

**Università di Pisa**

**Corsi di laurea in Computer Science e Business Informatics**

**Anno Accademico 2017/2018**

**Corso di Data Mining**

**Studenti:**

* Antonio Sisbarra (518552)
* Francesco Spinnato ()

****

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

# 1. Introduzione

Il presente elaborato si pone l’obiettivo di illustrare il modello di classificazione progettato per prevenire il rischio di acquisti errati alle aste di auto usate. Grazie al modello, infatti, i rivenditori d’auto potranno diminuire il rischio di comprare auto problematiche e invendibili al cliente finale.

I dati sono relativi alle vetture prodotte tra il 2001 e il 2010 e sono forniti dalla start-up americana Carvana. Il dataset fornisce una serie di caratteristiche di varia natura relative ai veicoli, come per esempio: l’assetto, la marca, il produttore, il tipo di cambio, il chilometraggio, la locazione geografica dell’asta e l’andamento nel mercato del settore automotive.

Sull’argomento non vi erano particolari conoscenze pregresse e sono state svolte alcune ricerche sull’andamento del mercato delle automobili negli Stati Uniti tra il 2001 al 2010. Si è scoperto, per esempio, che in quegli anni il mercato in USA subiva una forte crisi, specialmente delle grandi case GM, Chrysler e Ford[[1]](#footnote-1). Inoltre, per la valutazione delle auto usate vengono spesso presi in considerazione il chilometraggio, la marca, il colore, l’assetto e l’età.

Il progetto si dirama in quattro sezioni:

1. **Data Understanding**: analisi del training set, data quality, data semantic e preprocessing.
2. **Operazioni di clustering**: che hanno evidenziato ridondanze e particolari relazioni tra alcuni attributi del dataset, come ad esempio i valori di MMR, ossia il prezzo di acquisto dell’automobile e della garanzia.
3. **Scoperta di pattern frequenti e regole di associazione**
4. **Classificazione**: costruzione del modello di classificazione più adatto allo scopo dell’elaborato. Verranno presentati diversi modelli e la scelta di quello definitivo verrà accompagnata da una riflessione.

# 2. Data Understanding

## 2.1 Analisi qualitativa e quantitativa dei dati

Il training set presenta le seguenti caratteristiche:

N° di **record**: **72983** e N° di **attributi**: **33**. Tra gli attributi del training set se ne contano **18** di tipo numerico e **14** di tipo nominale. Nella tabella (0) sono riportati tutti gli attributi categorizzati per tipologia.

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

Dall’**analisi qualitativa** sono emersi alcuni errori nei valori degli attributi:

* il valore “Toyota Scion” dell’attributo Make[[2]](#footnote-2). Non esiste sul mercato una casa produttrice con entrambi i nomi e difatti nel dataset ricorre una sola volta, confermando la sua natura di errore semantico. 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 l’attributo Model sono stati identificati e corretti 26 errori di incompletezza per i modelli delle auto che terminavano con uno “/”, es.: “YARIS 1.5 L L4 EFI/”; inoltre, sono stati notati 31 errori di incompletezza della frase “unspecified” scritta come “unspecifi”; 6 errori per la parola Multiple scritta come “Multipl”.

Non si esclude che all’interno del training set potrebbero essere rimasti refusi ed errori di battitura.

L’individuazione di tutti gli errori avrebbe richiesto tempi computazionali molto elevati.

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

* Gli attributi AUCGUART e PRIMEUNIT[[3]](#footnote-3) 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 (MMRAcqu.Auc.AvgPrice). 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

Osservate le distribuzioni dei valori di tutti gli attributi, è stata fatta una scrematura delle più interessanti, illustrate di seguito.

**IsBadBuy: i buoni acquisti sono più frequenti**

Nel dataset si nota una distribuzione sbilanciata per valori di IsBadBuy: l’87,7% delle auto è un buon acquisto, mentre il restante 12,3% non lo è.

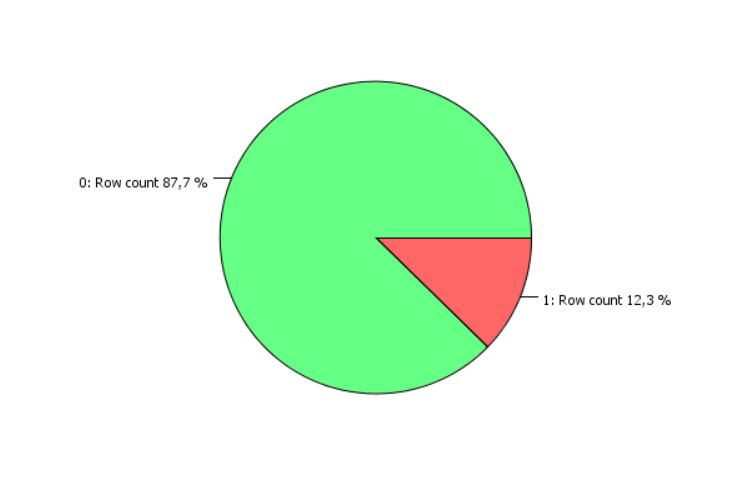
La sproporzione tra IsBadBuy=0 e IsBadBuy=1 dà un’informazione preziosa sul dataset: auto scadenti rappresentano eccezioni alle aste. Durante la costruzione e la valutazione dei modelli di classificazione, si dovranno bilanciare i valori di questo attributo così da poterne valutare l’efficienza con le diverse distribuzioni dei valori della classe. L’argomento verrà trattato approfonditamente nella quinta parte dell’elaborato.

Figura - Distribuzione IsBadBuy

**Make: Chevrolet è la casa automobilistica più frequente**

La distribuzione dei valori dell’attributo per le case produttrici presenta una preponderanza di automobili a marchio Chevrolet (23,81%), seguito poi da Dodge, Ford e Chrysler. Le auto americane rappresentano in totale l’83,77% dell’intero training set.

1. <http://www.sicurauto.it/esperto-di-sicurauto/news/crisi-del-mercato-auto-riflessioni-su-cause-responsabilita-e-rimedi.html> [↑](#footnote-ref-1)
2. *Casa produttrice, brand* in italiano. [↑](#footnote-ref-2)
3. Il tipo di garanzia sulla qualità dell’auto fornito dall’asta e il livello di richiesta sul mercato rispetto a un acquisto standard [↑](#footnote-ref-3)