ANALISI DEL RISCHIO DEL CREDITO

Giorgio Bini, Daniele Ciciani, Lorenzo Famiglini, Davide Lagano, Davide Mancino

Università degli Studi di Milano- Bicocca

Abstract
Data l'interconnessione dei diversi istituti bancari, la stabilità del sistema economico può essere compromessa dalle decisioni delle singole banche. L'analisi del rischio del credito può tutelare il settore bancario da eventuali crisi finanziarie, come la crisi dei subprime del 2007. In quella circostanza sono stati concessi prestiti ad alto rischio da parte degli istituti di credito in favore di clienti a forte rischio debitorio. Le nuove tecnologie possono supportare le banche in scelte decisionali critiche, come la concessione di un prestito. Per questo motivo, le banche raccolgono sempre più informazioni sui soggetti richiedenti un prestito al fine di poter applicare tecniche predittive, analisi statistiche per determinare i vari livelli di rischio. L'obiettivo di questo progetto è quello di costruire dei modelli in grado di prevedere se un debitore rispetterà i suoi impegni nei confronti di una banca, al fine di poter tutelare gli istituti di credito dal rischio di insolvenza.

Contents

1 Introduzione

Al fine di raggiungere il nostro obiettivo, si è deciso di analizzare il dataset "German.csv" presente sulla piattaforma "Uci".
Il dataset originario è composto da 1000 osservazioni e 21 variabili, quali:

Status of existing checking account: presenza del conto e il valore del conto stesso.
Duration in month: durata del prestito.
Credit history: storico relativo ai debiti contratti in precedenza.
Purpose: motivo del prestito.
Credit amount: ammontare del prestito.
Savings account/bonds: se il debitore possiede un conto di risparmio/obbligazioni.
Employment: numero degli anni di lavoro.
Installment rate: tasso rateale del prestito.
Personal status and sex: indica lo stato familiare rispetto al sesso.
Other debtors / guarantors: presenza di garanti.
Present residence since: indica da quanto tempo il debitore è residente.
Property: fonti di reddito.
Age: età del debitore.
Other installment plans: altri piani di rateizzazione
Housing: indica se l'alloggio è in affitto, di proprietà o regalato.
Number of existing credits at this bank: numero di crediti esistenti presso la banca in questione.
Job: posizione lavorativa.
Number of people being liable to provide maintenance for: persone disposte a sostenere il debito.
Telephone: se il numero telefonico è presente o meno nel database della banca.
Foreign worker: se il lavoratore è straniero o meno
Status of Loan: indica se è un buono o cattivo debitore.

È stata scelta "Status of Loan" come variabile target, di conseguenza l'obiettivo dell'analisi è quello di prevedere se un individuo sarà in grado o meno di restituire un prestito alla banca, quindi, se è un buono o cattivo debitore.
La relazione è strutturata nel modo seguente:
• Nel secondo capitolo è stata effettuata un'esplorazione dei dati ed è stato svolto un lavoro di discretizzazione e pulizia delle variabili. .
• Nel capitolo 3 abbiamo elencato i modelli adoperati per risolvere il problema di classificazione. Inoltre verrà l'operazione di Features Selection che è stata effettuata per ridurre la dimensionalità del dataset iniziale.
• Nel quarto capitolo verranno discussi i risultati delle procedure di Hold Out e Cross Validation al fine di individuare il modello più efficace nel rispondere alla nostra domanda target. Verrà inoltre esplicitata la fase di Validation, utile per identificare quali tecniche di Features Selection risultano ottimali per ciascun modello.

2 Esplorazione dei dati e preprocessing

2.1 Esplorazione

È stata effettuata un'analisi esplorativa per avere una visione generale dei dati a disposizione. Sono state utilizzate le librerie Pandas (per l'elaborazione dei dati) e Seaborn (per la visualizzazione) che sono disponibili su Pyhon. Rimandiamo al file "Credit Risk Exploration Data" per la visualizzazione completa dell'analisi esplorativa e riportiamo qui di seguito solo i risultati di maggior interesse.

In via preliminare è stata calcolata una matrice di correlazione di Pearson con lo scopo di individuare le variabili quantitative maggiormente correlate.

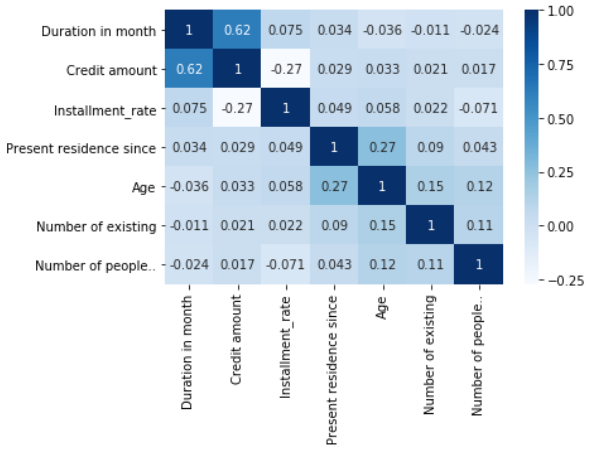

Fig. 1 Matrice di correlazione di Pearson

Dall'analisi visuale di questa heatmap emerge che tra le variabili in esame, si riscontrano solo correlazioni deboli ad eccezione degli attributi "Duration in Month" e "Credit Amount" tra i quali sussiste una correlazione positiva moderata.
Per individuare le relazioni insite tra le variabili categoriche e la variabile target, invece, è stato applicato il test di ipotesi di chi-quadro. Tale test è stato applicato per tutte le variabili qualitative che sono state, una ad una, sottoposte al test in coppia con la variabile "Status of Loan" al fine di verificarne l'indipendenza (ipotesi nulla).

VARIABLES	CHI SQUARED	DEGREES	P-VALUE
SAVING ACCOUNT/BONDS	36.09	4	2.7e^-7
STATUS OF EXISTING..	123.72	3	1.21e^-26
CREDIT HISTORY	61.69	4	1.27e^-12
PURPOSE	33.35	9	0.0001
EMPLOYMENT	18.36	4	0.001
PERSONAL STATUS AND SEX	9.6	3	0.022
OTHER DEBTORS/GUARANTORS	6.64	2	0.036
PROPERTY	23.71	3	2.85e^-5

OTHER INSTALLMENT PLANS	12.83	2	0.001
HOUSING	18.19	2	0.0001
JOB	1.88	3	0.596
TELEPHONE	1.17	1	0.278
FOREIGN WORKER	5.82	1	0.015
AGE	14.19	3	0.002
NUMBER OF PEOPLE LIAB.	0.0	1	1.0

Tab. 1 Valori relativi al test di chi-quadro

Dai risultati di tale test emerge che tutte le variabili categoriche, ad eccezione di "Job" e "Telephone", influenzano la variabile target "Status of Loan" ad un livello di significatività del 95%.

2.2 Preprocessing

Dall' esplorazione dei dati è stata inoltre riscontrata un'intensa sparsità nella distribuzione delle variabili continue "Age" e "Duration in Month". Tali attributi sono stati dunque discretizzati per quantili, che risulta essere la metodologia più rappresentativa nel suddividere sia le fasce di età, sia la durata del prestito in modo adeguato. Gli intervalli finali sono i seguenti:

- *Age:* 18-27 (young), 27-33 (adult), 33-42 (senior), 42+ (aged)
- *Duration in month*: 0-12 (short-term), 12-18 (middle-term), 18-24 (long-term), 24+ (extend-term)

Grazie alla discretizzazione, non solo è stato risolto il problema della sparsità, ma sono stati anche resi più interpretabili i dati.

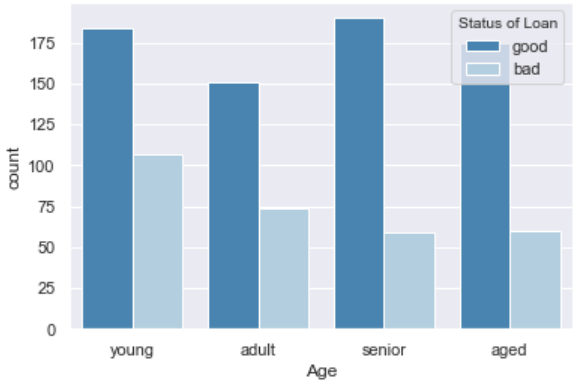

Fig. 2 Distribuzione dei valori di *Status of loan* raggruppati per *Age*

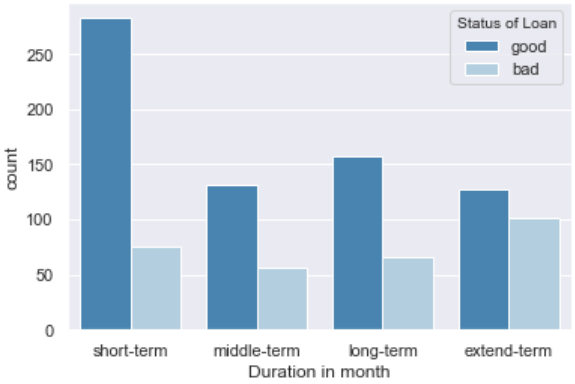

Fig. 3 Distribuzione dei valori di *Status of loan* raggruppati per *Duration in month*

Qui sopra viene fornita una rappresentazione di come si distribuisce la variabile "Status of Loan" rispetto alle variabili discretizzate "Age" e "Duration in Month". Come possiamo vedere dal grafico 1, con l'avanzare dell'età il numero di "cattivi debitori diminuisce". Il secondo grafico invece ci fornisce un'importante informazione: nei prestiti a breve termine risultano esserci molti più debitori appartenenti alla classe "Good" rispetto alla classe "Bad". Tale differenza risulta invece meno significativa nei prestiti in "extended-term" (nella quale la presenza di "cattivi debitori" risulta essere maggiore).

Si è deciso inoltre di convertire le variabili "Telephone" e "Foreign worker" da stringa ad intero e la variabile "Status of Loan" da intero a stringa.

Al fine di rendere comprensibile e chiaro il workflow, si è scelto di rinominare le modalità assunte da alcune variabili:

- *Status of existing checking account:* negative, 0-200, 200+, n.c.a.
- *Credit history*: A (prestiti mai richiesti o estinti), B (debiti presso la banca attuale già estinti), C (debiti esistenti finora regolarmente rimborsati), D (ritardi nei pagamenti passati), E (conto critico/altri crediti esistenti presso questa banca)
- *Purpose:* new car, used car, furniture, radio/TV, domestic appliances, repairs, education, vacation, retraining, business, others
- *Savings account/bonds:* 0-100, 100-500, 500-1000, 1000+, unknown/n.s.a.
- *Employment:* 0, 0-1, 1-4, 4-7, 7+
- *Personal status and sex*: male divorced/separated, female divorced/separated/married, male single, male married/widow
- *Other debtors / guarantors:* none, co-applicant, guarantor

- *Property:* A (beni immobiliari), B (assicurazione sulla vita/accordi di risparmio), C (macchine o altri), D (nessuna proprietà/non conosciuto)
- *Other instalment plans:* bank, stores, none
- *Housing:* rent, own, for free
- *Job:* A (lavoratore altamente qualificato), B (impiegato), C (disoccupati residenti), D (disoccupati non residenti)
- *Telephone:* Yes, No
- *Status of Loan:* Bad, Good

2.3 Partizionamento

È stato deciso di partizionare il dataset nel seguente modo: il 10% delle osservazioni è stato riservato al validation set, mentre il restante 90% è stato diviso in training set (67%) e test set (33%). Il procedimento di partizione del dataset viene effettuato tramite uno *Stratified sampling*, nel quale è stata scelta *Status of Loan* come variabile di riferimento per mantenere approssimativamente uguale la proporzione delle classi "good" e "bad" nei vari sottoinsiemi dei dati.

2.4 Oversampling data

Il dataset iniziale ci pone un problema di classe sbilanciata. All'interno della variabile target *Status of Loan*, infatti, troviamo il 70% dei valori di tipo 'good' e il 30% dei valori di tipo 'bad' e per questo motivo va opportunamente trattata. Per sviluppare modelli che generalizzino e prevedano bene soggetti con valore 'bad' nella variabile *Status of Loan*, è stato scelto di effettuare, in fase di addestramento, delle operazioni di oversampling dei dati per arricchire il training set. Il metodo utilizzato per effettuare l'oversampling è stato lo *SMOTE - Synthetic Minority Over-sampling Technique* [1] che aggiunge occorrenze sintetiche alla classe minoritaria. Così facendo, è stata bilanciata la cardinalità tra i valori della classe maggioritaria ('good') e quelli della classe minoritaria ('bad') della variabile *Status of Loan*. Tale procedura ci permetterà di avere prestazioni di classificazione migliori in tutti quegli algoritmi di apprendimento supervisionato che performano meglio quando la composizione delle classi è bilanciata.

3 Modelli e Features Selection

3.1 I modelli utilizzati

Per riuscire a classificare e prevedere i valori dell'attributo *Status of Loan* sono stati utilizzati i seguenti algoritmi di Machine Learning:

Decision Tree – J48: è l'implementazione Weka dell'albero di decisione C4.5 [2].

***Random Forest*:** classificatore d'insieme che è composto da alberi di decisione [3].

Naive Bayes: classificatore bayesiano che si basa sull'applicazione del teorema di Bayes [4].

Support Vector Machines – SVM: macchina a vettori di supporto [5].

Logistic: multinomial logistic regression con ridge estimator [6].

MLP: Multi Layer Perceptron [7].

3.2 La selezione delle variabili

Al fine di individuare attributi ridondanti o irrilevanti, abbiamo deciso di applicare la procedura di *Features Selection* ad ogni modello. Il metodo attraverso il quale sono stati individuati i sottoinsiemi di attributi ottimali per il problema di classificazione binaria, è stato il *Wrapper*. La porzione di dati con il quale sono stati addestrati i modelli è stata il training set, sottoposto a procedura di ricampionamento. Sono state sviluppate due tecniche di selezione automatica delle variabili: la *Forward inclusion* e la *Backward elimination*. In entrambi i casi, come misura di performance è stata scelta l'accuratezza.

4 Analisi e Risultati

I metodi utilizzati per ottenere una valutazione delle performance dei vari classificatori scelti sono stati: *Hold out* e *Cross validation*.

4.1 Hold out

Il metodo Hold Out è stato applicato a tutti i sei modelli di classificazione elencati in precedenza così da poterne calcolare gli indici di *accuracy*, *precision* e *recall*.

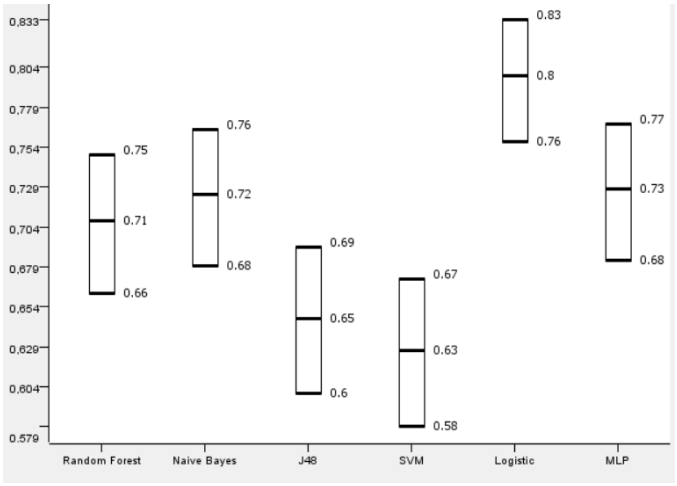

Fig. 4 *Box plot dell'Accuracy dei 6 modelli*

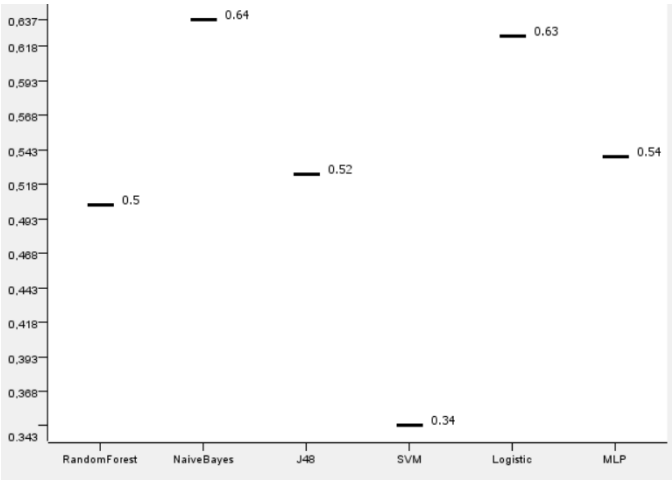

Fig. 5 F-measure dei 6 modelli

Nella figura 4, vengono rappresentati i box plot relativi all'intervallo di Wilson associato alla misura di Accuracy dei sei modelli a livello di confidenza del 95% (il valore centrale rappresenta la stima puntuale). La figura 5, invece, rappresenta il valore della stima della F-measure. Dalla prima figura si può evincere come i modelli che presentano l'accuracy migliore sono: Logistic (0.80) e MLP (0.73). Tuttavia, se prendiamo in considerazione anche la Fig.5, dal momento che l'efficacia di un modello per problemi di classe sbilanciata va valutata anche tenendo in considerazione l'F-measure, possiamo affermare che i modelli che performano meglio sono Logistic e Naive Bayes. Infatti, quest'ultimo classificatore registra una misura F nettamente superiore rispetto all'MLP (+10%), a fronte di uno scarto residuale per la misura di Accuracy trascurabile (0.01).

4.1.1 ROC Curve

Altri due elementi molto utili per la verifica delle performance di un modello sono il *Lift chart* e la *ROC Curve*, entrambi riportati nel workflow Knime. Pertanto, sono state confrontate le curve ROC dei due classificatori migliori inerenti la classe positiva *bad*.

L'*AUC* (Area under the curve) [9], parametro molto importante di una *ROC Curve*, risulta essere leggermente maggiore per il modello Naive Bayes. Inoltre, se fissiamo sull'asse x (percentuale dei FP) un intervallo compreso tra 0.3 e 0.6, il modello Naive Bayes è capace di prevedere con più efficacia il True Positive Rate nel dataset. Ad esempio, se una banca, utilizzando il classificatore Naive Bayes, dovesse accettare di classificare incorrettamente un buon debitore come cattivo nel 50% dei casi, allora il modello sarebbe in grado di prevedere correttamente i cattivi debitori nel 90% dei casi.

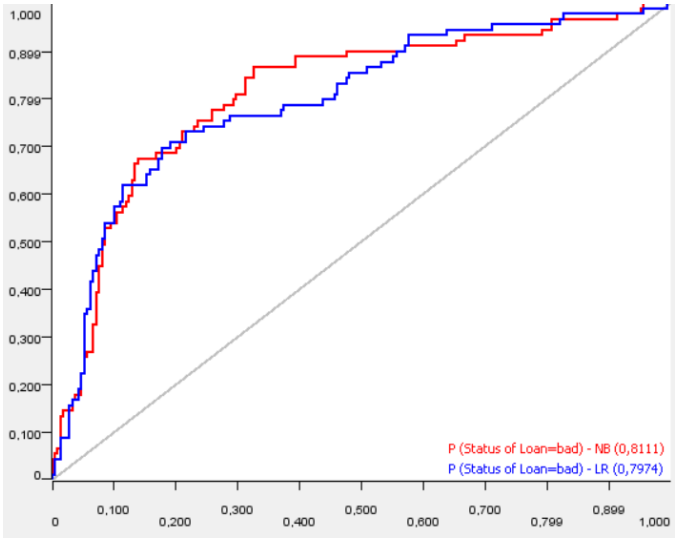

Fig. 6 ROC Curve – Naive Bayes (linea rossa) e Logistic (linea blu)

MODELLI	AUC
LOGISTIC	0,80
NAIVE BAYES	0,81

Tab. 2 Valori di AUC dei 2 modelli più performanti per l'approccio Hold out

4.2 Cross Validation

Il metodo Hold Out potrebbe essere soggetto al fenomeno di overfitting, per questo motivo si è deciso di utilizzare anche un altro metodo per valutare le performance dei classificatori: la *Cross validation*. Nel nostro caso abbiamo posto k (il numero di fogli) uguale a 5, e abbiamo stratificato il dataset rispetto alla variabile target *Status of Loan*. La *Cross validation* è stata applicata sia selezionando l'intero set di attributi per ogni classificatore, sia effettuando una procedura di *Features Selection* per ogni modello (verrà approfondito nel Capitolo 5 in che modo è stata effettuata la selezione delle variabili per ogni classificatore). I modelli che sono stati addestrati con procedura di Feature Selection registrano risultati non significativamente diversi rispetto ai modelli in cui la selezione delle variabili non è stata effettuata. Infatti, solo il J48 registra una perdita di accuratezza leggermente rilevante quando viene sottoposto ad una selezione delle variabili (-0.08), mentre gli altri classificatori non peggiorano la loro stima (nel caso dell'SVM la stima è anche leggermente migliorata). Per questo motivo, verranno elencati di seguito soltanto i risultati della procedura Cross Validation con Features Selection.

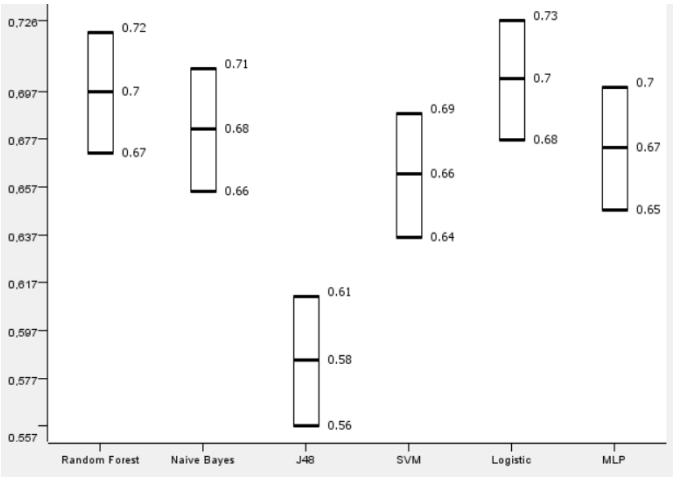

Fig. 7 Box plot dell'Accuracy per i 6 modelli – con Feature selection

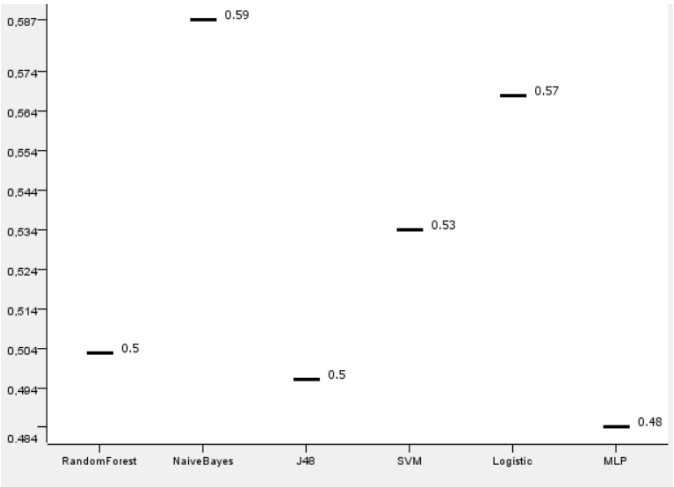

Fig. 8 F-measure dei 6 modelli - con Feature selection

Nella figura 7 gli estremi superiori ed inferiori dei box plot rappresentano i valori dell'intervallo di confidenza di Wilson a livello di significatività del 95% e il valore centrale rappresenta la stima puntuale dell'accuratezza. I modelli più efficienti risultano essere il Naive Bayes, il Logistic Regression e il Random Forest. Quest'ultimo classificatore, tuttavia, registra un valore di F-measure nettamente inferiore rispetto agli altri due, come possiamo osservare nella figura 8.

4.2.1 CI Precision e Recall

Nei problemi di classe sbilanciata, per affermare con sicurezza che un modello di classificazione sia migliore rispetto ad un altro, non possiamo fare affidamento solamente al valore di *accuracy* ma abbiamo il bisogno di prendere visione anche dei valori di *precision* e *recall*. Si è voluta sottolineare l'importanza di queste due misure non limitandosi a riportarne una stima puntuale, ma calcolando un intervallo di confidenza per averne una visione più precisa a livello statistico. Per calcolare gli intervalli di confidenza di *precision* e *recall* si è ricorsi all'approssimazione normale delle due misure. Tali misure si distribuiscono come delle Beta, come ben spiega lo studio "A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation" di Cyril Goutte and Eric Gaussier [8] sul quale ci siamo basati. Vengono assunte dunque le seguenti distribuzioni di probabilità a posteriori per precision e recall (per le spiegazioni più dettagliate di come si arriva a tale risultato, rimandiamo al paper):

- $p|D \sim Beta(TP+\lambda, FP+\lambda)$
- $r|D \sim Beta(TP+\lambda, FN+\lambda)$

Dalla Jeffrey's non-informative prior, abbiamo assunto per il parametro lambda il valore ½, in quanto esso garantisce la proprietà di invarianza per la ri-parametrizzazione. A questo punto si è ricorsi all'approssimazione normale di tali distribuzioni applicando il teorema del limite centrale. Gli intervalli di confidenza associati alle misure di precision e di recall risultano i seguenti:

$$\left[r \pm z_{\alpha/2} \sqrt{\frac{TP + \frac{1}{2}}{(TP + FN + 1)^2 (TP + FN + 2)}} \right]$$

$$\left[p \pm z_{\alpha/2} \sqrt{\frac{TP + \frac{1}{2}}{(TP + FP + 1)^2 (TP + FP + 2)}} \right]$$

Il livello di confidenza è stato fissato al 95% e qui di seguito vengono riportati risultati.

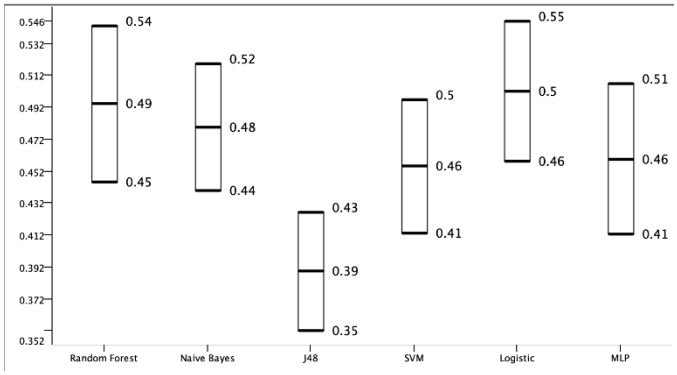

Fig. 9 Box plot della precision per i 6 modelli - con Feature selection

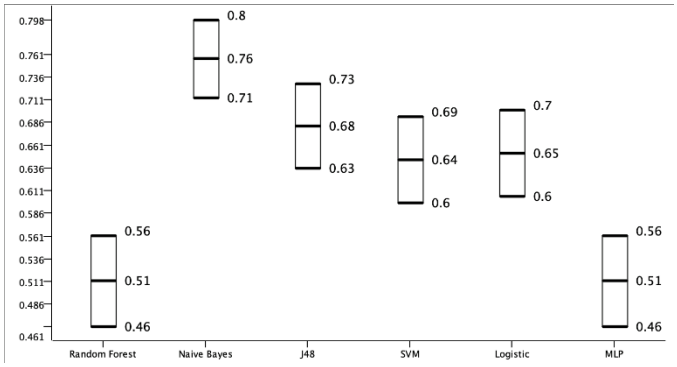

Fig. 10 Box plot della recall per i 6 modelli - con Feature selection

Oltre al calcolo dei CI (Confidence Intervals) derivanti da un'approssimazione normale sono stati calcolati su R (con il metodo *qbeta*) i quantili esatti a livello 0.05 e 0.95 delle due distribuzioni Beta a posteriori. I risultati che seguono (che sono pressoché uguali agli estremi del CI calcolato con il precedente metodo) sono stati ottenuti sulla base dei valori TP, FP, FN che sono stati imputati manualmente a seguito della medesima procedura di Cross Validation applicata in precedenza.

MODELS	LOWER Q. RECALL	UPPER Q. RECALL
RF	0.46	0.56
NB	0.71	0.80
J48	0.63	0.72
SVM	0.60	0.70
LOGISTIC	0.60	0.70
MLP	0.46	0.56

Tab. 3 Intervalli di confidenza per recall (metodo *qbeta R*)

MODELS	LOWER Q. PRECISION	UPPER Q. PRECISION
RF	0.44	0.54
NB	0.44	0.52
J48	0.35	0.42
SVM	0.41	0.50
LOGISTIC	0.45	0.54
MLP	0.41	0.50

Tab. 4 Intervalli di confidenza per precision (metodo *qbeta R*)

Alla luce di tali risultati, possiamo affermare che i due modelli migliori, dal punto di vista della recall, risultano essere il Logistic Regression, il Naive Bayes e il Random Forest. Se una banca è interessata principalmente a non commettere errori nel classificare incorrettamente come *buoni*, coloro che in realtà sono *cattivi* debitori, allora quelli appena elencati risultano essere i classificatori più efficaci. Tuttavia, il Random Forest registra un valore di precision di gran lunga minore rispetto sia al Naive Bayes, sia al Logistic Regression, che nel complesso, risultano dunque i modelli migliori.

4.2.2 ROC Curve

Per calcolare la ROC Curve e le relative AUC in Cross Validation (con Features Selection), abbiamo utilizzato le seguenti librerie in R: "dplyr" per la data manipulation, "caret" per il model-building, "DMwR" per l'implementazione dello Smote, "purrr" per lavorare con funzioni e vettori, e infine "pROC" per il calcolo della AUC.

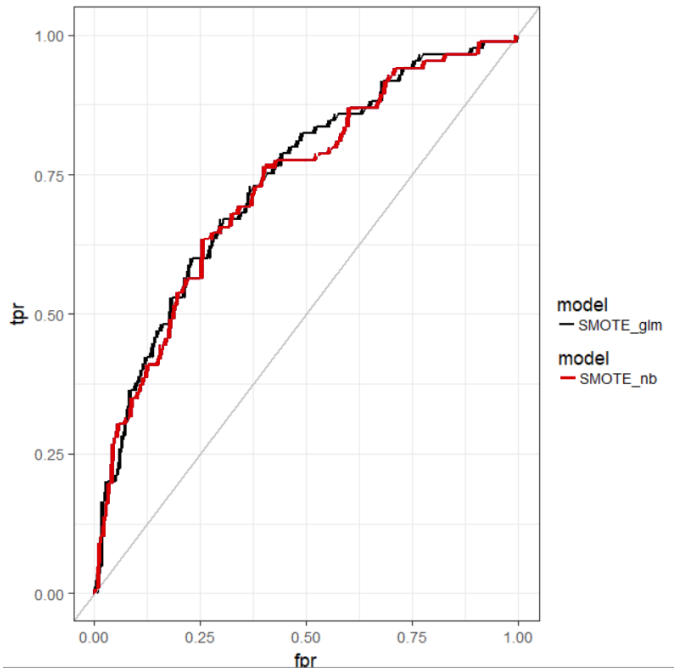

Fig. 11 ROC Curve per il Naive Bayes (SMOTE_nb) e per il Linear Regression Model (SMOTE_glm)

Le due curve ROC nella figura 11 si intersecano in più punti all'interno del grafico. Inoltre, le AUC risultano essere pari a 0,74 per il Logistic e 0,73 per il Naive Bayes. Le due curve non si discostano significativamente, ad eccezione dell'intervallo compreso tra 0.4 e 0.6 sull'asse x (percentuale dei FP). Per tutti i valori compresi in tale intervallo, infatti, il modello Logistic Regression risulta più efficace nell'individuare i True Positive.

4.2.3 Comparing Classifiers

I modelli che performano meglio, dai risultati della procedura di Cross Validation e dalla selezione delle variabili, risultano essere il Naive Bayes e il Logistic Regression. Ci siamo chiesti se la differenza tra gli errori di classificazione tra tali modelli (definita come *Logistic_error – NaiveBayes_error*) sia statisticamente significativa. È stato condotto un test d'ipotesi rappresentato dal seguente grafico:

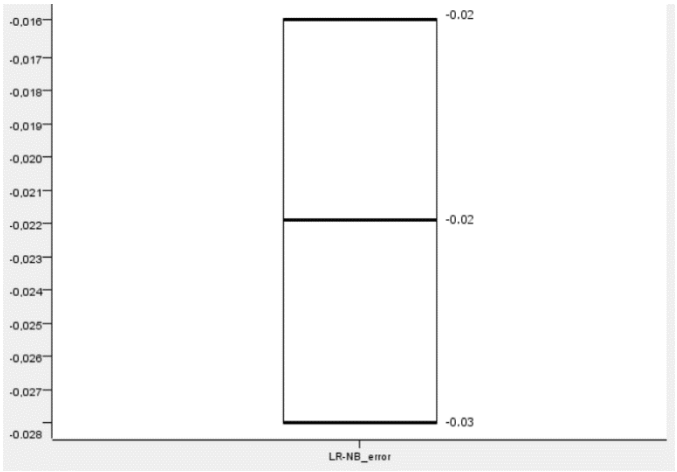

Fig. 12 Box plot della differenza degli errori tra *Logistic* e *Naive Bayes*

Come si evince dalla figura 12, l'estremo superiore del box plot è negativo, pertanto possiamo affermare che il *Logistic Regression* è un modello migliore, ad un livello di significatività del 95%.

4.3 Validation

Una porzione del dataset iniziale pari al 10% è stata riservata al *Validation set*. Questo dataset verrà utilizzato per selezionare quali tecniche di features selection risultano ottimali per ciascun modello. Per alcuni classificatori, l'approccio *Backward* è risultato più performante rispetto all'utilizzo del sottoinsieme di attributi selezionato tramite l'approccio *Forward*, mentre per altri è vero il contrario. Di seguito, vengono riportati i risultati delle misure di valutazione con i metodi *Forward inclusion* (Tab. 5) e *Backward elimination* (Tab. 6).

	RF	NB	J48	SVM	LOGISTIC	MLP
RECALL	0.40	0.60	0.60	0.77	0.60	0.53
PRECISION	0.57	0.47	0.40	0.49	0.53	0.38
F-MEASURE	0.47	0.53	0.48	0.60	0.56	0.44
ACCURACY	0.73	0.68	0.61	0.69	0.72	0.60

Tab. 5 Valori delle misure di valutazione nella fase di *Validation* con *Feature Selection Forward*

	RF	NB	J48	SVM	LOGISTIC	MLP
RECALL	0.50	0.80	0.57	0.67	0.63	0.40
PRECISION	0.50	0.56	0.41	0.48	0.59	0.67
F-MEASURE	0.50	0.66	0.48	0.56	0.61	0.50
ACCURACY	0.70	0.75	0.63	0.68	0.76	0.76

Tab. 6 Valori delle misure di valutazione nella fase di *Validation* con *Feature Selection Backward*

Dalle tabelle 5 e 6 si evince che l'approccio *Bacward elimination* risulta più efficace per i seguenti modelli: *Random Forest*, *Naive Bayes*, *Logistic Regression* e *Multi Layer Perceptron*. Per i modelli *J48* e *SVM*, invece, la *Forward inclusion* risulta la tecnica di selezione delle variabili più efficace. La scelta dell'approccio migliore è stata influenzata anche dal numero di attributi selezionati dai due metodi. Quando le misure di valutazione erano molto simili, è stato premiato l'approccio che comportava una riduzione della dimensionalità maggiore.

5 Conclusioni

Dopo aver effettuato le dovute analisi e verificato i risultati, possiamo affermare come il metodo *Cross Validation* sia il più indicato a riconoscere i modelli che performano meglio. L'*Hold out*, infatti, è soggetto ad introdurre bias dovuti alle specifiche partizioni dei dati di training set e di test set, che potrebbero portare al fenomeno di overfitting. Dai risultati della Cross Validation con Features selection siamo arrivati alla conclusione che i modelli che rispondono meglio all'obbiettivo di predire al meglio i cattivi debitori, sono il *Naive Bayes* e *Logistic Regression*. Effettuando un ultimo confronto tra i due modelli è stato verificato che il *Logistic Regression*, addestrato con 17 variabili, è il modello che performa in assoluto meglio. Per tanto, proiettando il lavoro in ottica reale di un'analisi del rischio del credito per una banca, si evince come il modello *Logistic Regression* classificherà con migliore efficacia i buoni e cattivi debitori e in particolare riuscirà ad individuare il 70% dei reali cattivi debitori.

Bibliografia

1. http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a.pdf
2. https://link.springer.com/content/pdf/10.1007%2FBF00993309.pdf
3. https://en.wikipedia.org/wiki/Random_forest
4. http://web.cs.iastate.edu/~honavar/bayes-continuous.pdf
5. https://www.csie.ntu.edu.tw/~cjlin/libsvm/
6. https://www.jstor.org/stable/2347628?seq=2#metadata_info_tab_contents
7. http://deeplearning.net/tutorial/mlp.html
8. https://pdfs.semanticscholar.org/e399/9a46cb8aaf71131a77670da5c5c113aad01d.pdf
9. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5