

Università Ca' Foscari di Venezia



Corso di Laurea in Economia Aziendale

Tesi di Laurea Triennale

Credit Scoring: alcuni aspetti teorici ed una applicazione

Relatore
Ch. Prof. Andrea Pastore

Laureanda
Elena Martarello

A tutte le donne che avrebbero voluto studiare

Nella vita non bisogna mai rassegnarsi, arrendersi alla mediocrità, bensì uscire da quella zona grigia in cui tutto è abitudine e rassegnazione passiva, bisogna coltivare il coraggio di ribellarsi.

Rita-Levi Montalcini

Ringraziamenti

Tantissime sono le persone che dovrei ringraziare a questo punto della mia vita, ma primi tra tutti sono mamma e papà che mi hanno cresciuta facendo di me una donna forte e decisa.

Ringrazio i miei genitori; che hanno sempre fatto il tifo per me ogni minuto della mia vita, che non hanno mai dubitato delle mie capacità, e che non mi hanno mai scoraggiata. Siete la mia ispirazione.

Grazie Linda per aver chiesto una sorellina, anche se negli anni si è dimostrata una peste. Forse più di un "grazie" dovrei dirti "scusa!", per fortunata c'è Nicola che ti aiuta con la sua dote da pacere.

Un grazie speciale va ai miei nonni; Gianni e Renata, che con le loro esperienze mi hanno insegnato più di quanto avrebbero potuto fare a parole.

Grazie zia Gina per essere la memoria di un passato che non potrò mai vivere; in questi anni mi hai insegnato la virtù del lavoro e del sacrificio. Resterei ore ad ascoltare qualsiasi cosa esca dalle tue labbra.

Un pensiero va a nonna Laura che mi tieni d'occhio da sempre proteggendomi da ogni mia scelta. Questo traguardo lo devo anche ad Alberto non solo la persona più importante in questi tre anni ma anche della mia vita.

Grazie per farmi vivere dei momenti magnifici che mi accapponano la pelle.

Ti ringrazio, infine, per avermi fatto conoscere i tuoi amici che col tempo sono diventati anche i miei. Perciò ringrazio gli "stecca friends" di Unipd che involontariamente mi avete spinto ad essere "l'update" di me stessa.

Ho avuto, infine, la fortuna di conoscere a San Giobbe delle persone splendide, grazie alle quali la mia esperienza universitaria si è trasformata in un speciale ricordo che porto alla luce quanto voglio sorridere.

Grazie al professore relatore di questa tesi; il Dr. Andrea Pastore che mi ha accompagnato in questa fase finale del mio percorso incornando un traguardo di cui sono orgogliosa e di cui non cambierei nulla. A questo punto sono in lacrime perciò concludo ringraziando tutti coloro che mi voglio bene per ogni momento, per ogni parola d'incoraggiamento, per ogni risata e per ogni caro ricordo.

Grazie, grazie grazie!

Sommario

La tesi propone l'implementazione di alcune tecniche di credit scoring con il linguaggio di programmazione R attraverso il software Rstudio.

Viene inizialmente introdotto il problema del credit scoring, assieme ad un richiamo delle principali nozioni statistico-matematiche che vengono utilizzate.

Successivamente, si passano in rassegna i metodi che sono più frequentemente impiegati dagli istituti di credito.

Viene infine proposta l'applicazione pratica del credit scoring con il modello logit utilizzando dati open source.

Parole chiave: Credit Scoring, Istituti di credito, Merito di credito, Regressione modello logit, Programmazione.

Indice

Elenco delle figure	IX
1 Introduzione alla disciplina del Credit Scoring	1
1.1 Definizione del Credit Scoring e dei suoi obiettivi	1
1.2 Vantaggi e limiti del Credit Scoring	4
1.3 Differenze tra il Credit Rating e Credit Scoring	7
2 Introduzione alle fasi del Credit Scoring	9
2.1 L'approccio al Credit Scoring	9
2.2 Identificazione della popolazione obiettivo	10
2.2.1 Tipologie di dati possibili	12
2.3 Composizione del dataset	13
2.4 Definizione del metodo: tipologie di analisi di regressione	13
2.4.1 Regressione lineare multipla	14
2.4.2 Regressione con il modello logit	16
2.5 Definizione dei parametri di scelta	18
2.6 Altri metodi statistici per il Credit Scoring	18
2.6.1 Le reti neurali	19
2.6.2 Gli alberi di classificazione	20
3 Applicazione al Dataset "Statlog (German Credit Data) Data Set"	21
3.1 L'obiettivo	21
3.2 Preparazione del dataset e creazione del file "lettura_dati.R"	22
3.2.1 Lettura del dataset	22
3.2.2 Assegnare un nome ad ogni osservazione	26
3.2.3 Trasformare la variabile risposta "responsegoodcredit" in termini binari	26
3.2.4 Identificare il vettore "responsegoodcredit" come un fattore	27
3.2.5 Train e Test split	27
3.3 Statistica descrittiva con il dataset: "German Credit Data".	28
3.3.1 Suddividere il dataframe "dedatacredit" in base alla variabile risposta	29
3.3.2 Studio della variabile "sex"	29
3.3.3 Studio della variabile "age"	31
3.3.4 Studio della variabile "duration"	33
3.3.5 Studio della variabile "purpose"	35

3.3.6	Le altre variabili	37
3.3.7	Studio di più variabili contemporaneamente	37
4	Modello logit per il data dataset: "German Credit Data"	40
4.1	Cenni ai modelli lineari generalizzati	40
4.2	Applicazione del modello Logit su tutte le variabili	40
4.2.1	Il problema dell'overfitting e dell'underfitting	41
4.2.2	La funzione <i>glm()</i>	41
4.2.3	Interpretazione dei "Coefficients"	43
4.2.4	Il metodo AIC	44
4.2.5	Altre considerazioni	45
4.3	Ottimizzazione del modello con variabili scelte tramite la Backward Stepwise Regression	48
4.3.1	Altre considerazioni	51
4.4	Applicazione del modello logit su variabili scelte tramite la Forward Stepwise Regression	52
4.5	Conclusioni	54
A	Codice file "lettura_dati.R"	55
A.1	Lettura del dataset	55
A.2	Assegnare un nome ad ogni osservazione	55
A.3	Trasformare la variabile risposta "responsegoodcredit" in termini binari	55
A.4	Identificare il vettore "responsegoodcredit" come un fattore	56
A.5	Train e Test split	56
A.6	Salvare le operazione del file "germandataset.RData"	56
B	Codice file "DescriptiveStat4GermanDataCredit.R "	57
B.1	Cariacamento del file d'archiviazione <i>germandataset.RData</i>	57
B.2	Suddividere il dataframe "dedatacredit" in base alla variabile risposta	57
B.3	Studio della variabile "sex"	57
B.3.1	Tabella e distribuzione di frequenza	58
B.3.2	Tabella di contingenza con la variabile risposta	58
B.4	Studio della variabile "age"	59
B.4.1	Tabella e distribuzione di frequenza	59
B.4.2	Tabella di contingenza con la variabile risposta	61
B.4.3	Tabella di frequenza per per i due dataset divisi per la variabile risposta	63
B.5	Studio della variabile "duration"	65
B.5.1	Tabella e distribuzione di frequenza	65
B.5.2	Tabella di contingenza con la variabile risposta	67
B.5.3	Tabella di frequenza per i due dataset divisi per la variabile risposta	68
B.6	Studio della variabile "purpose"	69
B.6.1	Tabella e distribuzione di frequenza	69
B.6.2	Tabella di contingenza con la variabile risposta	70
B.7	Studio di più variabili contemporaneamente	71
B.7.1	Caricamento del dataset in versione numerica	71

B.7.2	Creazione della matrice di correlazione	71
B.7.3	Creazione del Correlogram	72
B.7.4	Tabella di contingenza per le varaibili " <i>credit_history</i> " e " <i>property</i> " . . .	72
C	Codice file "RegressionModels.R "	74
C.1	<i>glm()</i> con il modello completo	74
C.1.1	Matrice di confusione per il modello completo	77
C.1.2	Alcuni indici di valutazione	77
C.1.3	Curva ROC e AUC.	78
C.2	<i>glm()</i> con le variabili selezione tramite la Backward Stepwise Regression	79
C.2.1	Matrice di confusione per il modello con le variabili selezionate	87
C.2.2	Alcuni indici di valutazione	87
C.2.3	Curva ROC e AUC.	88
C.3	<i>glm()</i> con le variabili selezione tramite la Forward Stepwise Regression	89
C.3.1	Matrice di confusione per il modello con le variabili selezionate	96
C.3.2	Alcuni indici di valutazione	97
C.3.3	Curva ROC e AUC.	98
	Riferimenti bibliografici	99

Elenco delle figure

1.1	Credit Scorecard: un esempio	2
1.2	Classi di appartenenza degli intervalli di score possibili	3
2.1	Iter procedurale per la definizione della griglia di scoring	10
2.2	Esempio di Output del Credit Scoring	11
2.3	Schema sulle varie tipologie di variabili	13
2.4	Mappa mentale delle classi di regressioni	14
2.5	Rappresentazione grafica della regressione lineare semplice	15
2.6	Mappa mentale delle interdipendenza variabile	17
2.7	Rappresentazione grafica della regressione logistica semplice.	18
2.8	Rete neurale: un esempio	19
2.9	Rete neurale: un esempio	20
3.1	Istogramma di frequenza della variabile "sex"	30
3.2	Istogramma di frequenza della variabile "age"	31
3.3	Distribuzione di frequenza della variabile "age"	32
3.4	Istogramma della frequenza della variabile "duration"	33
3.5	Distribuzione di frequenza della variabile "age"	34
3.6	Istogramma di frequenza della variabile "purpose"	36
3.7	Correlogram per tutte le variabili all'interno del dataset "Statlog (German Credit Data) Data Set" in versione numerica	38
4.1	La curva ROC per il modello logit " myglm "	47
4.2	La curva ROC per il modello logit " myglm_backward "	52
B.1	Istogramma di frequenza della variabile "sex"	58
B.2	Istogramma di frequenza della variabile "age"	60
B.3	Elegante visualizzazione della frequenza della variabile "age"	61
B.4	Istogramma di frequenza della variabile "age" in "splitted_bad"	64
B.5	Istogramma di frequenza della variabile "age" in "splitted_bad"	66
B.6	Elegante visualizzazione della frequenza della variabile "duration"	67
B.7	Istogramma di frequenza della variabile "duration" in "splitted_bad"	69
B.8	Isogramma di frequenza della variabile "purpose"	70

B.9	Correlogram per tutte le variabili all'interno del dataset " <i>Statlog (German Credit Data)</i> Data Set" in versione numerica	73
C.1	La curva ROC per il modello logit " <i>myglm</i> "	79
C.2	La curva ROC per il modello logit " <i>myglm_backward</i> "	89
C.3	La curva ROC per il modello logit " <i>myglm_backward</i> "	98

Capitolo 1

Introduzione alla disciplina del Credit Scoring

1.1 Definizione del Credit Scoring e dei suoi obiettivi

Un istituto finanziario è quell'ente che effettua transazioni finanziarie tra cui investimenti, prestiti e depositi. In questo elaborato si discuterà una delle attività principali degli istituti finanziari nonché l'erogazione del credito e la loro necessità di minimizzarne il rischio.

Tra le varie definizioni di rischio di credito si riporta quella di **Borsa Italiana** secondo la quale: «il rischio di credito si ha quando il debitore non è in grado di adempiere ai suoi obblighi di pagamento d'interessi e di rimborso del capitale»¹.

Gli istituti di credito mettono in atto raffinati sistemi di gestione del rischio che permettono di affidare crediti confacenti ai profili di rischio e rendimento voluti dalla strategia di business.

Il processo d'erogazione del credito viene articolato in modo personalizzato da ogni istituto in base alla strategia di corporate messa a punto dai manager.

In linea generale è possibile affermare che la gestione del credito, a prescindere delle fasi esclusive di ogni istituti, è suddivisibile in due macro processi sequenziali.

In prima istanza l'istituto dovrà concedere il credito e successivamente dovrà gestirlo attraverso un costante monitoraggio dal quale potrebbe seguire una revisione del rapporto creditizio.

In questa sede verrà approfondita, tra le varie tecniche istruttorie, l'indagine predittiva con cui le banche valutano il merito creditizio del cliente verificandone la capacità di rimborso.

Comune a tutti gli istituti, per valutare la bontà della promessa del cliente di rimborsare il credito, sono quelle analisi condotte su più dimensioni e nel caso di questa tesi si coinvolgeranno le discipline statistico-matematiche.

In generale le banche analizzano il profilo del cliente dal punto di vista quantitativo e qualitativo, oltre che l'aspetto andamentale del rapporto che il cliente potrebbe aver già intrattenuto con altri istituti di credito.

¹Per l'articolo completo si veda il seguente link: <https://www.borsaitaliana.it/borsa/glossario/rischio-di-credito.html>

La tecnica istruttoria che si approfondirà di seguito prende il nome di **Credit Scoring**, la quale è diventata una disciplina indispensabile per gli istituti di credito; data la sua capacità di ottenere una valutazione sulla rischiosità di una operazione, il tutto con una certa facilità e velocità di reperibilità dei dati.

Si definisce, sinteticamente, il credit scoring come quell'insieme di modelli predittivi e di tecniche che aiutano le istituzioni finanziarie nella concessione di crediti. L'output che sostanzia il credit scoring non è altro che un punteggio espresso in decimali assegnato ai richiedenti di credito che sintetizza la loro affidabilità creditizia nonché sul loro rischio d'insolvenza.

Nella figura 1.1 è riportato un esempio di Credit Scorecard, della quale si discuterà di seguito.



Figura 1.1: Credit Scorecard: un esempio

Una credit scorecard non è altro che una tabella riassuntiva di quelle caratteristiche che consentono di classificare ogni richiedente credito. Ogni osservazione ricadrà in più gruppi che coincidono con un determinato punteggio, la somma dei punteggi cumulati è lo score di credito. Il credit scoring non è altro che un giudizio sui richiedenti di credito che sintetizza la loro affidabilità creditizia nonché sul loro rischio d'insolvenza.

La valutazione sul merito di credito di ogni singolo richiedente è ricavata dall'analisi dei dati personali rilevanti. Alcuni esempi di dati personali rilevanti possono essere la posizione lavorativa occupata e il relativo RAL², oltre che eventuali redditi diversi, in generale si parla di documenti che attestino le informazioni relative al reddito. Nondimeno importanti sono gli eventuali rapporti che si esprimono sulla condotta del richiedente in occasioni di altre posizioni debitorie, come per esempio eventuali sofferenze, incagli o deterioramenti. Non si escludono dall'analisi anche i seguenti dati personali rilevanti quali; età, sesso, stato civile e nazionalità.

Se la valutazione ottenuta è sufficientemente positiva l'istituto procederà all'erogazione del credito indipendentemente dalla forma in cui esso viene concesso.

In riferimento alla Credit Scorecard vale la regola che più alto è lo score, più si è meritevoli di credito. Lo schema in figura 1.2 riassume ciò che è stato sostenuto finora.

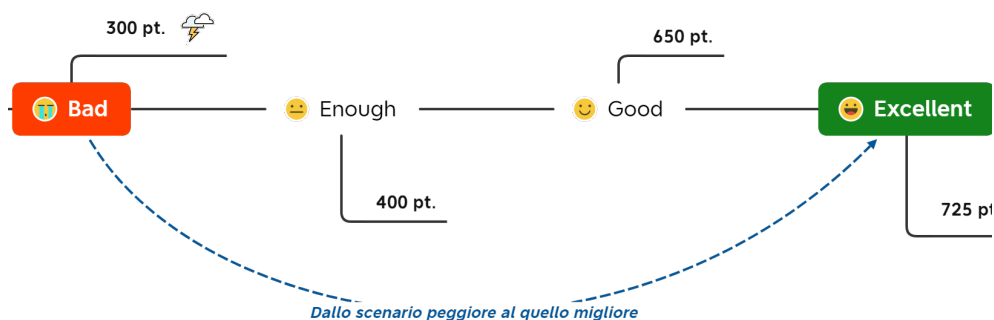


Figura 1.2: Classi di appartenenza degli intervalli di score possibili

Nel corso degli anni le tecniche utilizzate per raggiungere lo scopo poc'anzi descritto sono cambiate profondamente.

Un tempo il giudizio sulla sicurezza di una determinata operazione di credito veniva attribuito solamente con parametri di ordine soggettivo come l'esperienza e il buon senso. Adesso le banche e non solo si avvalgono di strumenti e tecniche sofisticate che hanno reso questo incarico escludibile a chi non è introdotto nelle discipline di matematica, statistica e probabilità.

Per gli istituti di credito è essenziale sviluppare algoritmi di decisione in modo da poter automatizzare il processo di affidamento, tuttavia non è sufficiente poiché è necessario un continuo aggiustamento del modello in modo da migliorarne l'affidabilità.

L'importanza che il credit scoring ha assunto lo ha reso una delle strategie operative dell'istituto di credito.

I teorici più autorevoli della disciplina del management aziendale sono di comune accordo nell'affermare che una buona strategia operativa consenta a qualsiasi organizzazione di ottenere un certo vantaggio competitivo, pertanto un funzionante algoritmo di decisione consente all'intermediario finanziario di erogare crediti di qualità superiore ai competitors.

Il concetto di qualità di credito verrà discusso nel prossimo paragrafo nel quale si discuteranno i

²Reddito Annuo Lordo.

limiti e i benefici che il credit scoring annovera.

Nell'Abstract è già stato anticipato che per questa tesi è stato preparato dall'autrice un esempio pratico, tuttavia non si è discusso come esso si sostanzierà.

Fin'ora è stato detto che le operazioni di credit scoring hanno come risultato finale la credit scorecard, tuttavia il processo necessario per ottenere tale output è molto lungo e articolato.

Convergono nella credit scorecard anche le scelte manageriali assunte dall'istituto, tuttavia il lavoro in oggetto non è caratterizzato da tale aspetto. In particolare, l'obiettivo del progetto non sarà la produzione di una credit scorecard ma bensì un giudizio positivo o negativo sui richiedenti credito.

In altre parole, data una serie di informazioni su una serie di richiedenti credito l'autrice s'impegna a prevedere se essi saranno buoni o cattivi debitori.

1.2 Vantaggi e limiti del Credit Scoring

Da quanto si può intendere finora le tecniche di credit scoring portano agli istituti finanziari una serie di vantaggi che rendono convenienti gli investimenti atti a migliorare tali tecniche.

È ragionevole annoverare tra i vantaggi più consistenti la maggiore velocità con cui è possibile prendere decisioni sulle possibili concessioni di credito. Questo aspetto positivo determina una maggiore produttività nel reparto crediti.

Ciò è possibile attraverso l'implementazione di algoritmi di decisione automatizzati che offrono come output una valutazione di veloce interpretazione per gli operatori finanziari.

La velocità di decisione non è l'unico vantaggio che deriva dall'automatizzazione del processo di affidamento del credito, infatti troviamo la coerenza decisionale e soprattutto la qualità decisionale.

Come già anticipato nel paragrafo introduttivo i modelli di credit scoring si avvalgono di discipline scientifiche che permettono lo sviluppo di un algoritmo standard che consente coerenza tra decisioni diverse prese da operatori diversi.

La coerenza e la velocità, tuttavia, non garantiscono una maggiore qualità sul credito erogato.

Un credito di qualità riduce il rischio di credito, che si ricorda essere l'obiettivo degli istituti finanziari. A questo punto sorge spontanea la domanda su come un istituto finanziario possa migliorare la qualità del credito distribuito perciò si procede definendo cosa si intende per qualità di credito. Il concetto di *credito di qualità* è definito secondo **BancoBPM** come «un'obbligazione di qualità che presenta una elevata sicurezza in termini di buon fine». Sempre secondo BancoBPM tale sicurezza «è misurata attraverso un'analisi su parametri diversi tra cui il rating assegnato al destinatario del credito»³.

In parole povere, un credito viene considerato qualitativamente migliore di un altro se lo stesso debitore a parità di condizioni ha una maggiore fattibilità di rimborso, si parla in questo caso di Probability of Default da qui in poi chiamata semplicemente PD.

³Per l'articolo completo si consulti la seguente pagina web: <https://www.bancobpm.it/magazine/glossario/qualita-del-credito/>

La probabilità d'insolvenza è una stima della probabilità che un debitore non sia in grado di soddisfare gli adempimenti e quindi l'obbligo di rimborso. Vale la regola generale che minore è la PD migliore sarà il credito stanziato dalla banca.

Questo non significa che gli istituti di credito non eroghino crediti a chi ha una bassa probabilità di rimborso. In quest'ultimo caso gli istituti di credito applicheranno alla obbligazione concessa un tasso d'interesse più elevato rispetto a chi ha una minore rischiosità.

Capiamo ora che per ottenere un portfolio di linee di credito qualitativamente migliori alla concorrenza è indispensabile per l'istituto finanziario avere un algoritmo che fornisca una accurata PD.

In altri termini la migliore qualità del credito concesso è garantito da una valutazione della PD più sicura e quindi con un margine di errore il più basso possibile in modo tale da poter aver maggior fiducia.

Le tecniche predittive adottate dalle banche, ma in generale da tutti gli istituti finanziari, sono molteplici e con un diverso grado di difficoltà e raffinatezza. Sebbene queste abbiano vantaggi e limiti diversi ciò che le accomuna tutte è l'obiettivo già ampiamente discusso. In riferimento alla PD si specifica che non sarà oggetto di questa tesi tuttavia è stato necessario introdurla per comprendere che la probabilità d'insolvenza di un debitore non è l'output del credit scoring ma lo la credit scorecard.

È sbagliata dunque l'interpretazione dello score di credito come la probabilità d'insolvenza del richiedente di credito. A questo punto è di facile comprensione il primo grande vincolo del credit scoring, il quale risulta essere incapace di esprimere un giudizio tra due richiedenti di credito su quale dei due sia più rischioso in termini d'insolvenza.

Da qui si comprende che le tecniche di credit scoring non siano prive di limitazioni perciò è opportuno sviscerare le cause che potrebbero rendere questi modelli predittivi inadeguati.

È importante identificare tra gli svantaggi del credit scoring la sua incapacità di dare valore ai dati economici, dove per dati economici si intendono sia i dati macroeconomici sia microeconomici. Si ricorda a fini accademici che la microeconomia è quella disciplina che ha per oggetto lo studio dei modi di agire dei singoli agenti di mercato. Questa branca si contrappone alla macroeconomia, la quale ha l'obiettivo di riconoscere modelli di comportamenti sistematici di agenti economici aggregati.

Queste due scienze non vengono considerate dagli istituti finanziari per il calcolo dello score di merito di credito e questo rende fallace il processo a meno che queste non producano conseguenze dirette sui richiedenti credito.

Per comprendere quanto questo limite possa essere rilevante si legga l'esempio riportato di seguito.

Esempio 1.1

A e B sono due soggetti richiedenti di credito e hanno rispettivamente uno score pari 450 e 625 ⁴

⁴Il punteggio del merito di credito è stato calcolato utilizzando i riferimenti in Figura 1.1.

Ora immaginiamo che l'economia del paese subisca una forte impennata entrando così in recessione⁵.

Ai fini del calcolo dello score quest'ultima informazione è del tutto irrilevante, poiché gli input del credit scoring sono dati pertinenti alla sfera personale ovvero il reddito, patrimonio e molti altri.

In questo esempio quindi la situazione di recessione non modificherà lo score fintantoché la posizione finanziaria del richiedente credito non muti come conseguenza di tale stato economico. Il cliente A ha un reddito da lavoro pari a 60.000€ mentre il cliente B percepisce 47.000€ all'anno di RAL.

In valori assoluti, per il credit scoring, il cliente A risulta essere più affidabile rispetto a B che percepisce 13.000 euro in meno.

Nello scenario più banale ipotizziamo che il cliente A perda il lavoro come ripercussione della recessione, a differenza del cliente B che mantiene la sua posizione lavorativa. Ora il cliente A non è più considerevole come maggiormente meritevole, ma lo è il cliente B per il semplice fatto di avere un reddito da lavoro. Nella situazione appena descritta abbiamo ipotizzato che la situazione economica del paese abbia avuto una influenza diretta sulle condizioni dei richiedenti credito. In seguito verrà presentata una situazione nella quale le conseguenze macroeconomiche possono considerarsi indirette per i due clienti. Supponiamo che l'economia dello stesso paese al termine del ciclo economico⁶ in atto entri in una fase di espansione, grazie alla quale il cliente A torna a percepire 60.000€ all'anno.

Di nuovo in valori assoluti il cliente A è più meritevole del cliente B che nel corso del tempo ha accresciuto per anzianità il suo RAL fino a 55.000€.

Supponiamo che il nuovo lavoro del cliente A sia a stretto contatto con i consumatori⁷ mentre il cliente B è ben inserito nel mercato delle materie prime⁸.

Si ricordi che sintomo della crescita economica è un generale aumento della capacità di spesa da cui ne consegue un generale aumento del tasso d'inflazione⁹. Nello specifico il rincaro dei prezzi è stato in misura percentuale più importante nel mercato finale, dovuto a un generale aumento della domanda di beni finali. Per A, operatore di questo mercato, significa un aumento del margine di profitto.

A ragion di logica, l'aumento dei beni richiesti ad A dai consumatori implica un aumento della domanda delle materie prima a B poiché è necessario per A aggiustare l'offerta.

Anche per l'operatore del mercato delle materie prime l'aumento della domanda e quindi del

⁵In macroeconomia la recessione è uno stato che l'economia di un paese assume quando i livelli di produzione sono bassi rispetto a quanto si avrebbe se tutti i fattori produttivi fossero efficientati in termini d'impiego. Si tenga a mente che la recessione è un scenario in antitesi con la crescita economica.

⁶Nella disciplina macroeconomica per ciclo economico si intende l'alternarsi di momenti caratterizzati da diversi stati di vigore dell'attività economica.

⁷Detto B2C o Business to Consumer

⁸Il mercato delle materie prime fa parte del B2B o Business to Business.

⁹In economia macroeconomica il generale aumento dei livelli medi dei prezzi è causa d'inflazione. Il continuo gonfiamento dei prezzi, ceteris paribus, erode nel lungo termine il potere d'acquisto dei consumatori.

Vale la regola che ogni unità di moneta potrà comprare sempre meno beni. Tale affermazione è testimoniata dalla scuola di Keynes secondo il quale l'inflazione dipende dalla domanda.

prezzo di vendita significa una maggiore marginalità di profitto.

Per contenere l'esplosività dell'inflazione in atto il governo decide di aumentare la pressione fiscale ¹⁰ accrescendo l'imposta sul valore aggiunto ¹¹ solamente nell'ultima fase di scambio.

Questa decisione rende i prezzi di vendita più elevati erodendo il potere d'acquisto dei consumatori, facendone diminuire la domanda cosicché il maggiore margine di profitto che A aveva incrementato grazie all'inflazione cala.

B non vedendo il suo mercato colpito dalle decisioni di politica fiscale continua a mantenere migliorato il suo margine di profitto.

In altri termini le entrate di A seppur accresciute in misura maggiore rispetto a B, subiscono una maggiore tassazione rispetto a B portandolo di fatto ad un potere di acquisto inferiore.

La banca nel valutare la rischiosità dei due soggetti non vaglia l'ipotesi secondo la quale B potrebbe essere più meritevole di A alla luce dei ragionamenti appena fatti.

1.3 Differenze tra il Credit Rating e Credit Scoring

In questo paragrafo si discuterà ciò che identifica il credit scoring rispetto al credit rating, dando una breve definizione di quest'ultimo.

Si distinguono le due tecniche per il semplice fatto che il credit scoring è parte del credit rating. Ciononostante, la differenza tra il credit scoring e il credit rating non si sostanzia nell'obiettivo bensì nei modi in cui essi lo perseguono.

La credit scorecard viene elaborata partendo dal credit scoring, il quale è una valutazione del rischio di credito di un potenziale debitore.

L'obiettivo del credit rating è quello di prevedere la capacità di rimborso di un debito da parte di un richiedente con una previsione implicita della probabilità di insolvenza dello stesso. Pertanto l'output del credit rating non è altro che la PD.

Il credit scoring, come già specificato, non calcola la PD ma profila un possibile debitore in base alla sua affidabilità creditizia.

L'affidabilità creditizia secondo **Findomestic** è «il grado di idoneità di un individuo di fare fronte al debito contratto restituendo il capitale e gli interessi nei tempi stabiliti»¹². Essa viene ponderata tramite un modello statistico che sarà, appunto, l'oggetto dell'applicazione pratica.

Lo score di credito di un soggetto, congiuntamente a quella parte di analisi qualitativa, producono il rating di credito. Per analisi qualitativa si intende quella parte d'esame compiuta da un analista finanziario su quella sfera che difficilmente sarebbe traducibile in termini statistici.

In altre parole, l'insieme delle tecniche istruttorie fornisce il credit rating. Essendo il credit scoring

¹⁰ La pressione fiscale è un indicatore della imposizione media.

¹¹ Il valore aggiunto è quel quantitativo di prezzo che eccede negli scambi tra le diverse fasi del ciclo di produzione come tra grossista e dettagliante.

¹² Nel seguente link è presente l'articolo appena citato: <https://www.findomestic.it/glossario/affidabilita-creditizia.shtml>.

parte del credit rating hanno in analisi gli stessi soggetti e cioè le persone fisiche, le imprese, le società e gli enti pubblici. Si ricordi che ciò che differenzia le persone fisiche dagli altri e che queste operano per finalità diversa dall'attività imprenditoriale o professionale.

In figura ?? viene rappresentato il concetto adottato per definire che il credit scoring è parte delle tecniche istruttorie e che l'output di esse è il credit rating.

Capitolo 2

Introduzione alle fasi del Credit Scoring

2.1 L'approccio al Credit Scoring

Nel capitolo uno è stato definito il credit scoring come l'insieme di metodi e modelli atti a identificare lo score del merito di credito.

A tal fine i dati utilizzati come input sono diversi, si annoverano come esempi il reddito, lo stato civile, il RAL e molto altro.

Nel corso della lettura del capitolo in oggetto verranno approfondite le fasi per la costruzione di una griglia di scoring, che di fatto saranno le stesse che l'autrice impiegherà per la stesura del suo progetto. Il tutto sotto l'ipotesi che la popolazione oggetto di studio rimanga omogenea nel tempo.

In figura 2.1, riportata nella pagina seguente, si dà al lettore un piccolo sunto che ha lo scopo di essere una mappa che permette di orientarsi nella prassi del credit scoring.

In questo elaborato, tra i vari metodi di credit scoring, verrà approfondito l'approccio al credit scoring mediante il modello logit, del quale verrà data una breve spiegazione già in questo capitolo.

Inoltre, è cura dell'autrice fornire, seppur con un grado di dettaglio più generico, altri metodi statistici utilizzati per il credit scoring. Infine verrà riportato un caso studio pensato appositamente per questo elaborato tuttavia si manifesta la necessità per il lettore di essere iniziato agli approcci utilizzati.

In questo capitolo e nel successivo si vedrà soddisfatta tale necessità attraverso quello che vorrebbe essere un breve prontuario.

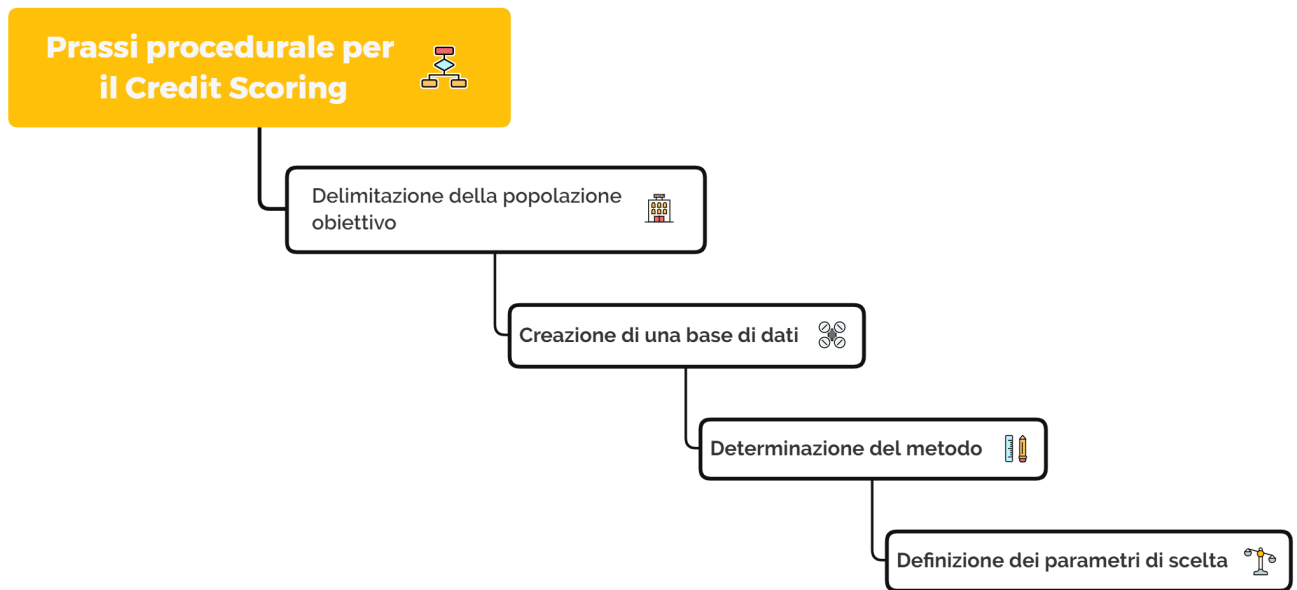


Figura 2.1: Iter procedurale per la definizione della griglia di scoring

2.2 Identificazione della popolazione obiettivo

Con il termine popolazione obiettivo si fa riferimento a un insieme di unità statistiche ¹, raggruppate avendo come denominatore comune almeno una qualità condivisa.

Nel caso della disciplina del calcolo del merito di credito le unità statistiche sono le persone fisiche, le persone dotate di personalità giuridica e ovviamente le imprese. Gli istituti finanziari potrebbero raggruppare le persone fisiche in base a quei dati che nel primo capitolo abbiamo definito rilevanti.

Si prenda in considerazione l'*Age Band 1* nel quale tutte le unità statistiche hanno in comune il fatto di essere tutti più giovani di venticinque anni.


Lo stesso vale per le società che potrebbero essere divise per ragione sociale o più banalmente tra società di capitali ² e di persone ³.

In figura 1.1, riportata nella pagina seguente, è possibile leggere tra la griglia dei dati di riferimento diversi livelli di reddito oltre che numerose fasce di età ed eventuali diritti di proprietà su immobili.

¹ L'unità statistica è definita come l'elemento inscindibile sul quale viene eseguita una rilevazione statistica.

² Tra le società di capitali si collocano le S.P.A. (Società per azioni), le S.R.L. (Società a responsabilità limitata) e le Sapa (Società in accomandita per azioni).

³ Tra le società di persone si annoverano le società semplici, le S.N.C. (Società in nome collettivo) e le S.A.S. (Società in accomandita semplice).

Credit score reference table 		
	Attributes	Score
Age Clusters	Age Band 1	<25
	Age Band 2	$25 \leq \text{Age} < 30$

Income Clusters	Income Band 1	< 30.000
	Income Band 2	$30.000 \leq \text{Age} < 50.000$

Asset Clusters	Home	Rent

Figura 2.2: Esempio di Output del Credit Scoring

I clusters non sono altro che unità statistiche raggruppate tra loro in popolazioni aventi come fattore comune l'età, il reddito e più in generale qualsiasi informazione rilevante allo scopo; come eventuali diritti di proprietà.

La difficoltà in questo primo passaggio sta nel capire quale qualità rende la ricerca rilevante.

Assodato che l'età può essere un buon indicatore della sicurezza finanziaria ci si deve chiedere quali intervalli di età sono meglio rappresentativi. Questo significa determinare i limiti superiori e inferiori del range e congiuntamente l'ampiezza.

Un ragionamento banale, circa l'età, è che il reddito tra ventenni e trentenni non è uguale perciò sorge spontanea questa prima divisione.

Tuttavia non si ha identificato quanti intervalli potrebbero intercorrere tra venti anni e trentanove anni.

Nell'esempio in Figura 1.1 si è provveduto a distinguere tre classi di età: meno di venticinque anni, tra i venticinque e i trenta anni e infine tra i trenta e i trentanove anni.

Non si discute l'efficacia della decisione presa dall'autrice di aver proceduto con la suddivisione appena enunciata, poiché si ricorda al lettore che l'esempio in figura 1.1 è di fantasia.

Della figura 2.2 abbiamo investigato le prime due colonne, nonché le categorie e gli attributi, la terza colonna relativa allo score è figlia di quell'analisi statistica che si discuterà di seguito.

Da questo ragionamento si comprende che non basta individuare la variabile sotto cui raggruppare le unità statistiche poiché occorre definire gli scaglioni della variabile stessa, indipendentemente che essa sia una variabile numerica o qualitativa.

2.2.1 Tipologie di dati possibili

Nella statistica descrittiva vengono definite variabili qualitative quelle che assumono un valore capace di definire una caratteristica che non può essere tradotta in un numero.

Tra le variabili qualitative si distinguono quelle cardinali e quelle nominali, dove le prime sono variabili qualitative che per loro origine possono essere ordinate; come per esempio i giudizi espressi in ordine di "pessimo", "medio" e "ottimo".

Tutte le altre variabili che non possono essere ordinate, come il sesso o la tipologia di lavoro, rientrano delle variabili qualitative nominali.

Le variabili quantitative permettono, invece, di esprimersi in termini numerici.

Si pensi per esempio all'età di un umano, esso può essere giovane, adulto o anziano ma allo stesso tempo si può definire l'età attraverso un valore numerico come dieci, trenta od ottant'anni.

Le variabili quantitative si distinguono a loro volta in discrete e continue. Le prime sono quelle che possono assumere in un valore definito in un intervallo numerico ordinabile e finito, le altre, invece, assumono valori che fanno parte dell'insieme dei numeri reali.

Il campione della popolazione obiettivo produce informazioni che possono assumere valore qualitativo o quantitativo, perciò è indispensabile utilizzare modelli di credit scoring capaci d'interpretare entrambe le tipologie di variabili.

Si vedrà nella teoria e nella pratica il modello di regressione logit, poiché esso è uno dei modelli più impiegati nell'analisi dei dati di risposta categoriali questo perché è in grado di stimare la probabilità che una risposta binaria si verifichi sulla base di un certo numero di variabili predittive.

In altre parole si giustifica l'uso della regressione logit con la sua capacità d'impiegare informazioni personali che assumono valore qualitativo piuttosto che quantitativo.

Data la potenza di questo approccio l'autrice ha deciso di utilizzarlo nell'applicazione pratica, perciò nei prossimi paragrafi verrà introdotta la regressione logit.

Tuttavia è prima necessario un breve confronto con la regressione lineare.

In figura 2.3 si trova uno schema che traduce in chiaro quanto appena spiegato sulle classi di variabili.

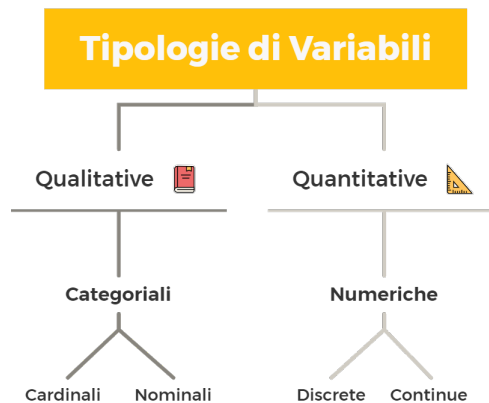


Figura 2.3: Schema sulle varie tipologie di variabili

2.3 Composizione del dataset

Gli studiosi della disciplina statistica sanno che operare su una intera popolazione è pressoché impossibile, motivo per cui si procede a identificare un campione casuale della popolazione stessa. Affinché un campione venga definito casuale deve possedere alcune qualità:

- Tutte le unità statistiche della popolazione abbiano la stessa probabilità di fare parte del campione.
- Tutti campioni abbiano la stessa probabilità di essere formati dalla popolazione.

Per la disciplina del credit scoring è valido come campione casuale l'insieme dei richiedenti credito in un determinato momento.

Per ognuno di loro vengono raccolte le informazioni necessarie e rappresentative della situazione finanziaria.

Nel caso delle imprese, per esempio, i bilanci di più anni sono un efficace rilevatore purché le informazioni in esso contenute siano veritiere.

Sorge, dunque, il problema della credibilità delle informazioni fornite dai richiedenti di credito, il quale richiederebbe una discussione più approfondita in un'altra circostanza. In riferimento al progetto pratico si anticipa che data la più operosa prassi per i soggetti giuridici si applicherà il credit scoring su un campione formato per lo più da personalità fisiche.

2.4 Definizione del metodo: tipologie di analisi di regressione

Alla fine di questo paragrafo sarà acquisita una conoscenza approssimativa sulle differenze tra due appartenenti alla famiglia di regressioni: quella logit e quella lineare.

In figura 2.4 si divide l'analisi di regressione in due classi, nonché quella semplice e multipla.

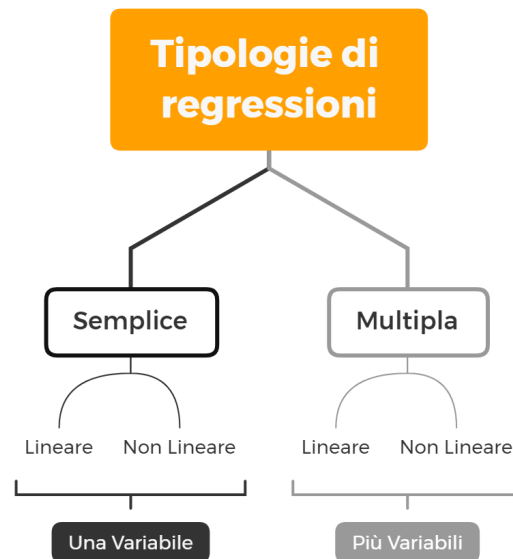


Figura 2.4: Mappa mentale delle classi di regressioni

Questo elaborato si concentra sulla regressione logit multivariata, nella quale verranno impiegate più di una sola variabile dipendente.

2.4.1 Regressione lineare multipla

Dando una definizione della regressione lineare si può asserire che è una pratica statistica atta a scoprire eventuali relazione tra una variabile "dipendente" ed una o più "covariate".

Nel caso in cui la variabile dipendente dovesse essere solo una si parla di regressione lineare semplice.

Il modello di regressione lineare semplice è descritto nella seguente equazione:

$$Y = a + bx + \epsilon \quad (2.1)$$

Questo significa che Y dipende in una certa misura da x e da ϵ :

- Dove x non è altro che la variabile interdependente X che ha assunto un determinato valore.
- Dove ϵ è quel valore avente capacità di esprimere la conseguenza di quegli elementi non osservati nel modello essendo esso di famiglia semplice.

In altri termini il valore che X prende (x), è in grado di condizionare il valore che Y assumerà. Lo stesso vale per la componente additiva ϵ , la quale si ipotizza avere valore atteso pari a zero, espresso in questo modo $E(\epsilon) = 0$, diventando di fatto influente. Di nuovo: Y acquisirà valore y dato il valore che X assumerà in x .

Si enuncia la formula della distribuzione di Y dato $X = x$, che si presenta graficamente in figura 2.5.

$$E(Y|X = x) = a + bx. \quad (2.2)$$

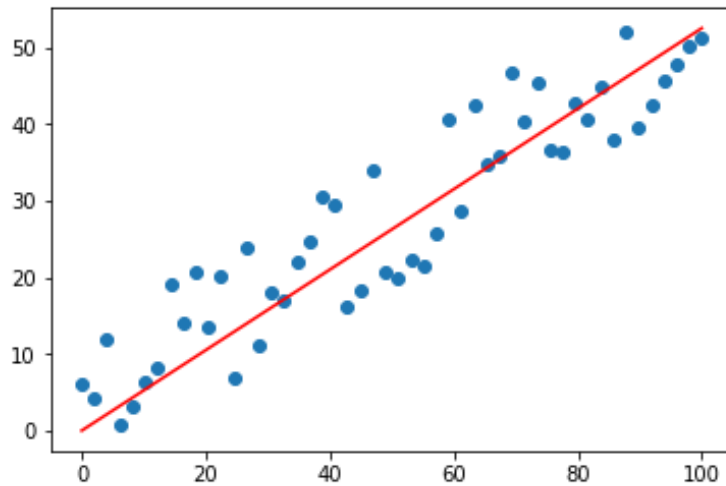


Figura 2.5: Rappresentazione grafica della regressione lineare semplice

Il modello di regressione lineare multipla, a differenza di quella semplice, è in grado di prendere in considerazione più di una variabile. In termini algebrici si sostanzia nel seguente modo:

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (2.3)$$

- Dove \tilde{y} è il valore che assume la variabile dipendente, nonché il risultato del modello.
- Dove β_0 è il valore che assumerebbe se tutti i valori di x_1 fossero nulli.
- Dove β_1 e β_n sono i coefficienti di x_1 e di x_n .
- Dove x_0, x_1, \dots, x_n solo l'insieme di tutte le variabili indipendenti che influenzano \tilde{y} .

Dall'equazione numero 2.3 si riconosce una retta che non si differenzia molto dall'equazione numero 2.2 che anch'essa risulta essere una retta.

2.4.2 Regressione con il modello logit

Nelle prossime righe verrà data una definizione non raffinata della regressione logit che ha lo scopo di avvicinare il lettore alla metodologia.

La regressione logit viene utilizzata per indicare la fattibilità di un evento e quindi la sua probabilità di accadimento.

È fondamentale specificare che l'evento oggetto di studio viene appreso dal metodo in formato binario, dove la risposta può assumere valore 1 o 0. Questo significa che la regressione logit può essere espressa in termini probabilistici.

Si può esprimere la funzione logit nei seguenti termini:

$$E(Y|X = x) = \frac{e^{(\alpha+\beta x)}}{1 + e^{(\alpha+\beta x)}} = \pi(x) \quad (2.4)$$

dove $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$.

Si noti, dunque, che $(\pi(x))$ indica la fattibilità di Y , nonché la differenza con il suo complementare.

Il rapporto tra la probabilità che un evento accada $(\pi(x))$ è il suo complementare $(1 - \pi(x))$ è definito **Odds**:

$$\frac{\pi(x)}{(1 - \pi(x))} \quad (2.5)$$

è detto Logit il logaritmo naturale dell'Odds, ovvero:

$$\text{Logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \quad (2.6)$$

Dove:

- α non è altro che il logaritmo dell'odds di Y dove $x = 0$
- β è la differenza tra il valore che assume Logit in $x = 1$ e in $x = 0$.

$$\text{Logit}[\pi(1)] - \text{Logit}[\pi(0)] = \beta \quad (2.7)$$

Questo significa che se $\beta > 0$ (< 0) la probabilità che $Y = 1$ dato X aumenta (diminuisca) transitando da $X = 0$ a $X = 1$.

Il fattore confondente

Nel caso in cui vi siano variabili diverse da X e Y in grado d'influenzare il modello; è necessario che esse vengano considerate al fine di evitare interpretazioni scorrette.

Questa terza variabile viene identificata come confondente, ed è quel fattore capace di condizionare sia X che Y grazie alla sua interdipendenza sequenziale con X e Y .

Si veda la figura 2.9 riportata di seguito.

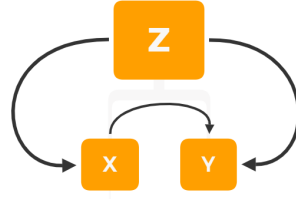


Figura 2.6: Mappa mentale delle interdipendenza variabile

La relazione che intercorre tra il fattore confondente e la variabile dipendente, oltre che interdipendente è chiamata spuria ed è presente quando due variabili non sono casualmente correlate e quindi presentato una associazione.

Questo significa che la variabile confondente è in grado di modificare l'esito di Y .

Il modello logit si adegua alla variabile confondente, nel seguente modo:

$$\log \frac{E(Y|X, Z)}{E(Y|X, Z)} = \log \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 X + \beta_2 Z. \quad (2.8)$$

Per ogni valore che le variabili indipendenti assumono si ottiene:

$$\begin{aligned} & \text{Se } Z = 0 \\ & \text{con } X = 0 \text{ allora: } \log\left(\frac{\pi(0,0)}{1-\pi(0,0)}\right) = \beta_0 \\ & \text{con } X = 1 \text{ allora: } \log\left(\frac{\pi(0,1)}{1-\pi(0,1)}\right) = \beta_0 + \beta_1 \\ & \text{Se } Z = 1 \\ & \text{con } X = 0 \text{ allora: } \log\left(\frac{\pi(1,0)}{1-\pi(1,0)}\right) = \beta_0 + \beta_2 \\ & \text{con } X = 1 \text{ allora: } \log\left(\frac{\pi(1,1)}{1-\pi(1,1)}\right) = \beta_0 + \beta_1 + \beta_2 \end{aligned}$$

Nella famiglia dell'analisi di regressione le differenze non sono solo procedurali ma anche grafiche. Quest'ultima considerazione appare banale visto il diverso output tra la regressione logit e lineare.

Ciò nonostante verrà dato un veloce chiarimento tra ciò che differenzia le due regressioni in termini di rappresentazione grafica.

Come suggerisce il nome la regressione lineare figura sul piano cartesiano come una retta. Per quanto riguarda la regressione logit la forma che assume sui quadranti del piano è paragonabile a una S più o meno accentuata, come si nota in figura 2.7.

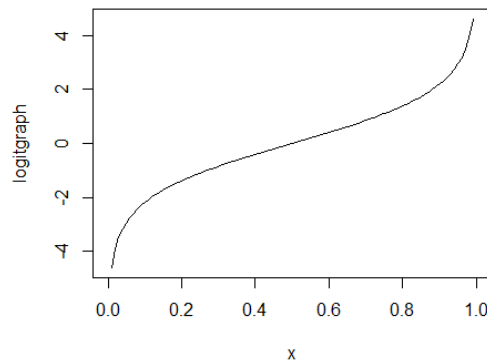


Figura 2.7: Rappresentazione grafica della regressione logistica semplice.

2.5 Definizione dei parametri di scelta

Se lo scopo del credit scoring è quello di determinare se un richiedente credito è meritevole si deve decidere quello che in disciplina viene chiamato valore soglia o cut-off.

Esso è in grado di minimizzare il valore atteso di C , nonché $E(C)$, dove C è la variabile costo, la quale è da considerarsi come conseguenza di eventuali errori di classificazione⁴. Gli errori possibili nella verifica d'ipotesi sono:

- **Tipo 1:** detto anche falso positivo; ovvero condannare un soggetto come cattivo debitore quando nella realtà non lo è.
- **Tipo 2:** detto anche falso negativo; ovvero rifiutare un soggetto meritevole di credito poiché l'analisi lo ha classificato come cattivo creditore.

In altri termini, il costo o l'errore di erogare un finanziamento ad un soggetto insolubile è più grave di escludere un soggetto solvibile da un finanziamento. Il valore soglia è frutto di scelte manageriali dettate da necessità aziendali come le strategie di core business.

Risulta, dunque, importante determinare la probabilità di compiere errori di classificazione al fine di ridurli il più possibile.

Tuttavia, per valutare la precisione del modello in generale sono necessari strumenti aggiuntivi.

2.6 Altri metodi statistici per il Credit Scoring

Sono molti i metodi statistici, oltre al modello di regressione logit, che si prestano alla disciplina del credit scoring. In particolare si accenneranno, solamente dal punto di vista teorico, le seguenti tecniche alternative: le *reti neurali* e gli *alberi di classificazione*.

⁴Per classificazione si intende il procedimento con cui si classifica un soggetto come un buon o un cattivo debitore.

2.6.1 Le reti neurali

Le reti neurali sono dei modelli non lineari che si sforzano di riconoscere le relazioni sottostanti in un insieme di dati attraverso un processo che imita l'operatività del cervello umano.

Questi algoritmi si basano su quelli che in disciplina vengono chiamati *neuroni*, i quali non sono altro che **variabili latenti**.

Una variabile latente è l'opposto di una variabile osservabile, infatti essa potrebbe non essere specificamente dichiarata o manifestata e quindi non è definita nell'ambito di un programma. In generale gli algoritmi delle reti neurali possono operare con una o più variabili latenti. Si usa il termine neurone in modo completamente intercambiabile alla cosiddetta variabile latente.

Un neurone, quindi, è una funzione che raccoglie e classifica le informazioni del modello sotto una predeterminata architettura.

I collegamenti tra neuroni, proprio come le sinapsi, sono in grado di trasferire dei *segnali*. Nello specifico i segnali che un neurone riceve da altri è in grado di elaborarli e segnalarli a sua volta ad altri neuroni. Tuttavia le connessioni non vengono chiamate sinapsi bensì *bordi*.

I bordi e i neuroni hanno una particolarità definita *peso*, il quale può essere usato come valore soglia per la trasmissione o meno di informazioni. Propria dei neuroni è la caratteristica di essere aggregati in *strati*, tra i quali i segnali viaggiano. I segnali viaggiano dal primo strato, nonché quello di input, fino all'ultimo strato, detto strato di output.

Per una visione di come appare una rete neurale si riporta il grafo della rete presa dell'articolo "*Neural network credit scoring models*"⁵ di D. West.

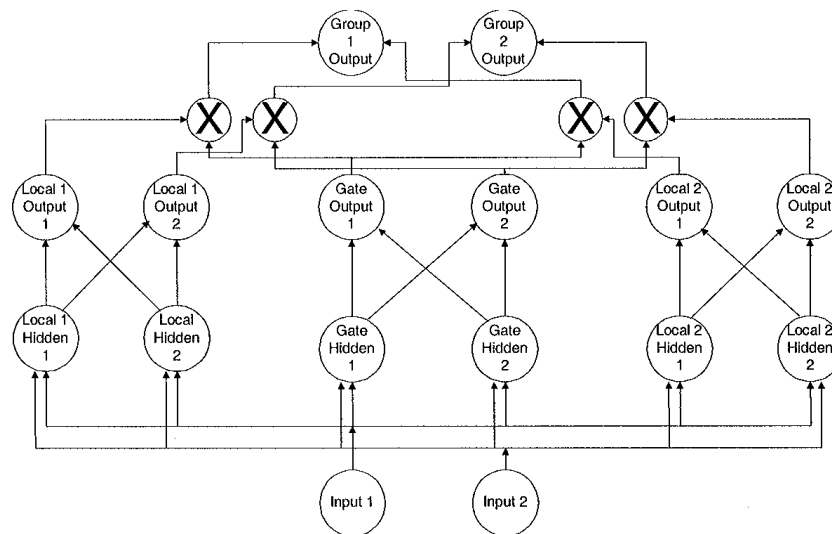


Figura 2.8: Rete neurale: un esempio

⁵L'articolo in oggetto è stato pubblicato da *Semantic Scholar* al seguente link:
<https://www.semanticscholar.org/paper/Neural-network-credit-scoring-models-West/7fc5bf2e6d63aeb6d10159d411752b5b67e573a5>.

Ciò che rende le reti neurali un potente mezzo è che essere, come il cervello dell'uomo, sia in grado di reagire ad una nuova informazione, adattandosi al cambiamento degli input senza dover ridisegnare i criteri degli output. In questo modo la rete neurale genera sempre il miglior risultato possibile.

2.6.2 Gli alberi di classificazione

Anche un albero di decisione si presenta come un grafo a supporto del processo decisionale. Di fatto non è altro che un percorso all'interno del quale vengono già esplicitate le possibili conseguenze, come per esempio il costo di misclassificazione.

Per comprendere come gli alberi decisionali funzionano basti pensare a loro come una gerarchia al fine di ripartire i dati di input. Gli alberi di classificazione sono impiegati per predire una risposta qualitativa e lo fanno corrispondendo la previsione alla classe più comune nelle osservazioni di training. In altri termini, una previsione per una osservazione è ottenuta usando la media o la moda computata nel training in riferimento alla regione in cui l'osservazione in oggetto ricade.

Un primo obiettivo degli alberi di classificazione è quello di ricercare una suddivisione tra nodi che minimizza la loro variabilità, al fine di avere una previsione migliore. Questo primo scopo rientra nella fase di crescita dell'albero seguita dal momento della potatura. Quest'ultimo passaggio si può porre diversi obiettivi, un esempio può essere quello di migliorare l'accuratezza della previsione, perciò si fa leva sul tasso di errata classificazione, cercando di minimizzarlo il più possibile.

Anche in questo caso viene riportato un esempio grafico di un albero di classificazione tratto dall'articolo "*Decision Tree vs. Random Forest – Which Algorithm Should you Use?*"⁶ di Abhishek Sharma.

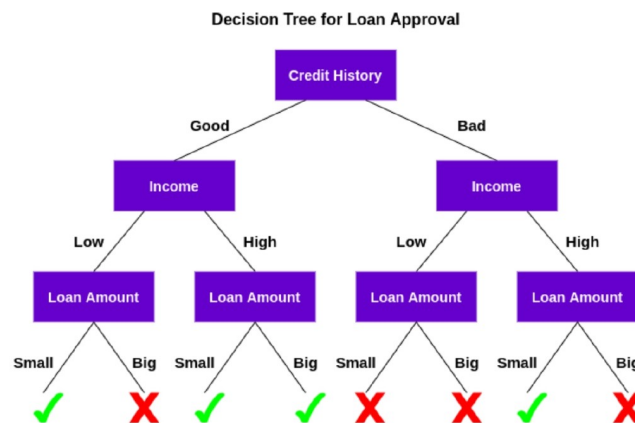


Figura 2.9: Rete neurale: un esempio

⁶L'articolo in oggetto è stato pubblicato da *Analytics Vidhya* al seguente link: <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/>

Capitolo 3

Applicazione al Dataset "Statlog (German Credit Data) Data Set"

3.1 L'obiettivo

Dopo aver introdotto il lettore nella disciplina del credit scoring è possibile procedere con la presentazione del progetto sviluppato appositamente per questo preparato.

Si anticipa che sotto consiglio del professore relatore di questa tesi, l'autrice ha scelto di lavorare con il dataset: **"Statlog (German Credit Data) Data Set"** ¹ fornito da UCI Machine Learning Repository ².

L'obiettivo che l'autrice si pone è quello di prevedere la capacità di rimborso di una serie di richiedenti credito analizzando quelle variabili che meglio li rappresentano.

Tuttavia è necessario, in prima battuta, sondare il dataset in modo tale da acquisire una profonda conoscenza sul contenuto dello stesso. Per questo motivo, nei prossimi paragrafi verrà discussa quella fase che nella disciplina di data analyst viene chiamata *"Exploratory Data Analysis"*, da qui in poi citata con l'acronimo EDA.

¹Professor Dr. Hans Hofmann, Institut Statistik und "Okonometrie Universit"at Hamburg, 1994-11-17, Disponibile al link: [https://archive.ics.uci.edu/ml/data sets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/data%20sets/statlog+(german+credit+data))

²Dua, Dheeru and Graff, Casey, 2017, UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml>, University of California, Irvine, School of Information and Computer Sciences

3.2 Preparazione del dataset e creazione del file "lettura_dati.R"

Innanzitutto si specifica che l'autrice per la stesura della applicazione pratica pensata per questo elaborato ha sviluppato con il linguaggio di programmazione R ³ nell'ambiente di RStudio ⁴

Le operazioni svolte in questa sezione del capitolo verranno computate in un file Rscript denominato **lettura_dati.R**, il cui contenuto è riportato in *appendice A*.

Questo file verrà caricato, tramite la funzione **load()**, in un ulteriore file di Rscript nel quale verrà praticata la statistica descrittiva oggetto della *sezione 3.3*.

3.2.1 Lettura del dataset

La prima cosa da fare nello script di RStudio è richiamare la funzione **read.table()**, mettendo come argomento il link dove il dataset è disponibile. In più si assegna al dataset il nome di **"dedatacredit"** ⁵, in modo tale da poter essere invocato più volte nel corso del progetto.

L'operazione di lettura del dataset non produrrà alcun output visibile ma consentirà al software di leggere il dataset in formato tabella e creare da esso un dataframe.

Si chiarisce che una frame di dati è una struttura dati organizzata in tabelle, somigliate per output grafico ad un foglio di calcolo.

Ora che il dataset è stato invocato e trasformato in dataframe, si interrompe momentaneamente la produzione di codice con lo scopo di conoscere l'effettivo contenuto del dataset grazie al documento "Data Set Description".

Tale documento è reperibile nella repository di UCI, e da esso si apprende che il dataset è composto da mille righe, le quali fanno riferimento alla numerosità campionaria.

Inoltre si evince che per ogni soggetto sono state raccolte venti osservazioni, nonché venti variabili di tipo qualitativo e quantitativo.

Per conoscere quali informazioni il dataset fornisce si legge nuovamente il documento "Data Set Description" vista la sua capacità descrittiva del dataset.

Con lo scopo di rendere la lettura più agevole si riporta il contenuto del file appena citato.

```

1      7. Attribute description for german
2      Attribute 1:  (qualitative)

```

³R è un linguaggio di programmazione che si presta per compiere azioni statistiche di qualunque genere. Non solo; R è dotato di grandi capacità grafiche. In questa tesi si vedrà solamente in parte la potenza di questo linguaggio di programmazione.

⁴R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2021, <https://www.R-project.org/>.

⁵"dedatacredit" è dunque un dataframe contenente una serie di colonne, nonché vettori che raccolgono tutte le variabili. Ogni variabile, contenuta nel vettore principale è essa stessa vettore dei valori che può assumere nelle varie osservazioni. Per esempio: all'interno della lista "dedatacredit" si trovano le osservazioni "age" e "sex". Queste due variabili sono a loro volta vettori che racchiudono i titoli assumibili, che in questo caso sono: "age"=c("19", "20", ..., "n") e "sex"= c("Male", "Female").


```

3           Status of existing checking account
4           A11 :      ... <      0 DM
5           A12 : 0 <= ... < 200 DM
6           A13 :      ... >= 200 DM /
7           salary assignments for at least 1 year
8           A14 : no checking account
9
10          Attribute 2: (numerical)
11                      Duration in month
12
13          Attribute 3: (qualitative)
14                      Credit history
15                      A30 : no credits taken/
16                      all credits paid back duly
17                      A31 : all credits at this bank paid back duly
18                      A32 : existing credits paid back duly till
19                          now
20                      A33 : delay in paying off in the past
21                      A34 : critical account/
22                          other credits existing (not at this
23                          bank)
24
25          Attribute 4: (qualitative)
26                      Purpose
27                      A40 : car (new)
28                      A41 : car (used)
29                      A42 : furniture/equipment
30                      A43 : radio/television
31                      A44 : domestic appliances
32                      A45 : repairs
33                      A46 : education
34                      A47 : (vacation - does not exist?)
35                      A48 : retraining
36                      A49 : business
37                      A410 : others
38
39          Attribute 5: (numerical)
40                      Credit amount
41
42          Attribute 6: (qualitative)
43                      Savings account/bonds
44                      A61 :      ... < 100 DM
45                      A62 : 100 <= ... < 500 DM
46                      A63 : 500 <= ... < 1000 DM
47                      A64 :      .. >= 1000 DM
48                      A65 : unknown/ no savings account
49
50          Attribute 7: (qualitative)
51                      Present employment since
52                      A71 : unemployed

```

```

50          A72 :      ... < 1 year
51          A73 : 1  <= ... < 4 years
52          A74 : 4  <= ... < 7 years
53          A75 :      .. >= 7 years
54
55      Attribute 8: (numerical)
56          Installment rate in percentage of disposable
                    income
57
58      Attribute 9: (qualitative)
59          Personal status and sex
60          A91 : male   : divorced/separated
61          A92 : female : divorced/separated/married
62          A93 : male   : single
63          A94 : male   : married/widowed
64          A95 : female : single
65
66      Attribute 10: (qualitative)
67          Other debtors / guarantors
68          A101 : none
69          A102 : co-applicant
70          A103 : guarantor
71
72      Attribute 11: (numerical)
73          Present residence since
74      Attribute 12: (qualitative)
75          Property
76          A121 : real estate
77          A122 : if not A121 : building society
                    savings agreement/
78                                life insurance
79          A123 : if not A121/A122 : car or other, not
                    in attribute 6
80          A124 : unknown / no property
81
82      Attribute 13: (numerical)
83          Age in years
84
85
86      Attribute 14: (qualitative)
87          Other installment plans
88          A141 : bank
89          A142 : stores
90          A143 : none
91
92      Attribute 15: (qualitative)
93          Housing
94          A151 : rent
95          A152 : own

```

```

96          A153 : for free
97
98      Attribute 16: (numerical)
99          Number of existing credits at this bank
100
101      Attribute 17: (qualitative)
102          Job
103          A171 : unemployed/ unskilled - non-resident
104          A172 : unskilled - resident
105          A173 : skilled employee / official
106          A174 : management/ self-employed/
107              highly qualified employee/ officer
108
109      Attribute 18: (numerical)
110          Number of people being liable to provide
111              maintenance for
112
113      Attribute 19: (qualitative)
114          Telephone
115          A191 : none
116          A192 : yes, registered under the customers
117              name
118
119      Attribute 20: (qualitative)
120          foreign worker
121          A201 : yes
122          A202 : no

```

Ora che sono note le variabili contenute del dataset si evince da una analisi sommaria che la maggior parte di esse è di tipo categoriale, le quali sono state chiamate nel precedente capitolo come variabili qualitative capaci di esprimere una caratteristica che non può essere tradotta in un numero.

Nello specifico tredici sono le variabili categoriali, mentre le restati sette sono di tipo numerico. Oltretutto si vuole anticipare al lettore la presenza di una ventunesima variabile, la quale dichiara se il richiedente credito in oggetto sia un creditore buono o cattivo. Essa è la variabile risposta del modello ⁶, la quale servirà per l'analisi sulla bontà del modello ottenuto dall'autrice per la previsione della capacità di rimborso dei mille soggetti presenti nel dataset **"Statlog (German Credit Data) Data Set"**. Per ora basti sapere al lettore che la colonna di riferimento a questa variabile verrà chiamata nel prossimo paragrafo come *"responsegoodcredit"*.

In generale, si consiglia al lettore di avere sempre a portata di mano il documento "Data Set Description", cosicché sarà più veloce individuare a cosa l'autrice, nel corso del progetto, si riferisce.

⁶Per variabile risposta si fa riferimento alla variabile dipendente, detta anche variabile di output di un modello statistico-predittivo.

3.2.2 Assegnare un nome ad ogni osservazione

A questo punto si procede assegnando ad ogni colonna, nonché variabile, un nome che ne faciliti l'identificazione, si parla in questo caso di *nomi parlanti*. Per operare nei termini appena dichiarati ci si avvale della funzione **colnames()**.

Quest'ultima ha la capacità d'impostare dei nomi a piacimento per ogni colonna del dataframe, pertanto si specifica ad argomento della funzione il dataset di riferimento.

Successivamente si assegna alla funzione un nuovo vettore contenente i nomi che si intende dare ad ogni colonna. Il vettore appena creato è definito in informatica come una lista ⁷ a cui sarà possibile accedervi nei modi che R permette.

Un modo per verificare l'avvenuta assegnazione per ogni colonna è quello di utilizzare la funzione **names()**, la quale ha come output l'elenco dei nomi delle colonne.

Si legga quanto riportato di seguito per conoscere i nomi che l'autrice ha attribuito ad ogni colonna.

```

1      names(dedatacredit)
2      > names(dedatacredit)
3          [1] "account_status"      "duration"
4              "credit_history"
5          [4] "purpose"             "credit_amount"
6              "saving_account"
7          [7] "present_employmentsince" "InstallmentRate"
8              "sex"
9          [10] "other_debtor"
10             "present_residencesince" "property"
11          [13] "age"                 "other_installplans"
12              "housing"
13          [16] "numexistingcredits"   "job"
14              "numpeople_maintenance"
15          [19] "telephone"           "foreignworker"
16              "responsegoodcredit"

```

3.2.3 Trasformare la variabile risposta "responsegoodcredit" in termini binari

Dal documento "Data Set Description" si apprende che l'output della variabile risposta è espresso in termini numerici, nonché 1 e 2, tuttavia per poter implementare il modello logit è necessario che sia formulata in termini binari.

Per verificare la veridicità di quanto appena detto, ci si avvale della funzione **head()** per visualizzare le prime righe del vettore "responsegoodcredit".

Un altro modo, che di fatto consente di ottenere lo stesso risultato, è quello di utilizzare la funzione **print()** che in generale è utile per ritornare i valori espressi nell'argomento della funzione.

⁷La lista è un insieme di dati finito. I valori che formano la lista sono elementi, i quali possono essere di diverse tipologie.

Si procede, dunque, visualizzando le prime venti righe. Ciò che ne consegue permette di considerare vero quanto detto prima.

```
1 > head(dedat acredit$responsegoodcredit)
2      [1] 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1
```

Per cambiare l'output della variabile "*responsegoodcredit*" da 1-2 a 0-1 è necessario sottrarre una unità a se stessa. In altre parole, i due valore che la variabile risposta può assumere saranno ridotti di una unità, ed ecco che l'output sarà binario.

In codice risulta essere:

```
1 dedat acredit$responsegoodcredit = dedat acredit$responsegoodcredit
  - 1
```

Anche in questo caso il codice non produrrà alcun output ma potrà essere verificato il buon fine dell'operazione nei modi sopra citati. Ora l'output delle prime righe sarà:

```
1 > head(dedat acredit$responsegoodcredit)
2      [1] 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
```

Si noti che il valore 0 si riferisce ai buoni richiedenti credito a differenza di quando l'osservazione "*responsegoodcredit*" assume valore di 1 in riferimento ai clienti che vengono considerati buoni creditori.

3.2.4 Identificare il vettore "*responsegoodcredit*" come un fattore

Tra i vari controlli preliminari è possibile effettuare una verifica sulla tipologia di dato che l'osservazione fornisce. In realtà, il documento "Data Set Description" esibisce già questa informazione tuttavia con la funzione *typeof()* all'interno di un *ciclo for* è possibile effettuare un ulteriore controllo.

Si noti che la variabile risposta è di tipo "*integer*". Ai fini del corretto funzionamento del modello è necessario specificare il vettore della colonna "*responsegoodcredit*" come fattore anziché come output numerico.

A tale scopo si utilizza la funzione *as.factor()*, la quale accetta come argomento una colonna di una classe o di un dataframe. Questo metodo farà convertire il vettore specificato come un fattore e il suo output non sarà visibile in console.

3.2.5 Train e Test split

Prima di procedere con l'implementazione del modello logit al dataset "*German Credit Data*" è necessario dividerlo in modo tale da avere una parte di dati utili all'allenamento del modello per poi testarlo con il rimanente.

L'autrice ha adottato la proporzione 70 : 30, dove 70 indica la percentuale di unità campionarie sul totale da utilizzare per l'allenamento del modello. Di conseguenza 30 si riferisce alla percentuale di soggetti richiedenti credito che si presteranno come gruppo di controllo.

Nella pratica si creano due variabili che identificheranno la proporzione di dataset destinata prima al "*train*" e poi al *test*.

Per preparare le due porzioni di dataset si è creato una variabile, chiamata "*test_index*", capace

di riferirsi ad un certo numero di righe del dataset tramite la funzione **sample()** che è in grado di estrarre dal dataset "*dedatacredit*" un certo numero di righe richieste.

Successivamente è stata creata una nuova variabile chiamata "*datatrain*", alla quale è stata assegnata la percentuale dichiarata nella variabile "*test_index*".

Per accedere alle righe del dataset si usano le parentesi quadrate ponendo ad oggetto la variabile "*test_index*" appena spiegata.

```
1 datatrain = dedatacredit[test_index,]
```

In altri termini "*dedatacredit[test_index,]*" è un modo di estrarre dal dataset "*dedatacredit*" il 70% delle righe. Ora, è necessario creare una terza variabile, la quale si dovrà riferire a quella porzione di dataset atta al test del modello.

Si noti che in questo caso la variabile "*datatest*" è stata moltiplicata per una unità negativa, creando di fatto il complementare della variabile *datatrain*.

Per controllare l'avvenuta divisione del dataset nelle due parti d'interesse è possibile utilizzare la funzione **nrow()** per identificare quante righe compongono di due nuovi dataset.

Ecco come appare l'output in console per la funzione appena richiamata per la variabile "*datatrain*":

```
1 > nrow(datatrain)
2 [1] 700
```

Il dataset su cui implementare il modello logit ha 700 righe, le quali corrispondono esattamente al 70% del totale campione statistico.

3.3 Statistica descrittiva con il dataset: "*German Credit Data*".

Ora che il file nel quale si prepara il dataset è stato creato si procede con l'esplorazione del dataset con l'obiettivo di estrarre una introduttiva conoscenza dallo stesso avendo alcuna conoscenza pregressa su di esso.

Avendo di base questo obiettivo l'autrice ha approcciato al dataset adottando parte della statistica descrittiva appresa durante il suo percorso di studio. La statistica descrittiva è quella branca della statistica che consente di rilevare elementi capaci di classificare, sintetizzare e rappresentare una serie dati raccolti da una popolazione o campione⁸.

Tutte le operazioni svolte in questa fase verranno salvate nel file chiamato dall'autrice come "*DescriptiveStat4GermanDataCredit.R*", il cui contenuto è copiato in *appendice B*.

⁸La popolazione statistica è l'aggregato di unità statistiche oggetto di studi. Il campione, invece, è parte della popolazione

3.3.1 Suddividere il dataframe "*dedatacredit*" in base alla variabile risposta

Se l'obiettivo è quello di prevedere, sulla base di determinate informazioni, se un richiedente credito sarà un buon o un cattivo creditore allora si rende necessaria la ricerca di quelle informazioni che possono essere definite come le più importanti tra il totale delle variabili offerte dal dataset. Il motivo per il quale si procede dividendo il dataframe tra *cattivi creditori* e *buoni creditori* è quello di cogliere segnali che aiutino a classificare le unità statistiche come tali. La porzione di dataframe contenente i *buoni creditori* è identificata come "*splitted_good*", il quale si ricorda essere ricavato dal dataframe chiamato in precedenza come "*dedatacredit*". Allo stesso modo per i *cattivi creditori*, il cui nome è *splitted_bad*.

3.3.2 Studio della variabile "*sex*"

Tra gli strumenti della statistica descrittiva si trovano le *tabelle di frequenza* e le *tabelle di contingenza*, che in questo caso verranno ricostruite per la variabile "*sex*".

Per ottenere una semplice tabella che esprima la frequenza di determinate osservazioni sarà sufficiente utilizzare la funzione *tab1()*, inserendo come argomento la variabile del dataset per la quale si intende studiare la frequenza.

Tuttavia è necessario scaricare la libreria "*epiDisplay*"⁹ tramite la funzione *install.packages()* e successivamente chiamarla tramite la funzione *library()*.

Inizialmente verrà prodotta la tabella di frequenza per la variabile "*sex*" con il suo relativo istogramma sul dataframe "*dedatacredit*", il quale si ricorda essere quello contenente tutte le unità statistiche. Successivamente, se meritevole di attenzione, verrà rielaborato il tutto su i due dataframe contenenti i *cattivi* e *buoni* creditori.

Si legga la sezione in *appendice B* per il codice nella sua interezza. Qui verrà riportato solamente parte dell'output.

```

1      > tab1(dedatacredit$sex, cum.percent = TRUE)
2      dedatacredit$sex :
3              Frequency Percent Cum. percent
4      A93              548    54.8         54.8
5      A92              310    31.0         85.8
6      A94              92     9.2         95.0
7      A91              50     5.0        100.0
8      Total          1000   100.0        100.0

```

Grazie alla tabella di frequenza appena elaborata si evince che la moda nella variabile "*sex*" è l'osservazione "**A93**", nonché uomini di sesso maschile non coinvolti in una relazione.

⁹Virasakdi Chongsuvivatwong, Virasakdi Chongsuvivatwong, Virasakdi Chongsuvivatwong, Epidemiological Data Display Packag, Package 'epiDisplay', versione "3.5.0.1", pagine "891–921", 2018-05-06, Package for data exploration and result presentation. Full 'epicalc' package with data management functions is available at '<<http://medipe.psu.ac.th/epicalc>>', "CRAN".

Dalla figura 3.1 si comprende ancora più velocemente quale sia il punto di massimo della distribuzione della variabile "sex".

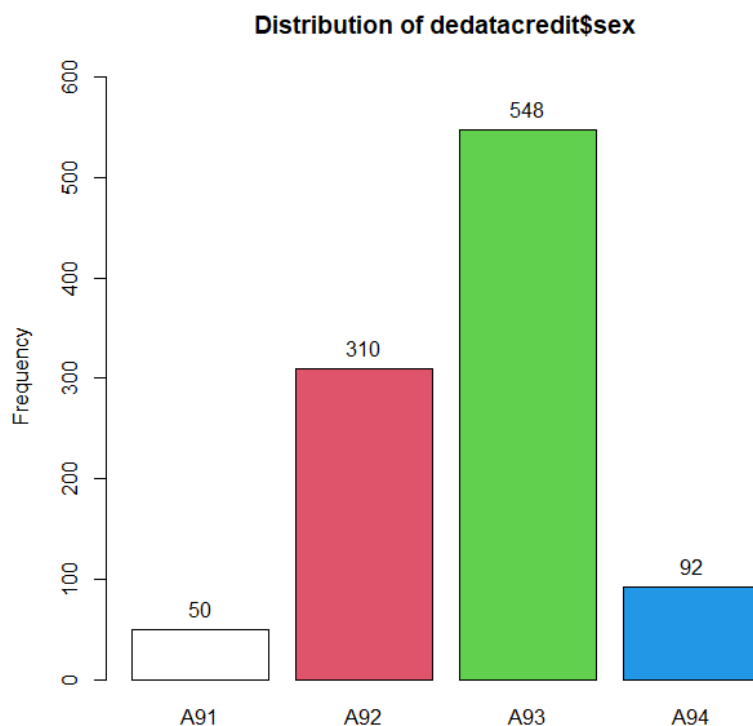


Figura 3.1: Istogramma di frequenza della variabile "sex"

In riferimento alla distribuzione di frequenza in figura 3.1 non è possibile esprimersi in merito alla sua forma. Questo impedimento è dovuto dal fatto che la variabile "sex" è di tipo qualitativo non ordinabile, ciò significa che non è possibile ordinare nell'asse orizzontale tutte le assunzioni che la variabile di riferimento assume.

L'analisi procede con l'elaborazione della *tabella di contingenza*, sempre in riferimento alla variabile oggetto di questa sezione.

Innanzitutto si specifica che una tabella di contingenza mette in relazione la frequenza congiunta di più variabili che in questo caso sarà la variabile risposta, oltre che alla variabile "sex".

Ciò che ne consegue è la seguente tabella:

1	> table(dedatacredit\$responsegoodcredit, dedatacredit\$sex)				
2					
3		A91	A92	A93	A94
4	0	30	201	402	67
5	1	20	109	146	25

Se il valore 0 si riferisce ai *buoni* creditori allora la maggior parte di essi è un maschio single. Vale lo stesso per i creditori *cattivi*, anch'essi per la maggior parte uomini maschi celibi.

Ne consegue che le mode per entrambe le sotto popolazioni siano identiche pertanto in questo caso non appare significativo studiare la distribuzione della variabile "sex" tra la popolazione di *cattivi* e *buoni* creditori poiché la distribuzione risulta la stessa osservata nell'analisi congiunta delle due tipologie di creditori.

3.3.3 Studio della variabile "age"

L'esplorazione del dataset attraverso le sue variabili procede con la variabile "age".

Poiché l'output della tabella di frequenza per la variabile in oggetto è troppo lungo verrà omesso, ciononostante sarà disponibile il suo relativo istogramma alla figura 3.2.

Dalla figura 3.2 si capisce che la massima frequenza, nonché la moda, è l'età di ventisette anni portata da 51 unità statistiche.

Si nota immediatamente che l'osservazione "age" non si distribuisce normalmente pertanto me-

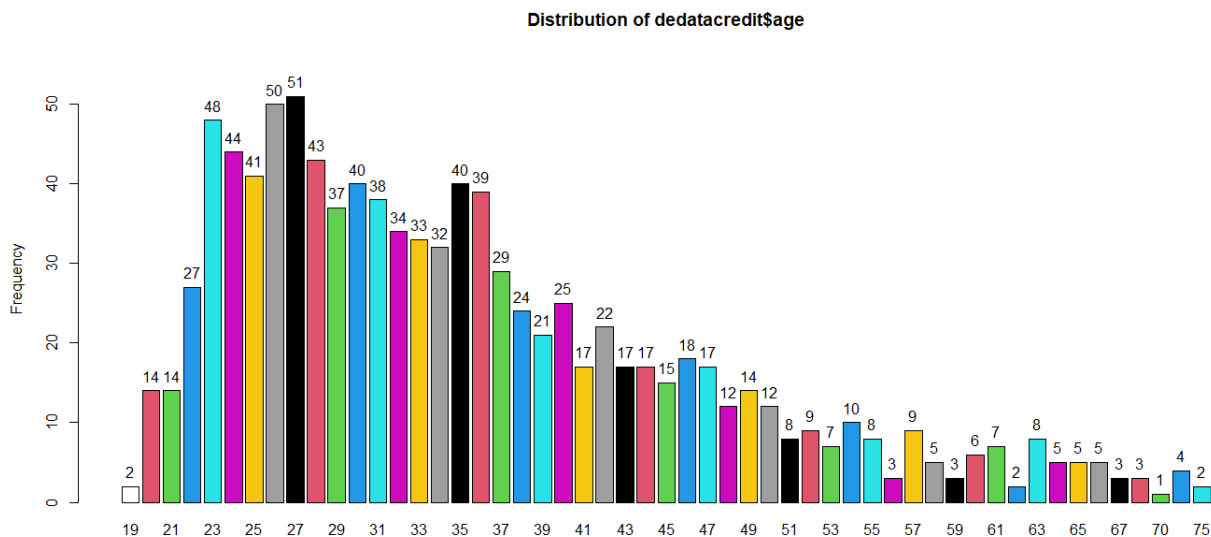


Figura 3.2: Istogramma di frequenza della variabile "age"

dia, mediana e moda non coincidono. Le caratteristiche che emergono dalla figura 3.2 lasciano intendere ad una distribuzione avente una curva asimmetrica positiva. Tanto è vero che la coda più lunga si trova alla destra del punto massimo della curva, il quale si riferisce alla moda.

Per una maggiore sicurezza sulla veridicità della tipologia di distribuzione della variabile "age", si calcola un indice, tra quelli possibili, per misurare l'asimmetria. In termini matematici l'indice scelto risulta essere:

$$\frac{(Media - Moda)}{DeviazioneStandard} \quad (3.1)$$

Si esplicita che la *media* d'età tra i tutti soggetti facenti parte del dataset "*dedatacredit*" è pari a 35.546 anni. La *moda*, come già discusso in precedenza, è pari a 27 anni.

Il valore che questo indice restituisce, una volta che è stato riportato a codice, è pari a 0.06604262, il quale essendo positivo e diverso da zero permette di confermare che la variabile "*age*" si distribuisca in modo asimmetrico e positivo.

Tuttavia, visto che tale indice ha un valore per cui si approssima allo zero si può affermare che tale asimmetria non renda la distribuzione fortemente discostata dalla normalità. Si riporta la figura 3.3, il cui codice è in **Appendice B**, per meglio visualizzare la distribuzione in oggetto.

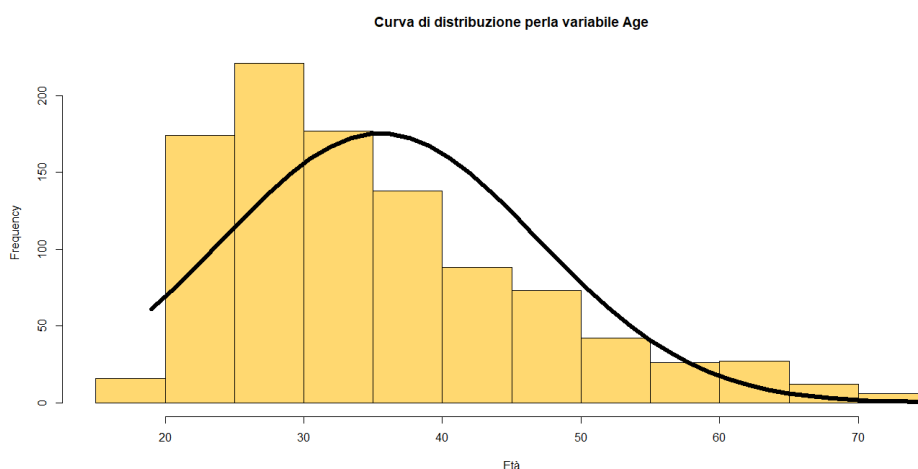


Figura 3.3: Distribuzione di frequenza della variabile "*age*"

Come per la variabile "*sex*" si elabora nelle stesse modalità la tabella di contingenza per l'età, il cui codice e output sarà, anche in questo caso, riportato in **Appendice B** data la tua modesta lunghezza.

Da esso si apprende che la maggior parte dei *cattivi* creditori hanno ventisette anni, mentre i *buoni* ventitré.

In questo caso, dunque, è interessante procedere con una analisi separata tra le due tipologie di clienti, avvalendosi dei due dataframe "*splitted_good*" e "*splitted_bad*", che si ricordano esse figli del dataframe "*dedatacredit*".

Elaborando la distribuzione di frequenza per la variabile "*age*" nei due dataframe, appare che entrambi si distribuiscono asimmetricamente verso sinistra. In altre parole la distribuzione d'età di tutte le unità statistiche non si differenzia se esse vengono divise in base all'osservazione della variabile risposta.

Non dovrebbe stupire tale conclusione visto che nella pratica le medie nei due tipi di clienti non si discostano molto l'uno dall'altra.

3.3.4 Studio della variabile "duration"

La durata del credito richiesto è la variabile che qui si andrà ad approfondire. Si imposta, anche in questo caso, la *tabella di frequenza*, la quale sarà parzialmente disponibile di seguito.

```

1      > tab1(dedatacredit$duration, cum.percent = TRUE)
2      dedatacredit$duration :
3      Frequency Percent Cum. percent
4      4          6      0.6          0.6
5      ..          ..      ...          ...
6      11         9      0.9         18.0
7      12        179     17.9         35.9
8      13         4      0.4         36.3
9      ..          ..      ...          ...
10     22         2      0.2         58.6
11     24        184     18.4         77.0
12     ..          ..      ...          ...
13     Total      1000    100.0        100.0

```

Assieme alla tabella di frequenza la variabile **tab1** eroga anche l'istogramma in riferimento ai valori sopra riportati. Da esso si vede immediatamente che la maggior parte dei soggetti che hanno richiesto un credito lo ha fatto per una durata di 24 mesi. Tuttavia anche 12 mesi sono un arco di tempo molto gettonato per la restituzione di un credito, in particolare 179 richiedenti credito sono impegnati con l'istituto finanziario per la durata di un anno.

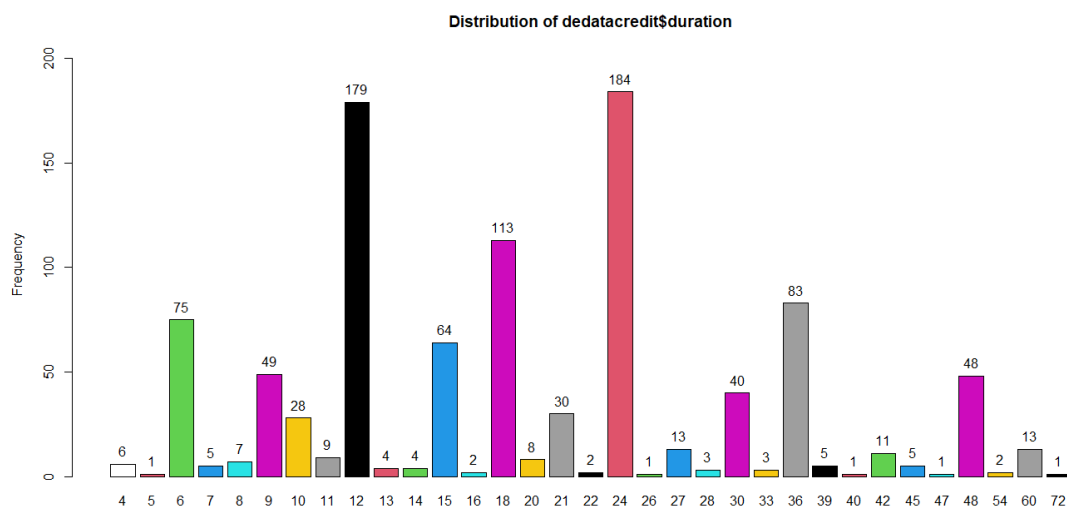


Figura 3.4: Istogramma della frequenza della variabile "duration"

Procedendo nelle stesse modalità adottate per le altre variabili si elabora la *distribuzione di frequenza* marcando con una linea continua la tipologia di distribuzione. Si veda la figura 3.5 riportata di seguito.

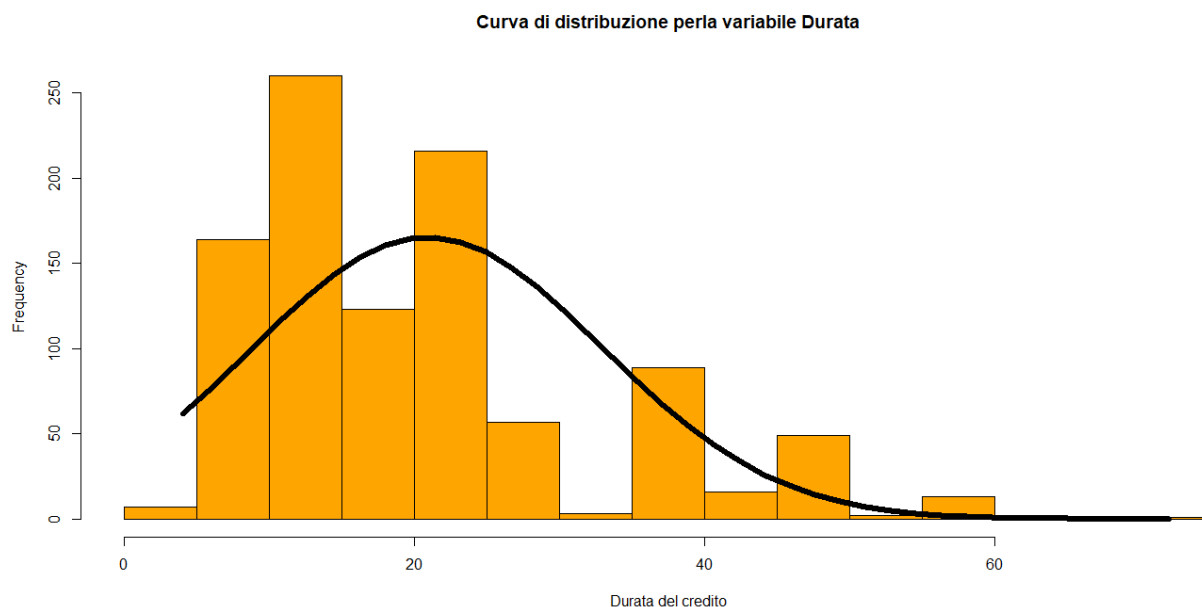


Figura 3.5: Distribuzione di frequenza della variabile "age"

Anche la variabile "*duration*" si distribuisce asimmetricamente con tendenza positiva, tale affermazione trova conferma nel valore che assume l'indice calcolato anche per la variabile "*age*" nell'equazione numero 3.1. Nello specifico l'indice assume un valore pari 0.02129766 ed essendo diverso da zero è indicativo di asimmetria.

Per capire se ha senso procedere con uno studio separato della variabile "*duration*" in funzione dei *cattivi* e *buoni* creditori si sviluppa la rispettiva *tabella di contingenza*, della quale anche in questo caso verrà riportato solo in parte l'output.

```

1      > table(dedatacredit$duration,
2              dedatacredit$responsegoodcredit)
3
4              0    1
5      4         6    0
6      5         1    0
7      ..      ..   ..
8      ..      ..   ..
9      12      130   49
10     13         4    0
11     14         3    1
12     15        52   12
13     ..      ..   ..

```

13	22	2	0
14	24	128	56
15	26	1	0
16

Dalla tabella di contingenza appare opportuno approfondire l'analisi tra le due classi di soggetti, visto che la durata del credito più frequente tra i *buoni* creditori è di 24 mesi. 130 *cattivi* creditori, invece, rendono 12 mesi la frequenza assoluta.

Banalmente ne consegue che i creditori buoni, in termini di frequenza, hanno una durata più lunga rispetto ai clienti classificati come cattivi, nello specifico si sono impegnati in un debito per il doppio del tempo.

Elaborando la distribuzione di frequenza per i clienti *cattivi* in "*splitted_bad*" e per i clienti *buoni* risultano entrambe asimmetriche positive con un indice di asimmetria pari 0.05871077 per i primi e 0.004874496 per i secondi.

A questo punto ci si chiede se vi sia correlazione tra durata del credito e il suo ammontare. Si proverà a rispondere a questo quesito nella sezione dedicata allo studio congiunto di più variabili.

3.3.5 Studio della variabile "*purpose*"

La prossima variabile oggetto di analisi è la variabile "*purpose*", la quale indica il motivo per il quale i clienti della banca hanno richiesto il credito.

Nelle stesse modalità adottate per le precedenti variabili si elabora la *tabella di frequenza* per la totalità delle unità statistiche, dalla quale emerge che il motivo "A43" è quello che più incoraggia i richiedenti credito.

Dal file *Data Set Description* si apprende che il codice "A43" si riferisce all'acquisto di "*radio/-television*".

Il motivo meno frequente è che l'"A49", il quale si riferisce a "*business*". Non appare strano che i richiedenti credito per "*business*" sia in misura inferiore visto che il dataset **German Credit Data** raccoglie informazioni su persone fisiche. Di seguito si riporta la tabella di frequenza oltre che all'istogramma che figurano quanto appena spiegato.

1	> tab1(dedatacredit\$purpose, sort.group = "decreasing",			
	cum.percent = TRUE)			
2	dedatacredit\$purpose :			
3		Frequency	Percent	Cum. percent
4	A43	280	28.0	28.0
5	A40	234	23.4	51.4
6	A42	181	18.1	69.5
7	A41	103	10.3	79.8
8	A49	97	9.7	89.5
9	A46	50	5.0	94.5
10	A45	22	2.2	96.7
11	A44	12	1.2	97.9
12	A410	12	1.2	99.1
13	A48	9	0.9	100.0
14	Total	1000	100.0	100.0

In figura 3.6 si può osservare la distribuzione di frequenza per la variabile oggetto di questa sezione.

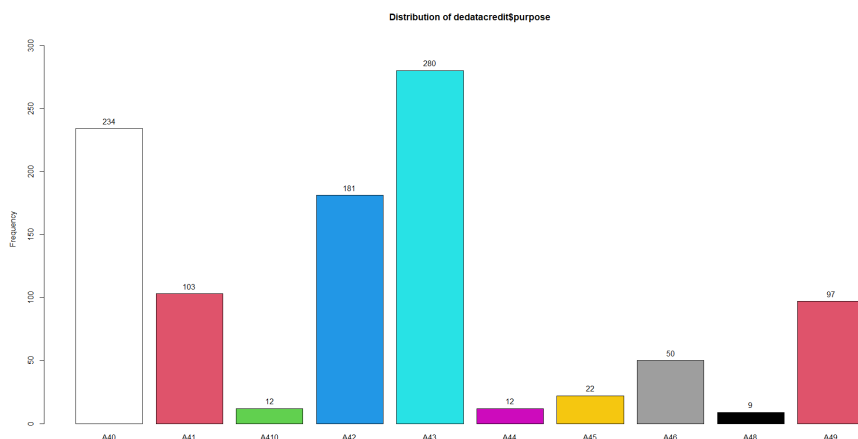


Figura 3.6: Istogramma di frequenza della variabile "purpose"

Ciò che restituisce la figura 3.6 non è utile a fini distributivi poiché anche la variabile "purpose" è di tipo qualitativo non ordinabile.

Ci siamo chiesti se la durata del credito sia in quale modo correlata al suo ammontare, ora ci si interroga su una possibile correlazione con anche il motivo. Banalmente, un credito per l'acquisto di un'auto è maggiore di quello fatto per l'acquisto di un televisore.

Si risponderà a questo quesito più avanti. Tuttavia adesso si proverà a capire se la maggior parte clienti *buoni* siano spinti a richiedere un credito per un motivo diverso rispetto a quelli *cattivi*.

La tabella di frequenza riportata di seguito indica che i clienti *cattivi* utilizzano il credito richiesto per comprare una radio o una televisione mentre quelli considerati *buoni* per comprare un'auto nuova ¹⁰.

```

1 > table(dedatacredit$responsegoodcredit, dedatacredit$purpose)
2
3      A40 A41 A410 A42 A43 A44 A45 A46 A48 A49
4 0 145  86    7 123 218   8  14  28   8  63
5 1  89  17    5  58  62   4   8  22   1  34

```

¹⁰Dal documento "Data Set Description" si evince che il motivo A43 è "radio/television" mentre il A40 è "car (new)".

3.3.6 Le altre variabili

Le altre variabili presenti nel dataframe non sono state oggetto di analisi poiché, come per la variabile "*sex*", non risulta statisticamente rilevante studiarle in funzione della variabile "*responsegoodcredit*".

3.3.7 Studio di più variabili contemporaneamente

Dopo aver analizzato singolarmente le variabili ritenute più opportune dall'autrice, si procede studiandole congiuntamente con l'obiettivo di trovare una quale relazione con la classe di cliente possibile.

Il dataset "**Statlog (German Credit Data) Data Set**" fornisce una variante di se stesso con le medesime variabili ma tutte in forma numerica, chiamato *German.Data-Numeric*.

Ci si avvale di esso per creare la **matrice di correlazione**, della quale verrà data anche una rappresentazione grafica tramite un *correlogram*¹¹.

In statistica l'indice di correlazione o dipendenza identifica qualsiasi relazione statistica, causale o meno, tra due variabili casuali. L'insieme delle correlazioni possibili tra tutte le variabili del dataset verranno visualizzate nella matrice di correlazione.

Una matrice di correlazione è una tabella che mostra i coefficienti di correlazione tra le variabili oggetto di studi. Ogni cella della tabella mostra la correlazione tra due variabili.

Data la grandezza, e quindi la sua complessità di interpretazione, che al matrice di correlazione assume per il dataset in oggetto non verrà riportata di seguito. Ciononostante viene inserita a pagina seguente la figura 3.7 che riporta il correlogramma della matrice di correlazione appena calcolata.

Per una migliore comprensione del correlogramma si noti che le correlazioni positive sono visualizzate in viola e quelle negative in arancione. Inoltre, l'intensità del colore e la dimensione del cerchio sono proporzionali ai coefficienti di correlazione.

In aggiunta nella parte destra del correlogramma viene fornita una legenda che mostra i coefficienti di correlazione e i colori corrispondenti.

Dalla figura 3.7 emerge a prima vista che nessuna variabile all'interno del dataset sia condizionata positivamente da un'altra. Esistono però delle correlazioni negative sufficientemente forti da meritare attenzione.

Prima di procedere con un'analisi più approfondita è necessario una distinzione tra le due tipologie di correlazione appena citate, ovvero quella positiva e quella negativa.

Si dice *correlazione positiva* se le due variabili si muovono nella stessa direzione. Ne è un esempio banale il voto di un esame e le ore di studio: più ore si dedicano allo studio più alto sarà il voto conseguito.

D'altro canto una *correlazione negativa* è tale se le due variabili si muovono in senso opposto, come la temperatura che si abbassa man mano che l'altitudine aumenta.

¹¹ Il correlogramma è un grafico che consente di visualizzare delle statistiche di correlazione.

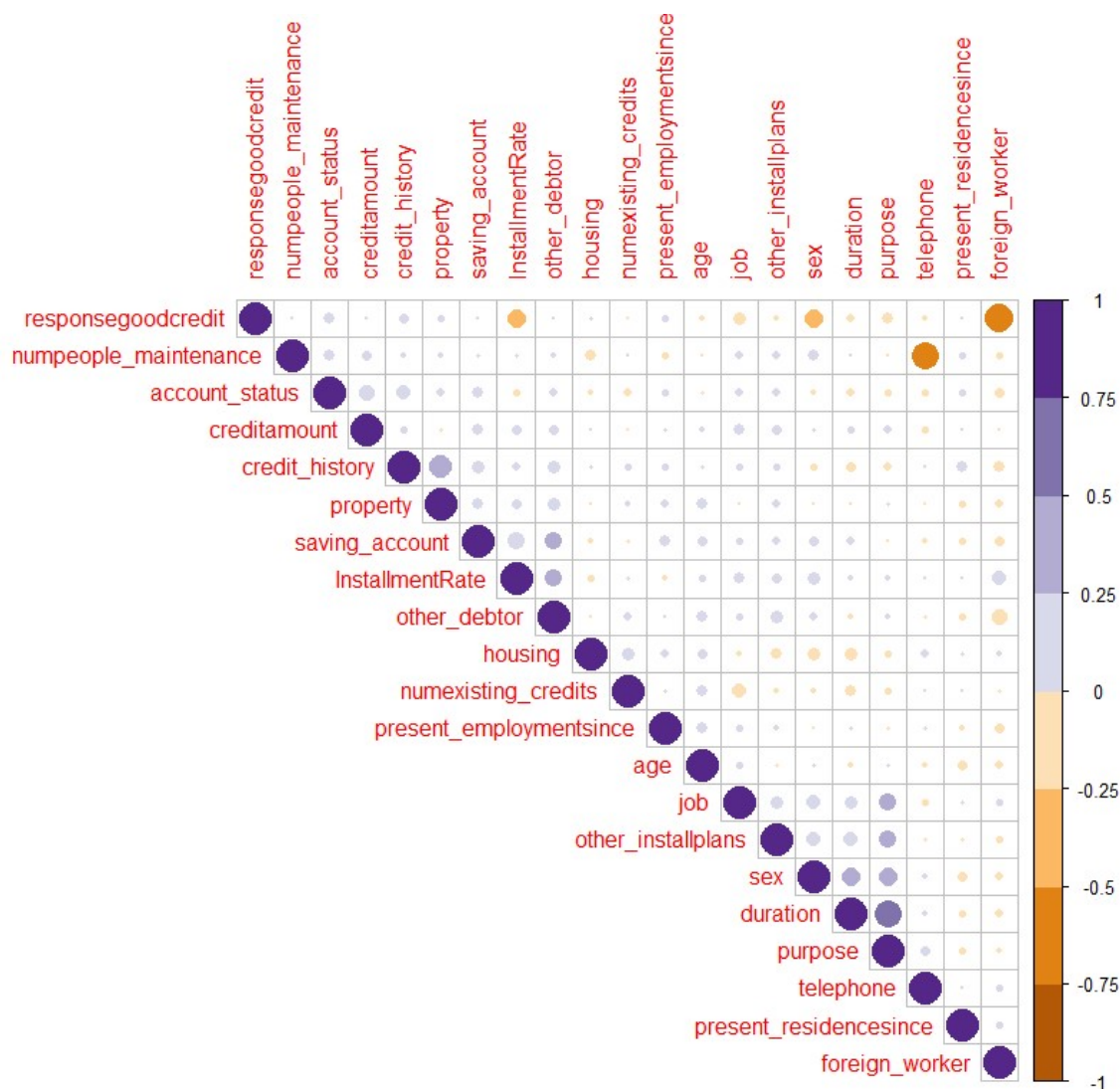


Figura 3.7: Correlogram per tutte le variabili all'interno del dataset "Statlog (German Credit Data) Data Set" in versione numerica

Le variabili "duration" e "purpose" hanno un indice di correlazione pari a 0.625, il quale indica una forte correlazione positiva. Questa affermazione risponde al quesito posto in *sezione 3.3.5*, cioè che la durata del credito è scelta in funzione al motivo per il quale è stato stipulato.

Per dare una risposta completa si deve verificare che tipo di dipendenza vi è tra la variabile "duration" e la variabile "creditamount". Dal risultato in console della matrice di correlazione si legge che l'indice in oggetto per le variabili di riferimento è pari a 0.048 che nella pratica è nullo. Un indice positivo tendente a zero è sinonimo di una dipendenza positiva molto debole se non nulla. Vale a dire che l'ammontare del credito richiesto incide leggermente sulla sua durata.

In ultima battuta si ricerca la correlazione tra il capitale richiesto a credito e il motivo per il quale

è stato richiesto, che è pari a 0.065. Come per le due variabili precedenti anche in questo caso la correlazione è debole perciò valgono le stesse considerazioni.

Tra le altre variabili avente una correlazione positiva meritevole di attenzione sono le variabili "*credit_history*" e "*property*", le quali hanno un indice di correlazione pari a 0.437. Nonostante il livello di tale correlazione sia moderato si procede con l'analisi producendo la tabella di contingenza di cui sotto.

```

1 > table(dedatacredit$credit_history, dedatacredit$property)
2
3           A121 A122 A123 A124
4      A30      5   10   17    8
5      A31     11    8   15   15
6      A32    158   122  179   71
7      A33     19    23   31   15
8      A34     89    69   90   45

```

Si nota immediatamente che nella maggior parte dei casi i clienti rientrano nella categoria *A32* ovvero sono clienti che possiedono altri crediti e che questi siano stati pagati regolarmente fino ad ora.

Le altre variabili sono collegate tra loro ancora più lievemente di quelle in qui analizzate, tanto è vero che nella maggior parte dei casi l'indice di correlazione è inferiore allo 0.5 che si ricorda essere rilevatore di una debole correlazione positiva.

Ciò che emerge dunque è che la maggior parte delle variabili sono poco o per nulla correlate tra loro e questo significa che è impossibile estrapolare informazioni reciprocamente tra le variabili. In altre parole, in riferimento al modello predittivo che si andrà a costruire nel prossimo capitolo, si deve tenere a mente che se due variabili hanno un elevato indice di correlazione allora l'informazione di una è contenuta nell'altra. Sotto questa ipotesi si può dire che l'omissione o l'aggiunta di una delle due variabili correlate non modifica le prestazioni del modello.

Tuttavia, data la generale indipendenza delle variabili appare che tutte le variabili siano necessarie a fini predittivi.

Questa affermazione può trovare conferma solo nel prossimo capitolo dove si produrrà il modello predittivo.

Capitolo 4

Modello logit per il data dataset: "*German Credit Data*"

4.1 Cenni ai modelli lineari generalizzati

Prima di procedere con la lettura di questo capitolo si invita il lettore a rileggere quanto detto nei primi capitoli, dove è stata data una definizione della regressione del modello logit.

In estrema sintesi il modello logit è utile quando si intende prevedere un risultato binario da un insieme di variabili predittive continue.

John Nelder¹ e Robert Wedderburn² hanno creato i "*modelli lineari generalizzati*"³ per unificare i diversi modelli statistici come, in questo caso, la regressione del modello logit.

Fortunatamente il linguaggio di programmazione R fornisce una serie di pacchetti che permettono di utilizzare i modelli lineari generalizzati tramite diverse funzioni. In questo elaborato si impiegherà la funzione *glm()*, che può essere tratta da diversi pacchetti di R.

4.2 Applicazione del modello Logit su tutte le variabili

In questa sessione si approccia al modello di regressione logit applicandolo alla totalità delle variabili contenute nel data frame "*dedatacredit*".

Si anticipa al lettore che l'uso di tutte le 21 variabili potrebbe creare un modello incapace di

¹John Ashworth Nelder è stato uno studioso molto influente nella statistica, egli è ricordato per il suo lavoro nell'analisi della varianza e nella "teoria statistica".

²Robert William MacLagan Wedderburn è ricordato per essere stato coautore di John Nelder nella metodologia del modello lineare generalizzato, per poi ampliarlo nella teoria della verosimiglianza.

³Tra le altre ne fanno parte: la regressione lineare, la regressione logit, la regressione di Poisson, la regressione binaria, la Log-log complementare e molte altre.

predire opportunamente la variabile risposta. Di seguito si discute il problema *dell'overfitting* e *dell'underfitting* e come essi possano essere evitati nell'applicazione pratica di questa tesi.

4.2.1 Il problema dell'overfitting e dell'underfitting

Ogni qualvolta che si intende costruire un modello predittivo si deve tenere in considerazione la possibilità che il modello in oggetto possa andare in sovradattamento.

In parole povere un modello viene considerato in overfitting quando egli si adatta troppo ai dati usati per l'addestramento e conseguentemente non è capace di prevedere in modo preciso i dati di test non visualizzati.

Tra le cause che possono indurre all'overfitting si riconosce l'uso di troppe variabili, le quali renderanno il modello troppo complesso e quindi difficile da interpretare.

Inoltre, mantenere tante o tutte le variabili di un dataset, renderanno le stime dei parametri estremamente instabili.

Se un modello in overfitting ha una prestazione buona in allenamento ma scarsa nella pratica allora un modello in underfitting avrà una pessima produttività sia in train che in test. L'underfitting non è problematico quanto l'overfitting poiché esso è facilmente diagnosticabile con una buona metrica delle prestazioni.

4.2.2 La funzione *glm()*

Per applicare i *modelli lineari generalizzati* al nostro dataframe di test, che si ricorda essere "*datatrain*", si utilizzerà la funzione *glm()*, la quale consente di applicare i seguenti modelli di regressione:

- Regressione gaussiana
- Regressione di Poisson
- Regressione binomiale (classificazione)
- Regressione binomiale frazionaria
- Regressione quasibinomiale
- Classificazione multinomiale
- Regressione gamma
- Regressione ordinale
- Regressione binomiale negativa
- Distribuzione Tweedie

Nel caso di questa applicazione verrà utilizzata la regressione binomiale di classificazione chiamata anche regressione logistica binomiale.

Una volta caricato il file contenuto nello script "*lettura_dati.R*" tramite la funzione *load()*, sarà sufficiente invocare la funzione *glm()* assegnandogli un nome a piacere, che in questo caso è

"myglm". Questa riga di codice non emetterà alcun output evidente, perciò è necessario usare la funzione **summary()**, mettendo come argomento il nome che è stato assegnato alla funzione **glm()**.

Ciò che si riporta di seguito è l'output parziale, completo solo in *Appendice C*:

```

1      > summary(myglm)
2
3      Call:
4      glm(formula = responsegoodcredit ~ ., family = "binomial",
5           data = datatrain)
6
7      Deviance Residuals:
8      Min       1Q   Median       3Q      Max
9      -2.0219  -0.6990  -0.3431   0.6794   2.8247
10
11     Coefficients:
12             Estimate Std. Error z value Pr(>|z|)
13 (Intercept)    9.404e-01  1.279e+00   0.736  0.46200
14 account_statusA12 -5.133e-01  2.660e-01  -1.929  0.05369 .
15 account_statusA13 -1.383e+00  4.837e-01  -2.859  0.00424
16      **
17 account_statusA14 -1.779e+00  2.785e-01  -6.387 1.69e-10
18      ***
19 duration         3.049e-02  1.147e-02   2.657  0.00788
20      **
21 credit_historyA31 -2.846e-01  7.015e-01  -0.406  0.68495
22 .....           .....           .....
23 .....           .....           .....
24 housingA153      -3.446e-01  6.012e-01  -0.573  0.56657
25 numexisting_credits 3.284e-01  2.309e-01   1.422  0.15490
26 jobA172          5.475e-01  7.717e-01   0.710  0.47799
27 jobA173          4.382e-01  7.365e-01   0.595  0.55189
28 jobA174          4.115e-01  7.443e-01   0.553  0.58032
29 .....           .....           .....
30 .....           .....           .....
31 numpeople_maintenance 4.114e-01  3.049e-01   1.349  0.17724
32 telephoneA192     -5.003e-01  2.460e-01  -2.033  0.04202 *
33 foreign_workerA202 -1.465e+00  8.306e-01  -1.764  0.07778 .
34 ---
35 Signif. codes:  0      ***    0.001    **    0.01    *    0.05
36                  .    0.1      1
37
38      (Dispersion parameter for binomial family taken to be 1)
39
40      Null deviance: 869.91  on 699  degrees of freedom
41      Residual deviance: 616.08  on 651  degrees of freedom
42      AIC: 714.08
43
44      Number of Fisher Scoring iterations: 14

```

Da una lettura sommaria si apprende che questo output non è di facile interpretazione perciò Verranno dedicate una serie di sezioni atte alla sua spiegazione.

4.2.3 Interpretazione dei "*Coefficients*"

Scorrendo l'output che RStudio presenta in console, si trovano i "*Coefficients*". Si ricorda che nei modelli di regressioni i coefficienti sono delle stime di quei parametri aventi capacità informativa sulla popolazione sconosciuta.

Essi descrivono la relazione tra una o più variabili predittive e la sua risposta. In altri termini i coefficienti sono i valori che moltiplicano i valori predittivi.

Il segno dei coefficienti indica la direzione della relazione tra le due variabili. Vale la regola che:

- Se il segno è *positivo* un aumento della variabile predittiva implica un aumento della variabile risposta.
- Se il segno è *negativo* vale viceversa.

Tra i coefficienti elaborati si trovano: gli "*Estimate*", lo "*Std. Error*", lo "*z value*" e la "*P(>|z|)*". Di seguito verranno interpretati alcuni risultati con lo scopo di dare al lettore una metodologia applicabile a tutte le stime calcolate dal modello logit.

Tuttavia prima di procedere è importante per una facile interpretazione arrotondare tutti i coefficienti, tramite la funzione ***round()***.

"L'Estimate"

Le stime indicano la relazione tra le variabili indipendenti e la variabile dipendente, dove la variabile dipendente si trova sulla scala logit.

Queste stime indicano di quanto aumentano le probabilità logaritmiche previste se l'*estimate* aumenta di una unità, mantenendo costanti tutti gli altri coefficienti. La stima della variabile "*duration*" è 0.3 e come è intuibile si approssima allo zero, di fatto la rende una variabile poco influente. Vale la regola, dunque, che maggiore è la stima maggiore sarà la sua influenza sulla variabile risposta.

Lo "Std. Error"

Il coefficiente "*Std. Error*"⁴ fornisce una quantità media che identifica di quanto la stima si scosta dal valore medio effettivo della nostra variabile di risposta.

Vale la regola che maggiore è l'errore standard meno sicura e precisa è la previsione su quella osservazione.

Per comprendere meglio quando appena spiegato si prenda in esame l'osservazione "*duration*", essa ha un errore standard pari 0.0115 che la rende una stima più sicura di "*credit_historyA31*"⁵, la quale ha un errore standard a 0.7015.

⁴Gli errori standard possono anche essere usati per calcolare gli intervalli di confidenza.

⁵la variabile *credit_historyA31* corrisponde nel file Data Set Description come "all credits at this bank paid back duly".

Lo "z value"

In parole semplici lo z-score identifica di quanto il valore ottenuto è distante dalla media, in termini di deviazioni standard⁶. Per esempio se lo z-score è pari a zero significa che la stima è pari alla media.

Considerando l'osservazione "*duration*" si evince che ha uno z-score pari a 2.997, questo significa che è a 2.657 deviazioni standard al di sopra della media.

Lo z-score dell'osservazione "*purposeA46*", che dal documento "Data Set Description" si apprende essere "*education*", è pari a -0.021 . Questo consente di affermare che la stima sulla motivazione della richiesta debito è nell'intorno della media.

In generale, più grande è il "*z value*" di una variabile più essa è importante.

La " $P(>|z|)$ "

Il "*p-value*" per la statistica Z deve essere interpretato nel seguente modo. Esso indica di quanto sia probabile che un risultato estremo o più estremo di quello osservato si sarebbe verificato sotto l'ipotesi nulla. Un valore elevato del "*p-value*" indicherebbe che la variabile in oggetto sia poco rilevante per il modello, tuttavia è sbagliato selezionare in questo modo le variabili da inserirvi, poiché è necessaria un'analisi incrociata con anche l'errore standard.

Tanto è vero che con un errore standard molto elevato non si sarebbe in grado di affermare che la variabile abbia o no effetto su modello.

A parità di errore standard maggiore è il p-value minore sarà il suo effetto e vale viceversa con un p-value basso.

4.2.4 Il metodo AIC

L'Akaike's Information Criteria non è altro che un mezzo capace di facilitare la selezione del modello poiché esso è in grado di stimarne la qualità.

L'AIC stima la quantità di informazioni che si sono perse lungo il modello, perciò vale la regola che minore è l'AIC maggiore è la qualità del modello in analisi.

Inoltre, l'AIC fornisce una misura della semplicità del modello.

Si comprende aver compreso che la sua potenza sta nell'essere in grado di porre insieme il rischio si overfitting e di underfitting, la cui discussione è già avvenuta.

Compresa l'importanza del criterio di Akaike appare evidente che l'interpretazione dell'AIC in valore assoluto non sia di alcuna rilevanza, ma che questo debba essere confrontato con l'AIC di altri modelli candidati.

Nei prossimi paragrafi saranno adottate nuove tecniche per la costruzione del modello di predizione e ne seguirà, dunque, un confronto tra i rispettivi Akaike's Information Criteria.

Si tenga a mente che una delle regole che Burnham e Anderson hanno fornito è che se due modelli hanno un AIC che si differenzia di due unità allora sono identici, ma se la differenza aumenta a cinque allora il modello con l'AIC minore è leggermente migliore rispetto all'altro. Per divari maggiori di dieci unità si può affermare che tale differenza è una prova abbastanza forte che il

⁶Si ricorda nuovamente che la deviazione standard è una misura della dispersione di un insieme di dati.

modello con l'AIC inferiore sia migliore.

Per completezza, si sottolinea che l'AIC del modello *"myglm"* è pari a 714.08.

4.2.5 Altre considerazioni

In questa parte verrà prodotta la matrice di confusione, la quale è idonea per mostrare la capacità previsionale del modello. La matrice di confusione è una matrice 2×2 che mette in relazione i dati reali con quelli previsti in funzione della variabile risposta.

Per far sì che l'output del modello *"myglm"* sia binario si implementa a codice una condizione per la quale un cliente verrà considerato *buono* se la variabile risposta assumerà nel modello un valore superiore a 0.5.

Si parla di valore soglia, di cui se ne discuterà più avanti.

Il codice, anche in questo caso, verrà riportato in *Appendice C*, tuttavia l'output in console è il seguente:

```

1      > table(datatrain$responsegoodcredit, predicted_myglm, dnn =
2              c("Truth", "Predicted"))
3              Predicted
4      Truth    0    1
5              0 425   56
              1  92 127

```

La matrice di confusione, sopra riportata, comunica le seguenti considerazioni:

- *Veri positivi*: 127
- *Veri negativi*: 425
- *Falsi positivi*: 56
- *Falsi negativi*: 92

L'utilità della matrice di confusione non si esaurisce nella mera esplicazione dei risultati del modello, poiché da essa è possibile calcolare la *Sensibilità*, la *Specificità*, la *Precisione* e l'*Accuratezza*.

La **sensibilità**, chiamato anche *Tasso di Vero Positivo* è quella parte di prestiti previsti a buon fine e che lo sono anche nella realtà. In termini algebrici risulterà:

$$Sensibilita' = \frac{A[1,1]}{A[1,1] + A[2,1]} \quad (4.1)$$

Dove A è la matrice di confusione, il cui argomento fa riferimento agli elementi di posto di se stessa.

Il codice per ottenere il seguente indice, ma anche i prossimi è in *Appendice C*, in ogni caso il tasso di sensibilità è pari a 0.8220503.

la **specificità** sempre per il modello *"myglm"* è pari a 0.6939891 e in termini algebrici risulterà:

$$Specificita' = \frac{A[2,2]}{A[2,2] + A[1,2]} \quad (4.2)$$

Per ora non viene data alcuna informazione su questi due indici ma se ne discuterà tra poche righe in riferimento alla curva di ROC.

Si prosegue l'analisi con il calcolo della **precisione** che come facilmente intuibile essa è la proporzione dei clienti *cattivi* correttamente previsti su tutte le unità statistiche, la quale risulta essere pari a 0.8835759. Di seguito è esplicita la funzione del tasso di precisione:

$$Precisione = \frac{A[1,1]}{A[1,1] + A[1,2]} \quad (4.3)$$

L'indice di precisione indica quanto ripetibili siano i suoi risultati, ovvero il suo grado di affidabilità e di coerenza.

La precisione è capace di quantificare il grado di efficacia con cui sono state effettuate le misure.

La precisione, dunque, si esprime sul modello, ma non è capace di giudicare il suo risultato.

Infine, si procede con il calcolo della **accuratezza** che rapporta i clienti correttamente valutati con il totale della popolosità campionaria.

A differenza della precisione, l'accuratezza è in grado di pronunciarsi sul modello di previsione ma in questo caso in termini esattezza, ovvero quanto una misura di avvicini al vero.

$$Accuratezza = \frac{A[1,1] + A[2,2]}{A[1,1] + A[2,2] + A[1,2] + A[2,1]} \quad (4.4)$$

Se un modello ha una precisione elevata e una accuratezza elevata si può dire che ha "centrato l'obiettivo".

Riassumendo il modelli di previsione "*myglm*" è sensibile allo 82%, è specifico al 69%, è preciso al 88% ed infine è accurato al 78%.

Un altro modo di visualizzare le prestazioni di un modello di classificazione è la curva **ROC**. Essa pone in relazione il Tasso di Vero Positivo con il Tasso di Falsi Positivi a diverse soglie di classificazione.

Il valore critico o soglia opera da "sparti acque" tra la positività e la negatività del test di assegnazione. Per ora si prenda come vero che un minore valore soglia classifica più elementi come positivi, aumentando così sia i falsi positivi che i veri positivi.

In altri termini, se la soglia di classificazione è bassa allora sarà più facile per una osservazione essere prevista come positiva.

La curva ROC indica quanto il modello è in grado di distinguere tra le classi le varie osservazioni, perciò la curva ROC non è altro che una curva di probabilità e quindi più si discosta dalla bisettrice migliore è il modello in analisi.

La ROC curve esaurisce la sua utilità nella mera visione del grafico, tuttavia da essa è possibile operare un calcolo integrale per determinare la sua area.

L'area sottesa alla curva ROC è detta **AUC** che per analogo ragionamento fatto per la curva ROC, maggiore è l'AUC migliore è il modello studiato.

Vale la regola che se un modello ha un AUC vicino alla unità significa che è eccellente, viceversa se l'AUC di un modello tende allo zero significa che sta classificando gli 0 come 1 e gli 1 come 0. Invece, quando l'AUC è 0.5, significa che il modello non ha alcuna capacità di separazione delle classi.

Inizialmete si crea la curca ROC che richiederà nella pratica l'uso del pacchetto "**ROCR**" ⁷, dal quale verranno utilizzate alcune funzioni per creare quanto in figura 4.1.

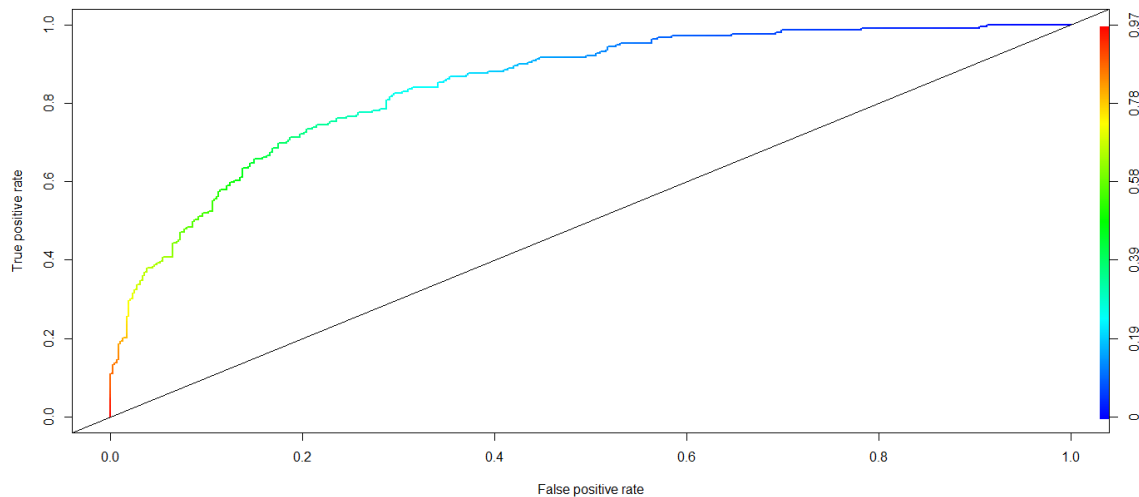


Figura 4.1: La curva ROC per il modello logit "**myglm**"

Interpretare la curva ROC non è semplice, tuttavia un primo giudizio consente di esprimere che sia abbastanza lontana dalla bisettrice. Per una accurata analisi si procede calcolando la sua area che è pari a 0.84529.

L'autrice vuole chiarire un ulteriore aspetto teorico della AUC, ovvero che essa misura la qualità delle previsioni del modello indipendentemente dalla soglia di classificazione scelta. Nel caso del modello "**myglm**", la soglia di classificazione, detta anche *cutoff* o valore critico, è stata imposta pari a 0.5.

Strettamente correlato all'analisi della curva ROC e del suo AUC sono la specificità e la sensibilità, rispettivamente al 82% e al 69%.

La **specificità** è la capacità di un test di classificare positivamente le osservazione effettivamente positivi. La **sensibilità** invece è identica per costituzione alla specificità ma in relazione ai clienti effettivamente negativi.

Maggiore è il tasso di sensibilità minore è il tasso di falsi negativi e ancora; maggiore è il tasso di specificità minore è il rischio di falsi positivi.

Questi due indici sono direttamente correlati al cutoff, questo significa che aumentando la soglia si otterranno un numero maggiore di valori negativi e quindi una maggiore specificità e una minore sensibilità.

A questo punto si comprende che un aumento della sensibilità implica una minore specificità. Si noti che vale viceversa.

⁷ROCR: visualizing classifier performance in R, T. Sing and O. Sander and N. Beerenwinkel and T. Lengauer, 2005, Bioinformatics, 21, 20, <http://rocr.bioinf.mpi-sb.mpg.de>

Prima di procedere con l'interpretazione di quanto fin'ora detto si ricorda che i clienti positivi, rientranti nella classe 1, sono *cattivi* debitori, mentre quelli in classe 0 sono i richiedenti credito *buoni*.

Se il modello è specifico all'82% significa che ha una buona capacità di classificare i clienti positivi come tali, ovvero è capace di prevedere quali prestiti non vedranno il buon fine. Si lascia al lettore l'interpretazione della sensibilità del modello "*myglm*" sulla falsa riga di quando fatto per la specificità.

4.3 Ottimizzazione del modello con variabili scelte tramite la Backward Stepwise Regression

Alla luce dei ragionamenti fatti per l'overfitting dovrebbe essere chiara l'esigenza di selezionare quelle variabili che meglio spiegano il dataframe. Tra le soluzioni possibili, in questa sede verranno approfondite quelle più conosciute nella disciplina di analisi dei dati.

La Backward selection o la backward elimination è una tecnica che consente di estrarre quelle variabili che meglio esprimono il dataset. La Backward selection parte dal modello completo, cioè con tutte le variabili e, da esso rimuove iterativamente quelle meno contributive fermandosi quando ha trovato tutte le variabili statisticamente significative.

Si usa, quindi, il modello "*myglm*", creato nel precedente paragrafo, come modello completo a cui applicare la selezione di tipo Backward. Si procede invocando la funzione *step()* nella quale deve essere specificato il modello di partenza e la direzione da seguire.

Questa funzione restituirà una serie di tabelle, alle quali corrispondono le variabili che la ricerca iterativa ritiene rilevanti. Ad ogni tabella corrisponde un AIC diverso ed è in misura minore man mano che la ricerca volge al termine, perciò vale la regola generale che bisogna prendere come migliore l'ultima selezione Backward poiché corrisponde al minore AIC.

Sotto questa logica si dovranno utilizzare le variabili selezionate nella tabella con AIC pari a 697.72.

```

1 > myglm_step=step(myglm, direction = "backward")
2 Start:  AIC=714.08
3   responsegoodcredit ~ account_status + duration +
   credit_history + purpose + creditamount + saving_account +
   present_employmentsince + InstallmentRate + sex +
   other_debtor + present_residencesince + property + age +
   other_installplans + housing + numexisting_credits + job +
   numpeople_maintenance + telephone + foreign_worker
4
5                               Df Deviance   AIC
6   - job                        3   616.62 708.62
7   - property                   3   618.21 710.21
8   .....
9   - credit_history             4   635.01 725.01
10  - account_status             3   665.38 757.38
11 [...]
12
13 Step:  AIC=698

```

```

14 responsegoodcredit ~ account_status + duration +
    credit_history + purpose + creditamount + saving_account +
    InstallmentRate + sex + other_debtor + other_installplans
    + numpeople_maintenance + telephone + foreign_worker
15
16                                Df Deviance    AIC
17      - numpeople_maintenance  1    631.72 697.72
18      <none>                                630.00 698.00
19      .....
20      - sex                      3    647.25 709.25
21      - account_status          3    684.32 746.32
22
23 Step:  AIC=697.72
24 responsegoodcredit ~ account_status + duration +
    credit_history +
25     purpose + creditamount + saving_account + InstallmentRate +
26     sex + other_debtor + other_installplans + telephone +
    foreign_worker
27
28                                Df Deviance    AIC
29      <none>                                631.72 697.72
30      - other_installplans  2    636.57 698.57
31      - foreign_worker     1    635.36 699.36
32      - telephone         1    636.70 700.70
33      - duration          1    638.39 702.39
34      - other_debtor      2    640.41 702.41
35      - creditamount      1    638.52 702.52
36      - purpose           9    655.03 703.03
37      - saving_account    4    646.14 704.14
38      - InstallmentRate   1    640.46 704.46
39      - sex               3    647.25 707.25
40      - credit_history    4    651.12 709.12
41      - account_status    3    686.21 746.21

```

Nel nostro modello di migliorato, che andremo a creare sotto il nome di "*myglm_backward*", si useranno le seguenti variabili:

- *account_status*
- *duration*
- *credit_history*
- *other_debtor*
- *purpose*
- *telephone*
- *creditamount*

- *saving_account*
- *InstallmentRate*
- *sex*
- *other_installplans*
- *foreign_worker*

Nella pratica non cambiano le modalità di attuazione del modello rispetto a quelle utilizzate nel "myglm" sennonché verranno esplicitate manualmente le variabili indipendenti rispetto alla dipendente. Il nuovo modello, "*myglm_backward*", restituisce il seguente output:

```

1      > summary(myglm_backward)
2
3      Call:
4      glm(formula = responsegoodcredit ~ +other_installplans +
5          foreign_worker +
6          telephone + duration + other_debtor + creditamount +
7          purpose +
8          saving_account + InstallmentRate + sex + credit_history +
9          account_status, family = "binomial", data = datatrain)
10
11      Deviance Residuals:
12      Min       1Q   Median       3Q      Max
13      -2.0515  -0.7086  -0.3656   0.7181   2.9138
14
15      Coefficients:
16
17      Estimate Std. Error z
18      value Pr(>|z|)
19
20      (Intercept)      1.713e+00  7.805e-01
21      2.195 0.028155 *
22      other_installplansA142 -1.468e-01  5.007e-01
23      -0.293 0.769436
24      other_installplansA143 -6.018e-01  2.837e-01
25      -2.121 0.033894 *
26      .....
27      .....
28      .....
29      ---
30      Signif. codes:  0      ***    0.001    **    0.01
31                      *    0.05    .    0.1      1
32
33      (Dispersion parameter for binomial family taken to
34      be 1)
35
36      Null deviance: 869.91  on 699  degrees of
37      freedom

```

```

26         Residual deviance: 631.72  on 667  degrees of
           freedom
27         AIC: 697.72
28
29         Number of Fisher Scoring iterations: 14

```

Si suggerisce al lettore di ricordare che l'AIC per il modello in oggetto è pari a 697.72. Sebbene ne verrà discusso nella parte dedicata alle conclusioni si anticipa che l'AIC del modello *"myglm"* è maggiore rispetto a quello per il modello *"myglm_backward"*, per ora quindi quest'ultimo risulta essere il candidato migliore tra i due modelli finora elaborati.

4.3.1 Altre considerazioni

Anche per questo modello si creerà la matrice di confusione con i suoi rispettivi indici capaci di esprimere un giudizio sulla bontà del modello. Si legga di seguito per la matrice di confusione.

```

1         > table(datatrain$responsegoodcredit, predicted_myglm_def,
2               dnn = c("Truth", "Predicted"))
3               Predicted
4         Truth    0    1
5         0  425   56
          1   96  123

```

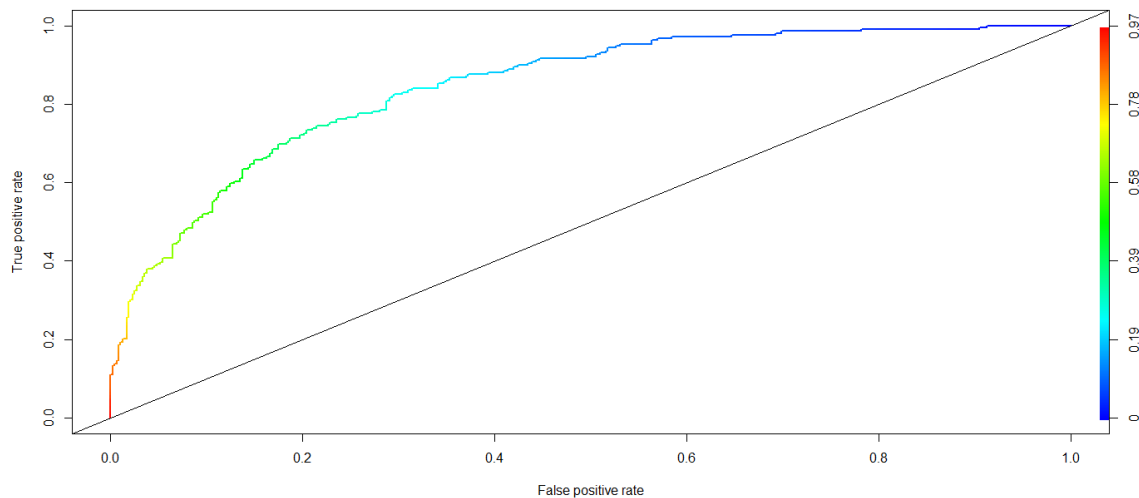
- *Veri positivi*: 123
- *Veri negativi*: 425
- *Falsi positivi*: 56
- *Falsi negativi*: 96

Da una lettura sommaria della tabella di contingenza si nota subito che è praticamente la stessa del modello *"myglm"*, il quale si ricorda essere in possesso di tutte le 21 variabili del dataset. Questo potrebbe consentire di dichiarare che il modello completo e il modello in analisi abbiano sostanzialmente stessa capacità predittiva.

Tuttavia si procede nelle stesse modalità adottate per il modello precedente al fine di trovare conferma di quanto appena detto.

Dai risultati in matrice di confusione si trova che il modello *"myglm_backward"* è sensibile all'82%, è specifico al 69%, è preciso al 88% ed infine è accurato al 78%. Questi risultati rafforzano l'ipotesi che i due modelli abbiano la stessa abilità nel predire le ipotesi.

In figura 4.2 si trova la curva ROC del modello *"myglm_backward"* la cui area, come per il modello *"myglm"* è pari a 0.84529. Si conclude che l'ottimizzazione del modello tramite la ricerca delle variabili più significative non abbia apportato una effettiva miglione in termini di competenza del modello.

Figura 4.2: La curva ROC per il modello logit *"myglm_backward"*

4.4 Applicazione del modello logit su variabili scelte tramite la Forward Stepwise Regression

La selezione in avanti ha lo stesso scopo della backward stepwise selection, ma si differenziano nelle modalità di esecuzione. Nel caso della Forward Stepwise selection si parte da un modello vuoto, senza variabili, al quale si aggiungeranno le variabili più significative una dopo l'altra. Dal punto di vista pratico sarà necessario creare un modello di partenza salvandolo sotto il nome di *"nullModel"*. A quest'ultimo si applicherà la funzione *step()* specificandone la direzione. La ricerca prende il nome di *"myglm_Foward"* e anche in questo caso si sceglierà la tabella avente l'AIC più basso, nonché l'ultima.

```

1      >
2      myglm_Foward=step(nullModel,scope=list(lower=nullModel,
3      Start: AIC=871.91
4      responsegoodcredit ~ 1
5      Df Deviance AIC
6      + account_status 3 767.41 775.41
7      + duration 1 826.24 830.24
8      ..... ..
9      <none> 869.91 871.91
10     + numexisting_credits 1 869.67 873.67
11     + numpeople_maintenance 1 869.80 873.80
12     + present_residencesince 1 869.91 873.91
13     + job 3 869.17 877.17

```

```

14 ..... ..
15 ..... ..
16
17 Step:  AIC=697.72
18      responsegoodcredit ~ account_status + duration +
19      credit_history +
20      other_debtor + purpose + saving_account + sex +
21      InstallmentRate +
22      creditamount + telephone + foreign_worker +
23      other_installplans
24
25      Df Deviance    AIC
26      <none>          631.72 697.72
27      + numpeople_maintenance 1    630.00 698.00
28      + numexisting_credits    1    630.01 698.01
29      + housing                2    628.45 698.45
30      + age                    1    631.17 699.17
31      + present_residencesince 1    631.71 699.71
32      + present_employmentsince 4    626.07 700.07
33      + property              3    629.14 701.14
34      + job                   3    631.62 703.62

```

Notiamo subito che le variabili scelte nella selezione Forward sono le stesse della ricerca di Backward, ciò significa che ci si aspettano gli stessi risultati e di conseguenza le stesse interpretazioni. Quanto appena detto è il motivo per il quale la costruzione del modello di previsione partendo da un modello nullo si conclude con questo paragrafo e si rimanda il lettore alla sezione dedicata alla Backward Stepwise Regression.

4.5 Conclusioni

Tra i modelli di previsione costruiti in questa tesi: nonché "*myglm*", "*myglm_Backward*" e "*myglm_Forward*" si preferisce quello con l'AIC minore, per i motivi già accennati precedentemente. Il modello completo ha un AIC pari a 714,08, il quale risulta essere superiore all'AIC che possiedono gli altri due modelli ottimizzati. Tanto è vero che l'AIC per "*myglm_Backward*" e "*myglm_Forward*" è pari a 679.72.

I due AIC, per il modello completo e quello per i due modelli ottimizzati, hanno una differenza di 34.36 unità.

Essendo un divario superiore a 10 si può scegliere con sicurezza il modello che contiene le variabili usate in "*myglm_Backward*" e "*myglm_Forward*", poichè esse permettono un modello migliore in termini di adattamento.

Non prevale, invece, la scelta di una metodologia particolare per selezione delle variabili rilevanti. Ne consegue che tutte le variabili contenute nel dataset "**Statlog (German Credit Data) Data Set**" siano rilevanti.

Appendice A

Codice file "lettura_dati.R"

A.1 Lettura del dataset

```
1 rm(list=ls(all=T))
2
3 dedatacredit =
  read.table("http://archive.ics.uci.edu/ml/machine-learning_
  -databases/statlog/german/german.data")
```

A.2 Assegnare un nome ad ogni osservazione

```
1
2 colnames(dedatacredit) = c("account_status", "duration",
  "credit_history", "purpose", "creditamount", "saving_account",
  "present_employmentsince", "InstallmentRate", "sex",
  "other_debtor", "present_residencesince", "property", "age",
  "other_installplans", "housing", "numexisting_credits", "job",
  "numpeople_maintenance", "telephone", "foreign_worker",
  "responsegoodcredit")
```

A.3 Trasformare la variabile risposta "responsegood-credit" in termini binari

```
1 dedatacredit$responsegoodcredit = dedatacredit$responsegoodcredit
  - 1
```

A.4 Identificare il vettore "responsegoodcredit" come un fattore

```
1 dedatacredit$responsegoodcredit <-  
2 as.factor(dedatacredit$responsegoodcredit)
```

A.5 Train e Test split

```
1 set.seed(400)  
2 test_index <- sample(nrow(dedatacredit), nrow(dedatacredit)*0.70)  
3 datatrain = dedatacredit[test_index,]  
4 datatest = dedatacredit[-test_index,]
```

A.6 Salvare le operazione del file "*germandataset.RData*"

```
1 save.image(file = "germandataset.RData")
```

Appendice B

Codice file "Descriptive-Stat4GermanDataCredit.R"

B.1 Caricamento del file d'archiviazione *germandataset.RData*

```
1 rm(list=ls(all=T))
2 load("germandataset.RData")
```

B.2 Suddividere il dataframe "datacredit" in base alla variabile risposta

```
1 splitted_good=Splitdatacredit$"1"
2 splitted_bad=Splitdatacredit$"0"
```

B.3 Studio della variabile *"sex"*

Installazione e caricamento pacchetto "epiDisplay"

```
1 Install.packages(epiDisplay)
2 library(epiDisplay)
```

B.3.1 Tabella e distribuzione di frequenza

```

1 tab1(dedatacredit$sex ,cum.percent = TRUE)

```

```

1      > tab1(dedatacredit$sex ,cum.percent = TRUE)
2      dedatacredit$sex :
3              Frequency Percent Cum. percent
4      A91             50      5.0           5.0
5      A92            310     31.0          36.0
6      A93            548     54.8          90.8
7      A94             92      9.2         100.0
8      Total          1000    100.0         100.0

```

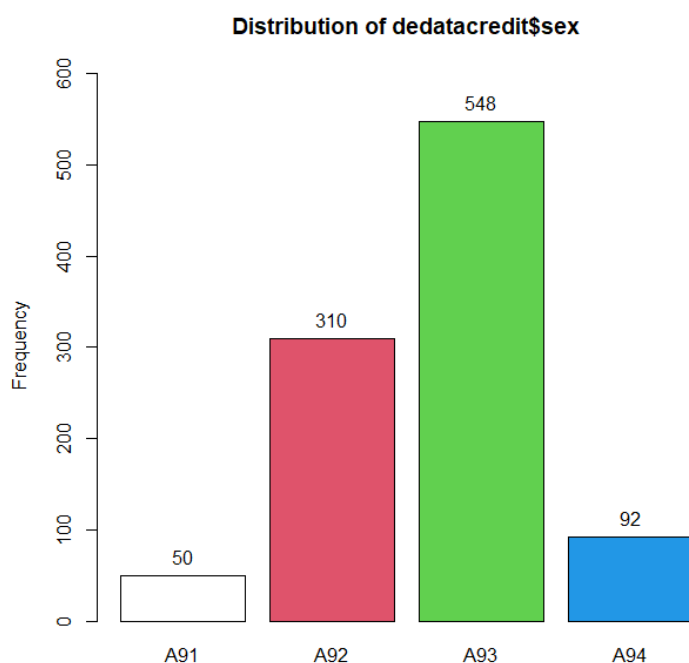


Figura B.1: Istogramma di frequenza della variabile "sex"

B.3.2 Tabella di contingenza con la variabile risposta

```

1 table(dedatacredit$responsegoodcredit , dedatacredit$sex)

```

```

1      > table(dedatacredit$responsegoodcredit ,
2              dedatacredit$sex)

```

```

2
3           A91 A92 A93 A94
4         0  30 201 402  67
5         1  20 109 146  25

```

B.4 Studio della variabile "age"

B.4.1 Tabella e distribuzione di frequenza

```

1 tab1(dedatacredit$age ,cum.percent = TRUE)

```

```

1      > tab1(dedatacredit$age ,cum.percent = TRUE)
2      dedatacredit$age :
3      Frequency Percent Cum. percent
4      19           2    0.2         0.2
5      20          14    1.4         1.6
6      21          14    1.4         3.0
7      22          27    2.7         5.7
8      23          48    4.8        10.5
9      24          44    4.4        14.9
10     25          41    4.1        19.0
11     26          50    5.0        24.0
12     27          51    5.1        29.1
13     28          43    4.3        33.4
14     29          37    3.7        37.1
15     30          40    4.0        41.1
16     31          38    3.8        44.9
17     32          34    3.4        48.3
18     33          33    3.3        51.6
19     34          32    3.2        54.8
20     35          40    4.0        58.8
21     36          39    3.9        62.7
22     37          29    2.9        65.6
23     38          24    2.4        68.0
24     39          21    2.1        70.1
25     40          25    2.5        72.6
26     41          17    1.7        74.3
27     42          22    2.2        76.5
28     43          17    1.7        78.2
29     44          17    1.7        79.9
30     45          15    1.5        81.4
31     46          18    1.8        83.2
32     47          17    1.7        84.9
33     48          12    1.2        86.1
34     49          14    1.4        87.5
35     50          12    1.2        88.7

```

36	51	8	0.8	89.5
37	52	9	0.9	90.4
38	53	7	0.7	91.1
39	54	10	1.0	92.1
40	55	8	0.8	92.9
41	56	3	0.3	93.2
42	57	9	0.9	94.1
43	58	5	0.5	94.6
44	59	3	0.3	94.9
45	60	6	0.6	95.5
46	61	7	0.7	96.2
47	62	2	0.2	96.4
48	63	8	0.8	97.2
49	64	5	0.5	97.7
50	65	5	0.5	98.2
51	66	5	0.5	98.7
52	67	3	0.3	99.0
53	68	3	0.3	99.3
54	70	1	0.1	99.4
55	74	4	0.4	99.8
56	75	2	0.2	100.0
57	Total	1000	100.0	100.0

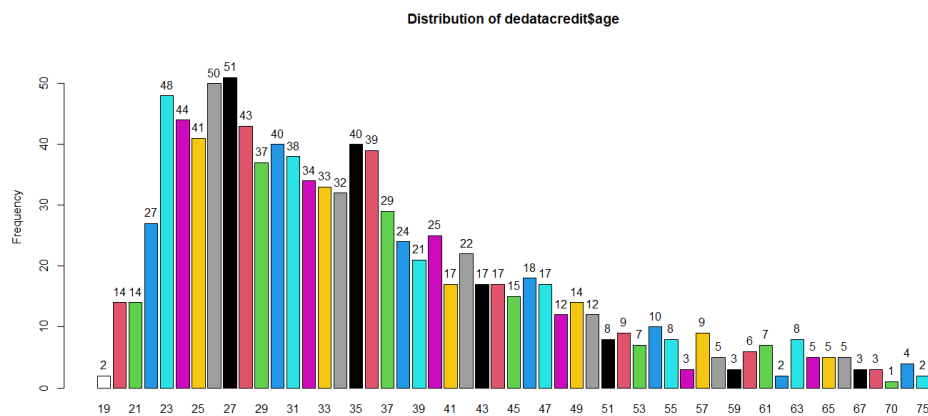


Figura B.2: Istogramma di frequenza della variabile "age"

Indice di asimmetria

Creare una funzione che sia in grado di computare la moda.

```

1 getm.ode <- function(e) {
2   unique <- unique(e)
3   unique[which.max(tabulate(match(e, unique)))]

```

```
4 }
```

Composizione dell'indica sotto la funzione chiamata *asymmetric_index*.

```
1 asymmetric_index1= function(x) {
2   ((mean(x) - getm.ode(x))/(var(x)))
3 }
4 asymmetric_index1(dedatacredit$age)

1 > asymmetric_index1(dedatacredit$age)
2 [1] 0.06604262
```

Migliore visualizzazione della distribuzione di frequenza

```
1 xAge=dedatacredit$age
2 propCurve=hist(xAge, col = "#ffd870", xlab = "Et ", main = "Curva di
  distribuzione per la variabile Age")
3 xfit_Age=seq(min(xAge), max(xAge), length= 40)
4 yfit_Age=dnorm(xfit_Age,mean=mean(xAge),sd=sd(xAge))
5 yfit_Age <- yfit_Age*diff(propCurve$mids[1:2])*length(xAge)
6 lines(xfit_Age, yfit_Age, col="black", lwd=6)
```

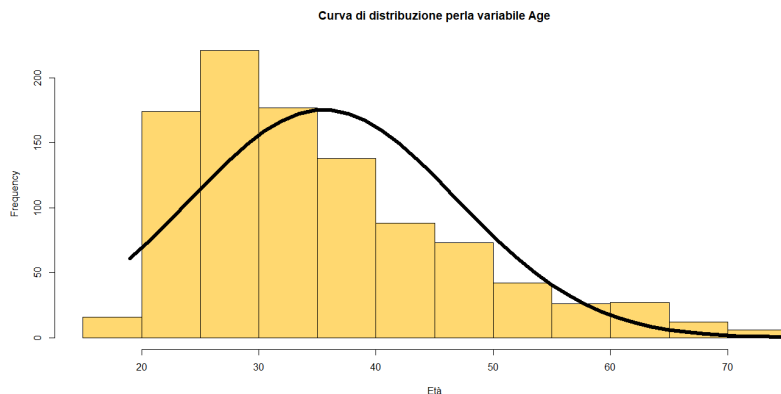


Figura B.3: Elegante visualizzazione della frequenza della variabile "age"

B.4.2 Tabella di contingenza con la variabile risposta

```
1 table(dedatacredit$responsegoodcredit , dedatacredit$age)

1 > table(dedatacredit$age ,
2         dedatacredit$responsegoodcredit)
```

3		0	1
4	19	1	1
5	20	9	5
6	21	9	5
7	22	16	11
8	23	28	20
9	24	25	19
10	25	22	19
11	26	36	14
12	27	38	13
13	28	28	15
14	29	22	15
15	30	29	11
16	31	27	11
17	32	25	9
18	33	20	13
19	34	21	11
20	35	34	6
21	36	33	6
22	37	21	8
23	38	20	4
24	39	15	6
25	40	19	6
26	41	13	4
27	42	14	8
28	43	12	5
29	44	12	5
30	45	12	3
31	46	14	4
32	47	12	5
33	48	9	3
34	49	13	1
35	50	9	3
36	51	7	1
37	52	8	1
38	53	2	5
39	54	8	2
40	55	5	3
41	56	3	0
42	57	6	3
43	58	3	2
44	59	2	1
45	60	3	3
46	61	4	3
47	62	2	0
48	63	7	1
49	64	5	0
50	65	4	1
51	66	3	2

```

52           67  3  0
53           68  1  2
54           70  1  0
55           74  3  1
56           75  2  0

```

B.4.3 Tabella di frequenza per per i due dataset divisi per la variabile risposta

```

1  tab1(splitted_bad$sex ,cum.percent = TRUE)

```

```

1  > tab1(splitted_bad$age ,cum.percent = TRUE)
2  splitted_bad$age :
3      Frequency Percent Cum. percent
4  19           1     0.1         0.1
5  20           9     1.3         1.4
6  21           9     1.3         2.7
7  22          16     2.3         5.0
8  23          28     4.0         9.0
9  24          25     3.6        12.6
10 25          22     3.1        15.7
11 26          36     5.1        20.9
12 27          38     5.4        26.3
13 28          28     4.0        30.3
14 29          22     3.1        33.4
15 30          29     4.1        37.6
16 31          27     3.9        41.4
17 32          25     3.6        45.0
18 33          20     2.9        47.9
19 34          21     3.0        50.9
20 35          34     4.9        55.7
21 36          33     4.7        60.4
22 37          21     3.0        63.4
23 38          20     2.9        66.3
24 39          15     2.1        68.4
25 40          19     2.7        71.1
26 41          13     1.9        73.0
27 42          14     2.0        75.0
28 43          12     1.7        76.7
29 44          12     1.7        78.4
30 45          12     1.7        80.1
31 46          14     2.0        82.1
32 47          12     1.7        83.9
33 48           9     1.3        85.1
34 49          13     1.9        87.0
35 50           9     1.3        88.3
36 51           7     1.0        89.3

```

37	52	8	1.1	90.4
38	53	2	0.3	90.7
39	54	8	1.1	91.9
40	55	5	0.7	92.6
41	56	3	0.4	93.0
42	57	6	0.9	93.9
43	58	3	0.4	94.3
44	59	2	0.3	94.6
45	60	3	0.4	95.0
46	61	4	0.6	95.6
47	62	2	0.3	95.9
48	63	7	1.0	96.9
49	64	5	0.7	97.6
50	65	4	0.6	98.1
51	66	3	0.4	98.6
52	67	3	0.4	99.0
53	68	1	0.1	99.1
54	70	1	0.1	99.3
55	74	3	0.4	99.7
56	75	2	0.3	100.0
57	Total	700	100.0	100.0

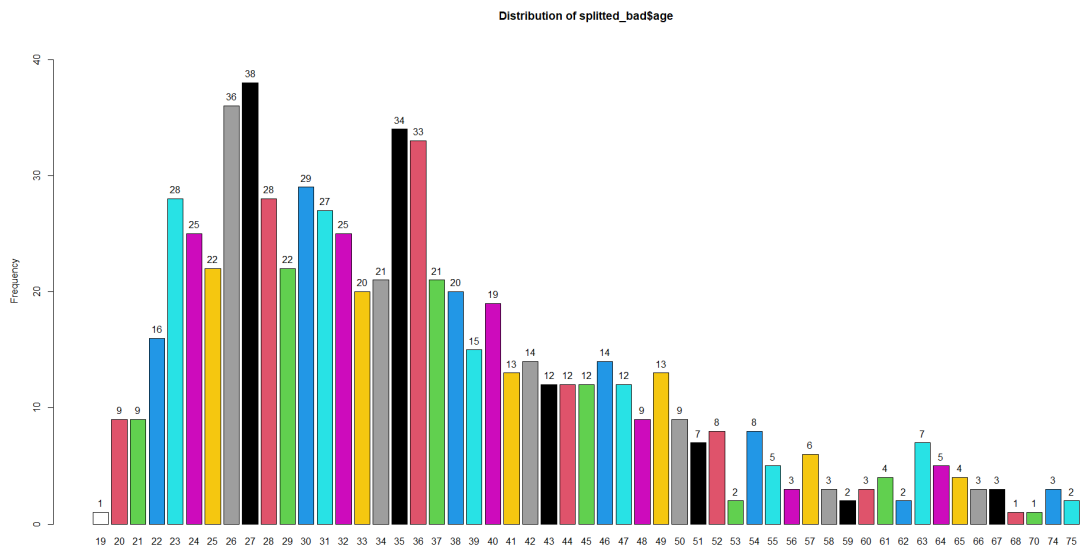


Figura B.4: Istogramma di frequenza della variabile "age" in "splitted_bad"

Si procede allo stesso modo per il dataset "splitted_good".

B.5 Studio della variabile "*duration*"

B.5.1 Tabella e distribuzione di frequenza

```

1 tab1(dedatacredit$duration ,cum.percent = TRUE)

```

```

1      > tab1(dedatacredit$duration ,cum.percent = TRUE)
2      dedatacredit$duration :
3      Frequency Percent Cum. percent
4      4           6     0.6         0.6
5      5           1     0.1         0.7
6      6          75     7.5         8.2
7      7           5     0.5         8.7
8      8           7     0.7         9.4
9      9          49     4.9        14.3
10     10          28     2.8        17.1
11     11           9     0.9        18.0
12     12         179    17.9        35.9
13     13           4     0.4        36.3
14     14           4     0.4        36.7
15     15          64     6.4        43.1
16     16           2     0.2        43.3
17     18         113    11.3        54.6
18     20           8     0.8        55.4
19     21          30     3.0        58.4
20     22           2     0.2        58.6
21     24         184    18.4        77.0
22     26           1     0.1        77.1
23     27          13     1.3        78.4
24     28           3     0.3        78.7
25     30          40     4.0        82.7
26     33           3     0.3        83.0
27     36          83     8.3        91.3
28     39           5     0.5        91.8
29     40           1     0.1        91.9
30     42          11     1.1        93.0
31     45           5     0.5        93.5
32     47           1     0.1        93.6
33     48          48     4.8        98.4
34     54           2     0.2        98.6
35     60          13     1.3        99.9
36     72           1     0.1       100.0
37      Total      1000   100.0       100.0

```

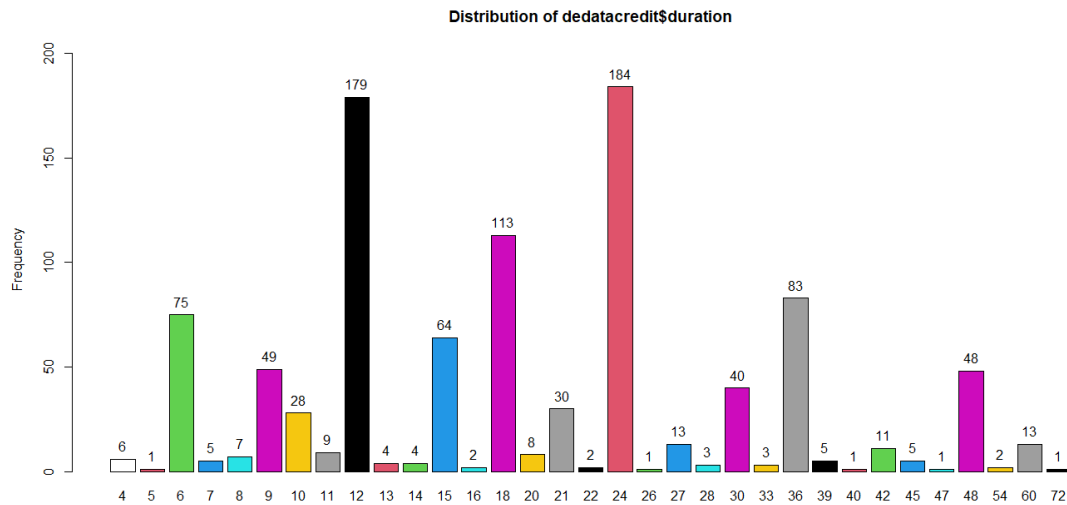


Figura B.5: Istogramma di frequenza della variabile "age" in "splitted_bad"

Indice di asimmetria

```

1  asymetric_index1(dedatacredit$age)

1  > asymetric_index1(dedatacredit$duration)
2  [1] 0.02129766

```

Migliore visualizzazione della distribuzione di frequenza

```

1  xduration=dedatacredit$duration
2  propCurve=hist(xduration, col = "#ffd870", xlab = "Durata del
   credito", main = "Curva di distribuzione per la variabile
   Duration")
3  xfit_duration=seq(min(xduration), max(xduration), length= 40)
4  yfit_duration=dnorm(xfit_Age, mean=mean(xduration), sd=sd(xduration))
5  yfit_duration <-
   yfit_duration*diff(propCurve$mids[1:2])*length(xduration)
6  lines(xfit_duration, yfit_duration, col="black", lwd=6)

```

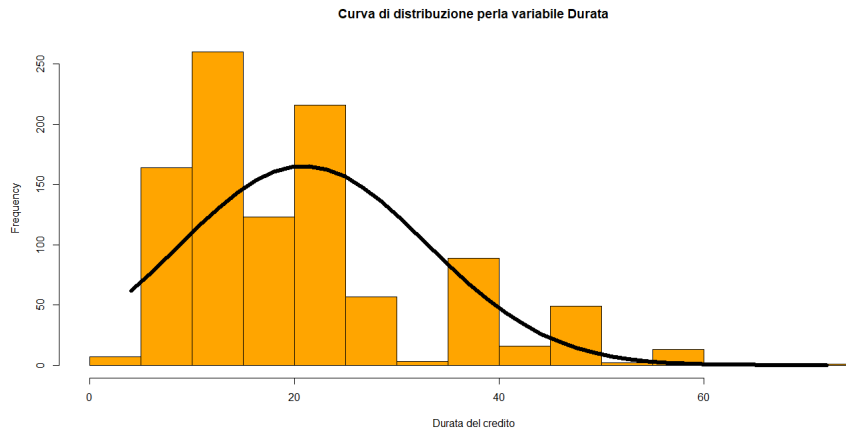


Figura B.6: Elegante visualizzazione della frequenza della variabile "duration"

B.5.2 Tabella di contingenza con la variabile risposta

```
1 table(dedatacredit$duration , dedatacredit$responsegoodcredit)
```

```
1 > table(dedatacredit$duration ,
2         dedatacredit$responsegoodcredit)
3           0    1
4    4      6    0
5    5      1    0
6    6     66    9
7    7      5    0
8    8      6    1
9    9     35   14
10   10     25    3
11   11      9    0
12   12    130   49
13   13      4    0
14   14      3    1
15   15     52   12
16   16      1    1
17   18     71   42
18   20      7    1
19   21     21    9
20   22      2    0
21   24    128   56
22   26      1    0
23   27      8    5
24   28      2    1
25   30     27   13
26   33      2    1
```

27	36	46	37
28	39	4	1
29	40	0	1
30	42	8	3
31	45	1	4
32	47	1	0
33	48	20	28
34	54	1	1
35	60	7	6
36	72	0	1

B.5.3 Tabella di frequenza per i due dataset divisi per la variabile risposta

```
1 tab1(splitted_bad$duration ,cum.percent = TRUE)
```

```
1 > tab1(splitted_bad$duration ,cum.percent = TRUE)
2 splitted_bad$duration :
3 Frequency Percent Cum. percent
4 4 6 0.9 0.9
5 5 1 0.1 1.0
6 6 66 9.4 10.4
7 7 5 0.7 11.1
8 8 6 0.9 12.0
9 9 35 5.0 17.0
10 10 25 3.6 20.6
11 11 9 1.3 21.9
12 12 130 18.6 40.4
13 13 4 0.6 41.0
14 14 3 0.4 41.4
15 15 52 7.4 48.9
16 16 1 0.1 49.0
17 18 71 10.1 59.1
18 20 7 1.0 60.1
19 21 21 3.0 63.1
20 22 2 0.3 63.4
21 24 128 18.3 81.7
22 26 1 0.1 81.9
23 27 8 1.1 83.0
24 28 2 0.3 83.3
25 30 27 3.9 87.1
26 33 2 0.3 87.4
27 36 46 6.6 94.0
28 39 4 0.6 94.6
29 42 8 1.1 95.7
30 45 1 0.1 95.9
31 47 1 0.1 96.0
```

32	48	20	2.9	98.9
33	54	1	0.1	99.0
34	60	7	1.0	100.0
35	Total	700	100.0	100.0

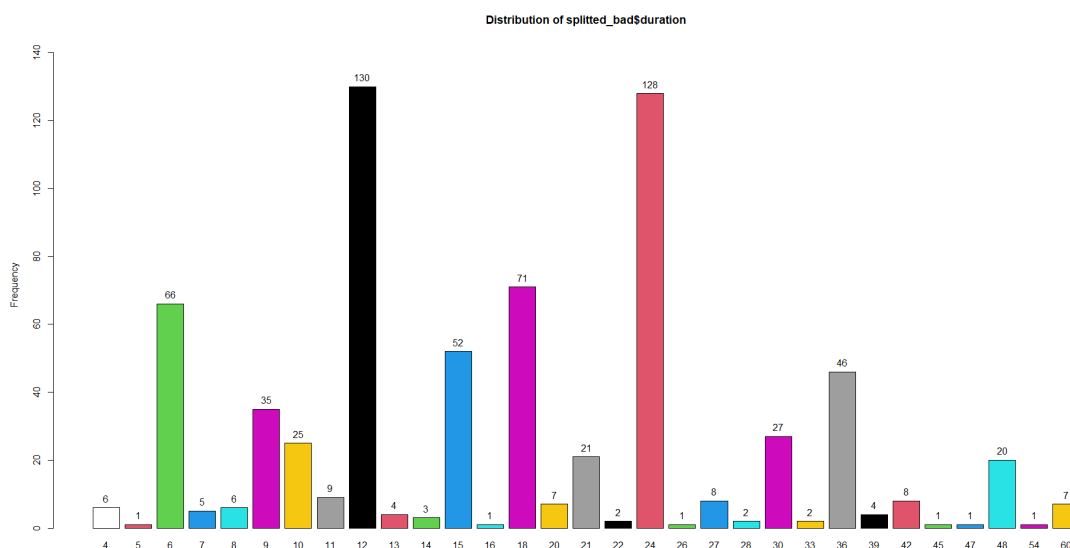


Figura B.7: Istogramma di frequenza della variabile "duration" in "splitted_bad"

Indice di asimmetria per i due dataset divisi per la variabile risposta

```

1 asymeric_index1(splitted_bad$duration)
2 asymeric_index1(splitted_good$duration)

```

```

1 > asymeric_index1(splitted_bad$duration)
2 [1] 0.05871077
3 > asymeric_index1(splitted_good$duration)
4 [1] 0.004874496

```

B.6 Studio della variabile "purpose"

B.6.1 Tabella e distribuzione di frequenza

```

1 tab1(dedatacredit$purpose ,cum.percent = TRUE)

```

```

1      > tab1(dedatacredit$purpose ,cum.percent = TRUE)
2      dedatacredit$purpose :
3          Frequency Percent Cum. percent
4      A40          234    23.4      23.4
5      A41          103    10.3      33.7
6      A410          12     1.2      34.9
7      A42          181    18.1      53.0
8      A43          280    28.0      81.0
9      A44           12     1.2      82.2
10     A45           22     2.2      84.4
11     A46           50     5.0      89.4
12     A48           9     0.9      90.3
13     A49           97     9.7     100.0
14     Total        1000   100.0     100.0

```

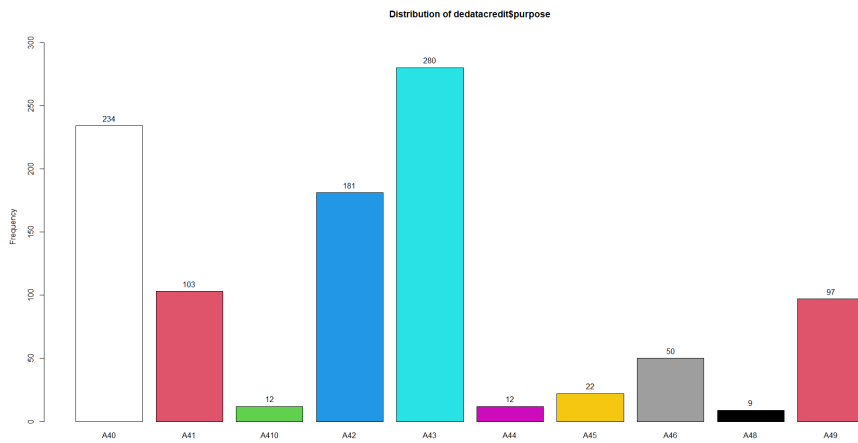


Figura B.8: Isogramma di frequenza della variabile "purpose"

B.6.2 Tabella di contingenza con la variabile risposta

```

1 table(dedatacredit$purpose , dedatacredit$responsegoodcredit)

```

```

1      > table(dedatacredit$purpose ,
2              dedatacredit$responsegoodcredit)
3
4              0    1
5      A40    145  89
6      A41     86  17
7      A410     7   5
8      A42    123  58
9      A43    218  62

```


9	A44	8	4
10	A45	14	8
11	A46	28	22
12	A48	8	1
13	A49	63	34

B.7 Studio di più variabili contemporaneamente

B.7.1 Caricamento del dataset in versione numerica

```
1
2 numeric_dedatacredit=
  read.table("http://archive.ics.uci.edu/ml/machine-learning_
  -databases/statlog/german/german.data-numeric")
3
4 colnames(numeric_dedatacredit) = c("account_status", "duration",
  "credit_history", "purpose", "creditamount", "saving_account",
  "present_employmentsince", "InstallmentRate", "sex",
  "other_debtor",
5                                     "present_residencesince", "property",
                                     "age", "other_installplans",
                                     "housing", "numexisting_credits",
6                                     "job", "numpeople_maintenance",
                                     "telephone", "foreign_worker",
                                     "responsegoodcredit")
7
8 numeric_dedatacredit=numeric_dedatacredit[-rev(seq_len(ncol
  (numeric_dedatacredit)))-21]
9
10 numSplitdatacredit= split(numeric_dedatacredit,
  numeric_dedatacredit$responsegoodcredit)
11 numsplitted_bad=numSplitdatacredit$"0"
12 numsplitted_good=numSplitdatacredit$"1"
```

B.7.2 Creazione della matrice di correlazione

```
1 cormatrix= cor(numeric_dedatacredit)
2 round(cormatrix,3)
```

B.7.3 Creazione del Correlogram

Installazione e caricamento dei pacchetti "corrplot"¹ e RColorBrewer².

```
1 install.packages(c(corrplot,RColorBrewer)
2 library(c(corrplot,RColorBrewer))
```

```
1 corrplot(cormatrix, type="upper", order="hclust",
2          col=brewer.pal(n=8, name="PuOr"))
```

B.7.4 Tabella di contiggenza per le varaibili "credit_history" e "property"

```
1 table(dedatacredit$credit_history , dedatacredit$property)
```

```
1 > table(dedatacredit$credit_history ,
2         dedatacredit$property)
```

	A121	A122	A123	A124
A30	5	10	17	8
A31	11	8	15	15
A32	158	122	179	71
A33	19	23	31	15
A34	89	69	90	45

¹Corrplot2021, R package "corrplot": Visualization of a Correlation Matrix, Taiyun Wei and Viliam Simko, 2021, Version 0.89, <https://github.com/taiyun/corrplot>.

²4. 1.1-2, R (2.0.0), 2014-12-07, Erich Neuwirth [aut, cre], Erich Neuwirth <erich.neuwirth at univie.ac.at>, Apache License 2.0, Graphics, Spatial, RColorBrewer results.

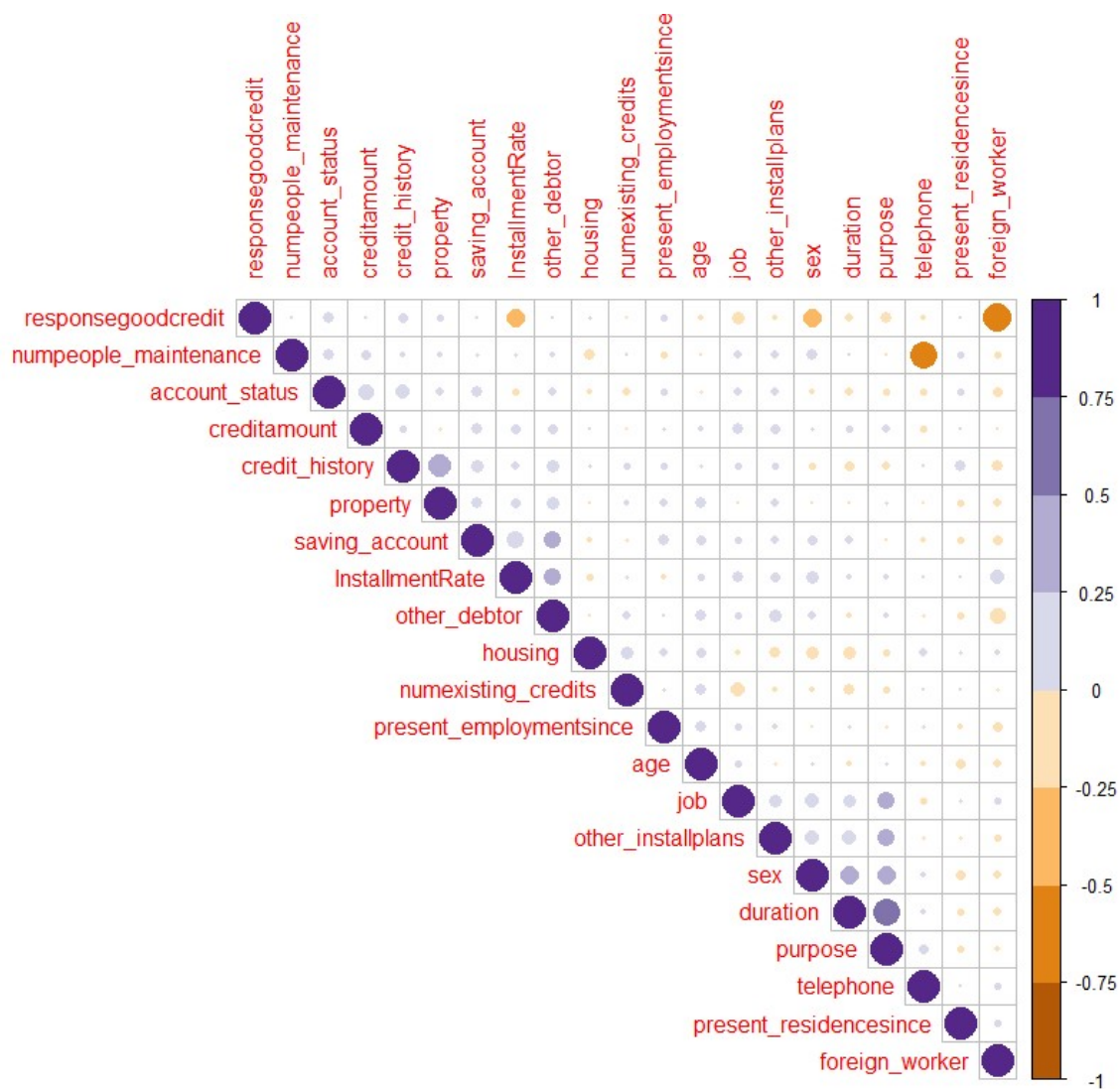


Figura B.9: Correlogram per tutte le variabili all'interno del dataset "Statlog (German Credit Data) Data Set" in versione numerica

Appendice C

Codice file "RegressionModels.R "

C.1 *glm()* con il modello completo

```
1 myglm=glm(responsegoodcredit~.,data=datatrain,family = "binomial")
2 summary(myglm)
```

```
1 > myglm=glm(responsegoodcredit~.,data=datatrain,family
2   = "binomial")
3   > summary(myglm)
4
5   Call:
6   glm(formula = responsegoodcredit ~ ., family =
7     "binomial", data = datatrain)
8
9   Deviance Residuals:
10      Min       1Q   Median       3Q      Max
11    -2.0219  -0.6990  -0.3431   0.6794   2.8247
12
13  Coefficients:
14              Estimate Std. Error z
15      value Pr(>|z|)
16  (Intercept)      9.404e-01  1.279e+00
17      0.736  0.46200
18  account_statusA12    -5.133e-01  2.660e-01
19      -1.929  0.05369 .
20  account_statusA13    -1.383e+00  4.837e-01
21      -2.859  0.00424 **
22  account_statusA14    -1.779e+00  2.785e-01
23      -6.387 1.69e-10 ***
24  duration            3.049e-02  1.147e-02
25      2.657  0.00788 **
```

18	credit_historyA31	-2.846e-01	7.015e-01
	-0.406 0.68495		
19	credit_historyA32	-1.226e+00	5.343e-01
	-2.295 0.02175 *		
20	credit_historyA33	-1.025e+00	5.812e-01
	-1.763 0.07789 .		
21	credit_historyA34	-1.927e+00	5.482e-01
	-3.515 0.00044 ***		
22	purposeA41	-1.444e+00	4.396e-01
	-3.285 0.00102 **		
23	purposeA410	-8.738e-01	1.038e+00
	-0.842 0.40002		
24	purposeA42	-6.132e-01	3.152e-01
	-1.946 0.05171 .		
25	purposeA43	-9.572e-01	3.001e-01
	-3.190 0.00142 **		
26	purposeA44	-4.014e-01	1.006e+00
	-0.399 0.68980		
27	purposeA45	-5.769e-01	6.324e-01
	-0.912 0.36169		
28	purposeA46	6.100e-02	4.914e-01
	0.124 0.90120		
29	purposeA48	-1.418e+01	6.707e+02
	-0.021 0.98313		
30	purposeA49	-1.063e+00	4.210e-01
	-2.525 0.01158 *		
31	creditamount	1.263e-04	5.498e-05
	2.297 0.02161 *		
32	saving_accountA62	-3.028e-01	3.640e-01
	-0.832 0.40542		
33	saving_accountA63	6.503e-02	4.504e-01
	0.144 0.88520		
34	saving_accountA64	-1.182e+00	5.967e-01
	-1.981 0.04764 *		
35	saving_accountA65	-8.550e-01	3.194e-01
	-2.677 0.00742 **		
36	present_employmentsinceA72	-6.958e-03	5.229e-01
	-0.013 0.98938		
37	present_employmentsinceA73	-3.299e-01	5.097e-01
	-0.647 0.51751		
38	present_employmentsinceA74	-7.903e-01	5.518e-01
	-1.432 0.15205		
39	present_employmentsinceA75	-1.450e-01	5.243e-01
	-0.277 0.78213		
40	InstallmentRate	3.052e-01	1.109e-01
	2.753 0.00591 **		
41	sexA92	-3.285e-01	4.644e-01
	-0.707 0.47942		

```

42      sexA93      -1.123e+00  4.652e-01
      -2.414  0.01577 *
43      sexA94      -4.809e-01  5.467e-01
      -0.880  0.37905
44      other_debtorA102      8.460e-01  5.282e-01
      1.602  0.10923
45      other_debtorA103      -8.772e-01  5.520e-01
      -1.589  0.11202
46      present_residencesince      -6.262e-02  1.059e-01
      -0.591  0.55423
47      propertyA122      3.493e-01  3.084e-01
      1.132  0.25743
48      propertyA123      3.570e-01  2.903e-01
      1.230  0.21881
49      propertyA124      5.873e-01  5.456e-01
      1.076  0.28176
50      age      -9.307e-03  1.135e-02
      -0.820  0.41232
51      other_installplansA142      -9.000e-02  5.134e-01
      -0.175  0.86086
52      other_installplansA143      -5.048e-01  2.930e-01
      -1.723  0.08489 .
53      housingA152      -4.599e-01  2.891e-01
      -1.591  0.11162
54      housingA153      -3.446e-01  6.012e-01
      -0.573  0.56657
55      numexisting_credits      3.284e-01  2.309e-01
      1.422  0.15490
56      jobA172      5.475e-01  7.717e-01
      0.710  0.47799
57      jobA173      4.382e-01  7.365e-01
      0.595  0.55189
58      jobA174      4.115e-01  7.443e-01
      0.553  0.58032
59      numpeople_maintenance      4.114e-01  3.049e-01
      1.349  0.17724
60      telephoneA192      -5.003e-01  2.460e-01
      -2.033  0.04202 *
61      foreign_workerA202      -1.465e+00  8.306e-01
      -1.764  0.07778 .
62      ---
63      Signif. codes:  0      ***      0.001      **      0.01
      *      0.05      .      0.1      1
64
65      (Dispersion parameter for binomial family taken to
      be 1)
66
67      Null deviance: 869.91  on 699  degrees of
      freedom

```

```
68           Residual deviance: 616.08  on 651  degrees of
           freedom
69           AIC: 714.08
70
71           Number of Fisher Scoring iterations: 14
```

C.1.1 Matrice di confusione per il modello completo

```
1 cutoff= 0.5
2 prob_myglm=predict(myglm, type = "response")
3 predicted_myglm=prob_myglm>cutoff
4 predicted_myglm=as.numeric(predicted_myglm)
5
6 confusion_table=table(datatrain$responsegoodcredit,
       predicted_myglm, dnn = c("Truth", "Predicted"))
7 confusion_matrix=as.matrix(confusion_table)
8 confusion_matrix
```

```
1           > confusion_matrix
2               Predicted
3           Truth    0    1
4               0 425   56
5               1  92 127
```

C.1.2 Alcuni indici di valutazione

```
1 sensitivity=  confusion_matrix[1,1] / (confusion_matrix[1,1] +
       confusion_matrix[2,1])
2 print(sensitivity )
3
4 Specificity = confusion_matrix[2,2] / (confusion_matrix[2,2] +
       confusion_matrix[1,2])
5 print(Specificity)
6
7 precision = confusion_matrix[1,1] / (confusion_matrix[1,1] +
       confusion_matrix[1,2])
8 print(precision)
9
10 accuracy = (confusion_matrix[1,1] + confusion_matrix[2,2]) /
       (confusion_matrix[1,1] +confusion_matrix[1,2]+
       confusion_matrix[2,1] + confusion_matrix[2,2])
11 print(accuracy)
```

```
1           > sensitivity=  confusion_matrix[1,1] /
       (confusion_matrix[1,1] + confusion_matrix[2,1])
```

```
2         > print(sensitivity )
3         [1] 0.8220503
4     > Specificity = confusion_matrix[2,2] /
5         (confusion_matrix[2,2] + confusion_matrix[1,2])
6         > print(Specificity)
7         [1] 0.6939891
8     > precision = confusion_matrix[1,1]/(confusion_matrix[1,1]
9         + confusion_matrix[1,2])
10        > print(precision)
11        [1] 0.8835759
12    > accuracy = (confusion_matrix[1,1] +
13        confusion_matrix[2,2]) / (confusion_matrix[1,1] +
14        confusion_matrix[1,2] + confusion_matrix[2,1] +
15        confusion_matrix[2,2])
16        > print(accuracy)
17        [1] 0.7885714
```

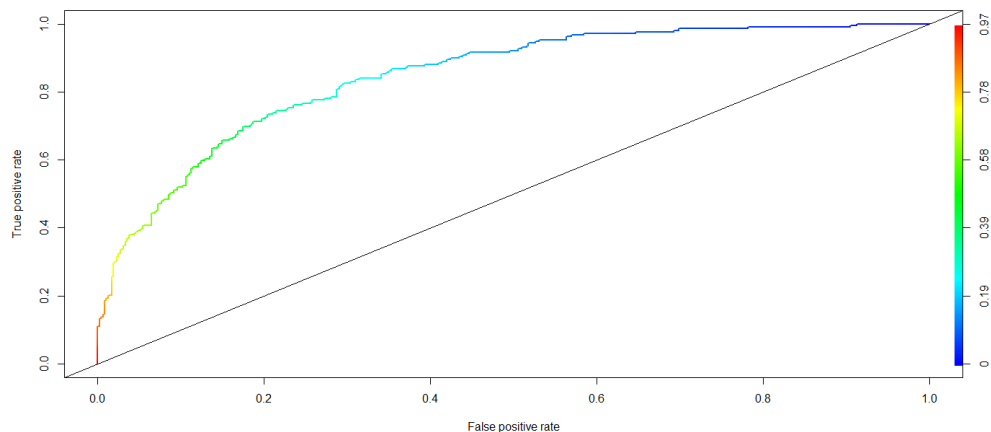
C.1.3 Curva ROC e AUC.

Installazione e caricamento pacchetto "**ROCR**"

```
1  install.packages(ROCR)
2  library(ROCR)

1  pred_myglm=predict(myglm, type = "response")
2  pred_myglm=prediction(pred_myglm, datatrain$responsegoodcredit)

1  ROC= performance(pred_myglm,"tpr","fpr")
2  plot(ROC, colorize = T, lwd = 2)
3  abline(a = 0, b = 1)
```

Figura C.1: La curva ROC per il modello logit "*myglm*"

```
1 auc = performance(pred_myglm, measure = "auc")
2 print(auc@y.values)
```

```
1 > auc = performance(pred_myglm, measure = "auc")
2 > print(auc@y.values)
3 [[1]]
4 [1] 0.84529
```

C.2 *glm()* con le variabili selezione tramite la Backward Stepwise Regression

```
1 n = length(resid(myglm))
2 myglm_step=step(myglm, direction = "backward")
```

```
1 > n = length(resid(myglm))
2 > myglm_step=step(myglm, direction = "backward")
3 Start: AIC=714.08
4 responsegoodcredit ~ account_status + duration + credit_history +
5   purpose + creditamount + saving_account +
6   present_employmentsince +
7   InstallmentRate + sex + other_debtor + present_residencesince +
8   property + age + other_installplans + housing +
9   numexisting_credits +
10  job + numpeople_maintenance + telephone + foreign_worker
```

	Df	Deviance	AIC
--	----	----------	-----

```

11 - job 3 616.62 708.62
12 - property 3 618.21 710.21
13 - present_employmentsince 4 621.56 711.56
14 - present_residencesince 1 616.43 712.43
15 - housing 2 618.60 712.60
16 - age 1 616.76 712.76
17 - other_installplans 2 619.34 713.34
18 - numpeople_maintenance 1 617.89 713.89
19 <none> 616.08 714.08
20 - numexisting_credits 1 618.13 714.13
21 - other_debtor 2 621.75 715.75
22 - foreign_worker 1 620.09 716.09
23 - telephone 1 620.29 716.29
24 - creditamount 1 621.38 717.38
25 - saving_account 4 627.77 717.77
26 - purpose 9 638.27 718.27
27 - duration 1 623.25 719.25
28 - InstallmentRate 1 623.88 719.88
29 - sex 3 628.28 720.28
30 - credit_history 4 635.01 725.01
31 - account_status 3 665.38 757.38
32
33 Step: AIC=708.62
34 responsegoodcredit ~ account_status + duration + credit_history +
35   purpose + creditamount + saving_account +
36   present_employmentsince +
37   InstallmentRate + sex + other_debtor + present_residencesince +
38   property + age + other_installplans + housing +
39   numexisting_credits +
40   numpeople_maintenance + telephone + foreign_worker
41
42 Df Deviance AIC
43 - property 3 618.66 704.66
44 - present_employmentsince 4 621.89 705.89
45 - present_residencesince 1 616.88 706.88
46 - housing 2 619.07 707.07
47 - age 1 617.25 707.25
48 - other_installplans 2 620.22 708.22
49 - numexisting_credits 1 618.45 708.45
50 - numpeople_maintenance 1 618.55 708.55
51 <none> 616.62 708.62
52 - other_debtor 2 622.21 710.21
53 - foreign_worker 1 620.54 710.54
54 - telephone 1 621.73 711.73
55 - creditamount 1 622.20 712.20
56 - purpose 9 638.85 712.85
57 - saving_account 4 628.91 712.91
58 - duration 1 623.75 713.75
59 - InstallmentRate 1 624.69 714.69

```

```

58 - sex                      3    628.79  714.79
59 - credit_history          4    635.51  719.51
60 - account_status          3    665.96  751.96
61
62 Step:  AIC=704.66
63 responsegoodcredit ~ account_status + duration + credit_history +
64   purpose + creditamount + saving_account +
65   present_employmentsince +
66   InstallmentRate + sex + other_debtor + present_residencesince +
67   age + other_installplans + housing + numexisting_credits +
68   numpeople_maintenance + telephone + foreign_worker
69
70 - present_employmentsince  4    624.40  702.40
71 - present_residencesince  1    618.80  702.80
72 - housing                  2    621.32  703.32
73 - age                      1    619.55  703.55
74 - numexisting_credits      1    620.39  704.39
75 - numpeople_maintenance   1    620.47  704.47
76 - other_installplans      2    622.66  704.66
77 <none>                     1    618.66  704.66
78 - foreign_worker          1    622.46  706.46
79 - telephone               1    623.01  707.01
80 - other_debtor            2    625.23  707.23
81 - creditamount            1    624.84  708.84
82 - saving_account          4    631.22  709.22
83 - purpose                  9    641.47  709.47
84 - sex                     3    630.68  710.68
85 - duration                 1    626.81  710.81
86 - InstallmentRate         1    627.63  711.63
87 - credit_history          4    637.82  715.82
88 - account_status          3    669.21  749.21
89
90 Step:  AIC=702.4
91 responsegoodcredit ~ account_status + duration + credit_history +
92   purpose + creditamount + saving_account + InstallmentRate +
93   sex + other_debtor + present_residencesince + age +
94   other_installplans +
95   housing + numexisting_credits + numpeople_maintenance +
96   telephone +
97   foreign_worker
98 - present_residencesince  1    624.56  700.56
99 - age                     1    624.99  700.99
100 - housing                 2    627.54  701.54
101 - numexisting_credits     1    626.08  702.08
102 - numpeople_maintenance  1    626.10  702.10
103 <none>                    1    624.40  702.40

```

```

104 - other_installplans      2    629.04  703.04
105 - foreign_worker         1    628.44  704.44
106 - telephone              1    629.04  705.04
107 - purpose                9    646.34  706.34
108 - other_debtor           2    632.56  706.56
109 - creditamount           1    630.75  706.75
110 - saving_account         4    637.28  707.28
111 - duration                1    631.32  707.32
112 - InstallmentRate        1    634.56  710.56
113 - sex                    3    638.95  710.95
114 - credit_history         4    642.86  712.86
115 - account_status         3    676.05  748.05
116
117 Step:   AIC=700.56
118 responsegoodcredit ~ account_status + duration + credit_history +
119     purpose + creditamount + saving_account + InstallmentRate +
120     sex + other_debtor + age + other_installplans + housing +
121     numexisting_credits + numpeople_maintenance + telephone +
122     foreign_worker
123
124               Df Deviance    AIC
125 - age          1    625.32  699.32
126 - housing      2    627.55  699.55
127 - numexisting_credits 1    626.18  700.18
128 - numpeople_maintenance 1    626.24  700.24
129 <none>                624.56  700.56
130 - other_installplans  2    629.22  701.22
131 - foreign_worker     1    628.55  702.55
132 - telephone          1    629.30  703.30
133 - purpose            9    646.59  704.59
134 - other_debtor       2    632.73  704.73
135 - creditamount       1    631.14  705.14
136 - duration           1    631.40  705.40
137 - saving_account     4    637.57  705.57
138 - InstallmentRate    1    634.65  708.65
139 - sex                3    639.26  709.26
140 - credit_history     4    643.25  711.25
141 - account_status     3    676.05  746.05
142
143 Step:   AIC=699.32
144 responsegoodcredit ~ account_status + duration + credit_history +
145     purpose + creditamount + saving_account + InstallmentRate +
146     sex + other_debtor + other_installplans + housing +
147     numexisting_credits +
148     numpeople_maintenance + telephone + foreign_worker
149
149               Df Deviance    AIC
150 - housing          2    628.40  698.40
151 - numexisting_credits 1    626.77  698.77

```

```

152 - numpeople_maintenance 1 626.90 698.90
153 <none> 625.32 699.32
154 - other_installplans 2 629.72 699.72
155 - foreign_worker 1 629.30 701.30
156 - telephone 1 630.67 702.67
157 - purpose 9 647.35 703.35
158 - other_debtor 2 633.36 703.36
159 - creditamount 1 632.09 704.09
160 - duration 1 632.43 704.43
161 - saving_account 4 638.67 704.67
162 - InstallmentRate 1 635.14 707.14
163 - sex 3 640.66 708.66
164 - credit_history 4 644.55 710.55
165 - account_status 3 677.18 745.18
166
167 Step: AIC=698.4
168 responsegoodcredit ~ account_status + duration + credit_history +
169     purpose + creditamount + saving_account + InstallmentRate +
170     sex + other_debtor + other_installplans + numexisting_credits +
171     numpeople_maintenance + telephone + foreign_worker
172
173               Df Deviance    AIC
174 - numexisting_credits 1 630.00 698.00
175 - numpeople_maintenance 1 630.01 698.01
176 <none> 628.40 698.40
177 - other_installplans 2 632.45 698.45
178 - foreign_worker 1 632.23 700.23
179 - telephone 1 633.75 701.75
180 - purpose 9 650.85 702.85
181 - other_debtor 2 637.32 703.32
182 - creditamount 1 635.42 703.42
183 - duration 1 635.43 703.43
184 - saving_account 4 642.29 704.29
185 - InstallmentRate 1 637.89 705.89
186 - sex 3 646.07 710.07
187 - credit_history 4 649.06 711.06
188 - account_status 3 683.00 747.00
189
190 Step: AIC=698
191 responsegoodcredit ~ account_status + duration + credit_history +
192     purpose + creditamount + saving_account + InstallmentRate +
193     sex + other_debtor + other_installplans +
194     numpeople_maintenance +
195     telephone + foreign_worker
196
196               Df Deviance    AIC
197 - numpeople_maintenance 1 631.72 697.72
198 <none> 630.00 698.00
199 - other_installplans 2 634.73 698.73

```

```

200 - foreign_worker      1    633.74  699.74
201 - telephone          1    635.02  701.02
202 - purpose            9    652.20  702.20
203 - duration           1    636.72  702.72
204 - other_debtor       2    639.00  703.00
205 - creditamount       1    637.15  703.15
206 - saving_account     4    644.51  704.51
207 - InstallmentRate    1    639.47  705.47
208 - credit_history     4    649.16  709.16
209 - sex                3    647.25  709.25
210 - account_status     3    684.32  746.32
211
212 Step:  AIC=697.72
213 responsegoodcredit ~ account_status + duration + credit_history +
214     purpose + creditamount + saving_account + InstallmentRate +
215     sex + other_debtor + other_installplans + telephone +
216         foreign_worker
217
217           Df Deviance    AIC
218 <none>           631.72  697.72
219 - other_installplans  2    636.57  698.57
220 - foreign_worker     1    635.36  699.36
221 - telephone          1    636.70  700.70
222 - duration           1    638.39  702.39
223 - other_debtor       2    640.41  702.41
224 - creditamount       1    638.52  702.52
225 - purpose            9    655.03  703.03
226 - saving_account     4    646.14  704.14
227 - InstallmentRate    1    640.46  704.46
228 - sex                3    647.25  707.25
229 - credit_history     4    651.12  709.12
230 - account_status     3    686.21  746.21

```

```

1 myglm_backward=glm(responsegoodcredit ~ account_status + duration
2     + credit_history +
3     other_debtor + purpose + saving_account + sex +
4     InstallmentRate +
5     creditamount + telephone + foreign_worker +
6     other_installplans,
7     data=datatrain,family = "binomial")
8 summary(myglm_backward)

```

```

1      > myglm_backward=glm(responsegoodcredit ~ account_status +
2        duration + credit_history +
3        + other_debtor + purpose +
4        saving_account + sex + InstallmentRate +
5        + creditamount + telephone +
6        foreign_worker + other_installplans,
7        + data=datatrain,family = "binomial")
8      > summary(myglm_backward)
9
10     Call:
11     glm(formula = responsegoodcredit ~ account_status +
12       duration +
13       credit_history + other_debtor + purpose +
14       saving_account +
15       sex + InstallmentRate + creditamount + telephone +
16       foreign_worker +
17       other_installplans, family = "binomial", data =
18       datatrain)
19
20     Deviance Residuals:
21         Min         1Q       Median         3Q        Max
22     -2.0515   -0.7086   -0.3656    0.7181    2.9138
23
24     Coefficients:
25
26             Estimate Std. Error z value
27             Pr(>|z|)
28 (Intercept)      1.713e+00  7.805e-01   2.195
29      0.028155 *
30 account_statusA12  -5.020e-01  2.561e-01  -1.960
31      0.050008 .
32 account_statusA13  -1.369e+00  4.667e-01  -2.934
33      0.003350 **
34 account_statusA14  -1.811e+00  2.704e-01  -6.697
35      2.13e-11 ***
36 duration          2.788e-02  1.088e-02   2.563
37      0.010385 *
38 credit_historyA31  -6.367e-01  6.624e-01  -0.961
39      0.336459
40 credit_historyA32  -1.465e+00  5.050e-01  -2.902
41      0.003710 **
42 credit_historyA33  -1.132e+00  5.667e-01  -1.997
43      0.045844 *
44 credit_historyA34  -1.966e+00  5.306e-01  -3.705
45      0.000211 ***
46 other_debtorA102    9.851e-01  5.278e-01   1.866
47      0.062004 .
48 other_debtorA103   -1.094e+00  5.386e-01  -2.031
49      0.042217 *

```

```

30      purposeA41      -1.286e+00  4.186e-01  -3.073
      0.002120 **
31      purposeA410     -8.362e-01  1.035e+00  -0.808
      0.419274
32      purposeA42     -5.752e-01  3.013e-01  -1.909
      0.056278 .
33      purposeA43     -9.879e-01  2.920e-01  -3.383
      0.000718 ***
34      purposeA44     -7.536e-01  9.838e-01  -0.766
      0.443693
35      purposeA45     -6.508e-01  6.211e-01  -1.048
      0.294762
36      purposeA46      2.829e-01  4.846e-01   0.584
      0.559348
37      purposeA48     -1.442e+01  6.382e+02  -0.023
      0.981972
38      purposeA49     -1.010e+00  4.060e-01  -2.487
      0.012889 *
39      saving_accountA62 -2.171e-01  3.453e-01  -0.629
      0.529495
40      saving_accountA63 -4.789e-03  4.396e-01  -0.011
      0.991308
41      saving_accountA64 -1.218e+00  5.626e-01  -2.164
      0.030438 *
42      saving_accountA65 -9.503e-01  3.109e-01  -3.057
      0.002236 **
43      sexA92         -2.770e-01  4.378e-01  -0.633
      0.526894
44      sexA93         -1.094e+00  4.384e-01  -2.495
      0.012583 *
45      sexA94         -4.923e-01  5.268e-01  -0.934
      0.350063
46      InstallmentRate  3.060e-01  1.051e-01   2.911
      0.003602 **
47      creditamount     1.375e-04  5.283e-05   2.603
      0.009234 **
48      telephoneA192   -4.858e-01  2.198e-01  -2.210
      0.027082 *
49      foreign_workerA202 -1.365e+00  8.110e-01  -1.683
      0.092455 .
50      other_installplansA142 -1.468e-01  5.007e-01  -0.293
      0.769436
51      other_installplansA143 -6.018e-01  2.837e-01  -2.121
      0.033894 *
52      ---
53      Signif. codes:  0      ***      0.001      **      0.01
      *      0.05      .      0.1      1
54

```

```

55         (Dispersion parameter for binomial family taken to be
56           1)
57         Null deviance: 869.91  on 699  degrees of freedom
58         Residual deviance: 631.72  on 667  degrees of freedom
59         AIC: 697.72
60
61         Number of Fisher Scoring iterations: 14

```

C.2.1 Matrice di confusione per il modello con le variabili selezionate

```

1 cutoff= 0.5
2 prob_myglm_def=predict(myglm_def, type = "response")
3 predicted_myglm_def=prob_myglm_def>cutoff
4 predicted_myglm_def=as.numeric(predicted_myglm_def)
5
6 confusion_table=table(datatrain$responsegoodcredit,
7   predicted_myglm_def, dnn = c("Truth", "Predicted"))
8 confusion_matrix=as.matrix(confusion_table)
9 confusion_matrix

```

```

1 > confusion_matrix
2
3           Predicted
4 Truth    0    1
5    0 425   56
6    1  96 123

```

C.2.2 Alcuni indici di valutazione

```

1 sensitivity= confusion_matrix[1,1] / (confusion_matrix[1,1] +
2   confusion_matrix[2,1])
3 print(sensitivity)
4
5 Specificity = confusion_matrix[2,2] / (confusion_matrix[2,2] +
6   confusion_matrix[1,2])
7 print(Specificity)
8
9 precision = confusion_matrix[1,1] / (confusion_matrix[1,1] +
10  confusion_matrix[1,2])
11 print(precision)
12
13 accuracy = (confusion_matrix[1,1] + confusion_matrix[2,2]) /
14   (confusion_matrix[1,1] + confusion_matrix[1,2] +
15   confusion_matrix[2,1] + confusion_matrix[2,2])

```

```
11 print(accuracy)
1      > sensitivity=  confusion_matrix[1,1] /
2      (confusion_matrix[1,1] + confusion_matrix[2,1])
3      > print(sensitivity )
4      [1] 0.815739
5      > Specificity = confusion_matrix[2,2] /
6      (confusion_matrix[2,2] + confusion_matrix[1,2])
7      > print(Specificity)
8      [1] 0.6871508
9      > precision = confusion_matrix[1,1] /
10     (confusion_matrix[1,1] + confusion_matrix[1,2])
11     > print(precision)
12     [1] 0.8835759
13     > accuracy = (confusion_matrix[1,1] +
14     confusion_matrix[2,2]) / (confusion_matrix[1,1] +
15     confusion_matrix[1,2] + confusion_matrix[2,1] +
16     confusion_matrix[2,2])
17     > print(accuracy)
18     [1] 0.7828571
```

C.2.3 Curva ROC e AUC.

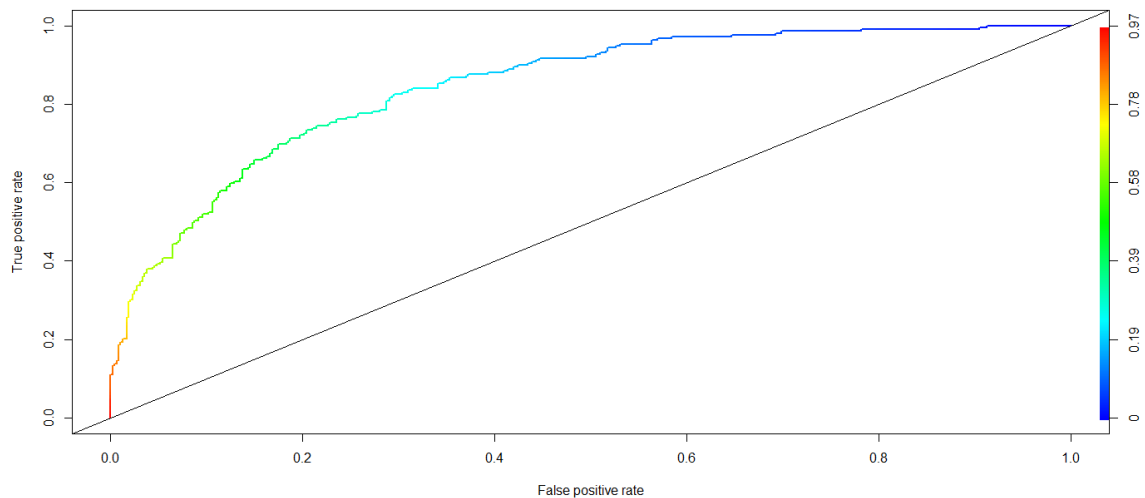
```
1 library(ROCR)
2 pred_myglm=predict(myglm_backward, type = "response")
3 pred_myglm=prediction(pred_myglm, datatrain$responsegoodcredit)
```

```
1 ROC= performance(pred_myglm,"tpr","fpr")
2 plot(ROC, colorize = T, lwd = 2)
3 abline(a = 0, b = 1)
```

Si veda la pagina seguente per la curva ROC in figura C.2.

```
1 auc = performance(pred_myglm, measure = "auc")
2 print(auc@y.values)
```

```
1      > print(auc@y.values)
2      [[1]]
3      [1] 0.8374012
```

Figura C.2: La curva ROC per il modello logit "*myglm_backward*"

C.3 *glm()* con le variabili selezione tramite la Forward Stepwise Regression

```

1 nullModel=glm(responsegoodcredit ~ 1, data=datatrain,family =
  "binomial")
2 myglm_Forward=step(nullModel,scope=list(lower=nullModel,
  upper=myglm), k = 2, direction="forward")

```

```

1 > myglm_Forward=step(nullModel,scope=list(lower=nullModel,
  upper=myglm), k = 2, direction="forward")
2 Start:  AIC=871.91
3 responsegoodcredit ~ 1
4
5
6      Df Deviance   AIC
7 + account_status    3   767.41 775.41
8 + duration          1   826.24 830.24
9 + credit_history    4   823.31 833.31
10 + saving_account   4   841.40 851.40
11 + creditamount     1   849.87 853.87
12 + property         3   850.06 858.06
13 + housing          2   853.24 859.24
14 + purpose          9   840.96 860.96
15 + foreign_worker   1   860.48 864.48
16 + other_installplans 2   860.35 866.35
17 + other_debtor     2   860.91 866.91
18 + present_employmentsince 4   857.48 867.48

```

```

18 + age                1    864.59 868.59
19 + sex                3    861.14 869.14
20 + InstallmentRate    1    866.62 870.62
21 + telephone          1    867.58 871.58
22 <none>                1    869.91 871.91
23 + numexisting_credits 1    869.67 873.67
24 + numpeople_maintenance 1    869.80 873.80
25 + present_residencesince 1    869.91 873.91
26 + job                3    869.17 877.17
27
28 Step:  AIC=775.41
29 responsegoodcredit ~ account_status
30
31                Df Deviance    AIC
32 + duration      1    733.78 743.78
33 + credit_history 4    735.48 751.48
34 + creditamount  1    748.76 758.76
35 + property      3    749.97 763.97
36 + other_debtor  2    754.31 766.31
37 + foreign_worker 1    756.88 766.88
38 + housing       2    757.95 769.95
39 + saving_account 4    755.11 771.11
40 + purpose       9    747.25 773.25
41 + sex           3    760.19 774.19
42 + other_installplans 2    762.20 774.20
43 + age           1    764.30 774.30
44 + InstallmentRate 1    764.68 774.68
45 + present_employmentsince 4    759.01 775.01
46 <none>          1    767.41 775.41
47 + telephone     1    766.45 776.45
48 + present_residencesince 1    767.17 777.17
49 + numexisting_credits 1    767.36 777.36
50 + numpeople_maintenance 1    767.37 777.37
51 + job           3    765.32 779.32
52
53 Step:  AIC=743.78
54 responsegoodcredit ~ account_status + duration
55
56                Df Deviance    AIC
57 + credit_history  4    710.80 728.80
58 + other_debtor    2    723.04 737.04
59 + sex             3    722.33 738.33
60 + present_employmentsince 4    720.90 738.90
61 + purpose         9    711.16 739.16
62 + foreign_worker  1    727.56 739.56
63 + housing         2    726.00 740.00
64 + saving_account  4    722.05 740.05
65 + telephone      1    730.56 742.56
66 + age            1    730.73 742.73

```

```

67 + InstallmentRate          1    731.49 743.49
68 <none>                     1    733.78 743.78
69 + property                 3    727.95 743.95
70 + other_installplans       2    730.24 744.24
71 + creditamount             1    732.99 744.99
72 + present_residencesince    1    733.45 745.45
73 + numexisting_credits       1    733.72 745.72
74 + numpeople_maintenance     1    733.73 745.73
75 + job                       3    733.59 749.59
76
77 Step:   AIC=728.8
78 responsegoodcredit ~ account_status + duration + credit_history
79
80                               Df Deviance    AIC
81 + other_debtor                2    699.30 721.30
82 + purpose                     9    685.90 721.90
83 + sex                         3    699.83 723.83
84 + present_employmentsince     4    698.63 724.63
85 + foreign_worker              1    705.88 725.88
86 + saving_account              4    699.91 725.91
87 + housing                     2    704.50 726.50
88 + telephone                  1    707.90 727.90
89 + InstallmentRate             1    708.24 728.24
90 <none>                        1    710.80 728.80
91 + age                         1    708.87 728.87
92 + numexisting_credits          1    709.12 729.12
93 + property                    3    705.62 729.62
94 + creditamount                1    710.17 730.17
95 + present_residencesince       1    710.75 730.75
96 + numpeople_maintenance       1    710.79 730.79
97 + other_installplans          2    709.27 731.27
98 + job                         3    710.55 734.55
99
100 Step:   AIC=721.3
101 responsegoodcredit ~ account_status + duration + credit_history +
102     other_debtor
103
104                               Df Deviance    AIC
105 + purpose                     9    676.44 716.44
106 + sex                         3    688.67 716.67
107 + saving_account              4    687.06 717.06
108 + present_employmentsince     4    689.57 719.57
109 + foreign_worker              1    695.59 719.59
110 + telephone                  1    696.07 720.07
111 + housing                     2    694.35 720.35
112 + InstallmentRate             1    696.66 720.66
113 <none>                        1    699.30 721.30
114 + age                         1    697.49 721.49
115 + numexisting_credits          1    697.55 721.55

```

```

116 + other_installplans      2    696.74  722.74
117 + creditamount           1    698.98  722.98
118 + present_residencesince  1    699.21  723.21
119 + numpeople_maintenance  1    699.22  723.22
120 + property                3    695.95  723.95
121 + job                     3    698.98  726.98
122
123 Step:   AIC=716.44
124 responsegoodcredit ~ account_status + duration + credit_history +
125     other_debtor + purpose
126
127               Df Deviance    AIC
128 + saving_account      4    663.60  711.60
129 + sex                 3    666.50  712.50
130 + foreign_worker      1    672.09  714.09
131 + present_employmentsince  4    666.59  714.59
132 + housing             2    671.51  715.51
133 + telephone           1    673.97  715.97
134 + InstallmentRate      1    674.26  716.26
135 + numexisting_credits   1    674.37  716.37
136 <none>                 1    676.44  716.44
137 + age                 1    674.66  716.66
138 + creditamount         1    675.88  717.88
139 + other_installplans   2    673.97  717.97
140 + present_residencesince  1    676.39  718.39
141 + numpeople_maintenance  1    676.44  718.44
142 + property             3    673.76  719.76
143 + job                  3    675.87  721.87
144
145 Step:   AIC=711.6
146 responsegoodcredit ~ account_status + duration + credit_history +
147     other_debtor + purpose + saving_account
148
149               Df Deviance    AIC
150 + sex                 3    653.45  707.45
151 + present_employmentsince  4    653.85  709.85
152 + foreign_worker      1    660.01  710.01
153 + telephone           1    661.06  711.06
154 + InstallmentRate      1    661.08  711.08
155 + housing             2    659.35  711.35
156 <none>                 1    663.60  711.60
157 + numexisting_credits   1    662.09  712.09
158 + age                 1    662.17  712.17
159 + other_installplans   2    660.61  712.61
160 + creditamount         1    662.83  712.83
161 + present_residencesince  1    663.59  713.59
162 + numpeople_maintenance  1    663.60  713.60
163 + property             3    660.95  714.95
164 + job                  3    663.42  717.42

```

```

165
166 Step:  AIC=707.45
167 responsegoodcredit ~ account_status + duration + credit_history +
168     other_debtor + purpose + saving_account + sex
169
170
171      Df Deviance    AIC
172 + InstallmentRate      1    648.78 704.78
173 + foreign_worker      1    649.93 705.93
174 + telephone          1    650.92 706.92
175 + other_installplans  2    649.26 707.26
176 + numexisting_credits 1    651.40 707.40
177 <none>                1    653.45 707.45
178 + present_employmentsince 4    645.88 707.88
179 + creditamount        1    652.40 708.40
180 + numpeople_maintenance 1    652.48 708.48
181 + housing            2    650.64 708.64
182 + age                1    653.07 709.07
183 + present_residencesince 1    653.45 709.45
184 + property           3    649.78 709.78
185 + job                3    652.99 712.99
186
187 Step:  AIC=704.78
188 responsegoodcredit ~ account_status + duration + credit_history +
189     other_debtor + purpose + saving_account + sex + InstallmentRate
190
191      Df Deviance    AIC
192 + creditamount      1    644.40 702.40
193 + foreign_worker    1    645.81 703.81
194 + other_installplans 2    644.39 704.39
195 + telephone        1    646.39 704.39
196 + numexisting_credits 1    646.62 704.62
197 <none>              1    648.78 704.78
198 + numpeople_maintenance 1    647.37 705.37
199 + housing           2    645.62 705.62
200 + present_employmentsince 4    642.03 706.03
201 + age              1    648.20 706.20
202 + present_residencesince 1    648.76 706.76
203 + property         3    645.69 707.69
204 + job              3    648.46 710.46
205
206 Step:  AIC=702.4
207 responsegoodcredit ~ account_status + duration + credit_history +
208     other_debtor + purpose + saving_account + sex +
209     InstallmentRate +
210     creditamount
211
212      Df Deviance    AIC
213 + telephone      1    640.13 700.13
214 + foreign_worker  1    641.46 701.46

```

```

213 + other_installplans      2    639.68 701.68
214 <none>                    644.40 702.40
215 + numexisting_credits     1    642.40 702.40
216 + numpeople_maintenance   1    642.72 702.72
217 + housing                 2    641.57 703.57
218 + present_employmentsince 4    637.74 703.74
219 + age                     1    643.78 703.78
220 + present_residencesince  1    644.40 704.40
221 + property                3    642.20 706.20
222 + job                     3    644.01 708.01
223
224 Step:   AIC=700.13
225 responsegoodcredit ~ account_status + duration + credit_history +
226   other_debtor + purpose + saving_account + sex +
227   InstallmentRate +
228   creditamount + telephone
229
229           Df Deviance    AIC
230 + foreign_worker      1    636.57 698.57
231 + other_installplans   2    635.36 699.36
232 + numexisting_credits  1    637.81 699.81
233 <none>                 640.13 700.13
234 + numpeople_maintenance 1    638.44 700.44
235 + housing              2    637.21 701.21
236 + present_employmentsince 4    633.52 701.52
237 + age                  1    639.89 701.89
238 + present_residencesince 1    640.11 702.11
239 + property             3    636.98 702.98
240 + job                  3    639.89 705.89
241
242 Step:   AIC=698.57
243 responsegoodcredit ~ account_status + duration + credit_history +
244   other_debtor + purpose + saving_account + sex +
245   InstallmentRate +
246   creditamount + telephone + foreign_worker
247
247           Df Deviance    AIC
248 + other_installplans   2    631.72 697.72
249 + numexisting_credits  1    634.14 698.14
250 <none>                 636.57 698.57
251 + numpeople_maintenance 1    634.73 698.73
252 + housing              2    633.57 699.57
253 + present_employmentsince 4    630.25 700.25
254 + age                  1    636.29 700.29
255 + present_residencesince 1    636.56 700.56
256 + property             3    633.52 701.52
257 + job                  3    636.27 704.27
258
259 Step:   AIC=697.72

```

```

260 responsegoodcredit ~ account_status + duration + credit_history +
261   other_debtor + purpose + saving_account + sex +
262     InstallmentRate +
263   creditamount + telephone + foreign_worker + other_installplans
264
265           Df Deviance    AIC
265 <none>           631.72 697.72
266 + numpeople_maintenance    1    630.00 698.00
267 + numexisting_credits      1    630.01 698.01
268 + housing                  2    628.45 698.45
269 + age                      1    631.17 699.17
270 + present_residencesince   1    631.71 699.71
271 + present_employmentsince  4    626.07 700.07
272 + property                 3    629.14 701.14
273 + job                      3    631.62 703.62

```

```

1 myglm_Forward= glm(responsegoodcredit~ account_status + duration +
2   credit_history +
3     other_debtor + purpose + saving_account + sex
4       + InstallmentRate +
5     creditamount + telephone + foreign_worker +
6       other_installplans ,data=datatrain,family
7       = "binomial")
8
9 summary(myglm_Forward)

```

```

1 > summary(myglm_Forward)
2
3 Call:
4 glm(formula = responsegoodcredit ~ account_status + duration +
5   credit_history + other_debtor + purpose + saving_account +
6   sex + InstallmentRate + creditamount + telephone +
7     foreign_worker +
8     other_installplans, family = "binomial", data = datatrain)
9
10 Deviance Residuals:
11     Min       1Q   Median       3Q      Max
12 -2.0515  -0.7086  -0.3656   0.7181   2.9138
13
14 Coefficients:
15 (Intercept)          1.713e+00  7.805e-01   2.195  0.028155 *
16 account_statusA12    -5.020e-01  2.561e-01  -1.960  0.050008 .
17 account_statusA13    -1.369e+00  4.667e-01  -2.934  0.003350 **
18 account_statusA14    -1.811e+00  2.704e-01  -6.697  2.13e-11 ***

```

```

19 duration                2.788e-02  1.088e-02   2.563 0.010385 *
20 credit_historyA31       -6.367e-01  6.624e-01  -0.961 0.336459
21 credit_historyA32       -1.465e+00  5.050e-01  -2.902 0.003710 **
22 credit_historyA33       -1.132e+00  5.667e-01  -1.997 0.045844 *
23 credit_historyA34       -1.966e+00  5.306e-01  -3.705 0.000211 ***
24 other_debtorA102         9.851e-01  5.278e-01   1.866 0.062004 .
25 other_debtorA103       -1.094e+00  5.386e-01  -2.031 0.042217 *
26 purposeA41              -1.286e+00  4.186e-01  -3.073 0.002120 **
27 purposeA410             -8.362e-01  1.035e+00  -0.808 0.419274
28 purposeA42              -5.752e-01  3.013e-01  -1.909 0.056278 .
29 purposeA43              -9.879e-01  2.920e-01  -3.383 0.000718 ***
30 purposeA44              -7.536e-01  9.838e-01  -0.766 0.443693
31 purposeA45              -6.508e-01  6.211e-01  -1.048 0.294762
32 purposeA46               2.829e-01  4.846e-01   0.584 0.559348
33 purposeA48              -1.442e+01  6.382e+02  -0.023 0.981972
34 purposeA49              -1.010e+00  4.060e-01  -2.487 0.012889 *
35 saving_accountA62       -2.171e-01  3.453e-01  -0.629 0.529495
36 saving_accountA63       -4.789e-03  4.396e-01  -0.011 0.991308
37 saving_accountA64       -1.218e+00  5.626e-01  -2.164 0.030438 *
38 saving_accountA65       -9.503e-01  3.109e-01  -3.057 0.002236 **
39 sexA92                  -2.770e-01  4.378e-01  -0.633 0.526894
40 sexA93                  -1.094e+00  4.384e-01  -2.495 0.012583 *
41 sexA94                  -4.923e-01  5.268e-01  -0.934 0.350063
42 InstallmentRate         3.060e-01  1.051e-01   2.911 0.003602 **
43 creditamount            1.375e-04  5.283e-05   2.603 0.009234 **
44 telephoneA192           -4.858e-01  2.198e-01  -2.210 0.027082 *
45 foreign_workerA202      -1.365e+00  8.110e-01  -1.683 0.092455 .
46 other_installplansA142  -1.468e-01  5.007e-01  -0.293 0.769436
47 other_installplansA143  -6.018e-01  2.837e-01  -2.121 0.033894 *
48 ---
49 Signif. codes:  0      ***      0.001      **      0.01      *      0.05
                    .      0.1      1
50
51 (Dispersion parameter for binomial family taken to be 1)
52
53 Null deviance: 869.91  on 699  degrees of freedom
54 Residual deviance: 631.72  on 667  degrees of freedom
55 AIC: 697.72
56
57 Number of Fisher Scoring iterations: 14

```

C.3.1 Matrice di confusione per il modello con le variabili selezionate

```

1 cutoff= 0.5
2 prob_myglmforward=predict(myglm_Forward, type = "response")
3 predicted_myglmforward=prob_myglmforward>cutoff

```

```

4 predicted_myglmforward=as.numeric(predicted_myglmforward)
5
6 confusion_table=table(datatrain$responsegoodcredit,
   predicted_myglmforward, dnn = c("Truth", "Predicted"))
7 confusion_matrix=as.matrix(confusion_table)
8 confusion_matrix

```

```

1      > confusion_matrix
2              Predicted
3      Truth    0    1
4              0 425  56
5              1  96 123

```

C.3.2 Alcuni indici di valutazione

```

1 sensitivity=
   confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[2,1])
2 print(sensitivity )
3
4 Specificity =
   confusion_matrix[2,2]/(confusion_matrix[2,2]+confusion_matrix[1,2])
5 print(Specificity)
6
7 precision =
   confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[1,2])
8 print(precision)
9
10 accuracy = (confusion_matrix[1,1] + confusion_matrix[2,2]) /
   (confusion_matrix[1,1] + confusion_matrix[1,2] +
   confusion_matrix[2,1] + confusion_matrix[2,2])
11 print(accuracy)

```

```

1      > sensitivity = confusion_matrix[1,1] /
   (confusion_matrix[1,1] + confusion_matrix[2,1])
2      > print(sensitivity )
3      [1] 0.815739
4      > Specificity = confusion_matrix[2,2] /
   (confusion_matrix[2,2] + confusion_matrix[1,2])
5      > print(Specificity)
6      [1] 0.6871508
7      > precision = confusion_matrix[1,1] /
   (confusion_matrix[1,1] + confusion_matrix[1,2])
8      > print(precision)
9      [1] 0.8835759
10     > accuracy = (confusion_matrix[1,1] +
   confusion_matrix[2,2]) / (confusion_matrix[1,1] +
   confusion_matrix[1,2] + confusion_matrix[2,1] +
   confusion_matrix[2,2])

```

```

11         > print(accuracy)
12         [1] 0.7828571

```

C.3.3 Curva ROC e AUC.

```

1 library(ROCR)
2 pred_myglm=predict(myglm_Forward, type = "response")
3 pred_myglm=prediction(pred_myglm, datatrain$responsegoodcredit)

```

```

1 ROC= performance(pred_myglm,"tpr","fpr")
2 plot(ROC, colorize = T, lwd = 2)
3 abline(a = 0, b = 1)

```

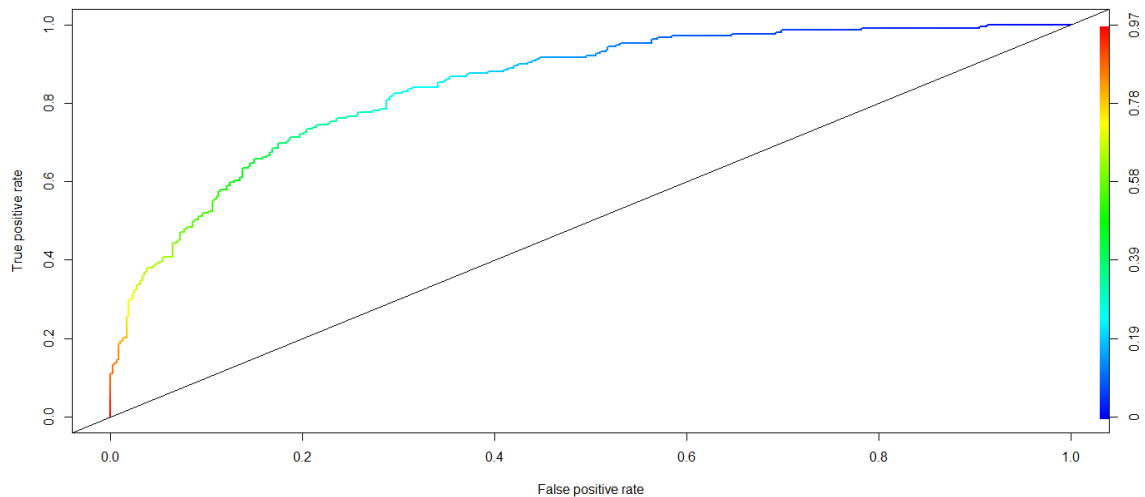


Figura C.3: La curva ROC per il modello logit *"myglm_backward"*

```

1 auc = performance(pred_myglm, measure = "auc")
2 print(auc@y.values)

```

```

1         > print(auc@y.values)
2         [[1]]
3         [1] 0.8374012

```

Bibliografia

- [1] DS (ds5jexcite.com). «Interdisciplinary Independent Scholar with 9+ years' experience in risk management». In: *Self Published* (2009).
- [2] Alan Agresti. *Categorical Data Analysis*. 2003.
- [3] BancoBPM. «GLOSSARIO - Qualità del credito». In: *bancobpm.it* (). URL: <https://www.bancobpm.it/magazine/glossario/qualita-del-credito/>.
- [4] Rebecca Bevans. «An introduction to the Akaike information criterion». In: *Scribbr Published* (2020).
- [5] Bluecology. «How do I interpret the AIC». In: *R blogger* (2018).
- [6] Thomas Brock. «What Is Credit Scoring?». In: *Investopedia* (2021).
- [7] Noel Capon. «Credit Scoring Systems: A Critical Analysis». In: *Issue published* (1982).
- [8] Bolton Christine. «Logistic regression and its application in credit scoring». In: *Dissertation (MSc)–University of Pretoria* (2010).
- [9] Nguyen Chi Dung. «An Application of Credit Scoring: Developing Scorecard Model for A Vietnam Commercial Bank». In: *RPubs* ().
- [10] Findomestic. «GLOSSARIO - Affidabilità Creditizia». In: *findomestic.it* (). URL: <https://www.findomestic.it/glossario/index.shtml>.
- [11] Peter K. Dunn Gordon K. Smyth. *Generalized Linear Models With Examples in R*. CRC Press, 2013.
- [12] Professor Dr. Hans Hofmann. «Statlog (German Credit Data) Data Set». In: *UCI, machine Learning Repository* (1994). URL: [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)).
- [13] S Hosmer D. & Lemeshow. *Applied Logistic Regression*. second edition. John Wiley & Sons, Inc, 2000.
- [14] Borsa Italiana. «GLOSSARIO FINANZIARIO - RISCHIO DI CREDITO». In: *borsaitaliana.it* (). URL: <https://www.borsaitaliana.it/borsa/glossario/rischio-di-credito.html>.
- [15] Yuho Kida. *Generalized linear models: Introduction to advanced statistical modelling*. Towards Data Science, 2019.
- [16] D Edelman L Thomas J Crook. «Credit scoring and its applications». In: *SIAM* (2017).

- [17] James Le. «Logistic Regression in R Tutorial». In: *DataCamp* (2018).
- [18] J. Scott Long. «Regression Models for Categorical and Limited Dependent Variables». In: *Sage Publications* (1997).
- [19] Govoni Lorenzo. «Come funziona un algoritmo di regressione logistica». In: *Self Published* ().
- [20] Akul Mahajan. «Logistic Regression and CART on German Credit Data». In: *Self Published* (2018).
- [21] Elizabeth Mays. *Handbook of Credit Scoring*. A cura di Ltf. The Glenlake Publishing Company, 2001.
- [22] Edgar C. Merkle e Michael Smithson. *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. Springer, 2018.
- [23] Jae H Min e Young-Chan Lee. «A practical approach to credit scoring». In: *Elsevier Ltd* (2007).
- [24] «Modello logit». In: *Wikipedia* (). URL: https://it.wikipedia.org/wiki/Modello%5C_logit.
- [25] Biz Nigatu. «CS871_IP3_BizNigatu». In: *Self Published* (2019).
- [26] «Overleaf Documentation and Tutorials». In: *Overleaf* (). URL: <https://www.overleaf.com/learn>.
- [27] Vidhi Rathod. «German Credit Scoring Data». In: *Self Published* (2020).
- [28] Carmona René. *Statistical Analysis of Financial Data in R*. Springer, 2014.
- [29] Elena Stanghellini. *Introduzione ai metodi statistici per il credit scoring*. Springer-Verlag Mailand, 2009.
- [30] James H. Stock e Mark W. Watson. *Regression with a Binary Dependent Variable, in Introduction to Econometrics*. Pearson, 2015.
- [31] Lumongga Bintang Yustisia. «Exploratory Data Analysis for German Credit Data (Part 1.)» In: *Medium article* (2019).