INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

PAP PROGRAMA DE MODELACIÓN MATEMÁTICA PARA EL DESARROLLO DE PLANES Y PROYECTOS DE NEGOCIO I

ITESO, Universidad Jesuita de Guadalajara

Primera Entrega

Proyecto 3: Modelado de Créditos con Machine Learning en Atrato

Presentan

Paulina Gomez Heredia 734173

Marian Montserrat Sedano Paz 734838

Profesores: Luis Felipe Gómez Estrada y Sean Nicolás Gonzáles Vázquez

Fecha: 08/02/2025

TABLA DE CONTENIDOS

GLOSARIO

Definiciones de términos clave utilizados en la presente investigación, con el objetivo de brindar claridad y uniformidad en su interpretación.

1. **Inteligencia Artificial:** rama de la informática que permite programar sistemas para realizar tareas de *inteligencia humana*, como aprender, razonar y tomar decisiones.
2. **Machine Learning:** rama de la Inteligencia Artificial que permite a desarrollar modelo y algoritmos capaces de aprender de los datos y hacer predicciones de estos, sin necesidad de ser programados paso a paso.
3. **Fintech:** empresas que combinan finanzas y tecnología para ofrecer servicios financieros *innovadores*.
4. **Collection Score Model:** modelo de machine learning diseñado para predecir la probabilidad de que un cliente en mora avance a un estado de atraso mayor.
5. **Bucket:** Técnica de agrupación que divide un conjunto de datos en rangos o categorías para su análisis.
6. **Funnel de conversión:** Representación del proceso por el cual una solicitud avanza desde su creación hasta su aprobación o rechazo.
7. **KPI (Key Performance Indicator):** Indicador clave de rendimiento utilizado para medir el éxito de un proceso.

PRIMERA ENTREGA

INTRODUCCIÓN

El *Machine Learning* (ML) es una rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y tomar decisiones sin necesidad de una programación explícita. Su capacidad para analizar grandes volúmenes de información y generar predicciones precisas lo ha convertido en una tecnología fundamental en diversas industrias, especialmente en el sector financiero, donde facilita la toma de decisiones basadas en datos y la optimización de procesos clave.

En la actualidad, el uso de ML en el ámbito empresarial, particularmente en *startups* y *fintechs*, ha transformado la manera en que se analizan los riesgos, se personalizan las ofertas y se optimizan operaciones. En **Atrato**, la implementación de modelos de ML ha representado una oportunidad clave para el desarrollo de soluciones innovadoras de procesos y aplicaciones claves:

- Modelos de riesgo crediticio
- Prevención de fraude
- Optimización de recuperación de cartera
- Personalización de ofertas
- Aprobación de créditos

Los Collection Scores, o modelos de cobranza, son modelos o algoritmos de Machine Learning diseñados para predecir la probabilidad de que un cliente en mora realice un pago dentro de un período determinado. Estos modelos se basan en técnicas de análisis estadístico y segmentación de clientes según su comportamiento de pago, permitiendo una gestión más eficiente de la cartera vencida.

Los modelos de cobranza son herramientas clave en la gestión de cartera vencida; pues nos permiten optimizar las estrategias empleadas para recuperar pagos de manera eficiente y minimizar pérdidas financieras para la organización. En el contexto de **Atrato**, la implementación de un Modelo de Collection Score tiene como objetivo predecir qué clientes en mora tienen mayor riesgo de avanzar a un estado de atraso más grave. Esto permite al equipo de cobranza priorizar esfuerzos, asignar recursos estratégicamente y optimizar la recuperación de pagos, reduciendo pérdidas financieras y mejorando la sostenibilidad del negocio.

Al adoptar un enfoque basado en datos, el uso de Machine Learning en la gestión de cobranza no solo mejora la eficiencia operativa, sino que también permite desarrollar

estrategias más efectivas y personalizadas, aumentando la probabilidad de recuperación de cartera y fortaleciendo la relación con los clientes.

CONTEXTO DEL PROBLEMA Y JUSTIFICACIÓN

En el sector fintech, la gestión eficiente del crédito es fundamental para garantizar la rentabilidad del negocio y minimizar pérdidas. Uno de los principales desafíos en este ámbito es el crecimiento de la cartera vencida, que impacta directamente la liquidez y sostenibilidad de la empresa.

El aumento de clientes en mora puede deberse a diversos factores: inestabilidad económica, cambios en el comportamiento de pago y falta de herramientas para gestión, etc. En muchas organizaciones, los procesos de recuperación de cartera aún dependen de reglas no escritas o estrategias genéricas que no consideran la probabilidad real de pago de cada cliente: desde condonación de intereses y comisiones hasta reducción de deuda.

La administración de la cartera vencida es uno de los principales desafíos en la gestión financiera de Atrato; el equipo de cobranza tiene como objetivo buscar recuperar la máxima cantidad de cuotas en atraso posibles. Para afrontar este reto, es fundamental tener a la mano herramientas que permitan priorizar y optimizar la recuperación de pagos. Un enfoque basado en datos y modelos predictivos puede marcar la diferencia en la eficiencia del proceso de cobranza y en la reducción de pérdidas financieras.

Ante esta problemática, se propone la implementación de un modelo de Machine Learning basado en Collection Scores, que permitirá optimizar las estrategias de cobranza mediante la segmentación de clientes según su perfil y comportamiento financiero. Este enfoque ayudará a priorizar esfuerzos, mejorar la recuperación de pagos y reducir pérdidas financieras.

En este contexto, el uso de Machine Learning en cobranza representa una ventaja competitiva clave para mejorar la rentabilidad y sostenibilidad del negocio.

OBJETIVOS

General

Diseñar e implementar un modelo de cobranza basado en Machine Learning, basado en el área de Ciencia de Datos, que permita predecir la probabilidad de roll-rate de los clientes en mora, con el fin de optimizar la estrategia de cobranza en Atrato.

Específicos

Analizar y estructurar la información histórica de cobranza, identificando variables relevantes para la predicción del modelo.

Desarrollar y entrenar modelos de Machine Learning que segmenten a los clientes según su probabilidad de roll-rate.

Evaluar el desempeño del modelo utilizando métricas a según lo analizado.

Implementar el modelo en el flujo operativo de cobranza, integrándose con las herramientas existentes de la empresa.

Medir el impacto del modelo de acuerdo a métricas claves.

BENEFICIOS ESPERADOS

Con el desarrollo del modelo de cobranza para Atrato se busca cumplir los siguientes beneficios:

1. Mejorar la toma de decisiones basadas en datos y predicciones, en lugar de depender de reglas manuales.
2. Optimizar la asignación de recursos del equipo de cobranza, de manera que puedan mejorar su rendimiento.
3. Reducir costos operativos asociados a la recuperación de cartera y contratación de despachos externos.
4. Incrementar la eficiencia en la gestión de morosidad, asegurando estabilidad financiera.

MÉTRICAS CLAVE

El dashboard se centrará en el monitoreo de las siguientes métricas, proporcionando una visión completa del ciclo de vida de las solicitudes de crédito y del comportamiento del usuario en Atrato en pagos pasados:

- **Tasa de aprobación de créditos:** Porcentaje de solicitudes aprobadas respecto al total recibido.
- **Demanda de solicitudes:** Número total de solicitudes de crédito recibidas en un periodo específico.
- **Estado de créditos activos:** Distribución de los créditos según su estado (vigente, en mora, en cobranza).
- **Evolución de pagos:** Seguimiento del comportamiento de pagos de los clientes a lo largo del tiempo.
- **Take Rate:** Porcentaje de los créditos aprobados que son finalmente firmados y otorgados.
- **Cartera vigente:** Total de créditos que están al día en sus pagos.
- **Cartera en mora:** Total de créditos que presentan algún grado de morosidad.
- **Probabilidad de Roll-Rate:** Predicción de la probabilidad de que los créditos en mora avancen a estados de mayor morosidad. (MODELO)
- **Población por bucket de mora:** Segmentación de la cartera según el número de días en mora (por ejemplo, 30, 60, 90 días).

** análisis prescripcitvo

Requerimientos del sistema de visualización

- Base de datos interna de Atrato para alimentar el dashboard.
- Modelado de datos para la construcción de métricas.
- Diseño del dashboard en Figma.
- Diseño y construcción del dashboard en Power BI.

CASOS DE USO

Caso de Uso: Monitoreo de Solicitudes de Crédito en Dashboard
ID único: CU01
Nombre: Monitoreo de Solicitudes de Crédito
Descripción: Los equipos de producto y comercial pueden visualizar en tiempo real el estado de las solicitudes de crédito en un dashboard de Power BI. Se presentan

métricas clave como el total de solicitudes, tasas de aprobación y conversión en cada etapa del proceso.

Precondiciones:

- Atrato debe contar con un sistema que registre las solicitudes de crédito y sus estados financieros.
- Los datos deben estar disponibles en la base de datos.
- Los usuarios deben tener acceso al dashboard con los permisos adecuados.

Post-condiciones:

- Los usuarios pueden visualizar métricas actualizadas hasta el cierre del día anterior.
- Se identifican cuellos de botella en el proceso de aprobación.
- Se facilita la toma de decisiones basada en datos.

Trigger:

- Un usuario accede al dashboard para consultar métricas de solicitudes de crédito.

Actores:

- Equipo Producto: Analiza el funnel de conversión y puntos de fricción.
- Equipo Comercial: Da seguimiento a clientes en base al estado de sus solicitudes.
- Analistas de Riesgo: Evalúan patrones en la aprobación/rechazo de solicitudes.

Escenario Principal:

1. El usuario accede al dashboard en Power BI.
2. Filtra por fecha o tipo de cliente.
3. Visualiza KPIs clave:
    - Total de solicitudes creadas, aprobadas, rechazadas.
    - Tiempos promedio en cada etapa del proceso.
    - Conversión en cada paso del funnel.
4. Identifica posibles cuellos de botella o tasas de abandono altas.
5. Toma decisiones basadas en los datos presentados.

Escenarios Alternativos:

- Los datos no están actualizados: Se muestra un mensaje indicando el último refresco de datos.
- El usuario no tiene permisos: Se le deniega el acceso y se le redirige a soporte.

Excepciones:

- La base de datos está caída o no responde.
- Power BI tiene un error en la carga del informe.

Reglas de Negocio:

- Los datos disponibles en el dashboard corresponden hasta el cierre del día anterior.
- Los accesos deben estar restringidos por usuarios y roles de Power BI.
- La visualización debe cumplir con los estándares definidos en el diseño de Figma

POBLACIÓN

Se definen 3 buckets de días de atraso conforme a los puntos de corte definidos por el equipo de cobranza de Atrato:

- 3 - 6 días
- 7 - 20 días
- 21 - 51 días

La base se construye seleccionando a los usuarios que pisen el atraso correspondiente al bucket de atraso en el que se encuentre, y la features se toman en el momento que pisan el atraso, hacia atrás.

Para entrenar el modelo buscamos atrapar en la base de entrenamiento a todos los clientes que puedan:

I. Clasificarse en algún bucket de atraso definido previamente
II. Que haya transcurrido el tiempo suficiente para determinar si migró al siguiente bucket o no (que tenga un target observable)

VARIABLES

Algunas de las variables consideradas y pensadas para el modelo son:

current_credit_delay_trend	Tendencia de atrasos en pagos en crédito actual de Atrato
average_current_credit_delay	Promedio de atrasos en pagos en crédito actual de Atrato
ratio_overdue_payments	Periodos entre atrasos, veces que se atrasó / número de parcialidades del crédito del crédito actual
max_overdue_amount	Máximo monto atrasado, este incluye la suma de comisiones, intereses moratorios, etc...
granted_credit_amount	Monto del crédito otorgado.
pct_credit_maturity	Porcentaje de avance en su crédito. (n° cuotas pagadas / totales) Madurez
risk_profile_v1	PR1

risk_profile_v2	PR2
num_atrato_credits	Número de créditos en Atrato (HIST)
early_settled_atrato_credits	Número de créditos en Atrato que se liquidaron de manera anticipada.
monthly_payment	monthly payment
owed_amount_cc	Monto adeudado en CC --> créditos , descartar tarjetas
outstanding_atrato_debt	Monto adeudado en ATRATO
has_mortgage_overdue	Atrasos en hipoteca --> binaria
monthly_payment_terms	Número de plazos pactados mensualizado--> (total)
age	Edad
gender	Género
avg_period_between_overdue	Overdue: periodos de atrasos cometidos (cada cuánto te atrasas)
payments_before_first_delay	N° pagos antes de primer atraso
payments_before_worst_delay	N° pagos antes de peor atraso
days_of_worst_overdue	días de peor atraso (de este crédito para atrás, incluye crédito actual mas no última parcialidad pagada)
has_mortgage_or_large_loans	tiene Hipotecas / créditos grandes --> autos, casas,
current_debt	Atraso entre deuda (monto que falta de pagar / monto crédito total)
num_bucket_1	N° veces que ha caído en bucket 1, contando todos sus créditos en Atrato
num_bucket_2	N° veces que ha caído en bucket 2, contando todos sus créditos en Atrato
num_bucket_3	N° veces que ha caído en bucket 3, contando todos sus créditos en Atrato
dashboard_inquiries	Consultas a dashboard
most_used_payment_method	Método de pago más usado
historical_overdue_atrato_credits	n° de créditos en los que se ha atrasado en Atrato
num_roll_up	n° de veces que el cliente migró de un bucket de atraso a uno más alto
num_roll_down	n° de veces que el cliente migró de un bucket de atraso a uno más bajo
num_expected_payments	n° pagos que debería llevar sin atrasos (basado en fecha de inicio de crédito, parcialidades pactadas, plazo pactado y fecha actual)
current_debt	% de lo que aún debe del crédito actual, al momento del corte de la información
max_overdue_amount	la cantidad mayor de atraso que ha tenido
montlhy_payment	el pago pactado mensual del crédito, sin comisiones ni otras penalizaciones
max_overdue_amount_to_monthly_payment_ratio	max_overdue_amount / montlhy_payment
dashboard_inquires	n° de veces que ha consultado su dashboard
days_of_worst_overdue	n° de días de su peor atraso en ATRATO, a nivel crédito NO USER

WOE & IV

El WOE y el IV son medidas utilizadas en algoritmos de ML para determinar qué tan fuerte o poderosa es una variable independiente. Nos ayudan a entender si una determinada clase de una variable independiente tiene una mayor distribución de buenos o malos, y su importancia. Su aplicación y uso es principalmente en modelos de riesgo crediticio para evaluar la capacidad predictiva de una variable independiente respecto a una variable objetivo binaria.

Weight of Evidence (WOE)

Evalúa la capacidad predictiva de una variable. Se basa en la separación entre buenos clientes (pagan) y malos clientes (no pagan). Para calcularlo, se usa el logaritmo natural del porcentaje de *buenos* dividido por el porcentaje de *malos*.

$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right) \qquad WOE = \ln\left(\frac{\%\text{ of non-events}}{\%\text{ of events}}\right)$$

Para interpretar el WOE debemos entender que un WOE positivo nos indica más buenos que malos clientes, y un WOE negativo, lo contrario. Es importante también tratar cada variable según sus particularidades:

1. variables continuas: dividir en bins (recomendable 10). Calcular el número y porcentaje de *buenos* y *malos* en cada bin. Obtener el WOE aplicando el logaritmo natural de la proporción obtenida
2. variables categóricas: agrupar categorías similares según el WOE.

Para la separación de bin y uso de WOE debemos tomar en cuenta algunas reglas:

- Cada bin debe tener al menos 5% de los datos para evitar inestabilidad.
- Ningún bin debe tener solo eventos o solo no eventos

  Si un bin no tiene eventos o no eventos, se ajusta con

$$ln\left(\frac{\frac{(Non-events + 0.5)}{Number\ of\ non-events}}{\frac{(Number\ of\ events\ in\ a\ group + 0.5)}{Number\ of\ events}}\right)$$

- WOE debe ser distinto por bin

- WOE debe ser monótono (creciente o decreciente)
- Valores faltantes deben agruparse en un bin separado

Algunas de las ventajas de usar WOE son mejor manejo de valores atípicos (outliers) y de valores nulos, reducción de dimensionalidad, facilita interpretación con log-odds...

Information Value (IV)

Mide la importancia de una variable en la predicción de la variable objetivo. Para calcularlo, se obtiene la suma, a través de todos los grupos o bins de una variable, del producto entre la diferencia de las proporciones de buenos y malos y el Weight of Evidence (WoE) de cada grupo.

$$IV = \sum (\%\ \text{of non-events} - \%\ \text{of events}) * WOE$$

Para la interpretación del IV tenemos un conjunto de reglas:

IV	Interpretación
< 0.02	No útil para predicción
0.02 - 0.1	Poder predictivo débil
0.1 - 0.3	Poder predictivo medio
0.3 - 0.5	Poder predictivo fuerte
> 0.5	Sospechoso (leakage)

*** los valores resaltados serían los mejores posibles*

Algunos puntos a considerar cuando utilizamos el IV son:

- El IV aumenta con el número de bins, hay que cuidar que al tener un alto número de bins (20+) pueden tener muy pocas observaciones, lo que puede afectar la calidad del análisis.
- Este método no es el mejor para la selección de variables en modelos de clasificación distintos a la regresión logística binaria (no óptimo para DT, RF, XGB, SVM)

WoE e IV se aplican en la fase de preprocesamiento de datos y selección de características, antes de entrenar el modelo. Al aplicar estos métodos podemos esperar mejor estabilidad y desempeño del modelo al reducir la variabilidad (hace que la relación entre variables sea más lineal), eliminación de variables irrelevantes (descartar con IV), reducción de colinealidad y mejora en la interpretabilidad. El IV y el WoE están diseñados específicamente para regresión logística, ya que están basados en las razones de probabilidades log-odd.

BIBLIOGRAFÍA

Iberdrola (s.A.). *Qué es el 'machine learning'* en Iberdrola. Disponible en: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

Deepanshu Bhalla (2015). *Weight of Evidence (WOE) and Information Value (IV) Explained* en Listen Data. Disponible en: https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html