Bank Subscription

Florencia Bellone
fbellone@frba.utn.edu.ar

Naara Cuniglio
ncuniglio@frba.utn.edu.ar

Abstract
A continuación, se detallará brevemente el análisis realizado de la base de datos de clientes de un banco para luego seleccionar un modelo que prediga como será la reacción de estos frente a una campaña de marketing.

1 INTRODUCCIÓN

¿Es nuestra campaña de marketing eficiente? Se desarrollará un modelo específico para analizar la performance de una campaña de marketing para nuevos subscritores a nuestros productos. A partir de la cartera de clientes, se analizarán las distintas variables, entenderán su importancia, para luego a partir de estos datos históricos obtener información útil y poder predecir el comportamiento de futuros clientes frente a nuevas campañas.

2 DATA SET

Para el análisis en cuestión se utilizó una base de datos de 45.211 clientes con 17 variables.

- Age: Edad del cliente
- Job: Tipo de empleo del cliente
- Estado civil
- Education: Educación máxima alcanzada por el cliente
- Deuda: Si tiene deuda de crédito o no
- Balance: Promedio de saldo en la cuenta en el año
- Seguro vivienda: Si tiene seguro de hogar o no
- Préstamo: Si tiene prestamos o no
- Contact: Tipo de contacto del cliente
- Dia de contacto: Ultimo día de contacto con el cliente en el mes
- Mes de contacto: Ultimo mes de contacto con el cliente en el año
- Duración: Duración del último contacto con el cliente medido en segundos
- Cantidad de contactos: Cantidad de contactos al cliente durante esta campaña, incluye el último contacto.
- Pdays: Cantidad de días que pasaron del último contacto con el cliente de una campaña anterior. -1 significa que no hubo contacto previo
- Previous: Cantidad de contactos previos a esta campaña para cada cliente
- Performance: Performance de la campaña de marketing anterior para este cliente
- Subscription: Si el cliente accede a la campaña o no.

2.1 Modelos

Dado el costo que cada modelo involucra, se realizaron dos modelos distintos, para luego seleccionar el que mejor se adaptaba a el caso.

- Suppor vector machine: modelo que maximiza la frontera de decisión a la hora de clasificar nuestros datos. Permitiéndonos, a su vez, flexibilidad a la hora de generar un modelo no lineal, a través de la utilización de un Kernel Gaussiano.
- KNN: Se utilizo este modelo en comparación con el SVM, considerando que es más sensible a los datos y gracias a la posibilidad de contar con los datos históricos de cómo se reaccionó a campañas anteriores. (1)

3 DESCRIPCIÓN DEL DATA SET

3.1 Introduction

Se comenzó haciendo un análisis exploratorio de los datos, para así conocer bien sobre las variables con las que estábamos tratando. Poder realizar una limpieza de información que podía estar erróneamente cargada e iba a entorpecer el análisis y poder entender cómo se relacionaban nuestras variables para poder enriquecer las conclusiones. Como así también diversas transformaciones para poder trabajar mejor los datos a analizar.

3.2 Variables

Para comprender con qué tipo de clientes se formó nuestra base se hizo una visualización de las variables que nos aportaban la mayor cantidad de información de los mismo.

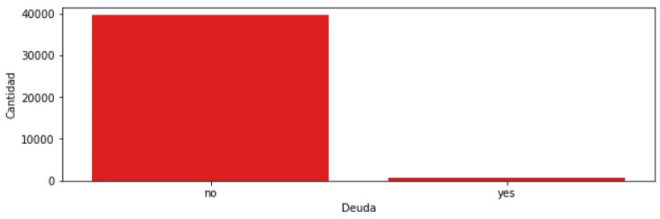

Figura 1 se puede apreciar la distribución de la variable Deuda.

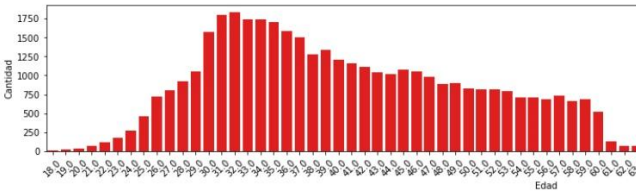

Figura 2 se puede apreciar la distribución de la variable Edad y como la misma esta sesgada hacia la derecha.

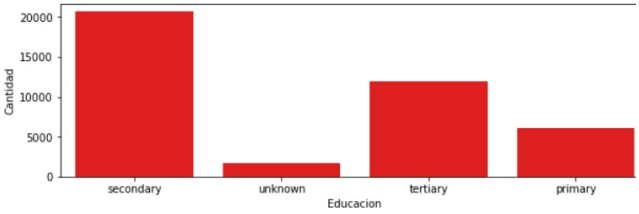

Figura 3 Distribución de la variable Educación

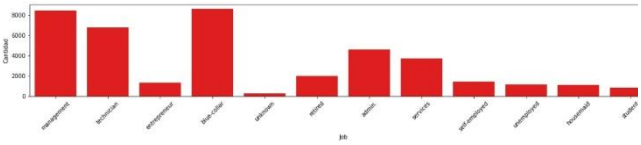

Figura 4 Distribución de la variable Job.

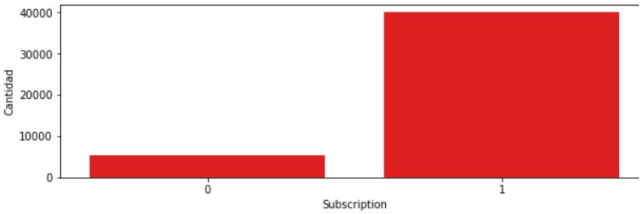

Figura 5 Distribución de la nuestra variable objetivo en los datos históricos.

En Figure 1 podemos apreciar como la distribución parece estar sesgada hacia la derecha, lo que significa que la mayoría de los clientes son de mediana edad. Como podemos observar en la información estadística [Tabla 1], con una media de 41 años.

	Age	Balance	Dia_contacto	Mes_contacto	Duracion	Cant_contacto
count	19586.000000	19586.000000	19586.000000	19586.000000	19586.000000	19586.00000
mean	40.951798	1370.929649	15.805320	6.147708	260.449737	2.74711
std	10.020491	2757.897096	8.307631	2.399746	240.118748	3.03019
min	18.000000	-6847.000000	1.000000	1.000000	1.000000	1.00000
25%	33.000000	106.250000	8.000000	5.000000	118.000000	1.00000
50%	40.951798	592.000000	16.000000	6.000000	221.000000	2.00000
75%	47.000000	1370.929649	21.000000	8.000000	279.000000	3.00000
max	92.000000	64343.000000	31.000000	12.000000	4918.000000	50.00000

Tabla 1. Descripción estadística

También podemos observar que la mayoría de los clientes que han adquirido nuestros productos en campañas anteriores son casados. [Figura 6]

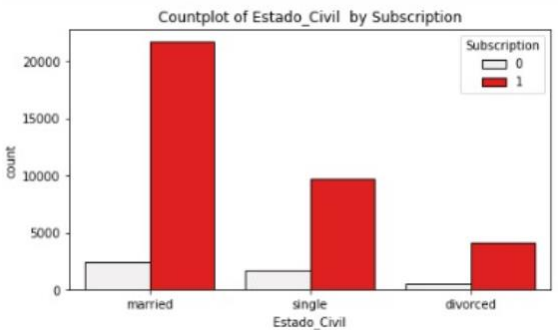

Figura 6 Suscripción por estado civil.

En contraposición se puede observar que el mismo clúster que mas suscripciones ha realizado es el que menos deuda crediticia posee.

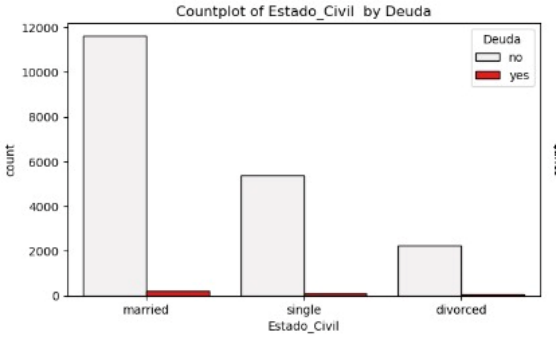

Figura 7 Deuda por estado civil

En adición, si buscamos entender en cuanto al balance anual y los estudios, podemos apreciar que las personas con secundario y terciario cuentan con los balances mas altos, mientras que en los cuales no se conoce el estudio, es muy bajo.

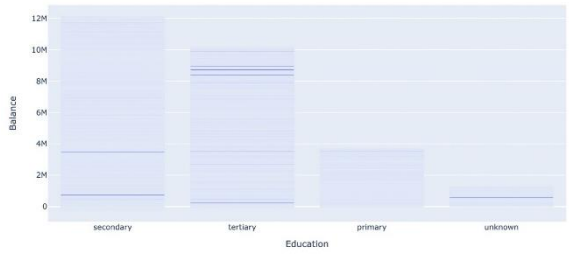

Figura 8 Balance por educación

La duración del contacto es un factor que condiciona positivamente la posibilidad de que esa persona se subscriba. A mayor duración más probabilidad. Pero esta es una variable que no sabemos hasta el momento que se realiza la llamada, por lo cual vamos a desestimar esta variable para futuras predicciones.

Así también como el canal por donde se realiza el contacto (sea por teléfono o celular).

3.3 Preparación de datos

A la hora de limpiar el ruido y tener la información lo más asertiva posible, como es de público conocimiento el Balance no debería ser negativo, por lo que estos datos menores a cero fueron eliminados.

La duración que está por encima de la media también será removida ya que funciona como un outlier y no es lo que suele suceder en la media de los casos.

Para reducir la dimensionalidad manualmente podemos ver que Pdays nos brinda información cuando la gente es nueva o no estuvo involucrada en una campaña anteriormente, misma información que nos provee la variable Previous cuando toma el valor 0.

Figura 9 Correlación de Pearson de variables continuas.

Como resultado del procesamiento de datos, obtuvimos una reducción del 40% de la información.

4 ENTRENAMIENTO DEL MODELO

Una vez trabajados y comprendidos nuestros datos, procedemos a entrenar los distintos algoritmos seleccionados para el análisis en cuestión, para posteriormente seleccionar el que mejor se adapte a nuestro caso de estudio.

Al ser un estudio de predicción donde nuestra variable objetivo es categórica se utilizó el modelo Support Vector Machine acorde a un caso de clasificación y no de regresión por la naturaleza de los datos.

4.1 Assumptions Train y Test

Gracias a la cantidad de datos provistos se seleccionó un test del 30% de nuestra muestra, para luego corroborar la veracidad del método.

4.2 SVM

Primero se realizó el análisis con un modelo de SVM, realizando validación de datos cruzados y grid search para la selección de los mejores parámetros para este caso. Como antes antelamos, se selecciono como hiper-parámetro un Kernel rbf, denotando que no sigue un modelo lineal. Con este modelo se obtuvo un desempeño alto, Accuracy del 88%. E independientemente del accuracy una curva ROC igual de precisa de 81%.

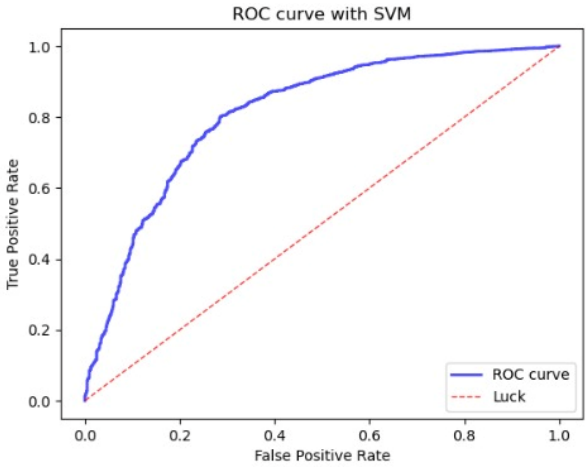

Figura 10 Curva ROC modelo SVM.

4.3 KNN

Previamente nombrado, se realizaron dos modelos distintos para comparar cual se adapta mejor. Como segunda opción se elegio un KNN el cual tiene un mejor costo computacional, con menor precisión. Ambas suposiciones fueron confirmadas obteniendo un accuracy menor de 72%, con una disminución en el tiempo del entrenamiento de los datos del 1260% (Se paso de 2hs 6 min a 6.94 seg)

En adición a la disminución de la curva ROC por igual.

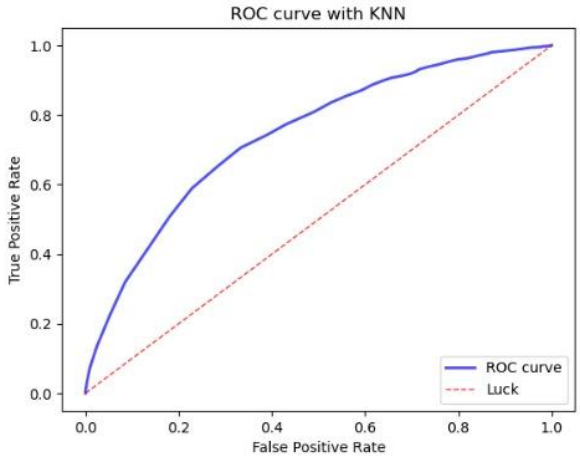

Figura 11 Curva ROC modelo KNN.

4.4 Reducción de la dimensionalidad

Nuestra muestra cuenta con 17 variables las cuales 5 de ellas son categorías, por lo que nuestra muestra a analizar luego de todas las transformaciones termino

siendo de dimensionalidad 31. Para mejorar la visualización, procedemos a realizar el análisis de reducción de la dimensionalidad.

En primera instancia vemos como es la variabilidad de nuestras features a través de un heatmap.

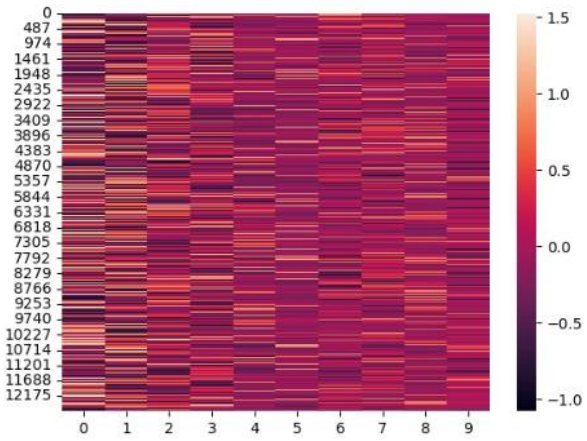

Figura 12 Heatmap de espacio latente.

Podemos observar que el 30% de nuestras variables nos están aportando información, por lo que no se va a poder reducir a una bidimensionalidad.

Con estas nuevas variables, realizadas a partir de nuestra muestra se procede a realizar nuevamente un entrenamiento para ver cuanto disminuye el desempeño. En el modelo SVM no se observa una disminución del Acurracy pero si observa una gran disminución de la curva ROC, descendiendo a un 61%. (2)

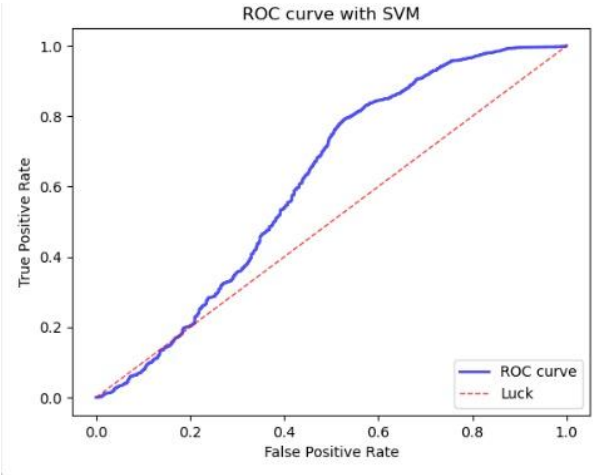

Figura 13 Curva ROC modelo SVM con PCA.

En contraposición el modelo KNN no sufrió mayores disminuciones de desempeño.

Resultando un Accuracy de 67% y una curva ROC de 70%

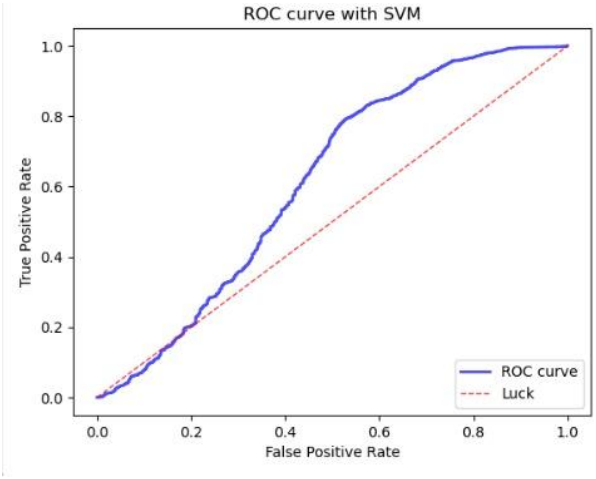

Figura 14 Curva ROC modelo KNN con PCA.

5 CONCLUSIÓN

Luego del análisis desarrollado, por la naturaleza observada en nuestras variables, consideramos desestimar el beneficio de realizar una reducción de la dimensionalidad dado que la misma trae en contraposición una gran disminución en la precisión.

Considerando este aspecto, se procede a seleccionar el modelo SVM quien nos brinda un accuracy de 88%.

6 CONTRIBUCIÓN

El equipo trabajo en conjunto realizando el análisis de los datos, tanto en la parte explotaría como procesamiento, selección del algoritmo e implementación del mismo. Como así también luego en el desarrollo y consolidación de los resultados en este informe.

7 REFERENCIAS

1. Introduction to Statistical Learning.
2. Deep learning book