

Auto Scout24 Data mining

Maria Vallarelli
Simone Zambetti
Giorgio Martelli
10/30/2020

Business target

Quale autoveicoli a KM0 sono soggetti ad un maggiore sconto?

- Dal punto di vista di un possibile acquirente, vogliamo individuare modelli e caratteristiche di autoveicoli in pronta consegna per le quali vi sia una maggiore convenienza.
- Ad oggi, alcuni portali di annunci di autoveicoli (in questo progetto vedremo autoscout24.it) già offrono modelli per la classificazione di annunci (come in immagine) ma solo per veicoli usati.

Valutazione del prezzo

Il prezzo indicato è stato confrontato tramite algoritmo con modelli dalle caratteristiche simili. La valutazione avviene in base ai dati del veicolo indicati dal venditore e prende in considerazione numerosi criteri. Particolarità riguardanti lo stato e le dotazioni del veicolo potrebbero in alcuni casi giustificare un prezzo superiore.

Informazioni sulla valutazione



Valutazione del prezzo

La valutazione del prezzo è disabilitata per i veicoli nuovi e a km 0

In che modo avviene il confronto?

La valutazione AutoScout24 confronta ciascun veicolo con offerte simili di privati e rivenditori attualmente presenti su AutoScout24 o inserite in passato. Tra i criteri di confronto rientrano per esempio la marca, il modello, l'anno di immatricolazione, la potenza, il cambio, il chilometraggio e le dotazioni.

Quali tecnologie si impiegano?

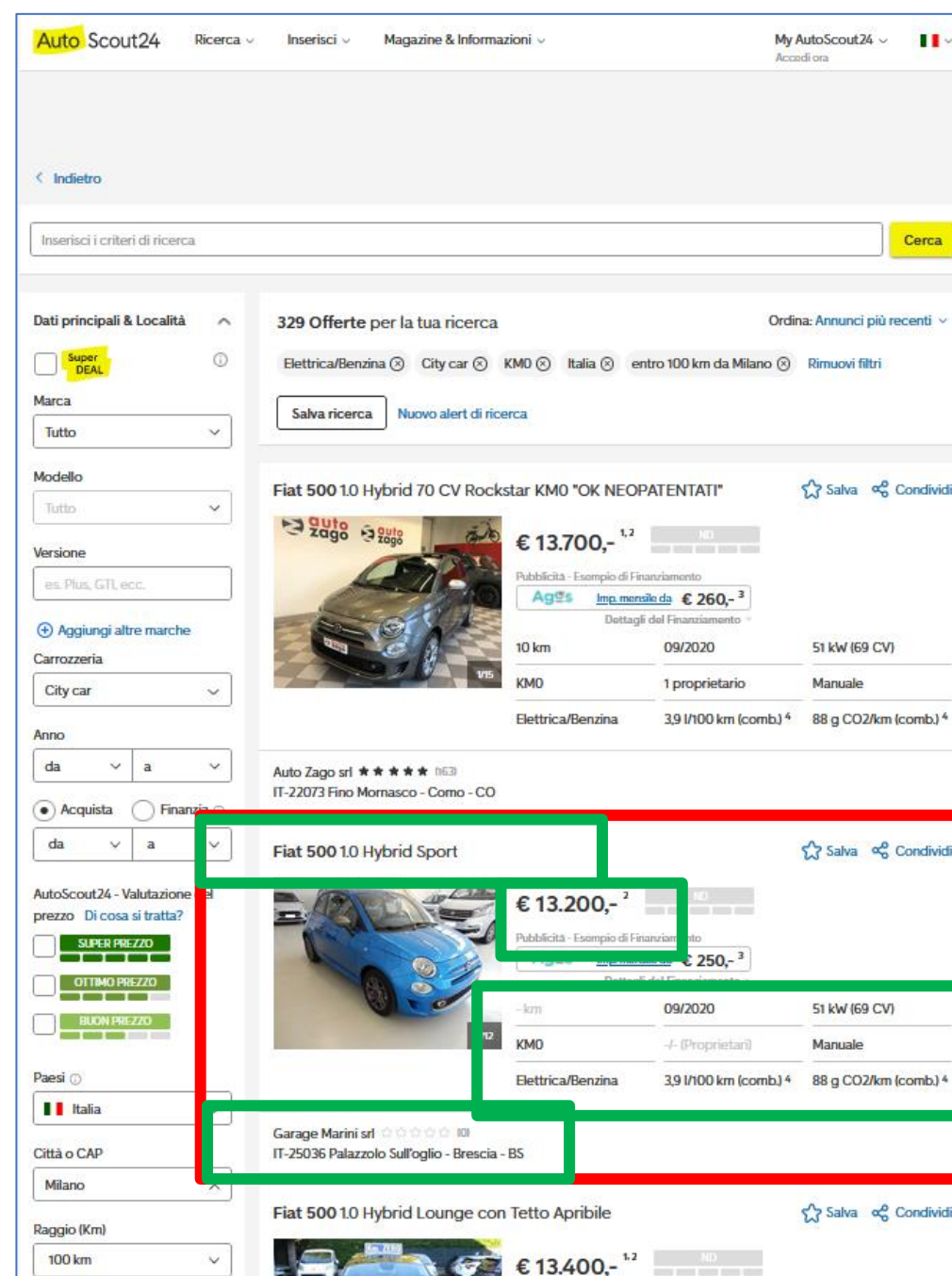
Il calcolo si avvale dei più recenti algoritmi dinamici di apprendimento automatico e di oltre 5 milioni di set dati nonché delle conoscenze specifiche dei nostri esperti. Il margine di prezzo così risultante consente di esprimere un giudizio affidabile sul rapporto qualità-prezzo.

Scraping

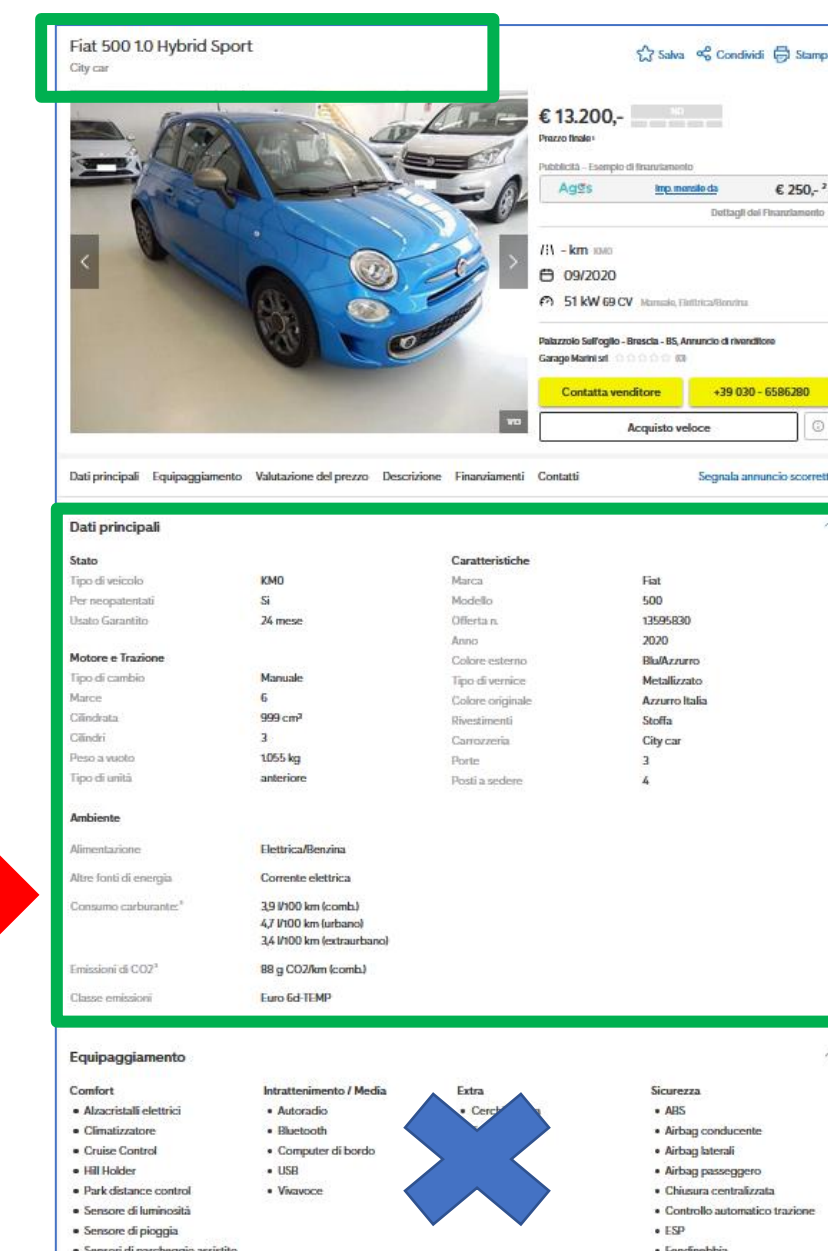


- Abbiamo formato il dataset scaricando gli annunci direttamente dal portale Autoscout24 (leader di mercato in Italia):
- In fase di scraping abbiamo filtrato gli annunci più recenti pubblicati **solo in provincia di milano e dintorni (100 km di raggio)**
- Siccome autoscout fornisce solo i primi 400 risultati per query, abbiamo utilizzato una combinazione di filtri come carrozzeria, alimentazione, tipo veicolo (KM0 e nuovo) per comporre il dataset di circa 7000 osservazioni.

Pagina di ricerca annunci



Dettaglio annuncio



Abbiamo cercato di considerare tutte le possibili variabili presenti negli annunci, scartando però equipaggiamenti inseriti opzionalmente dall'utente (i cosiddetti «optional»)

Preprocessing I



- Una volta scaricati i dati, abbiamo effettuato un'estensiva pulizia tramite OpenRefine. Ad esempio, sono stati risolti i seguenti problemi:
 - Trasformati i valori missing per veicoli elettrici dei campi numero di marce cilindrata ed emissioni
 - Uniformate le codifiche di alimentazione
 - Esclusi outlier nei prezzi (ad esempio annunci ad 1€)
 - Esclusi errori di annunci senza marca o modello (prettamente errori dello scraper)
 - Sistemazione formato dati (numerico, stringa, etc)
 - Uniformato consumo carburante
 - ... altri campi mancanti sempre tramite approccio deterministico
- Questo ci ha consentito di familiarizzare meglio con il dataset e le variabili in gioco

OpenRefine

analisi_mediana23_10_2020.xlsx

[Permalink](#)

Open...

Export ▾

Help

Facet / Filter

Undo / Redo 20 / 20

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.






Not sure how to get started?
[Watch these screencasts](#)

6778 rows

Show as: **rows** records

Show: 5 10 25 50 rows

« first < previous 1 - 10 next > last »

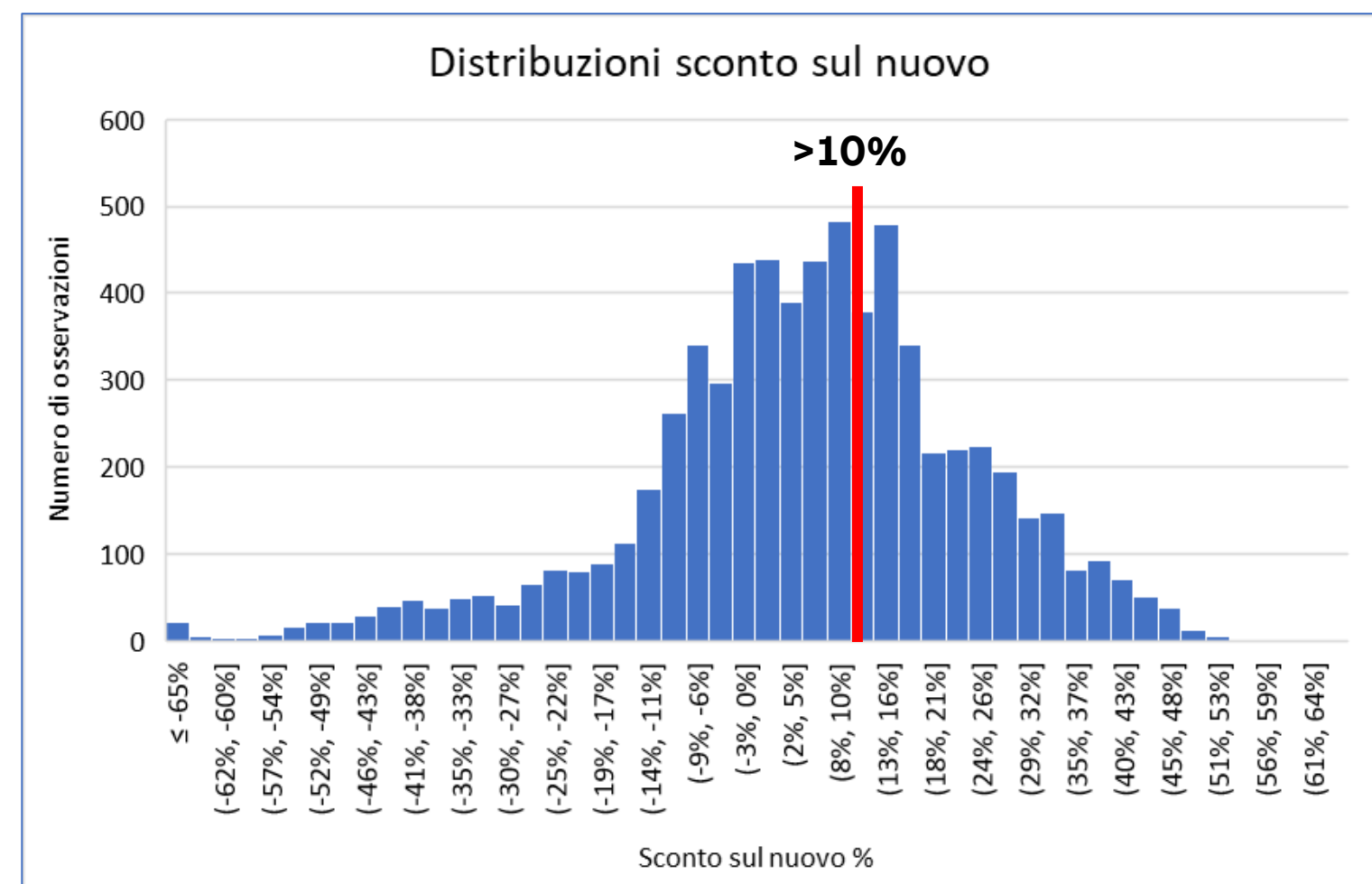
| ▼ All | ▼ Carrozzeria | ▼ Alimentazione | ▼ Tipo di veicolo | ▼ Anno | ▼ Marca | ▼ Modello | ▼ Porte | ▼ Posti a sedere | ▼ Tipo di cambio | ▼ Tipo di unita | ▼ |
|---|---------------|-----------------|-------------------|--------|---------|-----------|---------|------------------|------------------|-----------------|------------|
| ☆  | 1. | City car | Elettrica | KM0 | 2020 | BMW | i3 | 5 | 4 | Automatico | posteriore |
| ☆  | 2. | City car | Elettrica | KM0 | 2019 | BMW | i3 | 5 | | Automatico | posteriore |
| ☆  | 3. | Berlina | Elettrica | KM0 | 2020 | BMW | i3 | 5 | 4 | Automatico | posteriore |
| ☆  | 4. | Berlina | Elettrica | KM0 | 2019 | BMW | i3 | 5 | 4 | Automatico | posteriore |
| ☆  | 5. | Berlina | Elettrica | KM0 | 2019 | BMW | i3 | 5 | 4 | Automatico | posteriore |

In generale, abbiamo cercato di ridurre i missing tramite approcci deterministici il più possibile visto che nel nostro caso non rappresentano informazioni aggiuntive ma puramente mancanze da parte di chi ha pubblicato gli annunci

Preprocessing II



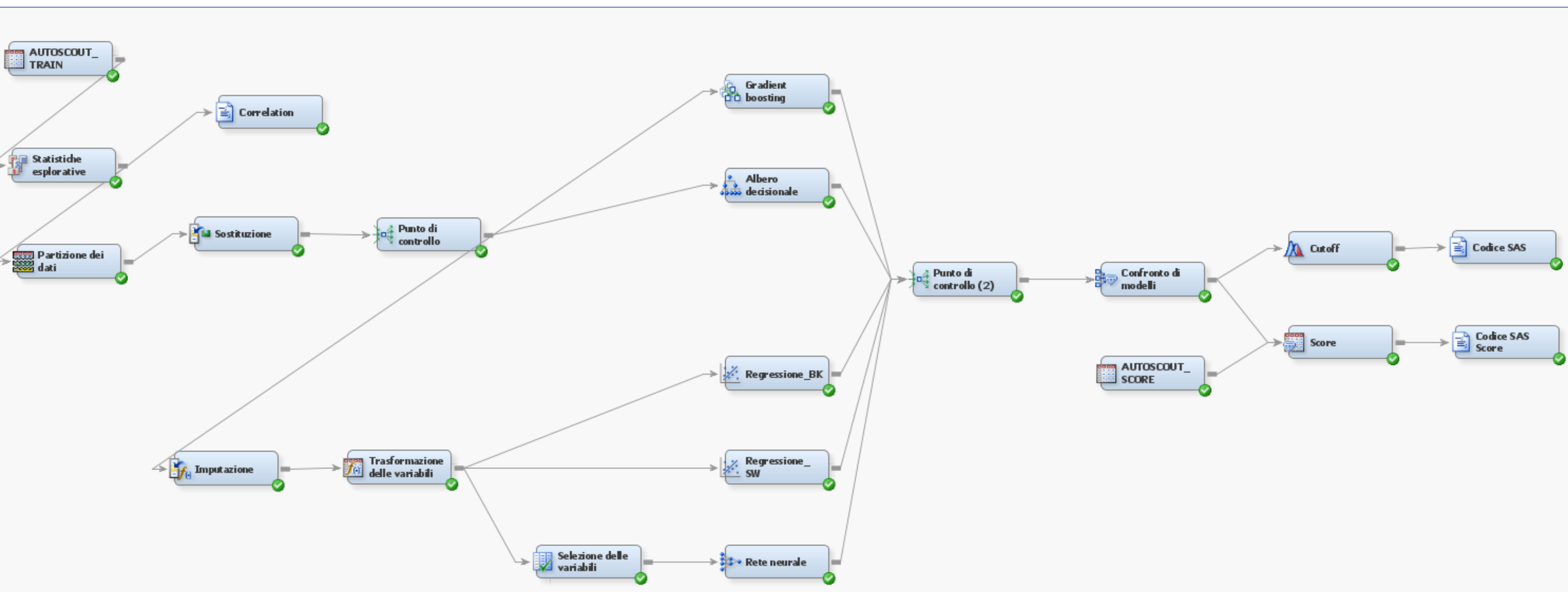
- **Individuazione variable target:** avendo scaricato gli annunci di veicoli nuovi e preimmatricolati abbiamo effettuato una inner join per mantenere solo i modelli presenti in entrambe le tipologie.
- Abbiamo dunque calcolato il prezzo mediano per modello di autoveicolo nuovo come baseline, e calcolato lo sconto per ogni singola osservazione a KM0. Dunque abbiamo calcolato le % di sconto sul prezzo nuovo per singola osservazione a KM0, le quali sono distribuite nel dataset come segue:



Dunque la variabile dipendente è stata dicotomizzata in base alla soglia del 10%.

Vi è un duplice motivo di business per questa scelta: abbiamo reputato che 10% sia una soglia accettabile di sconto, in più è lo sconto minimo necessario per accedere a incentivi della RL del decreto rilancio

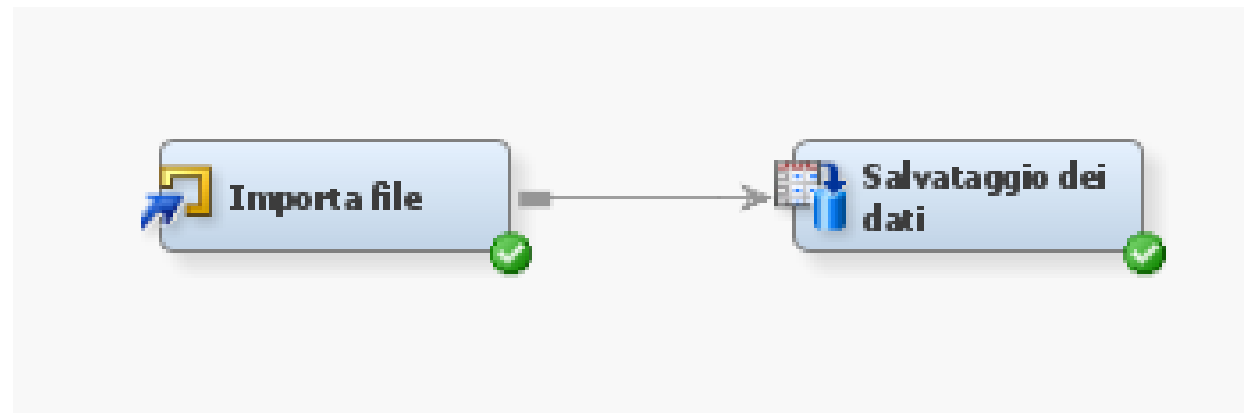
Overall diagram





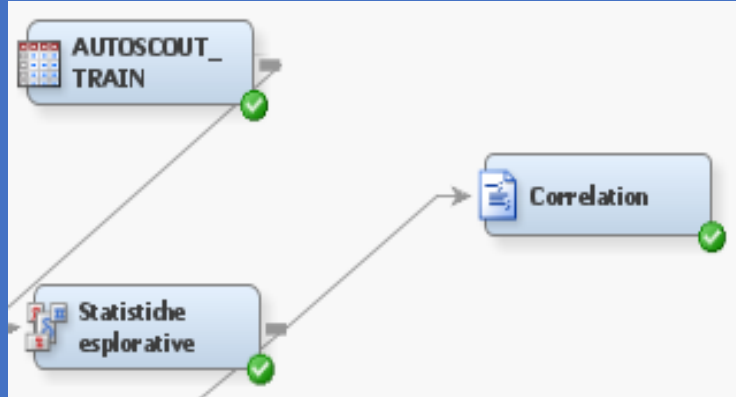
SAS

- Il dataset è stato dunque importato su SAS come xlsx e salvato come dataset di sas



- Le seguenti sono le variabili e target presenti nel dataset:

| Nome | Ruolo | Livello |
|---------------------------|-----------|----------|
| Alimentazione | Input | Nominale |
| Anno | Input | Nominale |
| Carrozzeria | Input | Nominale |
| Cilindrata | Input | Continuo |
| Cilindri | Input | Nominale |
| Classe emissioni | Input | Nominale |
| Colore esterno | Input | Nominale |
| Consumo Carburante Totale | Input | Continuo |
| Emissioni di CO2 | Input | Continuo |
| Marca | Input | Nominale |
| Marce | Input | Nominale |
| Modello | Input | Nominale |
| Per neopatentati | Input | Binario |
| Peso a vuoto | Input | Continuo |
| Porte | Input | Nominale |
| Posti a sedere | Input | Nominale |
| Tagliandi certificati | Input | Binario |
| Tipo di cambio | Input | Nominale |
| Tipo di unita | Input | Nominale |
| Tipo di vernice | Input | Nominale |
| Usato Garantito | Input | Binario |
| Veicolo per non fumatori | Rifiutata | Binario |
| conditional on price | Rifiutata | Nominale |
| price | Input | Continuo |
| target10 | Target | Binario |
| target20 | Rifiutata | Binario |
| target 35 | Rifiutata | Binario |
| target 5 | Rifiutata | Binario |
| vehicle user desc | Rifiutata | Nominale |



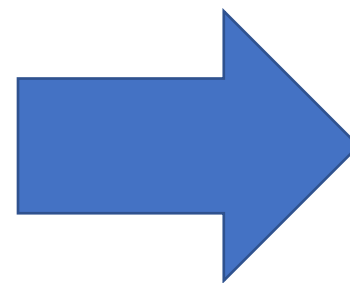
Esplorative

- Tramite il nodo di statistiche esplorative abbiamo effettuato un test X^2 di associazione X-Y:

Statistiche del chi-quadrato
(massimo 500 osservazioni stampate)

Ruolo dei dati=TRAIN Target=target10

| Input | Chi-quadrato | Df | Prob |
|---------------------------|--------------|-----|--------|
| Modello | 3355.4279 | 232 | <.0001 |
| Marca | 1577.2164 | 39 | <.0001 |
| Posti_a_sedere | 571.1949 | 8 | <.0001 |
| Peso_a_vuoto | 513.9707 | 5 | <.0001 |
| Carrozzeria | 509.0237 | 7 | <.0001 |
| Alimentazione | 440.9962 | 5 | <.0001 |
| Per_neopatentati | 334.0258 | 1 | <.0001 |
| Marce | 268.3107 | 7 | <.0001 |
| Emissioni_di_CO2 | 204.5847 | 5 | <.0001 |
| Cilindri | 175.6754 | 8 | <.0001 |
| Porte | 168.4790 | 2 | <.0001 |
| Anno | 163.3468 | 1 | <.0001 |
| Colore_esterno | 134.9968 | 14 | <.0001 |
| Tipo_di_unita | 111.6917 | 3 | <.0001 |
| Cilindrata | 92.8011 | 5 | <.0001 |
| price | 64.7145 | 4 | <.0001 |
| Tipo_di_cambio | 64.2160 | 3 | <.0001 |
| Classe_emissioni | 24.3085 | 7 | 0.0010 |
| Tipo_di_vernice | 15.1526 | 3 | 0.0017 |
| Tagliandi_certificati | 12.5810 | 1 | 0.0004 |
| conditional_on_price | 10.9375 | 9 | 0.2800 |
| Consumo_Carburante_Totale | 9.1097 | 3 | 0.0279 |
| Usato_Garantito | 7.2533 | 1 | 0.0071 |
| Veicolo_per_non_fumatori | 0.8340 | 1 | 0.3611 |

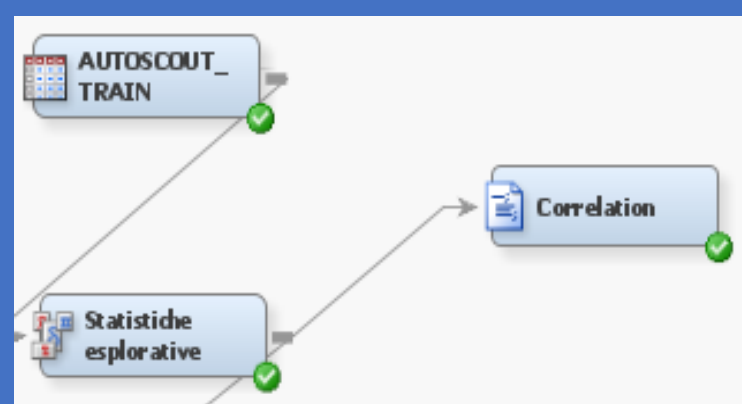


Statistiche del chi-quadrato
(massimo 500 osservazioni stampate)

Ruolo dei dati=TRAIN Target=target10

| Input | Chi-quadrato | Df | Prob |
|---------------------------|--------------|-----|--------|
| Modello | 3355.4279 | 232 | <.0001 |
| Marca | 1577.2164 | 39 | <.0001 |
| Posti_a_sedere | 571.1949 | 8 | <.0001 |
| Peso_a_vuoto | 513.9707 | 5 | <.0001 |
| Carrozzeria | 509.0237 | 7 | <.0001 |
| Alimentazione | 440.9962 | 5 | <.0001 |
| Per_neopatentati | 334.0258 | 1 | <.0001 |
| Marce | 268.3107 | 7 | <.0001 |
| Emissioni_di_CO2 | 204.5847 | 5 | <.0001 |
| Cilindri | 175.6754 | 8 | <.0001 |
| Porte | 168.4790 | 2 | <.0001 |
| Anno | 163.3468 | 1 | <.0001 |
| Colore_esterno | 134.9968 | 14 | <.0001 |
| Tipo_di_unita | 111.6917 | 3 | <.0001 |
| Cilindrata | 92.8011 | 5 | <.0001 |
| price | 64.7145 | 4 | <.0001 |
| Tipo_di_cambio | 64.2160 | 3 | <.0001 |
| Classe_emissioni | 24.3085 | 7 | 0.0010 |
| Tipo_di_vernice | 15.1526 | 3 | 0.0017 |
| Tagliandi_certificati | 12.5810 | 1 | 0.0004 |
| Consumo_Carburante_Totale | 9.1097 | 3 | 0.0279 |
| Usato_Garantito | 7.2533 | 1 | 0.0071 |

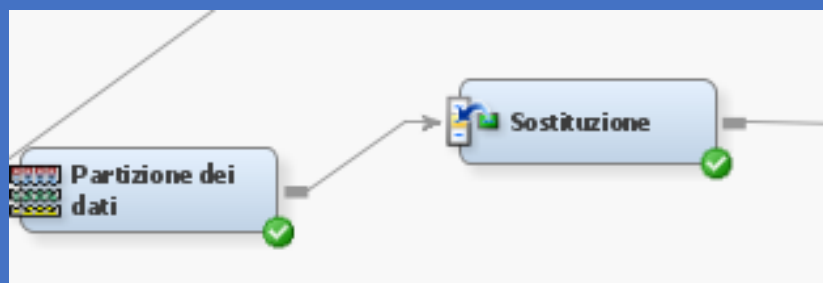
- Abbiamo dunque accettato l'ipotesi nulla (ovvero mancanza di dipendenza) per le covariate conditional_on_price e veicolo_per_non_fumatori, questo ha senso anche dal punto di vista del business.
- Inoltre, la covariabile marca è stata rifiutata, perché ridondante rispetto a Modello, visto che ad ogni modello corrisponde una ed una sola marca, e non vi sono problemi di missing o qualità di input.



SAS

- Tramite il nodo di codice SAS «correlation» abbiamo effettuato una proc corr per individuare eventuali correlazioni tra le variabili ma il coefficiente di correlazione non supera 0.60

| | | Coefficienti di correlazione di Pearson Num. di osservazioni | | | | | | | | | | | | | | | | | |
|----------------------|----------------------|---|------------------|------------------|------------------|--------------------------|------------------|------------------|------------------|------------------|------------------|--------------------|----------------------|------------------|---------------------|------------------|----------------------------------|-----------------------------------|---------------------------|
| | | Porte | Marce | Cilindrata | Cilindri | conditional_ on_price | price | target_35 | target_5 | target10 | target20 | Posti_a_ sedere | Emissioni_ di_CO2 | Peso_a_ vuoto | Usato_ Garantito | Per_neopatentati | Veicolo_ per_non_ fumatori | Consumo_ Carburante_ Totale | Tagliandi_ certificati |
| Porte | Porte | 1.00000 6489 | 0.10465 6105 | 0.22147 6419 | 0.20094 6254 | 0.03098 6489 | 0.20565 6489 | 0.09545 6489 | 0.16897 6489 | 0.15931 6489 | 0.12474 6489 | 0.62478 6377 | 0.25107 5932 | 0.51104 3783 | -0.02139 6489 | -0.38840 6489 | 0.02316 6489 | 0.12185 5915 | 0.02574 6489 |
| Marce | Marce | 0.10465 6105 | 1.00000 6124 | 0.42700 6104 | 0.37903 6030 | 0.05875 6124 | 0.52282 6124 | 0.00671 6124 | 0.12503 6124 | 0.12716 6124 | 0.05210 6124 | 0.20259 6085 | 0.30015 5641 | 0.51708 3729 | 0.06010 6124 | -0.24878 6124 | -0.01630 6124 | 0.05055 5678 | 0.06912 6124 |
| Cilindrata | Cilindrata | 0.22147 6419 | 0.42700 6104 | 1.00000 6679 | 0.59661 6272 | 0.07314 6679 | 0.49043 6679 | -0.08021 6679 | 0.13994 6679 | 0.12514 6679 | 0.07454 6679 | 0.24580 6448 | 0.36428 5898 | 0.43124 3788 | 0.10661 6679 | -0.17250 6679 | 0.13316 6679 | 0.05620 5927 | 0.42409 6679 |
| Cilindri | Cilindri | 0.20094 6254 | 0.37903 6030 | 0.59661 6272 | 1.00000 6287 | 0.05170 6287 | 0.34861 6287 | -0.07866 6287 | 0.15164 6287 | 0.10819 6287 | 0.03076 6287 | 0.23616 6215 | 0.59681 5774 | 0.39742 3785 | 0.02037 6287 | -0.11787 6287 | 0.04320 6287 | 0.16304 5804 | 0.05449 6287 |
| conditional_on_price | conditional_on_price | 0.03098 6489 | 0.05875 6124 | 0.07314 6679 | 0.05170 6287 | 1.00000 6778 | 0.04937 6778 | -0.02854 6778 | 0.00932 6778 | -0.01932 6778 | -0.02854 6778 | 0.02573 6469 | 0.03665 5948 | 0.07944 3788 | 0.09327 6778 | -0.00848 6778 | 0.04233 6778 | 0.07831 5930 | 0.01400 6778 |
| price | price | 0.20565 6489 | 0.52282 6124 | 0.49043 6679 | 0.34861 6287 | 0.04937 6778 | 1.00000 6778 | -0.01513 6778 | 0.00055 6778 | 0.00539 6778 | 0.00990 6778 | 0.25425 6469 | 0.18567 5948 | 0.74555 3788 | -0.02650 6778 | -0.33452 6778 | 0.00224 6778 | 0.01205 5930 | 0.04680 6778 |
| target_35 | target_35 | 0.09545 6489 | 0.00671 6124 | -0.08021 6679 | -0.07866 6287 | -0.02854 6778 | -0.01513 6778 | 1.00000 6778 | 0.21792 6778 | 0.27856 6778 | 0.45997 6778 | 0.04847 6469 | 0.09739 5948 | 0.05064 3788 | -0.02634 6778 | -0.11327 6778 | -0.00011 6778 | 0.03553 5930 | -0.02783 6778 |
| target_5 | target_5 | 0.16897 6489 | 0.12503 6124 | 0.13994 6679 | 0.15164 6287 | 0.00932 6778 | 0.00055 6778 | 0.21792 6778 | 1.00000 6778 | 0.78228 6778 | 0.47376 6778 | 0.13416 6469 | 0.19135 5948 | 0.24765 3788 | 0.04541 6778 | -0.20375 6778 | 0.00411 6778 | 0.04099 5930 | 0.04496 6778 |
| target10 | target10 | 0.15931 6489 | 0.12716 6124 | 0.12514 6679 | 0.10819 6287 | -0.01932 6778 | 0.00539 6778 | 0.27856 6778 | 0.78228 6778 | 1.00000 6778 | 0.60562 6778 | 0.10557 6469 | 0.16970 5948 | 0.23474 3788 | 0.03271 6778 | -0.22199 6778 | -0.01109 6778 | 0.03935 5930 | 0.04308 6778 |
| target20 | target20 | 0.12474 6489 | 0.05210 6124 | 0.07454 6679 | 0.03076 6287 | -0.02854 6778 | 0.00990 6778 | 0.45997 6778 | 0.47376 6778 | 0.60562 6778 | 1.00000 6778 | 0.03223 6469 | 0.13265 5948 | 0.17095 3788 | 0.00902 6778 | -0.19822 6778 | 0.00667 6778 | 0.02847 5930 | 0.03645 6778 |
| Posti_a_sedere | | 0.62478 6377 | 0.20259 6085 | 0.24580 6448 | 0.23616 6215 | 0.02573 6469 | 0.25425 6469 | 0.04847 6469 | 0.13416 6469 | 0.10557 6469 | 0.03223 6469 | 1.00000 6469 | 0.21486 5885 | 0.51259 3768 | -0.02892 6469 | -0.30131 6469 | 0.03622 6469 | 0.04392 5916 | 0.01435 6469 |
| Emissioni_di_CO2 | | 0.25107 5932 | 0.30015 5641 | 0.36428 5898 | 0.59681 5774 | 0.03665 5948 | 0.18567 5948 | 0.09739 5948 | 0.19135 5948 | 0.16970 5948 | 0.13265 5948 | 0.21486 5885 | 1.00000 5948 | 0.35062 3695 | -0.03439 5948 | -0.12350 5948 | 0.03392 5948 | 0.41455 5869 | 0.04074 5948 |
| Peso_a_vuoto | | 0.51104 3783 | 0.51708 3729 | 0.43124 3788 | 0.39742 3785 | 0.07944 3788 | 0.74555 3788 | 0.05064 3788 | 0.24765 3788 | 0.23474 3788 | 0.17095 3788 | 0.51259 3768 | 0.35062 3695 | 1.00000 3788 | -0.00316 3788 | -0.58067 3788 | -0.02833 3788 | 0.07922 3694 | 0.02043 3788 |
| Usato_Garantito | | -0.02139 6489 | 0.06010 6124 | 0.10661 6679 | 0.02037 6287 | 0.09327 6778 | -0.02650 6778 | -0.02634 6778 | 0.04541 6778 | 0.03271 6778 | 0.00902 6778 | -0.02892 6469 | -0.03439 5948 | -0.00316 3788 | 1.00000 6778 | 0.17645 6778 | 0.22129 6778 | -0.04461 5930 | 0.21220 6778 |
| Per_neopatentati | | -0.38840 6489 | -0.24878 6124 | -0.17250 6679 | -0.11787 6287 | -0.00848 6778 | -0.33452 6778 | -0.11327 6778 | -0.20375 6778 | -0.22199 6778 | -0.19822 6778 | -0.30131 6469 | -0.12350 5948 | -0.58067 3788 | 0.17645 6778 | 1.00000 6778 | 0.10649 6778 | -0.01583 5930 | 0.05063 6778 |

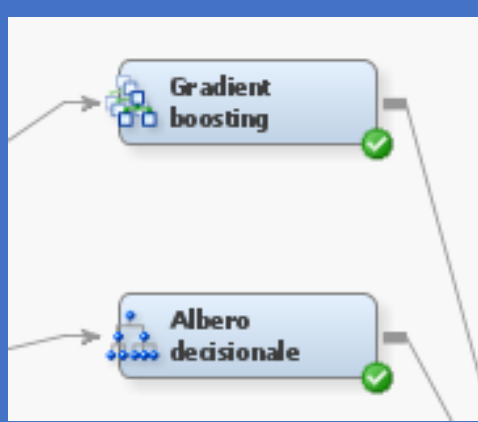


SAS II

- Dopo aver partizionato il dataset utilizzando una external holdout 70-30, è stata effettuata una sostituzione solo per le variabili categoriche missing. Non si è ritenuto opportuno sostituire le continue prevedendo di fare un'imputazione prima di applicare i modelli che non gestiscono i valori missing.

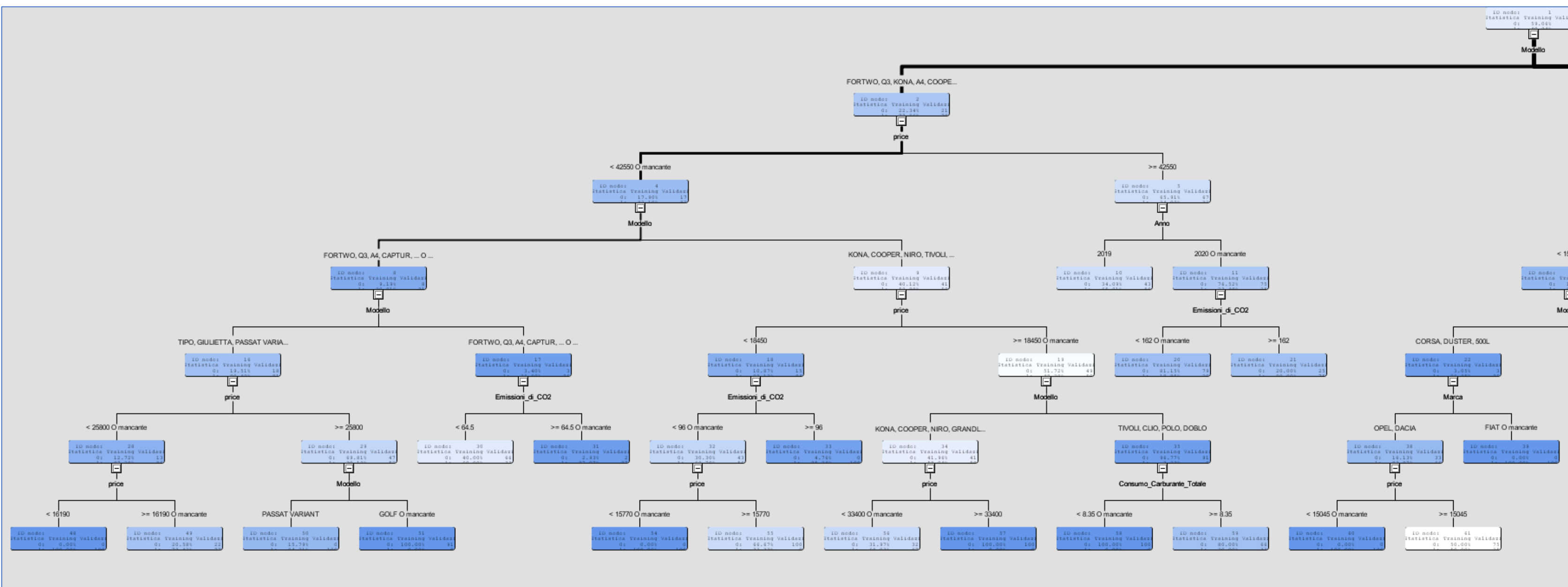
| Proprietà | Valore |
|-------------------------------|-----------------------------|
| Generale | |
| ID nodo | Repl |
| Dati importati | ... |
| Dati esportati | ... |
| Note | ... |
| Training | |
| Variabili continue | |
| Editor delle sostituzioni | ... |
| Metodo dei limiti predefinito | Nessuno |
| Valori di cutoff | ... |
| Variabili categoriche | |
| Editor delle sostituzioni | ... |
| Livelli sconosciuti | Ignora |
| Score | |
| Valori di sostituzione | Calcolati |
| Nascondi | No |
| Report | |
| Report di sostituzione | Sì |
| Stato | |
| Ora creazione | 19/10/20 23.46 |
| ID esecuzione | c1470f51-fc0b-46bb-a2d9-75d |
| Ultimo errore | |
| Ultimo stato | Completo |
| Ora ultima esecuzione | 28/10/20 19.35 |
| Durata esecuzione | 0 ore 0 min. 10,84 sec. |
| Host grid | |
| Nodo utente | No |

| Editor delle sostituzioni-WORK.OUTCLASS | | | | | | |
|---|-------------------|------------------------|------------------------|------|------------------------------------|-----------------|
| Variabile | Valore formattato | Valore di sostituzione | Conteggio di frequenza | Tipo | Valore alfanumerico non formattato | Valore numerico |
| Carrozzeria | Monovolume | | 552C | | Monovolume | . |
| Carrozzeria | Cabrio | | 166C | | Cabrio | . |
| Carrozzeria | CoupÃ© | Coupe | 66C | | CoupÃ© | . |
| Carrozzeria | _UNKNOWN_ | _DEFAULT_ | C | | | . |
| Cilindri | 4 | | 2926N | | | 4 |
| Cilindri | 3 | | 1246N | | | 3 |
| Cilindri | . | | 335N | | | . |
| Cilindri | 2 | | 102N | | | 2 |
| Cilindri | 0 | | 60N | | | 0 |
| Cilindri | 6 | | 38N | | | 6 |
| Cilindri | 1 | | 27N | | | 1 |
| Cilindri | 8 | | 8N | | | 8 |
| Cilindri | 5 | | 1N | | | 5 |
| Cilindri | _UNKNOWN_ | _DEFAULT_ | N | | | . |
| Classe_emissioni | Euro 6 | | 2610C | | Euro 6 | . |
| Classe_emissioni | Euro 6d-TEMP | | 1289C | | Euro 6d-TEMP | . |
| Classe_emissioni | | _MISSING_ | 653C | | | . |
| Classe_emissioni | Euro 6d | | 128C | | Euro 6d | . |
| Classe_emissioni | Elettrica | | 57C | | Elettrica | . |
| Classe_emissioni | Euro 6c | | 3C | | Euro 6c | . |
| Classe_emissioni | Euro 5 | | 2C | | Euro 5 | . |
| Classe_emissioni | Euro 4 | | 1C | | Euro 4 | . |
| Classe_emissioni | _UNKNOWN_ | _DEFAULT_ | C | | | . |
| Colore_esterno | Grigio | | 1304C | | Grigio | . |
| Colore_esterno | Bianco | | 1297C | | Bianco | . |
| Colore_esterno | Nero | | 810C | | Nero | . |
| Colore_esterno | Blu/Azzurro | | 465C | | Blu/Azzurro | . |
| Colore_esterno | Argento | | 371C | | Argento | . |
| Colore_esterno | Rosso | | 200C | | Rosso | . |
| Colore_esterno | Verde | | 91C | | Verde | . |
| Colore_esterno | | _MISSING_ | 84C | | | . |



Alberi decisionali e gradient boosting

- 2 tipi di alberi decisionali, uno con pruning automatico ed un gradient boosting
- Massima profondità di 6 e una misura di foglia minima di 10
- Non abbiamo trattato i missing visto che questi modelli li supportano



Imputazione e trasformazione

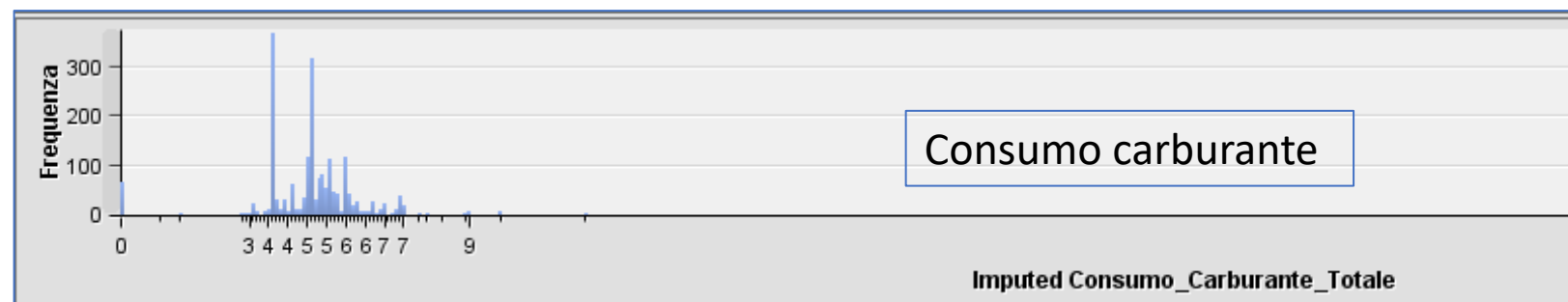
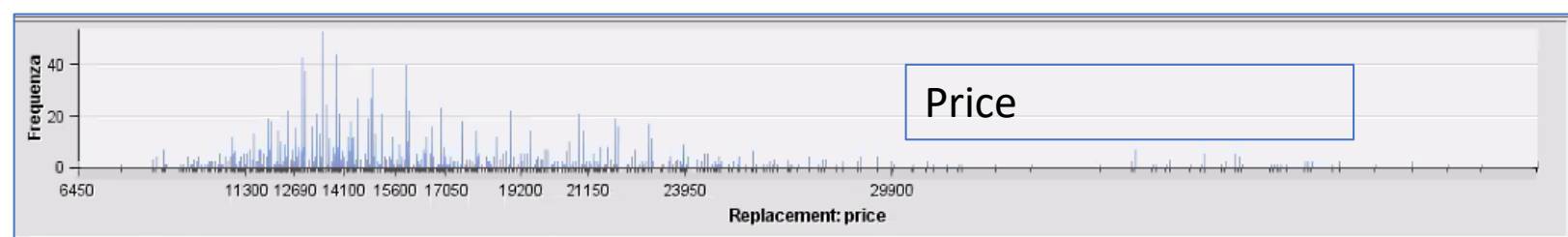
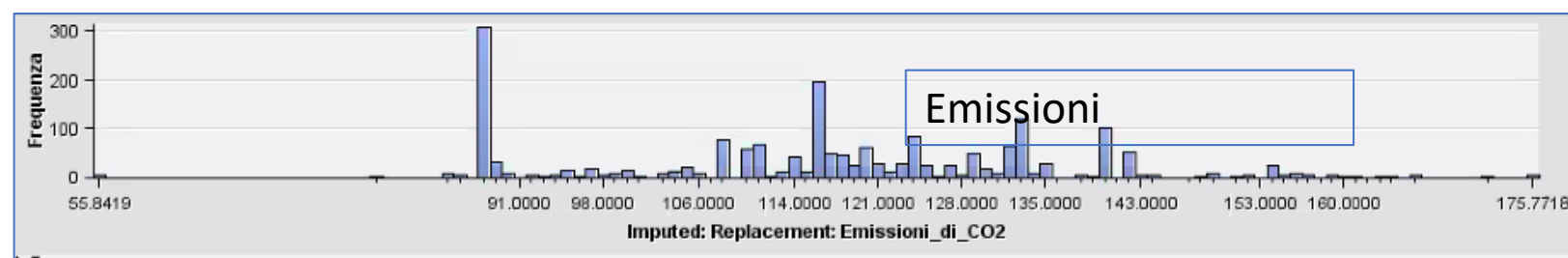


- Prima delle regressioni logistiche sui dataset, sono state imputate le covariate categoriche tramite alberi decisionali surrogati e le continue tramite mediana
- La trasformazione delle variabili è stata poi effettuata per redistribuire l'asimmetria della covariata «prezzo» ed abbiamo raggruppato pesi, emissioni, consumi che erano distribuiti più uniformemente, alcuni con una curtosi alta

Statistiche di riepilogo delle variabili continue
(massimo 500 osservazioni stampate)

Ruolo dei dati=TRAIN

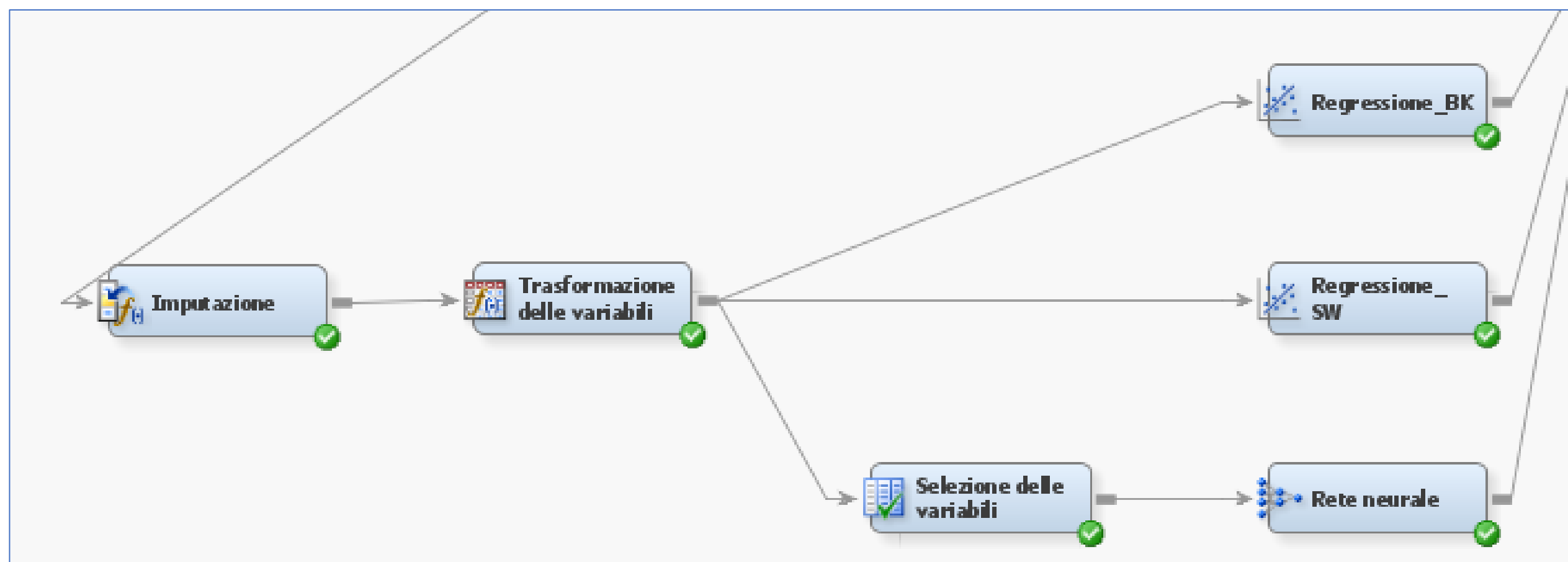
| Variabile | Ruolo | Media | Deviazione standard | Non mancanti | Mancanti | Minimo | Mediana | Massimo | Skewness | Curtosi |
|---------------------------|-------|----------|---------------------|--------------|----------|--------|---------|---------|----------|----------|
| Cilindrata | INPUT | 1452.299 | 582.2421 | 6679 | 99 | 0 | 1368 | 6000 | 2.127758 | 8.465619 |
| Consumo_Carburante_Totale | INPUT | 5.0857 | 2.741942 | 5930 | 848 | 0 | 4.9 | 65 | 14.00762 | 258.8244 |
| Emissioni_di_CO2 | INPUT | 115.132 | 27.49977 | 5948 | 830 | 0 | 116 | 284 | -0.54482 | 6.163999 |
| Peso_a_vuoto | INPUT | 1309.355 | 249.3186 | 3788 | 2990 | 350 | 1320 | 2585 | 0.772335 | 1.663317 |
| price | INPUT | 22446.52 | 13689.16 | 6778 | 0 | 6450 | 18700 | 189000 | 4.126061 | 30.4946 |

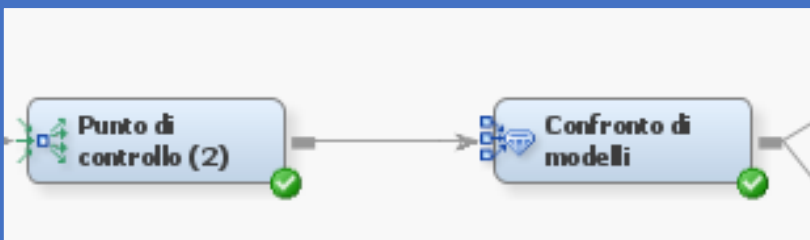


| Variabili - Trans | | | | |
|--|------------------------------------|--------------------------|-----------|----------|
| <div>(nessuno) <input type="checkbox"/> not Uguale <input type="checkbox"/> Mining <input type="checkbox"/> Base</div> | | | | |
| Colonne: | <input type="checkbox"/> Etichetta | | | |
| Nome | Metodo | Numero di raggruppamenti | Ruolo | Livello |
| IMP Emissioni di CO2 | Raggruppamento ottimale | 4 | Input | Continuo |
| IMP Cilindrata | Raggruppamento ottimale | 4 | Input | Continuo |
| IMP Peso a vuoto | Raggruppamento ottimale | 4 | Input | Continuo |
| price | Log 10 | 4 | Input | Continuo |
| Anno | Predefinito | 4 | Input | Nominale |
| IMP Cilindri | Predefinito | 4 | Input | Nominale |
| Alimentazione | Predefinito | 4 | Input | Nominale |
| Classe emissioni | Predefinito | 4 | Rifiutata | Nominale |
| Colore esterno | Predefinito | 4 | Rifiutata | Nominale |
| IMP Consumo Carburante Totale | Predefinito | 4 | Input | Continuo |
| IMP Porte | Predefinito | 4 | Input | Nominale |
| IMP Marce | Predefinito | 4 | Input | Nominale |
| IMP REP Classe emissioni | Predefinito | 4 | Input | Nominale |
| Carrozzeria | Predefinito | 4 | Rifiutata | Nominale |
| IMP Posti a sedere | Predefinito | 4 | Input | Nominale |
| IMP REP Tipo di cambio | Predefinito | 4 | Input | Nominale |
| Modello | Predefinito | 4 | Input | Nominale |
| IMP REP Tipo di unita | Predefinito | 4 | Input | Nominale |
| IMP REP Colore esterno | Predefinito | 4 | Input | Nominale |
| IMP REP Tipo di vernice | Predefinito | 4 | Input | Nominale |
| Marca | Predefinito | 4 | Input | Nominale |
| Per neopatentati | Predefinito | 4 | Input | Binario |
| Tagliandi certificati | Predefinito | 4 | Input | Binario |
| Usato Garantito | Predefinito | 4 | Input | Binario |
| Tipo di cambio | Predefinito | 4 | Rifiutata | Nominale |
| REP Carrozzeria | Predefinito | 4 | Input | Nominale |
| Tipo di unita | Predefinito | 4 | Rifiutata | Nominale |
| Tipo di vernice | Predefinito | 4 | Rifiutata | Nominale |
| target10 | Predefinito | 4 | Target | Binario |

Regressioni e rete

- Sono stati dunque creati vari nodi di regressione backward e stepwise
- Sono state anche selezionate le variabili per il ridurne il numero in import alla rete neurale





Confronto tra modelli

- Durante l'assessment, abbiamo scelto come metodo di selezione la ROC, visto che vogliamo massimizzare TPR e minimizzare FPR
- Il modello ad albero decisionale performa meglio su tutte le soglie, massimizzando AUC sempre
- ROC ed anche ASE sono consistenti tra train e validation per tutti i modelli

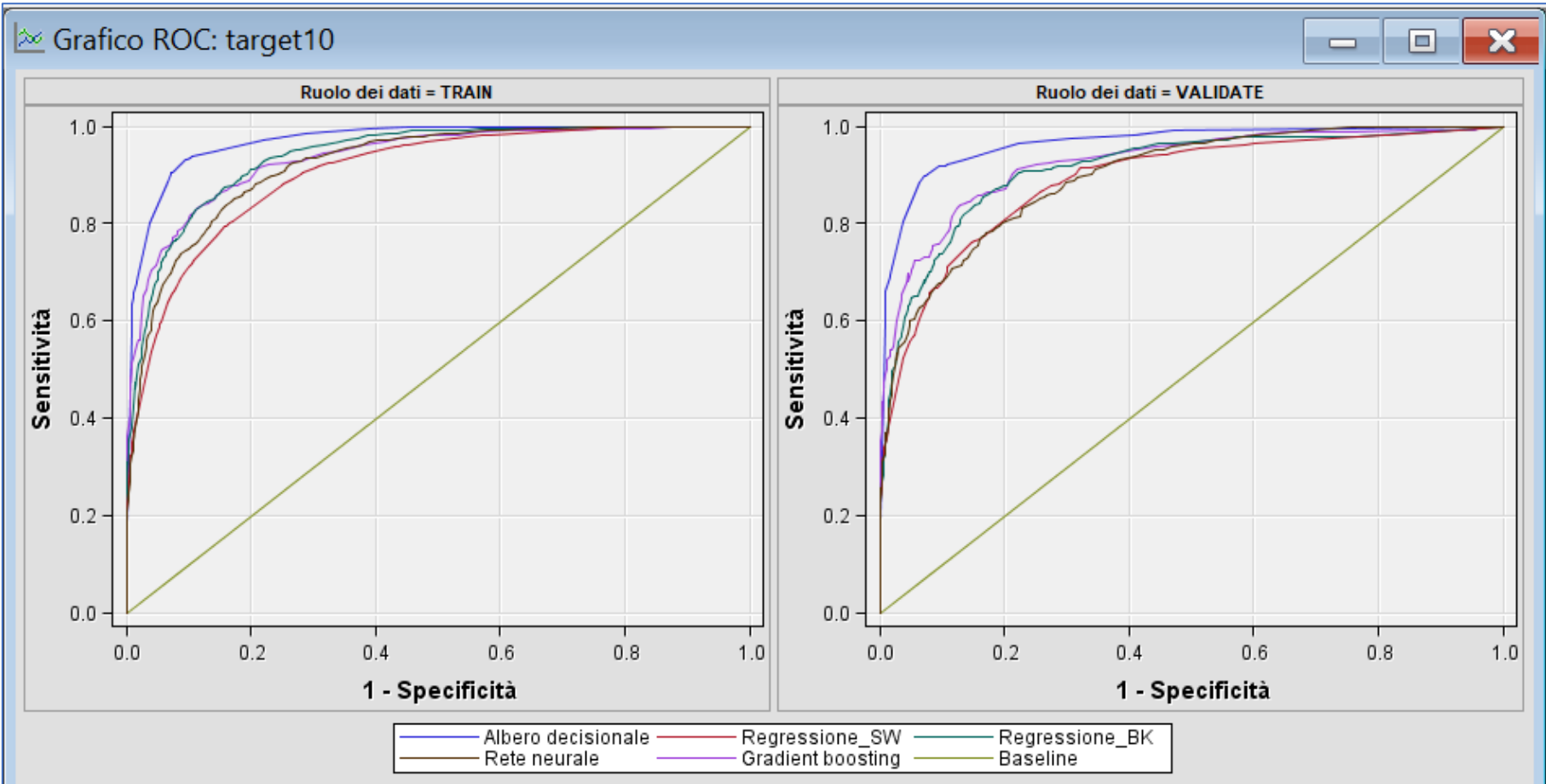
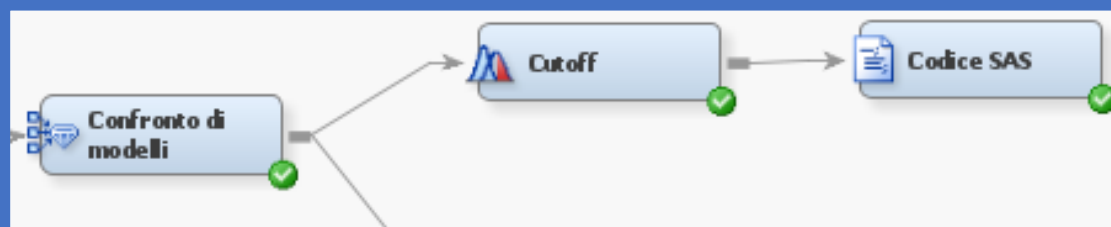


Tabella di classificazione degli eventi
Selezione del modello basata su Train: Indice ROC (_AUR_)

| Nodo del modello | Descrizione del modello | Ruolo dei dati | Target | Etichetta target | Falso negativo | Vero negativo | Falso positivo | Vero positivo |
|------------------|-------------------------|----------------|----------|------------------|----------------|---------------|----------------|---------------|
| Tree2 | Albero decisionale | TRAIN | target10 | target10 | 182 | 2592 | 209 | 1760 |
| Tree2 | Albero decisionale | VALIDATE | target10 | target10 | 83 | 1110 | 92 | 750 |
| Boost | Gradient boosting | TRAIN | target10 | target10 | 383 | 2529 | 272 | 1559 |
| Boost | Gradient boosting | VALIDATE | target10 | target10 | 183 | 1072 | 130 | 650 |
| Reg2 | Regressione_BK | TRAIN | target10 | target10 | 363 | 2507 | 294 | 1579 |
| Reg2 | Regressione_BK | VALIDATE | target10 | target10 | 172 | 1053 | 149 | 661 |
| Reg3 | Regressione_SW | TRAIN | target10 | target10 | 436 | 2396 | 405 | 1506 |
| Reg3 | Regressione_SW | VALIDATE | target10 | target10 | 207 | 1033 | 169 | 626 |
| Neural | Rete neurale | TRAIN | target10 | target10 | 400 | 2423 | 378 | 1542 |
| Neural | Rete neurale | VALIDATE | target10 | target10 | 220 | 1027 | 175 | 613 |

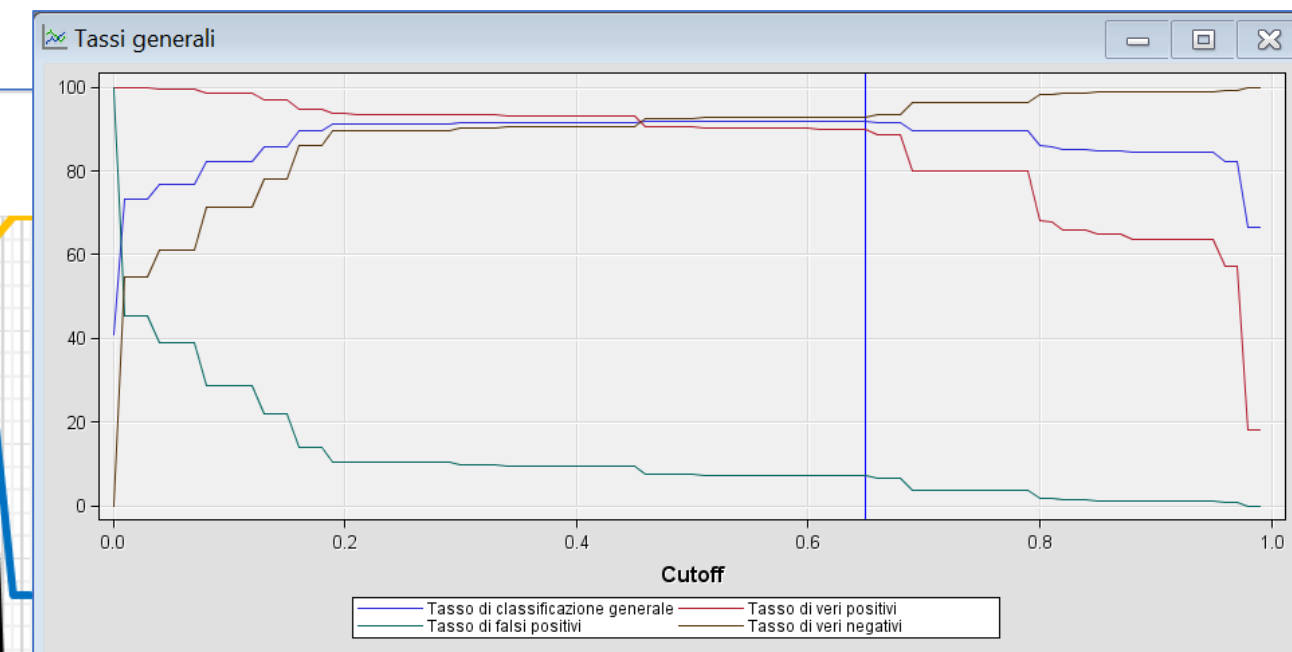
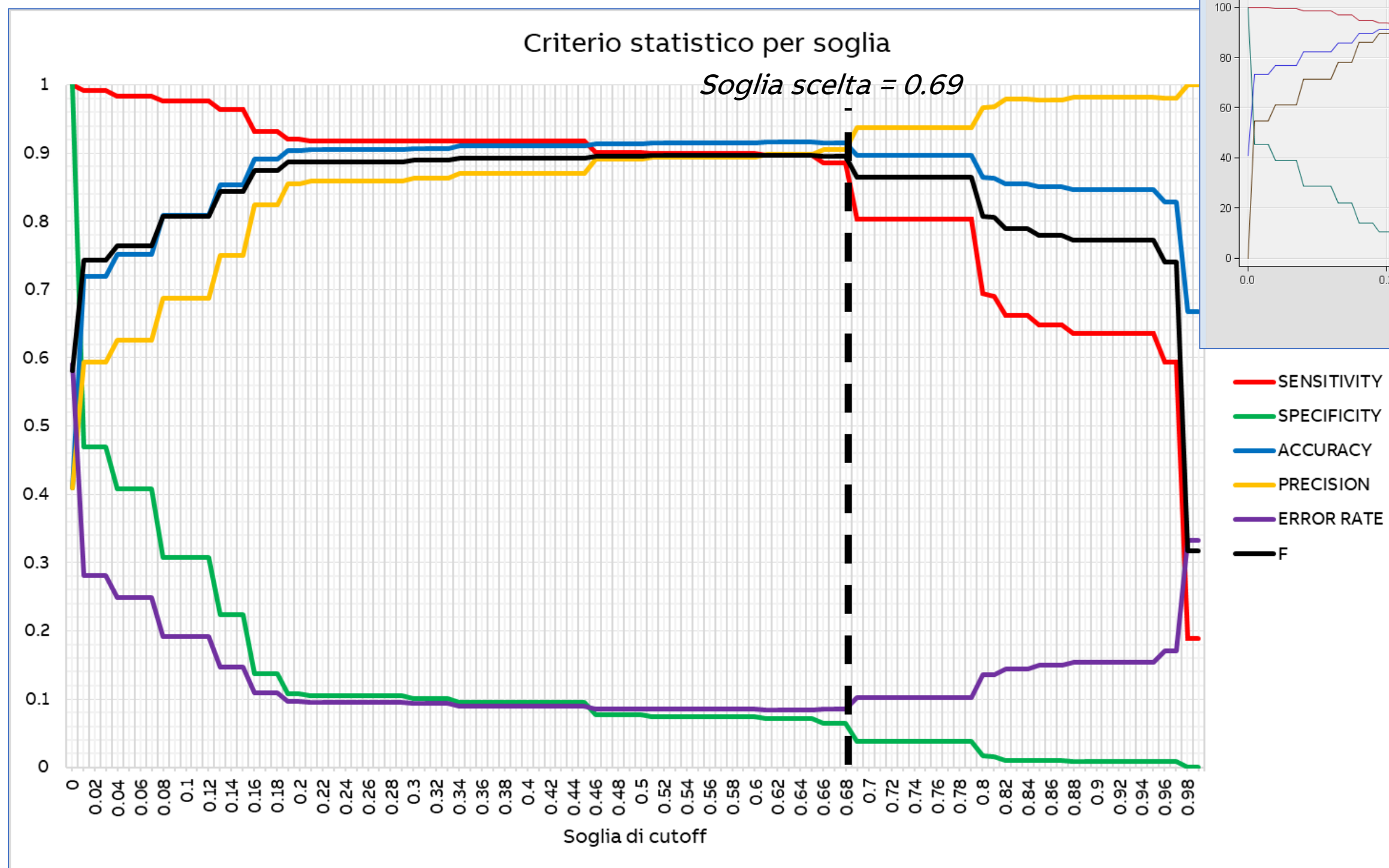
Statistiche di stima

| Modello | Nodo | Nodo | Descrizi | Variabile | Etichetta | Criterio | Train: | Train: | Train: | Train: | Train: | Train: | Train: | Valid: | Valid: | Valid: | Valid: | Valid: | Valid: | Valid: |
|---------|--------|---------|-----------------|-----------|-----------|---------------------------------|--------------------|---------------------------|-------------------------|-------------------------------|----------------------|---------------------------|------------------|--------------------|---------------------------|-------------------------|-------------------------------|----------------------|---------------------------|-------------------|
| seleto | ssore | modello | one del modello | target | target | di selezione: Train: Indice ROC | Somma di frequenze | Errore di classificazione | Errore assoluto massimo | Somma degli errori quadratici | Average Square Error | Root Average Square Error | Divisore per ASE | Somma di frequenze | Errore di classificazione | Errore assoluto massimo | Somma degli errori quadratici | Average Square Error | Root Average Square Error | Divisore per VASE |
| Y | Tree2 | Tree2 | Albero ... | target10 | target10 | 0.971 | 4743 | 0.0824... | 0.99827 | 598.58... | 0.0631... | 0.2512 | 9486 | 4743 | 0.0859... | 1 | 272.31... | 0.0669... | 0.2586... | 4070 |
| | Req2 | Req2 | Reares... | target10 | target10 | 0.938 | 4743 | 0.13852 | 0.9991... | 927.69... | 0.0977... | 0.3127... | 9486 | 4743 | 0.15774 | 1 | 459.53... | 0.1129... | 0.3360... | 4070 |
| | Boost | Boost | Gradie... | target10 | target10 | 0.935 | 4743 | 0.1380... | 0.9560... | 959.72... | 0.1011... | 0.3180... | 9486 | 4743 | 0.1538... | 0.9800... | 435.34... | 0.1069... | 0.3270... | 4070 |
| | Neural | Neural | Rete n... | target10 | target10 | 0.922 | 4743 | 0.1640... | 0.99747 | 1050.8... | 0.1107... | 0.3328... | 9486 | 4743 | 0.1941... | 0.9942... | 524.69... | 0.1289... | 0.3590... | 4070 |
| | Req3 | Req3 | Reares... | target10 | target10 | 0.905 | 4743 | 0.1773... | 0.9948... | 1153.8... | 0.1216... | 0.3487... | 9486 | 4743 | 0.1847... | 0.9999... | 525.95... | 0.1292... | 0.3594... | 4070 |



Scelta del cutoff

- Non avendo una matrice dei costi/profitti, è stato individuata la soglia in base al criterio statistico sul dataset di validation
- Nel nostro caso abbiamo voluto massimizzare la precision e sensitivity (ossia F)
- Abbiamo impostato $\text{presion} = \text{recall}$ nel nodo di cutoff in SAS, ed esportato i dati di output delle soglie plottando le statistiche di classificazione al variare della soglia di cutoff comprendendo anche F ed error rate.





Scoring

- Il dataset di scoring è stato scaricato tramite scraping due/tre settimane più tardi del dataset iniziale
- Abbiamo fatto preprocessing fuori da sas con le esatte stesse modalità del dataset di training/validation
- Abbiamo dunque scritto del codice sas per applicare il cutoff sul dataset di scoring, come riportato negli screenshot accanto
- Decisione rappresenta la classificazione di scoring, questo dato è stato esportato da sas e salvato come xls nel dataset di partenza (dimostrazione). Questo ci consente di inviarlo ad utenti con il link all'annuncio per contattare il venditore



| Codice di training | | | |
|--|--------|-------------|-----------|
| <pre> DATA goodcars; SET &EM_IMPORT_SCORE; obsnum=_N_; IF P_target101 >0.69 THEN decisione=1; else decisione=0 ; run; PROC PRINT data= goodcars; VAR obsnum P_target101 decisione; LABEL P_target101='Predicted*target101=Cheap*====='; TITLE "KMO Cheap Cars"; run; data em_goodcars ; set work.goodcars; run; </pre> | | | |
| KMO Cheap Cars | | | |
| Oss | obsnum | P_target101 | decisione |
| 1 | 1 | 0.00173 | 0 |
| 2 | 2 | 1.00000 | 1 |
| 3 | 3 | 1.00000 | 1 |
| 4 | 4 | 1.00000 | 1 |
| 5 | 5 | 0.00173 | 0 |
| 6 | 6 | 0.00173 | 0 |
| 7 | 7 | 0.00173 | 0 |
| 8 | 8 | 0.00173 | 0 |
| 9 | 9 | 0.00173 | 0 |
| 10 | 10 | 0.00173 | 0 |
| 11 | 11 | 0.97066 | 1 |
| 12 | 12 | 0.00173 | 0 |
| 13 | 13 | 0.00173 | 0 |
| 14 | 14 | 0.12963 | 0 |
| 15 | 15 | 0.97066 | 1 |
| 16 | 16 | 0.97066 | 1 |
| 17 | 17 | 0.97066 | 1 |
| 18 | 18 | 0.97066 | 1 |
| 19 | 19 | 0.97066 | 1 |
| 20 | 20 | 0.12963 | 0 |
| 21 | 21 | 0.02146 | 0 |
| 22 | 22 | 0.00000 | 0 |
| 23 | 23 | 0.97066 | 1 |
| 24 | 24 | 0.12963 | 0 |
| 25 | 25 | 0.97066 | 1 |
| 26 | 26 | 0.15789 | 0 |
| 27 | 27 | 0.80000 | 1 |
| 28 | 28 | 0.00173 | 0 |
| 29 | 29 | 0.12963 | 0 |
| 30 | 30 | 0.02146 | 0 |