Customer satisfaction from Booking

MAURIZIO ROMANO, LUCA FRIGAU, GIULIA CONTU

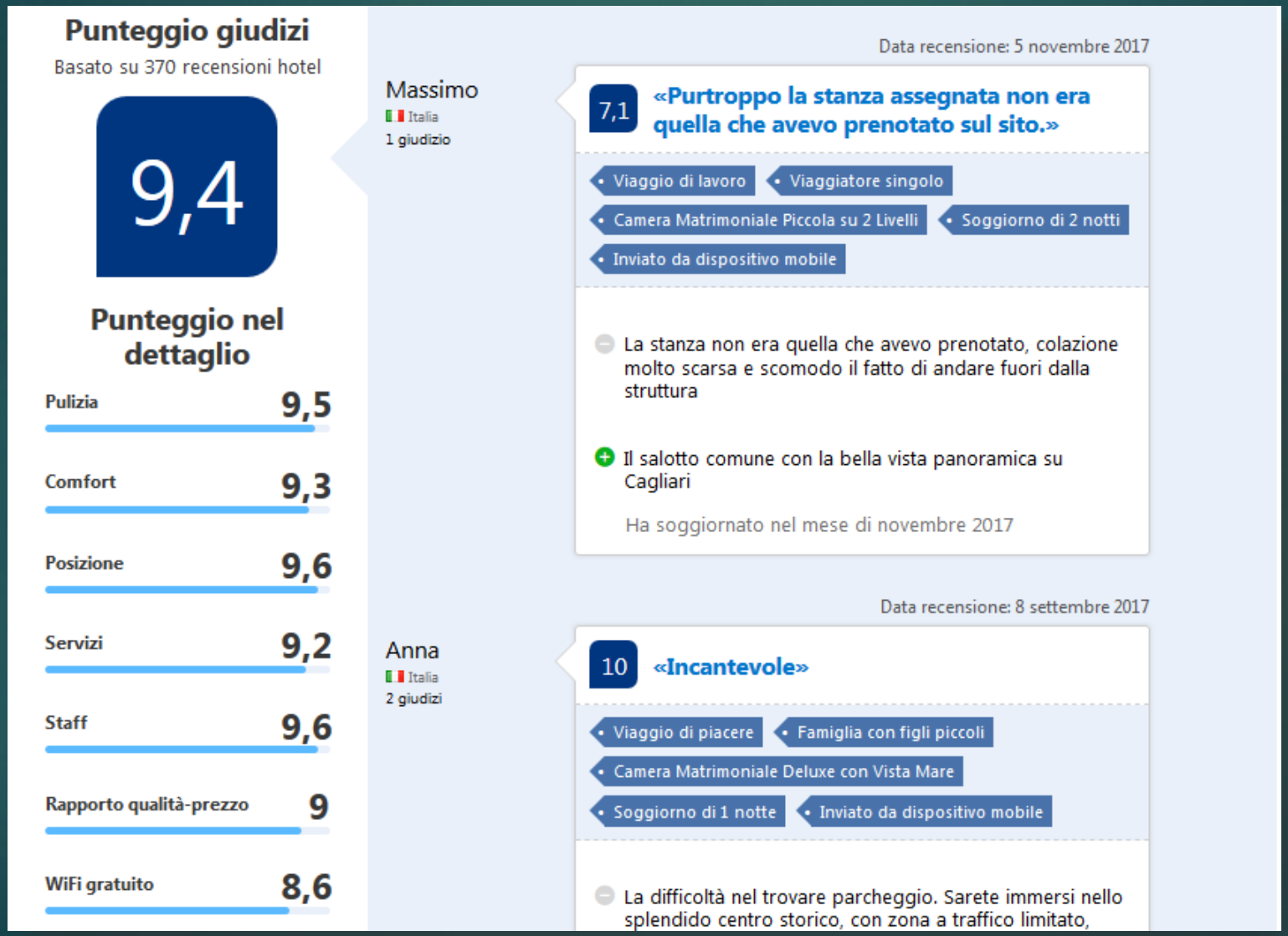
FRANCESCO MOLA, CLAUDIO CONVERSANO

UNIVERSITÀ DEGLI STUDI DI CAGLIARI

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI

Progetto P.I.A. "Realizzazione di una piattaforma ICT a supporto del settore turistico",
RAS Programmazione Unitaria 2007/2013 - P.O. FESR 2007/2013,
Interventi a sostegno della competitività e dell'innovazione: **Resp. Scientifico Prof. F. Mola**

Introduzione



Punteggio giudizi

Basato su 370 recensioni hotel

9,4

Punteggio nel dettaglio

Pulizia	9,5
Comfort	9,3
Posizione	9,6
Servizi	9,2
Staff	9,6
Rapporto qualità-prezzo	9
WiFi gratuito	8,6

Data recensione: 5 novembre 2017

Massimo
Italia
1 giudizio

7,1 **«Purtroppo la stanza assegnata non era quella che avevo prenotato sul sito.»**

- Viaggio di lavoro
- Viaggiatore singolo
- Camera Matrimoniale Piccola su 2 Livelli
- Soggiorno di 2 notti
- Inviato da dispositivo mobile

La stanza non era quella che avevo prenotato, colazione molto scarsa e scomodo il fatto di andare fuori dalla struttura

Il salotto comune con la bella vista panoramica su Cagliari

Ha soggiornato nel mese di novembre 2017

Data recensione: 8 settembre 2017

Anna
Italia
2 giudizi

10 **«Incantevole»**

- Viaggio di piacere
- Famiglia con figli piccoli
- Camera Matrimoniale Deluxe con Vista Mare
- Soggiorno di 1 notte
- Inviato da dispositivo mobile

La difficoltà nel trovare parcheggio. Sarete immersi nello splendido centro storico, con zona a traffico limitato,

I dati di Booking

- **619 Strutture totali**

Nome Struttura	Tipologia	Cap	Comune/Località	…
Struttura 1	Extralberghiere	09044	Sant' Isidoro	…
Struttura 2	3 Stelle	09049	Villasimius	…
Struttura 3	3 Stelle	07013	Mores	…
Struttura 4	4 Stelle	09123	Cagliari	…
…	…	…	…	…

- **66237 Recensioni complessive scindibili in 106800 commenti totali (ITA+ENG)**

Nome Struttura	ID Commento	ID Review	Commento	Neg-Pos	Score	…
Struttura 1	1	1	christina was the best…	Pos	10.0	…
Struttura 1	2	2	we booked an appartment…	Pos	10.0	…
Struttura 1	3	3	we travelled into cagliari…	Pos	9.2	…
Struttura 1	4	4	it was fantastic	Pos	10.0	…
…	…	…	…	…	…	…

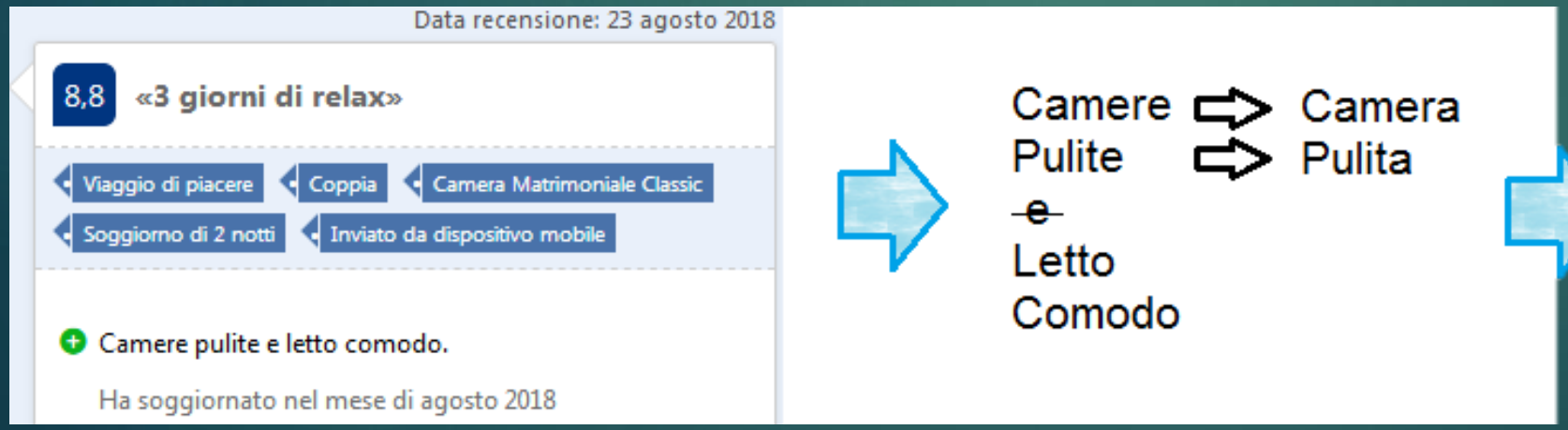
Data Cleaning

Word	New Word
Access	Accesso
Accesso	
Accessori	
Accogliente	
Accoglienti	Accogliente
Accoglienza	Accogliente
…	…

- **Cleaning**: sono state raggruppate fra loro le parole con il medesimo significato e sono stati ripuliti i commenti da congiunzioni, punteggiatura, numeri ed altre *stopwords*.
- **Accorpamento**: consiste nel sostituire nei commenti originali ogni parola accorpabile con una comune a ciascuna per significato.

Modello e strategia di analisi

- **Naive Bayes** per stimare la probabilità che un commento sia classificato come "Negativo" o "Positivo"
- **Cross Validation**
- Creazione del **Bag Of Words**



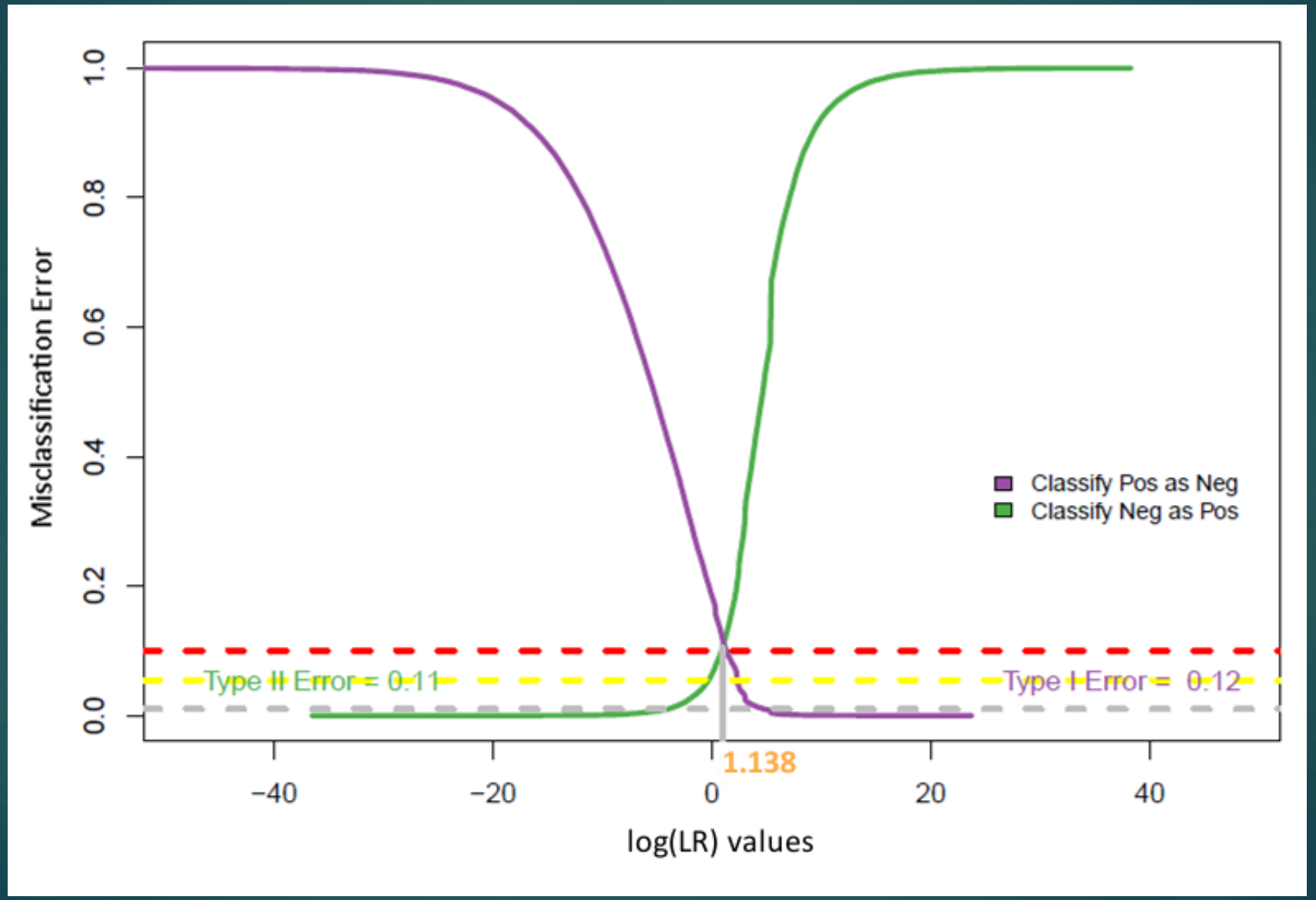
Camera	Pulita	Letto	Comodo
1	1	1	1

Classificazione Naive Bayes*

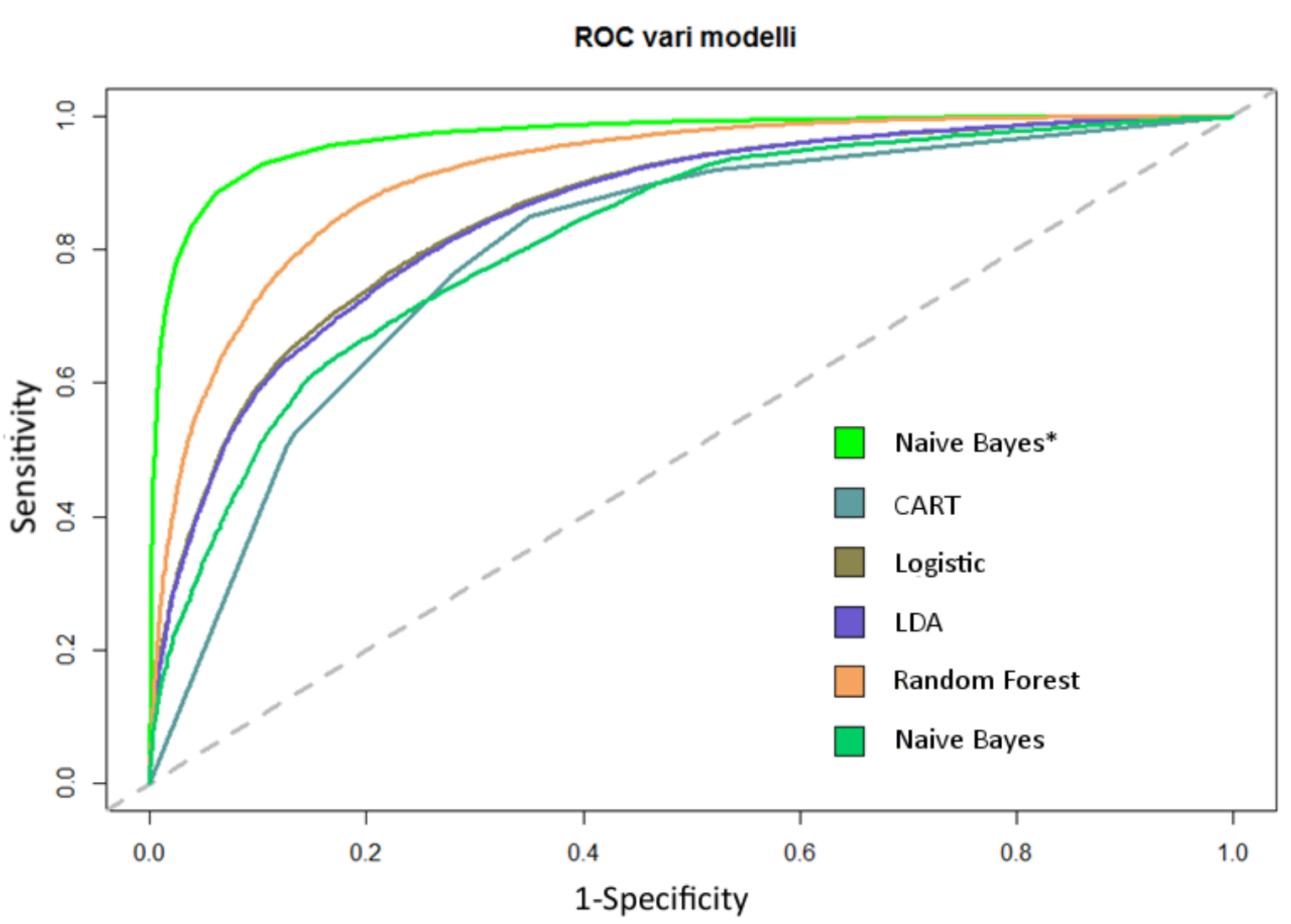
- La stima di P(neg | “contenuto”) si basa sul rapporto fra verosimiglianze LR:
- $LR = \frac{P(neg \mid "contenuto")}{P(pos \mid "contenuto")}$
- Confronto del valore di log(LR) approssimato con un valore soglia τ
  - Log(LR) > τ → il messaggio è classificato “Negativo”
  - Log(LR) ≤ τ → il messaggio è classificato “Positivo”

	"Assolutamente"	"Mare"	"Facilmente"	…
P(neg)	0,011	0,026	0,002	…
P(pos)	0,007	0,075	0,005	…
Log(LR) – Parola presente	0,411	-1,077	-1,006	…
Log(LR) – Parola assente	-0,004	0,052	0,003	…

Scelta di τ


Misclassification Error
1.0
0.8
0.6
0.4
0.2
0.0
Classify Pos as Neg
Classify Neg as Pos
Type II Error = 0.11
Type I Error = 0.12
1.138
−40
−20
0
20
40
log(LR) values

Benchmarking dei classificatori



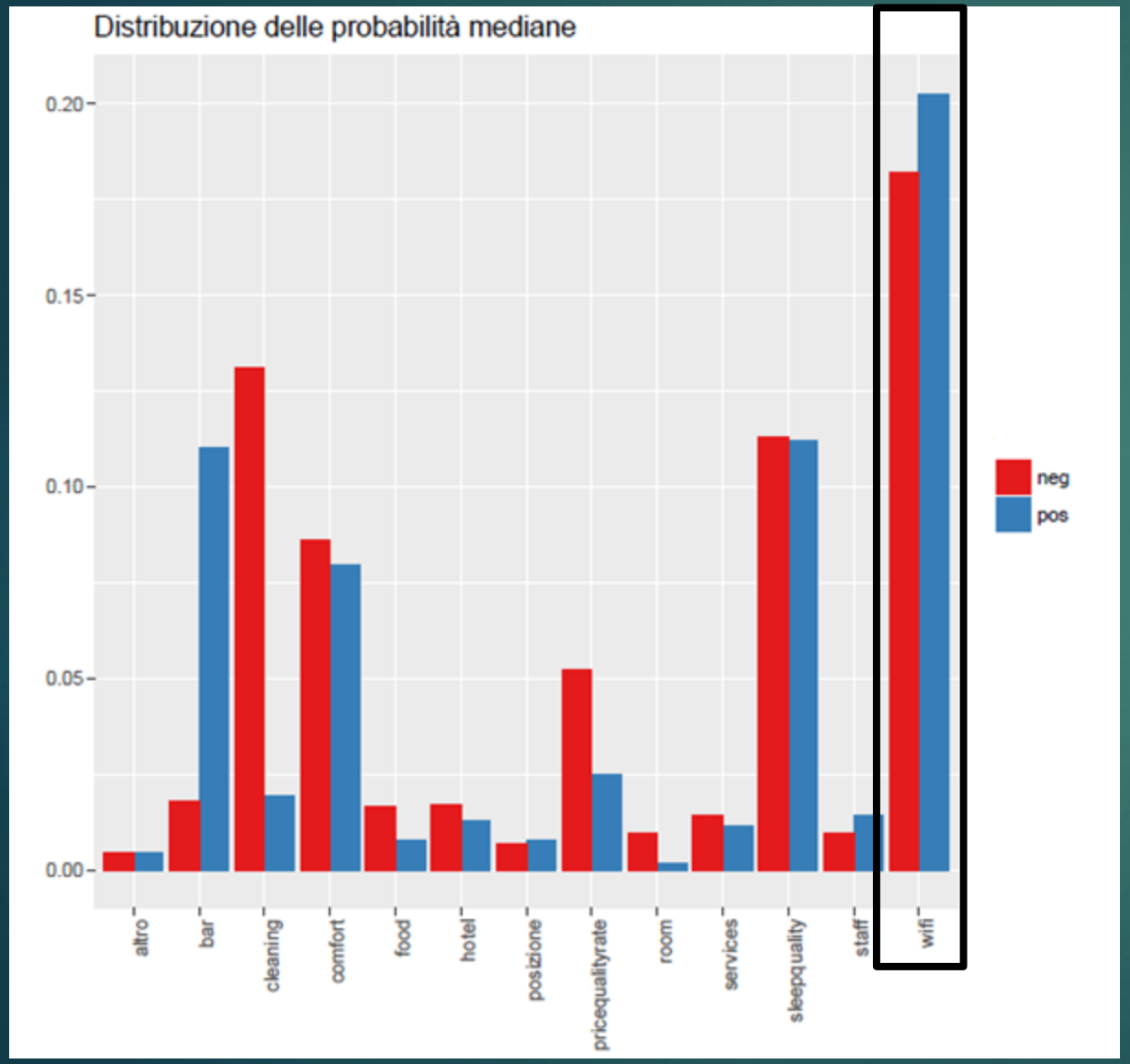
Model	Accuracy	F1 score	Matthews Correlation Coefficient
Naive Bayes*	**0,911**	**0,926**	**0,813**
Logistic	0,850	0,877	0,361
Random Forest	0,811	0,849	0,303
Naive Bayes (e1071)	0,806	0,834	0,390
Naive Bayes (klaR)	0,806	0,834	0,390
CART	0,768	0,815	0,272
LDA	0,764	0,816	0,246

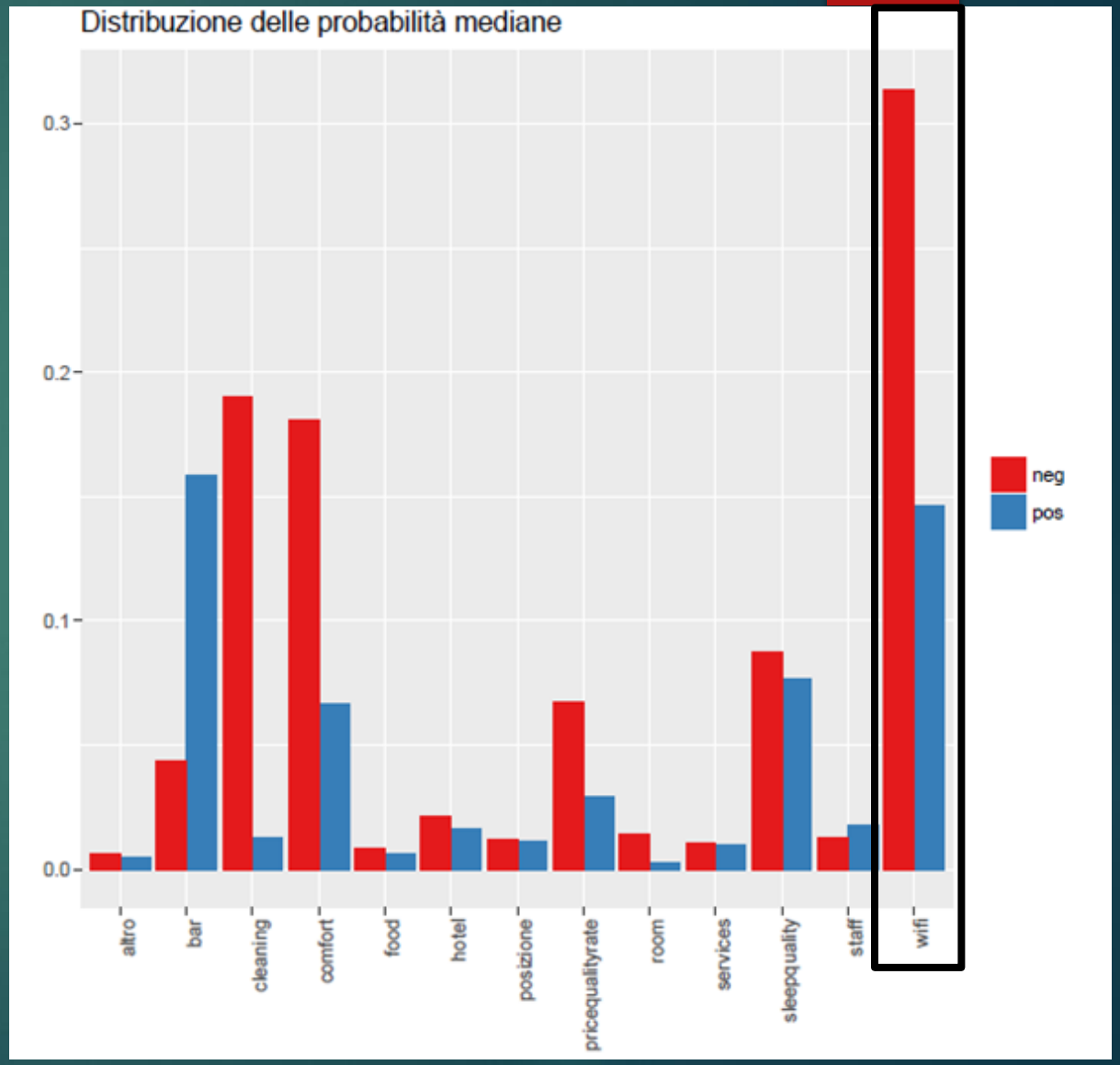
Interpretazione risultati classificazione Naive Bayes*

Word	Category
Colazione	Food
Ristorante	Food
Bread	Food
Cakes	Food
Mangiare	Food
Conto	Price-quality rate
Caro	Price-quality rate
Pagamento	Price-quality rate
Pay	Price-quality rate
Gestore	Staff
Stintino	Position
Orosei	Position
…	…

- Macro-categorie:
  1. **Bar**
  2. **Cleaning**
  3. **Comfort**
  4. **Food**
  5. **Hotel**
  6. **Position**
  7. **Price-quality rate**
  8. **Room**
  9. **Services**
  10. **Sleep quality**
  11. **Staff**
  12. **Wifi**
  13. **Other**

Strutture Cagliari vs Ex Carbonia-Iglesias





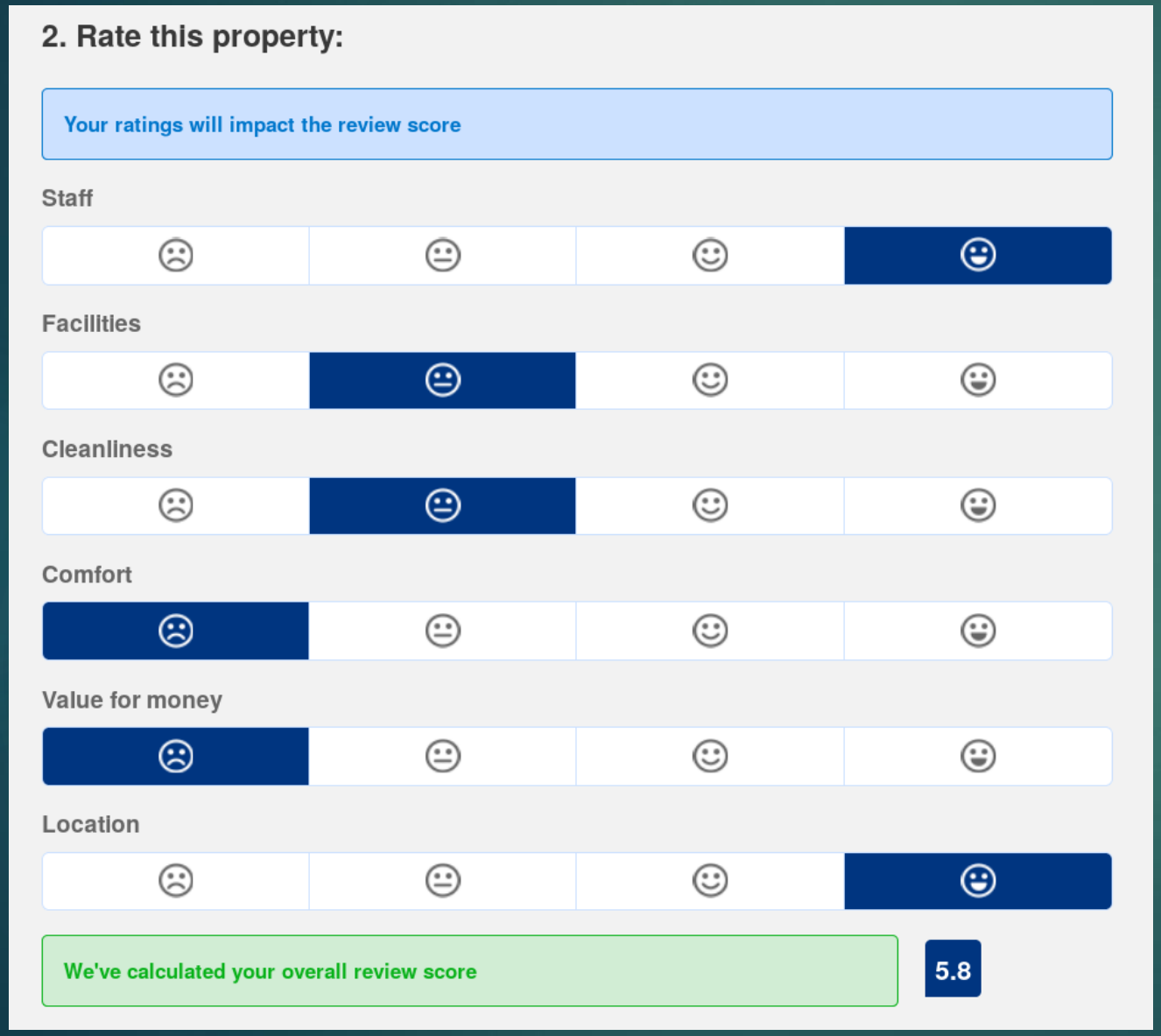
Previsione *score* di Booking

- Per procedere con l'applicazione di un modello di previsione (Y = "Score di Booking") sono state create delle variabili a supporto:
  - **review_position, review_bar, review_cleaning, …. , review_services ….**

    valori di log(LR) per ciascuna delle 13 categorie rilevate in precedenza
  - **Polarità:** classificazione in pos/neg sul log(LR) generale della recensione tramite **T**

- Miglior modello è risultato essere il **random forest**
  - **MSE**: 0.6704718

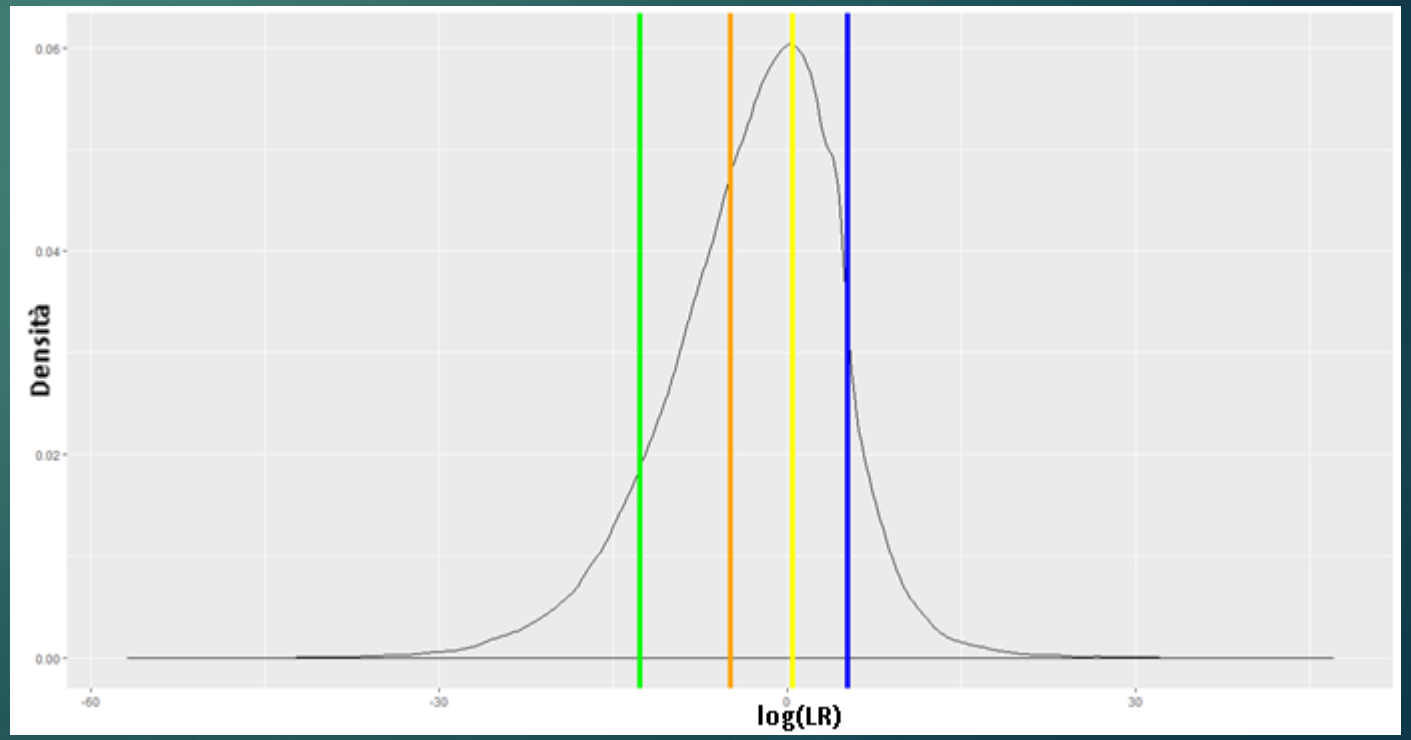
Focus sullo *score* di Booking



- **Booking <u>non</u> consente formulare il proprio giudizio attraverso un valore numerico**
- Si basa su di un sistema di "mood", ovvero viene associato il giudizio ad una fra 4 delle possibili risposte associate ai valori: 2.5, 5, 7.5, 10.
- Questo avviene per 6 categorie
- Lo *score* è pari alla media dei punteggi, arrotondata al primo decimale
- Fonte: ***https://partnerhelp.booking.com/hc/en-us/articles/213302185-How-is-my-review-score-calculated-***

Nuovo *score*

- Utilizzo dei quantili della distribuzione empirica dei log(LR) delle recensioni e individuazione di classi di soddisfazione attraverso i breakpoints
- Breakpoints utilizzati: 0, 0.1, 0.35, 0.65, 0.9, 1.
- Classi di soddisfazione:
  - Completamente insoddisfacente
  - Non soddisfacente
  - Potenzialmente buono
  - Buono
  - Ottimo



Nuovo *score*

- $$\frac{-1 \times \log(LR\_categoria)}{\left|\sum \log(LR\_categoria)\right|}$$
- Analizzando gli *score* **potenzialmente buoni**

review_position	review_cleaning	review_food	review_comfort
2.29885	0.86474	0.20404	0.19251
review_room	review_pricequalityrate	review_bar	review_staff
0.19124	0.10950	-0.09174	-0.09556
review_wifi	review_other	review_services	review_hotel
-0.31595	-0.32249	-0.56424	-0.79508
review_sleepquality			
-0.89018			

- Se il rapporto di una categoria è >1 il punteggio non è massimizzato e gli sforzi che vengono fatti per tenere quella categoria così in alto vengono vanificati dalle categorie con rapporto <<0.
- Non è un caso infatti che **completamente insoddisfacenti** presenti solo proporzioni negative mentre **ottimi** solo positive.

Conclusioni

- È stato creato un modello (affidabile al 91%) basato sui commenti dei clienti
- Modello che:
  - supporta le strutture ricettive nell'individuazione dei propri punti di forza e debolezza identificando gli aspetti sui quali operare per ottenere un miglioramento del servizio offerto
  - consente di identificare anche i punti di forza e debolezza della destinazione in cui opera la struttura ricettiva
  - può essere applicato in considerazione di differenti ambiti territoriali: i comuni, le provincie, le regioni o le specifiche destinazioni turistiche
- La sua applicazione consente di prevedere (con un errore accettabile) lo *score* di una recensione su Booking e, per ciascuna classe definita, i punti di forza/debolezza che la caratterizzano, ovvero su quali aspetti intervenire per migliorare il risultato ottenuto.

Sviluppi Futuri

- Test con siti affini a Booking
- *Word embeddings* per la rappresentazione delle parole e dei contenuti invece del BOW

Grazie per l'attenzione!

Classificazione Naive Bayes*

- La stima di P(neg | "contenuto") si basa sul rapporto fra verosimiglianze LR:
  - $LR = \frac{P(neg \mid "contenuto")}{P(pos \mid "contenuto")}$
  - $0 \leq LR < \infty$
- Si può calcolare attraverso il teorema di Bayes:

Sia contenuto = "a, b, c" allora $LR = \frac{P(neg \mid "contenuto")}{P(pos \mid "contenuto")} \approx$

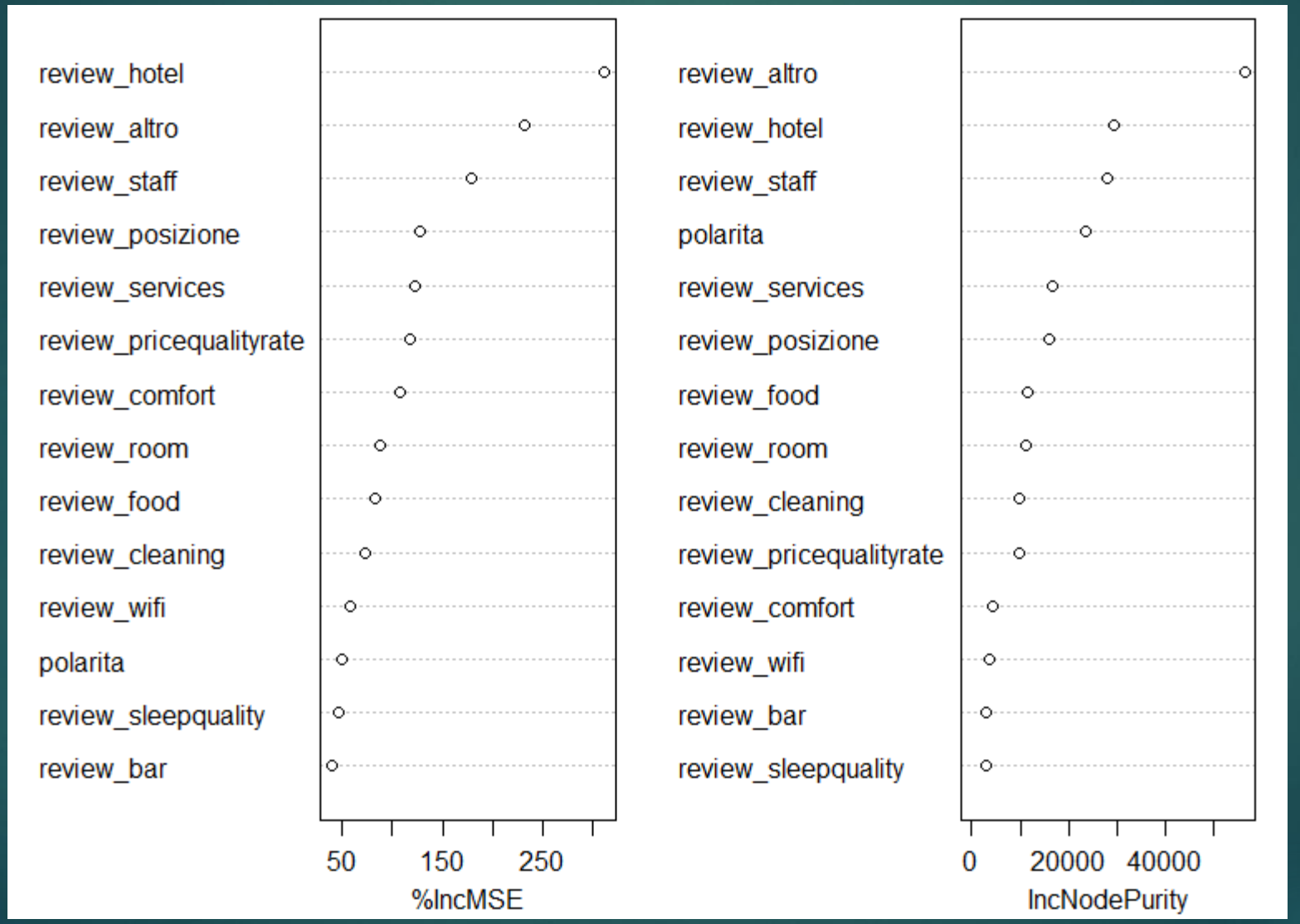
$$\frac{P("a"|neg)}{P("a"|pos)} \times \frac{P("b"|neg)}{P("b"|pos)} \times \frac{P("c"|neg)}{P("c"|pos)} \times \frac{P("\bar{a}"|neg)}{P("\bar{a}"|pos)} \times \frac{P("\bar{b}"|neg)}{P("\bar{b}"|pos)} \times \frac{P("\bar{c}"|neg)}{P("\bar{c}"|pos)} \times \frac{P(neg)}{P(pos)}$$

- Queste probabilità sono stimate attraverso le frequenze relative

Benchmarking

Model	Misclassification Error	Accuracy	Sensitivity	Fall-out	F1 score	Matthews Correlation Coefficient
Naive Bayes*	**0,089**	**0,911**	**0,929**	**0,117**	**0,926**	**0,813**
Logistic	0,150	0,850	0,884	0,532	0,877	0,361
Random Forest	0,189	0,811	0,873	0,591	0,849	0,303
Naive Bayes (e1071)	0,194	0,806	0,804	0,389	0,834	0,390
Naive Bayes (klaR)	0,194	0,806	0,804	0,389	0,834	0,390
CART	0,232	0,768	0,842	0,587	0,815	0,272
LDA	0,236	0,764	0,860	0,641	0,816	0,246

Previsione *score* di Booking



Focus sullo *score* di Booking