BANK MARKETING ANALYSIS

Matteo Paolo Dell'Acqua, matricola 875152
Stefano Madona, matricola 874799
Giorgio Oggioni, matricola 876170

Febbraio 2023

ABSTRACT
Il dataset in questione riguarda le informazioni in possesso da una banca sui suoi clienti.
La banca vuole iniziare una campagna marketing per proporre loro un deposito a termine.
Si è deciso di considerare solo variabili a disposizione della banca prima dell'inizio della campagna di marketing, in modo da creare un modello di classificazione che consenta alla banca di sapere quali clienti siano più propensi ad accettare l'offerta.
A seguito di analisi preliminari e trasformazioni delle variabili, i due modelli che sono stati provati sono la regressione logistica e il k nearest neighbors.
Dopo aver valutato le performance dei due si è deciso che il modello migliore fosse quello della regressione logistica.

1. INTRODUZIONE

Il marketing è il complesso delle tecniche intese a porre merci e servizi a disposizione del consumatore e dell'utente in un dato mercato nel tempo, luogo e modo più adatti, ai costi più bassi per il consumatore e nello stesso tempo remunerativi per l'impresa.
Nel caso in questione ci stiamo occupando di una banca che vuole rendere più efficace ed efficiente la propria campagna promozionale.
Le strategie attualmente più utilizzate per raggiungere questo obiettivo riguardano la pubblicità e la promozione del prodotto: ideare una campagna di advertising efficiente, creare applicazioni e siti che siano sia educational che user-friendly, raggiungere i clienti attraverso multipli canali di comunicazione sono sicuramente tra i metodi più remunerativi nel settore.
L'analisi dei dati è diventata nel tempo un supporto fondamentale per tutte queste strategie, garantendo che queste siano veicolate al target di popolazione più propenso ad accettare qualsivoglia proposta.
In queste analisi si considerano sia parametri che riguardano vari aspetti della persona, tra cui il lavoro e la situazione familiare, sia indicatori economici, come possono essere i tassi di interesse o i tassi di occupazione.
L'obiettivo dell'analisi qui riportata è quello di classificare, con i dati reperibili prima della campagna marketing, quali clienti siano più propensi a dire di sì alla proposta di

un deposito a termine e capire quali sono le variabili che più ne influenzano la decisione.

2. MATERIALI E METODI

Il dataset considerato si chiama “Bank Marketing Dataset” ed è disponibile su https://archive.ics.uci.edu/ml/datasets/bank+marketing.
È composto da 21 variabili e 41188 osservazioni, che riguardano i seguenti parametri:

- age: variabile numerica
- job: variabile categorica (i valori possibili sono: “admin.”, “blue-collar”, “entrepreneur”, “housemaid”, “management”, “retired”, “self-employed”, “services”, “student”, “technician”, “unemployed”, “unknown”)
- marital: variabile categorica (i valori possibili sono: “divorced”, “married”, “single”, “unknown”)
- education: variabile categorica (i valori possibili sono: “basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”)
- default: variabile categorica (i valori possibili sono: “yes”, “no”, “unknown”), indica se ha o no credito in default
- housing: variabile categorica (i valori possibili sono: “yes”, “no”, “unknown”), indica se il cliente ha aperto un mutuo.
- loan: variabile categorica (i valori possibili sono: “yes”, “no”, “unknown”), indica se il cliente ha aperto un prestito.
- contact: variabile categoria (i valori possibili sono: “cellular”, “telefono fisso”), indica con che metodo è stato/sarà contattato il cliente.
- month: variabile categorica (i valori possibili sono: “jan”, “feb”, …, “nov”, “dec”)
- day of the week: variabile categorica (i valori possibili sono: “mon”, “tue”, “wed”, “thu”, “fri”)
- duration: la durata dell’ultimo contatto effettuato con il cliente, ma la durata della chiamata non è conosciuta prima della campagna, è un valore che viene usato per motivi di benchmark, ma deve essere eliminato per l’uso di modelli predittivi.
- campaign: variabile numerica, numero di contatti avuti durante questa campagna.
- pdays: variabile numerica, numero di giorni passati dall’ultimo contatto con il cliente, se 999 il cliente non è mai stato contattato.
- previous: variabile numerica, numero di contatti avuti precedentemente con il cliente.
- poutcome: variabile categorica (i valori possibili sono: “failure”, “nonexistent”, “success”), output della precedente campgan di marketing.
- emp.var.rate: variabile numerica, variazione del numero di lavoratori.
- cons.price.idx: variabile numerica, indice dei prezzi al consumo.
- cons.conf.idx: variabile numerica , indice della fiducia dei consumatori

- euribor3m: variabile numerica, tasso euribor
- nr.employed: variabile numerica.
- output: Il cliente ha sottoscritto un deposito vincolato? variabile binaria (“yes” or “no”)

2.1 CONSIDERAZIONI SULLE VARIABILI

Data la natura a priori della nostra analisi, la decisione è di rimuovere fin da subito le variabili “month”, “day of the week”, “duration” e “campaign”. Esse infatti si riferiscono tutte ad informazioni recepite durante la campagna, di conseguenza non potremmo averle a disposizione prima dell’inizio della stessa.

Analizzando il dataset, si evince che apparentemente non sono presenti missing values (NA).

Nella fig 2.1 visualizziamo le classi delle variabili qualitative; si nota che alcune variabili contengono la modalità unknown; nel prosieguo dell’analisi queste osservazioni verranno trattate come missing values.

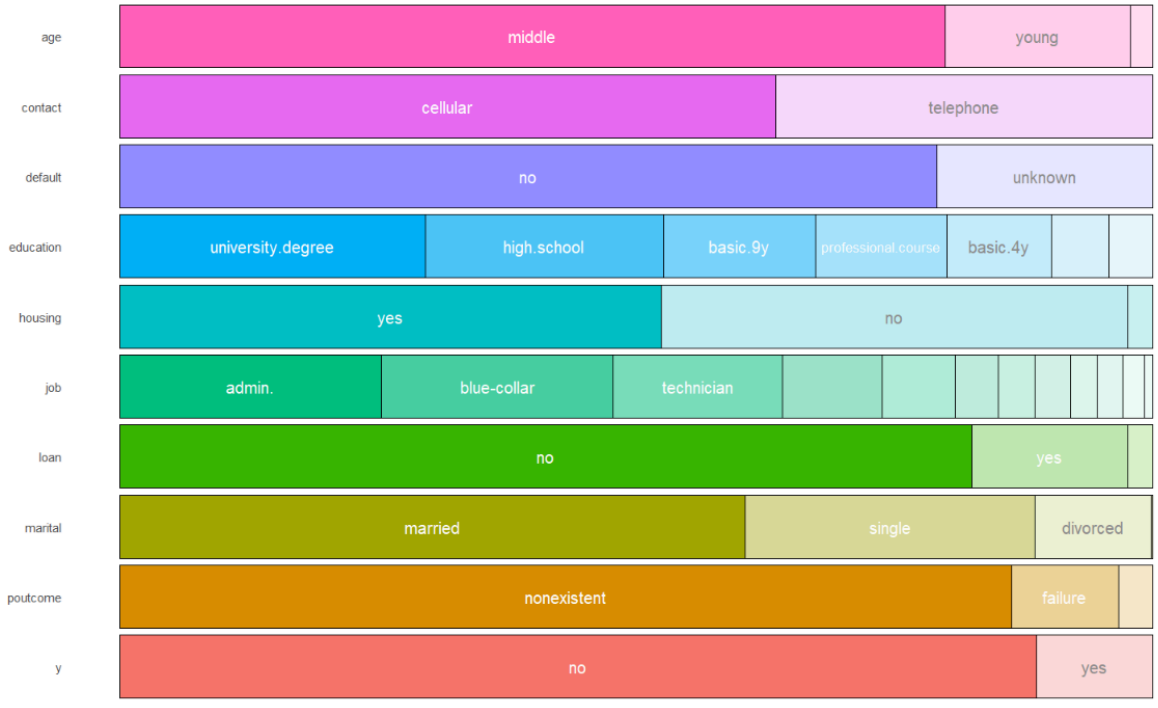

FIG. 2.1: frequenze relative delle variabili qualitative.

Di seguito elenchiamo le variabili su cui abbiamo fatto modifiche nella fase di esplorazione del dataset:

Age

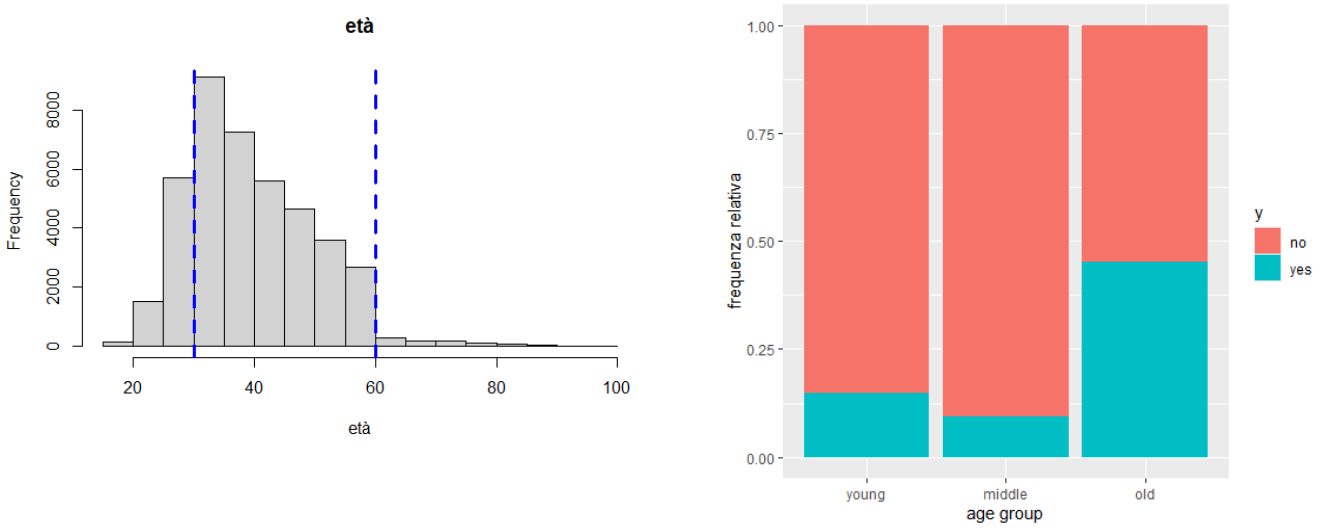

FIG 2.1.1: *istogramma delle frequenze assolute della variabile età. (fig. 2.1.1a) e frequenze relative dell'output condizionato alle classi di età (fig 2.1.1b)*

La variabile age è un valore numerico che rappresenta l'età della persona contattata. Come si vede nella fig. 2.1.1a, la decisione è stata quella di dividere l'età in 3 fasce: la predisposizione ad un deposito vincolato sarà sicuramente diversa tra qualcuno che ha appena iniziato a lavorare (categoria young), qualcuno che è nel pieno della propria carriera lavorativa (categoria middle) e qualcuno che è verso o nella pensione (fascia old).
La decisione presa è supportata dalla fig 2.1.1b, che sottolinea la distinzione netta dei livelli della variabile target tra i diversi gruppi di età.

Job

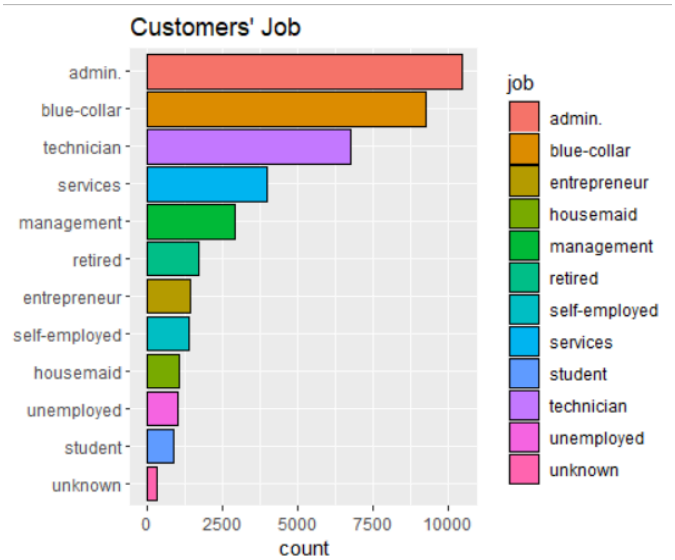

FIG 2.1.2: *istogramma delle frequenze assolute della variabile Job.*

Job è una variabile categorica. Come si vede nella fig 2.1.2 sono presenti molte modalità diverse tra loro; l'idea iniziale era quella di trovare un raggruppamento logico tra i dati, ma per via della diversità delle categorie non è stato possibile. Inoltre, sono presenti 330 osservazioni con valore unknown; dal momento che rappresentano una percentuale inferiore all'1% la decisione è stata di rimuovere queste osservazioni.

Marital

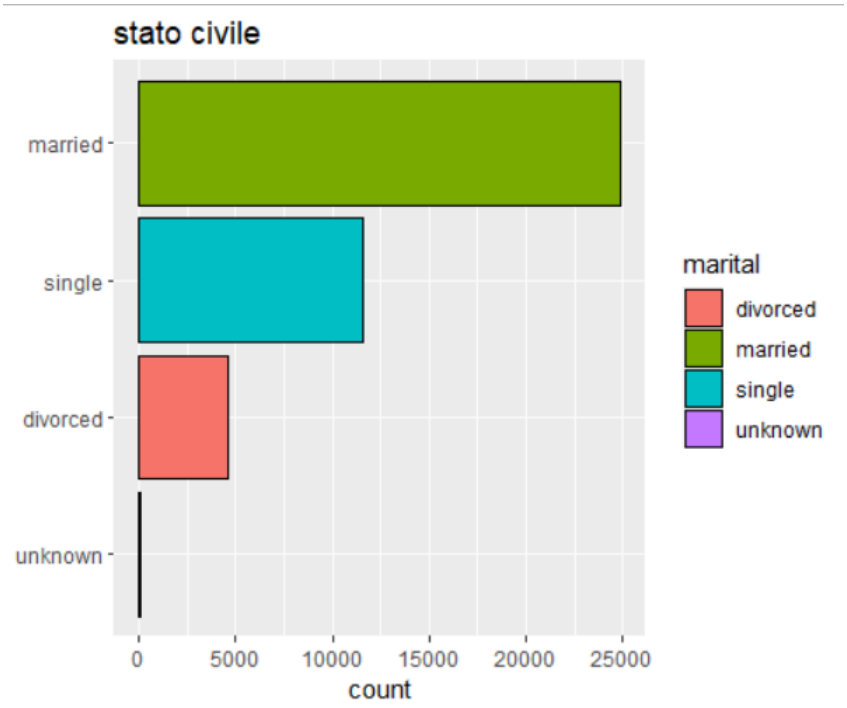

FIG 2.1.3: *istogramma delle frequenze assolute della variabile Marital.*

Come si vede nella fig. 2.1.3, la variabile Marital contiene 71 osservazioni unknown; siccome rappresentano una percentuale inferiore all'1% la decisione è ancora una volta quella di rimuovere queste osservazioni.

Education

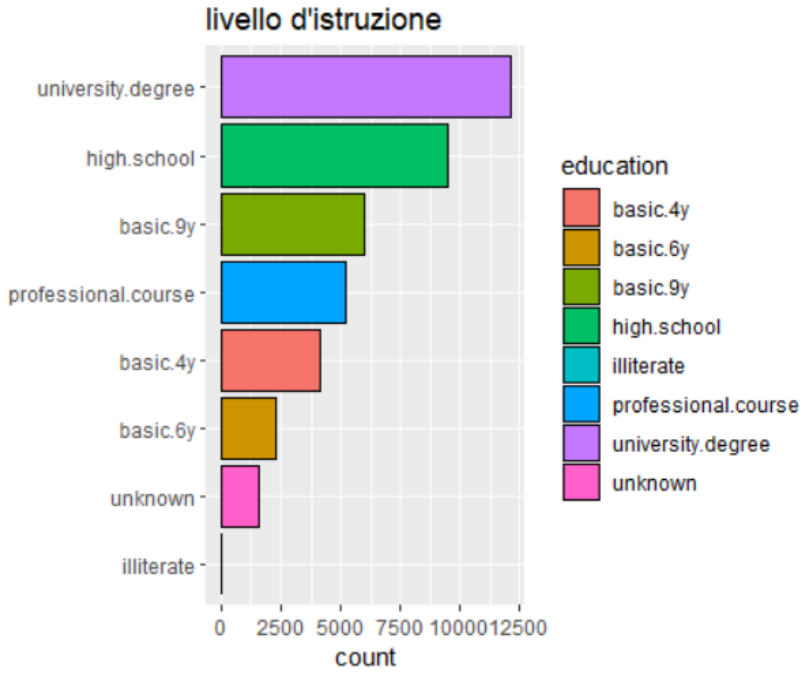

FIG 2.1.4: *istogramma delle frequenze assolute della variabile Marital.*

Per la variabile Education, rappresentata nella fig 2.1.4, abbiamo deciso di incorporare la categoria “illiterate” nella categoria “basic.4y” e di eliminare le osservazioni unknown.
Le motivazioni di queste decisioni sono:

- assegniamo alle persone senza educazione il minor livello più basso cercando di diminuire il numero di classi presenti nella variabile Education.
- le osservazioni unknown sono 1586, meno del 5%. Per questo decidiamo di eliminarle.

Default, Housing e Loan

Anche queste variabili contengono delle osservazioni unknown, tuttavia decidiamo per il momento di non eliminarle per via della quantità elevata di osservazioni unknown; altre scelte risolutive verranno prese nella fase di pre processing.

Pdays

La distribuzione della variabile Pdays si divide in due classi, coloro che hanno osservazione della variabile pari a 999 (ovvero coloro che non sono stati ancora chiamati), e quelli che invece hanno un valore numerico che indica il numero di giorni passati dall’ultimo contatto con il cliente.
Dopo aver notato una forte presenza del valore 999 nella distribuzione della variabile decidiamo di assegnarle due classi: “no” per le osservazioni che avevano valore 999 e “yes” per le altre.

Previous

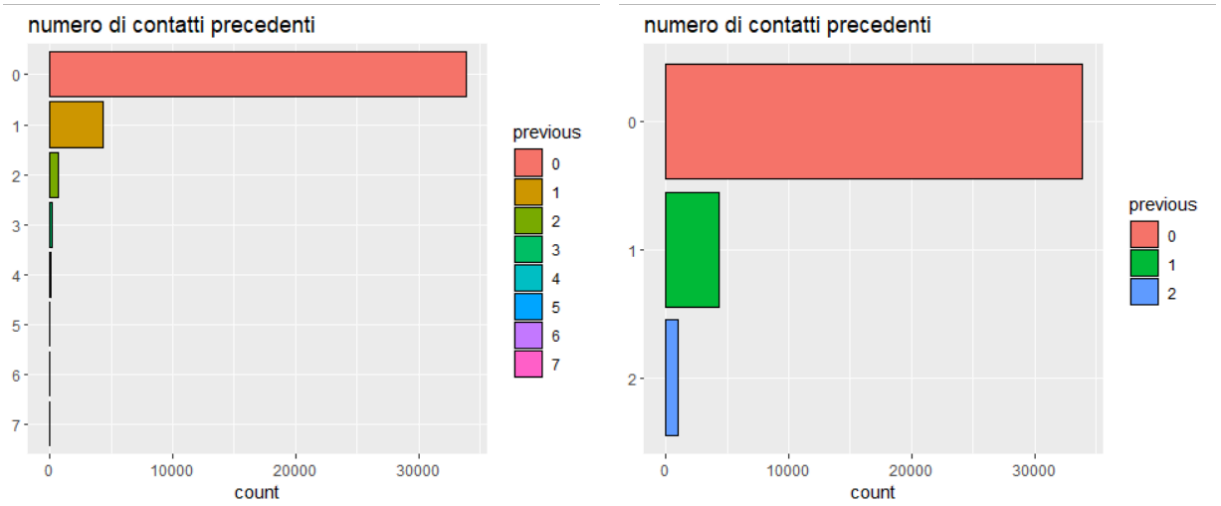

FIG 2.1.5: *istogramma delle frequenze assolute della variabile Previous prima e dopo il raggruppamento.*

Per la variabile Previous, siccome il numero delle osservazioni per le modalità maggiori o uguali a 2, abbiamo deciso di accorpare le osservazioni con tali valori in un'unica classe che simboleggia che il cliente ha precedentemente ricevuto almeno due chiamate, riconducendoci alla situazione illustrata nella figura 2.1.5

Ricapitolando, sono state rimosse 330 osservazioni che presentavano unknown nella variabile job, 71 osservazioni che contenevano questa modalità nella variabile marital e 1586 osservazioni sconosciute nella variabile education; si è dunque deciso di trattare queste osservazioni come dei missing values.
I dati considerati provengono da un'indagine telefonica, quindi è ragionevole pensare che siamo nel caso di MCAR, ovvero dati mancanti distribuiti in modo aleatorio. Siccome complessivamente queste osservazioni rappresentano meno del 5% dei dati (4.8% del totale), la decisione è stata di optare per una strategia passiva: casewise deletion.
Le frequenze relative della variabile target prima e dopo l'eliminazione di queste osservazioni sono uguali (arrotondando alla seconda cifra decimale); questo supporta ulteriormente l'assunzione che si tratti di MCAR e dunque la decisione di rimuovere queste osservazioni è ancor più giustificata.

SPLITTING DEI DATI

Suddividiamo tramite un'estrazione casuale il dataset in training, validation e test nelle seguenti percentuali:

- 60% training
- 20% validation
- 20% test

2.2 PRE PROCESSING

Da questo momento ci limitiamo a trarre le nostre considerazioni unicamente dall'analisi dei dati presenti nel training set.
In questa fase abbiamo valutato la significatività delle singole variabili per il nostro modello; per fare ciò abbiamo valutato le frequenze relative condizionate alla variabile target. Per quelle quantitative invece, essendo tutte variabili economiche, ci siamo concentrati a studiarne la correlazione.

Job

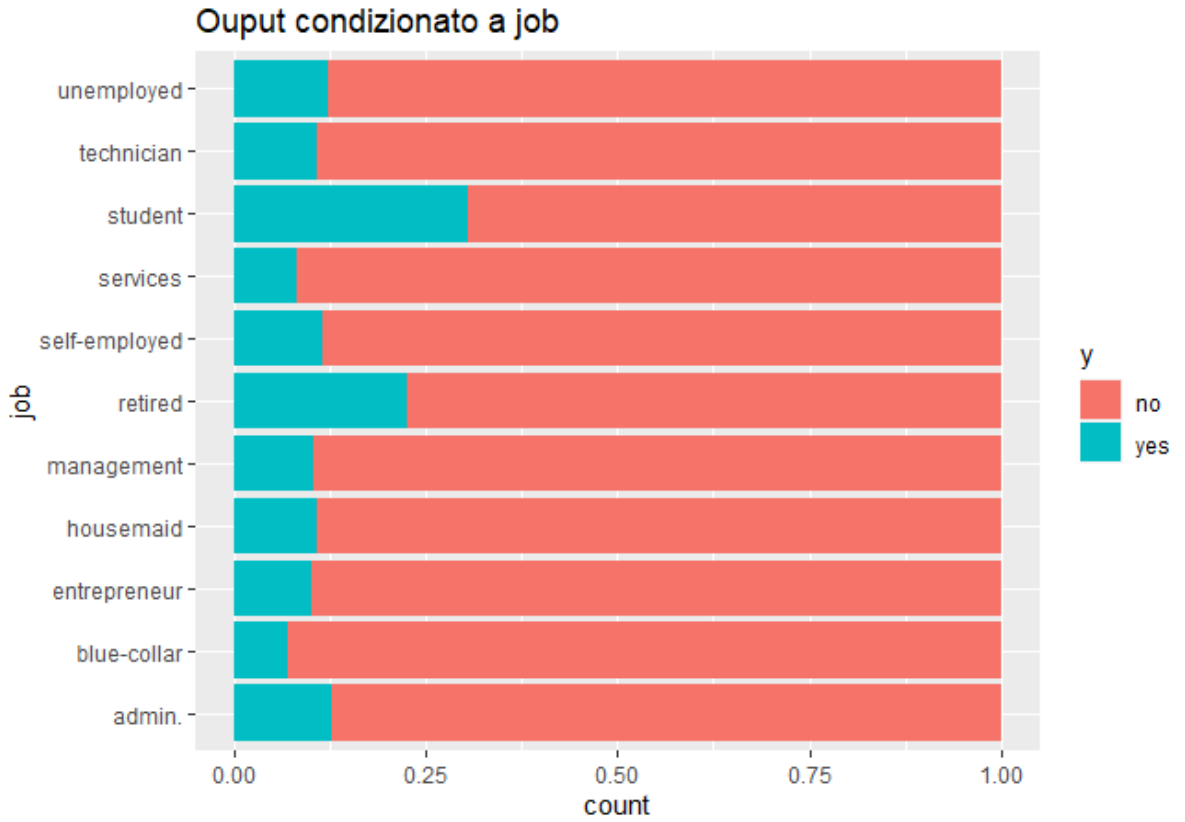

FIG 2.2.1: *istogramma delle frequenze relative della variabile target condizionate alla variabile Job.*

Come si evince dalla figura 2.2.1 il tasso di risposta cambia a secondo del lavoro che si svolge. Inoltre sempre dallo stesso grafico si nota come le classi student e retired hanno delle frequenze relative diverse rispetto alle altre; questo è una

dimostrazione aggiuntiva di come la divisione in classi della variabile età sia utile in termine di analisi.

Marital

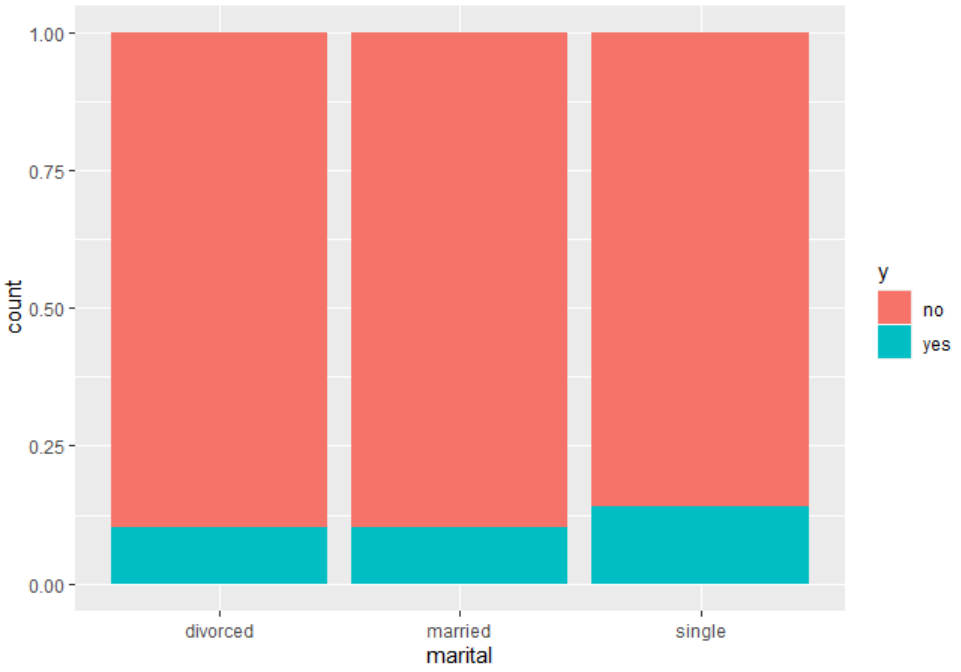

FIG 2.2.2: *istogrammi delle frequenze assolute e relative della variabile target condizionate alla variabile Marital.*

Si nota nella figura 2.2.2 si vede come la percentuale di output per persone single sia leggermente maggiore rispetto alle persone divorziate o sposate. Decidiamo di non valutare adesso se la variabile sia significativa o meno, ma sarà deciso in fase di modeling.

Education

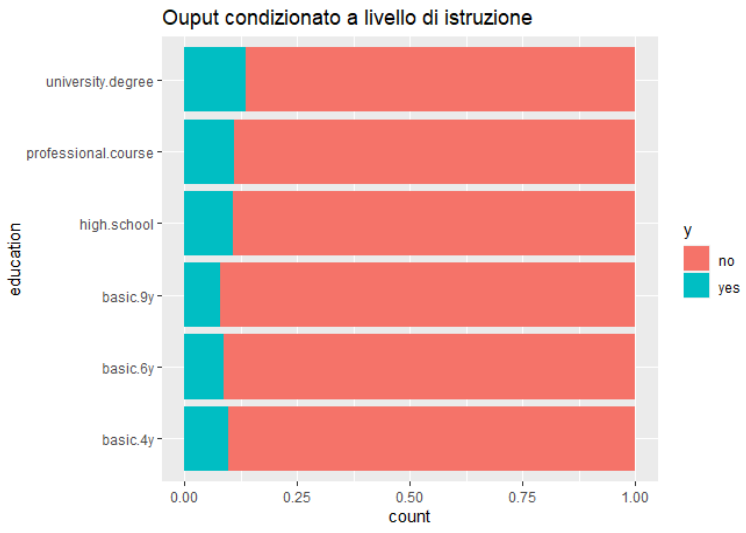

FIG 2.2.3 *Frequenze relative della variabile target condizionata al livello di istruzione*

La fig 2.2.3 sembra suggerire che un livello di istruzione maggiore influisca positivamente sulla variabile target (almeno da basic.9y in poi); da un punto di vista logico, sembra sensata l'idea di una maggiore disponibilità di denaro per coloro che hanno studiato di più, che si traduce nella possibilità di sottoscrivere un deposito vincolato.

Default

La variabile default assume nel dataset 3 modalità: yes, no e unknown. Siccome le osservazioni che hanno modalità yes sono solo 2, la variabile default non dà alcun valore informativo, perciò decidiamo di non considerarla nella fase di modeling.

Housing e Loan

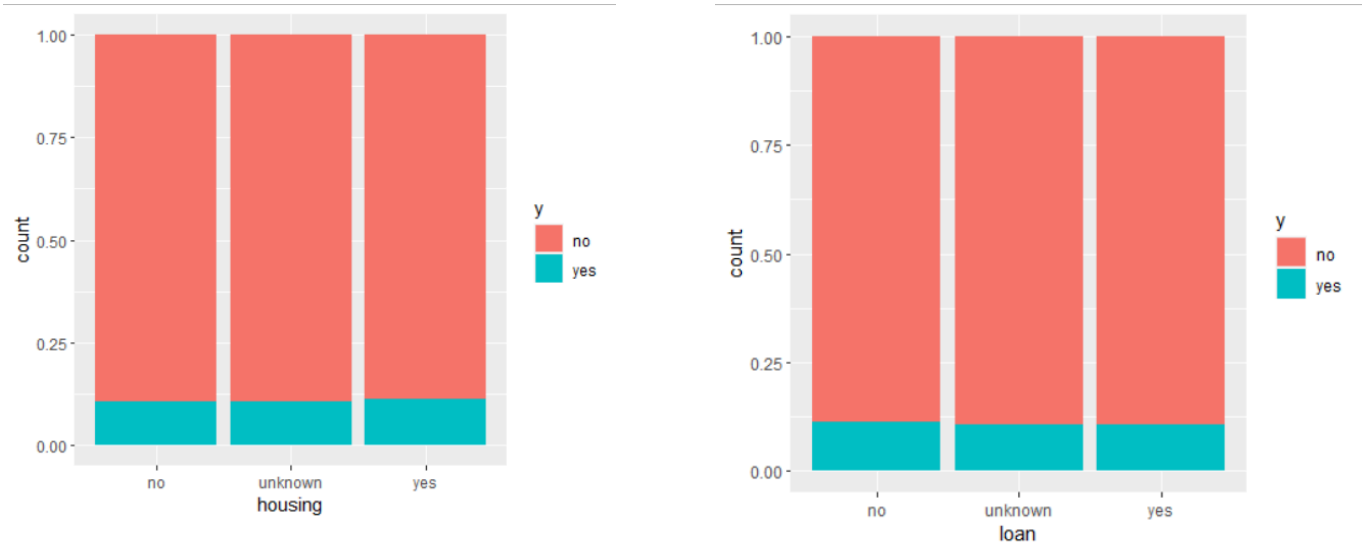

FIG 2.2.4: *output condizionato alle variabili housing (fig. 2.2.4a) e loan (fig 2.2.4b)*

Dalla figura 2.2.4 si evince come sia l'avere un mutuo per la casa che l'avere preso dei soldi a prestito non sia influente a fini predittivi dell'output; questo perchè le percentuali di sì e di no condizionate alle variabili housing e loan sono praticamente identiche. La scelta è dunque quella di eliminare le variabile housing e loan per la fase di modeling; questo risolve anche il problema delle osservazioni che contenevano la modalità unknown in queste variabili.

Contact

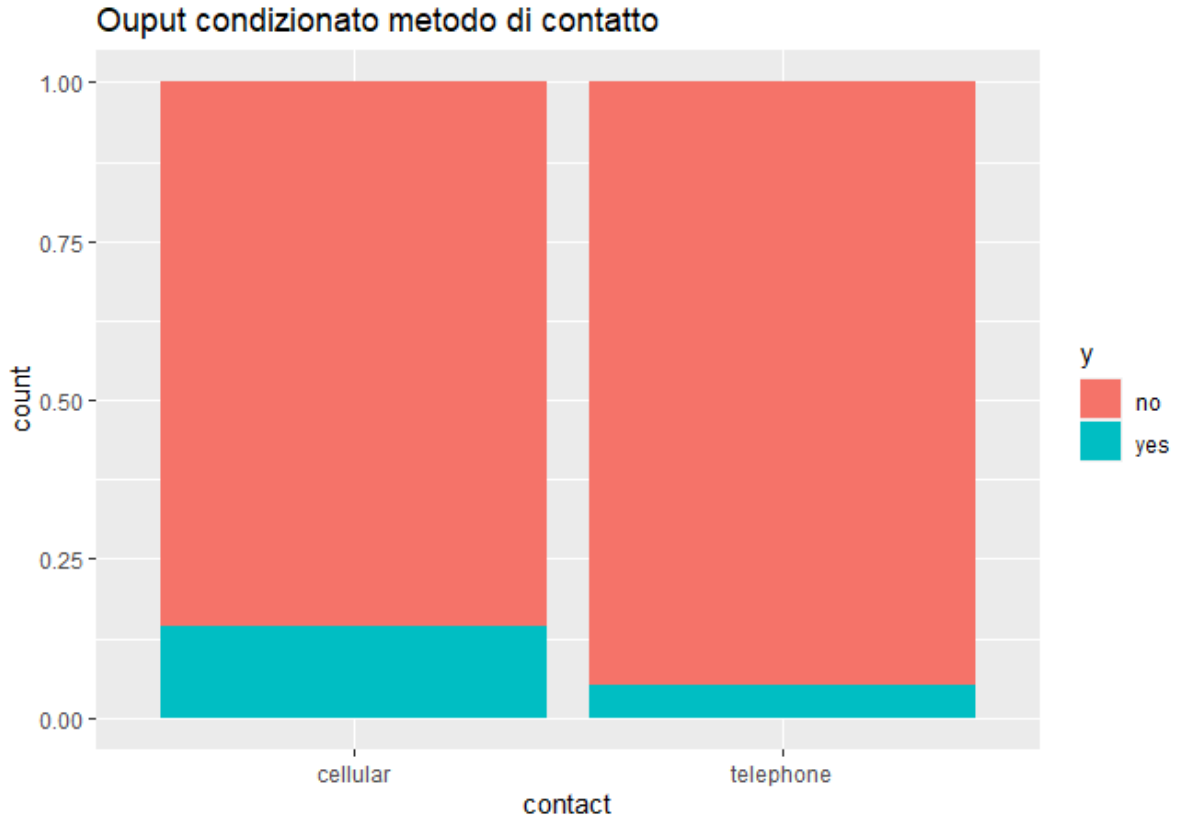

FIG 2.2.5: *frequenza relativa dell'output condizionata al metodo di contatto*

Nella figura 2.2.5 si evince che la differenza di output tra le persone contattate al cellulare e quelle contattate al telefono fisso è piuttosto netta: nella fase di modeling terremo in considerazione la variabile contact. Probabilmente questa differenza è dovuta al cambiamento dei tempi odierni: è molto più facile essere ascoltati contattando qualcuno sul cellulare, piuttosto che sul telefono fisso.

Pcontact e Previous

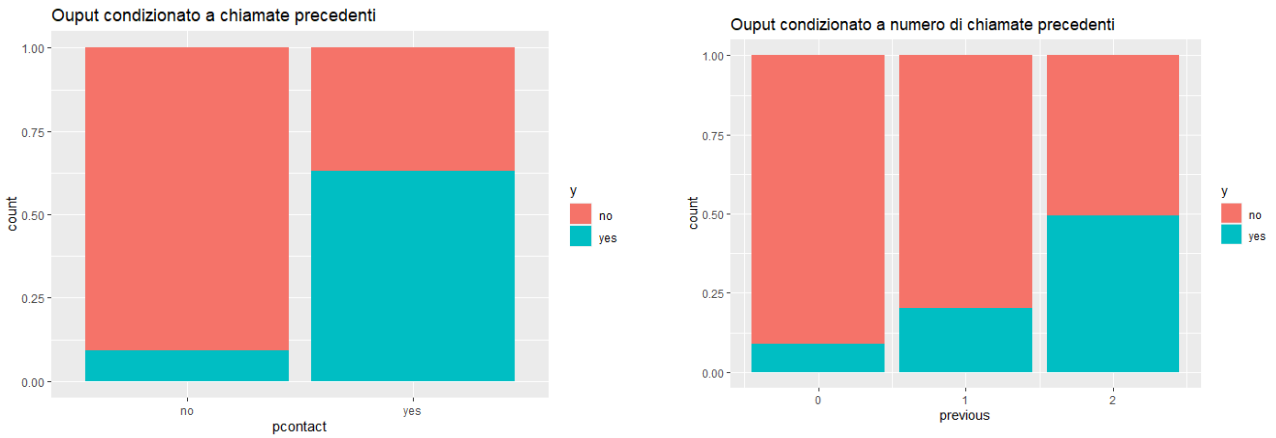

FIG 2.2.6: *output condizionato alla variabile pcontact (fig 2.2.6a) e alla variabile previous (fig 2.2.6b)*

Dalla fig 2.2.6a si evince che l'essere stati contatti in precedenza o meno influenza in modo importante la percentuale di output; inoltre, la probabilità di output positivo aumenta all'aumentare del numero delle chiamate effettuate in precedenza (fig 2.2.6b). Questo risultato ha senso a livello logico; infatti è sensato credere che un cliente che ha risposto a campagne precedenti sia più predisposto ad ascoltare una nuova proposta e a reagire positivamente a questa.

Contact response

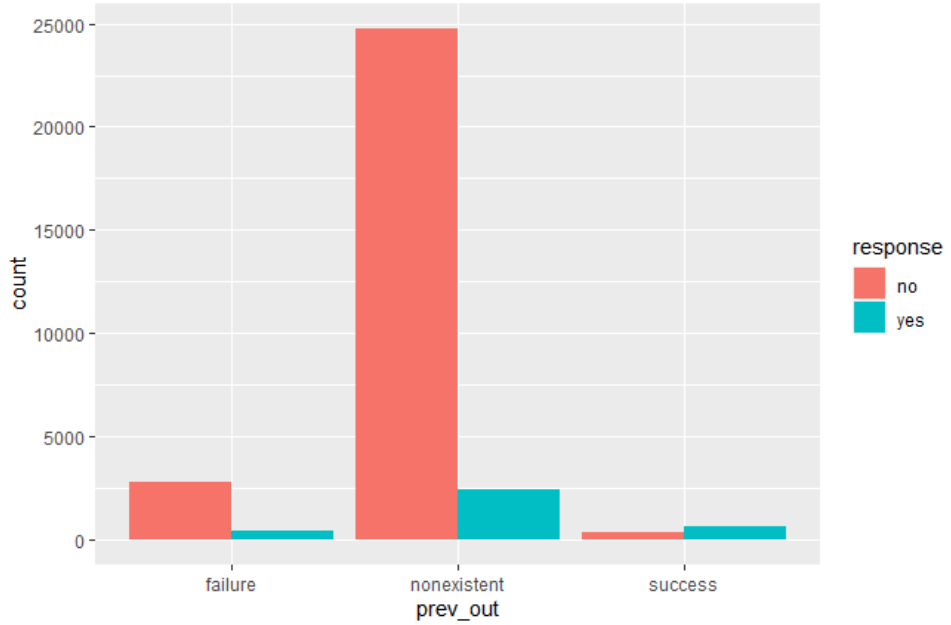

FIG 2.2.7: *frequenze assolute della variabile target condizionate alla risposta delle campagne precedenti*

La scelta di riportare le frequenze assolute e non quelle relative nella figura 2.2.7 è che questa conduce a due evidenti considerazioni:

1. La maggior parte dei clienti non è mai stato contattato in precedenza
2. Tra i clienti già contattati in precedenza, la propensione ad accettare l'offerta è ovviamente maggiore per quelli che avevano detto sì rispetto a quelli che avevano detto no.

Indicatori sociali ed economici

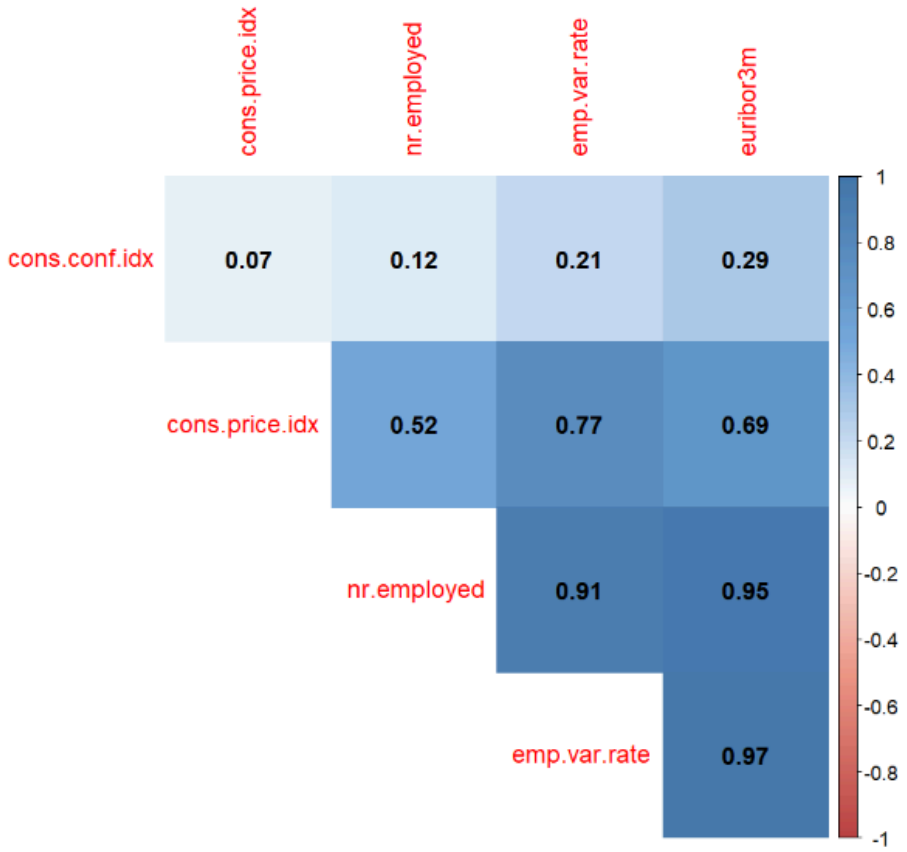

FIG 2.2.8: *grafico per la correlazione delle variabili socioeconomiche*

Le ultime 5 variabili del dataset sono indicatori socioeconomici. Dalla fig 2.2.8, si nota come il tasso euribor a 3 mesi sia quasi perfettamente correlato con le variabili emp. var. rate e nr. employed; per evitare il problema della multicollinearità, si è scelto di rimuovere la variabile euribor a 3 mesi. Da un punto di vista macroeconomico il tasso di interesse e il tasso di occupazione sono collegati tra loro tramite l'inflazione. Infatti nella curva di Phillips fig 2.2.9 si può vedere la relazione che è presente tra inflazione e tasso di disoccupazione (di segno opposto rispetto al tasso di occupazione).

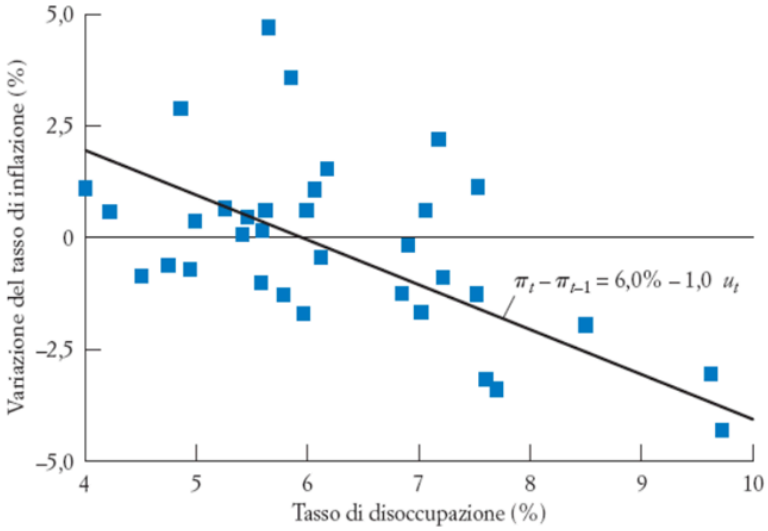

FIG 2.2.9 curva di Phillips

Se aumenta l'inflazione le banche centrali hanno come unico rimedio quello di alzare i tassi di interesse.
Riassumendo: un aumento del tasso di occupazione fa aumentare l'inflazione, da cui l'azione delle banche di alzare i tassi di interesse.

2.3 ASSUNZIONI E METODI

Assunzioni di normalità, varianza e covarianza comune

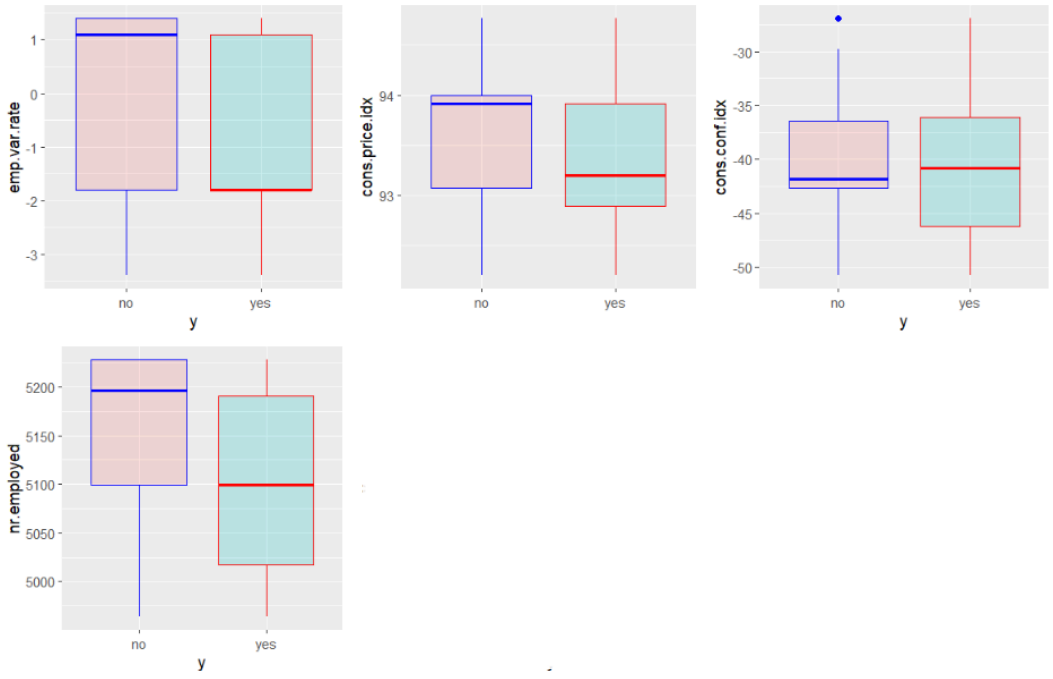

FIG. 2.3.1: boxplot delle variabili economiche condizionati alla variabile target.

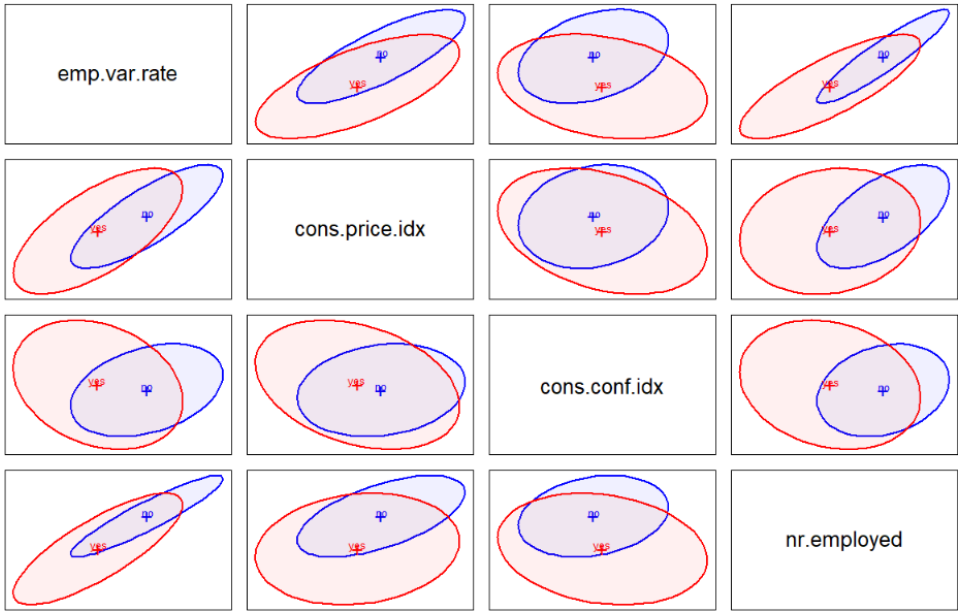

FIG. 2.3.2: Grafico covarianze variabili numeriche messe a confronto.

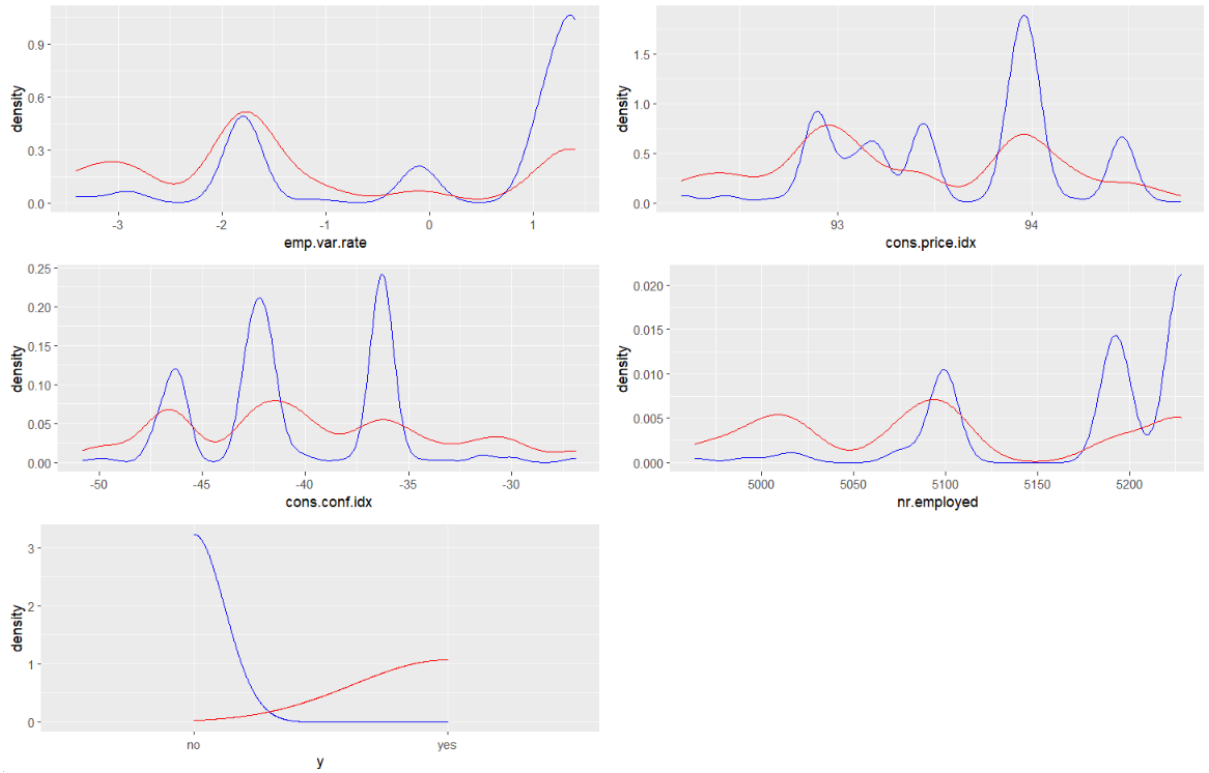

FIG. 2.3.3: Densità condizionate alla classe delle variabili numeriche.

Dopo queste analisi abbiamo scartato i metodi LDA e QDA, infatti le nostre variabili non hanno né varianza (fig. 2.3.1) né covarianza (fig. 2.3.2) comune e non hanno una distribuzione normale (fig. 2.3.3).

Regressione logistica

Data la presenza di una variabile target qualitativa dicotomica, come primo modello abbiamo testato la regressione logistica.
In primo luogo abbiamo stimato il modello sul training inserendo tutte le variabili rimaste e abbiamo applicato una selezione stepwise basata sul Akaike Information Criterion (AIC) al fine di identificare le variabili più informative a fini previsivi.
Il modello risultante comprende le variabili job, education, contact, pcontact, outcome, emp.var.rate, cons.price.idx e cons.conf.idx.
In seguito abbiamo verificato la presenza di punti influenti e dopo averli eliminati abbiamo ristimato i coefficienti del modello in modo tale da effettuare una verifica dell'accuratezza sul validation set.
Dopodiché abbiamo usato il train + validation per ottimizzare i coefficienti del modello utilizzando soltanto le variabili sopra indicate.
Il modello finale è stato poi ottimizzato (eliminazione dei punti influenti) e valutato tramite i dati del test.
Inoltre per migliorare la sensitivity abbiamo deciso di utilizzare una soglia del Bayes decision per l'attribuzione della classe "yes" pari a 0.13; anche questa scelta l'abbiamo effettuata nella fase train-validation e applicata successivamente al test.
La motivazione alla base di ciò è che il nostro intento è quello di voler predire il maggior numero di esiti positivi e abbassando tale soglia compensiamo lo sbilanciamento delle classi a favore della nostra variabile di interesse.
Dalla diagnostica della regressione logistica siamo anche riusciti a capire quali fossero le caratteristiche che più influiscono positivamente sull'esito del contatto con il cliente.

KNN

Il metodo KNN è un metodo non parametrico che non fa alcuna assunzione preliminare sulle distribuzioni di variabili input, e ha come unico parametro il parametro di tuning (k).
Per poter applicare questo metodo, tutte le variabili qualitative categoriche sono state trasformate in variabili dummies, ad eccezione della variabile job che è stata esclusa per via dell'alto numero di modalità (che avrebbe reso computazionalmente troppo intensiva la stima del modello). Inoltre, la variabile education è stata trasformata in una variabile numerica, poiché unica variabile qualitativa ordinale (un maggiore grado di educazione corrisponde, generalmente, ad una maggiore disponibilità economica e dunque possibilità di sottoscrivere depositi a termine)
Testando tutti i k tra 1 e 25 sul validation set si ottiene che il k che produce un error rate minore è k = 21.

Abbiamo deciso di non testare k maggiori di 25 siccome il nostro problema è sbilanciato e quindi con k grandi le nostre osservazioni sarebbero state classificate più frequentemente come “no”.
Durante il training ci siamo accorti che un k basso non era sufficiente per gestire lo sbilanciamento delle classi, da cui abbiamo deciso di abbassare la soglia di attribuzione della classe “yes” a 0.09 per avere una sensitivity simile a quella della regressione logistica e di poter così confrontare, a parità True Positive Rate, la specificity e l’accuracy.

3. RISULTATI

VALUTAZIONE CAPACITÀ DI CLASSIFICAZIONE- VALIDATION:

Valuteremo ora come si comportano i due modelli rispetto al Validation. Andremo quindi a prevedere i valori del Validation set e verificare quale approccio meglio classifica nuove istanze (osservazioni) del problema.

- **Regressione logistica**

Matrice confusione della regressione logistica

$Pr(y = 1|X = x) \geq 0.5$

Confusion Matrix		*True class*		*Total*
		0	1	
Prediction	0	6901	672	7573
	1	93	172	265
Total		6994	844	

Accuracy: 0.9024
Sensitivity: 0.20379
Specificity : 0.98670

Confronto ROC curve

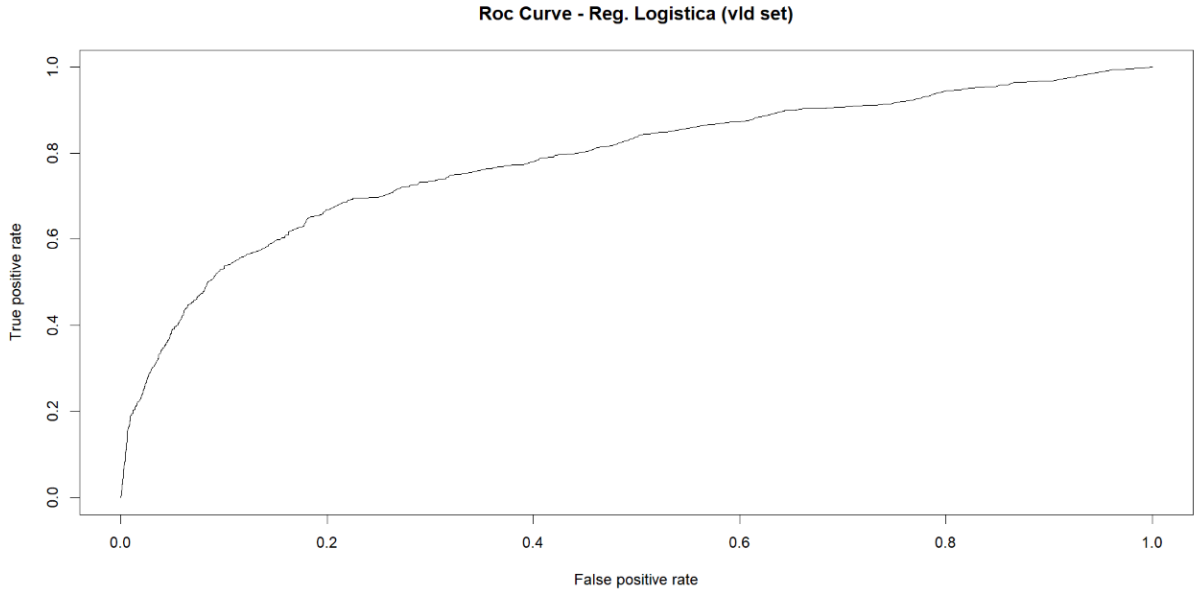

FIG. 3.1: Grafico Roc curve del modello di regressione logistica (Validation set)

La Roc curve traccia la probabilità di un risultato vero positivo rispetto ai positivi totali (sensitivity) in funzione della probabilità di un risultato falso positivo rispetto ai negativi totali per una serie di punti di cut-off determinati da soglie di bayesian decision differenti.

Come possiamo vedere dalla fig.3.1 la regressione logistica con soglia di assegnazione alla classe “yes” pari a 0,13 è più performante per il nostro scopo. Nell’analisi in questione, la sensitivity è la quantità che cerchiamo di massimizzare col modello stimato; infatti, per una banca l’errore più grave è quello di non contattare un cliente intenzionato ad accettare l’offerta piuttosto che quello di contattare un cliente non intenzionato ad accettare. Per questo preferiamo il modello con la soglia della probabilità di assegnazione pari a 0.13.

$Pr(y = 1|X = x) \geq 0.13$

Confusion Matrix		*True class*		*Total*
		0	1	
Prediction	0	5860	329	6189
	1	1134	515	1649
Total		6994	844	

Accuracy: 0.8133
Sensitivity : 0.61019
Specificity : 0.83786

Notiamo che utilizzando un valore soglia pari a 0.13 otteniamo un grosso miglioramento per quanto riguarda il valore di sensitivity (da 172 a 515 osservazioni classificate True Positive, cioè da 0.20 a 0.61 circa), a discapito della specificity e dell'accuracy che si riducono con una variazione meno importante (da 0.98 a 0.83 e da 0.90 a 0.81).

Dall'analisi della diagnostica della regressione lineare i coefficienti più significativi risultano:

- job retired: 0.308
- job student: 0.401

coerente con le medie marginali ottenute con il preprocessing, la ragione di scommessa per chi è in pensione o chi è studente è maggiore.

- contact telephone: -0.963

contattare un cliente al telefono dà meno probabilità di successo, è meglio contattarlo tramite cellulare.

- pcontctyes: 0.930

Un cliente che ha già accettato una nostra vecchia offerta di una vecchia campagna di marketing sarà sicuramente da ricontattare.

- poutcomenonexistent: 0.610
- poutcomesuccess: 0.994

Tenendo in considerazione che la baseline è poutcome = insuccess, questi coefficienti ci dicono che non conviene chiamare chi ha già detto di no a delle nostre offerte, ma è meglio optare per nuovi clienti o clienti che già hanno accettato in passato delle nostre offerte.

- emp.var.rate: -3.621
- cons.price.idx: 3.219
- cons.conf.idx: 1.142

I coefficienti delle variabili macroeconomiche dicono che la disoccupazione, l'indice dei prezzi al consumo e l'indice della fiducia dei consumatori sembrano avere una associazione positiva con la risposta.

- **KNN**

Testando tutti i k tra 1 e 25 sul validation set si ottiene che il k che produce un errore rate minore è k = 21, come lo dimostra la fig. 3.2.

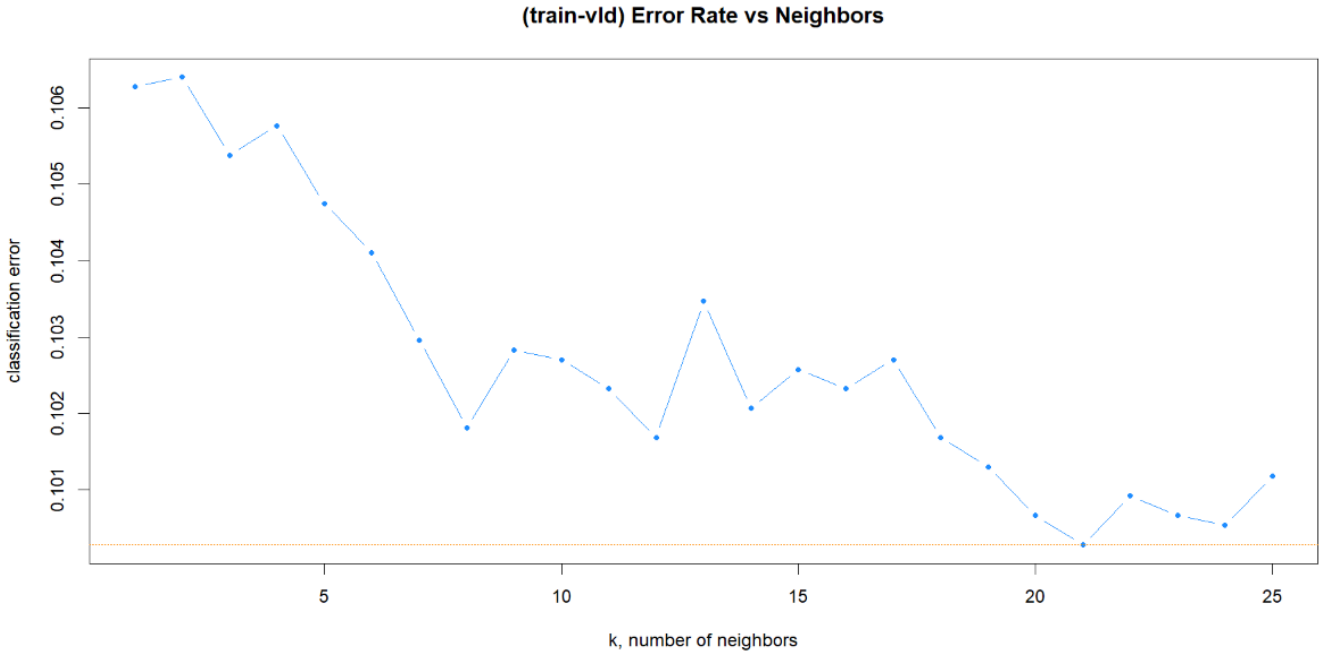

FIG. 3.2: Grafico (train-vld) Error Rate vs K neighbors.

Stimando il modello knn con k = 21 e usandolo per fare previsione sul validation, si ottiene un test error rate circa pari al 10%.

Matrice di confusione per KNN

$Pr(y = 1|X = x) \geq 0.5$

Confusion Matrix		*True class*		*Total*
		0	1	
Prediction	0	6876	678	7554
	1	118	166	284
Total		6994	884	

Accuracy: 0.8984
Sensitivity : 0.19668

Specificity : 0.98313

Confronto ROC curve

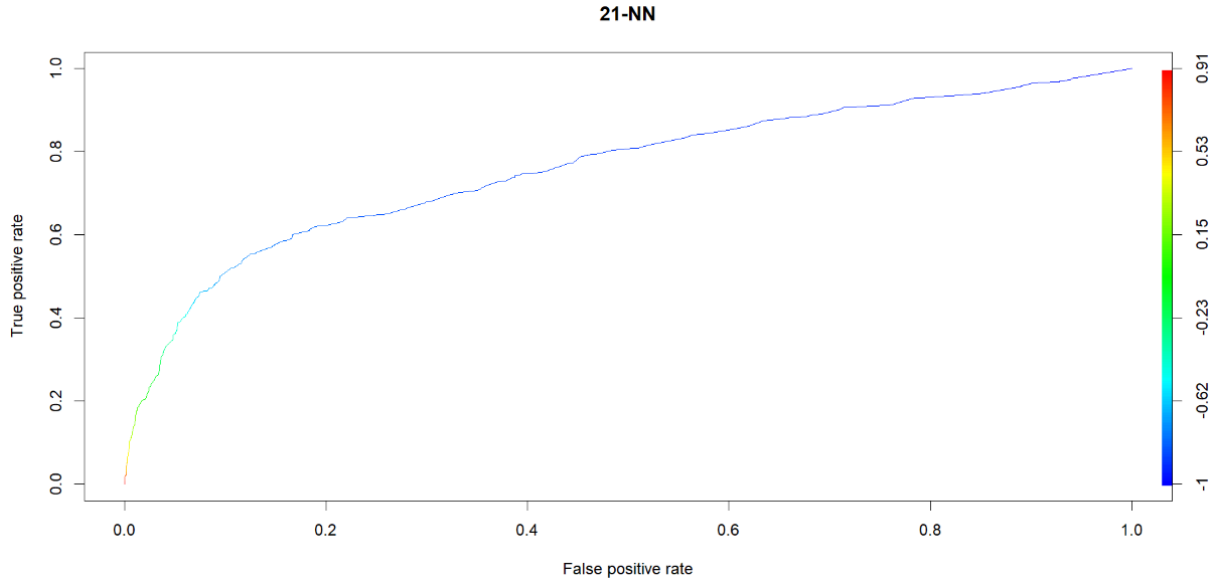

FIG. 3.3: Grafico Roc curve del modello 21-NN (Validation set)

La Roc curve del 21-NN (fig. 3.3) sembra piuttosto simile a quella della regressione logistica. Abbiamo dunque deciso di far variare la soglia di assegnazione alla classe 1 in modo da avere una sensitivity simile a quella ottenuta per la regressione logistica, per confrontare successivamente accuracy e specificity, scegliendo dunque il modello migliore.

$Pr(y = 1|X = x) \geq 0.09$

Confusion Matrix		*True class*		*Total*
		0	1	
Prediction	0	5491	321	5812
	1	1503	523	2026
Total		6994	884	

Accuracy:0.7673
Sensitivity:0.61967
Specificity:0.78510

A parità di sensitivity, il modello di regressione logistica performa leggermente meglio sia in termini di accuracy che in termini di specificity. Portiamo dunque entrambi i modelli alla fase di test.

VALUTAZIONE MODELLI - TEST:

Valuteremo ora come si comportano i due modelli rispetto al Test set. Andremo quindi a prevedere i valori del Test set, i quali non sono mai stati utilizzati nelle fasi di training dei modelli.

- **Regressione logistica**

Matrice di confusione della regressione logistica

Utilizziamo il valore soglia determinato nella fase train-validation.

$Pr(y = 1|X = x) \geq 0.13$

Confusion Matrix		*True class*		*Total*
		0	1	
Prediction	0	5855	308	6163
	1	1148	528	1676
Total		7003	836	

Accuracy: 0.8143
Sensitivity : 0.63158
Specificity : 0.83607

Il modello di regressione logistica ha performance molto simili sia nel test che nel training sia per quanto riguarda la matrice di confusione sia per quanto riguarda la ROC curve fig.3.4, il nostro modello non sembra avere problemi di overfitting o underfitting.

- **KNN**

Matrice di confusione per KNN

$Pr(y = 1|X = x) \geq 0.09$

Confusion Matrix		*True class*		*Total*
		0	1	
Prediction	0	5562	303	5865
	1	1441	533	1974
Total		7003	836	

Accuracy: 0.7775
Sensitivity : 0.63756
Specificity : 0.79423

- **Confronto Reg. Logistica e KNN tramite Roc Curve**

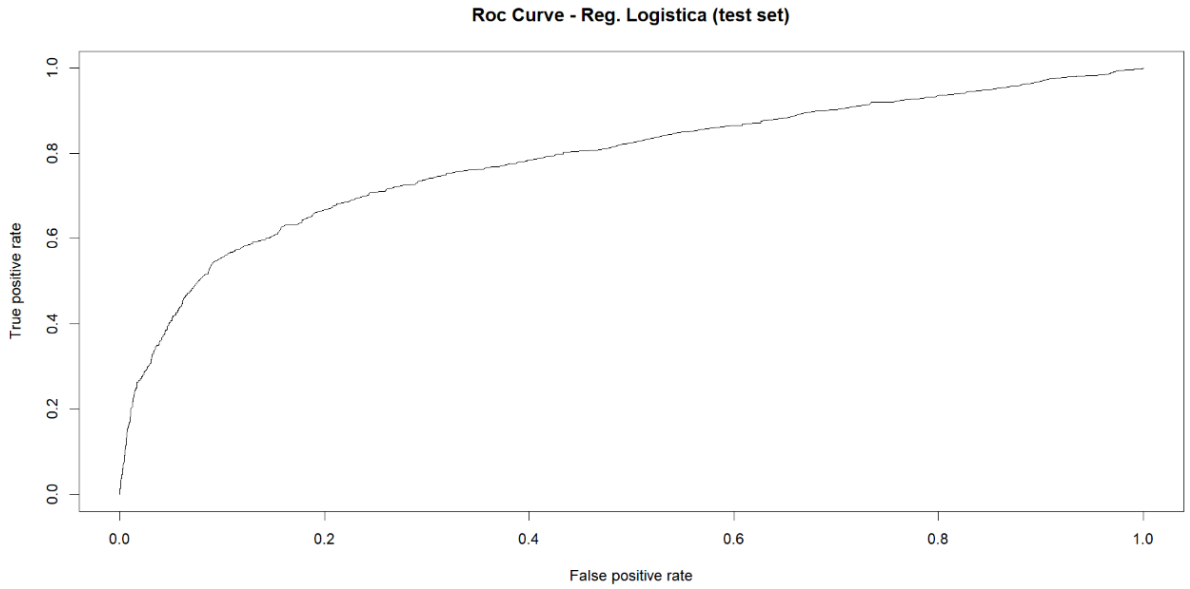

FIG. 3.4: Grafico Roc curve del modello di regressione logistica (Test set)

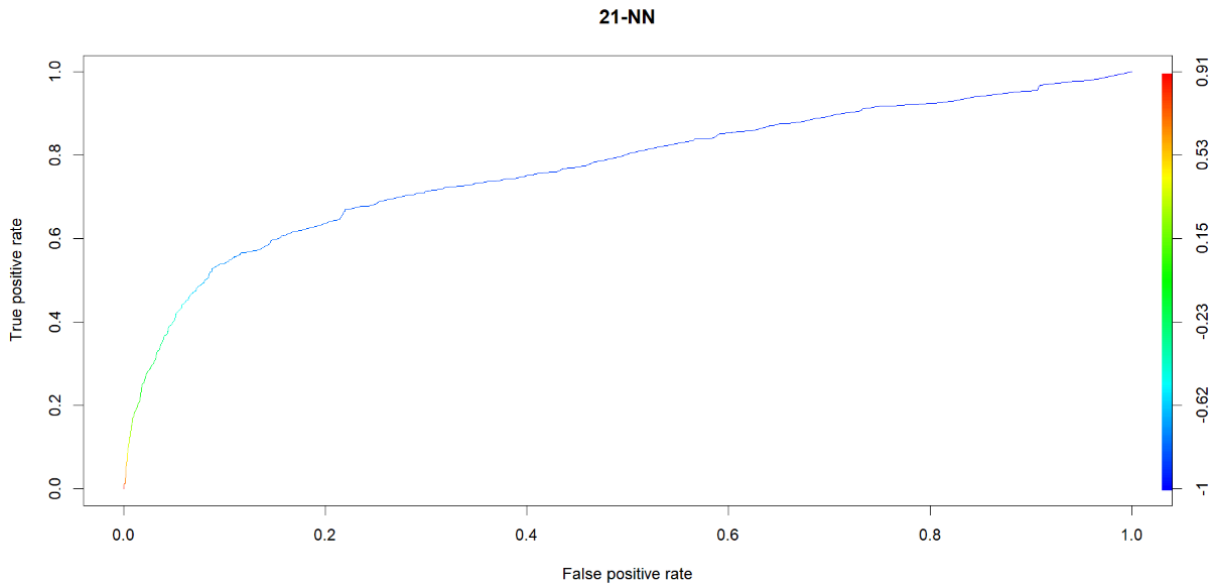

FIG. 3.5: Grafico Roc curve del modello 21-NN (test set)

Confrontando le figure 3.4 e 3.5 notiamo una performance leggermente migliore per il modello di regressione; le performance le andremo però a verificare più nel dettaglio confrontando i valori numerici dei due modelli.

- **Confronto Reg. Logistica e KNN in termini di Errore Rate e Sensitivity**

	Regressione logistica	*Knn*
Training Error (train+validation)	0.19228	0.21893
Test Error	0.18573	0.22247
Sensitivity (test)	0.63158	0.63756
Specificity (test)	0.83607	0.79423

Notiamo dunque che a parità di sensitivity, il modello di regressione logistica performa meglio sia per quanto riguarda il Training Error che per il valore del Test Error. Optiamo quindi per il modello di Regressione Logistica.

4. DISCUSSIONI

L'obiettivo iniziale era quello di ottimizzare la campagna marketing per un deposito a termine di una banca attraverso dati quantitativi e qualitativi dei clienti della banca stessa. L'intento dunque è stato quello di predire le persone che a seguito di una nostra chiamata avrebbero accettato la proposta, in modo tale da chiamare, in una possibile campagna futura, individui con caratteristiche simili.

Prima di iniziare l'analisi abbiamo eliminato alcune variabili poiché non sarebbero state conosciute a priori, e ne abbiamo eliminate altre durante la fase di pre processing poiché non sarebbero state utili a discriminare tra le classi della variabile target.
Il modello che si rivela più utile per il nostro fine è la regressione logistica, principalmente per questi motivi:

1. le assunzioni di normalità non sono rispettate.
2. nel dataset sono presenti variabili qualitative categoriche non ordinali.
3. a parità di sensitivity aveva migliori performance rispetto agli altri parametri, come specificity e accuracy.

L'accuracy e la sensitivity del modello risultano essere piuttosto buone, quindi può essere un'ottima strategia implementare nel modello di business della banca il nostro modello previsivo; infatti da una percentuale di "yes" dell'11% (frequenza relativa di "yes" del dataset) si può arrivare ad ottenere una percentuale pari al 31.5%, essendo 1676 (1148+528 = FP+TP) le persone che a seguito della nostra previsione opteremo di chiamare e 528 le persone che effettivamente risponderanno positivamente alla nostra offerta (TP).
Per migliorare ulteriormente le performance dei modelli di previsione servirebbe probabilmente avere più parametri sullo stato patrimoniale del cliente, come possono essere il suo conto corrente o il suo stipendio lavorativo annuo.

Bibliografia

- https://www.bankbound.com/blog/marketing-increase-deposits/

- https://archive.ics.uci.edu/ml/datasets/bank+marketing<