

Hello, my name is **INFLUENCER**

Data Science lab

a.a. 2017/2018

Alex Ceccotti		790497
Michela Sessi		777760
Stefano Fiorini		778379
David Govi		833653



Obiettivo



Predire il giudizio umano su
chi è ritenuto più **influyente** tra
due **utenti Twitter**

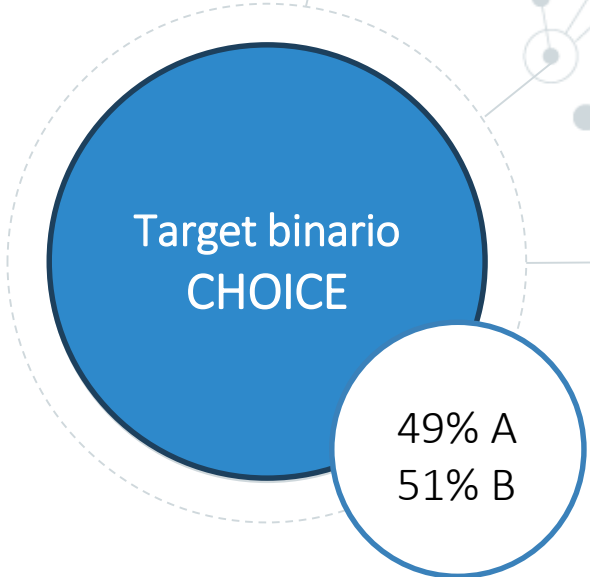


Riconoscere gli
Influencer
automaticamente

Dataset

Per ogni utente:

- Following_count
- Follower_count
- Retweets_received
- Retweets_sent
- Mentions_received
- Mentions_sent
- Posts
- Listed_count
- Network_feature_1
- Network_feature_2
- Network_feature_3



Target binario
CHOICE

49% A
51% B

Data manipulation



Creazione dei **rapporti** tra le variabili di A e B



Se il rapporto = **Inf** tengo A
Se il rapporto è NA metto 1



Creazione di ratio interni all'utente:
Folfol, Menmen, Retret

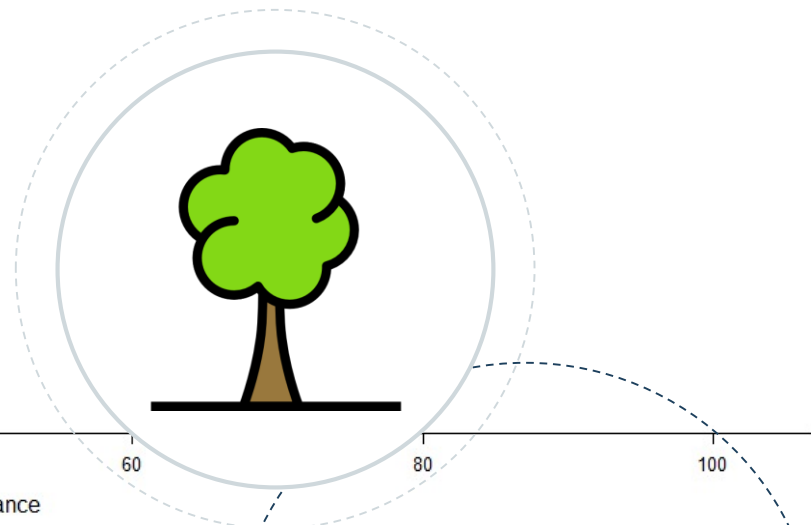


Dataset
5500 rows
42 variabili

Conteggio degli **zeri** per utente e **variabile HasZero**

rapp_listed_count
 rapp_network_feature_1
 rapp_follower_count
 rapp_mentions_received
 rapp_retweets_received
 A_ment_ratio
 B_posts
 A_listed_count
 A_network_feature_3
 B_follower_count
 A_network_feature_2
 rapp_retweets_sent
 B_mentions_sent
 A_retweets_received
 rapp_mentions_sent
 B_network_feature_3
 rapp_following_count
 A_foll_ratio
 rapp_network_feature_2
 B_network_feature_1
 B_mentions_received
 rapp_network_feature_3
 A_retw_ratio
 B_network_feature_2
 A_follower_count
 A_mentions_sent
 B_retweets_received
 A_posts
 rapp_posts
 B_foll_ratio
 A_listed_count
 B_retw_ratio
 B_following_count
 A_network_feature_1
 A_retweets_sent
 B_zeros
 B_ment_ratio
 B_retweets_sent
 A_mentions_received
 A_zeros
 has_zerosTRUE
 A_following_count

- Rapp listed_count
- Rapp network_feature_1
- Rapp follower_count
- Rapp mentions_received
- Rapp retweets_received

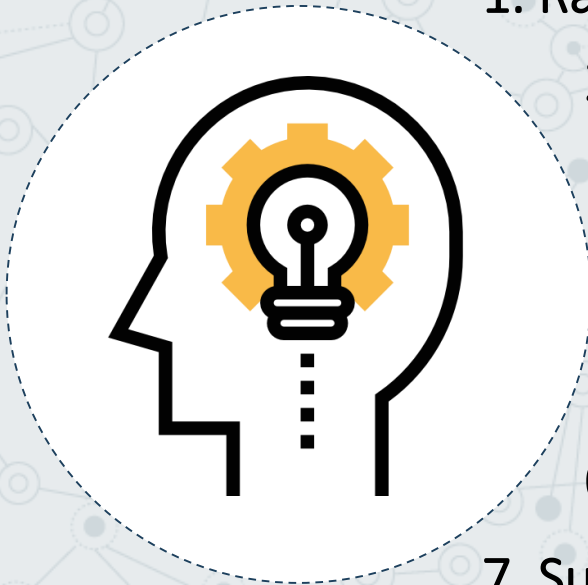


Selezione per:
 Neural Networks
 Support Vector Machine
 K Nearest Neighbors
 Logistica

Feature selection

Albero cross-validato 10folds
 Tuning con ROC = 0.8266

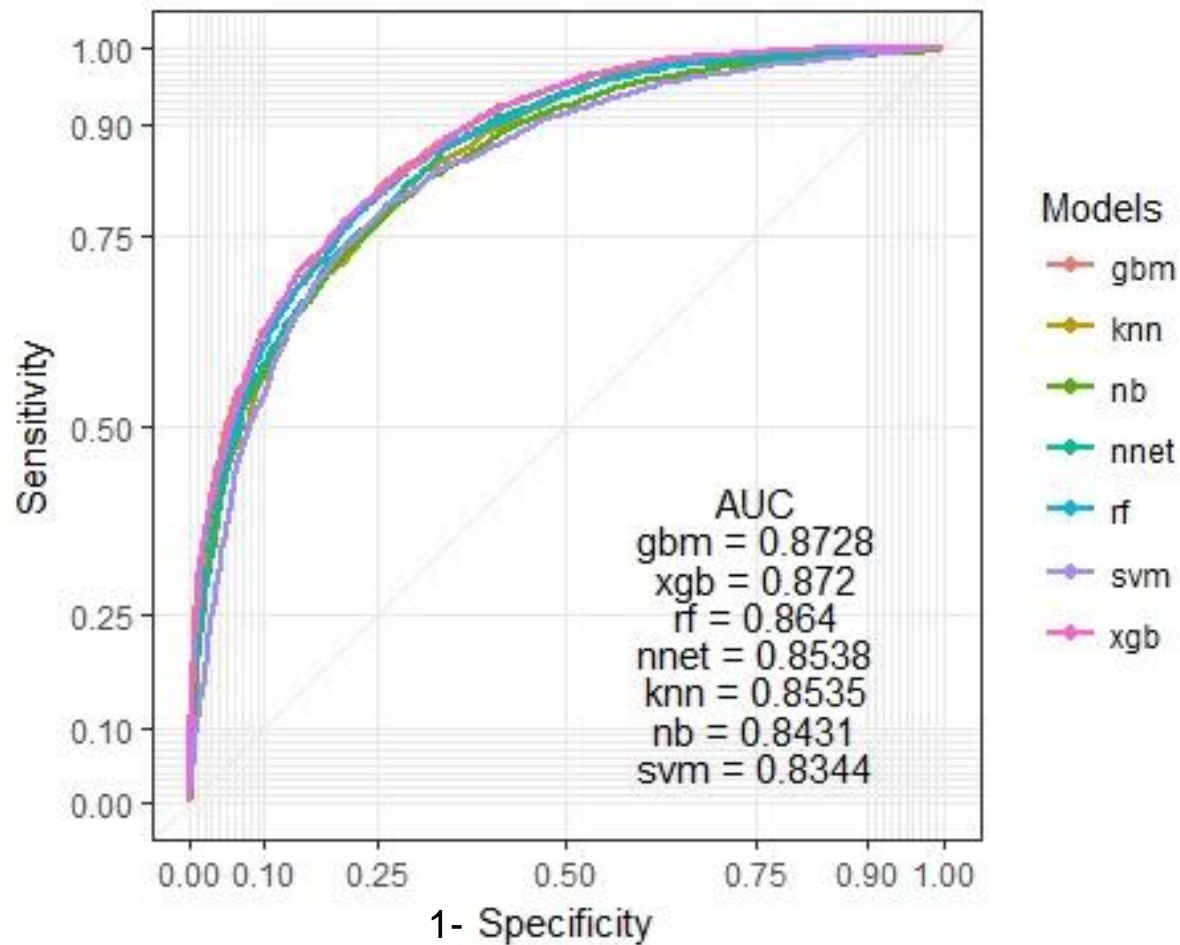
Modelli testati



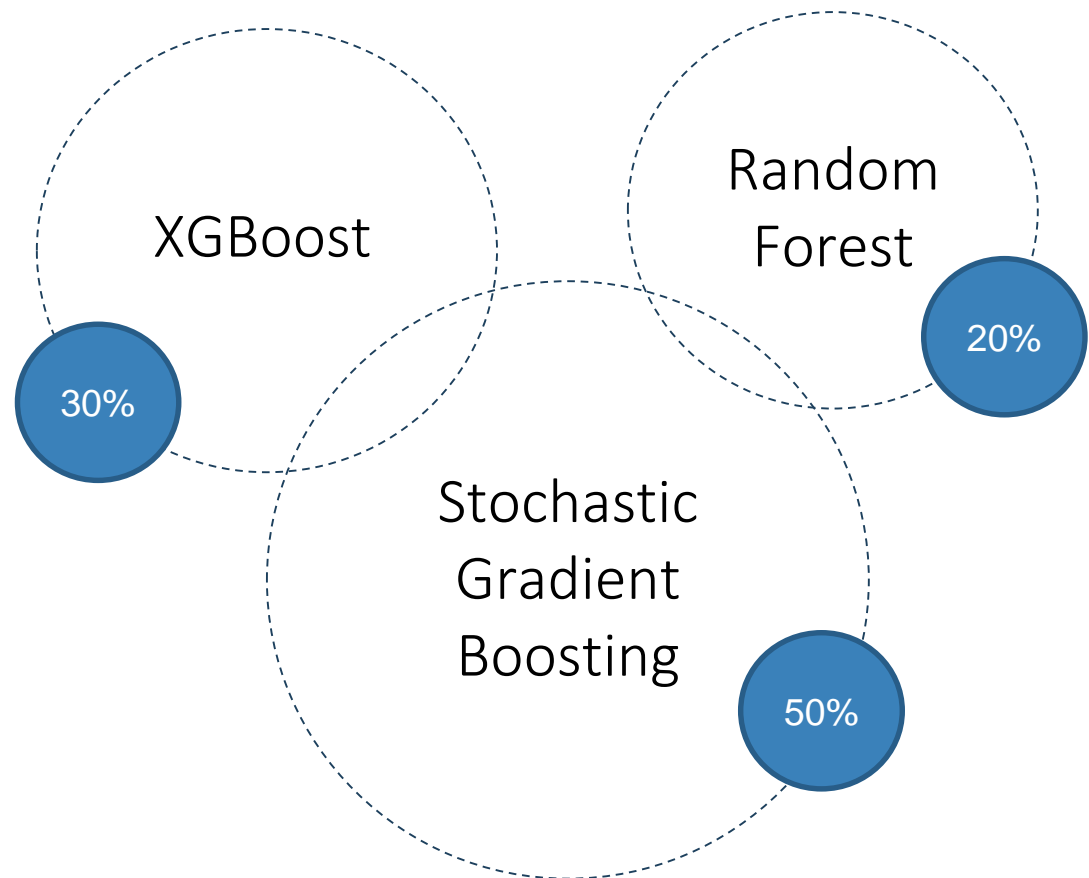
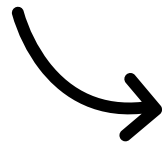
1. Random Forest
2. Naive Bayes
3. XGBoost
4. Stochastic Gradient Boosting
5. Neural Networks
6. K Nearest Neighbors
7. Support Vector Machine
8. Regressione logistica

Per tunare e scegliere il modello migliore è stata utilizzata la ROC

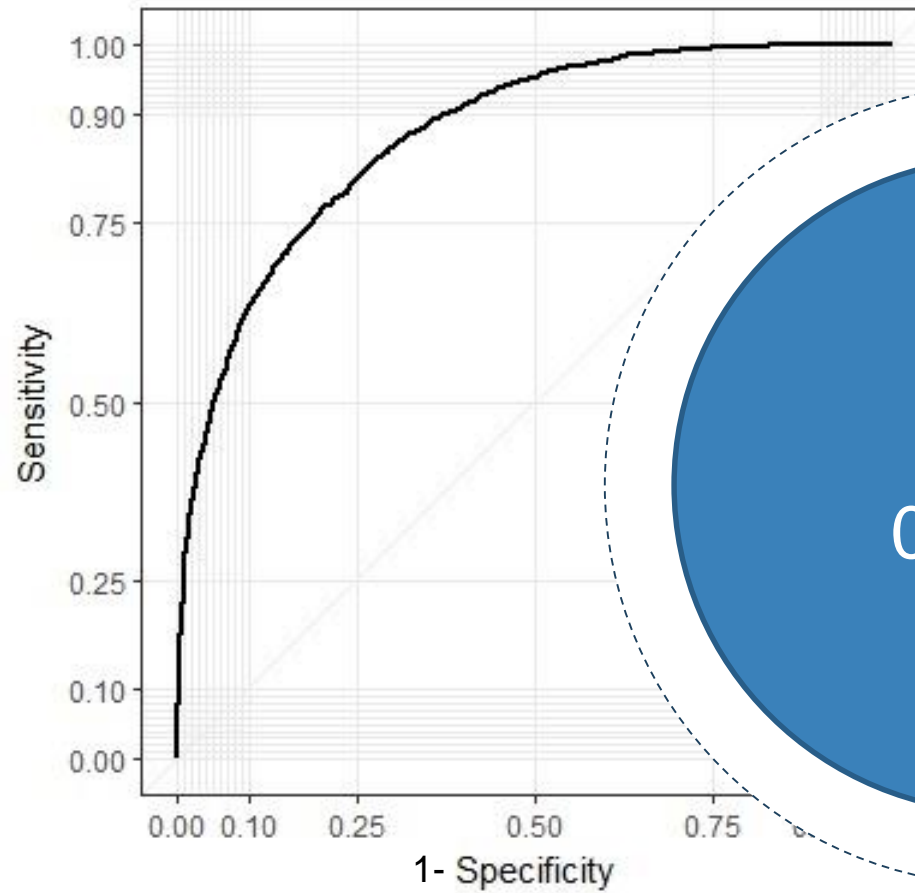
ROC curves



Ensamble method



Best model



ROC
0.8742

Qualche esempio

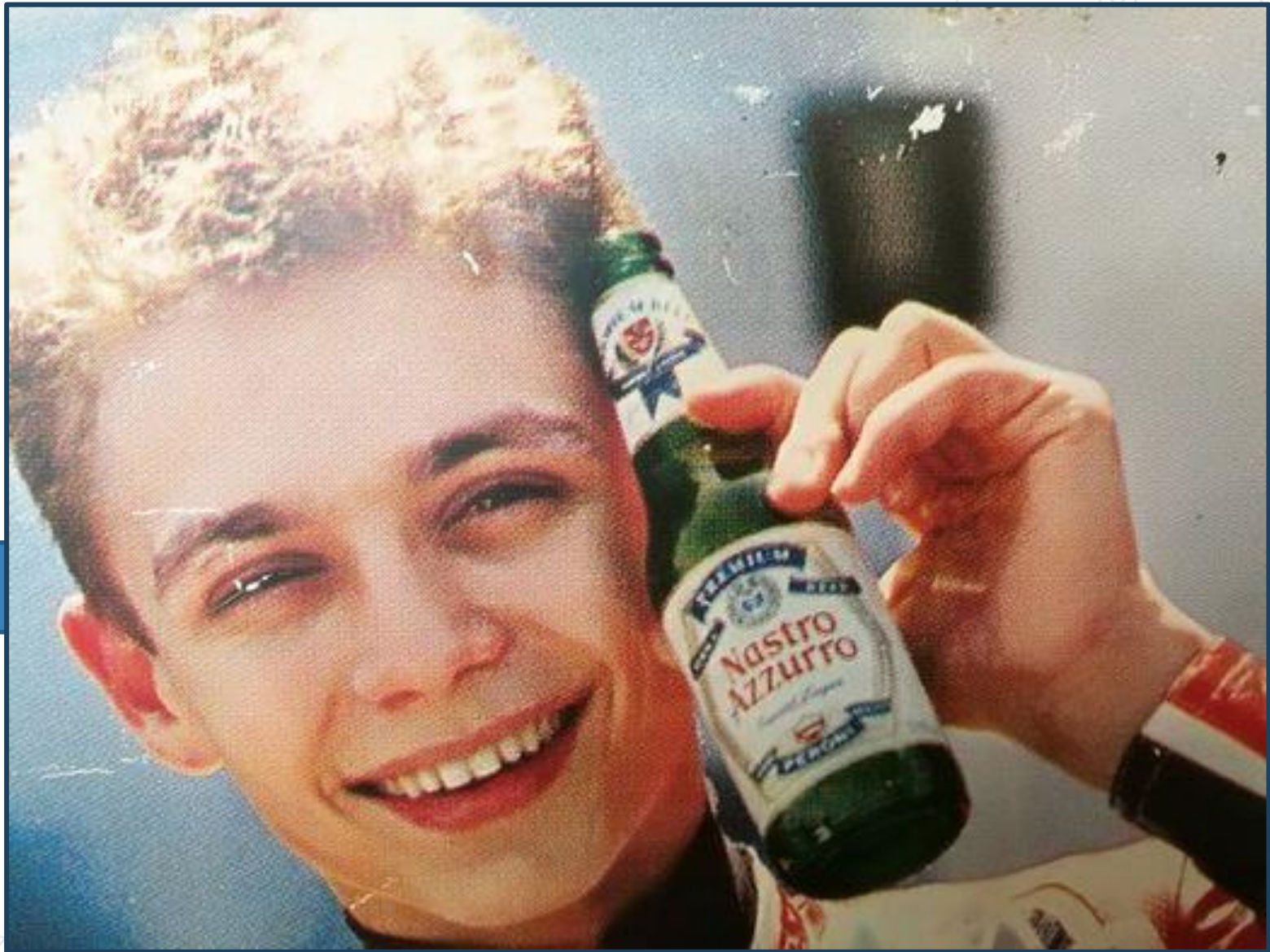




Prob. = 0.92



Prob. = 0.08





Prob. = 0.55



Prob. = 0.45



Prob. = 0.55



Prob. = 0.45



Grazie per l'attenzione

Data Science lab

a.a. 2017/2018

Alex Ceccotti		790497
Michela Sessi		777760
Stefano Fiorini		778379
David Govi		833653

