UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

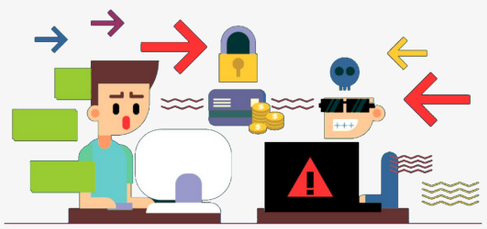
FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Modelo para detectar fraudes en tarjetas de créditos

Marco Oviedo/ Luis Navarro/Eliud Moreno / Nalleli Allende **Equipo 09 Grupo 002**

INTRODUCCIÓN

En resumidas cuentas, un fraude con tarjeta de crédito es un hecho mediante el cual unos delincuentes realizan operaciones como consumos en establecimientos físicos o virtuales, adelantos de efectivo, entre otras, para lo cual se aprovechan de la línea de crédito de los consumidores afectados. Un fraude con tarjeta de crédito puede ocurrir mediante el hurto o el robo de la tarjeta física y el PIN o clave secreta, o del único conocimiento de los datos confidenciales de esta, es decir su numeración completa, la fecha de vencimiento y el código CVV. En dicha medida, estos delincuentes pueden aprovechar y acumular diversos cargos con cargo a la línea de la tarjeta de crédito.



El comportamiento malicioso o fraude sigue patrones específicos y, por lo tanto, se puede predecir con base en ellos. A través del aprendizaje supervisado, se puede clasificar como fraudulenta o legítima a través de datos con una etiqueta clara sobre una pregunta base. Si la empresa financiera tiene acceso a todas las transacciones que se realizan con sus tarjetas, puede crear grandes conjuntos de datos y marcar los fraudes como tal.

OBJETIVO

Reducir el impacto de las transacciones fraudulentas encontrando el mejor modelo que me ayude a detectarlos

OBJETIVO SECUNDARIO

Aumentar la precisión de la detección de transacciones fraudulentas

RECURSOS

Software libre usado para el clasificador.
Python Lenguaje de programación interpretado

METODOLOGÍA

Nuestra base de datos cuenta con **284806 transacciones** de las cuales **0.17% son transacciones fraudulentas** y solo se nos proporciona variables numéricas es de aquí donde tenemos que encontrar que variables son las que nos ayuda a proporcionar una predicción de las transacción fraudulentas.
Como nuestros datos no están bien proporcionados tenemos que realizar una submuestra para evitar sesgos.
Usando el **método "Random Under Sampling"** que consiste en eliminar datos para tener un conjunto de datos más equilibrado y así evitar sobreajuste en los modelos.
Nuestra base de datos cuenta con 284806 transacciones de las cuales .17% son transacciones fraudulentas y solo se nos proporciona variables numéricas es de aquí donde tenemos que encontrar que variables son las que nos ayuda a proporcionar una predicción de las transacción fraudulentas.
Como nuestros datos no están bien proporcionados tenemos que realizar una submuestra para evitar sesgos.

Usando el **método "Random Under Sampling"** que consiste en eliminar datos para tener un conjunto de datos más equilibrado y así evitar sobreajuste en los modelos.
Antes de aplicar las técnicas de minería tenemos que usar **el modelo t-SNE** esto nos da una indicación de que los modelos predictivos adicionales funcionarán bastante bien para separar los casos de fraude de los casos que no lo son.
Posteriormente, capacitaremos a **cuatro tipos de clasificadores** y decidiremos qué clasificador será más efectivo para detectar transacciones fraudulentas. Antes, tenemos que dividir nuestros datos en conjuntos de prueba y entrenamiento y separar las características de las etiquetas.

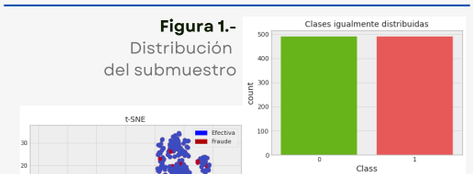


Figura 1.- Distribución del submuestro

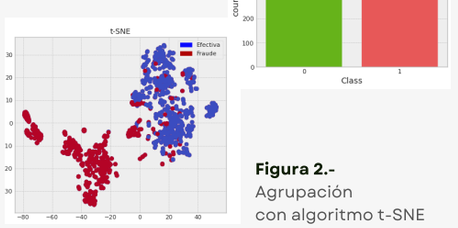


Figura 2.- Agrupación con algoritmo t-SNE

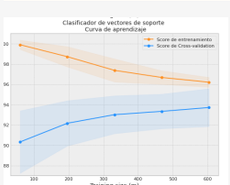


Figura 3.1.- Curva de aprendizaje SVC

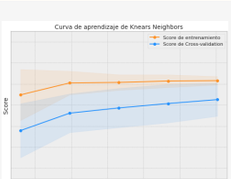


Figura 3.2.- Curva de aprendizaje KNN

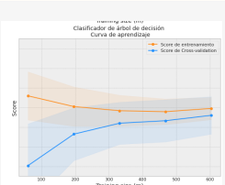


Figura 3.3.- Curva de aprendizaje DT

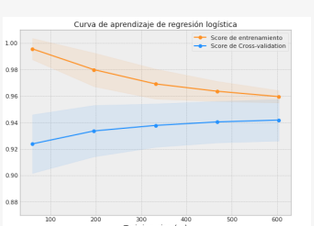


Figura 3.4.- Curva de aprendizaje Regresión Logistica

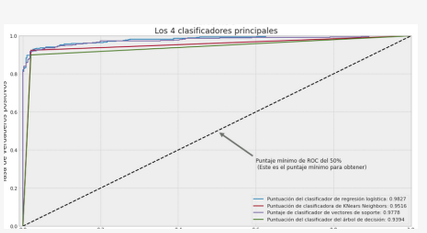


Figura 4.- Curva ROC

RESULTADOS

El clasificador de regresión logística muestra la mejor puntuación tanto en el entrenamiento como en los conjuntos de validación cruzada.

	Entrenamiento	Validación cruzada
LogisticRegression	95%	95.23%
knearsneighbors	94%	95.23%
SVC	94%	94.57%
DecisionTreeClassifier	91%	94.04%

El clasificador de **regresión logística es más preciso que los otros tres** clasificadores en la mayoría de los casos.
GridSearchCV se utiliza para determinar los parámetros que dan la mejor puntuación predictiva para los clasificadores.
La regresión logística tiene la mejor puntuación de característica operativa de recepción (ROC), lo que significa que la regresión logística **separa con bastante precisión las transacciones fraudulentas y no fraudulentas.**

TRABAJO A FUTURO

Mejorar o sustituir el método de submuestreo aleatorio pues cuenta con deficiencias a la hora de precisar un modelo

CONCLUSIÓN

La construcción de un sistema de detección de fraude de tarjetas de crédito preciso y eficiente es una de las tareas clave para las instituciones financieras
En este estudio, se utilizaron 4 métodos de clasificación para construir modelos de detección de fraude. Demostramos que **el mejor modelo para detectar fraudes con mayor precisión es Regresión Logística**

REFERENCIAS

Notebook

Dataset