Prometeo

Reporte: Proyecto Prometeo Cross-Selling

1. Introducción

El proyecto Prometeo Cross-Selling surge como respuesta a la necesidad de las instituciones financieras de optimizar sus estrategias de venta cruzada en el contexto del Open Banking. Hemos desarrollado un prototipo (modelo predictivo) que identifica clientes con alta propensión a adquirir productos financieros adicionales, específicamente seguros, mediante un enfoque de I+D complementado con metodologías de design thinking y lean. Hasta el momento, hemos completado las fases de investigación, análisis de datos, desarrollo del modelo predictivo y evaluación de resultados, quedando pendiente la implementación del dashboard interactivo.

2. Metodología

Nuestra metodología se estructuró en cinco fases principales:

- Observación e Investigación: Análisis de la propuesta de valor de prometeo y su competencia. Análisis de tendencias en Open Finance y mejores prácticas de cross-selling en el sector financiero.
-Definición y Focalización: Establecimiento de KPIs, métricas y alcance del prototipo.
- Ideación y Diseño: Diseño del modelo desde el EDA, fetature engineering, entrenamiento del modelo, evaluación del modelo y optimización del modelo. Aplicación de mejores prácticas de ventas crosselling, así como tendencias tecnológicas al diseño y flujo de trabajo del dashboard de crosselling.
- Prototipado y Validación. Implementación del dashboard con datos dummy. Presentación al CEO de Prometeo.
- Presentación de Conclusiones. Análisis de hallazgos y recomendaciones de negocio.

El enfoque se centró en optimizar la preparación de datos para maximizar el rendimiento predictivo, priorizando características con mayor relevancia comercial.

3. Investigación

Tendencias tecnológicas en Fintech y Open Finance

Las principales tendencias tecnológicas y más específicamente en Fintech y Open finance de acuerdo a nuestro estudio son:

- Agentes de IA: Asistentes virtuales capaces de percibir el entorno, aprender y automatizar procesos complejos en marketing y ventas.
- Hiperpersonalización: Uso intensivo de IA y analítica de datos para ofrecer productos y servicios financieros adaptados a cada cliente de forma única.

- Interoperabilidad y automatización de procesos: Integración de sistemas mediante APIs y flujos de trabajo inteligentes que permiten compartir datos en tiempo real y

agilizar las operaciones.

4. Propuesta de valor en el Open Finance

La tendencia global de Open Finance avanza hacia la estandarización de tres pilares clave: pagos cuenta a cuenta (A2A), validación de cuentas y agregación bancaria. El objetivo es ofrecer estas capacidades en múltiples mercados a través de una sola infraestructura. Empresas como Plaid, Tink, TrueLayer y Prometeo lideran este movimiento gracias a su cobertura regional y a su habilidad para orquestar datos financieros y habilitar pagos y validaciones de forma integrada.

En el caso particular de Prometeo, su valor diferencial está en adaptar estas mismas capacidades al contexto fragmentado de América Latina, con un enfoque especial en pagos internacionales. Pero su propuesta va más allá de simplemente conectar APIs: Prometeo mejora la calidad de los datos obtenidos y ofrece funcionalidades adicionales que realmente generan valor para el cliente.

Este enfoque debe replicarse en nuestro reto. No basta con conectar al banco y acceder a datos de calidad para predecir compras de productos financieros. Lo realmente valioso está en cómo se combinan esos datos con funcionalidades útiles que impacten positivamente al cliente.

5. Cross-selling y Mejores Prácticas

El cross-selling en finanzas consiste en ofrecer productos complementarios basados en las necesidades reales del cliente, generando valor y aumentando la rentabilidad por usuario. Para hacerlo efectivo, se deben aplicar tres principios clave: (1) conocer profundamente al cliente mediante análisis de datos y segmentación inteligente, (2) ofrecer productos relevantes en el momento adecuado —como seguros, líneas de crédito o inversión que complementen lo ya contratado— y (3) automatizar el proceso con recomendaciones personalizadas en canales digitales y CRM. Lo esencial no es solo vender más, sino hacerlo de forma útil y oportuna para fortalecer la relación con el cliente.

7. KPIS del proyecto

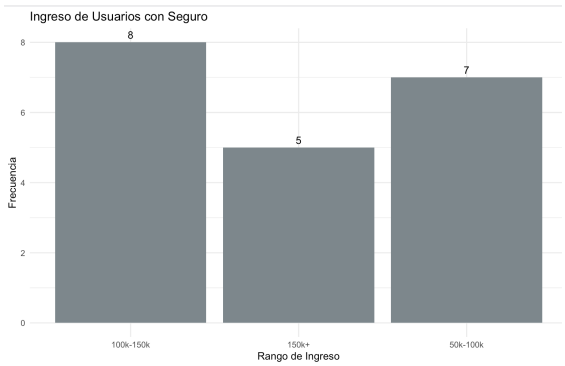
Este proyecto busca desarrollar un MVP funcional de venta cruzada en Open Finance, capaz de identificar clientes con alta propensión a adquirir productos financieros adicionales (como seguros), utilizando datos de APIs bancarias y modelos de machine learning. El éxito se medirá a través de KPIs técnicos como AUC-ROC, Accuracy, F1-Score, Recall y matriz de confusión; y de negocio como la tasa de conversión estimada, lista de clientes prioritarios, potencial de ingresos, eficiencia del modelo y segmentos con mayor valor.

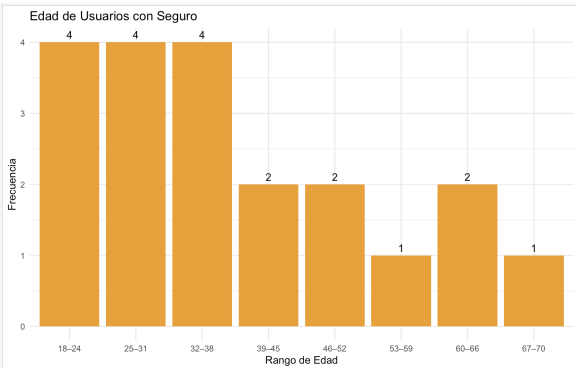
Prometeo

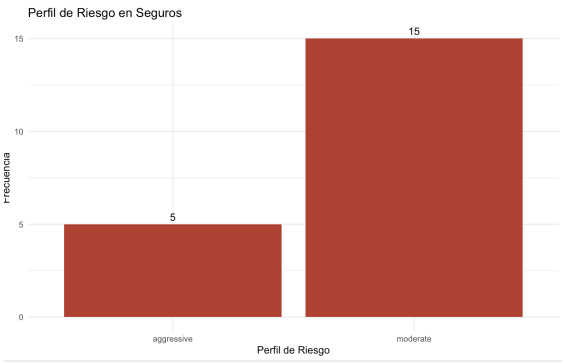
8.. Hallazgos del EDA

El análisis exploratorio de datos reveló patrones significativos respecto a la contratación de seguros:

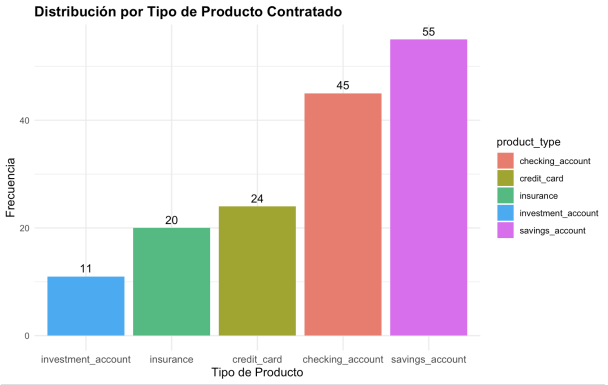
El análisis revela que los seguros son contratados principalmente por personas jóvenes entre 18 y 38 años, lo que marca una diferencia importante frente a otros productos financieros que se concentran en edades más avanzadas. Predomina el perfil de riesgo **moderado**, con presencia relevante de perfiles **agresivos**, mientras que los **conservadores** muestran una baja propensión a contratar seguros. En cuanto a ingresos, los contratantes se agrupan mayoritariamente en los rangos de **50k–100k** y **100k–150k**, con escasa participación en el segmento de **30k–50k**.

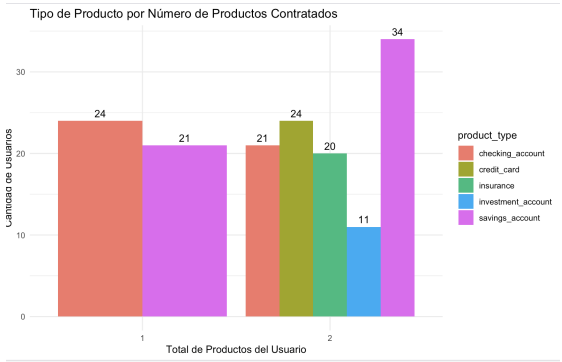


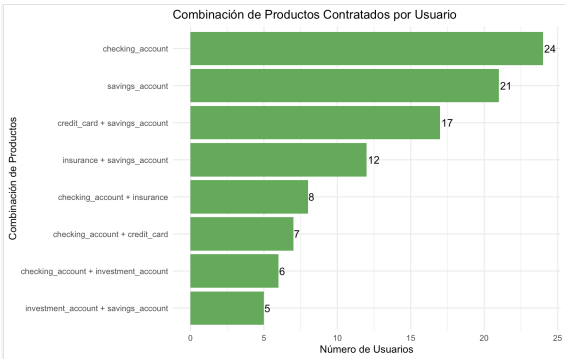




Las combinaciones más comunes incluyen “**cuenta de ahorro + seguro**” (12 usuarios) como el **segundo producto más contratado**, y “**cuenta corriente + seguro**” (8 usuarios) como el **tercero**, lo que confirma un patrón de **cross-selling** donde los seguros suelen adquirirse después de una cuenta bancaria básica.







Comparado con otros productos, como las **inversiones** (ligadas a mayores ingresos, edad y riesgo agresivo) o las **tarjetas de crédito** (predominantes entre 46 y 66 años), los seguros sobresalen por atraer a un segmento **más joven**, con ingresos **medios-altos** y perfil **moderado**.

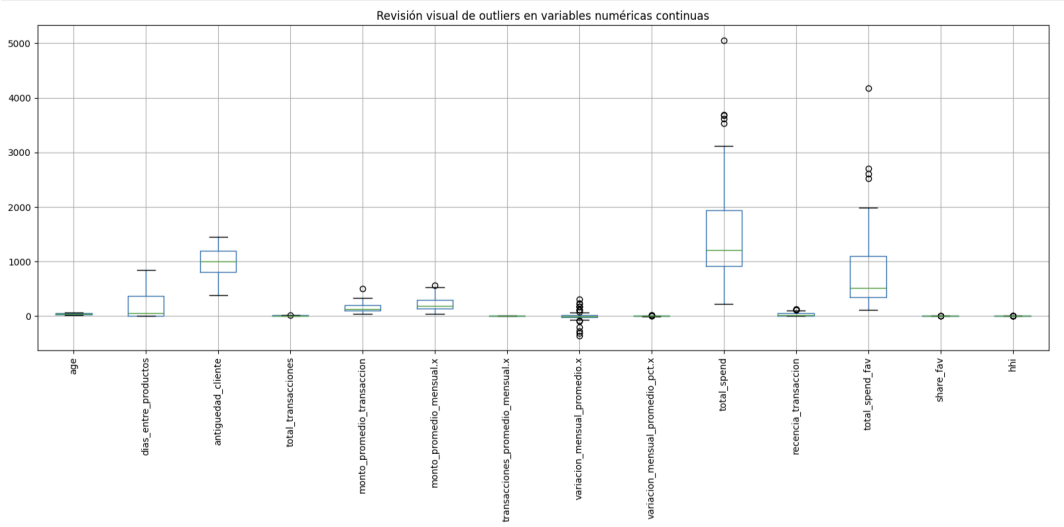
Estos hallazgos abren una oportunidad concreta para estrategias de **venta cruzada** dirigidas a **clientes jóvenes con cuentas activas**, en especial aquellos que se encuentran iniciando su vida profesional o en etapas tempranas de acumulación de patrimonio.

6. Limpieza

Los datos proporcionados estaban limpios en su totalidad. No obstante al crear nuevas variables en feature engineering se crearon variables con NA. La causa fue la creación de las variables para el cálculo de días de contratación entre un producto y otro (fecha_primer_producto - fecha_segundo_producto = dias_entre_productos). Como no todos tenian un segundo producto había NA u se imputaron como “none”, fecha dummy “01-01-1900” y 0.

Se revisaron los outliers, pero no hubo casos con outliers de gravedad. Es por ello que se optó por hacer logarítmicas algunas variables como total_spent. Mientras que las demás se normalizarán dependiendo el tipo de modelo.

Prometeo



6. Feature Engineering

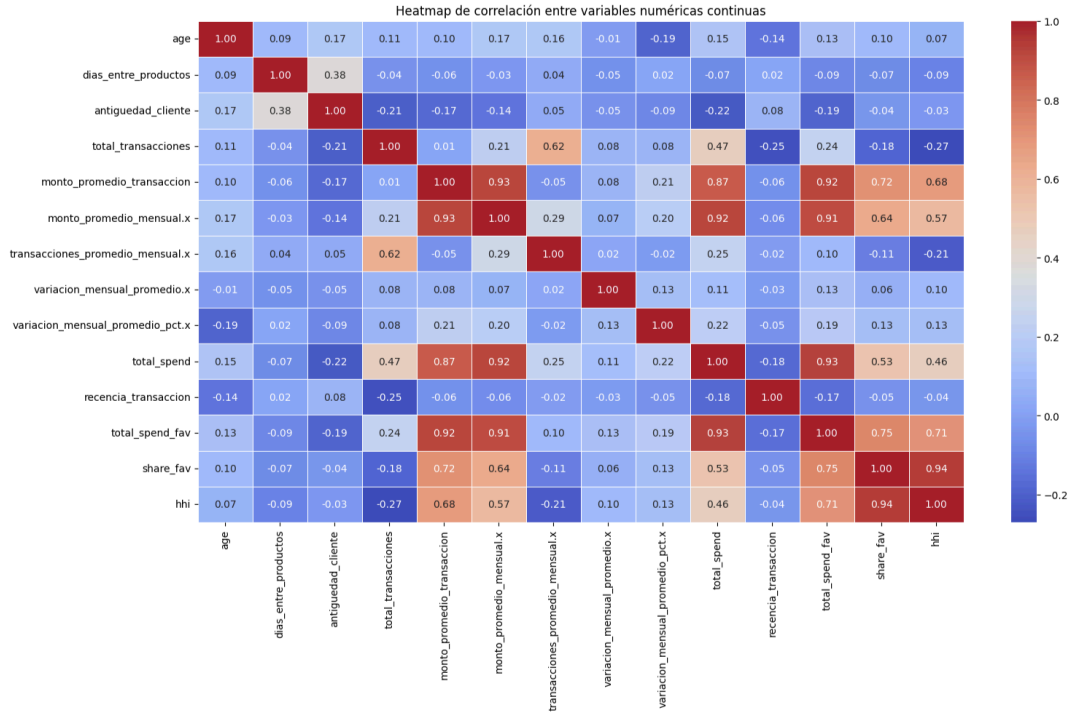
La ingeniería de características se centró en enriquecer y transformar los datos para maximizar el poder predictivo. Estas son las variables totales en las cuales hay variables que se crearon a partir de los datos existentes.

Variables de identificación: user_id
Variables demográficas: age, income_range, risk_profile, occupation, age_range_sturges
Variables de productos: primer_producto, fecha_primer_producto, segundo_producto, fecha_segundo_producto, checking_account, savings_account, credit_card, insurance, investment, numero_productos, combinacion_productos
Variables transaccionales: entertainment_count, food_count, health_count, shopping_count, supermarket_count, transport_count, travel_count, total_transacciones, monto_promedio_transaccion, mes_mas_compras, mes_mayor_monto, monto_promedio_mensual, transacciones_promedio_mensual, variacion_mensual_promedio, variacion_mensual_promedio_pct, total_spend, categoria_favorita_monto, share_fav, hhi
Variables temporales: n_meses_activos, recencia_transaccion, antiguedad_cliente, dias_entre_productos

Las que destacan es la antigüedad del cliente, la combinación de los productos, **categoría favorita**, **promedio mensual del monto de las transacciones**, variación en el monto de las trsansacciones mensuales, recencia y concentración del gasto.

Posteriormente se buscó quitar las variables que están correlacionadas y tienen colinealidad.

Prometeo



Resultados del heatmap:

Correlación alta: monto_promedio_transaccion vs monto_promedio_mensual.x -monto_promedio_mensual.x vs total_spend -total_spend_fav vs monto_promedio_mensual.x -share_fav vs total_spend_fav -share_fav vs hhi

Las menos correlacionadas son edad, recencia, dias_entre_productos y variacion_mensual_promedio

	Variable	VIF
0	const	289.308340
10	total_spend	67.991098
6	monto_promedio_mensual.x	59.743131
5	monto_promedio_transaccion	39.436096
12	total_spend_fav	39.065034
14	hhi	12.948370
13	share_fav	12.133946
4	total_transacciones	10.097715
7	transacciones_promedio_mensual.x	8.122448
3	antiguedad_cliente	1.378885
2	dias_entre_productos	1.215362
1	age	1.186664
11	recencia_transaccion	1.166270
9	variacion_mensual_promedio_pct.x	1.159112
8	variacion_mensual_promedio.x	1.073107

.

Prometeo

Se analizó la colinealidad y se eliminaron: total_spen", monto_promedio_mensual.x, monto_promedio_transaccion, hhi, share_fav y total_transacciones. Se verificó la correlación y colinealidad con resultados positivos.

Posteriormente se hicieron las transformaciones dependiendo el modelo

-Fecha: En todos los modelos las variables de fecha de hicieron TIMESTAMP.
-Cuantitativas Las variable total_spend_fav la convertimos en Log y las demás se normalizaron.
-Binarias: Se transformaton en Booleanl.
-Categoricas: Encoding one-hot para la regresión logística y label encoder en los otros modelos.

Finalmente en cada modelo se separó la variable target del resto del dataset para el entrenamiento del modelo. El target es la variable de insurance que son los usuarios que contrataron un seguro.

8. Resultados del Modelo

-Planteamiento:
Desarrollamos tres modelos base (Regresión Logística, Random Forest y XGBoost) para predecir la propensión a contratar seguros, evaluados mediante validación cruzada 5-fold.

-Resultados de modelos base:
Los mejores resultados los obtuvol XGBoost. Un ROC AUC de 0.95, una exactitud del 95%, precisión del 96%, recall del 80% y un F1-score de 0.84. La matriz de confusión (79 verdaderos negativos, 1 falso positivo, 4 falsos negativos y 16 verdaderos positivos) demuestra que este modelo consigue el mejor equilibrio entre detectar correctamente a los clientes positivos y minimizar los errores.

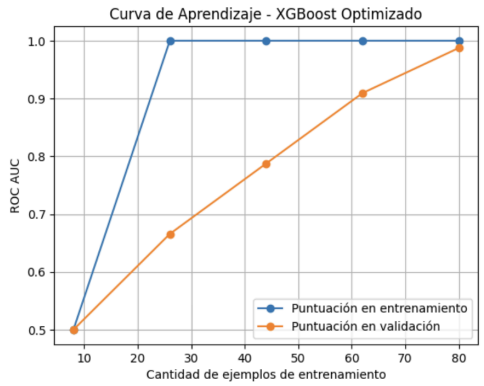
```
Evaluando modelo base: LogisticRegression
Métricas: {'roc_auc': np.float64(0.740625), 'accuracy': np.float64(0.65), 'precision': np.float64(0.3577777777777778), 'recall': np.float64(0.9), 'f1': np.float64(0.5091575091575091)}
Matriz de Confusión:
 [[47 33]
 [ 2 18]]

Evaluando modelo base: RandomForest
Métricas: {'roc_auc': np.float64(0.9109375), 'accuracy': np.float64(0.8299999999999998), 'precision': np.float64(0.7), 'recall': np.float64(0.25), 'f1': np.float64(0.36)}
Matriz de Confusión:
 [[79  1]
 [16  4]]

Evaluando modelo base: XGBoost
Métricas: {'roc_auc': np.float64(0.953125), 'accuracy': np.float64(0.95), 'precision': np.float64(0.96), 'recall': np.float64(0.8), 'f1': np.float64(0.8444444444444443)}
Matriz de Confusión:
 [[79  1]
 [ 4 16]]
```

-Elección y optimización del modelo:
Seleccionamos XGBoost por su superior desempeño y robustez ante datos heterogéneos. La optimización de hiperparámetros mediante grid search (learning_rate=0.05, max_depth=6, subsample=0.8) mejoró el rendimiento.

Prometeo

```
Métricas de XGBoost optimizado:
{'roc_auc': np.float64(0.99375), 'accuracy': np.float64(0.93), 'precision': np.float64(0.95), 'recall': np.float64(0.7), 'f1': np.float64(0.7880952380952381)}
Matriz de Confusión de XGBoost optimizado:
[[79  1]
 [ 6 14]]
```

Curva de aprendizaje para XGBoost optimizado:

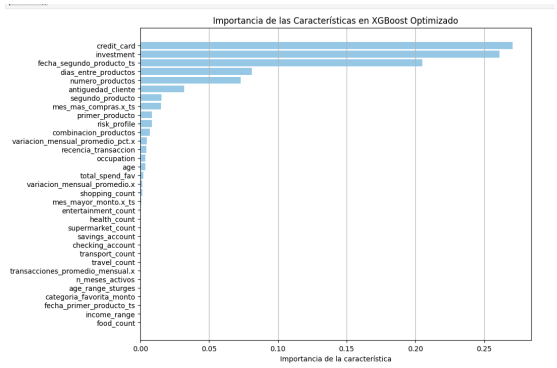


Con esta configuración, el modelo alcanzó un ROC AUC de aproximadamente 0.9935 en validación, lo que indica una capacidad discriminativa casi perfecta entre los clientes que contratan el seguro y los que no. La exactitud fue del 94%, con una precisión del 95%, lo que significa que la gran mayoría de las predicciones positivas son realmente correctas. Aunque el recall (70%) muestra que aún se pierden algunos clientes potenciales, el balance global (F1-score de 0.80) es muy sólido.

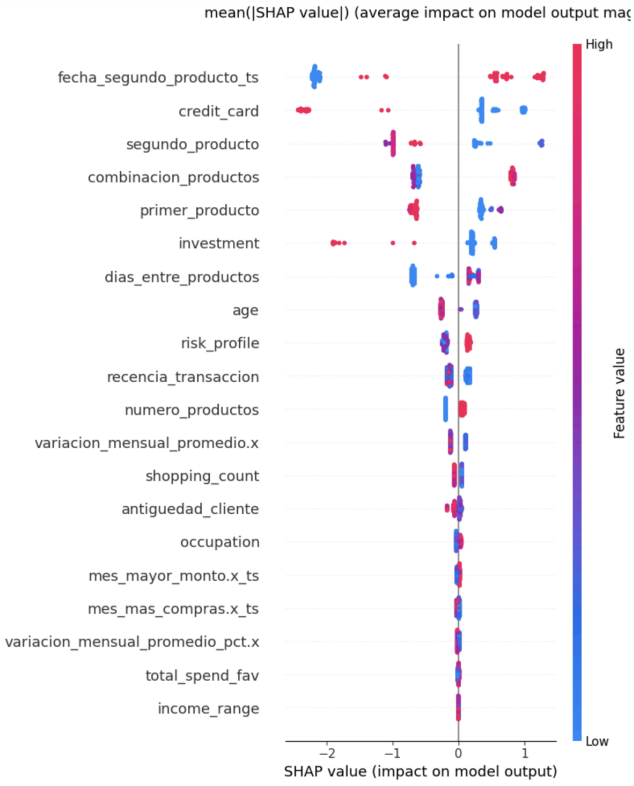
Además, al graficar la curva de aprendizaje pude notar que tanto el desempeño en el conjunto de entrenamiento como en el de validación se encuentran muy cerca y en valores elevados, lo cual sugiere que el modelo no presenta un sobreajuste significativo y está aprendiendo de manera estable.

-Importancia de las características:
Se realizó el análisis para saber qué variables predicen el comportamiento del target. En otras palabras que variables son importantes para predecir que un cliente va a contratar un seguro.

Prometeo



1. credit_card y investment - determinan altamente el comportamiento
2. fecha_segundo_producto_ts, dias_entre_productos, numero_productos - determina el comportamiento los dias que pasan para adquirir un nuevo producto. Se descarta fecha_seguno_producto.ts debido a la imputación.
3. antiguedad_cliente y mes_mas_compras_ts
4. Variables como primer_producto, risk_profile, combinacion_productos
5. Otras variables (gasto, recencia, ocupación, etc.)



El análisis SHAP reveló que tanto clientes con productos financieros consolidados (como tarjetas o inversiones) como aquellos sin ellos pueden tener alta probabilidad de contratar un seguro, sugiriendo la existencia de dos segmentos distintos a considerar. También mostró que

variables como la rapidez en adquirir productos o la antigüedad del cliente deben interpretarse con cuidado, ya que en algunos casos pueden indicar saturación o comportamientos atípicos por imputaciones. Por otro lado, se confirma que perfiles con cierto historial de productos, perfil de riesgo adecuado y comportamientos de gasto recientes siguen siendo los más propensos. Variables secundarias como ocupación, recencia o gasto mensual ayudan a afinar la segmentación, aunque con menor impacto.

9. Recomendaciones (menos de 3 páginas)

9.1 Características más importantes para predecir la contratación de seguros

A partir del análisis del modelo (XGBoost), el EDA y los valores SHAP, se destacan las siguientes variables como las más influyentes, con matices relevantes en la interpretación:

1. **Crédito e inversión (credit_card, investment)**
   Son indicadores de mayor bancarización y apertura a nuevos productos. Al mismo tiempo, SHAP revela que la ausencia de estos productos en determinados segmentos no implica desinterés en seguros; puede sugerir que el seguro sería el “producto de entrada” para clientes menos experimentados.

2. **Días transcurridos entre productos (dias_entre_productos)**
   El intervalo de tiempo para adquirir un segundo producto se ha relacionado con mayor propensión, pero existe la hipótesis de que un intervalo muy corto también puede reflejar “saturación” o necesidad de cautela al interpretar los datos. Clientes de perfil moderado podrían tardar más en convencerse de contratar un seguro, por lo que la velocidad de contratación no siempre significa mayor disposición.

3. **Número de productos (numero_productos)**
   Quienes pasan de 1 a 2 productos muestran una mayor probabilidad de adquirir seguro, posiblemente porque están empezando a diversificar su portafolio. Sin embargo, el dataset carece de casos con 3 o más productos, lo que limita la conclusión y sugiere que el efecto SHAP podría estar sobre-representando la distinción entre “uno vs. dos” productos. Este hallazgo conviene validarlo con datos futuros más completos.

4. **Antigüedad del cliente (antiguedad_cliente)**
   Indica un doble fenómeno:
   -Clientes jóvenes, que tienden a tener conciencia temprana sobre la importancia de un seguro.
   -Clientes con mayor tiempo en la institución y varios productos contratados, que también suelen ser receptivos a una oferta de cobertura adicional.

En ambos escenarios, la fidelidad y el conocimiento de la marca facilitan la venta cruzada.

5. **Edad (age)**
   El EDA mostró que las edades más bajas (aprox. 18 a 38 años) son las que más contratan seguros, posiblemente por cultura financiera emergente o etapas de vida en las que desean protegerse.

6. **Perfil de riesgo y combinación de productos (risk_profile, combinacion_productos)**
   Un perfil de riesgo moderado/alto puede ver al seguro como un mecanismo complementario de protección o un refuerzo a su portafolio. Las combinaciones que incluyen cuentas corrientes y/o tarjetas de crédito suelen asociarse a una actitud más abierta a nuevos productos.

7. **Primer producto (primer_producto)**
   Los clientes cuyo primer producto fue una cuenta corriente tienden a mostrar una mayor tasa de conversión en la contratación de seguros, posiblemente porque ya manejan flujos de efectivo más elevados o tienen más contacto con servicios bancarios.

9.2. Segmentos y estrategias de campaña

Basados en las características anteriores, se sugieren cuatro segmentos principales y líneas de acción:

1. **Jóvenes con perfil de riesgo moderado**

   Hallazgo clave: Edades entre 18 y 38 años, con un riesgo moderado y, en algunos casos, sin productos de crédito o inversión.

   Estrategia de Awareness:
   -Comunicaciones de concientización sobre la importancia de un seguro para proteger su inicio de vida financiera.
   -Mensajes o llamadas periódicas con frecuencia moderada, sin ser invasivas; reforzar la idea de seguridad y accesibilidad.
   -Ofrecer planes flexibles, con coberturas básicas que faciliten la adopción del seguro.

2. **Clientes sin crédito ni inversión (venta directa)**

   Hallazgo clave: SHAP indica que incluso si no poseen tarjeta de crédito o inversión, ciertos clientes pueden interesarse en un seguro como producto inicial o alternativo.

   Estrategia de Venta Directa:
   -Campañas que destaquen la facilidad de contratación y la tranquilidad que brinda un

seguro, especialmente en canales digitales y móviles.
-Venta directa y agresiva
-Ofrecer asesoría simplificada para desmitificar la contratación y posicionar el seguro como un paso lógico antes o en paralelo a otros productos.

3. **Clientes con ingreso >100k, edad media y perfil de riesgo alto**

   Hallazgo clave: Ingresos por encima de 100k y un perfil de riesgo mayor sugieren interés en proteger patrimonio o salud, además de estar dispuestos a gastos de primas más elevadas.

   Estrategia Premium:
   -Ofrecer productos de seguro con coberturas más amplias, alineadas a la protección de patrimonio o salud integral.
   -Destacar beneficios adicionales (por ejemplo, asistencia VIP o coberturas internacionales) que apunten a su mayor capacidad económica y perfil de riesgo agresivo.

4. **Clientes recién incorporados con tarjeta (nuevos en la institución)**

   Hallazgo clave: Si el primer producto del cliente es una cuenta corriente o tarjeta, y no tienen aún otra cobertura, la probabilidad de contratar un seguro crece con abordajes oportunos.

   Estrategia de "Onboarding Plus":
   -Incluir la oferta de seguro durante las primeras interacciones poscontratación (primeras semanas)
   -Aprovechar su disposición inicial para simplificar la venta cruzada: "Ahora que ya tienes tu tarjeta, obtén un seguro con beneficios inmediatos y evita procesos adicionales".

10. Siguientes Pasos y Tareas Pendientes

- Enriquecimiento con datos adicionales de APIs de Open Finance
- Implementación del dashboard interactivo con ranking de clientes por probabilidad de contratación
- Desarrollo de agente de IA para recomendaciones personalizadas
- Mejoramiento del modelo y hacer pruebas continuamente.

El avance actual demuestra la viabilidad técnica y potencial valor comercial del proyecto. La implementación del dashboard consolidará estos resultados, permitiendo su aplicación práctica inmediata en estrategias de cross-selling.