



Artículo

Detección Multivariante de Outliers

C.Rondan

Detección de outliers multivariados

C. Rondan¹²

1. Futura, Data & Analítica Avanzada

2. carlos.rondan.poma@gmail.com

Resumen

Este estudio compara tres métodos de detección de outliers en datos multivariados: la distancia de Mahalanobis, su versión robusta y el algoritmo Isolation Forest. Su efectividad se evalúa sobre un conjunto de datos real de fuga de clientes bancarios y, posteriormente, se entrenan modelos de clasificación para examinar el impacto de la limpieza de outliers. Los resultados muestran diferencias relevantes tanto en la detección como en su efecto sobre el desempeño de los modelos, y ofrecen una guía práctica para seleccionar el método de detección de outliers en análisis de datos avanzados.

Palabras clave: detección de outliers, análisis multivariante, distancia de Mahalanobis, Mahalanobis robusta, Isolation Forest

Introducción

La detección de outliers es un desafío crítico en el análisis de datos multivariados, con implicaciones significativas en diversas aplicaciones, desde la detección de fraude hasta el análisis de datos biomédicos. Los outliers pueden proporcionar insights valiosos o, por el contrario, pueden llevar a conclusiones erróneas si no se manejan adecuadamente. En este estudio, profundizamos en tres metodologías prominentes para la identificación de outliers: la tradicional distancia de Mahalanobis [1], su adaptación robusta [2, 3] y el método Isolation Forest [4]. Cada una de estas técnicas ofrece un enfoque único para abordar la detección de outliers, y su comparativa directa busca iluminar sus fortalezas y limitaciones en un contexto práctico. Este análisis no solo es crucial para entender la naturaleza y el comportamiento de los outliers en conjuntos de datos complejos, sino también para optimizar la precisión y la eficacia de los modelos predictivos en la ciencia de datos.

Metodología

La metodología de este estudio involucra la aplicación de tres técnicas distintas para la detección de outliers en un conjunto de datos multivariante y la evaluación de su impacto en la clasificación de datos.

Dataset

Para este estudio se utilizó el conjunto de datos «Churn Modelling» obtenido de Kaggle, una plataforma popular para competencias de ciencia de datos. Este conjunto de datos es ampliamente utilizado en el análisis predictivo de la rotación de clientes, especialmente en el sector bancario.

El conjunto de datos consta de 10,000 registros y 14 atributos, que incluyen información demográfica, de productos y de comportamiento de los clientes de un banco. Los campos clave son:

- **CustomerId:** identificador único del cliente.
- **Surname:** apellido del cliente.
- **CreditScore:** puntuación de crédito del cliente.
- **Geography:** país de residencia del cliente.
- **Gender:** género del cliente.
- **Age:** edad del cliente.
- **Tenure:** número de años que el cliente ha estado con el banco.
- **Balance:** saldo bancario del cliente.
- **NumOfProducts:** número de productos bancarios utilizados por el cliente.
- **HasCrCard:** indica si el cliente tiene tarjeta de crédito.
- **IsActiveMember:** indica si el cliente es un miembro activo.
- **EstimatedSalary:** salario estimado del cliente.
- **Exited:** indica si el cliente ha abandonado el banco (variable objetivo).

Técnicas de detección de outliers

Distancia de Mahalanobis

La distancia de Mahalanobis [1] es una medida de distancia entre un punto P y una distribución D , y se define como:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}$$

donde $\vec{x}$ es el vector de observación en un espacio multivariado, $\vec{\mu}$ es el vector de medias de la muestra, calculado para cada variable, y Σ es la matriz de covarianza de la muestra, que captura las covarianzas

entre cada par de variables. Esta distancia es útil para identificar outliers en datos multivariados, ya que tiene en cuenta la correlación entre las variables.

Distancia de Mahalanobis robusta

La distancia de Mahalanobis robusta mejora la resistencia de la distancia de Mahalanobis tradicional frente a los outliers [3]. Se basa en estimadores robustos para $\vec{\mu}$ y Σ . Una forma común de obtenerlos es mediante el algoritmo de determinante de covarianza mínimo (MCD) [2]. La distancia robusta se calcula de la misma manera que la tradicional, pero utilizando las estimaciones robustas de la media y la covarianza:

$$D_{MR}(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu}_{rob})^T \Sigma_{rob}^{-1} (\vec{x} - \vec{\mu}_{rob})}$$

```
1 import numpy as np
2 from sklearn.covariance import MinCovDet
3
4 # Minimum Covariance Determinant
5 robust_cov = MinCovDet().fit(data[
6     numerical_cols])
7
8 # Media y covarianza del estimador robusto
9 robust_mean = robust_cov.location_
10 robust_cov_matrix = robust_cov.covariance_
11 robust_cov_matrix_inv = np.linalg.inv(
12     robust_cov_matrix)
13
14 def mahalanobis(x, mean, inv_cov_matrix):
15     x_minus_mu = x - mean
16     return np.dot(np.dot(x_minus_mu,
17         inv_cov_matrix), x_minus_mu.T)
```

Código 1: Cálculo de la distancia de Mahalanobis robusta en Python.

Isolation Forest

Isolation Forest [4] es un algoritmo basado en árboles para la detección de outliers. A diferencia de los métodos de distancia, Isolation Forest aísla los outliers en lugar de medir distancias. Funciona construyendo árboles de decisión aleatorios: en cada árbol, un punto se aísla eligiendo un atributo al azar y seleccionando un valor de división entre el máximo y el mínimo de dicho atributo. El número de divisiones requeridas para aislar un punto se utiliza como medida de anomalía; menos divisiones indican un outlier potencial:

$$s(\vec{x}, n) = 2^{-\frac{E(h(\vec{x}))}{c(n)}}$$

donde $s(\vec{x}, n)$ es el puntaje de anomalía de la observación $\vec{x}$, $E(h(\vec{x}))$ es el número promedio de divisiones necesarias para aislar $\vec{x}$ y $c(n)$ es un factor de normalización que depende del tamaño de la muestra n .

Modelos de clasificación

Regresión logística

La regresión logística es un modelo estadístico que se utiliza para predecir la probabilidad de una variable categórica binaria. En el contexto de la clasificación,

estima la probabilidad de que una observación pertenezca a una de las dos categorías:

$$P(Y = 1 | \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}$$

donde $P(Y = 1 | \mathbf{x})$ es la probabilidad de que la observación pertenezca a la clase 1, $\beta_0, \beta_1, \dots, \beta_n$ son los coeficientes del modelo y $x_1, x_2, \dots, x_n$ son las características de la observación.

Máquinas de vectores de soporte (SVM)

Las SVM son un conjunto de métodos de aprendizaje supervisado utilizados para clasificación y regresión. En clasificación, las SVM buscan el hiperplano que mejor separa las clases en el espacio de características:

$$\mathbf{w}^T \mathbf{x} + b = 0$$

donde $\mathbf{w}$ es el vector de pesos, perpendicular al hiperplano, $\mathbf{x}$ es el vector de características y b es el término de sesgo.

Evaluación del rendimiento

Para evaluar el rendimiento de los modelos se utilizan la precisión, el recall, el puntaje F1 y el área bajo la curva ROC, definidos a continuación en términos de verdaderos positivos (VP), falsos positivos (FP) y falsos negativos (FN).

Precisión

Proporción de verdaderos positivos entre todas las predicciones positivas. Mide la exactitud del modelo al clasificar casos positivos:

$$\text{Precisión} = \frac{VP}{VP + FP}$$

Recall (sensibilidad)

Proporción de casos positivos reales que el modelo identifica correctamente:

$$\text{Recall} = \frac{VP}{VP + FN}$$

Puntaje F1

Media armónica de la precisión y el recall; resume el balance entre ambas en un único número:

$$F1 = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Área bajo la curva ROC (AUC)

El AUC mide la capacidad del modelo para distinguir entre clases. Un AUC de 1 indica un modelo perfecto, mientras que un AUC de 0.5 indica un modelo no mejor que el azar:

$$AUC = \int_0^1 ROC(t) dt$$

Resultados

Esta sección detalla los hallazgos obtenidos del análisis del conjunto de datos. Antes de presentar los resultados, es fundamental considerar algunos aspectos preliminares relativos al tratamiento de los datos.

Consideraciones

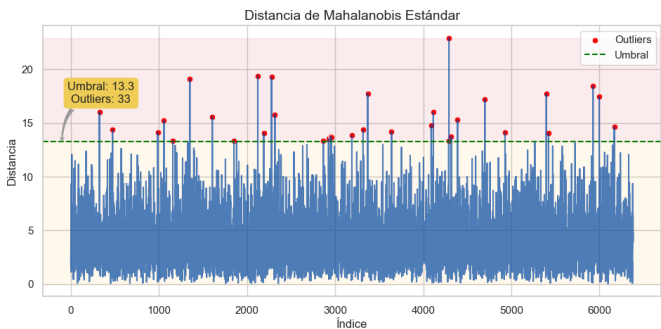
Los datos fueron segmentados según el saldo de la cuenta (Tabla 1).

Tabla 1: Distribución de la variable *Balance*.

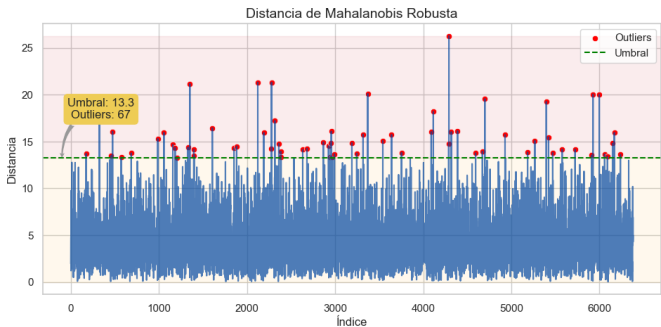
Balance (saldo)	Cantidad	Proporción (%)
Saldo positivo	6,383	63.83
Saldo nulo (0)	3,617	36.17

La considerable proporción de cuentas con saldo nulo justifica la división del conjunto de datos en dos grupos distintos: aquellos con saldo y aquellos sin saldo. Los criterios adoptados para el análisis fueron:

- a) Para facilitar la comprensión del análisis, nos centramos en el Grupo 1, es decir, las cuentas con saldo positivo, para realizar la limpieza de datos.
- b) La variable edad presenta una distribución asimétrica positiva; por ello se aplicó una transformación logarítmica antes de detectar anomalías con la distancia de Mahalanobis, tanto en su forma estándar (Figura 1a) como robusta (Figura 1b).
- c) Para la identificación de outliers mediante Isolation Forest se empleó PCA con el fin de visualizar mejor el patrón de los datos anómalos (Figura 1c).



(a) Detección de 33 outliers con la distancia de Mahalanobis estándar. La línea discontinua verde denota el umbral de identificación.



(b) Identificación de 67 outliers con la distancia de Mahalanobis robusta.

Se evaluaron dos modelos predictivos, regresión logística y SVM, aplicando las tres técnicas de preprocesamiento mencionadas. La efectividad de los modelos se midió mediante precisión, recall, puntaje F1 y AUC.

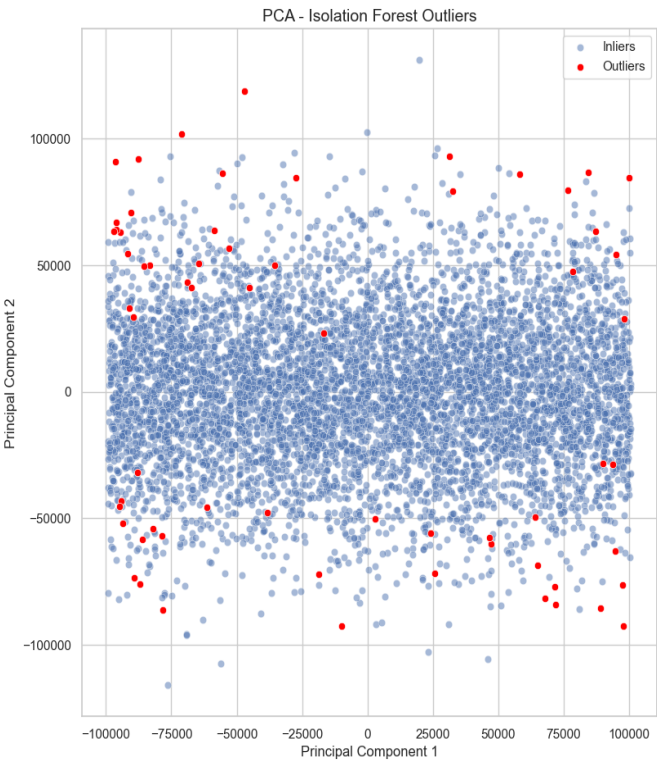
Impacto en la regresión logística

La distancia de Mahalanobis estándar ofreció un AUC de 0.752, mientras que la distancia de Mahalanobis robusta e Isolation Forest produjeron AUC de 0.774 y 0.772, respectivamente, evidenciando una ligera mejora en la discriminación del modelo (Figura 2). Al considerar la precisión, el recall y el puntaje F1, la distancia de Mahalanobis estándar presentó un rendimiento más consistente, siendo particularmente superior en recall y F1, lo que resalta su relevancia al predecir la fuga de clientes en el contexto bancario.

La superioridad de la distancia de Mahalanobis robusta en términos de AUC subraya la importancia de una selección meticulosa del método de tratamiento de outliers para optimizar modelos predictivos, especialmente cuando los datos pueden contener anomalías significativas.

Impacto en SVM

Con la distancia de Mahalanobis estándar, el modelo SVM logró una precisión de 55.21 %, un recall de 63.84 % y un puntaje F1 de 59.21 %. El AUC fue de 80.8 %, lo que denota una capacidad moderada de distinción de clases.



(c) Isolation Forest identifica aproximadamente 100 puntos como outliers en el espacio PCA. Los puntos rojos representan outliers claramente diferenciados del conjunto principal.

Figura 1: Comparativa del volumen de puntos clasificados como outliers por tres técnicas: distancia de Mahalanobis estándar y robusta, e Isolation Forest combinado con análisis de componentes principales (PCA).

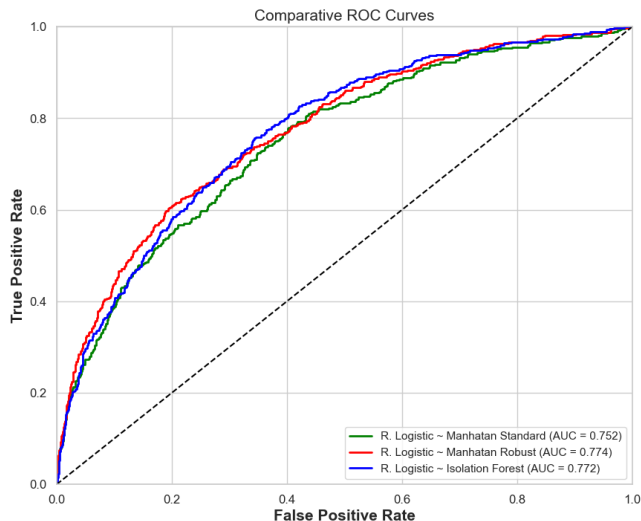


Figura 2: Curvas ROC de la regresión logística tras excluir outliers con distancia de Mahalanobis estándar (verde, AUC = 0.752), distancia de Mahalanobis robusta (rojo, AUC = 0.774) e Isolation Forest (azul, AUC = 0.772). La línea punteada denota clasificación aleatoria.

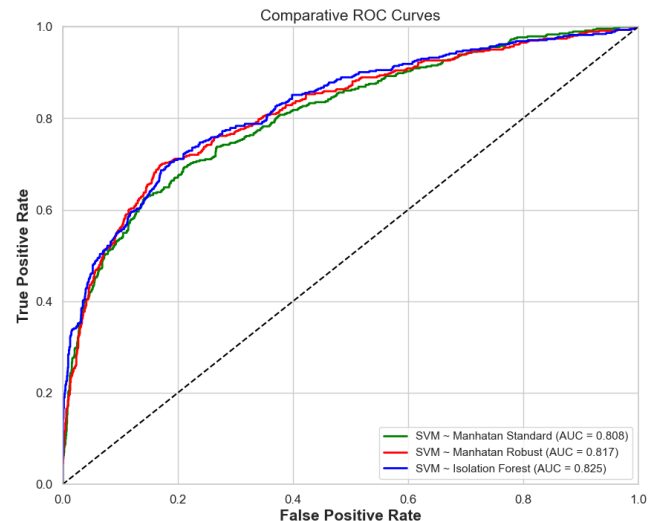


Figura 3: Curvas ROC de SVM tras excluir outliers con distancia de Mahalanobis estándar (verde, AUC = 0.808), distancia de Mahalanobis robusta (rojo, AUC = 0.817) e Isolation Forest (azul, AUC = 0.825). La línea punteada denota clasificación aleatoria.

El uso de Isolation Forest redujo la precisión a 50.54 %, mientras que el recall aumentó a 72.68 % y el puntaje F1 se incrementó ligeramente a 59.62 %. El AUC observado fue de 82.5 %, una mejora sobre la distancia de Mahalanobis estándar.

La distancia de Mahalanobis robusta equilibró la precisión y el recall, con valores de 55.76 % y 70.28 %, respectivamente, y alcanzó el puntaje F1 más alto, con 62.21 %. El AUC resultante fue de 81.8 %, lo que confirma su buena capacidad discriminativa (Figura 3).

Aunque las técnicas de preprocesamiento difieren en su rendimiento, la distancia de Mahalanobis robusta se destaca como la más equilibrada. La elección del método adecuado puede variar según el contexto y los objetivos de la aplicación, pero los resultados favorecen a la distancia de Mahalanobis robusta por su puntaje F1 superior y un AUC consistente.

Conclusiones

Este estudio ha investigado el impacto de distintas técnicas de preprocesamiento en la eficacia de los modelos de regresión logística y SVM para la predicción de fuga de clientes en el sector bancario. Los resultados indican que la selección adecuada del método de tratamiento de outliers es fundamental para optimizar los modelos predictivos.

En la regresión logística, si bien la presencia de outliers puede degradar el desempeño del modelo, técnicas como la distancia de Mahalanobis robusta mejoran su robustez y capacidad predictiva. En SVM, la distancia de Mahalanobis robusta emergió como la técnica superior, alcanzando el mejor balance entre precisión y sensibilidad, reflejado en un puntaje F1 más alto y un AUC consistente.

Las implicaciones prácticas son claras: aplicar técnicas de preprocesamiento apropiadas puede mejorar significati-

vamente la identificación de clientes con alta probabilidad de abandono, permitiendo a las instituciones bancarias intervenir de manera proactiva para retenerlos.

Referencias

1. Mahalanobis PC. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* 1936; 2:49-55
2. Rousseeuw PJ y Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999; 41:212-23. DOI: 10.1080/00401706.1999.10485670
3. Cabana E, Lillo RE y Laniado H. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers* 2021; 62:1583-609. DOI: 10.1007/s00362-019-01148-1
4. Liu FT, Ting KM y Zhou ZH. Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*. 2008 :413-22. DOI: 10.1109/ICDM.2008.17