Descrizione del problema da affrontare durante l'hackathon:

Durante questo secondo giorno di hackathon, i partecipanti dovranno affrontare la seguente sfida:

- *Un problema di regressione:*
  I partecipanti dovranno sviluppare modelli in grado di predire il numero di giorni rimanenti fino al default per le aziende del dataset.

Per fare ciò, avranno a disposizione un training set su cui addestrare i modelli e un test set su cui valutare le prestazioni dei modelli.

Descrizione dataset:

I dataset forniti sono file CSV denominati "train_set.CSV" e “test_set.CSV” che contengono vari indici relativi ad alcune aziende. Nel file train vengono inserite anche le due variabili che indicano se le aziende sono entrate in default e, in caso affermativo, dopo quanti giorni, così da permettere il training del modello elaborato. Elenchiamo di seguito le variabili presenti nel dataset:

Ecco l'elenco delle caratteristiche con le relative descrizioni nel formato richiesto:

- *application_ID*: Identificativo dell'applicazione
- *decision_date*: Data della decisione
- company_ID: Identificativo dell'azienda
- *external_score_ver01*: Score fornito da società esterna circa la bontà dell'azienda (nella scala, 1 è meglio)
- *external_score_ver02*: Score fornito da società esterna circa la bontà dell'azienda (nella scala, 1 è meglio)
- *late_payment_score*: Score fornito da società esterna circa la bontà dell'azienda - è legato ai ritardi nei pagamenti (nella scala, 0 è meglio)
- *external_score_late_payment_integrated*: Score che integra gli score external_score_ver01 e late_payment_score
- *external_score_moderate*: Score fornito da società esterna circa la bontà dell'azienda - definito in uno scenario macroeconomico futuro moderato (nella scala, 1 è meglio)
- *external_score_adverse*: Score fornito da società esterna circa la bontà dell'azienda - definito in uno scenario macroeconomico futuro avverso (nella scala, 1 è meglio)
- *external_score_ver03*: Score fornito da società esterna circa la bontà dell'azienda (nella scala, A è peggio)
- *age*: Età dell'azienda (in anni)
- *province*: Provincia
- *juridical_form*: Forma giuridica
- *industry_sector*: Settore industriale

- *gross_margin_ratio*: The ability of a company to generate excess funds from its operations, given a certain amount of assets
- *core_income_ratio*: The ratio between the adjusted earnings before interests and taxes (EBIT) and the revenues
- *cash_asset_ratio*: A conservative measure of a company's ability to pay its short-term liabilities, as it measures how much of the current assets would be available to pay off interests on debt
- *consolidated_liabilities_ratio*: Ratio between non-current (long term) liabilities and total liabilities. It represents how much of a company's liabilities are due in the next 12 months, highlighting pressing short-term debt positions in the balance sheet
- *tangible_assets_ratio*: Portion of non-current assets represented by tangible items, such as real estate and machinery, as opposed to intellectual property and goodwill. This number provide a high-level assessment of the quality and solidity of strategic balance sheet assets
- revenues: Fatturato dell'azienda
- *cr_available*: Booleano che indica la disponibilità di informazioni da Centrale Rischi (database di Banca d'Italia)
- region: Regione
- *geo_area*: Area geografica
- *last_statement_age*: Tempo trascorso (in anni) dal deposito dell'ultimo bilancio
- *overrun_freq_a_revoca_autoliquidanti*: Percentuale di mesi con importi non pagati nell'ultimo anno per la classe di prodotti "a revoca / autoliquidante" (o nell'ultimo periodo osservabile se non disponibile un anno intero)
- *avg_tension_a_revoca_autoliquidanti*: Media su orizzonte temporale annuale del rapporto tra l'utilizzato e l'accordato per la classe di prodotti "a revoca / autoliquidante" (i.e. l'azienda si è accordata con il sistema bancario per poter usufruire di un determinato ammontare di denaro, la tensione indica la percentuale di denaro utilizzato)
- *std_tension_a_revoca_autoliquidanti*: Deviazione standard su orizzonte temporale annuale della variabile sopra
- *max_tension_a_revoca_autoliquidanti*: Massimo valore osservato su orizzonte temporale annuale della variabile sopra
- *last_tension_a_revoca_autoliquidanti*: Ultimo valore osservato della variabile sopra
- *avg_rel_used_a_revoca_autoliquidanti*: Media su orizzonte temporale annuale del rapporto tra l'utilizzato e il fatturato per la classe di prodotti "a revoca / autoliquidante"
- *std_rel_used_a_revoca_autoliquidanti*: Deviazione standard su orizzonte temporale annuale della variabile sopra
- *max_rel_used_a_revoca_autoliquidanti*: Massimo valore osservato su orizzonte temporale annuale della variabile sopra
- last_rel_used_a_revoca_autoliquidanti: Ultimo valore osservato della variabile sopra

- *overrun_freq_a_scadenza*: Percentuale di mesi con importi non pagati nell'ultimo anno per la classe di prodotti "a scadenza" (o nell'ultimo periodo osservabile se non disponibile un anno intero)
- *avg_rel_used_a_scadenza*: Media su orizzonte temporale annuale del rapporto tra l'utilizzato e il fatturato per la classe di prodotti "a scadenza"
- *std_rel_used_a_scadenza*: Deviazione standard su orizzonte temporale annuale della variabile sopra
- *max_rel_used_a_scadenza*: Massimo valore osservato su orizzonte temporale annuale della variabile sopra
- *last_rel_used_a_scadenza*: Ultimo valore osservato della variabile sopra
- *avg_count_enti_affidanti:* Media su orizzonte temporale annuale del numero di richieste di enti che affidano denaro all'azienda
- *std_count_enti_affidanti:* Deviazione standard su orizzonte temporale annuale della variabile sopra
- *max_count_enti_affidanti:* Massimo valore osservato su orizzonte temporale annuale della variabile sopra
- *last_count_enti_affidanti:* Ultimo valore osservato della variabile sopra
- *avg_count_numero_prima_info*: Media su orizzonte temporale annuale del numero di richieste di affidamento al sistema bancario da parte dell'azienda
- *std_count_numero_prima_info:* Deviazione standard su orizzonte temporale annuale della variabile sopra
- *max_count_numero_prima_info:* Massimo valore osservato su orizzonte temporale annuale della variabile sopra
- *last_count_numero_prima_info:* Ultimo valore osservato della variabile sopra
- days_to_default: Giorni impiegati dall'azienda per andare in default.
- target: Variabile binaria che indica se l'azienda è andata in default (1) o meno (0)

Valutazione dei risultati:

- Per il problema di regressione, la metrica sarà l'errore medio assoluto (Mean Absolute Error). NB. Il Mean Absolute Error su test set è calcolato come:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Dove:

- (yi) sono i tempi di default osservati nel test set.

- (yi) sono i tempi di default predetti nel test set.

- (n) è il numero totale di osservazioni nel test set.

Consegna della Soluzione su Open Data Playground:

Alcune specifiche per il formato della soluzione che dovrete consegnare su Open Data Playground. Per ogni giorno di challenge vi sarà richiesto di caricare sulla piattaforma un file .zip con i seguenti file al suo interno:

(i) file .csv di predizione (con una sola colonna composta dai valori previsti e nessun header)
(ii) il codice riproducibile utilizzato per risolvere il problema
(iii) notebook commentato per verificare la riproducibilità del codice

Inoltre, il file .zip dovrà rispettare i seguenti requisiti di formattazione:

(i) Il file zip dovrà contenere il nome del team e non possono essere usati caratteri speciali e punteggiatura [i.e. nometeam.zip]
(ii) Il file .zip non dovrà contenere cartelle al suo interno altrimenti non verrà recepito dalla piattaforma

Su Open Data Playground avrete la possibilità di caricare diverse soluzioni per verificare quale modello sia il più performante; la piattaforma manterrà solamente lo score più alto raggiunto. Inoltre, dopo aver effettuato la sottomissione di una soluzione è necessario attendere un periodo di 30 secondi prima di consegnarne una nuova.

Valutazione generale hackathon:

Una volta che l'hackathon sarà concluso, sarà creata una classifica temporanea risultante dalla media ponderata tra i migliori risultati ottenuti dal team nei due giorni di challenge.

Specificamente il miglior risultato ottenuto nel DAY1 peserà il 60% del punteggio totale, mentre il miglior risultato ottenuto nel DAY2 peserà il restante 40%.