DESCRIZIONE DELLE FUNZIONI CONTENUTE IN funzioni.R

|  |  |
| --- | --- |
| Pacchetti necessari | randomForest, quantregForest, intervals |
|  |  |
| Nome Funzione |  |
| CONTROLLODATI | **Input**: Dati di QMI  **Output**: Visualizza a video anomalie ed errori.  **Descrizione**: La funzione esegue una serie di controlli sintattici precodificati, per individuare eventuali anomalie da correggere. Lo script non scrive nulla nel file di dati, ma segnala solo i problemi.  Si evidenziano:   * Presenza di valori mancanti o vuoti; * Anomalie (Redemption > 100, Quantità di buoni emessi negativa) |
| PREPARAZIONEDATI | **Input:** Dati di QMI;  **Output**: Dati di input per la foresta certificati da noi.  **Descrizione**: La funzione a seconda della tipologia di campagna scelta (Cartaceo o Print@home) esegue le seguenti operazioni:   * Seleziona il set di variabili su cui verrà fatta la foresta (indicare i nomi delle variabili); * Esclude variabili dalla selezione; * Corregge alcuni errori di battitura presenti nel file fornito il 10 Marzo 2015; * Modifica la classe R della variabile dove necessario; * Seleziona solo le campagne indicate come definitive e non anomale; * Seleziona le campagne nell’intervallo di anni indicato. * Modifica il nome della variabile target in “outcome” (di default la variabile target è promo\_redemption) * Salva il file di dati nella directory indicata con nome indicato. |
| RANDOMFOREST | **Input**: Dati di input per la foresta certificati da noi  **Output**: Stima della foresta casuale per la media  **Descrizione**: La funzione applica sui dati di input la funzione randomForest e salva l’oggetto R nella directory indicata con nome indicato.  Sono parametri della funzione il numero di alberi e il numero di variabili di split. |
| QUANTILERANDOMFOREST | **Input**: Dati di input per la foresta certificati da noi  **Output**: Stima della quantile random forest per mediana e rischio.  **Descrizione**: La funzione applica sui dati di input la funzione quantregForest e salva l’oggetto R nella directory indicata con nome indicato.  Sono parametri della funzione il numero di alberi e il numero di variabili di split. |
| PROFILICAMPAGNE | **Input**: Dati di input per la foresta certificati da noi+ output della funzione RANDOMFOREST  **Output**: dataframe con il profilo dei nodi di ciascun albero della foresta per ogni campagna di input.  **Descrizione**: La funzione restituisce il dataset di output (Numero righe= numero campagne input, Numero colonne= numero alberi) e salva l’oggetto R nella directory indicata con nome indicato. |
| PREVISIONE | **Input**: Profilo completo di una nuova campagna + Output delle funzioni RANDOMFOREST QUANTILERANDOMFOREST + Valore di break even della redemption  **Output**: lista composta da:   * Stima delle media della redemption per la nuova campagna * Stima delle media della redemption per la nuova campagna * Stima delle rischio che la redemption stimata superi il valore BreakEven   **Descrizione**: La funzione predice tramite la foresta e la foresta quantile la stima media della redemption, la stima della mediana e la stima della probabilità che il valore stimato superi la soglia indicata nel parametro BreakEven.  Attenzione: il vettore di profilo della campagna nuova deve contenere valori per tutte le variabili su cui è costruita la foresta e la foresta quantile. Eventuali valori mancanti causano errore e blocco della funzione. |
| SOMIGLIANZE | **Input**: Profili delle campagne di training (output di PROFILICAMPAGNE) + profilo della nuova campagna (output di PREVISIONE) + Dati di training (output di PREPARAZIONEDATI)  **Output**: Vettore di somiglianze tra campagna nuova e campagne di training + prime N campagne ordinate per grado decrescente di somiglianza alla campagna nuova.  **Descrizione**: La funzione esegue le seguenti operazioni:   * Viene calcolato il grado di somiglianza tra la campagna nuova e il set di campagne di training come percentuale di nodi in comune. * Viene estratto il sottoinsieme dei dati di input costituito dalle N (parametro, di default 20) campagne più simili alla nuova. * Viene creato il grafico scatterplot grado di somiglianza-redemption delle N campagne selezionate. |
| Funzioni sussidiarie |  |
| X.VarExp | **Input**: vettore di valori osservati + vettore di valori osservati  **Output**: Percentuale di varianza spiegata  **Descrizione**: Calcola la percentuale di varianza spiegata (da un modello) e la visualizza a video. |
| importanzaVar | **Input**: oggetto random forest  **Output**: Variabili usate nella random forest con grado di importanza + Grafico variabili-importanza  **Descrizione**: A partire dall’oggetto random forest calcola l’importanza delle variabili usate per la costruzione della foresta ordinate in senso decrescente rispetto all’importanza nella classificazione e ne fa il grafico ordinato. |
| resultsRF | **Input**: oggetto random forest + dati di input foresta  **Output**: Percentuale di varianza spiegata dalla foresta  **Descrizione**: Applica la funzione X.VarExp al vettore outcome osservato e ai valori previsti dalla foresta sui dati di training. |