Bank Subscriptions

1. Introducción y objetivos

En el presente trabajo se buscará predecir potenciales suscriptores a campañas de marketing lanzadas por un banco. El banco, el cliente, proveyó de un dataset de 45211 personas con 17 variables que muestran algunas características de sus clientes.

2. Descripción del dataset

Variable	Significado
Age	Edad.
Job	Tipo de empleo.
Marital Status	Estado civil.
Education	Educación máxima alcanzada.
Credit	Si tiene deuda de crédito o no.
Balance Euros	Promedio del saldo de la cuenta en el año.
Hosing Loan	Si tiene seguro del hogar o no.
Personal Loan	Si tiene prestamos o no.
Contact	Tipo de contacto del cliente.
Last Contact Day	Último día de contacto en el mes.
Last Contact Month	Último mes de contacto en el año.
Last Contact Duration	Duración del último contacto medido en segundos.
Campaign	Cantidad de contactos durante esta campaña, incluye el último contacto.
PDays	Cantidad de días que pasaron del último contacto de una campaña anterior. -1 significa que no hubo contacto previo.
Previous	Cantidad de contactos previos a esta campaña.
POutcome	Performance de la campania de marketing anterior
Subscription	Si el cliente accede a la campania o no.

Tabla 1. Diccionario dataset

2.1. Registros con información faltante

Del total de 45211 registros del dataset, se decidió eliminar 34468 registros por falta de información en los mismos, quedando un total de 10743 registros.

2.2. Información conflictiva

Dentro del dataset, la variable "PDays", marca la cantidad de días que pasaron desde el último contacto, y "-1" si no hubo contacto previo. Al estudiar el dataset, se halló discrepancia con dicha descripción provista por el banco. Se encontraron múltiples registros donde figuraba "-1", es decir, sin contacto previo, pero aún así se registraban valores en "Last Contact Day", "Last Contact Month" y "Contact Duration", dando a entender que efectivamente hubo un contacto previo. Al haber encontrado varios registros con esta falla, se decidió descartar dicha variable.

2.3. Outliers

Al observar cada variable, naturalmente se encontraron outliers en varias de ellas, de todas formas, no se consideró necesario ni pertinente eliminar estos outliers a excepción de un único registro, que marcaba en la variable "Previous", un total de 275 contactos realizados a dicho cliente. En contraste con el resto de los registros, y por la naturaleza de la variable, no se considera algo normal, que pueda, o deba volver a darse. Se decidió eliminar dicho registro del análisis.

3. Análisis exploratorio de datos

3.1. Variables numéricas

	age	balance_euros	last_contact_day	last_contact_duration	campaign	previous
count	10742	10742	10742	10742	10742	10742
mean	40.85	1381.54	15.81	262.95	2.73	0.56
std	10.54	3010.81	8.31	265.40	2.99	1.89
min	18	-6847	1	1	1	0
25%	33	76.25	8	104	1	0
50%	39	454	16	180	2	0
75%	48	1436.75	21	326	3	0
max	92	64343	31	4918	50	38

Tabla 2. Descripción variables numéricas

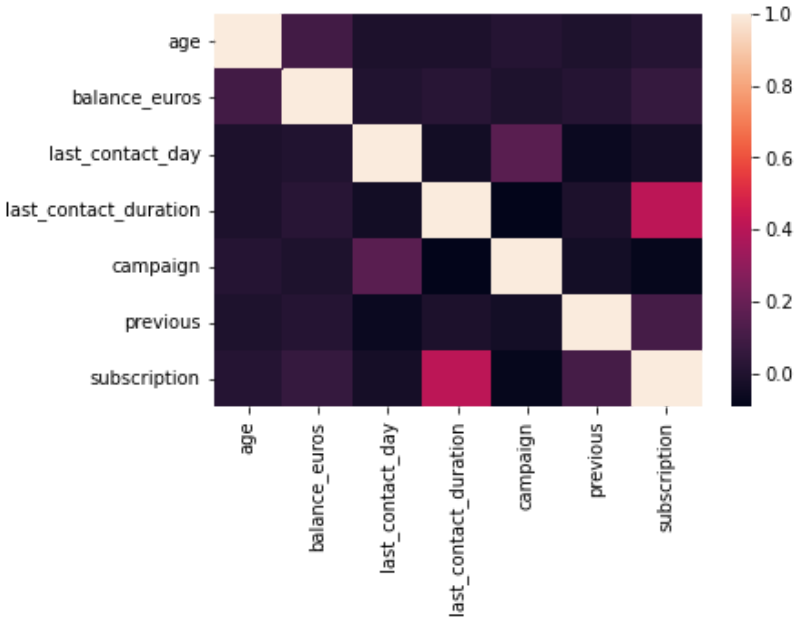


Figura 2. Matriz de correlaciones variables numéricas

A continuación, se visualiza la relación entre estas variables numéricas, coloreadas por clientes suscriptos (puntos negros) y no suscriptos (puntos claros).

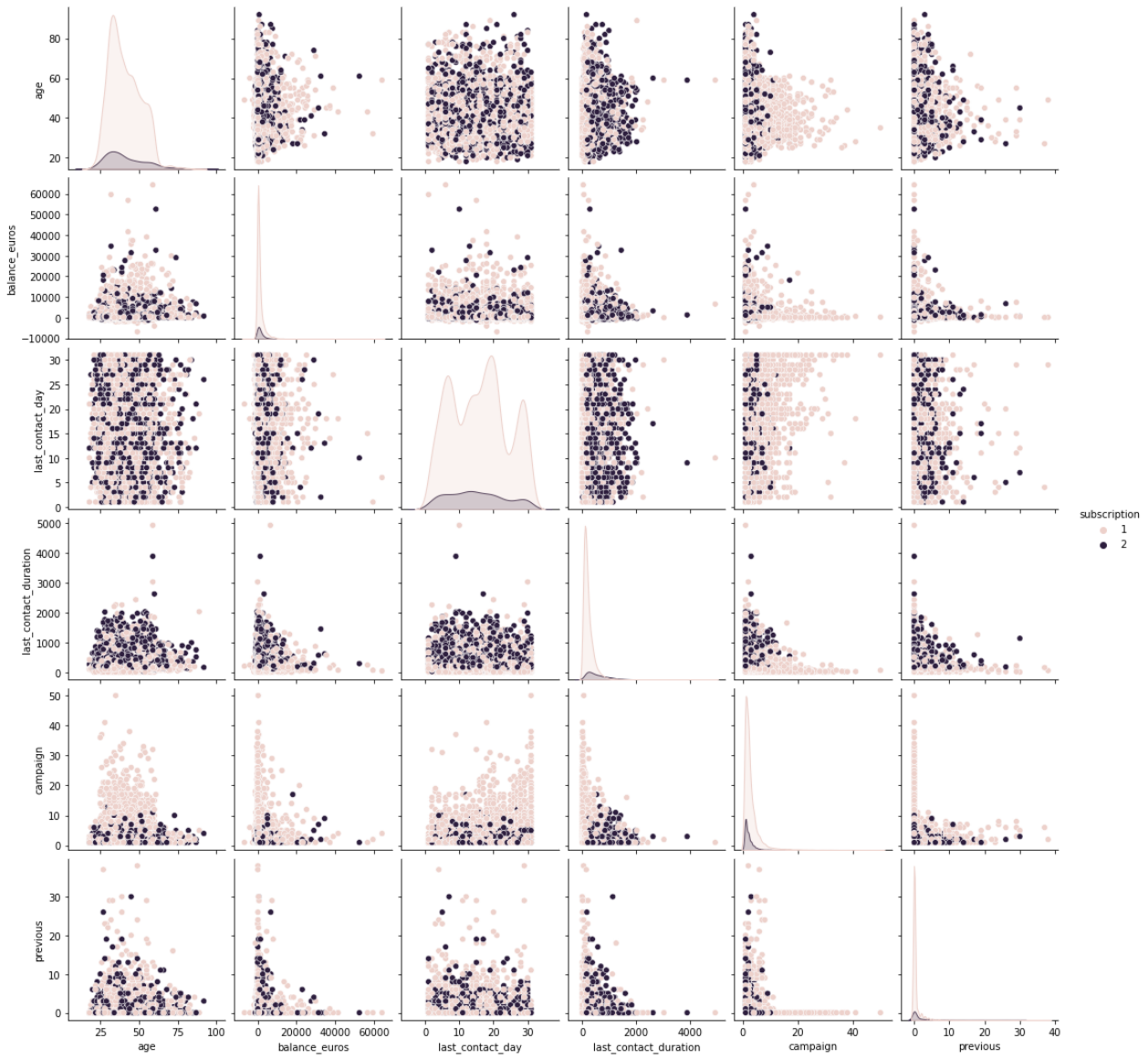


Figura 1. PairPlot variables numéricas

En varias de estas relaciones, se puede apreciar algún patrón aparente. Por ejemplo, a mayor duración de la llamada, se puede observar una mayor concentración de suscriptos.

3.2. Variables no numéricas

3.2.1. Variables binarias

Se identificaron variables que por naturaleza serían binarias, y que estaban expresadas como “yes” o “no” en el dataset. A éstas se las transformó en 1 y 0 respectivamente. Siendo estas variables: “Credit”, “Hosing Loan”, y “Personal Loan”.

3.2.2. Variables categóricas

Para las variables categóricas “Job”, “Marital Status”, “Education”, “Contact”, “Last Contact Month”, “Previous”, y “POutcome”, se crearon variables “dummies”.

4. Materiales y métodos

4.1. StandardScaler - Preparación dataset

Al contener el dataset a las variables binarias y dummies, se buscará separar estas variables para que no sean afectadas por el StandardScaler, para luego, re-unificar las variables transformadas junto con las mencionadas y conseguir efectivamente el set de variables que se inyectarán en el modelo. Por otro lado, se transforma la columna de suscriptos en 0 y 1 en vez de 1 y 2. De aquí quedarán 3 conjuntos separados

- X → Contiene las variables numéricas
- Dummies → Contiene las variables binarias y dummies
- Y → Contiene las etiquetas, correspondiente a la columna suscription

4.2. StandardScaler - Test y Train

Se separó los 3 grupos en Test y Train, bajo la misma estructura de modo de una vez realizada la transformación, poder volver a unir las columnas de X y Dummies para formar el conjunto de datos que se inyectará al modelo.

4.3. Modelo - SVM

Al ser un desafío de clasificación supervisada con etiquetas provistas por el banco, se optó por un modelo Support Vector Machine, provisto por Sklearn.

4.3.1. Selección de Hiperparámetros

Se realizó un GridSearch del modelo SVM con los siguientes hiperparámetros:

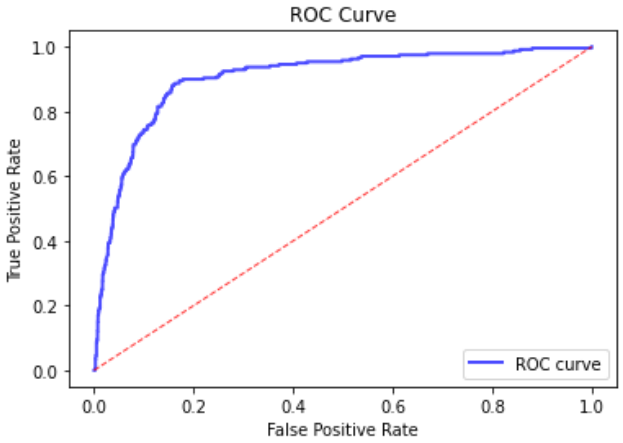
- Kernel
  - rbf
  - Linear
- C
  - 0.1
  - 1
  - 10
  - 100
- Gamma
  - 0.1
  - 0.01
  - 0.001
  - 0.0001

Tras los resultados del GridSearch, se realizó el modelo SVM con la combinación:

- Kernel → rbf
- C → 1
- Gamma → 0.1

4.3.2. Resultados Modelo Predictivo

Accuracy = 0.9069 **AUC = 0.9044**



5. Discusión y conclusiones

A partir de los datos provistos por el cliente, se ha logrado conformar un modelo relativamente confiable, dado que, evidentemente, existen relaciones bien ligadas entre los puntos de información registrados por el banco y la suscripción o no del cliente a la campaña. Dicho esto, el modelo se construyó con menos del 25% de los datos provistos, es decir, más del 75% de los datos no fueron útiles, principalmente por no contar con la información completa. Por otro lado, dado el error de la columna “PDays”, recomendaría revisar el registro de la información por parte del cliente, de modo de asegurar datos más completos y confiables.
Fuera de estas sugerencias, el modelo final consigue, en forma consistente, clasificar a los clientes que se suscribirían a la campaña en cuestión. Con un accuracy del **XXX** y AUC del **XXX**, el banco puede utilizar el modelo propuesto para un mejor planeamiento de sus campañas, de modo de hacerlas más predecibles y programables alineado a sus objetivos.

6. Referencias

Liu, Y. (., Mirjalili, V., Raschka, S., & Dzhulgakov, D. (2022). *Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python*. Packt Publishing.

scikit-learn. (n.d.). scikit-learn: machine learning in Python — scikit-learn 1.1.3 documentation. Retrieved December 1, 2022, from https://scikit-learn.org/

Witten, D., Hastie, T., James, G., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US.