Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica

Progetto Data Mining
Credit Card Default

Studenti:
Alessandro Berti
Luca Corbucci
Eugenio Paluello

Anno Accademico 2018/2019

Indice

Capitolo 1

Data Understanding

In questa sezione eseguiamo il processo del data understanding sul dataset. In particolare, nella sottosezione 1.1 analizziamo gli attributi che definiscono ciascun record. Nella sottosezione 1.2, mostriamo con l'ausilio dei grafici la distribuzioni degli attributi e delle relative statistiche. Nella sottosezione 1.3, forniamo la valutazione della qualità dei dati nel dataset. Nella sottosezione 1.4, illustriamo come abbiamo trasformato i valori degli attributi per ottenere delle analisi più significative. Infine, nella sottosezione 1.5, esplicitiamo la correlazione degli attributi attraverso il Pearson's coefficient.

1.1 Semantica dei dati

Ogni record presente all'interno del dataset viene descritto da 24 attributi, in particolare $credit\_default$ è la nostra variabile dipendente. Nella Tabella 1.1 elenchiamo per ciascun attributo il nome, la descrizione, il tipo ed il dominio ad esso associato.

Tabella 1.1: Descrizione degli attributi del dataset

Nome	Descrizione	Tipo	Dominio
Limit	Quantità di dollari NT spendibili dal cliente. Comprende sia il credito del cliente sia quello della famiglia (ove presente)	Numerico, discreto	$N_{>0}$
Sex	Sesso del cliente	Stringa, categorico	{Male, Female}
Education	Livello di istruzione del cliente	Stringa, ordinale	{Graduate School, High School, University, others}
Status	Stato civile del cliente	Stringa, non ordinale	{Married, Single, Other}
Age	Età del cliente	Numerico, discreto	$N_{>0}$
Ps-apr...Ps-sep	Status del pagamento del cliente per ogni mese	Numerico, discreto	[-2,8]

Continued on next page

Tabella 1.1

Nome	Descrizione	Tipo	Dominio
Ba-apr...Ba-sep	Quantità di soldi che il cliente ha speso in ognuno dei mesi	Numerico, discreto	N
Pa-apr...Pa-apr	Quantità di dollari NT pagati dal cliente, il pagamento si riferisce alla spesa del mese precedente	Numerico, discreto	$N_{>0}$
Credit default	Indica se la persona è in credit default	Numerico, categorico	{no,yes}

1.2 Distribuzione degli attributi e statistiche

In questa sezione descriviamo nel dettaglio il dataset, prima studieremo le statistiche e le distribuzione degli attributi singolarmente, mentre nella seconda parte confronteremo quanto ciascun attributo incide sulla classificazione del credit default per ciascun utente. Abbiamo analizzato gli attributi categorici come *Sex*, *Status* e *Education* producendo dei barchart (Figura 1.1) per studiare le tipologie di utenti della banca. Osserviamo che, per quanto riguarda il sesso, la maggior parte dei clienti risulta donna, mentre non c'è un significativo distacco tra la frequenza dei consumatori sposati e single; ne sono poi presenti altri, con bassa frequenza, che non rientrano in queste due prime categorie e che vengono classificati come "others". Discorso simile può essere fatto con l'attributo *Education*, nel quale troviamo una prevalenza di utenti con un grado di istruzione universitario.

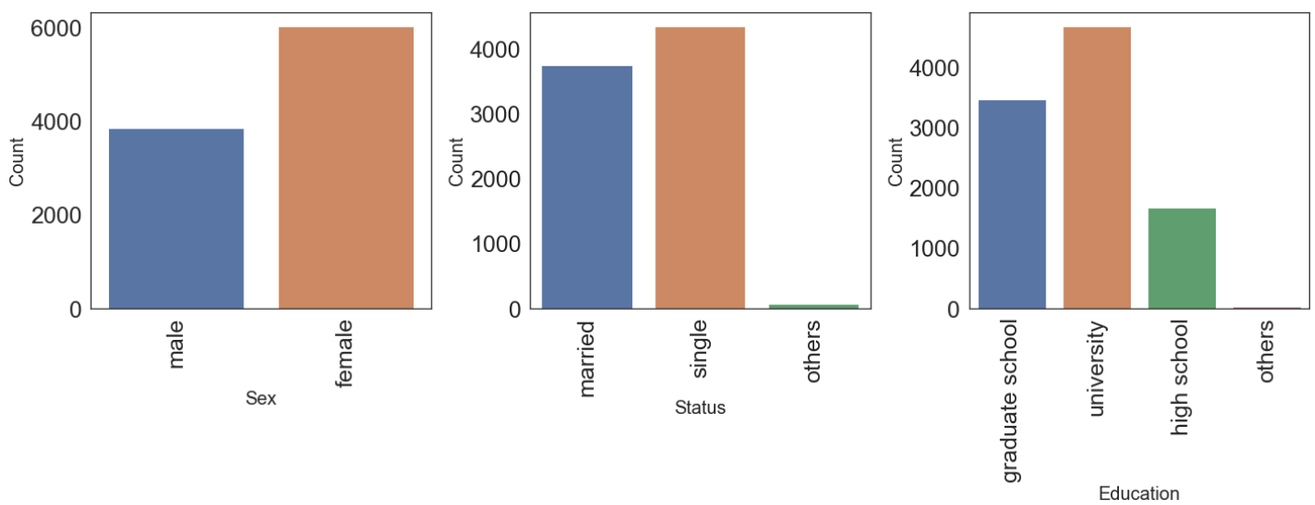
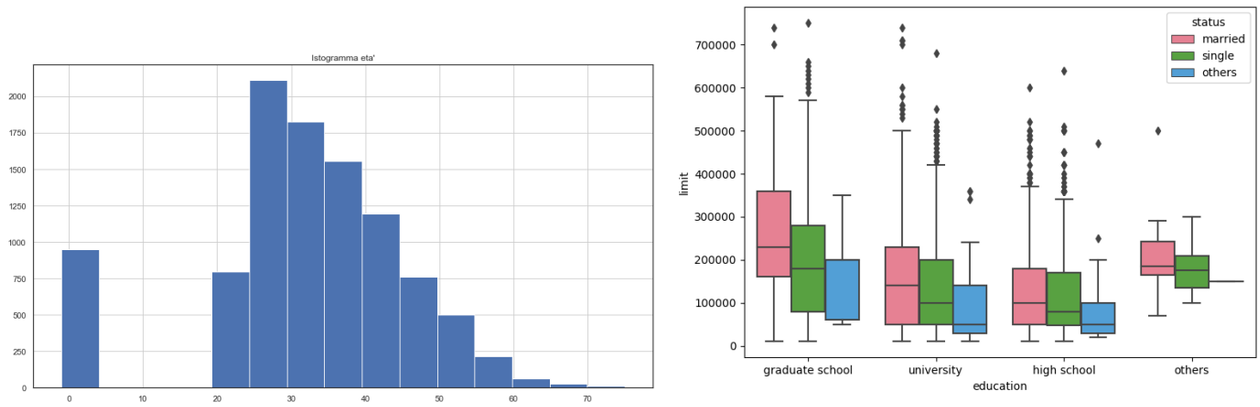


Figura 1.1: Distribuzione delle Variabili

Per quanto riguarda, l'età, nella Figura 1.2a osserviamo che il cliente della banca più frequente è quello in un'età compresa tra i 25 ed i 30 anni. Sono inoltre presenti 1000 valori (-1) che non soddisfano il vincolo dell'attributo. Abbiamo analizzato la distribuzione degli attributi *status* ed *education* rispetto a *limit*. Nella Figura 1.2b notiamo come, in media, un livello di educazione maggiore corrisponda ad un limite maggiore, allo stesso modo i clienti sposati hanno un limite più alto. Questo indica che la banca è più incline ad offrire un limite di spesa più alto a chi offre delle garanzie sociali maggiori.

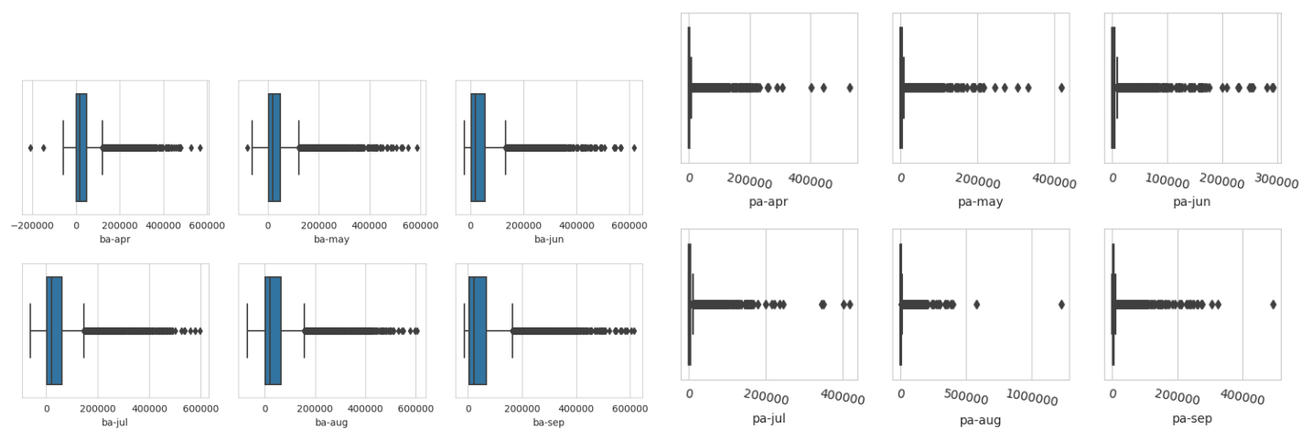


(a) Distribuzione *age*

(b) BoxPlot Education Status

Figura 1.2

Le distribuzioni dei Bill Amount (Figura 1.3a) e dei Payment Amount (Figura 1.3b), sono state rappresentate con dei boxplot per comprendere meglio quali fossero gli outlier in modo da utilizzarli per analisi future. Confrontando i due grafici di Figura 1.3a e 1.3b osserviamo che nel primo caso, a differenza del secondo, troviamo anche valori negativi. Abbiamo interpretato questa informazione come una sorta di credito che la banca offre ai clienti. Un altro dettaglio interessante è che se il Bill Amount è compreso in media tra 0 e 50000 Dollari NT allora il Payment Amount è molto più basso (compreso tra 0 e 4000). Ciò sembrerebbe riflettere un comportamento dei clienti incline al debito, dal momento che, in media, è restituito alla banca meno del dovuto. Tuttavia, questa tendenza verso il debito non è troppo accentuata come osservato in Figura 1.4a. Infatti, lo stato del pagamento, in media, è -1 o 0, ovvero: l'importo speso nel mese precedente è stato restituito completamente (-1) o in parte (0). Nonostante ciò, cercando di collegare queste conoscenze, sembra che il cliente medio della banca taiwanese adotti un approccio "borderline", cioè, si fronteggia continuamente contro il rischio di indebitamento. Generalizzando questo, possiamo supporre che la situazione potrebbe essere presente anche in altre banche taiwanesi, suggerendo un malessere economico del paese.



(a) Box Plot Bill Amount

(b) Box Plot Payment Amount

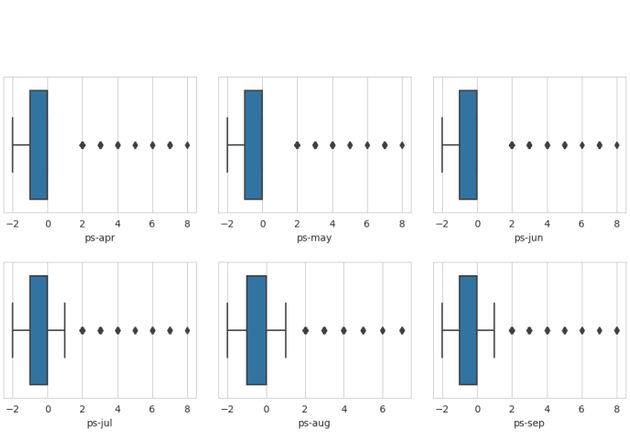
Figura 1.3

Dopo questa prima analisi sulle distribuzioni, ci siamo concentrati nell'identificare come gli attributi presi in considerazione incidessero sulla classificazione del credit default di ciascun utente. Inizialmente abbiamo preso in esame la percentuale dei clienti in credit default (Figura 1.4b) che si aggira intorno al 22%. Tale informazione successivamente ci tornerà utile nei grafici normalizzati per validare o meno, eventuali ipotesi riguardo l'influenza di un attributo sulla classificazione degli utenti.

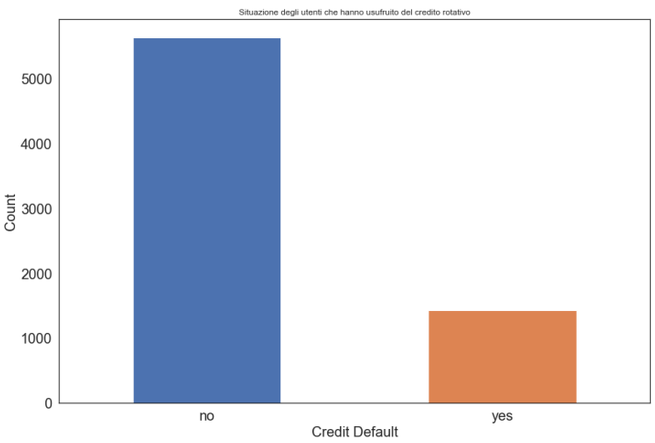
Passando al grafico della distribuzione dell'età rispetto al credit default, in particolare in quello normalizzato 1.4d, notiamo che per ciascuna età, la percentuale di clienti in credit default è comparabile

alla percentuale degli utenti in credit default su tutto il dataset indipendentemente dall'età (22%). Ciò ci suggerisce che l'attributo *age* non funge da discriminante per stabilire la classificazione di un utente.

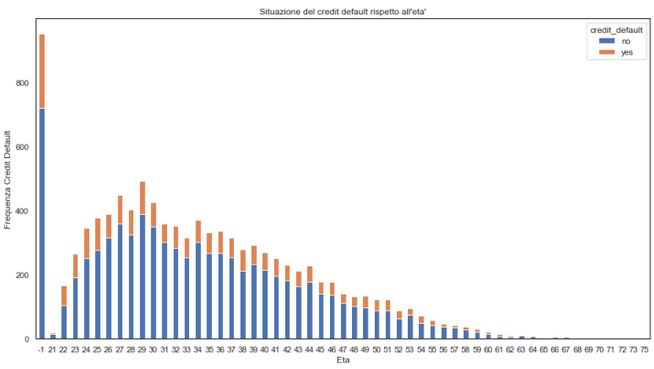
Osservando i grafici 1.4e, 1.4f, 1.4g, anch'essi normalizzati comparati rispetto al credit default, possiamo trarre la stessa conclusione riguardante l'età: ovvero non sembrano essere attributi influenzanti la classificazione di un utente. Una piccola nota di attenzione comunque rimane: nel grafico 1.4e, sembra che i clienti aventi *education* = "others", siano meno propensi ad andare in credit default, vale invece l'opposto per quanto riguarda il valore "others" nel grafico 1.4g. Tale conclusione è invalidata dal fatto che solo il 3.6% degli utenti ha come *education* "others" e solo il 7.5% ha come *status* "others", rendendoli di fatto irrilevanti.



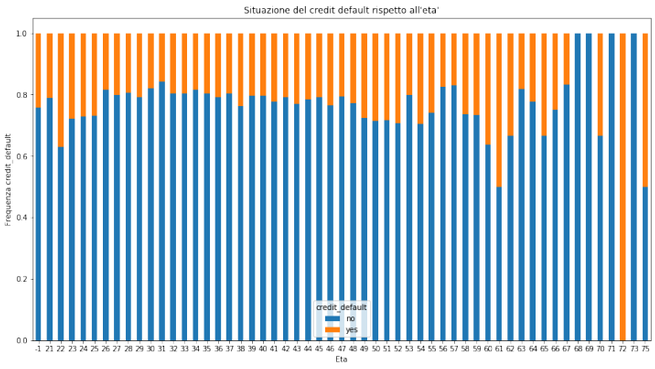
(a) BoxPlot Payment Status



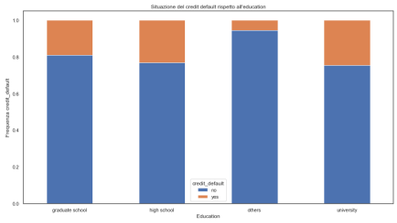
(b) Credit Default



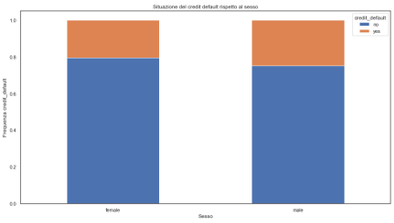
(c) Distribuzione *age* rispetto al *credit default*



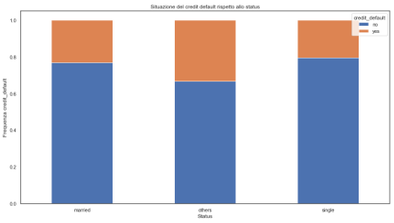
(d) Distribuzione *age* normalizzato rispetto al *credit default*



(e) Credit Default rispetto ad *education*



(f) Credit Default rispetto a *sex*



(g) Credit Default rispetto a *status*

Figura 1.4

Osservato che gli attributi *sex*, *status*, *education* e *age* non sembrano essere validi discriminanti per la classificazione del "credit default", ci siamo concentrati sugli attributi rimanenti: Payment Amount, Bill-Amount e Payment-Status. Sulla base della nostra conoscenza dei dati, abbiamo verificato per la prima volta che i valori di Bill-Amount e Payment-Amount di per sé non hanno troppa influenza sul credit default. Quindi, abbiamo rappresentato graficamente questi due attributi e non è stato trovato alcun tipo di legame significativo per la classificazione del cliente. Abbiamo quindi considerato l'attributo "Payment Status" che, in un certo senso, rappresenta il riepilogo di Bill Amount e Payment Status mese per mese. Abbiamo scoperto che, in ogni mese, la concentrazione dei valori è compresa tra [-2,0], in quanto rappresenta una riduzione del debito completa o parziale. Come possiamo vedere nella Figura 1.5, la colonna "yes", con il passare dei mesi, ha sempre un numero maggiore di colori rispetto alla

colonna no nell'intervallo [2,8], questo indica un ritardo nel pagamento del debito. Pertanto, riteniamo che lo stato di pagamento di settembre potrebbe essere un fattore discriminante per la classificazione di insolvenza del cliente.

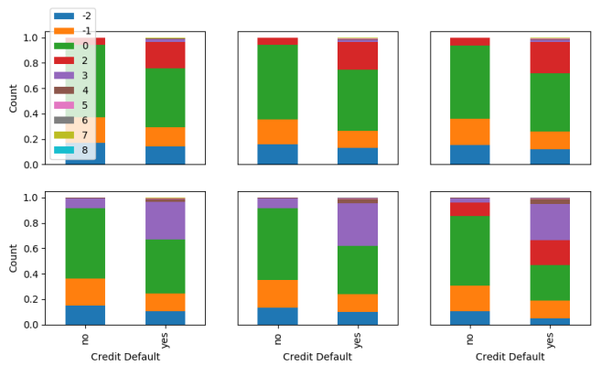


Figura 1.5: Distribuzione Payment Status normalizzato rispetto al *credit default*

1.3 Valutazione della qualità dei dati

Nella seguente sezione andremo ad analizzare la qualità dei dati in nostro possesso. La Figura 1.6 ci fornisce una prima informazione riguardante il posizionamento dei dati mancanti (null values) rispetto alle features. Quello che osserviamo è che solamente le colonne di *education* e *status* presentano dei *null values*. Dove *status* presenta una significativa frequenza di missing values, circa 1822 su un totale di 10000 righe. Infine si osserva, come indicato in basso a destra della figura Figura 1.6, che su un totale di 24 features, ciascuna entry ha al più 2 missing values. In particolare, sono 2007 il numero totale di entry alla quale manca almeno un attributo.

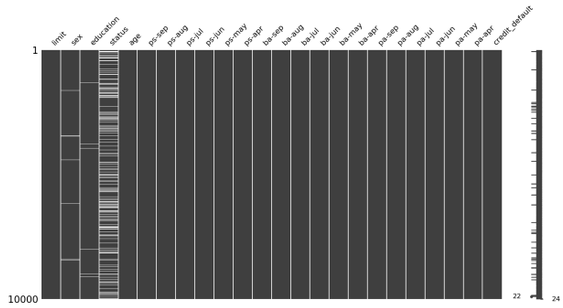


Figura 1.6: Distribuzione Missing Values

Comunque, la Figura 1.6 appena presentata non è sufficientemente esaustiva per l'obbiettivo di questa sezione in quanto non indica, invece, quanti sono per ciascuna features i valori che non rispettano i vincoli. Basandoci sulla nostra conoscenza del significato di ciascuna feature, l'unica assunzione che possiamo fare senza danneggiare le future analisi, sono relative al campo *age*, dove sicuramente il valore -1 indica l'assenza dell'informazione piuttosto che l'esistenza di persone con età negativa, Figura 1.2a.

1.4 Correlazione tra gli attributi ed eventuali variabili ridondanti

L'indice di correlazione di Pearson tra le coppie di attributi del dataset è rappresentato tramite una HeatMap nella Figura 1.7.

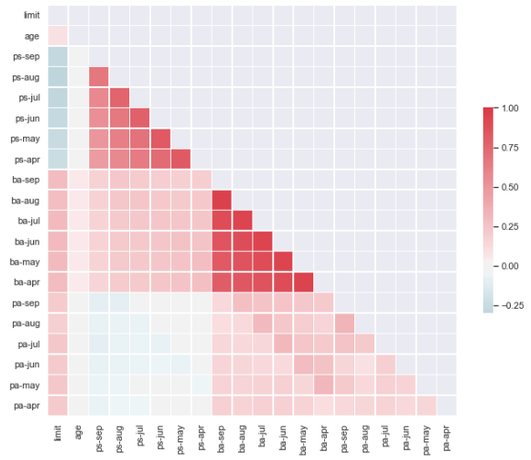


Figura 1.7: Matrice di correlazione tra gli attributi del dataset

Possiamo subito notare una correlazione superiore al 75% tra gli attributi bill amount *ba* con il bill amount del mese precedente.

Per questa ragione, essendo l'attributo ridondante abbiamo deciso di effettuare le nostre analisi considerando solamente l'attributo *ba-aug* (senza considerare i mesi precedenti) in modo da avere anche il relativo *ps-sep*. In particolare, abbiamo definito le equazioni 1.1 e 1.2.

$$Sia\, x \in [4,5,6,7,8] \in \mathbf{Z}.$$

$$Siano\, ba_{\mathrm{i}}(x)\, e\, pa_{\mathrm{i}}(x)\, due\, funzioni\, che\, restituiscono\, il\, Bill/PaymentAmount\, delmese\, x\, per\, la\, i^{\mathrm{th}}\, entry.$$

Allora:

$$ba_{\mathrm{i}}(x) - pa_{\mathrm{i}}(x+1) = debt_{\mathrm{i}}(x+1) \tag{1.1}$$

$$ba_{\mathrm{i}}(x) - debt_{\mathrm{i}}(x) = outlay_{\mathrm{i}}(x) \tag{1.2}$$

Notiamo poi che anche per gli attributi payment status ovvero *ps* abbiamo una situazione molto simile a quella dei *ba*, in questo caso però la correlazione è sopra al 50% ma non supera il 75%. Anche in questo caso abbiamo adottato la stessa tecnica utilizzata in precedenza con i ba ovvero siamo andati ad utilizzare solamente l'ultimo *ps*.

1.5 Gestione (e trasformazione) degli attributi mancanti o errati

Lo studio del dataset ci ha permesso di analizzare meglio i record trovando delle informazioni mancanti o errate. In particolare abbiamo notato che gli attributi *sex*, *education* e *status* presentano dei missing values. Rispettivamente 100, 127 e 1822 record sono quindi sprovvisti di un valore per questi attributi. Dato che i missing value relativi agli attributi *sex* ed *Education* sono pochi rispetto al numero complessivo di record (10.000) e visto che, stando ad alcuni test effettuati, lo status non incide sul *credit default*, abbiamo deciso di inserire al posto dei valori mancanti il valore "Mode" della colonna. Se avessimo semplicemente eliminato i circa 2000 record, avremmo perso delle informazioni utili nella fase di training del modello dell'albero di decisione. Un altro dettaglio interessante riguarda i valori errati presenti all'interno dei record. In particolare per quanto riguarda l'attributo *age* abbiamo notato la presenza di valori -1. Trattandosi di attributi numerici, li abbiamo sostituiti con la mediana dell'età della popolazione del nostro dataset. L'attributo *credit default* è stato poi modificato e abbiamo sostituito ai label "yes" e "no" gli interi 0 e 1. Queste trasformazioni degli attributi sono necessarie per l'utilizzo di algoritmi di clustering.

Capitolo 2

Analisi dei Cluster

2.1 Kmeans

2.1.1 Attributi utilizzati per il clustering con KMeans

Per eseguire la clusterizzazione con K-Means abbiamo preso in considerazione solamente i seguenti attributi: *limit, ba-aug, pa-sep.* Abbiamo escluso tutti gli attributi categorici e anche i *pa* e *ba* precedenti all'ultimo mese disponibile.

2.1.2 Ricerca del numero corretto di cluster per KMeans

Per utilizzare l'algoritmo KMeans abbiamo dovuto scegliere il numero di cluster K in cui suddividere il dataset . La scelta di K è importante perché con valori troppo bassi o troppo alti si ottengono risultati poco significativi. Lo studio del valore di K da utilizzare è stato svolto prendendo in considerazione l'andamento dell'SSE e della silhouette al variare di K.

Riportiamo nel grafico di Figura 2.1 l'andamento di questi due valori al variare del numero K di cluster.

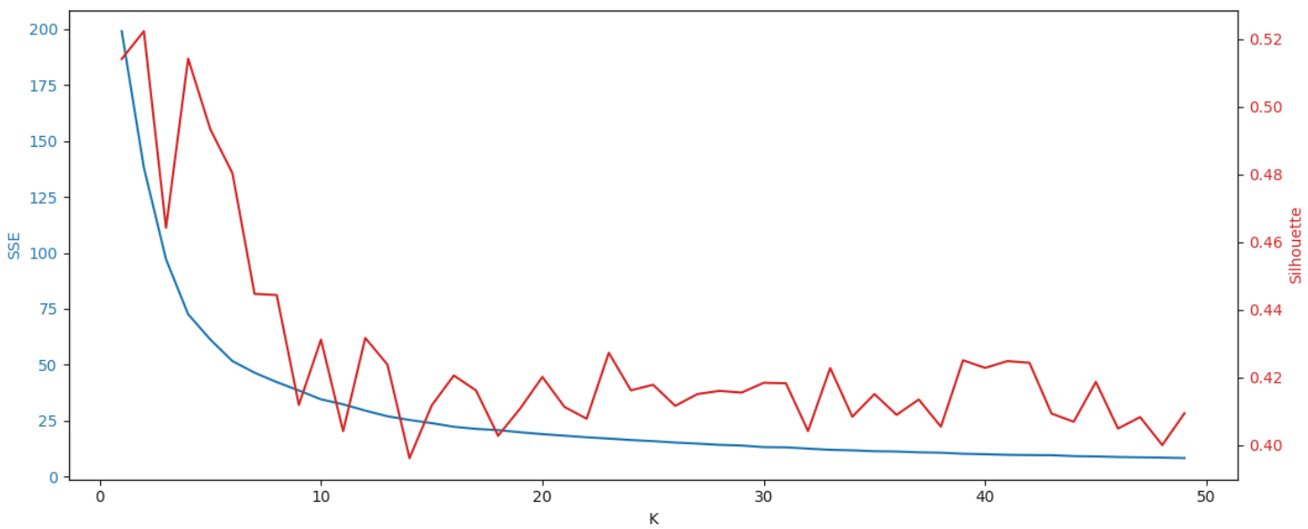


Figura 2.1: Confronto tra la variazione del valore dell'SSE e della Silhouette rispetto a K

Per la scelta del parametro abbiamo considerato il K per cui si verifica una maggiore diminuzione del valore dell'SSE e un aumento della silhouette. Volendo evitare una K troppo piccola e una troppo grande che facesse diminuire troppo l'SSE, analizzando il grafico abbiamo scelto come parametro K = 10 perchè si trova in prossimità del punto di gomito.

2.1.3 Analisi del Cluster ottenuto con KMeans

La caratterizzazione dei cluster consiste nell'analisi dei centroidi e della distribuzione dei vari cluster confrontata con quella dell'intero dataset.

Nella figura 2.2 possiamo vedere che gli attributi dei cluster 0,1,10,9 e 3,6,2,7 seguono all'incirca lo stesso andamento, questo indica che con una K più bassa questi cluster potrebbero unirsi. I cluster 3,9,0,10,6,1 hanno un valore basso per l'attributo *limit*, lo stesso attributo assume un valore più alto nei cluster rimanenti. Il cluster 4 è l'unico che ha un valore di *ba-aug* piuttosto alto. Particolare è il cluster 5 che mantiene un valore costante per tutti gli attributi presi in considerazione. Tutti i cluster hanno poi un valore molto simile per l'attributo *pa-sep*. Questo potrebbe voler significare che gli attributi *ba-aug* e *pa-sep* sono correlati.

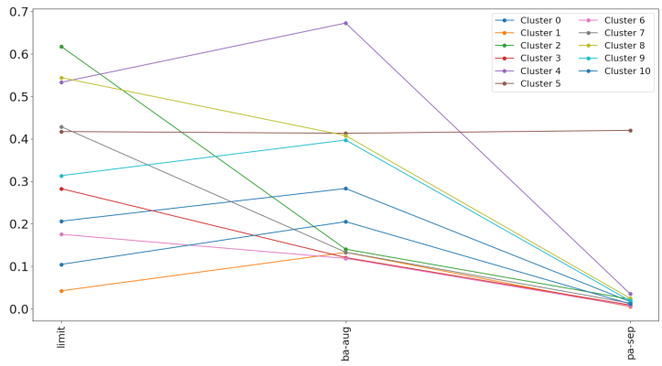
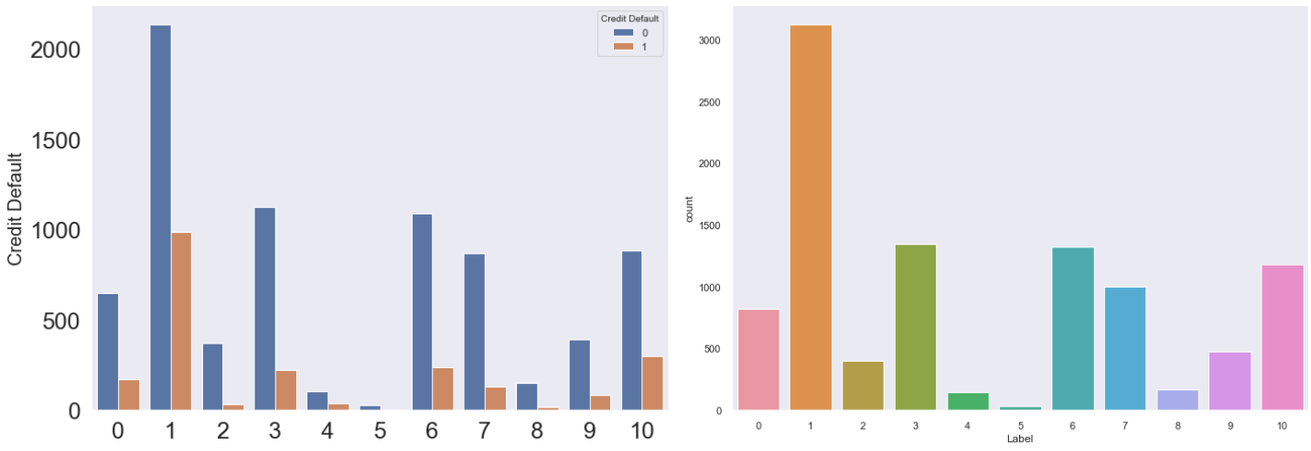
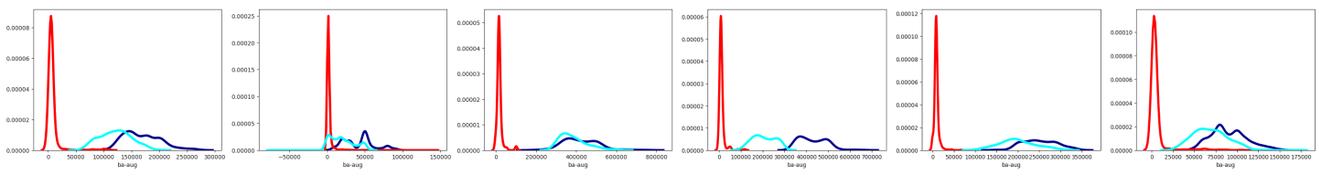


Figura 2.2: Analisi dei centroidi

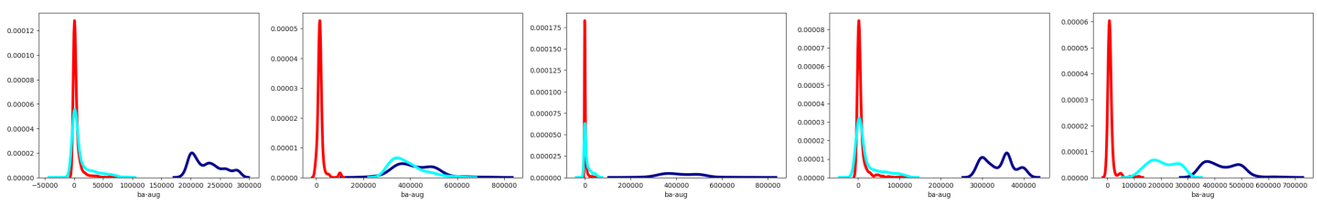
Nella figura 2.3a mostriamo la distribuzione degli attributi nell'intero dataset. In particolare la figura 2.3b mostra la distribuzione dell'attributo *credit default* nei cluster che sono stati prodotti, da questa non notiamo una distinzione netta nei vari cluster tra gli utenti in credit default e quelli non in credit default. Studiando la distribuzione delle variabili nei vari cluster abbiamo notato che l'attributo *pa-sep* segue sempre la stessa distribuzione che ha nel grafico del dataset completo. Per i cluster 1,2,5,9,10,11 le distribuzioni di *ba-aug* e di *limit* sono molto simili tra loro. Nei cluster 3,4,6,7,8 invece la variabile *ba-aug* segue una distribuzione normale.



(a) Distribuzione degli attributi nel dataset completo e dell'attributo credit Default nei vari cluster



(b) Distribuzione degli attributi nei cluster 1,2,5,9,10,11



(c) Distribuzione degli attributi nei cluster 3,4,6,7,8

Figura 2.3: Analisi delle distribuzioni degli attributi

2.2 DBScan

2.2.1 Ricerca dei parametri MinPoints ed epsilon

Per ottenere un risultato interessante dall'esecuzione di DBScan è necessario stimare il valore dei parametri MinPoints ed Epsilon. La scelta di Epsilon non è banale, selezionando un valore troppo piccolo avremmo troppi punti indicati come outlier mentre con un valore troppo alto si andrebbe a creare un unico cluster. Per stimare il valore di Epsilon abbiamo utilizzato il metodo K-Nearest neighbor. Nel grafico 2.4 mostriamo la curva ottenuta con il metodo K-Nearest neighbor, per stimare Epsilon abbiamo identificato il punto di gomito e il corrispondente valore di Epsilon.

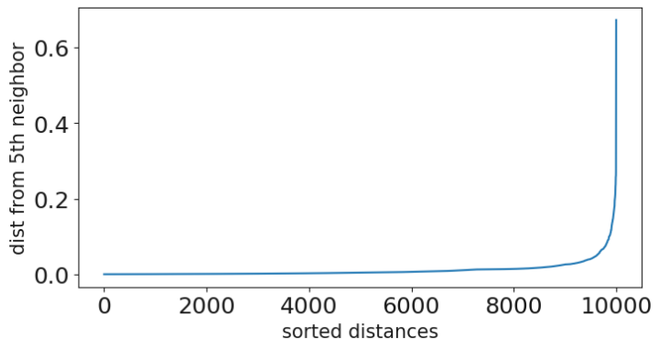


Figura 2.4: Media delle distanze tra ogni punto e i 5 nodi più vicini.

Dal grafico possiamo notare che il punto di gomito si trova nel range compreso tra $[0.05, 0.20]$, quindi abbiamo suddiviso questo range in 20 punti e per ogni possibile valore di Epsilon è stata calcolata la Silhouette corrispondente. Dovendo stimare anche il valore del parametro MinPoint, per ogni possibile valore della Epsilon abbiamo anche provato a modificare questo valore con l'intenzione di selezionare la combinazione dei due parametri che ci fornisce la maggiore Silhouette.

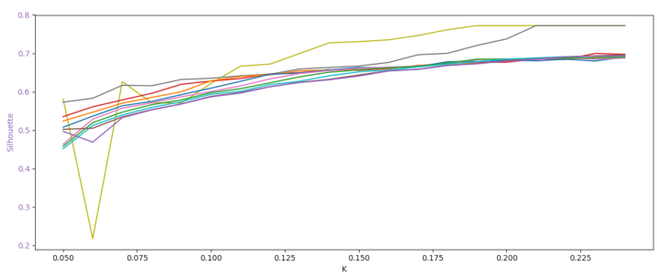


Figura 2.5: Variazione del valore della Silhouette rispetto al valore di Epsilon e di MinPoints

Abbiamo provato svariati valori di MinPoints compresi tra 5 e 250, stando al grafico 2.5 il valore migliore della Silhouette si ottiene fissando MinPoints a 240 e Epsilon a 0.225. Con questi parametri però DBScan produce un unico cluster con 9965 punti andando a separare 35 punti che rappresentano il rumore. Per ottenere un risultato più significativo, è stato creato un secondo dataset eliminando i 35 "noise point" e su questo è stato nuovamente eseguito DBScan.

2.2.2 Interpretazione dei cluster ottenuti con DBScan

Per la scelta dei parametri da utilizzare su questo dataset "modificato" abbiamo utilizzato lo stesso metodo descritto in precedenza, inoltre sono stati studiati i possibili cluster creati dall'algoritmo al variare dei parametri. Con Epsilon compresa tra 0.01 e 0.02 BScan ci restituisce una grande quantità di cluster molto piccoli con una silhouette minore di 0. Con la silhouette compresa tra 0.02 e 0.04 abbiamo dei cluster leggermente più grandi in cui però molti punti vengono identificati come rumore. Per epsilon compresa tra 0.05 e 0.15 abbiamo un unico cluster e pochi noise point. Studiando i cluster che sono stati generati dall'algoritmo abbiamo scelto come parametri $MinPoints = 200$ e $Epsilon = 0.06$. DBScan ci

ha restituito due cluster, uno con 8631 punti e uno con 144 punti, i restanti 1190 punti sono stati segnati come *noise point*.

2.3 Clustering Gerarchico

Abbiamo eseguito il Clustering gerarchico utilizzando tre metodi differenti, scegliendo quelli che meno sono suscettibili a rumore ed outlier ovvero *Complete, Average e Ward*.

2.3.1 Analisi dei dendrogrammi ottenuti

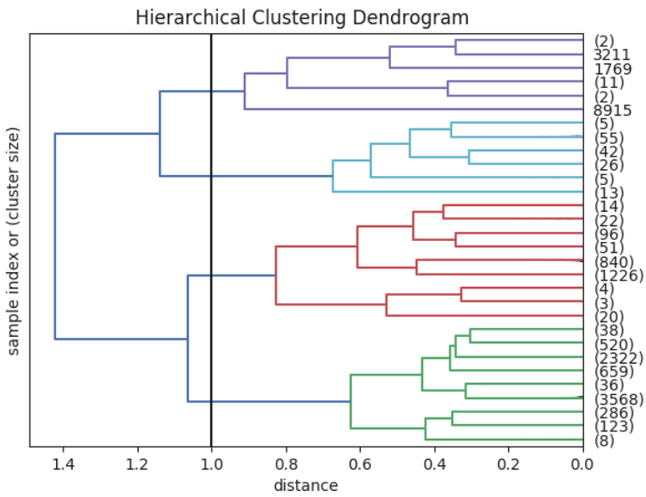


Figura 2.6: Dendrogramma per metodo Complete

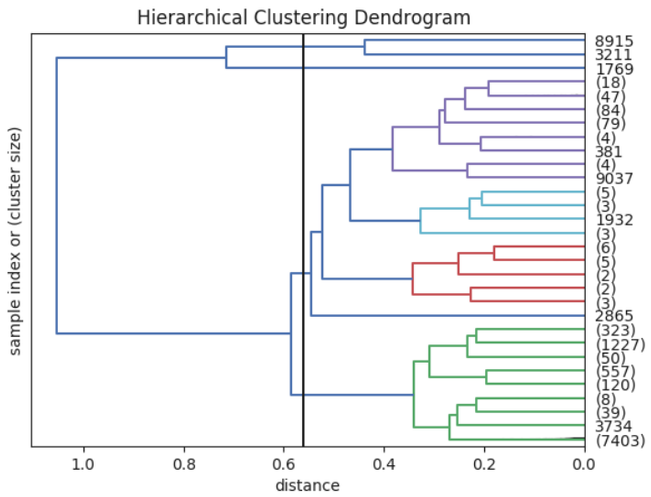


Figura 2.7: Dendrogramma per metodo Average

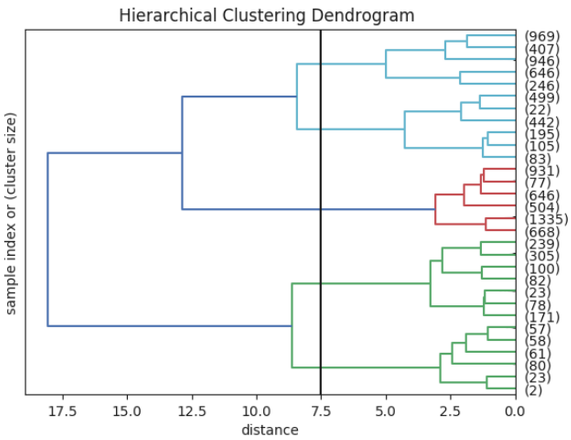


Figura 2.8: Dendrogramma per metodo Ward

Complete linkage: La figura 2.6 mostra il dendrogramma ottenuto con il metodo Complete Linkage. A distanza minore di 0.2 otteniamo 30 cluster. Il dendrogramma suggerisce la presenza di 4 cluster ottenuti con una linea di taglio a distanza 1.0.

Average Linkage: La figura 2.7 mostra il dendrogramma ottenuto con il metodo Average Linkage. Nella parte alta della figura abbiamo un singleton che si unisce ad un agglomerato di due cluster, dato che l'unione avviene ad altezza 0.7 circa, abbiamo pensato che questo possa essere un outlier. Nella parte bassa della figura si creano due grandi agglomerati di cluster. Dato che l'unione avviene ad altezza 0.6 abbiamo deciso di porre la linea di taglio ad altezza 0.56 in modo da ottenere i due cluster creati dai due agglomerati e poi i due cluster che potrebbero essere "outlier".

Ward Linkage: La figura 2.8 mostra il dendrogramma ottenuto con il metodo Average Linkage e in questo caso le altezze assumono valori più elevati rispetto ai metodi precedentemente analizzati. Abbiamo nella parte alta della figura due agglomerati di cluster (rosso e celeste) che si uniscono ad una

altezza pari a 12 circa. Dato che i cluster ottenuti in questo caso sembrano tutti abbastanza equilibrati abbiamo deciso di fare un taglio all'altezza 7.5 per ottenere 5 cluster.

Per il metodo gerarchico abbiamo calcolato la silhouette per tutti e tre le possibili metriche (Complete, Average, Ward) e per un numero di cluster compreso tra 2 e 10. Alla fine dell'analisi il risultato migliore è stato ottenuto con la metrica Average e 3 cluster.

2.4 Valutazione del migliore algoritmo per il Clustering

Una volta eseguito il clustering con KMeans, DBScan e con il metodo gerarchico, abbiamo cercato di confrontarli per capire quale fosse la migliore clusterizzazione ottenuta.

Algoritmo	Silhouette
K-Means	0.430
DBScan	0.432
Gerarchico	0.732

Tabella 2.1: Descrizione degli attributi del dataset

Nella tabella 2.1 sono riportati i valori della Silhouette ottenuti con i tre algoritmi appena citati, ricordiamo che i parametri utilizzati sono stati i seguenti:

- Per Kmeans abbiamo usato K = 11
- In DBScan abbiamo usato MinPoints = 200 e Epsilon = 0.06

Nel caso dell'algoritmo Gerachico abbiamo considerato per il confronto la silhouette calcolata con il metodo Average e con 3 cluster. L'algoritmo che ha dato risultati migliori è il clustering gerarchico usando il metodo average linkage che ha evidenziato la presenza di 3 cluster rispettivamente formati da 2249, 133 e 7170 elementi. La dimensione di questi cluster è simile a quella dei cluster ottenuti con DBScan mentre con KMeans abbiamo un comportamento del tutto differente visto che vengono creati 11 cluster.

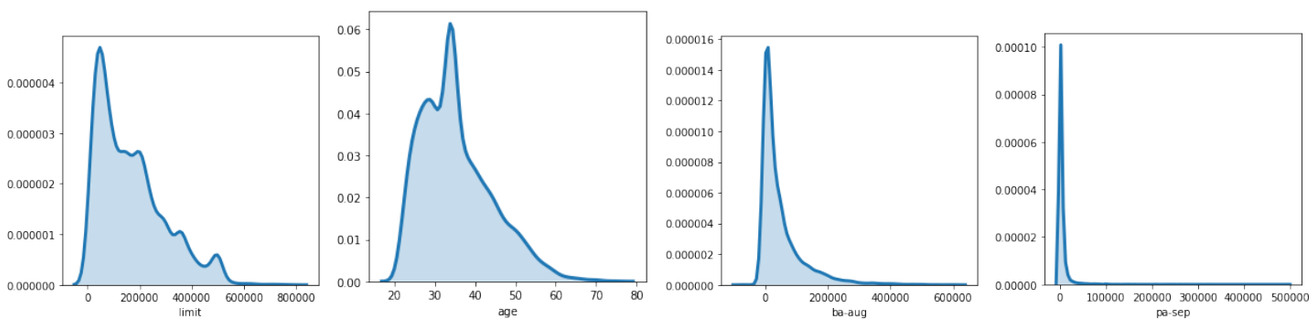
Capitolo 3

Pattern e Association Rules mining

In questa sezione descriviamo il processo di analisi delle association rules. Nella prima sezione si discutono le operazioni preliminarie effettuate sul dataset per preparare i dati alle operazioni successive. Nelle sezioni successive si estraggono gli itemset frequenti e poi, da questi itemset, le association rules.

3.1 Operazioni preliminari

Una operazione necessaria per l'estrazione degli itemset frequenti è la discretizzazione degli attributi numerici ovvero *limit*, *age*, *ba-aug* e *pa-sep*. Per discretizzare questi attributi abbiamo preso in considerazione i grafici delle loro distribuzioni e di conseguenza è stato scelto il numero di intervalli da creare. Per l'attributo *limit* abbiamo creato 8 intervalli, questa suddivisione è ispirata dal fatto che la distribuzione non è regolare ed è presente un pinnacolo tra 0 e 150.000. Per l'attributo *age* è stata effettuata una suddivisione in 5 intervalli, anche in questo caso, studiando la distribuzione, abbiamo notato che è presente un pinnacolo tra 30 e 40 e questo ci ha portato a scegliere 5 intervalli. Anche per *ba-aug* vale lo stesso discorso dei precedenti attributi, in questo caso è presente un pinnacolo tra 0 e 100.000, quindi abbiamo deciso di creare 10 intervalli. L'attributo *pa-sep* presenta un pinnacolo compreso tra 0 e 20.000, in questo caso il numero di intervalli che abbiamo deciso di creare è pari a 10.



(a) Distribuzione Limit (b) Distribuzione Age (c) Distribuzione Ba-Aug (d) Distribuzione Pa-Sep

Figura 3.1: Analisi delle distribuzioni degli attributi

Prima di procedere con l'estrazione degli itemset frequenti abbiamo anche Modificato i valori dell'attributo *pa-sep* assegnando una stringa per ogni possibile valore numerico.

3.2 Estrazione degli itemset frequenti

Descriviamo ora il procedimento di estrazione delle diverse tipologie di itemset, in particolare massimali, chiudi e frequenti.

3.2.1 ItemSet Massimali

L'estrazione degli itemset massimali è stata svolta utilizzando l'algoritmo APriori utilizzando il parametro $MinSup = 20\%$ e selezionando solamente i set contenenti almeno 4 elementi. Nella tabella 3.1 riportiamo gli itemset frequenti con un support compreso tra il 20% e il 35%. Possiamo notare dagli itemset 3 e 4 che l'utilizzo del Revolving Credit nel mese di Settembre non va a comportare un Credit Default. Il valore "No" del credit default lo troviamo in più di una occasione in un set con l'attributo "married", questo mi indica che spesso il cliente sposato non va in credit default.

#	ItemSet	Supporto(%)
1	$Female, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	35.87%
2	$university, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	26.00%
3	$RevolvingCredit, Female, No, [0.0, 49335.8)Pa - Sep, married$	25.80%
4	$RevolvingCredit, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	25.48%
5	$[10000.0, 106250.0)Limit, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	25.30%
6	$[31.8, 42.6)Age, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	24.52%
7	$[21.0, 31.8)Age, Female, [0.0, 49335.8)Pa - Sep, married$	23.98%
8	$university, Female, No, [0.0, 49335.8)Pa - Sep, married$	22.65%
9	$[10000.0, 106250.0)Limit, RevolvingCredit, [0.0, 49335.8)Pa - Sep, married$	22.28%
10	$[21.0, 31.8)Age, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	21.88%

Tabella 3.1: Tabella con gli itemset Massimali più frequenti

3.2.2 ItemSet chiusi

Nel caso degli ItemSet chiusi abbiamo utilizzato come parametro un $MinSupp$ pari al 30% andando a selezionare i soli set con almeno 4 elementi. Nella tabella 3.2 riportiamo la lista degli ItemSet chiusi più frequenti. Anche in questo caso troviamo il set 4 in cui sono presenti i valori "Married", "Revolving Credit" e anche il valore "No" relativo all'attributo credit default.

#	ItemSet	Supporto(%)
1	$[-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	57.11%
2	$[-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	47.62%
3	$Female, [-2205.0, 65367.0)Ba - Aug, [0.0, 49335.8)Pa - Sep, married$	45.31%
4	$RevolvingCredit, No, [0.0, 49335.8)Pa - Sep, married$	41.77%
5	$[10000.0, 106250.0)Limit, [-2205.0, 65367.0)Ba - Aug, [0.0, 49335.8)Pa - Sep, married$	36.39%
6	$Female, [-2205.0, 65367.0)Ba - Aug, No, married$	36.13%
7	$Female, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep$	36.08%
8	$Female, [-2205.0, 65367.0)Ba - Aug, No, [0.0, 49335.8)Pa - Sep, married$	35.87%
9	$university, No, [0.0, 49335.8)Pa - Sep, married$	35.69%
10	$university, [-2205.0, 65367.0)Ba - Aug, [0.0, 49335.8)Pa - Sep, married$	34.75%

Tabella 3.2: Tabella con gli itemset Chiusi più frequenti

3.2.3 ItemSet Frequenti

Per gli ItemSet Frequenti abbiamo utilizzato come parametro un $MinSupp$ pari al 30% andando a selezionare i soli set con 4 elementi. Nella tabella 3.3 riportiamo la lista degli ItemSet più frequenti. La

lista degli item set coincide in gran parte con la lista degli ItemSet "Closed", questo perchè gli ItemSet frequenti sono un sovra insieme dei "closed" elencati in precedenza.

#	ItemSet	Supporto(%)
1	$[-2205.0, 65367.0)Ba-Aug, No, [0.0, 49335.8)Pa-Sep, married$	57.11%
2	$Female, No, [0.0, 49335.8)Pa-Sep, married$	47.62%
3	$Female, [-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep, married$	45.31%
4	$RevolvingCredit, No, [0.0, 49335.8)Pa-Sep, married$	41.77%
5	$[10000.0, 106250.0)Limit, [-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep, married$	36.39%
6	$Female, [-2205.0, 65367.0)Ba-Aug, No, married$	36.08%
7	$Female, [-2205.0, 65367.0)Ba-Aug, No, [0.0, 49335.8)Pa-Sep$	35.87%
8	$Female, [-2205.0, 65367.0)Ba-Aug, No, [0.0, 49335.8)Pa-Sep, married)$	35.87%
9	$university, No, [0.0, 49335.8)Pa-Sep, married$	35.69%
10	$university, [-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep, married$	34.75%

Tabella 3.3: Tabella con gli itemset Frequenti

3.3 Estrazione delle association rules

Nelle tabelle 3.4, **??** e 3.6 riportiamo la lista delle association rules. Abbiamo deciso di estrarre queste regole considerando solamente gli itemset frequenti che avessero una lunghezza maggiore o uguale a 3.

La tabella 3.4 mostra le association rules che hanno un valore della confidence compreso tra 80% e 100%. In nessuna di queste regole abbiamo come post condizione il valore "No" o "Yes" relativo al Credit Default. Le post-condizioni delle 6 regole sono equamente suddivise tra "Married" e "[0.0, 49335.8)Pa-Sep". La regola 1 indica che una persona che non è in credit Default e che nel mese di Settembre paga una cifra compresa tra [0.0, 49335.8) è anche sposata. Questa regola è simile alla 6 in cui però possiamo notare che una persona sposata e *non* in credit default tende a pagare a settembre una cifra compre tra [0.0, 49335.8). Per le regole 3 e 5 vale lo stesso discorso fatto con la 1 e la 6, in questo caso però al posto del valore "No" relativo al *credit default* troviamo "Female". La regola 4 invece indica che una spesa nel mese di agosto compresa tra [-2205.0, 65367.0) combinata con un pagamento a settembre compreso tra [0.0, 49335.8) comporta la presenza di un cliente sposato.

#	Precondizione	PostCondizione	Confidence(%)	Lift
1	(No, [0.0, 49335.8)Pa-Sep)	married	99.35%	1.00109544262
2	([-2205.0, 65367.0)Ba-Aug, married)	[0.0, 49335.8)Pa-Sep	99.35%	1.00981594846
3	(Female, [0.0, 49335.8)Pa-Sep)	married	99.30%	1.00054119424
4	([-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep)	married	99.22%	0.999724174439
5	(Female, married)	[0.0, 49335.8)Pa-Sep	98.37%	0.999838573642
6	(No, married)	[0.0, 49335.8)Pa-Sep	98.11%	0.997186782673

Tabella 3.4: Association rules con confidence compresa tra 70% e 78%

La tabella **??** mostra le association rules che hanno un valore della confidence compreso tra 70% e 80%. In questo caso abbiamo ben 6 regole in cui la post condizione è relativa all'attributo "Credit Default, in tutti e sei i casi si tratta del valore "No". La prima regola riprende quelle indicate in precedenza,

la presenza di un cliente donna e sposata determina un'assenza di credit default. Alla stessa post condizione si arriva anche quando il cliente della banca è donna e paga a settembre una quantità di denaro compresa tra [0.0, 49335.8) oppure quando il cliente è spostato, paga a settembre tra [0.0, 49335.8) dollari e spende ad agosto [-2205.0, 65367.0) dollari. Le successive 6 regole hanno come post condizione il Bill Amount del mese di agosto compreso tra [-2205.0, 65367.0) dollari. Le regole sono simili alle precedenti e compare spesso nella precondizione il cliente "Female" e "Married".

#	Precondizione	PostCondizione	Confidence(%)	Lift
1	(Female, married)	No	79.70%	1.0233
2	(Female, [0.0, 49335.8)Pa-Sep)	No	79.44%	1.0200
3	([0.0, 49335.8)Pa-Sep, married)	No	77.74%	0.9982
4	([-2205.0, 65367.0)Ba-Aug, married)	No	77.15%	0.9907
5	([-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep, married)	No	77.14%	0.9905
6	([-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep)	No	77.05%	0.9893
7	([0.0, 49335.8)Pa-Sep, married)	[-2205.0, 65367.0)Ba-Aug	75.81%	1.0094
8	(Female, [0.0, 49335.8)Pa-Sep)	[-2205.0, 65367.0)Ba-Aug	75.62%	1.0070
9	(No, [0.0, 49335.8)Pa-Sep)	[-2205.0, 65367.0)Ba-Aug	75.23%	1.0018
10	(No, [0.0, 49335.8)Pa-Sep, married)	[-2205.0, 65367.0)Ba-Aug	75.22%	1.0016
11	(Female, married)	[-2205.0, 65367.0)Ba-Aug	74.95%	0.9980
12	(No, married)	[-2205.0, 65367.0)Ba-Aug	74.29%	0.9892

Tabella 3.5: Association rules con confidence compresa tra 70% e 78%

La tabella 3.6 mostra le association rules che hanno un valore della confidence compreso tra 60% e 70%. In questo caso tutte le post condizioni sono relative all'attributo "Sex" con valore "Female" e le precondizioni sono più o meno le stesse viste nelle tabelle precedenti. Anche in questo caso si evidenzia il fatto che il cliente non in credit default e sposato è spesso una donna.

#	Precondizione	PostCondizione	Confidence(%)	Lift
1	(No, [0.0, 49335.8)Pa-Sep, married)	Female	62.72%	1.02289497579
2	(No, married)	Female	62.71%	1.02277338882
3	(No, [0.0, 49335.8)Pa-Sep)	Female	62.71%	1.02273816716
4	([0.0, 49335.8)Pa-Sep, married)	Female	61.34%	1.00035104138
5	([-2205.0, 65367.0)Ba-Aug, married)	Female	61.25%	0.998915900474
6	([-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep, married)	Female	61.20%	0.998123237534
7	([-2205.0, 65367.0)Ba-Aug, [0.0, 49335.8)Pa-Sep)	Female	61.14%	0.997139900258

Tabella 3.6: Association rules con confidence compresa tra 60% e 70%

Capitolo 4

Classificazione

Nel seguente capitolo andremo a presentare il processo di classificazione. In tale processo vogliamo identificare se un determinato utente della banca risulta in credit default, oppure no.
Grazie alle analisi svolte nei capitoli precedenti, abbiamo potuto fare delle prime assunzioni su quali fossero gli attributi significativi per una corretta classificazione; abbiamo quindi mantenuto tutti i *Payment Status, ba-aug* e *pa-sep*.
Fin da subito ci siamo dovuti scontrare con un set di dati piuttosto sbilanciato rispetto all'attributo oggetto della classificazione; in particolare su circa 10000 istanze, solo $\frac{1}{4}$ hanno il valore "yes" sotto all'attributo credit default.
Per cercare di arginare tale problematica, abbiamo utilizzato principalmente due tecniche:

- *StratifiedKFold*: garantisce che ciascun fold sia un buon rappresentativo dell'intero dataset.
- *Oversampling con SMOTE*: arricchisce il dataset con dati fittizi appartenenti alla minority class.

Quest'ultima tecnica è stata applicata sul training set in modo che, l'introduzione dei dati artificiali creati, non andasse ad introdurre un qualche bias rendendo poco significativi gli score del test set.
Le ricerche degli iperparametri per i seguenti classificatori sono state eseguite attraverso una Grid Search. I modelli che andremo a presentare sono stati scelti prestando attenzione a non cadere in overfitting, come possibile osservare dalle Figure: 4.3, 4.4 e 4.8.

4.1 Classificazione tramite Decision Trees

Nella seguente sezione andremo a discutere i modelli di Decision Trees individuati più significativi in funzione degli attributi e della diversa configurazione degli iperparametri, quali: *Criterion, Max depth, Min samples split, Min samples leaf.*

4.1.1 Modello 1

In questo modello abbiamo scelto di considerare come attributi, esclusivamente i "Payment-Status". Come si può osservare dal grafico sulla rilevanza delle feature 4.1, *ps-sep* risulta ricoprire un importante ruolo nella classificazione. Area definita dalla curva ROC: 0.76

criterion	gini
max depth	5
min samples leaf	2
min samples split	102

Tabella 4.1: Iperparametri Decision Tree 1

Set	Accuracy	Recall	Precision	F1-Score
Train	0.715	0.715	0.734	0.710
Test	0.791	0.791	0.799	0.794

Tabella 4.2: Score Decision Tree 1

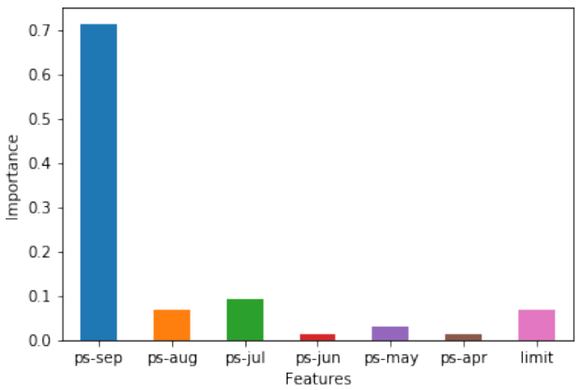


Figura 4.1: Importanza delle Feature Decision Tree 1

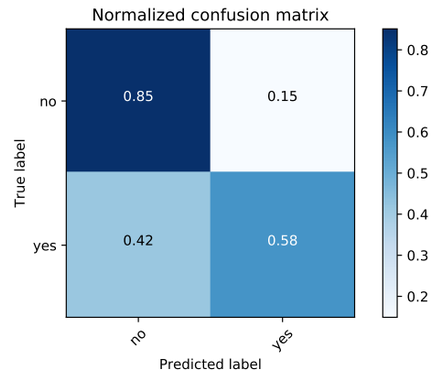


Figura 4.2: ConfusionMatrix Modello 1

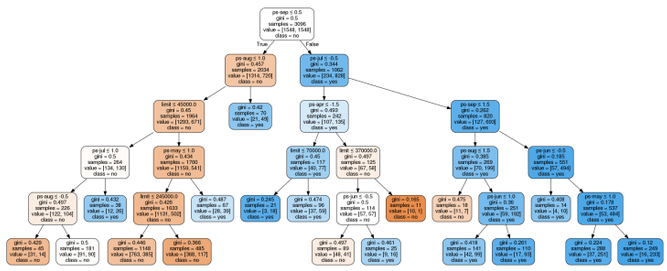


Figura 4.3: Decision Tree 1

4.1.2 Modello 2

Avendo osservato nel modello precedente l'importante rilevanza dell'attributo *ps-sep*, abbiamo provato a creare l'albero di decisione basandoci principalmente su questo attributo. Tale esperimento, come possibile osservare dalla Confusion Matrix in Figura 4.5, ha rilevato che l'attributo *ps-sep*, di per sé, aiuta nella corretta classificazione delle istanze con valore "No" sul credit default, mentre non aiuta nel riconoscimento degli utenti con credit default a "Yes".
Questo modello nonostante i buoni score sul Test Set ottenuti (Tabella 4.4), risulta essere affetto da un "bias" sui dati, dovuto ad una maggiore percentuale di istanze con *credit default* uguale a "no". Area definita dalla curva ROC: 0.69

criterion	gini
max depth	2
min samples leaf	52
min samples split	2

Tabella 4.3: Iperparametri Decision Tree 2

Set	Accuracy	Recall	Precision	F1-Score
Train	0.777	0.777	0.777	0.742
Test	0.838	0.838	0.817	0.813

Tabella 4.4: Score Decision Tree 2

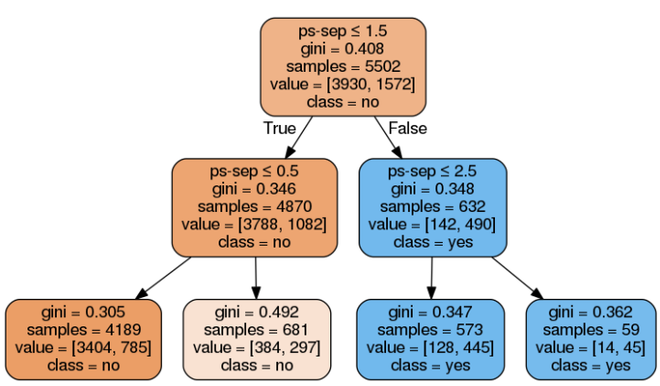


Figura 4.4: Decision Tree 2

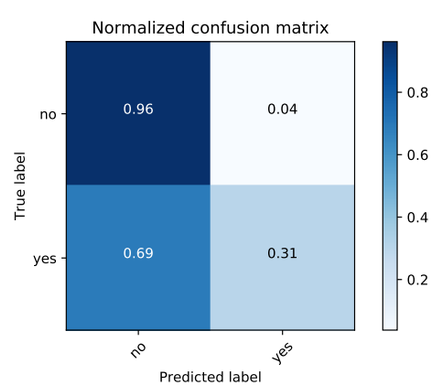


Figura 4.5: ConfusionMatrix Decision Tree 2

4.1.3 Modello 3

Riflettendo sulla semantica dei "Payment Status", ovvero che i valori maggiori di 1 indicano un ritardo di pagamento del debito, abbiamo ritenuto ragionevole contare per ciascuna entry, il numero di *ps* aventi un valore superiore ad 1 ed aggiungere il valore ottenuto in una nuova colonna denominata "g3". Tale operazione aveva l'aspettativa di controbilanciare l'effetto dell'attributo *ps-sep*, utile, invece, per predirre con una certa sicurezza che un utente non sia in default. Oltre a *ps-sep* e *g3* abbiamo incluso anche l'attributo *ps-apr* che in altri modelli si è rivelato rilevante nella classificazione.
Dalla Figura 4.6 osserviamo che "g3" si è rivelata un attributo significativo per la classificazione. Il modello ottenuto risulta essere meno soggetto al "bias" discusso in 4.1.2, in quanto riesce a riconoscere maggiormente i *True Positive* (Figura 4.11). Area definita dalla curva ROC: 0.76

criterion	gini
max depth	6
min samples leaf	252
min samples split	2

Tabella 4.5: Iperparametri Decision Tree 3

Set	Accuracy	Recall	Precision	F1-Score
Train	0.71	0.71	0.72	0.71
Test	0.774	0.774	0.800	0.784

Tabella 4.6: Score Decision Tree 3

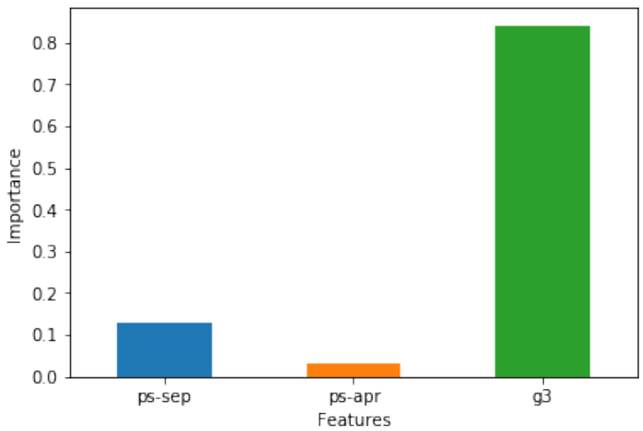


Figura 4.6: Importanza delle Features Decision Tree 3

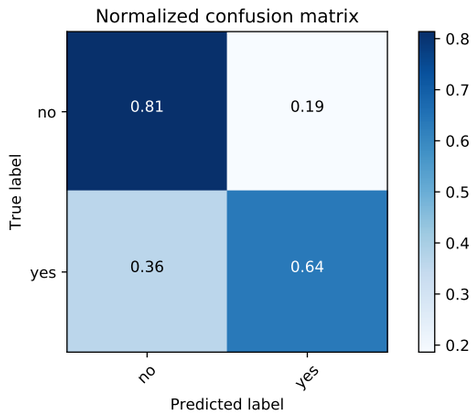


Figura 4.7: ConfusionMatrix Decision Tree 3

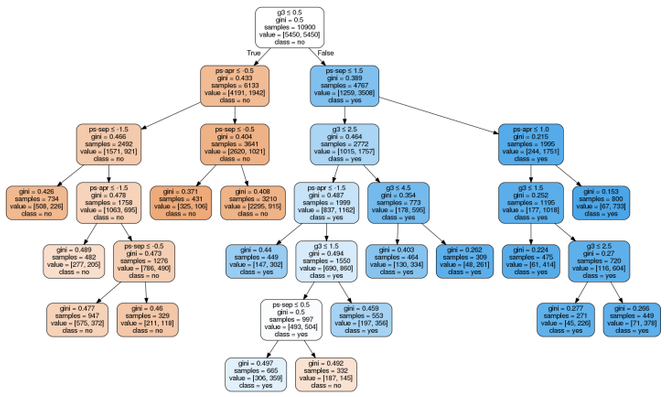


Figura 4.8: Decision Tree 3

4.2 Classificazione tramite Random Forest

In questa sezione andremo a discutere dei risultati ottenuti sul dataset con un metodo di ensemble learning, il Random Forest.
Abbiamo confrontato i risultati ottenuti dai Decision Tree della sezione 4.1 con quelli del Random Forest utilizzando lo stesso sottoinsieme di attributi. Nel caso del Random Forest, gli iperparametri considerati sono: N° Estimators, Criterion, Max depth, Min samples split, Min saples leaf.

Una caratteristica interessante scaturita dall'utilizzo di un esemble method, è la valutazione più bilanciata dell'importanza delle Feature come possibile osservare nelle Figure 4.10 e 4.12. Il confronto con il modello della sottosezione 4.1.2 è stato rimosso poiché utilizzando un singolo attributo *ps-sep*, le performance dei due classificatori sono risultate le medesime.

4.2.1 Modello 1

Per il training dei due classificatori sono stati utilizzati solamente gli attributi *ps-sep, ps-aug, ps-jul, ps-jun, ps-may, ps-apr* e *"limit"*. La tabella 4.7 mostra come il Random Forest riesca a generalizzare maggiormente, guadagnando circa il 6% su Accuracy, Recall, Precision ed F1-Score.

Score	Albero di Decisione 1	Random Forest 1
Accuracy	0.715	0.784
Recall	0.715	0.784
Precision	0.733	0.795
F1-Score	0.709	0.788
ROC	0.75	0.77

Tabella 4.7: Score Decision Tree 1 Vs Random Forest 1

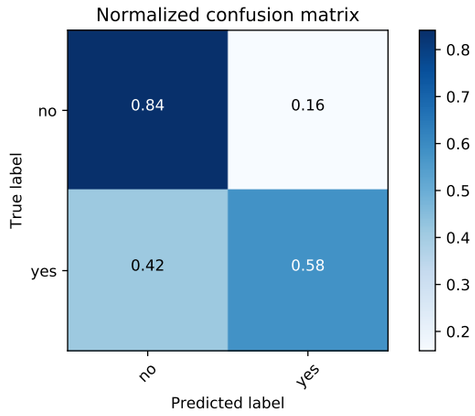


Figura 4.9: ConfusionMatrix Random Forest 1

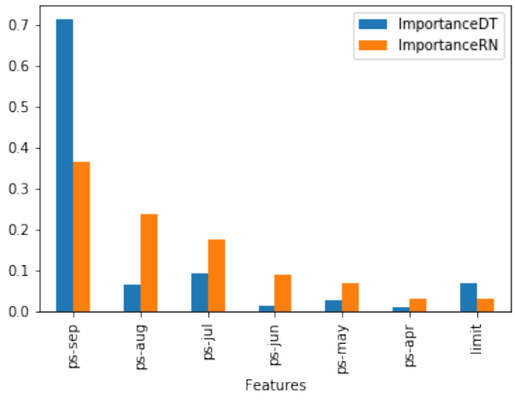


Figura 4.10: Importanza Feature Decision Tree 1 Vs Random Forest 1

4.2.2 Modello 3

Per il training dei due classificatori abbiamo considerato *ps-sep, ps-apr* e *g3*. Anche in questo caso, come nella sottosezione 4.2.1, osserviamo un miglioramento del 6% su tutti gli score (Figura 4.8).

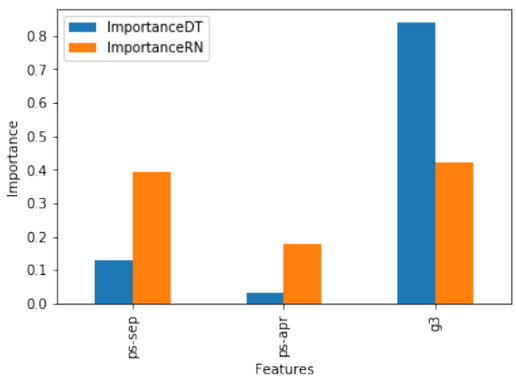


Figura 4.12: Importanza Feature Decision Tree 3 Vs Random Forest 3

Score	Albero di Decisione 3	Random Forest 3
Accuracy	0.710	0.756
Recall	0.710	0.756
Precision	0.717	0.795
F1-Score	0.707	0.769
ROC	0.76	0.76

Tabella 4.8: Score Decision Tree 1 Vs Random Forest 3

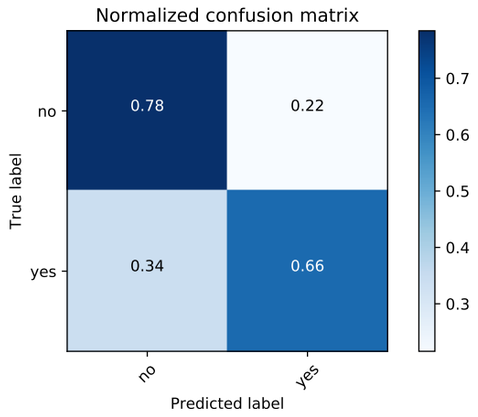


Figura 4.11: ConfusionMatrix Random Forest 3

4.3 Classificazione tramite Multi Layer Perceptron

L'ultimo classificatore considerato è il Multi Layer Perceptron. Gli iperparametri presi in esame sono: Learning Rate Init, Activation, Batch Size, Learning Rate, Hidden Layer Sizes. Come possiamo osservare dai risultati ottenuti dai 3 modelli, nonostante il Modello 1 (sottosezione 4.3.1) ed il Modello 2 (sottosezione 4.3.2) abbiano ottenuto maggiori score rispetto al Modello 3 (sottosezione 4.3.3), si osserva dalle Confusion Matrix 4.13 e 4.14 che entrambi presentano un "bias" che li spinge a classificare un percentuale elevata di istanze con credit default uguale "no" anche quando dovrebbe essere classificata come credit default a "yes".

4.3.1 Modello 1

Attributi considerati: *ps-sep, ps-aug, ps-jul, ps-jun, ps-may, ps-apr* e *limit.*
Area definita dalla curva ROC: 0.67.

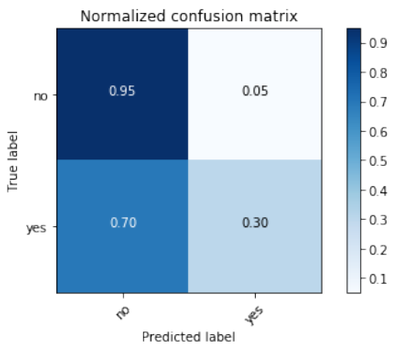


Figura 4.13: Confusion Matrix MLP Modello 1

Set	Accuracy	Recall	Precision	F1-Score
Train	0.70	0.70	0.72	0.65
Test	0.81	0.81	0.79	0.78

Tabella 4.9: Score MLP 1

4.3.2 Modello 2

Attributo considerato: *ps-sep*. Area definita dalla curva ROC: 0.71.

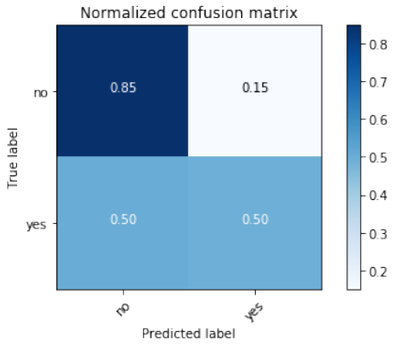


Figura 4.14: Confusion Matrix MLP Modello 2

Set	Accuracy	Recall	Precision	F1-Score
Train	0.73	0.73	0.73	0.72
Test	0.77	0.77	0.78	0.78

Tabella 4.10: Score MLP 2

4.3.3 Modello 3

Attributi considerati: *ps-sep, ps-apr* e *g3*. Area definita dalla curva ROC: 0.75.

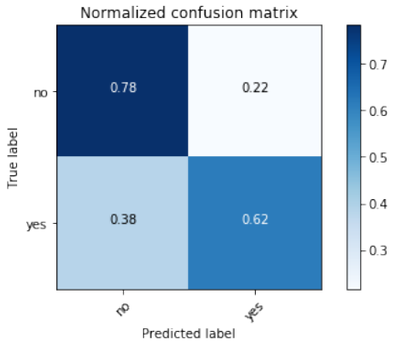


Figura 4.15: Confusion Matrix MLP Modello 3

Set	Accuracy	Recall	Precision	F1-Score
Train	0.71	0.71	0.71	0.71
Test	0.75	0.75	0.79	0.76

Tabella 4.11: Score MLP 3

4.4 Miglior Modello

La scelta del miglior modello deve essere fatta cercando di calarci nel contesto del dataset, ovvero quello di una banca che deve individuare quando un proprio cliente è in credit default oppure no. Ciò che desideriamo è, in generale, avere un modello accurato che permetta di predirre il più correttamente possibile la situazione di un certo utente, evitando di etichettare ingiustamente un cliente come in "credit default", quindi cercando di massimizzare la Precision di un modello piuttosto che la Recall.
Abbiamo individuato nove modelli, di cui sei, nonostante gli alti score, sembra che prediligano classificare un utente come *non* in default. Allo stesso tempo questi modelli hanno un'alta percentuale di False Positive, ovvero classificano un utente come in "credit default" quando in realtà non lo è. Tale aspetto è ben visibile osservando le confusion matrix mostrate precedentemente.
Sulla base di questo ragionamento, i modelli che riteniamo più corretti al contesto sono quelli descritti nelle sottosezioni: 4.1.3, 4.2.2 e 4.3.3 che risultano più bilanciati nella classificazione. In particolare il modello della sottosezione 4.1.3 risulta il migliore.