UNIVERSIDAD DE GRANADA

INTELIGENCIA DE NEGOCIO

JANUARY 28, 2021

Detección de anomalías

David Alberto Martín Vela

davidmv1996@correo.ugr.es
Doble Grado Ingeniería Informática y Matemáticas

Curso 2020-2021

Contents

Descripción y análisis del problema

¿Alguna vez nos hemos preguntado como los bancos detectan fraudes o en las redes sociales cuando sospechan que un inicio de sesion es fraudulento? Esto se realiza principalmente a través del proceso denominado Detección de anomalías (*Anomaly Detection*).

Una anomalía, por definición, es algo que se desvía de lo que es estándar, normal o esperado. La detección de anomalías o la detección de valores atípicos es el proceso de identificación de elementos raros, observaciones, patrones, valores atípicos o anomalías que diferirán significativamente de los elementos o patrones normales. Las anomalías a veces se denominan valores atípicos, novedades, ruido, desviaciones o excepciones.

Se dice que la información es poder, y cada vez se tiene más en cuenta que esa información en la sociedad actual viene dada por los datos. Ahora bien, una gran cantidad de datos conlleva poder si se manejan correctamente.

Según un artículo de Forbes [For] el **61%** de los vendedores planean usar el aprendizaje automático como parte de su estrategia de datos, dado que todavía hay empresas que se están perdiendo esta ventaja con el resto de los competidores. Se remarca el hecho de que, entre otros, pueden ayudar a descubrir palabras clave y otros elementos de las campañas de marketing que no se están aprovechando, prevenir las violaciones y amenazas a la seguridad y detectar amenazas y problemas antes de que causen daños.

En el mismo estudio de Forbes se menciona el caso de la empresa de consultoría Accenture. Casi el **10%** de sus 25 millones de procesos anuales de líneas de gastos estaban siendo marcados por incumplimiento o fraude. Mientras que su sistema basado en reglas funcionaba hasta cierto punto, Accenture implementó un algoritmo de aprendizaje automático para optimizar el proceso. Se utilizó para reducir los falsos positivos, detectar los valores anómalos y crear una solución no supervisada.

Por supuesto es difícil saber que pasa exactamente con los datos, pero es ahí donde entra la inteligencia artificial. Herramientas como por ejemplo Google Analytics, Facebook Ads y Shopify no son capaces de abordar todos los datos en grandes empresas. Y es aquí donde un negocio debe apostar por mecanismos de detección de anomalías con algoritmos de aprendizaje automático.

Al principio para orientar esta práctica alternativa a un tema específico, las primeras dos opciones que se me venían a la cabeza eran dos áreas del conocimiento, una primera opción, **la detección de terremotos**, debido a la gran cantidad de los mismos ocurrido últimamente en Granada [Ter], y otra opción, detección de anomalias en el campo de la bolsa, debido otro acontecimiento reciente donde **GameStop se**

dispara en bolsa tras la compra de acciones de usuarios de Reddit [Red].

Planteamiento del problema

Un resumen rápido de esta situación es que un grupo de inversores decidió apostar fijo por la caída de las acciones de GameStop sin arriesgarse, pues esta tienda llevaba perdiendo en bolsa desde bastante tiempo principalmente por el auge de las ventas digitales frente al formato físico. De esta forma, acordaron vender sus acciones a un precio fijo dentro de un periodo de tiempo determinado dando por sentado que las acciones de la tienda seguirían cayendo y deseando así que las acciones cayeran para obtener beneficio.

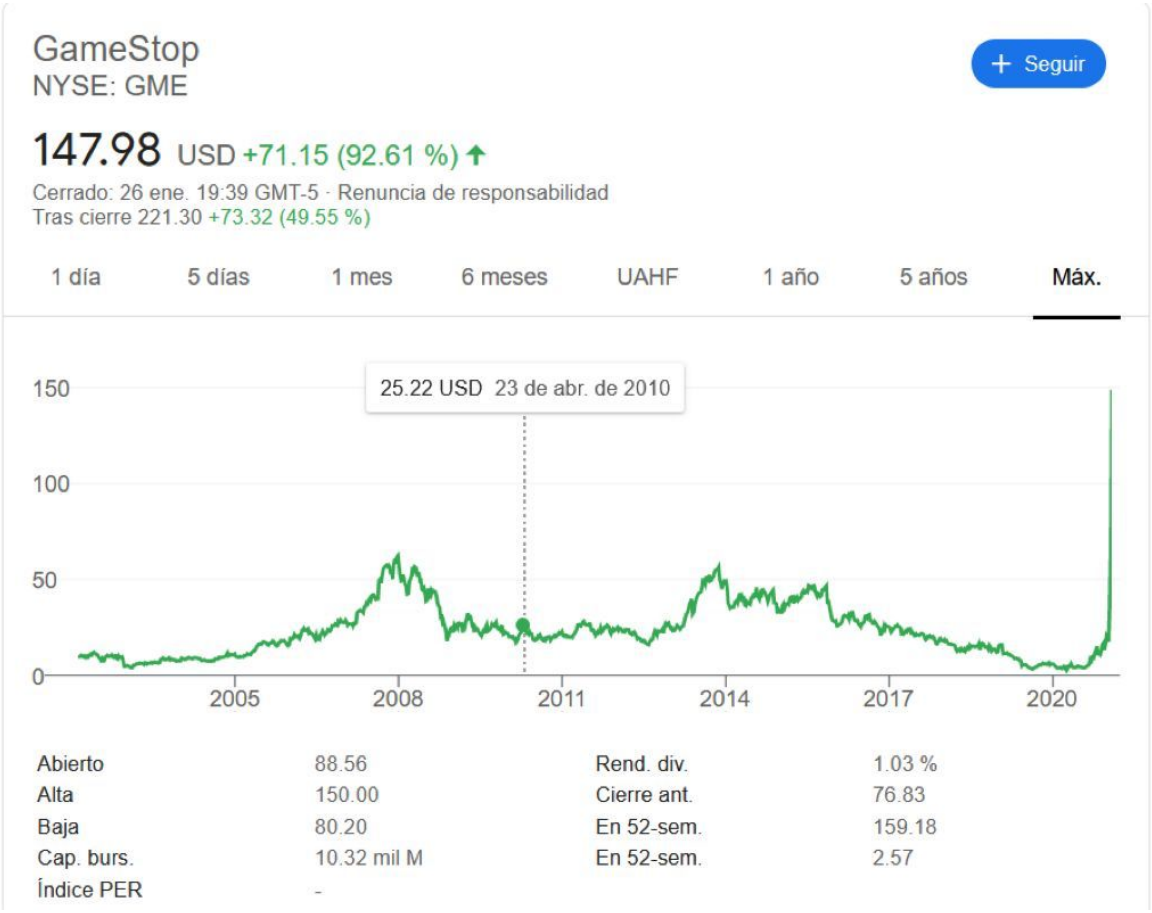

Figure 1: Gamstop stock from the last days

Sin embargo, esta estrategia llegó a oídos de los miembros del subreddit *wallstreetbets*, a quienes les pareció mal cómo se estaban comportando estos inversores. Decidieron tomar la decisión de comprar estas acciones baratas, inflando rápidamente el valor mucho más allá de lo que esperaban los administradores de fondos de cobertura, de manera que algunos miembros de Reddit han llegado a pagar miles de dólares con el objetivo de reventar los planes de los inversores mencionados anteriormente [1]. Mientras que *wallstreetbets* celebran la locura y dicen que no van a

vender y seguir comprando (incluso hay gente que compro hace año y medio una call con 50k y si la ejecuta se llevaría 36 millones ahora mismo) [Cal]. Ahora los compradores de Reddit deben calcular cuándo vender sus acciones para obtener beneficio, el cual podría ser de hasta 3.000 veces lo que compraron. Además, están explorando otros informes de **AMC** y **BlackBerry**, una cadena de cines estadounidense y una empresa de tecnología canadiense, para llevar a cabo acciones similares. **Otro tema interesante podría ser que debido a este fenómeno hay gente aplicando análisis de opiniones/sentimientos** en este foro de reddit para ver cuál puede ser el próximo objetivo pero esto ya se sale de nuestro tema elegido que es la detección de anomalias.

Para los datos, vamos a utilizar el paquete *pandas-datareader* [Pan] [1] donde extraeremos los datos trading volume data de Yahoo Finance. En nuestro caso, nuestras características de entrada serán una lista de símbolos ETF, los comentados anteriormente que corresponderán a Gamestop (GME), Blackberry (BB), Nokia (NOK) y AMC. Definiremos este entorno como nuestro ”mercado”, aunque en la práctica podríamos hacer que sea mucho, mucho más grande. Cogemos fechas desde hace 5 años hasta hoy (28 de Enero de 2020), donde han ocurrido los acontecimientos recientes. Mostramos imagenes del trading volume y del precio de cierre en la figuras [2] y [3] respectivamente.

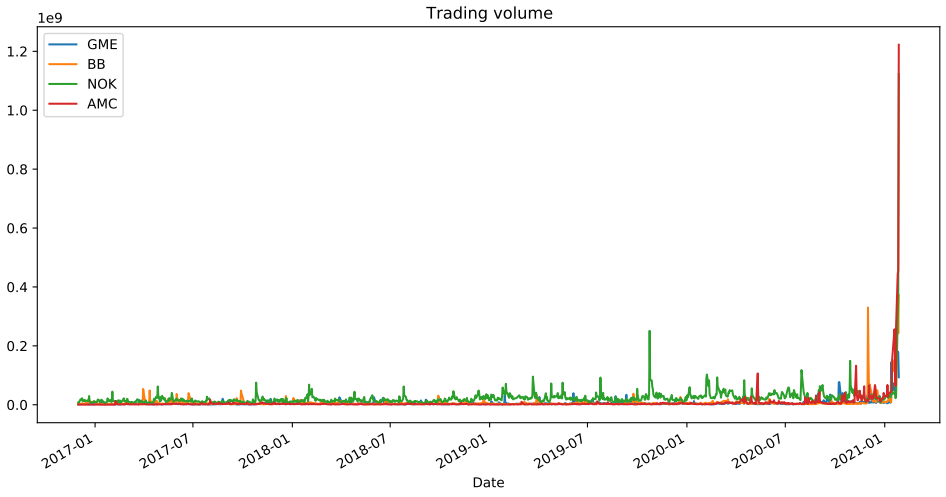

Figure 2: Trading volume data

[1]The Pandas datareader is a sub package that allows one to create a dataframe from various internet datasources, currently including: Yahoo! Finance. Google Finance.

Figure 3: Closing Price

En el comercio como en la vida, a menudo es extremadamente valioso determinar si el entorno actual es anómalo o no de alguna manera. Si las cosas están actuando ”normalmente”, sabemos que nuestras estrategias pueden operar de cierta manera. Por ejemplo, si nos encontramos en un entorno comercial normal, podríamos emplear una estrategia de volatilidad en corto. Por otro lado, si identificamos que estamos en un mercado anormalmente emocionante, podría ser necesario emplear una estrategia que haga exactamente lo contrario: buscar oportunidades para el comercio basado en el impulso, por ejemplo. En ese tipo de mercado, acortar la volatilidad podría ser muy peligroso. El objetivo será aplicar una serie de algoritmos para determinar cuándo el volumen de operaciones de nuestra lista de símbolos se encuentra en un estado anómalo. Esto podría significar, por ejemplo, que estamos detectando un pico en el volumen de operaciones.

Descripción de los algoritmos

Vamos a utilizar y comparar algoritmos de las bibliotecas PyOD [Pyo] y Scikit-Learn Outlier Detection [Skl], primero, vamos a comentar algunos de ellos. El módulo Python Outlier Detection (PyOD) facilita el modelado de detección de anomalías. Recopila una amplia gama de técnicas que van desde el aprendizaje supervisado hasta las técnicas de aprendizaje no supervisado. No es necesario probar todas las técnicas para encontrar anomalías. Dependiendo de los datos, algunas técnicas funcionan mejor que otras.

Isolation Forest

Un Isolation Forest *aísla* las observaciones seleccionando al azar una característica y luego seleccionar al azar un valor de división entre el máximo y mínimo valores mínimos de la característica seleccionada.
Dado que la partición recursiva puede ser representada por una estructura de árbol, el número de divisiones necesarias para aislar una muestra es equivalente a la longitud del camino desde el nodo raíz hasta el último nodo. Esta longitud de la ruta, promediada sobre un bosque de árboles tan aleatorios, es una medida de la normalidad y nuestra función de decisión. La división aleatoria produce caminos notablemente más cortos para las anomalías.
Por lo tanto, cuando un bosque de árboles al azar produce colectivamente entre los nodos longitudes de camino más cortas para muestras particulares, es muy probable que sean anomalías.
La idea de identificar una observación normal frente a una anormal puede observarse en un punto normal, requiere que se identifiquen más particiones que una anomalía. Al igual que con otros métodos de detección de anomalías, se requiere un punto anómalo para la toma de decisiones. En el caso del Isolation Forest, se define como:

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

donde $h(x)$ es la longitud del camino de observación x, $c(n)$ es la longitud media del camino de búsqueda fallida en un Árbol de Búsqueda Binario y n es el número de nodos externos.
A cada observación se le da una puntuación de la anomalía y se puede tomar la siguiente decisión en base a ella:

- Una puntuación cercana a 1 indica anomalías.

- Una puntuación mucho menor que 0.5 indica observaciones normales.
- Si todas las puntuaciones se acercan a 0.5, entonces la muestra completa no parece tener anomalías claramente diferenciadas.

Estudio experimental

Planteamiento de futuro

Referencias

[For] *How To Apply Anomaly Detection And Reap These Three Benefits*. visited on 27-01-2021. URL: https://www.forbes.com/sites/forbesagencycouncil/2020/02/03/how-to-apply-anomaly-detection-and-reap-these-three-benefits/#5a289a5114bf.

[Ter] *La noche de los 40 terremotos en Granada*. URL: https://elpais.com/espana/2021-01-27/la-noche-de-los-40-terremotos-en-granada-ibamos-buscando-espacios-sin-arboles-ni-edificios.html.

[Pan] *Pandas Datareader*. URL: https://pandas-datareader.readthedocs.io/en/latest/.

[Pyo] *PyOD*. URL: https://github.com/yzhao062/pyod/.

[Red] *Reddit group blew up GameStop stock*. URL: https://edition.cnn.com/2021/01/27/investing/gamestop-reddit-stock/index.html.

[Cal] *Reddit user*. URL: https://www.reddit.com/r/wallstreetbets/comments/l5nphz/gme_yolo_update_jan_26_2021/.

[Skl] *Scikit-Learn Outlier Detection*. URL: https://scikit-learn.org/stable/modules/outlier_detection.html.