Tesi Metodologie Statistiche per i Big Data

"To host or not to host, that is the question"

Andrea Baldinini

Jacopo Bonanno

Sabrina Infante

CAPITOLO 1. ANALISI ESPLORATIVA

1.1 *Scelta del dataset e analisi esplorativa*

Il dataset è stato scaricato da Kaggle.com e riguarda le cancellazioni di prenotazioni alberghiere afferenti ad un'unica struttura. La scelta è ricaduta su questo dataset in quanto ci ha dato la possibilità di studiare il fenomeno dell'overbooking, ovvero la possibilità data alle strutture ricettive di accettare un numero superiore di prenotazioni rispetto alla loro vera capacità. In questo contesto diventa cruciale riuscire a classificare correttamente una prenotazione che verrà cancellata da una che invece non verrà cancellata. Abbiamo lavorato con 36275 osservazioni e 11 caratteristiche:

- **number of adults:** numero di adulti [numerica]
- **number of children:** numero di bambini [numerica]
- **number of weekend nights:** numero di notti nel weekend [numerica]
- **number of week nights:** numero di notti nella settimana [numerica]
- **lead time:** tempo espresso in numero di giorni tra il momento della prenotazione e la data della prenotazione [numerica]
- **arrival month:** mese di arrivo [categorica]
- **number of previous cancellations:** numero di cancellazioni precedenti [numerica]
- **number of previous bookings not canceled:** numero di prenotazioni precedenti non cancellate [numerica]
- **average price per room:** prezzo medio per stanza [numerica]
- **number of special requests:** numero di richieste speciali [numerica]
- **booking status:** stato della cancellazione [categorica, 0: non cancellata; 1: cancellata]

La variabile dicotomica di nostro interesse risulta essere "booking status" mentre le altre caratteristiche sono variabili quantitative, tranne "arrival month" che viene utilizzata per la creazione di subset mensili.

Vista le differenti scale di misura di queste variabili, procediamo con la standardizzazione del dataset, a parte la dicotomica e il mese della prenotazione.

number of adults	number of children	number of weekend nights	number of week nights	lead time	arrival month	number of previous cancellations
Min. :0.000	Min. : 0.0000	Min. :0.0000	Min. : 0.000	Min. : 0.00	Min. : 1.000	Min. : 0.00000
1st Qu.:2.000	1st Qu.: 0.0000	1st Qu.:0.0000	1st Qu.: 1.000	1st Qu.: 17.00	1st Qu.: 5.000	1st Qu.: 0.00000
Median :2.000	Median : 0.0000	Median :1.0000	Median : 2.000	Median : 57.00	Median : 8.000	Median : 0.00000
Mean :1.845	Mean : 0.1053	Mean :0.8107	Mean : 2.204	Mean : 85.23	Mean : 7.424	Mean : 0.02335
3rd Qu.:2.000	3rd Qu.: 0.0000	3rd Qu.:2.0000	3rd Qu.: 3.000	3rd Qu.:126.00	3rd Qu.:10.000	3rd Qu.: 0.00000
Max. :4.000	Max. :10.0000	Max. :7.0000	Max. :17.000	Max. :443.00	Max. :12.000	Max. :13.00000

number of previous bookings not canceled	average price per room	number of special requests	booking status
Min. : 0.0000	Min. : 0.00	Min. :0.0000	0:24390
1st Qu.: 0.0000	1st Qu.: 80.30	1st Qu.:0.0000	1:11885
Median : 0.0000	Median : 99.45	Median :0.0000	
Mean : 0.1534	Mean :103.42	Mean :0.6197	
3rd Qu.: 0.0000	3rd Qu.:120.00	3rd Qu.:1.0000	
Max. :58.0000	Max. :540.00	Max. :5.0000	

Tabella 1.1. Summary per colonna

In seguito, è stato controllato il numero di NAs, i quali risultavano del tutto assenti.

A questo punto è risultato necessario osservare nello specifico le nostre caratteristiche e mediante la funzione “ggpairs” le abbiamo studiate dal punto di vista sia univariato che bivariato.

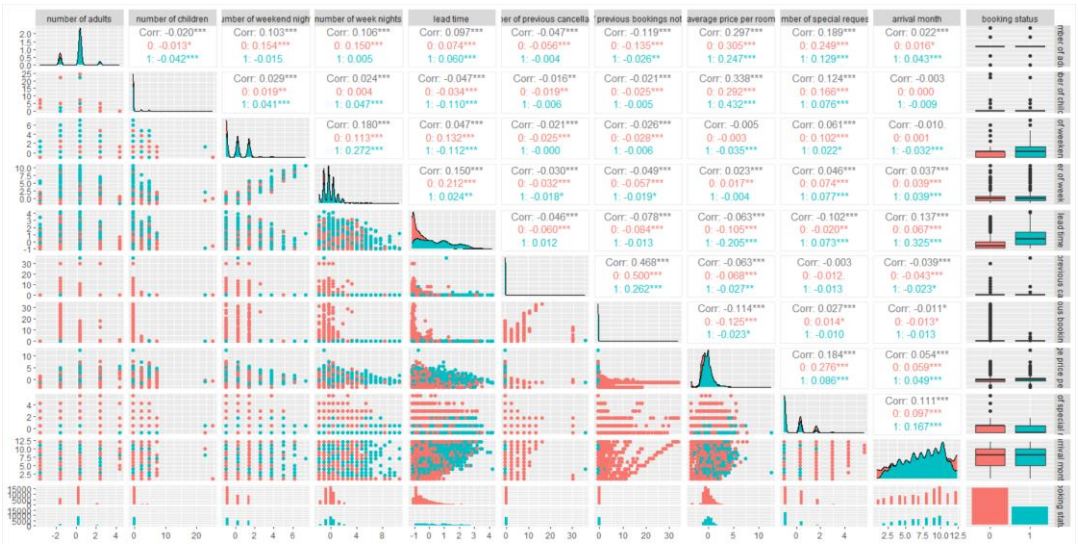

Figura 1.1. Grafico GGpairs

Si possono effettuare diverse annotazioni in merito all’output grafico. In primis, per quanto riguarda l’analisi univariata mediante le distribuzioni sulla diagonale principale, le variabili appaiono per lo più dissimili da caratteristiche normalmente distribuite, il che viene analizzato in seguito mediante calcolo di skewness e kurtosis. La colonna di destra presenta i boxplot, in grado di mostrare come per alcune variabili il numero di outliers risulta elevato – un esempio è il numero di precedenti non cancellazioni - , mentre per altre non risultano presenti in maniera numerosa.

Possiamo notare la distribuzione non eccessivamente squilibrata tra i due valori della variabile dicotomica, pari a 32,76% per la cancellazione e 67,24% per la non cancellazione. Tramite questo grafico otteniamo anche i valori riferiti ai test inferenziali svolti in merito alle correlazioni così come gli scatterplot aventi sugli assi le variabili prese in coppa. Nonostante ciò abbiamo ritenuto comunque utile effettuare il plot delle correlazioni per semplificare la lettura delle stesse.

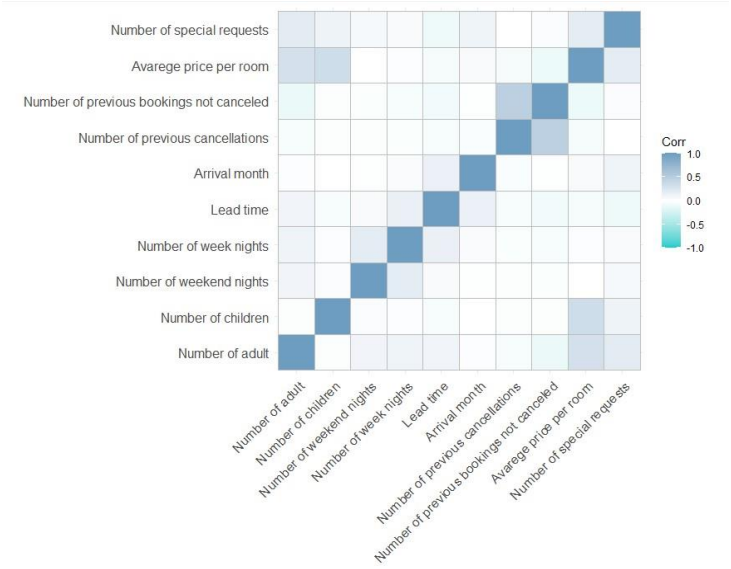

Figura 1.2. Grafico delle correlazioni

Si può notare una moderata e positiva correlazione tra number of previous bookings not canceled e number of previous cancellations (0.468 da "ggpairs" precedente), evidentemente caratterizzanti un cliente frequente della struttura; inoltre una inferiore correlazione tra il number of children ed average price per room (0.338). Si nota una diffusa incorrelazione o leggermente negativa tra le altre variabili. Quelle correlate si può supporre dovranno essere rimosse dall'analisi in quanto ridondanti, ma questa decisione deriva da una PCA o da test più specifici legati alla collinearità.

In seguito abbiamo analizzato la skewness e la kurtosis delle caratteristiche e ne deriva la seguente tabella:

	Skewness	Kurtosis
number of adults	-0.3332674	0.8143339
number of children	4.7099600	36.9743892
number of weekend nights	0.7375550	0.2984691
number of week nights	1.5992181	7.7964484
lead time	1.2923846	1.1790357
number of previous cancellations	25.1977919	732.5939585
number of previous bookings not canceled	19.2485987	457.2914998
average price per room	0.6670777	3.1531856
number of special requests	1.1449861	0.8809361
arrival month	-0.3482001	-0.9333397

Tabella 1.2. Asimmetria e della curtosi

Una skewness maggiore di 0 indica asimmetria a destra, mentre minore di 0 una a sinistra; una Kurtosis maggiore di 3 indica distribuzione leptocurtica, mentre minore di 3 una platocurtica. Queste caratteristiche delle variabili, già evidenti dall'output del "ggpairs", sono ora confermate quantitativamente.

1.2 *Principal Components Analysis*

L'analisi viene condotta sulla matrice delle correlazioni poiché si è osservata una diversa variabilità tra le variabili. Ciò è stato fatto per evitare che variabili con varianza maggiore diano un contributo maggiore delle altre alterando i risultati dell'analisi.

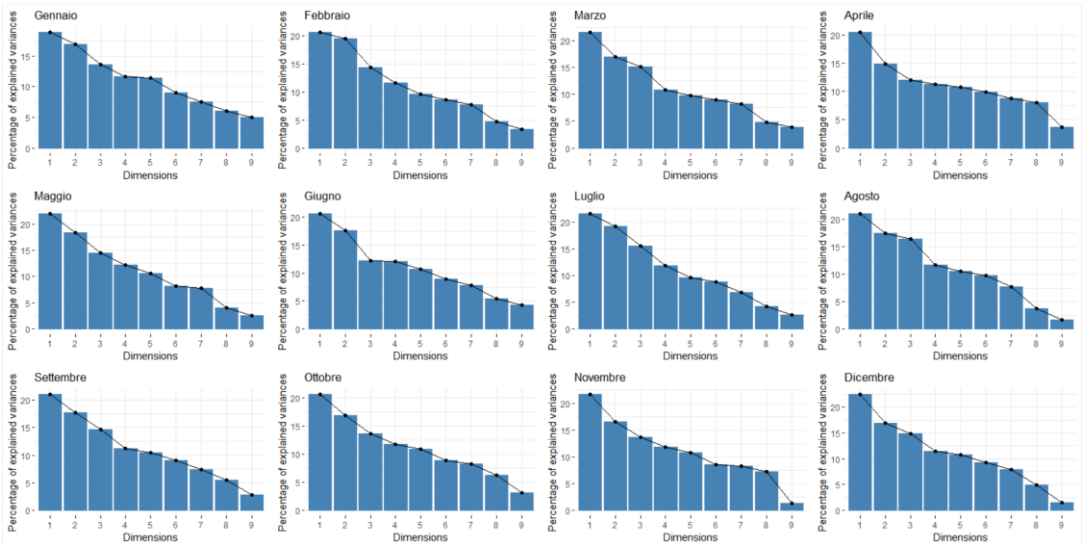

Figura 1.3. Grafico delle varianze spiegate per mese

Si evince che la varianza spiegata dalle componenti principali varia leggermente al variare del mese. In generale, soffermandosi sulla prima componente principale, si può affermare che spiega all'incirca il 20% della varianza totale in tutti i mesi. La seconda componente spiega mediamente il 17%. Dunque si può affermare che le prime due componenti principali spiegano una bassa percentuale (circa 37%) della variabilità del dataset originario. Ne deriva che un numero sufficiente di componenti principali in grado di spiegare all'incirca il 70% della variabilità totale del dataset originario è di 5 componenti in ciascun mese, tranne per il mese di aprile in cui la soglia del 70% della variabilità totale viene superata scegliendo in favore di 6 componenti principali.

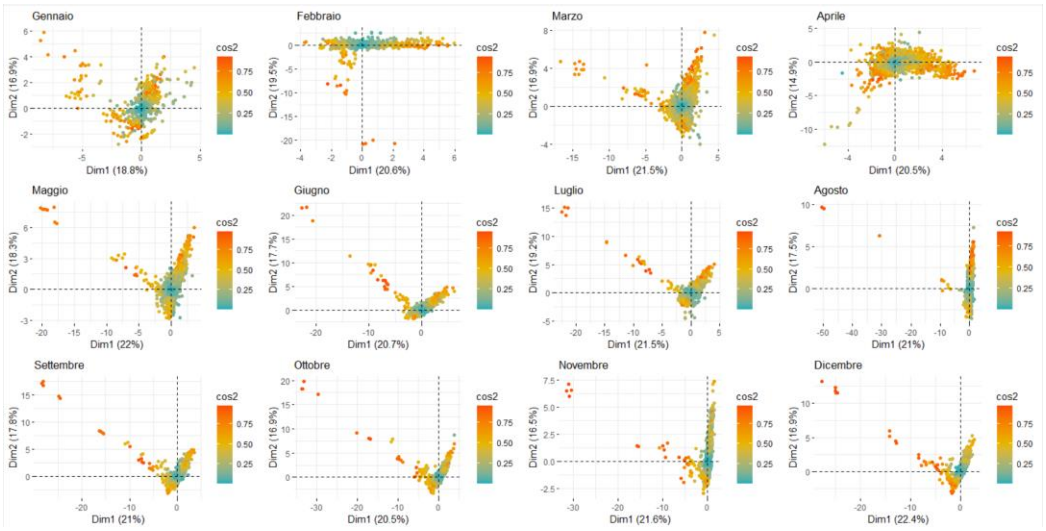

Figura 1.4. Mappa fattoriale della prima e seconda componente principale per mese

Dalla mappa fattoriale si evince tendenzialmente la presenza di un agglomerato nell'intorno dell'origine, stando a significare la bassa correlazione delle osservazioni con le due componenti principali. Si può notare comunque la presenza di outliers disposti lungo una traiettoria lineare nel secondo quadrante, giustificabile solamente mediante successiva analisi del biplot.

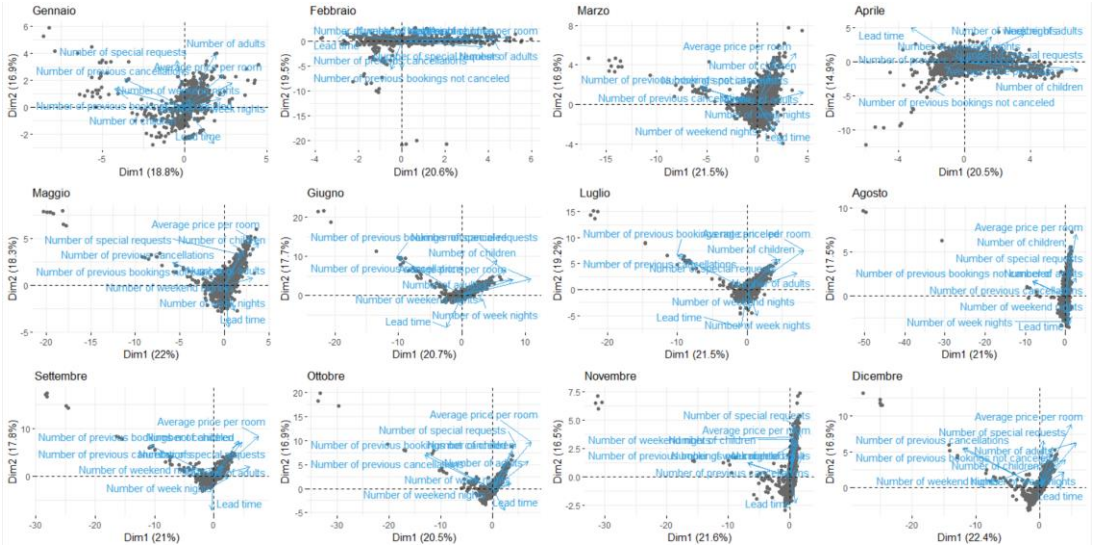

Figura 1.5. Biplot per mese

Osservando la distribuzione delle variabili, si può attribuire la struttura di cui sopra alla presenza di numerosi outliers correlati alle direzioni delle variabili "number of previous cancellation" e "number of previous bookings not canceled". Abbiamo, dunque, deciso di rimuovere dall'analisi queste variabili, affinché venissero rimossi questi outliers.

Ripetendo la PCA e focalizzando l'attenzione riguardo ai mesi giugno, luglio e agosto, di maggior interesse, notiamo quanto segue.

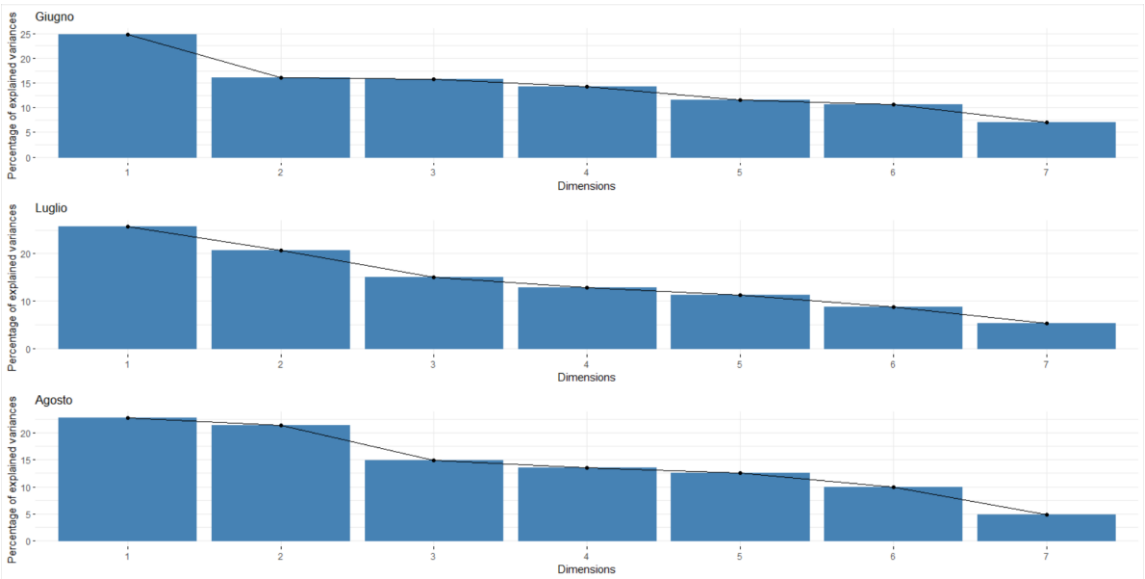

Figura 1.6. Grafico delle varianze spiegate giugno, luglio, agosto

Si può notare un leggero aumento della capacità di spiegare la varianza delle componenti principali. Le prime due, infatti, risultano poter spiegare mediamente il 40% della variabilità totale. Inoltre, risulta essere 4 il numero sufficiente di componenti principali in grado di spiegare almeno il 70% della variabilità totale.

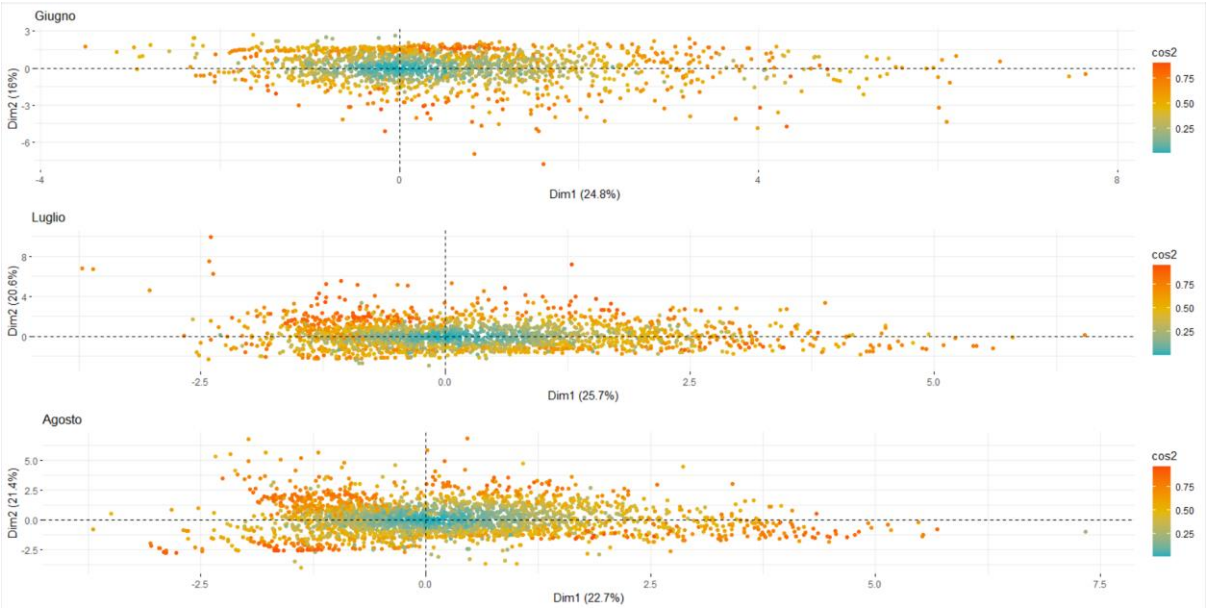

Figura 1.7. Mappa fattoriale della prima e seconda componente principale giugno, luglio, agosto

A conferma di ciò, possiamo notare che le mappe fattoriali presentano un numero molto minore di outliers ed una maggiore concentrazione attorno all'origine.

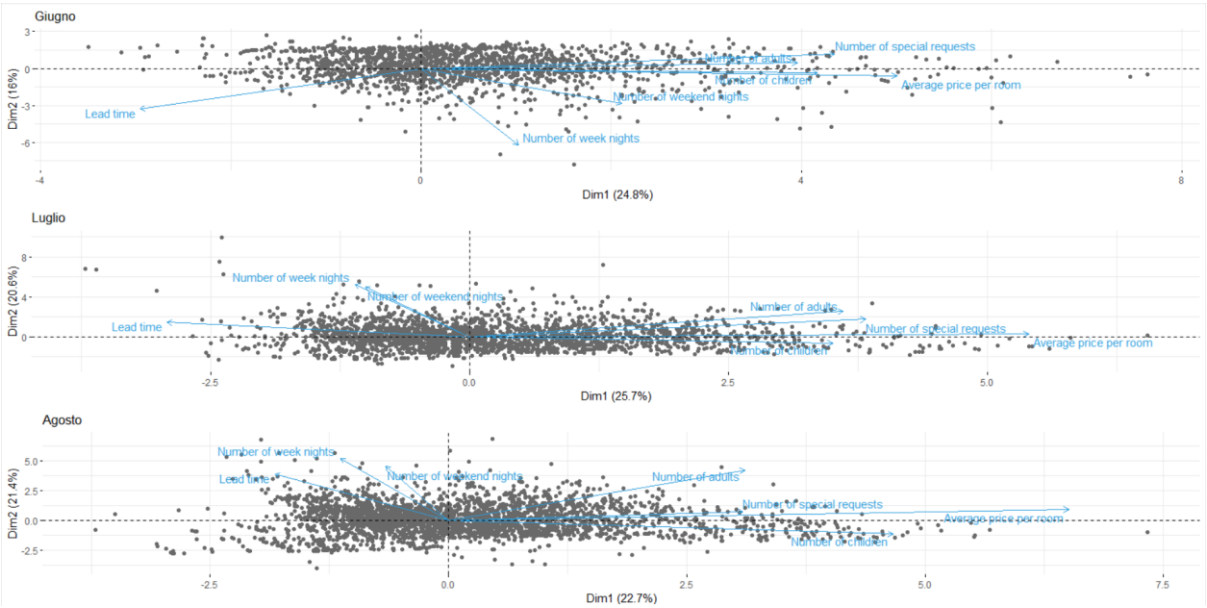

Figura 1.8. Biplot giugno, luglio, agosto

Infine, possiamo osservare come, ora, le variabili si dispongono in maniera simile per luglio e agosto, mentre giugno risulta leggermente diverso, presentando alcune variabili con direzione opposta rispetto ai primi due mesi citati. Inoltre, per ogni mese, si può notare la disposizione delle osservazioni in una struttura ampia, piuttosto che stretta, dato che le variabili tendono a disporsi maggiormente lungo la prima componente.

1.3 *Cluster Analysis*

Il nostro obiettivo risultava essere quello di verificare se la classificazione mediante k-means, svolta mese per mese, rispecchiasse la suddivisione dicotomica data dalla variabile di interesse booking status.

Abbiamo proceduto utilizzando il dataset scalato, privo della caratteristica oggetto di studio e delle due variabili rimosse in sede di PCA.

Inizialmente abbiamo individuato il numero ottimale di gruppi mediante analisi della silhouette, tramite la funzione "fviz_nbclust". Occorre sottolineare che il numero migliore di cluster raramente sia pari a 2, presentando valori per i diversi mesi da 2 a 10. Questo fatto ci ha portato a concludere che la cluster analysis non possa rispecchiare la nostra suddivisione dicotomica.

Nonostante ciò abbiamo proseguito per valutare la presenza di raggruppamenti significativi.

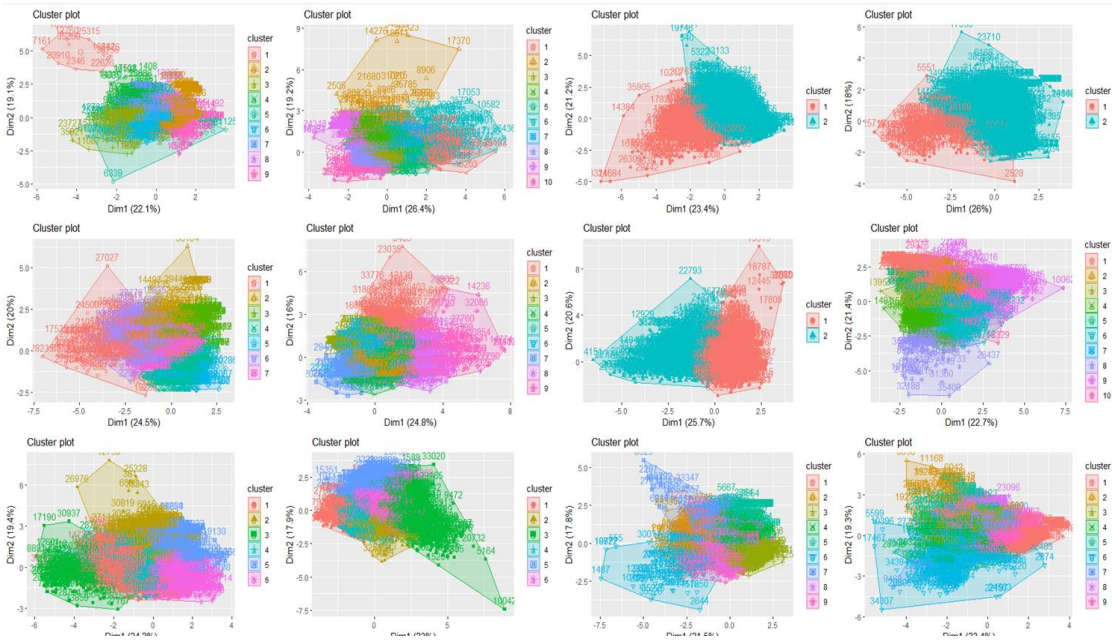

Figura 1.9. K-means plot tramite mappa fattoriale per mese

Come è possibile notare dai plot dei nostri gruppi, questi non riescono mai ad individuare delle suddivisioni particolarmente convincenti, infatti queste risultano essere caratterizzate da un numero solitamente molto alto di cluster e numerosi fenomeni di sovrapposizione delle nostre unità.

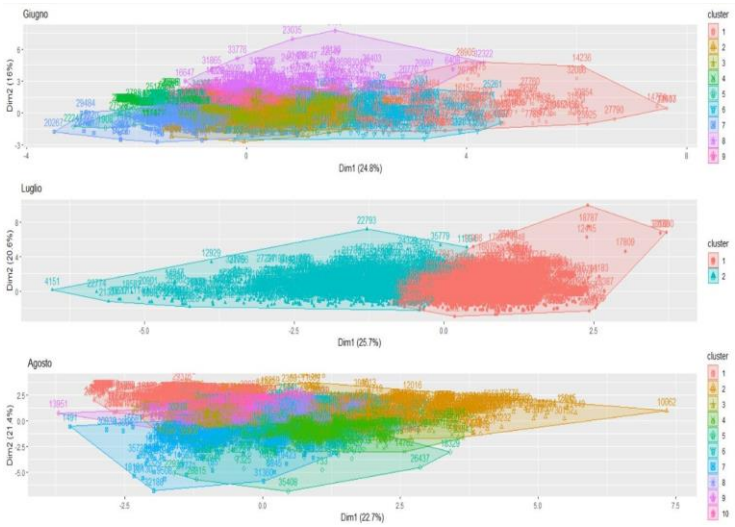

Figura 1.10. K-means plot tramite mappa fattoriale per mesi target

CAPITOLO 2. MODELLI DI CLASSIFICAZIONE

2.1 *Note metodologiche*

Tutti i modelli adottati vengono applicati solamente ai mesi di alta stagione (giugno, luglio, agosto) questo perché sono i mesi presentanti una situazione più equilibrata in merito alla distribuzione della variabile dicotomica.
Abbiamo implementato un algoritmo in grado di misurare mese per mese gli score: precision, accuracy, sensitivity e specificity.
Per ogni mese, vengono testati i threshold compresi tra 0.05 e 0.95 (estremi inclusi) e per ciascuno di questi il modello viene validato per un numero molteplice di volte, per garantire la robustezza dei modelli al variare del training e test set.
Abbiamo mantenuto sia la proporzione tra training-set e test-set, ovvero 80-20, per ciascun modello e sia la proporzione in entrambi i set della distribuzione della variabile dicotomica.
Una volta effettuato il training del modello, viene calcolata la confusion matrix, e da questa ne traiamo tutti i valori necessari per gli score desiderati.

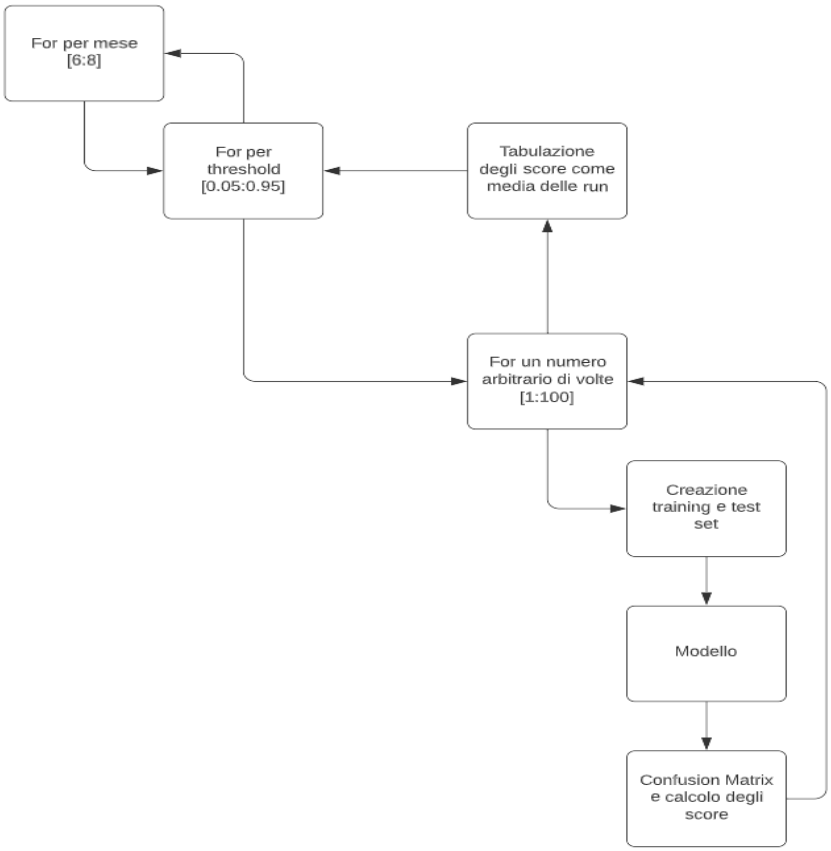

Figura 2.1. Diagramma rappresentante il funzionamento del nostro algoritmo

Abbiamo, in seguito, tabellato i valori ottenuti per i diversi mesi in funzione dei vari threshold, considerando la media di ogni valore score per le molteplici run.
Infine abbiamo cercato, per ogni mese, il valore soglia, per cui ogni diverso score presentasse il valore maggiore.

Nei successivi paragrafi vengono presentati i risultati delle tecniche di classificazione implementate: LDA, regressione logistica, k-nearest-neighbor, SVM e random forest. In alcuni modelli abbiamo preferito utilizzare una versione del nostro algoritmo leggermente rivisitata, ciononostante il fine rimane quello della robustezza.

2.2 *Linear Discriminant Analysis*

Abbiamo inizialmente utilizzato per il modello LDA un dataset formato da tutte le variabili a nostra disposizione. Di seguito vengono riportate tutte le tabelle risultanti.

Threshold	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
0.05	0.45	0.50	0.99	0.17	0.46	0.48	1.00	0.05	0.42	0.46	1.00	0.12
0.10	0.49	0.58	0.98	0.32	0.48	0.51	0.99	0.13	0.45	0.53	0.96	0.25
0.15	0.52	0.63	0.94	0.41	0.51	0.56	0.98	0.22	0.49	0.60	0.94	0.38
0.20	0.54	0.66	0.92	0.48	0.54	0.62	0.94	0.36	0.53	0.65	0.90	0.49
0.25	0.57	0.68	0.89	0.54	0.57	0.65	0.90	0.44	0.57	0.70	0.87	0.59
0.30	0.62	0.73	0.86	0.64	0.60	0.68	0.87	0.52	0.60	0.72	0.83	0.65
0.35	0.65	0.75	0.83	0.69	0.64	0.71	0.84	0.61	0.63	0.74	0.81	0.70
0.40	0.69	0.78	0.80	0.76	0.66	0.73	0.80	0.67	0.65	0.75	0.77	0.74
0.45	0.70	0.77	0.73	0.79	0.69	0.73	0.73	0.74	0.68	0.77	0.74	0.78
0.50	0.75	0.78	0.69	0.84	0.71	0.73	0.66	0.78	0.71	0.77	0.69	0.82
0.55	0.76	0.78	0.66	0.86	0.74	0.72	0.59	0.83	0.74	0.77	0.64	0.85
0.60	0.82	0.79	0.60	0.91	0.76	0.71	0.51	0.87	0.76	0.76	0.55	0.89
0.65	0.86	0.79	0.58	0.94	0.78	0.70	0.45	0.90	0.81	0.75	0.48	0.93
0.70	0.86	0.78	0.53	0.94	0.80	0.68	0.39	0.92	0.83	0.73	0.38	0.95
0.75	0.87	0.74	0.42	0.96	0.81	0.65	0.30	0.94	0.87	0.71	0.29	0.97
0.80	0.87	0.71	0.32	0.97	0.81	0.63	0.23	0.95	0.89	0.68	0.21	0.98
0.85	0.94	0.68	0.23	0.99	0.87	0.62	0.18	0.98	0.88	0.65	0.13	0.99
0.90	0.97	0.66	0.16	1.00	1.00	0.59	0.08	1.00	0.90	0.64	0.09	0.99
0.95		0.60	0.01	1.00		0.55	0.00	1.00		0.61	0.01	1.00

Tabella 2.1. Matrice score per ogni threshold

	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
Value	0.97	0.79	0.99	1.00	1.00	0.73	1.00	1.00	0.90	0.77	1.00	1.00
Threshold	0.90	0.65	0.05	0.95	0.90	0.45	0.05	0.95	0.90	0.55	0.05	0.95

Tabella 2.2. Matrice score massimi e relativi threshold

In seguito, ci siamo domandati se il modello avente solamente le variabili maggiormente discriminanti, potesse o meno avere risultati migliori.
Abbiamo, dunque, mediante partition plot, selezionato per ogni mese le due variabili maggiormente discriminanti sulla base dell'AER minore.

Figura 2.2. Partition plot delle variabili più discriminanti

Da subito abbiamo osservato che l'AER per ogni mese risultasse peggiore rispetto ai risultati ottenuti dal modello generale.

Nonostante ciò, abbiamo proceduto mediante validazione dei tre modelli ridotti. I risultati ottenuti hanno confermato la nostra ipotesi, mediamente i modelli costruiti in questo modo risultano essere peggiori, dunque abbiamo scelto di mantenere come modello migliore quello generale.

2.3. *Regressione Logistica*

Per quanto concerne la regressione logistica occorre soffermarsi sulla modalità di selezione delle variabili.
Dapprima abbiamo testato il modello generale, notando che le variabili Number of children, Number of previous cancellations e Number of previous bookings not canceled presentavano p-value superiori al 10%.
Per questo motivo abbiamo operato attraverso il confronto mediante LR test ed AIC, ponendo in relazione il modello generale e quello nested, rimuovendo di volta in volta la variabile con p-value maggiore.
Al termine dell'analisi preliminare, siamo giunti alla conclusione attesa, ovvero che il modello migliore fosse quello privo di suddette variabili non significative.

Di seguito i risultati della validazione del modello usando il dataset ridotto.

Threshold	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
0.05	0.47	0.54	0.99	0.24	0.46	0.47	1.00	0.05	0.43	0.47	0.99	0.14
0.10	0.51	0.61	0.97	0.37	0.48	0.52	0.99	0.13	0.46	0.54	0.96	0.27
0.15	0.53	0.64	0.93	0.45	0.51	0.56	0.98	0.22	0.50	0.60	0.94	0.39
0.20	0.56	0.67	0.91	0.51	0.54	0.61	0.93	0.34	0.53	0.65	0.90	0.49
0.25	0.60	0.72	0.89	0.60	0.57	0.65	0.89	0.45	0.57	0.69	0.87	0.58
0.30	0.63	0.74	0.86	0.66	0.60	0.68	0.87	0.52	0.62	0.73	0.84	0.67
0.35	0.66	0.76	0.84	0.70	0.63	0.71	0.85	0.60	0.64	0.75	0.81	0.71
0.40	0.70	0.78	0.82	0.76	0.66	0.73	0.80	0.67	0.66	0.76	0.77	0.75
0.45	0.74	0.80	0.77	0.81	0.71	0.74	0.74	0.75	0.68	0.76	0.74	0.78
0.50	0.75	0.78	0.69	0.84	0.72	0.73	0.66	0.79	0.71	0.77	0.68	0.82
0.55	0.82	0.81	0.67	0.90	0.75	0.73	0.59	0.84	0.74	0.76	0.61	0.86
0.60	0.84	0.80	0.63	0.92	0.77	0.71	0.52	0.87	0.78	0.76	0.55	0.90
0.65	0.85	0.79	0.59	0.93	0.78	0.69	0.45	0.89	0.82	0.75	0.47	0.93
0.70	0.86	0.78	0.54	0.94	0.80	0.68	0.37	0.92	0.85	0.73	0.38	0.96
0.75	0.83	0.70	0.33	0.95	0.80	0.65	0.30	0.94	0.87	0.71	0.29	0.97
0.80	0.91	0.70	0.28	0.98	0.81	0.63	0.22	0.96	0.89	0.68	0.21	0.98
0.85	0.93	0.68	0.22	0.99	0.85	0.62	0.18	0.97	0.90	0.66	0.14	0.99
0.90	0.94	0.66	0.18	0.99	0.98	0.58	0.08	1.00	0.91	0.64	0.09	0.99
0.95	0.92	0.61	0.03	1.00		0.55	0.00	1.00		0.62	0.02	1.00

Tabella 2.3. Matrice score per ogni threshold

	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
Value	0.94	0.81	0.99	1.00	0.98	0.74	1.00	1.00	0.91	0.77	0.99	1.00
Threshold	0.90	0.55	0.05	0.95	0.90	0.45	0.05	0.95	0.90	0.50	0.05	0.95

Tabella 2.4. Matrice score massimi e relativi threshold

Inoltre abbiamo studiato il valore medio, per ogni mese, dei regressori.

	Giugno	Luglio	Agosto
(Intercept)	-1.2805607	-0.72234467	-0.96901904
`number of adults`	0.5766291	0.09649001	0.03533986
`number of weekend nights`	0.4136497	0.07940087	0.06983604
`number of week nights`	0.1797151	0.09460238	0.16189579
`lead time`	1.4267361	0.97648139	1.29687323
`average price per room`	0.9084261	0.47594634	0.66046189
`number of special requests`	-1.5161704	-1.00368096	-0.83316479

Tabella 2.5. Matrice dei beta

Possiamo notare come tutti i regressori contribuiscano positivamente alla scelta di cancellare la prenotazione, a parte il number of special requests.

2.4 *K-Nearest-Neighbor*

Per quanto concerne l'implementazione di questa tecnica, occorre sottolineare che il modello mediante la tecnica del K-folds, nel nostro caso abbiamo optato per k = 10 ed i risultati sono i seguenti.

Threshold	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
0.05	0.82	0.86	0.82	0.88	0.77	0.78	0.75	0.81	0.70	0.77	0.72	0.80
0.10	0.82	0.85	0.81	0.88	0.76	0.77	0.73	0.81	0.72	0.79	0.73	0.82
0.15	0.84	0.87	0.84	0.89	0.75	0.77	0.75	0.79	0.71	0.77	0.71	0.81
0.20	0.83	0.87	0.85	0.88	0.79	0.81	0.78	0.83	0.71	0.77	0.71	0.81
0.25	0.85	0.88	0.86	0.90	0.74	0.77	0.75	0.78	0.73	0.79	0.75	0.82
0.30	0.83	0.88	0.86	0.88	0.74	0.77	0.74	0.79	0.71	0.77	0.70	0.81
0.35	0.83	0.87	0.85	0.88	0.78	0.80	0.77	0.83	0.75	0.80	0.74	0.84
0.40	0.85	0.88	0.87	0.90	0.75	0.77	0.75	0.80	0.70	0.77	0.71	0.80
0.45	0.82	0.86	0.85	0.87	0.74	0.76	0.72	0.79	0.76	0.80	0.73	0.85
0.50	0.87	0.88	0.84	0.91	0.74	0.76	0.74	0.78	0.76	0.82	0.79	0.84
0.55	0.86	0.88	0.85	0.91	0.74	0.79	0.81	0.77	0.74	0.81	0.78	0.83
0.60	0.84	0.87	0.85	0.89	0.80	0.79	0.72	0.85	0.74	0.81	0.79	0.82
0.65	0.87	0.89	0.85	0.91	0.79	0.80	0.75	0.84	0.75	0.81	0.76	0.84
0.70	0.86	0.88	0.82	0.91	0.74	0.77	0.74	0.79	0.72	0.79	0.74	0.82
0.75	0.83	0.86	0.82	0.89	0.76	0.78	0.75	0.80	0.75	0.79	0.68	0.86
0.80	0.85	0.87	0.83	0.90	0.74	0.74	0.67	0.80	0.74	0.79	0.72	0.84
0.85	0.86	0.88	0.85	0.91	0.73	0.75	0.71	0.78	0.73	0.79	0.73	0.82
0.90	0.85	0.87	0.83	0.90	0.77	0.76	0.68	0.83	0.74	0.79	0.72	0.84
0.95	0.84	0.86	0.80	0.90	0.77	0.78	0.73	0.82	0.72	0.77	0.70	0.82

Tabella 2.6. Matrice score per ogni threshold

	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
Value	0.87	0.89	0.87	0.91	0.80	0.81	0.81	0.85	0.76	0.82	0.79	0.86
Threshold	0.65	0.65	0.40	0.65	0.60	0.20	0.55	0.60	0.45	0.50	0.50	0.75

Tabella 2.7. Matrice score massimi e relativi threshold

2.5 *Support Vector Machine*

Anche per questo modello abbiamo optato, per la validazione tramite k-folds selezionando un numero k = 10. Di seguito i risultati.

Threshold	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
0.05	0.43	0.46	1.00	0.09	0.45	0.46	1.00	0.01	0.40	0.42	0.99	0.05
0.10	0.47	0.55	0.97	0.28	0.49	0.54	0.99	0.16	0.42	0.46	0.96	0.13
0.15	0.55	0.66	0.87	0.52	0.54	0.62	0.95	0.34	0.51	0.63	0.93	0.43
0.20	0.71	0.81	0.88	0.76	0.64	0.73	0.88	0.60	0.64	0.75	0.84	0.69
0.25	0.78	0.85	0.87	0.84	0.67	0.76	0.90	0.63	0.66	0.77	0.83	0.72
0.30	0.78	0.85	0.85	0.84	0.75	0.80	0.83	0.77	0.66	0.76	0.80	0.74
0.35	0.84	0.86	0.79	0.90	0.76	0.80	0.82	0.78	0.74	0.81	0.80	0.82
0.40	0.82	0.84	0.78	0.89	0.76	0.81	0.82	0.79	0.73	0.80	0.76	0.82
0.45	0.82	0.84	0.78	0.88	0.74	0.77	0.73	0.80	0.77	0.80	0.70	0.87
0.50	0.86	0.85	0.75	0.92	0.79	0.80	0.75	0.84	0.75	0.78	0.67	0.86
0.55	0.85	0.84	0.72	0.92	0.78	0.79	0.75	0.83	0.79	0.79	0.63	0.89
0.60	0.88	0.83	0.68	0.93	0.81	0.79	0.71	0.86	0.84	0.81	0.62	0.92
0.65	0.90	0.84	0.68	0.95	0.83	0.79	0.68	0.89	0.83	0.80	0.60	0.92
0.70	0.89	0.84	0.69	0.94	0.80	0.74	0.56	0.89	0.85	0.77	0.50	0.94
0.75	0.91	0.83	0.63	0.96	0.84	0.77	0.60	0.91	0.87	0.76	0.44	0.96
0.80	0.94	0.81	0.57	0.98	0.90	0.75	0.49	0.96	0.92	0.76	0.42	0.98
0.85	0.90	0.83	0.64	0.95	0.98	0.65	0.22	1.00	0.97	0.72	0.28	0.99
0.90	0.94	0.71	0.31	0.99	0.98	0.63	0.17	1.00	0.96	0.69	0.22	0.99
0.95	0.96	0.66	0.17	0.99		0.55	0.00	1.00	0.97	0.64	0.09	1.00

Tabella 2.8. Matrice score per ogni threshold

	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
Value	0.96	0.86	1.00	0.99	0.98	0.81	1.00	1.00	0.97	0.81	0.99	1.00
Threshold	0.95	0.35	0.05	0.95	0.85	0.40	0.05	0.95	0.85	0.35	0.05	0.95

Tabella 2.9. Matrice score massimi e relativi threshold

2.6 *Random Forest*

Per quanto riguarda la validazione della random forest non abbiamo utilizzando né il numero arbitrario di run né il k-folds, in quanto il modello adotta il sistema di bootstrap con la divisione in bag e out of the bag samples per la validazione. Di seguito le tabelle risultanti dalla random forest con CP pari a 0 e un numero di alberi pari a 1000.

Threshold	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
0.05	0.60	0.73	0.99	0.55	0.58	0.68	0.98	0.43	0.50	0.62	0.97	0.39
0.10	0.67	0.79	0.95	0.69	0.62	0.71	0.94	0.52	0.59	0.72	0.96	0.57
0.15	0.75	0.86	0.97	0.78	0.67	0.77	0.95	0.61	0.65	0.77	0.93	0.68
0.20	0.80	0.88	0.93	0.84	0.71	0.81	0.94	0.69	0.68	0.79	0.87	0.74
0.25	0.79	0.86	0.91	0.83	0.75	0.82	0.92	0.74	0.74	0.83	0.86	0.81
0.30	0.80	0.87	0.92	0.84	0.77	0.83	0.89	0.78	0.74	0.82	0.83	0.81
0.35	0.87	0.91	0.90	0.91	0.80	0.85	0.88	0.82	0.76	0.82	0.79	0.84
0.40	0.88	0.90	0.88	0.92	0.81	0.84	0.84	0.84	0.82	0.85	0.78	0.89
0.45	0.87	0.88	0.84	0.91	0.83	0.86	0.87	0.86	0.85	0.87	0.79	0.91
0.50	0.90	0.91	0.86	0.93	0.84	0.84	0.81	0.87	0.88	0.88	0.80	0.93
0.55	0.89	0.89	0.82	0.93	0.83	0.85	0.84	0.86	0.85	0.85	0.75	0.92
0.60	0.94	0.91	0.83	0.96	0.88	0.85	0.78	0.92	0.89	0.86	0.73	0.94
0.65	0.94	0.88	0.74	0.97	0.84	0.83	0.77	0.88	0.89	0.85	0.72	0.94
0.70	0.96	0.87	0.71	0.98	0.89	0.83	0.72	0.93	0.89	0.84	0.67	0.95
0.75	0.96	0.84	0.64	0.98	0.94	0.82	0.65	0.97	0.94	0.86	0.69	0.97
0.80	0.95	0.83	0.62	0.98	0.98	0.84	0.65	0.99	0.91	0.81	0.58	0.96
0.85	0.99	0.84	0.62	0.99	0.96	0.78	0.54	0.98	0.95	0.80	0.52	0.98
0.90	1.00	0.83	0.58	1.00	0.98	0.77	0.50	0.99	0.98	0.79	0.47	0.99
0.95	1.00	0.80	0.50	1.00	0.98	0.71	0.37	0.99	0.94	0.75	0.39	0.98

Tabella 2.10. Matrice score per ogni threshold

	Giugno				Luglio				Agosto			
	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity
Value	1.00	0.91	0.99	1.00	0.98	0.86	0.98	0.99	0.98	0.88	0.97	0.99
Threshold	0.90	0.60	0.05	0.90	0.80	0.45	0.05	0.95	0.90	0.50	0.05	0.90

Tabella 2.11. Matrice score massimi e relativi threshold

Per ogni mese abbiamo osservato il peso delle variabili del decremento dell'accuracy e dell'indice di Gini, qualora ogni variabile venisse rimossa dal modello.

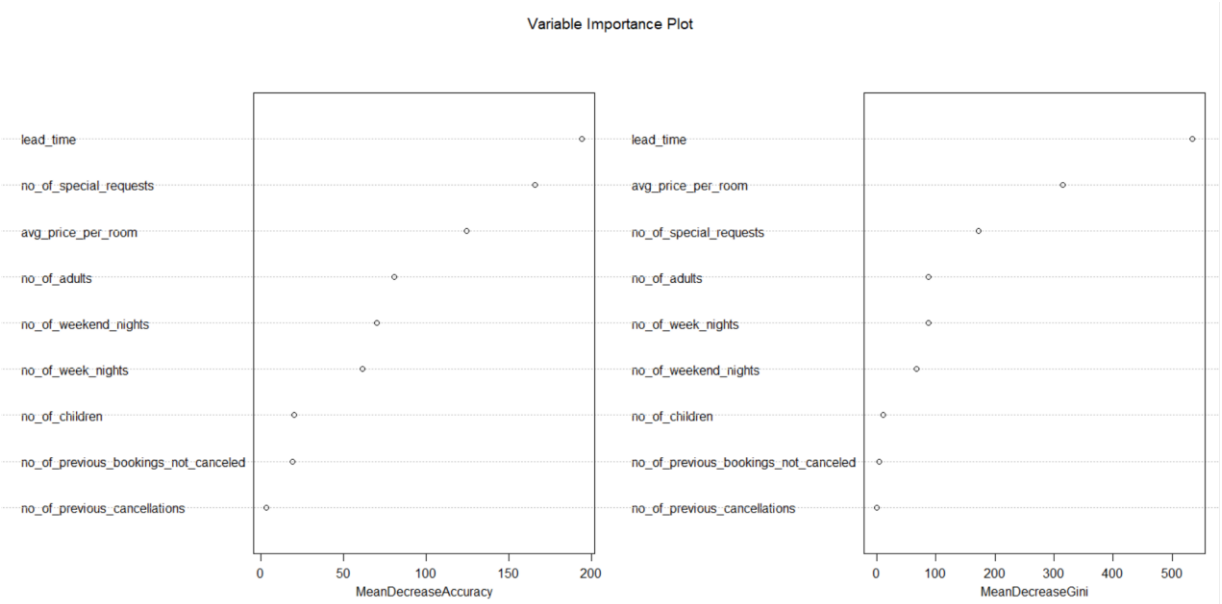

Figura 2.3. Variable Importance Plot Giugno

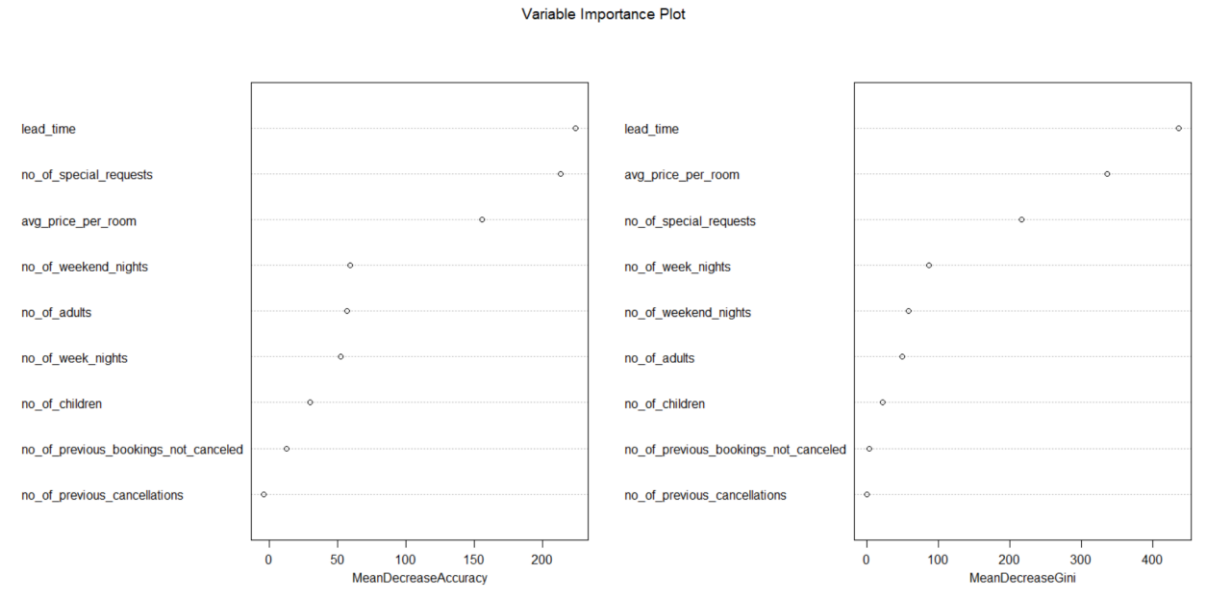

Figura 2.4. Variable Importance Plot Luglio

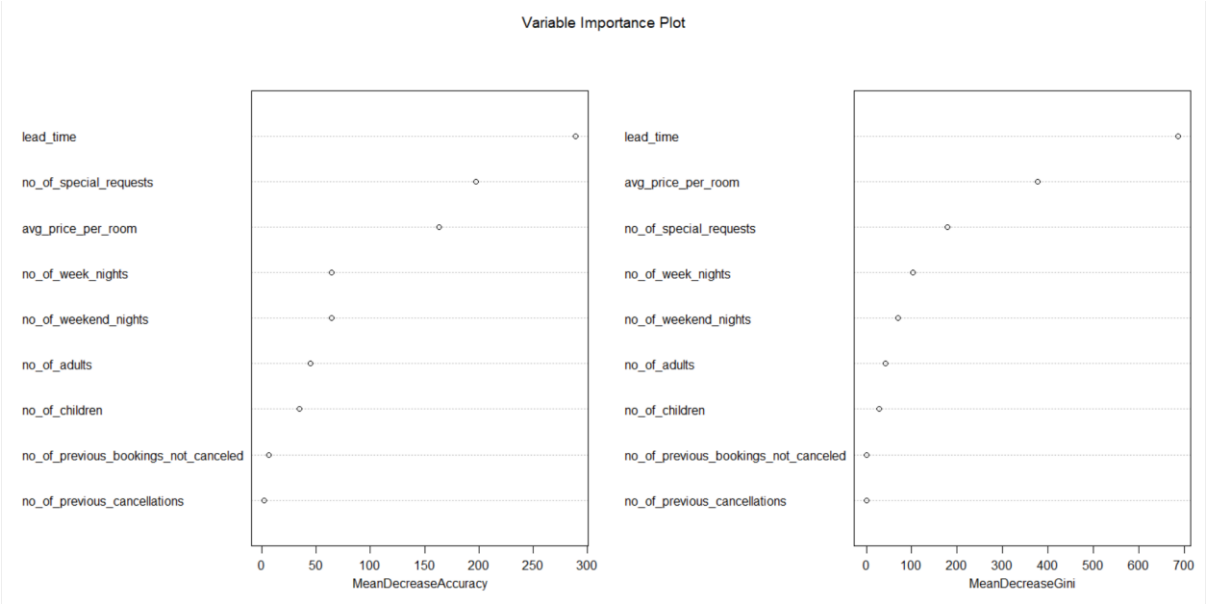

Figura 2.5. Variable Importance Plot Agosto

Notiamo che le variabili più importanti risultano Lead time, Number of special requests e Average price per room.

Inoltre abbiamo misurato l'andamento dell'error rate. Si può notare che occorrono meno di duecento alberi per poter raggiungere un valore abbastanza stabile. Per il mese di giugno il valore dell'errore per la classificazione della cancellazione si attesta intorno al 13%. Il mese di luglio presenta un valore del 17%. Infine, per quanto riguarda agosto circa il 22%.

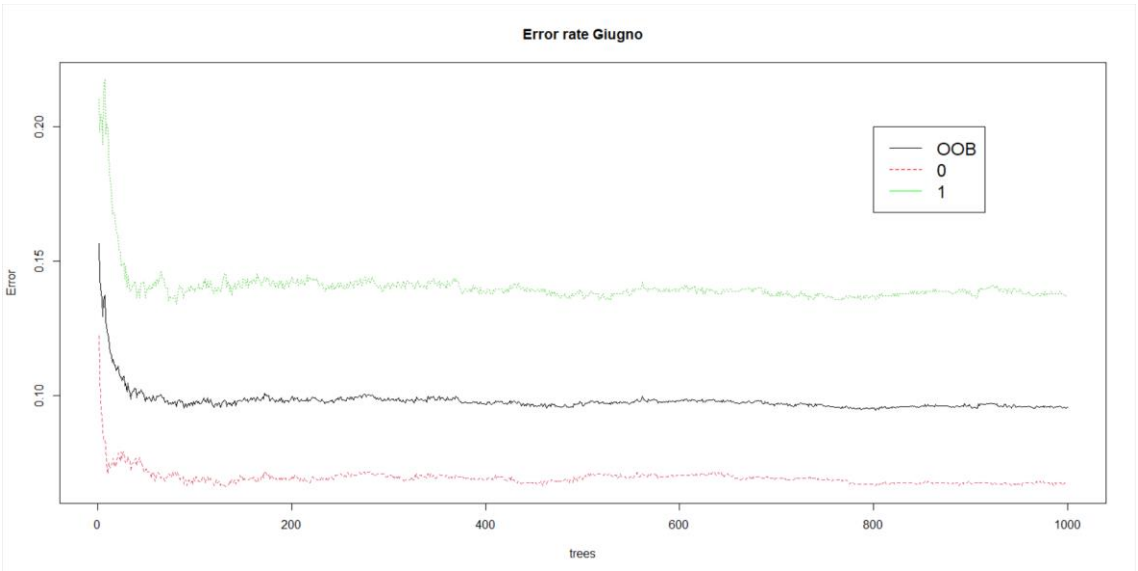

Figura 2.6. Error rate plot giugno

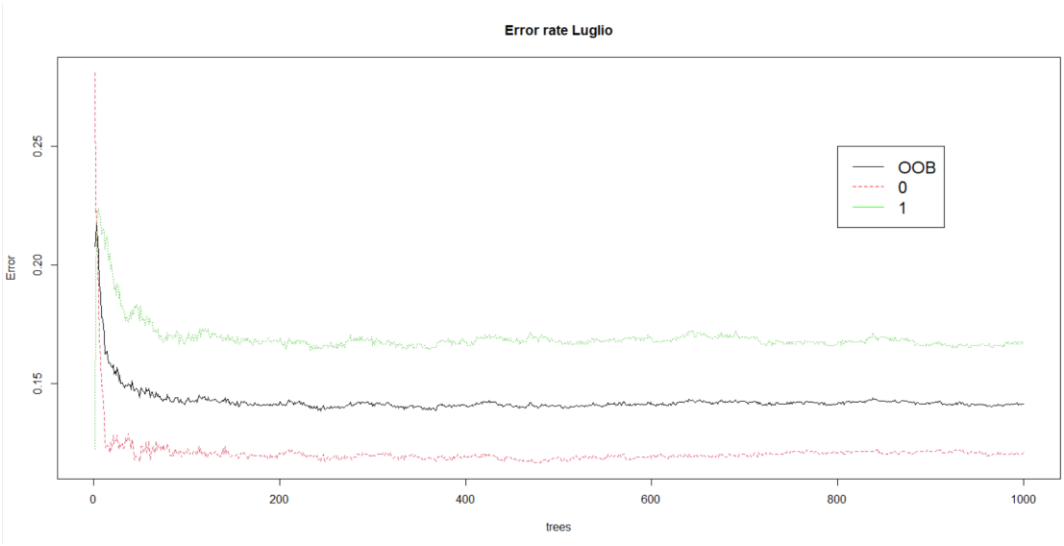

Figura 2.7. Error rate plot luglio

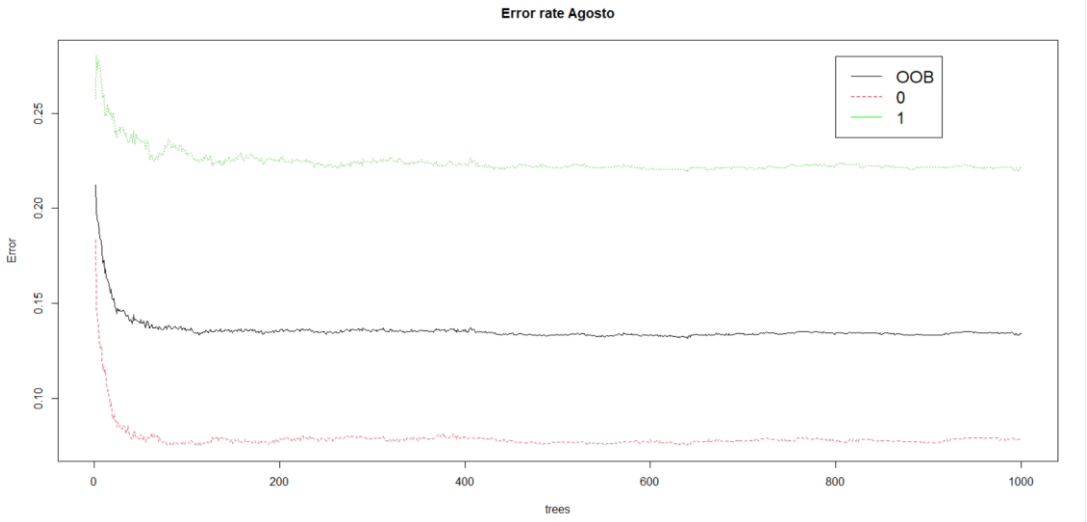

Figura 2.7. Error rate plot agosto

CAPITOLO 3. CONFRONTO RISULTATI OTTENUTI E CONCLUSIONI

3.1 Precision score

Possiamo osservare in tabella che i casi di cancellazione della prenotazione risultano essere facilmente individuabili per ogni mese (valori maggiori o uguali del 90%). Occorre notare, soprattutto, il modello che meglio discrimina i casi di cancellazione, ovvero la Random forest, dal quale si ottiene un valore del 99% di precision in media.

Modello	Giugno	Luglio	Agosto	Media modello
LDA	0.97	1.00	0.90	0.95
GLM	0.94	0.98	0.91	0.94
KNN	0.87	0.80	0.76	0.81
SVM	0.96	0.98	0.97	0.97
RF	1.00	0.98	0.98	0.99
Media mese	0.95	0.95	0.90	

Figura 2.6. Tabella riassuntiva Precision

3.2 Accuracy score

In tabella possiamo leggere che l'accuracy per ogni mese si attesta intorno all'80%, con un picco all'85% per giugno. Il modello che discrimina meglio i casi risulta essere nuovamente la Random forest, dal quale si ottiene un valore dell'88% in media.

Modello	Giugno	Luglio	Agosto	Media modello
LDA	0.79	0.73	0.77	0.76
GLM	0.81	0.74	0.77	0.77
KNN	0.89	0.81	0.82	0.84
SVM	0.86	0.81	0.81	0.82
RF	0.91	0.86	0.88	0.88
Media mese	0.85	0.79	0.81	

Figura 2.6. Tabella riassuntiva Accuracy

3.3 Sensitivity e Specificity score

Abbiamo scelto di commentare i due score insieme, data il loro stretto rapporto.
Possiamo notare come la specificity risulti leggermente migliore per quanto riguarda i singoli mesi.
Nonostante ciò, mediante la maggior parte dei modelli si riescono ad ottenere valori di ambedue gli score elevati (intorno al 100%).

Modello	Giugno	Luglio	Agosto	Media modello
LDA	0.99	1.00	1.00	0.99
GLM	0.99	1.00	0.99	0.99
KNN	0.87	0.81	0.79	0.82
SVM	1.00	1.00	0.99	1.00
RF	0.99	0.98	0.97	0.98
Media mese	0.97	0.96	0.95	

Figura 2.6. Tabella riassuntiva Sensitivity

Modello	Giugno	Luglio	Agosto	Media modello
LDA	1.00	1.00	1.00	1.00
GLM	1.00	1.00	1.00	1.00
KNN	0.91	0.85	0.86	0.87
SVM	0.99	1.00	1.00	1.00
RF	1.00	0.99	0.99	1.00
Media mese	0.98	0.97	0.97	

Figura 2.6. Tabella riassuntiva Specificity

3.4 Conclusioni

L'obiettivo dello studio riguardava la validazione di un modello tanto più efficace possibile nell'individuazione delle prenotazioni che vengono cancellate.
Osservando i risultati ottenuti, si può affermare che il modello di Random forest risulti il migliore, in quanto per la quasi totalità degli score, e nello specifico per l'accuracy, presenta il valore maggiore.
L'accuracy per i mesi estivi, rispettivamente del 91%, 86% e 88% risulta essere particolarmente soddisfacente, considerando la complessità del modello, concernente nove variabili.
In merito a questo modello, abbiamo anche individuato le variabili che maggiormente influiscono nella scelta di cancellare o meno la prenotazione (lead time, number of special requests, average price per room), così da poterle attenzionare nel migliore dei modi.