Capitolo 1

Data Understanding

Credit Card Default è il dataset utilizzato per questa analisi che si pone l'obiettivo di classificare ed identificare i clienti di una banca di Taiwan in base ai loro default payments o pagamenti in stato di insolvenza, relativi alla carta di credito personale. Dal punto di vista della gestione del rischio, il risultato della classificazione sarà prezioso per identificare clienti credibili o non credibili. Il periodo a cui si riferiscono i dati in possesso va da Aprile a Settembre 2005 e la valuta di riferimento è il dollaro taiwanese (NT$). Il datasetet contiene 10.000 records descritti da 24 colonne di attributi.

1.1 Analisi degli attributi

Di seguito (Tabella 1.1) vengono riportati e descritti gli attributi raggruppandoli in base alla loro tipologia e indicando per ciascuno di essi il loro dominio.

Tipologia		Attributo	Descrizione	Dominio
Numerici	Discreti	age	Indica l'età anagrafica del titolare della carta di credito.	[21, 75] ∩ $\mathbb{N}$
		limit	Indica l'importo limite in NT$ della carta di credito di ciascun titolare.	[10 000, 780 000] ∩ $\mathbb{N}$
		ba-X, X ∈ {apr, may, jun, jul, aug, sep}	Bill Amount: indica l'importo in NT$ da pagare relativo alle spese effettuate nel mese X. (Valori negativi indicano di avere un credito positivo sulla carta al momento del conto)	[-209 051, 616 836] ∩ $\mathbb{Z}$.
		ps-X, X ∈ {apr, may, jun, jul, aug, sep}	Payment Status: indica lo stato di pagamento relativo alle spese del mese X. Legenda dei valori: -2: nessun utilizzo -1: pagamento per intero 0: utilizzo del credito rotativo, rateizzando l'importo [1,9]: indica di quanti mesi è in ritardo un pagamento	[-2, 8] ∩ $\mathbb{Z}$.
		pa-X, X ∈ {apr, may, jun, jul, aug, sep}	Payment Amount: indica l'importo effettivo pagato nel mese X+1 relativo a ba-X.	[0, 1 227 082] ∩ $\mathbb{N}$
Categorici	Nominali	Status	Indica lo stato civile del titolare della carta di credito.	{single, married, others}
		Sex	Indica il sesso del titolare della carta di credito	{male, female}
	Ordinali	Education	Indica l'istruzione del titolare della carta di credito	{ others, graduate school, high school, university }
	Binari	Credit default	Indica se il titolare di carta di credito è in situazione di insolvenza del credito fornito dalla banca.	{yes, no}

Tabella 1.1: Attributi del dataset

1.2 Analisi della qualità dei dati

È stata analizzata la qualità dei dati all'interno del dataset fornito secondo i seguenti parametri:

Accuratezza sintattica
È stata verificata la correttezza delle stringhe nei valori di attributi categorici `sex`, `education`, `status` e `credit_default`. Sono stati analizzati i valori degli attributi numerici `limit`, `ps−X`, `ba−X` e `pa−X` ($X \in \{apr, may, jun, jul, aug, sep\}$) verificando che corrispondessero a valori presenti nel dominio fornito nei metadati del dataset. I dati forniti non presentavano errori a livello sintattico.

Accuratezza Semantica
Dal punto di vista semantico sono stati analizzati nello specifico gli attributi `pa−X`, `ba−X` e `ps−X`. L'attributo `Payment Amount` è stato confrontato con l'attributo `Bill Amount` e `Payment Status` per verificare a quale importo di pagamento si riferisse il valore indicato `pa−X`, ovvero se `pa−apr`, per esempio, indicasse un pagamento effettuato ad Aprile (quindi relativo al conto di Marzo) o un pagamento del conto di Aprile (indicato da `ba−apr`) e quindi effettuato a Maggio. Attraverso tale valutazione si è concluso che il valore `pa−X` indica il pagamento della cifra indicata dal `Bill Amount` del mese precedente. Questo corrisponde alla realtà poichè utilizzando una carta di credito, il pagamento effettuato in un mese X è relativo all' importo da pagare del mese precedente. Riprendendo l'esempio precedente, il pagamento relativo al conto di Aprile (`ba−apr`) è indicato da `pa−may`. Da sottolineare la presenza di valori negativi nell'attributo `ba−X`, rappresentanti dal punto di vista semantico una situazione di credito e non di debito nei confronti della banca.

	limit	ps-sep	ps-aug	ps-jul	ps-jun	ps-may	ps-apr	ba-sep	ba-aug	ba-jul	ba-jun	ba-may	ba-apr	pa-sep	pa-aug	pa-jul	pa-jun	pa-may	pa-apr
157	460000	-1	-1	-1	-1	0	-1	1637	-196	4594	1517	1517	1306	0	4790	1517	0	1306	0

Figura 1.1: un esempio di record che ha permesso facilmente di riconoscere il valore semantico degli attributi relativi alle transazioni monetarie. Si può notare grazie ai valori di `ps−X`, come la cifra da pagare nel mese di Aprile, per esempio, sia equivalente alla cifra indicata nel pagamento di Maggio e lo stato di pagamento di aprile sia effettivamente -1, ovvero pagato completamente. Anche nei successivi mesi è possibile vedere un comportamento che corrisponde alla semantica degli attributi sopradescritta.

Da sottolineare però che per verificare il corretto valore degli attributi `pa−X`, `ba−X` e `ps−X` sarebbe necessario conoscere tutta la cronologia delle transazioni di una specifica carta di credito così da poter calcolare attraverso semplici operazioni aritmetiche i valori effettivi di spese e pagamenti. Pertanto si è deciso di assumere come corretti i dati considerando inoltre che essendo informazioni raccolte da un istituto bancario non dovrebbero essere presenti dati monetari errati.

Gestione di Outliers e Missing Values
Per la ricerca di outliers è stato prodotto per ciascun attributo il relativo Boxplot, ma per la natura degli attributi del dataset, i valori rappresentati come outliers nel boxplot, non sono considerabili errati. Infatti per gli attributi numerici che rappresentano somme di denaro (`pa−X`, `ba−X` e `limit`), sarebbe scorretto considerare outlier, per esempio, una singola somma di denaro elevata poichè seppur sia un valore isolato, è un valore non errato ma concreto di uno specifico titolare di carta di credito in base a diversi fattori come il reddito. lo stesso ragionamento vale per gli attributi `age` e `ps−X`, poichè valori di età anagrafica o stati di pagamento che si discostano molto dall'andamento medio dei dati non possono considerarsi valori da eliminare.
L'utilizzo dei boxplot ha sollevato una problematica più facilmente osservabile del dataset, ovvero la presenza dei Missing Values:

attributi numerici: nessun attributo numerico relativo al plafond (`limit`), allo stato del pagamento (`ps−X`) o alle transazioni della carta (`pa−X`, `ba−X`) presentano missing values. Questi ultimi, presentano valori negativi o il valore 0, contenuti nel dominio di tali attributi. Per verificare aritmeticamente se i valori negativi o lo zero siano missing values, dovrebbe essere nota la cronologia completa delle transazioni relative a una carta ma avendo solamente dati relativi a un intervallo di tempo (Aprile 2005−Settembre 2005), non è stato possibile verificare se specifici valori nascondessero Missing Values. Un attributo numerico che presenta un effettivo outliers è l'attributo `age`, infatti all'interno del dataset sono presenti per tale attributo molteplici valori -1, valore non coerente con un attributo che indichi l'età anagrafica. È facilmente intuibili come tale valore indichi un valore mancante. Pertanto, tale valore è stato sostituito con la media dell'età anagrafica dei record del dataset, ovviamente non tenendo conto dei valori -1. La media corrisponde a 35.4 che (per coerenza con il tipo di dato intero dell'età) è stata arrotondata a 35 e settata in sostituzione del valore -1.

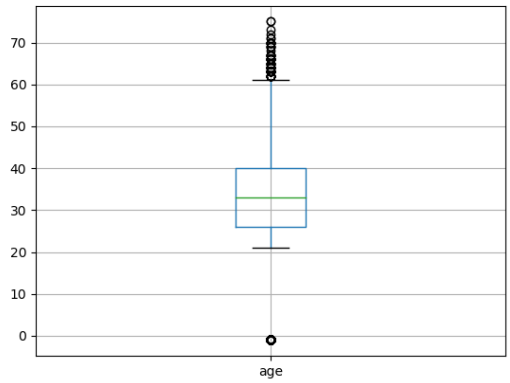

Figura 1.2: Boxplot dell'attributo age

attributi categorici: ad eccezione di `credit_default`, in ciascuno degli altri attributi categorici sono presenti missing values (nan), sostituiti dalla moda dell'attributo:

- `sex`: moda "female"
- `education`: moda "university"
- `status`: moda "single"

1.3 Normalizzazione delle variabili

Gli attributi `limit`, `ba−X` e `pa−X` rappresentando somme di denaro e pagamenti, hanno un ampio range di valori con massimo e minimo molto discostanti. Per utilizzare e visualizzare al meglio questi 3 attributi è stata utilizzata una normalizzazione **min-max**, trasformando il loro dominio in un range di valori continui [0,1].

1.4 Distribuzione delle variabili e analisi statistiche

Attributi Numerici
Per la visualizzazione della distribuzione di attributi numerici sono stati utilizzati degli istogrammi con curva gaussiana. Per gli attributi `limit`, `ba−X`, `pa−X`" e `age` si è scelto un numero di bins ottimale pari a 15 utilizzando la regola di Sturge, mentre per gli attributi `ps−X`, avendo un ristretto range di valori [-2, 8] si è utilizzato un numero di bins pari al numero di valori dell'attributo.

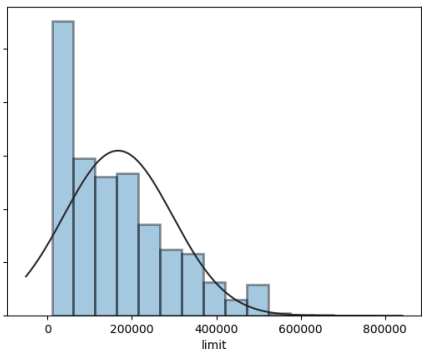

Figura 1.3: Distribuzione attributo `limit`

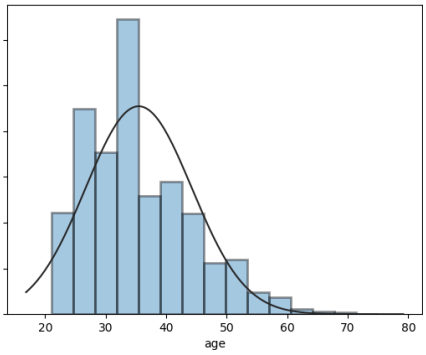

Figura 1.4: Distribuzione attributo `age`

Attributo	Media	Moda	Mediana	STD
limit	167197.0	50000	140000	128975.49
age	35.44	35	35	8.78

Tabella 1.2: valori statistici `limit e age`

Attraverso l'istogramma dell'attributo `limit` (Figura 1.3) è possibile vedere come vi sia una distribuzione inversamente proporzionale alla somma di denaro che la banca mette a disposizione del titolare di carta di credito. Un comportamento dei dati che rispecchia la realtà, evidenziando come le persone più facoltose e con limiti di carta di credito più alte siano anche le meno numerose. L'attributo `age` segue una distribuzione normale, infatti si può notare in Figura 1.4 come il picco dell'istogramma coincida con la curva distribuzione, esattamente sul valore 35.

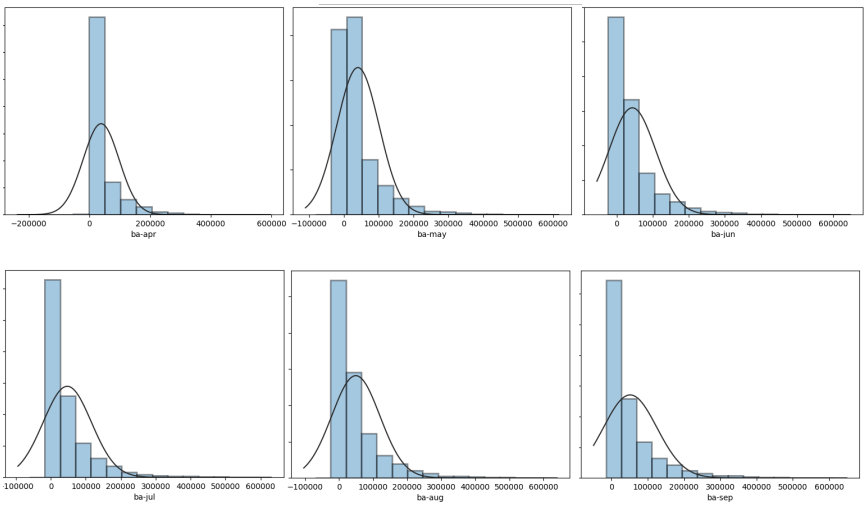

Figura 1.6: Distribuzione attributo `ba-X`

Attributo	Media	Moda	Mediana	STD
ba-apr	38621.58	0	16977	59325.34
ba-may	40182.13	0	18071	60732.33
ba-jun	43306.11	0	19072	64519.91
ba-jul	46957.47	0	19905.5	68948.63
ba-aug	49239.44	0	21202	70777.47
ba-sep	51490.70	0	22246	73740.38

Tabella 1.3: valori statistici `ba`$-X$

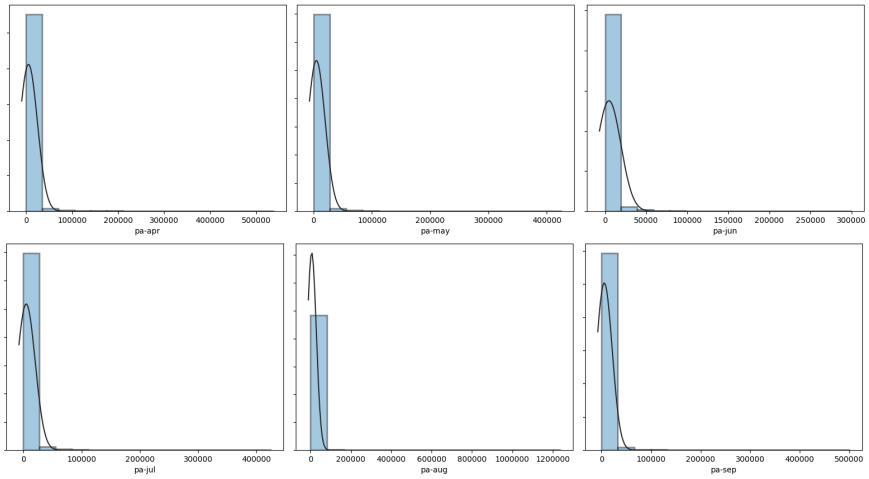

Figura 1.8: Distribuzione attributo `pa-X`

Attributo	Media	Moda	Mediana	STD
pa-apr	5480.15	0	1500	19361.41
pa-may	4734.70	0	1500	14912.37
pa-jun	4719.77	0	1500	14483.41
pa-jul	5131.90	0	1800	15416.40
pa-aug	5973.68	0	2000	22511.75
pa-sep	5651.34	0	2081.5	15835.84

Tabella 1.4: valori statistici `pa`$-X$

Per i gli attributi relativi al `Bill Amount` e al `Payment Amount` è possibile osservare picchi elevati per valori bassi. E' osservabile come il valore più ricorrente in entrambi i casi sia 0, ovvero che la maggior parte dei titolari di carta ogni mese non abbiano da pagare nessuna somma di denaro. Si può come nonostante siano presenti valori medi di `Bill Amount` piuttosto alti, per ogni mese il Payment Amount ha un valore medio di circa 5000. Osservando gli istogrammi in figura 1.10 e la Tabella 1.5 relativi ai Payment Status, si può giustificare la differenza di valori medi degli attributi `Bill Amount` e `Payment Amount` da un alto numero di titolari di carta che ricorrono al metodo di pagamento **revolving credit**, rappresentato dal valore 0 degli attributi `Payment Status`, valore più ricorrente. Inoltre è da sottolineare come gli attributi `Payment Status` seguano a grandi linee una distribuzione normale, con picco sul valore 0.

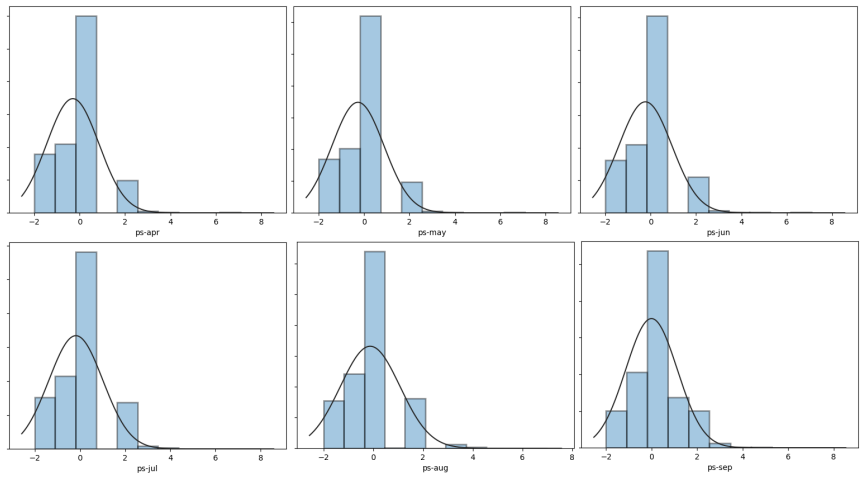

Figura 1.10: Distribuzione attributo `ps-X`

Attributo	Media	Moda	Mediana	STD
ps-apr	-0.30	0	0	1.15
ps-may	-0.26	0	0	1.15
ps-jun	-0.23	0	0	1.17
ps-jul	-0.17	0	0	1.19
ps-aug	-0.13	0	0	1.20
ps-sep	0.00	0	0	1.13

Tabella 1.5: valori statistici `ps`$-X$

Attributi Categorici

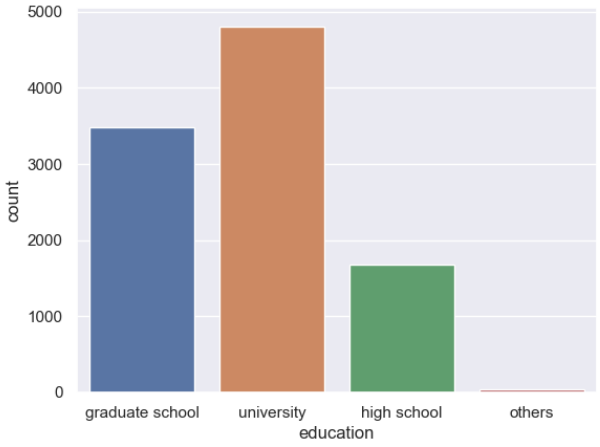

Figura 1.12: Istogramma attributo `education`

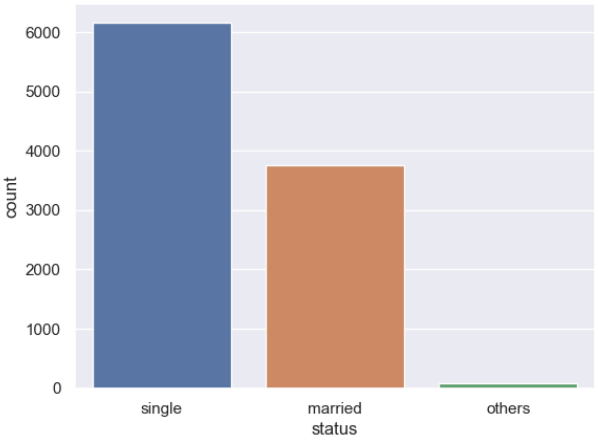

Figura 1.13: Istogramma attributo `status`

Per gli attributi categorici `education` e `status` sono stati generati gli istogrammi riportati in figura 1.12 e 1.13, mentre per gli attributi binari `sex` e `credit_default` si è preferito l'utilizzo di grafici a torta.

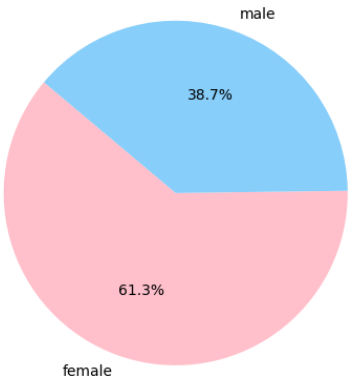

Figura 1.14: Istogramma attributo `sex`

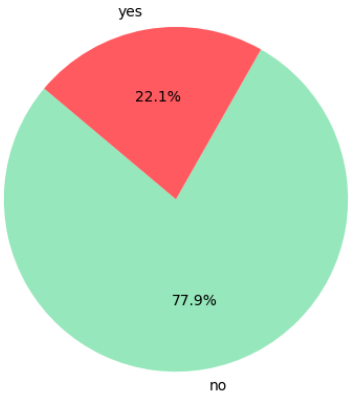

Figura 1.15: Istogramma attributo `credit_default`

1.5 Correlazione tra variabili

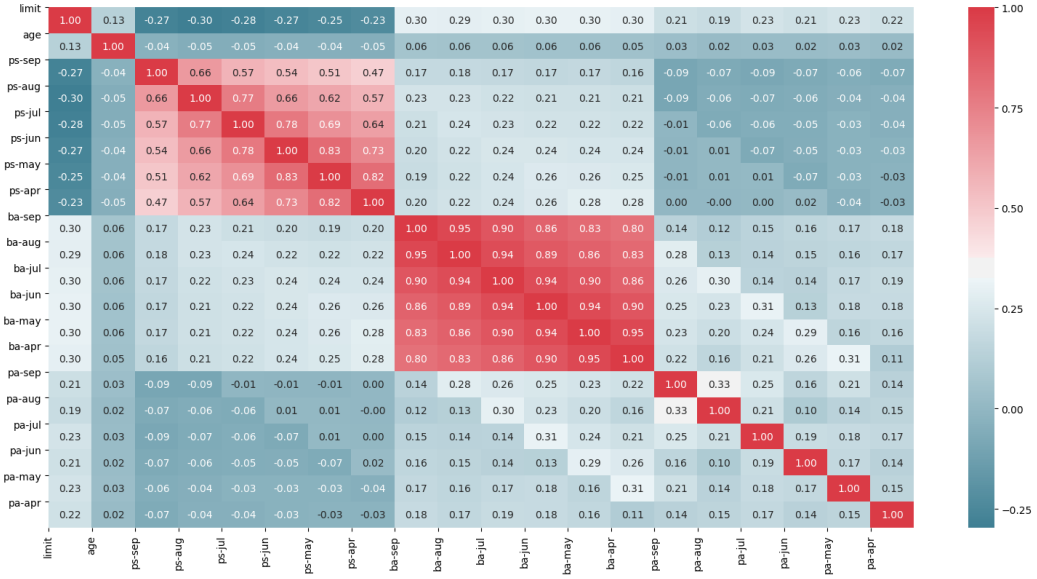

Figura 1.16: Heatmap delle correlazioni tra attributi

Scrivere baggianate sulla correlazione

Capitolo 2

Clustering

In questo capitolo mostreremo il comportamento degli algoritmi di clustering KMeans, DBSCAN e Hierarchical applicati al nostro insieme di dati.

Il dataset utile per questa fase è stato ottenuto eliminando gli attributi categorici `education`, `sex` e `status` e l'attributo `credit_default` poichè il clustering rientra tra gli addestramenti di tipo non supervisionato.

2.1 KMeans

Il miglior parametro k con cui eseguire KMeans è stato stimato calcolando la *SSE* variando k su un range da 2 a 20, con metrica di distanza euclidea. In figura 2.1 sono riportati i risultati ottenuti. Si nota che il cambio di pendenza della curva si trova quando k vale 7. A partire da ciò abbiamo eseguito l'algoritmo per $k \in [3, 10]$, reiterando per ogni passo 50 volte KMeans per evitare che la scelta casuale dei centroidi influenzasse i risultati e calcolando in questo caso anche l'indice di Silhouette per i cluster trovati. In seguito a queste analisi abbiamo constatato come i cluster ottenuti per $k = 4$ siano i più significativi, la nostra considerazione è stata rafforzata dal fatto che il clustering ottenuto per $k = 4$ presenta il valore dell'indice di Silhouette più alto trovato.

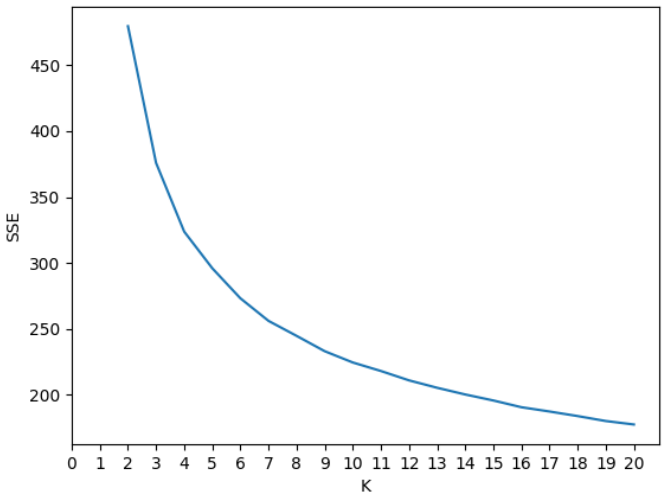

Figura 2.1: SSE

Di seguito analizziamo i cluster trovati, assegnando loro un nome che contraddistingua le caratteristiche di ogni cluster.

Senza rischio Cluster costituito da 3032 persone. Pagano in modo puntuale le loro spese ogni mese senza registrare ritardi e compiono solitamente spese di bassa entità. Non costituiscono alcuna minaccia per la banca.

Piccoli pagatori Gruppo di 4832 persone. Sono soliti fare un uso abbastanza intensivo del revolving credit per pagare le loro spese. Anch'essi registrano spese di bassa entità e non commettono gravi ritardi nel pagamento, per questo motivo rientrano tra clienti credibili per la banca.

Grandi pagatori Cluster di 1083 persone. Come per i *piccoli pagatori*, anch'essi sono soliti usare la modalità di pagamento rateizzata anche se le loro spese sono generalmente molto elevate. Non si registrano comunque grossi ritardi ed é per questo che rientrano comunque tra il gruppo di clienti credibili.

Ritardatari Gruppo formato da 1053 persone. Registrano spese di media entità ma al contrario degli altri tre gruppi si verificano gravi ritardi nei pagamenti. Sono il cluster di persone che, infine, finisce in credit default e non sono clienti credibili per la banca.

In figura 2.2 si sono plottate una selezione delle coordinate dei centroidi, mostrando chiaramente dove i quattro cluster trovati differiscono maggiormente.

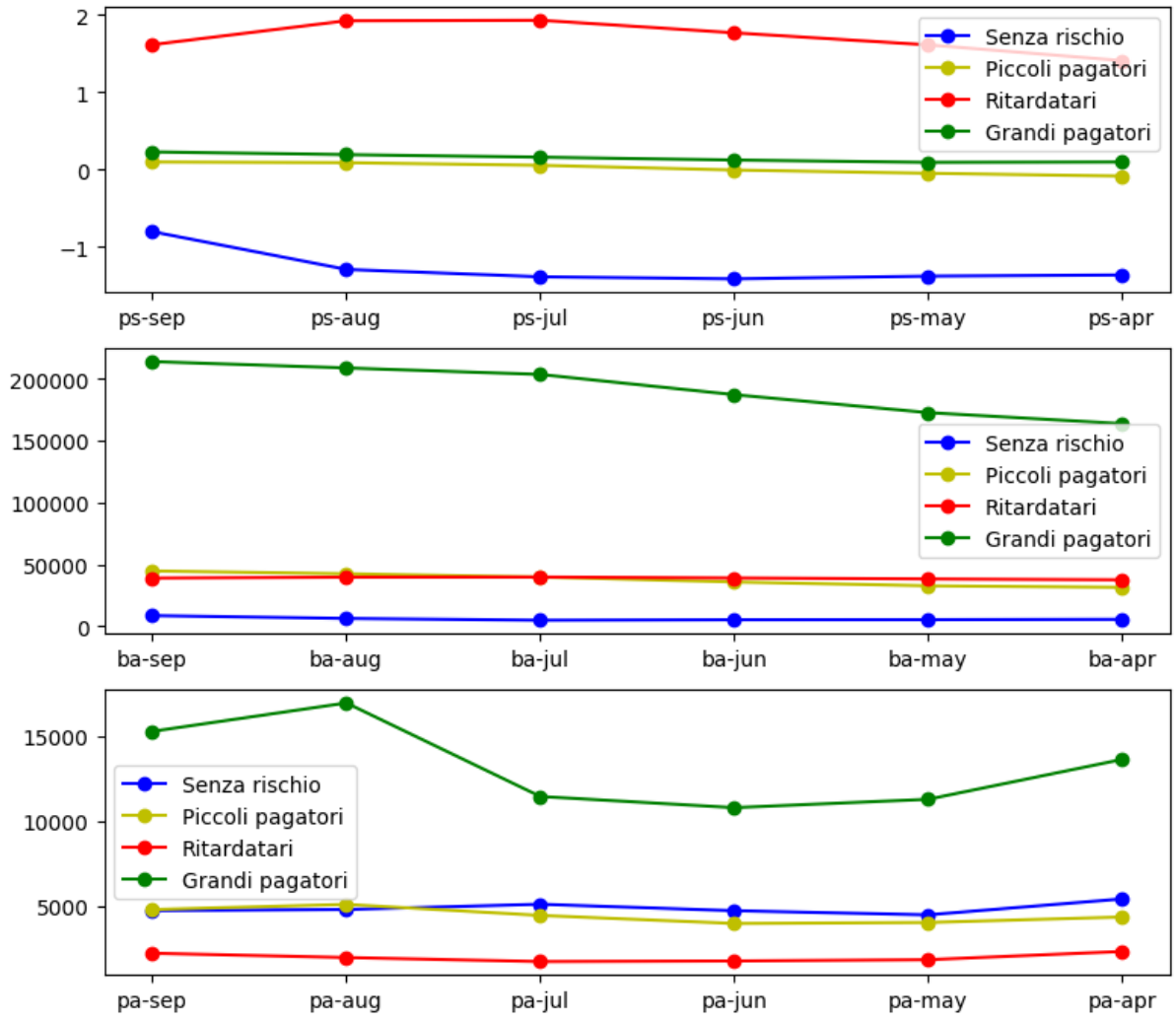

Figura 2.2: Caratteristiche dei cluster

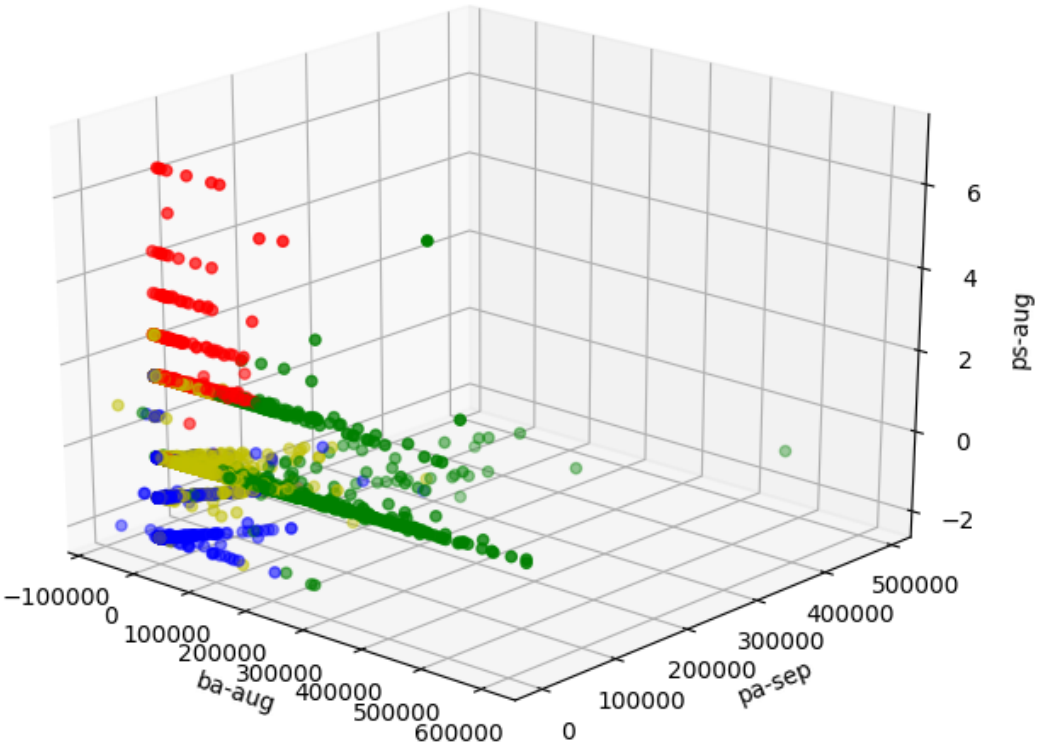

Figura 2.3: Distribuzione dei cluster su un pagamento mensile

Infine si è plottato su un un grafico 3D (figura 2.3 la distribuzione dei cluster su un pagamento mensile (in questo caso si e' preso il pagamento del mese di agosto) ed abbiamo notato che la distribuzione è rispettata per ogni terna di attributi validi costruibili sull'insieme dei mesi disponibili. In particolare il cluster dei *ritardari* si posiziona sempre a valori molto elevati di `payment status`, mentre vale esattamente il contrario per il cluster dei *senza rischio*. Le tre dimensioni scelte per l'esempio sono `billing amount august`, `payment status august` e `payment amount september`.

Infine riportiamo una tabella contenente la media e la deviazione standard di ogni cluster individuato. Per ragioni di spazio per gli attributi `payment status`, `payment amount` e `billing amount` mostriamo i valori degli ultimi due mesi.

Cluster	**ps-sep**	**ps-aug**	**pa-sep**	**pa-aug**
Senza rischio	$-0.8(\pm 1.0)$	$-1.3(\pm 0.7)$	$4.7k(\pm 12.0k)$	$4.8k(\pm 14.3)$
Piccoli pagatori	$0.09(\pm 0.71)$	$0.08(\pm 0.71)$	$4.8k(\pm 10k)$	$5k(\pm 13.3k)$
Grandi pagatori	$0.22(\pm 0.75)$	$0.19(\pm 0.71)$	$15k(\pm 35k)$	$16k(\pm 56k)$
Ritardatari	$1.60(\pm 1.18)$	$1.90(\pm 1.03)$	$2k(\pm 3.4k)$	$1.8k(\pm 3.0k)$
	ba-sep	**ba-aug**	**limit**	**age**
Senza rischio	$8.7k(\pm 21k)$	$6.4k(\pm 16k)$	$215k(\pm 126k)$	$36(\pm 8)$
Piccoli pagatori	$44k(\pm 39k)$	$42k(\pm 35k)$	$130k(\pm 113k)$	$35(\pm 9)$
Grandi pagatori	$213k(\pm 93k)$	$208k(\pm 88k)$	$281k(\pm 115k)$	$37(\pm 8)$
Ritardatari	$38k(\pm 35k)$	$39k(\pm 36k)$	$79k(\pm 68k)$	$34(\pm 8)$
	sex	**status**	**education**	**default**
Senza rischio	F	Single	University	17%
Piccoli pagatori	F	Single	University	18%
Grandi pagatori	F	Single	University	19%
Ritardatari	F	Single	University	61%

Si noti come il gruppo dei ritardatari ha un limite imposto dalla banca molto più basso rispetto agli altri cluster. Segno che la banca ha valutato ottimamente i profili nella scelta di concessione del credito.