Axel Andrei Zetina Cuevas

Evaluación Científico de Datos Jr

SECCION 1

Caso de Uso: El fondo voluntario de pensiones es un producto que permite a sus clientes ahorro y optimización de su capital mediante inversión diversificada. Dado un caso hipotético de crisis en los mercados y aumento de la inflación, la compañía observa un aumento en los retiros de capital en dicho producto. En busca del bienestar sus clientes, quienes estarían descapitalizando al realizar retiros, la compañía requiere un modelo predictivo que muestre la probabilidad de que un CLIENTE retire el 70% o más de su saldo en los siguientes 3 meses.
Para esto, usted tendrá acceso a tres bases de datos en formato parquet:

1. Transacciones/Transferencias: Aportaciones y retiros DIARIOS por CONTRATO y por PRODUCTO en 22 meses.
2. Saldos: Saldo mensual por CLIENTE, CONTRATO y PRODUCTO en 22 meses.
3. Clientes: Información demográfica del CLIENTE. Podrá acceder a las bases en los siguientes vínculos:

https://datamxdevsa.blob.core.windows.net/crskmex000/DSTest/0saldos.parquet
https://datamxdevsa.blob.core.windows.net/crskmex000/DSTest/0transferencias.parquet
https://datamxdevsa.blob.core.windows.net/crskmex000/DSTest/0clientes.parquet

Recuerde que un cliente puede tener más de un contrato activo y que a su vez, un contrato puede tener más de un producto. Adicionalmente, el modelo debe diseñarse a nivel de CLIENTE. Tenga en cuenta los retiros netos, es decir, si un cliente realiza aportaciones y retiros en el mismo periodo de tiempo, usted SÓLO deberá tener en cuenta el total neto de la suma de estos. Desarrolle el modelo requerido en GitHub donde se cuente con lo siguiente:

1. Código comentado desde la lectura de los parquets orígenes hasta la construcción del modelo y su validación (entregar url al código en GitHub).
   https://github.com/AxelZetina/skandia-predictive-model.git
2. Presentación de no más de 5 diapositivas que contenga:
   a. Nombre y apellido del aspirante
   b. Tipo del modelo
   c. Métricas estadísticas utilizadas y métricas del resultado del modelo
   d. Breve descripción del proceso de modelado
   e. Breve descripción del porqué de su selección del modelo
   f. Hallazgos relevantes y recomendaciones para ejecutar a partir del modelo.

El modelo se puede realizar en el lenguaje de su preferencia (Python, R, M, SPSS)

SECCION 2

Preguntas abiertas

1. ¿Qué diferencia hay entre datos estructurados y no estructurados?

   Los datos estructurados están altamente organizados y formateados de tal manera que se pueden buscar fácilmente en bases de datos relacionales. Los datos no estructurados no tienen un formato u organización predefinidos, lo que hace que sea mucho más difícil de recopilar, procesar y analizar. Básicamente los datos estructurados están organizados en tablas o estructuras definidas, mientras que los datos no estructurados no siguen un formato predefinido y pueden ser más variados en su naturaleza y contenido.

2. Se tiene un grupo de clientes a los cuáles se quiere dar aviso sobre las aportaciones pendientes en su producto de ahorro. Se desea probar tres encabezados distintos que podrían tener los correos electrónicos en los que se les notifique y evaluar si existe impacto positivo y cómo difiere el impacto entre cada uno de ellos. La base de clientes actuales es de 45,000, de los cuales el 10% deja de aportar mensualmente. Se consideran eficaces las campañas que generan una diferencia de al menos el 5 % y se quiere un nivel de confianza del 95 %.

   a. ¿Con que tipo de población se está trabajando?

      Con el tipo de población de clientes que actualmente están registrados en el producto de ahorro y que reciben correos electrónicos de notificación sobre aportaciones pendientes. es decir, la población se compone de los 45,000 clientes de la base de datos.

   b. ¿De qué tamaño debe ser la muestra de cada grupo para obtener resultados útiles para el experimento para los parámetros indicados?

      Para determinar el tamaño de la muestra necesaria para cada grupo en el experimento, podemos utilizar la fórmula para el cálculo del tamaño de la muestra en pruebas de proporciones y calcular el tamaño de muestra necesario para cada grupo, necesitamos conocer la proporción de éxito esperada en cada grupo (p), el nivel de confianza deseado (Z), el tamaño del efecto mínimo significativo (d), y la proporción de la población que deja de aportar mensualmente, entonces la formula es:

$$n = \frac{N \cdot p \cdot (1-p)}{(N-1) \cdot d^2 + p \cdot (1-p)}$$

      Por lo tanto, el nivel de confianza (Z) nos proporciona los valores críticos de la distribución normal estándar. Entonces para un nivel de confianza del 95%, Z es aproximadamente 1.96.

c. ¿Qué métrica medirías?
Empezaría por medir la tasa de apertura y tasa de cancelación por su producto de ahorro

d. ¿Qué prueba estadística aplicarías y por qué?
Utilizaría prueba de análisis de varianza porque estamos comparando más de dos grupos y si la prueba indica que hay diferencias significativas entre los grupos, podríamos seguir con pruebas de comparaciones múltiples e identificar cuales grupos difieren entre ellos.

3. Dentro de la compañía tengo una base de clientes de 50,000. Mensualmente se cancelan en promedio 250 contratos con una desviación estándar de 50. Quiero crear un modelo que me prediga los casos en que el cliente me cancelará ¿Qué preprocesamiento básico requiere la base y que consideraciones se deberían tener?

   En primera instancia, la base de datos debe estar limpia, con formato adecuada y ordenada, una vez que se puede manejar la base de datos es de suma importancia identificar las características relevantes que podrían influir en la cancelación de un contrato y eliminar esas características irrelevantes o redundantes que no aporten información útil para el modelo. Una consideración a realizar, es el análisis exploratorio de datos para comprender mejor las características y su relación con las cancelaciones.

4. Explique que es la ley de los grandes números.

   La ley de los grandes números es una inferencia estadística y menciona que conforme tomemos muestras muy grandes de una población, podemos esperar que las estimaciones basadas en esas muestras se acerquen más y más al verdadero valor de la población. Por lo tanto, es fundamental para la toma de decisiones basada en datos.

5. ¿Cómo implementarías un modelo de machine learning en producción? ¿Cuáles son los puntos fundamentales a analizar y asegurar?

   Procesamiento de datos escalables, seguridad y privacidad, mantenimiento, pruebas y la documentación.

6. ¿Qué consideraciones debes tener en cuenta al trabajar con datos personales?

   Las consideraciones a tener en cuenta son éticas, legales y de seguridad para garantizar el manejo adecuado y responsable de la información, por lo que no habrá fuga y perdida de datos, estos mismos serán seguros, anónimos y transparentes.

7. A partir de los siguientes datos: transacciones, calificación de satisfacción, interacciones con el portal, quejas y llamadas de servicio; se solicita realizar un modelo que prediga si el cliente va a realizar una aportación extra, representado por el valor 1; o no, representado por el valor 0. Mensualmente un 5% de los clientes hace una aportación extra, se entrenaron 3 modelos distintos con las siguientes métricas de rendimiento. ¿Cuál elegirías y por qué?
¿Qué recomendarías? El entrenamiento se realizó con 70,000 registro y el testeo con 30,000

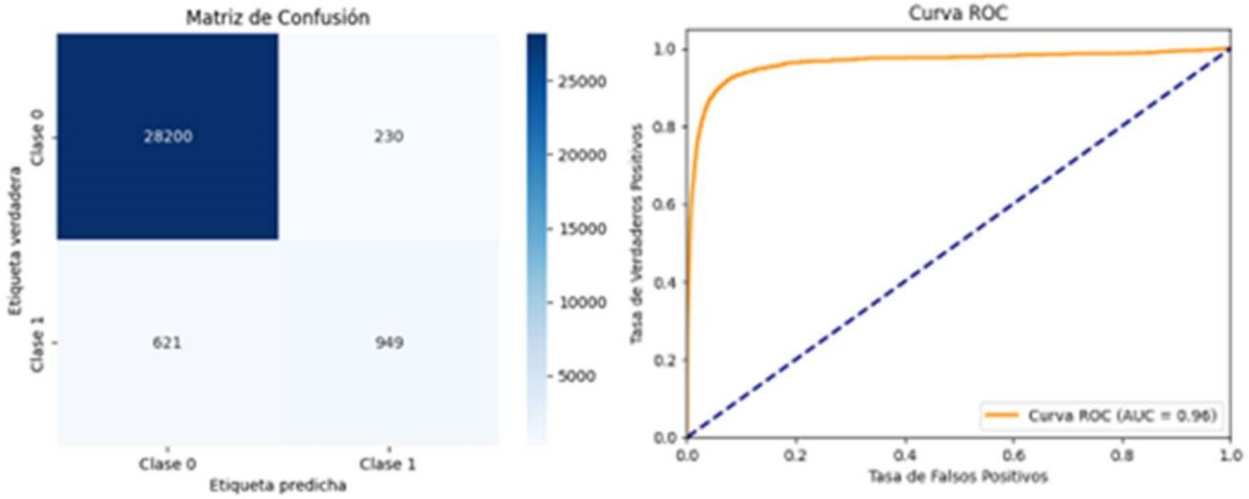
Observando las precisiones, matrices de confusión y las métricas, el segundo modelo parece tener un mejor rendimiento en general. Tiene una precisión global más alta y una precisión de la clase positiva (Recall) más alta en comparación con los otros dos modelos. Además, tiene menos falsos positivos (FP) y falsos negativos (FN) en comparación con los otros modelos.

Por lo tanto, el segundo modelo es el más óptimo para la predicción de aportaciones extra de clientes.

Recomendaría hacer una ajuste a los parámetros y considerar técnicas de balanceo de clases para obtener un modelo robusto y generalizable que pueda predecir con precisión las aportaciones extra de los clientes.

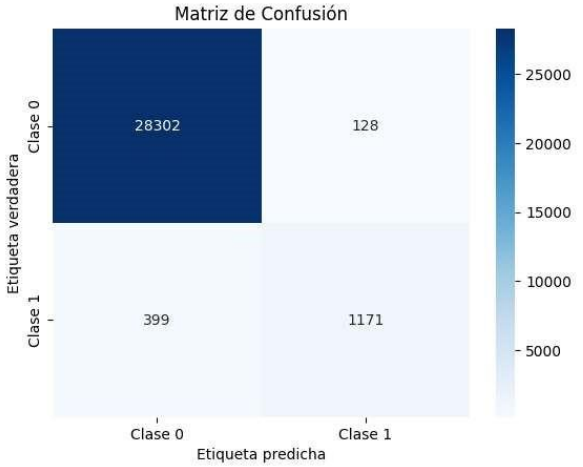
Modelo A:

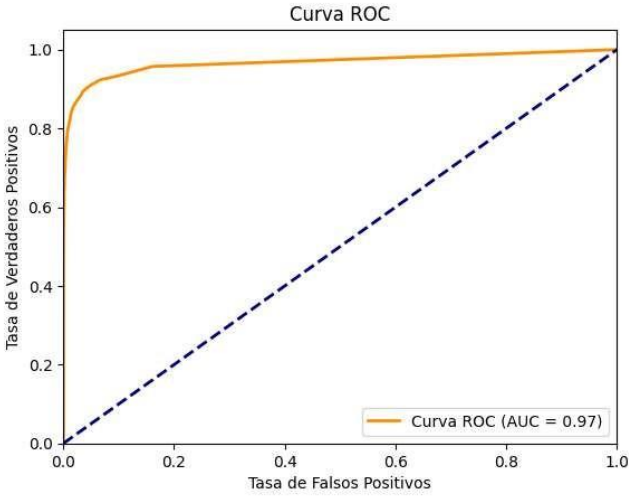
	precision	recall	f1-score	support
0	0.98	0.99	0.99	28430
1	0.80	0.60	0.69	1570
accuracy			0.97	30000
macro avg	0.89	0.80	0.84	30000
weighted avg	0.97	0.97	0.97	30000



Modelo B:

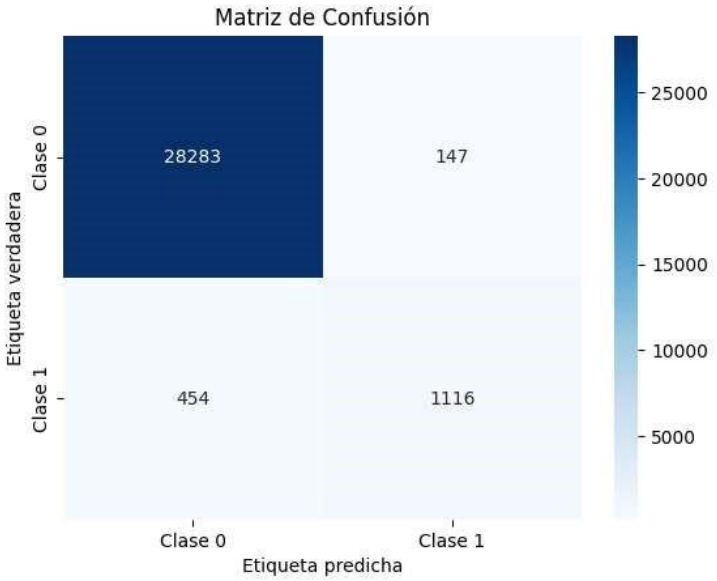
	precision	recall	f1-score	support
0	0.99	1.00	0.99	28430
1	0.90	0.75	0.82	1570
accuracy			0.98	30000
macro avg	0.94	0.87	0.90	30000
weighted avg	0.98	0.98	0.98	30000

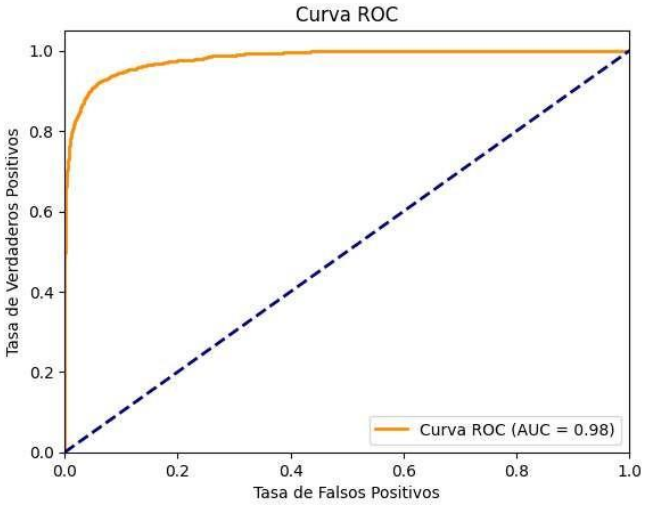




Modelo C:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	28430
1	0.88	0.71	0.79	1570
accuracy			0.98	30000
macro avg	0.93	0.85	0.89	30000
weighted avg	0.98	0.98	0.98	30000





8. La empresa ha tenido un aumento significativo en los retiros, te piden que identifiques la razón y hagas la predicción de los próximos 3 meses para poder hacer estrategias específicas con los puntos de dolor clave ¿Qué es mejor interpretabilidad o precisión del modelo a desarrollar?

Ambos factores son importantes y tienen pros y contras, no obstante, como es un problema donde se requiere identificar la razón detrás del aumento significativo en los retiros y hacer predicciones para futuros retiros, podría ser beneficioso priorizar la interpretabilidad del modelo. Esto permitirá comprender mejor los factores que contribuyen a los retiros y explicar los resultados de manera clara a las partes interesadas, sin hacer un lado la precisión del modelo.