

Il modello vs il processo di modellizzazione

Data Mining

CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- KJ §1.2.3 §1.2.4 §1.2.6 §1.2.7
- KS § 1.5

Il processo di analisi dei dati

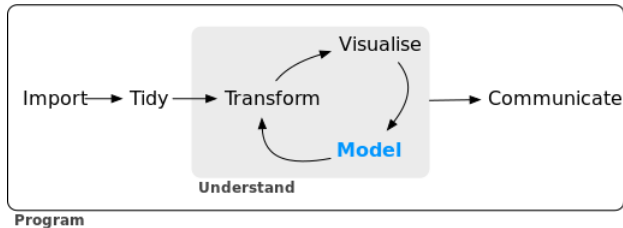
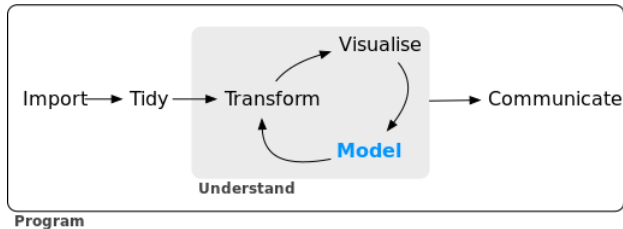


Figure 1.2 del libro KS

Prima di tutto, bisogna considerare il processo (spesso sottovalutato) di pulizia dei dati

Successivamente bisogna capire i dati. Questa fase viene chiamata analisi esplorativa dei dati.



Exploratory Data Analysis (termine coniato da J. Tukey, abbreviato in EDA) comprende le operazioni di

- visualizzazione
- trasformazione
- modellizzazione

Tuttavia non ci sono regole ben definite. EDA è fundamentalmente un processo creativo

Visualizzazione, trasformazione e modellizzazione dei dati

Il libro di Wickham e Grolemund (2016) (WG) illustra queste operazioni

La visualizzazione dei dati è un ottimo punto di partenza: ci consente di costruire grafici informativi che aiutano a comprendere i dati (WG, sezione 3)

La trasformazione dei dati ci consente di creare nuove variabili (*feature engineering*), di escludere le osservazioni anomale, etc. (WG, sezione 5)

La modellizzazione dei dati rende (matematicamente) precisa la relazione tra le variabili (WG, sezioni 22-25)

EDA

EDA (WG, section 7) è un processo iterativo

- Poniti delle domande sui tuoi dati.
- Cerca le risposte visualizzando, trasformando e modellando i tuoi dati.
- Usa ciò che impari per perfezionare le tue domande e/o generare nuove domande.

Il processo di modellizzazione

Il processo di modellizzazione è anch'esso un processo iterativo

Il modello per sé rappresenta una minima parte del processo di modellizzazione.

Le fasi principali comprendono:

- *Exploratory data analysis*
- *Feature engineering*
- *Model fitting / tuning*
- *Model evaluation*

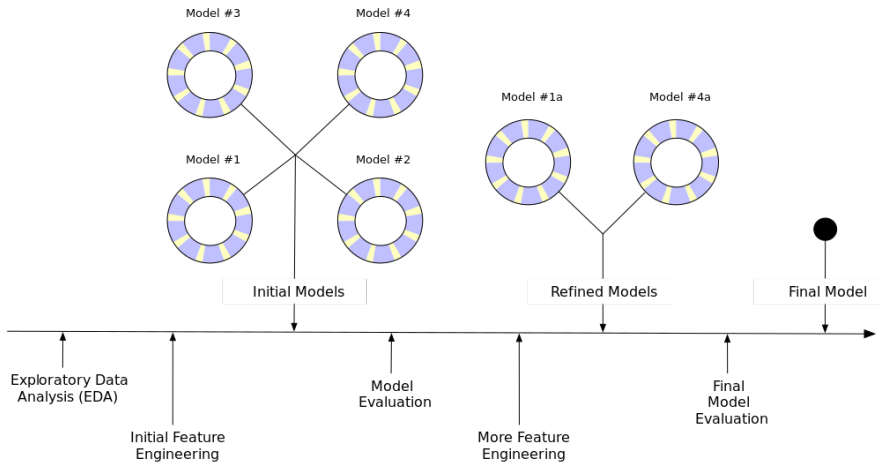


Figure 1.3 del libro KS

Chicago Train Ridership Data

Nel capitolo 4 di KJ (e capitoli successivi), viene discussa la modellizzazione del numero di utenti giornalieri alla stazione di Clark/Lake del sistema ferroviario pubblico di Chicago

I predittori utilizzati sono le date del calendario, la serie storica del numero di utenti nelle varie stazioni, il tempo atmosferico e altri fattori

Il processo di modellizzazione può essere esemplificato con il seguente monologo interiore

EDA I valori del numero di utenti nelle diverse stazioni sono estremamente correlati.

EDA I valori del numero di utenti nei giorni feriali e nel fine settimana sono molto diversi.

EDA Il giorno 11 Giugno 2010 ha un valore estremamente elevato di utenze.

EDA Quali stazioni presentano i valori più bassi?

Feat.eng. Le date dovrebbero essere codificate come giorno della settimana e anno.

Feat.eng. Forse i predittori fortemente correlati potrebbero essere rappresentati con una PCA.

Feat.eng. Le registrazioni meteorologiche orarie potrebbero essere riassunte in misurazioni giornaliere.

Mod.fit. Iniziamo con una regressione lineare, k -vicini più vicini e un boosting di alberi decisionali.

- Mod.tun. Quanti vicini k usare?
- Mod.tun. Quante iterazioni di boosting? Poche o tante?
- Mod.eval. Quali modelli hanno il MSE più basso?
- EDA Quali giorni sono stati previsti in modo non soddisfacente?
- Mod.eval. I punteggi di importanza delle variabili indicano che le informazioni meteorologiche non sono predittive. Li scarteremo dalla prossima serie di modelli.
- Feat.eng. Le registrazioni meteorologiche orarie potrebbero essere riassunte in misurazioni giornaliere.
- Mod.eval. Sembra che dovremmo concentrarci su molte iterazioni di boosting
- Feat.eng. Abbiamo bisogno di codificare le festività per migliorare le previsioni su (e intorno a) quelle date
- Mod.eval. Eliminiamo k -NN dall'elenco dei modelli