CLUSTER AI – Informe Horowitz - Amaya

PREDICCIÓN DE SUSCRIPCIÓN A DEPÓSITOS A PLAZO EN CAMPAÑAS DE MARKETING BANCARIO

INTRODUCCIÓN Y OBJETIVOS

Las entidades financieras utilizan campañas de marketing directo para ofrecer productos como depósitos a plazo fijo. Sin embargo, la mayoría de los contactos no se convierten en suscripciones efectivas, lo que implica costos elevados y una utilización ineficiente de los recursos comerciales.

El objetivo de este trabajo es construir modelos de clasificación que permitan **predecir la probabilidad de suscripción** a un depósito a plazo por parte de un cliente contactado en una campaña de telemarketing. En particular, se busca:

- Analizar en profundidad el dataset disponible mediante un **análisis exploratorio de datos (EDA)**.
- Evaluar la calidad de los datos (faltantes, outliers, desbalanceo de clases).
- Definir una estrategia de **preprocesamiento** y **codificación de variables categóricas**.
- Comparar el desempeño de modelos obtenidos con **AutoML (mljar-supervised)** sobre:
    - El **dataset completo**.
    - **Tres segmentos de clientes** definidos según su historial de contacto previo.
    - Versiones del dataset **con y sin reducción de dimensionalidad mediante PCA**.
- Seleccionar el enfoque que logre mejor compromiso entre discriminación (AUC), precisión y recall para la clase positiva (suscripción).

DESCRIPCIÓN DEL DATASET

El dataset reúne **45.211 contactos** de una campaña de marketing directo de un banco europeo, en la que se ofrecía un depósito a plazo por teléfono y celular. Incluye tanto variables **sociodemográficas** (edad, ocupación, estado civil, nivel educativo), como **financieras** (saldo promedio en cuenta, presencia de préstamos hipotecarios o personales, existencia de crédito en mora) y de **interacción comercial** (día, mes y duración del último contacto, cantidad de llamadas en la campaña actual, historial de contactos previos y resultado de campañas anteriores). En general, se observa una población adulta de alrededor de **41 años** en promedio, con saldos bancarios muy heterogéneos y fuertemente sesgados a la derecha (pocos clientes con saldos muy altos). La mayoría de los contactos se realizan por **teléfono celular**, concentrados en algunos meses específicos, y sólo una fracción reducida de clientes presenta historial de campañas previas exitosas. La variable objetivo, **Subscription**, indica si el cliente

terminó o no suscribiendo el producto, con una clara desproporción a favor de la clase negativa (alrededor del **12 %** de los casos corresponde a suscripciones efectivas).

ANÁLISIS EXPLORATORIO DE DATOS

En primer lugar, el análisis de calidad de datos mostró un conjunto relativamente limpio en cuanto a duplicados (no se detectaron filas repetidas), pero con **faltantes no triviales**: alrededor de 17 % en variables ligadas a la campaña (Personal Loan, Pdays, Housing Loan y Last Contact Duration) y cerca de 11 % en variables sociodemográficas y financieras centrales (Age, Credit, Education, Job, Marital Status y Balance). Esto obligó a plantear desde el inicio una estrategia sistemática de imputación (mediana para numéricas y categoría más frecuente para categóricas), dado que muchas de estas variables son potencialmente relevantes para la predicción. La variable objetivo, en cambio, no presenta faltantes.

La **distribución de la variable objetivo** evidenció un problema de desbalance importante: solo alrededor del 11,7 % de los contactos termina en suscripción, mientras que el 88,3 % restante corresponde a rechazos. Esto implica que una métrica como la accuracy puede ser engañosa y justifica el foco en medidas como AUC ROC y en la lectura cuidadosa de precision/recall sobre la clase positiva. En las variables numéricas se observan patrones bastante marcados: la edad se concentra en adultos de mediana edad (promedio cercano a 41 años, con algunos casos de edades altas), mientras que el saldo en cuenta presenta una distribución muy asimétrica, con la mayor parte de los clientes con saldos bajos o moderados y unos pocos con saldos muy elevados que aparecen como outliers. La duración del último contacto muestra una cola larga (muchos llamados breves y unos pocos muy extensos) y resulta ser la variable numérica con mayor correlación con la suscripción. Variables como Campaign, Pdays y Previous confirman que la mayoría de los clientes fue contactada pocas veces y no tenía un historial previo intenso, aunque existe una fracción con seguimiento más frecuente, que también se identifica como outlier bajo el criterio IQR.

En las **variables categóricas** se detectan perfiles y segmentos con comportamiento claramente diferenciado. Ciertas ocupaciones (por ejemplo, estudiantes y retirados) y niveles educativos más altos exhiben tasas de suscripción superiores al promedio, lo que sugiere que el contexto socioeconómico influye en la propensión a contratar el producto. Algo similar ocurre con las características financieras: los clientes sin hipoteca muestran una probabilidad de suscripción mayor que quienes tienen préstamo de vivienda. Desde el punto de vista de la interacción comercial, el canal celular es el más utilizado y el de mejor desempeño, y algunos meses específicos (como marzo, octubre o diciembre) combinan menor volumen absoluto con tasas de conversión muy elevadas. Finalmente, la variable Poutcome (resultado de campañas anteriores) se destaca como una de las más informativas: aunque la mayoría de los clientes tiene resultado “unknown”, en el subgrupo con campañas previas exitosas la probabilidad de suscripción actual se dispara a niveles muy altos (cerca de dos tercios), lo que refuerza la idea de que el historial de respuesta es un predictor clave. La matriz de correlación confirma que, más allá de la fuerte relación entre duración de la llamada y suscripción, no hay

multicolinealidad grave entre las explicativas numéricas, por lo que se dispone de un conjunto de variables relativamente complementarias para el modelado.

MATERIALES Y MÉTODOS

En este trabajo se aborda un problema de **clasificación binaria supervisada**: a partir de información sociodemográfica, financiera y de interacción comercial de cada cliente, se busca estimar la **probabilidad de suscripción** al depósito a plazo (Subscription = 1). Dado el fuerte desbalance de clases, la métrica central para comparar modelos es el **AUC ROC**, complementada con precision, recall y F1 de la clase positiva. El dataset se dividió en conjuntos de entrenamiento y prueba (70 % / 30 %) mediante `train_test_split` con estratificación en la variable objetivo para preservar la proporción de suscriptores.

Para capturar diferencias de comportamiento según el historial de contacto, se definieron tres **segmentos de clientes** a partir de las variables Pdays, Previous y Poutcome: (A) clientes sin historial de contacto previo (Pdays = −1 y Previous = 0), que representan la mayoría del dataset; (B) clientes con historial, pero sin éxito en campañas anteriores; y (C) clientes con campañas previas exitosas (Poutcome = "success"). Además de entrenar un modelo global sin segmentación, se ajustaron modelos independientes para cada segmento con el fin de evaluar si la personalización según el historial mejora el rendimiento predictivo.

En cuanto al **preprocesamiento**, se trabajó con dos enfoques paralelos. En el primero, “sin PCA”, se imputaron faltantes numéricos con la mediana y categóricos con la moda, y luego las variables categóricas se codificaron mediante One-Hot Encoding, siguiendo las recomendaciones del análisis exploratorio. Los modelos se entrenaron directamente sobre estas features transformadas. En el segundo enfoque, “con PCA”, se aplicó un `ColumnTransformer` que imputa y escala las variables numéricas, imputa y codifica las categóricas, y sobre la matriz expandida (51 columnas) se aplicó un **PCA con n_components = 0.9**, reteniendo 18 componentes principales que explican aproximadamente el 90 % de la varianza total, los cuales se usaron como entrada a los modelos.

El **ajuste de modelos** se realizó con la librería **mljar-supervised**, utilizando AutoML en modo explicativo (`mode="Explain"`) y optimizando la métrica AUC (`eval_metric="auc"`). Para cada escenario (global y por segmento, con y sin PCA) se fijó un límite de tiempo total de entrenamiento entre 300 y 900 segundos y una semilla (`random_state=42`), guardando los resultados en rutas diferenciadas por segmento. AutoML probó de manera sistemática varios algoritmos base (Baseline, Árbol de Decisión, Random Forest, XGBoost y una red neuronal densa simple) y luego construyó un **modelo ensamblado (Ensemble)** con los mejores candidatos, que en la práctica fue el modelo final seleccionado. La evaluación se realizó sobre el conjunto de test, generando para cada caso la matriz de confusión, el classification report (precision, recall, F1) y el valor de **AUC ROC** calculado a partir de las probabilidades predichas (`predict_proba`).

EXPERIMENTOS Y RESULTADOS

En el **modelo global sin PCA**, el AutoML logró un desempeño muy bueno: un **AUC ROC de 0,91** y una **accuracy de 0,90**. Esto indica que el modelo distingue bastante bien entre quienes suscriben y quienes no. Sin embargo, con el umbral estándar de 0,5 la **recall de la clase positiva** es sólo **0,36** (frente a una **precision de 0,66**), lo que sugiere que el modelo tiende a ser conservador: acierta bastante cuando predice que alguien va a suscribir, pero deja pasar muchos suscriptores potenciales (muchos falsos negativos).

Al segmentar sin PCA, el **Segmento A (sin historial previo)** reproduce casi el mismo patrón que el modelo global (**AUC ≈ 0,91, accuracy ≈ 0,92, recall ≈ 0,32–0,33**). Es decir, la segmentación no ofrece una ganancia clara de performance: el problema de identificar a la minoría que suscribe sigue siendo similar. En el **Segmento B (historial no exitoso)** el desempeño baja un poco (**AUC ≈ 0,86, accuracy ≈ 0,87**), algo coherente con el hecho de que son clientes que ya rechazaron campañas previas y, por lo tanto, resultan “más difíciles” de predecir. El **Segmento C (historial exitoso)** se comporta distinto: el modelo logra un **recall muy alto (≈ 0,90)** para la clase positiva, pero con un **AUC más bajo (≈ 0,68)** y una accuracy de alrededor de 0,67. En la práctica, en este grupo pequeño y bastante homogéneo, el modelo “agarra casi a todos los que suscriben”, pero con menos capacidad de discriminar finamente.

Cuando se repiten los experimentos trabajando sobre las **18 componentes principales de PCA**, el resultado general es una **caída sistemática del AUC**: el modelo global pasa de 0,91 a ≈ 0,88, el Segmento A de ≈ 0,91 a ≈ 0,87, y el Segmento B de ≈ 0,86 a ≈ 0,83. Solo el Segmento C se mantiene prácticamente igual (AUC ≈ 0,68 tanto con como sin PCA). Esto indica que, para este problema, la reducción de dimensionalidad **no aporta mejora en performance** y, de hecho, sacrifica algo de capacidad predictiva a cambio de tener una representación más compacta. En síntesis, los **mejores resultados** se obtienen con el **modelo global sin PCA**, seguido muy de cerca por el modelo del **Segmento A sin PCA**; el **Segmento B** rinde algo peor y el **Segmento C** es el más desafiante, tanto por su tamaño reducido como por tratarse de clientes ya fuertemente sesgados hacia la suscripción.

DISCUSIÓN Y CONCLUSIONES

El EDA mostró un dataset fuertemente desbalanceado (sólo ~11,7 % de suscripciones) y confirmó que **Last Contact Duration** es, por lejos, la variable más informativa para predecir Subscription, seguida por las variables asociadas a campañas anteriores (Pdays, Previous y especialmente Poutcome cuando hubo éxito previo). No se detectó multicolinealidad importante entre las variables numéricas, por lo que resulta razonable utilizarlas conjuntamente en los modelos.

La **segmentación por historial de contacto** (segmentos A/B/C) tiene sentido desde el negocio, porque distingue entre clientes sin historial, insistidos y previamente exitosos. Sin embargo, en términos de métricas, los modelos segmentados no superan al modelo global: el mejor AUC se obtiene con el modelo global sin PCA (0,909), muy cercano al Segmento A

(0,906), mientras que los segmentos B y, sobre todo, C presentan rendimientos algo inferiores o menos estables. Desde una perspectiva práctica, esto sugiere que es más eficiente trabajar con un **modelo global** y, si se desea personalizar, ajustar el **umbral de decisión por segmento** en lugar de entrenar modelos separados.

El uso de **PCA** permitió reducir de 51 variables a 18 componentes explicando alrededor del 90 % de la varianza, pero en todos los casos el AUC empeoró levemente respecto de los modelos sin PCA. Esto es consistente con el hecho de que PCA es no supervisado y optimiza varianza total, no capacidad predictiva. Dado que el entrenamiento sin PCA es computacionalmente manejable, la reducción de dimensionalidad no se justifica.

En síntesis, el **modelo recomendado** es el AutoML global sin PCA, entrenado sobre todo el dataset tras imputación y One-Hot encoding, basado en un ensemble de árboles, Random Forest, XGBoost y redes neuronales, con AUC ≈ 0,91 y accuracy ≈ 0,90. Para un despliegue real se sugiere ajustar el umbral para aumentar el recall en la clase positiva cuando el costo de contacto sea bajo, y complementar con técnicas de re-muestreo y reglas de negocio por segmento. Como líneas de trabajo futuro, resultaría útil incorporar costos y beneficios para optimizar métricas económicas, calibrar las probabilidades predichas y estudiar la estabilidad temporal del modelo entre distintas campañas.

REFERENCIAS

A modo de marco teórico y para fundamentar las decisiones metodológicas, se consideraron las siguientes referencias:

1. **Moro, S., Cortez, P., & Rita, P. (2014).**
   A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31.
2. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).**
   The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
3. **Géron, A. (2019).**
   Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.
4. **Piotrowski, P. (2020).**
   Documentación y ejemplos de *mljar-supervised* (AutoML en Python).