PROYECTO FINAL ESTRATEGIAS EMPRESARIALES BASADAS EN CIENCIA DE DATOS

DETECCIÓN DE FRAUDE

ESTEBAN MAYEN SOTO
BRIAN ANTONIO ARANDA MEJÍA
CARLOS CABRERA CASTREJÓN

Introducción

- El mundo está experimentando un rápido crecimiento en el campo de las transacciones digitales. El uso extendido de tarjetas de crédito y sistemas de pago en línea ha simplificado la vida de millones, pero también ha abierto nuevas vías para el fraude financiero. Estos actos delictivos no solo resultan en pérdidas financieras directas sino que también erosionan la confianza en las instituciones financieras.
- **Importancia del Problema**
- **Para las Empresas**: El fraude de tarjetas de crédito puede llevar a pérdidas financieras significativas y daños a la reputación que pueden ser difíciles de recuperar.
- **Para los Clientes**: La seguridad financiera y la confianza en los sistemas de pago digitales están en juego. Un solo incidente puede llevar a la pérdida de confianza y posiblemente a la pérdida de clientes.

Objetivos

01

Detección y
Prevención de Fraude

Objetivos

Objetivos

03

Mejora de la Seguridad de las Transacciones

Objetivos

OBJETIVO

S

Crear un algoritmo de detección de fraudes que identifique al menos el 90% de las transacciones fraudulentas en un conjunto de 1 millón de transacciones simuladas, manteniendo la tasa de falsos positivos por debajo del 1%

M

Mediremos el éxito del objetivo mediante la tasa de detección de fraudes y la tasa de falsos positivos obtenidos después de implementar el algoritmo

A

Lograr un modelo de regresión logistica que sea capaz de recibir datos y poder determinar de que tipo de transacción se trata

R

La detección precisa de fraudes protege los activos financieros y la confianza de los clientes, reduciendo pérdidas y mejorando la reputación del negocio

T

El objetivo debe cumplirse en tres meses a partir del inicio del proyecto, permitiendo evaluación y ajustes antes de la fecha límite: 21 de Octubre de 2023.

Timeline

A lo largo de 7 semanas hemos preparado el proyecto para poder presentar el resultado final.

Documentación y Presentación

Desarrollo y Entrenamiento del Algoritmo

Exploración y Preparación de Datos

Definición y Planificación

METODOLOGÍA

1.Entendimiento del Negocio
- Identificación del problema y del impacto del fraude en las transacciones de tarjetas.
- Definición de los objetivos y métricas clave.

2.Entendimiento de los Datos
- Exploración inicial de los datos para identificar características y etiquetas.
- Estadísticas descriptivas.

3.Preparación de los Datos
- Limpieza de datos: manejo de valores nulos, duplicados, etc.
- Tuning de datos.

4.Modelado

- Selección de algoritmos: Regresión logística.
- Entrenamiento y ajuste de parámetros.

5.Evaluación del Modelo
- Validación cruzada, matrices de confusión, métricas de clasificación (Precisión, Recall, F1-Score).

6.Despliegue
- Integración del modelo en un sistema de detección de fraudes en tiempo real.

Beneficios

Protección del Cliente

Detección temprana de actividades fraudulentas protege a los clientes.

Eficiencia Operativa

Automatización mejora la eficiencia en la identificación de actividades sospechosas.

Cumplimiento Legal

Cumplir regulaciones evita sanciones legales y protege la integridad del negocio.


ENTENDIMIENTO
DE LOS DATOS

ANÁLISIS EXPLORATORIO DE DATOS

Recopilación de datos iniciales

Obtención de conjuntos de datos crudos, la recopilación de información sobre las variables, y la preparación de los datos para su posterior análisis

Descripción de los datos

Esto implica examinar las características clave de los datos, como estadísticas descriptivas, distribuciones, tendencias y patrones

Verificación de la calidad de los datos

Implica la identificación y manejo de problemas como valores atípicos, datos faltantes, errores, duplicados o inconsistencias en los datos

Exploración de los datos

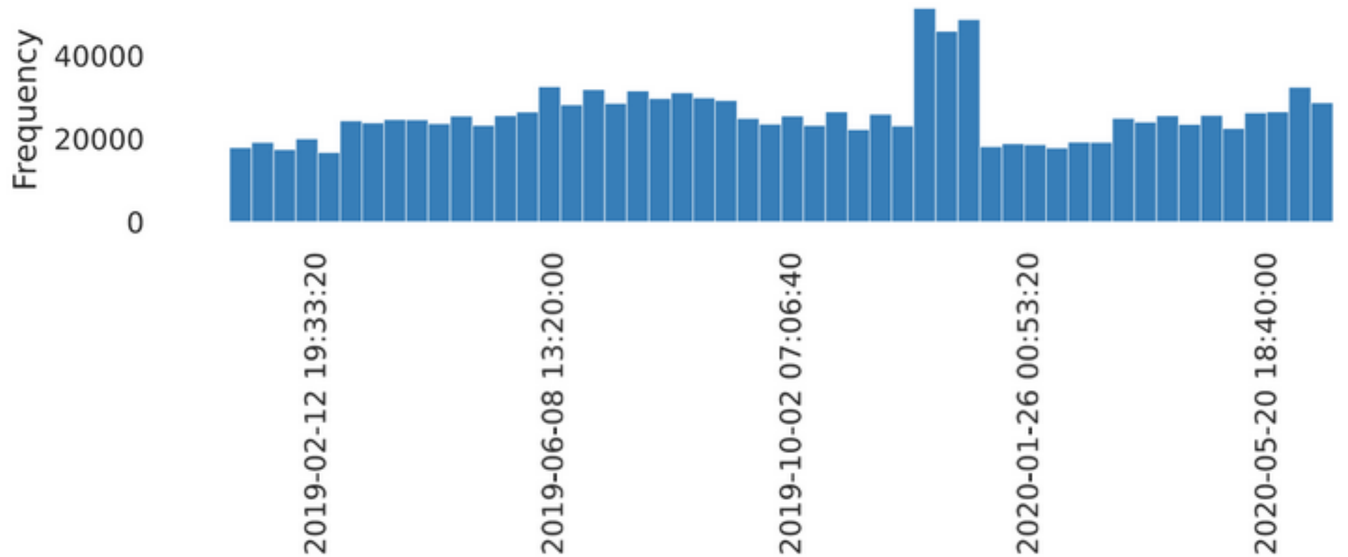
El objetivo es descubrir patrones, tendencias, relaciones y características clave en los datos.

trans_date_trans_time

Date

Distinct	1274791	**Minimum**	2019-01-01 00:00:18
Distinct (%)	98.3%	**Maximum**	2020-06-21 12:13:37
Missing	0		
Missing (%)	0.0%		
Memory size	9.9 MiB		

Histogram



Histogram with fixed size bins (bins=50)

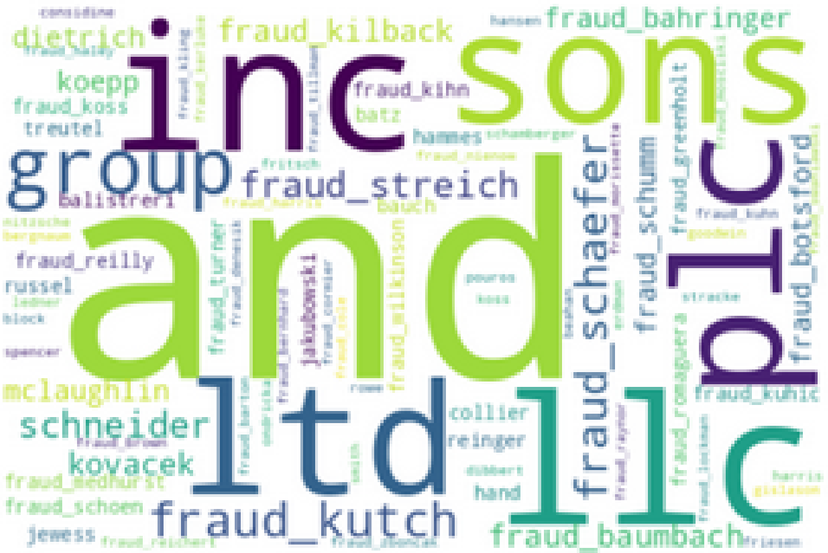
merchant

Text

Distinct	693
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
and	474111	15.7%
llc	97780	3.2%
inc	91939	3.0%
sons	73145	2.4%
ltd	70853	2.3%
plc	66475	2.2%
group	50447	1.7%
fraud_kutch	10560	0.3%
fraud_schaefer	9394	0.3%
fraud_streich	9250	0.3%
Other values (804)	2069403	68.4%



category

Categorical

Distinct	14
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Categories | Words | Characters

Common Values

Value	Count	Frequency (%)
gas_transport	131659	10.2%
grocery_pos	123638	9.5%
home	123115	9.5%
shopping_pos	116672	9.0%
kids_pets	113035	8.7%
shopping_net	97543	7.5%
entertainment	94014	7.3%
food_dining	91461	7.1%
personal_care	90758	7.0%
health_fitness	85879	6.6%
Other values (4)	228901	17.7%

amt

Real number (ℝ)

SKEWED

Distinct	52928	**Minimum**	1
Distinct (%)	4.1%	**Maximum**	28948.9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	70.351035	**Memory size**	9.9 MiB

first

Text

Distinct	352
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
christopher	26669	2.1%
robert	21667	1.7%
jessica	20581	1.6%
james	20039	1.5%
michael	20009	1.5%
david	19965	1.5%
jennifer	16940	1.3%
william	16371	1.3%
mary	16346	1.3%
john	16325	1.3%
Other values (342)	1101763	85.0%

last

Text

Distinct	481
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
smith	28794	2.2%
williams	23605	1.8%
davis	21910	1.7%
johnson	20034	1.5%
rodriguez	17394	1.3%
martinez	14805	1.1%
jones	13976	1.1%
lewis	12753	1.0%
gonzalez	11799	0.9%
miller	11698	0.9%
Other values (471)	1119907	86.4%

gender

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB


54.7%
(709863)
45.3%
(586812)
F
M

street

Text

Distinct	983
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
apt	327791	6.4%
suite	305467	5.9%
island	22954	0.4%
michael	18967	0.4%
common	17978	0.3%
station	17957	0.3%
islands	17917	0.3%
david	17476	0.3%
brooks	16991	0.3%
fields	16321	0.3%
Other values (1940)	4376722	84.9%

city

Text

Distinct	894
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
city	21314	1.3%
west	19473	1.2%
north	14425	0.9%
saint	14363	0.9%
falls	12794	0.8%
new	11842	0.7%
mount	11375	0.7%
lake	11249	0.7%
san	10260	0.6%
springs	8727	0.5%
Other values (918)	1482445	91.6%

state

Text

Distinct	51
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
tx	94876	7.3%
ny	83501	6.4%
pa	79847	6.2%
ca	56360	4.3%
oh	46480	3.6%
mi	46154	3.6%
il	43252	3.3%
fl	42671	3.3%
al	40989	3.2%
mo	38403	3.0%
Other values (41)	724142	55.8%

zip

Real number (ℝ)

HIGH CORRELATION

Distinct	970	**Minimum**	1257
Distinct (%)	0.1%	**Maximum**	99783
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	48800.671	**Memory size**	9.9 MiB

lat

Real number (ℝ)

HIGH CORRELATION

Distinct	968	**Minimum**	20.0271
Distinct (%)	0.1%	**Maximum**	66.6933
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	38.537622	**Memory size**	9.9 MiB

long

Real number (ℝ)

HIGH CORRELATION

Distinct	969	**Minimum**	-165.6723
Distinct (%)	0.1%	**Maximum**	-67.9503
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	1296675
Infinite (%)	0.0%	**Negative (%)**	100.0%
Mean	-90.226335	**Memory size**	9.9 MiB

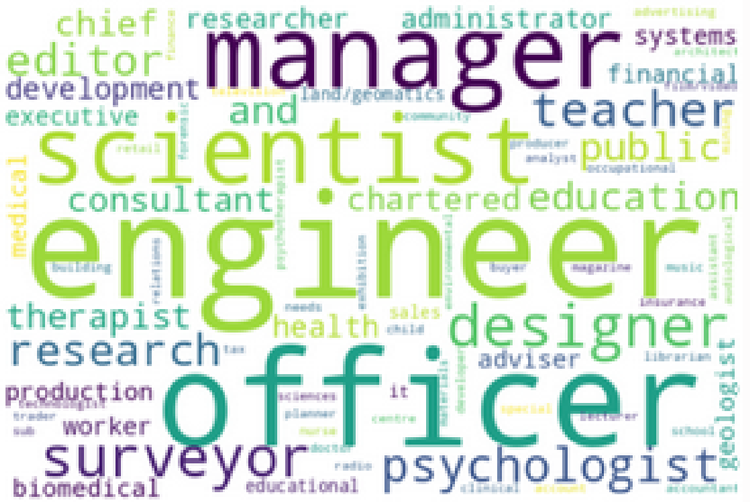
job

Text

Distinct	494
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Overview | Words | Characters

Value	Count	Frequency (%)
engineer	131756	4.6%
officer	110915	3.9%
manager	61124	2.1%
scientist	55878	1.9%
designer	52218	1.8%
surveyor	49062	1.7%
teacher	38126	1.3%
psychologist	32600	1.1%
research	29754	1.0%
editor	28725	1.0%
Other values (456)	2289024	79.5%

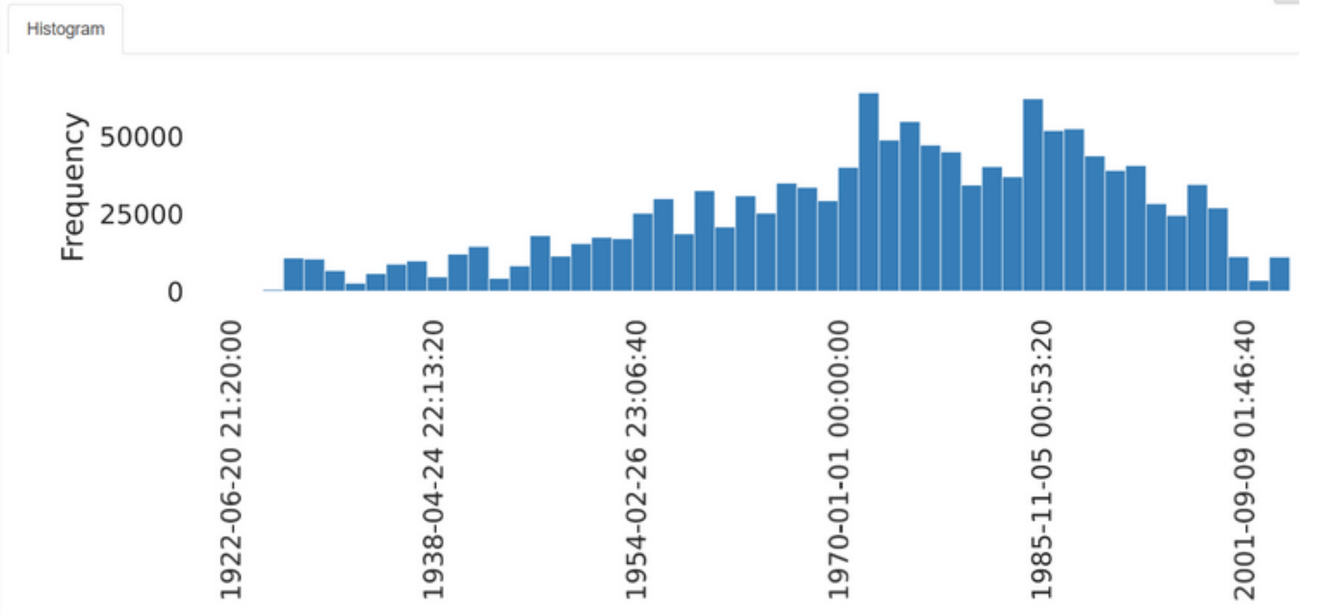


dob

Date

Distinct	968
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

Minimum	1924-10-30 00:00:00
Maximum	2005-01-29 00:00:00



trans_num

Text

UNIQUE

Distinct	1296675
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

merch_lat

Real number (ℝ)

HIGH CORRELATION

Distinct	1247805	**Minimum**	19.027785
Distinct (%)	96.2%	**Maximum**	67.510267
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	38.537338	**Memory size**	9.9 MiB

merch_long

Real number (ℝ)

HIGH CORRELATION

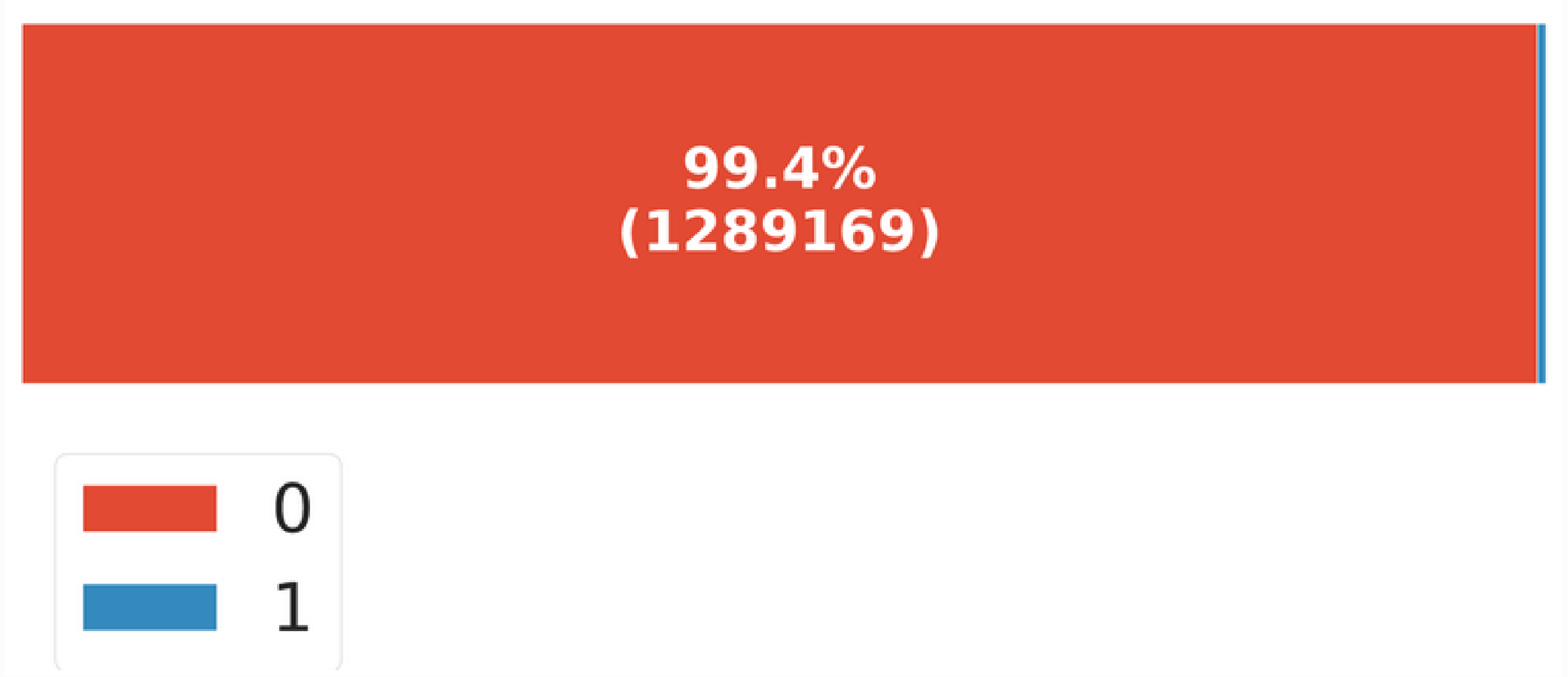
Distinct	1275745	**Minimum**	-166.67124
Distinct (%)	98.4%	**Maximum**	-66.950902
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	1296675
Infinite (%)	0.0%	**Negative (%)**	100.0%
Mean	-90.226465	**Memory size**	9.9 MiB

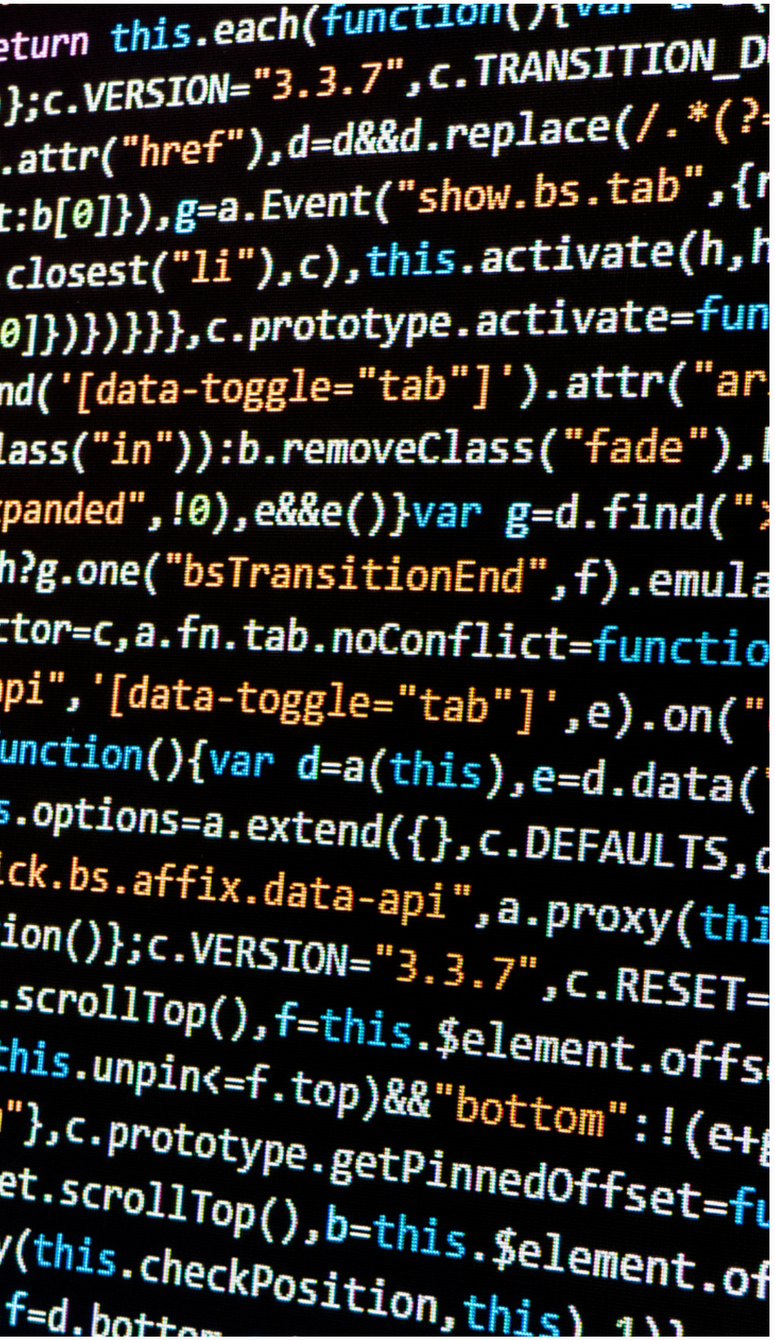
is_fraud

Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB


99.4%
(1289169)
0
1



Limpieza de Datos

¿Esta relacionado cada atributo al objetivo del negocio y de la minería?

Atributo	
trans_date_trans_time	ok
cc_num	x
merchant	x
category	ok
first	x
last	x
gender	x
street	x
city	x
state	ok
zip	x
lat	ok
long	ok
city_pop	ok
job	x
dob	x
trans_num	x
unix_time	x
merch_lat	x
merch_long	x
is_fraud	ok

¿Esta relacionado cada atributo al objetivo del negocio y de la minería?

Objetivo del negocio

Lograr un decremento de las transacciones fraudulentas

Objetivo de la minería

Predecir transacciones fraudulentas

```python
#Distribucion 40-60
# Counting the instances of each class in the 'is_fraud' column
class_counts = df['is_fraud'].value_counts()

# Finding the minority and majority class
majority_class = class_counts.idxmax()
minority_class = class_counts.idxmin()

# Separating the majority and minority class instances
df_majority = df[df['is_fraud'] == majority_class]
df_minority = df[df['is_fraud'] == minority_class]

# Over-sampling the minority class to match the majority class count (40% created data)
df_minority_oversampled = df_minority.sample(int(class_counts[majority_class] * 0.4), replace=True, random_state=42)

# Combining the original majority class with the oversampled minority class (60% original data)
df = pd.concat([df_majority, df_minority_oversampled], axis=0)

# Shuffling the dataset
df = df.sample(frac=1, random_state=42).reset_index(drop=True)

# Displaying some information about the modified dataset
df['is_fraud'].value_counts(), df.head()
```

	Unnamed: 0	trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	street	...	lat	long	city_pop	job	dob
0	0	2019-01-01 00:00:18	2703186189652095	fraud_Rippin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F	561 Perry Cove	...	36.0788	-81.1781	3495	Psychologist, counselling	1988-03-09
1	1	2019-01-01 00:00:44	630423337322	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F	43039 Riley Greens Suite 393	...	48.8878	-118.2105	149	Special educational needs teacher	1978-06-21
2	2	2019-01-01 00:00:51	38859492057661	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M	594 White Dale Suite 530	...	42.1808	-112.2620	4154	Nature conservation officer	1962-01-19
3	3	2019-01-01 00:01:16	3534093764340240	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M	9443 Cynthia Court Apt. 038	...	46.2306	-112.1138	1939	Patent attorney	1967-01-12
4	4	2019-01-01 00:03:06	375534208663984	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	...	38.4207	-79.4629	99	Dance movement psychotherapist	1986-03-28

```
df['trans_date_trans_time'] = df['trans_date_trans_time'].astype('datetime64[ns]')
df = df.assign(trans_date = df['trans_date_trans_time'].dt.strftime("%d/%m/20%y"))
df = df.assign(trans_time = df['trans_date_trans_time'].dt.strftime("%H:%M:%S"))
```

```
# Convertir la columna 'trans_time' a tipo datetime.time y luego a un string para manipularla más fácilmente
df['trans_time'] = pd.to_datetime(df['trans_time'], format='%H:%M:%S').dt.time.astype(str)

# Redondear los minutos a la hora más cercana (eliminando los minutos y segundos)
df['trans_time_rounded'] = df['trans_time'].str[:2] + ':00'
```

```
# Convertir la columna 'trans_date' a tipo datetime
df['trans_date'] = pd.to_datetime(df['trans_date'], format='%d/%m/%Y')

# Descomponer la columna 'trans_date' en componentes temporales
df['year'] = df['trans_date'].dt.year
df['month'] = df['trans_date'].dt.month
df['day'] = df['trans_date'].dt.day
df['weekday'] = df['trans_date'].dt.weekday
```

```
df = df.drop(['cc_num'],axis=1)
df = df.drop(['first'],axis=1)
df = df.drop(['last'],axis=1)
df = df.drop(['gender'],axis=1)
df = df.drop(['city'],axis=1)
df = df.drop(['zip'],axis=1)
df = df.drop(['street'],axis=1)
df = df.drop(['job'],axis=1)
df = df.drop(['dob'],axis=1)
df = df.drop(['trans_num'],axis=1)
df = df.drop(['unix_time'],axis=1)
df = df.drop(['merchant'],axis=1)
df = df.drop(['merch_lat'],axis=1)
df = df.drop(['merch_long'],axis=1)
df = df.drop(['trans_date_trans_time'],axis=1)
```

```
df = df.drop(['year'],axis=1)
```

Resultado Final de la Limpieza

	category	amt	state	lat	long	city_pop	is_fraud	trans_time_rounded	month	day	weekday
0	entertainment	170.06	FL	28.5697	-80.8191	54767	0	16:00	12	5	3
1	gas_transport	76.92	MI	44.2529	-85.0170	1126	0	09:00	8	19	0
2	grocery_pos	268.73	WI	42.8035	-88.4092	9679	1	01:00	7	3	2
3	grocery_pos	345.91	AL	32.2844	-86.9920	800	1	03:00	4	12	4
4	grocery_pos	290.95	KY	37.1046	-83.5706	467	1	00:00	1	13	0

Modelo e implementación

Regresión Logística

1. Manejo de Características Categóricas y Numéricas:
    - La regresión logística puede manejar tanto variables categóricas (como el estado y la categoría de la transacción) como numéricas (como la cantidad y la población de la ciudad), que son prominentes en nuestro conjunto de datos.
2. Buen Rendimiento con Datos Desbalanceados:
    - La regresión logística puede ser eficaz incluso cuando hay un desequilibrio en las clases, como suele ser el caso en la detección de fraudes.

RESULTADOS