

# Stats under the stars<sup>3</sup>

Markov Unchained

*"the second lag is silent"*

June 28, 2017

**Descrizione del dataset** Il problema di previsione affrontato consisteva nel selezionare su un test set i 10mila clienti della banca Findomestic che risultano più propensi ad accettare una proposta commerciale in seguito a una telefonata.

Il problema si qualificava come un'analisi supervisionata con un *training set* (40mila obs.) che riportava il valore della variabile target, in questo caso una dummy che assumeva valore 1 in caso il cliente chiamato avesse accolto la proposta ricevuta e 0 in caso contrario.

Tale *training set* presentava inoltre un forte sbilanciamento per quanto riguardava la distribuzione della variabile *target*: infatti ben il 95% dei casi aveva valore target pari a 0.

**Undersampling** Per ovviare all'inconveniente dello sbilanciamento del dataset si è ricorso ad una procedura di *undersampling*: una volta campionate delle osservazioni con valore target 1 si campionato lo stesso numero di osservazioni tra quelle con target 0. Il dataset ottenuto aveva dunque lo stesso numero di 0 e 1.

Oltre a risolvere il problema dello sbilanciamento, l'undersampling ci ha permesso di valutare la bontà di modelli diversi dato che un minor numero di osservazioni riduceva sensibilmente il peso computazionale.

**Feature engineering** Nella prima parte del lavoro ci siamo concentrati sulla costruzione di variabili che aiutassero a migliorare la previsione. Inizialmente abbiamo rimosso alcune variabili a causa dell'elevato numero di dati mancanti quali anzianità bancaria (ANZ\_BAN), reddito familiare

Per quanto riguarda la creazione di nuove variabili il criterio utilizzato è stato quello della riduzione del numero di regressori da dare in pasto ai modelli che via via abbiamo considerato.

## Random Forest