Università degli Studi di Cagliari

PROGETTO LABORATORIO DI BIG DATA

RISCHIO DI CRISI BANCARIA IN CONTESTI DI DIFFICOLTA' ECONOMICA

Ludovico Sanna 11/82/00398

A.A. 2024/2025

INDICE

Obiettivi del progetto

Il problema che si è andato a studiare in questo progetto può essere identificato come un problema di classificazione. È stato condotto anche uno studio sulla correlazione che c'è tra la variabile target "contcrisis" e le altre features. La variabile target è binaria e numerica, mentre le restanti features sono tutte variabili numeriche fatta eccezione per la variabile "exporter".

Per svolgere questa analisi sono stati utilizzati il modello Random Forest e la Regressione Logistica, i quali sono stati addestrati con il metodo train-test split classico. Prima di procedere con l'addestramento di questi due modelli però, è stato necessario lavorare sui dati per renderli più completi e interpretabili in quanto il dataset presentava valori nulli o incompleti e un forte sbilanciamento, insieme a uno studio della correlazione tra ciascuna variabile indipendente con la variabile dipendente.

Per valutare le prestazioni dei modelli sono state utilizzate come misure principali Accuracy, Recall, Precision, F1 e l'MCC. Inoltre, le prestazioni sono state rappresentate graficamente tramite l'utilizzo della curva ROC.

Scelta del dataset

Il dataset "finaldataset_1" raccoglie una serie di dati riguardanti l'andamento della crescita dell'export di diversi settori, in diversi paesi, durante periodi di crisi finanziaria e del rischio di contrarre una crisi bancaria.

È stato scelto questo dataset perché, in un'economia sempre più indirizzata verso un concetto di globalizzazione, si prestava meglio per il tipo di analisi che si voleva svolgere.

Il dataset si presenta con uno schema composto da 44 features e 39588 osservazioni. La variabile target "contcrisis" esprime la presenza o meno di una crisi bancaria mentre le altre variabili riguardano i paesi esportatori, i codici dei prodotti, l'anno e il valore delle trades, la crescita delle esportazioni, gli indicatori delle situazioni bancarie e indicatori macroeconomici.

Tipologia del problema e architettura della soluzione

Procedendo con l'analisi, è stata necessaria una prima fase di preparazione dei dati nella quale si sono implementate le funzioni per la gestione dei valori nulli o mancanti e per la trasformazione di alcuni tipi di dati, come l'indicizzazione delle features categoriche in modo tale che potessero essere utilizzate dai modelli. Più nello specifico, nella fase di trasformazione delle variabili categoriche, è stato necessario, non solo indicizzare le variabili

categoriche, ma anche eseguire l'encoding di alcune variabili come "exporter". Per quest'ultima è stata necessaria una trasformazione in formato binario per evitare che venissero create troppe classi che avrebbero potuto compromettere le prestazioni dei modelli successivamente. Inoltre, questo processo è stato necessario perché l'elevato numero di valori mancanti o nulli può introdurre dei bias rendendo i dati più complicati da gestire, di difficile interpretazione e di conseguenza la creazione di modelli meno performanti e meno affidabili.

Per avere una situazione più chiara sulla relazione tra le features e la variabile target è stato condotto uno studio preliminare della correlazione tra le singole variabili indipendenti e la variabile target. Studio che è stato poi tenuto in considerazione anche nella fase di valutazione con l'utilizzo dell'MCC. L'impiego di questa misura può risultare utile perché identifica il coefficiente di correlazione tra le predizioni dei modelli e i valori reali della variabile target tenendo conto dello squilibrio della classe. Tutto questo procedimento è servito a evidenziare quali variabili fossero maggiormente correlate. In particolare, nel caso specifico di classificazione binaria, la predittività di ogni features è proporzionale alla sua distanza dallo zero.

Alcune delle variabili più predittive e meno predittive:

```
blanguar: 0.6782   GDPgr: -0.1180
policytot: 0.6510  recession: 0.0870
BANK: 0.6475       year: -0.0551
forbb: 0.6076      stmktcap: -0.0495
liqsup: 0.6068     developing: -0.0349
```

Successivamente a questo, definendo la variabile target, si è valutato lo sbilanciamento del dataset.

Come si può notare dal risultato ottenuto, la classe target è risultata fortemente sbilanciata.

```
+-----+-----+
|label|count|
+-----+-----+
|    0|38837|
|    1|  751|
+-----+-----+
```

Con la presenza di un forte sbilanciamento si può prevedere che alcune delle misure delle prestazioni dei modelli restituiscano valori fuorvianti. Come ad esempio, l'Accuracy, che potrebbe avere valori elevati perché predice solo la classe maggioritaria; oppure, la coppia

Precision-Recall, che fornisce una valutazione più affidabile perché si concentra sulla capacità dei modelli di individuare correttamente la classe minoritaria.

Come soluzione si è optato per assegnare dei pesi maggiori alle osservazioni della classe minoritaria. In questo modo, il dataset non viene alterato, mantiene tutte le informazioni e risulta essere più efficiente con Spark. Inoltre, questo metodo è semplice da implementare, non genera ulteriori dati e funziona bene con i modelli di Random Forest e di Regressione Logistica che si andranno a utilizzare. Questo succede perché all'interno dei modelli viene data più importanza alla classe minoritaria e di conseguenza si potrebbe ridurre la quantità di falsi negativi e contenere i falsi positivi.

Per quanto riguarda l'addestramento dei modelli si è optato per il metodo del train-test split classico. È stato scelto tale metodo perché, visto il forte sbilanciamento della classe, oltre ad adattarsi a entrambi i modelli scelti, risulta essere anche una tecnica robusta e semplice da implementare.

La scelta dei modelli per studiare la correlazione è ricaduta sul Random Forest e sulla Regressione Logistica perché, come detto in precedenza, si adattano bene ai dataset bilanciati con i pesi.

Il primo risulta essere un modello robusto e funziona bene anche con features di diversi tipi. La struttura di questo modello supporta direttamente i pesi.

Per quanto riguarda la Regressione Logistica, invece, è un modello lineare semplice e interpretabile. All'interno dell'analisi, permette di gestire gli squilibri di classe e infine è facilmente integrabile con il metodo di addestramento dei modelli scelto.

I due modelli permettono quindi una valutazione comparativa bilanciata.

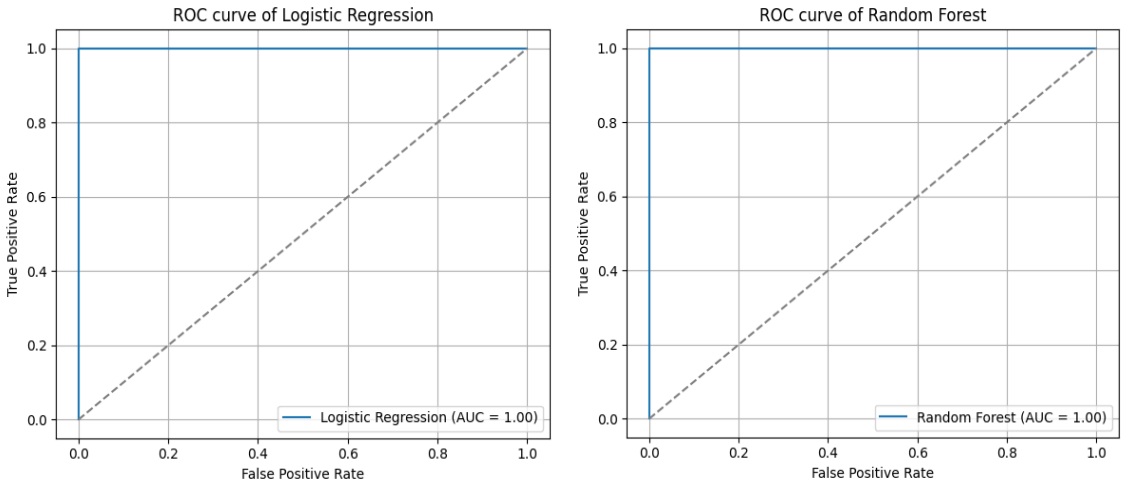
Valutazione risultati

Per la valutazione della performance dei due modelli sono state utilizzate le misure principali come l'Accuracy, la Precision, la Recall, l'F1, l' AUC e la curva ROC. Visto il forte sbilanciamento e il tipo di dati presi in considerazione, è stato utilizzato anche l'MCC, il quale misura molto bene i casi di problemi binari come in questo caso.

```
Random Forest:      Logistic Regression:
Precision: 1.0000   Precision: 1.0000
Recall: 1.0000      Recall: 1.0000
Accuracy: 1.0000    Accuracy: 1.0000
F1: 1.0000          F1: 1.0000
Auc: 1.0000         Auc: 1.0000
MCC: 1.0000         MCC: 1.0000
```

Esaminando i valori ottenuti dalle misure di performance dei modelli utilizzati, come anche dell'MCC, si può concludere che entrambi i modelli hanno classificato correttamente ogni esempio. Allo stesso tempo, però, valori perfetti possono risultare sospetti.

Inoltre, questi valori sono anche confermati dal grafico della curva ROC.



Infatti, il valore perfetto della AUC, lo si può evidenziare graficamente all'interno dei grafici dei due modelli della curva ROC. Un valore AUC = 1.0 sta a significare che entrambi i modelli sono in grado di separare perfettamente le classi in base alle probabilità predette ma questo valore può evidenziare anche un problema.

Questi risultati perfetti trovano conferma nello studio preliminare della correlazione delle variabili, che ha evidenziato la presenza di alcune features con una forte associazione positiva. Questo supporta l'ipotesi che il dataset possa contenere pattern facilmente individuabili dai modelli.

Conclusioni

Per lo svolgimento di ciascuna fase di questo progetto è stata utilizzata la libreria Spark. Le funzioni contenute al suo interno hanno permesso una gestione più semplice di alcune problematiche relative alla struttura del dataset, a cominciare dalla fase di data cleaning.

Dall'analisi della correlazione si è potuto capire che i modelli avrebbero avuto buone prestazioni. In particolare, in virtù della presenza di diverse variabili con forte correlazione rispetto alla variabile target.

In effetti, le misure di performance utilizzate, cioè Accuracy, Precision, Recall, F1, MCC e AUC, hanno restituito valori perfetti, confermando quello che si era ipotizzato nella fase esplorativa. Questo potrebbe indicare che la semplice assegnazione dei pesi alle classi, adottata per compensare lo sbilanciamento della variabile target, non sia stata sufficiente a gestire adeguatamente la complessità del problema, suggerendo l'opportunità di adottare tecniche più sofisticate.