Business Game 2018

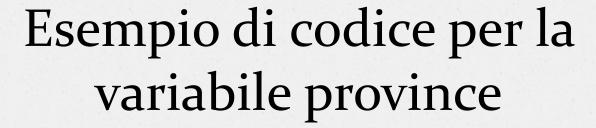


- price : variabile numerica (variabile risposta da prevedere)
- rewiew_score : variabile numerica
- rewiew : descrizione del vino fatta da un sommelier. Ho deciso di eliminare questa variabile
- country
- region_2 : ho eliminato questa variabile perchè presenta il 52% di valori NA



Ho ricategorizzato le seguenti variabili conservando solo le modalità con la maggior frequenza e accorpando le altre all'interno della modalità "altro":

- province : di cui ho tenuto 30 livelli
- region_1 : di cui ho tenuto 30 livelli
- winery: di cui ho tenuto 25 livelli
- variety : di cui ho tenuto 25 livelli
- o designation : di cui ho tenuto 20 livelli



#PROVINCE

```
levels(dati_stima$province) <- c(levels(dati_stima$province), "Altro")
selected_province<- names(sort(table(dati_stima$province), de = TRUE)[1:30])
dati_stima$province[!(dati_stima$province %in% selected_province)] <- "Altro"
dati_stima$province <- factor(dati_stima$province)
levels(dati_previsione$province) <- c(levels(dati_previsione$province), "Altro")
dati_previsione$province[!(dati_previsione$province %in% selected_province)]
<- "Altro"
dati_previsione$province <- factor(dati_previsione$province)
```

Divisione del dataset

Ho diviso in modo casuale il dataset in due parti:

- stima con 50598 osservazioni e verifica con 25297
- a sua volta il dataset stima è diviso casualmente in due parti uguali

Modello finale

Il modello finale è una RANDOM FOREST, i cui parametri di regolazione corrispondono a 200 per il numero di alberi e a 3 per il numero di variabili estratte casualmente a ogni nodo.

Esempio codice