PREDICTION TAXI TIP Capozio Sara Matricola:853600

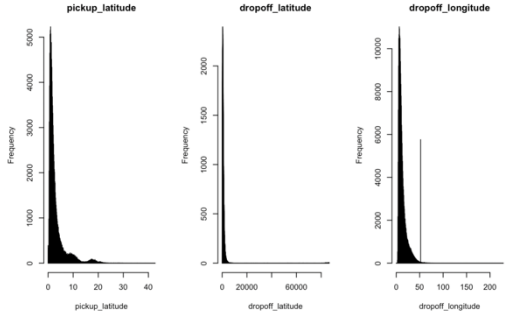
ANALISI GENERALE DEL DATASET

Il dataset in esame complessivamente comprende 243179 osservazioni e 21 predittori, di cui 6 variabili qualitative e le restanti 15 quantitative. In primo luogo, è stata svolta un'attenta fase di pre-processing col fine di identificare la natura delle variabili in esame e la presenza di eventuali anomalie all'interno dei dati.
Dall'analisi è stata esclusa, oltre alla variabile identificativo, l'unica variabile affetta da zero-variance ovvero la variabile 'PICKUP_MONTH' in quanto caratterizzata da un solo livello coincidente con il mese a cui corrispondono le rilevazioni delle osservazioni, ovvero il mese di maggio. Dopo aver unito i dataset di training e di test al fine di effettuare un pre-processing comune e si è svolta un'analisi sulle caratteristiche delle variabili qualitative e successivamente sulle variabili quantitative.

ANALISI DELLE VARIABILI QUALITATIVE:

- VARIABILI WDAY, WEEK, PICK_DOY: Dopo aver trasformato le variabili wday, week e hours come factors si è analizzata la coerenza tra le variabili temporali in modo da individuare eventuali errori di codifica. Tramite un'attenta analisi si è osservato che il livello '1' relativo alla variabile wday, che dovrebbe identificare il primo giorno della settimana, in realtà rappresenta il primo giorno della prima settimana del mese che coincide con venerdì. Si è deciso quindi di ricodificare i giorni della settimana dandogli la corretta codifica per lo più a fine interpretativo: (1=Venerdì, 2=Sabato, 3 = domenica ecc..).
  In seguito alla codifica della variabile 'wday', si è deciso di eliminare dall'analisi la variabile 'pick_doy' poiché combinazione delle variabili 'wday'' e 'week'. È stata effettuata tale scelta poiché a fini previsivi sarebbe più interessante andare a valutare se una tratta effettuata in un particolare orario o in particolari giorni della settimana incida significativamente sulla mancia ricevuta piuttosto che considerare i singoli giorni dell'anno.
- PICKUP_HOUR: La trasformazione di tale variabile è stata pensata inseguito alla lettura della documentazione relativa alla descrizione delle tariffe dei taxi fornita dal sito governativo di NYC.
  Per tutte le tratte erogate durante gli orari notturni (8 p.m to 6 p.m) e gli orari di punta (4 p.m to 8.pm) sembrerebbero esserci degli aumenti nelle tariffe rispettivamente di 1$ e 2.5$. Per tale ragione si è deciso di raggruppare i livelli di tale variabile in modo da riassumerne l'informazione seguendo un criterio oggettivo. Tramite raggruppamento si sono ridotti i livelli della variabile passando da 24 livelli a quattro livelli: 6:00- 12:00--> orario mattina, 12:00-16:00--> orario pomeriggio, 16:00-20:00-->orario di punta, 20:00- 6.00-->orario notturno.
- DROPOFF_NTA_CODE E PICKUP_NTA_CODE: Utilizzare le seguenti variabili in fase di stima, senza apportare alcun tipo di modifiche, genererebbe un modello estremamente complesso con un numero di paramenti pari al numero di livelli escluso il livello baseline. Le seguenti variabili risultano essere maggiormente informative rispetto alle variabili Boro Code in quanto contengono informazioni più capillari sulle tratte effettuate da ogni soggetto. Leggendo la documentazione, gli NTA CODE sono codici identificativi di quartieri che a loro volta appartengono a dei macro-distretti. Poiché Manhattan è la tratta più frequente nel nostro dataset, si è deciso di aggregare i suoi NTA in macroaree ovvero nei seguenti livelli:'Manhattan_Facilities','Lower Manhatta','Upper Manhattan' e 'Upper Mid Manhattan', mentendo invece per le altre aree territoriali le modalità di Boro Code.

ANALISI DELLE VARIABILI QUANTITATIVE:

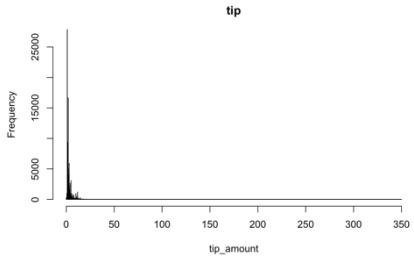


La distanza in miglia, il tempo di percorrenza e la quota di pagamento mostrano una distribuzione fortemente asimmetrica positiva con forti punti outliers presenti sulle code di destra che aumentano il range di variazione delle variabili in esame. Questi valori anomali verranno analizzati nel dettaglio successivamente. Da notare la presenza di una forte sistematicità nella distribuzione di 'fare_amount' in prossimità di una quota pari a 52$. In primo luogo, si sono analizzate le tipologie di tratte effettuate dai seguenti soggetti e si è riscontrato che il 20% di essi hanno percorso le seguenti tratte: QN98-MN17 e MN17-QN98. Dal raggruppamento dei livelli della variabile NTA_code è emerso che il codice

QN98 identifica soggetti che sono partiti o hanno raggiunto l'aeroporto John F. Kennedy o La Guardia situazioni nel Queens. Leggendo la documentazione fornita dal sito governativo di NYC, sembrerebbe che i taxi che percorrono le tratte aeroportuali presentino delle tariffe fisse che non dipendono dal tempo o dalla distanza per corsa. A tal proposito si è deciso di inserire nel modello una variabile dummy chiamata 'airport' costruita in prossimità di livelli di NTA_code coincidenti con 'QN98':

$$airport = \begin{cases} 0\ il\ soggetto\ non\ ha\ effettuato\ una\ tratta\ aeroportuale \\ 1\ l\ soggetto\ \ ha\ effettuato\ una\ tratta\ aeroportuale \end{cases}$$

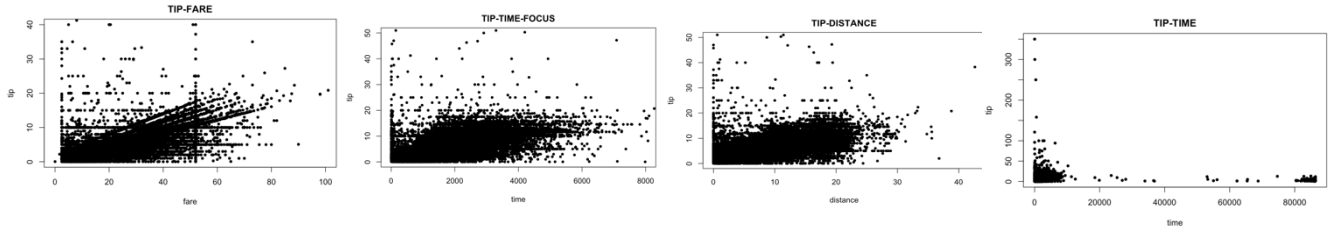
ANALISI DEL TARGET E ANALISI GRAFICHE



La distribuzione della variabile dipendente è fortemente asimmetrica. Importante notare la presenza di forti picchi in corripondenza di alcuni valori delle mance. I picchi sono in corrispondenza delle tre mance più comuni ovvero 1$, 2$e 1.5$. Generalmente secondo la documentazione se la tariffa risulta essere minore di 10$, si lascia una mancia non inferiore a un dollaro. Sembrerebbero esistere quindi delle mance fisse. Inoltre, non è un caso che tip_amount si distribuisca come fare_amount. Generalmente un soggetto lascia una mancia pari al 15,20,25 % della tariffa totale.

Tra le due variabili, infatti, esiste certa corrispondenza a livello di distribuzione e di picchi che le caratterizzano.

Si analizza ora la relazione tra la variabile dipendente e le variabili esplicative tramite delle analisi grafiche in modo da individuare a priori le variabili con maggior potere esplicativo.



I seguenti plot sono stati ottenuti restringendo il dominio e codominio delle variabili di riferimento. Questo perché la forte presenza di punti outliers tende a coprire la vera relazione tra la mancia e le tre variabili in esame. Tra le tre covariate e tip sembrerebbe esserci una relazione lineare crescente che viene mascherata da effetti sistematici dovuti alla presenza di quote e mance fisse. La variabile length_time in particolare presenza un gruppo di punti outliers significativo che impatta fortemente sull'andamento distributivo di lengh_time. Sono presenti, infatti, soggetti con lenght_time superiore a 60000 secondi = 16 ore. Caso anomalo, in quanto leggendo la documentazione il numero massimo di ore che può effettuare un tassista è pari a 12 ore. Analizzando la relazione tra le variabili qualitative e la variabile target non sembrerebbero esserci delle variabili particolarmente discriminanti.

OUTLIER DETECTION E POSSIBILI SOLUZIONI

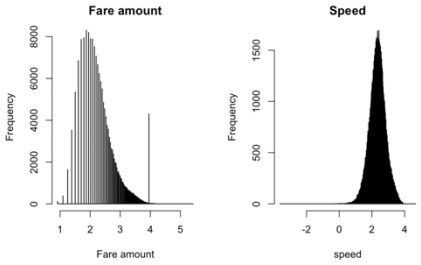
Sono stati individuati i seguenti valori anomali:
- valori di lenght time >= 50000 (13 ore)
- valori di length_time nulli con distanza positiva valori delle longitudini e latitudini diverse andata-ritorno
- valori di distanza non nulla con coordinate di latitudine e longitudine andata-ritorno uguali e fare=2.50: queste osservazioni presentano degli errori computazionali in quanto se il soggetto ha coordinate uguali e presenta una tariffa pari a 2.50 che corrisponde alla tariffa di attivazione, allora la distanza deve essere nulla e non diversa da zero.
- valori di fare_amount <2.5, in quanto leggendo la documentazione il costo di attivazione di una tratta non può essere inferiore ai 2.5$

Procedendo con l'analisi sono stati identificati altri valori anomali ed errori computazionali difficili da analizzare caso per caso. Per risolvere tale problematica si è deciso di introdurre una variabile controllo:

$$speed = distance/(time/3600)$$

In questo caso costruendo una nuova misura riassuntiva si è stati in grado di identificare con maggior facilità valori anomali aggiuntivi di soggetti che hanno viaggiato ad una velocità superiore alle 120 miglia/oraria. Nell'analisi verrà mantenuta questa variabile al posto di trip_distance e length_time poiché riassuntiva del loro apporto informativo. I casi appena elencati sono stati imputati in modo robusto utilizzando gli alberi di regressione. Gli outliers sono stati gestiti in due modi:

1. Introducendo opportune trasformazioni alle variabili fare e speed.



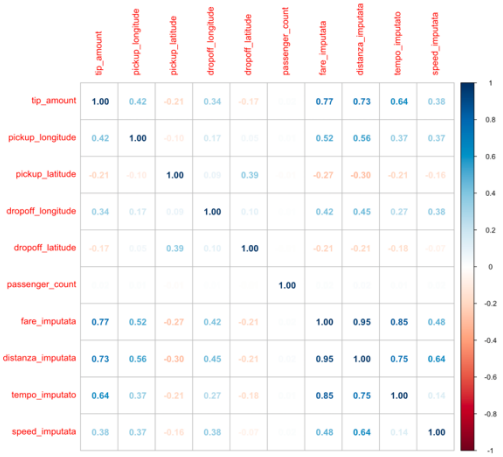
Si è optato per la trasformazione logaritmica in quanto, oltre ad essere facilmente interpretabile in termini di incrementi percentuali, essa risulta ottimale nella riduzione del range di variabilità delle variabili in esame. Comprimendo la distribuzione e rendendola più simmetrica, l'effetto di quei punti che erano presenti sulle code positive in questo modo è meno impattante.

2. Applicazione della Robust regression, famiglia di strumenti di regressione che gestiscono in modo robusto gli outliers. Tale metodologia consiste nel minimizzare una nuova funzione di perdita che è funzione dei residui, assegnando ad ognuno di essi una funzione peso che tende a minimizzare quei valori con magnitudine maggiore tramite l'utilizzo di uno stimatore IRLS (Iteratively reweighted least squares). Riassumendo l'algoritmo IRLS procede nel seguente modo:
   - Esso 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: $b = (X'WX)^{-1}X'Wy$
   - Ripetere l'algoritmo fino a convergenza

   In fase di stima si possono utilizzare differenti funzioni di perdita con le rispettive funzioni peso, come ad esempio il metodo Huber, Tuckey e Lar. Tra le tre metodologie si è optato per utilizzare il modello LAR con funzione di perdita $\rho(e) = |e|$ e funzione pesi $w(e) = \frac{1}{|e|}$.

   Questo perché il modello Huber presenta una funzione peso che per costruzione non in grado di penalizzare forti punti oultiers, mentre il modello Tuckey, al contrario, risulta essere troppo penalizzante annullando completamente l'effetto dei punti outliers elevati a cui vengono assegnati peso nullo. L'obiettivo fissato non è quello di annullare l'apporto informativo dei punti outliers ma di moderarne l'effetto.

MODELLISTICA



Tramite l'analisi del correlogramma risulta possibile valutare se le variabili quantitative in esame sono correlate con la variabile dipendente e se sono presenti eventuali casi di collinearità. In generale le variabili non risultano molto correlate tra loro, fatta eccezione per 'distance', 'time' e 'fare' che presentano un forte legame lineare in quanto quest'ultima è costruita in loro funzione. Già dall'analisi delle correlazioni si evince che l'applicazione di modelli come le componenti e la ridge regression potrebbero non essere ottimali in quanto essi risultano particolarmente performanti in presenza di variabili molto correlate tra loro, risultato confermato in fase di stima. Si è dunque deciso di modellare il logaritmo di tip_amount in modo da assicurarci che le sue previsioni non assumano mai valore negativi. Le variabili complessive mantenute in esame risultano essere: log(fare_imputata),pickup_hour,pickup_week,pickup_wday,pickup_longitude,pickup_latitude , dropoff_longitude, dropoff_latitude, vendor_id , airport , NTA_pickup_zone , NTA_droppof_zone, log(speed_imputata).

In primo luogo, è stato stimato un modello lineare completo su cui si è condotta un'analisi dei punti influenti sul modello. Step necessario in quanto la regressione robusta non è efficiente in presenza di punti di leva o punti influenti ma nel nostro caso tale problematica non è stata riscontrata.
Successivamente sono state identificate tramite l'applicazione di modelli GAM le forme funzionali ottimali che esprimono l'andamento delle variabili esplicative in funzione della variabile risposta.
Le splines univariate stimate non presentano particolari andamenti funzionali per cui si è deciso di non effettuare alcun tipo di trasformazione sulle esplicative. I modelli fittati che hanno generato le migliori performance previsive sono:

1.Forward Stepwise Regression cross validata con pesi LAD:
Tale modello è stato applicato sul modello lineare completo pesato tramite pesi LAD ed è stato implementato al fine di individuare il modello ottimale in grado di minimizzare il MAE cross-validato. Il risultato ottenuto risulta essere più parsimonioso del modello completo con un numero di covariate totali pari a nove: log(fare_imputata), airport, pickup_hour, NTA_pickup_zone, pickup_latitude , pickup_wday , vendor_id, NTA_droppof_zone, pickup_longitude. Una volta identificata la complessità ottimale del modello, sono state introdotte delle interazioni tra le variabili 'longitudine' e 'latitudine 'e tra le variabili 'orari' e 'giorni della settimana'. Tale procedura ha permesso di ridurre significativamente l'errore di previsione sul dataset di validation. Nella fase della Stepwise le prime variabili selezionate sono state: 'fare_amount' e 'airport'.

2.Lasso con pesi LAD: La scelta della griglia dei valori di lambda iniziali, è stata effettuata tramite varie simulazioni. Il valore di labda ottimale ottenuto tramite cross-validation risulta essere pari a 0.0356. Tale valore indica che la penalizzazione, e di conseguenza il grado di distorsione inserito è molto basso. Il modello lasso presenta delle performance previsive non ottimali questo perché, con un valore di lambda pressoché nullo, viene modellato un modello lineare completo pesato senza ridurre la magnitudine dei coefficienti di regressione. Analizzando il percorso Lasso effettuato da ciascuna variabile, anche in questo caso si è osservato che 'fare_amount' è la variabile che contribuisce maggiormente alla previsione di tip_amount in quanto il suo coefficiente ha una velocità di convergenza più lenta rispetto alle altre variabili esplicative.

3.Elastic Net con pesi LAD: Nel caso dell'Elastic Net il grado di distorsione ottimale cross-validato inserito nel modello risulta essere maggiore rispetto a quello del lasso con un valore di lambda pari a 0.05. Anche in questo caso il parametro di penalizzazione è prossimo al valore nullo.

4.Gam: I modelli GAM, metodi semi-parametrici, sono stati utilizzati non solo in fase iniziale al fine di identificare le trasformazioni ottimali delle variabili esplicative ma anche come modello previsivo utilizzando anche in questo caso i pesi LAD. In questo caso le covariate utilizzate sono state quelle identificate dalla forward regression. Come atteso, le sue performance previsive risultano molto simili rispetto al modello lineare pesato, questo perché dall'analisi degli andamenti funzionali, essi sembrerebbero di tipo lineare.

Modelli fittati	MAE Validation
Forward Stepwise regression cross validata con pesi LAD	0.590
Lasso con pesi LAD	0.633
Elastic Net pesato LAD	0.640
Modelli GAM	0.591

In conclusione, il modello migliore ottenuto al fine di prevedere la mancia dei tassisti di NYC sulla base di particolari caratteristiche delle corse effettuate risulta essere la Forward Stepwise Regression applicata su un modello lineare con pesi robusti e le variabili che hanno contribuito maggiormente a migliorarne le performance previsive sono: log(fare_imputata), airport, pickup_hour*pickup_wday, NTA_pickup_zone, vendor_id, NTA_droppof_zone, pickup_longitude*pickup_latitude.

Il link della documentazione presa come riferimento è il seguente:
https://www.nyc.gov/site/tlc/passengers/taxi-fare.page