INDIVIDUAL PROJECT: LORENZO GIUDICI, 850832, APPELLO D'ESAME DEL 21/11/2023

Scopo di questo lavoro è prevedere l'importo delle mance (*tip_amount*) relative ad un insieme di corse di taxi nella città di New York. Si ha a disposizione un training set, contenente 243.179 osservazioni e 21 variabili (tra le quali la variabile risposta *tip_amount*) e un test set, contenente 243.180 osservazioni e le 20 variabili esplicative, su cui effettuare le previsioni finali.

1. Analisi esplorativa e pre-processing

Una volta importati i dati ci si è resi immediatamente conto che due delle variabili (*ID* e *pickup_month*) non sono di alcuna utilità per l'analisi: la prima è un semplice codice identificativo sequenziale della corsa mentre la seconda assume un unico valore, in quanto tutte le osservazioni fanno riferimento al mese di maggio. Esse vengono quindi eliminate. Successivamente, dato anche il gran numero di osservazioni a disposizione, si decide di suddividere i dati in tre sottoinsiemi: training set, validation set e test set (**Tabella 1**). Si vuole nel seguito utilizzare il primo per "allenare" i vari modelli, per poi confrontarne le diverse performance attraverso le previsioni effettuate sul validation set. Infine, con il test set si vuole quantificare l'errore finale del modello. Si procede quindi con un'analisi esplorativa delle variabili del training set.

	% osservazioni	**Numero di osservazioni**
Training set	50%	121.589
Validation set	25%	60.795
Test set	25%	60.795

Tabella 1: *dimensioni degli insiemi di training, validation e test*

a. Variabili numeriche

Inizialmente si decide di considerare le variabili numeriche (*length_time*, *trip_distance*, *fare_amount* e la variabile risposta *tip_amount*). Si decide di convertire *trip_distance* in kilometri, per facilità di interpretazione dei risultati. Osservando i range dei valori si nota subito come *length_time* e *trip_distance_km* abbiano in certe osservazioni valori nulli. Inoltre, la variabile *length_time* assume talvolta valori eccessivamente alti o bassi mentre in un caso *fare_amount* è pari a 0.01$. Si decide, in via preliminare, di eliminare le osservazioni

- che abbiano valori nulli di *trip_distance*
- che abbiano una *length_time* inferiore a 10 secondi o superiore a 10800 secondi
- che abbiano *fare_amount* inferiore a 2.5$

A questo punto, volendo trovare eventuali osservazioni che siano classificabili come outliers con riferimento a tutte le variabili numeriche, si decide di applicare il concetto di distanza di Mahalanobis (**Definizione 1**).

Definizione 1: si definisce distanza di Mahalanobis tra l'i-sima unità statistica $\boldsymbol{x}_i$ e il suo baricentro $\overline{\boldsymbol{x}}$

$$d_M(\boldsymbol{x}_i, \overline{\boldsymbol{x}}) = \sqrt{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T S^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})}$$

dove S^{-1} è l'inversa della matrice di varianze e covarianze.

Come detto, si considerano in questa fase le sole variabili numeriche e si calcola la distanza di Mahalanobis al quadrato, ossia ${d_M}^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$, tra ogni osservazione

$$\boldsymbol{x}_i = (x_{i,length\_time}, x_{i,trip\_distance\_km}, x_{i,fare\_amount}, x_{i,tip\_amount})^T$$

e il vettore di medie

$$\overline{\boldsymbol{x}} = (\bar{x}_{length\_time}, \bar{x}_{trip\_distance\_km}, \bar{x}_{fare\_amount}, \bar{x}_{tip\_amount})^T$$

Assumendo che le osservazioni così definite siano realizzazioni indipendenti di un'unica Normale *p*-variata, si può dimostrare che ${d_M}^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$ si distribuisce come una χ^2 con *p* gradi di libertà. Si può quindi classificare un'unità statistica come outlier se

$${d_M}^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}}) > q_{0.95}$$

dove $q_{0.95}$ è il quantile 0.95 di una distribuzione χ^2 con *p* gradi di libertà (ovvero, nel caso in esame, 4 gradi di libertà). Si può inoltre definire il valore atteso di outliers come

$$valore\ atteso\ di\ outliers = n \times 0.05$$

Applicando tale criterio, sono state classificate come outliers 8351 osservazioni, un numero sostanzialmente più elevato di quello outliers attesi, pari, secondo la definizione di cui sopra, a 6064. Volendo stimare un modello che performi bene sui dati "ordinari" e che non sia influenzato da osservazioni anomale, si decide di rimuovere dal training set questi outliers. Si può notare come, dopo l'eliminazione, i range dei valori delle variabili considerate si siano notevolmente ridotti (in particolare l'estremo superiore). Concentrandosi poi in maniera specifica sulla variabile risposta *tip_amount*, si osserva, tramite un boxplot, come vi siano altre osservazioni che potrebbero essere classificate come anomale. Si decide di calcolare una variabile aggiuntiva *tip_perc*, definita come il rapporto tra la mancia (*tip_amount*) e la tariffa (*fare_amount*), e di eliminare le osservazioni che abbiano valori inferiori al 10% o superiori al 60%. A questo punto dell'analisi, il numero di unità statistiche facenti parte del training set è pari a 107.735.

b. Variabili "di tempo"

Si procede ora all'analisi delle variabili che identificano il momento in cui è avvenuta la corsa (*pickup_hour*, *pickup_week*, *pickup_doy*, *pickup_wdoy*). Tali variabili, seppur anch'esse assumano valori numerici, sono state codificate come delle factor in quanto, di fatto, variabili discrete. Dopo aver osservato i boxplot della variabile risposta condizionati a ciascuna di queste variabili, si decide di tenere in considerazione le sole *pickup_hour* e *pickup_wday* e di trascurare *pickup_week* e *pickup_wdoy* (che non sembrano avere un grande impatto su *tip_amount*)

c. Variabili geografiche

Per quanto riguarda le variabili che identificano il punto di partenza e il punto di arrivo della corsa, si decide di trascurare le variabili *pickup_NTACode*, *dropoff_NTACode* e *pair*, in quanto il numero di valori assunti da tali variabili è molto alto e, in particolare, alcuni valori sono presenti con una frequenza troppo bassa. Allo stesso modo si eliminano le quattro variabili relative alla latitudine e longitudine. Si tengono invece in considerazione le variabili *pickup_BoroCode* e *dropoff_BoroCode*: esse assumono cinque possibili valori ("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"). Poiché le osservazioni relative a Bronx e Staten Island sono poche, meno dell'1% del totale, si decide di accorpare questi due livelli in un generico "Other".

d. Altre variabili

La variabile *passenger_count*, seppur numerica, è stata codificata come una factor con due livelli: "one" se in quella corsa vi era un solo passeggero, oppure "two or more" se vi erano almeno due passeggeri. È stata inoltre lasciata nel set di variabili *vendor_id*, una variabile factor a due livelli che identifica il taxi provider che ha effettuato la corsa.

2. Stima dei modelli e confronto sul validation set

Come prima cosa si procede nel modificare il validation set in analogia a quanto effettuato in precedenza: si eliminano le variabili non considerate in quest'analisi e si effettua la ricodifica di *pickup_BoroCode*, *dropoff_BoroCode* e *passenger_count*. Inizialmente si decide di stimare alcuni modelli lineari: uno estremamente semplice, che contiene la sola esplicativa *fare_amount*, un secondo contenente anche *trip_distance_km* e *length_time*, e infine un terzo con tutte le variabili. Di seguito le performance ottenute in base al MEAN ABSOLUTE ERROR (MAE), calcolato sulle previsioni del validation set (**Tabella 2**).

	MAE
Modello 1 (solo *fare_amount*)	0.6134
Modello 2 (variabili numeriche)	0.6334
Modello 3 (tutte le variabili)	0.6241

***Tabella 2**: Mean Absolute Error dei modelli lineari calcolato sul validation set*

Ciò che balza immediatamente all'occhio è che il modello migliore è quello che contiene la sola variabile relativa alla tariffa della corsa: ciò suggerisce che questa variabile sia estremamente importante, anche con riferimento a tutte le altre, per la previsione della mancia. Ad ogni modo, non essendo del tutto soddisfatti e volendo provare a migliorare le previsioni ottenute, si decide di considerare i metodi della ridge regression, del LASSO e dell'elastic net (si ricorda che i primi due possono essere visti come casi particolari dell'ultimo). Si utilizza il metodo della cross-validation per scegliere il parametro λ ottimale, ottenendo le seguenti performance dei modelli (**Tabella 3**).

	MAE (λ minimo)	**MAE** (λ 1-standard-error)
Ridge regression	0.6058	0.6440
LASSO	0.6056	0.6077
Elastic net (con $\alpha = 0.5$)	0.6057	0.6318

***Tabella 3**: Mean Absolute Error di ridge, LASSO, elastic-net calcolato sul validation set*

In generale, i tre modelli hanno errori medi assoluti abbastanza simili tra loro. Si nota che i valori ottimali di λ risultano essere molto vicini allo zero. Ciò significa che nel caso della ridge regression la penalità utilizzata per stimare i parametri è molto bassa e dunque i coefficienti stimati sono abbastanza simili a quelli ottenuti con una regressione lineare semplice. Per quanto riguarda invece il LASSO, ne consegue che, almeno nel caso in cui λ è scelto come valore minimo, viene effettuato uno shrinkage piuttosto ridotto (solo 5 coefficienti vengono annullati). Questo è probabilmente conseguenza del fatto che già nelle fasi preliminari è stato deciso di trascurare alcune variabili. Con riferimento ai modelli ottenuti con λ scelto tramite la regola dell'1-standard-error, il LASSO è senza dubbio quello che prevede meglio: pur effettuando uno shrinkage ben più marcato (i coefficienti annullati sono in questo caso 24), il suo errore assoluto di previsione aumenta in modo quasi impercettibile, a differenza invece di quelli della regressione ridge e dell'elastic net. Essendo soddisfatti del miglioramento ottenuto, si decide di utilizzare proprio il LASSO come metodo di previsione finale.

3. Stima del LASSO su training e validation e stima finale dell'errore di previsione

A questo punto si procede considerando come nuovo insieme di stima del modello l'unione tra il precedente training set e il validation set. Per la stima finale dell'errore medio assoluto si utilizzerà invece il test set, mai considerato fin'ora nell'analisi (**Tabella 4**).

	% osservazioni	**Numero di osservazioni**
Training set	75%	182.384
Test set	25%	60.795

***Tabella 4**: dimensioni degli insiemi di training e di test*

Si procede effettuando tutte le operazioni di pre-processing sul nuovo training set: in particolare, dopo l'eliminazione dei valori anomali si ottiene un insieme di 161.906 osservazioni. Si stima il LASSO su questo insieme e si ottiene la stima dell'errore medio assoluto sul test set (**Tabella 5**).

	MAE (λ minimo)	**MAE** (λ 1-standard-error)
LASSO	0.6030	0.6051

***Tabella 5**: Mean Absolute Error del LASSO calcolato sul test set*

4. Stima del LASSO su tutto l'insieme a disposizione, calcolo delle previsioni finali e conclusioni

Si decide ora di utilizzare tutto l'insieme di dati a disposizione. Dopo avere effettuato le operazioni di eliminazione degli outliers e di ricodifica delle variabili, si ottiene un training set contenente 215.902 osservazioni. A questo punto si considera come test set l'insieme di 243.180 unità, delle quali non si conosce la variabile risposta e, dopo aver stimato il LASSO sul nuovo insieme di training, si ottengono le previsioni finali di *tip_amount*. Sebbene, come affermato in precedenza, la scelta di λ con la regola dell'1-standard-error produca un modello più parsimonioso, si decide di utilizzare il λ minimo con l'obiettivo ottenere un errore di previsione (seppur lievemente) più basso. Considerando che in precedenza la stima del LASSO su un insieme di training più ampio (punto 3) ha fornito un errore medio di previsione più basso del precedente (punto 2), è ipotizzabile che l'errore assoluto medio finale, non quantificabile in quanto non si conoscono i veri valori della variabile risposta nel test set, sia inferiore a 0.6030.