Práctica – Parte 1

Alumno: Marc Román Porras

Diciembre 2024

El conjunto de datos seleccionado para la realización es el conjunto de préstamos totales emitidos por Lending Club entre 2007 y 2015, el cual se puede encontrar accesible desde entorno Kaggle:

Lending Club Loan Data

1. [10%] Justificad brevemente vuestra selección, sea por motivos personales o profesionales.

Lending Club es una plataforma estadounidense de préstamos entre particulares (P2P, *peer-to-peer*) que conecta directamente a prestatarios con inversores. Fue fundada en 2006 con la definición de un modelo de negocio innovador que permitió que personas solicitar préstamos personales sin intermediación bancaria, mientras que los inversores obtenían retornos competitivos al financiar dichos préstamos.

Aunque tuvo un crecimiento rápido, también enfrentó desafíos regulatorios y de gestión de riesgos. El dataset refleja la información histórica de préstamos emitidos por la plataforma entre los años 2007 y 2015, siendo estos los principales años de auge de este "modelo innovador", que además sufrió de manera casi paralela los efectos de una gran crisis inmobiliaria. Por tal razón, el conjunto es una fuente rica para el análisis de crédito y comportamiento financiero, concebido desde el punto de vista de explicación/visualización.

Por ende, la temática del conjunto de datos ha sido seleccionada por un interés personal y profesional subyacente. Desde hace 4 años trabajo en el sector bancario, específicamente gestionando y calibrando datos regulatorios de riesgos de crédito. Por ello, si bien el conjunto seleccionado refleja un entorno similar al que gestiono, permitiendo aplicar conocimientos teóricos y prácticos en un contexto relevante, se distancia de la nominalidad al tratarse de información de un negocio fuera del "core" convencional bancario, fuertemente impactado por la crisis financiera del 2008 y sobre un paradigma de regulación financiera escaso como es el mercado americano.

2.[10%] La relevancia del conjunto de datos en su contexto. ¿Son datos actuales? ¿Tratan un tema importante por algún colectivo concreto? ¿Se ha tenido en cuenta la perspectiva de género?

Tal y como se ha hecho notar en el punto anterior, el conjunto de datos abarca datos históricos de préstamos emitidos por Lending Club entre los años 2007 y 2015. Por esta razón, aunque en la actualidad el análisis no pueda ser representativo por cómo ha evolucionado el sector y, en concreto, el mercado, el interés viene fuertemente impactado sorbe cómo esta plataforma consiguió salir adelante con un modelo innovador en un contexto macroeconómico totalmente desfavorable.

El conjunto reporta datos múltiples que son idiosincráticos para la gestión del riesgo de crédito, la cual es crucial para instituciones financieras y plataformas fintech, pues impactan directamente en sus estados contables y en la rentabilidad de su negocio.

Finalmente, aunque el dataset no incluye explícitamente datos sobre género, pues estos datos discrecionales no pueden usarse en términos de gestión del riesgo, al igual que cualquier dato sobre identidad racial, nacionalidad, etc… Se puede analizar si existen sesgos indirectos en métricas como tasas de aprobación o impago, utilizando variables proxy como ubicación geográfica o categorías de ingresos.

3. [25%] La complejidad (medida, variables disponibles, tipos de datos, etc.). Debe tener del orden de miles de registros mínimo. Y debe tener un mínimo del orden de decenas de variables. ¿Combina datos categóricos y cuantitativos? ¿Incluye otros tipos de datos? La riqueza en tipología de variarles os puede ayudar a realizar un trabajo más brillante: valores discretos, continuos, fecha u hora, lógicos, cartográficos.

1. Tamaño del conjunto de datos:
   - Contiene aprox 2.26 millones de registros (operaciones relacionadas con un préstamo).
   - Incluye más de 140 variables que combinan datos categóricos, numéricos y de fecha. Tipos de variables:
     - **Numéricas:** Monto/saldo de préstamos, tasas de interés, ingresos declarados, etc.
     - **Categóricas:** Estado del préstamo (*Fully Paid, Charged Off*), nivel de educación, empleo, etc.
     - **Fechas:** Fecha de apertura del préstamo, fecha del último pago, etc.
     - **Variables calculadas y/o calificaciones:** Ratio deuda/ingresos (DTI), calificación crediticia, etc.
2. **Ejemplo en términos de riqueza:** Permite analizar tanto métricas financieras (tasas de interés, saldos pendientes) como contextos sociodemográficos (ubicación del prestatario, nivel de empleo).
3. **Complejidad adicional:** Requiere limpieza, manejo de valores faltantes y transformación de datos para análisis avanzados.

4. [25%] La originalidad. Se valora no repetir los conjuntos de datos clásicos o muy trabajados Links to an external site. ni temas ya muy tratados (p. ej. Covid-19, tráfico, criminalidad...) Podéis combinar o mejorar el conjunto de datos. En el primer caso, enriquecer el conjunto de datos con otros diferentes para dar un enfoque nuevo. En el segundo caso, generando nuevas métricas o indicadores con las variables existentes mediantes transformaciones. ¿Hay otras visualizaciones basadas en este conjunto de datos? ¿Es una evolución o actualización de un conjunto anterior? ¿Habéis enriquecido un conjunto de datos ya existente?

Aunque Lending Club es un tema conocido dentro de la industria financiera/bancaria, así como tratarse de un conjunto de datos de préstamos "tipo", es altamente interesante siempre la visualización/explicación que puede derivarse de un conjunto de datos agregados de crédito, desde el mero entendimiento de la información de cada acreditado, a como este es juzgado para potencialmente conceder o no el crédito según los patrones de comportamiento de los clientes.

Asimismo, el enfoque puede ser innovador al diferenciar de un análisis más nominal datos enriquecidos (combinar con información externa, como tasas de desempleo o indicadores económicos por estado en EE.UU.), nuevas métricas transformadas (por ejemplo, crear un índice de riesgo ajustado por localización o analizar la probabilidad de impago según patrones temporales), visualizaciones existentes que vayan más allá de la tasa de aprobación o pérdida, como el análisis de impactos sociodemográficos o tendencias temporales poco exploradas y, finalmente, incluso uso de herramientas analíticas para visualización para visualizar agrupaciones y detectar patrones de comportamiento de prestatarios (UMAP, T-SNE, etc...).

La visualización de Lending Club, no ha sido altamente visualizada, pues únicamente ha interesado a expertos dentro del negocio bancario para entender el funcionamiento de dicha plataforma.

5.[30%] Las cuestiones que responderéis con la visualización de datos, ¿Tienen en cuenta los puntos anteriores? ¿Han sido planteadas en otras visualizaciones u otros proyectos? ¿Son adecuadas para el conjunto de datos elegido? En este punto, elaborar un diccionario de las variables, su significado y si es un hecho a estudiar o una dimensión que lo mide, os puede ayudar.

Según la tipología de datos que nos proporciona el conjunto, nos podemos formular las siguientes preguntas, que potencialmente nos llevarían a un escenario de análisis y visualización afrontando de este modo la práctica de la asignatura:

Preguntas clave

- ¿Qué características tienen los prestatarios con mayor probabilidad de impago?
  - Variables relevantes: loan_amnt, annual_inc, emp_length, grade, dti.
- ¿Cómo varía el riesgo de crédito según la localización geográfica?
  - Variables relevantes: addr_state, loan_status, tasas externas de desempleo.
- ¿Qué relación existe entre la tasa de interés y la puntuación crediticia (FICO)?
  - Variables relevantes: int_rate, fico_range_high, fico_range_low.
- ¿Se observan patrones temporales en los impagos?
  - Variables relevantes: issue_d, loan_status.
- ¿Qué variables predicen mejor el cumplimiento del préstamo?
  - Variables relevantes: Combinación de datos numéricos y categóricos.

Diccionario de variables

Variable	Descripción	Tipo	Rol
loan_amnt	Monto del préstamo solicitado.	Numérica	Hecho
int_rate	Tasa de interés aplicada al préstamo.	Numérica	Dimensión
grade	Calificación asignada al prestatario (A, B, C, etc.).	Categórica	Dimensión
loan_status	Estado del préstamo (Fully Paid, Charged Off, Late).	Categórica	Hecho
annual_inc	Ingresos anuales declarados por el prestatario.	Numérica	Dimensión
addr_state	Estado de residencia del prestatario en EE.UU.	Categórica	Dimensión
issue_d	Fecha en que se emitió el préstamo.	Fecha	Dimensión
dti	Ratio deuda/ingresos del prestatario.	Numérica	Dimensión
fico_range_high	Límite superior del rango de puntuación crediticia.	Numérica	Dimensión
emp_length	Años de experiencia laboral declarada.	Categórica	Dimensión

Posibles visualizaciones para generar:

- **Mapas geográficos:** Mostrar tasas de impago por estado.
- **Gráficos de dispersión:** Relación entre tasa de interés y FICO.
- **Series temporales:** Evolución de tasas de impago a lo largo del tiempo.
- **Diagramas de caja y bigotes:** Comparar ingresos anuales por estado de préstamo.
- **Proyecciones con UMAP:** Agrupaciones de prestatarios según características.