TECNOLÓGICO DE MONTERREY
Escuela de Ingeniería y Ciencias

Maestría en Inteligencia Artificial Aplicada (MNA)

Análisis de Créditos Bancarios usando PySpark

Lectura, Escritura y Manejo de Archivos de 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:
27 de Abril del 2025

Contexto y propósito del análisis

El análisis de riesgo crediticio es fundamental en el sector financiero para evaluar la capacidad de pago de los solicitantes y minimizar pérdidas por impago; se trata de un tema de alta relevancia en la gestión de riesgos que requiere manejar grandes volúmenes de datos que reflejan condiciones reales del mercado, permitiendo así el desarrollo de soluciones escalables y de impacto directo.

Debido a esta necesidad, en el presente proyecto se ha definido como objetivo utilizar técnicas de Big Data con PySpark para procesar grandes volúmenes de datos de préstamos bancarios, identificar patrones de incumplimiento y desarrollar modelos de scoring crediticio, fortaleciendo así la toma de decisiones financieras basadas en datos.

Selección y descripción del Dataset

Para cumplir con el objetivo del proyecto se seleccionó el "Lending Club Loan Data (2007-2020 Q3)", publicado por LendingClub para uso académico y disponible en Kaggle. El dataset, en formato CSV comprimido (~450 MB) y descomprime a ≈ 1.2 GB, contiene 2 925 493 registros de préstamos emitidos entre 2007 y 2020, con 142 variables por registro. Entre sus principales campos destacan el estado final del préstamo, montos y tasas de interés, características socioeconómicas de los prestatarios, historial de morosidad y métricas internas de riesgo (grade, sub-grade, FICO ranges).

Revisión del Dataset

1. Estructura y tamaño

El conjunto de datos contiene aproximadamente 2.9 millones de registros y 142 variables, combinando información numérica (monto del préstamo, ingreso anual), categórica (propósito, grado crediticio, tipo de vivienda) y temporal (fecha de emisión, último pago).

2. Valores faltantes

En cuanto a los valores faltantes, el dataset muestra una tasa global aproximada del **25.97%** de valores nulos en las 142 columnas.

Las cinco columnas con mayor cantidad de vacíos son aquellas relacionadas con programas de indulgencia, como *hardship_loan_status*, *hardship_reason*, *hardship_status*, *hardship_dpd* y *hardship_length*, que tienen más del **95%** de valores nulos. También hay otras columnas opcionales, como *description*, que superan el **50%** de vacíos. Esto es común en bases de datos longitudinales, donde el esquema cambia con el tiempo (Han, Kamber & Pei, 2011)

3. Distribución del Estado de los Préstamos

La distribución del estado de los préstamos muestra que:

- El **63.8%** de los préstamos concluidos han sido **completamente pagados** (*Fully Paid*)
- El **19.5%** de los préstamos ha sido **cargado a pérdida** (*Charged Off* o *Default*).
- El resto de los préstamos se encuentra en diferentes etapas, como *Current*, *Late* o *In Grace Period*.

Este panorama sugiere que, al modelar la probabilidad de incumplimiento, es recomendable centrarse únicamente en los préstamos finalizados, ya que de lo contrario se corre el riesgo de introducir un sesgo de supervivencia en el modelo (Provost & Fawcett, 2013).

4. Calidad de los datos

Además de los valores faltantes, se detectaron otros problemas:

- **Codificación heterogénea**: Algunos campos como *emp_title* (título del empleo) y *emp_length* (antigüedad laboral) contienen texto libre, lo que puede generar inconsistencias a la hora de procesar estos datos.
- **Outliers**: Se identificaron valores extremos en diversas columnas. Por ejemplo, se encuentran outliers en *out_prncp_inv* (2.70% de outliers), *delinq_2yrs* (2.43%), y *num_accts_ever_120_pd* (2.21%). Estos valores extremos sugieren que se requieren técnicas de tratamiento de datos, como la winsorización o transformaciones logarítmicas (Aggarwal, 2015), para evitar que estos afecten la calidad del análisis.

Análisis del Dataset

1. Variables de importancia

Se identificaron variables esenciales para el análisis del riesgo crediticio, incluyendo *loan_amnt*, *funded_amnt*, *int_rate*, *grade*, *sub_grade*, *annual_inc*, *dti*, *home_ownership*, *purpose* y los rangos de puntaje FICO. Estas variables reflejan tanto la situación financiera del prestatario como su perfil de riesgo, y son fundamentales para construir modelos predictivos de incumplimiento.

2. Estadísticos descriptivo:

Utilizando pyspark para procesar el volumen masivo de datos, se revela la siguiente información:

- El análisis indica que el monto **promedio** de los préstamos otorgados es **$15 358.78** y la **mediana** se mantiene en **$12 800.00**, lo que confirma una ligera asimetría hacia montos altos.

- Las **tasas de interés** oscilan entre **5.31 % y 30.99 %**, con un promedio de 13.05 % (dato que ahora sí está disponible tras la inferencia de tipos en PySpark).Lo que refleja la diversidad en los perfiles de riesgo de los prestatarios.

- El **ingreso anual promedio** de los prestatarios es **$79 937.29**, mientras que la **relación deuda/ingreso (DTI) promedia 19.30** %. Se observa un rango anómalo de DTI de –1 % a 999 %, lo que sugiere valores atípicos o registros mal capturados.

```
📄 Detectados archivo(s):
  • Loan_status_2007-2020Q3.gzip
📊 Registros: 2,925,493 | Columnas: 142

=== Top-10 columnas con más nulos ===
                            nulls  pct_nulls
emp_length_num            1151489      39.36
emp_length                 205221       7.01
tot_coll_amt                70277       2.40
num_accts_ever_120_pd       70277       2.40
dti                          3109       0.11
delinq_2yrs                    30       0.00
pub_rec                        30       0.00
annual_inc                      5       0.00
grade                           1       0.00
int_rate                        1       0.00

=== Distribución de loan_status ===
+-----------------------------------------------+------+
|loan_status                                    |count |
+-----------------------------------------------+------+
|Fully Paid                                     |1497783|
|Current                                        |1031016|
|Charged Off                                    |362548 |
|Late (31-120 days)                             |16154  |
|In Grace Period                                |10028  |
|Late (16-30 days)                              |2719   |
|Issued                                         |2062   |
|Does not meet the credit policy. Status:Fully Paid |1988   |
|Does not meet the credit policy. Status:Charged Off|761    |
|Default                                        |433    |
|NULL                                           |1      |
+-----------------------------------------------+------+


=== Estadísticos descriptivos ===
+-------+------------------+------------------+------------------+------------------+
|summary|        loan_amnt|      int_rate_num|       annual_inc|               dti|
+-------+------------------+------------------+------------------+------------------+
|  count|           2925492|           2925492|           2925488|           2922384|
|   mean|15358.775510517889|13.045447041578173|79937.27401521573|19.296479634435542|
| stddev|9478.383451419413| 4.860814827474883|111747.6087097762|15.719772862825048|
|    min|               500|              5.31|               0.0|              -1.0|
|    max|             40000|             30.99|            1.1E8|             999.0|
+-------+------------------+------------------+------------------+------------------+
```

Figura 1. Código en PySpark para calcular las estadísticas descriptivas del dataset.

3. Problemas detectados y consideraciones de calidad

Se detectaron varias consideraciones críticas para la calidad de los datos:

- **Valores faltantes**: Se confirma que el **25.97%** de los datos están incompletos, lo que requiere técnicas de imputación para manejar estos vacíos.
- **Outliers**: Se encontró que varias columnas contienen valores extremos que deben ser gestionados adecuadamente. Esto puede hacerse a través de técnicas de **winsorización** o **transformaciones logarítmicas**, que ayudan a mitigar el impacto de estos outliers en los modelos.
- **Inconsistencias en formato**: Algunos campos, como *emp_length*, contienen datos en formato texto que necesitan ser estandarizados para evitar la pérdida de información durante el preprocesamiento.

4. Potencial para analisis predictivo

A pesar de los retos de calidad, el dataset ofrece un gran potencial para el desarrollo de modelos predictivos de riesgo crediticio. PySpark permitirá realizar procesamiento en paralelo, facilitar el entrenamiento y validación de modelos de clasificación, generar nuevas variables mediante feature engineering y evaluar la importancia relativa de los atributos, optimizando así los resultados sin necesidad de muestreo intensivo.

Conclusión

El conjunto de datos de préstamos de LendingClub seleccionado constituye una opción idónea para proyectos de análisis de riesgo financiero utilizando PySpark. Su volumen masivo y la diversidad de variables reflejan la complejidad real del otorgamiento de crédito, ofreciendo un entorno ideal para explorar técnicas de análisis predictivo de Big Data.
Si bien se identifican retos importantes de calidad y estructura de datos, estos pueden ser abordados mediante metodologías adecuadas de limpieza y transformación. En conclusión, el dataset provee los elementos necesarios para cumplir el objetivo general del proyecto: analizar factores de riesgo crediticio y desarrollar modelos de scoring basados en datos, potenciando así la toma de decisiones en instituciones financieras.

Referencias

- Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.
- GES Comunicación. (2023, 1 de junio). Análisis de riesgo crediticio: La clave para una gestión financiera sólida. Noticias ESEC – Universidad Galileo. Recuperado de https://www.galileo.edu/esec/noticias/analisis-de-riesgo-crediticio-la-clave-para-una-gestion-financiera-solida/
- Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
- LendingClub. (2020). Lending Club Loan Data (2007-2020Q3) [Conjunto de datos]. Kaggle. Recuperado de https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1
- Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.