CAPITOLO 3

ANALISI STATISTICA PER IL CALCOLO DELLA PROBABILITY OF DEFAULT

3.1 Presentazione dei dati e data pre-processing

Il database selezionato agli scopi dell'analisi è stato scaricato dal sito web di Kaggle, una community online di data scientists amatoriali e professionisti, all'indirizzo: https://www.kaggle.com/datasets/marceloventura/the-berka-dataset. Il database in questione è stato preparato da Petr Berka and Marta Sochorova ed è stato utilizzato per la challenge PKDD'99. Contiene i dati finanziari di una banca ceca, raccolti in otto relazioni, di seguito descritte nelle tabelle dalla 3.1 alla 3.8. Lo scopo dello studio che si intende condurre prevede la costruzione di un modello che sia in grado di prevedere i casi di insolvenza di un prestito sulla base dei dati a disposizione. A tal fine è risultato indispensabile costruire, a partire dal database, un unico dataset contenente tutte le informazioni necessarie su ogni account a cui è stato ceduto un prestito, e quindi operare un merge di tali relazioni. Successivamente alla fase di merge, sono state fatte delle considerazioni e delle trasformazioni, che sono descritte nel seguito.

La relazione *account* contiene 4500 record e ognuno descrive le caratteristiche di un account.

Tabella 3.1. Relazione *account*.

Colonna	**Descrizione**	**Note**
account_id	Codice identificativo dell'account	Chiave primaria
district_id	Codice identificativo del distretto	Chiave esterna

date	Data di creazione dell'account nella forma "AAMMGG"	Questo attributo è stato utilizzato per calcolare il numero di mesi che intercorre tra l'apertura dell'account e la richiesta del prestito, ottenendo il nuovo attributo "acc_m"
frequency	Frequenza di rilascio delle dichiarazioni: "POPLATEK MESICNE" sta per "monthly issuance" "POPLATEK TYDNE" sta per "weekly issuance" "POPLATEK PO OBRATU" sta per "issuance after transaction"	Per facilitare la comprensione dei risultati, le osservazioni sono state tradotte in lingua inglese come "Monthly", "Weekly" e "AfterTrans"
acc_m	Numero di mesi che intercorre tra l'apertura dell'account e la richiesta del prestito	N/A

La relazione *card* contiene 892 record e ognuno descrive una carta di credito rilasciata ad un account.

Tabella 3.2. Relazione *card.*

Colonna	**Descrizione**	**Note**
card_id	Codice identificativo della carta di credito	Chiave primaria
disp_id	Codice identificativo della disposizione	Chiave esterna
type (*card_type*)	Tipo della carta di credito: "junior", "classic" o "gold"	Per facilitare la comprensione dei risultati, l'attributo è stato rinominato "card_type" Siccome per oltre il 75% degli account richiedenti un prestito questo attributo assume valore NA, questo è stato rimosso

issued	Data di emissione della carta di credito nella forma “AAMMGG”	Il formato è stato trasformato in AAAA-MM-GG

La relazione *client* contiene 5369 record e ognuno descrive le caratteristiche di un cliente.

Tabella 3.3. Relazione *client.*

Colonna	**Descrizione**	**Note**
client_id	Codice identificativo del cliente	Chiave primaria
district_id	Codice identificativo del distretto	Chiave esterna
birth_date	Data di nascita del cliente nella forma “AAMMGG” per gli uomini e “AA50+MMGG” per le donne	Il formato è stato trasformato in AAAA-MM-GG Questo attributo è stato utilizzato per calcolare il sesso e l’età del cliente alla richiesta del prestito, ottenendo i nuovi attributi “gender” e “age”
gender	Sesso del cliente richiedente il prestito: “M” per gli uomini “F” per le donne	N/A
age	Età del cliente al momento della richiesta del prestito	N/A

La relazione *disposition* contiene 5369 record e ognuno mette in relazione un cliente a un account.

Tabella 3.4. Relazione *disposition.*

Colonna	**Descrizione**	**Note**
disp_id	Codice identificativo della disposizione	Chiave primaria
client_date	Codice identificativo del cliente	Chiave esterna
account_id	Codice identificativo dell'account	Chiave esterna
type	Tipo della disposizione: "OWNER" sta per "owner" "DISPONENT" sta per "user"	Solo gli account di tipo *owner* possono richiedere prestiti, pertanto non ci occuperemo di quelli di tipo *user*

La relazione *district* contiene 77 record e ognuno descrive le caratteristiche demografiche di un distretto.

Tabella 3.5. Relazione *district.*

Colonna	**Descrizione**	**Note**
district_id	Codice identificativo del distretto	Chiave primaria
A2 (*district_name*)	Nome del distretto	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato "district_name"

A3 (*region*)	Regione: “west Bohemia” “east Bohemia” “central Bohemia” “Prague” “south Bohemia” “south Moravia” “north Moravia” “north Bohemia”	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “region”
A4 (*inhab*)	Numero di abitanti	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “inhab” L’unità di misura è stata cambiata dalle unità alle decine di migliaia
A5 (*mun499*)	Numero di comuni con un numero di abitanti non superiore a 499	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “num499”
A6 (*mun1999*)	Numero di comuni con un numero di abitanti compreso tra 500 e 1999, estremi inclusi	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “mun1999”
A7 (*mun9999*)	Numero di comuni con un numero di abitanti compreso tra 2000 e 9999, estremi inclusi	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “mun9999”
A8 (*mun10000*)	Numero di comuni con un numero di abitanti non superiore a 10000	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “mun10000”
A9 (*cities*)	Numero di città	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “cities”
A10 (*ratio_urban*)	Tasso di abitanti urbani	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “ratio_urban”

A11 (*avg_salary*)	Salario medio	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “avg_salary” L’unità di misura è stata cambiata dalle unità alle centinaia
A12 (*unempl95*)	Tasso di disoccupazione nel 1995	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “unempl95”
A13 (*unempl96*)	Tasso di disoccupazione nel 1996	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “unempl96”
A14 (*ratio_entr*)	Numero di imprenditori ogni 1000 abitanti	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “ratio_entr”
A15 (*crimes95*)	Numero di crimini commessi nel 1995	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “crimes95” L’unità di misura è stata cambiata dalle unità alle centinaia
A16 (*crimes96*)	Numero di crimini commessi nel 1996	Per facilitare la comprensione dei risultati, questo attributo è stato rinominato “crimes96” L’unità di misura è stata cambiata dalle unità alle centinaia

La relazione *loan* contiene 682 record e ognuno descrive un prestito fornito ad un account.

Tabella 3.6. Relazione *loan.*

Variabile	**Descrizione**	**Note**
loan_id	Codice identificativo del prestito	Chiave primaria
account_id	Codice identificativo dell'account	Chiave esterna
date	Data di concessione del prestito nella forma "AAMMGG"	Il formato è stato trasformato in AAAA-MM-GG
amount	Ammontare del prestito	L'unità di misura è stata cambiata dalle unità alle decine di migliaia
duration	Numero di rate: 12 mesi 24 mesi 36 mesi 48 mesi 60 mesi	N/A
payments	Ammontare di ogni rata	L'unità di misura è stata cambiata dalle unità alle centinaia
status	Stato del pagamento del prestito: "A" sta per "contract finished, no problems" "B" sta per "contract finished, loan not payed" "C" sta per "running contract, OK so far" "D" sta per "running contract, client in debt"	Questo attributo è stato reso dicotomico assegnando le categorie A e C alla classe 0 e le categorie B e D alla classe 1

La relazione *order* contiene 6471 record e ognuno descrive le caratteristiche di un ordine di pagamento. Tuttavia, nessuno degli account richiedenti un prestito ha effettuato ordini di pagamento presenti in questa relazione. Per completezza, se ne riportano gli attributi.

Tabella 3.7. Relazione *order.*

Colonna	**Descrizione**	**Note**
order_id	Codice identificativo dell'ordine	Chiave primaria
account_id	Codice identificativo dell'account che emette l'ordine	Chiave esterna
bank_to	Codice identificativo della banca del destinatario	N/A
account_to	Codice identificativo dell'account del destinatario	N/A
amount	Importo addebitato	N/A
k_symbol	Caratterizzazione del pagamento: "POJISTNE" sta per "insurance payment" "SIPO" sta per "household" "LEASING" sta per "leasing" "UVER" sta per "loan payment"	Per facilitare la comprensione dei risultati, le osservazioni sono state tradotte in lingua inglese come "insurance", "household", "leasing" e "loan_payment"

La relazione *trans* contiene 1056320 record e ognuno descrive una transazione.

Tabella 3.8. Relazione *trans.*

Colonna	**Descrizione**	**Note**
trans_id	Codice identificativo della transazione	Chiave primaria
account_id	Codice identificativo dell'account ordinante la transazione	Chiave esterna
date	Data della transazione nella forma "AAMMGG"	Il formato è stato trasformato in AAAA-MM-GG
type	Tipo della transazione "PRIJEM" sta per "credit" "VYDAJ" e "VYBER" stanno per "withdrawal"	Per facilitare la comprensione dei risultati, le osservazioni sono state tradotte in lingua inglese come "credit" e "withdrawal" Per ogni account richiedente un prestito, sono state fatte delle aggregazioni di amount; si veda la nota della variabile amount
operation	Tipo di operazione: "VYBER KARTOU" sta per "credit card withdrawal" "VKLAD" sta per "credit in cash" "PREVOD Z UCTU" sta per "collection from another bank" "VYBER" sta per "withdrawal in cash" "PREVOD NA UCET" sta per "remittance to another bank"	Per facilitare la comprensione dei risultati, le osservazioni sono state tradotte in lingua inglese come "credit_card_w", "cash_c", "collection", "cash_w" e "remittance" Per ogni account richiedente un prestito è stato calcolato il numero di transazioni per ogni operazione, ottenendo i nuovi attributi "credit_card_w", "cash_c", "collection", "cash_w" e "remittance"

amount	Ammontare della transazione	Questo attributo è stato utilizzato per calcolare, per ogni account richiedente un prestito, l'ammontare medio per ogni type, ottenendo gli attributi "avg_amtC" e "avg_amtW"
balance	Saldo dopo la transazione	Questo attributo è stato utilizzato per calcolare il bilancio medio, ottenendo l'attributo "avg_balance"
k_symbol	Caratterizzazione della transazione: "POJISTNE" sta per "insurance payment" "SLUZBY" sta per "payment for statement" "SIPO" sta per "household" "UVER" sta per "loan payment" "UROK" sta per "interest credited" "SANKC. UROK" sta per "sanction interest if negative balance" "DUCHOD" sta per "old-age pension"	Per facilitare la comprensione dei risultati, le osservazioni sono state tradotte in lingua inglese come "insurance", "statement", "household", "loan", "interest_credited", "sanction_negBalance" e "pension"
bank	Codice identificativo della banca del partner della transazione	N/A
account	Codice identificativo del partner della transazione	N/A
avg_amtC	Ammontare medio di ogni transazione di tipo "credit"	L'unità di misura è stata cambiata dalle unità alle centinaia
avg_amtW	Ammontare medio di ogni transazione di tipo "withdrawal"	L'unità di misura è stata cambiata dalle unità alle centinaia

avg_balance	Bilancio medio	L'unità di misura è stata cambiata dalle unità alle migliaia
collection	Numero di transazioni di operazione "collection"	Siccome per oltre il 69% degli account richiedenti un prestito questo attributo assume valore NA, questo è stato rimosso
remittance	Numero di transazioni di operazione "remittance"	N/A
cash_c	Numero di transazioni di operazione "cash_c"	N/A
cash_w	Numero di transazioni di operazione "cash_w"	N/A
credit_ card_w	Numero di transazioni di operazione "credit_card_w"	Siccome per oltre il 78% degli account richiedenti un prestito questo attributo assume valore NA, questo è stato rimosso

Bisogna tener conto, durante la lettura dei dati, che:

- un cliente può avere più di un account e un account può essere condiviso da più di un cliente;
- un account può essere associato a più carte di credito;
- solo i clienti con un account di tipo *owner* possono emettere ordini e richiedere prestiti;
- ad un account può essere concesso al massimo un prestito.

3.1.1 Costruzione del dataset e analisi delle singole variabili

Per ognuno dei 682 account richiedenti un prestito sono state estratte le variabili di risposta *status* e altre 27 variabili elencate nelle tabelle dalla 3.9 alla 3.13, divise in base alla relazione di provenienza. Siccome 8 osservazioni presentavano valori mancanti, queste sono state rimosse, riducendo a 674 il numero di osservazioni. Per ogni variabile diversa da *status* sono stati effettuati dei test per valutare l'eventuale rapporto di dipendenza con la variabile di risposta: per le variabili numeriche sono stati eseguiti il t-test e l'F-test, mentre

per le variabili categoriche è stato effettuato il test Chi Squared ed è stato calcolato l'indice di V-Cramer:

- Il t-test è stato utilizzato per confrontare la differenza osservata tra le medie condizionate dei due gruppi, *status* = 0 e *status* = 1. L'ipotesi nulla H_0 afferma che la differenza è pari a 0, mentre l'ipotesi alternativa H_1 afferma che la differenza è diversa da 0.
- L'F-test è stato utilizzato per confrontare il rapporto tra le varianze osservate condizionate dei due gruppi, *status* = 0 e *status* = 1. L'ipotesi nulla H_0 afferma che il rapporto è pari a 1, mentre l'ipotesi alternativa H_1 afferma che il rapporto è diverso da 1.
- Il test Chi Squared è stato utilizzato per verificare se l'associazione tra la variabile esplicativa e *status* è statisticamente significativa, e ciò è fatto confrontando le frequenze osservate con quelle attese. L'ipotesi nulla H_0 afferma che le due variabili sono indipendenti, mentre l'ipotesi alternativa H_1 afferma che esiste un rapporto di associazione.
- L'indice V-Cramer, costruito a partire dal test del Chi Squared, è stato utilizzato per valutare il grado dell'associazione tra la variabile esplicativa e *status*.

Considerando un livello di significatività α, fissato a 0.05, si pone un asterisco accanto ai *p-value* che portano al rifiuto dell'ipotesi nulla H_0.

Tabella 3.9. Variabili provenienti dalla relazione *loan.*

Variabile	**Distribuzione**	**Test**
status	0: 89% 1: 11%	N/A
amount	Min.: 0.50 1st Qu.: 6.65 Median: 11.63 3rd Qu.: 15.03 Max.: 59.08	t-test: p-value = 0.0004* F-test: p-value = 0.0205*

duration	12: 19.29% 24: 20.03% 36: 19.29 % 48: 20.18% 60: 21.22%	Chi Squared Test: p-value = 0.8389 V-Cramer: 0.0461
payments	Min.: 3.04 1st Qu.: 24.51 Median: 39.00 3rd Qu.: 57.77 Max.: 99.10	t-test: p-value = 1.067e-05* F-test: p-value = 0.6714

Tabella 3.10. Variabili provenienti dalla relazione *account.*

Variabile	**Distribuzione**	**Test**
frequency	AfterTrans: 4.60% Weekly: 13.35% Monthly: 82.05	Chi Squared Test: p-value = 0.5713 V-Cramer: 0.0408
acc_m	Min.: 3.0 1st Qu.: 8.0 Median: 13.0 3rd Qu.: 19.0 Max.: 23.0	t-test: p-value = 0.0432* F-test: p-value = 0.7587

Tabella 3.11. Variabili provenienti dalla relazione *client.*

Variabile	**Distribuzione**	**Test**
gender	F: 50.74% M: 49.26%	Chi Squared Test: p-value = 0.6306 V-Cramer: 0.0233
age	Min.: 13.0 1st Qu.: 27.0 Median: 37.0 3rd Qu.: 48.75 Max.: 61.0	t-test: p-value = 0.9955 F-test: p-value = 0.4711

Tabella 3.12. Variabili provenienti dalla relazione *district.*

Variabile	**Distribuzione**	**Test**
region	west Bohemia: 8.46% east Bohemia: 12.46% central Bohemia: 13.35% Prague: 12.46% south Bohemia: 8.9% south Moravia: 19.14% north Moravia: 16.17% north Bohemia: 9.05%	Chi Squared Test: p-value = 0.1869 V-Cramer: 0.1220
inhab	Min.: 4.57 1st Qu.: 9.39 Median: 12.49 3rd Qu.: 22.61 Max.: 120.49	t-test: p-value = 0.5066 F-test: p-value = 0.2155
mun499	Min.: 0.0 1st Qu.: 8.0 Median: 36.0 3rd Qu.: 65.0 Max.: 151.0	t-test: p-value = 0.7917 F-test: p-value = 0.7311
mun1999	Min.: 0.0 1st Qu.: 10.0 Median: 23.0 3rd Qu.: 33.0 Max.: 70.0	t-test: p-value = 0.6569 F-test: p-value = 0.6913
mun9999	Min.: 0.0 1st Qu.: 2.0 Median: 5.0 3rd Qu.: 8.0 Max.: 20.0	t-test: p-value = 0.2317 F-test: p-value = 0.2631
mun10000	Min.: 0.0 1st Qu.: 1.0 Median: 1.0 3rd Qu.: 2.0 Max.: 5.0	t-test: p-value = 0.7921 F-test: p-value = 0.8722

cities	Min.: 1.0 1st Qu.: 4.0 Median: 6.0 3rd Qu.: 7.0 Max.: 11.0	t-test: p-value = 0.3742 F-test: p-value = 0.6762
ratio_urban	Min.: 33.9 1st Qu.: 52.7 Median: 62.1 3rd Qu.: 87.7 Max.: 100.0	t-test: p-value = 0.8831 F-test: p-value = 0.9370
avg_salary	Min.: 81.1 1st Qu.: 85.46 Median: 89.92 3rd Qu.: 98.97 Max.: 125.41	t-test: p-value = 0.4207 F-test: p-value = 0.5588
unempl95	Min.: 0.29 1st Qu.: 1.51 Median: 2.77 3rd Qu.: 3.97 Max.: 7.34	t-test: p-value = 0.6944 F-test: p-value = 0.4012
unempl96	Min.: 0.43 1st Qu.: 1.96 Median: 3.49 3rd Qu.: 4.72 Max.: 9.40	t-test: p-value = 0.7809 F-test: p-value = 0.2357
ratio_entr	Min.: 81.0 1st Qu.: 105.0 Median: 115.0 3rd Qu.: 134.2 Max.: 167.0	t-test: p-value = 0.2048 F-test: p-value = 0.6733
crimes95	Min.: 8.18 1st Qu.: 21.66 Median: 37.33 3rd Qu.: 69.49 Max.: 856.77	t-test: p-value = 0.4822 F-test: p-value = 0.1873

crimes96	Min.: 8.88 1st Qu.: 23.05 Median: 38.68 3rd Qu.: 68.72 Max.: 991.07	t-test: p-value = 0.4622 F-test: p-value = 0.1797

Tabella 3.13. Variabili provenienti dalla relazione *trans.*

Variabile	**Distribuzione**	**Test**
avg_balance	Min.: 6.69 1st Qu.: 34.67 Median: 45.58 3rd Qu.: 55.43 Max.: 79.27	t-test: p-value = 2.364e-09* F-test: p-value = 0.6111
avg_amtC	Min.: 20.57 1st Qu.: 78.99 Median: 119.83 3rd Qu.: 155.27 Max.: 267.73	t-test: p-value = 0.0287* F-test: p-value = 0.0033*
avg_amtW	Min.: 8.96 1st Qu.: 41.17 Median: 62.07 3rd Qu.: 88.75 Max.: 154.31	t-test: p-value = 0.0906 F-test: p-value = 0.0442*
remittance	Min.: 1.0 1st Qu.: 21.0 Median: 45.0 3rd Qu.: 89.0 Max.: 287.0	t-test: p-value = 0.0014* F-test: p-value = 0.2179
cash_c	Min.: 1.00 1st Qu.: 11.00 Median: 39.00 3rd Qu.: 63.75 Max.: 148.00	t-test: p-value = 0.0015* F-test: p-value = 0.8530

cash_w	Min.: 13.0 1st Qu.: 70.0 Median: 99.0 3rd Qu.: 161.8 Max.: 324.0	t-test: p-value = 0.0108* F-test: p-value = 0.9574

Solo 9 delle 23 variabili numeriche su cui sono stati effettuati i test t-test e F-test hanno restituito *p-value* minori del livello di significatività $\alpha = 0.05$ per almeno uno dei due test (*amount*, *payments*, *acc_m*, *avg_balance*, *avg_amtC*, *avg_amtW, remittance, cash_c, cash_w*) e nessuna delle 4 variabili categoriche su cui è stato effettuato il test Chi Squared hanno restituito *p-value* minori di tale livello di significatività α.

3.1.2 Rimozione della multicollinearità

Successivamente, si è fatta un'analisi per rilevare l'eventuale presenza di alta correlazione tra le variabili esplicative, una condizione che non crea problemi nella previsione, ma rende instabili le stime dei coefficienti di regressione. Essendo possibile calcolare la correlazione, ovvero la forza del legame lineare, solo tra variabili numeriche, queste sono state estratte e ne è stata calcolata la matrice di correlazione, per poi individuare le coppie con correlazione maggiore di 0.8 in valore assoluto.

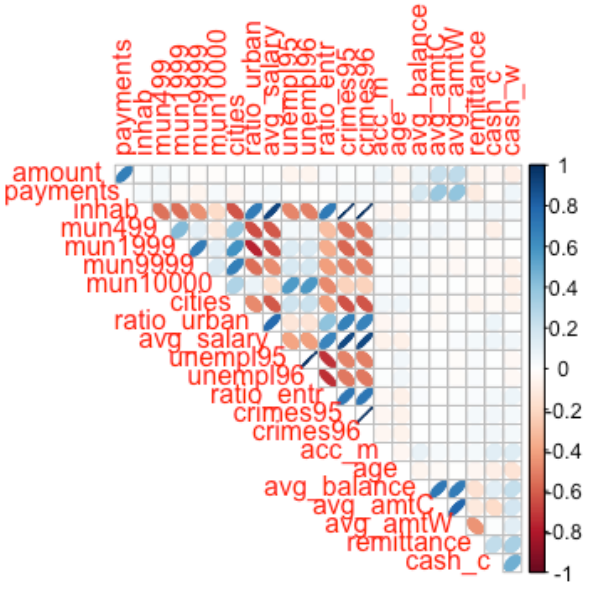

Figura 3.1. Matrice di correlazione delle variabili esplicative numeriche.

Sono state individuate 7 coppie con correlazione maggiore di 0.8 in valore assoluto, e la multicollinearità è stata risolta rimuovendo 4 variabili: *inhab, unempl96, crimes95* e *crimes96*. Il dataset risultante, che verrà utilizzato per la costruzione dei modelli, consta di 24 variabili, la variabile di target *status* e 23 variabili esplicative, di cui 4 categoriche e 19 numeriche.

3.1.3 Exploratory Data Analysis

Con l'obiettivo di individuare caratteristiche rilevanti dei dati, indagare relazioni e valutare le distribuzioni, si procede con un'analisi esplorativa, meglio nota come *exploratory data anlysis* o EDA. Questo tipo di analisi, sempre utile, può avvalersi di diversi strumenti e può raggiungere diversi livelli di dettaglio. In questa sede si è deciso di avvalersi prevalentemente di strumenti grafici. Prima si procede con lo studio della distribuzione univariata della variabile di risposta *status* e poi con lo studio delle distribuzioni univariate e condizionate delle variabili esplicative, concentrandosi principalmente sulle variabili i cui test di cui alle tabelle dalla 3.9 alla 3.13 hanno portato al rifiuto dell'ipotesi nulla H_0.

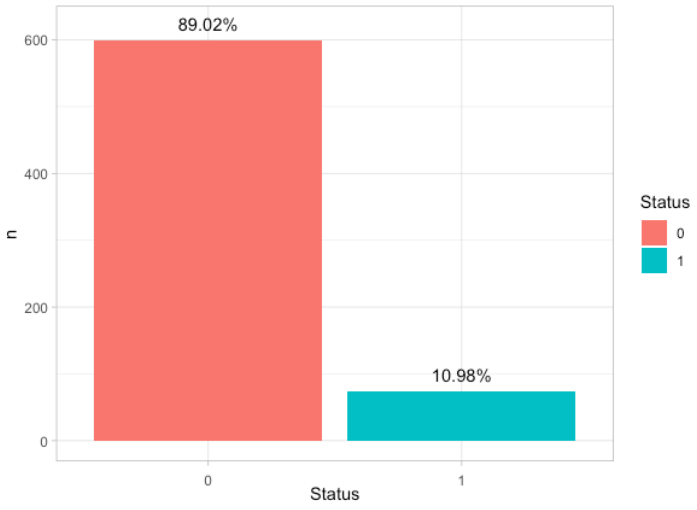

Figura 3.2. Barplot della variabile di risposta *status*.

La figura 3.2 mostra come la variabile di risposta *status* presenta un forte sbilanciamento tra le classi, con 600 osservazioni appartenenti alla classe 0, quindi account solventi, e 74 osservazioni appartenenti alla classe 1, quindi insolventi. In percentuale, alla classe 0 appartiene circa l'89% del campione, mentre alla classe 1 appartiene circa l'11% del campione. Il forte sbilanciamento presente deve essere necessariamente preso in

considerazione sia nella fase di divisione dei dati in training set e test set, sia in fase di modellazione.

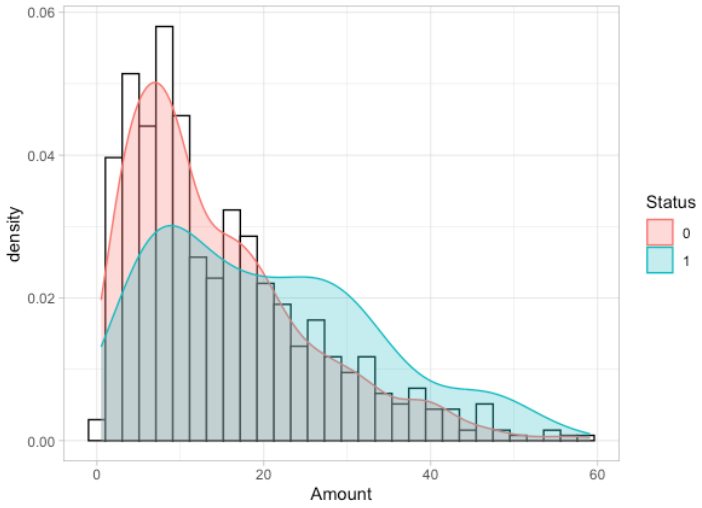

Figura 3.3. Istogramma della variabile *amount* con funzioni di densità condizionate rispetto alla variabile *status*.

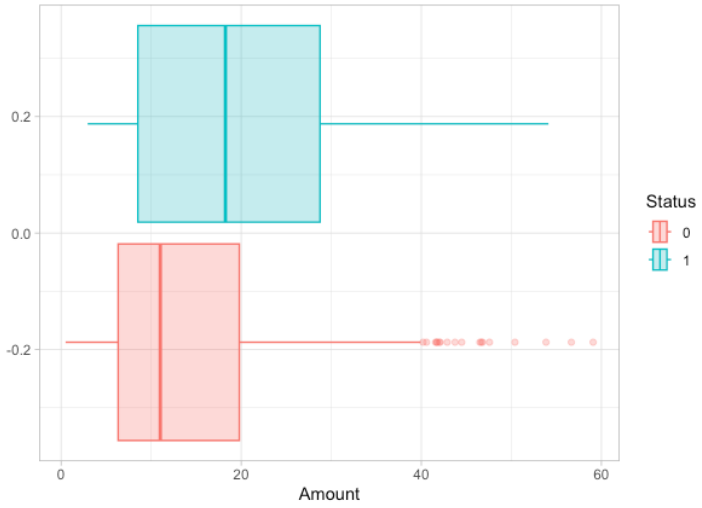

Figura 3.4. Boxplot della variabile *amount* condizionata rispetto alla variabile *status*.

Dalla figura 3.3 si evince che la distribuzione univariata di *amount* risulta essere asimmetrica, nello specifico si tratta di un'asimmetria positiva. La geometria dell'istogramma indica che per circa il 50% delle osservazioni l'ammontare del prestito è minore di 115000 corone ceche. La coda destra della distribuzione indica, però, la presenza

di qualche osservazione che supera le 500000 corone ceche, fino ad arrivare a quasi 600000 corone ceche. La distribuzione condizionata sulla variabile *status*, mostrata nelle figure 3.3 e 3.4, mette in luce come gli account insolventi hanno una distribuzione di *amount* spostata verso destra, a indicare valori di ammontare del prestito più elevati per questa categoria. Inoltre, si noti come la distribuzione di *amount* per i clienti insolventi risulti essere meno asimmetrica, a indicare un numero maggiore di clienti che richiede prestiti di ammontare elevato.

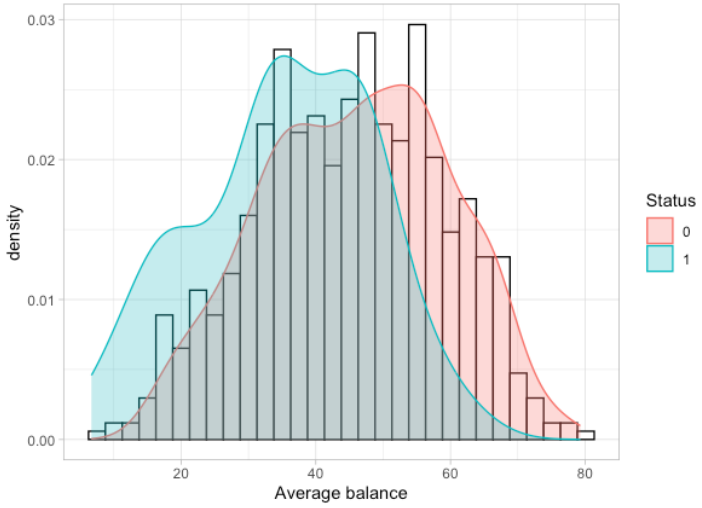

Figura 3.5. Istogramma della variabile *avg_balance* con funzioni di densità condizionate rispetto alla variabile *status*.

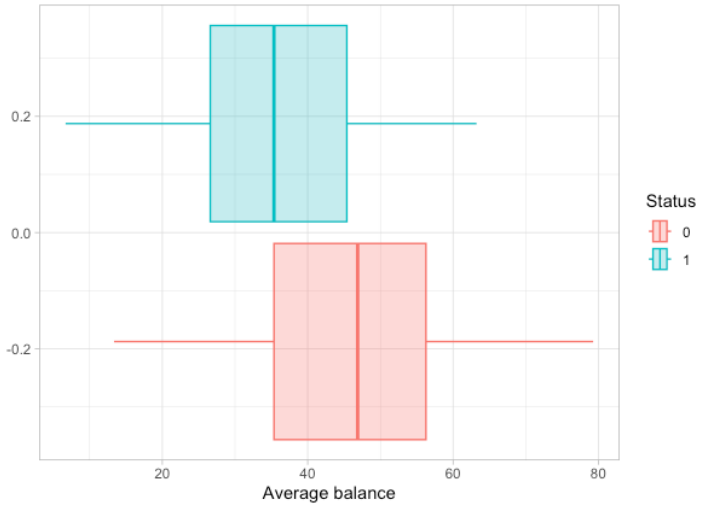

Figura 3.6. Boxplot della variabile *avg_balance* condizionata rispetto alla variabile *status*.

La figura 3.5 mostra come la distribuzione di *avg_balance* univariata risulta essere quasi simmetrica, ma la distribuzione condizionata, visibile alle figure 3.5 e 3.6, mostra che per i clienti solventi questa è traslata verso destra rispetto a quelli insolventi, suggerendo che questi ultimi hanno un bilancio medio inferiore, con uno scarto di circa 10000 corone ceche in mediana.

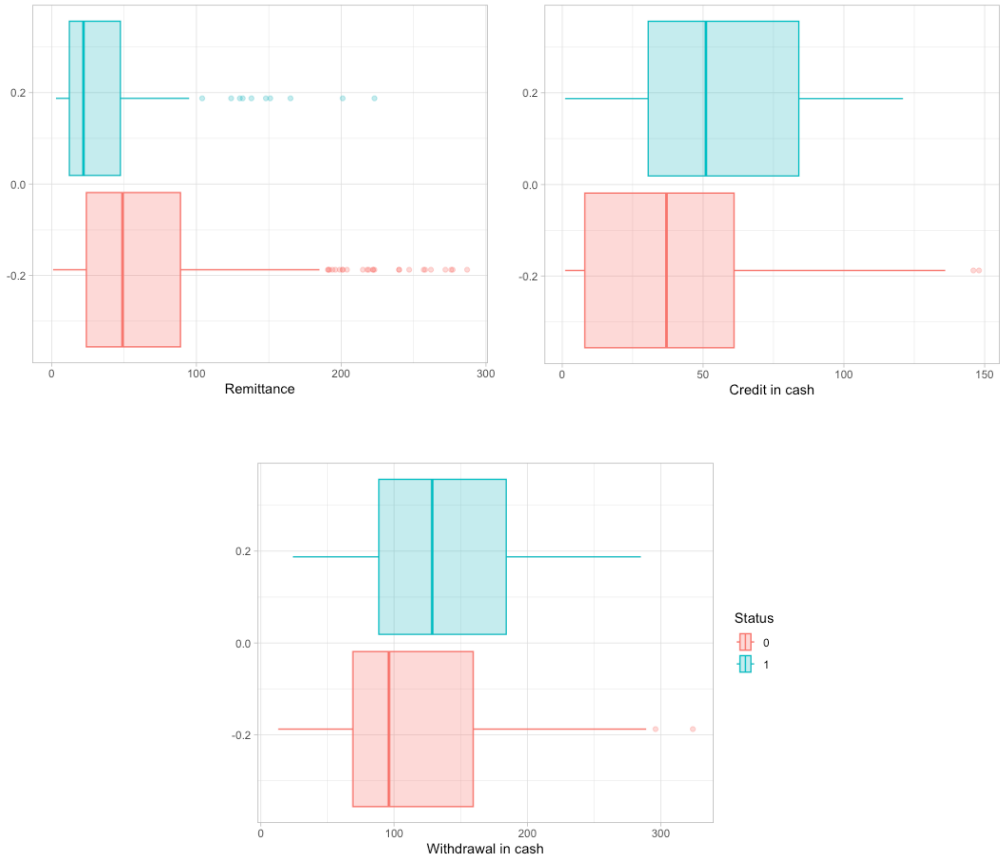

Figura 3.7. Boxplot delle variabili *remittance, cash_c* e *cash_w* condizionate rispetto alla variabile *status*.

I boxplot relativi alle variabili *remittance*, *cash_c* e *cash_w*, alla figura 3.7, mostrano come il comportamento finanziario degli account sia diverso tra utenti solventi e insolventi. Nello specifico, per quel che riguarda *remittance*, la distribuzione appare asimmetrica positiva, con asimmetria più marcata per il gruppo di clienti insolventi, i quali sembrano fare un numero di transazioni di tipo “remittance to another account” inferiore, con uno scarto in mediana di circa 30 operazioni. Si noti, inoltre, il gran numero di outliers per entrambi i gruppi. Per quel che riguarda *cash_c* e *cash_w*, il numero di transazioni “credit

in cash" e "withdrawal in cash" è superiore per i clienti insolventi piuttosto che per quelli solventi. Nello specifico, la variabile *cash_c*, risulta leggermente asimmetrica positiva per gli account che risultano insolventi, a differenza di ciò che accade per quelli solventi e, tra le due distribuzioni, vi è uno scarto in mediana di circa 12 transazioni. Per la variabile *cash_w*, entrambe le distribuzioni condizionate risultano leggermente asimmetriche positive e lo scarto in mediana tra le due è di circa 30 transazioni.

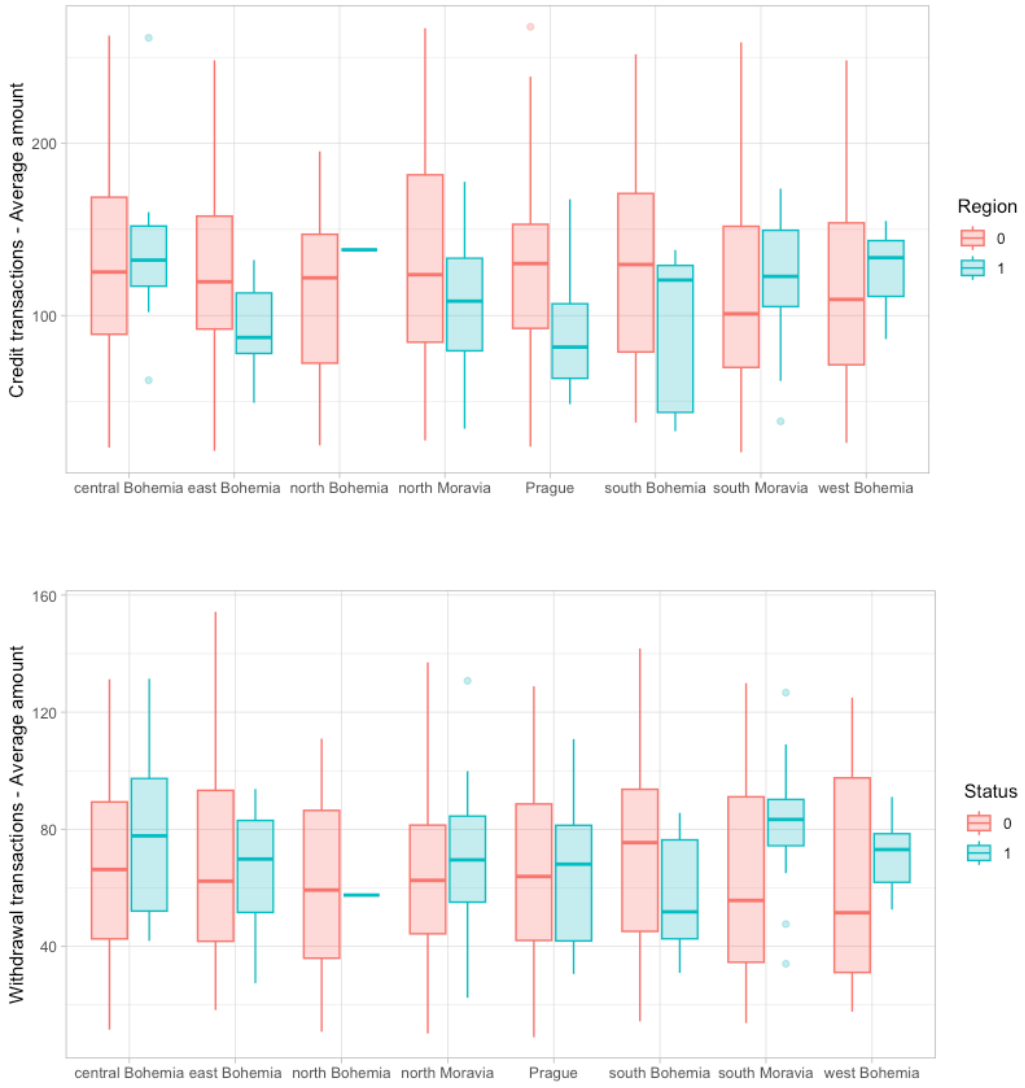

Figura 3.8. Boxplot delle variabili *avg_amtC* e *avg_amtW* condizionate rispetto alle variabili *region* e *status*.

I boxplot condizionati rispetto alla regione all'esito del prestito delle variabili *avg_amtC* e *avg_amtW*, alla figura 3.8, mostrano come il comportamento economico degli account cambi sia in base alla regione che all'esito del prestito. Per le regioni central Bohemia, south Moravia e west Bohemia gli account insolventi mostrano un ammontare medio di transazioni maggiore sia nel caso di operazioni di tipo "credit" che di tipo "withdrawal" rispetto a quelli solventi, il contrario di ciò che accade in south Bohemia. Per le regioni east Bohemia, noth Moravia e Prague, gli account insolventi sembrano fare transazioni di tipo "credit" di ammontare medio inferiore rispetto agli account solventi e transazioni di tipo "withdrawal" di ammontare medio superiore rispetto agli account solventi. Si noti il ridotto numero di account insolventi nella regione east Bohemia.

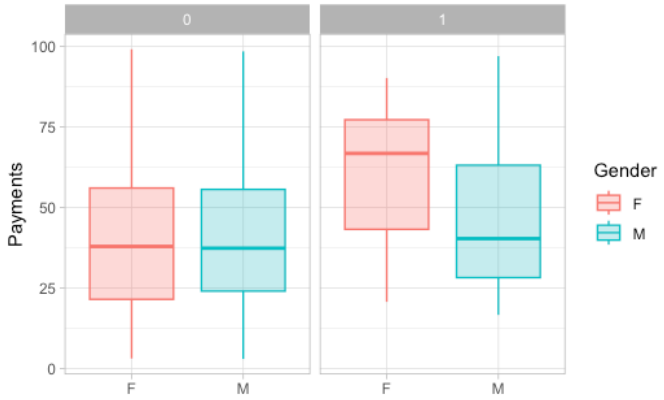

Figura 3.8. Boxplot della variabile *payements* condizionata rispetto alle variabili *gender* e *status*.

La distribuzione di *payments* condizionata rispetto al sesso e all'esito del prestito, alla figura 3.8, mette in luce come, per gli account solventi, l'ammontare mensile della rata si distribuisce pressoché allo stesso modo tra uomini e donne. Discorso diverso vale per gli account insolventi, dove la distribuzione condizionata per le donne presenta un'asimmetria negativa, mentre quella per gli uomini un'asimmetria positiva. Inoltre, l'ammontare della rata per le donne insolventi è maggiore rispetto a quella per gli uomini, con uno scarto in mediana di circa 2500 corone ceche.

3.2 Costruzione del modello

Siccome lo scopo dello studio è la previsione della variabile dicotomica *status*, il problema si riduce ad un problema di classificazione binaria. I modelli applicati sono il modello *logit* e il modello *complementary log-log*, presentati al Capitolo 2. Dato il forte sbilanciamento della variabile di risposta, il partizionamento dei dati in training set e test set è applicato mantenendo la stessa proporzione tra classi dei dati originali. La percentuale di istanze assegnata al training set è l'80%, mentre quella assegnata al test set è il 20%. Un aspetto importante è dato dalla selezione delle variabili da inserire all'interno del modello di regressione, pertanto, per ciascuno dei due modelli costruiti, sono state confrontate misure legate alla bontà di adattamento e all'accuratezza previsiva calcolate considerando diversi criteri di selezione delle variabili, avvalendosi anche del *likelihood ratio test* per i modelli annidati. Successivamente, sui singoli parametri dei modelli scelti, è stato eseguito il test di Wald per valutare la significatività dei coefficienti stimati. I principi di tali strumenti sono stati tutti presenti al Capitolo 2.

3.2.1 Il modello logit

La selezione delle variabili da utilizzare per la costruzione del modello *logit* è stata fatta confrontando i modelli ottenuti inserendo:

a) tutte le variabili di cui alle tabelle dalla 3.9 alla 3.13, escluse quelle rimosse al paragrafo 3.1.2 del presente capitolo ai fini della risoluzione della multicollinearità;
b) solo le variabili statisticamente significative risultanti dal modello (a): *acc_m, avg_balance, avg_amtW, remittance, cash_c, cash_w*;
c) le variabili selezionate dalla *stepwise selection* con criterio di informazione AIC come criterio di selezione: *amount, mun9999, ratio_entr, acc_m, avg_balance, avg_amtW, remittance, cash_c, cash_w*;
d) le variabili i cui test di cui alle tabelle dalla 3.9 alla 3.13 sono risultati statisticamente significativi: *amount, payments, acc_m, avb_balance, avg_amtC, avg_amtW, remittance, cash_c, cash_w*;
e) solo le variabili statisticamente significative risultanti dal modello (d): *amount, avg_balance, avg_amtW, remittance, cash_c, cash_w*;

f) quelle selezionate dalla *stepwise selection* con criterio di informazione AIC come criterio di selezione tra tutte quelle considerate nel modello (d): *amount, acc_m, avg_balance, avg_amtW, remittance, cash_c, cash_w*.

Tabella 3.14. Misure per la valutazione dei modelli *logit* a confronto.

	AIC	**BIC**	**AUC (training set)**	**AUC (test set)**
a)	262.7486	408.662	0.9318	0.8214
b)	245.0067	275.0477	0.8966	0.8601
c)	235.0852	278.0009	0.9105	0.8690
d)	242.8182	285.7339	0.9064	0.8744
e)	240.5584	270.5994	0.8997	0.8827
f)	239.9068	274.2393	0.9034	0.8798

I modelli c) e f) sono quelli che, rispetto agli altri considerati, presentano i livelli di AIC più bassi e di AUC più alti. Il criterio di informazione BIC risulta, invece, più basso per il modello e), ma ciò è dovuto al fatto che, essendo un criterio più restrittivo, porta alla selezione del modello con numero di variabili inferiore. Essendo i modelli c) ed f) annidati, il modello f) nel modello c), per stabilire quale dei due utilizzare è possibile ricorrere al *likelihood ratio test*.

Tabella 3.15. *Likelihood ratio test* per il confronto tra i modelli c) ed f).

	#Df	**LogLik**	**Df**	**Pr(>Chisq)**
Modello c) (completo)	10	-107.54		
Modello f) (ridotto)	8	-111.95	-3	0.01215 *
Signif. codes: 0 ‘***’, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’				

Siccome il *p-value* è minore del livello di α, fissato a 0.05, si rifiuta l’ipotesi nulla del test e, pertanto, si preferisce il modello c), che include, rispetto al modello f), anche le variabili *mun9999* e *ratio_entr*.

3.2.1.1 Il modello logit selezionato

Il modello selezionato contiene variabili provenienti da diverse relazioni. L’unica informazione strettamente legata al prestito emesso è *amount*, ovvero il capitale ceduto;

mentre l'unica informazione relativa all'account è *acc_m*, ovvero il numero di mesi intercorso tra la creazione dell'account e la cessione del prestito. Per quel che riguarda le informazioni di tipo demografico, sono presenti *ratio_entr*, ovvero il numero di imprenditori ogni 1000 abitanti e *mun9999*, ovvero i comuni con un numero di abitanti compreso tra 2000 e 9999. Le restanti variabili selezionate dal modello scelto sono tutte provenienti dalla relazione *trans*: il bilancio medio, l'ammontare medio delle transazioni di tipo "withdrawal", il numero totale di transazioni di tipo "credit in cash", il numero totale di transazioni di tipo "withdrawal in cash" e il numero totale di transazioni di tipo "remittance". Le stime dei coefficienti sono riportare nella seguente tabella.

Tabella 3.16. Stime dei coefficienti del modello *logit*.

	Estimate	**Std. Error**	**Pr(>|z|)**
intercetta	3.285608	1.418731	0.02056 *
amount	0.041694	0.015323	0.00651 **
mun9999	-0.071872	0.048932	0.14189
ratio_entr	-0.025361	0.009018	0.00492 **
acc_m	-0.066763	0.034306	0.05164 .
avg_balance	-0.213364	0.027606	1.08e-14 ***
avg_amtW	0.007222	0.001255	8.60e-09 ***
remittance	-0.008317	0.004882	0.08845 .
cash_c	0.021846	0.007094	0.00207 **
cash_w	0.009238	0.003703	0.01261 *
Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'			

I risultati dei t-test sui parametri del modello mostrano che non tutti i coefficienti sono statisticamente significativi, pertanto, si procede con il confronto con un modello ridotto, che esclude la varibile *mun9999*, con l'applicazione del *likelihood ratio test*.

Tabella 3.17. *Likelihood ratio test* per il confronto tra il modello completo c) di cui alla tabella 3.16 e il modello ridotto.

	#Df	**LogLik**	**Df**	**Pr(>Chisq)**
Modello completo	10	-107.54		
Modello ridotto	9	-108.70	-1	0.1276
Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'				

Essendo il *p-value* maggiore del livello α fissato a 0.05, si accetta l'ipotesi nulla del test e, pertanto, si preferisce il modello ridotto. Le stime del modello ridotto sono le seguenti.

Tabella 3.18. Stime dei coefficienti del modello *logit* ridotto.

	Estimate	**Std. Error**	**Pr(>|z|)**
intercetta	2.241170	1.230312	0.06851 .
amount	0.042212	0.015253	0.00565 **
ratio_entr	-0.020994	0.008661	0.01536 *
acc_m	-0.060684	0.033862	0.01536 *
avg_balance	-0.212175	0.027443	1.06e-14 ***
avg_amtW	0.072201	0.001246	6.85e-09 ***
remittance	-0.007959	0.004836	0.09984 .
cash_c	0.022402	0.007080	0.00155 **
cash_w	0.009063	0.003691	0.01408 *
Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'			

Le stime dei coefficienti ottenuti dal modello *logit* sono espressione dell'effetto della variabile indipendente presa in considerazione sul *logit* della probabilità $P\{Y = 1\}$. Per poter fare considerazioni approfondite è opportuno calcolare gli *odds-ratio*, basati sull'esponenziale dei coefficienti stimati.

Tabella 3.19. *Odds-ratio* del modello *logit* selezionato.

	Odds Ratio	**Lower CI**	**Upper CI**
intercetta	9.404	0.893	113.668
amount	1.043	1.013	1.075
ratio_entr	0.979	0.962	0.995
acc_m	0.941	0.879	1.005
avg_balance	0.809	0.763	0.851
avg_amtW	1.075	1.050	1.103
remittance	0.992	0.982	1.001
cash_c	1.023	1.009	1.038
cash_w	1.009	1.002	1.016

Alle variabili le cui stime dei coefficienti hanno segno positivo, sono associati *odds-ratio* maggiori di 1, viceversa per le variabili i cui coefficienti hanno segno negativo. Per quel che riguarda le informazioni strettamente relative al prestito emesso, per ogni incremento di 10000 corone ceche dell'ammontare di quest'ultimo, la probabilità di insolvenza aumenta del 4.3% rispetto alla probabilità di solvenza; mentre per quel che riguarda informazioni strettamente legate all'account del richiedente, per ogni mese in più che intercorre tra l'apertura dell'account e la richiesta del prestito, la probabilità di insolvenza diminuisce del 5.9% rispetto alla probabilità di solvenza. Circa le informazioni demografiche, per ogni incremento unitario del numero di imprenditori ogni 1000 abitanti, la probabilità di insolvenza diminuisce circa del 2.1% rispetto alla probabilità di insolvenza. Le restanti informazioni, tutte relative all'analisi del comportamento finanziario del soggetto richiedente il prestito, ci suggeriscono che: la probabilità di insolvenza rispetto a quella di insolvenza diminuisce del 19.1% per ogni aumento di 1000 corone ceche del bilancio medio e dello 0.8% per ogni incremento unitario del numero di transazioni di tipo "remittance", mentre aumenta dello 7.5% per ogni aumento di 100 corone ceche dell'ammontare medio delle transazioni di tipo "withdrawal", del 2.3% per ogni incremento unitario del numero di transazioni di tipo "credit in cash" e dello 0.9% per ogni incremento unitario del numero di transazioni di tipo "withdrawal in cash".

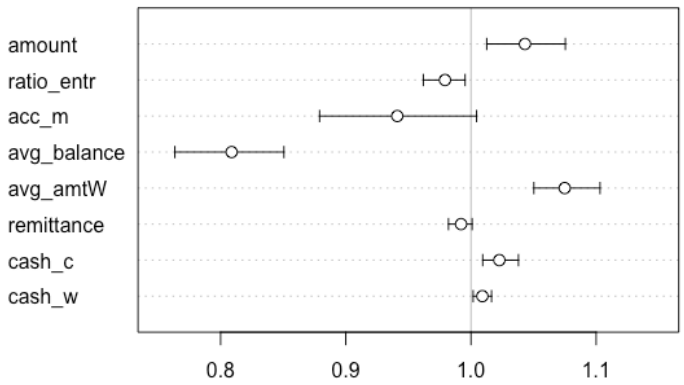

Figura 3.9. Odds-ratio del modello *logit* selezionato.

Un utile strumento per l'interpretazione dell'effetto delle variabili indipendenti sulla variabile di risposta è lo studio degli effetti marginali, ovvero l'effetto della variazione della covariata X_j sulla probabilità $P\{Y = 1\}$, e non sul suo *logit*. Ciò, in termini matematici, si traduce nel calcolo della derivata prima del valore atteso di Y rispetto a X_j. Essendo l'effetto marginale calcolabile per ogni variabile, è possibile stimare sia l'effetto marginale per il valore medio delle variabili sia la media degli effetti marginali. Nello specifico, di seguito si calcola la media degli effetti marginali per ogni variabile.

Tabella 3.20. Effetti marginali del modello *logit* selezionato.

	EM	**Lower CI**	**Upper CI**
amount	0.0025	0.0008	0.0042
ratio_entr	-0.0012	-0.0022	-0.0003
acc_m	-0.0036	-0.0075	0.0003
avg_balance	-0.0126	-0.0150	-0.0101
avg_amtW	0.0043	0.0030	0.0056
remittance	-0.0005	-0.0010	0.0001
cash_c	0.0013	0.0005	0.0021
cash_w	0.0005	0.0001	0.0010

È possibile, a partire da queste stime, fare considerazioni sia in termini di direzione che in termini di magnitudo dell'effetto. La direzione di tale effetto è data dal segno dell'effetto marginale: siccome per le variabili *amount, avg_amtW*, *cash_c* e *cash_w* il segno è positivo,

un incremento della variabile esplicativa è associato ad un aumento della probabilità di insolvenza, mentre per le variabili *ratio_entr, acc_m, avg_balance* e *remittance* vale il contrario. L'entità di tale effetto è tanto maggiore quanto maggiore è il valore assoluto della stima. La variabile *avg_balance* è quella che ha l'effetto maggiore, seguita da *avg_amtW,* e *acc_m*.

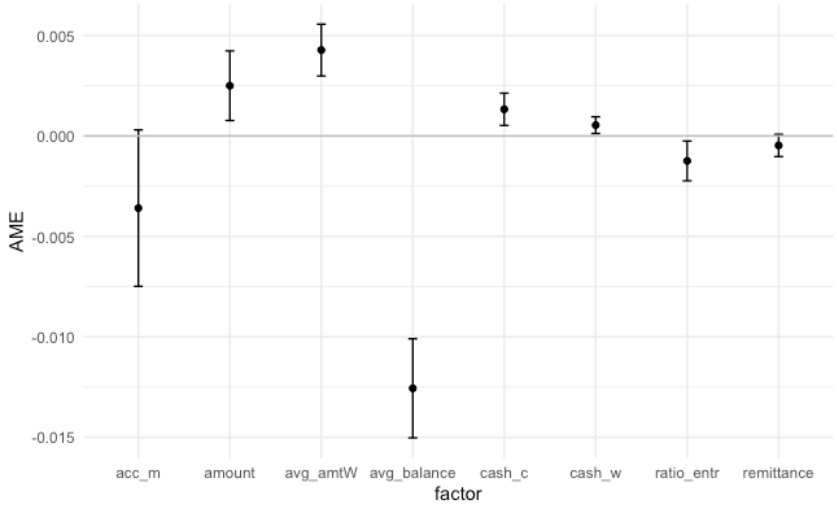

Figura 3.9. Effetti marginali medi per il modello *logit*.

Il motivo dell'instabilità della stima di *avg_balance* è legata alla correlazione di quest'ultima con la variabile *avg_amtW*. Al paragrafo 3.1.2 del presente capitolo è stata illustrata la risoluzione della multicollinearità, che ha riguardato tutte le variabili la cui correlazione con almeno una delle restanti fosse pari o superiore a 0.8. Essendo la correlazione tra le due variabili in questione pari a 0.69, ed essendo queste variabili non altamente correlate a nessuna delle restanti, entrambe sono state inserite nel modello. Per quel che concerne l'instabilità associata alla variabile *acc_m*, si può ipotizzare che sia dovuta alla ridotta dimensione del campione.

Per valutare graficamente l'adattamento del modello ai dati è possibile utilizzare i marginal model plots, alla figura 3.10, che mostrano la variabile di risposta sull'asse delle ordinate e la variabile esplicativa sull'asse delle ascisse. La linea continua blu rappresenta il reale andamento dei dati, mentre quella tratteggiata rossa rappresenta l'andamento del modello. Quanto più le linee sono vicine, tanto maggiore è la capacità del modello di riprodurre i dati. Il modello *logit* costruito, nel complesso, sembra essere capace di

riprodurre bene i dati, così come mostrato dall'ultimo grafico in basso a destra, che considera l'intera componente lineare, tuttavia, ci sono dei problemi di adattemento per le variabili *avg_amtW* e *cash_c*.

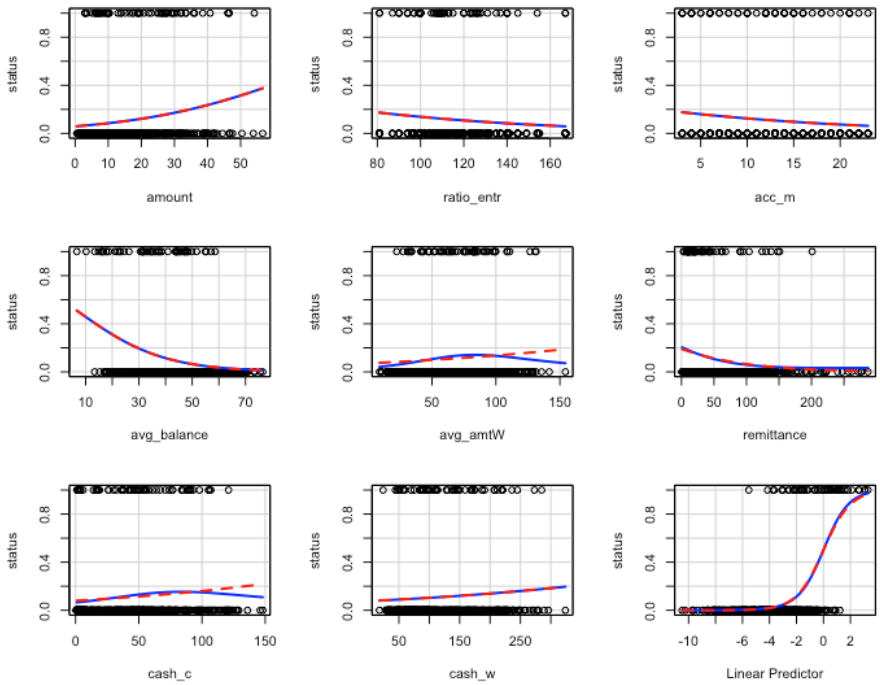

Figura 3.10. Marginal model plots per il modello *logit*.

3.2.2 Il modello complementary log-log

Anche per il modello *complementary log-log* (*c-loglog* nel seguito) la selezione delle variabili da utilizzare per la modellazione è stata fatta confrontando i modelli ottenuti inserendo:

a) tutte le variabili di cui alle tabelle dalla 3.9 alla 3.13, escluse quelle rimosse al paragrafo 3.1.2 del presente capitolo ai fini della risoluzione della multicollinearità;
b) solo le variabili statisticamente significative risultanti dal modello (a): *duration, region, acc_m, avg_balance, avg_amtW, remittance, cash_c, cash_w*;
c) le variabili selezionate dalla *stepwise selection* con criterio di informazione AIC come criterio di selezione: *amount, mun9999, ratio_entr, acc_m, avg_balance, avg_amtW, remittance, cash_c*, *cash_w*;

d) le variabili i cui test di cui alle tabelle dalla 3.9 alla 3.13 sono risultati statisticamente significativi: *amount, payments, acc_m, avb_balance, avg_amtC, avg_amtW, remittance, cash_c, cash_w*;
e) solo le variabili statisticamente significative risultanti dal modello (d): *amount, acc_m, avb_balance, avg_amtW, remittance, cash_c, cash_w*;
f) quelle selezionate dalla *stepwise selection* con criterio di informazione AIC come criterio di selezione tra tutte quelle considerate nel modello (d): *amount, acc_m, avg_balance, avg_amtW, remittance, cash_c, cash_w*.

Tabella 3.21. Misure per la valutazione dei modelli *c-loglog* a confronto.

	AIC	**BIC**	**AUC (training set)**	**AUC (test set)**
a)	258.7722	404.6855	0.9274	0.8339
b)	244.0744	321.3227	0.9149	0.8333
c)	230.3594	273.2751	0.9091	0.8714
d)	240.5513	283.467	0.9047	0.8792
e)	237.2247	271.5573	0.9030	0.8845
f)	237.2247	271.5573	0.9030	0.8845

I due modelli derivati con la *stepwise selection*, i modelli c) ed f), si rivelano, anche in questo caso, quelli che rappresentano il miglior compromesso tra AIC e AUC, e quindi tra bontà di adattamento e accuratezza previsiva. Si noti che, in realtà, il modello selezionato al punto e) è lo stesso di cui al punto f). Ancora una volta, essendo i modelli annidati, il modello f) nel modello c), per stabilire quale dei due utilizzare è possibile ricorrere al *likelihood ratio test*.

Tabella 3.22. *Likelihood ratio test* per il confronto tra il modello completo c) di cui alla tabella 3.21 e il modello ridotto.

	#Df	**LogLik**	**Df**	**Pr(>Chisq)**
Modello c) (completo)	10	-105.18		
Modello f) (ridotto)	8	-110.61	-2	0.004371 **
Signif. codes: 0 ‘***’, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’				

Essendo il *p-value* minore del livello di significatività 0.05, si rifiuta l'ipotesi nulla del test e si preferisce il modello c), che include, rispetto al modello f), anche le variabili *mun9999* e *ratio_entr*.

3.2.2.1 Il modello complementary log-log selezionato

Come per il modello selezionato al paragrafo 3.2.2 del presente capitolo, il modello *complementary log-log* selezionato contiene variabili provenienti da diverse relazioni. Nello specifico, alle variabili selezionate per il modello *logit*, in questo si aggiunge la variabile *mun9999* proveniente dalla relazione *district*, che rappresenta il numero di comuni con un numero di abitanti compreso tra 2000 e 9999. Le stime dei coefficienti sono riportare nella seguente tabella.

Tabella 3.23. Stime dei coefficienti del modello *c-loglog*.

	Estimate	**Std. Error**	**Pr(>|z|)**
intercetta	2.735111	1.164952	0.01888 *
amount	0.035526	0.012118	0.00337 **
mun9999	-0.074499	0.042121	0.07695 .
ratio_entr	-0.023671	0.007495	0.00159 **
acc_m	-0.057720	0.027584	0.03639 *
avg_balance	-0.185086	0.021572	< 2e-16 ***
avg_amtW	0.061490	0.009853	4.36e-10 ***
remittance	-0.007775	0.004011	0.05255 .
cash_c	0.018609	0.005881	0.00155 **
cash_w	0.008872	0.002966	0.00278 **
Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'			

Le stime dei coefficienti ottenuti dal modello *c-loglog* sono espressione dell'effetto della variabile indipendente presa in considerazione sul logaritmo di meno il logaritmo del complemento ad uno della probabilità $P\{Y = 1\}$, ovvero sul logaritmo di meno il logaritmo della probabilità $P\{Y = 0\}$. Nel caso del modello *c-loglog* non è possibile ricavare una nozione di *odds-ratio* comparabile a quelli ottenuti per il modello logistico, tuttavia, spesso viene calcolato l'esponenziale delle stime dei coefficienti per valutare l'effetto delle variabili indipendenti su meno il logaritmo della probabilità $P\{Y = 0\}$.

Tabella 3.24. Esponenziale dei coefficienti del modello *c-loglog*.

	Exp(Estimate)	**Lower CI**	**Upper CI**
intercetta	15.411	1.550	167.429
amount	1.036	1.012	1.060
mun9999	0.928	0.851	1.006
ratio_entr	0.977	0.962	0.991
acc_m	0.944	0.891	0.998
avg_balance	0.831	0.795	0.866
avg_amtW	1.063	1.044	1.084
remittance	0.992	0.983	1.000
cash_c	1.019	1.008	1.031
cash_w	1.009	1.003	1.015

Esattamente come per gli *odds-ratio*, per le proprietà della funzione esponenziale, alle variabili le cui stime dei coefficienti hanno segno positivo, sono associati esponenziali maggiori di 1, viceversa per le variabili i cui coefficienti hanno segno negativo. Per quel che riguarda le informazioni relative al prestito emesso, per ogni incremento di 10000 corone ceche dell'ammontare di quest'ultimo, la probabilità di insolvenza aumenta; mentre per quel che riguarda informazioni legate all'account del richiedente il prestito, per ogni mese in più che intercorre tra l'apertura dell'account e la richiesta del prestito, la probabilità di insolvenza diminuisce. Soffermandoci sulle informazioni demografiche, si deduce che per ogni incremento unitario del numero di imprenditori ogni 1000 abitanti, la probabilità di insolvenza diminuisce, così come diminuisce per ogni incremento unitario del numero di comuni con numero di abitanti compreso tra 2000 e 9999. Le restanti informazioni sono relative all'analisi del comportamento finanziario del soggetto richiedente il prestito e ci suggeriscono che: la probabilità di insolvenza diminuisce per ogni aumento di 1000 corone ceche del bilancio medio e per ogni incremento unitario del numero di transazioni di tipo "remittance", mentre aumenta per ogni aumento di 100 corone ceche dell'ammontare medio delle transazioni di tipo "withdrawal", per ogni incremento unitario del numero di transazioni di tipo "credit in cash" e per ogni incremento unitario del numero di transazioni di tipo "withdrawal in cash".

Come nel caso del modello *logit*, è possibile avvalersi dello studio degli effetti marginali per valutare l'effetto delle variabili indipendenti sulla variabile di risposta. Si ricorda che si

tratta dell'effetto della variazione della covariata X_j sulla probabilità $P\{Y = 1\}$, e quindi della derivata prima del valore atteso di Y rispetto a X_j. Anche in questo caso si sceglie di calcolare la media degli effetti marginali per ogni osservazione e non l'effetto marginale per il valore medio delle variabili.

Tabella 3.25. Effetti marginali del modello *c-loglog* selezionato.

	EM	**Lower CI**	**Upper CI**
amount	0.0025	0.0008	0.0041
mun9999	-0.0051	-0.0108	0.0005
ratio_entr	-0.0016	-0.0026	-0.0006
acc_m	-0.0040	-0.0077	-0.0003
avg_balance	-0.0128	-0.0150	-0.0106
avg_amtW	0.0042	0.0030	0.0054
remittance	-0.0005	-0.0011	-0.0000
cash_c	0.0013	0.0005	0.0021
cash_w	0.0006	0.0002	0.0010

Come già detto precedentemente, è possibile, a partire da queste stime, fare considerazioni sia in termini di direzione che in termini di magnitudo dell'effetto. Siccome per le variabili *amount, avg_amtW*, *cash_c* e *cash_w* il segno è positivo, un incremento della variabile esplicativa è associato ad un aumento della probabilità di insolvenza, mentre per le variabili *mun9999*, *ratio_entr, acc_m, avg_balance* e *remittance* vale il contrario. L'entità di tale effetto è tanto maggiore quanto maggiore è il valore assoluto della stima. Anche in questo modello, la variabile *avg_balance* è quella che ha l'effetto maggiore, seguita da *mun9999,* e *acc_m*.

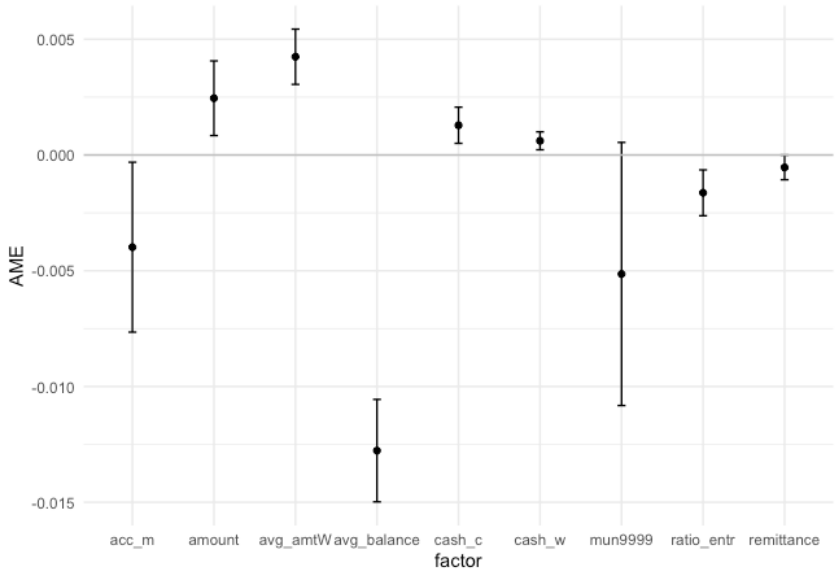

Figura 3.11. Effetti marginali del modello *c-loglog*.

Come mostrato per il caso del modello *logit*, l'instabilità della stima di *avg_balance* è legata alla correlazione di quest'ultima con la variabile *avg_amtW*. Per quel che concerne l'instabilità associata alle variabili *mun9999* e *acc_m*, si può ipotizzare anche in questo caso che ciò sia dovuto alla ridotta dimensione del campione.

Per valutare graficamente l'adattamento del modello ai dati è possibile, anche per questo modello prendere in considerazione i marginal model plots, alla figura 3.12. Come per il modello *logit*, ci sono dei problemi di adattamento per le variabili *avg_amtW* e *cash_c*, ma dall'ultimo grafico centrale pare che il modello *c-loglog*, nel complesso, riesca a riprodurre meglio i dati rispetto al modello *logit*, in quanto le due curve sembrano combaciare perfettamente.

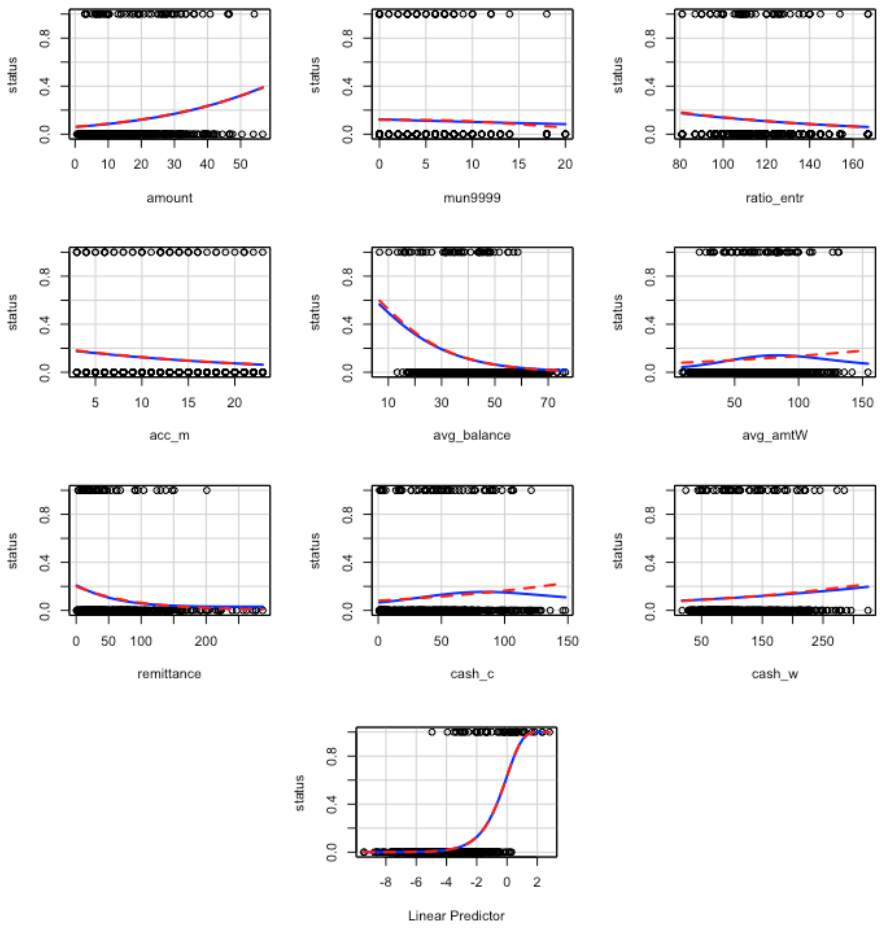

Figura 3.12. Marginal model plots per il modello *c-loglog*.

3.2.2.2 Il modello di regressione GEV

Come sottolineato al paragrafo 2.5 del Capitolo 2, l'utilizzo della distribuzione di Gumbel fa un'assunzione sul parametro di forma della distribuzione GEV, assumendo che questo tenda a 0. La formulazione generale [55] del capitolo precedente ha il vantaggio di rimuovere la necessità di ipotesi a priori sulla distribuzione esatta da adottare, permettendo che siano i dati stessi a determinare il comportamento della coda della distribuzione. Il presente studio ha incluso questo passaggio e il parametro selezionato, prossimo allo 0, ha confermato che l'utilizzo della distribuzione di Gumbel è appropriato.

3.3 Confronto tra i modelli

I modello costruiti vengono ora confrontati sia in termini di bontà di adattamento che di accuratezza previsiva. Gli strumenti utilizzati per valutare la bontà d'adattamento dei modelli sono: il criterio di informazione di AIC, la devianza residua e gli indici pseudo-R^2. Tutti e tre gli strumenti scelti confermano che il modello che si adatta meglio ai dati è il modello *complementary log-log*.

Tabella 3.26. Misure per il confronto tra i modelli *logit* e *c-loglog* selezionati.

	AIC	**Devianza Residua**	**R^2 Mc Fadden**	**R^2 Nagelkerke**	**R^2 CoxSnell**
Logit	235.41	217.4062	0.4229258	0.5087337	0.2555132
C-loglog	230.36	210.3594	0.4416304	0.5279512	0.2651653

Dei modelli costruiti viene ora valutata la capacità di prevedere la probabilità di insolvenza su nuove osservazioni. Le capacità previsive del modello sono state testate sul set di dati non utilizzato nella fase di addestramento del modello stesso. Una prima misura della capacità previsiva del modello è data dalla curva ROC e dall'AUC, presentati al paragrafo 2.8.1 del Capitolo 2. L'AUC calcolato sul test set del modello *logit* è pari a 0.8655, mentre l'AUC calcolato sul test set del modello *c-loglog* è pari a 0.8714, mostrando che entrambi i modelli sono altamente accurati, ma il modello *complementary log-log* risulta leggermente migliore in termini di accuratezza previsiva.

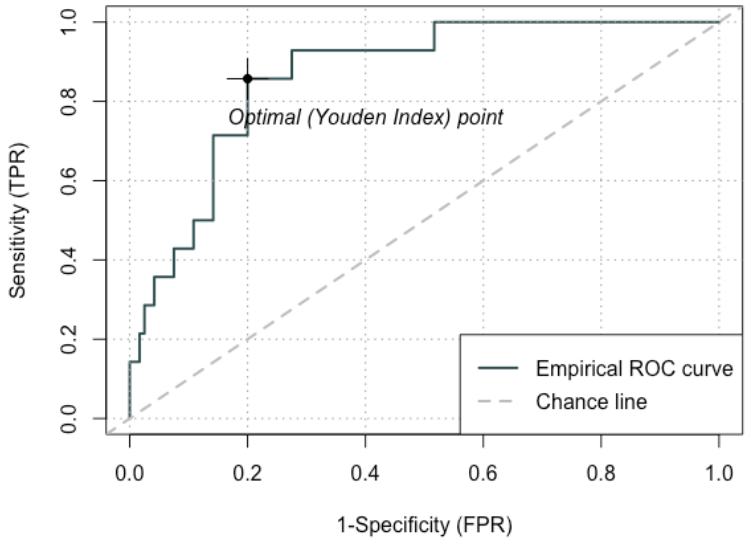

Figura 3.13. Curva ROC del modello *logit*.

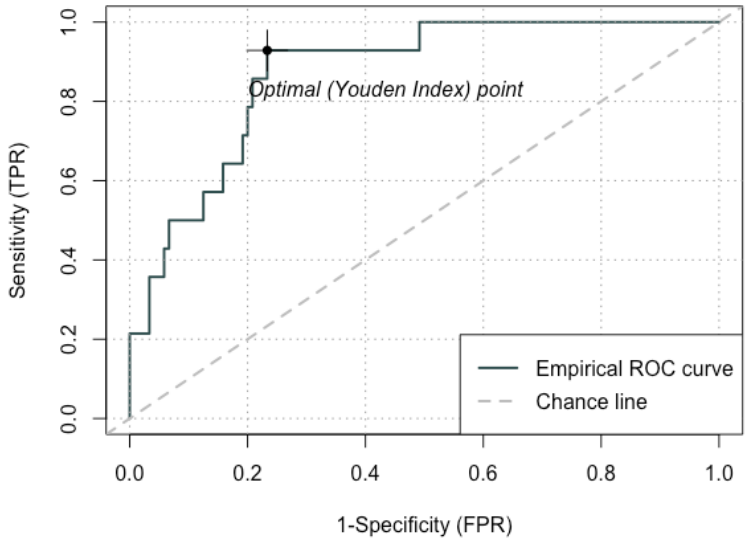

Figura 3.14. Curva ROC del modello *c-loglog*.

Per le considerazioni fatte al paragrafo 2.8 del Capitolo 2, non sempre la *threshold* ottimale è quella canonicamente fissata a 0.50, pertanto la classificazione è stata testata anche sulla

soglia 0.11, pari alla percentuale di osservazioni per cui *status* = 1, e sulla soglia calcolata sulla base dell'indice di Youden, ovvero la soglia che massimizza la funzione di *sensitivity* e quella di *specificity*. La soglia ottimale calcolata sulla base dell'indice di Youden assume valore 0.0988 per il modello *logit* e valore 0.0672 per il modello *complementary log-log*. I risultati sono di seguito riportati.

Tabella 3.27. Misure di accuratezza per il confronto tra i modelli *logit* e *c-loglog* selezionati.

Misure	Soglia 0.50		Soglia 0.11		Soglia di Youden	
	logit	*cloglog*	*logit*	*cloglog*	*logit*	*cloglog*
True Negative	87.31%	88.06%	73.13%	74.63%	71.64%	68.66%
True Positive	2.99%	2.24%	7.46%	6.72%	8.21%	8.95%
False Negative	7.46%	8.21%	2.99%	3.73%	2.24%	1.49%
False Positive	2.24%	1.49%	16.42%	14.93%	17.91%	20.90%
Accuratezza	90.30%	90.30%	80.60%	81.34%	79.85%	77.61%
Sensitivity	28.57%	21.43%	71.44%	64.29%	78.57%	85.71%
Specificity	97.50%	98.33%	81.67%	83.3%	80.00%	76.67%

Si noti come al variare della soglia, andando dal valore più alto, pari 0.5, a quello più basso, dato dall'indice di Youden, le percentuali di True Negative e False Negative, diminuiscono, al contrario di quel che accade con le percentuali di True Positive e False Positive, e ciò risulta vero per entrambi i modelli. La causa di ciò è il forte sbilanciamento che caratterizza i dati, il quale porta ad errori di classificazione a sfavore della classe meno rappresentata, in questo caso la classe dei positivi e quindi degli account insolventi. Abbassando la soglia di classificazione, si aumenta artificialmente la probabilità di assegnare le osservazioni alla classe dei positivi. Si può notare, inoltre, che il modello *complementary log-log* presenta valori di *specificity* maggiori rispetto al modello *logit* sia per la soglia pari a 0.5 che per la soglia pari a 0.11, mentre, considerando come soglia l'indice di Youden, a migliorare è la *sensitivity*, ovvero la frazione dei positivi correttamente classificati. Nonostante l'indice AUC calcolato su questo modello sia maggiore rispetto a quello calcolato per il modello *logit*, ciò non si traduce in un netto miglioramento della capacità previsiva, ma è da tenere a mente che il presente studio ha un limite legato al ridotto numero di osservazioni presenti. Sebbene l'accuratezza maggiore sia quella relativa alla soglia 0.5, bisogna considerare anche il contesto di applicazione del modello e gli effetti dati dalla misclassificazione

relativa a una classe piuttosto che all'altra. Nel caso della previsione di insolvenza, bisogna ponderare i costi relativi alla cessione di un prestito che risulterà insolvente e il mancato guadagno derivante dalla cattiva classificazione di un potenziale cliente che in realtà si sarebbe rivelato solvente. In sintesi, non esiste una scelta corretta a priori, bensì dipende dalle scelte di gestione. Volendo massimizzare la *specificity*, ovvero la frazione dei negativi correttamente classificati, bisognerebbe scegliere la soglia 0.5, mentre volendo massimizzare la *sensitivity*, ovvero la frazione dei positivi correttamente classificati, bisognerebbe scegliere la soglia data dall'indice di Youden.

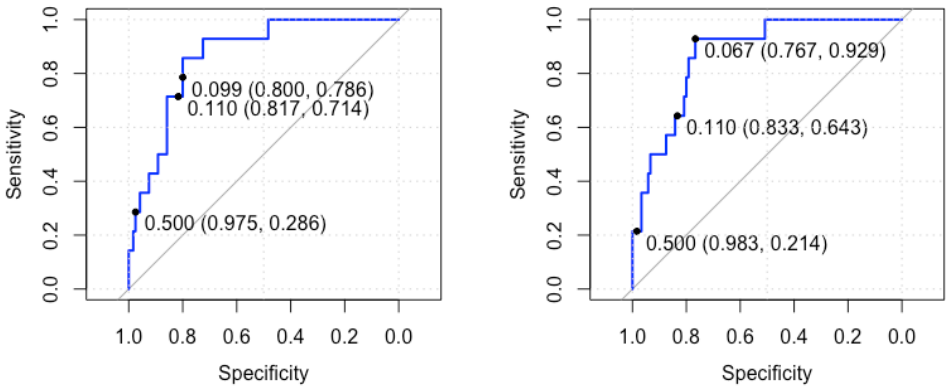

Figura 3.15. Curva ROC con i diversi livelli di *threshold* per il modello *logit* a destra e per il modello *c-loglog* a sinistra.