REPORTE FINAL: MODELO DE RIESGO CREDITICIO Y TASAS DINÁMICAS

1. Objetivo

El objetivo principal de este proyecto fue desarrollar un modelo predictivo de riesgo crediticio para Bankaya, utilizando información interna y externa de los clientes, con el fin de generar un score de riesgo y, en base a éste, establecer una asignación dinámica de tasas de interés. Esto permite una mejor segmentación de clientes, optimización de ingresos y control de morosidad.

2. Descripción de Variables

Dataset Principal (main_dataset.parquet)

Contiene información detallada de las solicitudes de préstamos de Bankaya, incluyendo datos del cliente y su interacción con la plataforma.

- customer_id: Identificador único del cliente.
- loan_id: Identificador único del préstamo.
- ACC_CREATION_DATETIME Fecha de creación de la cuenta del cliente.
- APPLICATION_DATETIME Fecha en la que se solicitó el préstamo.
- LOAN_ORIGINATION_DATETIME: Fecha en que el préstamo fue aprobado o iniciado.
- max_days_late: Máximo número de días que el cliente se atrasó en un pago.
- target: Variable objetivo original (0: buen comportamiento, 1: mal comportamiento).
- account_to_application_days: Días entre la creación de cuenta y la solicitud del préstamo.

- n_sf_apps: Número de solicitudes previas en la plataforma “SF” (no siempre presente).
- first_app_date: Fecha de la primera solicitud de crédito registrada.
- last_app_date: Fecha de la última solicitud de crédito registrada.
- n_bnpl_apps: Número de aplicaciones tipo “Buy Now Pay Later” hechas por el cliente.
- n_bnpl_approved_apps: Número de esas aplicaciones que fueron aprobadas.
- first_bnpl_app_date: Fecha de la primera solicitud BNPL.
- last_bnpl_app_date: Fecha de la última solicitud BNPL.
- n_inquiries_l3m: Número de consultas de crédito en los últimos 3 meses.
- n_inquiries_l6m: Número de consultas de crédito en los últimos 6 meses.

Dataset de Reportes de Crédito (credit_reports.parquet)

Este dataset contiene el historial crediticio externo de los clientes, donde cada fila representa un registro de crédito específico del cliente con diversas entidades financieras. Un mismo customer_id puede tener múltiples entradas en este dataset.

- -customer_id: Identificador único del cliente (clave de unión con main_dataset).
- -REPORT_DATE: Fecha de generación o actualización del reporte de crédito.
- -LOAN_OPENING_DATE: Fecha de apertura del crédito externo.
- -LOAN_CLOSING_DATE: Fecha de cierre o terminación del crédito externo.
- -CREDIT_TYPE: Tipo de crédito (ej., tarjeta de crédito, préstamo personal, hipoteca).
- -PAYMENT_FREQUENCY: Frecuencia de los pagos de este crédito (ej., mensual, semanal).
- -MAX_CREDIT: Monto máximo de crédito aprobado para esta línea de crédito.
- -CREDIT_LIMIT: Límite de crédito asignado para esta línea de crédito.
- -PAYMENT_AMOUNT: Monto del pago más reciente registrado.
- -CURRENT_BALANCE: Saldo actual pendiente de pago en esta línea de crédito.
- -BALANCE_DUE: Monto total vencido o adeudado.
- -BALANCE_DUE_WORST_DELAY: Monto máximo que estuvo vencido o adeudado en el peor momento de atraso.
- -DELAYED_PAYMENTS: Número de pagos que el cliente ha atrasado en esta cuenta.
- -WORST_DELAY: El peor número de días de atraso registrado para este crédito.

- -WORST_DELAY_DATE: Fecha en que se registró el peor atraso.
- -TOTAL_PAYMENTS: Número total de pagos realizados para esta cuenta.
- -TOTAL_REPORTED_PAYMENTS: Número total de pagos reportados a las agencias de crédito.
- -UPDATE_DATE: Fecha de la última actualización de este registro de crédito.
- -LAST_PURCHASE_DATE: Fecha de la última compra o disposición de crédito.
- -LAST_PAYMENT_DATE: Fecha del último pago registrado.

- Variables clave:

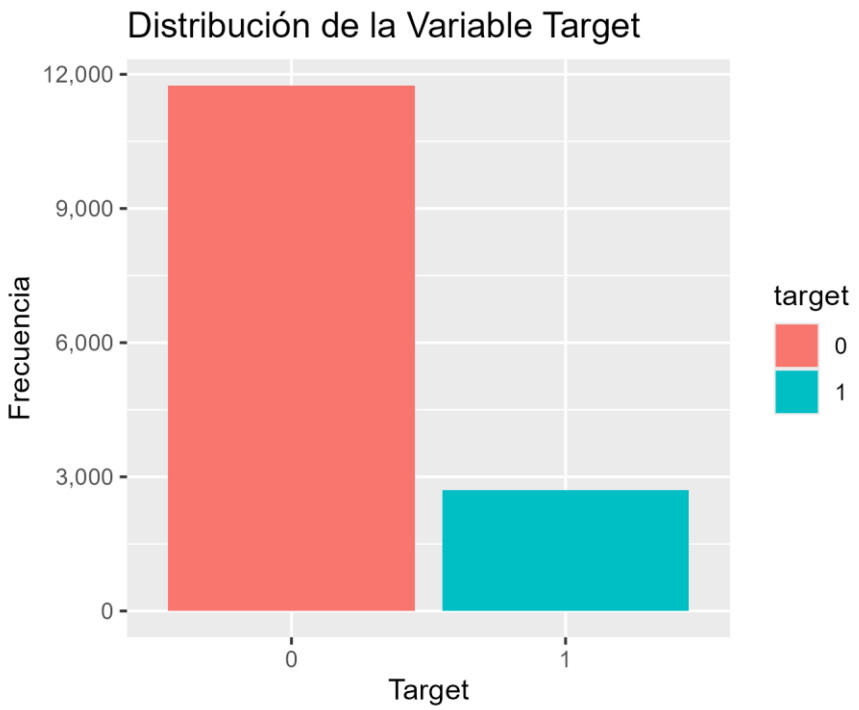
- max_days_late
- n_sf_apps`, `n_bnpl_apps`, `n_inquiries_l3m`: Actividad previa del cliente.
- `uso_credito_pct`, `pagos_tarde_pct`: Derivadas del historial crediticio.

3. Limpieza y Tratamiento de Datos

- Valores negativos en `max_days_late` fueron reemplazados por 0.
- Variables con NAs significativos que indican ausencia de actividad se imputaron con 0.
- Se eliminaron variables no informativas como IDs únicos (`customer_id`, `loan_id`) y fechas exactas.
- Variables categóricas con más de 53 niveles fueron descartadas por limitaciones técnicas del modelo Random Forest.
- Datos de crédito externo fueron agregados a nivel `customer_id` para integrarse con el dataset principal.

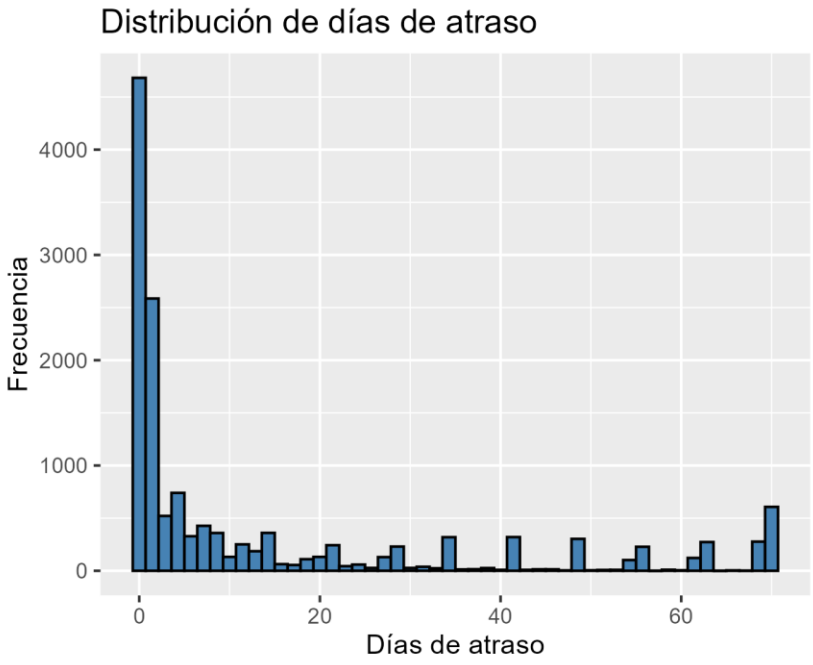
4. Visualizaciones y Análisis Exploratorio

- Distribución de la variable target:



Muestra un desbalance significativo. La mayoría de los clientes están clasificados como "buen comportamiento" (target = 0), lo cual indica la necesidad de balancear las clases para el entrenamiento del modelo.

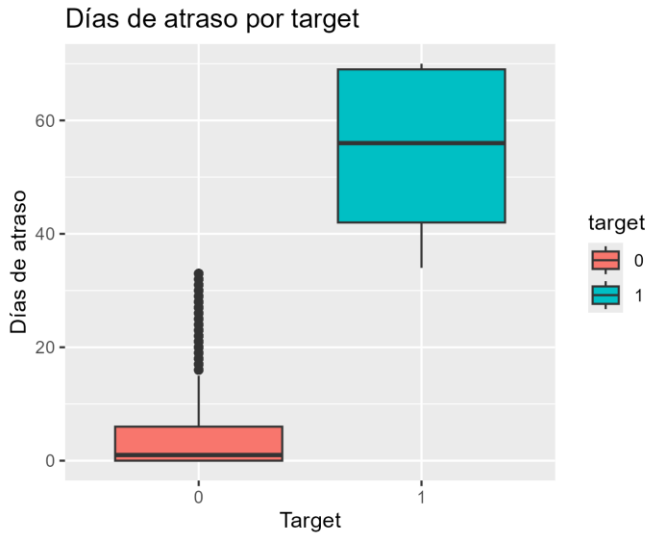
- Histograma de “max_days_late”



e`:

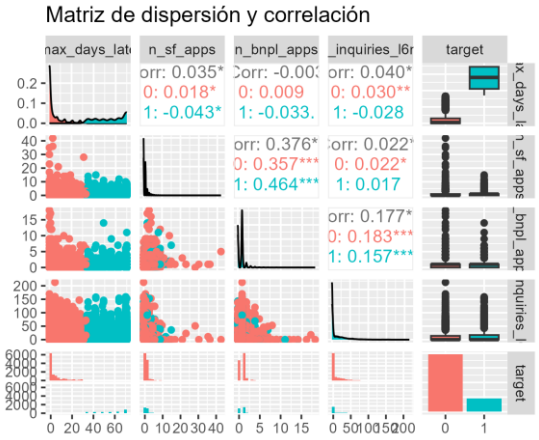
La mayor parte de los clientes se atrasan muy poco en sus pagos. Existen pocos casos con altos números de días de atraso, lo que sugiere una distribución sesgada hacia la izquierda y posible presencia de valores atípicos.

- Boxplot de `max_days_late` por `target`:



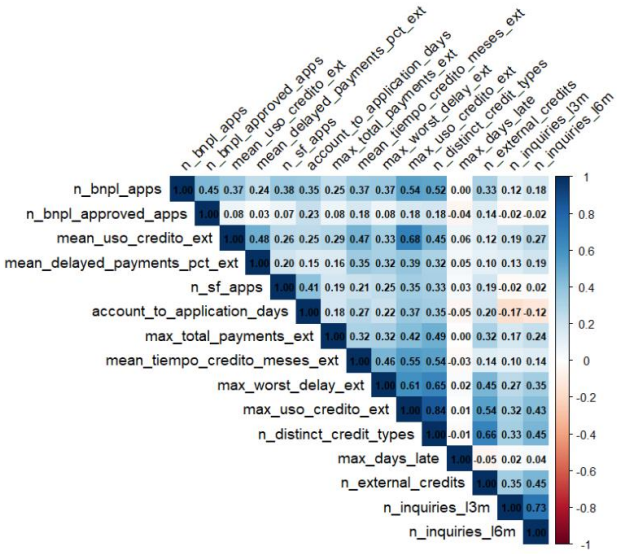
Se observan valores extremos principalmente en la categoría target = 0. Esto puede dificultar la clasificación, ya que hay traslape entre los grupos. La mediana es baja para ambos grupos, pero la categoría 1 presenta mayor simetría.

- Matriz de correlación y pairplot:



Las variables no presentan multicolinealidad extrema. `max_days_late` tiene la mayor separabilidad entre clases, por lo que es una de las más relevantes para la clasificación.

-Matriz de correlación del main_set con las variables externas agregadas



Las únicas variables correlacionadas fuertemente son: max_uso_crédito_ext y n_distinct_credit_types, sin embargo, no es tan alto como para preocuparnos, se mantendrán todas las variables.

5. Modelado Predictivo

Se construyeron y evaluaron los siguientes modelos:

- Regresión Logística

- Árbol de Decisión (rpart)

- XGBoost

- Random Forest

- Red Neuronal (nnet)

Tabla Comparativa de Modelos (con características de external credit)

Modelo	Accuracy	Kappa	F1	Precision	Recall
Regresión Logística	0.9927335640138408	0.9764403088497328	0.9809264305177112	0.9625668449197861	1
Árbol de Decisión	0.9913494809688581	0.9720313558501886	0.9773755656108597	0.9557522123893806	1
XGBoost	0.9889273356401385	0.9643748651730409	0.9712230215827338	0.9440559440559441	1
Random Forest	0.986159169550173	0.955715599141894	0.9642857142857143	0.9310344827586207	1
Neural Net	0.9826989619377162	0.9450256800456532	0.9557522123893806	0.9152542372881356	1

Previous 1 Next

Todos los modelos lograron buenos resultados

5.1 Evaluación de Resultados y Matrices de Confusión

Interpretación general:

La matriz de confusión nos permite observar los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Un número alto de falsos negativos (clientes con mal comportamiento clasificados como buenos) es el error más costoso, ya que implica conceder crédito a quien probablemente no pagará. Por ello, métricas como el Recall son prioritarias.

Errores y su impacto:

- Falsos negativos (FN): Riesgo de pérdida financiera al aprobar clientes malos.
- Falsos positivos (FP): Rechazo o penalización a buenos clientes, afectando la experiencia del usuario.

Ejemplos:

Regresión logística

```
          Reference
Prediction    0    1
         0 2329    0
         1   21  540
```

Tenemos 21 errores clasificando como rechazo cuando se debía aceptar.

Árbol de decisión

```
          Reference
Prediction    0    1
         0 2325    0
         1   25  540
```

25 errores del mismo tipo que regresión logística.

6. Cálculo del Score de Riesgo utilizando los resultados de XGBoost

El modelo XGBoost entrega una probabilidad estimada de que un cliente tenga mal comportamiento (`target = 1`). Esta probabilidad se asigna como `riesgo_score_final`, un valor entre 0 y 1.

- Bajo score = bajo riesgo de incumplimiento.

- Alto score = mayor riesgo de incumplimiento.

7. Asignación Dinámica de Tasas de Interés

La tasa de interés por cliente se definió como una función creciente del `riesgo_score_final`:

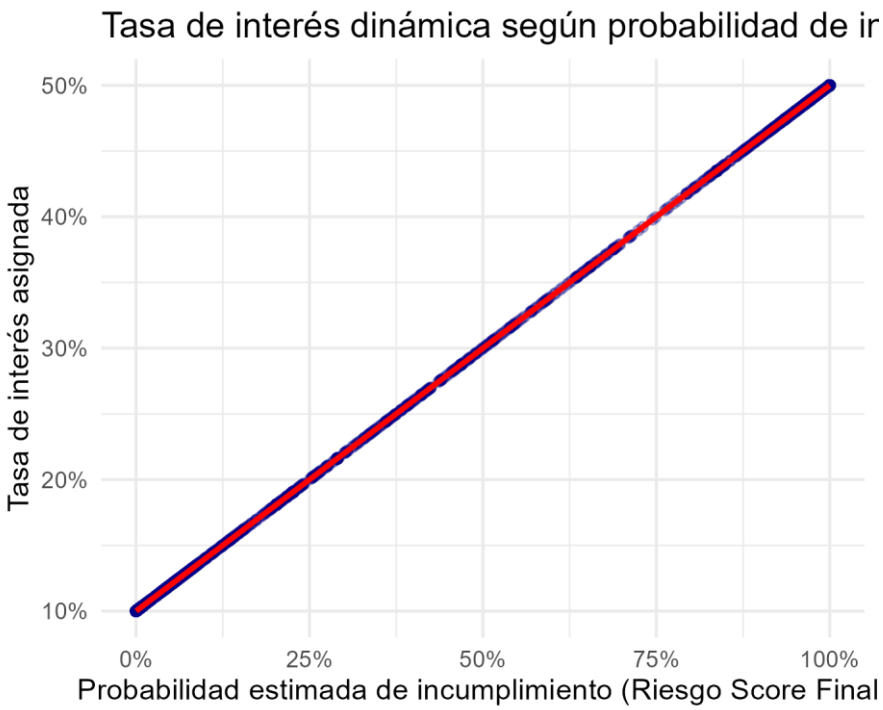
```r
# Tasa de interés dinámica
tasa_base <- 0.10  # Mínimo 10%
tasa_max <- 0.50   # Máximo 50%

# Cálculo dinámico de la tasa individual basado en el riesgo_score_final (probabilidad de incumplimiento)
df_final_model <- df_final_model %>%
  mutate(tasa_interes_dinamica = tasa_base + riesgo_score_final * (tasa_max - tasa_base))
```

Esto permite:

- Clientes de bajo riesgo reciben tasas bajas.

- Clientes de alto riesgo son penalizados con tasas más altas.



Podemos observar como se va asignando una tasa de interés de forma lineal dependiendo del score de riesgo

8. Conclusiones

- Se construyó un pipeline robusto de limpieza, integración, modelado y evaluación.
- En los modelos de aprendizaje automático se ve un error en la clasificación de las variables de tipo 1, mientras que para 0 prácticamente todos tienen el 100% de efectividad.
- Seleccionamos el modelo XGBoost para continuar con los cálculos y asignación de tasas de interés
- Este modelo proporciona a Bankaya una herramienta analítica poderosa para automatizar y mejorar las decisiones crediticias. La capacidad de discernir el riesgo con mayor precisión se traducirá en:
- Reducción de Morosidad: Al identificar mejor a los clientes riesgosos, se pueden rechazar solicitudes o asignar tasas que compensen el riesgo.
- Expansión Cautelosa del Mercado: Permite aprobar clientes con un riesgo moderado a una tasa adecuada, ampliando la base de clientes de forma controlada.

- Mejora de la Experiencia del Cliente: Los clientes de bajo riesgo se benefician de tasas más bajas.

Anexos

Ejemplos de Riesgo Score y Tasa de Interés Dinámica por Préstamo (Árbol de Decisión)

	customer_id	loan_id	riesgo_score	tasa_interes_dinamica
1	1223	1	0.001	0.1
2	5190	2	0	0.1
3	5194	3	0	0.1
4	3978	4	0	0.1
5	4535	5	0	0.1
6	3604	6	0	0.1
7	271	7	0	0.1
8	5430	8	0	0.1
9	5128	9	0	0.1
10	4402	10	0	0.1

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 … 10 Next

Ejemplos de Riesgo Score y Tasa de Interés Dinámica por Préstamo (Árbol de Decisión)

	customer_id	loan_id	riesgo_score	tasa_interes_dinamica
11	1329	11	0	0.1
12	5323	12	0	0.1
13	3460	13	0	0.1
14	1881	14	1	0.5
15	3597	15	0	0.1
16	2888	16	1	0.5
17	5101	17	0	0.1
18	4798	18	0	0.1
19	1338	19	0	0.1
20	1148	20	1	0.5

Showing 11 to 20 of 100 entries

Previous 1 2 3 4 5 … 10 Next

Ejemplos de Riesgo Score y Tasa de Interés Dinámica por Préstamo (Árbol de Decisión)

	customer_id	loan_id	riesgo_score	tasa_interes_dinamica
21	2092	21	0.999	0.5
22	3149	22	0	0.1
23	4030	23	0	0.1
24	3662	24	0	0.1
25	3614	25	0.001	0.1
26	3750	26	0	0.1
27	4368	27	0.001	0.1
28	701	28	0	0.1
29	5204	29	0	0.1
30	4710	30	0.001	0.1

Showing 21 to 30 of 100 entries

Previous 1 2 3 4 5 … 10 Next

Autor

Carlos Walls Salcedo

Fecha

24 de junio de 2025