UNIVERSITAT POMPEU FABRA

La potencialidad de Big Data para los servicios financieros

José García Montalvo
Catedrático de Economía

6º Congreso Anual de Crédito y Recobro
Barcelona, 20 de noviembre de 2014

Esquema

- Introducción
- El estado del "big data" y la e-science
- Big data y economía
- La potencialidad de *big data* en los servicios financieros y los seguros
  - Calificaciones crediticias
  - Detección de fraude en tarjetas de crédito
  - Big data y seguros
- Los límites del *big data:* privacidad, consentimiento y correlación
- Conclusiones

Introducción

- “Nos ahogamos en información y, a la vez, estamos hambrientos de conocimiento” John Naisbitt

Introducción

- "Nos ahogamos en información y, a la vez, estamos hambrientos de conocimiento" John Naisbitt

- "*Big data* es como el sexo adolescente: todos hablan de ello, nadie realmente sabe cómo hacerlo, todo el mundo piensa que todos los demás lo están haciendo y, por tanto, todo el mundo asegura que ellos también lo hacen" Dan Ariely

Introducción: el caso Target

- En estos tiempos hay una pelea enorme por contratar matemáticos y estadísticos para los departamentos de "predictive analytics" de las empresas que sepan utilizar todo tipo de algoritmos
- Pole trabajaba justamente en Target's Guest Marketing Analytics department": economista y estadístico
- Pole fue asignado a buscar aquellos momentos únicos en la vida de un consumidor cuando sus hábitos son más flexibles y la adecuada publicidad o cupón le causa comprar en nuevas formas -> el nacimiento de un hijo es uno de esos momentos
- …pero todos lo saben y Target quería evitar llegar tarde (cuando ya ha nacido el niño): mejor momento el segundo trimestre de embarazo

Introducción: el caso Target

- En esos momentos es cuanto más vulnerables son a la publicidad (cambio de casa, nacimiento de un hijo, divorcio, etc.)
- Identificar futuros padres supone unas ganancias millonarias
- Identificaron 25 productos (suplementos de calcio, magnesio, cinc, jabón sin aroma, bolsas grandes de algodón, etc.) que se compran en las primeras 20 semanas para predecir embarazos…
- Y otros algoritmos refuerzan el "habit looping"

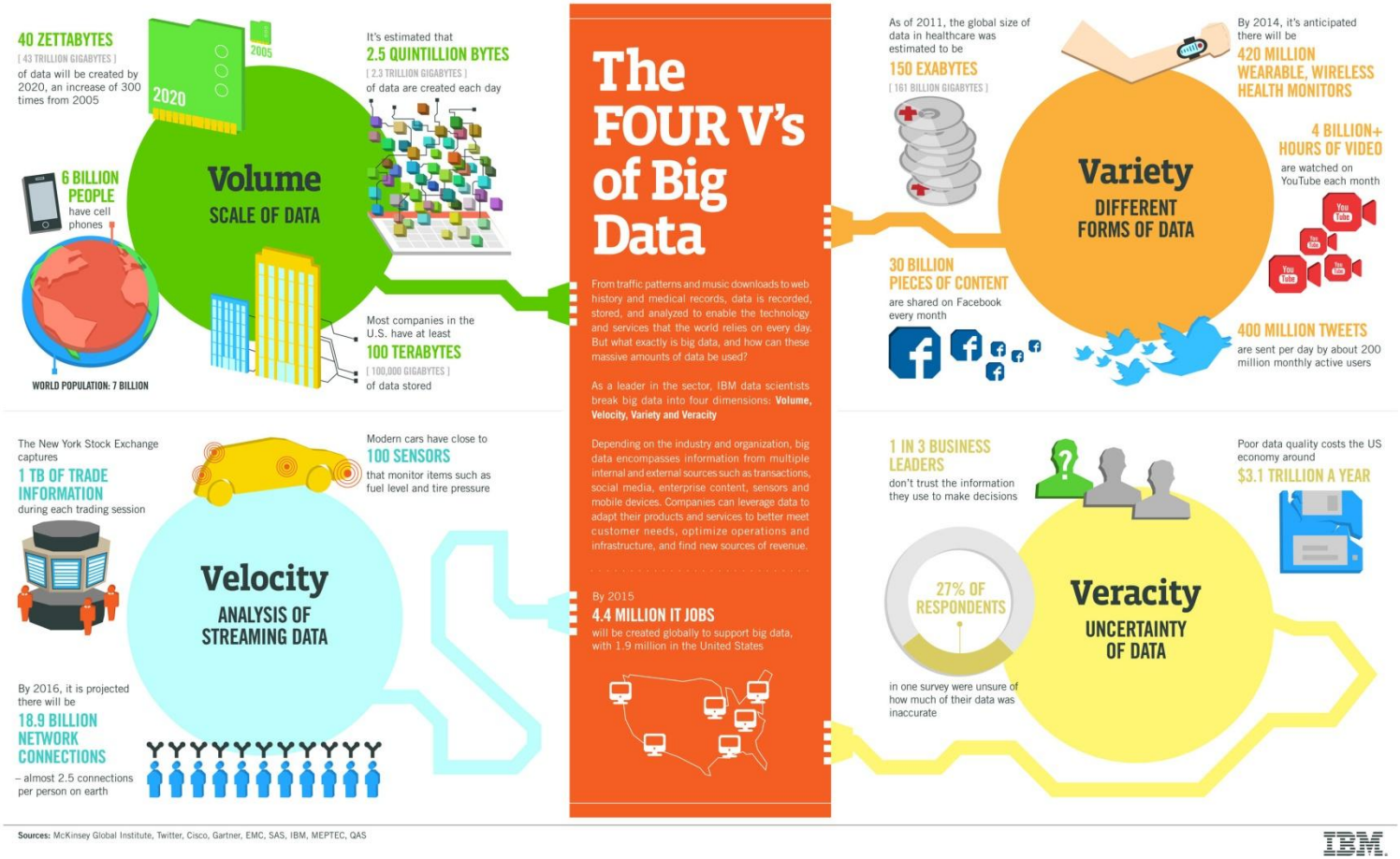
Introducción: el caso Amazon

- Amazón empleó hasta 2001 a docenas de críticos y editores para sugerir títulos a sus clientes-> “Amazon voice” fue considerado por el WSJ como el crítico más influyente de EEUU
- Jeff Bezos se pregunto si no sería mejor hacer recomendaciones basadas en libros específicos comprados por los clientes:
  - Primero se hizo utilizando muestras y buscando similaridades entre la gente
  - Linden propone solución nueva: filtro “item-by-item”
  - El ordenador no necesita saber por que el comprador de el Quijote le gustaría comprar también una tostadora

Introducción: el caso Amazon

- "Amazon voice" o "machine learning"? Críticos o algoritmos? -> el sistema de recomendación basado en datos únicamente ganó por goleada
- Todos lo críticos fueron despedidos -> hoy una tercera parte de las ventas de Amazon son resultado del sistema personalizado de recomendación
- El sistema de Linden ha sido adoptado por la mayoría de los grandes comercios digitales (por ejemplo Netflix, la compañía de alquiler de películas)
- Datos y PERs -> creación de valor a partir de la información

Big Data



barcelona gse
graduate school of economics

Big data

- Ingente cantidad de información: “la muestra es la población”. No se desperdicia nada
- Enorme heterogeneidad de formatos: sensores, GPS, clicks, logs de servidores, correos electrónicos, imágenes, voz, etc
- Bajo nivel de señal sobre ruido
- Reutilización de los datos
- No pretende explicaciones causales sino meramente predictivas -> causalidad es irrelevante, solo la correlación importa.
- Importancia de la visualización
- Reacción en tiempo real – modificación de modelos no supervisada y continua

Big data

Multiples of bytes v·d·e				
SI decimal prefixes		Binary usage	IEC binary prefixes	
Name (Symbol)	Value		Name (Symbol)	Value
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}

Processor or Virtual Storage	Disk Storage
· 1 Bit = Binary Digit	· 1 Bit = Binary Digit
· 8 Bits = 1 Byte	· 8 Bits = 1 Byte
· 1024 Bytes = 1 Kilobyte	· 1000 Bytes = 1 Kilobyte
· 1024 Kilobytes = 1 Megabyte	· 1000 Kilobytes = 1 Megabyte
· 1024 Megabytes = 1 Gigabyte	· 1000 Megabytes = 1 Gigabyte
· 1024 Gigabytes = 1 Terabyte	· 1000 Gigabytes = 1 Terabyte
· 1024 Terabytes = 1 Petabyte	· 1000 Terabytes = 1 Petabyte
· 1024 Petabytes = 1 Exabyte	· 1000 Petabytes = 1 Exabyte
· 1024 Exabytes = 1 Zettabyte	· 1000 Exabytes = 1 Zettabyte
· 1024 Zettabytes = 1 Yottabyte	· 1000 Zettabytes = 1 Yottabyte
· 1024 Yottabytes = 1 Brontobyte	· 1000 Yottabytes = 1 Brontobyte
· 1024 Brontobytes = 1 Geopbyte	· 1000 Brontobytes = 1 Geopbyte

barcelona gse
graduate school of economics

Big data

- **PETABYTE**
  - **1 petabyte** – Tres años de datos del Earth Observing System (EOS) (sobre el año 2001).
  - **2 petabytes** – Todas las bibliotecas de investigación académica en Estados Unidos.
  - **8 petabytes** – Toda la información que existe en Internet.
  - **20 petabytes** – Producción de discos duros en 1995.
  - **20 petabytes** – La información que Google rastrea cada día.
  - **200 petabytes** – Todo el material impreso o toda la producción de cintas digitales magnéticas en 1995.
- **EXABYTE**
  - **5 exabyte** – todas las palabras habladas por la humanidad en su existencia
  - **27 exabytes** – Toda la información creada en el planeta en una semana de 2011
  - **295 exabytes** – Toda la información creada por la humanidad entre 1986 y 2007

Big data

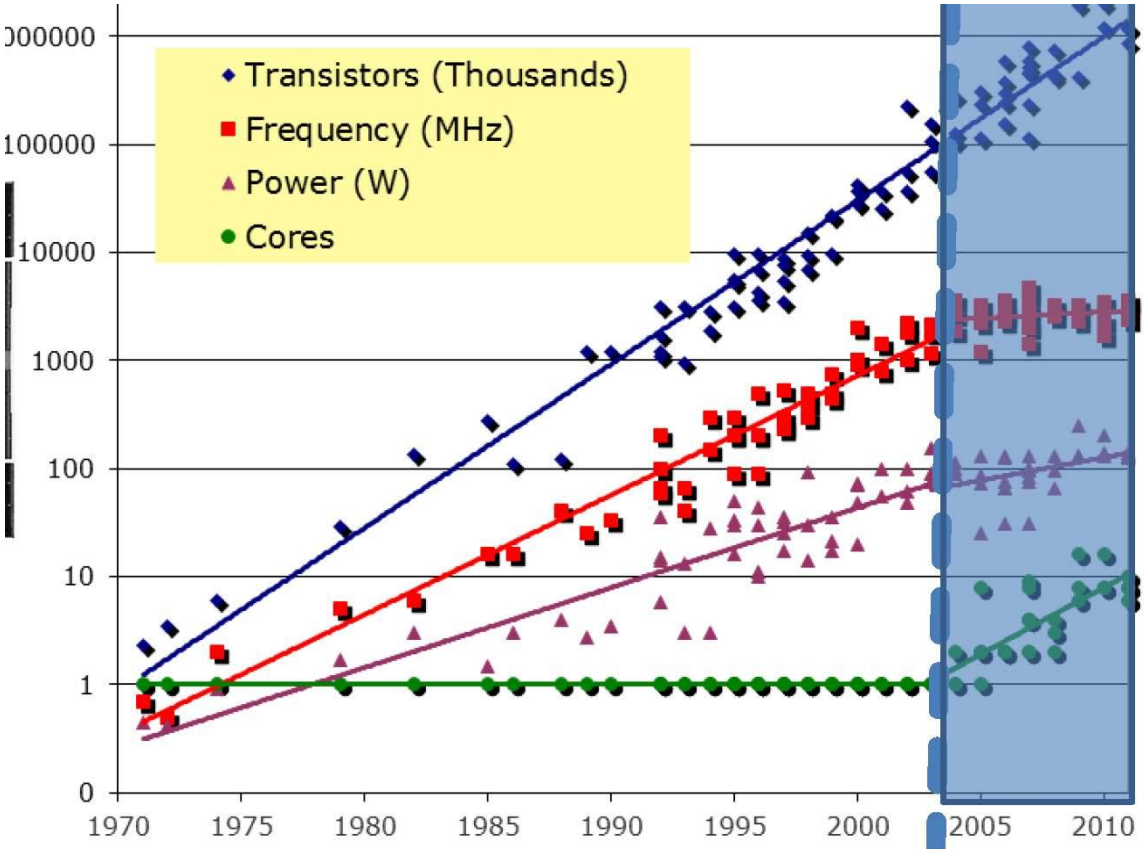
- En 1 minuto en Internet
  - 204 millones de emails
  - 2 millones de solicitudes de búsqueda en Google
  - 100.000 twits
  - Se crean 571 webs
  - Se suben 48 horas de video a Youtube
  - Se bajan 47000 apps de Apple

Big data

- Algunos hechos básicos sobre big data
  - La mayor cantidad de datos en la actualidad no la genera Internet sino los sensores y las grandes infraestructuras científicas: el Large Hadron Collider del CERN produce 600 TB/sec con sus 15 millones de sensores y, después de filtrado, necesita almacenar 25 PB/año
  - El problema fundamental en la actualidad no es la capacidad de computación sino la creación de información a un ritmo más rápido que la capacidad de almacenarla y la energía necesaria para mover la información entre el procesador y el dispositivo de almacenaje de la información (Islandia y Finlandia)

Big data: el problema del almacenaje

La ley de Moore + muro de memoria + muro energético



barcelona gse
graduate school of economics

Big data

- Algunos hechos básicos sobre big data
  - Necesidad de pasar de un modelo centrado en el ordenador en un modelo centrado en los datos con computación por paralelismo masivo (muchos cores y aceleradores) y memoria persistente – nuevas arquitecturas para aliviar el problema del gasto energético (minería de bitcoin en Islandia)
  - Necesidad de escalado y paralelismo (“distributed computing”) y los nuevos tipos de datos ha llevado al desarrollo de nuevas herramientas que pueden trabajar con bases de datos no relacionales (NoSQL) como por ejemplo MapReduce

Big data

- Algunos hechos básicos sobre big data
  - Nuevas soluciones de big data y data science han reducido significativamente el coste de procesos muy complejos como la secuenciación de genomas o la microsegmentación
  - Demanda de estadísticos, matemáticos e informáticos ha crecido significativamente: la inserción de los universitarios en 1999 y en 2014

Big data y estadística

- Hay algunas características del "big data" que lo diferencian de la estadística clásica:
  - Elevada dimensionalidad: k>>n: necesidad de utilizar métodos para reducir la dimensionalidad como LASSO (o cualquier otro que penalice el número de parámetros)
  - Mientas muchas ciencias avanzan hacía la formulación de métodos que permitan establecer causalidad (como la experimentación o la quasiexperimentación), los métodos estadísticos en big data se basan fundamentalmente en técnicas de "machine learning" donde se enfatizan los aspectos predictivos no causales
  - La crítica de Lucas a los modelos predictivos con un ejemplo sobre el fraude en tarjetas de crédito

Big data y economía

- La disponibilidad creciente de enormes bases de datos, en muchos casos geocodificadas, que fusionan información de procedencia diversa hace de la economía una disciplina cada vez más científica
  - The Billion Prices Project: estimación en tiempo real de la evolución de los precios utilizando millones de precios de tiendas on-line. Con este proyecto se muestra que la evolución oficial de la inflación y la de precios on-line sigue patrones parecidos en Brasil, Chile, Venezuela o Colombia pero no en Argentina (diferencia acumulada entre 2007 y 2011 un 65%)
  - 24 millones de créditos
  - Seguimiento de los participantes en el proyecto STAR
  - Choi y Varian (2014): complementar la información de pasado de una serie con la búsqueda en algunas categorías -> ejemplo: modelo AR(1) de subsidio de desempleo completado con Google Trends para palabras como jobs, welfare o unemployment

Big data y servicios financieros: motivación

- Necesidad de ganar reputación y recuperar la confianza de los clientes
- EBA, IMF, etc. insisten en que la banca europea tiene un problema fundamental: la rentabilidad -> necesidad de un nuevo modelo de negocio
- Competencia creciente de nuevos actores en la intermediación financiera que pueden absorber parte de la cadena de valor bancaria

1. Recuperar la confianza de los clientes

- ¿Podría la banca hacer como Amazon y recomendar productos casi individualizados a sus clientes?
- Las técnicas de big data abren la posibilidad de adaptarse a las necesidades de cada cliente
- Objetivo: mejorar el acceso de familias de renta medio baja y baja a productos financieros a un coste adecuado a su perfil de ingresos, capacidad de pago y nivel de aversión al riesgo

1. Recuperar la confianza de los clientes

- En muchos países, incluido Estados Unidos, hay una proporción importante de clientes que por no tener historial crediticio, o historial insuficiente (comportamental versus concesional) no pueden acceder a los servicios bancarios
- En Estados Unidos muchos de estos potenciales clientes acaban en empresas de "payday loan" con altos tipos de interés y plazos breves

2. Nuevo modelo de negocio

- La reducción de la rentabilidad de la banca unida al aumento constante en la regulación y al elevado nivel de endeudamiento actual requieren un esfuerzo de mejora de eficiencia
- La sostenibilidad del modelo de negocio bancario se puede basar en aprovechar, utilizando “big data”, las bolsas de ineficiencias que existen en el sector y aumentar la satisfacción de los clientes con la utilización intensiva de “big data”

3. Hacer frente a nuevos competidores

- La desintermediación también está afectando a las cuentas de resultados de las entidades financieras.
- Aunque en el pasado reciente este fenómeno se ha concentrado fundamentalmente en medios de pago (modelas criptográficas, pagos vía móvil, monedas complementarias, etc.) la competencia hacia otras partes de la cadena de valor se está moviendo rápidamente (préstamos “peer to peer”, préstamos al consumo, etc.)

Big data y finanzas: aplicaciones

- Las aplicaciones son muy variadas:
  - Medición del riesgo de crédito
  - Optimizar las relaciones con clientes
  - Mejorar las funciones financieras (mesas y fondos)
  - Asegurar el cumplimiento normativo
  - En recuperaciones un mejor conocimiento de las circunstancias de los clientes puede mejorar el "targeting" y aumentar las tasas de recuperación
  - Mejorar la segmentación de clientes: predicción del siguiente producto que adquirirá
  - Gestión de carteras de activos / inmuebles (okupados?)
  - Modificación de primas en seguros

1. Big data y banca: scoring

- FICO o modelos propios de cada entidad
- Modelos comportamentales para clientes vinculados y concesionales (basados en características demográficas básicamente) para el caso de no clientes:
  - Versiones en función del nivel de vinculación del cliente
  - Poco sofisticados en el uso de la información
  - Muy anticuados: hay más datos pero también hay mejores técnicas de análisis (“machine learning”)
  - ¿Cuál es el valor añadido en clientes con alto nivel de vinculación de la información de Internet? ¿Y en los no clientes o con poca vinculación?

1. Big data y banca: scoring

- Un ejemplo para cliente vinculado
  - Activos medios/pasivos medios
  - Importe debe/importe haber
  - Salario
  - Número de meses con saldo en exceso
  - Antigüedad del contrato de tarjeta más antiguo
  - Meses desde la última retirada de efectivo
  - Importe en efectivo/importe movimientos
- Claramente mejorable sin necesidad de introducir información de las redes sociales

1. Big data y banca: scoring

- Los credit scores de los bureaus proporcionan un ranking de clientes ordenados por su morosidad y tasa de impago que es muy útil como benchmark para la predicción de "machine learning"

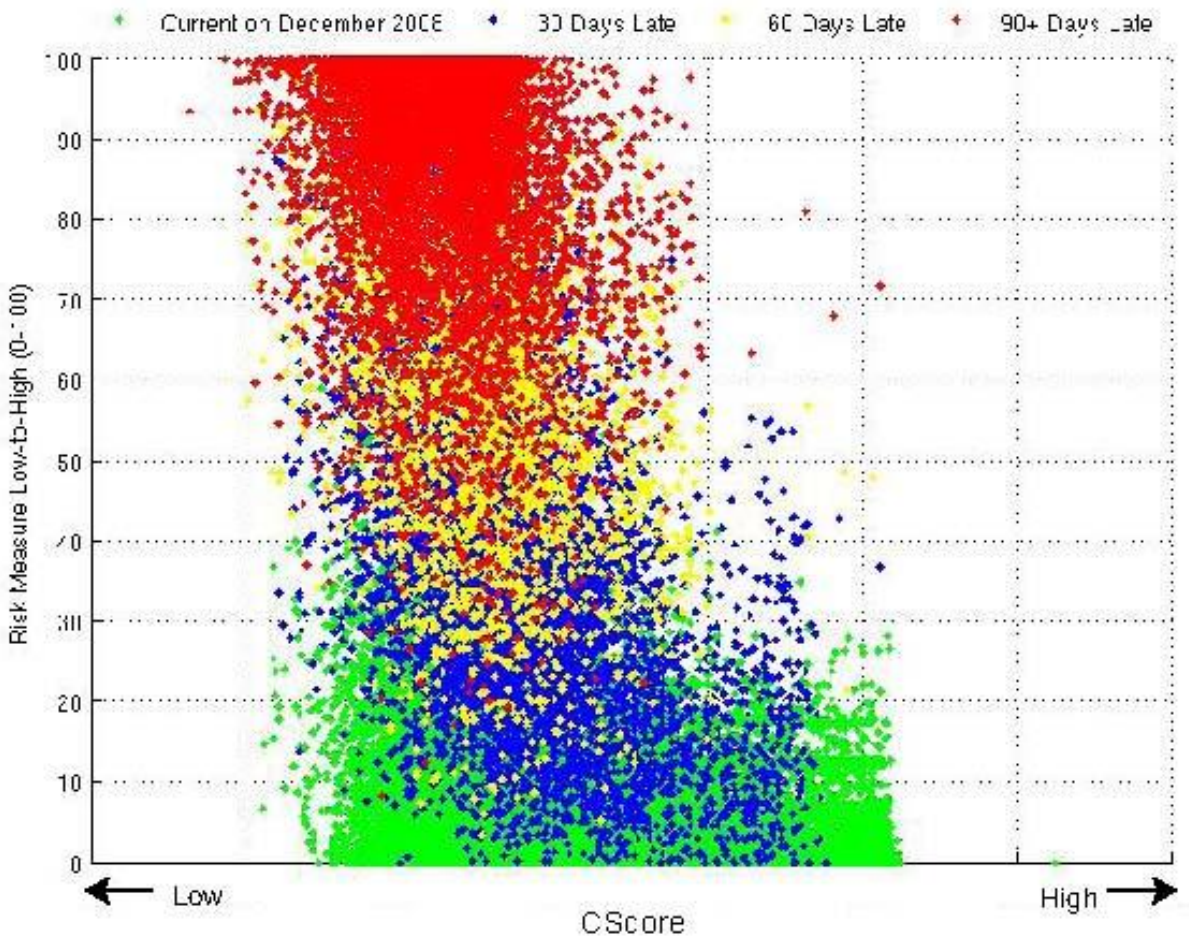
FICO

- Factors to calculate the credit score. Approximate weights:
  - 35% — Payment History – Late payments on bills, such as a mortgage, credit card or automobile loan, can cause a consumer's FICO score to drop.
  - 30% — Credit Utilization – The ratio of current revolving debt (such as credit card balances) to the total available revolving credit (credit limits).
  - 15% — Length of Credit History – As consumer's credit history ages, assuming they pay their bills, it can have a positive impact on their FICO score.
  - 10% — Types of Credit Used (installment, revolving, consumer finance) – Consumers can benefit by having a history of managing different types of credit.
  - 10% — Recent search for credit and/or amount of credit obtained recently – Multiple credit inquiries for a consumer seeking to open new credit, such as credit cards, retail store accounts, and personal loans, can hurt an individual's score.

1. Big data y banca: scoring

Model Inputs	
Credit Bureau Data	**Transaction Data (Cont.)**
Total Number of Trade Lines	Total expenses at discount stores
Number of Open Trade Lines	Total expenses at big-box stores
Number of closed trade lines	Total recreation expenses
Number and balance of auto loans	Total clothing stores expenses
Number and balance of credit cards	Total department store expenses
Number and balance of home line of credits	Total other retail stores expenses
Number and balance of home loans	
Number and balance of all other loans	Total utilities expenses
Number and balance of all other lines of credit	Total cable TV & Internet expenses
Number and balance of all mortgages	Total telephone expenses
Balance of all auto loans to total debt	Total net flow from brokerage account
Balance of all credit cards to total debt	Total net flow from dividends and annuities
Balance of all home line of credit to total debt	
Balance of all home loans to total debt	Total gas station expenses
Balance of all other loans to total debt	Total vehicle related expenses
Balance of all other lines of credit to total debt	
Ratio of total mortgage balance to total debt	Total logging expenses
	Total travel expenses
Total credit-card balance to limits	
Total home line of credit balances to limits	Total credit-card payments
Total balance on all other lines of credit to limits	Total mortgage payments
	Total outflow to car and student loan payments
Transaction Data	
	Total education related expenses
Number of Transactions	
Total inflow	**Deposit Data**
Total outflow	
Total pay inflow	Savings account balance
	Checking account balance
Total all food related expenses	CD account balance
Total grocery expenses	Brokerage account balance
Total restaurant expenses	
Total fast food expenses	
Total bar expenses	

1. Big data y banca: scoring



1. Big data y banca: ejemplos scoring

- El modelo utilizando solo variables transaccionales de los clientes y credit scores generados por bureaus -> Mejora la predicción de morosidad e impago alcanzando el 85%
- Un considerable proporción del ciclo crediticio se puede predecir 6-12 meses antes
- Usando supuestos conservadores sobre costes y beneficios de gestionar las líneas de crédito usando predicciones de “machine learning” se estima un ahorro entre el 6 y el 25% de las pérdidas totales

1. Big data y banca: scoring

- La banca tiene un enorme volumen de información que no utiliza eficientemente.
- Cuando no se recopilaba dicha información no había lugar a preguntarse si podría utilizarse: si se almacena hay que utilizarla -> pre-screening de créditos y límites
- Mejor y mayor utilización de los datos ya disponibles -> automatizar concesión de créditos con un modelo supervisado sin forzajes
- ¿Qué aporta “big data”?

1. Big data y banca: scoring

- J. P. Morgan en 1912 ante una comisión del congreso: "El factor más importante para conseguir un crédito no es la riqueza sino la reputación del solicitante. Un hombre en quien no confío no obtendría dinero de mí ni con toda la riqueza de la Cristiandad" -> reputación social -> credit score social o status social online
- Fuentes básicas para medir esa reputación: Facebook, Twitter, LinkedIn, etc.

1. Big data y banca: ejemplos scoring

- Neo Finance (Palo Alto) especializado en créditos para adquisición de vehículos de solicitantes jóvenes con poco historial laboral que tendría que pagar unos tipos altos en una financiera
  - Utiliza el número y la calidad de las conexiones en LinkeId del solicitante con los trabajadores de la empresa para predecir la estabilidad del empleo en el futuro y los ingresos
  - También estima sus contactos en otras empresas para estimar la probabilidad de encontrar un empleo una vez perdido -> objetivo último medir la estabilidad en el empleo

1. Big data y banca: ejemplos scoring

- Keditech
  - Utiliza el lugar de residencia de los amigos y sus trabajos para obtener un crédito. Si tiene amigos con créditos impagados tiene menor probabilidad de conseguir el crédito
  - El algoritmo tiene un tiempo de resolución de 8 segundos (media) y una tasa de mora menor del 10%
- Lenddo obtiene un capital social online en un score de 0 a 1000 usando el número de seguidores en Facebook, las características de los mismo, su nivel educativo, y su empleador e historial crediticio de sus amigos. Si un amigo deja de pagar mi "credit score" empeora -> parecido a microcréditos

1. Big data y banca: ejemplos scoring

- Nuevas variable con poder predictivo significativo: capitalizar en una solicitud -> el cliente que escribe toda la solicitud en mayúsculas o minúsculas tiene una probabilidad de mora significativamente superior

1. Big data y banca: ejemplos scoring

- Experian (Extended View) y Equifax (Vantage Score) están abiertos a utilizar otro tipo de información diferente de la habitual o a tratarla con pesos diferentes. FICO ha señalado que en principio no tiene interés
- Algunos lo denominan reinventar la rueda de las calificaciones crediticias: "it's the Wild West like the early days of FICO" Pete Fader
- Las entidades financieras tienen datos de alta cualidad basados en experiencia de muchos años de vinculación que difícilmente podrán ser sustituidos por las redes sociales que tienen mucho ruido

1. Big data y banca: ejemplos scoring

- BBVA: visualización gasto TPV's en tiempo real y servicios a terceros sobre posibilidades de negocio (ligadas a créditos)
- MasterCard y el Spending Pulse: datos en tiempo real sobre consumo en diferentes categorías comerciales
- VISA genera predicciones periódicas a partir de encuestas económicas
- Moody's Analytics predice cada mes el empleo en el sector privado utilizando unas 500.000 empresas a las que se facilita un software para las nóminas

2. Big data y banca: fraude en tarjetas

- El motivo es claro: un procedimiento sofisticado de detección de fraude en tarjetas de crédito puede ahorrar cientos de millones de euros a un banco. En este caso es fundamental una de las V de "big data": la velocidad.
- La cantidad de datos utilizados para la detección de fraude es ingente: datos sobre empleados, aplicaciones, fallecidos, encarcelados, listas negras, IRS, etc. así como patrones que pueden extraerse de la distribución geográfica de los pagos, las características del sector del negocio, de establecimientos similares, etc.

2. Big data y banca: fraude en tarjetas

- Cuatro aproximaciones:
  - Basada en reglas (patrones conocidos)
  - Detección de anomalías o outliers (patrones desconocidos)
  - Análisis predictivo en búsqueda de patrones complejos
  - Modelos híbridos

- "Big data, big time" -> es importante evitar falsos positivos por el gran efecto que puede tener sobre el cliente -> tacto de las pulsaciones sobre el pad en tiempo real

3. Otras utilidades en el sector financiero

- Peer to peer lending: RateSetter, Zopa, Lending Club, etc.
- Plataformas como Social finance, CommonBond o Upstart para deuda de graduados universitarios ->no son bancos sino que hacen de intermediarios entre inversores y solicitantes de crédito. Los solicitantes aceptan que estas plataformas obtengan datos sobre ellos de empleadores y redes sociales.
- La prima de los seguros de coche y los sensores

Limitaciones del big data

- Si bien es cierto que "big data" proporciona herramientas muy útiles en un ambiente de mayor incertidumbre, regulación y desconfianza de los consumidores en el sector financiero, no es menos cierto que la transformación de un proyecto de "big data" en un programa de éxito no está garantizada
  - Rendimientos decrecientes de la acumulación de información
  - Datos no proporcionan ventaja si no se analizan correctamente
  - Coste-Beneficio: ROI

Limitaciones del big data

- La existencia de grandes cantidades de datos no puede hacer olvidar los fundamentos de la ciencia estadística, la influencia de los errores de medida o la precaución contra la utilización de correlaciones espurias.
- Además del conocimiento técnico hace falta estar dispuestos a analizar constantemente la capacidad predictiva de los modelos y hacer ajustes a medida que el sistema pierde potencia explicativa. La experiencia de Google Trends (Google Flu Trends)

Limitaciones con los algoritmos

- Predicciones de la gripe ("Google Flu Trends"): big data y algoritmos tienen sus limitaciones -> en los últimos 3 años ha sobreestimado la gripe un 50% (artículos en Science y Nature)
- Princeton versus Facebook
  - Facebook perderá 80% de sus usuarios usando modelo epidémico sobre el número de veces que la palabra Facebook se busca en Google y usando MySpace como comprobación
  - "Utilizando el principio correlación implica causalidad y los mismos criterios de Princeton la universidad desaparecerá próximamente"

Limitaciones del big data

- Privacidad de los datos y reutilización. Nuevas cláusulas de consentimiento
- La posibilidad de que los errores en la captura, fusión o limpieza de los datos generen consecuencias negativas para los ciudadanos a partir de la aplicación de técnicas de "big data" a problemas concretos.
- Un ejemplo es la industria de generación de "credit scores" a partir de "big data" captado en Internet-> NCLC (2014) (National Consumer Law Center): "Big data disappointment for scoring consumer credit risk

Limitaciones del big data

- "Big data disappointment for scoring consumer credit risk":
  - Fair credit reporting act
  - Equal credit opportunity act
- ¿Cuánto valen mis datos? ¿Es suficiente para compensarme el dejarme navegar gratis?

Conclusiones

- Todavía hay mucho que ganar exprimiendo los datos transaccionales de las entidades financieras
- … pero no hay que cerrarse a la posibilidad de extraer información de fuentes poco convencionales (Facebook, Twitter, etc.) aunque siempre siendo conscientes del enorme ruido que se introduce y de la necesidad de nuevas herramientas estadísticas
- La información en las redes puede servir para medir la satisfacción de los clientes y complementar las encuestas -> bonus es función de satisfacción
- Realizar análisis coste beneficio antes de comenzar una actuación basada en “big data”

UNIVERSITAT POMPEU FABRA