Clasificador de ventas de tarjetas de crédito

Índice

Objetivo

Un banco regional nos contrató con el objetivo de crear un modelo de clasificación que busca identificar a los posibles clientes (prospectos de acá en adelante) que más probabilidad tienen de aceptar el producto de **Tarjeta de Crédito (TC).**

De esta forma esperan poder llevar adelante campañas de venta más asertivas y a su vez, en caso de necesidad (como puede ser el cumplimiento del plan comercial), hacer pre-embozados (imprimir tarjeta antes de concretar venta) y enviarlos al domicilio asegurándose una alta tasa de aceptación del plástico, es decir que la variable Target será “Venta”.

Características del Dataset

Para pronosticar el valor de nuestra variable Target “Venta”, se cuenta con información socioeconómica, de bancarización y el resultado de la gestión de los 28.844 prospectos que fueron precalificados y gestionados (a través de llamadas salientes) para la venta de una TC.

Los features que se utilizan son los siguientes:

- Edad y Sexo.
- Oferta: margen/límite crediticio ofrecido.
- Nivel socioeconómico.
- Actividad laboral: monotributista (clasificación por categoría de A a K), autónomo, empleado, informal (no registrado) o pasivo (jubilado).
- Cantidad de entidades financieras con las que tiene deuda vigente en algún producto activo como TC o préstamo.
- Monto adeudado a otras entidades entidades financieras.
- Principal entidad con la que tiene deuda vigente.

Preparación de datos

Variables monetarias

Teniendo en cuenta el contexto inflacionario de Argentina, no es conveniente trabajar con valores nominales, por lo que se modificarán las variables monetarias de los datos, el límite crediticio ofrecido y el monto endeudado del prospecto.

La transformación se realiza dividiendo dichos montos por el Salario Mínimo Vital y Móvil (SMVM) de julio del 2021.

SMVM = $ 27.216

Variables ordinales

El **nivel socioeconómico** está expresado de forma categórica en un orden de relación, se transforma a una variable discreta ordinal, asignando el valor más alto a la categoría más alta.

Variables categóricas a numéricas

Se transforma la variable **sexo** del prospecto.

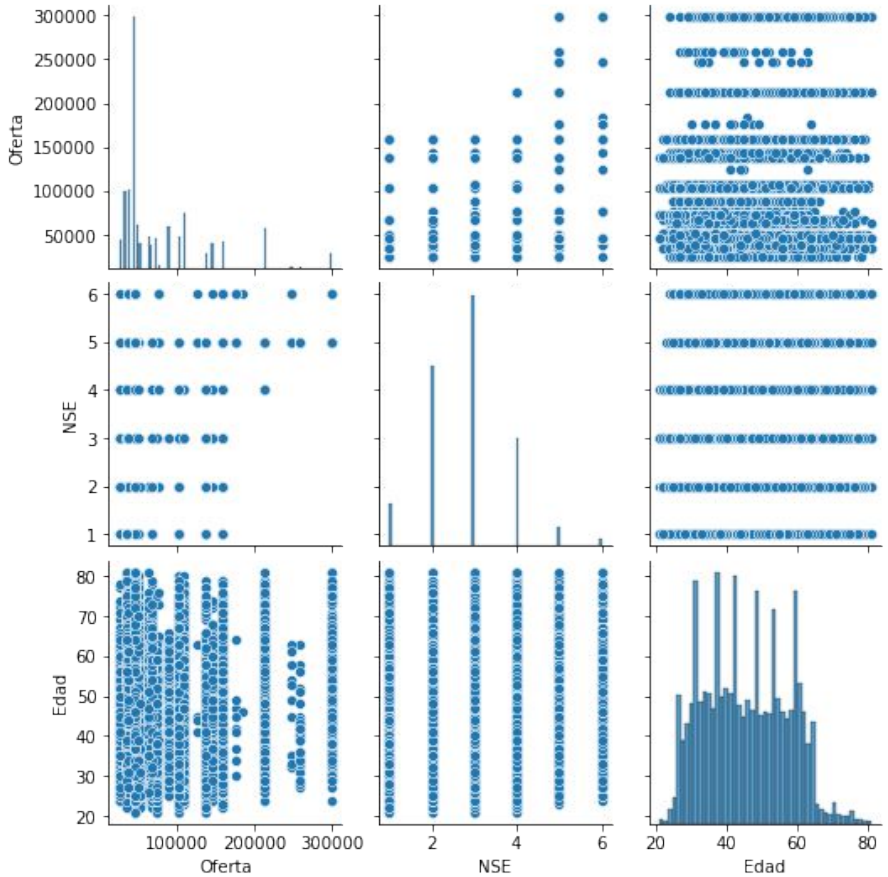
Análisis EDA

Se comprueba que los datos están levemente desbalanceados (55% - 45%).

Se verifica que ninguna variable cuente con valores nulos.

Se grafica un pair plot de las variables con las características de los prospectos contactados.

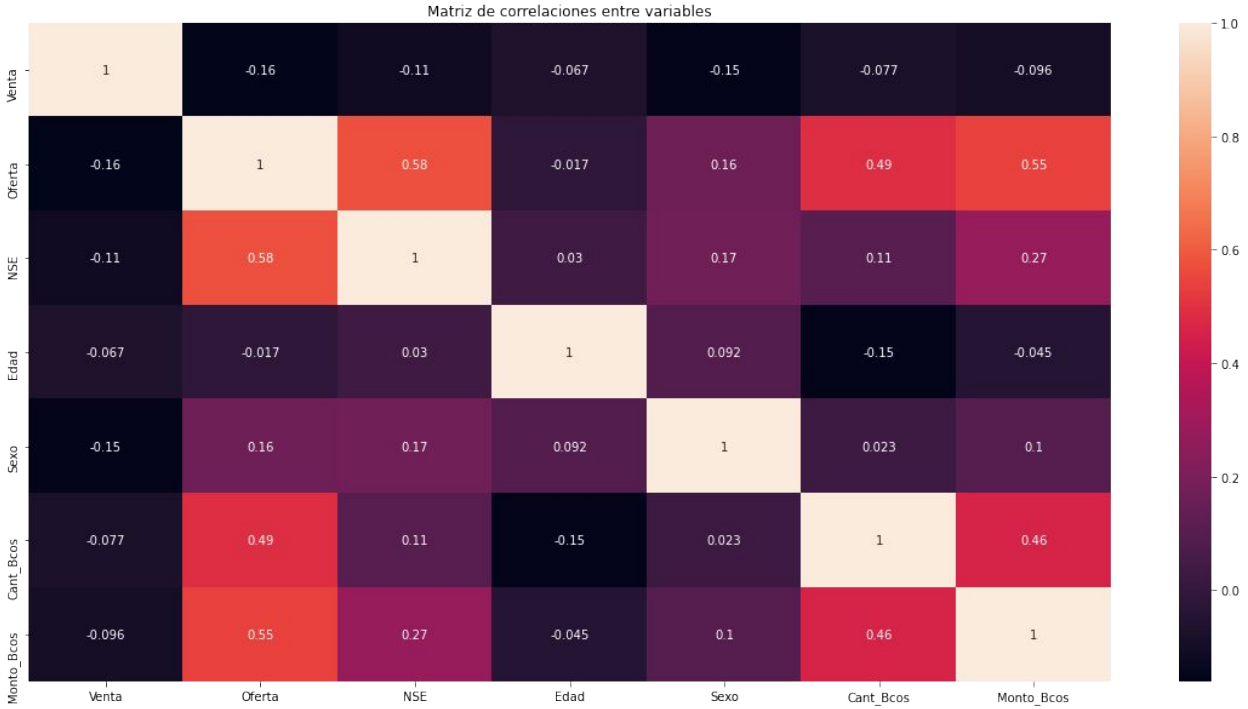
Se puede ver que hay un amplio rango de edades, niveles socioeconómicos. Los datos son heterogéneos.



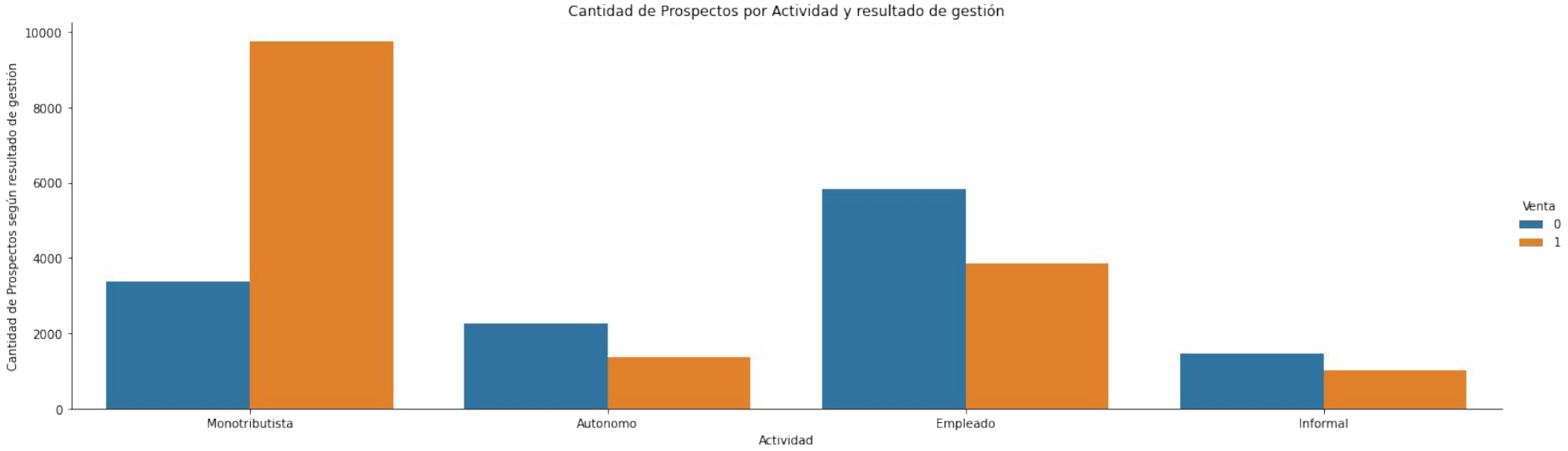
Análisis EDA

La correlación medida a través del coeficiente de correlación (R2) indica la asociación entre las variables, mientras más cercano a 1 en valor absoluto, más fuerte la relación. El signo indica el sentido de esta.

Las features NSE (Nivel Socio Económico), Oferta, y Sexo, son las que más asociación tienen con el target (Venta).



Análisis EDA

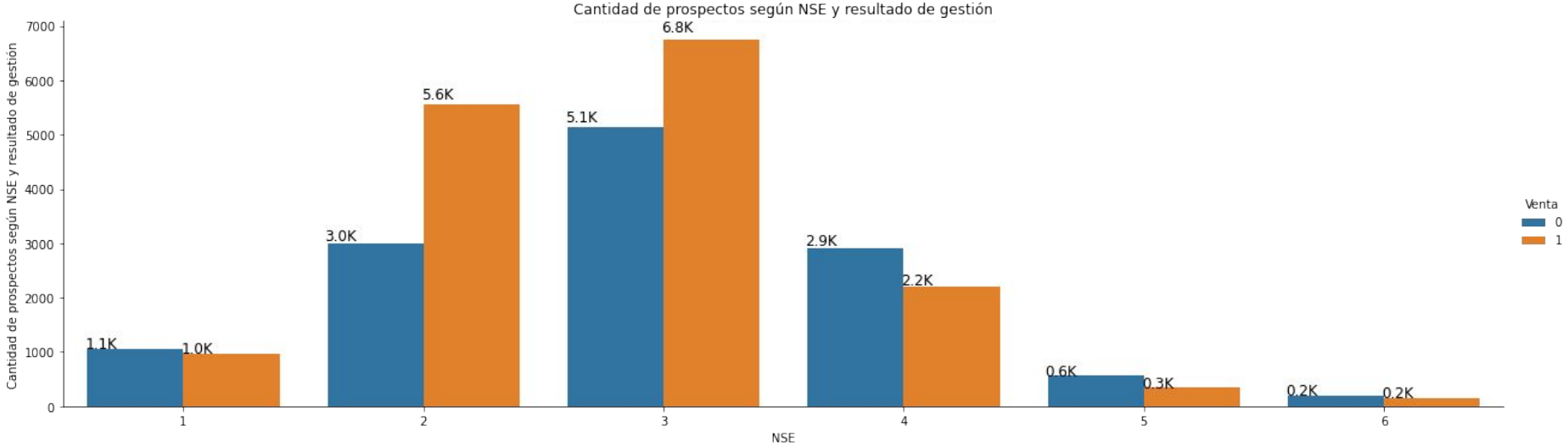


Distribución de venta por actividad laboral. Se observa que se constata la hipótesis de que lo que menos tasa de rechazo presentan, son los que están menos bancarizados, es decir, los informales

Análisis EDA

Se analiza la cantidad de ventas según el nivel socioeconómico de los prospectos.

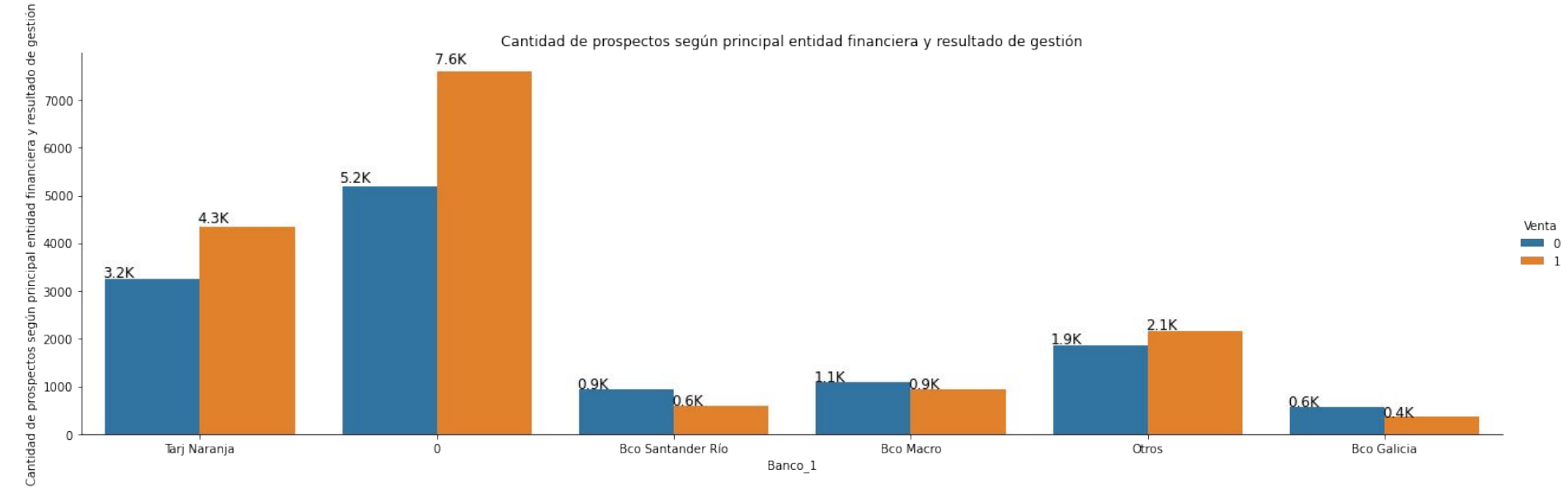
Se observa que los segmentos 2 y 3 son los más propicios a aceptar el producto. Esto se relaciona directamente con los prospectos con menor acceso a crédito financiero



Análisis EDA

Principales entidades de los prospectos que aceptaron el producto:

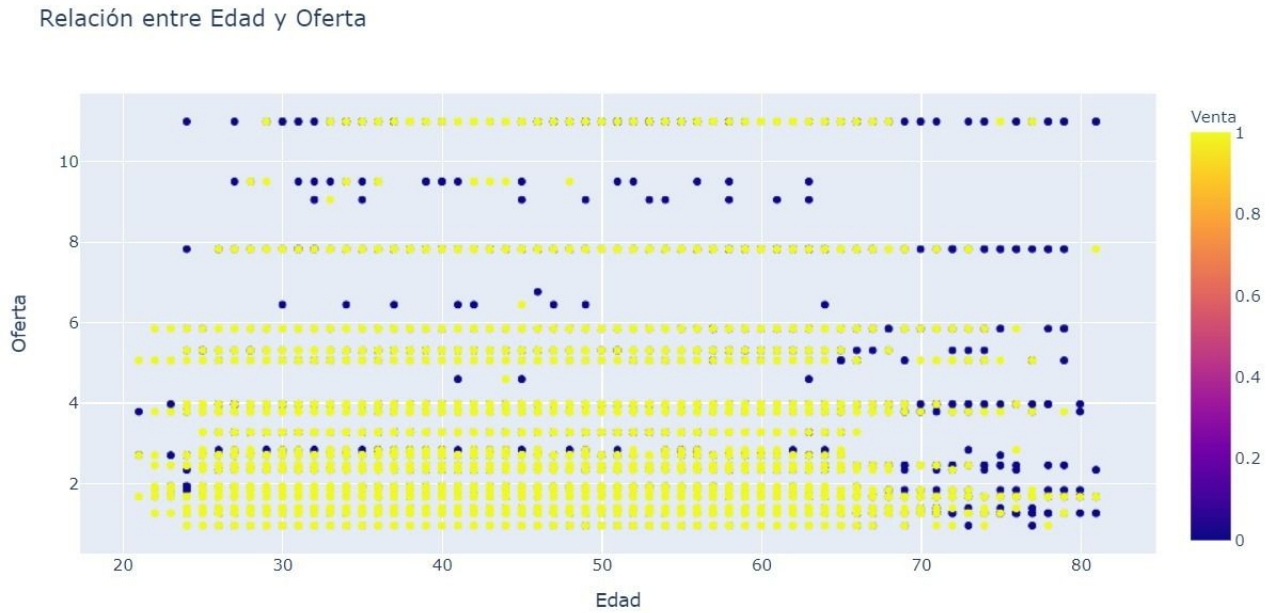
De nuevo, se observa que los individuos no bancarizados (“0”) son los que más aceptan la oferta, seguido por clientes de Tarjeta Naranja, principal competidor en el segmento al que apunta el banco.



Análisis EDA

Se relaciona la edad del prospecto y oferta (límite crediticio ofrecido) por la entidad.

Se observa que la **no** aceptación del producto se centra en personas mayores a los 70 años incluso con altos límites crediticios ofrecidos.



Evaluación de Modelos

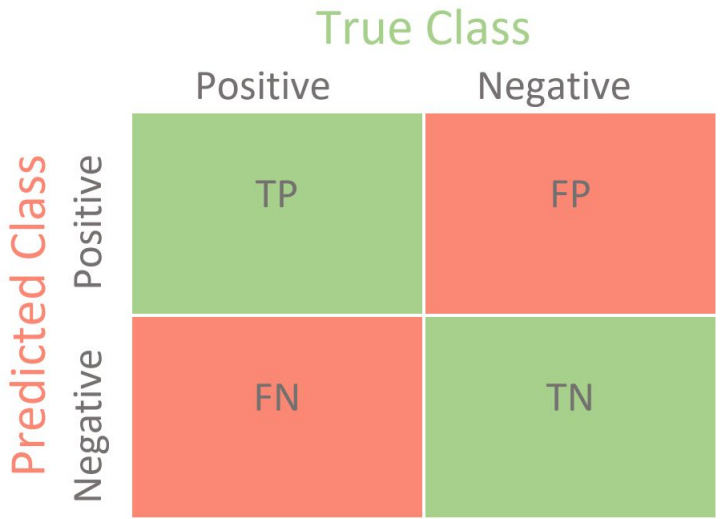
Los Output posibles de nuestro modelo son los siguientes:

1) Verdaderos positivos (TP): El modelo predice que un prospecto acepta la tarjeta y efectivamente la toma

2) Verdaderos negativos (TN): El modelo predice que un prospecto no acepta la tarjeta y efectivamente no la toma.

3) Falsos positivos (FP): El modelo predice que un prospecto acepta la tarjeta, pero en realidad no la toma.

4) Falsos negativos (FN): El modelo predice que un prospecto no acepta la tarjeta, pero en realidad la toma.



Evaluación de Modelos

Para el Objetivo del problema es de interés maximizar los TP y minimizar los FP y los FN y por ello utilizamos la métrica de **F1-Score** ya que combina las medidas de **Precision y Recall** en un solo valor.

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

Modelos utilizados y métricas de performance

Modelos	F1-Score
KNN	0,6735
Regresión Logística	0,6745
Naive Bayes sin estandarizar (Gaussiano)	0,6588
Regresión Logística Random Search CV	0,6745
KNN Random Search CV	**0,6921**

Ensamble

Los modelos de ensamble se basan en la combinación de las decisiones de varios modelos para mejorar su rendimiento general, son técnicas de aprendizaje supervisado. Los modelos de ensamble ayudan a minimizar las principales causas de error (ruido, sesgo y varianza).

Se distinguen dos familias de métodos de ensamble:

- Los métodos de averaging, que consisten en construir estimadores de forma independiente y luego hacer un promedio de sus predicciones. Como por ejemplo métodos de **Bagging** y su implementación particular **Random Forest**.

- Los métodos de Boosting, que consiste en construir estimadores de manera secuencial y uno trata de reducir el sesgo del estimador combinado, priorizando aquellos casos en los que se observa una peor performance. Como por ejemplo métodos de **ADA Boost**, **Gradient Boost**.

Ensamble - métricas de performance

Modelos	F1-Score
Árbol de Clasificación	0,6889
Bagging	0,6889
ADA Boost	0,6911
Gradient Boosting	**0,7247**

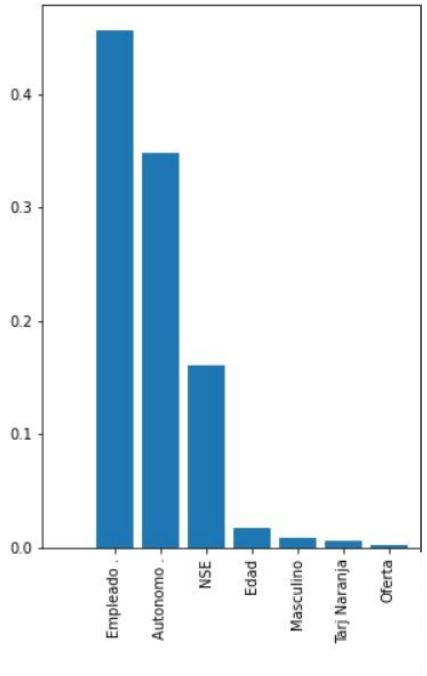
Gradient boosting es un método de aprendizaje lento donde los sucesivos modelos de árboles de decisión son entrenados para predecir los residuales del árbol antecesor permitiendo que los resultados de los modelos subsiguientes sean agregados y corrijan los errores promediando las predicciones.

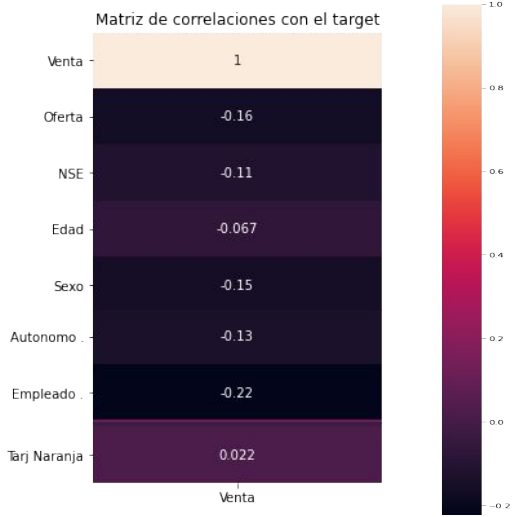
Árboles de decisión con Gradient boosting es uno de los modelos más poderosos y más utilizados para problemas de aprendizaje supervisado.

Feature importance

Es importante conocer cuáles son las features que contribuyen al modelo y cuál es su importancia en la predicción.

Conocer las features más importantes nos ayuda a reducir el conjunto de features que componen el modelo y así disminuir su costo computacional.





Conclusiones

Luego de un largo análisis de los datos y la evaluación de distintos modelos, se llega a la conclusión de que el mejor modelo de predicción es el **Gradient Boosting** con un **F1-Score** de **0,7247**.

A su vez, se determinó que **las variables más relevantes sobre la predicción son**: “Empleado”, “Autónomo”, “Nivel Socioeconómico”, “Edad”, “Masculino”, “Tarjeta Naranja”, “Oferta”.

En otras palabras, el banco tiene la posibilidad de mejorar la eficiencia del proceso de venta de tarjetas de crédito gracias a que el modelo le ayuda a conocer mejor el perfil de los futuros clientes. De esta manera es posible ahorrar costos de prospección y venta.

¡Muchas Gracias!

Grupo 8:

Bozzano, Santiago
Pérez Curiel, Joaquín
Devoto, Marcos Tomás
Florentín, Alejo
Hernandez, Maria Paz
Lopera Erazo, Jose Alejandro