Statlog - German credit
Report per l'Esame di Fondameti di Machine Learning

Alessandro Pansera

152543
Ing. Informatica
282423@studenti.unimore.it

Abstract

Il sistema bancario moderno si basa sulla cessione di prestiti ai propri correntisti al fine di realizzare profitto attraverso interessi e fidelizzare il cliente. Alla base del modello di business descritto c'è la necessità di ponderare scrupolosamente l'affidabilità creditizia di un cliente affinché non risulti in futuro insolvente. Il Machine Learning permette di risolvere problemi di questa natura mediante la classificazione tramite la quale offrire un supporto al normale svolgimento delle mansioni di un operatore bancario.

1 Dataset

Il dataset è composto da dati bancari raccolti tra il 1973 e 1975 in Germania, è stato donato all'università di Amburgo e pubblicato nel 2019 all'interno dell'archivio dell'università della California. All'interno del dataset troviamo sia features categoriche che numeriche, non sono presenti sample con attributi nulli. Essendo il dataset, per la natura dei features che lo compongono, multivariato, occorre applicare delle tecniche di replacing per permettere il training di un classificatore. E' stata realizzata, a scopo puramente dimostrativo, la funzione che esegue la standardizzazione delle features ma non viene applicata in quanto degrada le performance. Segue una visualizzazione delle features che più risultano essere correlate tra loro per il problema in questione.

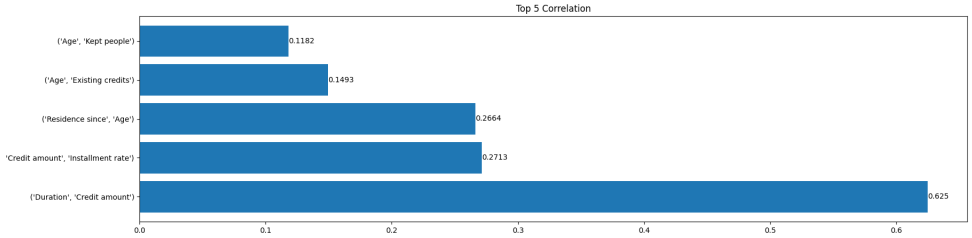

Figure 1: TOP 5 correlated features

Le 13 variabili categoriche verranno tutte rimpiazzate mediante operazioni di replacing basate sul codice presente per ogni categoria presente:

```
# original category
cat = "A201"
# replaced category
new_cat = int(cat[-2:])
```

Per consentire la modifica dinamica delle codifiche con cui eseguire il replacing è stato predisposto un sistema basato su file .json importabili i quali indicano per ogni feature le regole di sostituzione. L'obbiettivo è quello non di variare il codice ma di cambiare il .json importato per il replace rendendo il processo più rapido e meno soggtto a potenziali errori.

1.1 Var. Target

La variabile target presente all'interno del dataset ammette 2 valori interi che sono 1 ('Good borrower') e 2 ('Bad borrower'). Considerando la natura del problema si possono eseguire le predizioni mediante un modello di classificazione binaria. Non è necessario eseguire il replace per i valori della variabile target.

1.2 Considerazioni

Prima di scegliere il tipo di algoritmo da impiegare per la classificazione binaria occorre studiare con attenzione il dataset che si ha a disposizione. Il primo problema facilmente osservabile risiede nel mancato bilanciamento delle classi all'interno del dataset. La classe 'Good Borrower' riporta 700 samples mentre 'Bad Borrower' 300 samples; l'entropia rilevata per i dataset vale:

$$H(T) = -\sum_{n=1}^{i} p_n * \log_2 p_n = 0,88$$

Per evitare problemi legati a predizioni 'positive' sbagliate occorre applicare delle tecniche di over-sampling o under-sampling che verranno elencate alla voce 2.1.

Per poter creare un algoritmo di Machine Learning realmente applicabile ad un contesto lavorativo è importante valutare anche aspetti indipendenti dalla pura natura dei dati. Il classificatore che si vuole realizzare deve essere in grado predirre con precisione 'Good borrower' e 'Bad borrower'; a livello applicativo per un istituto di credito è più importante predirre con precisione i casi appartenenti alla prima classe che alla seconda. Se infatti sbagliamo la predizione legata ad un 'Finto good borrower' rischiamo di incorrere in perdite economiche al contrario di un 'Finto bad borrower' per cui la perdita economica è solo ipotetica.
Sarà necessario in fase di training andare allora a considerare le problematiche appena citate per andare a realizzare un classificatore in grado di fornire delle buone performance.

2 Modello

Come già detto il problema di Machine Learning trattato è di classficazione binaria. Per realizzare un modello di classificazione binaria sono disponibili numerose tecniche ma quelle implementate per la realizzazione del progetto sono la Logistic Regression, il K-NearestNeighbour e il Random Forest. A livello pratico i modelli implementati sono stati scelti in maniera empirica cercando di raccogliere tutti i modelli affrontati a lezione: la Logistic Regression è stata introdotta per fornire un esempio di modello parametrico, il KNN per portare un esempio di modelli non parametrici e il Random Forest per portare un esempio di ensembling.

2.1 Pre-processing

Come detto in fase di descrizione del dataset sarà necessario eseguire delle operazioni di preprocessing sui dati oltre al replacing. All'interno del progetto sono state implementate diverse soluzioni per far fronte allo sbilanciamento tra classi che costituiscono il dominio della variabile target:

- Under-sampling: tecnica basata sul ridurre i samples appartenenti alla classe dominante.
  - Near miss: è un algoritmo di under-sampling basato sulla rimozione randomica dei samples appartenenti, per il nostro problema, alla classe 'Good borrower' che è in maggioranza.

- Over-sampling: consiste nel generare fedelmente dei samples sulla base di assunzioni relative alla classe di appartenenza. Gli algoritmi implementati per il progetto appartengono alla famiglia dei 'Synthetic Minority Oversampling Technique' . Gli algoritmi SMOTE non operano semplicemente duplicando i dati bensì selezionano i samples più simili basandosi sulle features e ne generano di nuovi sulla base dell'interpolazione tra i samples selezionati nello spazio multidimensionale.
  - K SMOTE: attraverso l'algoritmo K-means vengono generati dei cluster all'interno dei quali vengono selezionati i sample che verranno usati come base per l'over-sampling. Successivamente vengono scartati i cluster in cui abbiamo una sproporzione tra sample appartenenti alla classe maggioritaria e minoritaria. Per i cluster rimanenti applico KNN per classificare.
  - SVM SMOTE: l'algoritmo è un successore di SMOTE e in questo caso genera dei dati sintetici sul confine tra due classi, confine rilevato grazie alle SVM.
  - ADASYN: il principio di funzionamento è differente dai precedenti due in quanto vengono creati dei dati sintetici negli spazi multidimensionali in cui la classe minoritaria è poco densa.

2.2 Performance

Per visualizzare i risultati comparativi tra i vari modelli, dopo aver eseguito main.py si può consultare il file log/master.log che riporta per ogni misura delle performance analizzata i 10 migliori risultati registrati tra tutti i modelli. La tecnica di cross-validation utilizzata per partizionare il dataset è la stratificazione.
Per ogni algoritmo vengono testate tutte le tecniche di bilanciamento del dataset. Sarà possibile eseguire il training e la relativa verifica delle performance in 2 modalità:

- heavy: viene eseguita l'ottimizzazione degli iper-parametri ed inoltre per ogni coppia ¡modello-tecnica di sampling¿ viene eseguita la media tra 5 fit separati. I dati che seguono faranno riferimento a questo tipo di test in quanto le misure raggiunte sono migliori e matematicamente più affidabili.
- light: offre la possibilità di eseguire il codice in poco tempo ed è utile per comprendere il funzionamento dello script.

2.3 Analisi performance

Come già asserito alla voce 1.2 è altamente vincolante la precisione con cui vengono predetti i 'Good borrower'. Di seguito con FP verranno indicati i 'False good borrower' e con TP i 'True good borrower'. All'interno di una Confusion Matrix il risultato desiderato si traduce in un gap, il più elevato possibile, tra gli elementi all'interno della prima colonna (rappr. PyCharm).
E' comunque importante avere una buona precisione in fase di predizione dei TN altrimenti il modello, a livello generale, perderebbe di precisione ed è un risultato non accettabile.
Fissato l'obbiettivo che si vuole raggiungere si può procedere allo studio di quelli che sono gli score a cui si è maggiormente interessati per la valutazione della bontà del modello.
Se l'F1 score è più un indicatore generico che è sempre buona norma valutare per stimare la bontà generale di un classificatore F.D.R., Precision e Recall sono specifici per il tipo di problema in questione. L'F.D.R., 'False discovery rate', permette di valutare il rapporto tra FP e TP; la metrica è perciò una delle principali da considerare in quanto valuta esattamente quello che è il principale vincolo del problema di classificazione in questione.

$$F1score = \frac{2 * precision * recall}{precision + recall}$$

$$F.D.R. = \frac{FP}{TP + FP}$$

Precision e Recall misurano entrambe la qualità con cui vengono rilevati i TP e FP. Se la Precision misura la precisione delle predizioni positive invece la Recall ne misura la completezza, di seguito sono riportate le formule di calcolo.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{FP}{TP + FN}$$

Precision e Recall sono inversamente proporzionali ed è possibile osservarlo anche dai test eseguiti sul classificatore. L'esempio più apprezzabile è legato al confronto tra i risultati ottenuti con la Logistic Regression e il Random Forest applicando l'SVM SMOTE per l'over-sampling.
Più si è precisi a definire ciò che si cerca minore è la completezza delle predizioni positive. Per la tipologia di problema in questione, considerando soprattutto il settore in cui si opera, si predilige una buona Precision a discapito di una miglior Recall.

Le soluzioni migliori per la tipologia di problema considerato sono raggiunte applicando l'algoritmo di Logistic Regression impiegando l'SVM SMOTE in fase di pre-processing per l'over-sampling oppure impiegando il Random Forest facendo uso di K-SMOTE o SVM SMOTE per l'over-sampling. Le strade appena citate presentano differenze legate agli score che seguono:

- FDR
- Precision/Recall

Gli score citati servono per valutare il giusto tradeoff per la soluzione che si vuole applicare. Per quanto concerne l'F1-score i tre modelli indicati riportano uno score compreso tra l'85% e l'82%.
Il Random Forest presenta, con SVM/K-SMOTE, una Recall molto elevata a discapito di una Precision nettamente inferiore (circa del 10%) rispetto al medesimo valore raggiunto mediante la Linear Regression.
L'upgrade legato all'applicazione della Logistic Regression, combinata con l'SVM SMOTE, è riconducibile ad un incremento della Precision e ad una riduzione dell'FDR.
Applicando la Logistic Regression con l'SVM SMOTE in fase di pre-processing è possibile apprezzare una Precision media superiore del 5% rispetto alla miglior performance media indicata all'interno del sito che rende fruibile il dataset.
Per ricapitolare, considerando tutti i modelli generati, quelli che hanno generato i risultati migliori fanno uso della Logistic Regression impiegando una Support Vector Machine per l'over-sampling producendo risultati soddisfacenti per tutti gli score di riferimento; ciò che maggiormente influisce nella valutazione sono il massimo punteggio ottenuto mediamente per la Precision ed il minimo relativamente all'FDR.
Un'ulteriore soluzione apprezzabile, ma meno puntale, è prodotta del Random Forest applicando, per l'over-sampling, l'SVM-SMOTE oppure il K-SMOTE.
Eseguendo la cross-validation può capitare che alcuni modelli generino risultati sporadicamente ottimi per singole metriche perciò non sono stati menzionati in quanto non stabili e globalmente non affidabili. Per esempio il Random Forest con l'under-sampling ha prodotto, in fase di sperimentale, un FDR del 12%, a discapito di F1-Score nettamente inferiore rispetto ai modelli citati in precedenza, peccando di costanza nelle rilevazioni successive. Ciò significa che i modelli proposti, a seguito di numerosi test, hanno mantenuto delle performance costanti sotto tutti i punti di vista; è preferibile optare per soluzioni costanti e bilanciate rispetto a soluzioni che presentano score sbilanciati.
Per completezza, all'interno dei log prodotti dal programma per ogni classificatore, è riportata anche l'accuracy in quando rappresenta una metrica che è sempre buona norma considerare pur non essendo fondamentale per questo problema.
I grafici riportati alla fine del documento illustrano le performance raggiunte riportando, per ogni metrica, le top 5 valutazioni raggiunte tra le 12 possibili combinazioni disponibili. E' inoltre presente un istogramma che, tramite una media pesata tra tutti gli score, valuta i top 5

modelli applicabili. Tutte le codifiche presenti all'interno degli istogrammi sono riportate alla voce 2.4.

2.4 Implementazione

Il modello di classificazione può essere addestrato attraverso differenti combinazioni tra algoritmi e tecniche di under-sampling/over-sampling. Per permettere un rapido interscambio tra tecniche e algoritmi usati è stato implementato, tramite incapsulamento, il Factory Pattern che permette di pilotare tramite 2 stringhe tutti i fattori che portano all'addestramento del modello. E' stata inoltre prevista la possibilità di salvare il classificatore affinché sia utilizzabile in seguito al training.
Seguono le codifiche presenti all'interno del progetto per rappresentare modelli, tecniche di over-sampling e under-sampling:

- Modelli:
  - RF: Random Forest
  - LR: Logistic Regression
  - KNN: K-Nearest Neighbour
- Tecniche di sampling:
  - US: under-sampling con Near Miss
  - OS_K: over-sampling con K-SMOTE
  - OS_SVM: over-sampling con SVM-SMOTE
  - OS_ADASYN: over-sampling con ADASYN

Ogni classificatore realizzato verrà salvato in formato .joblib all'interno della cartella classifier/.
Per i classificatori è stata prevista la possibilità di eseguire il tuning degli iper-parametri. Di seguito sono riportati, per ogni modello, gli iper-parametri per cui viene eseguita l'ottimizzazione.

- Random Forest:
  - max. depth
  - min. samples split
  - n. estimators
- Logistic Regression
  - c
- KNN
  - n. of neighbours

E' possibile visualizzare, tramite due file excel contenuti nella cartella dataset/ il dataset scaricato pre e post normalizzazione (data.xlsx e data_normalized.xlsx). Si noti che l'over-sampling viene fatto solo sul dataset di training, non su quello usato in fase di validazione, per avere dati puri in fase di testing.
Ogni modello addestrato produce un file di log dedicato all'interno della cartella log/ con i riferimenti alle performance e agli iper-parametri ottimizzati ricavati per ogni test (se la modalità di esecuzione lo prevede).
Ci aspetteremo, per quanto già detto in merito alla fase di training, che per ogni file siano presenti 5 test per avere una media delle performance operando in modalità 'heavy'. L'overall delle performance e i dati riassuntivi sono presenti all'interno del file log/master.log
Il comando con cui si consiglia di eseguire un test rapido è il seguente:

```
python main.py --mode light --source dataset/data.csv --verbose 1
```

Il tempo per eseguire lo script in modalità 'heavy' si aggira intorno ai 35 minuti con un processore AMD Ryzen 5. Tramite l'utilizzo di parametri sarà possibile stabilire in che modalità operare, configurare il path da cui leggere il dataset, settare il livello di verbosità ed in caso indicare se salvare il dataset intermedio normalizzato.

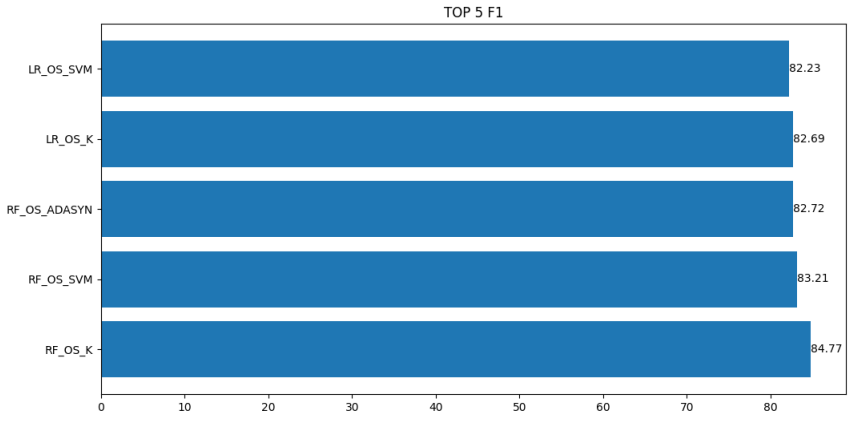

Figure 2: TOP 5 F1-Score

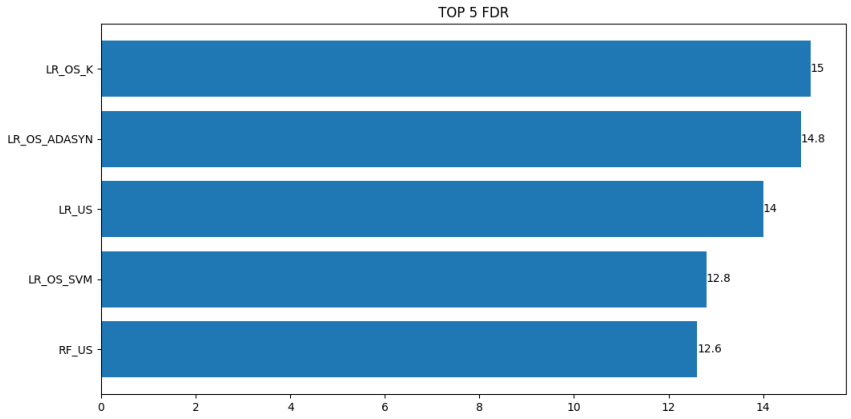

Figure 3: TOP 5 F.D.R.

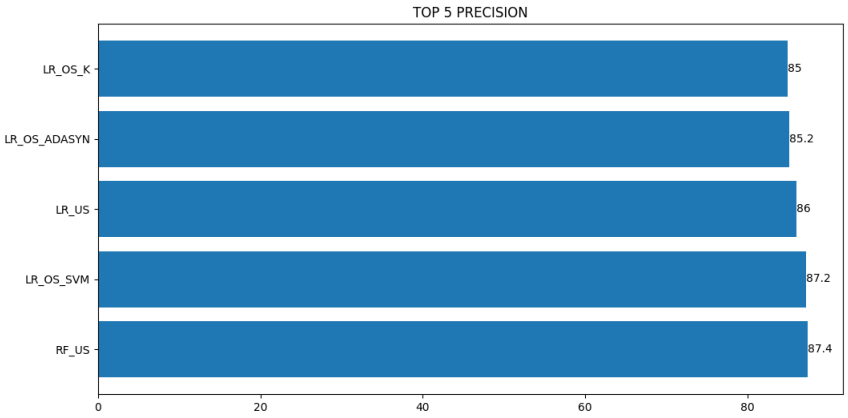

Figure 4: TOP 5 Precision

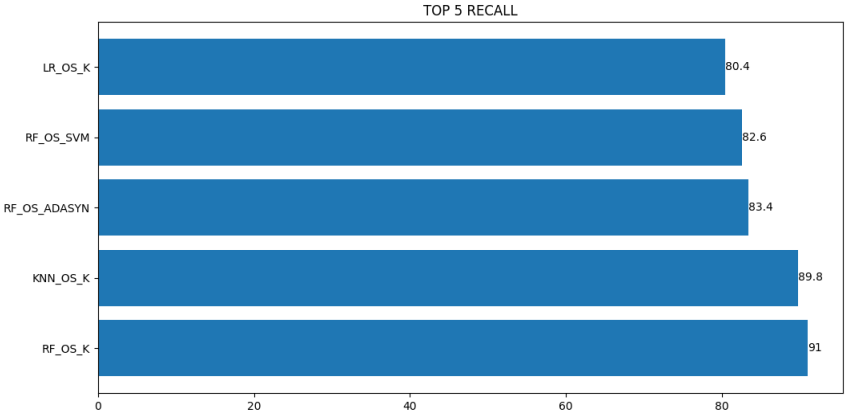

Figure 5: TOP 5 Recall

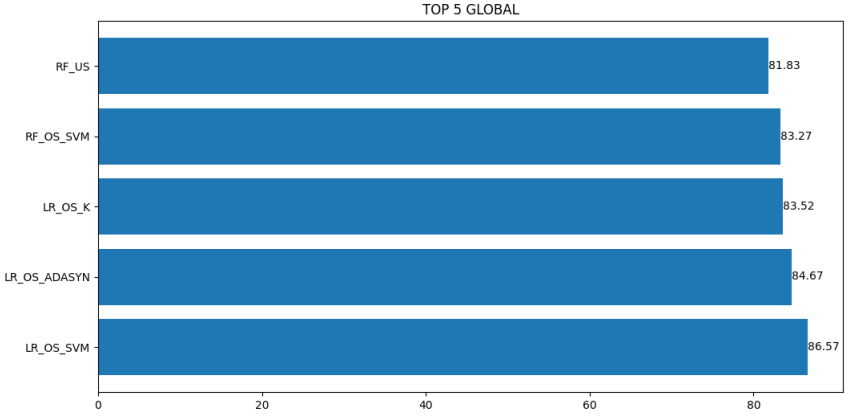

Figure 6: TOP 5 Global