INFORME: BANK SUBSCRIPTION

APLICACIÓN DE MODELOS DE CLASIFICACIÓN
Y ANÁLISIS EXPLORATORIO DE DATOS

Universidad Tecnológica Nacional
Cátedra de Ciencia de Datos – 2023
Franco, Olivia *Muñoz Little, Santiago Juan*

Abstract: En el presente informe desarrollamos un análisis exploratorio de datos para entender la naturaleza de los clientes del banco y su respuesta a la campaña de marketing. Por otro lado, determinamos un modelo de aprendizaje supervisado de clasificación para predecir la aceptación de la campaña ante futuros clientes.

Introducción

El principal propósito del informe es conocer la naturaleza de los clientes de un banco y su aceptación a la campaña de marketing realizada. Entendemos que medir la eficiencia de las campañas es fundamental para aplicar mejoras en futuros desarrollos y garantizar la satisfacción de las necesidades de los clientes.

Para lograr dichos objetivos, desarrollaremos a lo largo del trabajo:

- Un Análisis Exploratorio de datos (EDA) sobre el data set bancario mediante el uso de herramientas como histogramas, boxplots, scatterplots y gráficos de torta.
- Un modelo de clasificación de datos para realizar una predicción de la variable categórica 'Subscription'
- Un método de PCA (reducción de dimensionalidad) para volver a hacer las predicciones del modelo con el data set reducido y evaluar si mejora su eficiencia.

Descripción del Data set

El data set utilizado pertenece a un banco con una cartera de clientes de 45.211 personas. Por cada individuo se presentan 17 variables que brindan información sobre las características de la persona, su relación con el banco y su reacción frente a las campañas de marketing.

Las primeras variables son aquellas que describen las propiedades intrínsecas del cliente, tales como su edad, estado civil, trabajo y educación. Para los trabajos, obtenemos información sobre las profesiones y oficios de las personas, pudiendo también incluir si son estudiantes o jubilados. En el caso de educación, obtenemos la educación máxima alcanzada por el usuario. Estas características nos permiten analizar la demografía de la cartera de clientes.

Las siguientes son las variables que vinculan al cliente con el banco. El balance, determina la cantidad de dinero depositado en el banco, el crédito nos dice si el individuo tiene o no deudas de crédito. Luego, se indican los tipos de préstamos que son tomados por los usuarios y se dividen entre personales y de viviendas.

Por último, obtenemos variables que brindan información sobre la relación entre los clientes y las campañas de marketing realizadas actualmente y en el pasado. Conocemos la cantidad de contactos durante y previos a la campaña de marketing y el modo (teléfono o celular), cuál fue el día, mes y duración del último contacto con el usuario, cuanto tiempo pasó desde la última comunicación, y el éxito o fracaso de las campañas anteriores y actuales.

Es importante mencionar que el data set posee una combinación de variables categóricas y numéricas que deben ser tenidas en cuenta para desarrollar el análisis de los datos. Las variables categóricas son el estado civil, trabajo, educación, crédito, préstamos, seguros del hogar, contacto, performance de campañas anteriores y actuales, y último mes de contacto. Por otra parte, las variables numéricas son la edad, balance, último día de contacto y duración, contactos durante y previos a la campaña y días que pasaron desde el último contacto.

A su vez, una característica importante del data set es la presencia de valores nulos y “unknowns”, que es vital darles un tratamiento para evitar que generen errores en la interpretación de los datos y en el aprendizaje del modelo de clasificación.

Análisis Exploratorio de Datos

Luego de pre-procesar los datos, comenzamos analizando las variables numéricas y su relación entre ellas. Al obtener las métricas más importantes, llegamos a las siguientes conclusiones:

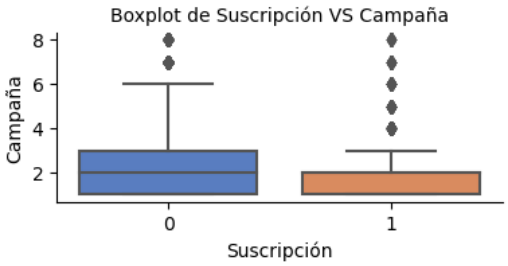
- La variable Pdays demuestra que hay una amplia concentración de valores -1, lo que indica que no hubo contacto previo con gran

cantidad de los clientes de la campaña anterior. Esto puede generar inconvenientes a la hora de aplicar el modelo de aprendizaje.

- La variable del balance indica la presencia de outliers, ya que el máximo es muy superior a la media y mediana de la distribución.
- La variable Campaign muestra una distribución entre 1 y 3 días, aunque presenta valores anómalos. Es una variable interesante para ver su relación con la suscripción.

	count	mean	std	min	25%	50%	75%	max
Age	10630.0	41.089276	10.652741	18.0	33.0	39.0	49.0	95.0
Balance (euros)	10630.0	1357.860960	3028.454521	-2604.0	72.0	446.0	1454.0	81204.0
Last Contact Day	10630.0	15.696331	8.328843	1.0	8.0	16.0	21.0	31.0
Last Contact Duration	10630.0	256.620978	259.270058	0.0	104.0	180.0	316.0	4918.0
Campaign	10630.0	2.780245	3.113898	1.0	1.0	2.0	3.0	55.0
Pdays	10630.0	40.912888	100.954838	-1.0	-1.0	-1.0	-1.0	828.0
Previous	10630.0	0.563311	1.861234	0.0	0.0	0.0	0.0	51.0
Subscription	10630.0	0.112982	0.316586	0.0	0.0	0.0	0.0	1.0

Se avanzó con un pairplot del cual la única información relevante para el análisis que pudimos obtener fue que a mayor cantidad de contactos (campaign) pareciera que hay mayor rechazo a la campaña. Este razonamiento fue acompañado por el siguiente Boxplot:

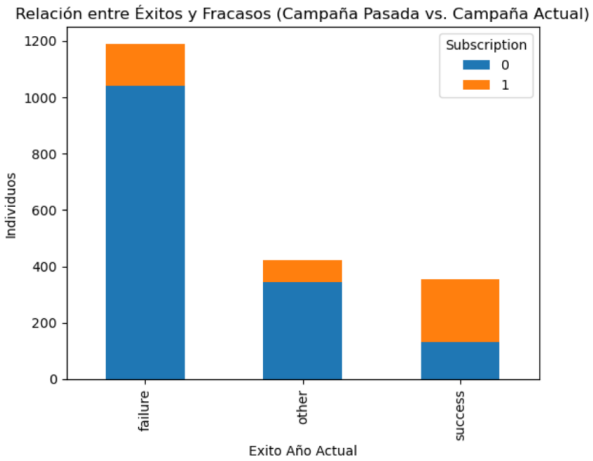


Continuamos el análisis, con la matriz de correlación donde sólo el par de variables Pdays y Previous demuestran una correlación significativa positiva, con un valor de 0,54. Esto se explica porque hay gran cantidad de Pdays=-1 (no hubo contactos), por lo que Previous va a ser 0.

El foco del análisis ahora estará sobre la variable suscripción.

Primero se determinó la distribución de clientes suscriptos a la campaña dando un resultado de 88,7% que no se suscribieron y un 11,3% que sí lo hicieron. Resultó relevante, continuar el análisis determinando la relación entre la performance de la campaña anterior y la actual: se observa que hay gran aceptación de la campaña actual por los individuos que aceptaron la campaña previa.

Asimismo, se puede analizar qué llevó a los clientes a suscribirse a la campaña actual, habiendo rechazado la anterior y analizar si esos factores pueden ser un patrón que se magnifique y logre revertir más clientes que habían rechazado para que acepten la campaña actual.



Otra variable importante que se analizó ligada a las suscripciones fueron los créditos, y se llegó a la conclusión de que la campaña actual tiene mal rendimiento para los clientes que poseen créditos y préstamos, mientras que es más equitativa para el caso de los clientes que tienen seguro para el hogar. Además, denotamos que hay muy pocos clientes con créditos (este podría ser el foco para una campaña futura) y préstamos.

Para conocer la demografía de nuestros clientes, analizamos la variable **Edad**. El 50% de los usuarios tiene entre 33 y 49 años, con una media en 41 años. A su vez, hay un fuerte descenso en la cantidad de clientes mayores a 60 años. Al visualizar la distribución de edades por cada trabajo, vemos que el rango de edades indicado coincide con todos los clientes que tienen un oficio o profesión, destacándose por encima, los jubilados (con una mediana en 60 años) y por debajo los estudiantes (con una mediana en 25 años).

Siguiendo el análisis según los **trabajos**, observamos que los trabajos más frecuentes en la cartera de clientes son administradores, gerentes, obreros y técnicos, donde, junto con los jubilados, encontramos la mayor concentración de suscripciones a la campaña. El tipo de trabajo puede ser un foco relevante para asociarse con distintas empresas. Por otra parte, la distribución del Balance según los trabajos es relativamente estable para todos ellos, salvo para los jubilados que poseen una mediana que llega a duplicar las medianas de los demás empleos.

Continuando con el análisis del **balance**, observamos una gran disparidad en los montos. Para empezar, el 50% de los clientes tienen un balance entre 70 y 1.400 euros, pero se destaca la existencia de anomalías que alcanzan los 81.000 euros. Por otra parte, al comparar el balance con la edad, vemos que los balances más elevados se concentran entre los 40 y 60 años; y que, a mayor balance, menor suscripción a la campaña actual.

La variable **educación** también fue analizada junto a distintas variables. En cuanto a las suscripciones,

el estudio nos permitió ver que la mayor cantidad de suscriptores se encuentran en clientes cuya educación máxima fue secundaria o terciaria. Al analizar educación con el balance, contrario a lo que se podría esperar, la mediana de todos los clientes con distintos tipos de educación se encuentra al mismo nivel, teniendo una distribución un poco más amplia y elevada para los terciarios, pero sin significar una gran diferencia respecto de las otras dos educaciones máximas.

A continuación, se determinó el estudio de los contactos **(Last contact day y Last Contact Duration)**, pudiendo observar que la mediana de la duración del último contacto se encuentra alrededor de los 3 minutos y el 75% no superaba los 10 minutos. Sin embargo, hay gran presencia de outliers, lo que indica que hay contactos de mucha mayor duración que pueden ser perjudiciales para el modelo. Para el día de contacto, vimos que la distribución fue bastante uniforme, teniendo contactos todos los días del mes, con una mayor concentración entre los días 16 y 21, pero sin ser lo suficientemente notorio.

Por otro lado, se hizo un boxplot de las campañas y se observó que el 75% de los clientes fueron contactados menos de 4 veces, aunque, por la presencia de outliers, notamos que hubo clientes que fueron contactados muchas veces más.

Por último, nos pareció importante analizar las suscripciones respecto de la distribución de los estados civiles y los meses de contacto. En cuanto a los estados civiles, observamos que el 60% de los usuarios están casados y el 25% están solteros, y son quienes mayor concentración de suscriptos poseen.

El análisis de los meses refleja que la mayor cantidad de últimos contactos fue realizada entre Mayo y Agosto, con mayor concentración de suscriptos, aunque no creemos que el éxito esté directamente relacionado con el mes.

Materiales y Métodos

Antes de entrar en detalle sobre los modelos de aprendizaje utilizados, debemos tener en cuenta que necesitamos datos standard, lo que requiere de diversos pasos de pre-procesamiento como limpieza, filtro, selección y escalado de variables.[1]

Fillna con media: se realiza un reemplazo de los nulos presentes en las variables numéricas por la media con el objetivo de eliminar la dispersión de la base de datos.

One-Hot Encoder: para las variables que toman valores categóricos aplicamos una transformación para obtener una nueva variable que tome valores binarios por cada categoría existente.

StandardScaler: en este enfoque, cada valor en la característica se transforma restando la media de la característica y dividiendo por la desviación estándar. La fórmula es:

$$Z = \frac{x - \mu}{\sigma}$$

donde *z* es el valor estandarizado, *x* es el valor original, *μ* es la media de la característica y *σ* es la desviación estándar de la característica. Esto permite suavizar el efecto de los outliers.

Ahora sí, desarrollaremos los tres modelos utilizados para el modelo clasificador, donde el objetivo es asignar una etiqueta a una instancia de datos basada en sus características observadas.

Logistic Regression: es un clasificador lineal, que utiliza una regresión lineal precedida de una función de activación sigmoide, lo que genera que el output sea binario. A cada muestra clasificada, le asigna una probabilidad de pertenecer a cada clase y, de superar el umbral, la asigna a una clase.

$$Sigmoide = \sigma = \frac{1}{1 + e^{-x}}$$

$$Fun\ Cost = Lce = -[y.\log y' + (1 - y).\log(1 - y')]$$

SVC: el modelo SVM es una extensión del clasificador que resulta de ampliar el espacio de características de una manera específica, utilizando kernels.[2] El modelo intenta encontrar un hiperplano que mejor separe las instancias de diferentes clases en un espacio de características. La implementación específica para la clasificación se llama Support Vector Classification (SVC).

Linear SVC: similar al SVC con kernel lineal, pero implementado en término de liblinear en vez de libsvm, por lo que tiene mayor flexibilidad en las penalidades y función de costos, y suele escalar mejor ante gran cantidad de muestras.[3]

Por otra parte, para evaluar el desempeño de los modelos utilizamos los siguientes métodos:

Curva ROC-AUC: la curva característica operativa del receptor (ROC) es un gráfico que ilustra el rendimiento de un modelo clasificador binario o múltiple en valores de umbral variables. El área bajo esta curva (AUC) se usa como un indicador de la calidad del clasificador[4]. La curva ROC se obtiene formando estas predicciones y calculando las tasas de falsos positivos y verdaderos positivos

[1] (Manjurul Ahsan, Parvez Mahmud, Kumar Saha, Datta Gupta, & Siddique)

[2] (James, Witten, Hastie, & Tibshirani, 2023)
[3] (Scikit Learn, s.f.)
[4] (Díaz Barrios, 2015)

para un rango de valores de t. Un clasificador óptimo abrazará la esquina superior izquierda del gráfico de la curva ROC.

Matriz de confusión: es un elemento para evaluar los resultados de la clasificación. En cada posición se cuentan los True Positive (TP), True Negative (TN), False Positive (FP), False negative (FN). Luego se obtienen los siguientes coeficientes:

- Accuracy = (TN+TP) / Total
- Sensitivity = TP/(TP+FN)
- Specificity = TN/(TN+FP)

Por último, para realizar la reducción de dimensionalidad se utilizó el siguiente modelo:

PCA: el análisis de componentes principales se refiere al proceso mediante el cual se calculan los componentes principales y el uso posterior de estos componentes para comprender los datos. Además de producir variables derivadas para su uso en problemas de aprendizaje supervisado, PCA también sirve como una herramienta para la visualización de datos (visualización de las observaciones o visualización de las variables) y para la imputación de datos, es decir, para completar los valores faltantes en una matriz de datos.[5]

Experimentos y Resultados

Para comenzar con el desarrollo de los modelos, realizamos un pre-procesamiento de los datos. En primer lugar, se analizaron los tipos de variables y los valores únicos de cada una. Aquí se decidió eliminar las columnas 'Poutcome' y 'Contact', ya que poseían gran cantidad de 'unknowns', por lo que no brindan información y variabilidad.

Por otro lado, 'Pdays' fue transformada ya que más del 75% de los registros son -1 (valor arbitrario), por lo que los datos no representaban correctamente la naturaleza de la distribución. La transformación se llevó a cabo generando 3 rangos: 'Sin Contacto' para -1, 'Contacto Cuatrimestre' para menores de 120 y 'Contacto Largo Plazo' para mayores a 120. Esto se decidió luego de haber eliminado 'Pdays' en pruebas anteriores y obtener mal desempeño del modelo.

Para el caso de nulos en variables numéricas se tomó la decisión de imputar la media, mientras que para los nulos y 'unknowns' de las variables categóricas se decidió eliminar los registros, ya que imputar la moda puede ser dañino para la variabilidad del modelo.

La segunda parte de la ingeniería de atributos consta de la eliminación de las anomalías presentes en Balance y Last Contact Day que fueron observadas en el EDA. Se utilizaron los percentiles 99 y 1 para realizar el filtrado.

El último paso previo a la aplicación de los modelos fue convertir la variable categórica 'Last Contact Month' en una numérica, mediante el uso de un diccionario; y la generación de Dummies para las demás variables categóricas.

Con todo el data set pre-procesado, avanzamos con el pipeline de Entrenamiento, donde el primer paso es dividir el Train (70%) y Test (30%) y escalarlos con un Standard Scaler.

A continuación, se aplican los mismos pasos para los modelos de Logistic Regression, SVC y Linear SVC, que son: hacer GridSearch para hallar los mejores hiper-parámetros del modelo y Cross Validation con el Train; luego, seleccionamos el mejor modelo e hiper-parámetros y hacemos las predicciones con el Test. Por último, evaluamos los resultados con confusion matrix y la curva ROC.

Para el modelo de Logistic Regression se validaron los siguientes hiper-parámetros: penalidad Lasso y Ridge, C (costo) de 0,001, 0,01, 0,1, 1, 10 y 100, y un solver liblinear y saga. De estos, los mejores hiper-parámetros resultaron ser: C de 0,001, penalty lasso y solver saga. Al analizar las métricas, obtuvimos:

AUC: 0,82	Sensibility: 0,6349
Accuracy: 0,8116	Specifiticy: 0,8336

Por más de obtener unos valores de AUC, Accuracy y Specificity relativamente buenos, la sensibility no es adecuada por ser baja.

Para el modelo de SVC se validaron los siguientes hiper-parámetros: gamma de 0,01, 0,1, 1 y 'auto', C (costo) de 0,1, 1, 10, y kernel linear y rbf. De estos, los mejores hiper-parámetros resultaron ser: C de 1, gamma de 0,01 y kernel rbf. Al analizar las métricas, obtuvimos:

AUC: 0,86	Sensibility: 0,7748
Accuracy: 0,8084	Specifiticy: 0,8126

Este modelo implica un amplio aumento en la sensibilidad y AUC, a costa de una pequeña reducción en el accuracy y specificity. Por lo tanto, creemos que el SVC presenta una importante mejora respecto de LR.

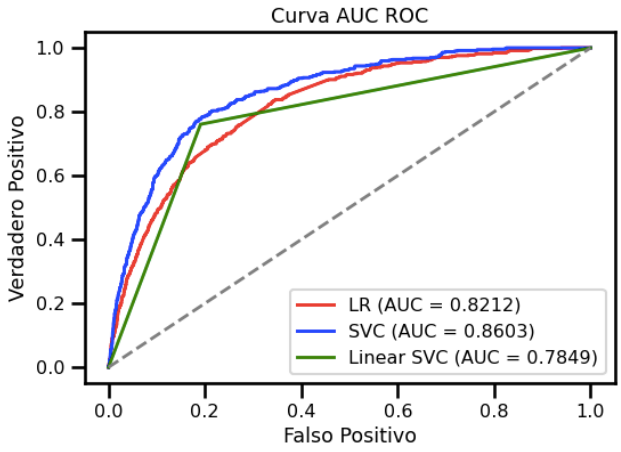
Por último, para el modelo de Linear SVC solamente se validó el hiper-parámetro de costo (C) en valores 0,1, 1 y 10, para el cuál 0,1 resultó ser el mejor. Al analizar las métricas, obtuvimos:

AUC: 0,78	Sensibility: 0,7606
Accuracy: 0,8037	Specifiticy: 0,8091

[5] (James, Witten, Hastie, & Tibshirani, 2023)

Este modelo no significa un progreso ante lo obtenido por el SVC, ya que se reduce de manera importante el AUC y las demás métricas también tienen un pequeño descenso.

A continuación, se presenta una visualización del conjunto de métricas previamente mencionadas:

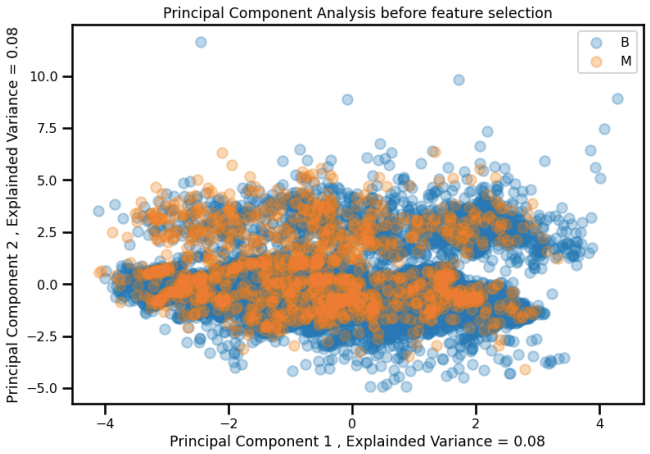


Métrica	LR	SVC	LinearSVC
Sensitivity	0.634888	0.774848	0.760649
Specificity	0.833586	0.812626	0.809091
Accuracy	0.811588	0.808444	0.803728

Por el momento, SVC resulta ser el mejor modelo para la clasificación ya que posee, en conjunto, la mejor combinación de métricas.

Para complementar el desarrollo, se adicionó una etapa de reducción de dimensionalidad (PCA), donde se determinó extraer 26 componentes, ya que pruebas realizadas con menores indicaban gran variabilidad en todas sus features. Recién a partir de 27 features es cuando detectamos que la feature 27 no explica variabilidad, con una varianza de 5,47e-31, mientras las anteriores tienen una varianza entre 2,78 y 0,39.

Al visualizar el PCA en 2D, vemos que las dos primeras columnas no son suficientes para detectar una distribución distintiva para el fenómeno de suscripciones.



Para entrenar el modelo con la reducción de dimensionalidad, utilizamos SVC, ya que fue la herramienta que mejores resultados arrojó en la etapa previa. Sin embargo, contrario a lo que se podría esperar del modelo con PCA, los resultados obtenidos fueron críticamente inferiores a los del SVC sin PCA, con las siguientes métricas:

AUC: 0,57	Sensibility: 0,3976
Accuracy: 0,7083	Specifiticy: 0,7470

Es por eso, que se decide desarrollar finalmente el modelo SVC sin reducción de la dimensionalidad, con un accuracy del 80% y un AUC del 0,86.

Conclusiones

El rubro bancario se apalanca fuertemente en las campañas publicitarias para conseguir que sus clientes adquieran créditos, plazos fijos u otros servicios que ofrecen. Por lo tanto, analizar la eficiencia de dichas campañas y los factores de éxito es vital para aplicar medidas correctivas y potenciar los beneficios.

Como se mostró en el análisis exploratorio de datos, hay factores que presentan cierta relación, como balance y cantidad de contactos con las suscripciones, donde es necesario analizar si está relación representa una causalidad que se pueda utilizar para maximizar las suscripciones.

Por otra parte, en el modelo de aprendizaje observamos que el modelo radial (rbf) se adapta mejor a los datos que los modelos lineales planteados, obteniendo parámetros buenos, cercanos al 80%, pero que se podrían mejorar utilizando otras herramientas más complejas, como redes neuronales.

En conclusión, estamos conformes con los resultados obtenidos, fruto de diversos experimentos realizados, y creemos que la información provista puede ser de gran utilidad para la toma de decisiones en el banco y para futuras campañas de marketing.

Referencias

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R.*

Manjurul Ahsan, Parvez Mahmud, Kumar Saha, Datta Gupta, & Siddique. (s.f.). *Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance.*

Díaz Barrios, A.-R. C.-H. (2015). *Machine Learning algorithms for Splice Sites classification in genomic sequences.*

Scikit Learn. (s.f.). Obtenido de https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html