Data Mining A.A. 2023-2024

Progetto Finale

NYC Taxi Tip Prediction

Autore: Luca De Simone

Indice

---

1. Introduzione

L'obiettivo dell'analisi consiste nel prevedere l'ammontare della mancia dei tassisti della città di New York. I dati sono tratti dalla NYC Taxi and Limousine Commission e filtrati per contenere le sole transazioni effettuate con carta di credito nel maggio 2015. Il campione in studio è costituito da 243179 osservazioni descritte da 21 predittori, di cui sei qualitativi e quindici quantitativi.

2. Analisi Zero Variance e Codifica delle Variabili

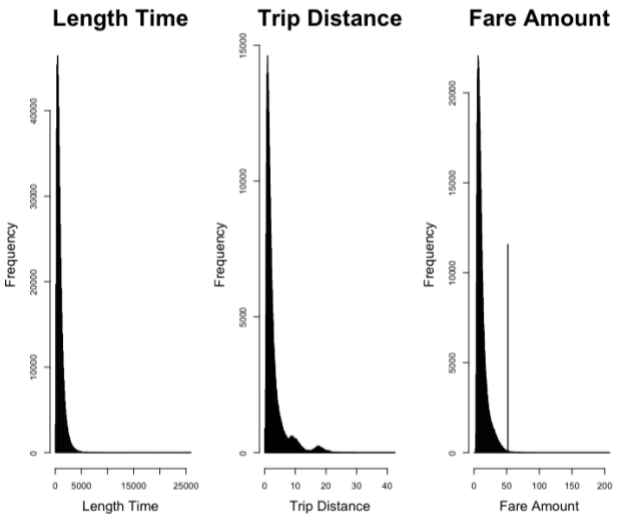
In primo luogo, è stata eseguita un'analisi esplorativa dei dati al fine di identificare la natura delle variabili a disposizione e la presenza di eventuali anomalie. Le variabili `pickup_week` e `pickup_wday` sono state trasformate in factors e ne si è analizzata la coerenza temporale incrociandole con `pickup_doy`. Da questo approfondimento è emerso che al valore '1' di `pickup_wday` corrisponde a venerdì e quindi, per una maggior facilità interpretativa, si è deciso di passare da una codifica numerica ad una nominale. Inoltre, si è scelto di escludere dallo studio `pickup_doy` in quanto ridondante e di dettaglio informativo inferiore rispetto `pickup_week` e `pickup_wday`. Successivamente anche `pickup_month` verrà rimossa in quanto degenere.

Tramite il sito governativo dello stato di New York è stato possibile reperire della documentazione ufficiale riguardante le tariffe orarie degli yellow cab. Da questa è emerso che le tratte effettuate fascia notturna, dalle 8pm alle 6am, e in orario di punta, dalle 4pm alle 8pm, sono sottoposte a supplemento. Per tale ragione si è deciso di ricodificare la variabile `pickup_hour` accorpandone i livelli al fine di massimizzarne l'apporto informativo e ridurre il numero di parametri in fase di stima. Inoltre, la covariata è stata rinominata in `timeslots` e il raggruppamento ha prodotto i seguenti livelli: from 6am to 12am → 'morning time', from 12am to 4pm → 'afternoon time', from 4pm to 8pm → 'rush time' e from 8pm to 6am → 'night time'.

Le variabili `pickup_NTACode` e `dropoff_NTACode` risultano essere maggiormente informative delle rispettive `pickup_BoroCode` e `dropoff_BoroCode` in quanto descrivono il dato in maniera capillare. D'altra parte, però, senza apportare alcun tipo di modifica alla loro codifica, se utilizzate in fase di stima, produrrebbero dei modelli estremamente variabili a causa dell'elevata numerosità dei parametri necessari a stimare il loro effetto. Dunque, si è deciso di creare delle nuove variabili, rispettivamente `pickup_Area` e `dropoff_Area`, in cui i codici NTA sono stati aggregati al fine di creare delle macroaree geografiche più informative dei BoroCode. Questo ha permesso di suddividere il distretto di Manhattan in 'Manhattan_Facilities', 'Lower Manhattan', 'Middle-Lower Manhattan', 'Middle-Upper Manhattan' e 'Upper Manhattan' e il distretto di Brooklyn in 'Northern Brooklyn' e 'Southern New York', avendogli aggregato insieme il distretto di Staten Island. Le altre aree territoriali sono state codificate esattamente con medesime le modalità dei BoroCode.

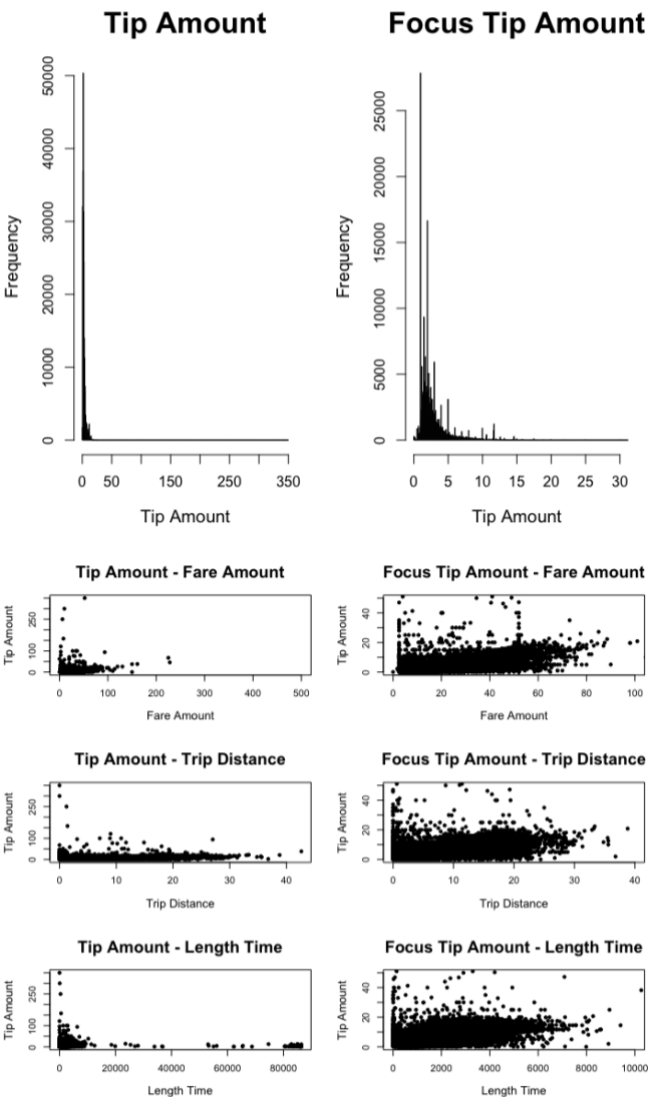
3. Analisi Grafiche

I grafici riportati evidenziano come le covariate `length_time`, `trip_distance`, e `fare_amount` presentino delle distribuzioni fortemente asimmetriche positive. Sulle code di destra, è possibile notare, per tutte le variabili, una forte presenza di punti outliers che ne aumentano notevolmente il range di variazione. Questi verranno analizzati nello specifico in un paragrafo successivo dedicato. Inoltre, per quanto riguarda la distribuzione di `fare_amount`, si nota un'importante sistematicità in prossimità del valore 52USD. Uno studio approfondito ha mostrato come tutte le tratte che coinvolgono il distretto di Manhattan e l'NTA Code QN98 siano affette da tale sistematicità. Il codice identifica gli aeroporti della città di New York, il 'JFK Airport' e il 'LaGuardia Airport', e tramite documentazione ufficiale è stato possibile risalire al



fatto che queste tratte presentano una tariffa fissa indipendente dalla direzione, dalla distanza e dal tempo di percorrenza. A tal proposito si è deciso di disgregare tale informazione e di costituire una variabile ad hoc di identifica per queste tratte:

$$\texttt{airport} = \begin{cases} 1 \ il\ soggetto\ ha\ percorso\ una\ tratta\ aeroportuale \\ 0 \ altrimenti \end{cases}$$



La distribuzione della dipendente `tip_amount` è fortemente asimmetrica e anch'essa è affetta, sulla coda di destra, da una massiccia presenza di outliers che ne aumentano il range di variabilità. Ristringendo il campo di variazione, è possibile osservare la presenza di forti sistematicità in corrispondenza di alcuni valori delle mance. Infatti, è possibile notare che questi picchi sono in corrispondenza dei valori 1USD, 2USD e 5USD, e leggendo la documentazione ufficiale è emerso che esistono delle mance fisse in funzione di alcuni importi. Tale affermazione è possibile riscontrarla anche a livello visivo, infatti, dai grafichi sotto riportati, si evidenziano sia delle forti sistematicità a fronte di alcuni importi e sia un notevole legame tra `fare_amount` e `tip_amount` che, a causa degli outliers, risulta essere parzialmente offuscato. I plot sono stati ottenuti prima in scala originale e poi riducendo i domini delle covariate di riferimento al fine di evidenziare i legami che altrimenti risulterebbero mascherati. Tra tutte le esplicative e `tip_amount` sembrerebbe esserci una relazione di tipo lineare e questa sembrerebbe essere dovuta al fatto che, generalmente, la quota di mancia lasciata da un individuo è circa pari al 15,20,25 % della tariffa totale, informazione ricavata tramite consulto della documentazione ufficiale. Inoltre, è noto che il prezzo della corsa è definito dal tempo e dalla distanza percorsa, dunque, essendo `length_time` e `trip_distance` correlate positivamente con `fare_amount`, anch'esse risultano legate linearmente con `tip_amount`.

4. Analisi Casi Anomali e punti Outliers

Nel precedente paragrafo, i grafici di `length_time`, `trip_distance`, e `fare_amount` hanno fatto emergere diverse anomalie, le cui principali sono:

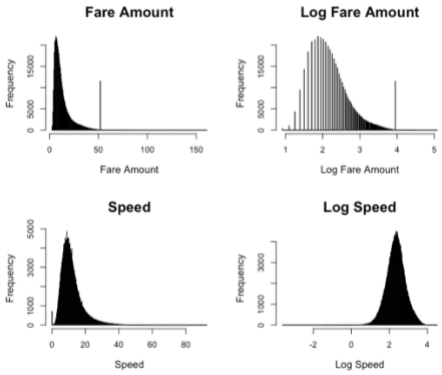
- Durate di alcune corse superiori a 6000 secondi, superiori alle 16 ore, dato abbastanza illogico contestualizzato alla città di New York e al fatto che, legalmente, il numero massimo di ore che un tassista può effettuare è pari a 12.
- Distanze percorse diverse da zero in corrispondenza di durate nulle e viceversa.
- Tariffe minori di 2.5USD. La documentazione ufficiale afferma che le tariffe hanno un addebito iniziale di 2.5USD; dunque, il costo totale non può essere inferiore di tale importo.

Per compiere un'analisi più completa ed approfondita per identificare eventuali incongruenze spazio-temporali si è deciso di computare la covariata `geodetic_distance`, al fine di sfruttare l'informazione contenuta all'interno di `pickup_longitude`, `pickup_latitude`, `dropoff_longitude` e `dropoff_latitude`. Procedendo con l'analisi sono emerse diverse anomalie ed errori computazionali difficili da rilevare caso per caso e per questo motivo si è deciso di introdurre un'ulteriore variabile di controllo:

$$\texttt{speed} = \texttt{trip\_distance} \,/\, (\texttt{length\_time} \,/\, 3600)$$

Questa ha permesso di identificare con maggior facilità ulteriori valori anomali, ovvero osservazioni riportanti una velocità superiore alle 100 miglia/oraria. Successivamente, tale variabile verrà utilizzata in fase di stima

perché riassuntiva dell'apporto informativo di `length_time` e `trip_distance` e più funzionale poiché potrebbe essere un segnalatore, seppur molto grezzo, della presenza di traffico. I casi appena citati sono stati poi imputati in modo robusto tramite alberi di regressione.



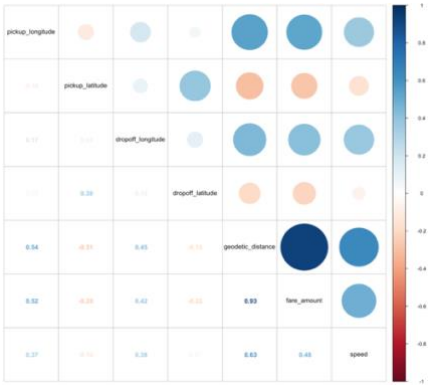
Gli outliers sono stati gestiti in due modi:

1. Tramite opportune trasformazioni delle covariate `speed` e `fare_amount`. In particolare, si è optato per una trasformazione logaritmica in modo da ridurre il range di variazione al fine di smorzare l'effetto di eventuali outlier e per eliminare, quanto possibile, le forti asimmetrie per massimizzarne la linearità con la dipendente. Inoltre, la trasformata facilita l'interpretazione dell'effetto (elasticità).
2. Applicazione, in fase di stima della Robust Regression, uno strumento regressivo in grado di gestire in modo robusto i valori outliers, ridurne l'impatto e l'influenza sui coefficienti. La RR consiste nel minimizzare una generica funzione di perdita, funzione dei residui, assegnando ad ognuno di essi un peso, tramite l'ausilio di uno stimatore IRLS (Iteratively Reweighted Least Squares).
   L'algoritmo IRLS procede nel seguente modo:
   - Viene inizializzato penalizzando una determinata funzione di perdita
   - Vengono stimati i residui
   - Ad ogni residuo viene assegnata una funzione peso
   - Vengono stimati i coefficienti di regressione utilizzando il metodo WLS:
     $$\beta = (X^T W X)^{-1} X^T W Y$$
   - I passaggi vengono ripetuti fino a convergenza

   In fase di stima si possono utilizzare differenti funzioni di perdita con le rispettive funzioni peso, come ad esempio quella di Huber, Tuckey e Assoluta. Tra le tre metodologie si è optato per utilizzare una LAR (Least Absoulte Regression) con funzione di perdita $\rho(\varepsilon) = |Y - X\beta|$ e pesi $w(\varepsilon) = \frac{1}{|\varepsilon|}$.
   Questo perché il modello Huber presenta una funzione peso che, per costruzione, non è in grado di penalizzare forti punti oultiers, il Tuckey, al contrario, risulta essere troppo penalizzante annullando completamente l'effetto dei punti outliers elevati assegnandogli peso nullo. Infatti, il nostro obiettivo non è quello di annullare l'apporto informativo dei outliers ma di moderarne l'effetto. L'unico limite della RR è la presenza di un numero cospicuo di punti ad alto leverage che lo renderebbe inefficiente. Per questo motivo verrà condotta, in fase di stima, un'analisi specifica, che però ha dato esito negativo.

In ultimo luogo, anche la dipendente `tip_amount` è stata trasformata in scala logaritmica al fine di ridurne il range di variabilità, renderla la più simmetrica possibile e, essendo una quantità naturalmente positiva, per assicurare, in fase di stima, previsioni tutte positive.

5. Fase di Stima e Conclusioni

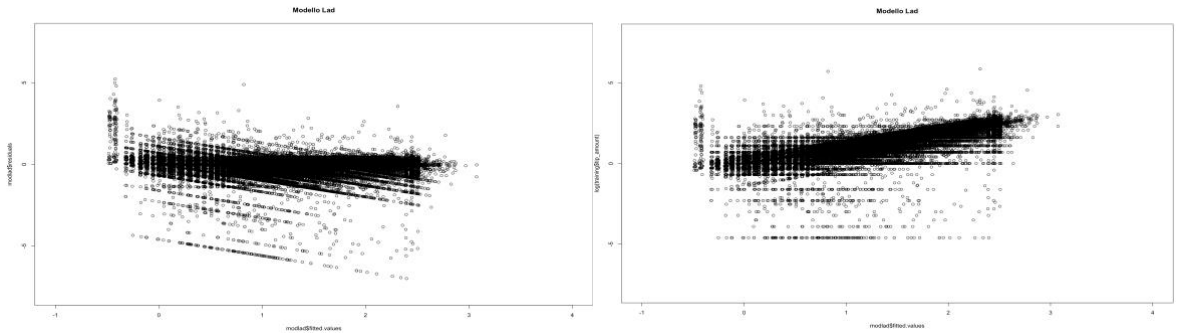


L'analisi delle correlazioni e delle connessioni tra esplicative consente di valutare la forza e la direzione dei legami presenti tra queste. Il correlogramma non evidenza forti correlazioni tra le variabili, se non tra `fare_amount` e `geodetic_distance`, pressoché collineari. Dallo studio si evince che l'applicazione di modelli PCR e la Ridge Regression potrebbero risultare non ottimali in questo contesto perché particolarmente performanti in presenza di variabili molto correlate tra loro, risultato confermato in fase di stima. Anche la tabella del chi-quadrato tra le qualitative non fa emerge nulla di rilevante, se non una lieve connessione tra le variabili `pickup_Area`, `dropoff_Area` e `airport`.

La fase esplorativa appena conclusa ha permesso di selezionare il seguente set di esplicative da utilizzare in fase di stima: `speed`, `timeslots`, `pickup_wday`, `airport`, `pickup_longitude`, `pickup_latitude`,

`dropoff_longitude`, `dropoff_latitude`, `vendor_id`, `passenger_count`, `fare_amount`, `pickup_Area` e `dropoff_Area`.

Per prima cosa è stato stimato un modello lineare completo su cui si è condotta un'analisi dei punti influenti sul modello per valutare l'efficienza della RR. I residui del modello completo LAD, come osservabile, non si disperdono in modo casuale a causa della presenza di mance fisse, singolarità evidenziata in precedenza. Inoltre, emerge come il modello fittato sia troppo ottimistico, producendo un'asimmetria della distribuzione dei residui.



Per definire al meglio le forme funzionali delle esplicative da inserire in fase di stima si sono utilizzati i modelli GAM che, però, non hanno evidenziato forme funzionali rilevanti. I modelli fittati che hanno generato le migliori performance previsive sono:

Forward Stepwise Regression cross validata con pesi LAD: Al modello lineare completo, pesato tramite pesi LAD, si è operata una Robust forward regression cross-validata con funzione di perdita MAE, al fine di individuare il set ottimale di predittori da includere nel modello in grado di minimizzare la funzione di perdita assoluta. Le covariate ottimali risultano essere le seguenti: `log(fare_amount)`, `airport`, `timeslots`, `pickup_longitude`, `pickup_latitude`, `pickup_Area`. Inoltre, si è inserito nel modello finale l'interazione tra `pickup_longitude` e `pickup_latitude`. Tale procedura ha permesso di ridurre significativamente l'errore di previsione.

Conseguentemente, si è valutato l'utilizzo di modelli di tipo Signal Sparsity, principalmente perché si ha ragion di credere, avendo osservato il correlogramma e i grafici delle distribuzioni delle covariate incrociate con il target, che l'informazione risiede solo in poche esplicative.

Lasso con pesi LAD: La scelta della griglia dei valori di lambda iniziali, è stata effettuata tramite varie simulazioni. Il valore di lambda ottimale ottenuto tramite cross-validation risulta essere pari a 0.04151011. Tale valore indica che la penalizzazione, e di conseguenza il grado di distorsione inserito, è molto basso.

Elastic Net con pesi LAD: Il valore di lambda ottimale ottenuto tramite cross-validation risulta essere pari a 0.04366043. Anche in questo caso il parametro di penalizzazione è prossimo al valore nullo.

Gam: I modelli GAM sono dei metodi semi-parametrici, utilizzati non solo al fine di identificare le trasformazioni ottimali delle variabili esplicative, ma anche come veri e propri modelli previsivi. In questo caso, l'analisi degli andamenti funzionali delle esplicative, non segnala particolari forme funzionali. Le relazioni tra le esplicative e la dipendente sono di tipo lineare, dunque le performance previsive del modello risultano essere molto simili a quelle del modello lineare pesato

METODOLOGIA	MAE
Forward Stepwise regression cross validata con pesi LAD	0.5954
Lasso con pesi LAD	0.6864
Elastic Net pesato LAD	0.6680
GAM	0.5954

In conclusione, il modello ottimale al fine di prevedere la mancia dei tassisti di NYC sulla base delle caratteristiche in esame risulta essere la Forward Stepwise Regression, applicata su un modello lineare con pesi robusti.

La documentazione presa come riferimento è la seguente: https://www.nyc.gov/site/tlc/passengers/taxi-fare.page