Detección de Fraude en Logística Combinando Aprendizajes Supervisados y No Supervisados

Sofia Nahir Sapienza – sofisapi66@gmail.com

***Abstract*— Los aprendizajes supervisados suelen ser muy utilizados en detección de fraude, ya que partiendo de casos marcados como fraudulentos, se puede aprender qué variables explican tal comportamiento y utilizarlas para prevención. Sin embargo, para poder emplear estas técnicas es preciso contar con una muestra marcada, que separe entre aquella población normal y aquella anómala para que el modelo pueda aprender a diferenciarlas. En este trabajo se presenta una técnica híbrida, que combina aprendizaje no supervisado para detección de anomalías en envíos realizados a través de plataformas de comercio electrónico, y a posteriori, un aprendizaje supervisado sobre tal muestra para generación de una marca de fraude.**

I. Introducción

En los envíos generados a partir de compras en plataformas de comercio electrónico, se considera que un envío fue fraudulento cuando se envía algo distinto a lo declarado. El conocimiento de que esto ha sucedido puede inferirse cuando se observa la diferencia entre el pesaje estimado según el contenido declarado por el remitente, versus el pesaje real de ese paquete. De existir una diferencia en el pesaje, esto se traduce a una pérdida económica, ya que esa diferencia monetaria no puede ser transferida al cliente.

El primer propósito de este trabajo es establecer qué diferencias en pesajes son comunes y cuáles anómalas, ya que, por ejemplo, una diferencia absoluta de 500 gramos en un envío estimado originalmente en 5 kg podría ser explicado por cuestiones como el envoltorio, pero una diferencia de 500 gramos en un paquete originalmente estimado en 1 kilogramo representaría un 50% más de pesaje, que podría ser explicado también por el envoltorio, o por estar enviado algo distinto a lo declarado. La hipótesis inicial indica que según de qué estimación de pesaje inicial se parta, el desvío de pesaje considerado "normal" irá aumentando conforme aumente el valor del pesaje estimado. A su vez, se asume que los comportamientos aislados en la distribución de diferencias de pesaje de envíos coinciden con los fraudulentos. Para esta primera etapa, se emplea una técnica no supervisada entrenando un Isolation Forest [1].

La segunda parte de este trabajo utiliza una técnica supervisada de Regresión Logística [2] que, tomando como valor de entrada la muestra marcada del método no supervisado, construye una regla que permite generalizar el umbral de diferencia de pesaje considerado anómalo para construir una marca de fraude por envío.

Existen trabajos en donde se han utilizado técnicas híbridas para detección de fraude, como es en [3], orientado a detección de transacciones fraudulentas con tarjetas de crédito, o [4], en el que K. Yamanishi y J. Takeuchi generalizan esta técnica para detección de anomalías en muestras. La originalidad de este trabajo consiste en aplicar esta técnica a la detección de fraude en logística de comercio electrónico.

II. Método

A. Aprendizaje No Supervisado

En esta sección se detallan los métodos utilizados para el entrenamiento del Isolation Forest, que busca separar aquellas diferencias de pesaje anómalas de las comunes.

1) Construcción del set de datos

Se toma una base de 205,993 envíos, de los que se conoce su pesaje estimado y su pesaje real final, denominado a partir de ahora "peso medido", y se construyen dos variables:

$$Diferencia\ Absoluta = Peso\ Medido - Peso\ Estimado$$

$$Ratio\ Pesos = \frac{Peso\ Medido}{Peso\ Estimado}$$

Se conservan este set de datos tanto diferencias absolutas positivas como negativas, ya que la hipótesis inicial busca encontrar las anomalías detectando aquellos valores más desviados de toda la distribución de envíos.

2) Entrenamiento Isolation Forest

Se separa el set de entrenamiento, el cual solo contiene las variables creadas de Diferencia Absoluta y Ratio Pesos. Utilizando la librería de Python Sklearn [5], se entrena un Isolation Forest utilizando 300 estimadores -árboles que se construirán - y 100 muestras máximas por árbol. Este modelo arroja dos resultados por dato: una clase, que es -1 cuando el dato es considerado anómalo, o 1 cuando el dato no lo es, y un puntaje.

3) Selección del Puntaje de Corte y Validación

Dado que el objetivo de esta primera parte es obtener una etiqueta de anomalía robusta para luego entrenar un modelo supervisado, se analiza el histograma de puntajes (Fig. 1) de aquellos envíos considerados anómalos con diferencias de pesaje positivas y se decide definir un corte en -0.15, dado que puntajes menores o iguales a ese valor concentran el 50% de la distribución. Se descartan las diferencias negativas consideradas anómalas de esta muestra ya que todo el trabajo busca concentrarse en el comportamiento fraudulento, que surge cuando la diferencia es positiva, ya que se incurre en pérdidas monetarias.

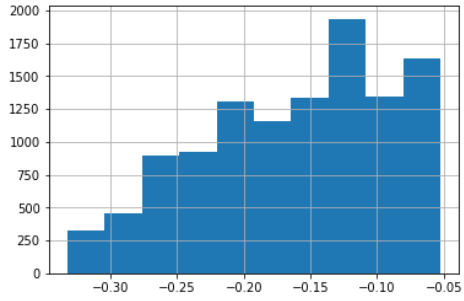

Fig.1 *Distribución de puntaje para envíos con diferencia de pesaje positiva y clasificados como anómalos.*

Este 50% de la distribución de envíos más anómalos arroja un total de 5,700 casos, de los cuales se han validado manualmente 30 casos para comprobar la eficiencia en la clasificación.

B. Aprendizaje Supervisado

En esta segunda parte del trabajo, se elige utilizar un análisis de regresión logística, que emplea la función logística para estimar la probabilidad de ocurrencia de un determinado evento, que, en el caso de dicho análisis, es la ocurrencia de un pesaje anómalo. Dicha probabilidad de ocurrencia queda establecida mediante la siguiente ecuación:

$$Marca_{peso} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * V_1 + \beta_2 * V_2 + \beta_3 * V_3)}}$$

Lo que se busca con este aprendizaje es definir los parámetros β_i que forman parte de la ecuación y acompañan a las variables V_i.

1) Construcción del Set de Datos

Se toma como base los mismos 205,993 envíos que se utilizaron para el entrenamiento no supervisado, etiquetados con su clase final: la anomalía es 1 cuando el puntaje es menor o igual a -0.15 y la diferencia absoluta de pesajes es positiva, de lo contrario es 0. A su vez, se crea la variable que refleja la diferencia relativa entre los pesos, para reemplazar a la variable "Ratio Pesos" utilizada en el modelo no supervisado.

$$Diferencia\ Relativa = \frac{Peso\ Medido - Peso\ Estimado}{Peso\ Estimado}$$

2) Entrenamiento Regresión Logística

Se separa el set de entrenamiento, el cual solo contiene las variables de Diferencia Absoluta y Diferencia Relativa. Utilizando la librería de Python Statsmodel [6] se entrena con Logit (función para entrenar regresión logística) utilizando todo el conjunto de datos.

Para contrastar resultados, se decide también entrenar otra regresión utilizando la técnica de Split-Test (Separación y Testeo) utilizando la función train_test_split de la librería Sklearn, con un porcentaje de test del 25%.

3) Validación

Utilizando el paquete "Metrics" de la librería Sklearn, se obtuvieron las métricas en test del modelo con Split-Test.

Para evaluar el desempeño del modelo realizado con todo el conjunto de datos, se utiliza K-Fold [7] de la librería Sklearn, con 5 divisiones de los datos, para obtener las métricas de Exactitud (Accuracy), Precisión, Exhaustividad (Recall) [8] y Valor-F (F1).

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Se valida también el Variance Inflation Factor (VIF) [9] utilizando la función variance_inflation_factor de Sklearn y se mide correlación utilizando la función Corr de la librería Pandas [10].

También, se grafica la curva ROC [11] (Receiver Operating Characteristic) para el modelo utilizando roc_auc_score de Sklearn.

III. Resultados

A. Aprendizaje No Supervisado

La Tabla 1 muestra los resultados obtenidos del Isolation Forest, donde se puede observar que un 5,5% de la muestra resulta anómala con diferencia de pesaje positiva.

Tabla 1

Total de Casos Analizados	205,993
Casos Anómalos con Diferencia Positiva	11,315
% Casos Anómalos con Diferencia Positiva	5.5%

A su vez, la Tabla 2 busca contrastar las medias y medianas de las variables utilizadas para el entrenamiento, entre la población anómala con diferencias positivas y la población total con diferencias positivas. Se puede ver cómo los valores de la clase anómala son entre 2,5 y 5 veces más grandes que los de la población total.

Tabla 2

Indicador	**Población con Diferencia de Pesaje Positiva**	
	Clase anómala	***Ambas clases***
Media de Diferencia Absoluta	7.45	2.98
Media de Ratio Diferencia	9.6	1.88
Mediana de Diferencia Absoluta	2.8	0.41
Mediana de Ratio Diferencia	3.18	1.19

En la Fig. 2 se puede observar una gráfica de dispersión de una muestra de 30,000 datos, donde se presentan el peso estimado y el peso medido, separando por la clase anómala (-1) y la clase no anómala (1). La diagonal corresponde a casos donde el pesaje estimado fue igual al medido, y hacia arriba de la diagonal se encuentran aquellos casos que generaron diferencias de pesaje positivas (peso medido mayor a peso

estimado), foco de este análisis. La Fig. 3 muestra la misma gráfica pero con las variables limitadas a valores menores a 10 kilogramos, lo que permite observar el cumplimiento de la hipótesis inicial: según de qué estimación de pesaje inicial se parta, el desvío de pesaje considerado "normal" irá aumentando conforme aumente el valor del pesaje estimado. Esto solo se cumple de la diagonal hacia arriba, ya que de la diagonal hacia abajo (estimaciones superiores al valor medido) hasta el rango de los 9 kg estimados es usual que los paquetes pesen menos de los estimado.

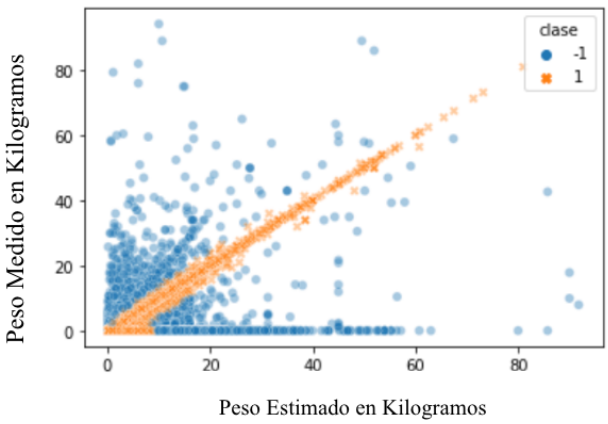

Fig. 2 *Gráfica de dispersión que contrasta peso medido versus peso estimado, separando por clase anómala y no anómala.*

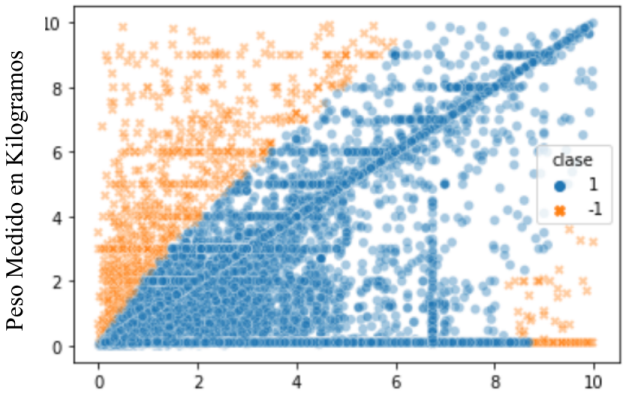

Fig. 3 *Gráfica de dispersión que contrasta pesos medidos versus pesos estimados menores a 10 kg, separando por clase anómala y no anómala.*

En la tabla 3 anexada a continuación, se muestran los valores finales luego de seleccionar al 50% de registros con puntaje de anomalía más alta (puntaje menor o igual a -0.15, Fig.1), de los cuales se revisaron en forma individual 30 casos y arrojaron una precisión de un 100%.

TABLA 3

Total de Casos Analizados	205,993
Casos Anómalos con Diferencia Positiva con Puntaje <= 0.15	5,658
% Casos Anómalos con Diferencia Positiva con Puntaje <= 0.15	2,74%

B. Aprendizaje Supervisado

La regresión logística entrenada con la muestra total, es decir, los 205,993 envíos clasificados en anómalos y no anómalos arroja los siguientes coeficientes β_i:

$$\beta_{\text{DIFERENCIA\_PESO\_ABSOLUTA}} = \mathbf{0.134}$$

$$\beta_{\text{DIFERENCIA\_PESO\_RELATIVA}} = \mathbf{1.778}$$

$$\beta_0 = \mathbf{0}$$

Con estos coeficientes se configura la fórmula final que arroja una probabilidad de anomalía en el pesaje de un envío utilizando las variables de diferencia de peso absoluta y relativa:

$$Marca_{pesoV1} = \frac{1}{1 + e^{-(0.134 * dif.peso.absoluta + 1.778 * dif.peso.relativa)}}$$

La marca final considera como anómalo todo aquel pesaje que supere una probabilidad de 0.99 a partir de esta fórmula.

La regresión tiene un P-Valor igual a 1, lo cual indica que los resultados son estadísticamente significativos.

Para analizar la correlación, se calcula el VIF, que da un valor de 1 para ambas variables (valores menores a 5 son aceptables). A la vez, se puede ver en la figura 4 la correlación entre variables utilizando un mapa de calor, con un valor de 0.22 entre la diferencia absoluta y la clase, y 0.14 para la relativa:

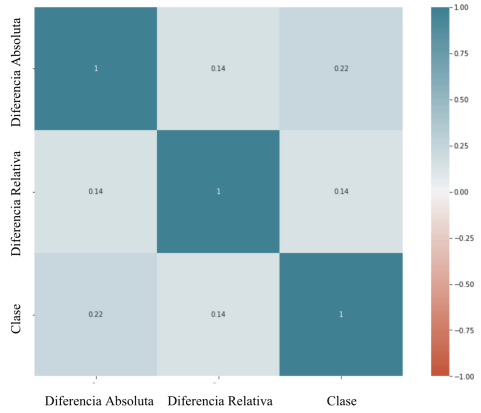

Fig.4 *Mapa de calor del set de datos utilizado para la regresión logística*

El K-Fold de 5 divisiones arroja las siguientes métricas de performance del modelo:

```
accuracy kfold:  0.9936745434219871
f1 kfold:  0.8839372378222506
precision kfold:  0.8992543421737091
recall kfold:  0.8696117948282254
```

En la fig. 5 se observa la curva ROC de dicho modelo, la cual contrasta el ratio de Falso Positivo y Verdadero Positivo a lo largo de los diferentes cortes en el puntaje de la regresión.

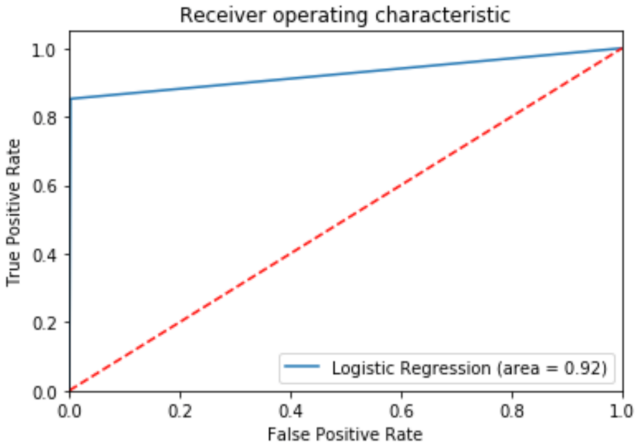

Fig.5 *Curva ROC Regresión Logística*

La regresión logística entrenada con la técnica de Split-Test, arroja los siguientes coeficientes, que varían tan solo en centésimas con respecto al entrenamiento con la muestra total:

$$\beta_{\text{DIFERENCIA\_PESO\_ABSOLUTA}} = \mathbf{0.134}$$

$$\beta_{\text{DIFERENCIA\_PESO\_RELATIVA}} = \mathbf{1.773}$$

$$\beta_0 = \mathbf{0}$$

Y sus métricas en test resultan:

```
accuracy_test:  0.9935921085846328
f1_test:  0.8854166666666667
precision_test:  0.9081196581196581
recall_test:  0.8638211382113821
```

Por último, se valida la presición de la regla final presentada al comienzo (Marca Peso V1, considerando una probabilidad mayor al 99%) en nuevos envíos y se toma una muestra de 45 casos. La precisión de esta muestra arroja un valor de 98%.

IV. Discusión

La forma de la curva ROC se explica por la composición de la muestra y el tipo de problema: el balance de la muestra con la que se entrena (en este caso 96.3% de clase 0 y 2.7% de clase 1) lleva a que los cortes del score que tengan un buen desempeño estén cercanos a esa proporción de desbalance (97%, 98% o 99%). Esto explica el porqué de una curva ROC que comienza casi en vertical, con una tasa de falso positivo casi constante en 0 hasta el corte que comienza a detectar las anomalías. Por las características del problema, y buscando para una primera versión de este análisis priorizar una buena precisión, se decide realizar el corte del puntaje en 99%.

El hecho de que los coeficientes de la regresión entrenada con la población total varíen en tal solo centésimas con respecto de la regresión entrenada utilizando la técnica de Split-Test, es una prueba de la robustez del modelo realizado.

También es destacable el valor del área bajo la curva ROC de 0.92 (valores cercanos a 1 representan un excelente desempeño) y las métricas generales del modelo, obteniendo valores superiores al 87% en todas ellas.

Como punto a mejorar, se podrían haber tomado muestras más grandes a la hora de revisar casos manualmente para validar las eficiencias en términos conceptuales. Sin embargo, se ha optado por esos tamaños ya que al estar en ambos modelos conservando los casos más anómalos, las diferencias absolutas y relativas eran considerables y de forma rápida se podía confirmar que lo enviado era distinto a lo declarado tan solo observando la diferencia entre los pesajes y contrastando tal información con lo que el remitente había declarado que enviaría.

Cabe destacar que, dado que la variable de peso medido se conoce a posteriori de haberse realizado el envío, esta marca no sirve para prevención de fraude, sino para identificar el falso negativo tiempo después. Aun así, este análisis resulta de gran valor para poder generalizar fácilmente una marca de fraude sin depender de validaciones manuales, y poder con ella entrenar modelos que sirvan para detección en-línea y prevención.

V. Conclusión

El método de aprendizaje híbrido arroja buenos resultados para la resolución de la problemática de encontrar una fórmula que fácilmente pueda diferenciar a los envíos con diferencias de pesaje significativas del resto de la población. Esto puede verse tanto en las métricas de performance del modelo, como también reflejado en las precisiones asociadas a las validaciones de casos en forma manual para ambas iteraciones.

Utilizar modelos de inteligencia artificial puede resultar común para las clásicas casuísticas de prevención de fraude, como lo pueden ser fraude con tarjeta de crédito, pero resulta sumamente innovador en el ámbito de la logística de comercio electrónico, ya que esta es una casuística que ha cobrado importancia en los últimos años gracias al crecimiento de las plataformas de compraventa digitales, y no se han encontrado artículos de divulgación sobre estas cuestiones aún. Este trabajo, que puede replicarse para problemáticas similares, siembra entonces la primera semilla en la temática, que se espera seguir profundizando en trabajos futuros.

References

[1] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.

[2] Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression (p. 536). New York: Springer-Verlag.

[3] Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information sciences, 557, 317-331.

[4] Yamanishi, K., & Takeuchi, J. I. (2001, August). Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 389-394).

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[6] Seabold, S., & Perktold, J. (2010, June). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (Vol. 57, No. 61, pp. 10-25080).

[7] Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The 'K'in K-fold cross validation. In 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (pp. 441-446). i6doc. com publ.

[8] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

[9] Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. Quality engineering, 14(3), 391-403.

[10] McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

[11] Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve?. Emergency Medicine Journal, 34(6), 357-359.