

USO DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE TRANSACCIONES FRAUDULENTAS CON TARJETAS DE CRÉDITO

Presentado por: Sara Estefanía Chamseddine Fajardo



CONTENIDO

- Resumen
- Introducción
- El problema
- Objetivo general
- Enfoque
- Herramientas
- Datasets
- Transacciones fraudulentas vs no fraudulentas
- Características de los modelos aplicados
- Resultados
- Conclusiones
- Áreas de mejora



RESUMEN

La tesis se centra en la detección de fraudes en transacciones con tarjetas de crédito en entornos distribuidos, utilizando diversos modelos de aprendizaje automático. Se abordan múltiples etapas, desde la extracción y procesamiento de datos hasta la implementación y entrenamiento de modelos en la plataforma Azure Databricks con Python. La prioridad es garantizar la detección de transacciones fraudulentas, dada la desigualdad en los datos, enfocándose en la métrica de sensibilidad para minimizar los falsos negativos. Se evalúan tres modelos (SVM, Regresión Logística e Isolation Forest) y se determina que SVM es la mejor opción debido a su alta sensibilidad, aunque con una precisión más baja, lo que puede resultar en más falsos positivos.



INTRODUCCIÓN

En la era actual, con el auge de las nuevas tecnologías, las tarjetas de crédito se han convertido en un componente crucial de la vida cotidiana. Esto se debe a su versatilidad para su uso en una amplia gama de transacciones, tanto en compras en línea como en establecimientos físicos. Sin embargo, este progreso tecnológico y financiero también ha dado lugar a un aumento en la creatividad de individuos maliciosos que buscan aprovecharse de estos avances para su propio beneficio. Se ha vuelto más evidente que nunca el robo y la falsificación de tarjetas de crédito.

01

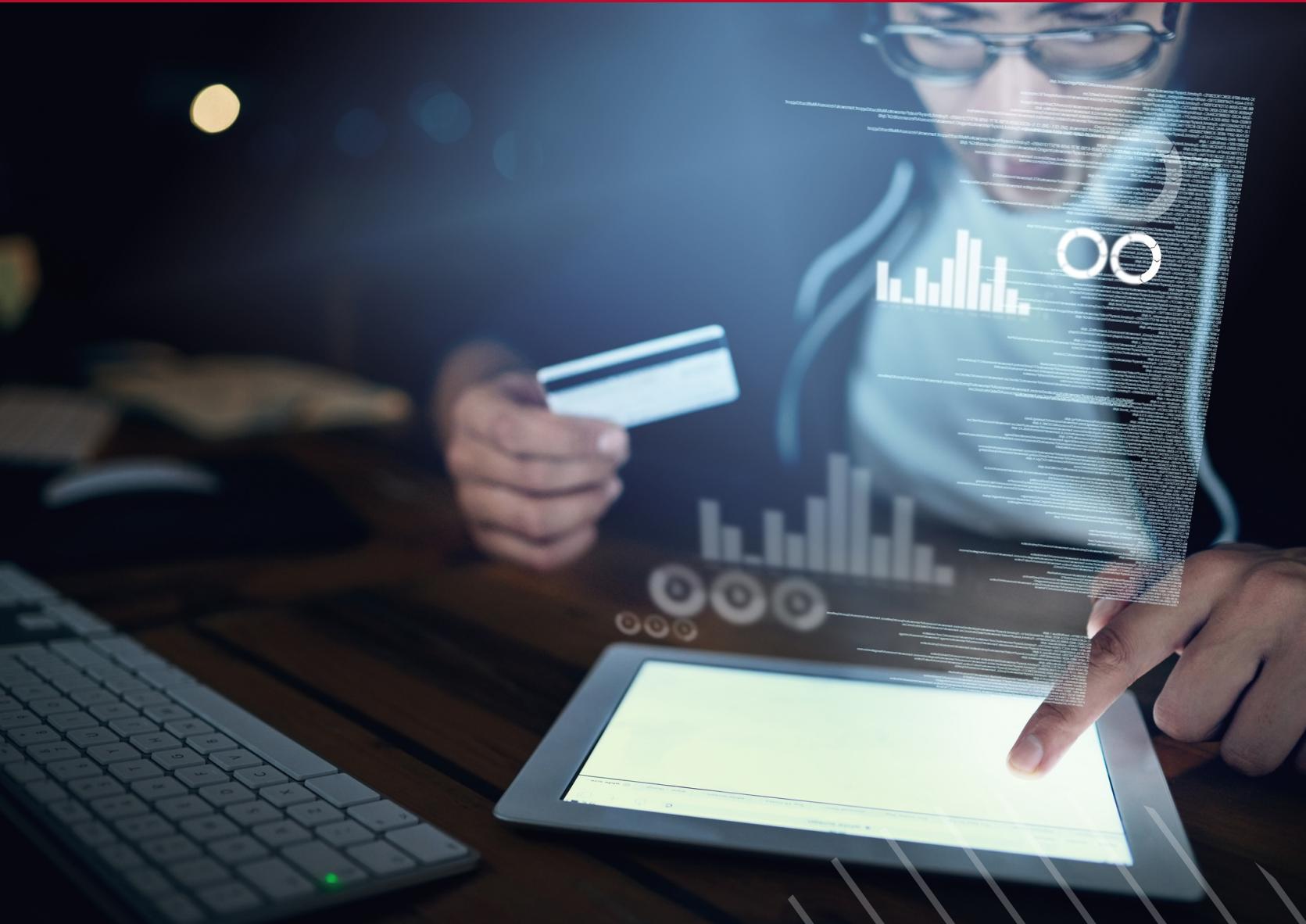




INTRODUCCIÓN

Por lo tanto, es de suma importancia implementar nuevos procedimientos destinados a la identificación de transacciones potencialmente fraudulentas. Este enfoque resulta fundamental para combatir esta creciente amenaza. Al hacerlo, podemos salvaguardar los intereses tanto de los consumidores como de las entidades financieras, evitando pérdidas financieras significativas. Asimismo, contribuye a fortalecer y consolidar la confianza en el sistema financiero en su conjunto, fomenta el cumplimiento de la normativa vigente y permite una utilización responsable de las avanzadas tecnologías disponibles en la actualidad.

02



EL PROBLEMA



Construir modelos de Machine Learning para detectar fraude financiero presenta varios desafíos:

● Tamaño de los datos

● Desequilibrio de clases

El desequilibrio de clases en conjuntos de datos grandes suele causar que los algoritmos de aprendizaje automático se centren en maximizar la precisión general, lo que significa clasificar la mayoría de las observaciones como pertenecientes a la clase mayoritaria. En la detección de fraudes financieros, esto lleva a una baja capacidad de predecir transacciones fraudulentas, que son la clase minoritaria de interés, ya que se pasan por alto en favor de la clase mayoritaria.



OBJETIVO GENERAL DE LA TESIS

Implementar y evaluar múltiples modelos de aprendizaje automático en un entorno distribuido para determinar cuál proporciona un rendimiento óptimo en la detección de anomalías en transacciones con tarjetas de crédito.





ENFOQUE

Priorizamos evitar omitir transacciones fraudulentas en lugar de etiquetar erróneamente transacciones legítimas como fraudulentas. Para escenarios de desequilibrio de clases, la métrica más relevante es la sensibilidad (recall o tasa de verdaderos positivos), ya que mide la capacidad del modelo para identificar con precisión las transacciones fraudulentas en comparación con las transacciones fraudulentas reales.

Se decide maximizar la sensibilidad para minimizar la omisión de transacciones fraudulentas, lo que resulta crítico cuando se prioriza la prevención de estos errores.



HERRAMIENTAS



MLlib

kaggle

DATASETS

Credit Card Transaction

Transaction

User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Errors?	Is Fraud?	
0	0	0	2002	9	1	06:21	\$134.09	Swipe Transaction	3527213246127876953	La Verne	CA	91750.0	5300	NaN	No
1	0	0	2002	9	1	06:42	\$38.48	Swipe Transaction	-727612092139916043	Monterey Park	CA	91754.0	5411	NaN	No
2	0	0	2002	9	2	06:22	\$120.34	Swipe Transaction	-727612092139916043	Monterey Park	CA	91754.0	5411	NaN	No
3	0	0	2002	9	2	17:45	\$128.95	Swipe Transaction	3414527459579106770	Monterey Park	CA	91754.0	5651	NaN	No
4	0	0	2002	9	3	06:23	\$104.71	Swipe Transaction	5817218446178736267	La Verne	CA	91750.0	5912	NaN	No

Cards

User	CARD INDEX	Card Brand	Card Type	Card Number	Expires	CVV	Has Chip	Cards Issued	Credit Limit	Acct Open Date	Year PIN last Changed	Card on Dark Web	
0	0	0	Visa	Debit	4344676511950444	12/2022	623	YES	2	\$24295	09/2002	2008	No
1	0	1	Visa	Debit	4956965974959986	12/2020	393	YES	2	\$21968	04/2014	2014	No
2	0	2	Visa	Debit	4582313478255491	02/2024	719	YES	2	\$46414	07/2003	2004	No
3	0	3	Visa	Credit	4879494103069057	08/2024	693	NO	1	\$12400	01/2003	2012	No
4	0	4	Mastercard	Debit (Prepaid)	5722874738736011	03/2009	75	YES	1	\$28	09/2008	2009	No

Users

Person	Current Age	Retirement Age	Birth Year	Birth Month	Gender	Address	Apartment	City	State	Zipcode	Latitude	Longitude	Per Capita Income - Zipcode	Yearly Income - Person	Total Debt	FICO Score	Num Credit Cards	
0 Hazel Robinson	53	66	1966	11	Female	462 Rose Lane		Nan	La Verne	CA	91750	34.15	-117.76	\$29278	\$59696	\$127613	787	5
1 Sasha Sadr	53	68	1966	12	Female	3606 Federal Boulevard		Nan	Little Neck	NY	11363	40.76	-73.74	\$37891	\$77254	\$191349	701	5
2 Saanvi Lee	81	67	1938	11	Female	766 Third Drive		Nan	West Covina	CA	91792	34.02	-117.89	\$22681	\$33483	\$196	698	5
3 Everlee Clark	63	63	1957	1	Female	3 Madison Street		Nan	New York	NY	10069	40.71	-73.99	\$163145	\$249925	\$202328	722	4
4 Kyle Peterson	43	70	1976	9	Male	9620 Valley Stream Drive		Nan	San Francisco	CA	94117	37.76	-122.44	\$53797	\$109687	\$183855	675	1

Transacciones generadas a partir de una simulación realizada por IBM. Los datos abarcan 2000 consumidores (sintéticos) residentes en Estados Unidos, pero que viajan por todo el mundo. Los datos abarcan también décadas de compras e incluyen varias tarjetas de muchos de los consumidores.

DATASETS

MCC (Código de Categoría Mercante)

mcc	edited_description	combined_description	usda_description	irs_description	irs_reportable
742	Veterinary Services	Veterinary Services	Veterinary Services	Veterinary Services	Yes
763	Agricultural Co-operatives	Agricultural Co-operatives	Agricultural Co-operatives	Agricultural Cooperative	Yes
780	Horticultural Services, Landscaping Services	Horticultural Services, Landscaping Services	Horticultural Services	Landscaping Services	Yes
1520	General Contractors-Residential and Commercial	General Contractors-Residential and Commercial	General Contractors-Residential and Commercial	General Contractors	Yes
1711	Air Conditioning Contractors – Sales and Installation, Heating Contractors – Sales, Service, Installation	Air Conditioning Contractors – Sales and Installation, Heating Contractors – Sales, Service, Installation	Air Conditioning Contractors – Sales and Installation	Heating, Plumbing, A/C	Yes
1731	Electrical Contractors	Electrical Contractors	Electrical Contractors	Electrical Contractors	Yes
1740	Insulation – Contractors, Masonry, Stonework Contractors, Plastering Contractors, Stonework and Masonry Contractors, Masonry Contractors, Tile Settings Contractors	Insulation – Contractors, Masonry, Stonework Contractors, Plastering Contractors, Stonework and Masonry Contractors, Tile Settings Contractors	Insulation – Contractors	Masonry, Stonework, and Plaster	Yes
1750	Carpentry Contractors	Carpentry Contractors	Carpentry Contractors	Carpentry Contractors	Yes
1761	Roofing – Contractors, Sheet Metal Work – Contractors, Siding – Contractors	Roofing – Contractors, Sheet Metal Work – Contractors, Siding – Contractors	Roofing - Contractors	Roofing/Siding, Sheet Metal	Yes
1771	Contractors – Concrete Work	Contractors – Concrete Work	Contractors – Concrete Work	Concrete Work Contractors	Yes
1799	Contractors – Special Trade, Not Elsewhere Classified	Contractors – Special Trade, Not Elsewhere Classified	Contractors – Special Trade, Not Elsewhere Classified	Special Trade Contractors	Yes
2741	Miscellaneous Publishing and Printing	Miscellaneous Publishing and Printing	Miscellaneous Publishing and Printing	Miscellaneous Publishing and Printing	Yes
2791	Typesetting, Plate Making, & Related Services	Typesetting, Plate Making, & Related Services	Typesetting, Plate Making, & Related Services	Typesetting, Plate Making, and Related Services	Yes
2842	Specialty Cleaning, Polishing, and Sanitation Preparations	Specialty Cleaning, Polishing, and Sanitation Preparations	Specialty Cleaning, Polishing, and Sanitation Preparations	Specialty Cleaning	Yes
3000	UNITED AIRLINES	UNITED AIRLINES	UNITED AIRLINES	Airlines	Yes
3001	AMERICAN AIRLINES	AMERICAN AIRLINES	AMERICAN AIRLINES	Airlines	Yes
3002	PAN AMERICAN	PAN AMERICAN	PAN AMERICAN	Airlines	Yes
3003	Airlines	Airlines	null	Airlines	Yes
3004	TRANS WORLD AIRLINES	TRANS WORLD AIRLINES	TRANS WORLD AIRLINES	Airlines	Yes

Repositorio público
de MCC (Código de
Categoría Mercante)
en diferentes
formatos de fácil
lectura

PROCEDIMIENTO

Para la elaboración del modelo se llevan a cabo una lista de actividades en distintos Notebooks de Databricks (desplegado en una instancia de Azure), utilizando diferentes clústeres dependiendo de las necesidades del proceso a realizar. A continuación, se muestran las actividades generales realizadas para construir y escoger el modelo de detección de fraude.

Configuración del entorno de trabajo

- Instancia de Azure Databricks.
- Se configuran 5 clústers de diferentes tamaños en Databricks.
- Se descargan y configuran las credenciales de la cuenta de Kaggle.

Análisis Exploratorio

Identificar patrones, abordar problemas de calidad de datos, generar visualizaciones y calcular estadísticas que confirman el desequilibrio de clases en el conjunto de datos.

Modelado de ML

- Particionamiento del DataFrame para su procesamiento.
- Codificación con One-Hot-Encoding y String Indexer
- División estratificada de datos en train y test.
- Reducción de dimensionalidad con PCA
- Oversampling para mejorar la representación de la clase minoritaria.
- Entrenamiento de (3) modelos: SVM, Regresión Logística e Isolation Forest.

1

2

3

4

5

6

Extracción y procesamiento

Los datos de transacciones fraudulentas, tarjetas de crédito, usuarios y MCC se descargan y almacenan en Databricks mediante Python para su posterior tratamiento.

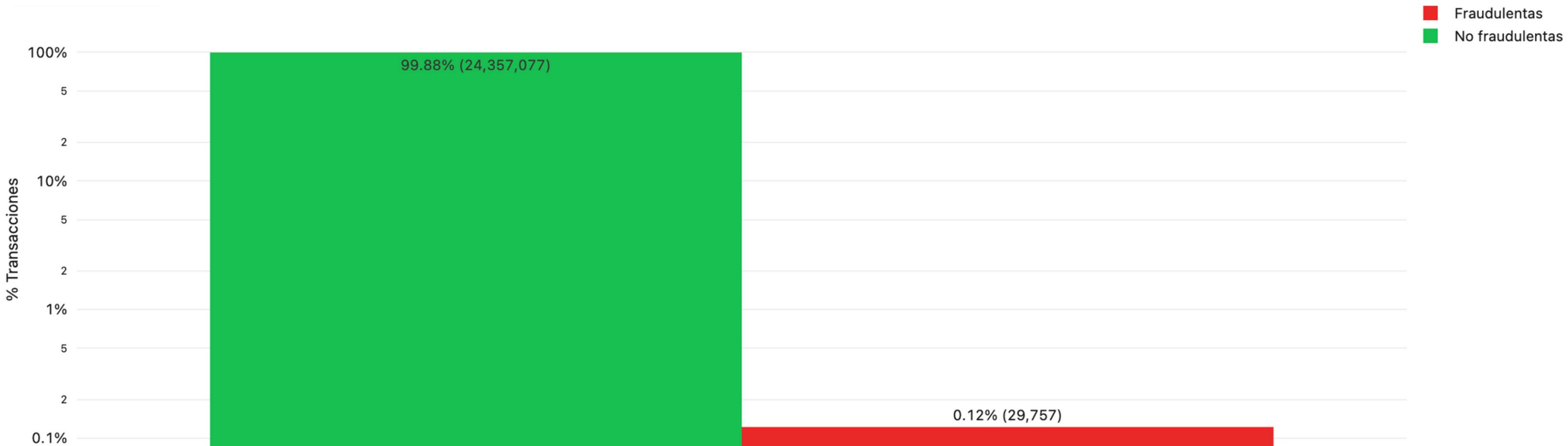
Transformación y feature engineering

Transformaciones en los datos y creación de nuevas características (features) relevantes.

Evaluación

El SVM es el mejor en detectar fraudes, pero con riesgo de falsos positivos. La regresión logística es menos sensible y también con alto riesgo de falsos positivos. Isolation Forest tiene baja sensibilidad pero menos falsos positivos.

TRANSACCIONES FRAUDULENTAS VS NO FRAUDULENTAS



Luego del análisis exploratorio, el resultado más relevante es comprender que el conjunto de datos presenta desbalance de clases.

CARACTERÍSTICAS DE LOS MODELOS APLICADOS

Se llevaron a cabo diversas tareas que involucraron la creación de características, la aplicación de transformaciones, codificaciones, reducción de dimensiones y selección de variables. El proceso para obtener un modelo adecuado implicó probar distintas combinaciones de parámetros, modelos, codificadores, transformaciones y variables en distintos experimentos. Esto quiere decir que la creación del modelo es un proceso iterativo.

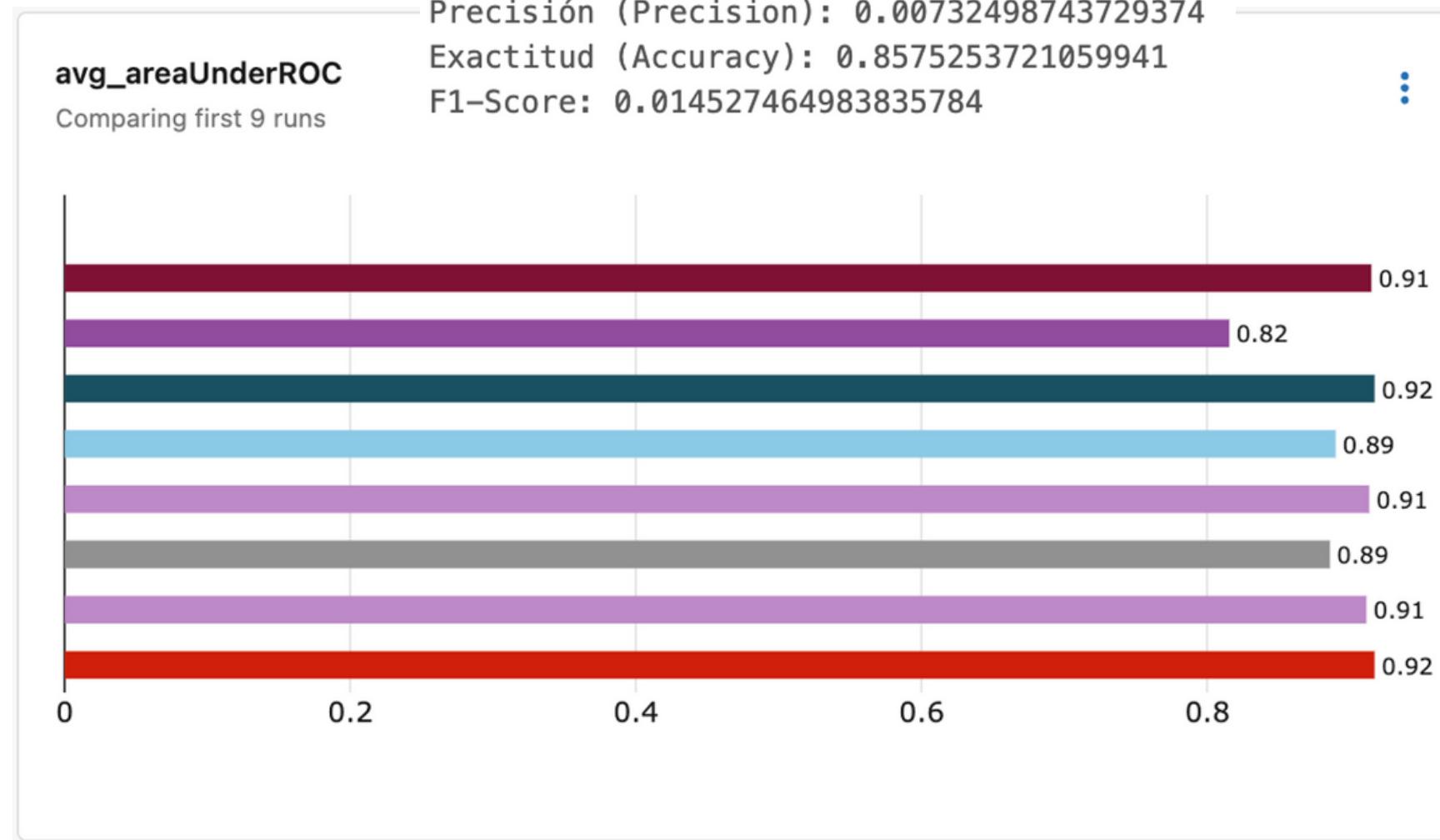
Sin embargo, como muchos de esos modelos no fueron óptimos, el enfoque que se detallará a continuación resultó ser el experimento más efectivo y que se tomará como resultado final.

Particionamiento	Particionamiento con una columna de salt generada con números aleatorios de una distribución uniforme.
Transformaciones básicas adicionales	<ul style="list-style-type: none">• Eliminación de variables que generan ruido: user_id y card_id.• Crear nuevas features de fecha: month, day, day_of_year, day_of_month, day_of_week.• Eliminación de la variable merchant_city, pues esta tenía más de 10,000 valores, se podía utilizar el merchant_state que tenía menos valores.
Codificador de variables	StringIndexer y OneHotEncoder (dependiendo de la cantidad de posibles valores de la columna).
Selector de variables	No aplicado.
Reducción de dimensionalidad	PCA.
Estandarización de datos	No aplicado.
Unión de variables en un vector	VectorAssembler.
Separación train-test	División estratificada con randomSplit. Se aplicó adicionalmente Oversampling para equilibrar las clases.
Modelo	LinearSVC, Logistic Regression e IsolationForest.
Ejecución del modelo	ParamGridBuilder, CrossValidator.

RESULTADOS

SVM

Sensibilidad (Recall): 0.8686114880144731
Precisión (Precision): 0.00732498743729374
Exactitud (Accuracy): 0.8575253721059941
F1-Score: 0.014527464983835784



label	prediction	count_svm
1	0	1162
0	0	6265219
1	1	7682
0	1	1041057

Como el enfoque del experimento es la sensibilidad (recall), podemos observar que el modelo responde muy bien al valor, aunque la precisión no es muy óptima.

Isolation Forest

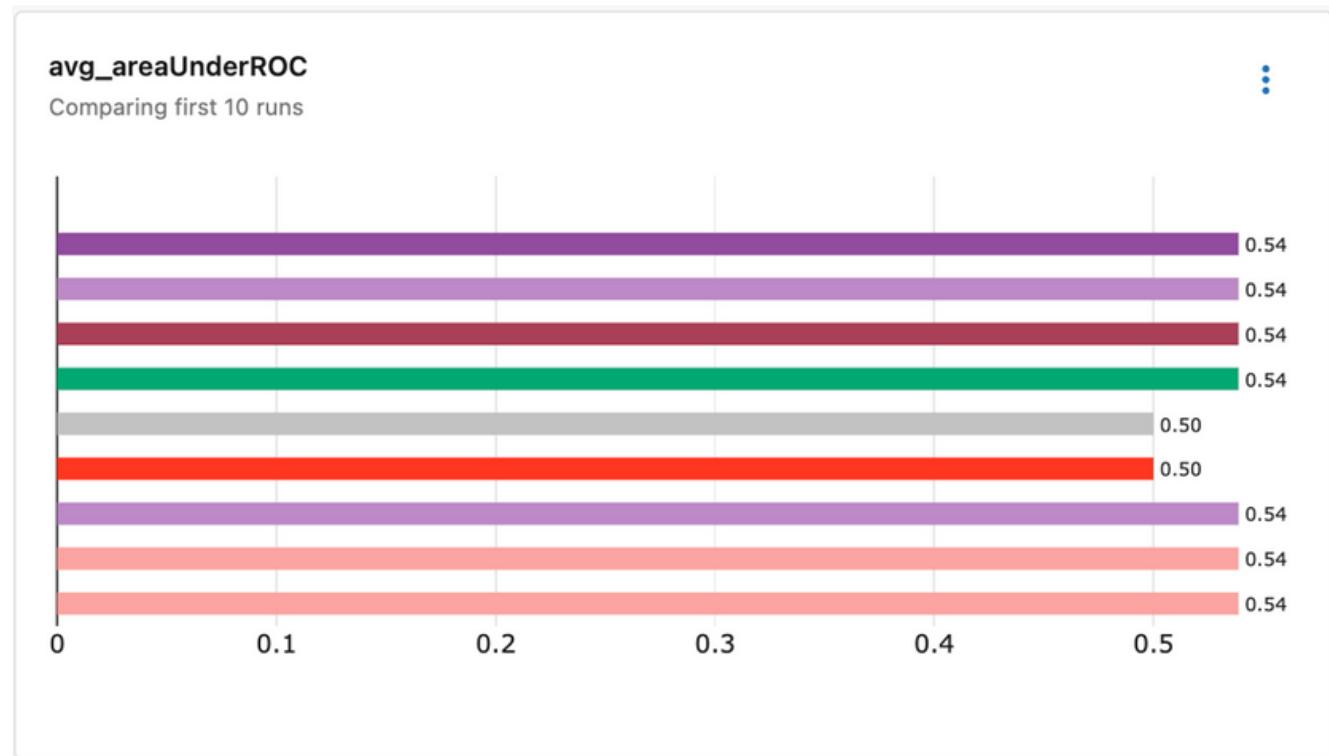
label	prediction	count_i_forest
1	0	7981
0	0	7230387
1	1	863
0	1	75889

Este modelo es no supervisado, pero como realmente sí tenemos las variables de salida, excepcionalmente se pueden calcular las métricas.

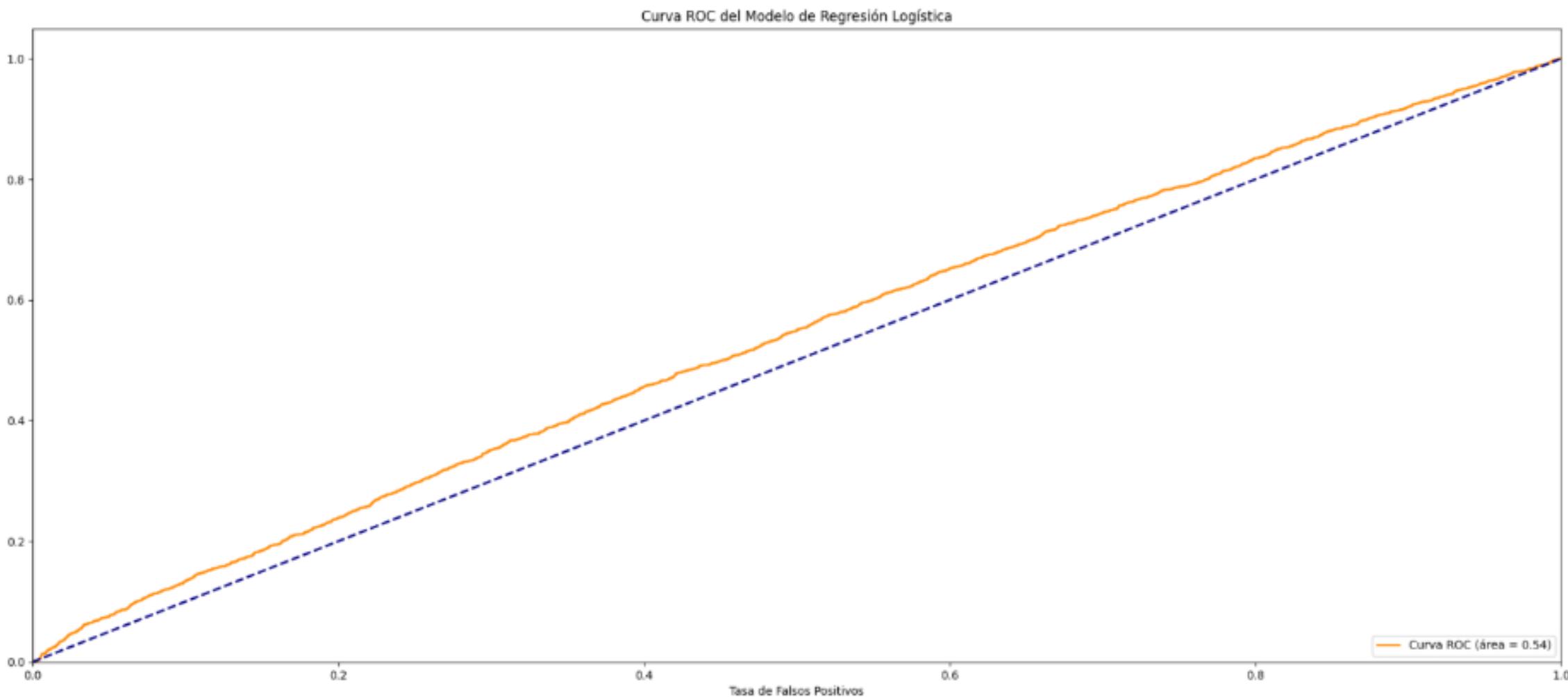
Se muestra que el modelo tiene un valor muy bajo para la sensibilidad. La única métrica que muestra un valor significativamente alto es la precisión, pero esta no es la prioridad.

RESULTADOS

Regresión Logística



label	prediction	count_lr
1	0	3506
0	0	3264981
1	1	5338
0	1	4041295



Sensibilidad (Recall): 0.6035730438715513
Precisión (Precision): 0.0013191213534807828
Exactitud (Accuracy): 0.44706293266549285
F1-Score: 0.0026324893471224224

En la gráfica anterior de ROC, se observa con más detalle que el modelo es prácticamente aleatorio. También se observa que el modelo tiene un valor bastante regular de 0.6, pero tampoco son muy buenos los valores de precisión, exactitud o F1-Score.

RESULTADOS

Métrica	SVM	Regresión Logística	Isolation Forest
ROC	0.92	0.54	-
Sensibilidad	0.8686	0.6035	0.0975
Precisión	0.0073	0.0013	0.0112
Exactitud	0.8575	0.4470	0.9885
F1-Score	0.0145	0.0026	0.0201

SVM tiene la sensibilidad más alta (0.8686), lo que significa que es el mejor modelo para identificar transacciones fraudulentas. Sin embargo, la precisión (0.0073) es baja, lo que indica que puede haber un número significativo de falsos positivos.

La regresión logística tiene una sensibilidad (0.6035) menor en comparación con SVM, pero aun así es significativa. La precisión (0.0013) es aún más baja, lo que sugiere una mayor cantidad de falsos positivos.

Para el caso de Isolation Forest, aunque este modelo es no supervisado, se poseen las etiquetas de los datos, por lo que solo para este caso es posible calcular las métricas que normalmente no serían posibles. Siendo así, el modelo tiene una sensibilidad muy baja (0.0975), lo que indica que es menos efectivo en la detección de transacciones fraudulentas. Sin embargo, la precisión (0.0112) es más alta en comparación con los otros modelos.

CONCLUSIONES

- Dado que la prioridad es minimizar la cantidad de transacciones fraudulentas pasadas por alto, SVM es la mejor opción entre estos tres modelos debido a su alta sensibilidad. Sin embargo, también se debe considerar que la precisión de SVM es baja, lo que significa que puede generar más falsos positivos.
- El oversampling mejora los resultados al reducir los falsos negativos, aunque puede aumentar los falsos positivos. El enfoque se centra en generar alertas para detectar comportamientos sospechosos en lugar de centrarse solo en transacciones fraudulentas, lo que es beneficioso.
- Databricks proporciona un entorno robusto para abordar desafíos en el procesamiento de grandes conjuntos de datos. Permite la creación de clústers, partición de datos y acceso a datos desde data lakes y bases de datos en un solo lugar, lo que simplifica la integración de datos y mejora la eficiencia en la detección de fraudes en transacciones con tarjetas de crédito. Utilizar Databricks fue beneficioso para este proyecto.



ÁREAS DE MEJORA

- Considerar alternativas a Databricks y MLlib, como TensorFlow, para abordar grandes conjuntos de datos de manera más eficiente en la detección de fraudes.
- Explorar librerías y lenguajes de programación que ofrezcan un espectro más amplio de modelos y técnicas de aprendizaje automático en entornos distribuidos.
- Investigar modelos de aprendizaje automático no supervisados, que pueden mejorar la detección de anomalías sin depender de etiquetas de datos.
- Buscar un equilibrio más adecuado entre las métricas de recall, precisión y exactitud en el rendimiento del modelo.
- Validar los resultados en conjuntos de datos reales para evaluar la idoneidad de las características en un entorno más representativo.

