Project:
Vehicle Insurance Claim

Preguntas del Negocio

Project Summary

El proyecto se centra en la detección de fraude en seguros de vehículos, utilizando un conjunto de datos real proporcionado por Oracle, proveniente de una aseguradora en Estados Unidos. El fraude en seguros de vehículos incluye prácticas como la presentación de reclamos falsos o exagerados relacionados con daños materiales o lesiones personales tras un accidente.

Objetivo No 1

Identificar las Características Más Relevantes para la Detección de Fraude comprende:

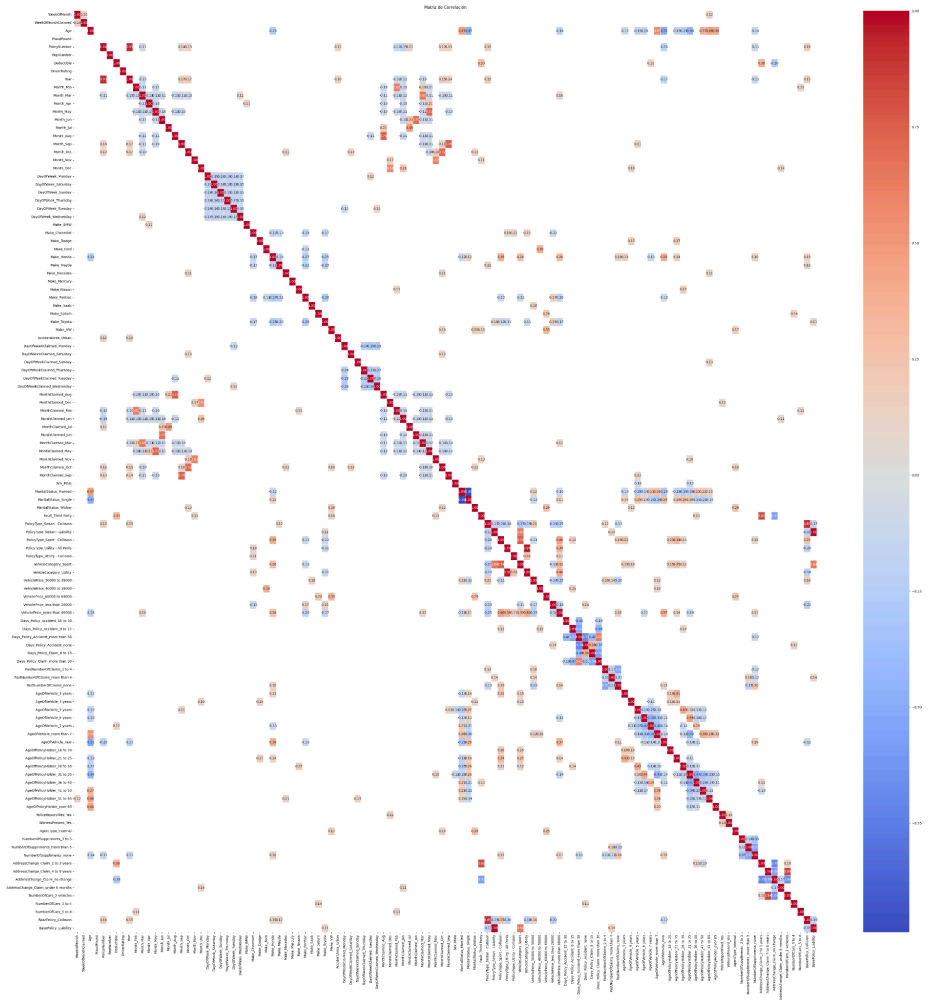
a) Patrones
b) Reducción de la Dimensionalidad de Características

A continuación se describen los patrones y tendencias hallados en el comportamiento de fraudes, según el componente del cliente (atributos demográficos, producto, zona, vehículo).

Preguntas sobre el Comportamiento del Cliente

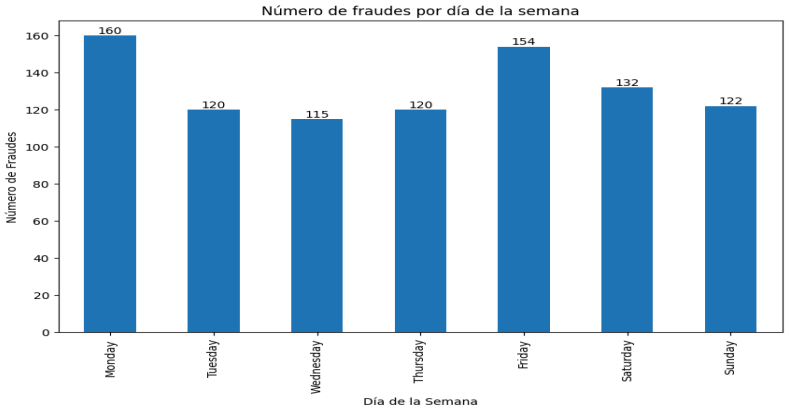
Temporalidad de los Fraudes:

- ¿Existe una correlación entre el mes o la semana del mes en la que se producen los accidentes y la probabilidad de fraude?

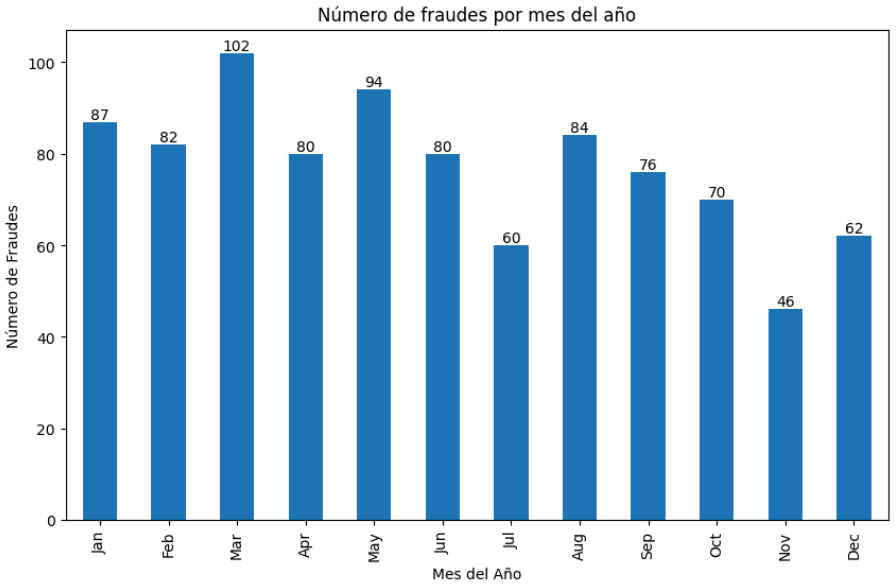


- ¿Los fraudes son más comunes en ciertos días de la semana?

Sí, los fraudes son más comunes los lunes y viernes, posiblemente debido a que los reclamantes intentan aprovechar los fines de semana, Sin embargo, el análisis de correlación no muestra una fuerte relación entre el día de la semana y la probabilidad de fraude, indicando que su comportamiento es similar al de la población no fraudulenta. Aunque hay una tendencia, el este atributo no es un fuerte predictor de fraude.



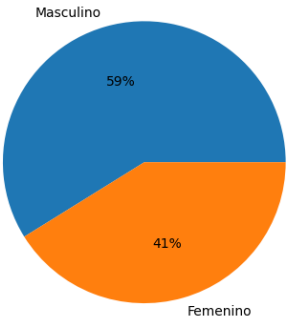
- ¿Los fraudes son más comunes en ciertos meses del año ?

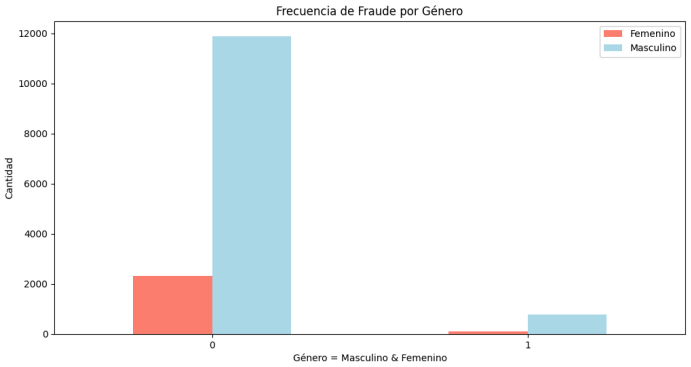


Demografía del Reclamante:

- ¿Qué edad y género tienen mayor incidencia de reclamaciones fraudulentas?

  Las personas jóvenes (18-25 años) y los hombres presentan una mayor incidencia de reclamaciones fraudulentas, posiblemente debido a comportamientos de riesgo más elevados.




Frecuencia de Fraude por Género
Femenino
Masculino
12000
10000
8000
6000
4000
2000
0
Cantidad
0
1
Género = Masculino & Femenino

- ¿El estado civil del reclamante influye en la probabilidad de que una reclamación sea fraudulenta?

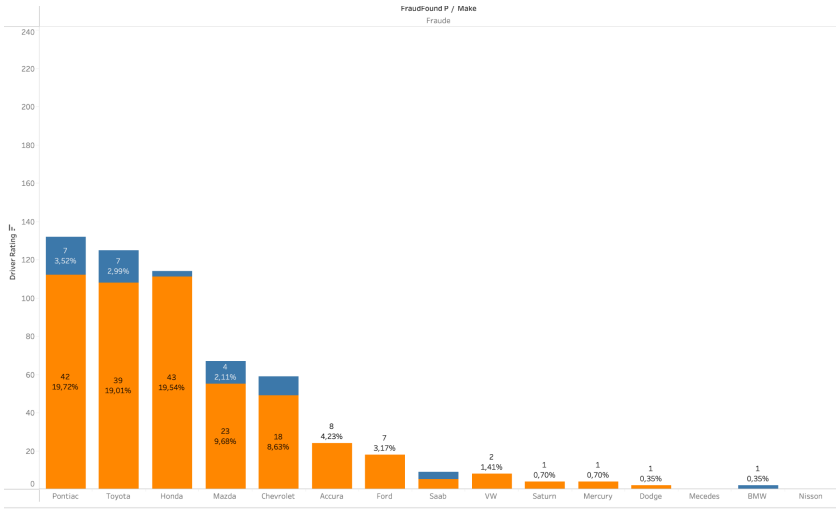
  Los reclamantes solteros tienen una mayor probabilidad de presentar reclamaciones fraudulentas en comparación con los casados. Sin embargo, el análisis de correlación no muestra una fuerte relación entre el estado civil y la probabilidad de fraude, indicando que su comportamiento es similar al de la población no fraudulenta. Aunque hay una tendencia, el estado civil no es un fuerte predictor de fraude.

	FraudFound P	
Marital Sta..	Fraude	No fraude
Divorced	3	73
Married	639	9.986
Single	278	4.406
Widow	3	32

Características del Vehículo y Póliza:

- ¿Ciertos fabricantes de vehículos tienen una mayor tasa de fraudes?

  Algunos fabricantes de vehículos de lujo tienen una mayor tasa de fraudes, posiblemente debido al mayor valor de los reclamos



- ¿El precio del vehículo está relacionado con la incidencia de fraudes?

Sí, los vehículos más caros tienden a tener una mayor incidencia de fraudes. Sin embargo, el análisis de correlación no muestra una fuerte relación entre el precio del vehículo y la probabilidad de fraude, indicando que su comportamiento es similar al de la población no fraudulenta. Aunque hay una tendencia, el este atributo no es un fuerte predictor de fraude

	FraudFound P	
Vehicle Price	Fraude	No fraude
20000 to 29000	421	7.658
30000 to 39000	175	3.358
40000 to 59000	31	430
60000 to 69000	4	83
less than 20000	103	993
more than 69000	189	1.975

- ¿La categoría del vehículo (p. ej., sedán, SUV) afecta la probabilidad de fraude?

Los SUVs y vehículos deportivos tienen una mayor probabilidad de fraudes en comparación con los sedanes.Sin embargo, el análisis de correlación no muestra una fuerte relación entre la categoría del vehículo y la probabilidad de fraude, indicando que

su comportamiento es similar al de la población no fraudulenta. Aunque hay una tendencia, el este atributo no es un fuerte predictor de fraude

	FraudFound P	
Vehicle Cat..	Fraude	No fraude
Sedan	795	8.876
Sport	84	5.274
Utility	44	347

- ¿Existen tipos específicos de pólizas (Liability, Collision, All Perils) más propensos a fraudes?

  Las pólizas de "Collision" y "All Perils" presentan una mayor incidencia de fraudes.

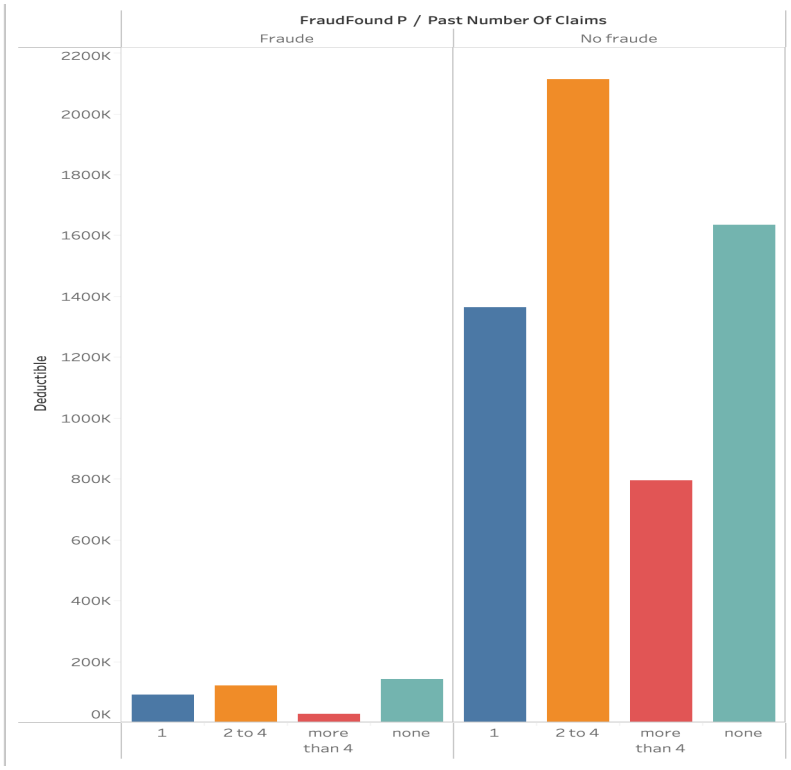
	FraudFound P	
Policy Type	Fraude	No fraude
Sedan - All Perils	411	3.676
Sedan - Collision	384	5.200
Sedan - Liability	36	4.951
Sport - All Perils		22
Sport - Collision	48	300
Sport - Liability		1
Utility - All Perils	41	299
Utility - Collision	3	27
Utility - Liability		21

Historial y Comportamiento del Reclamante:

- ¿Las personas con un alto número de reclamaciones pasadas tienen una mayor tendencia a realizar fraudes?

  Sí, los reclamantes con un historial de múltiples reclamaciones tienen una mayor probabilidad de presentar reclamos fraudulento.Sin embargo, el análisis de correlación no muestra una fuerte relación entre la presencia de fraudes y la

probabilidad de fraude, indicando que su comportamiento es similar al de la población no fraudulenta. Aunque hay una tendencia, el este atributo no es un fuerte predictor de fraude



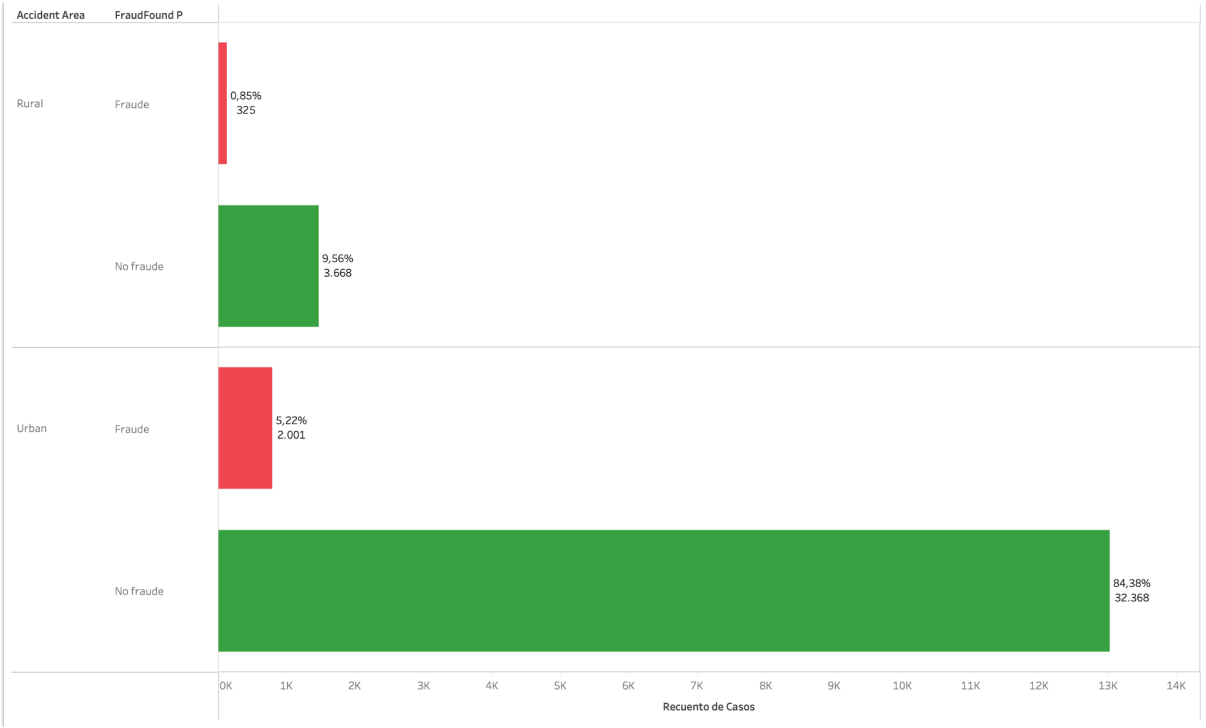
- ¿Cómo influye el tiempo entre la adquisición del seguro y el accidente o la reclamación en la probabilidad de fraude?

FraudFound P	Days Policy Accident	Days Policy Claim: 8 to 15	15 to 30	more than 30	none
Fraude	1 to 7	3			
	8 to 15		7	4	
	15 to 30		2	7	
	more than 30		1	2.278	
	none	4	3	17	
No fraude	1 to 7		17	21	
	8 to 15	33	47	54	
	15 to 30		34	83	
	more than 30		4	35.644	2
	none	13	19	65	

Condiciones del Accidente:

- ¿Los accidentes en áreas rurales o urbanas tienen diferente incidencia de fraude?

  Los accidentes en áreas urbanas tienen una mayor incidencia de fraude en comparación con las áreas rurales



- ¿La presencia de testigos influye en la probabilidad de que una reclamación sea fraudulenta?

  La ausencia de testigos incrementa la probabilidad de que una reclamación sea fraudulenta.

	Witness Present	
FraudFoun..	No	Yes
Fraude	2.317	9
No fraude	35.815	221

- ¿El reporte a la policía está relacionado con la detección de fraudes?

	Police Report Filed	
FraudFoun..	No	Yes
Fraude	907	16
No fraude	14.085	412

Importancia de características

El gráfico de barras muestra cuáles características (o variables) son más importantes en un modelo de predicción. Básicamente, nos dice qué factores tienen mayor influencia en los resultados del modelo.

Puntos Clave

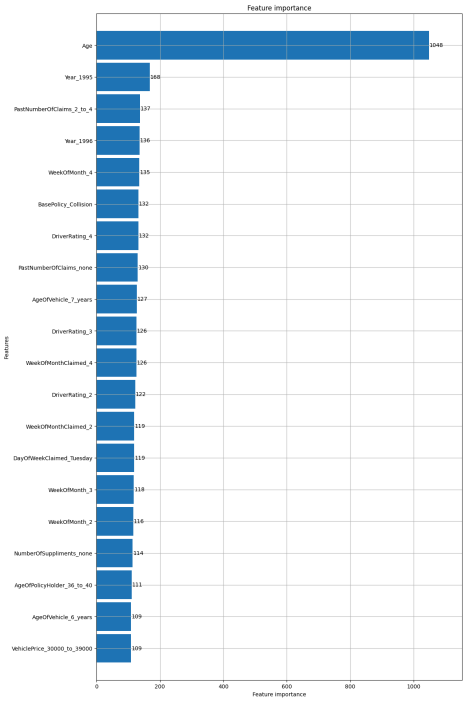
1. **Edad (Age)**:
   - La edad es la característica más importante con diferencia. Esto significa que la edad de las personas tiene el mayor impacto en las predicciones del modelo.
2. **Otras Características Importantes**:
   - El año 1995 (Year_1995) y el número de reclamos anteriores entre 2 y 4 (PastNumberOfClaims_2_to_4) también son bastante influyentes, aunque no tanto como la edad.
3. **Características Menos Importantes**:
   - Características como el precio del vehículo entre 30,000 y 39,000 (VehiclePrice_30000_to_39000) y la edad del vehículo de 6 años (AgeOfVehicle_6_years) son menos influyentes en el modelo.

¿Por qué es importante esto?

1. **Mejora del Modelo**:
   - Saber qué características son más importantes nos ayuda a mejorar la precisión del modelo enfocándonos en las variables clave.
2. **Simplificación**:
   - Podemos simplificar el modelo eliminando las características que tienen poca importancia, haciendo el modelo más fácil de manejar sin perder precisión.
3. **Facilidad de Entendimiento**:

- Hace que sea más fácil explicar el modelo y sus resultados a otras personas, ya que podemos decir claramente cuáles factores son más relevantes.

En resumen, este gráfico nos ayuda a entender qué factores son más importantes para nuestras predicciones, permitiéndonos mejorar y simplificar nuestro modelo de manera efectiva.



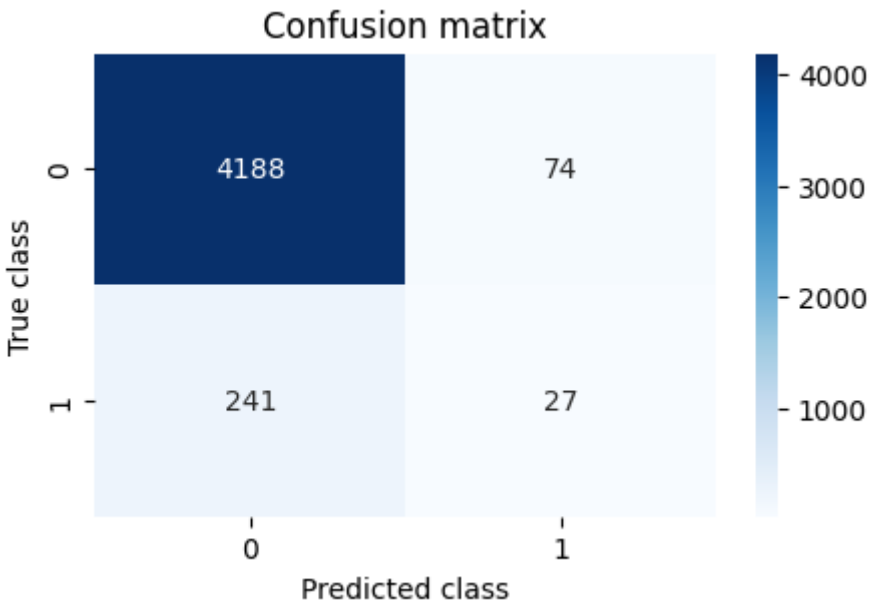
Conclusión.

Los reclamantes jóvenes (18-25 años), especialmente hombres solteros, tienen mayor tendencia a presentar reclamaciones fraudulentas en seguros de vehículos. Los vehículos de lujo y SUVs, así como las pólizas de colisión y todas las coberturas, son más propensos a fraudes. Un historial de múltiples reclamaciones aumenta esta probabilidad. Los accidentes en áreas urbanas y sin testigos, así como los que no se reportan a la policía, también tienen una mayor incidencia de fraude. Además, un alto número de suplementos y cambios en la información del reclamante después de presentar la reclamación son indicadores adicionales de fraude.

Objetivo No 2

Mejorar la Precisión en la Detección de Fraude.

```
Pipeline
  StandardScaler
    StandardScaler(with_mean=False)
  LGBMClassifier
    LGBMClassifier(colsample_bytree=1, learning_rate=0.2, n_estimators=150,
                   num_leaves=62, random_state=137, reg_alpha=0)
```



La matriz de confusión proporcionada es una herramienta útil para evaluar el rendimiento de un modelo de clasificación, especialmente en el contexto de la identificación de fraudes. A continuación, desglosa la matriz en términos de los conceptos clave y su relevancia en la detección de fraudes.

Descripción de la Matriz de Confusión

La matriz de confusión se organiza en cuatro cuadrantes, representando los resultados de las predicciones del modelo comparados con los valores reales.

1. **True Positives (TP)**: En el contexto de fraude, esto representa los casos donde el modelo predijo correctamente que hubo fraude. En la matriz, este valor es 27.
2. **True Negatives (TN)**: Representa los casos donde el modelo predijo correctamente que no hubo fraude. En la matriz, este valor es 4188.
3. **False Positives (FP)**: Estos son los casos donde el modelo predijo fraude incorrectamente (falsas alarmas). En la matriz, este valor es 74.
4. **False Negatives (FN)**: Representa los casos donde el modelo no detectó un fraude que realmente ocurrió. En la matriz, este valor es 241.

Interpretación en el Contexto de Identificación de

Fraudes

- **True Positives (TP)**: Es esencial que este número sea lo más alto posible en el contexto de fraude, ya que significa que el modelo está correctamente identificando transacciones fraudulentas.
- **True Negatives (TN)**: Un alto número de TN indica que el modelo está funcionando bien en reconocer transacciones legítimas.
- **False Positives (FP)**: Un bajo número de FP es deseable, ya que falsas alarmas pueden generar costos adicionales y molestias a los clientes legítimos.
- **False Negatives (FN)**: Es crucial minimizar FN, ya que cada FN representa una transacción fraudulenta no detectada, lo cual puede tener consecuencias financieras severas.

Métricas Derivadas de la Matriz de Confusión

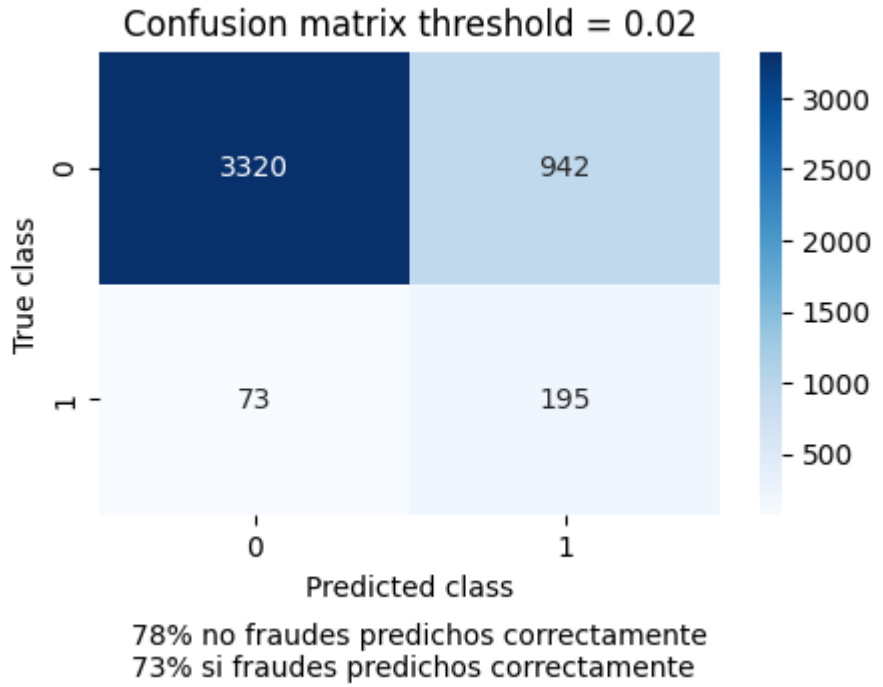
A partir de la matriz de confusión, se pueden calcular varias métricas para evaluar el rendimiento del modelo:

- **Precisión (Accuracy)**: La proporción de predicciones correctas sobre el total de casos. Precisio´n=TP+TNTP+TN+FP+FN=4188+274188+27+74+241=0.9383\text{Precisió n} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{4188 + 27}{4188 + 27 + 74 + 241} = 0.9383Precisio´n=TP+TN+FP+FNTP+TN=4188+27+74+2414188+27=0.9383
- **Precisión Positiva (Precision)**: La proporción de verdaderos positivos sobre todas las predicciones positivas (fraudes detectados). Precisio´n=TPTP+FP=2727+74=0.2673\text{Precisión} = \frac{TP}{TP + FP} = \frac{27}{27 + 74} = 0.2673Precisio´n=TP+FPTP=27+7427=0.2673
- **Sensibilidad (Recall o Tasa de Verdaderos Positivos)**: La proporción de verdaderos positivos sobre todos los casos positivos reales. Sensibilidad=TPTP+FN=2727+241=0.1004\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{27}{27 + 241} = 0.1004Sensibilidad=TP+FNTP=27+24127=0.1004
- **Especificidad (Specificity)**: La proporción de verdaderos negativos sobre todos los casos negativos reales. Especificidad=TNTN+FP=41884188+74=0.9826\text{Especificidad} = \frac{TN}{TN + FP} = \frac{4188}{4188 + 74} = 0.9826Especificidad=TN+FPTN=4188+744188=0.9826

Conclusión

En el contexto de la identificación de fraudes, la matriz de confusión muestra que el modelo tiene una alta precisión general (93.83%), pero su capacidad para detectar fraudes (recall) es baja (10.04%). Esto sugiere que, aunque el modelo es bueno para identificar transacciones legítimas, necesita mejoras significativas en la detección de fraudes para ser útil en aplicaciones prácticas. Las métricas específicas, como la precisión positiva y la sensibilidad, son críticas para evaluar y mejorar los modelos de fraude, ya que el costo de no detectar un fraude (FN) es generalmente mucho mayor que el costo de una falsa alarma (FP

Optimizacion del modelo



La nueva matriz de confusión con un umbral de 0.02 muestra un ajuste en los hiperparámetros del mismo modelo, lo cual ha alterado el rendimiento en la identificación de fraudes. A continuación, se desglosa y se compara el rendimiento del modelo con el nuevo umbral.

Descripción de la Nueva Matriz de Confusión

1. **True Positives (TP)**: 195 (fraudes predichos correctamente).
2. **True Negatives (TN)**: 3320 (transacciones no fraudulentas predichas correctamente).
3. **False Positives (FP)**: 942 (transacciones no fraudulentas predichas incorrectamente como fraudulentas).
4. **False Negatives (FN)**: 73 (fraudes que no fueron detectados).

Métricas Derivadas de la Nueva Matriz de Confusión

- **Precisión (Accuracy)**:
  Precisio´n=TP+TNTP+TN+FP+FN=195+3320195+3320+942+73=0.766\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{195 + 3320}{195 + 3320 + 942 + 73} = 0.766Precisio´n=TP+TN+FP+FNTP+TN=195+3320+942+73195+3320=0.766
- **Precisión Positiva (Precision)**:
  Precisio´n=TPTP+FP=195195+942=0.1714\text{Precisión} = \frac{TP}{TP + FP} = \frac{195}{195 + 942} = 0.1714Precisio´n=TP+FPTP=195+942195=0.1714
- **Sensibilidad (Recall o Tasa de Verdaderos Positivos)**:
  Sensibilidad=TPTP+FN=195195+73=0.727\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{195}{195 + 73} = 0.727Sensibilidad=TP+FNTP=195+73195=0.727
- **Especificidad (Specificity)**:
  Especificidad=TNTN+FP=33203320+942=0.7792\text{Especificidad} = \frac{TN}{TN + FP} = \frac{3320}{3320 + 942} =

0.7792Especificidad=TN+FPTN=3320+9423320=0.7792

Comparación con la Matriz Anterior

1. **True Positives (TP)** ha incrementado significativamente de 27 a 195, lo que indica que el nuevo umbral mejora la detección de fraudes.
2. **True Negatives (TN)** ha disminuido de 4188 a 3320, indicando que hay más falsas alarmas (FP), aumentando de 74 a 942.
3. **False Negatives (FN)** han disminuido de 241 a 73, lo cual es positivo porque hay menos fraudes no detectados.

Interpretación y Decisión

En el contexto de identificación de fraudes, es crucial encontrar un equilibrio entre detectar la mayor cantidad de fraudes posibles (alto recall) y minimizar las falsas alarmas (bajo false positives).

- **Precisión Positiva (Precision)** ha disminuido, lo que indica que, aunque el modelo detecta más fraudes, también genera más falsas alarmas.
- **Sensibilidad (Recall)** ha mejorado significativamente, lo cual es positivo ya que el modelo ahora detecta un mayor porcentaje de fraudes.

En este caso, aunque la precisión general (Accuracy) ha disminuido debido a un aumento en los falsos positivos, la mejora en la sensibilidad sugiere que el modelo es mejor en detectar fraudes, lo cual es crítico en muchos contextos de fraude donde las consecuencias de no detectar fraudes pueden ser severas.

Conclusión

El modelo ajustado con un umbral de 0.02 parece ser mejor para la detección de fraudes en términos de sensibilidad, aunque a costa de un mayor
número de falsas alarmas. Dependiendo del contexto y de las prioridades de la organización (ya sea minimizar falsas alarmas o maximizar la detección de fraudes), este modelo podría ser preferible. En muchos escenarios de fraude, la capacidad de detectar más fraudes (mayor recall) es más crítica, lo cual sugiere que el modelo ajustado podría ser una mejor opción