UNIVERSITÀ DI PISA

DISTRIBUTED DATA ANALYSIS AND MINING

anno accademico 2020/2021

Ferri Lorenzo (607828)
Ilic Ema (602796)
Pappolla Roberta (534109)

BANK LOAN STATUS DATASET

Link: https://www.kaggle.com/zaurbegiev/my-dataset

Dataset content

- Vengono forniti due files: ‘credit_train.csv’ e ‘credit_test.csv’. Il primo ha 19 colonne, il secondo ne ha 18: nel secondo non è presente l’attributo ‘Loan Status’.
- Ogni riga rappresenta un prestito di uno specifico cliente: la combinazione di attributi {‘Loan ID’, ‘Customer ID’} è Superchiave della tabella.

IN SUPREMÆ DIGNITATIS 1343
UNIVERSITÀ DI PISA

- Gli attributi forniscono le caratteristiche di ciascun prestito e cliente e l'esito finale. Ad esempio il 'Credit Score', l''Annual Income' e l''Home Ownership' del cliente; ed il 'Loan Status', il 'Term' ed il 'Purpose' del prestito.
- Tramite questi attributi lo scopo principale del Dataset è fornire un modello di classificazione per la predizione dell'attributo categorico binario 'Loan Status', ma si presta anche ad analisi predittive di altro tipo. In primis la costruzione di un modello di regressione per l'attributo 'Credit Score', mediante tutti gli attributi diversi da 'Loan Status', cosa che può rilevarsi utile per il riempimento dei numerosi Missing Values che esso contiene.

Loan status description

L'attributo 'Loan Status' assume due possibili valori: ['Fully Paid', 'Charged Off'].'

- Fully Paid' rappresenta la casistica in cui il prestito è andato a buon fine. In particolare il prestito è stato ripagato interamente rispettando termini e condizioni previste dal contratto.
- 'Charged Off' rappresenta il caso di un prestito non andato a buon fine. Più precisamente il finanziatore non ha ricevuto i pagamenti delle rate dovute per più di 180 giorni e di conseguenza il conto corrente dove è stato addebitata la somma del prestito viene dichiarato in Default.

Insightful & Vast USA Statistics

Dataset content

- Unico file csv, questo è composto da 39029 righe e 80 variabili, di queste solo 6 sono di tipo categorico. Sono presenti diversi null value.
- Singolo record si riferisce alla situazione finanziaria, e non solo, di un soggetto e tutti i dati relativi alla zona in cui abita (es. city, zip_code, area_code, lat...).
- La variabile "BLOCKID" è composta interamente da NaN, "SUMLEVEL" e "primary" contengono un unico valore ("140" e "tract") -> eliminare.

UNIVERSITÀ DI PISA

Obiettivo

- Creazione di una variabile di output:
  bad_debt= second_mortgage+home_equity-home_equity_second_mortgage
- Scelta di una soglia per creare un output binario (0.5).
- Dataset sbilanciato (99% con valore 0).

→ Trovare le features più importanti, aggiustare la soglia per aumentare la precisione ed equilibrare il dataset.