Report SUS5

Team - SC1

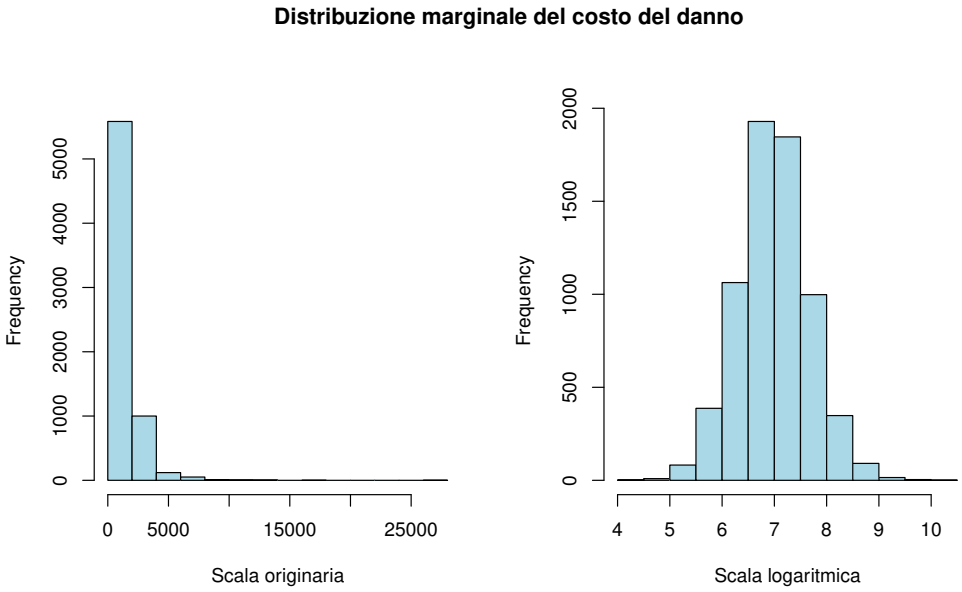
18 June 2019

1 Analisi preliminari

Le operazioni di pulizia del dataset hanno posto l'attenzione sui seguenti aspetti:

- rimozione di "weight", perché identicamente uguale a 1
- rimozione di "Province_code"
- introduzione di un nuovo predittore: la differenza tra la data dell'incidente e la data di inizio pratica, in giorni trascorsi (quantitativa)
- introduzione di un nuovo predittore rappresentante la differenza tra la data dell'incidente e la data di immatricolazione del veicolo, sotto ipotesi che un veicolo più recente ha un maggiore valore economico (quantitativa)
- introduzione di un nuovo predittore, "fault", che a partire da "Bareme_table_code_customer" e "Bareme_table_code_other_driver", rappresenta se il cliente ha ragione ("R") o torto("W"), se la colpa è condivisa ("S") o se non è verificabile ("NV"). (qualitativa)
- introduzione di un nuovo predittore, "resp", che incrocia i valori di "Bareme_table_code_customer" e "Bareme_table_code_other_driver" (qualitativa)
- valori mancanti: le variabili con un numero contenuto di valori mancanti sono stati imputati con la mediana (per le variabili quantitative) e con la moda (per le variabili qualitative). Per un numero di dati mancanti elevato (maggiore di 500) non si è considerata la variabile

Inoltre abbiamo condotto l'analisi considerando la trasformata logaritmica della variabile risposta per renderla simmetrica.



2 Modelli

Essendo l'accuratezza misurata in termini di errore medio assoluto, abbiamo optato per adattare ai dati modelli che tenessero conto di questa caratteristica:

1. Foresta Casuale Quantilica
2. Boosting con funzione di perdita di pseudo-Huber

$$L(a) = \delta^2(\sqrt{1 + (a/\delta)^2} - 1), \quad a = \hat{y} - y$$

3. Boosting con funzione di perdita quantilica

$$L(a) = \begin{cases} \tau a \\ (\tau - 1)a \end{cases}, \quad \tau \text{ quantile e } a = \hat{y} - y$$

La proposta numero 2 è risultata essere la migliore; per implementarla abbiamo usato il pacchetto Xgboost di R, modificando manualmente la funzione di perdita. Inoltre, per la scelta del parametro δ abbiamo scelto con una procedura di stima-verifica il valore finale pari a 3.

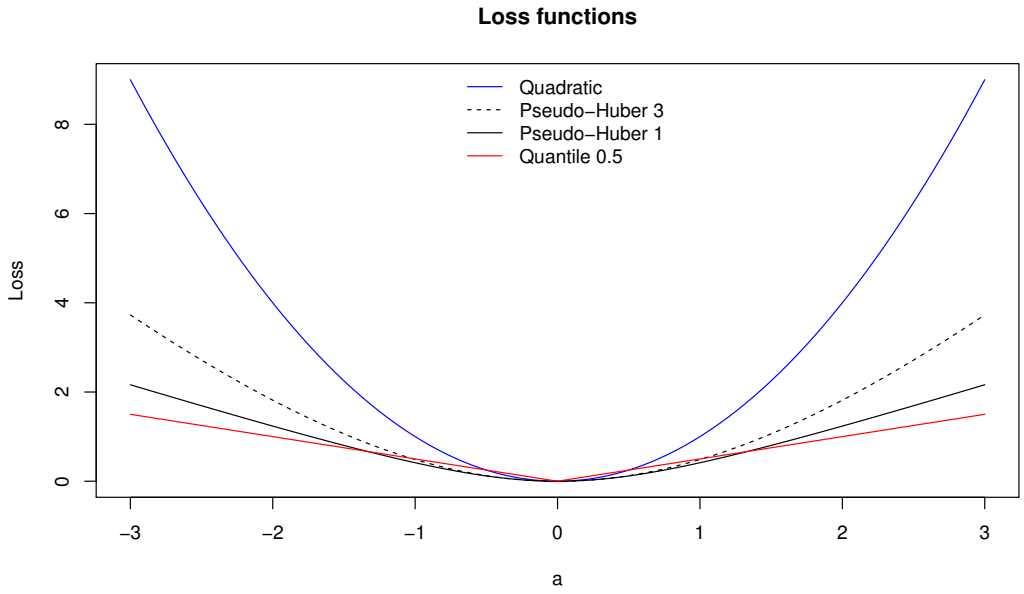


Figure 1: Come si nota dalla figura, l'utilizzo della solita funzione di perdita quadratica porterebbe a grosse perdite in accuratezza per valori grandi dell'errore di previsione; la soluzione che minimizzerebbe la perdita in accuratezza sarebbe utilizzare la funzione quantilica, ma per problemi di differenziabilità in zero abbiamo scelto un'approssimazione, cioè la funzione di pseudo-huber. All'aumentare del parametro di regolazione δ con pseudo-huber mi avvicino alla funzione di perdita quadratica. Quindi per fissarlo abbiamo portato avanti una procedura di regolazione basata su insiemi di stima e verifica.