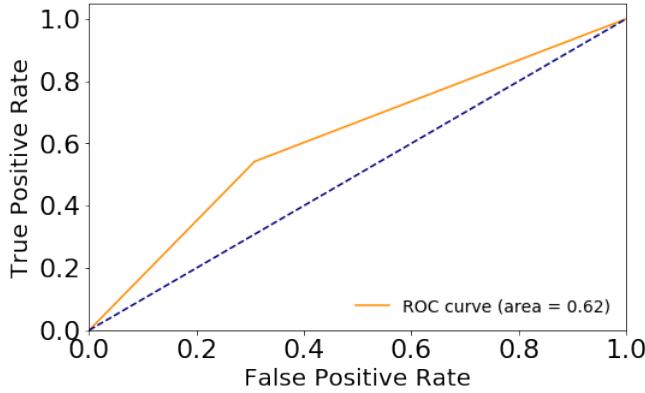


Figura 4.2: Area della ROC Curve del modello

4.2.1 Interpretazione dell'albero

Il risultato del miglior Decision Tree è mostrato nella Figura 5.1. Il primo nodo divide i veicoli in maniera sbilanciata rispetto alla variabile VehicleAge. In generale, i nodi di colore tendente al blu si riferiscono alla classificazione in Bad-Buy e i nodi tendenti al rosso alla classificazione dei veicoli in non Bad-Buy. Guardando il Decision Tree risulta che i veicoli con età ≤ 4.5 anni possono essere classificati come non Bad-Buy. In particolare, il 94.7% dei veicoli che presentano un valore di VehBCost > 6262.5 dollari ed età ≤ 2.5 anni viene classificato come non Bad-Buy. Vengono classificati come veicoli non Bad-Buy anche quelli che presentano un valore di VehBCost ≤ 6262.5 e una domanda non al di sopra della media ($PRIMEUNIT \leq 0.5$): i veicoli non Bad-Buy sono infatti il 97.7%.

I veicoli con età > 4.5 anni, il cui livello di garanzia è 'RED' (livello più basso) e VehBCost > 4657.5 dollari vengono classificati come Bad-Buy. Al contrario, veicoli con età > 4.5 anni, il cui livello di garanzia è 'GREEN' (livello più alto) e

VehBCost ≤ 4202.5 dollari sono classificati come non Bad-Buy.

Grazie al Decision Tree è stato possibile, infine, estrarre informazioni utili circa l'importanza delle variabili del Dataset. Ciò che è emerso è che le variabili VehicleAge, VehBCost e AUCGUART hanno dato un contributo rispettivamente del 50%, del 17% e del 9% alla classificazione. Tutte le altre variabili, invece, hanno contribuito al più al 2%.

4.2.2 Confronto accuracy senza il filtraggio del Dataset

La cross-validation condotta dalla grid-search, con il filtraggio dei veicoli considerati rumore secondo DBScan, ha prodotto uno score di 0.619 (come detto in precedenza). Consideriamo, però, il valore più preciso di 0.61936.

Per valutare se il filtraggio del rumore ha migliorato il processo di classificazione, è stata effettuata una grid-search anche con il dataset di training non filtrato. Il miglior modello trovato per i parametri criterion=Gini, max_depth=6, min_samples_leaf=100, min_samples_split=300 con uno score di 0.61908. Considerando, invece, gli stessi parametri del miglior modello con Dataset filtrato, è ottenuto uno score di 0.61876.

Filtrare il rumore ha portato, quindi, a trovare un modello con un accuracy più alta.

4.3 Random Forest

Con i parametri sopra elencati, la ricerca del modello per il miglior Random Forest ha restituito i seguenti risultati:

- **Modello con rank 1:**

- **mean validation score:** 0.629 (std: 0.005);
- **criterion:** Gini;
- **max_depth:** 30;
- **min_sample_split:** 200;
- **min_sample_leaf:** 30.

- **Modello con rank 2:**

- **mean validation score:** 0.627 (std: 0.007);
- **criterion:** Gini;
- **max_depth:** 30;
- **min_sample_split:** 200;
- **min_sample_leaf:** 30.

- **Modello con rank 3:**

- **mean validation score:** 0.627 (std: 0.008);
- **criterion:** Entropy;
- **max_depth:** 40;
- **min_sample_split:** 300;
- **min_sample_leaf:** 15.

Anche per il Random Forest sceglio il modello con il punteggio di cross-validation più alto, e calcoliamo le performance. Le performance sul training set e test set sono riportate rispettivamente nelle tabelle 4.3 e 4.4.

Target	Precision	Recall	F1-Score
IsNotBadBuy	0.941	0.724	0.818
BadBuy	0.256	0.677	0.371

Tabella 4.3: Performance del Random Forest sul training set

Target	Precision	Recall	F1-Score
IsNotBadBuy	0.920	0.701	0.796
BadBuy	0.204	0.559	0.300

Tabella 4.4: Performance del Random Forest sul test set

Di nuovo valgono tutte le considerazione fatte per il Decision Tree, riportiamo Confusion Matrix in Figura 4.3; l'area definita dalla ROC Curve è pari a 0.63.

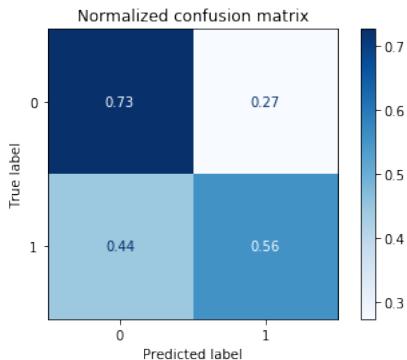


Figura 4.3: Confusion Matrix normalizzata del Random Forest

Stando al risultato ottenuto dal cross-validation del grid search, per questo Dataset il Random Forest risulta essere un classificatore migliore rispetto al Decision Tree. Il miglior modello trovato per il classificatore Random Forest presenta un cross-validation-score di 0.629, contro uno score di 0.619 ottenuto dal miglior modello trovato per il classificatore Decision Tree.

5 Conclusioni

Il Dataset proposto risulta fortemente sbilanciato per quanto riguarda la variabile target Is BadBuy e con una maggiore presenza di variabili categoriali. Tali variabili sono risultate molto utili per l'estrazione delle regole associative e per la classificazione. Per le tecniche di clustering sono state utilizzate, invece, esclusivamente variabili quantitative.

Mediane le tecniche di clustering utilizzate è emerso che la maggior parte di queste conduceva a clusters fortemente sbilanciati e, per tale motivo, ritenute non appropriate allo scopo del nostro progetto. I clusters ottenuti nella Sezione 2 mediante le tecniche K-Means e metodo di Ward hanno condotto a risultati simili, evidenziando delle caratteristiche comuni tra i veicoli sia per i veicoli Bad-Buy che non. La tecnica DBscan è risultata, invece, utile nell'individuazione degli outliers e proprio per questo è stata di supporto per la costruzione del Decision Tree.

Le regole associative estratte nella Sezione 3 hanno permesso di ottenere dei buoni risultati per predire i veicoli non Bad-Buy, sia in termini di support che in termini di confidence. Lo stesso non si può dire, invece, per i veicoli Bad-Buy.

Per quanto riguarda la classificazione della Sezione 4, infine, risulta particolarmente importante sottolineare l'importanza di bilanciare il peso delle classi della variabile IsBadBuy in quanto, altrimenti, la grid-search porta a un modello che classifica sempre IsBadBuy=0.

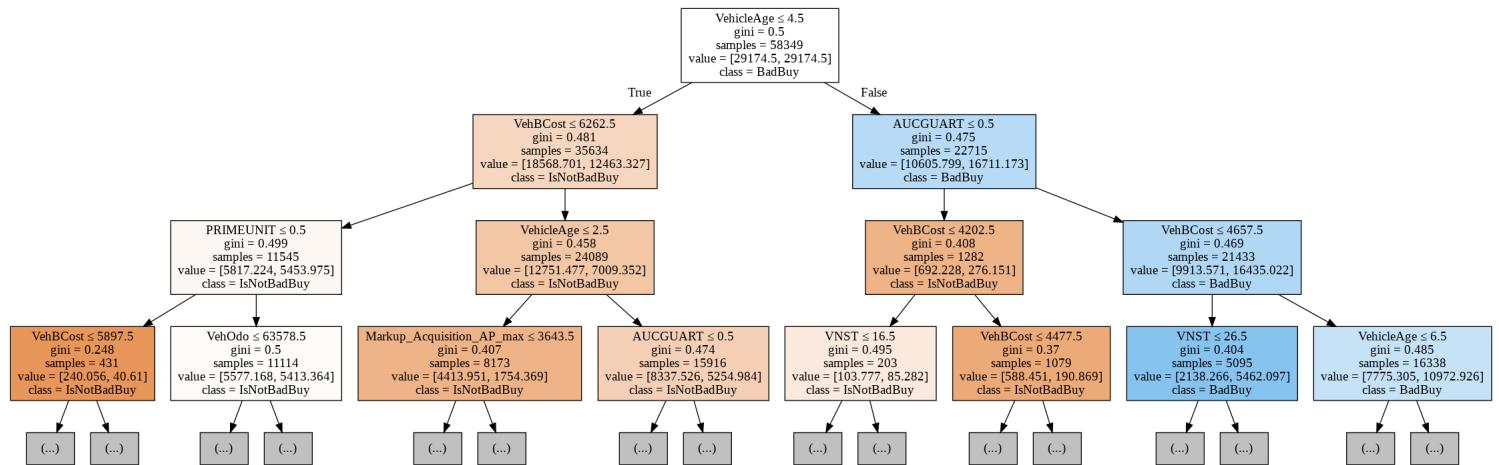


Figura 5.1: Decision Tree di Rank 1