Módulo:

Análisis exploratorio para la construcción de modelos de riesgo de crédito

Agenda

DATA ANALÍTICA

1. Introducción al riesgo de crédito
2. Contexto del caso de uso
3. Análisis exploratorio
    a. Tratamiento de missings, outliers y transformación de variables
    b. Optimal binning y Transformación WoE
4. Construcción de un modelo de riesgo de crédito
    a. Regresión logística

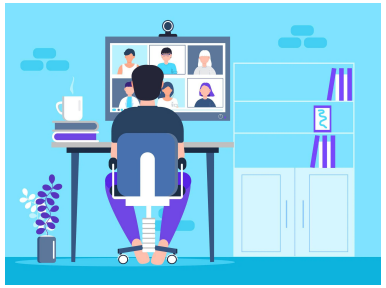
Lenguaje de programación: Python

Reglas del Juego

DATA ANALÍTICA

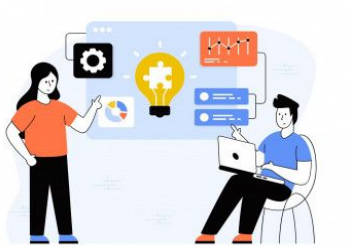
Mantener el micrófono apagado en caso de que no vayamos a hablar.

Nos encantaría verte.
Ten tu cámara encendida y conozcámonos virtualmente.



Preguntar en caso que tengan dudas.

Disfruta de este espacio.
Desconecta del resto y participa.



Por cada clase tendremos 10 min o 15 min de receso.

Modo de Evaluación

20%

Evaluación continua:

Notebooks de ejercicios, formularios de ejercicios o tareas (challenges).

20%

Pruebas de Entrada:

Exámen de 5 a 10 preguntas que serán tomados después del receso de la **4ta sesión**.

60%

Proyecto Final:

Caso aplicativo haciendo uso de las herramientas y aprendizajes obtenidos a lo largo del curso. Presentación y exposición final en la **5ta sesión.**

Proyecto Final

Objetivo:
- Consolidar lo aprendido durante el curso.

Contenido:
- Contexto y casos de uso (Importancia)
- Datos
- Flujo de desarrollo (Limpieza de Datos, Análisis)
- Conclusiones

Tiempo de Exposición: 10 min por grupo

Grupos: De 4 o 5 personas

Entregable:
- Diapositivas
- Notebooks (ordenado y replicable)

Fecha de presentación: Viernes 21 de abril (sesión 5)

Repositorios de Datos:

Kaggle

https://www.kaggle.com/datasets?topic=trendingDataset

UCI ML Repository:

https://archive.ics.uci.edu/ml/datasets.php

https://archive-beta.ics.uci.edu/datasets

INTRODUCCIÓN AL RIESGO DE CRÉDITO

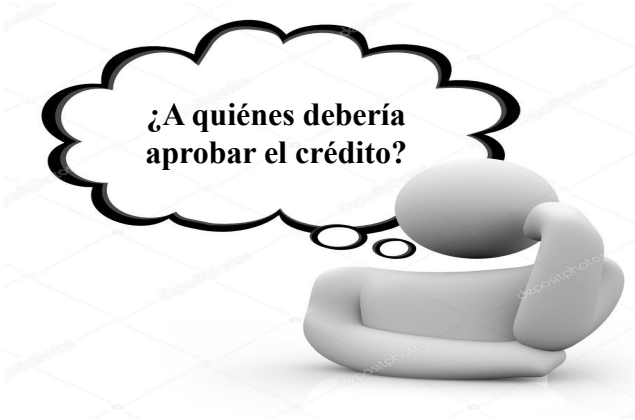
¿Cuál es el negocio de un banco?

Tarjeta de Crédito

Créditos Personales

Créditos Vehiculares

Créditos Hipotecarios



Nota: El deudor o prestatario es aquella persona natural o jurídica que tiene la obligación de satisfacer una cuenta por pagar contraída a la parte acreedora o prestamista

Tipos de modelos de riesgo de crédito

- **Objetivo:** Calcular la **probabilidad de que un cliente incumpla** con sus obligaciones de pago

Applicant	**Behavior**
Calcula la probabilidad de incumplimiento **antes** de ser cliente del banco	Calcula la probabilidad de incumplimiento en el **transcurso** de ser cliente del banco
Modelos de originación o admisión	Modelos de comportamiento

- **Importancia:** Reduce la exposición a riesgos financieros y minimiza las pérdidas por impagos. Un modelo bien diseñado puede identificar los prestatarios de alto riesgo y permitir que las empresas tomen medidas preventivas, como exigir mayores garantías o reducir el monto del préstamo.

Nota: Un modelo es un simplificación de la realidad mediante una representación matemática

Ejemplos de modelos riesgo de crédito

Applicant Hipotecario

1. Variables a probar:
   - ❏ Profesión
   - ❏ Ingreso bruto
   - ❏ Estado civil
   - ❏ Localidad
   - ❏ Promedio depósito en CtaAho 6UM
   - ❏ Antigüedad laboral
   - ❏ Edad, etc

2. Variable objetivo:
   - ❏ La persona se atrasa más de 60 días en los próximos 18 meses en su crédito hipotecario (target binario)

Behavior Crédito personal

1. Variables a probar:
   - ❏ Número de cuotas pagadas 6UM
   - ❏ Ratio entre último pago y cuota
   - ❏ Ratio entre deuda UM y deuda U6M
   - ❏ Clasificación en el SSFF
   - ❏ Promedio saldo en CtaAhorr 3UM
   - ❏ Promedio depósito en CtaAho 6UM
   - ❏ Máximo días atraso U12M, etc

2. Variable objetivo:
   - ❏ La persona se atrasa más de 60 días en los próximos 12 meses en su crédito personal (target binario)

Nota: Un modelo de clasificación se compone de un conjunto de variables y una variable objetivo (binaria o multiclase) a predecir

Caso de uso

Aprobación de tarjetas de crédito

Caso de uso: Aprobación de tarjetas de crédito

- **Fuente**: [Credit Card Approval Prediction | Kaggle]

- **Objetivo**: Construir un modelo de aprendizaje automático para predecir si un solicitante es un cliente "bueno" o "malo", a diferencia de otras tareas, no se da la definición de "bueno" o "malo". Deberá utilizar alguna técnica, como el análisis vintage, para construir su etiqueta. Además, el desequilibrio de los datos es un gran problema en esta tarea.

Nota: Un cliente se dice malo cuando inclumple o se atrasa cierta cantidad de días en un plazo de ciertos meses

Caso de uso: Aprobación de tarjetas de crédito

- **Target o variable objetivo:** Se realizó un análisis para determinar qué variable objetivo nos permitirá medir el incumplimiento adecuadamente.

Target = Persona se atrasa 30 días en una ventana de 12 meses (cliente malo)

0	+1	+2	+3	+4	…	+9	+10	+11	+12	target
ID-1	0	0	5	12	…	0	4	5	5	0
ID-2	0	13	24	**35**	…	25	12	0	5	1

REVISAR EL ARCHIVO: 01.VintageAnalysis.ipynb

Otra solución: EDA & Vintage Analysis | Kaggle

Nota: Un cliente se dice malo cuando inclumple o se atrasa cierta cantidad de días en un plazo de ciertos meses

Caso de uso: Aprobación de tarjetas de crédito

- **Columnas:** 1 identificador y 17 variables

Columnas	Descripción
ID	Número de cliente
CODE_GENDER	Género
FLAG_OWN_CAR	Flag si tiene carro
FLAG_OWN_REALTY	Flag si tiene propiedad (casa)
CNT_CHILDREN	Número de hijos
AMT_INCOME_TOTAL	Ingresos anuales
NAME_INCOME_TYPE	Categoría de ingresos
NAME_EDUCATION_TYPE	Nivel de educación
NAME_FAMILY_STATUS	Estado civil
NAME_HOUSING_TYPE	Modo de vivir (Condición vivienda)

Columnas	Descripción
DAYS_BIRTH	Días de nacimiento (Cuenta hacia atrás desde el día actual (0), -1 significa ayer)
DAYS_EMPLOYED	Antigüedad laboral (Contar hacia atrás desde el día actual(0). Si es positivo, significa que la persona está actualmente desempleada.)
FLAG_MOBIL	Flag si tiene celular
FLAG_WORK_PHONE	Flag si tiene teléfono de trabajo
FLAG_PHONE	Flag si tiene teléfono
FLAG_EMAIL	Flag si tiene email
OCCUPATION_TYPE	Ocupación
CNT_FAM_MEMBERS	Tamaño de la familia

Análisis exploratorio

Aprobación de tarjetas de crédito

Tratamiento de missings

¿Por qué faltan datos en el conjunto de datos?

- Mantenimiento inadecuado de la información
- Error en la digitación de los datos
- Se han proporcionado datos incorrectamente
- Se negó a responder

Ojo: En ciertos contextos un missing tiene un significado o sentido de negocio.

¿Por qué debemos preocuparnos por el manejo de los datos faltantes?

- Falta de precisión en el análisis estadístico
- Muchos algoritmos fallan si el conjunto de datos contiene valores faltantes

Nota: Los modelos clásicos de regresión (lineal y logística) no admiten variables con missing (se les debe colocar un valor). Mientras que los modelos relacionados con árboles (árbol de decisión, random forest, xgboost, etc) admiten variables con missings

Tratamiento de atípicos

¿Qué causa los valores atípicos?

- Error en la digitación del dato
- Dato natural (ejem: personas con altos ingresos)
- Sesgo en la recolección de datos (ejem: en la altura de deportistas se incluyeron datos de jugadores de baloncesto).

Nota: Esto se detecta con un histograma o gráfico de densidad donde veamos 2 o más modas.

¿Por qué debemos preocuparnos por los valores atípicos?

- Puede influir y afectar a nuestras estimaciones en los modelos clásicos
- Afecta a los estadísticos descriptivos como promedio, desviación estándar, correlaciones, etc.
- Afecta a los supuestos que pueda tener un modelo.

Transformación de variables

¿Qué es la transformación de variables?

Reemplazar una variable por una transformación de la misma (ejem: Raíz cuadrada, logaritmo, estandarización, WoE, etc)

¿Cuándo debemos utilizar la transformación de variables?

- Se quiere corregir la distribución (sesgada a simétrica)
- Cambiar o uniformizar la escala **!Muy importante en los modelos clásicos de regresión!**

¿Cuáles son los métodos más comunes?

- Logaritmos
- Raíz cuadrada o cúbica
- Binning (Transformar variables numéricas en categóricas)

¿Qué buscaremos con el análisis exploratorio?

Procedimiento clásico

- Tratamiento de atípicos según percentiles
- Tratamiento de missings según imputación por algún valor (mediana, moda, etc)
- Transformar de variables según una estandarización (numéricas) y variables dummy u otra transformación (categóricas)

Procedimiento para modelos de riesgo

- Utilizar la función **OptimalBinning** (Agrupaciones óptimas) para tratar atípicos y missings, además de transformar variables según valores WoE (weight of evidence) y obtener su poder predictivo según el IV (information value). Este procedimiento sirve para variables numéricas y categóricas.

¿Qué finalidad tiene el OptimalBinning?

- Por medida regulatoria de la SBS se pide que las variables de un modelo de riesgo de crédito tengan **sentido de económico esperado (tendencia monótona con el target)**
- La función **OptimalBinnig** genera grupos óptimos de tal manera que se respeta la tendencia monótona del target, es robusto a valores atípicos, se le asigna un valor (WoE) para transformar la variable original, los valores missing pueden ser imputados según estos valores y además se reporta el poder predictivo de la variable según el IV.

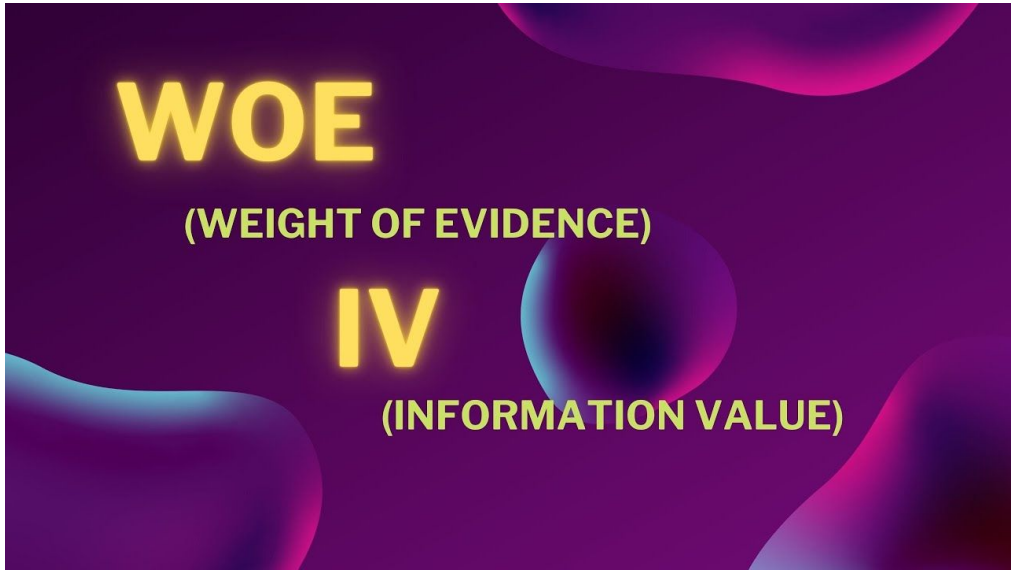
EDAD_CLIENTE	Count	Count (%)	Non-event	Event	Event rate	WoE	IV
0. (-inf, 29.50)	83,113	15%	71,539	11,574	**14%**	**0.676**	9%
1. [29.50, 40.50)	219,774	40%	201,901	17,873	**8%**	**0.073**	0%
2. [40.50, 57.50)	183,664	34%	173,263	10,401	**6%**	**-0.315**	3%
3. [57.50, inf)	57,741	11%	56,210	1,531	**3%**	**-1.106**	8%
4. Special	0	0%	0	0	0%	0	0%
5. Missing	0	0%	0	0	0%	0	0%
Total	544,292	100%	502,913	41,379	8%		**21%**

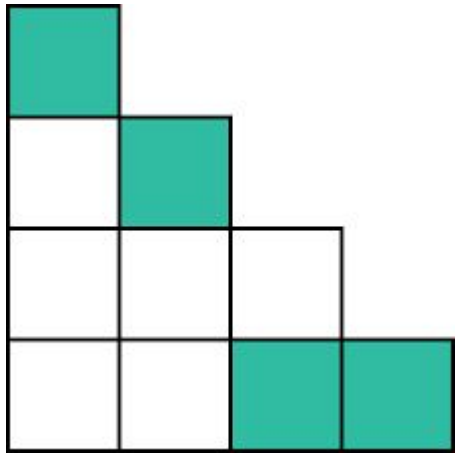
Information Value	Predictive Power
<0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
> 0.5	Suspicious or too good predictor

Nota: Tendencia monótona quiere decir que la variable sigue una tendencia creciente o decreciente

A codear!!!

Referencias





- Weight of Evidence (WOE) and Information Value (IV) Explained
- Weight of Evidence (WoE) and Information Value (IV) - how to use it in EDA and Model Building?
- Understand Weight of Evidence and Information Value! - Analytics Vidhya

Tutorial: optimal binning with binary target — optbinning

QUIZ TIME

GRACIAS

DATA ANALÍTICA