Il ruolo dell'informazione nel mercato del credito nell'era dei Big Data

Emanuele Luzzi

Relatore: Professor Marco Spallone

Giugno 2017

B.Sc. Economia e Management

Indice

1 Introduzione

Da sempre nell'elaborare modelli economici ci si è trovati di fronte a delle barriere invalicabili.
Data l'impossibilità di abbatterle, si sono trovati modi per aggirarle, quali l'utilizzo di "assiomi" come la razionalità degli agenti o la perfetta informazione.
Nel 1952 Milton Friedman spiegò come questo metodo assiomatico fosse in grado di fare dell'economia una materia guidata dal metodo scientifico. Infatti, si è ora in grado di produrre ipotesi, sperimentarle seguendo certe variabili ed interpretarne i risultati.
L'errore prodotto dai modelli viene considerato "accettabile", con il limite di dare delle interpretazioni probabilistiche e di tendenza ai risultati ottenuti.

Al giorno d'oggi, con l'avvento dell'informazione digitale, i dati permettono l'adattamento di nuove tecniche al campo. L'Econometria, la materia indirizzata all'analisi dei dati economici, e la Microeconomia sono diventate base portante della Macroeconomia (dalla critica "Microfundation of Macroeconomics" *Lucas 1970*).
Inoltre, nel corso degli anni, gli strumenti analitici e di stima sono migliorati esponenzialmente. Varie branche matematico-quantitative sono state applicate alla scienza economica nel tentativo di conferirle più attendibilità.

Al momento, il più importante sviluppo in corso è quello sullo studio dei Big Data, ossia delle grandi moli dati spesso non strutturati donateci dagli archivi digitali.

Lo scopo di questa ricerca riguarderà proprio quest'ultimo punto.

La prima parte si soffermerà infatti sull'informazione in termini economici e sulla delucidazione del mondo intorno ai Big Data.
Faremo innanzitutto riferimento al modello di Stiglitz & Weiss sul razionamento del credito e lo rivisiteremo traendone le conclusioni inerenti ai nostri scopi.
Successivamente, spiegheremo il processo di analisi dei Big Data a partire dalle sorgenti fino all'interpretazione qualitativa. Ci ricollegheremo al modello tramite l'utilità degli strumenti introdotti nell'area degli intermediari finanziari. Soprattutto, vedremo nel dettaglio i *Credit Score* e i metodi per discriminare gli agenti in base al rischio.

La seconda parte sarà invece dedicata ad una simulazione di un possibile modello per la predizione del rischio nelle banche.
Tramite l'analisi di un dataset saremo infatti in grado di stabilire quali variabili ci interessano ai fini del raggiungimento del nostro obiettivo e come vadano pesate ed interpretate.
Inoltre, andremo ad osservare come le conclusioni del modello S&W si riflettano nell'economia reale in base ai nostri risultati.

2 Informazione e Teoria Economica

Che cos'è l'informazione in termini economici?

Immaginiamo un'azienda produca scarpe e le venda al prezzo di 12€. I consumatori stimano che le scarpe di alta qualità abbiano un valore di 14€, mentre le altre di 8€. Ora, il consumatore *non ha modo di sapere* quali scarpe siano di alta o bassa qualità prima di acquistarle.
Di conseguenza, si affiderà alla qualità media, il quale prezzo è:

$$p = 14q + 8(1 - q)$$

Dove q è la probabilità di alta qualità.
Possiamo ora riflettere su 3 casi:

- Tutti le scarpe sono di bassa qualità ($q = 0$):

  $p = 8 < 12$

  Nessun paio di scarpe viene venduto.

- Tutte le scarpe sono di alta qualità ($q = 1$);

$p = 14 > 12$

Tutte le scarpe vengono vendute e i consumatori ne ricevono un surplus pari a:

$SP = 2$

- Alcune scarpe sono di alta qualità e altre no, nel qual caso dobbiamo definire la probabilità minima che permetterà all'azienda di vendere almeno un paio di scarpe:

  $12 = 14q + 8 - 8q \rightarrow q = \frac{2}{3}$

  ossia almeno 2 scarpe su 3 devono essere di alta qualità affinché i consumatori siano disposti a comprarne almeno un paio.

Quest'esempio, il quale è una versione ripresa dal "Market for Lemons" di Akerlof, è un esempio teorico di Adverse Selection.
Essa si verifica ogniqualvolta si ha una scelta da fare in condizioni di informazioni nascoste. Nell'esempio, il consumatore ha bisogno di comprare un paio di scarpe, ma non potrà mai sapere quali sono di buona qualità e quali di cattiva.
Allo stesso modo, una banca vorrebbe sapere quali clienti ripagheranno il debito e quali no, un'assicurazione quali agenti verranno derubati più facilmente e quali invece no etc...
La selezione avversa è il motivo per cui si parla di rischio, è tra le ragioni principali per cui molti modelli non posso dare risultati esatti, e allo stesso tempo per cui molti altri cercano di capirla e ridurla.

Nel nostro esempio abbiamo anche accennato un metodo elementare per contrastarla, l'imposizione di un'*incidenza* media.
Il prezzo delle scarpe è la media tra due possibili stati, così come il premio assicurativo:

$P_h > \bar{P} > P_l$

Dove gli agenti ad alto rischio (P_h), saranno avvantaggiati con un surplus derivante dal pagare $\bar{P}$, e gli agenti a basso rischio (P_l) potranno godere di un prezzo più vicino al loro budget ottimale.

Un altro fenomeno caratteristico dell'asimmetria informativa è il *Moral Hazard*.
Esso descrive il comportamento umano conseguente dal non usare i propri soldi (egoismo naturale).
Se la nostra bici è assicurata abbondantemente contro il furto, mancheranno incentivi per l'agente a preoccuparsi che essa non venga rubata.
Di conseguenza, la probabilità di essere considerati rischiosi da parte della società assicurativa aumenta nel momento in cui ci procuriamo l'assicurazione stessa.
I metodi attuariali devono quindi tenerne conto quando vanno ad impostare il premio.
Se l'assicurazione è troppo elevata renderà il furto futile alla prospettiva del cliente, il quale non ha alcun interesse nella massimizzazione del profitto della società; se invece è troppo bassa, taglierà fuori una potenziale parte di Domanda.
Questo circolo di azioni nascoste crea quindi un paradosso:

l'assicurazione stabilirà un prezzo minore al prezzo ottimale per assicurarsi un incentivo alla cura del bene, nonostante gli agenti siano disposti a pagare di più. Infatti, se il prezzo e la copertura assicurativa aumentassero, avere meno cura del bene sarebbe una scelta razionale.

Ultimo ma non meno importante è il *Cost of the Verification Process*.
Essi sono i costi affrontati nel verificare se gli accordi e gli incentivi derivanti dal contratto sono in atto.
Si ricollega al concetto di informazioni nascoste nella Selezione Avversa e per quanto possa sembrare un metodo per contrastarla, il fatto che ci siano dei costi coinvolti li rende oggetto di studio nella valutazione dei trade-off informativi.

2.1 Il Modello

Prima di cimentarci nell'analisi concreta dei Big Data, non possiamo non selezionare un modello di riferimento. Per lo scopo della nostra ricerca, rivisiteremo un modello riguardante l'informazione imperfetta e le sue conseguenze nel complesso del mercato del credito.
Nello specifico, faremo riferimento al fenomeno del Credit Rationing (*Stiglitz & Weiss 1981*)

Intuizione

Il più noto e affermato fondamento della scienza economica dimostra come siano i prezzi i veri protagonisti del raggiungimento dell'equilibrio. Dunque, *da dove deriva la necessità di razionare il credito?*

Questa è la domanda con cui si apre il modello.

Sicuramente una prima "risposta" possiamo ricercarla nel periodo di transizione tra due diversi equilibri, ad esempio a seguito di uno shock esogeno. In questo intervallo di tempo si osserveranno delle inspiegabili frizioni sul mercato del lavoro e del capitale, intervallo nel quale si introdurrà il razionamento.
Durante questo periodo le *banche* saranno ostacolate nel loro obiettivo di massimizzare i tassi d'interesse a causa dell'informazione imperfetta. Infatti, la mera scelta dei tassi potrebbe comportare un rischio derivante da:

1. Scelta degli agenti *(selezione avversa)*
2. Azzardo morale

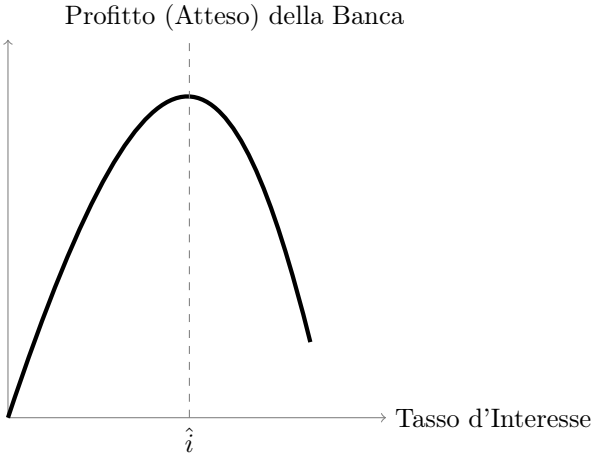
Nello specifico, la selezione avversa è caratterizzata dalla necessità degli intermediari finanziari di riconoscere gli agenti con alta probabilità di ripagare il finanziamento. In un ipotetico stato del mondo con informazione perfetta sarebbe infatti sufficiente applicare diversi tassi per le diverse tipologie di committenti:

- $R_{\mathrm{H}} \Rightarrow$ Agenti ad alto rischio
- $R_{\mathrm{L}} \Rightarrow$ Agenti a basso rischio

Nel momento però in cui si presentano frizioni nasce il bisogno di utilizzare dei meccanismi di *screening*, ossia gli strumenti addetti alla discriminazione tra "buoni" e "cattivi" agenti.

Struttura

Uno di questi strumenti è proprio il tasso. Al crescere del tasso $\bar{R}$, si verificherà un aumento della rischiosità media degli agenti coinvolti, in quanto essi percepiranno più bassa la loro probabilità di ripagare il finanziamento. Tuttavia,la nostra intuizione ci dice che i progetti rischiosi saranno anche i più redditizi.



La banca sarà incapace di monitorare direttamente il progetto di finanziamento. Quindi, formulerà il contratto in maniera tale da indurre l'agente a comportarsi entro certi vincoli. La conseguenza di ciò sarà un profitto atteso della banca più lento al crescere del tasso d'interesse, fino ad arrivare ad un punto di massimo dopo il quale esso decrescerà (per ragioni che andremo a spiegare a breve).

Innanzitutto se F è l'ammontare del finanziamento al tasso d'interesse $\hat{i}$, definiamo la condi-

zione di default nel momento in cui il ritorno R dell'agente insieme ad eventuali garanzie collaterali C non è sufficiente a coprire la quantità desiderata:

$$R + C \leq F(1 + \hat{i}) \tag{1}$$

Ora per facilità assumiamo che la banca conosca il ritorno medio R di due progetti ma non il grado Φi di rischiosità degli stessi. *Soprattutto*, utilizziamo un modello di rischio noto come *Mean Preserving Spread*, che ci permette di comparare due diverse distribuzioni (una lo spread della PDF dell'altra) fermo restando che il valore atteso non cambi. Definiamo la distribuzione dei ritorni come $F(R, \Phi)$ e la relativa funzione di densità come $f(R, \Phi)$.
Quindi per $\Phi_1 > \Phi_2$:

$$\int_0^\infty Rf(R, \Phi_1)dR = \int_0^\infty Rf(R, \Phi_2)dR$$

E per $t \geq 0$:

$$\int_0^t F(R, \Phi_1)dR \geq \int_0^t F(R, \Phi_2)dR$$

Infine, assumiamo che gli agenti siano neutrali al rischio, i progetti abbiano un costo fisso e non siano divisibili.

A questo punto abbiamo una **struttura** che ci permetterà di derivare diversi Teoremi sul funzionamento del mercato dei crediti con informazione imperfetta.
Tuttavia, per lo scopo di questa ricerca, ci concentreremo su ciò che ci sarà utile per l'applicazione successiva dei Big Data.

Ottimizzazione e Teoremi

Chiaramente ora sappiamo che il ritorno $\pi(R, \hat{i})$ dell'agente sarà definito come:

$$\pi(R, \hat{i}) = max(R - F(1 + \hat{i}); -C) \tag{2}$$

Ossia il ricavato dal progetto in caso di successo o la collaterale in caso di fallimento.
Di risposta la banca si aspetterà:

$$\rho(R, \hat{i}) = min(R + C; F(1 + \hat{i})) \tag{3}$$

Dove l'agente ripagherà la somma promessa o il massimo che può ripagare $R + C$ (analiticamente: la banca si limiterà a profittare dal ritorno solo e soltanto nel momento in cui è minore della somma dovuta).
Possiamo ora trarre diverse conclusioni.

Teorema 1: *"Per un certo tasso d'interesse $\hat{i}$ (profitti attesi nulli), esiste un certo livello di rischiosità $\hat{\phi}$ tale che se $\phi \leq \hat{\phi}$ il progetto non verrà realizzato."*

Questo teorema segue logicamente la convessità della (2). Di conseguenza, i profitti degli agenti crescono con la rischiosità del progetto. Calcolando il valore di $\hat{\phi}$ per il quale i profitti attesi sono nulli otteniamo:

$$\Pi(\hat{i}, \hat{\phi}) = \int_0^\infty \pi(R, \hat{i})dF(R, \hat{\phi}) = 0$$
$$\text{dove:} \qquad dF(R, \hat{\phi}) = f(R, \hat{\phi})dR$$

E differenziando abbiamo:

$$\frac{d\hat{\phi}}{d\hat{i}} > 0 \tag{4}$$

Quindi:

Teorema 2: *"Al crescere del tasso d'interesse il valore critico $\hat{\phi}$ cresce".*

Quindi l'aumento del tasso di interesse attirerà agenti più rischiosi, ma data la relativa

concavità della (3),l'aumentare del rischio porterà il profitto atteso della banca a decrescere:

$$\frac{d\rho}{d\phi} < 0 \tag{5}$$

Analiticamente la concavità della funzione in esame porterà ad un'area decrescente all'aumentare della variabile:

Teorema 3: *"Il profitto atteso della banca è una funzione decrescente della rischiosità del progetto".*

I Teoremi **2** e **3** ci mostrano esplicitamente l'effetto della *selezione avversa*. Normalmente ci aspetteremmo infatti un aumento dei profitti attesi (della banca) al seguire di un aumento dei tassi, invece adesso sappiamo che l'aumento dei tassi ridurrà significativamente l'arco degli agenti vincolandolo a quelli più rischiosi, i quali di tutta risposta tenderanno a diminuire i profitti in quanto in situazione di default (1).

Possiamo ora illustrare il grafico mostrato in precedenza tramite il:

Teorema 4: *"Dato un numero discreto di agenti ognuno con un ϕ differente, $\rho(i)$ non sarà una funzione monòtona di i, in quanto per ogni agente che lascerà il mercato ci sarà una caduta dei profitti".*

Ma allora la banca sarà interessata ad applicare un tasso che massimizzi $\rho(i)$ *anche se ciò significherà lasciare una parte della domanda insoddisfatta* :

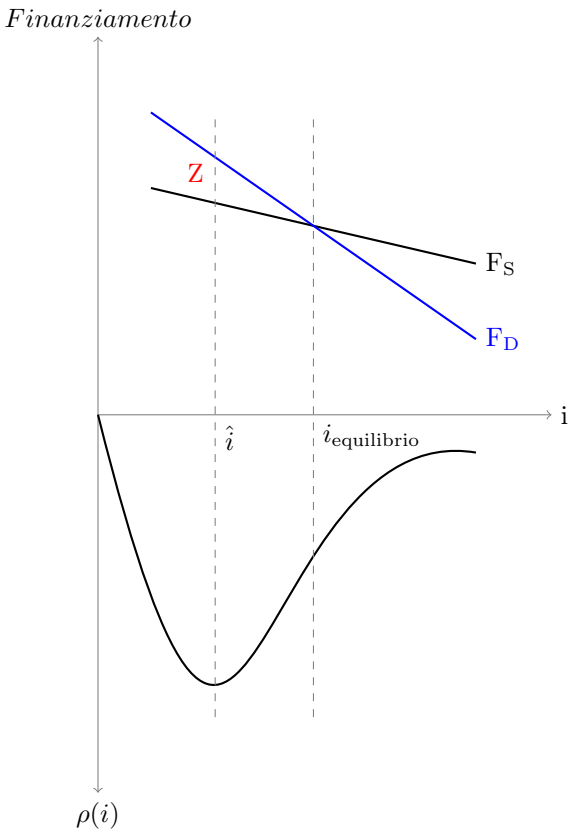
Il che dimostra:

Teorema 5: *"Se la funzione dei profitti $\rho(i)$ ha un punto di massimo (globale), esiste un'offerta di fondi tale che l'equilibrio competitivo implichi il razionamento del credito".*

Dove Z è la misura dell' *eccesso di Domanda.*

Se invece la curva è caratterizzata da più massimi (locali e globali), l'equilibrio di mercato potrà avere:

1. Un tasso al massimo uguale a i_e
2. Due diversi tassi, con un eccesso di domanda per il minore.



credito è confermato su dati reali e tende ad imporsi in maniera simile a come descritto.

Andremo ora ad osservare un esempio recente del fenomeno nella nostra Economia.

2.2 Conclusioni

Come volevasi dimostrare, in presenza di informazione imperfetta e rischio, la banca massimizza i suoi profitti (e quindi la sua utilità) ponendo un tasso/dei tassi non congruenti con quelli dell'equilibrio di mercato, razionando il credito.
Importante, si avrà una parte della domanda di credito insoddisfatta.

Questo modello è l'esempio perfetto per la nostra ricerca. Infatti, per quanto esso faccia delle assunzioni non indifferenti, il razionamento del

2.3 Il Credit Rationing nell' Economia Italiana

Uno dei casi più celebri e attuali di razionamento del credito è quello delle piccole e medie imprese italiane.
Per anni si è supposto che le piccole imprese fossero meno interessate ad essere finanziate da intermediari, non il contrario. Tuttavia, nel 2017 Banca d'Italia afferma:

> *"Nel 2015 i prestiti bancari sono cresciuti per le imprese di maggiore dimensione mentre hanno continuato a contrarsi per quelle più piccole; questo divario si osserva anche per aziende appartenenti allo stesso settore di attività economica o con condizioni di bilancio simili. Stime econometriche confermano che, a parità di numerose caratteristiche di impresa (redditività, liquidità, dinamica del fatturato, spesa per investimenti, settore di attività economica e area geografica), il credito si è ridotto soprattutto per le microimprese e per le aziende più rischiose. La maggiore fragilità finanziaria delle microimprese, dovuta in particolare al più elevato indebitamento, spiega oltre il 70 per cento della differenza nel tasso di variazione dei prestiti con le grandi aziende e circa il 40 di quello con le imprese di piccola e media dimensione."*

L'analisi è stata svolta su un campione di 260.000 imprese società di capitali, di cui 197.620 micro imprese.
Il fattore più evidente per il contrarsi dei crediti è la rischiosità delle imprese in questione.
Tuttavia, è stato rilevato un fattore ombra (non presente nelle regressioni) influente ma non correlato con il rischio.
Infatti:

Tabella 1: SME

Categoria	Dipendenti	Turnover	Totale Passivo
Medie	< 250	$\leq$ €50 m	$\leq$ €43 m
Piccole	< 50	$\leq$ €10 m	$\leq$ €10 m
Micro	< 10	$\leq$ €2 m	$\leq$ €2 m

> *"I risultati indicano infine che vi è una componente della minor crescita del credito delle microimprese non spiegata dagli indicatori inclusi nelle regressioni. Ciò potrebbe riflettere una minore propensione delle banche a finanziare clientela di piccola dimensione a causa della maggiore incidenza dei costi fissi oppure le difficoltà ad adattare i metodi di valutazione del merito di credito basati sull'informazione qualitativa ai rilevanti cambiamenti tecnologici e regolamentari in corso. [...] Potrebbero anche incidere fattori dal lato dei costi, che rendono non adeguatamente remunerativo per le banche l'erogazione e la gestione di fidi di importo contenuto."*

Inoltre, è interessante notare le differenze a livello internazionale.
Ricordando che l'italia ha un rapporto pmi / Grandi Imprese sopra la media mondiale, l'Organizzazione per la Cooperazione e lo Sviluppo Economico ha elaborato un modello generale.

Sull'asse delle ordinate troviamo la quantità di prestiti concessi alle SME, ossia le aziende che rientrano nella definizione di medie, piccole o micro (Tabella 1)

Sull'asse delle ascisse abbiamo invece la variabile associata al PIL.
La correlazione positiva non è certo una sorpresa.
Tuttavia, non possiamo trarre conclusioni sulla causalità basandoci solo sui risultati illustrati.

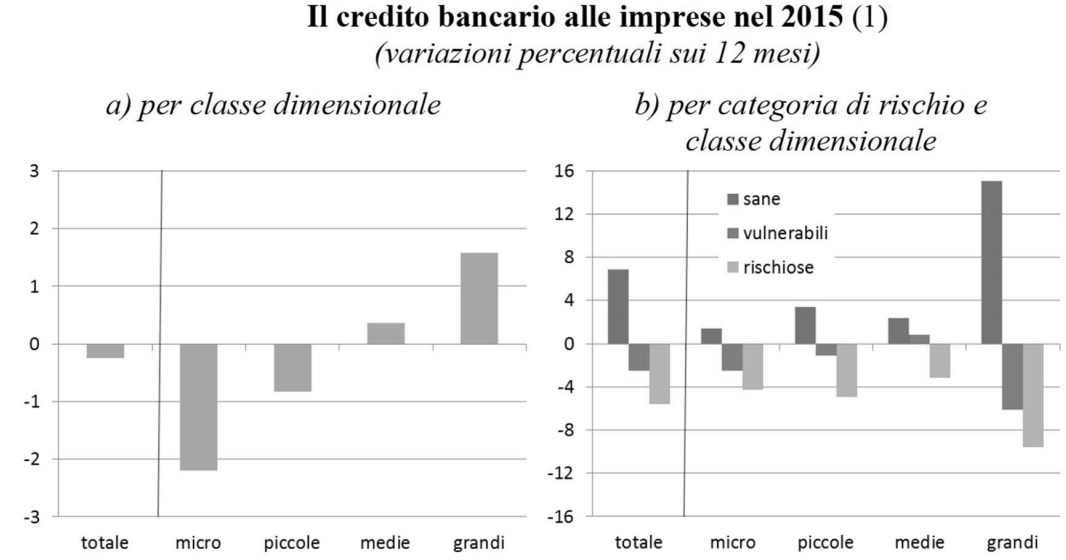


Figura 1: ***Credito Bancario***

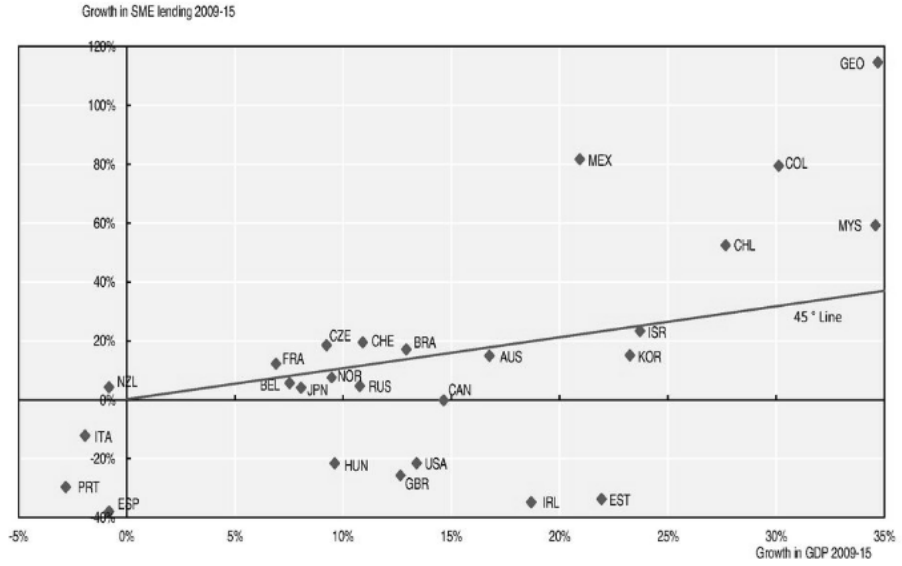


Figura 2: ***SME nel Mondo e Correlazione con il PIL***

3 L'Era dei Big Data

3.1 Il Contesto

Non è certo una novità che i dati siano oggetto di estensiva ricerca da parte di numerose discipline ed imprese, in particolare nei contesti business, economici e delle scienze sociali.
Tuttavia, fino a neanche due decenni fa la quantità di dati rappresentativi presenti sull'attività economica era terribilmente scarsa. Le procedure di raccoglimento informazioni erano nella maggior parte efficienti quanto lente e i risultati venivano "spremuti" pesantemente al fine di ricavarne fino all'ultimo carattere significativo.
Sia per l'avvento di Internet, sia per l'avanzamento tecnologico o per l'attenzione crescente data alla ricerca di informazioni, la situazione si è completamente ribaltata negli ultimi anni. Per questo motivo, tutti ormai sappiamo come ogni ricerca, transazione, operazione o semplicemente click è registrato nei rigorosi database delle grandi imprese (e non solo). Questi enormi aggregati di dati (spesso non strutturati) sono risultati talmente ricchi di preziose informazioni da aver meritato una categoria dedicata, i cosiddetti *"Big Data"*.

Il motivo per cui è divenuto fondamentale introdurre questo fenomeno in una ricerca economica è una conseguenza dell'incredibile potenziale che se ne può trarre. Infatti, non è certo una sorpresa che figure quali il Data Scientist o il Data Modeler siano diventate centrali per le imprese, le quali non solo non fanno segreto della loro crescente domanda per talenti simili, ma anzi stanno implementando figure manageriali apposite (e.g. Chief Data Officer).
Inoltre, modelli considerati fondamenti della scienza economica stanno andando incontro ad un processo di rivalutazione radicale e l'Econometria sta diventando uno strumento brillante per trarre conclusioni realistiche sulla base di metodi statistici.

Sorgenti:

Innanzitutto, dove sono questi Big Data? Probabilmente le prime piattaforme che identifichiamo con il fenomeno sono i Social Media, ma perché imporre questo limite? *Qualsiasi* informazione in *qualsivoglia* contesto è considerata eligibile per l'aggregato. Ricordo infatti che stiamo parlando di smisurati insiemi dati caratterizzati da assenza (totale o parziale) di strutture o tabulati.
Possiamo infatti cominciare a scalfirne la superficie con:

- Archivi di documenti scansionati
- Log di Sistema
- Documenti elettronici (xls, pdf, email, html, xml, json, etc.)

E ancora:

- Applicazioni di Business (residuals e logs)
- Data Storage (possono essere strutturati ma comunque di proporzioni ingenti, come i *portali intranet*)

Ma anche:

- Media (qualunque manifestazione)
- Web
- Deep Web (siti non indicizzati)

Tuttavia, il fatto che siano facilmente individuabili non significa siano anche altrettando accessibili.
Infatti questa vera e propria sfida affonda le sue basi nelle gerarchie sociali, al cui apice troviamo

i cosiddetti *Senior Researchers*, rinominati chiaramente in ambito accademico.
Un esempio lo abbiamo nella ricerca del team in collaborazione con Raj Chetty, professore di Economics all'università di Stanford e laureatosi ad Harvard. La ricerca riguardava quanto presto l'educazione in tenera età avesse effetto sul successo futuro. Grazie al network di Chetty, il team aveva accesso al US Internal Revenue Service Data e, per quanto complicata e prolungata fosse la successiva analisi, l'apertura di questa porta gli ha permesso di aggirare l'incombente ostacolo iniziale della ricerca. La conseguenza di ciò è lampante, abbiamo infatti grandi economisti accademici quali Hal Varian e Huberman in allontamento dal mondo accademico per lavorare in imprese globalizzate alla ricerca di accessi (rispettivamente in Google e Hewlett-Packard).

La domanda non è quindi *dove* trovare i dati, ma *cosa* cercare e *come* accedervi.

Invero, una volta circoscritto lo scopo della ricerca si pone il secondo dilemma della questione: l'*Analisi* e l'*Interpretazione*.

3.2 L'Analisi

L'analisi della Business Intelligence nel contesto Big Data può essere suddivisa in quattro macroaree:

- *Descriptive Analytics*:
  Dr. Micheal Wu, Chief Scientist di Lithium Technologies, afferma che l'80% delle operazioni BI ricadono in quest'area. Si tratta della forma più semplice di analytics, volta a condensare i BD in pacchetti ordinati di informazione. Se la parte descrittiva ha successo, allora si potrà cominciare ad identificare andamenti (*patterns*) e a trarre conclusioni.

- *Predictive Analytics*:
  Questa parte più tecnica riguarda l'applicazione di metodi statistici (quali vedremo più avanti) volti a predire lo sviluppo futuro di variabili. Tra questi non possiamo non citare la *regressione* (lineare, logistica, non lineare, multivariata etc.) e lo *studio delle serie temporali*.
  Un esempio attinente sono i cosiddetti *Credit Score*, ossia dei veri e propri punteggi quantitativi assegnati dagli intermediari finanziari agli individui in modo da distinguerli in base alla rischiosità.
  Detto ciò, quest'area verrà ripresa più avanti nella nostra ricerca in quanto rientra nell'applicazione BD al modello di nostro interesse.

- *Prescriptive Analytics*:
  Chiaramente, una volta identificato un *probabile* andamento si cerca di ottimizzarlo nella maniera più rigorosa possibile. I software di navigazione (GPS) utilizzano da anni queste tecniche ogniqualvolta ce ne serviamo per ricercare la via più breve. Il mondo del Business non fa certo eccezione, da portafogli finanziari alla ricerca ed analisi di mercato si presuppone sempre un'ottimizzazione. Il sistema potrebbe per esempio rilevare come per aumentare del 10% il Conversion Lift bisogni diminuire del 35% la campagna email, aumentare del 25% la campagna social media e abbassare al 10% le survey telefoniche. A livello di analisi questa è la stessa procedura addetta a rilevare se girare a destra o a sinistra al prossimo incrocio sul GPS, con l'unica differenza che ora non ci troviamo in un campo geospaziale.
  Infine, è giusto ricordare che per quanto le prescrizoni permettano di migliorare l'efficacia di una manovra, esse non possono giudi-

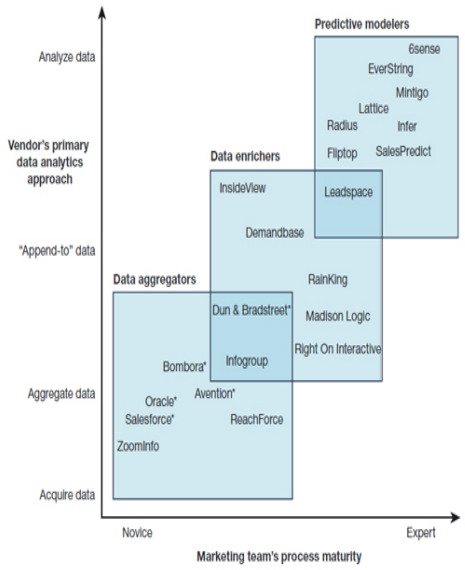


Figura 3: *Evoluzione nell' Analisi dei Dati Fonte: Forrester Research (2015)*

care se era prima di tutto corretto svolgerla. Il KPI (Key Performance Indicator) di un business è un fattore che può essere intuito solo e soltanto dalle menti umane.

- *Automated Analytics*:
  Ultime ma non meno importanti sono le procedure prettamente algoritmiche che avvengono all'interno dei software d'analisi una volta impostati i parametri arbitrari (ad esempio tramite *machine learning*). Sono fondamentali dal punto di vista tecnico/ingegneristico dell'analisi.

Infine, in *Figura 1* possiamo vedere come qualitativamente l'offerta si possa dividere tra tre delle quattro suddette macroaree. Prendendo in considerazione la maturità del team Marketing (da novizi ad esperti) e l'approccio analitico dei fornitori (dalla semplice acquisizione dati all'analisi), possiamo individuare diversi strumenti, alcuni dei quali polivalenti tra più aree. In generale:

- ***Aggregatori:*** ossia il mero raccoglimento e a volte pulizia di dati aziendali ed esterni.
- ***Arricchitori:*** i quali tendono invece ad estrapolare informazioni mirate.
- ***Modellatori:*** i quali costruiscono strutture basate su algoritmi e distribuzioni di probabilità, le quali andranno poi a prevedere andamenti futuri.

Errori nell'Interpretazione:

Per quanto ovvio, è opportuno rammentare che la vastità dei dati in questione tende a confondere facilmente anche gli analisti più esperti. Spesso infatti, dopo aver realizzato una Big Data Architecture articolata e funzionante, ci si accorge di essere andati fuori strada rispetto alle direttive manageriali. Gli errori che si commettono a livello interpretativo sono comuni a quelli dei dati strutturati, in quanto si presentano dopo la fase di aggregazione dei dati. Infatti, ha senso parlare dei più comuni errori nell'interpretazione dei BD ancor prima di spiegare come effettivamente vadano implementati nelle decisioni. Dividiamo quindi gli errori in due gruppi: *quantitativi* e *qualitativi*.

Quantitativamente si commettono sbagli nel momento in cui il modello, per quanto funzionante, proceda su assunzioni incorrette. Nel dettaglio avremo:

- ***Confirmation Bias***: nel momento in cui gli analisti usano un campione eccessivamente limitato di dati per testare un'ipotesi che reputano intuitivamente corretta. Se inoltre i dati sono selezionati in maniera soggettiva e non oggettiva, si parla anche di *Selection Bias*.
- ***Non-Normality***: assumere la distribuzione normale non è soltanto una convenzione, essa affonda le sue radici nella Teoria delle

Probabilità applicata alla Statistica Inferenziale, specialmente nei riguardi del Teorema del Limite Centrale.
Tuttavia, non sempre le variabili casuali in esame fanno riferimento ad una popolazione adatta o presentano le caratteristiche per essere approssimate ad una Gaussiana, ciò crea una distorsione dei risultati.

- ***Overfitting e Underfitting***: rispettivamente, l'utilizzo di un modello esageratamente complicato o eccessivamente semplice risulta nell'inefficienza di arrivare a risultati attendibili.

- ***Confounding Variables***: variabili che spiegano la correlazione sia con la variabile indipendente (con la cui già forma un bias) sia con la variabile dipendente. Esse tendono a confondere i risultati in quanto mirano a distorcere l'interazione tra i fattori osservati.

Supponendo che l'analisi sia stata realizzata impeccabilmente e in maniera efficiente, l'errore di interpretazione può allora risiedere nella revisione qualitativa dei diretti interessati *esterni* alle procedure tecniche.
Non soltanto i dirigenti e committenti tenderanno infatti ad *intuire* il significato dei risultati, ma apprenderanno e divulgheranno l'informazione che più li aggrada.

Simpson's Paradox: questo fenomeno è adatto a descrivere la situazione.
Esso pone in luce come per quanto diversi gruppi osservati possano presentare un trend, esso può dissolversi (o ripresentarsi opposto) nel momento in cui i dati vengano combinati. L'importanza di questo fenomeno è immensa e merita di essere osservata più nello specifico. Un esempio semplice è il seguente (inspirato dal "*UC Berkeley Gender Bias 1973*"):

Supponiamo di voler dimostrare un *gender gap* tra gli studenti ammessi ad una certa università, essi risultano:

	Popolazione	Ammessi
M	6650	**44%**
F	4000	**34%**

A chiunque venga presentata una tabella simile verrà data l'irremovibile (e pericolosa) idea che i maschi siano favoriti alle femmine nel processo di immatricolazione dell'università.
Ciò che non viene mostrato però è il contesto della stessa; infatti è completamente sorvolato il fatto che l'università sia divisa in diversi dipartimenti e che chiaramente le scelte degli stessi non siano egualmente distribuite.
Di seguito allarghiamo i risultati ai dipartimenti più significativi:

	M		F	
Dip.	Popolazione	Ammessi	Popolazione	Ammessi
A	1800	67%	450	**86%**
B	500	50%	1600	**68%**
C	850	**49%**	250	38%
D	1700	58%	300	**70%**
E	500	28%	1500	**30%**
F	1300	**30%**	100	8%

Paradossalmente rileviamo ora una leggera tendenza a vantaggio delle femmine, il che modifica radicalmente le conclusioni che avevamo tratto dalla tabella precedente.
Di questo passo potremmo rivalutare molte fuorvianti statistiche non solo in termini di Business ma anche in termini sociali.

Un altro grande errore nella revisione qualitativa è la confusione tra *Correlazione* e *Causalità*. La Correlazione (nella maggioranza dei casi lineare) è condizione *necessaria* ma **non** *sufficiente* ad affermare che due eventi siano uno la causa dell' altro. Solo dopo aver dimostrato una *cronologia* tra i due e, soprattutto, la *stabilità* tramite introduzione di terze variabili (correlate)

possiamo giustificare questa affermazione.

In conclusione, questo tipo di errori derivano più dall'intuizione di chi cerca di interpretare le analisi al di fuori dell'ambiente tecnico (ad esempio gli stessi Managers), piuttosto che dagli operatori stessi.
Tuttavia, difficilmente si potranno trarre conclusioni valide da un un modello basato su strumenti inattendibili.

3.3 Previsione dei Default Finanziari

A questo punto, è opportuno riportare la nostra attenzione sull'informazione imperfetta nell'Economia. Il Razionamento del Credito definito nel modello di Stiglitz & Weiss è infatti una diretta conseguenza della selezione avversa, ossia dell'impossibilità di distinguere tra diverse tipologie di agenti. Tuttavia, abbiamo accennato come l'analisi dei Big Data possa assumere funzioni predittive e di forecasting; sarebbe quindi attinente supporre che i modelli siano in grado di "prevedere" la rischiosità di un dato individuo? Ebbene, andremo ora nel dettaglio dei già citati "Credit Scores".

I Credit Score si presentano come dei veri e propri servizi sia agli agenti che agli intermediari finanziari, assegnati da società ad hoc come una sorta di "etichetta" la quale quantifica il potenziale rischio di un cliente. Ogni Score è derivato e fondato su diverse variabili, fornite a loro volta da altre agenzie specializzate nell'archiviazione di informazioni sui crediti passati (ad esempio *Equifax, Experian e TransUnion* negli USA).

Prenderemo ora in esame i tre CS più importanti degli Stati Uniti:

- FICO Score
- VantageScore
- Plus Score

FICO Score

Il Fair Isaac Co. Score è stato ed è tutt'ora il più conosciuto ed utilizzato. Esso è indirizzato principalmente ai prestatori finanziari (anche se non mancano gli utilizzi individuali). Il punteggio è dato da una scala da 300 a 850, dove più alto è il numero e meno rischioso è considerato il soggetto. Per essere considerati idonei per il FICO è richiesta una storia contabile e un conto aperto da almeno sei mesi.

Le variabili principali usate nella derivazione dell'indice sono:

- ***Storia dei pagamenti [35%]***: l'andamento dei pagamenti registrati sul conto in esame. L'analisi si focalizza su tre punti: *Recency*, gli ultimi pagamenti, *Frequency*, la frequenza degli stessi, e *Severity*, la differenza tra i pagamenti portati a termine e quelli mancati.
- ***L'ammontare dei debiti [30%]***: sia dovuti che in corso.
- ***Intervallo di tempo della Storia dei Crediti [15%]***: variabile indirizzata a capire l'esperienza che il soggetto possiede nel sistema dei crediti. Più l'esperienza risulta durevole più la variabile esprime risultati significativi.
- ***Nuovi Crediti e inchieste [10%]***: per inchieste si intendono quelle a valenza quantitativamente importante. Sono esclusi infatti i casi in cui si tratta di semplici richieste di informazioni aggiuntive da parte della banca.
- ***Numero di conti e tipi di credito [10%]***: Ultima ma non meno importante è la varietà con cui il soggetto si impegna in termini finanziari. Chiaramente si cerca nuovamente di riconoscere l'esperienza dell'agente.

Inoltre, è interessante notare come invece molte variabili non rientrino nel modello citato, ad esempio:

- *Informazioni sull'occupazione*
- *Residenza*
- *Nucleo familiare*
- *Inchieste Secondarie*
- *Tassi d'interesse sul conto*
- *Riscossioni inferiori ai 100$*

Vantage Score

Il Vantage Score è un alternativo sistema di credit scoring e, negli ultimi anni, è riuscito a farsi valere tra i modelli più utilizzati.
L'ultima versione (3.0) usa dei voti rappresentativi da A a F e riprende sia la scala del FICO, 300 - 850, sia alcune delle variabili principali. Nonostante ciò, si distingue dal FICO sotto vari punti di vista sia di peso dei fattori sia nella selezione degli stessi, i quali vedremo nel dettaglio.

Innanzitutto, il Vantage Score si costituisce come:

- ***Storia dei pagamenti [Estremamente Influente]***: come nel FICO, è l'andamento dei pagamenti registrati sul conto in esame.
- ***Età e tipologia dei Crediti [Molto Influente]***: la storia dei crediti e quanta varietà esiste tra i diversi conti del soggetto.
- ***Utilizzazione dei Crediti [Molto Influente]***: è un ratio calcolato mettendo a rapporto l'attuale saldo e il credito disponibile. E' consigliato a livello di punteggio di tenerlo al di sotto del 30%.
- ***Saldi Totali [Moderatamente Influente]***: sia legittimi che illegittimi (derivanti da debiti non ancora pagati).
- ***Comportamento recente [Poco Influente]***: la gravità delle inchieste ricevute negli ultimi mesi.
- ***Quantità di Crediti disponibili [Lievemente Influente]***

Naturalmente, anche il VS non guarda ad informazioni personali quali la residenza, l'occupazione, la religione etc.

Ora, le principali differenze con il FICO riguardano il peso delle inchieste, le quali vengono "deduplicate" in maniera differente. Infatti, multiple inchieste su un tipo unico di credito vengono raggruppate in una singola inchiesta. Però, FICO si interessa ad un intervallo di tempo di 45 giorni, mentre VS di 14 giorni. Inoltre, VS utilizza lo stesso trattamento su tutti i tipi di crediti, mentre FICO solo su mutui e prestiti per auto e studio. Queste discrepanze non creano cambiamenti sostanziali nel voto, anche se tuttavia, essendo il punteggio diffuso in tutta la nazione, è chiaro che uno stacco di pochi punti possa comunque fare la differenza. E' infine importante notare che il VS non mette a completa disposizione pubblica le percentuali usate per le variabili. Quelle che risultano dalla tabella sono infatti indotte da esempi reali.

Plus Score

Infine, il Plus Score da Experia si distingue dagli altri non in termini di variabili (i quali pesi non sono pubblicati) ma in termini di Target. Questo indice è infatti usato esclusivamente dagli individui, presupponendo un'interesse alla loro situazione finanziaria. Gli agenti bancari non ne fanno uso e normalmente lo correlano fortemente con lo score FICO.

Nonostante ciò, la distribuzione del Plus è leggermente più simmetrica di quella del FICO, probabilmente a causa dell'influenza incognita di certe variabili.

Tabella 2: Tabella di Comparazione

	FICO	**Vantage (v3)**	**Plus**
Minimo	350	350	330
Massimo	850	850	830
Mediana	723	NA	724
CRAs	EX, EQ & TU	EX, EQ, TU	EX
Disponibile agli Individui	Si	Si	Si
Usato dalle Banche	Si	Si	No
	Variabili		
Storia dei Pagamenti	35%	28%	NA
Utilizzo dei Crediti	30%	23%	NA
Storia dei Crediti	15%	9%	NA
Tipo di Crediti	10%	0%	NA
Nuovi Crediti	10%	30%	NA
Saldo	0%	9%	NA
Crediti Disponibili	0%	1%	NA

4 Simulazione

4.1 Introduzione

Avendo esaminato e spiegato il rapporto tra l'informazione economica ed i Big Data, vedremo ora come questi concetti si evolvano in una situazione reale.

Utilizzeremo un dataset estrapolato dal "*Lending Club*", una banca peer-to-peer americana concentrata su *prestiti di tipo familiare* e a fini *consumistici*.
Ci concentreremo principalmente su 2 Obiettivi:

1. Definire la Probabilità di Default

2. Analizzare il comportamento della banca e il Credit Rationing.

Il primo obiettivo verrà affrontato tramite un'analisi di default basata sulle variabili significative a disposizione. Il secondo è invece indirizzato ad osservare come la banca si ponga diversamente nei confronti delle diverse tipologie di agenti, come teorizzato in parte nel Modello di Stiglitz & Weiss.
La scelta della tipologia bancaria è giustificata dalla grande quantità di informazioni facilmente reperibili. Di fatti, ci troviamo di fronte a 887.379 osservazioni a 75 Variabili ciascuna. Per quanto (parzialmente) strutturati, una tale mole dati ricade nella definizione di Big Data.

Lo strumento usato per l'analisi e l'esplicazione del modello è il linguaggio di programmazione *R*.

4.2 Statistiche Descrittive

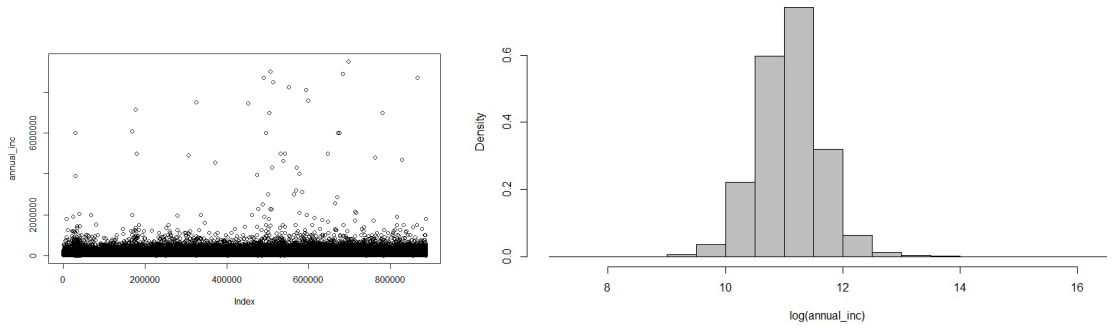
Dopo un *data cleaning* volto alla pulizia dalle variabili parziali e/o fuorvianti per considerare solo le osservazioni rappresentative, otteniamo 15 Variabili analizzabili.

Esse riguardano:

- Reddito Annuale *(annual_ inc)*

  Per la variabile reddito ho ritenuto necessario mettere a confronto il grafico a dispersione (a) con l'istogramma del logaritmo (b). Possiamo infatti notare come il reddito dei soggetti presi in considerazione si distribuisca in maniera concentrata tra $e^{10} < x < e^{12}$.

  A causa di questa omogeneità presente sul campione, il dato non gode di un'alta rappresentatività. Nonostante ciò, è comunque interessante osservarlo in relazione alle altre variabili.



(a) (b)

- Delinquenza negli ultimi due anni *(delinq_ 2yrs)*

  Anche la variabile delinquenza appare concentrata in 0 con forte asimmetria e coda a destra. Tuttavia, l'effetto sulla rischiosità del soggetto è decisamente più rilevante rispetto al reddito annuale.
  Di conseguenza, la terremo in considerazione.

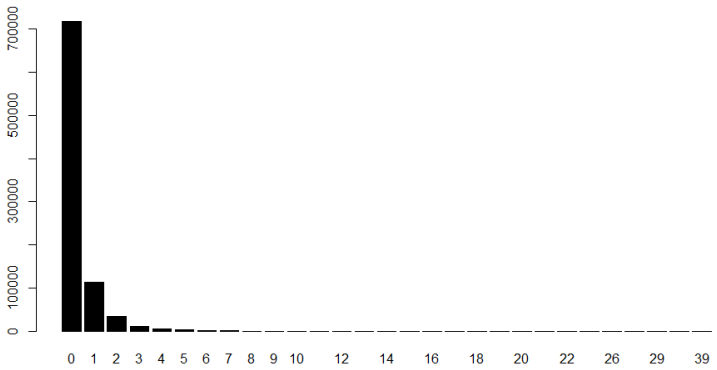


Figura 5: Delinquenza *(ultimi 2 anni)*

- Occupazione *(emp_ length)*

  La variabile originaria è divisa in varie classi per anni. Data l'omogeneità delle frequenze sotto la moda dei +10 anni ho deciso di creare una variabile binaria *emp*, la quale:

$$emp = \begin{cases} 1, & \text{if } emp\_lenght = 10+ \text{ years} \\ 0, & \text{else} \end{cases}$$

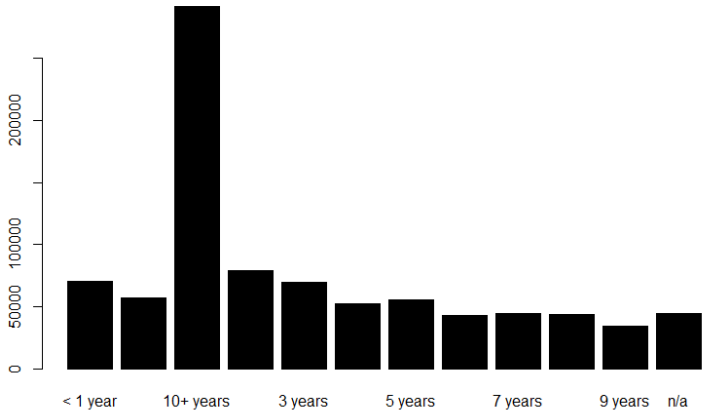


Figura 6

- Quantità già ripagata *(funden_ amnt)*

  La variabile non è direttamente rappresentativa al livello di rischiosità dell'agente. Tuttavia, la precisione ed esattezza temporale con la quale è formata la rendono un peso necessario nella formazione delle nostre regressioni sul dataset.
  Inoltre, è interessante notare come incida indirettamente sulla variabile dipendente.

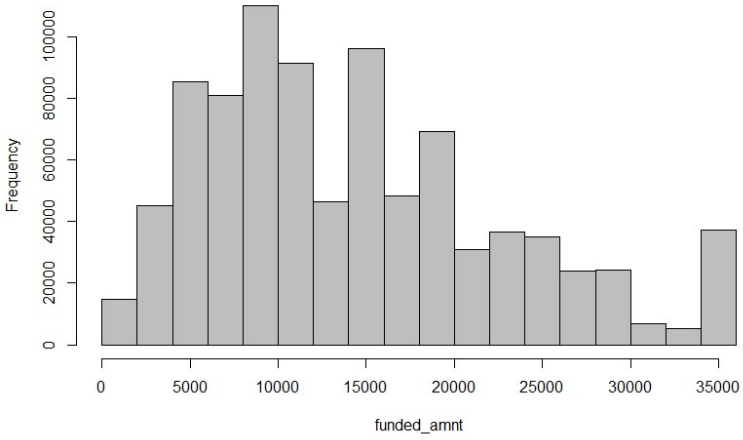


Figura 7

- Proprietà di Residenza *Home Ownership*:

  Forse una delle variabili categoriche più rappresentative, essa si consolida come la tipologia di finanziamento per l'abitazione dell'agente.
  Per motivi esplicativi, anch'essa rientra tra le variabili da me reindirizzate in forma dicotomica:

$$home = \begin{cases} 1, & \text{if } home\_ownership = Mortgage \\ 0, & \text{else} \end{cases}$$

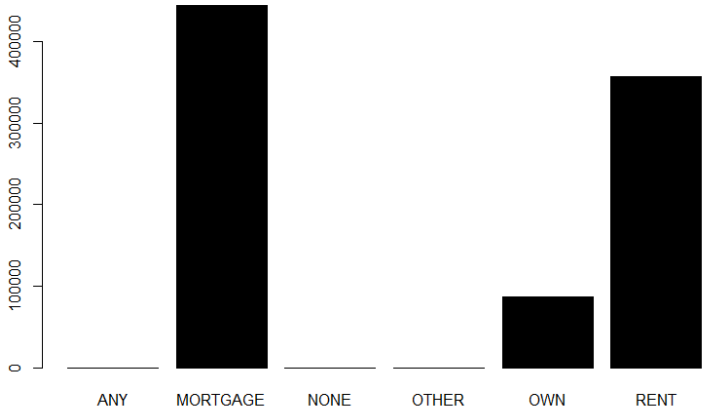


Figura 8

- Tasso d'interesse *(int_rate)*

  Il tasso è fondamentale non tanto per l'analisi di default.
  Esso fungerà in un'analisi successiva come termine di paragone tra il nostro caso e il razionamento del credito definito dal modello di Stiglitz & Weiss.

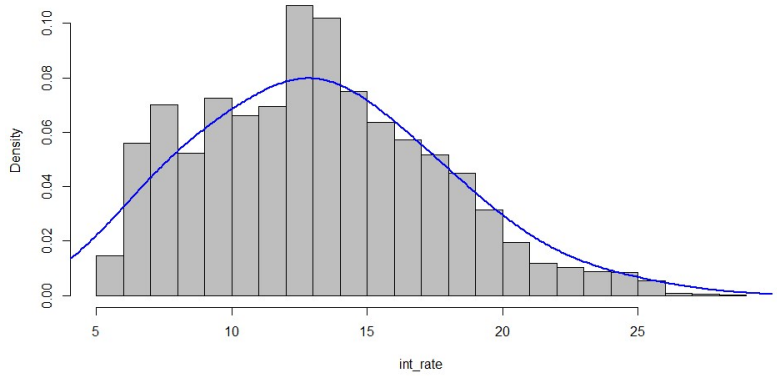


Figura 9

- Ultimo Pagamento *(last_ pymnt_ amnt)*

  Questa Variabile fa parte del gruppo delle meno rappresentative ed omogenee.
  Tuttavia, la rappresentazione grafica logaritmica osserva un' interessante asimmetria con coda a destra.

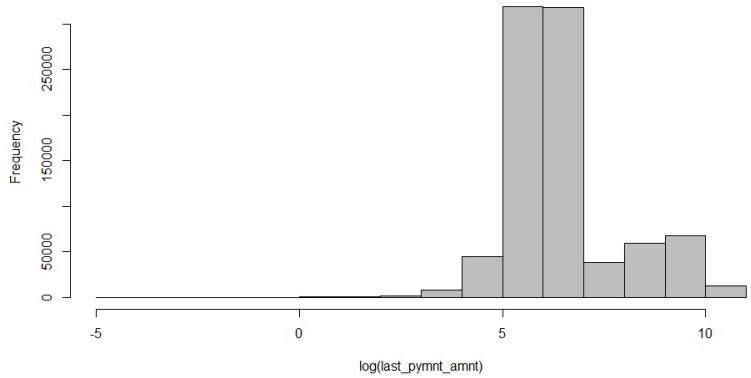


Figura 10

- Ammontare del Prestito *(loan_ amnt)*

  Per questa variabile il nostro interessa di sofferma sulla direzione della probabilità di default, in quanto l'intuizione non basta nello stabilire se la quantità prestata sia un vantaggio o uno svantaggio.

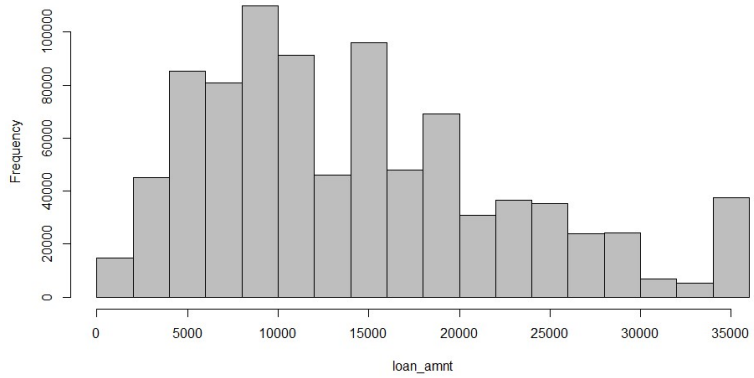


Figura 11

- Stato del Prestito *(loan_status)*

  Questo è il fattore da cui deriviamo la nostra *Variabile Dipendente.*
  Importante, nella nostra analisi consideriamo un ritardo nel pagamento alla stessa stregua di un default:

$$default = \begin{cases} 1, & \text{if } loan\_status = Default \text{ or if } loan\_status = Late\ XX\ days \\ 0, & \text{else} \end{cases}$$

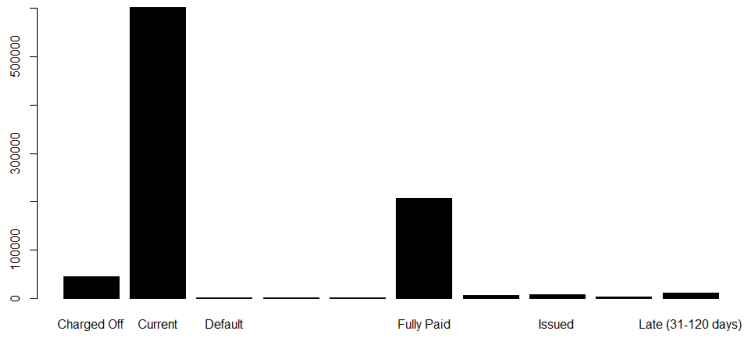


Figura 12

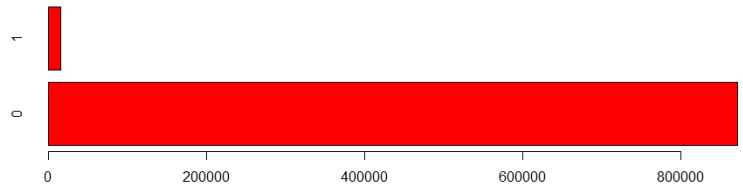


Figura 13: **default**

- Pagamento rimanente *(out_ prncp)*

  La forte asimmetria di questa variabile e la sua varietà alla fonte la rende rappresentativa.
  E' importante annotare che il motivo per cui la moda è 0 riguarda anche gli agenti ai quali è stato negato il prestito.

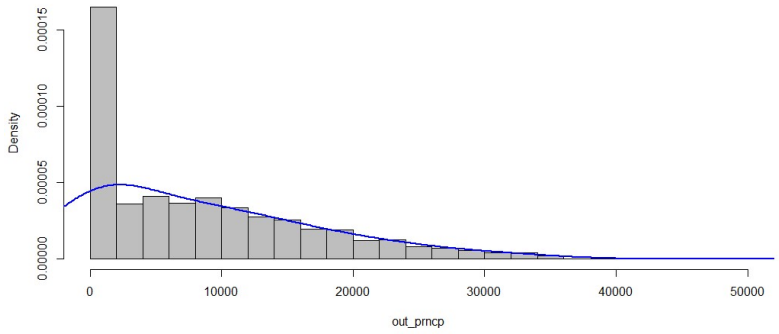


Figura 14

- Tasso di utilizzazione del credito *(revol_ util)*

  Questa variabile quantifica il comportamento degli agenti riguardo all'utilizzo del credito ricevuto.
  Inoltre, espressa in forma logaritmica evidenzia delle interessanti frequenze anomale da analizzare nel modello.

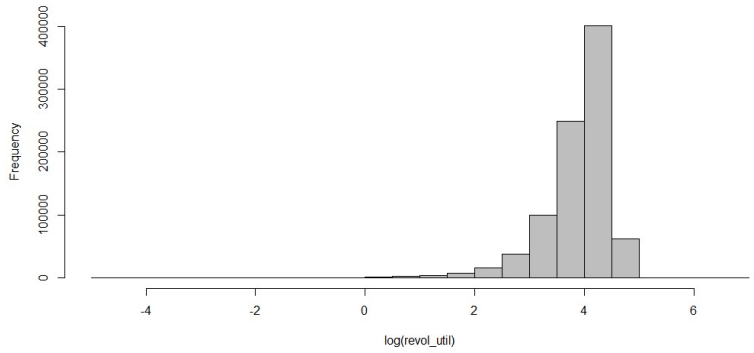


Figura 15

- Termine *(term)*

  La variabile è categorica e dicotomica.

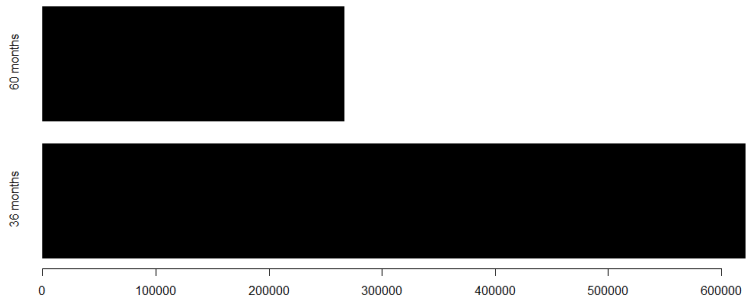
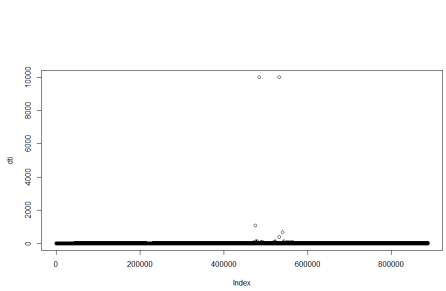


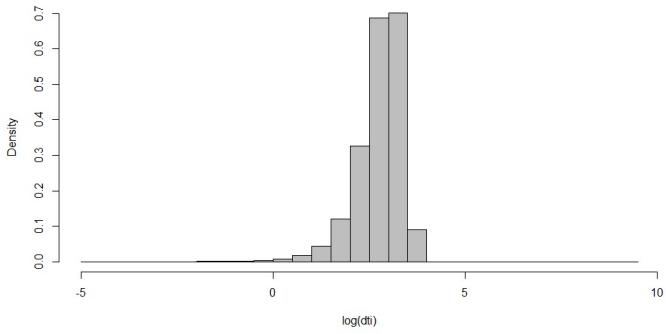
Figura 16

- Rapporto Debito / Reddito *(dti)*

  I due grafici ci indicano come sebbene questa variabile sia ben distribuita su tutte le osservazioni, essa rappresenti un valore significativo solo per alcune di queste.
  Comunque, in forma logaritmica mostra in maniera più esplicita i valori meno frequenti.



(a)



(b)

- Pagamenti ricevuti *(total_ pymnt)*

  Variabile fortemente asimmetrica verso destra.
  Da notare come circa il 30% delle osservazioni si concentri in 0.

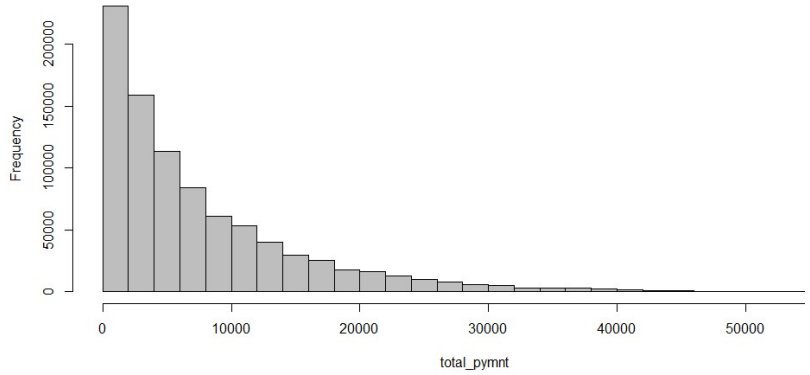


Figura 18

- Accertata delinquenza del conto *(acc_ now_ delinq)*

  Questa variabile indica *il numero* di accertate delinquenze sul conto degli agenti.
  Essa è volta all'analisi comportamentale dei clienti in quanto valuta, nel nostro modello, la probabilità che *anche* il conto in questione vada in default.

4.3 Analisi

I Modelli

A seguito della definizione degli obiettivi è necessario scegliere una metodologia di approccio all'analisi.
Come abbiamo osservato nelle statistiche descrittive, ho deciso di definire una variabile binaria "*default*", la quale prende valore "1" in caso di ritardo o fallimento del pagamento e "0" altrimenti (seguendo l'andamento della variabile originaria *loan_status*).
Di conseguenza, avendo un campione sufficientemente rappresentativo, possiamo vedere l'intervallo:

$$0 < default < 1$$

come la probabilità di *non* ripagare il debito, la quale assume ora accezione *predittiva*.

Ora abbiamo bisogno di un modello in grado di spiegare una variabile compresa tra 0 e 1, ossia una probabilità.
Potremmo usare un modello lineare, ma la geometria della retta tende a risultare in valori non compresi tra 0 e 1, i quali sarebbero non interpretabili in termini statistici (ad esempio una probabilità uguale a -0.7).
Dunque, vogliamo delle tipologie di regressione non lineari tali da non ammettere valori della variabile dipendente al di fuori dell'intervallo.

A nostra disposizione si presentano due modelli (equivalenti): il *Modello Probit,* e il *Modello Logit (o Regressione Logistica).*

Il modello *Probit* si definisce come:

$$P(Y_i|X_i) = \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_n X_{ni} + u_i) \tag{6}$$

Dove $\Phi(x)$ si riferisce alla Funzione di distribuzione cumulativa (cdf) della normale:

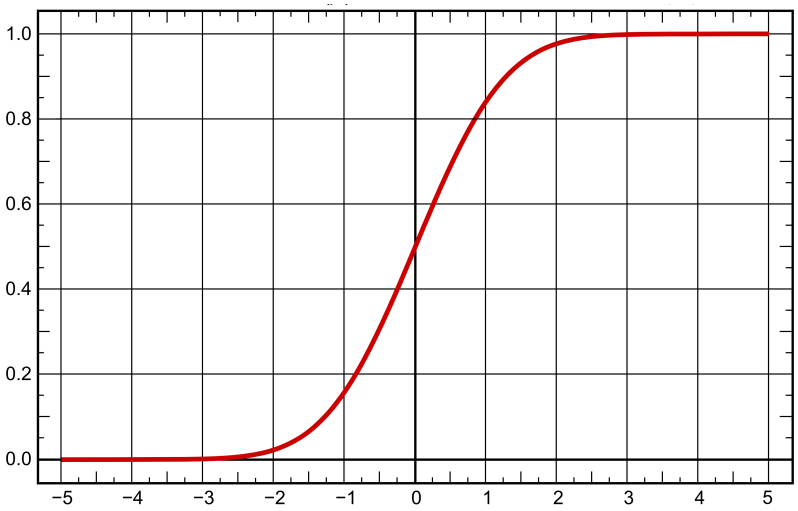


Figura 19: Distribuzione Cumulativa Normale

Dato che la funzione (continua) cumulativa di probabilità è *per definizione* compresa tra 0 e 1, il modello risponderà come voluto al nostro input.

Il modello *Logit* si definisce invece come:

$$P(Y_i|X_i) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_{1i}+...+\beta_n X_{ni}+u_i)}} \tag{7}$$

Graficamente:

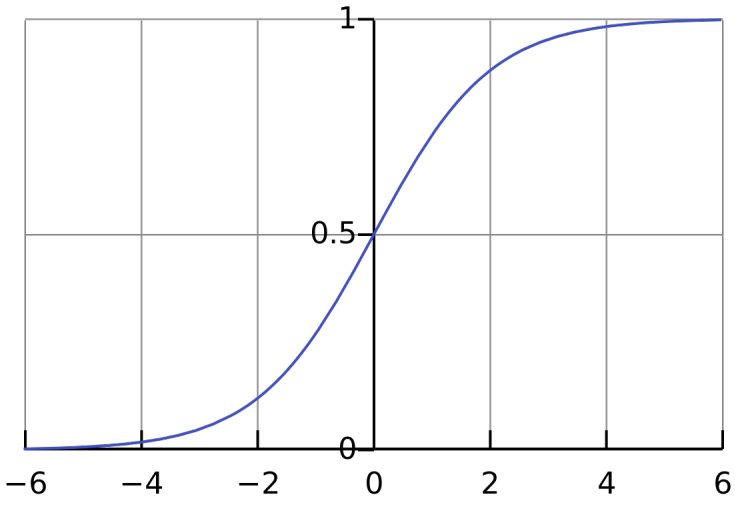


Figura 20: Logistic Curve

Il logit è infatti una funzione con dominio definito tra 0 e 1.

Analiticamente è il logaritmo del rapporto di probabilità (*odds*):

$$logit(p) = ln(\frac{p}{1-p})$$

La funzione inversa è invece la nostra funzione logistica, la quale invertendo la logit *ottiene codominio compreso tra 0 e 1*(dato che la variabile del logit è p) avendo come input un qualsiasi numero α :

$$logit^{-1}(\alpha) = logistic(\alpha) = \frac{1}{1+e^{-\alpha}}$$

Inoltre, prima di presentarvi i risultati è importante notare come non sia possibile dare immediata interpretazione marginale ai coefficienti dei regressori, come invece accade nel modello lineare. Infatti, i β_i contengono informazioni solo sulla *direzione* dell'effetto.

Ciononostante, potremo osservare gli effetti in termini quantitativi tramite il calcolo della media degli *Effetti Marginali*, i quali tengono conto della geometria non lineare del modello.

Tabella 3: **Logit Model**

	Dependent variable:
	default
annual_inc	−0.00000***
	(0.00000)
int_rate	0.109***
	(0.003)
delinq_2yrs	0.055***
	(0.008)
dti	0.0002
	(0.0002)
emp1	−0.120***
	(0.018)
funded_amnt	−0.0002**
	(0.0001)
home1	−0.147***
	(0.017)
installment	0.003***
	(0.0002)
last_pymnt_amnt	−0.0005***
	(0.00002)
loan_amnt	−0.0001*
	(0.0001)
out_prncp	0.0002***
	(0.00001)
revol_util	−0.0003
	(0.0004)
term 60 months	−0.066
	(0.046)
total_pymnt	0.0001***
	(0.00000)
acc_now_delinq	0.105
	(0.075)
Constant	−5.216***
	(0.043)
Observations	886,877
Log Likelihood	−69,748.910
Akaike Inf. Crit.	139,529.800
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 4: **Probit Model**

	Dependent variable:
	default
annual_inc	−0.00000***
	(0.00000)
int_rate	0.046***
	(0.001)
delinq_2yrs	0.024***
	(0.003)
dti	0.0002*
	(0.0001)
emp1	−0.051***
	(0.008)
funded_amnt	−0.0001***
	(0.00003)
home1	−0.060***
	(0.007)
installment	0.001***
	(0.0001)
last_pymnt_amnt	−0.0002***
	(0.00001)
loan_amnt	−0.0001*
	(0.00003)
out_prncp	0.0001***
	(0.00000)
revol_util	−0.0002
	(0.0002)
term 60 months	−0.011
	(0.019)
total_pymnt	0.0001***
	(0.00000)
acc_now_delinq	0.053
	(0.035)
Constant	−2.623***
	(0.018)
Observations	886,877
Log Likelihood	−69,639.770
Akaike Inf. Crit.	139,311.500
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 5: **Effetti Marginali *(Logit)***

logitmfx(formula = glm1, data = dataset, atmean = FALSE)
Marginal Effects:

	dF/dx	Std. Err.	z	P >|z|	
annual_inc	**-0.0000000174787**	0.0000000035791	-4.8836	0.000001041884191	***
int_rate	**0.0017849314452**	0.0000474543928	37.6136	<2.2e-16	***
delinq_2yrs	**0.0008964384700**	0.0001309533715	6.8455	0.000000000007622	***
dti	**0.0000040843846**	0.0000026251796	1.5558	0.11974	
emp1	**-0.0019295630358**	0.0002903848809	-6.6448	0.000000000030353	***
funded_amnt	**-0.0000028990674**	0.0000012347580	-2.3479	0.01888	*
home1	**-0.0024015532988**	0.0002842567433	-8.4485	<2.2e-16	***
installment	**0.0000542238947**	0.0000037427560	14.4877	<2.2e-16	***
last_pymnt_amnt	**-0.0000079826865**	0.0000003268610	-24.4223	<2.2e-16	***
loan_amnt	**-0.0000021544106**	0.0000012223561	-1.7625	0.07798	.
out_prncp	**0.0000037194530**	0.0000001036794	35.8746	<2.2e-16	***
revol_util	**-0.0000053908083**	0.0000060088301	-0.8971	0.36964	
term 60 months	**-0.0010732451457**	0.0007442372729	-1.4421	0.14928	
total_pymnt	**0.0000024302737**	0.0000000707283	34.3607	<2.2e-16	***
acc_now_delinq	**0.0017341600457**	0.0012283500839	1.4118	0.15801	

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
dF/dx is for discrete change for the following variables:
[1] "emp1" "home1" "term 60 months"

Tabella 6: **Effetti Marginali *(Probit)***

probitmfx(formula = glm2, data = dataset, atmean = FALSE)
Marginal Effects:

	dF/dx	Std. Err.	z	P >|z|	
annual_inc	**-0.0000000144933**	0.0000000034265	-4.2298	0.000023393798397	***
int_rate	**0.0017834769460**	0.0000466857318	38.2018	<2.2e-16	***
delinq_2yrs	**0.0009343843750**	0.0001364612527	6.8473	0.000000000007528	***
dti	**0.0000064715915**	0.0000034460021	1.8780	0.0603812	.
emp1	**-0.0019464492358**	0.0002888047783	-6.7397	0.000000000015875	***
funded_amnt	**-0.0000035338437**	0.0000010468476	-3.3757	0.0007363	***
home1	**-0.0023511025367**	0.0002827345543	-8.3156	<2.2e-16	***
installment	**0.0000568824368**	0.0000037561656	15.1438	<2.2e-16	***
last_pymnt_amnt	**-0.0000063159547**	0.0000002496195	-25.3023	<2.2e-16	***
loan_amnt	**-0.0000019962612**	0.0000010324848	-1.9335	0.0531804	.
out_prncp	**0.0000040712993**	0.0000001059264	38.4352	<2.2e-16	***
revol_util	**-0.0000094266943**	0.0000059777720	-1.5770	0.1148052	
term 60 months	**-0.0004331900714**	0.0007541968353	-0.5744	0.5657156	
total_pymnt	**0.0000027589782**	0.0000000753679	36.6068	<2.2e-16	***
acc_now_delinq	**0.0020674820089**	0.0013588720254	1.5215	0.1281421	

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
dF/dx is for discrete change for the following variables:
[1] "emp1" "home1" "term 60 months"

4.4 Risultati

Come anticipato, i due modelli convergono agli stessi risultati.

Possiamo ora estrapolare l'interpretazione più adatta per ogni fattore.
Inoltre, per non creare confusione, definirò un effetto negativo come "meno rischioso" (meno probabilità di non ripagare) e viceversa.

Facendo riferimento alle Tabelle 3-6, i fattori responsabili di una *diminuzione* della rischiosità di un individuo sono:

- Il Reddito annuale (*annual_inc*)

  Come ci aspettavamo, il reddito annuale non ha un effetto importante nel modello, probabilmente dovuto alla forte omogeneità dello stesso tra le osservazioni.
  Tuttavia, esso tende ad avere un leggero effetto negativo, ossia all'aumentare del reddito la probabilità di non ripagare il debito decresce.
  Da ciò potremmo inferire che, in generale, il reddito ha un effetto positivo sulla rischiosità degli agenti, il che non è una particolare sorpresa.

- Quantità già ripagata *(funded_amnt)*,
  Ammontare dell'ultimo pagamento*(last_pymnt_amnt)*
  Ammontare del Prestito *(loan_amnt)*

  Al crescere di queste tre variabili si ha un comune (modesto) effetto nella riduzione del rischio.
  La quantità ripagata e l'ammontare dell'ultimo pagamento confermano l'intuizione qualitativa.
  Tuttavia, come mai il tasso di utilizzazione e l'ammontare del prestito tendono ad avere questo effetto?
  L'ammontare risulta variabile non significativa al livello del 5% ($P > 0.05$), mentre il tasso è non significativa al 10% (modello Probit); quindi non possiamo rigettare l'ipotesi che esse abbiano effetto pari a 0.
  Ma, intuitivamente, una cifra più alta da ripagare e/o il consumo più alto della stessa dovrebbero deteriorare la probabilità di pagamento. Invece, dalla nostra analisi risulta lievemente l'opposto (o un effetto nullo).
  Se la banca trattasse prestiti di tipo societario potremmo dedurne che l'utilizzazione del credito sia correlata ad investimenti generalmente profittevoli, capaci di aumentare le speranze di ripagamento.
  Ma il Lending Club è indirizzato fortemente al consumo.
  Una possibile interpretazione potrebbe quindi riguardare il fatto che chi consuma di più si trova, di norma, in una posizione di stabilità e sicurezza finanziaria.
  In ogni caso, se vogliamo dedurre informazioni più specifiche, abbiamo bisogno di più dati.

- Occupazione *(emp)*,
  Finanziamento della Casa *(home)*,

Scadenza *(term)*.

Similmente al caso precedente, l'occupazione superiore a 10 anni o una scadenza più vantaggiosa (60 mesi invece che 36) ricadono nell'intuizione e non hanno bisogno di particolare spiegazione. Importante notare però che l'effetto dell'occupazione è il secondo più forte sulla diminuizione del rischio ($\approx 0.2\%$)

Il primo è il finanziamento della casa ($\approx 0.23\%$), in caso esso avvenga tramite mutuo.
Le interpretazioni possono essere molteplici.
Probabilmente, i due motivi principali per cui ciò si verifica sono:

- *Esperienza Finanziaria*: come nel calcolo dei Credit Score, l'esperienza nel finanziarsi e la storia dei crediti sul conto sono indici positivi sulla probabilità di default.
- *Avversione al Rischio*: in caso il mutuo sia ancora in corso, raramente una famiglia si impegnerebbe in un ulteriore debito se non fortemente rassicurata da altri fattori. Ciò è una diretta conseguenza di una generale avversione al rischio degli agenti volti al consumo.

Passiamo ora a valutare i fattori che vanno invece ad *aumentare* la probabilità di default e la conseguente rischiosità:

- Tasso d'interesse (*int_rate*)

  Questa variabile non può essere interpretata come significativa in quanto la causalità con la variabile dipendente è biunivoca. Essa sarà protagonista del prossimo paragrafo incentrato sul comportamento della banca e sul razionamento del credito.

- Delinquenza (*delinq_2yrs* e *acc_now_delinq*)

  La delinquenza è stata misurata in due modi: crimini negli ultimi 2 anni e conti considerati delinquenti.
  Entrambi hanno effetti forti sulla capacità di ripagare da parte di un individuo, e la delinquenza per conto ha il valore più alto relativo all'aumento del rischio ($\approx 0.2\%$).

- Pagamenti ricevuti (*total_pymnt*)
  Pagamento rimanente (*out_prncp*)

  Troviamo un effetto inaspettato riguardo la variabile sui pagamenti ricevuti.
  Ci saremmo infatti aspettati un effetto di diminuzione del rischio e invece ci ritroviamo di fronte a un leggero effetto di aumento della probabilità di default.
  La variabile è decisamente significativa con un p-value prossimo allo 0.
  Senza ulteriori analisi potremmo riflettere sul fatto che il 30% delle osservazioni ha un totale di 0 pagamenti ricevuti, i quali probabilmente hanno ripagato completamente il debito. Di conseguenza, sono presi in considerazione solo coloro il quale debito è ancora in corso, tra cui quelli in ritardo (i quali sono considerati come default nella nostra analisi).

Pertanto, senza andare più a fondo e data la lievità della variabile, direi che il risultato è un'effetto collaterale delle assunzioni del modello.

4.5 Tasso d'interesse come strumento per il Credit Rationing

In precedenza abbiamo discusso come gli intermediari finanziari facciano largo uso di meccanismi quali i *Credit Score*.
Quindi, ci interessa ora esaminare il comportamento della banca nei confronti dei clienti da lei considerati più rischiosi.

Il dataset ci fornisce il punteggio FICO e Vantage Score per una porzione più che rappresentativa delle osservazioni, riassunti poi sotto il nome di "*grade*".
La variabile *grade* è categorica ed espressa in lettere dalla A(punteggio più alto) alla G, le quali corrispondono ai punteggi numerici degli score.
Osserveremo come il tasso d'interesse vari al variare del punteggio e come il credito venga razionato sistematicamente.

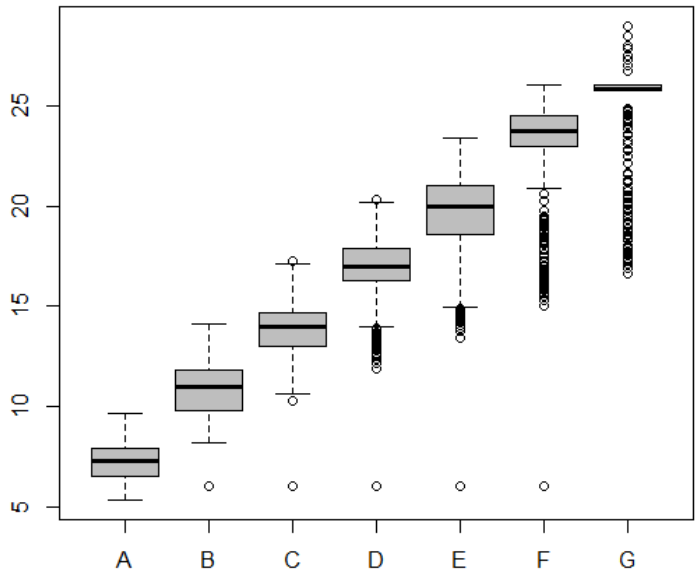


Figura 21

Il box-plot ci permette di notare come il tasso d'interesse mediano cresca quasi *linearmente* al diminuire del voto (o del punteggio) con le dovute eccezioni nel caso dei voti più bassi *(figura 21)*.
Per quantificare questa variazione costruiamo un modello lineare:

$$int\_rate = \beta_0 + \beta_1 D_B + \beta_2 D_C + \beta_3 D_D + \beta_4 D_E + \beta_5 D_F + \beta_6 D_G$$

dove:

$$D_X = \begin{cases} 1 & \text{if } grade = X \\ 0 & \text{else} \end{cases}$$

e D_A è omessa per evitare conflitti con l'intercetta (*dummy variable trap*)

Tramite la nostra regressione lineare *(Tabella 7)* scopriamo come il tasso cresca di circa 3 punti percentuali per ogni voto. Questa variazione sistematica è quindi intenzione diretta della banca, la quale sta utilizzando il tasso d'interesse come *strumento* per controllare gli agenti rischiosi.
Di conseguenza, *non possiamo utilizzare il tasso come indice di rischiosità.*
Ossia, non possiamo affermare che l'aumentare del tasso *causi* un aumento della probabilità di default (anche se sia il modello precedente che l'intuizione ci dicono il contrario).
Infatti, il tasso d'interesse è impostato dalla banca come *conseguenza* della stimata rischiosità degli individui, calcolata in maniera simile al nostro modello di default.
Non c'è quindi dubbio che ci sia *correlazione* tra agenti non paganti e tassi alti, ma la *causalità* è *invertita.*

Quindi, abbiamo appena dimostrato con dati reali quello che Stiglitz & Weiss spiegavano analiticamente nel loro modello sul Credit Rationing.
In un' economia dove l'informazione è imperfetta, le banche utilizzano le risorse a loro disposizione per individuare gli agenti più o meno rischiosi (contro la Selezione Avversa).
Arrivati a conclusioni statistiche attendibili, le banche utilizzano il tasso d'interesse (e le garanzie collaterali) come strumento per schermarsi dai più pericolosi.

Così facendo stanno tenendo fuori una porzione di Domanda (nel modello S&W chiamata **Z**) la quale però gli permette di massimizzare i profitti attesi.

Tabella 7: Effetto dei Credit Score sui Tassi d'interesse

	Dependent variable:
	int_rate
gradeB	3.586***
	(0.004)
gradeC	6.737***
	(0.004)
gradeD	9.933***
	(0.005)
gradeE	12.654***
	(0.006)
gradeF	16.339***
	(0.009)
gradeG	18.383***
	(0.018)
Constant	7.243***
	(0.003)
Observations	887,379
R^2	0.912
Adjusted R^2	0.912
Residual Std. Error	1.297 (df = 887372)
F Statistic	1,540,170.000*** (df = 6; 887372)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.6 Tabelle

Variabili	***Descrizione***
annual_inc	The self-reported annual income provided by the borrower during registration.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
funded_amnt	The total amount committed to that loan at that point in time.
home_ownership	Mortgage, Rent etc...
int_rate	Interest Rate on the loan
last_pymnt_amnt	Last total payment amount received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
out_prncp	Remaining outstanding principal for total amount funded
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_pymnt	Payments received to date for total amount funded
acc_now_delinq	The number of accounts on which the borrower is now delinquent.

Altre Variabili	***Descrizione***
Risk_Score	For applications prior to November 5, 2013 the risk score is the borrower's FICO score. For applications after November 5, 2013 the risk score is the borrower's Vantage score.
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
grade	LC assigned loan grade
sub_grade	LC assigned loan subgrade

5 Conclusioni

All'inizio di questa ricerca ci siamo chiesti quali conseguenze avessero i Big Data sullo studio dell'informazione nel mercato del credito. Dopo aver selezionato un modello di riferimento, abbiamo dimostrato come sia possibile arrivare a risultati di rilevanza sia pratica che teorica partendo da archivi dati disponibili pubblicamente.
I dati a disposizione sono stati raccolti da una banca peer to peer americana volta al servizio di prestiti di tipo consumistico e familiare.
Dopo aver ripulito il dataset dalle osservazioni non rappresentative e dalle variabili non correlate con la v. dipendente, ci siamo concentrati sulla quantificazione dell'effetto dei fattori significativi sulla probabilità di default.
Nello specifico, usando strumenti algoritmico-statistici, abbiamo presentato un possibile modello di previsione del rischio e interpretato in quest'ottica ognuna delle variabili con i relativi effetti marginali.
Infine, abbiamo realizzato come il tasso d'interesse fosse usato dalla banca come strumento per schermarsi dai clienti più rischiosi, collegandoci dall'analisi empirica alla teoria sull'informazione e sul razionamento del credito, costruendo quindi un ponte tra la teoria economica e l'economia reale.

Il nostro approccio alla soluzione è però solo uno tra i tanti possibili.
Un'analisi differente e più tecnica poteva ad esempio prevedere l'applicazione di strumenti quali le reti neurali artificali (*Applying artificial neural networks to bank-decision simulations (Witkowska, 1999)*) o altri metodi derivanti dal machine learning.
Inoltre, gli intermediari finanziari non sono solo interessati all'applicazione di questi metodi nel campo dei crediti.
La possibilità di minimizzare l'informazione imperfetta è estremamente preziosa anche in sede di risk management e banche d'investimento. Questo lato merita infatti di essere studiato più a fondo mettendo a confronto la teoria economico-quantitativa dei mercati finanziari con l'analisi dei dati d'investimento.

Tuttavia, una domanda sorge allora spontanea: se l'analisi dei Big Data è in continua evoluzione e così versatile, che bisogno si ha di continuare a lavorare sulla teoria?
La risposta è implicita nella presente ricerca. I metodi analitici non avranno mai una funzione talmente esaustiva da rendere la teoria obsoleta. Anzi, grazie a questi nuovi strumenti, saranno proprio gli economisti e i ricercatori accademici ad iniziare un processo di rivalutazione ed evoluzione dell'economia e della finanza, ridefinendo i modelli esistenti sotto un'ottica empirica più accurata.

Riferimenti bibliografici

[1] STIGLITZ, J. E. E WEISS, A.
CREDIT RATIONING IN MARKETS WITH IMPERFECT INFORMATION, *The American Economic Review, Volume 71, Issue 3 (Jun. 1981), 393-410*

[2] VARIAN, H. R.
Intermediate microeconomics
In-text: (Varian, 2003)
Bibliografia: Varian, H. (2003). Intermediate microeconomics. 1st ed. New York ; London: W. W. Norton & Company.

[3] EINAV, L. E LEVIN, J.
Economics in the age of big data
In-text: (Einav and Levin, 2014)
Bibliografia: Einav, L. and Levin, J. (2014). Economics in the age of big data. Science, 346(6210), pp.1243089-1243089.

[4] STIGLITZ, J. E. E WEISS, A.
Asymmetric Information in Credit Markets and its Implications for Macro - Economics
In-text: (STIGLITZ and WEISS, 1992)
Bibliografia: STIGLITZ, J. and WEISS, A. (1992).
Oxford Economic Papers, 44(4), pp.694-724.

[5] EMILIA BONACCORSI DI PATTI & PAOLO FINALDI RUSSO
Questioni di Economia e Finanza: Fragilità delle imprese e allocazione del credito
Banca d'Italia
Online: https://www.bancaditalia.it/pubblicazioni/qef/2017-0371/QEF_371.pdf

[6] NICOLA, G.
Il ruolo dei big data nel cambiamento del business model delle banche - Statistica e Società
In-text: (Nicola, 2017)
Bibliografia: Nicola, G. (2017). Il ruolo dei big data nel cambiamento del business model delle banche - Statistica e Società. [online] Statistica e Società. Available at: http://www.rivista.sis-statistica.org/cms/?p=59 [Accessed 9 Jun. 2017].

[7] LONGO, A.
A fianco del cliente con big data
In-text: (Longo, 2017)
Bibliografia: Longo, A. (2017). A fianco del cliente con big data. [online] Nova 24. Available at: http://nova.ilsole24ore.com/esperienze/a-fianco-del-cliente-con-big-data/ [Accessed 9 Jun. 2017].

[8] LE FONTI DEI BIG DATA E LE LORO CARATTERISTICHE
In-text: (Vincos, 2017)
Bibliografia: Vincos. (2017). Le fonti dei Big Data e le loro caratteristiche.
Available at: https://vincos.it/2013/11/28/le-fonti-dei-big-data-e-le-loro-caratteristiche/ [Accessed 9 Jun. 2017].

[9] Zerounoweb.it, 2017
Bibliografia: Zerounoweb.it. (2017). Cos'è Big data analytics? Tutto quello che serve sapere sull'analisi dei dati.
Available at: http://www.zerounoweb.it/approfondimenti/big-data/come-fare-big-data-analysis-e-ottenere-valore-per-le-aziende.html Accessed 9 Jun. 2017.

[10] JEFF BERTOLUCCI
Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive - InformationWeek In-text: (Bertolucci, 2017)
Bibliografia: Bertolucci, J. (2017). Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive - InformationWeek. [online] InformationWeek.
Available at: http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279 [Accessed 9 Jun. 2017].

[11] Prescriptive Analytics – Let Me See You Work
In-text: (Community, Blog and Prescriptive Analytics – Let Me See You Work, 2017)
Bibliografia: Community, L., Blog, S. and Prescriptive Analytics – Let Me See You Work, W. (2017). Prescriptive Analytics – Let Me See You Work, Work, Work. [online] Lithium Community.
Available at: https://community.lithium.com/t5/Science-of-Social-Blog/Prescriptive-Analytics-Let-Me-See-You-Work-Work-Work/ba-p/255119 [Accessed 9 Jun. 2017].

[12] Statsoft.com, 2017)
Bibliografia: Statsoft.com. (2017). Credit Scoring, Scorecard, Statistics, Risk Management.
Available at: http://www.statsoft.com/Textbook/Credit-Scoring#Classic_credit_scoring [Accessed 9 Jun. 2017].

[13] BRINKLEY-BADGETT
What Does FICO Stand For? What is a FICO Score?
In-text: (Brinkley-Badgett, 2017)
Bibliografia: Brinkley-Badgett (2017). What Does FICO Stand For? What is a FICO Score?. [online] Credit.com. Available at: https://www.credit.com/credit-scores/what-does-fico-stand-for-and-what-is-a-fico-credit-score/ [Accessed 29 May 2017].

[14] PLUS SCORE - DOCTOR OF CREDIT
In-text: (Doctor Of Credit, 2017)
Bibliografia: Doctor Of Credit. (2017). PLUS Score - Doctor Of Credit. [online] Available at: http://www.doctorofcredit.com/credit-scores/plus-score/ [Accessed 29 May 2017].

[15] WHY THE AGE OF YOUR CREDIT HISTORY MATTERS | Mike Goldstein
In-text: (Creditkarma.com, 2017)
Bibliografia: Creditkarma.com. (2017). Why the age of your credit history matters | Credit Karma. [online] Available at: https://www.creditkarma.com/article/age-of-credit-history [Accessed 29 May 2017].

[16] SIMONETTA, B.
L'app che valuta il rischio del credito
In-text: (Simonetta, 2017)
Bibliografia: Simonetta, B. (2017). L'app che valuta il rischio del credito.
Il Sole 24 ORE. Available at:
http://www.ilsole24ore.com/art/tecnologie/2016-05-25/l-app-che-valuta-rischio-credito-173858.shtml?uuid=ADwSdYP&refresh_ce=1

[17] PAPERNO, B.
FICO v. VantageScore: 5 Differences You Should Understand
In-text: (Paperno, 2017)
Bibliografia: Paperno, B. (2017). FICO v. VantageScore: 5 Differences You Should Understand. [online] Credit.com.
Available at: http://blog.credit.com/2013/01/fico-v-vantagescore-5-differences-you-should-know-64279/ [Accessed 9 Jun. 2017].

[18] CHARLES, W.
FICO Score Vs VantageScore - Doctor Of Credit
In-text: (Charles, 2017)
Bibliografia: Charles, W. (2017). FICO Score Vs VantageScore - Doctor Of Credit. [online] Doctor Of Credit.
Available at: http://www.doctorofcredit.com/fico-score-vs-vantagescore/ [Accessed 9 Jun. 2017].

[19] WITKOWSKA, D.
Applying artificial neural networks to bank-decision simulations In-text: (Witkowska, 1999)
Bibliografia: Witkowska, D. (1999). Applying artificial neural networks to bank-decision simulations. International Advances in Economic Research, 5(3), pp.350-368.