
Desarrollo Modelo de Clasificación para la Identificación de Clientes que Realizan Grandes Retiros en el Fondo de Pensiones Voluntarias

Prueba Técnica – Data Scientist –

Nombre del Aspirante: Cindy Zulima Alzate Román

Presentado a: Skandia

Enero 2023



Objetivo → Mostrar la probabilidad de que un cliente del fondo voluntario de pensiones retire el 70% o más de su saldo en los siguientes 3 meses.

El modelo de clasificación será implementado en búsqueda de lograr el objetivo planteado, mediante la definición de la siguiente variable objetivo:

$$var_{obj} = \begin{cases} 1 & \text{si } \frac{Saldo_{oct2022}}{Saldo_{jul2022}} \leq 0.3 \\ 0 & \text{en otro caso} \end{cases}$$

Para el desarrollo del modelo se estructuró una base de datos consolidada a nivel CLIENTE con la siguiente información:

- Identificación del cliente
- Saldo del cliente para 19 periodos
- Transacciones del cliente (valores netos de los aportes o retiros a las distintas cuentas) para 19 periodos
- Variable creada "act_positiva" → Cuenta aquellos meses donde los APORTES fueron superiores a los RETIROS
- Variable creada "act_negativa" → Cuenta aquellos meses donde los RETIROS fueron superiores a los APORTES
- Variable objetivo (binaria 1 o 0)

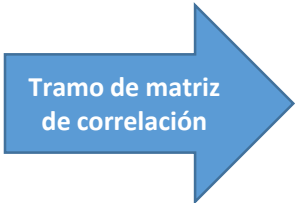
NroDocum	SALDO_202101	SALDO_202102	...	SALDO_202207	TRANS_202101	TRANS_202102	...	TRANS_202207	act_positiva	act_negativa
1000001634	47,092,300	46,315,660	...	49,877,450	2,467,200	-600,000	...	-600,000	5	14
1000014510	20,260,190	20,195,410	...	5,979,061	0	6,000,000	...	3,000,000	14	1
1000031391	86,798,040	86,355,390	...	85,632,600	10,000,000	-2,000,000	...	0	10	3
1000053307	105,372,000	106,214,600	...	115,459,400	1,150,000	1,300,000	...	0	16	0

Ventana de tiempo utilizada:
Enero2021 a Julio2022

Métricas Estadísticas

Correlación

Medida de dependencia lineal entre dos variables aleatorias cuantitativas. Para la base del ejercicio, los saldos muestran una alta correlación entre los distintos periodos.



	SALDO_2021O1	SALDO_2021O2	SALDO_2021O3
SALDO_2021O1	1.000000	0.982750	0.968064
SALDO_2021O2	0.982750	1.000000	0.981018
SALDO_2021O3	0.968064	0.981018	1.000000
SALDO_2021O4	0.959862	0.972635	0.986598
SALDO_2021O5	0.930472	0.942929	0.954288

Multicolinealidad

Dada la alta correlación de los saldos y la evaluación de colinealidad con el cálculo del *Variance Inflation Factor*¹, se determina la eliminación de algunas de las variables que inicialmente fueron contempladas.

VIF > 10 indica un serio problema de colinealidad

	VIF
SALDO_2021O1	638.089810
SALDO_2021O2	692.310062
SALDO_2021O3	425.831903
SALDO_2021O4	356.203713
SALDO_2021O5	1666.495036
SALDO_2021O6	1693.293583

Métricas del Resultado del Modelo

Accuracy

Este indicador de desempeño muestra el cociente entre los individuos que el modelo clasificó de manera correcta y el total de individuos evaluados.

AUC ROC

Habilidad del modelo de clasificar y distinguir entre las 2 clases de clientes que se generaron con la variable objetivo.

Ambos indicadores exhiben un mejor desempeño del modelo cuando sus valores son cercanos a 1.

1 VIF. Cuando su valor es alto, indica una fuerte asociación lineal con otra variable, es decir, que su aporte al modelo ya se está teniendo en cuenta con la inclusión de otra variable, en este caso lo más prudente, es excluir aquella variable con VIF alto.




Las opciones de modelos que fueron desplegadas como posible solución al planteamiento del objetivo, son:
Regresión Logística, Árbol de Decisión, Random Forest y XGBOOST.

Descripción del por qué de su selección del modelo

El modelo de clasificación se adaptaba muy bien para lograr el objetivo planteado, adicionalmente, la implementación del XGBOOST obtuvo los resultados más altos en los indicadores de desempeño como se muestra en el comparativo de los indicadores para los diferentes modelos ejecutados que se exhibe a continuación:

	Accuracy	AUC ROC
Logistic Regression	0.73	0.54
Decision Tree	0.85	0.57
Random Forest	0.92	0.68
XGBOOST	0.93	0.71



Hallazgos relevantes y recomendaciones

- Hallazgos → Identificación de las variables que más aportan en la predicción de la variable objetivo, a saber, SALDO_202207 (correspondiente al último periodo evaluado –julio de 2022), act_positiva y act_negativa.
- Recomendaciones →
 - * Incluir una nueva variable “Sin_Actividad” que realice el conteo de los periodos donde no se presentaron transacciones (ni aportes, ni retiros).
 - * Identificar la combinación óptima de hiperparámetros del XGBOOST que maximice el desempeño del modelo.
 - * Implementar modelos adicionales y comparar su desempeño.