TECNOLÓGICO DE MONTERREY

Escuela de Ingeniería y Ciencias

Maestría en Inteligencia Artificial Aplicada (MNA)

Análisis de Créditos Bancarios usando PySpark

Técnicas de muestreo orientadas a Big data

Integrantes:

Karla Merino Ángeles - A01795859

Jorge Andrés Moya Pacheco – A00813287

Héctor Raúl Peraza Alavez – A01795125

Materia:

Análisis de Grandes Volúmenes de Datos (Grupo 10)

Fecha:

4 de mayo del 2025

Introducción

El presente análisis se basa en un conjunto de datos históricos proveniente de LendingClub, una plataforma de préstamos entre pares (peer-to-peer), que contiene información detallada sobre millones de préstamos personales emitidos entre 2007 y el tercer trimestre de 2020. El objetivo del análisis es comprender las características sociodemográficas y financieras de los solicitantes, así como el comportamiento de los préstamos en términos de riesgo crediticio y desempeño. Para ello, se realizó una caracterización de las variables más representativas, dividiéndolas en dos grandes grupos: variables categóricas y variables numéricas.

Caracterización de la población

Variables numéricas

Las variables numéricas seleccionadas son representativas del perfil financiero y crediticio de los solicitantes:

- loan_amnt: Monto solicitado del préstamo.
- int_rate: Tasa de interés asignada, indicador del nivel de riesgo percibido.
- installment: Cuota mensual a pagar, reflejo de la carga financiera asumida.
- fico_range_low / fico_range_high: Rango de puntuación FICO, medida estándar de solvencia crediticia.
- annual_inc: Ingreso anual reportado, fundamental para evaluar capacidad de pago.
- dti (debt-to-income ratio): Relación deuda-ingreso, usada ampliamente en análisis de riesgo crediticio.
- open_acc / total_acc: Número de cuentas de crédito abiertas y totales, reflejan historial crediticio.
- revol_bal: Saldo de crédito rotativo pendiente.
- revol_util: Porcentaje de utilización del crédito rotativo, indicador de comportamiento financiero.

Estas variables son frecuentemente empleadas en modelos predictivos de scoring crediticio, ya que permiten cuantificar el riesgo potencial de un prestatario y construir perfiles financieros detallados.

Variable	Valores Únicos	Mínimo	Máximo	Media
loan_amnt	1 572,00	500,00	40 000,00	15 577,42
int_rate	704,00	5,31	30,99	13,03
installment	97 895,00	4,93	1 719,83	458,19
fico_range_low	48,00	610,00	845,00	700,24
fico_range_high	49,00	614,00	850,00	704,24
annual_inc	92 319,00	0,00	110 000 000,00	82 107,53
dti	11 458,00	-1,00	999,00	19,01
open_acc	93,00	1,00	104,00	11,79
total_acc	157,00	1,00	176,00	24,16
revol_bal	111 731,00	0,00	2 904 836,00	17 247,56

revol_util	1 437,00	0,00	7 540,00	49,35

Observaciones de calidad de datos :

- La columna dti presenta un valor máximo de 999 % y un mínimo de −1 %, indicios de captura errónea. Estos outliers se etiquetarán para futura limpieza. * ≈ 26 % de los 142 campos totales del dataset presentan nulos; el análisis posterior descarta variables con >50 % nulos.

Variables Categóricas

Se eligieron variables categóricas clave que permiten describir aspectos cualitativos relevantes del prestatario o del préstamo:

- term: Indica la duración del préstamo (36 o 60 meses), fundamental para analizar la relación entre plazo y riesgo.
- grade: Asignación interna de riesgo crediticio de LendingClub, crucial para segmentar el perfil de los solicitantes.
- emp_length: Refleja la estabilidad laboral del prestatario, factor importante en la evaluación del riesgo de incumplimiento.
- home_ownership: Representa la situación habitacional del prestatario, lo que puede estar vinculado con su estabilidad financiera.
- verification_status: Señala si los ingresos del prestatario fueron verificados, lo que afecta la confiabilidad de la información crediticia.
- purpose: Describe el propósito declarado del préstamo, lo cual permite identificar patrones de uso del crédito.
- loan_status: Es el resultado del préstamo (pagado, vigente, moroso, incumplido, etc.), variable objetivo clave para modelos predictivos de riesgo.

Estas variables son esenciales para entender comportamientos cualitativos del prestatario y el contexto del préstamo, y son comúnmente utilizadas en modelos de clasificación y segmentación.

Variable	Categoría	Frequencia
term	**36 months**	1 899 703,00
	60 months	817 879,00
grade	**B**	796 324,00
	C	744 780,00
	A	611 109,00
	D	384 498,00
	E	129 949,00
	F	39 420,00
	G	11 502,00
emp_length	**10+ years**	945 379,00
	< 1 year	271 703,00
	2 years	261 642,00
	3 years	232 108,00

	1 year	193 968,00
	5 years	182 216,00
	4 years	176 008,00
	6 years	130 614,00
	7 years	115 977,00
	8 years	112 614,00
	9 years	95 352,00
	reactors"	1,00
home_ownership	**MORTGAGE**	1 344 604,00
	RENT	1 082 750,00
	OWN	287 044,00
	ANY	2 958,00
	OTHER	177,00
	NONE	48,00
	2 years	1,00
verification_status	**Source Verified**	1 106 712,00
	Not Verified	981 453,00
	Verified	629 416,00
	38000	1,00
purpose	**debt_consolidation**	1 525 886,00
	credit_card	648 722,00
	home_improvement	176 078,00
	other	162 032,00
	major_purchase	58 370,00
	medical	31 499,00
	small_business	29 029,00
	car	27 578,00
	vacation	18 352,00
	house	17 960,00
	moving	17 695,00
	wedding	2 321,00
	renewable_energy	1 647,00
	educational	412,00
	https://lendingclub.com/browse/loanDetail.action?loan_id=61400928	1,00
loan_status	**Fully Paid**	1 410 196,00
	Current	945 077,00
	Charged Off	331 567,00
	Late (31-120 days)	14 284,00
	In Grace Period	9 178,00
	Late (16-30 days)	2 334,00
	Does not meet the credit policy. Status:Fully Paid	1 935,00
	Issued	1 901,00
	Does not meet the credit policy. Status:Charged Off	740,00

	Default	369,00
	oct-15	1,00

Interpretación rápida: la gran proporción de debt consolidation y Fully Paid indica que el portafolio está dominado por prestatarios que buscan refinanciar deudas y logran liquidar sus préstamos, aunque un 11 % termina en Charged Off.

Particionamiento

El proceso de particionamiento aplicado al conjunto de datos de LendingClub tiene como objetivo organizar y dividir la información en subconjuntos significativos según combinaciones específicas de variables clave. Esto permite realizar análisis más enfocados y eficientes, así como construir modelos predictivos ajustados a segmentos particulares de la población.

Se definieron 68 estratos combinando grade y loan_status. A continuación se muestran las primeras 20 filas completas de la distribución poblacional y la distribución de la muestra.

Tabla 1. Distribución poblacional — primeros 10 estratos

grade_status	Registros	% población
A_Charged Off	19 957	0.73
A_Current	281 277	10.35
A_Default	28	0.00
A_Does not meet the credit policy. Status:Fully Paid	77	0.00
A_Fully Paid	305 488	11.24
A_In Grace Period	1 271	0.05
A_Issued	1 095	0.04
A_Late (16-30 days)	305	0.01
A_Late (31-120 days)	1 605	0.06

Tabla 2. Población vs. muestra — primeros 10 estratos

grade_status	Registros	% población	Muestra	% muestra
A_Charged Off	19 957	0.73	1 997	10.01
A_Current	281 277	10.35	28 034	9.97
A_Default	28	0.00	3	10.71
A_Does not meet the credit policy. Status:Fully Paid	77	0.00	8	10.39

A_Fully Paid	305 488	11.24	30 475	9.98
A_In Grace Period	1 271	0.05	147	11.57
A_Issued	1 095	0.04	110	10.05
A_Late (16-30 days)	305	0.01	22	7.21
A_Late (31-120 days)	1 605	0.06	161	10.03
B_Charged Off	66 916	2.46	6 707	10.02

Variables utilizadas para el particionamiento

- grade: Representa una evaluación integral de riesgo basada en el perfil del prestatario. Es una variable generada por la propia plataforma y agrupa múltiples factores como historial crediticio, ingresos, y deuda. Utilizar grade como criterio de partición permite analizar cómo se comportan prestatarios con niveles de riesgo similares ante diferentes condiciones de pago.
- loan_status: Esta variable refleja el resultado del préstamo, permitiendo distinguir entre casos exitosos (Fully Paid) y problemáticos (Charged Off, Default, Late). Al combinar esta variable con grade, se puede observar cómo los prestatarios con la misma calificación se comportan a lo largo del tiempo, lo cual es crucial para construir modelos de predicción del estado del préstamo.

Funcionamiento:

```
#FILTRADO DE PARTICIÓN

from pyspark.sql.functions import count, avg, min, max

print("Extracción de submuestras por combinación:\n")

for grade_val, loan_status_val in combinations:
    subset_df = df.filter((col("grade") == grade_val) & (col("loan_status") == loan_status_val))
    count_val = subset_df.count()

    print(f"\n Combinación: grade = {grade_val}, loan_status = {loan_status_val}")
    print(f"Total de registros: {count_val}")
```

1. Se genera una lista de combinaciones únicas (combinations) entre grade y loan_status.
2. En cada iteración, se filtra el DataFrame para seleccionar solo los registros que cumplan con ambos criterios simultáneamente.
3. Se almacena la submuestra correspondiente y se calcula el número total de registros (count()), lo cual sirve como verificación de la distribución de los datos en cada partición.

Para observar empíricamente la estructura de los datos hemos generado una representación visual de los subconjuntos generados, en ella se muestra la distribución de registros para cada combinación entre las variables grade y loan_status, las cuales se utilizaron como base para las reglas de particionamiento.

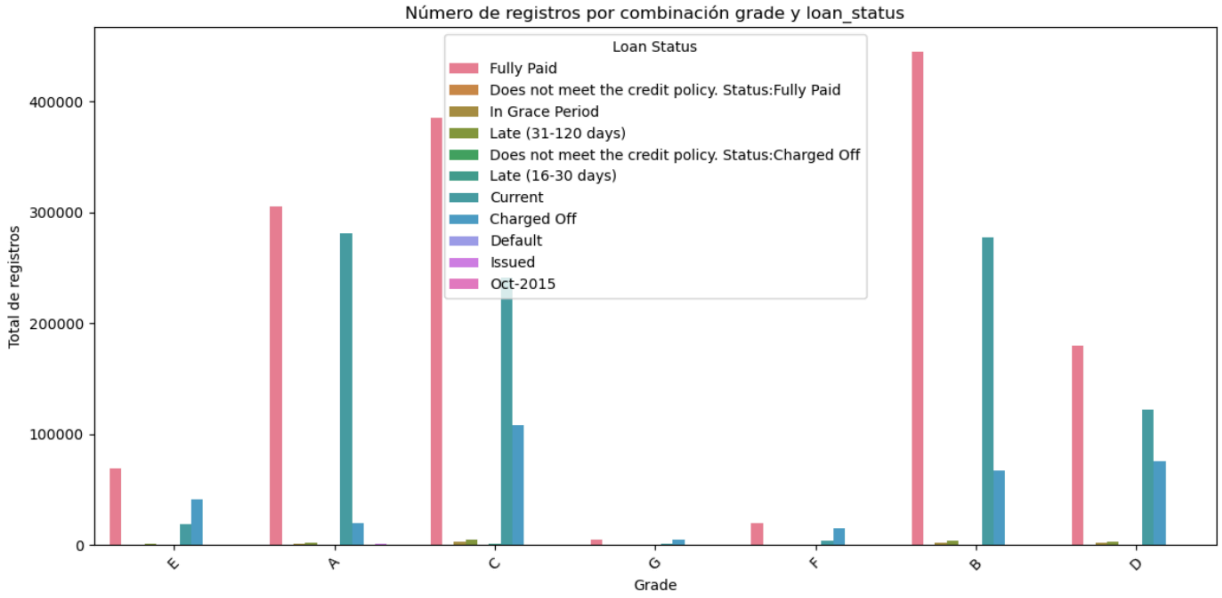

Figura 1. *Distribución de registros por combinación de las variables grade y loan_status.*

Esta visualización permite identificar patrones de densidad en la base de datos original, resaltando qué combinaciones son más frecuentes y, por tanto, potencialmente más representativas de la población. Por ejemplo, se observa que combinaciones como grade A y loan_status 'Fully Paid' dominan ampliamente en frecuencia, lo cual podría sesgar los modelos si no se compensa adecuadamente. En contraste, combinaciones como grade G y 'Default' presentan baja ocurrencia, lo que puede limitar el entrenamiento de modelos para la detección de comportamientos de alto riesgo.

Por lo tanto, permite detectar combinaciones menos frecuentes que podrían requerir un tratamiento especial en las etapas de muestreo y análisis posterior, como en el balanceo de clases o el ajuste de tamaños muestrales.

Esta observación resulta clave para guiar la selección de técnicas de muestreo estratificado, garantizando que cada estrato tuviera representación adecuada. Así, se fortalecen las bases para un análisis posterior más equilibrado y robusto, en particular en escenarios de clasificación o predicción del estado de crédito.

Enlace a los archivos de ejecución junto con la base de datos

Proyecto

Proceso de Muestreo

El muestreo es una etapa esencial en el procesamiento de grandes volúmenes de datos, especialmente cuando se pretende entrenar y validar modelos predictivos. En contextos de big data, donde procesar la totalidad del conjunto puede ser ineficiente o innecesario, se requiere una estrategia de muestreo bien fundamentada para preservar la representatividad y reducir sesgos (Kim & Wang, 2019). En este caso, el proceso de muestreo se realiza a partir de las particiones generadas por las combinaciones de las variables grade y loan_status.

Partición Inicial de los Datos

Como primer paso, el conjunto de datos fue segmentado en subconjuntos a partir de las combinaciones únicas de las variables categóricas grade (nivel crediticio del prestatario) y loan_status (estado del préstamo). Esta estrategia de particionamiento se alinea con el principio de estratificación, que busca agrupar datos homogéneos para facilitar análisis más precisos (Ahmed, 2024).

Este enfoque permite capturar la diversidad inherente a las combinaciones de riesgo crediticio y desempeño del préstamo, lo cual es fundamental para construir modelos robustos que puedan generalizar adecuadamente sobre los distintos perfiles de prestatarios.

Selección de la Técnica de Muestreo

Dado que el conjunto de datos es amplio y presenta una alta variabilidad entre las distintas particiones, se seleccionó el muestreo aleatorio estratificado como técnica principal. Esta técnica consiste en dividir el conjunto completo en estratos (en este caso, las combinaciones de grade y loan_status) y luego seleccionar aleatoriamente una muestra dentro de cada estrato.

Este método tiene ventajas claras en cuanto a representatividad y control del error de estimación, ya que asegura que cada subgrupo relevante esté proporcionalmente incluido en la muestra (Lakens, 2022). Además, facilita la comparación entre estratos y mejora la eficiencia del muestreo al reducir la varianza total de la estimación (Ahmed, 2024).

Muestreo Aleatorio Estratificado

Definición: El muestreo aleatorio estratificado consiste en dividir el conjunto de datos en subgrupos (estratos) que son homogéneos con respecto a una o más características clave. Luego, se seleccionan muestras aleatorias dentro de cada estrato.

Justificación: En el contexto del dataset de LendingClub, las combinaciones de grade y loan_status actúan como estratos, ya que reflejan variabilidad tanto en el riesgo crediticio como en el estado del préstamo. La combinación de ambos crea subgrupos que pueden diferir significativamente en su distribución de características, por lo que es crucial asegurarse de que todos los estratos estén representados proporcionalmente en la muestra.

Proceso: Para cada partición (es decir, para cada combinación de grade y loan_status), se realiza un muestreo aleatorio para seleccionar un número representativo de registros, que puede ser el 10%, 20%, o cualquier otro porcentaje adecuado dependiendo de los objetivos del análisis. Este enfoque asegura que cada subgrupo dentro de la partición tenga una representación proporcional dentro de la muestra final.

Método de Selección dentro de las Particiones

El muestreo dentro de cada partición se realiza de la siguiente forma:

1. Determinación de la proporción de la muestra: Se define una fracción del total de registros dentro de cada partición que se seleccionará para el análisis. Esto podría ser, por ejemplo, el 20% de cada partición.
2. Muestreo aleatorio dentro de la partición: Se realiza una selección aleatoria de los registros dentro de cada partición, garantizando que cada registro dentro de esa partición tenga la misma probabilidad de ser seleccionado.
3. Tamaño total de la muestra: Dependiendo de los objetivos del análisis o del modelo, el tamaño de la muestra de cada partición puede variar. Esto también dependerá de la distribución de los datos en las particiones, ya que algunas combinaciones de grade y loan_status pueden tener más registros que otras.

Código en Pyspark de ejemplo para el muestreo aleatorio:

sampled_df = subset_df.sample(fraction=0.2, seed=42) # 20% de los registros de cada partición

Este código selecciona aleatoriamente el 20% de los registros de cada partición, asegurando que cada subgrupo sea representado de manera proporcional en el muestreo.

Conclusión

En conclusión, se realizó un análisis detallado que permitió caracterizar de manera precisa a la población objetivo (P), a partir de variables clave como loan_amnt, grade y loan_status, las cuales ofrecen una visión integral del comportamiento financiero de los solicitantes de crédito. Estas variables facilitaron la definición de reglas de particionamiento que derivaron en subconjuntos estadísticamente representativos, manteniendo la estructura y proporciones de la población original.

Asimismo, se establecieron técnicas de muestreo específicas para cada partición, empleando principalmente el muestreo estratificado como estrategia para reducir al mínimo los sesgos en la selección de instancias. Esta metodología asegura que cada grupo esté adecuadamente representado, respetando la heterogeneidad de la población y permitiendo que las conclusiones derivadas de los modelos de aprendizaje automático sean tanto robustas como confiables

Referencias bibliográficas

Ahmed, S. K. (2024). *How to choose a sampling technique and determine sample size for research: A simplified guide for researchers*. *Oral Oncology Reports, 12*, 100662. https://doi.org/10.1016/j.oor.2023.100662

Kim, J. K., & Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review, 87*(S1), S177–S191. https://doi.org/10.1111/insr.12267

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267