Universidad Austral

Maestría en Ciencia de Datos

Introducción al Data Mining

INTRODUCCIÓN A LA MINERÍA DE DATOS

TRABAJO PRÁCTICO FINAL

Docentes

Mg. Leandro Kovalevski
Mg. Pablo Beltramone

Universidad Austral
Maestría en Ciencia de Datos
Introducción al Data Mining

Trabajo Práctico Final

Fecha de entrega: 23/02/2025

1. Introducción

ESTUDIO DE LAS DEUDAS REGISTRADAS DEL SISTEMA FINANCIERO ARGENTINO

El Banco Central de la República Argentina (BCRA) publica mensualmente un informe consolidado de deudas actuales e históricas (24 meses), denominado 'Central de Deudores del Sistema Financiero' elaborado en función de los datos recibidos de distintos tipos de entidades financieras (entidades financieras, empresas no financieras emisoras de tarjetas, fideicomisos financieros, otros proveedores no financieros de crédito, etc.), las cuales deben obligatoriamente remitir mensualmente al BCRA, detallando la totalidad de las financiaciones con la correspondiente situación de cada deudor.

Cada deuda informada al BCRA es acompañada de su situación que es una aproximación a la cantidad de días de atraso en el cumplimiento de pago:

Situación 1 | Situación normal: atraso en el pago que no supere los 31 días.
Situación 2 | Riesgo bajo: atraso en el pago de más de 31 días y hasta 90 días.
Situación 3 | Riesgo miedo: atraso en el pago de más de 90 días y hasta 180 días.
Situación 4 | Riesgo alto: atraso en el pago de más de 180 días a un año.
Situación 5 | Irrecuperable: atraso superior a un año.
Situación 6 | Irrecuperable por disposición técnica: deuda con una ex entidad.

Se cuenta con una muestra aleatoria de 19.737 cuits de personas físicas que tenían al menos una de deuda en el sistema financiero en Junio de 2019, que se encontraban en situación crediticia 1 o 2 (es decir, que no tuvieran atrasos mayores a 90 días) y cuyo monto total adeudado en ese momento no superaba los 100.000 pesos argentinos.

Para los cuits de la muestra aleatoria se registraron y resumieron las deudas en todas las entidades en Junio de 2019 y también 6 meses hacia atrás. También se registraron las deudas de esos cuits entre Julio 2019 y Junio 2020 para poder evaluar su evolución.

En el conjunto de datos 'df_bcra_individuals.rds' se encuentra la información registrada, con las siguientes 30 variables:

	variable	detalle
1	id_individuo	identificación del individuo (anonimizada).
2	tipo_persona	primeros dos dígitos del cuit (20: hombres; 27: mujeres).
3	n_deudas_actual	cantidad de entidades en las que el cuit tenía al menos una deuda en Jun-2019.
4	proxy_edad_actual	tres primeros números del dni.
5	deuda_total_actual	monto total de deuda en Jun-2019 (expresada en miles de pesos).
6	deuda_con_garantia_actual	monto total de deuda garantizada en Jun-2019 (expresada en miles de pesos).
7	situacion_mes_actual	situación crediticia más grave en todas las deudas del cuit en Jun-2019.
8	prop_con_garantia_actual	proporción de la deuda garantizada en Jun-2019
9	tiene_garantia_actual	variable indicadora (0: no, 1: si) de si el cuit tenía al menos una deuda garantizada en Jun-2019.
10	mora_30_dias_mes_actual	variable indicadora (0: no, 1: si) de si el cuit estaba en situación 2 en Jun-2019.
11	n_meses_seg_bcra	cantidad de meses en los que el cuit tenía al menos una deuda informada en el sistema financiero, entre Dic-2018 y Jun-2019.
12	media_deuda_total	promedio de la deuda total entre Dic-2018 y Jun-2019.
13	media_deuda_situacion_1	promedio de la deuda en situación 1 entre Dic-2018 y Jun-2019.
14	media_deuda_situacion_2	promedio de la deuda en situación 2 entre Dic-2018 y Jun-2019.
15	media_deuda_con_garantia	promedio de la deuda garantizada entre Dic-2018 y Jun-2019.
16	media_deuda_sin_garantia	promedio de la deuda no garantizada entre Dic-2018 y Jun-2019.
17	media_deuda_en_default	promedio de la deuda en defaut (situación 3 o peor) entre Dic-2018 y Jun-2019.
18	max_situacion_mes	maxima situación entre Dic-2018 y Jun-2019.
19	max_sit_mes_con_garantia	maxima situación en las deudas garantizadas entre Dic-2018 y Jun-2019.
20	max_sit_mes_sin_garantia	maxima situación en las deudas no garantizadas entre Dic-2018 y Jun-2019.
21	media_prop_situacion_1	promedio de la proporción de deuda en situación 1 entre Dic-2018 y Jun-2019.
22	media_prop_situacion_2	promedio de la proporción de deuda en situación 2 entre Dic-2018 y Jun-2019.
23	media_prop_default	promedio de la proporción de deuda en default entre Dic-2018 y Jun-2019.
24	media_prop_con_garantia	promedio de la proporción de deuda garantizada entre Dic-2018 y Jun-2019.
25	prop_tuvo_garantia	proporción de meses en los cuales el cuit tuvo deuda garantizada, entre Dic-2018 y Jun-2019.
26	prop_mora_30_dias_seg	proporción de meses en los cuales el cuit estuvo en situación 2, entre Dic-2018 y Jun-2019.
27	prop_default_seg	proporción de meses en los cuales el cuit estuvo en default, entre Dic-2018 y Jun-2019.
28	peor_situacion_respuesta	situación crediticia más grave en todas las deudas del cuit entre Jul-2019 y Jun-2020.
29	default	situación crediticia más grave mayor o igual 3 en todas las deudas del cuit entre Jul-2019 y Jun-2020.
30	mora_mayor_30_dias	situación crediticia más grave igual 2 en todas las deudas del cuit entre Jul-2019 y Jun-2020.

2. Consignas

a. Describa la distribución univariada de las variables presente en el conjunto de datos. ¿Se evidencian *outliers* en alguna de ellas?
b. Calcule e interprete la matriz de correlaciones de variables disponibles a Jun-2019 (las posibles predictores de default en el período Jul-2019 a Jun-2020).
c. Realice un análisis de componentes principales sobre las mismas variables. ¿Qué porcentaje de la variabilidad total logran explicar las dos primeras componentes? ¿Es posible realizar una interpretación sobre los componentes? ¿Cuál? ¿Logran esas componentes diferenciar a los cuits según el tipo de persona?

d. ¿Existen distintos subgrupos de cuits en los datos? ¿Cuántos logra identificar? ¿Qué características tienen? Explique la metodología utilizada.

e. Construya un modelo predictivo para a variable respuesta 'default' utilizando sólo las variables disponibles a Jun-2019. ¿Qué capacidad predictiva tiene ese modelo?

f. ¿Utilizaría el modelo construido para evaluar futuros solicitantes de un crédito? Justifique su respuesta.