****

Proyecto final Data Science



# Modelo de clasificación para identificar transacciones con riesgo de fraude

# 

[Modelo de clasificación para identificar transacciones con riesgo de fraude](#_heading=h.1fob9te) 0

[Integrantes del equipo](#_heading=h.qmfgzdxk57ks) 2

[Sobre nosotros](#_heading=h.imwjsle8uxvg) 3

[Antecedentes de la organización en la que se basa el proyecto](#_heading=h.wg1sdzstp7w4) 3

[Problema a resolver](#_heading=h.2et92p0) 4

[Objetivo](#_heading=h.1t3h5sf) 4

[Data Acquisition](#_heading=h.q0j75gkctb6z) 4

[Estadística descriptiva](#_heading=h.vtpw1acl6gib) 7

[El dataset está conformado por 27 columnas y 25,560 registros.](#_heading=h.j3n5k3qcarv5) 7

[Data Wrangling](#_heading=h.r2w755665wcm) 8

[Análisis Univariado](#_heading=h.nu3x92rf88ax) 12

[Análisis bivariado y multivariado](#_heading=h.4c9hw3gerjls) 14

[Dataset final](#_heading=h.rsf24nak76xs) 19

[Modelos Machine Learning](#_heading=h.jvfkb36dpkjr) 20

[Conclusiones y futuras líneas](#_heading=h.6bbih5thkjcd) 28

# 

# Integrantes del equipo

Joaquín Ugarte

Lucas Lis

Issis Lobato

Juan Flores

# Sobre nosotros

Somos una consultora independiente que brinda soluciones a través del análisis de datos y modelos de Machine Learning para la prevención de posibles fraudes.

# Antecedentes de la organización en la que se basa el proyecto

La empresa es una startup de medios de pago que cuenta con tarjetas tanto de crédito como de débito físicas y digitales, que se solicitan desde su aplicación. Una que se puede usar como método de pago “normal” para comprar en casi todas las tiendas; simplemente, funciona como un préstamo para que no se use tu dinero inmediatamente y, al terminar el mes, se suman todas las compras que se realizan para liquidar la tarjeta; es decir, que se pague solo lo que se compra en ese periodo.

La tarjeta tiene 2 versiones:

Tarjeta digital: vive en la app, dentro de la sección de “Mis tarjetas”, ahí se encuentran los 16 dígitos, la fecha de vencimiento y el CVV dinámico (los 3 dígitos que se encontraban atrás de las tarjetas).

Tarjeta física: te la mandan a tu dirección por paquetería. Es una tarjeta sin nombre ni datos impresos y esto la hace prácticamente inclonable. Además, cuenta con tecnología de pago sin contacto o contactless*,* con la que solo tienes que pasarla sobre la terminal para pagar.

# Problema a resolver

Es necesario que la empresa cuente con un alertamiento de transacciones que pueden ser posibles fraudes a través de un modelo de machine learning para tomar acciones preventivas y estratégicas para disminuir el número de aclaraciones que reciben relacionadas a un fraude.

# Objetivo

Implementar un modelo de clasificación que nos ayude a indicar las transacciones con altas probabilidades de fraude para la toma de acciones preventivas y/o estratégicas.

# Data Acquisition

**Dataset de transacciones reclamadas por los usuarios de la aplicación que proporciona la start up, descripción de los campos:**

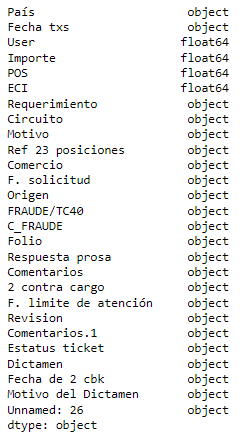
|  |  |
| --- | --- |
| **País** | País en donde está registrado el usuario |
| **Fecha Txs** | Fecha en la que se realiza la transacción |
| **User** | Número de usuario asignado dentro de la aplicación |
| **Importe** | Monto por el que se realizó la transacción |
| **POS** | Código numérico para indicar cómo fue procesada la transacción, dependiendo el valor podemos identificar si la tarjeta estuvo presente o fue una compra electrónica. |
| **ECI** | ECI (Indicador de Comercio Electrónico), código que permite informar al comercio si el usuario logró completar la autenticación y ha sido verificado por su banco, dependiendo del valor podemos identificar si el comercio es seguro, es decir, contó con una validación 3DS. |
| **Requerimiento** | Solicitud que se realiza dentro del proceso: Quebranto, Contracargo, Importe menor. |
| **Circuito** | Visa, Mastercard |
| **Motivo** | Indica la razón por la que se está generando la solicitud de aclaración: ATM, No reconoce, Rechazado, Devolución, Duplicado, Pago por otro medio, Excede importe, Mercancía/Servicio no recibido. |
| **Ref 23 posiciones** | Identificador único que se le asigna a cada transacción, consta de 23 dígitos. |
| **Comercio** | Nombre del comercio en donde se efectuó la transacción. |
| **F. solicitud** | Fecha en el que se registra un contracargo, no todos los caso tienen ya que hay casos en los que de acuerdo al monto o análisis no se genera contra cargo, y se resuelve. |
| **Origen** | Determina si la transacción fue realizada en el extranjero o en territorio nacional. |
| **Fraude TC40** | Fecha del registro del reporte a la marca derivado de un cargo no reconocido. |
| **C Fraude** | Se confirma o no la transacción como producto de un fraude |
| **Folio** | Folio que genera el portal de VISA para dar seguimiento a un contracargo registrado. |
| **Respuesta prosa** | Respuesta al contracargo registrado. |
| **Comentarios** | Alguna observación importante del contracargo registrado |
| **2 contra cargo** | Fecha cuando se declina un prearbitraje en respuesta a un contracargo. |
| **F. límite de atención** | Es la fecha máxima para solución (fecha de solicitud + 31 días) |
| **Revisión** | Fecha de cierre del caso |
| **Estatus del ticket** | Estado del ticket dentro de la plataforma de reclamos de la empresa. |
| **Dictamen** | Resultado del reclamo registrado por el usuario. |
| **Motivo del dictamen** | Detalle del tipo de fraude resultado del reclamo registrado por el usuario. |

# Estadística descriptiva

# El dataset está conformado por 27 columnas y 25,560 registros.

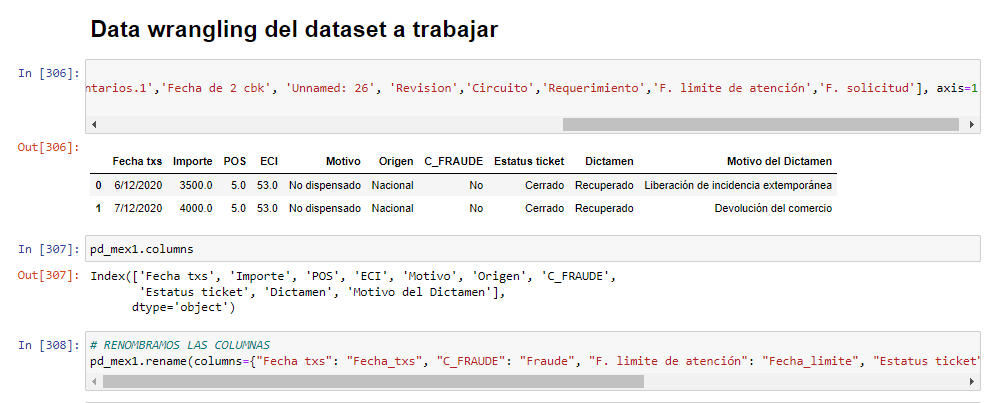
# 

Dataset



# Data Wrangling

Eliminamos columnas que no necesitamos para el analisis y renombramos algunas para identificarlas más fácilmente.



Detalle de la depuración en las columnas del dataset

Modificamos las variables categóricas a numéricas.





Cambios en las variables categóricas

Detalle de todas las variables modificadas:

https://drive.google.com/drive/folders/1DKtaxYEDoeejjOC4-0KU0hjIMoiDRXnw

Modificamos columna por columna hasta lograr tener variables numéricas que nos ayuden a crear el modelo de clasificación.

1. ['Fraude']: No = 0 \ Si = 1
2. ['POS']: INDICADOR DEL MODO DE ENTRADA (POS): TRANSACCIÓN CON TARJETA PRESENTE = 5 \ TRANSACCIÓN CON TARJETA DIGITAL = 1
3. ['Motivo']: CONTROVERSIA(NO FRAUDE) = 0 \ CARGO NO RECONOCIDO (FRAUDE) = 1
4. ['Origen']: NACIONAL = 0 \ INTERNACIONAL = 1
5. ['ECI']: INDICADOR DE COMERCIO ELECTRÓNICO (ECI): COMERCIO ELECTRÓNICO SEGURO = 5 \ COMERCIO ELECTRÓNICO NO SEGURO = 7
6. ['Dictamen']: RECUPERADO = 0 \ EN CONTRA = 1 \ NO RECUPERADO = 2 \ EN PROSA/VISA = 3 \ VISA = 4 \ NO 0 = 4 \ nan = 0

# Análisis Univariado

Analizamos el principal motivo por el que se genera una aclaración. Así identificamos que el principal motivo de reclamaciones de la empresa son cargos no reconocidos relacionados a un posible fraude.

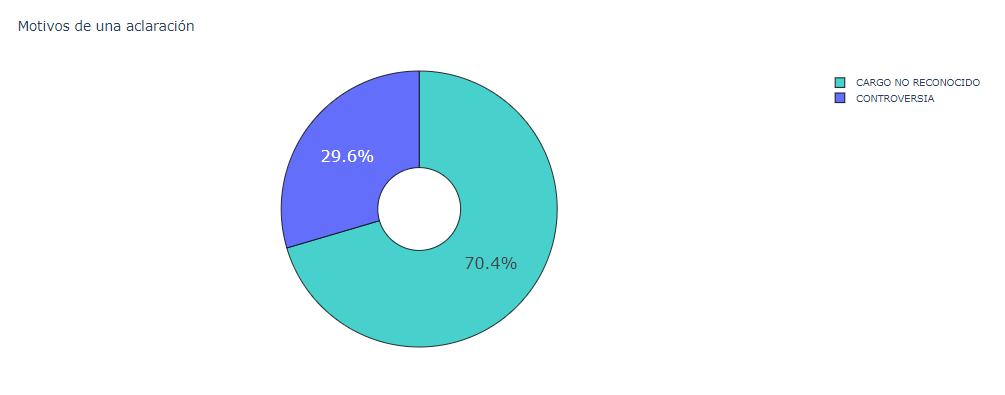


Figura 5 Gráfico para identificar el motivo de una aclaración

Realizamos un gráfico para identificar la distribución de las transacciones con base al origen de la misma. Las transacciones de origen nacional son las más frecuentes.

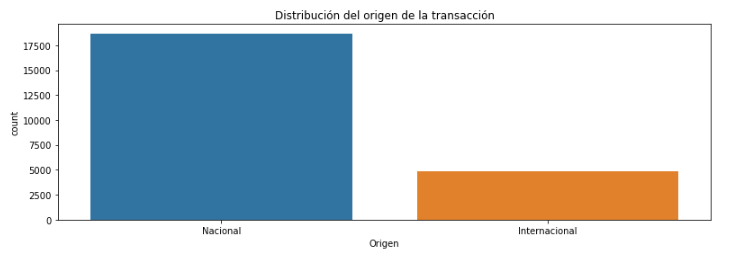


Figura 6 Gráfico del origen de las transacciones

Realizamos un gráfico para identificar el principal modo de entrada en una transacción en donde tenemos. Así comprobamos que las transacciones de las aclaraciones que recibimos son principalmente realizadas con tarjetas digitales.

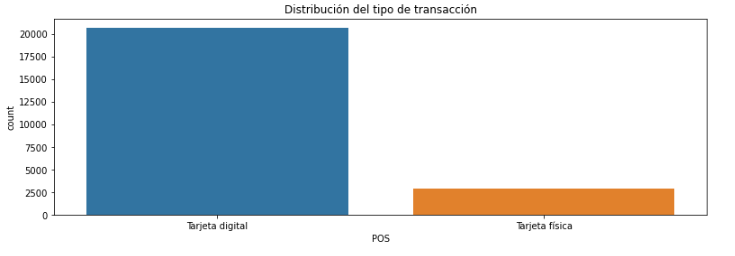


Figura 7 Gráfico del modo de entrada de las transacciones

# Análisis bivariado y multivariado

Analizamos la relación entre el origen de la transacción y el motivo del reclamo. Así identificamos que la mayoría de las transacciones relacionadas a un cargo no reconocido son nacionales.

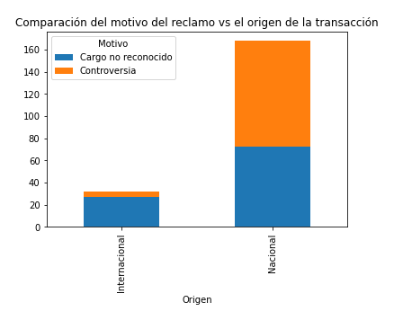


Figura 8 Gráfico para identificar el motivo del reclamo vs. el origen de la transacción

Examinamos la relación entre el origen de la transacción y el modo de entrada de la misma. Así identificamos que la mayoría de las transacciones relacionadas a un cargo no reconocido son realizadas con la tarjeta digital y son de origen internacional.

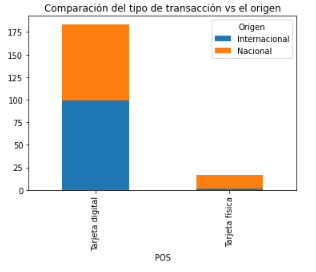


Figura 9 Gráfico para identificar el modo de entrada vs el origen de la transacción

Identificamos la relación entre el importe de la transacción y el motivo del reclamo. Los importes más elevados están relacionados a un cargo no reconocido.

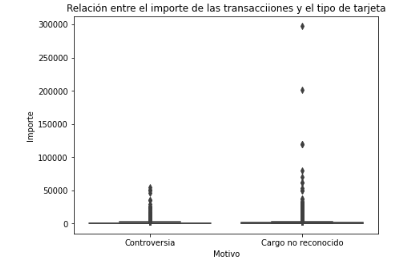


Figura 10 Gráfico para identificar la relación entre el importe y el motivo de reclamo

Analizamos la relación entre el importe y el modo de entrada de la transacción. Los importes más elevados están relacionados a las compras con tarjetas digitales.

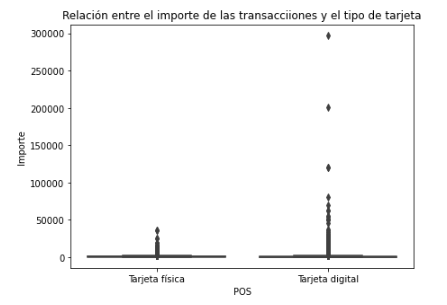


Figura 11 Gráfico para identificar la relación entre el importe y el modo de entrada

Observamos la relación entre el importe y el indicador de comercio electrónico de la transacción. Los importes más elevados están relacionados a las compras realizadas en comercios electrónicos NO seguros.

Figura 12 Gráfico para identificar la relación entre el importe y tipo de comercio electrónico

Por último revisamos la correlación entre cada una de las columnas del Dataset. Identificando las más altas entre el Motivo y POS (modo de entrada) y Fraude confirmado con ECI (indicador de comercio electrónico).

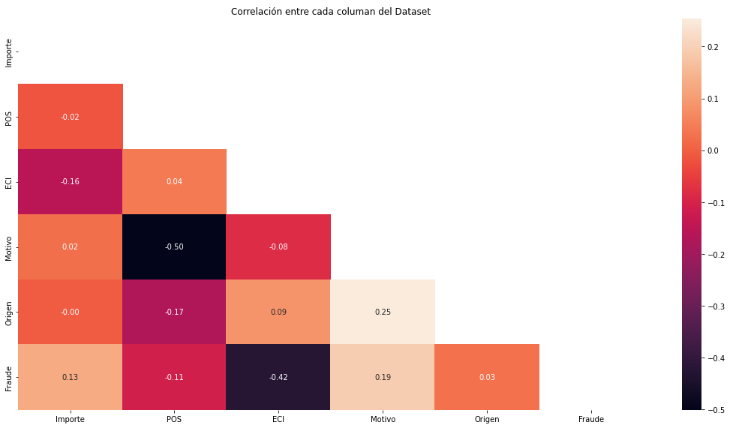
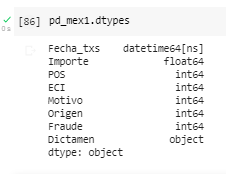
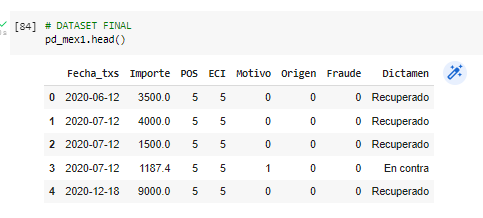


Figura 13 Correlación entre las variables que conforman el Dataset a utilizar

# 

# Dataset final

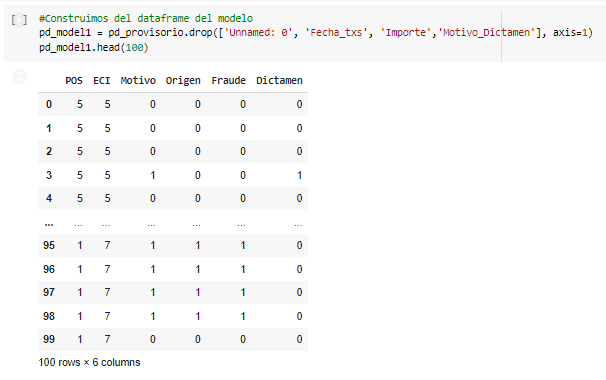




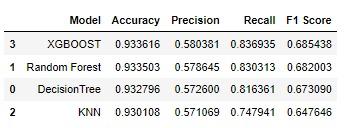


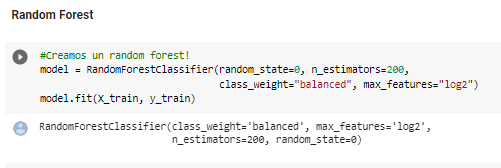
# Modelos Machine Learning

Creamos una nueva variable pd\_model1, es el DataFrame que usaremos para entrenar los Modelos de clasificación.



Investigamos 5 tipos de modelos: KNN, regresión logística, Random Forest, Tree Classifier y XGBBOOST este último es el que elegimos con base a sus resultados para aplicar en el Proyecto definido.

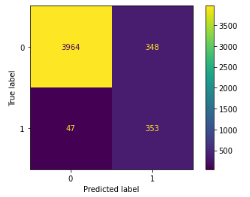




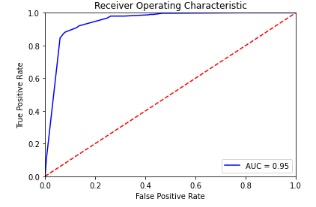


Entrenamos y balanceamos el Dataset del primer modelo.

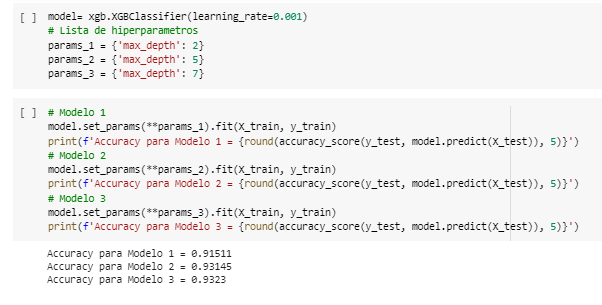
Matriz de confusión



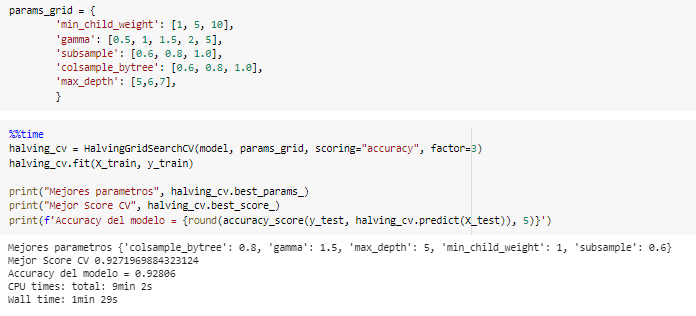
Curva ROC



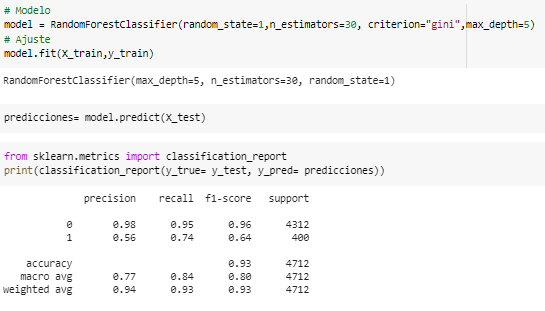
Mejoras de los parámetros 1



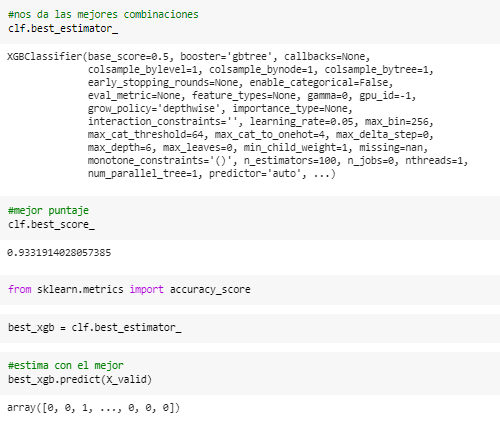
Mejoras de los parámetros 2

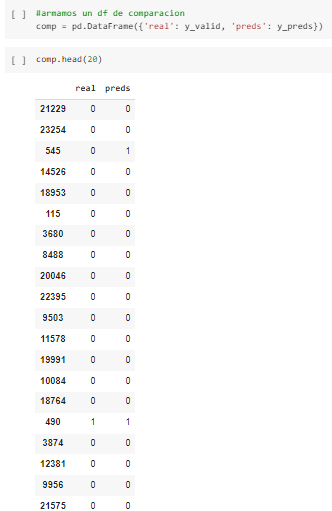


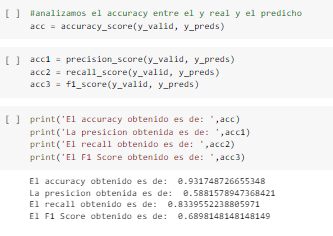
Mejoras del modelo



XGBOOST







# Conclusiones y futuras líneas

# Después de haber realizado la limpieza y análisis de datos que presenta la empresa, se recomienda llevar un control más estricto de los registros de las transacciones, es decir, contar con listas desplegables y conceptos unificados al momento de llenar los registros, esto con la finalidad de tener mejores prácticas y que la calidad de la información mejore considerablemente para cualquier tipo de análisis o proyecto a futuro.

# Una vez realizadas las pruebas con los distintos tipos de modelos utilizaremos el XGBBOOST ya que nos ayuda a mejorar la precisión del mismo dándonos una precisión del 0.98 cuando no es fraude lo que indica que tiene un alto porcentaje de acierto en estos escenarios. Al tratarse los casos de fraude mejora la precisión un 0.58 con un margen de error del 32% pero nos permite reducir el volumen de las transacciones que se tendría que revisar por parte de los analistas de la empresa para confirmar o no los casos de fraude con análisis específicos o contactando a los usuarios para hacer la confirmación de dicha transacción. Reduciendo así los casos que pasarían a una revisión “manual” para ser analizados a detalle reduciendo así la carga de trabajo de los analistas de monitoreo de fraude y probablemente optimizar los recursos con los que se cuentan actualmente y no realizar la revisión de todas las transacciones generadas y reportadas por los usuarios como se hace actualmente. Con la implementación del modelo podemos eficientar la revisión de transacciones alertadas y llegar a reducir analistas en la plantilla reduciendo así costos importantes para la empresa.