DATA MINING

APP STORE

INDICE

IL DATASET

- DATASET «MOBILE APP STORE»
- FONTE: WWW.KAGGLE.COM
- 7197 OSSERVAZIONI
- 16 VARIABILI INDIPENDENTI: 11 NUMERICHE, 5 CATEGORIALI
- 1 VARIABILE DIPENDENTE «CLASS» = C0 -> APPLICAZIONE DI NON SUCCESSO
  = C1 -> APPLICAZIONE DI SUCCESSO

LE VARIABILI

NUMERO LIVELLI (CATEGORIALI):

- TRACK_NAME: 7195
- CURRENCY: 1
- VER: 1590
- CONT_RATING: 4
- PRIME_GENRE: 23

VARIABILE TARGET:

- Class=0 -> 2416 (33,6%)
- Class=1 -> 4781 (66,4%)

```
> str(AppleStore)
tibble [7,197 x 17] (S3: tbl_df/tbl/data.frame)
 $ X1              : num [1:7197] 1 2 3 4 5 6 7 8 9 10 ...
 $ id              : num [1:7197] 2.82e+08 2.82e+08 2.82e+08 2.83e+08 2.83e+08 ...
 $ track_name      : chr [1:7197] "PAC-MAN Premium" "Evernote - stay organized" "WeatherBug - Local
Weather, Radar, Maps, Alerts" "eBay: Best App to Buy, Sell, Save! Online Shopping" ...
 $ size_bytes      : num [1:7197] 1.01e+08 1.59e+08 1.01e+08 1.29e+08 9.28e+07 ...
 $ currency        : chr [1:7197] "USD" "USD" "USD" "USD" ...
 $ price           : num [1:7197] 3.99 0 0 0 0 0.99 0 0 9.99 3.99 ...
 $ rating_count_tot: num [1:7197] 21292 161065 188583 262241 985920 ...
 $ rating_count_ver: num [1:7197] 26 26 2822 649 5320 ...
 $ user_rating_ver : num [1:7197] 4.5 3.5 4.5 4.5 5 4 4.5 4.5 5 4 ...
 $ ver             : chr [1:7197] "6.3.5" "8.2.2" "5.0.0" "5.10.0" ...
 $ cont_rating     : chr [1:7197] "4+" "4+" "4+" "12+" ...
 $ prime_genre     : chr [1:7197] "Games" "Productivity" "Weather" "Shopping" ...
 $ sup_devices.num : num [1:7197] 38 37 37 37 37 47 37 37 37 38 ...
 $ ipadSc_urls.num : num [1:7197] 5 5 5 5 5 5 0 4 5 0 ...
 $ lang.num        : num [1:7197] 10 23 3 9 45 1 19 1 1 10 ...
 $ vpp_lic         : num [1:7197] 1 1 1 1 1 1 1 1 1 1 ...
 $ Class           : num [1:7197] 1 1 0 1 1 1 1 1 1 1 ...
```

DIVISIONE DATASET

ESSENDO IL DATASET COSTITUITO DA UN ELEVATO NUMERO DI OSSERVAZIONI E NON AVENDO A DISPOSIZIONE UN DATASET DI SCORE ABBIAMO PROCEDUTO CON TALE SUDDIVISIONE:

- 10% SCORE (719 OBS)

- 90% DATASET: - 70% TRAINING (4535 OBS)
  - 30% VALIDATION (1943 OBS)

PRE-PROCESSING

- MISSING VALUES

  NEL DATASET NON SONO PRESENTI VALORI MANCANTI

- NZV

  RIMUOVIAMO LE VARIABILI «CURRENCY» E «VPP_LIC»

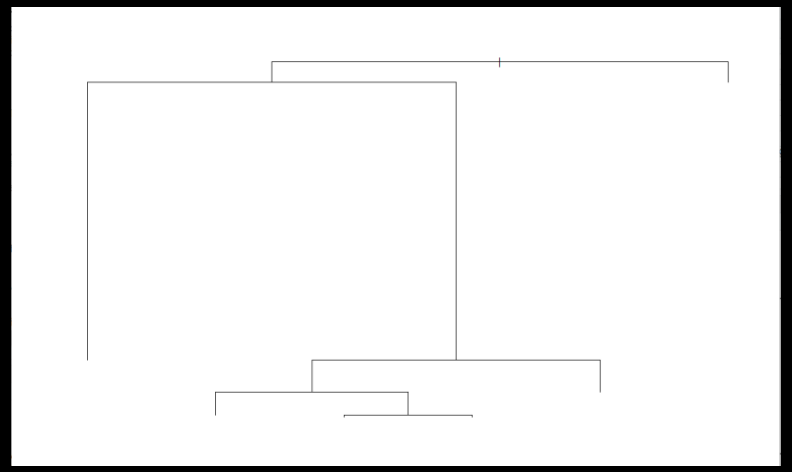
- CORRELAZIONI/COLLINEARITA'

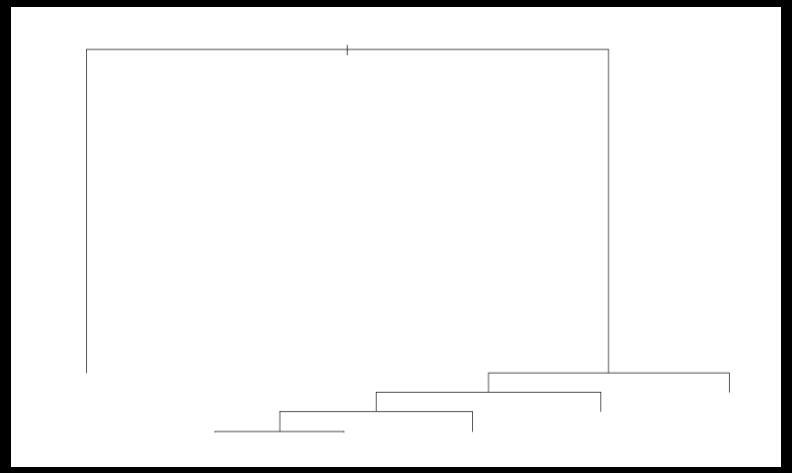
  CUTOFF = 0.7
  ELIMINIAMO LA VARIABILE «X1»

OPTIMAL GROUPING

ABBIAMO USATO IL METODO «TREE» DEL PACCHETTO «CARET» E ABBIAMO OTTENUTO LE SEGUENTI RIDUZIONI:

- VER: 1590 -> 6
- RATING_COUNT_TOT: 3185 -> 6

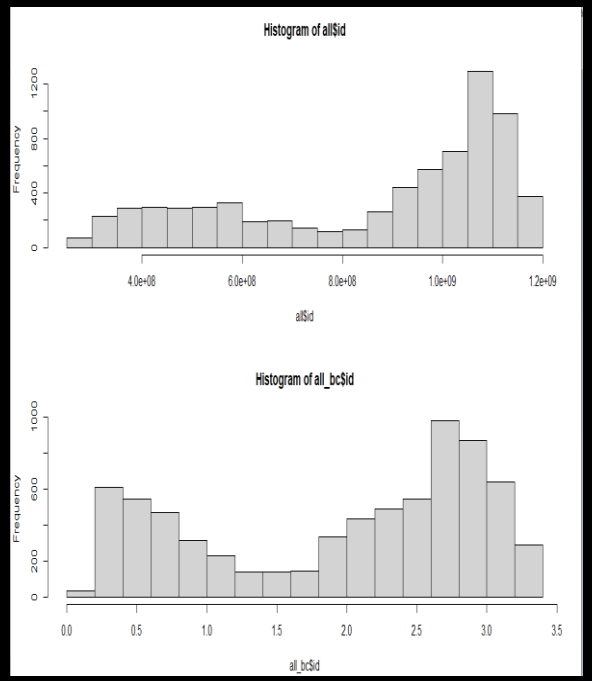




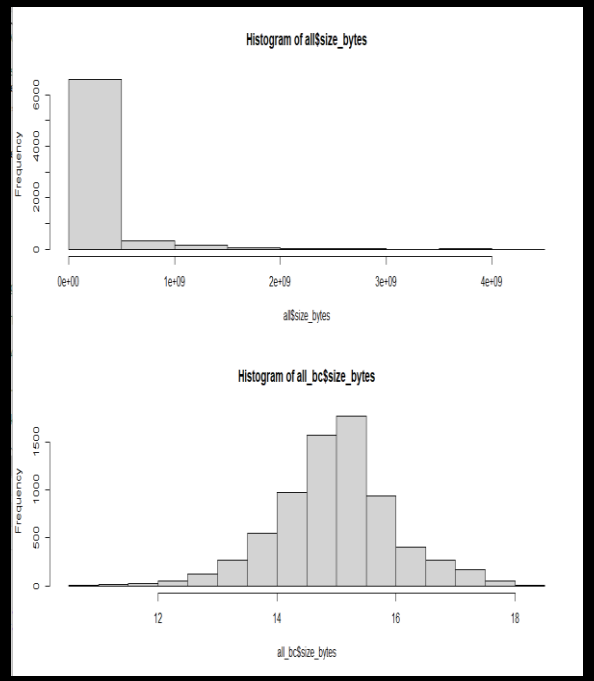
TRASFORMAZIONI

LE SEGUENTI VARIABILI NUMERICHE SONO STATE TRASFORMATE E RESE PIU' SIMMETRICHE:

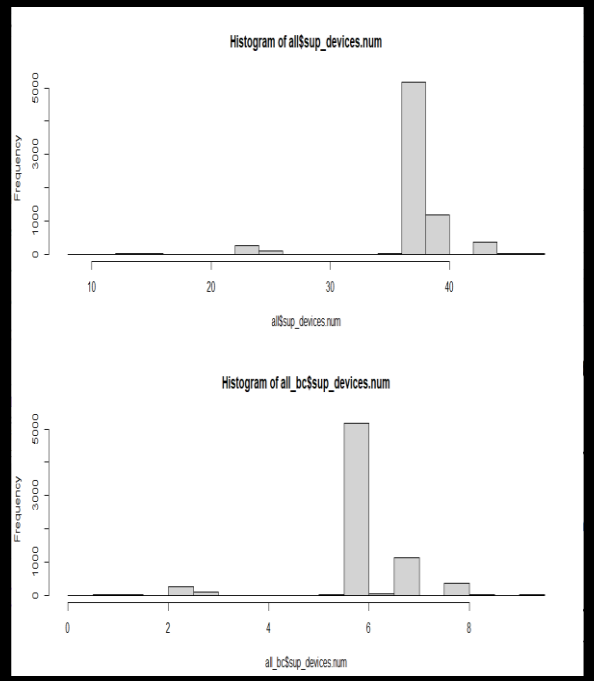
- ID



- SIZE_BYTES



-SUP_DEVICES.NUM

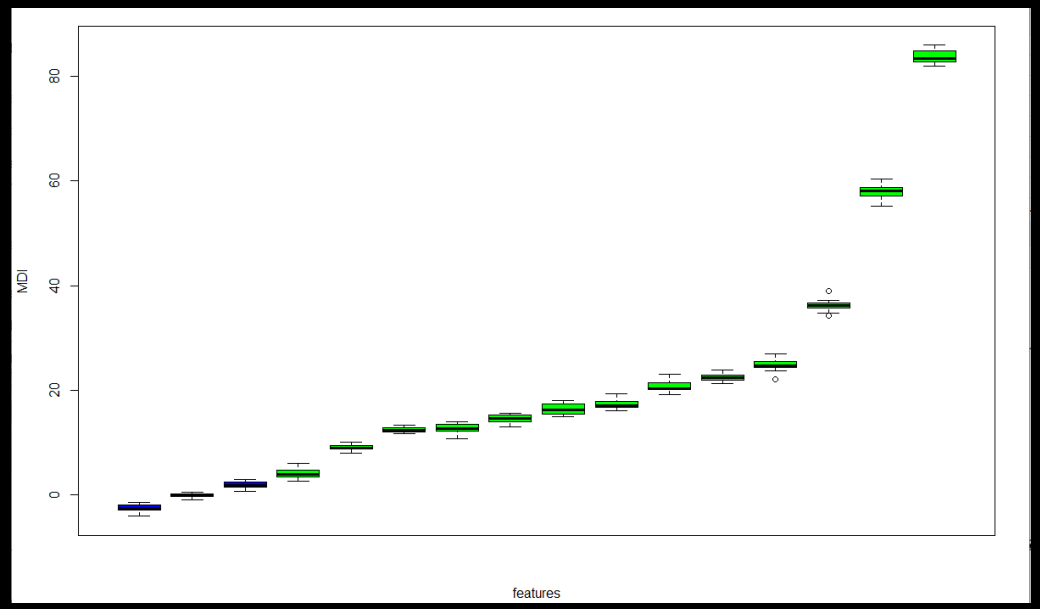


GANDINI MATTIA DONZELLA MARCO GALBIATI DANIELE

MODEL SELECTION

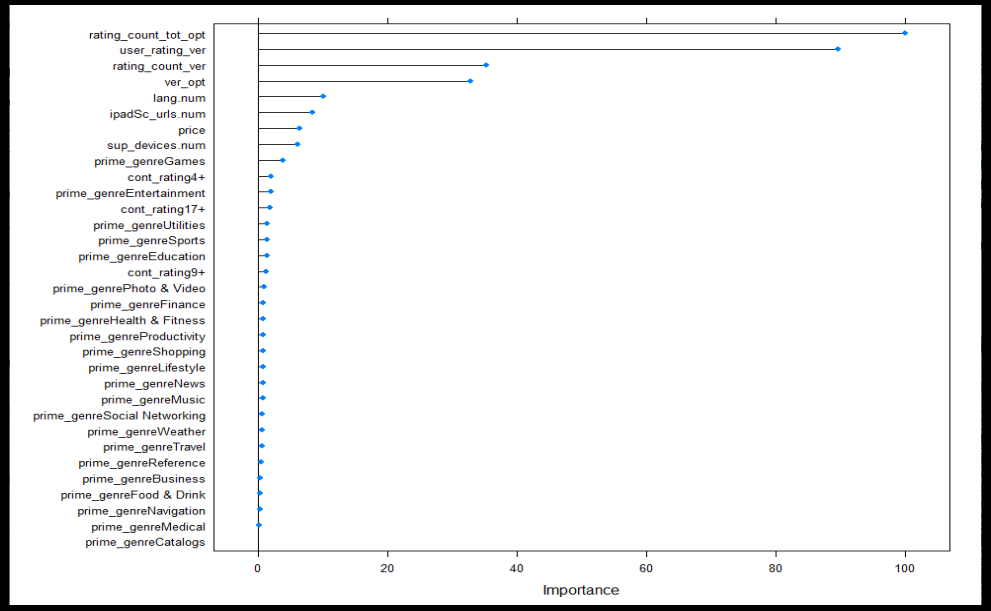
- BORUTA

BORUTA NON ESLUDE
NESSUNA VARIABILE



- RANDOM FOREST

METRICA USATA: AUC
SOGLIA DI IMPORTANZA: 15%

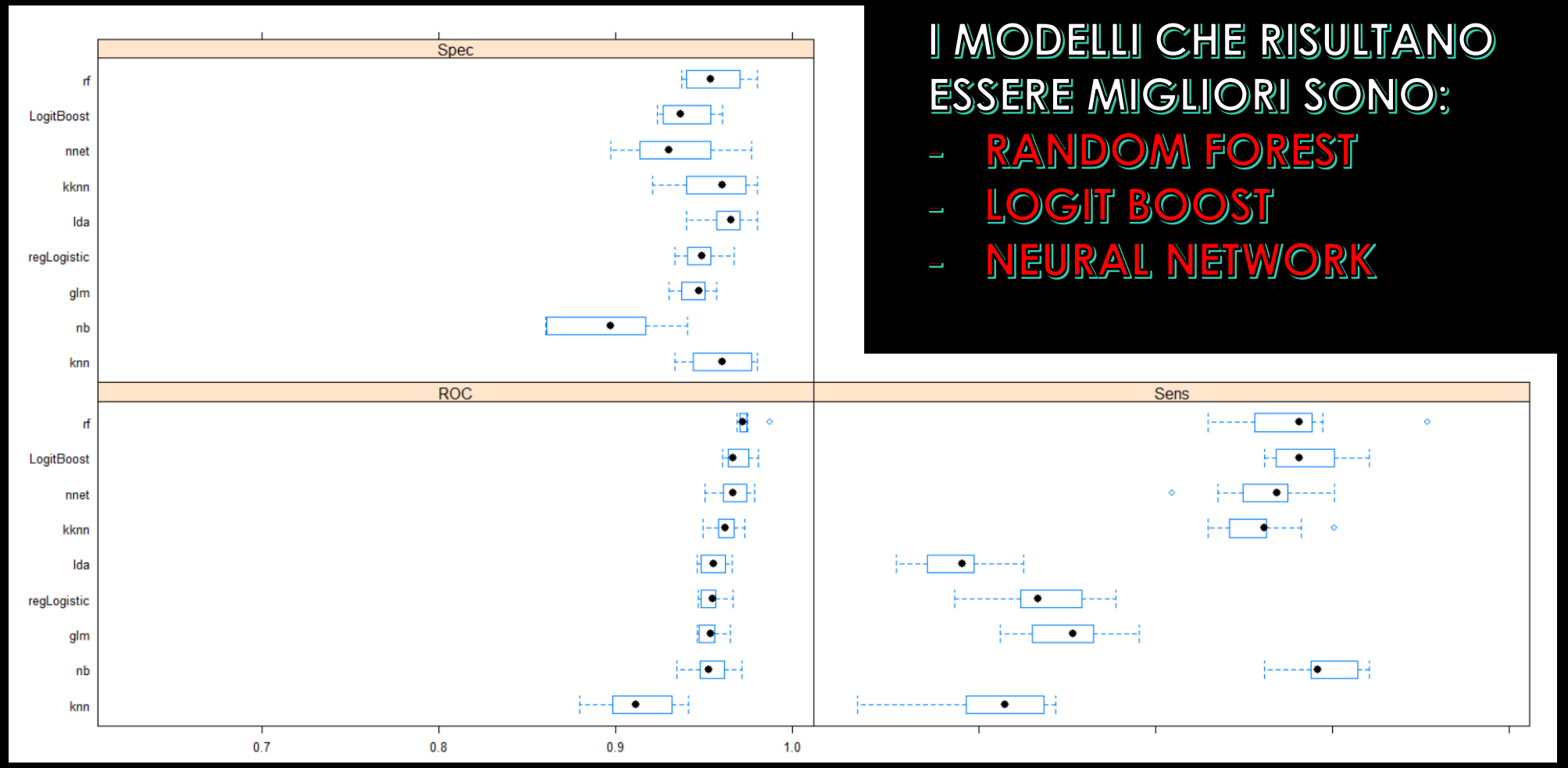


GANDINI MATTIA DONZELLA MARCO GALBIATI DANIELE

CHECK DATASET

- APPLESTORE_PRE:
  NZV + COLLINEARITA'
- APPLESTORE_OPT_TR:
  NZV + COLLINEARITA' + OPTIMAL GROUPING + TRANSFORMATION
- APPLESTORE_PRE_MS:
  NZV + COLLINEARITA' + MODEL SELECTION
- APPLESTORE_OPT_TR_MS:
  NZV + COLLINEARITA' + OPTIMAL GROUPING + TRANSFORMATION + MODEL SELECTION

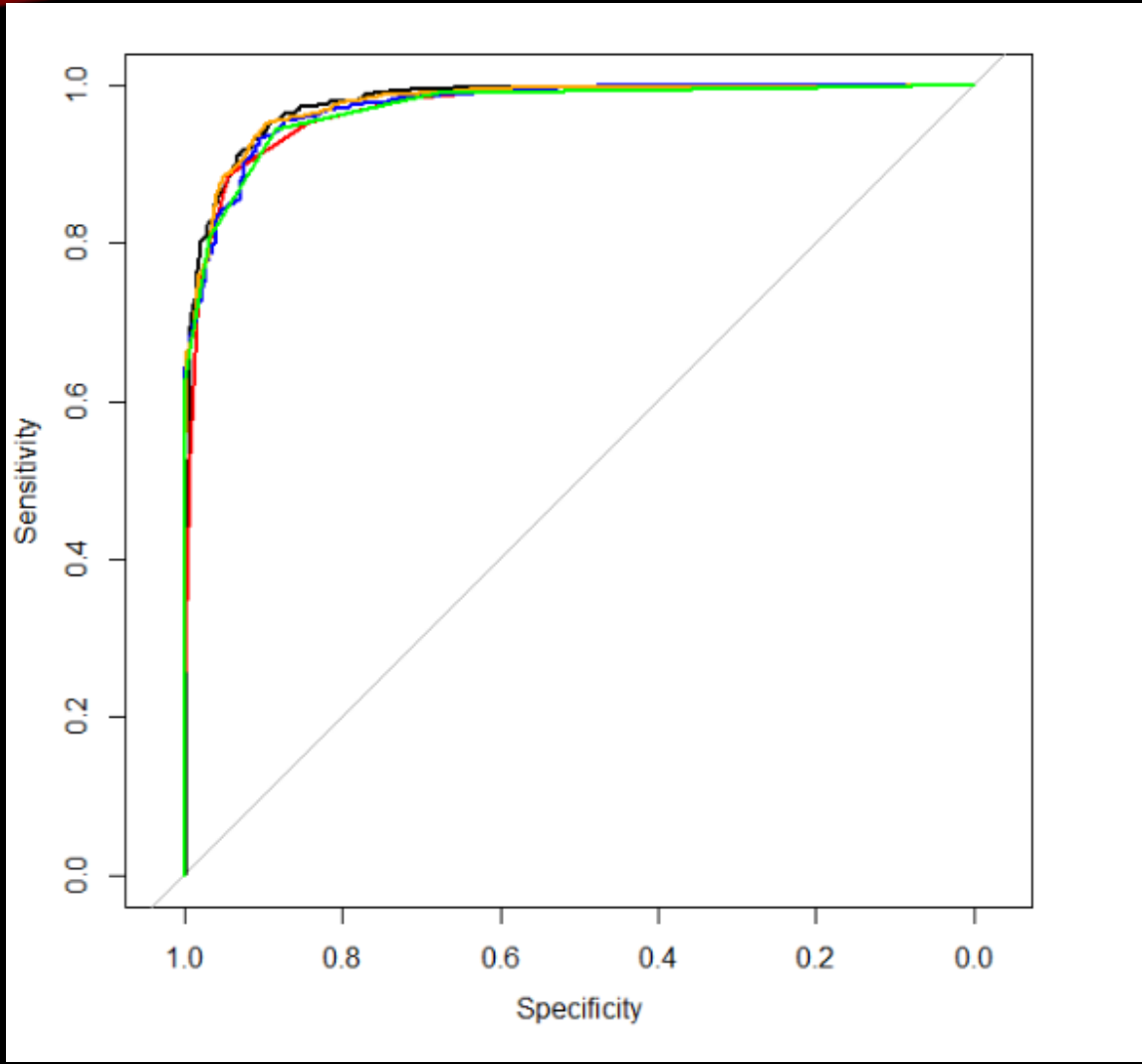
CONFRONTO SU TRAINING



I MODELLI CHE RISULTANO ESSERE MIGLIORI SONO:

- RANDOM FOREST
- LOGIT BOOST
- NEURAL NETWORK

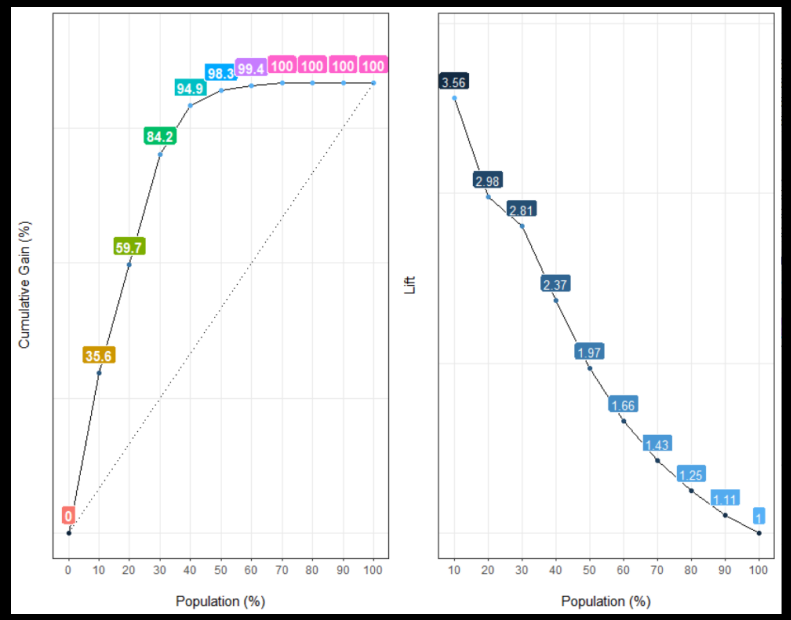
ASSESSMENT - ROC



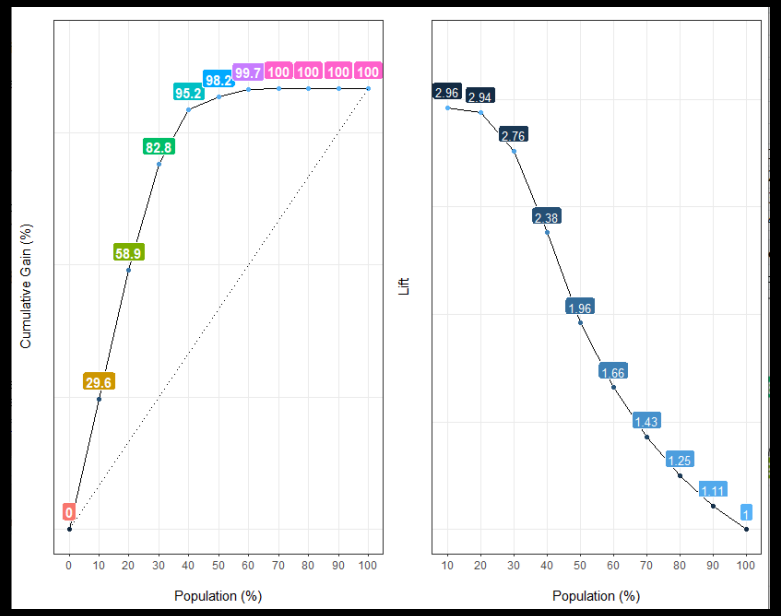
MODELLI MIGLIORI
1) **RANDOM FOREST**
[APPLESTORE_OPT_TR_MS]
2) **NEURAL NETWORK 1**
[APPLESTORE_PRE_MS + syze_bytes]
3) **NEURAL NETWORK 2**
[APPLESTORE_OPT_TR_MS]
4) **LOGIT BOOST 1**
[APPLESTORE_OPT_TR_MS]
5) **LOGIT BOOST 2**
[APPLESTORE_OPT_TR]

ASSESSMENT - LIFT

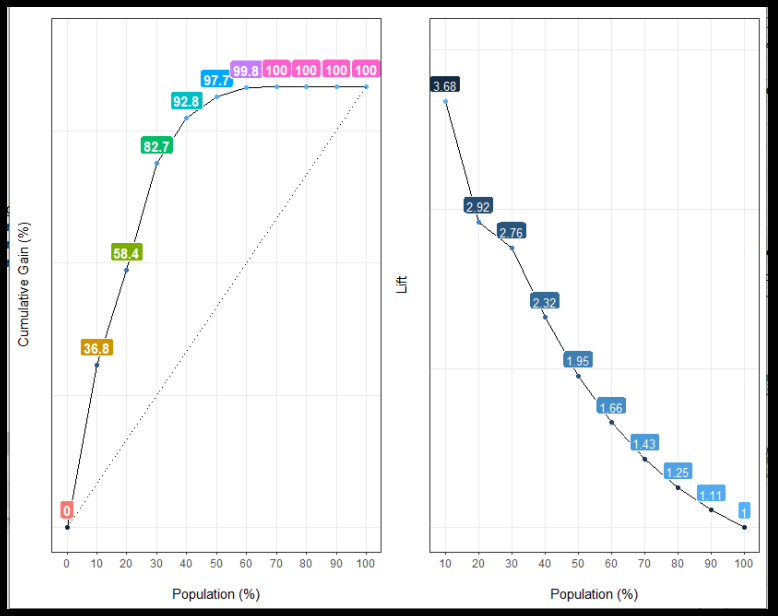
RANDOM FOREST



NEURAL NETWORK 1

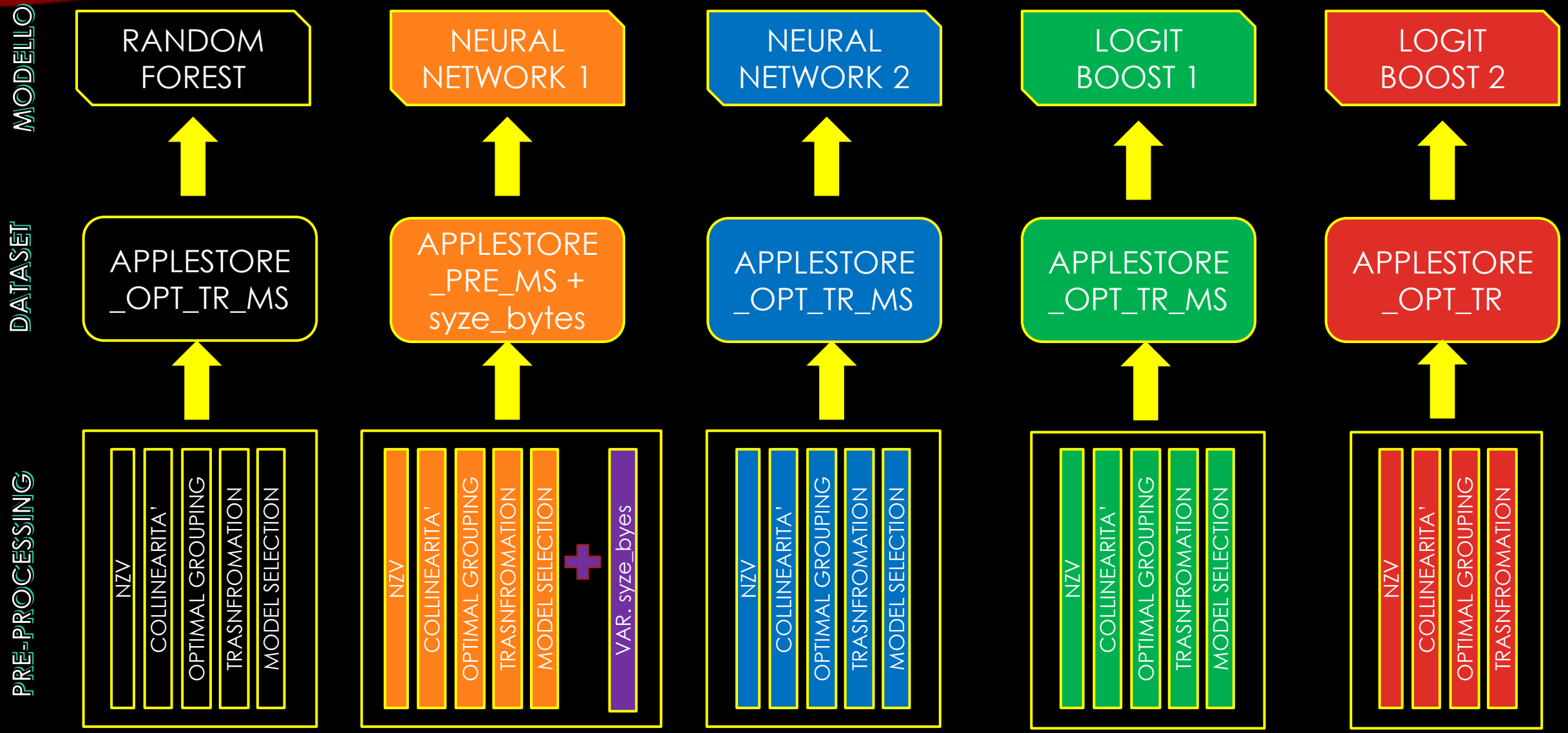


NEURAL NETWORK 2



OSSERVANDO LE CURVE LIFT E PRENDENDO COME RIFERIMENTO IL TEZO DECILE, RANDOM FOREST, SI CONFERMA COME MODELLO PIU' PERFORMANTE.

MODELLI MIGLIORI

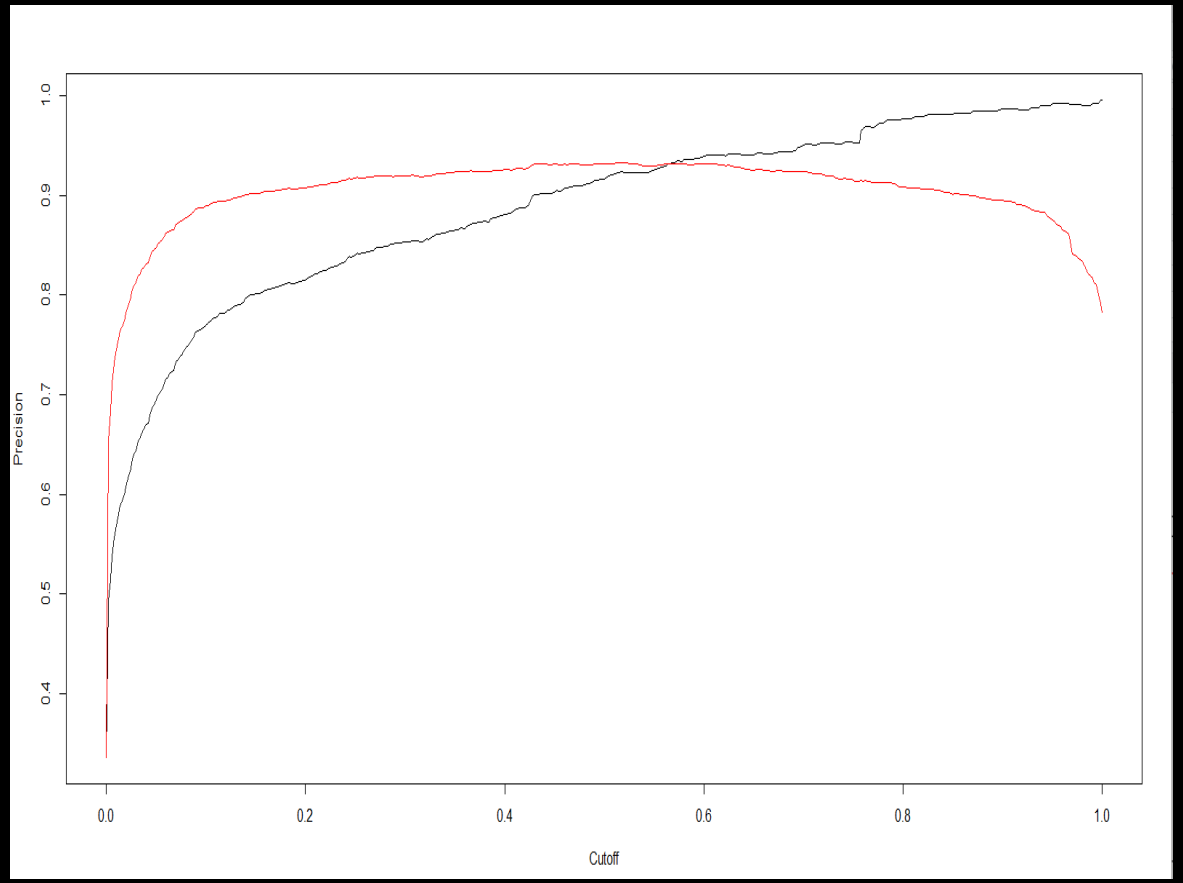


MODELLO VINCENTE

- RANDOM FOREST

SOGLIA UTILIZZATA: 0,6

OTTENUTA MASSIMIZZANDO PRECISION E ACCURACY



MODELLO VINCENTE

- RANDOM FOREST

MATRICE DI CONFUSIONE:

ACCURACY: 0,9269
KAPPA: 0,8338

```
Confusion Matrix and Statistics

          Reference
Prediction   c0   c1
        c0  563   53
        c1   89 1237
```

CONCLUSIONI

- IL MODELLO VINCENTE E' STATO UTILIZZATO PER CLASSIFICARE 719 OSSERVAZIONI.
- LA VARIABILE TARGET E' COSI' DISTRIBUITA NEL DATASET DI SCORE:

	C0	C1
VALORI ASSOLUTI	222	497
VALORI RELATIVI	0,3088	0,6912

- LA VARIABILE RISPOSTA SERVE A STABILIRE SE UN'APPLICAZIONE DI APP STORE HA RISCOSSO SUCCESSO OPPURE NO.

CONCLUSIONI

LE CARATTERISTICHE DI UN'APP DA CONSIDERARE SONO:

- NUMERO DI RECENSIONI DELL'ULTIMA VERSIONE
- NUMERO DI RECENSIONI TOTALE
- NUMERO DI LINGUE IN CUI è DISPONIBILE L'APPLICAZIONE
- VOTO MEDIO DELL'ULTIMA VERSIONE DELL'APP
- VERSIONE DELL'APPLICAZIONE
- (DIMENSIONE DELL'APP)