Conoscenza ed utilizzo degli strumenti FinTech

Antonio Cola

1 Introduzione

1.1 Descrizione dell'indagine

L'indagine statistica in esame ha come obiettivo quello di valutare il livello di conoscenza ed utilizzo degli strumenti FinTech da parte del campione selezionato. I risultati di questa ricerca forniranno un'immagine precisa della diffusione e dell'utilizzo di queste tecnologie finanziarie innovative e aiuteranno a comprendere come le persone interagiscono con esse. La raccolta dei dati è stata effettuata attraverso un questionario online e l'analisi dei risultati è stata condotta utilizzando tecniche statistiche appropriate. Questa indagine è di particolare importanza in un mondo sempre più digitale e in rapida evoluzione, dove le FinTech stanno giocando un ruolo sempre più importante nell'industria finanziaria.

1.2 Descrizione del dataset

Dopo aver letto il questionario, è stato possibile operare sul dataset costituito da 78 variabili di diversa natura e 625 osservazioni, ognuna delle quali rappresenta un'unità di dati raccolti per ogni variabile. In totale, il dataset include 48750 dati. Tuttavia, un solo dato mancante è stato identificato, il che significa che una singola unità di informazione non è disponibile per una specifica variabile. Questo dato è risultato non rilevante ai fini dell'analisi.

2 Analisi Esplorativa

2.1 Individuazione delle variabili d'interesse

La prima fase dell'analisi esplorativa ha riguardato l'identificazione delle variabili dipendenti ed esplicative. Grazie al questionario, è stato possibile rapidamente identificare 33 possibili variabili dipendenti raggruppate in 5 batterie di domande, concernenti la conoscenza di strumenti FinTech. Dopo un'attenta valutazione sono state scelte 6 variabili dipendenti, ordinali con modalità da 0 a 6, appartenenti alla stessa batteria di domande. Tali variabili fanno riferimento al livello di conoscenza di: *crowdfunding, cryptocurrencies, instantinsurance, instantpayments, roboadvisor, p2p.*
Tra le restanti 45 variabili molte sono state scartate. Alcune a causa della variabilità quasi nulla, come la variabile dicotomica *italia* relativa alla residenza in Italia che presenta solo 11 osservazioni di 625 per la modalità *"non residente in Italia"*. Altre in quanto mutabili sconnesse come la variabile *areastudio* relativa all'area disciplinare di appartenenza. Altre ancora per la presenza di una modalità che ne ha compromesso l'utilizzo, come la variabile *reddito* che presenta la modalità *"Preferisco non rispondere"*, che seppur trattata come dato mancante non risulta significativa.
Dopo questa fase è stato più semplice individuare le variabili esplicative: *genere, eta, laureati, lavoratori, finemese, domande1, domande2.* La variabile *genere* è una variabile dicotomica che intuitivamente fa riferimento al genere dell'intervistato e presenta modalità 0 per gli uomini e modalità 1 per le donne. La variabile *eta* è una variabile numerica che appunto fa riferimento all'età dell'individuo e assume valori compresi tra 18 e 67. La variabile *laureati* è anch'essa una variabile dicotomica ottenuta dicotomizzando la variabile *studio* che in origine presentava 5 modalità, e successivamente modalità 0 e 1 rispettivamente per i non laureati e laureati. Anche la variabile *lavoro* è stata dicotomizzata alla stesso modo, utilizzando la modalità 0 per i non lavoratori e la modalità 1 per i lavoratori, in origine la variabile presentava 7 modalità. La variabile *finemese* invece è una variabile ordinale che presenta modalità da 1 a 10 e fa riferimento al livello di benessere economico, ossia con che facilità si arriva a finemese. Le variabili *domande1* e *domande2* sono il risultato di una combinazione lineare

di più variabili e presentano rispettivamente modalità da 0 a 6 e modalità da 3 a 18. Nello specifico la variabile *domande1* è una variabile quantitativa discreta ed è stata ottenuta sommando 6 variabili dicotomiche, relative alle risposte di quesiti a risposta multipla, con modalità 0 per risposte sbagliate e modalità 1 per risposta esatta. La variabile *domande2* invece è una variabile ordinale ottenuta dalla somma di 3 variabili ordinali, relative alle risposte di quesiti di preferenza, con modalità da 1 a 6, dove 1 rappresenta una minore confidenza dell'argomento o materia e 6 un livello di confidenza maggiore.
Questa fase iniziale ha permesso di comprendere meglio la struttura del dataset e di preparare la base per ulteriori analisi quantitative. L'identificazione delle variabili dipendenti ed esplicative è un passo cruciale per la comprensione dei dati e per l'elaborazione di conclusioni valide sui trend e sulle relazioni presenti nel dataset.

2.2 Distribuzione delle variabili

Conoscere la distribuzione delle variabili in un dataset è importante, poiché permette di comprendere la struttura dei dati e di valutare la presenza di eventuali anomalie o pattern. Questo a sua volta può aiutare a prendere decisioni riguardo alle tecniche di analisi da utilizzare e a formulare conclusioni affidabili sul dataset.
Le prime distribuzioni ad essere analizzate sono quelle relative alle variabili dipendenti come mostrano i *barplot* riportati in figura:

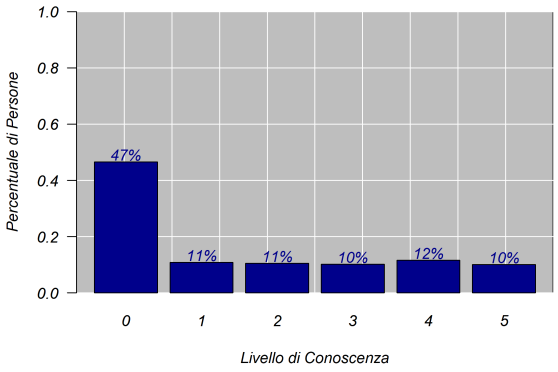

Figura 1: Conoscenza *Crowdfunding*

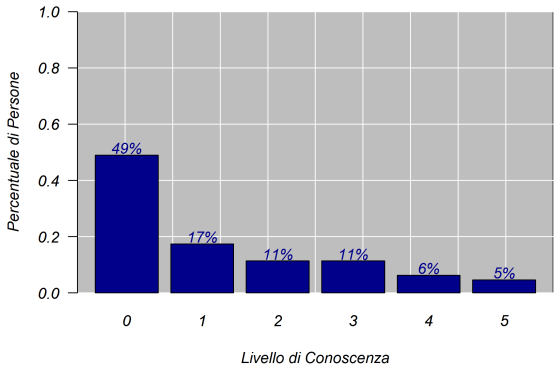

Figura 2: Conoscenza *Cryptocurrencies*

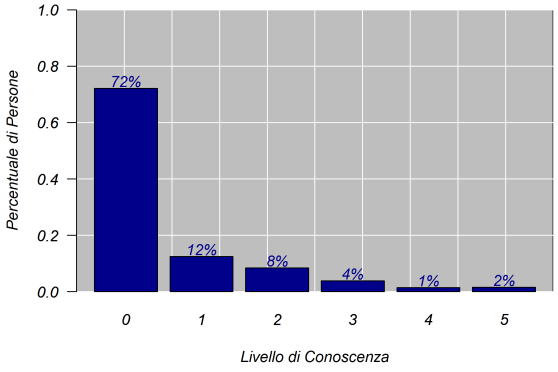

Figura 3: Conoscenza *Instantinsurance*

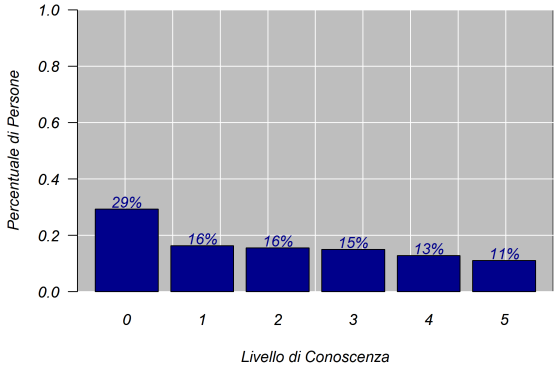

Figura 4: Conoscenza *Instantpayments*

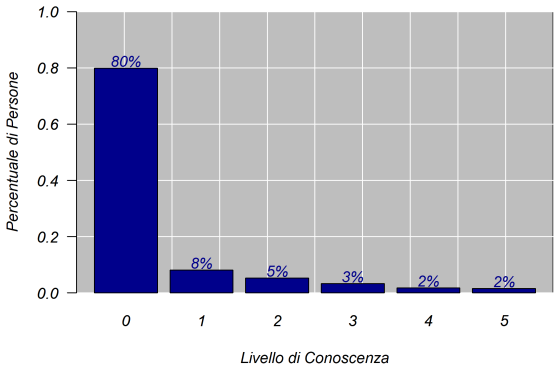

Figura 5: Conoscenza *Roboadvisor*

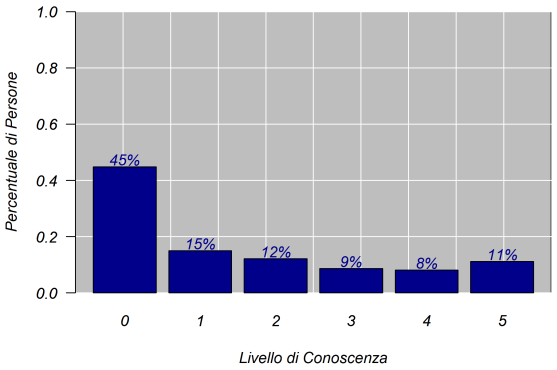

Figura 6: Conoscenza *p2p*

I grafici riportati evidenziano una distribuzione altamente disomogenea, ovvero con molte osservazioni con modalità 0 e pochi valori più alti, con una maggiore evidenza per le variabili: *instantinsurance* e *roboadvisor*. È possibile quindi affermare che la maggior parte delle persone intervistate non ha alcuna conoscenza dei seguenti strumenti FinTech: *crowdfunding, cryptocurrencies, instantinsurance, roboadvisor, p2p*. Lo strumento *instantpayments* è sicuramente più noto alla popolazione, ma presenta ugualmente una prevalenza di modalità 0, tuttavia meno marcata. Questo risultato motiva la scelta di proseguire con dei *Two-Part Models* che verranno analizzati successivamente.
La distribuzione delle variabili dipendenti è stata analizzata anche rispetto a tutte le variabili esplicative.
Rispetto alla variabile *genere* di seguito sono riportati alcuni dei grafici più significativi:

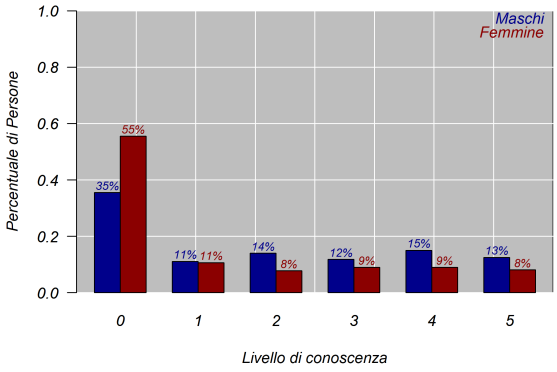

Figura 7: Conoscenza *Crowdfunding* rispetto al *Genere*

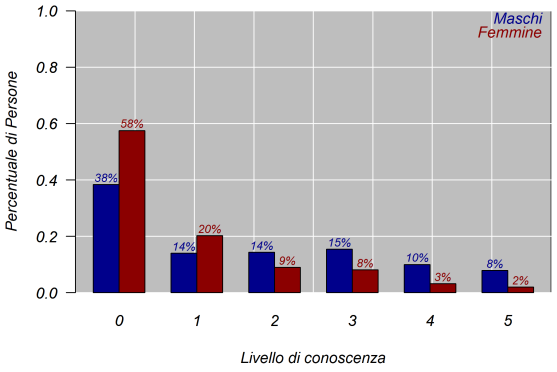

Figura 8: Conoscenza *Cryptocurrencies* rispetto al *Genere*

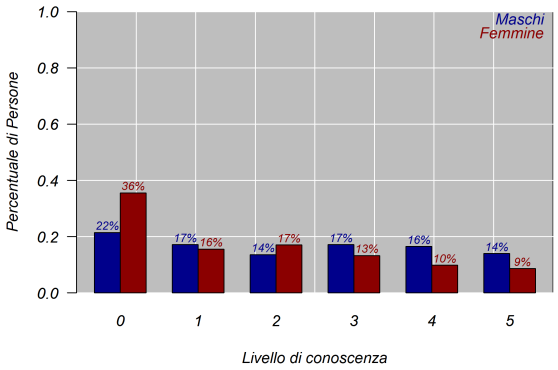

Figura 9: Conoscenza *Instantpayments* rispetto al *Genere*

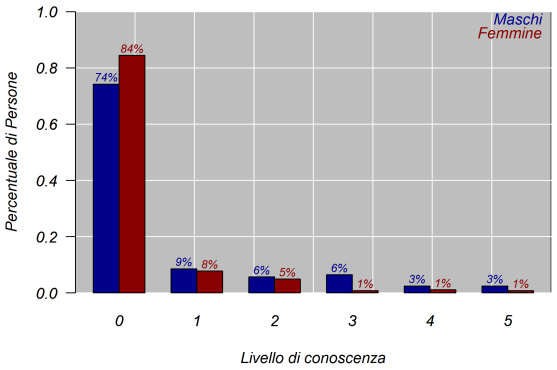

Figura 10: Conoscenza *Roboadvisor* rispetto al *Genere*

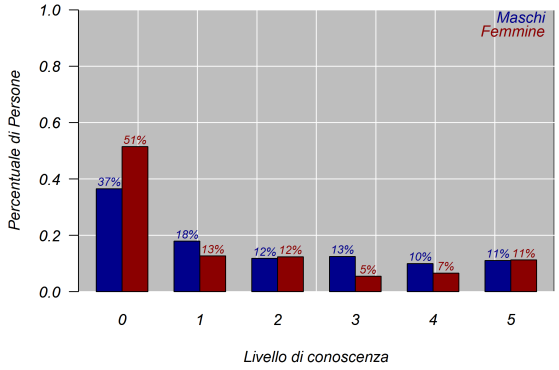

Figura 11: Conoscenza *p2p* rispetto al *Genere*

Si può notare dai precedenti barplot una visiva differenza tra le frequenze delle modalità della variabile *genere*, che è stata verificata con dei test sulle proporzioni. Il test sulle proporzioni è un test statistico per verificare se la differenza tra le frequenze relative delle modalità di una variabile dicotomica è significativa o meno. Il risultato ottenuto dai test, su ciascuna delle distribuzioni precedenti, conferma che le frequenze relative non sono le medesime nelle popolazioni di riferimento, con un'elevata significatività dettata da un valore di $p\text{-}value < 0.001$ in tutti i casi. È possibile quindi affermare che le donne sono più impreparate degli uomini, rispetto alla conoscenza degli strumenti FinTech considerati. Questo risultato suggerisce che la variabile *genere* potrebbe risultare significativa in un successivo modello.
Lo stesso processo, effettuato sulle altre variabili esplicative, ha prodotto un risultato interessante solo nel caso della variabile *laureati*:

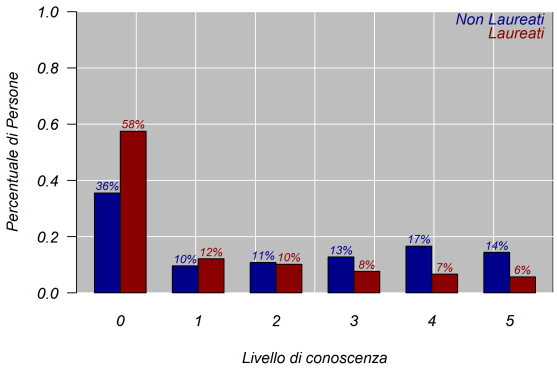

Figura 12: Conoscenza *Crowdfunding* rispetto ai *Laureati*

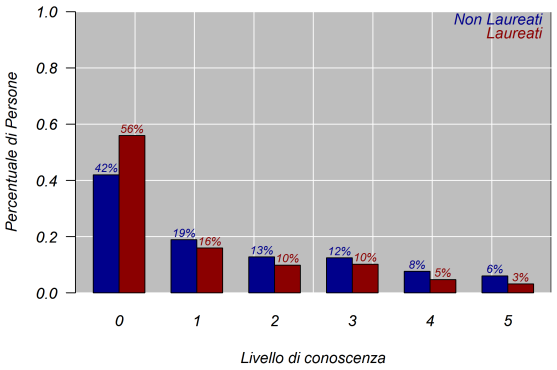

Figura 13: Conoscenza *Cryptocurrencies* rispetto ai *Laureati*

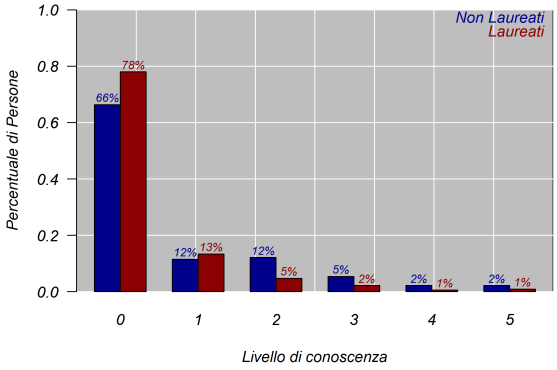

Figura 14: Conoscenza *Instantinsurance* rispetto ai *Laureati*

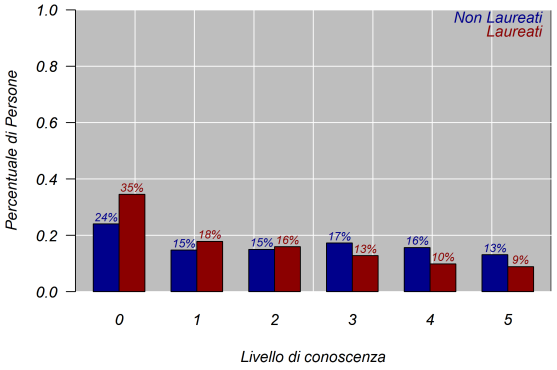

Figura 15: Conoscenza *Instantpayments* rispetto ai *Laureati*

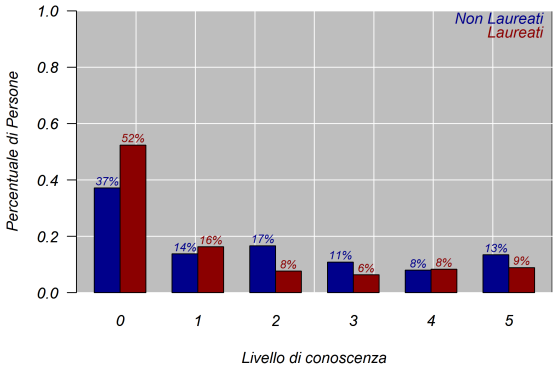

Figura 16: Conoscenza *p2p* rispetto ai *Laureati*

Anche in questo caso è immediata una visiva differenza tra le frequenze delle modalità della variabile *laureati*. Il test sulle proporzioni eseguito sulle distribuzione ha prodotto come risultato una conferma della precedente affermazione, con una forte significatività dettata da un valore di $p\text{-}value < 0.02$ in tutti i casi. È possibile quindi affermare che vi è una differenza nella conoscenza degli strumenti FinTech in

relazione inversa al grado d'istruzione. Il risultato prodotto, tutt'altro che intuitivo, potrebbe suggerire un buon impatto della variabile *laureati* in un successivo modello.

2.3 Correlazione

La conoscenza della *correlazione* tra variabili è importante, poiché ci permette di identificare la relazione tra due o più variabili quantitative. Quest'informazione è utile per comprendere il comportamento dei dati e può essere utilizzata per prevedere il comportamento di una variabile in base alla conoscenza di un'altra. In questo caso sono state analizzate le correlazioni tra le sole variabili dipendenti. I risultati sono riportati nella seguente *matrice di correlazione*:

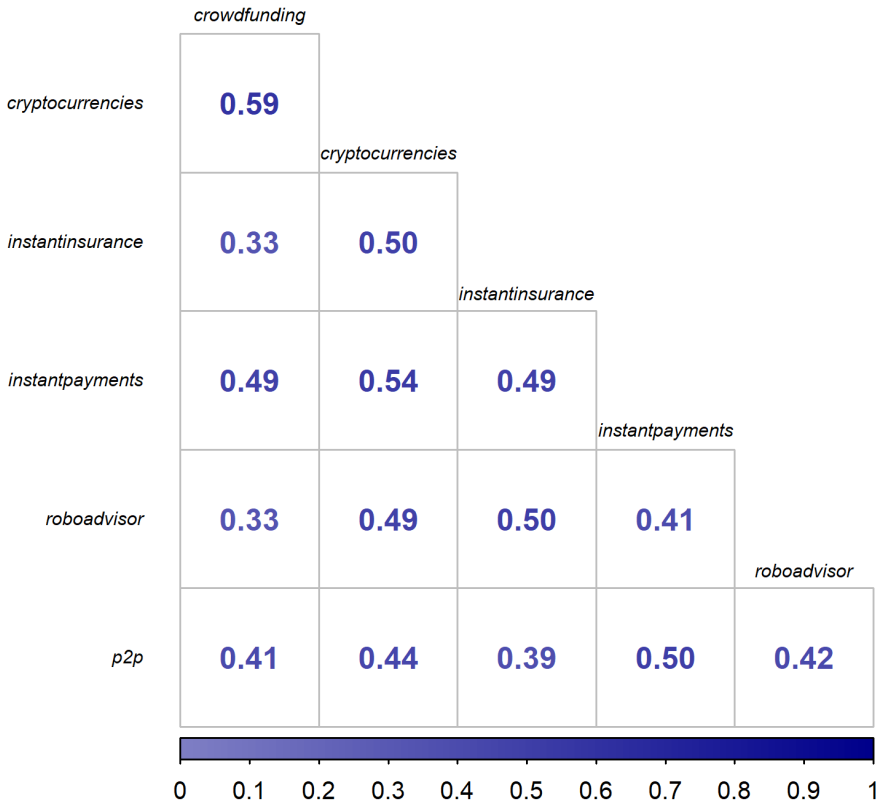

Figura 17: Matrice di Correlazione

Generalmente la matrice di correlazione è una *matrice simmetrica*, ossia una matrice quadrata che è uguale alla sua trasposta. Ciò significa che l'elemento *(i,j)* nella matrice è uguale all'elemento *(j,i)* nella stessa matrice. In questo caso, per evitare dati ridondanti, la matrice di correlazione è stata ridotta ad una triangolare inferiore. Successivamente sono stati eliminati i dati, con valore unitario, presenti sulla diagonale principale. I valori delle correlazioni delle variabili dipendenti sono tutti positivi, con un minimo di 0.33 e un massimo 0.59. Si può quindi affermare che vi è una dipendenza lineare positiva tra tutte le variabili dipendenti a due a due. In altre parole una variabile dipendente aumenta all'aumentare di un'altra, in misura del coefficiente di correlazione, ricordando che quest'ultimo è: $-1 < \rho < 1$.

3 Modelli

3.1 Two-Part Models

I *Two-Part Models* vengono utilizzati per analizzare dati che presentano una distribuzione altamente disomogenea, ovvero con molte osservazioni che sono zero o molto piccole rispetto a pochi valori più alti. Possono essere stimati utilizzando modelli separati per la distribuzione binomiale e i dati ordinali. In questo caso sono stati scelti: il *modello di regressione logistica* per la distribuzione binomiale e l' *ordered logistic regression* per i dati ordinali.

3.2 Modelli di regressione logistica

La regressione logistica è un tipo di modello di regressione che viene utilizzato per prevedere la probabilità di un evento dicotomico, sulla base di una o più variabili esplicative. Il modello utilizza una funzione logistica per trasformare la predizione lineare in una probabilità. Per proseguire con questo modello sono state dicotomizzate le variabili dipendenti, in modo che assumano valore 1 per modalità 0, e valore 0 altrove. Lo step successivo è stato adottare una strategia sequenziale (*stepwise regression*) per la selezione delle variabili esplicative. In particolar modo è stata scelta la strategia *backward elimination* (eliminazione all'indietro), ossia una procedura *step-down*, detta anche "a ritroso". Si parte dal modello di regressione più completo, con tutte le variabili esplicative, eliminando progressivamente le variabili che risultano non significative. Poi, si ristima di volta in volta un nuovo modello con le variabili significative del passo precedente. Infine la procedura si ferma quando tutte le variabili che sono presenti nel modello sono risultate significative ad un certo livello, ossia quando tutte le statistiche-test t calcolate superano una prefissata soglia critica ($t < 0.05$).
Il primo modello è relativo alla variabile *crowdfunding*:

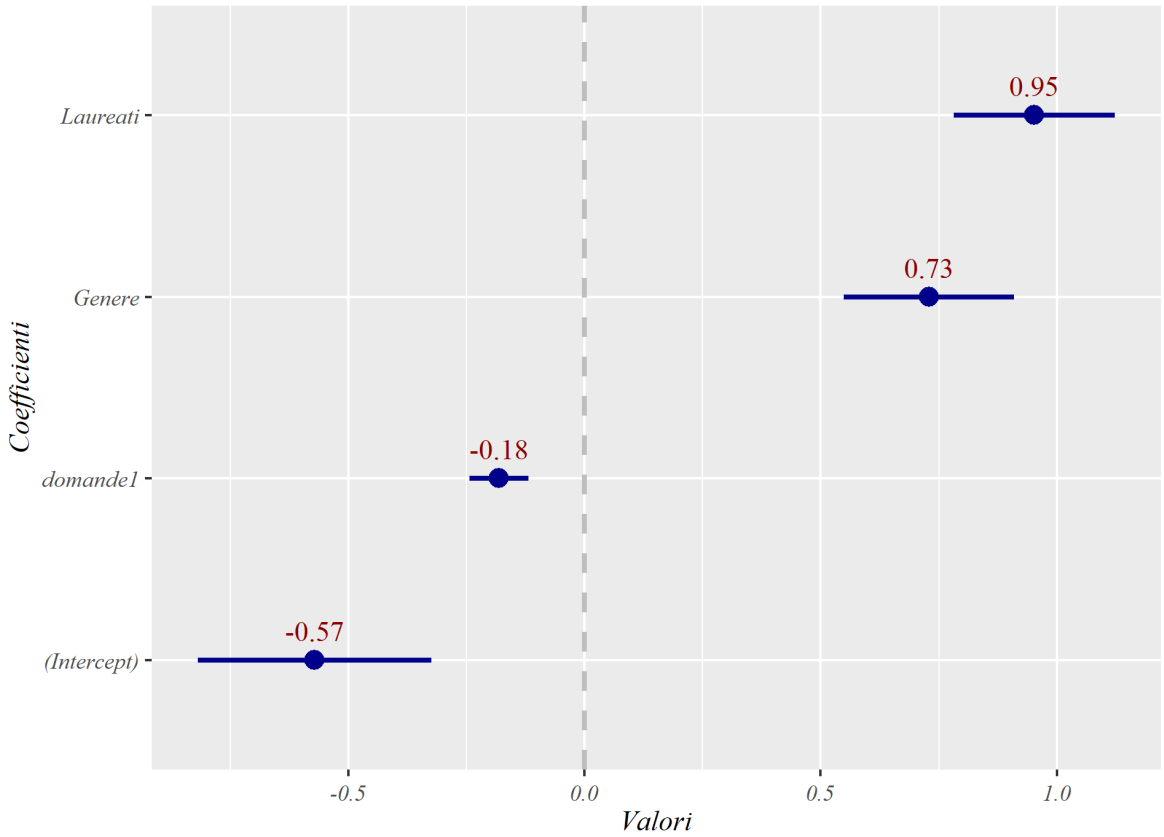

Figura 18: Modello di Regressione Logistica: *crowdfunding*

Il modello di regressione logistica non è immediato da interpretare perché non si specifica una relazione tra variabili osservate, ma tra la probabilità di un evento ed una o più variabili esplicative. Uno dei possibili modi di interpretare tale modello è tramite *odds ratio*, ossia il rapporto tra la probabilità di un evento e la probabilità della sua negazione:

$$odds(E) = \frac{Pr(E)}{Pr(\overline{E})} = \frac{Pr(E)}{1 - Pr(E)}$$

Applicando il logaritmo si ottiene il *log-odds*:

$$log - odds(E) = \log(odds(E)) = \log\left[\frac{Pr(E)}{1 - Pr(E)}\right] = logit(Pr(E))$$

Il modello ottenuto si presenta nella seguente forma:

$$logit(Pr(crowdfunding{=}1)) = -0.57 - 0.18\, domande1 + 0.73\, genere + 0.95\, laureati$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.18} = 0.83; \qquad e^{\beta_2} = e^{0.73} = 2.07; \qquad e^{\beta_3} = e^{0.95} = 2.59$$

La probabilità che si verifichi l'evento nullo (nessuna conoscenza dello strumento *crowdfunding*) si riduce del 17% per ogni risposta esatta relativa ai quesiti a risposta multipla. Passando da uomini a donne la probabilità che si verifichi l'evento nullo aumenta del 107%. Passando da non laureati a laureati la probabilità che si verifichi l'evento nullo aumenta del 159%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	842.84	820.84
AIC	807.35	803.10
log-Lik (ℓ)	-395.68	-397.55

Avendo utilizzato il metodo *backward elimination* per la selezione delle variabili esplicative, ci troviamo nel caso di modelli annidati. Nella tabella sono riportati 3 criteri di valutazione: *BIC* (Bayesian Information Criterion), *AIC* (Akaike Information Criterion) e *log-Lik* (log-Likelihood). Si preferisce il modello con tali indici più bassi. Poiché:

$$BIC = AIC + (p+1)[\log n - 2]$$

Sarà sempre:

$$BIC > AIC \quad se \quad n > 8$$

Per cui l'indice BIC è più "severo" nella scelta tra più modelli. L'indice AIC tende a sovra-parametrizzare i modelli che seleziona, per cui si preferisce BIC quando è importante utilizzare un modello parsimonioso. L'indice AIC è distorto mentre l'indice BIC per i modelli più comuni è consistente e, al crescere di n, specifica il modello corretto con probabilità 1. Tuttavia l'indice BIC e l'indice AIC sono utilizzati per confrontare modelli non annidati.
Un altro metodo di confronto, per modelli annidati, consiste nel *test del rapporto di verosimiglianza (LRT)*:

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-397.55 - 395.68) = 3.74 \rightarrow \chi^2_{g=4} \qquad p - value = 0.44$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.

La *matrice di confusione* è uno strumento utilizzato per valutare la performance di un modello di classificazione, come la regressione logistica. La matrice di confusione mostra il numero di previsioni corrette e sbagliate per ogni classe. Nel caso della regressione logistica binaria, la matrice di confusione ha due colonne e due righe, con le colonne che rappresentano le previsioni del modello e le righe che rappresentano le classi reali . La matrice di confusione serve a calcolare diversi indici di performance, come l'accuratezza, la precisione, il recall e il f1-score, che possono aiutare a valutare la capacità del modello di effettuare previsioni corrette. Inoltre, la matrice di confusione può anche essere utilizzata per identificare eventuali pattern nei risultati del modello, come la tendenza a sovrastimare o sottostimare una classe specifica. Queste informazioni possono essere utilizzate per migliorare ulteriormente il modello.

Figura 19: Matrice di confusione: *crowdfunding*

Un altro metodo di valutazione del modello è la *curva ROC*. La curva ROC (*Receiver Operating Characteristic*) è un grafico utilizzato per valutare la performance di un modello di classificazione binaria, come la regressione logistica. La curva ROC rappresenta la relazione tra la sensibilità (*True Positive Rate, TPR*) e la falsa positività (*False Positive Rate, FPR*) del modello per diverse soglie di classificazione. La TPR rappresenta la percentuale di osservazioni positive effettivamente identificate dal modello, mentre la FPR rappresenta la percentuale di osservazioni negative che vengono classificate come positive. La curva ROC mostra come la performance del modello cambia con la soglia di classificazione, permettendo di valutare la capacità del modello di equilibrare la sensibilità e la falsa positività. In generale, una curva ROC che si avvicina all'angolo superiore sinistro indica un modello con un'elevata sensibilità e una bassa falsa positività, ovvero un modello con una buona capacità di identificare le osservazioni positive e di evitare falsi positivi. Una superficie sotto la curva ROC (*Area Under the Curve, AUC*) più grande indica una performance superiore rispetto a una superficie più piccola.

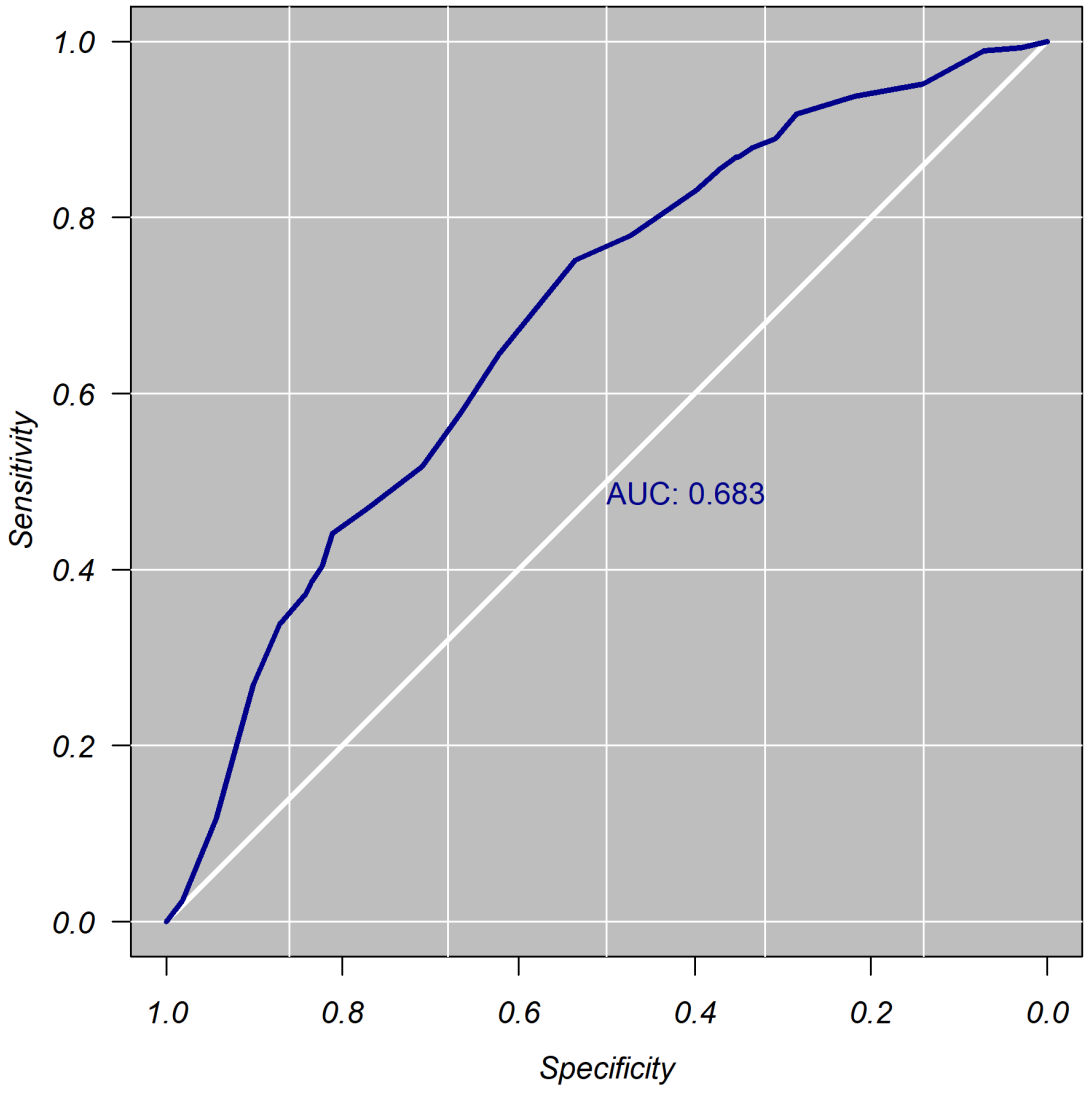

Figura 20: Curva ROC: *crowdfunding*

Grazie a questi ultimi due risultati è possibile affermare che il modello stimato ha una buona performance.

Il secondo modello è relativo alla variabile *cryptocurrencies*:

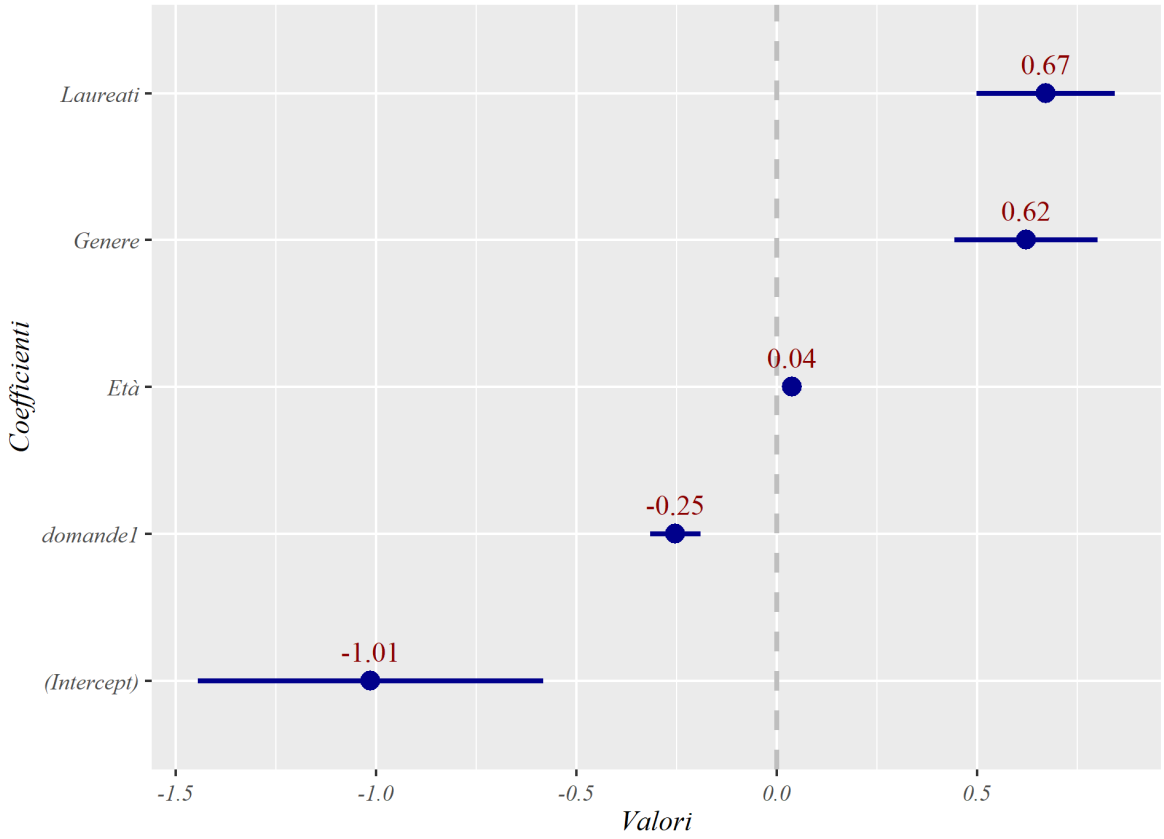

Figura 21: Modello di Regressione Logistica: *cryptocurrencies*

Il modello ottenuto si presenta nella seguente forma:

$$logit(Pr(cryptocurrencies{=}1)) = -1.01 - 0.25\, domande1 + 0.04\, eta + 0.62\, genere + 0.67\, laureati$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.25} = 0.78; \qquad e^{\beta_2} = e^{0.04} = 1.04; \qquad e^{\beta_3} = e^{0.62} = 1.86: \qquad e^{\beta_4} = e^{0.67} = 1.96$$

La probabilità che si verifichi l'evento nullo (nessuna conoscenza dello strumento *cryptocurrencies*) si riduce del 22% per ogni risposta esatta relativa ai quesiti a risposta multipla ed aumenta del 4% all'aumentare di un anno di età. Passando da uomini a donne la probabilità che si verifichi l'evento nullo aumenta del 86%. Passando da non laureati a laureati la probabilità che si verifichi l'evento nullo aumenta del 96%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	849.85	833.03
AIC	814.36	810.85
log-Lik	-399.18	-400.43

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-400.43 - 399.18) = 2.49 \rightarrow \chi^2_{g=3} \qquad p - value = 0.48$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.

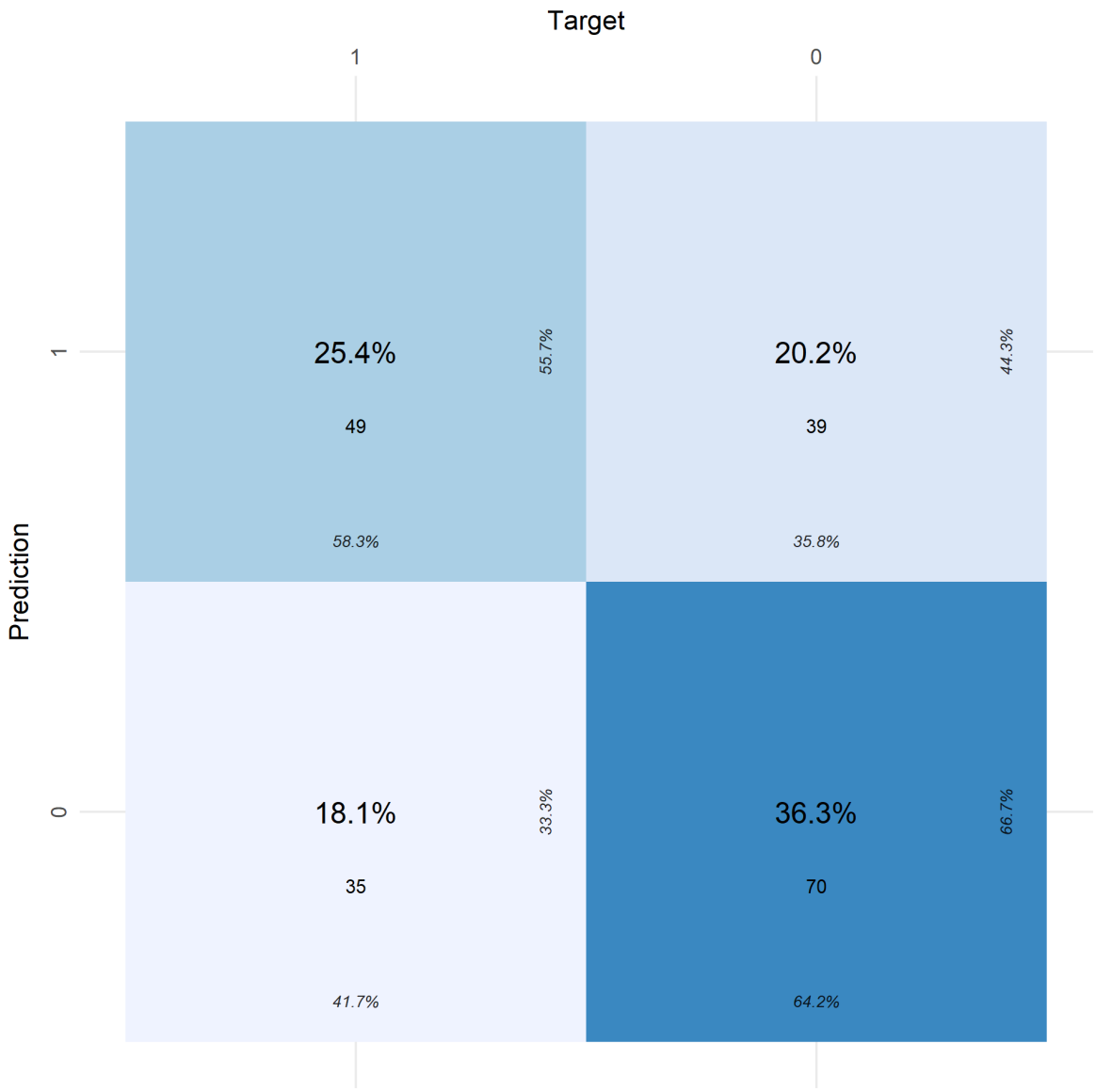

Figura 22: Matrice di confusione: *cryptocurrencies*

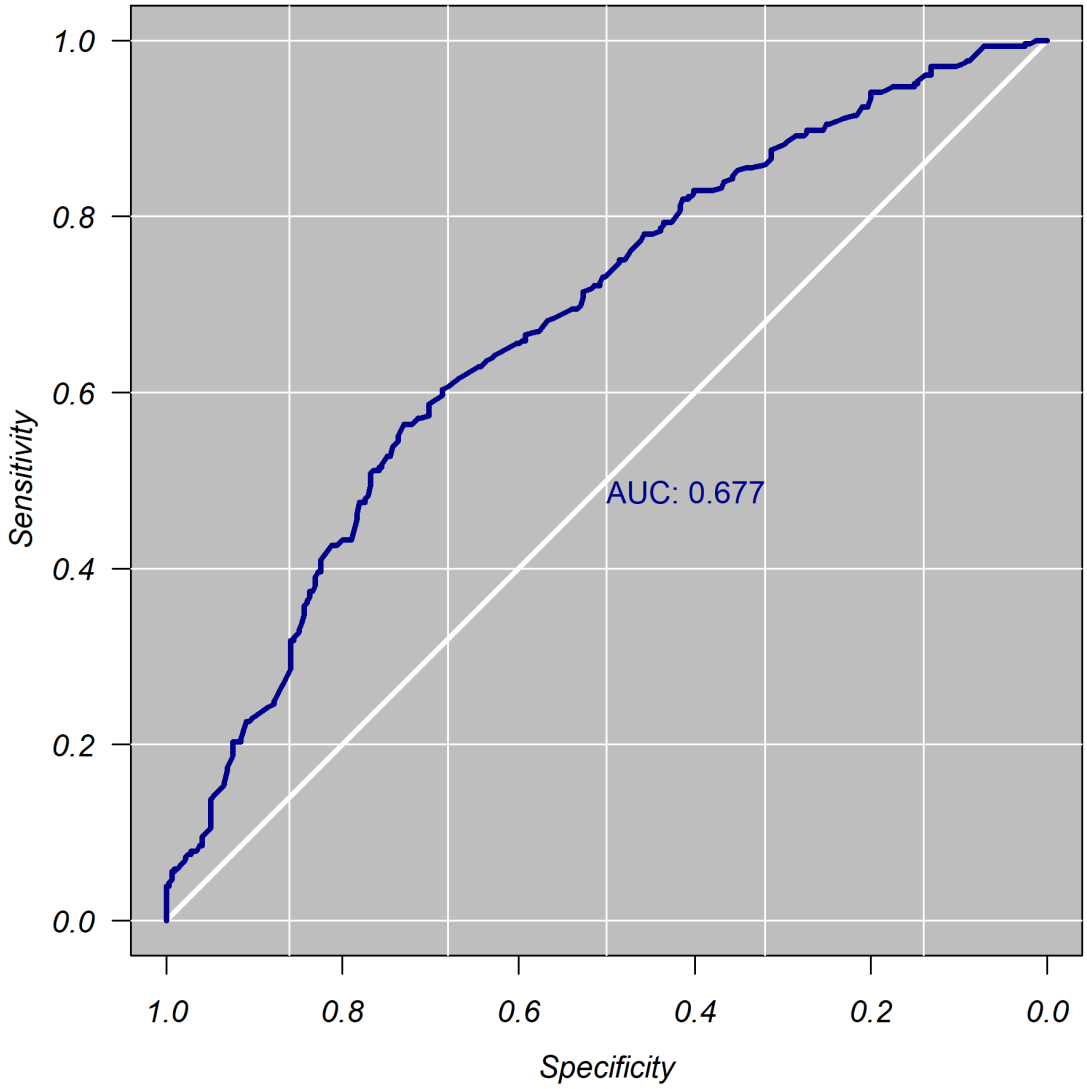

Figura 23: Curva ROC: *cryptocurrencies*

Grazie a questi ultimi due risultati è possibile affermare che il modello stimato ha una buona performance.

Il terzo modello è relativo alla variabile *instantinsurance*:

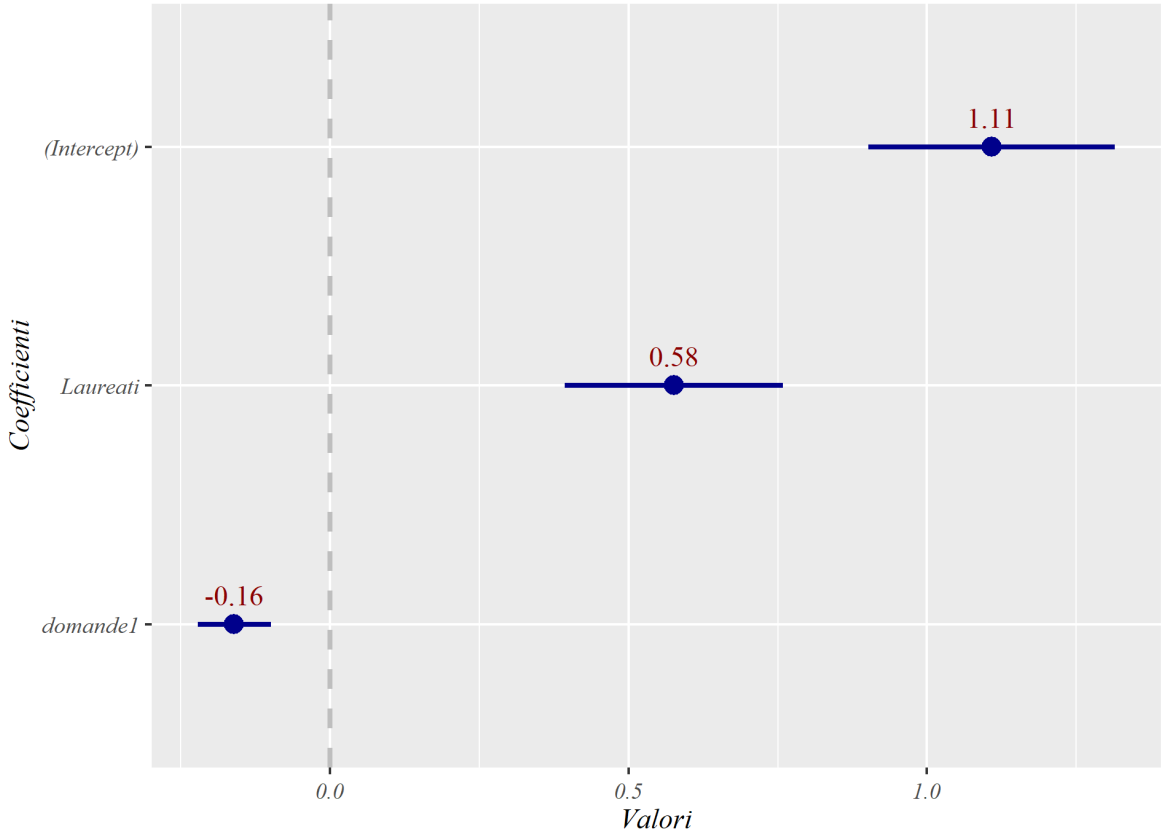

Figura 24: Modello di Regressione Logistica: *instantinsurace*

Il modello ottenuto si presenta nella seguente forma:

$$logit(Pr(instantinsurance{=}1)) = 1.11 - 0.16\, domande1 + 0.58\, laureati$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.16} = 0.85; \qquad e^{\beta_2} = e^{0.58} = 1.78$$

La probabilità che si verifichi l'evento nullo (nessuna conoscenza dello strumento *instantinsurance*) si riduce del 15% per ogni risposta esatta relativa ai quesiti a risposta multipla. Passando da non laureati a laureati la probabilità che si verifichi l'evento nullo aumenta del 78%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	771.31	740.67
AIC	735.82	727.36
log-Lik	-359.91	-360.68

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-360.68 - 359.91) = 1.54 \rightarrow \chi^2_{g=5} \qquad p - value = 0.91$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.

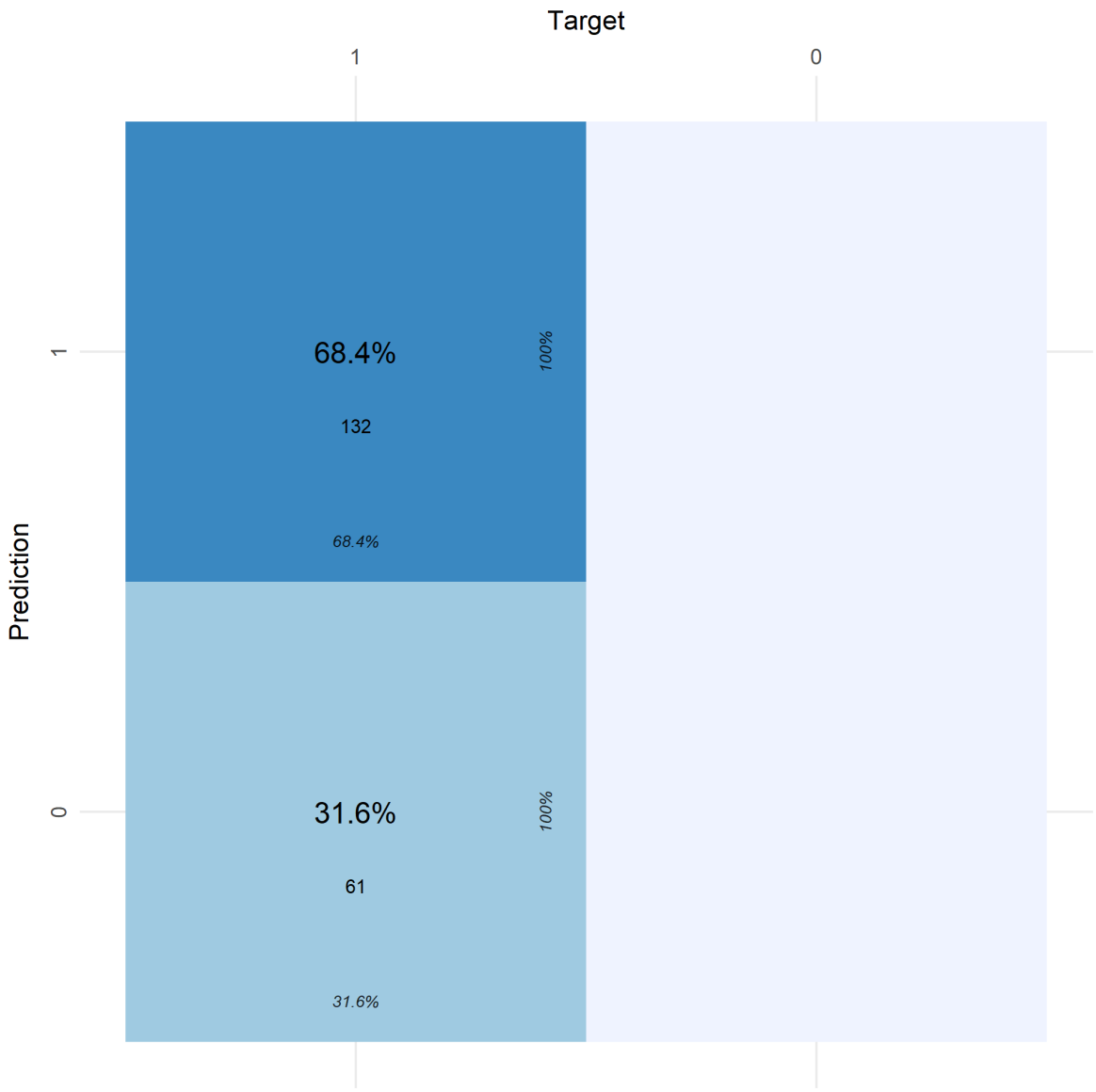

Figura 25: Matrice di confusione: *instantinsurance*

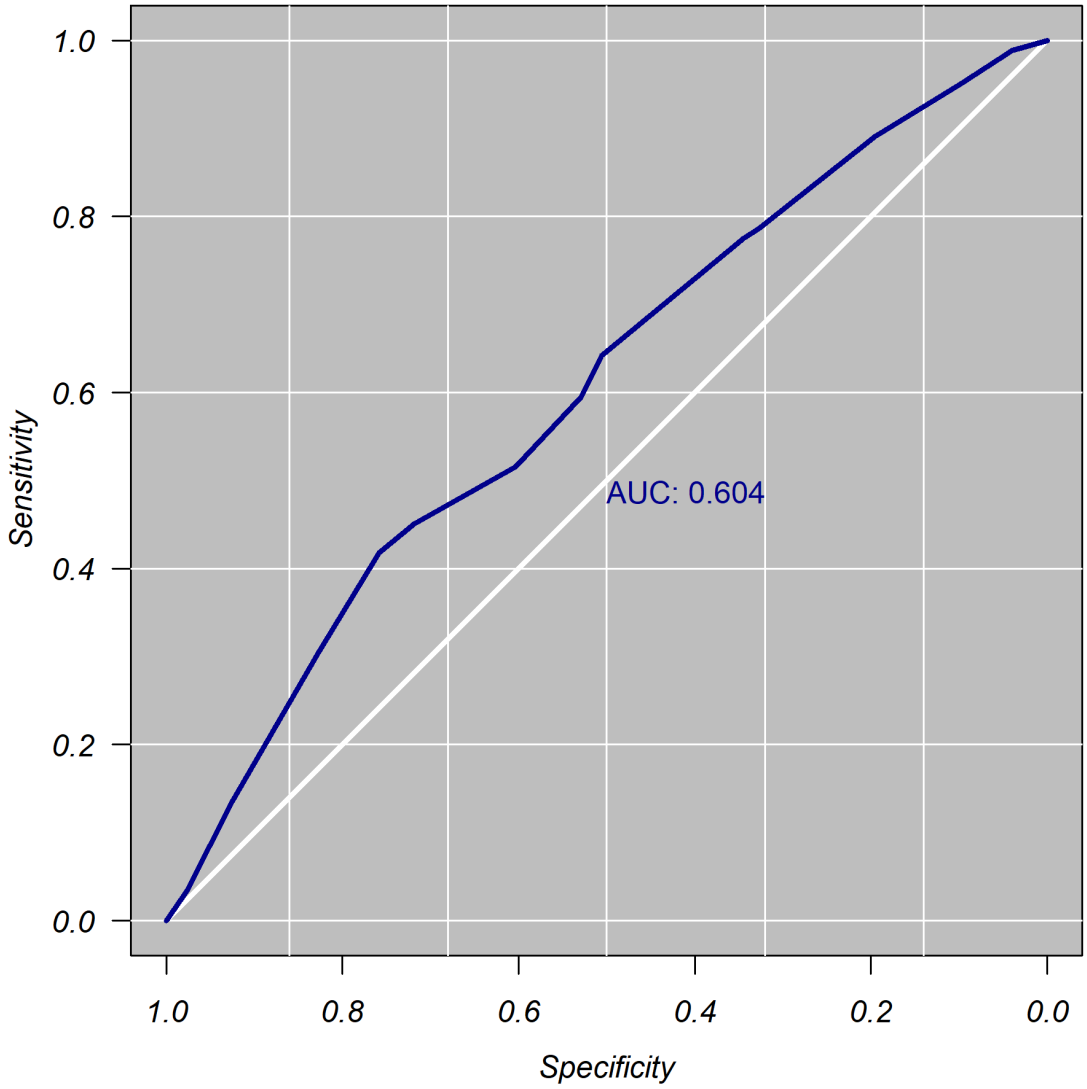

Figura 26: Curva ROC: *instantinsurance*

Grazie a questi ultimi due risultati è possibile affermare che il modello stimato ha una buona performance.

Il quarto modello è relativo alla variabile *instantpayments*:

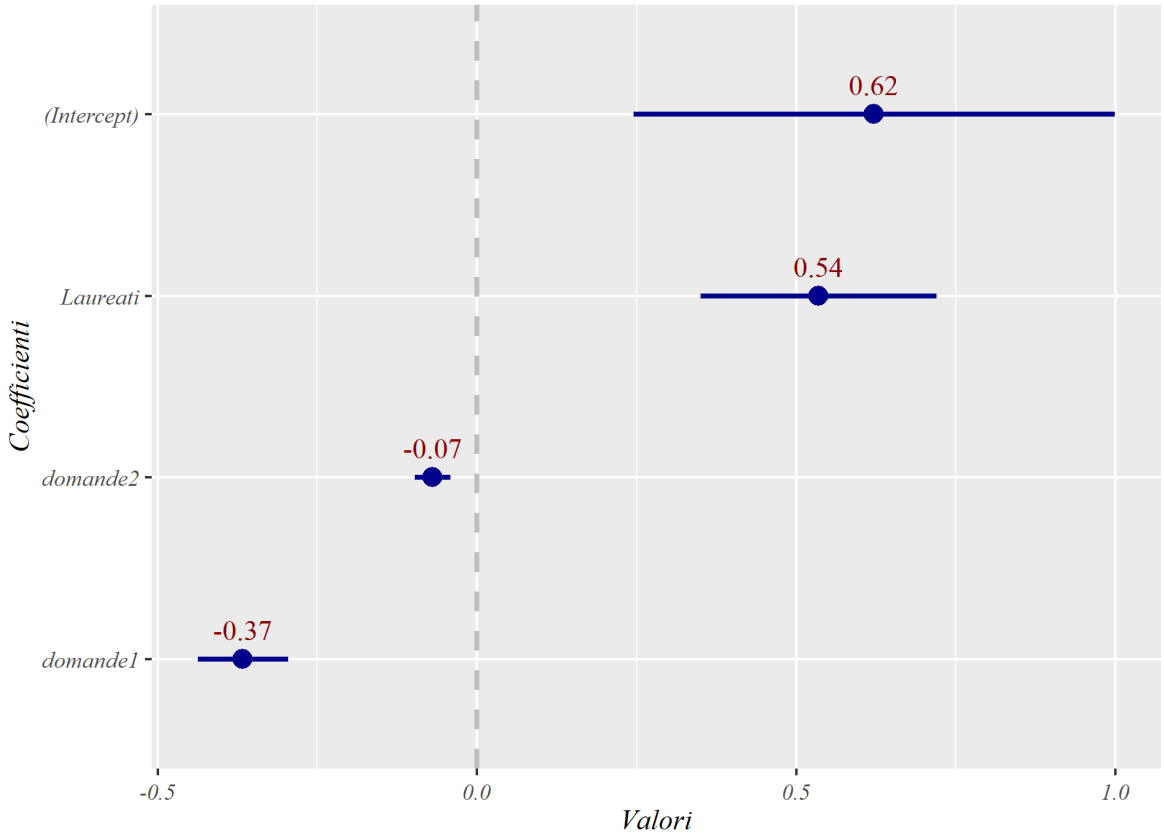

Figura 27: Modello di Regressione Logistica: *instantpayments*

Il modello ottenuto si presenta nella seguente forma:

$$logit(Pr(instantpayments{=}1)) = 0.62 - 0.37\, domande1 - 0.07\, domande2 + 0.54\, laureati$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.37} = 0.69; \qquad e^{\beta_2} = e^{-0.07} = 0.93; \qquad e^{\beta_3} = e^{0.54} = 1.71$$

La probabilità che si verifichi l'evento nullo (nessuna conoscenza dello strumento *instantpayments*) si riduce del 31% per ogni risposta esatta relativa ai quesiti a risposta multipla e si riduce del 7% per ogni incremento del livello di confidenza relativo ai quesiti preferenziali. Passando da non laureati a laureati la probabilità che si verifichi l'evento nullo aumenta del 71%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	743.27	726.53
AIC	707.78	708.79
log-Lik	-345.89	-350.40

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-350.40 - 345.89) = 9.01 \rightarrow \chi^2_{g=4} \qquad p - value = 0.06$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.

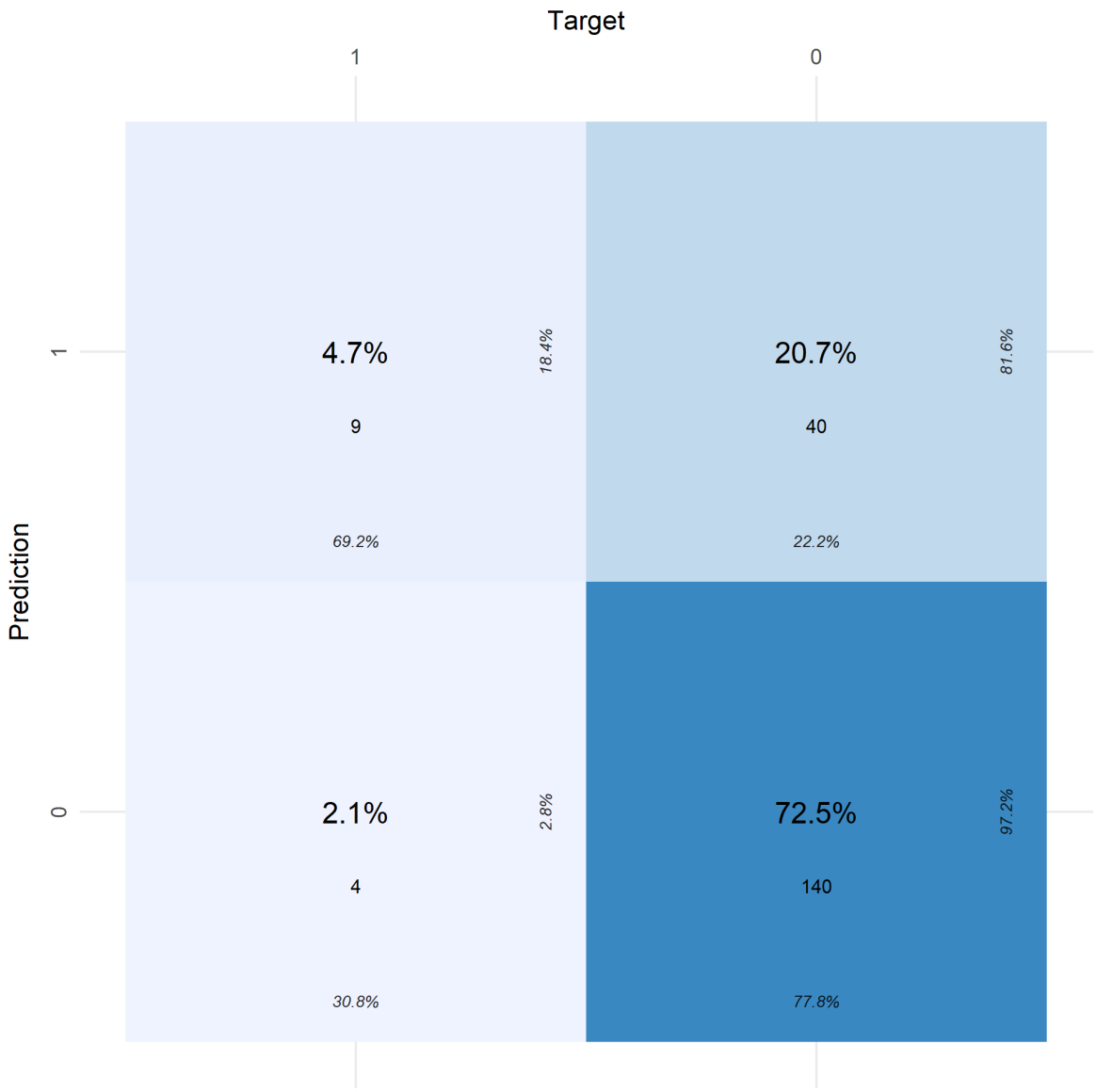

Figura 28: Matrice di confusione: *instantpayments*

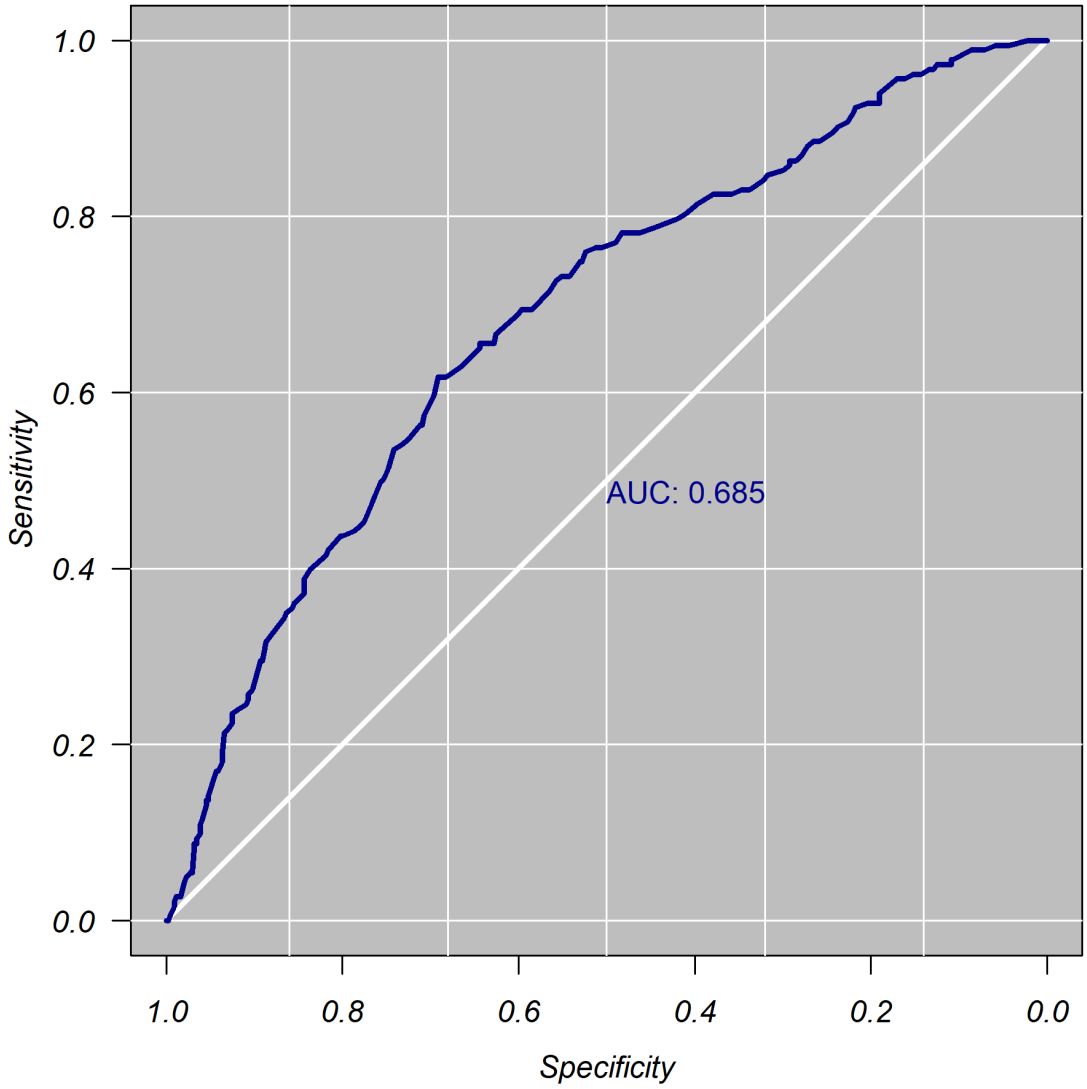

Figura 29: Curva ROC: *instantpayments*

Grazie a questi ultimi due risultati è possibile affermare che il modello stimato ha una buona performance.

Il quinto modello è relativo alla variabile *roboadvisor*:

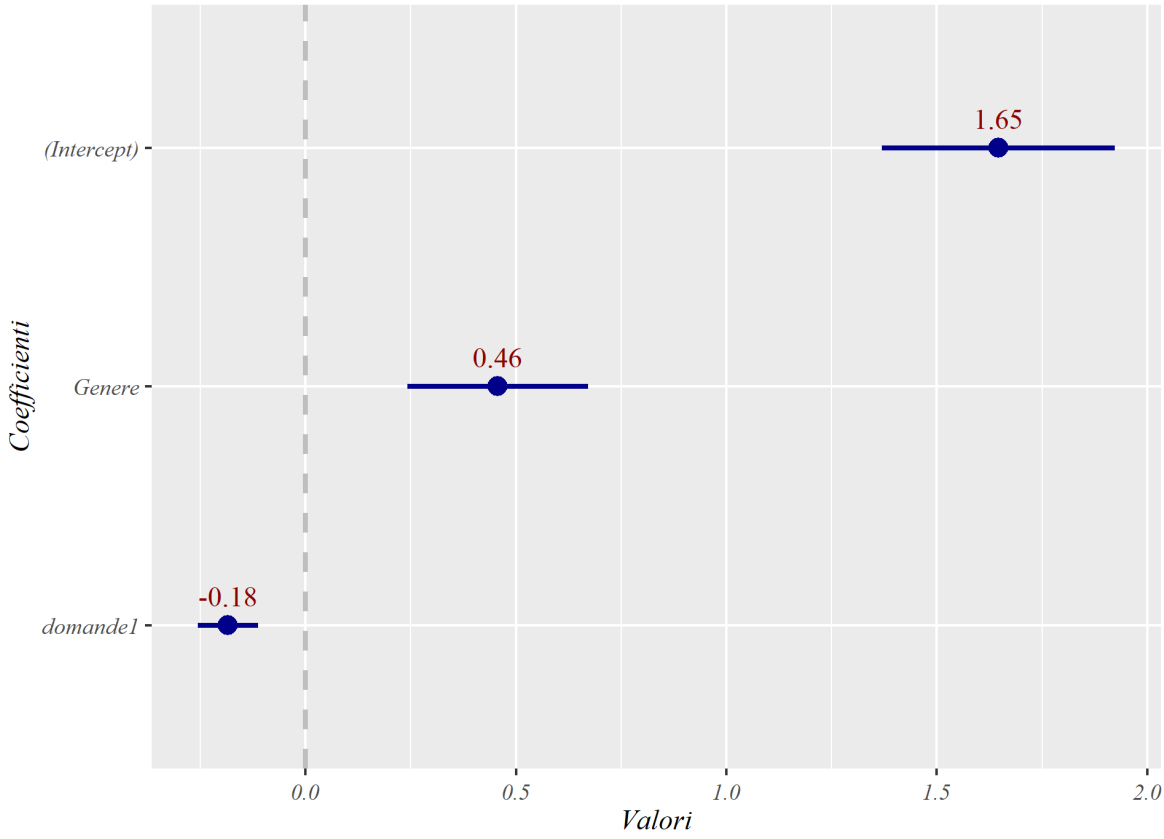

Figura 30: Modello di Regressione Logistica: *roboadvisor*

Il modello ottenuto si presenta nella seguente forma:

$$logit(Pr(roboadvisor{=}1)) = 1.65 - 0.18\, domande1 + 0.46\, genere$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.18} = 0.83; \qquad e^{\beta_2} = e^{0.46} = 1.58$$

La probabilità che si verifichi l'evento nullo (nessuna conoscenza dello strumento *roboadvisor*) si riduce del 17% per ogni risposta esatta relativa ai quesiti a risposta multipla. Passando da uomini a donne la probabilità che si verifichi l'evento nullo aumenta del 58%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	653.02	630.47
AIC	617.53	617.16
log-Lik	-300.76	-305.58

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-305.58 - 300.76) = 9.63 \rightarrow \chi^2_{g=5} \qquad p - value = 0.09$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.

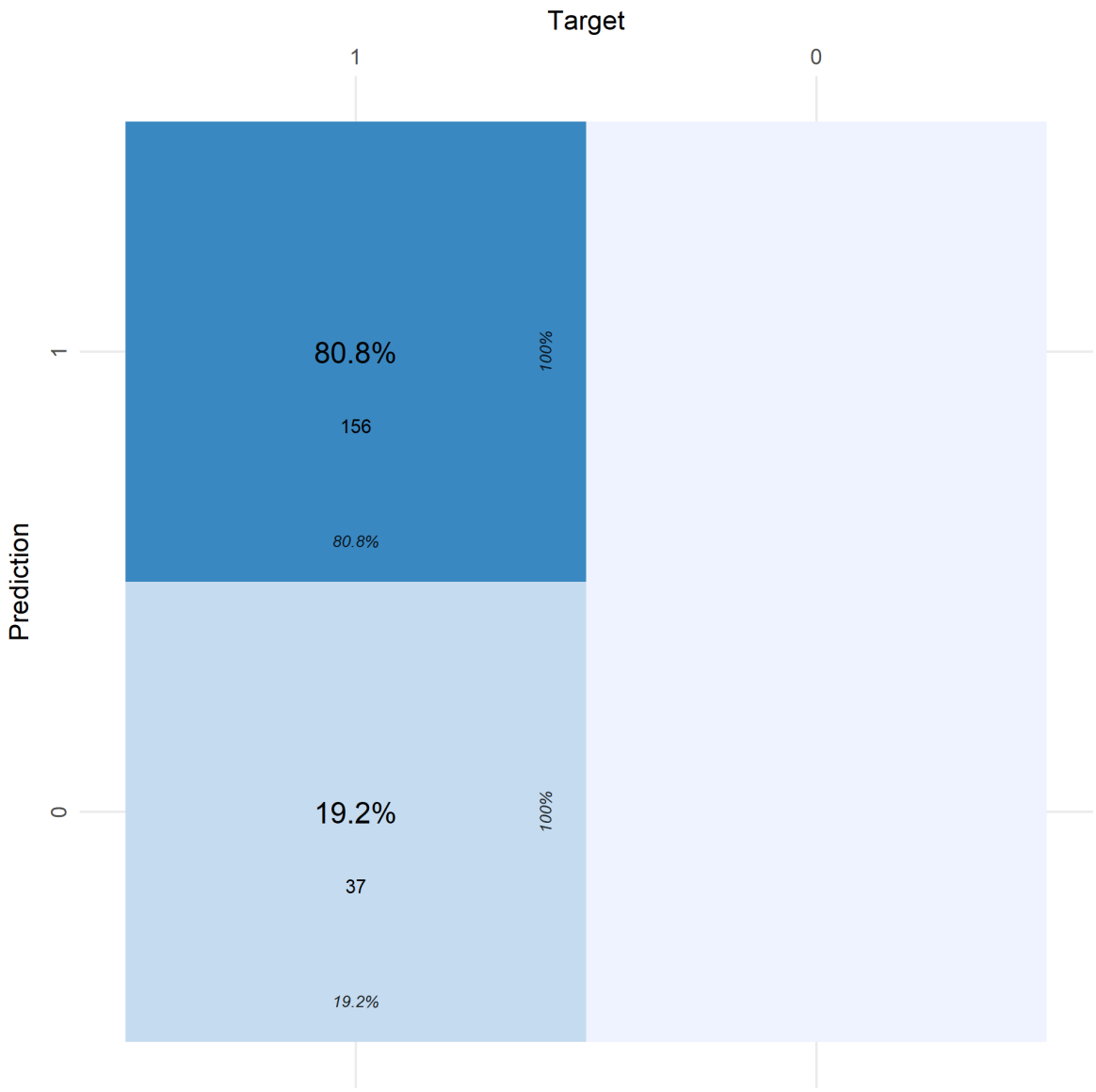

Figura 31: Matrice di confusione: *roboadvisor*

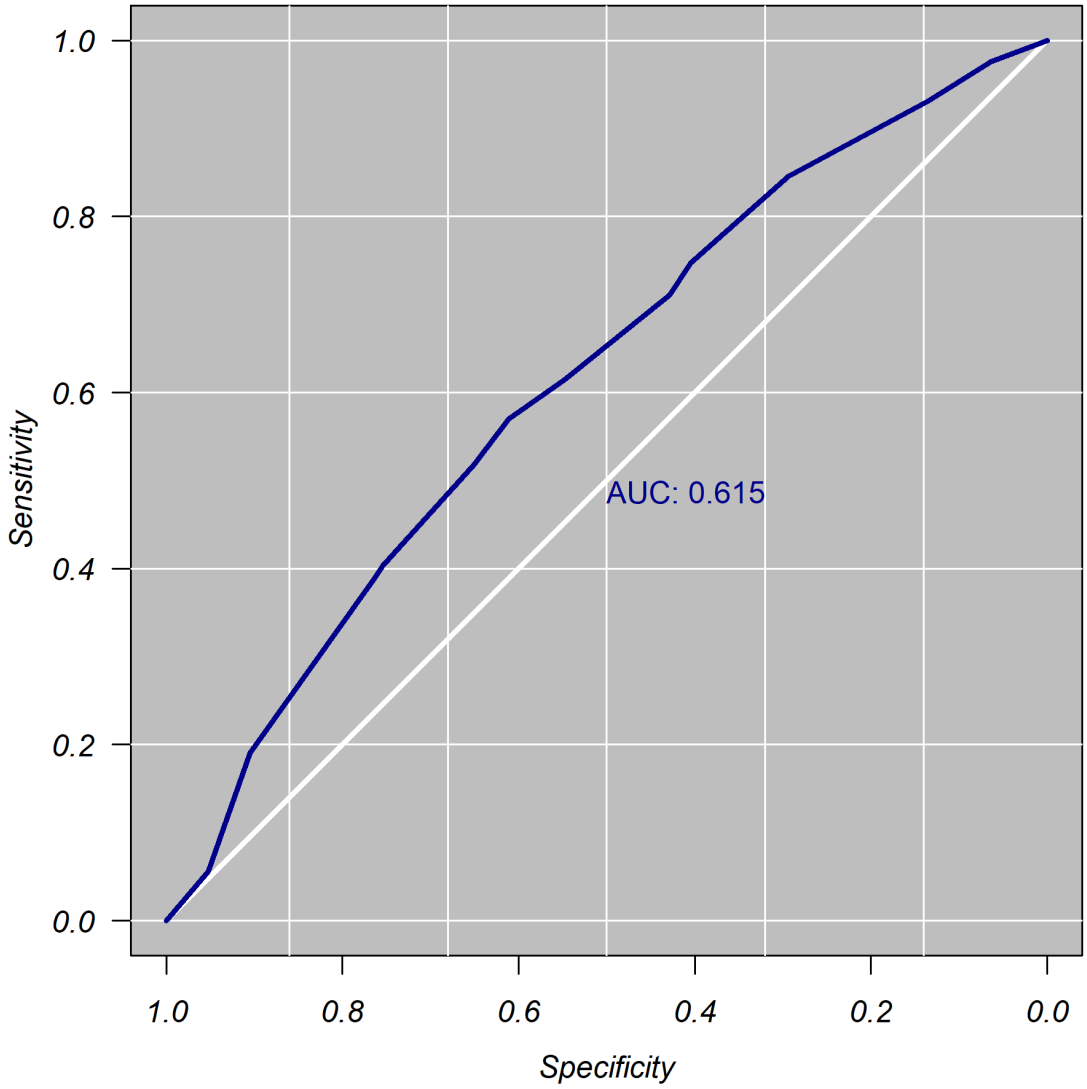

Figura 32: Curva ROC: *roboadvisor*

Grazie a questi ultimi due risultati è possibile affermare che il modello stimato ha una buona performance.

Il sesto modello è relativo alla variabile *p2p*:

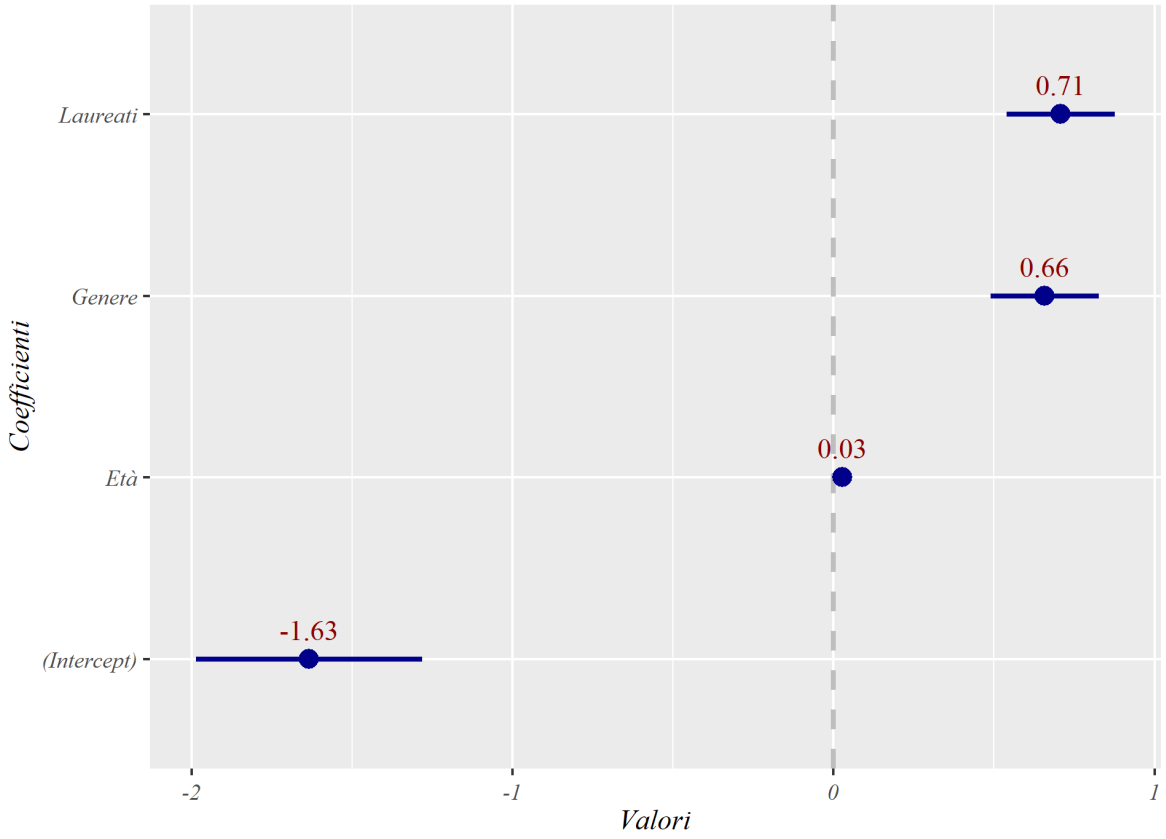

Figura 33: Modello di Regressione Logistica: *p2p*

Il modello ottenuto si presenta nella seguente forma:

$$logit(Pr(p2p{=}1)) = -1.63 + 0.03\, eta + 0.66\, genere + 0.71\, laureati$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{0.03} = 1.03; \qquad e^{\beta_2} = e^{0.66} = 1.93; \qquad e^{\beta_3} = e^{0.71} = 2.03$$

La probabilità che si verifichi l'evento nullo (nessuna conoscenza dello strumento *p2p*) aumenta del 3% all'aumentare di ogni anno d'età. Passando da uomini a donne la probabilità che si verifichi l'evento nullo aumenta del 93%. Passando da non laureati a laureati la probabilità che si verifichi l'evento nullo aumenta del 103%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	870.93	849.90
AIC	835.44	832.15
log-Lik	-409.72	-412.07

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-412.07 - 409.72) = 4.70 \rightarrow \chi^2_{g=4} \qquad p - value = 0.46$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.

Figura 34: Matrice di confusione: *p2p*

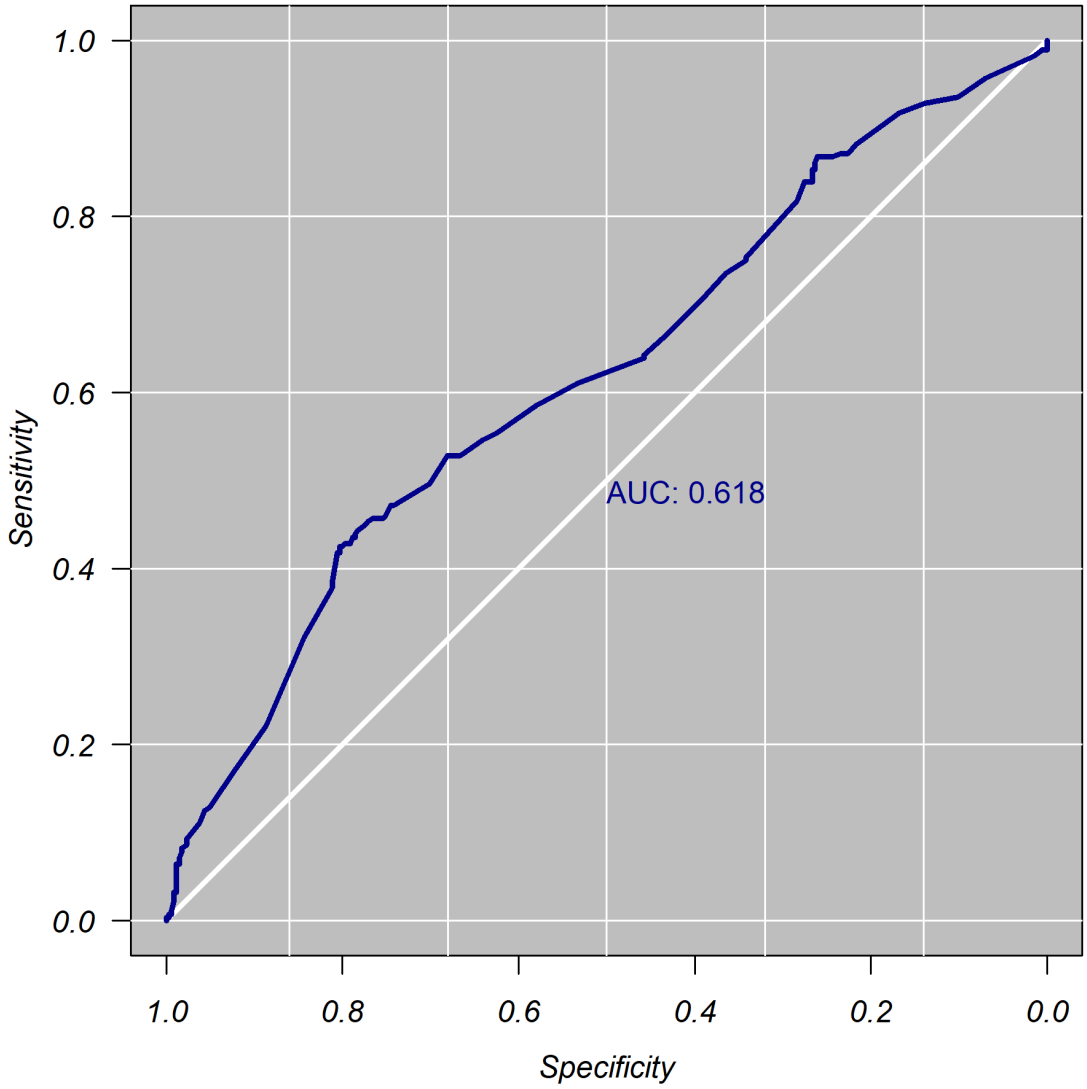

Figura 35: Curva ROC: *p2p*

Grazie a questi ultimi due risultati è possibile affermare che il modello stimato ha una buona performance.

3.3 Ordered logistic regression

L'*ordered logistic regression* è un modello di regressione utilizzato per la classificazione ordinale di variabili dipendenti. La differenza principale rispetto alla regressione logistica ordinaria è che l'ordered logistic regression tiene conto dell'ordine tra le categorie della variabile dipendente, assegnando un peso diverso a ciascuna categoria e modellando la probabilità di appartenenza ad una categoria rispetto alle altre. Questo modello è particolarmente utile in situazioni in cui la variabile dipendente ha più di due categorie e l'ordine tra di esse è importante.Per la selezioni delle variabili esplicative, anche in questo caso, è stata scelta la strategia *backward elimination*, con la differenza che la valutazione della significatività delle variabili è stata effettuata tramite p-value ($p - value < 0.05$).
Il primo modello è relativo alla variabile *crowdfunding*:

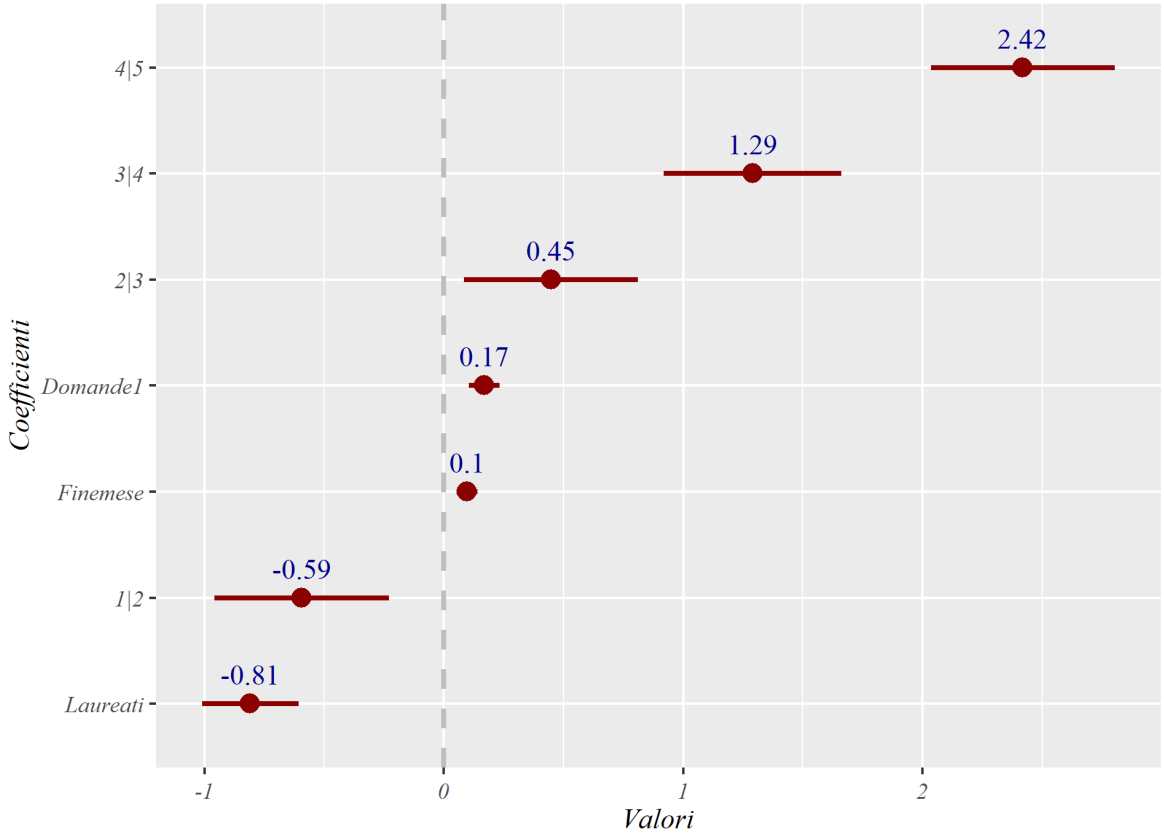

Figura 36: Modello ordered logistic regression: *crowdfunding*

Il modello ottenuto si presenta nella seguenti forme:

$$logit(Pr(crowdfunding \leq 1)) = -0.59 - 0.81\, laureati + 0.10\, finemese + 0.17\, domande1$$
$$logit(Pr(crowdfunding \leq 2)) = 0.45 - 0.81\, laureati + 0.10\, finemese + 0.17\, domande1$$
$$logit(Pr(crowdfunding \leq 3)) = 1.29 - 0.81\, laureati + 0.10\, finemese + 0.17\, domande1$$
$$logit(Pr(crowdfunding \leq 4)) = 2.42 - 0.81\, laureati + 0.10\, finemese + 0.17\, domande1$$

Come per il modello di regressione logistica, anche in questo caso è possibile interpretare il modello utilizzando gli *odds*:

$$e^{\beta_1} = e^{-0.81} = 0.44; \qquad e^{\beta_2} = e^{0.10} = 1.11; \qquad e^{\beta_3} = e^{0.17} = 1.19$$

Passando da non laureati a laureati la probabilità, che il livello di conoscenza di crowdfunding sia inferiore ad una certa quantità, si riduce del 56%. La stessa probabilità aumenta rispettivamente del

11% e 19%, per ogni incremento del benessere economico e risposte esatte ai questi a risposta multipla. Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	1103.90	1084.87
AIC	1061.98	1058.19
log-Lik	-519.99	-522.10

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-522.10 - 519.99) = 4.22 \rightarrow \chi^2_{g=4} \qquad p - value = 0.38$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.
Il *test di Brant-Wald* è condotto sul confronto tra gli odd proporzionali ed i modelli generalizzati. Un test di Brant-Wald è un test di ipotesi sulla significatività della differenza nei coefficienti del modello, che produce una statistica chi-quadrato. Un p-value basso in un test di Brant-Wald è un indicatore del fatto che il coefficiente non soddisfa l'ipotesi della probabilità proporzionale.
Test di Brant-Wald:

	χ^2_g	g	*p-value*
Omnibus	15.19	9	0.09
laureati	1.10	3	0.78
finemese	6.13	3	0.11
domande1	6.75	3	0.08

Dai risultati ottenuti si può non rifiutare l'ipotesi di probabilità proporzionali.

Il secondo modello è relativo alla variabile *cryptocurrencies*:

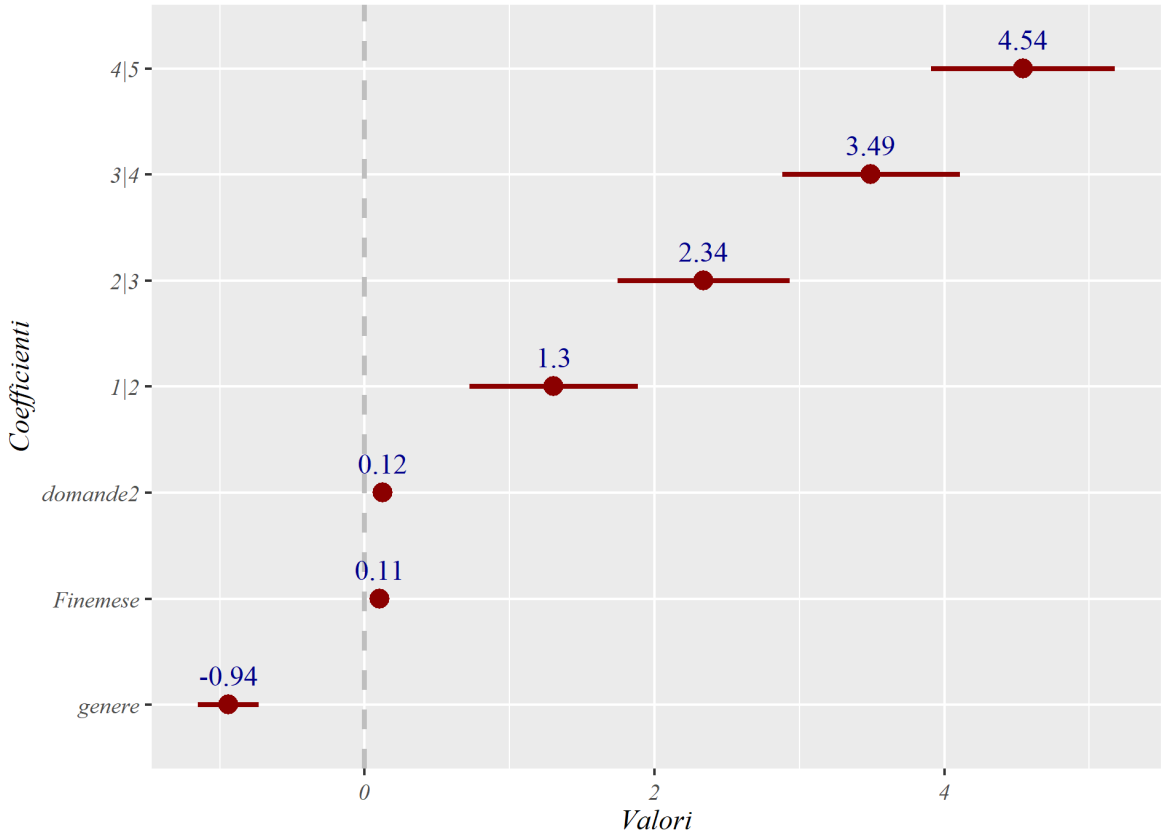

Figura 37: Modello ordered logistic regression: *cryptocurrencies*

Il modello ottenuto si presenta nella seguenti forme:

$$logit(Pr(cryptocurrencies \leq 1)) = 1.30 - 0.94\, genere + 0.11\, finemese + 0.12\, domande2$$
$$logit(Pr(cryptocurrencies \leq 2)) = 2.34 - 0.94\, genere + 0.11\, finemese + 0.12\, domande2$$
$$logit(Pr(cryptocurrencies \leq 3)) = 3.49 - 0.94\, genere + 0.11\, finemese + 0.12\, domande2$$
$$logit(Pr(cryptocurrencies \leq 4)) = 4.54 - 0.94\, genere + 0.11\, finemese + 0.12\, domande2$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.94} = 0.93; \qquad e^{\beta_2} = e^{0.11} = 1.12; \qquad e^{\beta_3} = e^{0.12} = 1.13$$

Passando da uomini a donne la probabilità, che il livello di conoscenza di cryptocurrencies sia inferiore ad una certa quantità, si riduce del 7%. La stessa probabilità aumenta rispettivamente del 12% e 13%, per ogni incremento del benessere economico e livello di confidenza con argomenti o materie.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	976.37	957.54
AIC	934.95	931.18
log-Lik	-456.47	-458.59

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-458.59 - 456.47) = 4.24 \rightarrow \chi^2_{g=4} \qquad p - value = 0.37$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.
Test di Brant-Wald:

	χ^2_g	g	*p-value*
Omnibus	12.32	9	0.20
genere	1.15	3	0.76
finemese	2.69	3	0.44
domande2	8.60	3	0.04

Dai risultati ottenuti si può non rifiutare l'ipotesi di probabilità proporzionali.

Il terzo modello è relativo alla variabile *instantinsurance*:

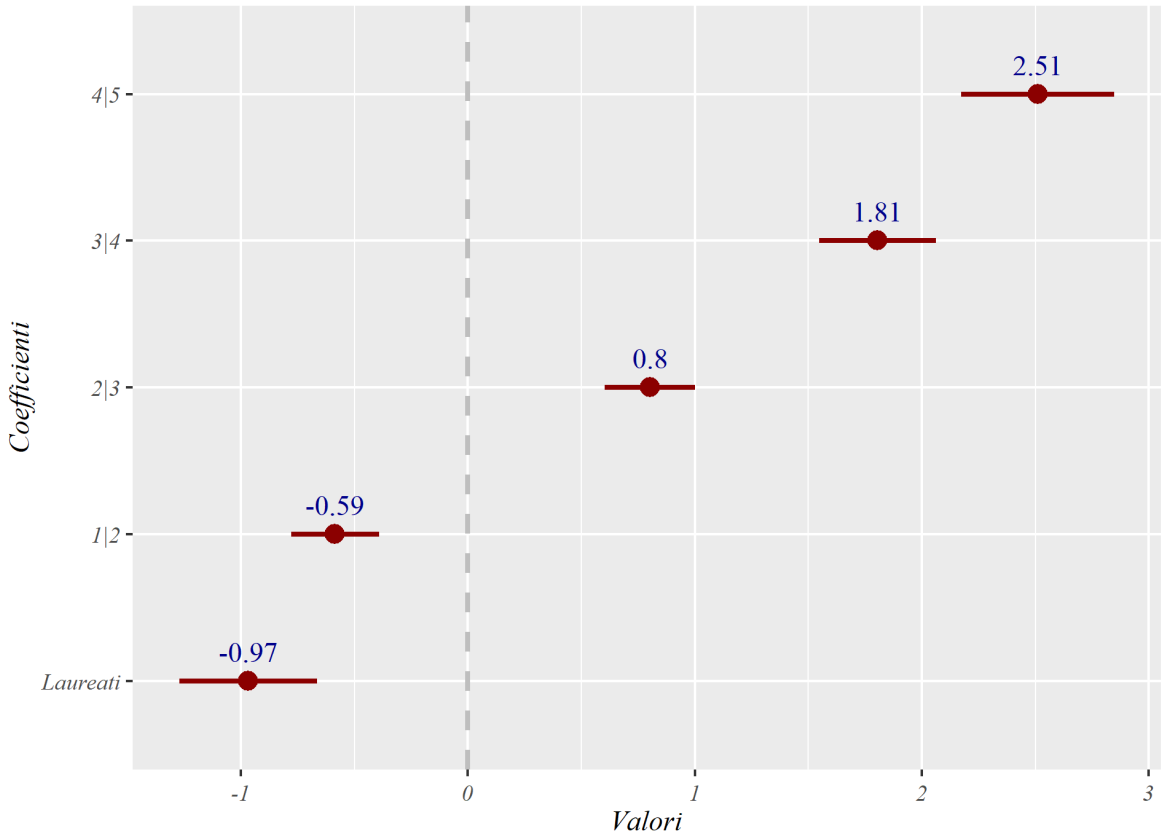

Figura 38: Modello ordered logistic regression: *instantinsurance*

Il modello ottenuto si presenta nella seguenti forme:

$$logit(Pr(instantinsurance \leq 1)) = -0.59 - 0.97\, laureati$$
$$logit(Pr(instantinsurance \leq 2)) = 0.80 - 0.97\, laureati$$
$$logit(Pr(instantinsurance \leq 3)) = 1.81 - 0.97\, laureati$$
$$logit(Pr(instantinsurance \leq 4)) = 2.51 - 0.97\, laureati$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.97} = 0.38$$

Passando da non laureati a laureati la probabilità, che il livello di conoscenza di instantinsurance sia inferiore ad una certa quantità, si riduce del 62%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	495.47	471.86
AIC	460.72	456.07
log-Lik	-219.36	-223.03

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-223.03 - 219.36) = 7.34 \rightarrow \chi^2_{g=6} \qquad p - value = 0.29$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.
Test di Brant-Wald:

	χ^2_g	g	*p-value*
Omnibus	1.63	3	0.65
laureati	1.63	3	0.65

Dai risultati ottenuti si può non rifiutare l'ipotesi di probabilità proporzionali.

Il quarto modello è relativo alla variabile *instantpayments*:

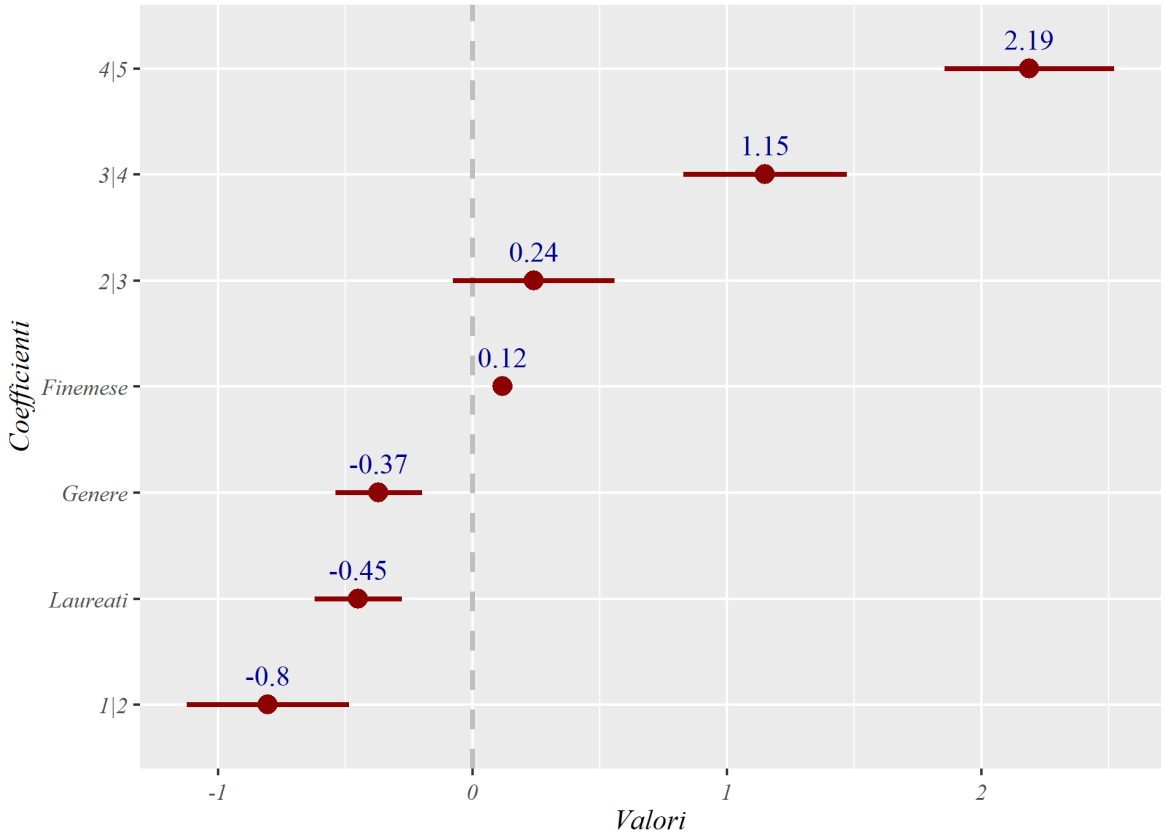

Figura 39: Modello ordered logistic regression: *instantpayments*

Il modello ottenuto si presenta nella seguenti forme:

$$logit(Pr(instantpayments \leq 1)) = -0.80 - 0.45\, laureati - 0.37\, genere + 0.12\, finemese$$
$$logit(Pr(instantpayments \leq 2)) = 0.24 - 0.45\, laureati - 0.37\, genere + 0.12\, finemese$$
$$logit(Pr(instantpayments \leq 3)) = 1.15 - 0.45\, laureati - 0.37\, genere + 0.12\, finemese$$
$$logit(Pr(instantpayments \leq 4)) = 2.19 - 0.45\, laureati - 0.37\, genere + 0.12\, finemese$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.45} = 0.64; \qquad e^{\beta_2} = e^{-0.37} = 0.69; \qquad e^{\beta_3} = e^{0.12} = 1.13$$

Passando da non laureati a laureati la probabilità, che il livello di conoscenza di instantpayments sia inferiore ad una certa quantità, si riduce del 36%. Passando a uomini a donne la probabilità si riduce del 31%. La stessa probabilità aumenta del 13%, per ogni incremento del benessere economico.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	1453.21	1432.41
AIC	1408.23	1403.79
log-Lik	-693.12	-694.90

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-694.90 - 693.12) = 3.56 \rightarrow \chi^2_{g=2} \qquad p - value = 0.46$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.
Test di Brant-Wald:

	χ^2_g	g	*p-value*
Omnibus	13.63	9	0.14
genere	3.43	3	0.33
laureati	1.06	3	0.79
finemese	8.99	3	0.03

Dai risultati ottenuti si può non rifiutare l'ipotesi di probabilità proporzionali.

Il quinto modello è relativo alla variabile *roboadvisor*:

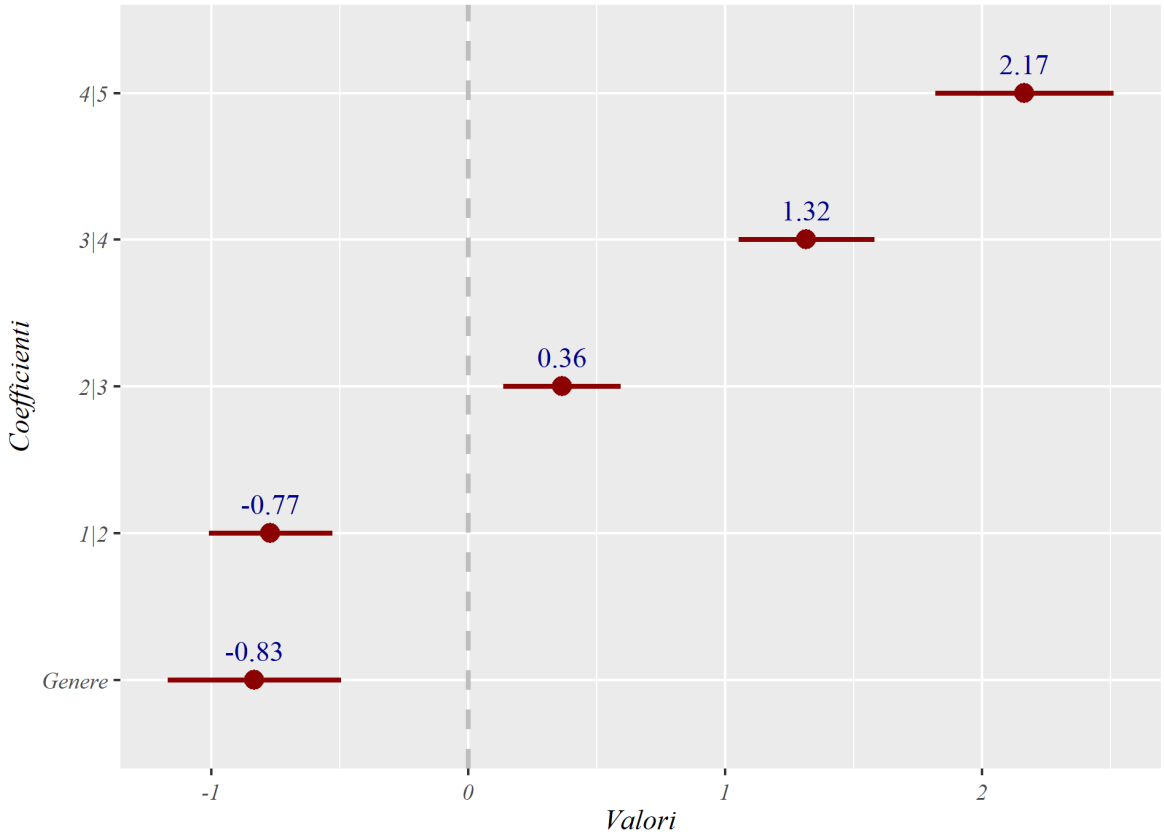

Figura 40: Modello ordered logistic regression: *roboadvisor*

Il modello ottenuto si presenta nella seguenti forme:

$$logit(Pr(roboadvisor \leq 1)) = -0.77 - 0.83\, genere$$
$$logit(Pr(roboadvisor \leq 2)) = 0.36 - 0.83\, genere$$
$$logit(Pr(roboadvisor \leq 3)) = 1.32 - 0.83\, genere$$
$$logit(Pr(roboadvisor \leq 4)) = 2.17 - 0.83\, genere$$

Utilizzando gli odds per interpretare il modello rispetto alle variabili esplicative risulta che:

$$e^{\beta_1} = e^{-0.83} = 0.44$$

Passando da uomini a donne la probabilità, che il livello di conoscenza di crowdfunding sia inferiore ad una certa quantità, si riduce del 56%.
Di seguito alcuni indici per confrontare il modello completo (M_c) con il modello finale (M_f):

	Modello Completo	*Modello Finale*
BIC	399.79	378.17
AIC	368.59.44	363.99
log-Lik	-173.29	-176.99

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-176.99 - 173.29) = 3.56 \rightarrow \chi^2_{g=4} \qquad p - value = 0.47$$

Dato che il p-value>0.05 non si rifiuta l'ipotesi nulla. Questo significa che il modello completo e il modello finale si adattano ugualmente bene ai dati. Pertanto, dovremmo utilizzare il modello finale

perché le variabili predittive aggiuntive nel modello completo non offrono un miglioramento significativo nell'adattamento.
Test di Brant-Wald:

	χ^2_g	g	*p-value*
Omnibus	6.53	3	0.09
genere	6.53	3	0.09

Dai risultati ottenuti si può non rifiutare l'ipotesi di probabilità proporzionali.

Riferimenti bibliografici

[1] Agresti A. *Analysus of Ordinal Categorical Data.* Wiley, 2 edition, 2010.

[2] Agresti A. *Categorical Data Analysis.* Wiley, 3 edition, 2013.

[3] Dobson A. *An Introduction to Generalized Linear Model.* Springer, 2 edition, 1990.

[4] Piccolo D. *Statistica.* Il Mulino, 3 edition, 2010.