Progetto di Machine Learning: Modelli di Classificazione

Corso di Laurea Magistrale in Informatica (LM - 18)
Appello 15 Febbraio 2023

Bancora Davide – M. 905588

Donato Benedetta – M. 905338

Dubini Emanuele – M. 904078

Contesto ed obiettivi

L’obiettivo del progetto ha lo scopo di identificare i modelli di classificazione che possano contribuire a prevedere se un cliente cancellerà o meno una prenotazione effettuata presso un determinato hotel.

Il Dataset

- È composto da 19 attributi e circa 36 mila istanze.
- Non presenta valori mancanti e/o nulli.
- L'attributo ***Booking_ID*** non ha alcun valore predittivo per il modello di classificazione ed è stato rimosso.

```
> # Identificazione dei valori mancanti
> missing_values <- colSums(is.na(data))
> print(missing_values)
                    Booking_ID                   no_of_adults
                             0                              0
                no_of_children                no_of_weekend_nights
                             0                              0
               no_of_week_nights                type_of_meal_plan
                             0                              0
        required_car_parking_space               room_type_reserved
                             0                              0
                     lead_time                      arrival_year
                             0                              0
                arrival_month                      arrival_date
                             0                              0
           market_segment_type                    repeated_guest
                             0                              0
 no_of_previous_cancellations no_of_previous_bookings_not_canceled
                             0                              0
             avg_price_per_room            no_of_special_requests
                             0                              0
                booking_status
                             0
```

Conversioni covariate

- Alcune delle covariate sono state convertite in tipo *factor*: booking_status, type_of_meal_plan, market_segment_type, room_type_reserved, required_car_parking_space, repeated_guest. Assumevano valori appartenenti a un insieme di categorie.
- Le altre covariate sono rimaste invariate (tipo interno) in quanto utilizzate per rappresentare quantità numerabili.

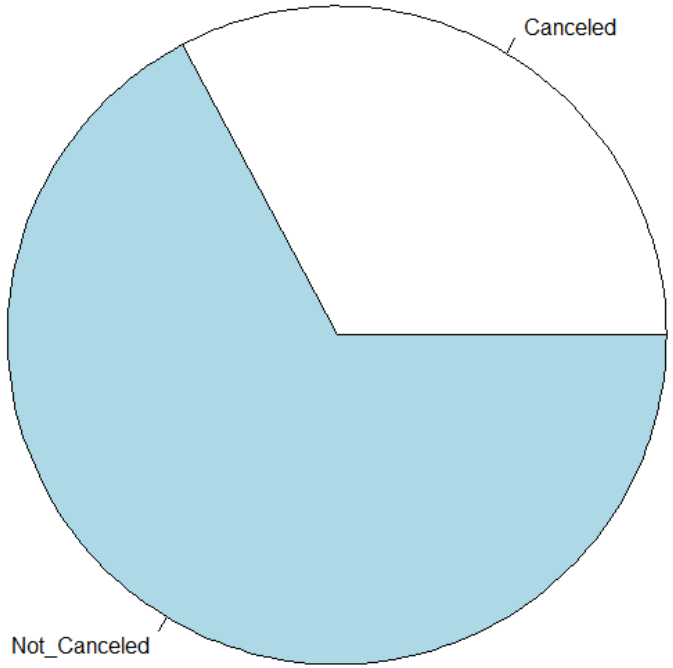
```
> str(data)
'data.frame':   36275 obs. of  18 variables:
 $ no_of_adults                      : int  2 2 1 2 2 2 2 2 3 2 ...
 $ no_of_children                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_weekend_nights              : int  1 2 2 0 1 0 1 1 0 0 ...
 $ no_of_week_nights                 : int  2 3 1 2 1 2 3 3 4 5 ...
 $ type_of_meal_plan                 : Factor w/ 4 levels "Meal Plan 1",..: 1 4 1 1 4 2 1 1 1 1 ...
 $ required_car_parking_space        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ room_type_reserved                : Factor w/ 7 levels "Room_Type 1",..: 1 1 1 1 1 1 1 4 1 4 ...
 $ lead_time                         : int  224 5 1 211 48 346 34 83 121 44 ...
 $ arrival_year                      : int  2017 2018 2018 2018 2018 2018 2017 2018 2018 2018 ...
 $ arrival_month                     : int  10 11 2 5 4 9 10 12 7 10 ...
 $ arrival_date                      : int  2 6 28 20 11 13 15 26 6 18 ...
 $ market_segment_type               : Factor w/ 5 levels "Aviation","Complementary",..: 4 5 5 5 5 5 5 5 4 5
 $ repeated_guest                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ no_of_previous_cancellations      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ avg_price_per_room                : num  65 106.7 60 100 94.5 ...
 $ no_of_special_requests            : int  0 1 0 0 0 1 1 1 1 3 ...
 $ booking_status                    : Factor w/ 2 levels "Canceled","Not_Canceled": 2 2 1 1 1 1 2 2 2 2 ...
>
```

Analisi delle covariate

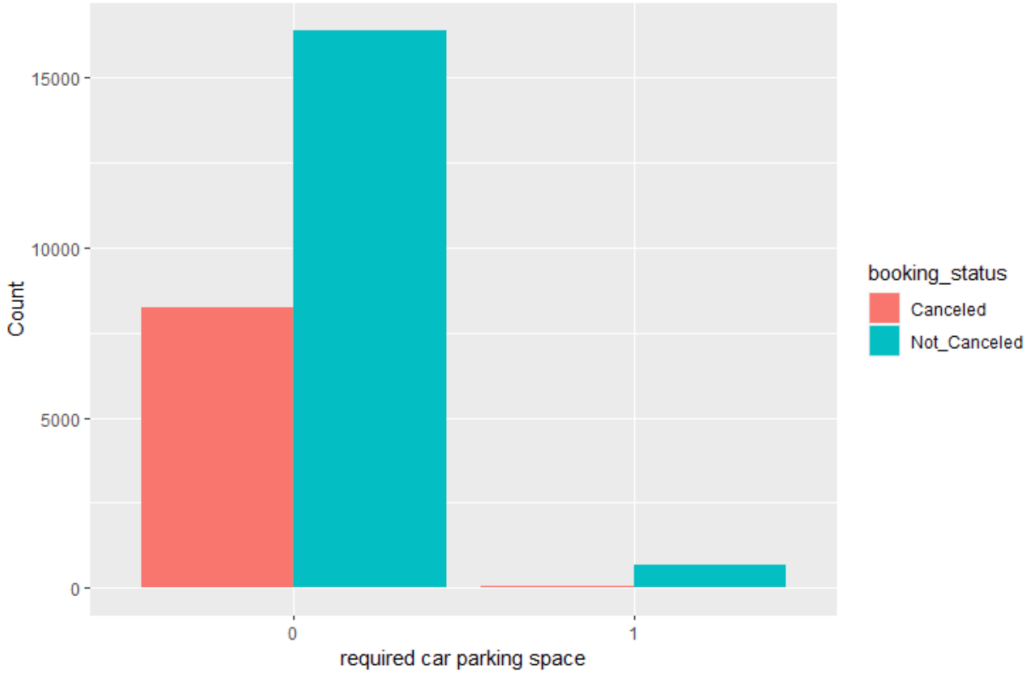
1. Le prenotazioni non cancellate sono superiori alle cancellate.

2. Le prenotazioni dei clienti che richiedono un posto auto auto hanno meno probabilità di essere cancellate.

1.



2.

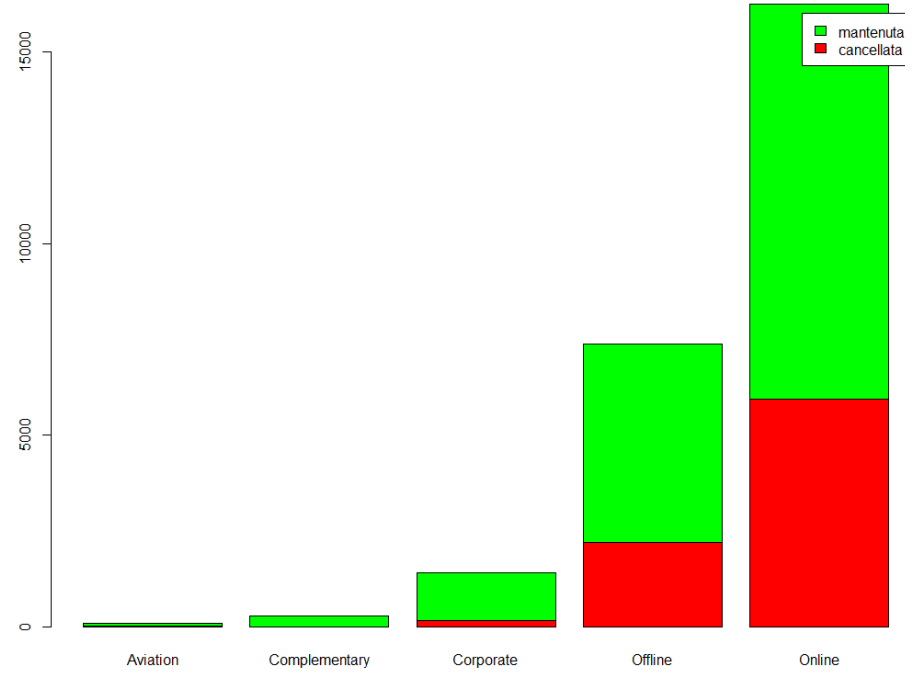


Analisi delle covariate

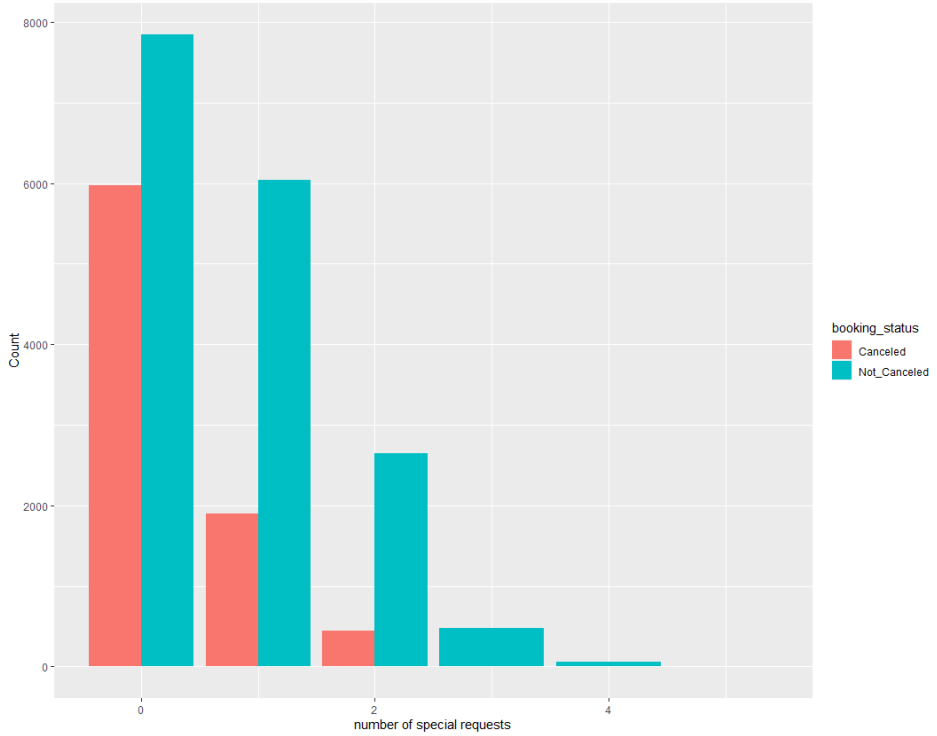
1. Le prenotazioni Complementary hanno una probabilità di cancellazione pari allo 0%.

2. Le prenotazioni che contengono 3 o piú richieste speciali non vengono mai cancellate.

1.



2.



Modelli di Machine Learning scelti

Sono state implementate 3 differenti tecniche di classificazione:

Alberi decisionali

- Permettono la gestione di dati numerici e categorici.
- Individuano relazioni tra diverse variabili con il target creando una gerarchia di decisioni consentendo di prevedere la classe di appartenenza delle istanze.

Random Forrest

- Modello più preciso rispetto a quello che si otterrebbe utilizzando un singolo albero di decisione.
- Sono robusti ai rumori presenti nei dati.
- Capaci di gestire problemi di overfitting.

Reti Neurali

- Le reti neurali possono rappresentare relazioni complesse nei dati, grazie alla loro architettura di più strati di neuroni.
- Possono essere progettate per risultare robuste ai rumori nei dati.

Alberi Decisionali

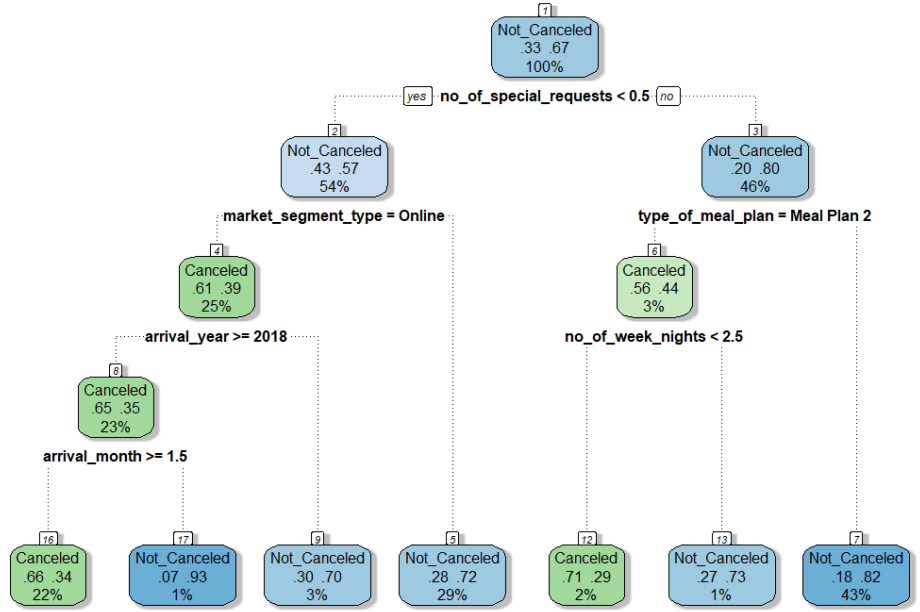
Inizialmente, è stato ottenuto tale albero decisionale.

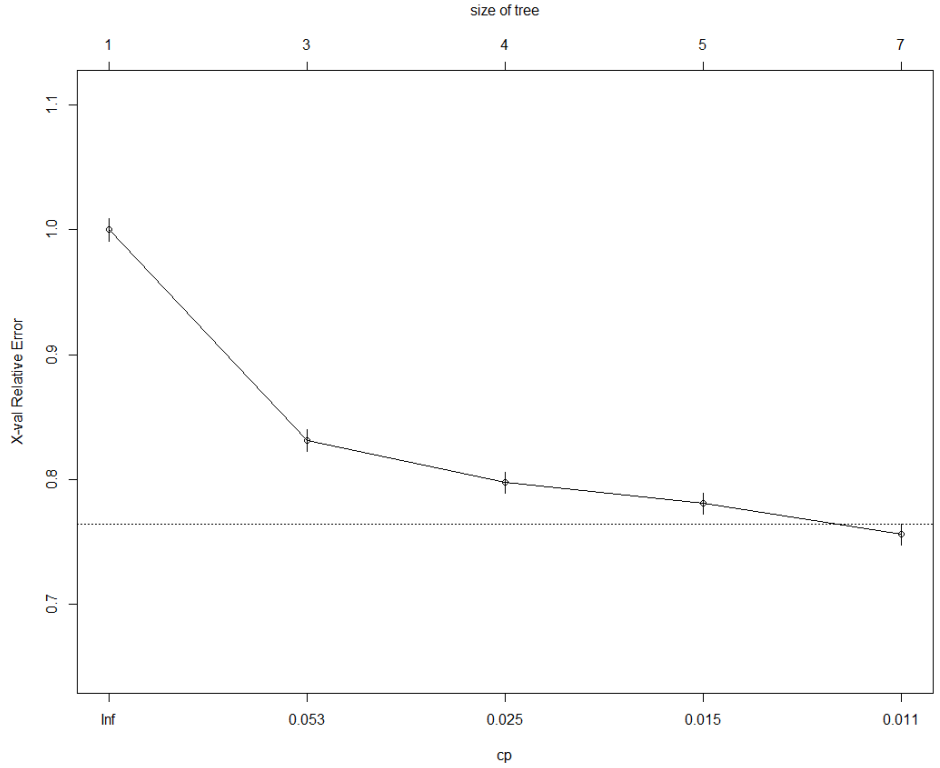
Accuratezza: 75,28%

Kappa: 0,398

Tali valori non risultano essere molto soddisfacenti.

Per evitare overfitting, si analizza il valore di cp impostandolo da 0,011 a 0,015.



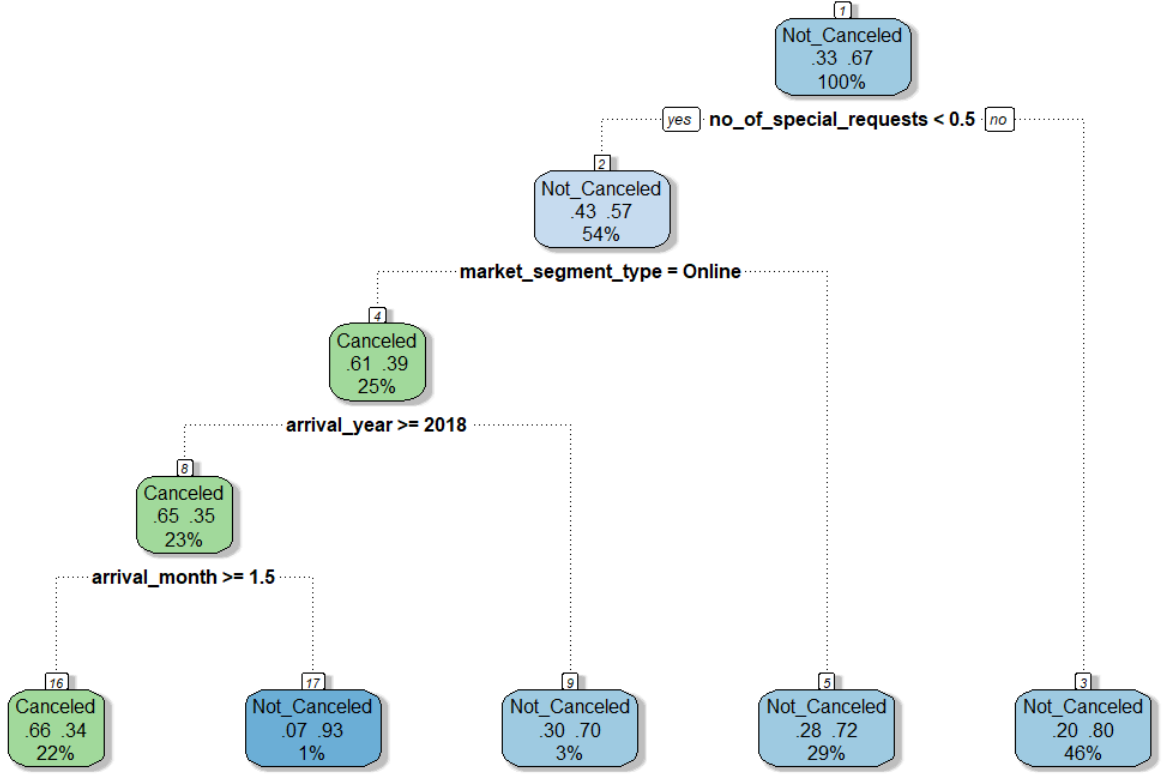


Alberi Decisionali

Albero decisionale con cp: 0,015.

L'accuratezza e Kappa sono entrambi diminuiti di poco, ma è aumentata la leggibilità dell'albero.

Previene overfitting.



Valori ottenuti considerando la classe positiva "Canceled"

```
Precision : 0.6642
   Recall : 0.4511
       F1 : 0.5373
```

Valori ottenuti considerando la classe positiva "Not_ Canceled"

```
Precision : 0.7687
   Recall : 0.8889
       F1 : 0.8244
```

```
                Accuracy : 0.7455
                  95% CI : (0.7372, 0.7536)
     No Information Rate : 0.6724
     P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.3704
```

Random Forest

Il modello è stato costruito utilizzando 500 alberi e ogni albero utilizza 3 variabili scelte in modo casuale per suddividere i nodi.

Per ottenere risultati migliori sono state normalizzate alcune variabili.

Accuratezza: 89,35%

Kappa: 0,75

Buon livello di precisione.

Matrice di confusione

Prediction	Reference Canceled	Not_Canceled
Canceled	2824	418
Not_Canceled	741	6899

```
          Accuracy : 0.8935
            95% CI : (0.8875, 0.8992)
No Information Rate : 0.6724
P-Value [Acc > NIR] : < 2.2e-16

             Kappa : 0.7525
```

Random Forest

I risultati mostrano che il modello ha una buona capacità di classificare correttamente sia i casi "*Canceled*" che i casi "*Not_canceled*".

Confrontando questi valori con quanto ottenuto analizzando il modello *Decision Tree* deduciamo che si ha un **notevole miglioramento.**

Valori ottenuti considerando la classe positiva "Canceled"

```
Precision : 0.8675
   Recall : 0.7935
       F1 : 0.8289
```

Valori ottenuti considerando la classe positiva "Not_ Canceled"

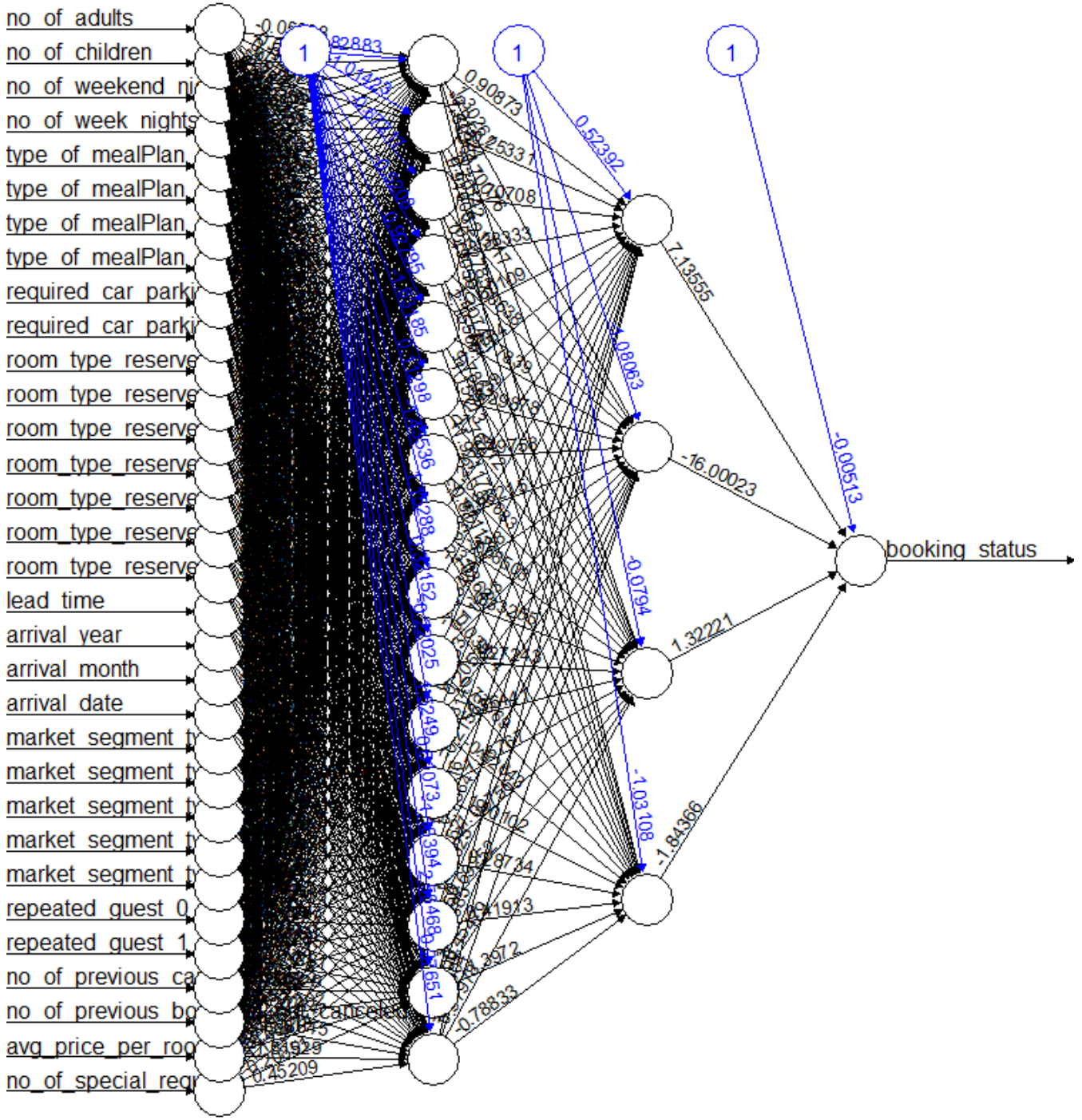
```
Precision : 0.9034
   Recall : 0.9410
       F1 : 0.9218
```

Rete Neurale

La rete neurale utilizza 16 nodi nel primo strato e 4 nel secondo strato.

È stato necessario applicare il processo di binarizzazione per gli attributi categorici.

Accuratezza: 84,33%
Kappa: 0,63
Precision: 82,30%



Rete Neurale

Valori di performance:

- *Precision:* 82,30%
- *Recall:* 81,17%
- *F1-measure:* 81,68%

Possiamo dedurre che è un buon modello di classificazione.

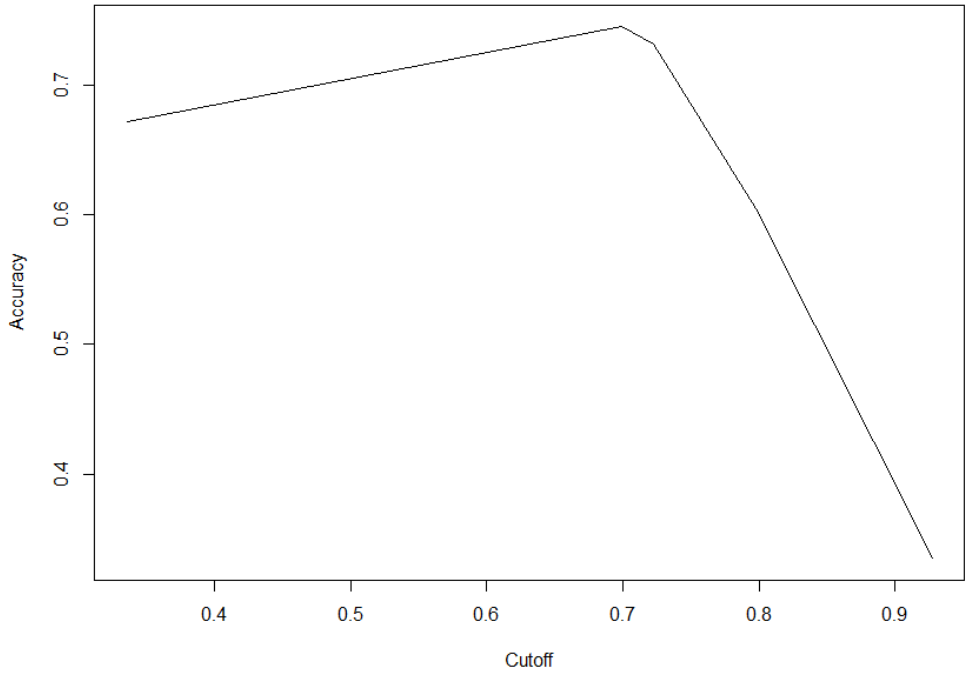
```
                  Accuracy : 0.8433
                    95% CI : (0.8117, 0.8715)
No Information Rate : 0.68
P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.634
```

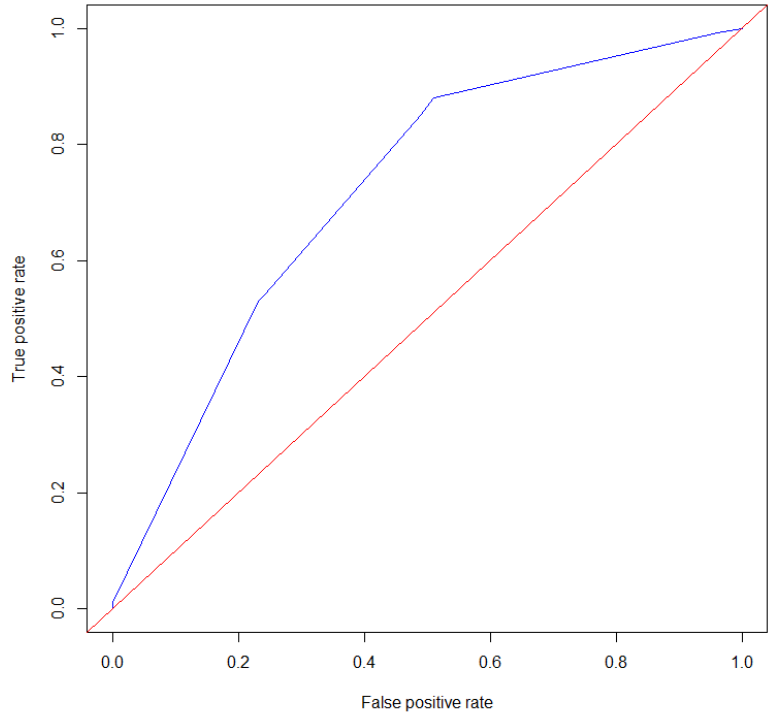
Misure di Performance – Alberi Decisionali

Si ottiene un valore di accuratezza migliore quando il *cutoff* è pari a circa 0,7.

Un modello ideale avrà una *curva ROC* che si avvicina il più possibile al punto in alto a sinistra della figura.

Il seguente modello di predizione non è molto preciso nella scelta del target corretto.





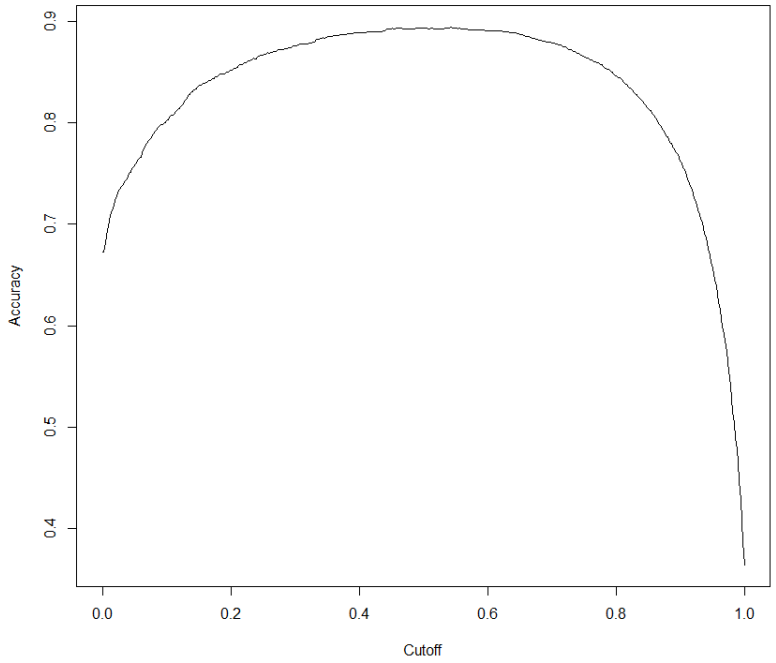
Misure di Performance – Random Forest

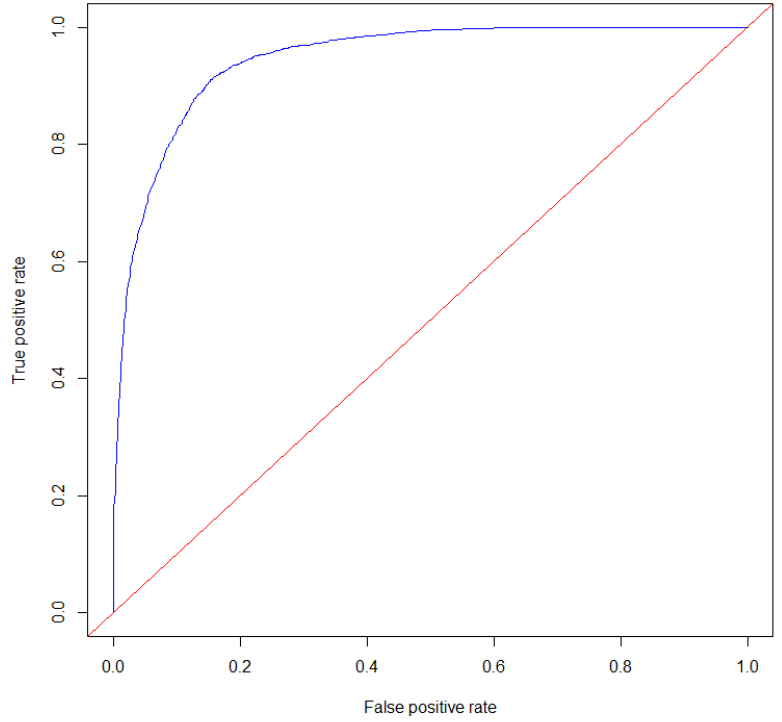
Il valore di *cutoff* ottimale è attorno a 0,5.

L'accuratezza del modello assume valore pari a circa 0,9.

Curva ROC: la curva si avvicina molto a quella ottimale.

Il modello Random Forest ha alta precisione e ottima capacità di discriminazione.





Misure di Performance – Rete Neurale

Il valore di *cutoff* ottimale si aggira intorno a 0,5 e corrisponde ad un'accuratezza pari a 0,84.

La curva ROC si avvicina molto a quella ottimale.