Apoyo en la Compra y Venta de Acciones de la Bolsa de Valores Estadounidense Utilizando Técnicas de Aprendizaje por Refuerzo

Elkin Jadier Narvaez Paz
Pontificia Universidad Javeriana Cali
elkinnarvaez@javerianacali.edu.co

20 de noviembre del 2023

Abstract

This research explored the development of a reinforcement learning-based trading agent, aiming to evaluate its performance in comparison to the *buy-and-hold* strategy within the dynamic environemnt of financial markets. The core problem at the heart of this investigation was to discern whether an agent trained through reinforcement learning techniques could effectively navigate the complexities of trading, marked by market volatility, evolving trends, and financial uncertainties. Our approach involved a deliberate and iterative progression of experiments, beginning with preliminary phases designed to extract qualitative insights and identify trends. These early experiments were constrained in terms of computational resources and training duration but proved instrumental in steering us toward optimal hyperparameters and training configurations for the last experiment. This last experiment, characterized by an extensive training period, stands as the focal point of our findings, showcasing the culmination of our efforts. Within this extended training, the agent exhibited a remarkable capacity to adapt to dynamic market conditions, resulting in performance that competes favorably with the *buy-and-hold* strategy. These outcomes underscore the agent's adaptability and learning capabilities within the context of trading, revealing its potential for practical application in real financial markets. In conclusion, this research sheds light on the promise of reinforcement learning in trading, emphasizing the importance of a systematic approach to experimentation and paving the way for future refinements to enhance the model's robustness for real-world financial applications.

Keywords: Reinforcement learning, trading agent, buy-and-hold strategy, technical analysis indicators, stock market, financial positions, deep learning

Resumen

Este proyecto de investigación se adentró en el desarrollo de un agente de trading basado en aprendizaje por refuerzo, con el objetivo de evaluar su desempeño en comparación con la estrategia *buy-and-hold* en el dinámico entorno de los mercados financieros. El problema central de esta investigación era evaluar si un agente entrenado mediante técnicas de aprendizaje por refuerzo podía navegar eficazmente por las complejidades de la compra y venta de activos, marcadas por la volatilidad del mercado, tendencias cambiantes e incertidumbres financieras. Nuestro enfoque implicó una progresión deliberada e iterativa de experimentos, comenzando con fases preliminares diseñadas para extraer información cualitativa e identificar tendencias. Estos primeros experimentos estaban limitados en términos de recursos computacionales y duración del entrenamiento, pero fueron fundamentales para dirigirnos hacia hiperparámetros óptimos y configuraciones de entrenamiento para el último experimento. Este último experimento, caracterizado por un período de entrenamiento extenso, se destaca como el punto central de nuestros hallazgos, exhibiendo la culminación de nuestros esfuerzos. Dentro de este entrenamiento extendido, el agente demostró una notable capacidad para adaptarse a las dinámicas cambiantes del mercado, lo que se tradujo en un desempeño que compite favorablemente con la estrategia *buy-and-hold*. Estos resultados destacan la adaptabilidad y las capacidades de aprendizaje del agente en el contexto de compra y venta de activos, revelando su potencial para su aplicación práctica en los mercados financieros reales. En conclusión, esta investigación arroja luz sobre la promesa del aprendizaje por refuerzo en el trading, enfatizando la importancia de un enfoque sistemático para la experimentación y dejando el camino para futuros refinamientos destinados a mejorar la robustez del modelo para aplicaciones financieras del mundo real.

Palabras clave: Aprendizaje por refuerzo, agente de trading, estrategia buy-and-hold, indica- dores técnicos, posiciones financieras, aprendizaje profundo

1. Introducción

La inversión en acciones ha sido reconocida durante mucho tiempo como un poderoso vehículo para la acumulación de riqueza, ofreciendo a los inversionistas una perspectiva de retornos sustanciales a largo plazo. La parte atractiva de las inversiones en el mercado de valores radica en la oportunidad que brinda a las personas para hacer crecer su capital, asegurar su futuro financiero y cumplir sus aspiraciones de inversión. El rendimiento histórico del mercado de valores corrobora su papel como base fundamental de muchas carteras de inversión. Con el potencial de la creación de riqueza y la seguridad financiera, actúa como un motor clave para el bienestar financiero individual.

Sin embargo, detrás de la promesa de ganancias financieras a largo plazo, invertir en acciones puede presentar un gran desafío, especialmente cuando se trata de inversiones a corto plazo. Las acciones son conocidas por su inherente volatilidad, caracterizada por fluctuaciones rápidas y a menudo impredecibles en los precios en periodos cortos. Por ejemplo, considere los abruptos cambios de precios exhibidos por gigantes tecnológicos como Amazon y Tesla, o la inestabilidad experimentada por las compañías farmacéuticas mientras navegan por los resultados de ensayos clínicos cruciales. Estos movimientos abruptos de precios pueden provocar incertidumbre y riesgo significativo en el panorama de inversión, generando preguntas sobre la idoneidad de la participación a corto plazo en el mercado de valores.

Aún así, las implicaciones de la volatilidad del mercado de valores se extienden mucho más allá del ámbito de los inversionistas individuales. Las repercusiones afectan a economías enteras, afectando a millones de personas y sectores más allá del alcance directo de los mercados financieros. El bienestar de los ciudadanos de una nación, la estabilidad de su moneda e incluso el progreso de sus iniciativas de desarrollo están relacionados con el desempeño de sus mercados de valores. En consecuencia, mitigar los desafíos planteados por la volatilidad del mercado de valores no es solo una preocupación para los inversionistas, sino una cuestión de importancia nacional.

En respuesta a estos desafíos, la aparición de agentes de compra y venta de activos impulsados por tecnologías de aprendizaje por refuerzo presenta una oportunidad transformadora. Estos sofisticados agentes, diseñados con capacidades avanzadas de inteligencia artificial, poseen el potencial de descifrar las complejidades de la dinámica del mercado de valores. Pueden ofrecer un apoyo a los inversionistas para tomar decisiones más informadas, estratégicas y basadas en datos. Al aprovechar el análisis colectivo incorporado en datos históricos del mercado y adaptarse dinámicamente a las condiciones del mercado en tiempo real, estos agentes están especialmente preparados para navegar por la inestabilidad a corto plazo de los mercados de valores. Al hacerlo, permiten a los inversionistas aprovechar oportunidades mientras gestionan prudentemente los riesgos asociados con la volatilidad de los cambios abruptos en los precios de los activos. En las siguientes secciones, profundizamos en el mundo de las inversiones en el mercado de valores, examinando los desafíos planteados por la volatilidad a corto plazo y explorando el papel fundamental de los agentes de trading basados en aprendizaje por refuerzo en promover la estabilidad financiera y la prosperidad económica.

2. Fundametación teórica

2.1. Aprendizaje por refuerzo

De acuerdo con Fischer [1], la mayor parte de la investigación con respecto a la aplicación de aprendizaje automático en la predicción del mercado de acciones se ha dedicado al aprendizaje supervisado. Entre los métodos más usados para esta tarea se encuentran: máquinas de vectores de soporte [2, 3], algoritmos basados en árboles [4, 5, 6, 7] y clasificación de vecinos más cercanos [8]. Según Fischer [1], la idea general de todos estos trabajos se basa en entrenar un modelo predictivo basado en datos históricos para prever el cambio de precio de una acción usando un conjunto de variables explicativas (características), para luego alimentar un modulo de trading que deriva la acción que se debe realizar, e.g., comprar el activo financiero en caso de que la predicción sobrepase un cierto umbral. Sin embargo, Fischer señala que este enfoque tiene distintas limitaciones que pueden llevar a que se tenga un rendimiento subóptimo [9].

La naturaleza interactiva del aprendizaje por refuerzo hace que sea particularmente adecuado para el dominio de compra y venta de activos dentro de un portafolio de inversión. El aprendizaje por refuerzo trata de modelar un aprendizaje dirigido por un agente que interactúa con un ambiente, generalmente estocástico, del que tiene una información incompleta o desconocida. El objetivo es que el agente pueda tomar decisiones que permitan lograr un objetivo a largo plazo al aprender el valor de estados y acciones basadas en una señal de recompensa. Este aprendizaje continuo por medio de recompensas deriva una política que codifica las reglas de comportamiento del agente asignando estados a acciones. Este tipo de aprendizaje difiere del aprendizaje supervisado ya que optimiza el comportamiento de un agente basado en experiencia de intento y error por medio de una señal de recompensa, en lugar de generalizar usando ejemplos representativos correctamente etiquetados de la variable objetivo. De igual

manera, el aprendizaje por refuerzo no solo se limita en hacer predicciones, sino que también incluye acciones y sus consecuencias dentro del ambiente.

2.2. Q-learning

Q-learning fue un algoritmo desarrollado por Chris Watkins en su tesis doctoral [10] (1989). Un proceso de decisión de Markov (MDP, por sus siglas en inglés) es un modelo matemático utilizado en aprendizaje por refuerzo, en el cual un agente toma decisiones secuenciales en un entorno incierto. En el trabajo realizado por Chris Watkins se introduce la programación dinámica incremental para aprender a controlar un MDP sin necesidad de modelar las matrices de transiciones y recompensas, eliminando la limitación que los métodos de programación dinámica tradicionales presentan.

El algoritmo de Q-learning mejora progresivamente la función de valor estado-acción después de una inicialización aleatoria por un número dado de episodios. En cada paso que el agente realiza, selecciona una acción basada en la política ϵ-greedy, y usa una taza de aprendizaje α para actualizar la función de valor de la siguiente manera:

$$Q(S_t, A_t) \longleftarrow Q(S_t, A_t) + \alpha(R_t + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t)) \tag{1}$$

En la ecuación anterior γ representa el factor de descuento aplicado a la función de valor para el siguiente estado y α es la tasa de aprendizaje. El usar la estimación de la función de valor para mejorar su misma estimación se conoce como *bootstrapping* [11].

2.3. Aprendizaje por refuerzo profundo

La función de valor estado-acción descrita por la ecuación 1 nos permite calcular la recompensa a largo plazo en un ambiente con estados y acciones discretas. Sin embargo, aunque las acciones del agente dentro de nuestro ambiente de compra y venta de activos son discretas, el conjunto de estados es continuo, lo cual nos impide hacer uso esta función en su versión original. De este modo, es necesario modificar la función de valor teniendo estados continuos y manteniendo las acciones discretas.

El hecho de que los estados sean continuos implica que no podemos simplemente tabular en un arreglo parejas estado-acción con su respectivo valor. En su lugar, necesitamos aproximar q^* usando una red neuronal (NN), generando así una *deep Q-network*.

2.4. Q-learning profundo

El algoritmo Deep Q-learning es una adaptación del algoritmo Q-learning que nos permite hacer uso de estados continuos en nuestro ambiente. Su objetivo es estimar el valor de las acciones disponibles para un estado dado usando una red neuronal profunda.

El algoritmo Deep Q-learning aproxima la función de valor aprendiendo un conjunto de pesos θ de una red Q profunda (DQN) que relaciona estados con acciones de tal forma que

$$q(s, a, \theta) \approx q^*(s, a)$$

El algoritmo aplica gradiente descendente estocástico basado en una función de pérdida que calcula la diferencia al cuadrado entre el objetivo estimado por la DQN y la predicción del valor Q de la pareja estado-acción actual. La ecuación 3 expresa esta función de pérdida.

$$y_i = \mathbb{E}[r + \gamma max_{a'} Q(s', a'; \theta_{i-1} | s, a)] \tag{2}$$

$$L_i(\theta_i) = (y_i - Q(s, a; \theta))^2 \tag{3}$$

Podemos notar que en la función de pérdida tanto el objetivo como la estimación actual dependen de los pesos de la DQN. Esto es precisamente lo que lo diferencia del aprendizaje supervisado, donde los objetivos son fijos antes del entrenamiento.

2.5. Análisis técnico

Las características originales del conjunto de datos no le dan suficiente información a los agentes para que puedan tener una visión más amplia durante cada paso que realizan en el ambiente, teniendo en cuenta que durante cada paso solo tendrán acceso a un registro del conjunto de datos, el cual representará *su estado actual*. Por esta razón, para el desarrollo de este proyecto se ha decido hacer uso de análisis técnico, específicamente centrándose en **indicadores técnicos**. Tal y como es expuesto por S. Donadio y S. Ghosh [12], los indicadores técnicos se basan en cálculos matemáticos para pronosticar la dirección del mercado financiero. De esta manera, cuándo el agente se encuentre en un punto dado en el ambiente (i.e., su estado actual), va a tener características que lo ayuden a tener un pronóstico más amplio para la toma de decisiones. Existe una larga lista de indicadores técnicos, categorizados en 8 tipos (e.g., volatilidad, volumen, momentum, ...) y 4 clases (e.g., acumulativos, índices, osciladores y de superposición). Sin embargo, en este trabajo solo se usó un subconjunto de estos indicadores, entre los cuales se incluyen indicadores de *momentum*, *volatilidad* y *volumen*. Mas precisamente, los indicadores usados fueron: Relative Strength Indicator (RSI), Moving Average Convergence/Divergence (MACD), Average True Range (ATR), Slow Stochastic (STOCH), Ultimate Oscillator (ULTOSC), Bollinger Bands (BBANDS) y On Balance Volume (OBV).

3. Resultados

Los activos usados como medio de evaluación en este trabajo fueron Apple Inc, Microsoft, Amazon Inc y Pepsico Inc. Las figuras 1, 3, 2 y 4 muestran, para cada uno de estos activos, la media móvil sobre los últimos 100 episodios de los valores de retorno acumulativo para 960 periodos de entrenamiento, así como también la media móvil de la proporción de los últimos 100 episodios en los que los agentes superaron a la estrategia *buy-and-hold*.

De manera general se puede decir que los agentes lograron capturar muy bien las tendencias del mercado para los activos trabajados, lo cual se tradujo en una toma de decisiones inteligente que lo llevó, en algunos casos, a sobrepasar las retribuciones que se lograron con una estrategia *buy-and-hold*.

Figura 1: Media móvil de rendimientos anuales AAPL

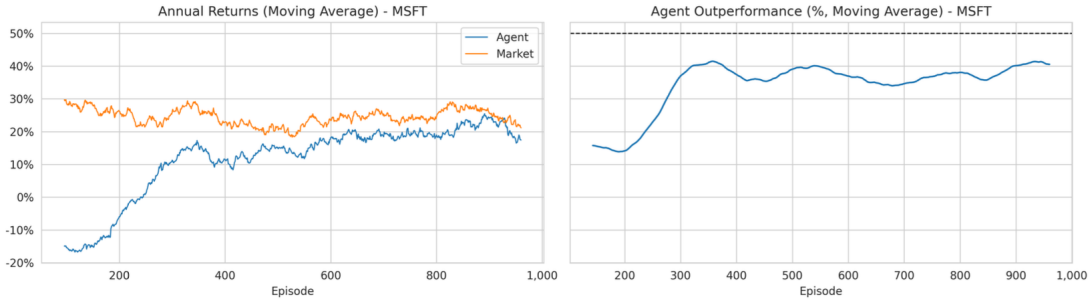

Figura 2: Media móvil de rendimientos anuales MSFT

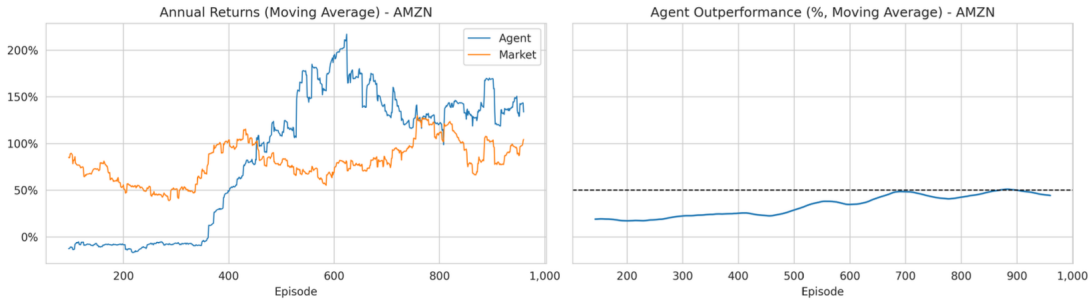

Figura 3: Media móvil de rendimientos anuales AMZN

Figura 4: Media móvil de rendimientos anuales PEP

4. Discusión y conclusiones

En este proyecto logramos desarrollar un agente de trading utilizando un marco de aprendizaje por refuerzo, específicamente implementando Double-Deep Q Learning. Nuestro objetivo principal fue evaluar su capacidad para lograr resultados competitivos en comparación con una estrategia de *buy-and-hold* en el contexto de compra y venta de activos. A lo largo de este trabajo surgieron varias observaciones e ideas clave.

El ambiente de trading fue diseñado para responder a algunas complejidades del trading del mundo real, aunque con ciertas simplificaciones. El agente tenía la capacidad de tomar decisiones en cada paso de tiempo, incluyendo comprar, vender en corto o mantener su capital. El espacio de estados se enriqueció con indicadores técnicos y retribuciones en diferentes ventanas de tiempo. Sin embargo, es importante reconocer que nuestro ambiente tenía algunas limitaciones, como la necesidad de utilizar el capital completo para comprar o vender en corto. Además, nuestra función de recompensa solo consideraba los retornos generados por las acciones del agente, sin tener en cuenta los riesgos inherentes asociados con las ventas en corto.

A lo largo del proyecto, realizamos una evaluación exhaustiva del rendimiento de nuestro agente. Inicialmente, los resultados eran consistentes con el desafió de superar una estrategia de *buy-and-hold*. Esto destacó la resistencia de este enfoque, que a menudo es preferido para inversiones a largo plazo por su capacidad de generar rendimientos constantes durante periodos prolongados. Sin embargo, a medida que el agente recibió un amplio entrenamiento a través 1000 episodios (en el último experimento), comenzó a demostrar un progreso notable. Importante destacar que logró resultados competitivos e incluso superó, en algunos casos, los rendimientos de la estrategia *buy-and-hold*. Esto demuestra el potencial de nuestro enfoque de aprendizaje por refuerzo.

Las implicaciones de este logro son significativas. La estrategia de *buy-and-hold*, particularmente en el contexto de la inversión a largo plazo, ha demostrado consistentemente su efectividad al capitalizar el crecimiento inherente de los mercados financieros con el tiempo. La capacidad de nuestro agente de aprendizaje por refuerzo para lograr resultados competitivos destaca su adaptabilidad y capacidad de aprendizaje en respuesta a las dinámicas del mercado, incluso frente a una estrategia a largo plazo establecida y exitosa. Sin embargo, es crucial reconocer que el modelo actual no replica completamente las complejidades de la inversión a largo plazo, como las consideraciones de dividendos e impuestos. Estos factores deben incorporarse en futuras iteraciones para proporcionar una evaluación más completa del rendimiento.

En el futuro, los próximos pasos del proyecto deben centrarse en refinar el modelo e incorporar estas complejidades del mundo real. Las estrategias de gestión de riesgos deben integrarse para tener en cuenta los desafíos únicos asociados con las ventas en corto. Además, el ajuste de hiper-parámetros y la exploración de algoritmos alternativos de aprendizaje por refuerzo pueden mejorar aún más el rendimiento del agente. Realizar análisis de sensibilidad en diferentes condiciones del mercado y evaluar su robustez con datos fuera de muestra será fundamental para medir su eficacia en la aplicación práctica en los mercados financieros reales. En conclusión, nuestro proyecto demuestra el potencial del aprendizaje por refuerzo en el contexto de trading, pero destaca la importancia de la investigación y el desarrollo continuo para hacer que el modelo sea más robusto y aplicable a escenarios del mundo real.

5. Trabajo futuro

5.1. Incorporación de dimensionamiento de posiciones

En futuras iteraciones de esta investigación, una mejora fundamental a considerar es la incorporación de dimensionamiento de posiciones dentro del entorno de trading. Actualmente, las acciones del agente de comprar y vender en corto se limitan a utilizar todo el capital. Para proporcionar estrategias de trading más realistas y flexibles, podemos modificar el ambiente para permitir que el agente asigne una parte del capital, en lugar de la totalidad, al tomar estas acciones. Este ajuste permitiría al agente emplear tamaños de posición diversos, reflejando la práctica del mundo real de asignación parcial de capital. La implementación de dimensionamiento de posiciones no solo mejoraría el realismo del entorno, sino que también permitiría al agente explorar un espectro más amplio de estrategias de trading, potencialmente conduciendo a un mejor rendimiento y gestión del riesgo.

5.2. Integración del riesgo en la función de recompensa

Vender en corto conlleva un gran riesgo porque no existe un límite superior para cuánto puede aumentar el precio de un activo, lo que podría resultar en pérdidas ilimitadas si el precio sube significativamente.

Con el fin de crear una función de recompensa más completa, el trabajo futuro debe centrarse en incluir el riesgo asociado con la venta en corto. La función de recompensa actual depende principalmente de los rendimientos generados por las acciones tomadas en el paso anterior, pasando por alto los riesgos que se puedan generar por los cambios abruptos que puedan ocurrir en el precio de un activo. Para abordar esta limitación, podemos introducir medidas de riesgo como el Valor en Riesgo (VaR) o el Valor en Riesgo Condicional (CVaR) en la función de recompensa. Al cuantificar el riesgo potencial de las posiciones cortas, podemos alentar al agente a adoptar estrategias conscientes del riesgo, fomentando así un enfoque de trading más prudente y equilibrado.

5.3. Registrar la cantidad real de posiciones

Actualmente, el entorno rastrea el NAV del capital, lo que proporciona información valiosa sobre el rendimiento general de la cartera. Sin embargo, una ampliación valiosa del entorno implicaría el seguimiento de la cantidad real de activos en posesión, lo que permitiría una gestión de posiciones más precisa. Para lograr esto, podemos implementar un sistema de registro que supervise la cantidad de cada activo en posesión del agente, además del NAV. Esta expansión del entorno facilitaría un análisis más detallado del comportamiento de trading, incluida la capacidad de medir asignaciones de activos y liquidez. Además, permitiría al agente ejecutar decisiones de trading más detalladas, como ajustar el tamaño de posiciones individuales, lo cual es un elemento crucial en la gestión de carteras.

Incorporar estas mejoras en el entorno de trading y el marco de aprendizaje por refuerzo representa vías prometedoras para investigaciones futuras. Al abordar el dimensionamiento de posiciones, las consideraciones de riesgo y el seguimiento de las cantidades reales de posiciones, podemos elevar aún más el realismo y la eficacia del agente de trading, convirtiéndolo en una herramienta más valiosa tanto para la exploración académica como para la aplicación práctica en los mercados financieros reales.

Referencias

[1] Fisher T. Reinforcement learning in financial markets - a survey. 2018.

[2] Kim J. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2):307–319, 2003.

[3] Huang W, Nakamori Y, and Wang S. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10):2513–2522, 2005.

[4] Kumar M and Thenmozhi M. Forecasting stock index movement: A comparison of support vector machines and random forest. *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*, 2006.

[5] Booth A, Gerding E, and Mcgroarty F. Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications*, 41(8):3651–3661, 2014.

[6] Moritz B and Zimmermann T. Deep conditional portfolio sorts: The relation between past and future stock returns. *LMU Munich and Harvard University*, 2014.

[7] Krauss C, Do X, and Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702, 2017.

[8] Teixeira L and Oliveira A. A method for automatic stock trading combining technical analysis and nearest neighbor classification. *Expert Systems with Applications*, 37(10):6885–6890, 2010.

[9] Moody J, Wu L, Liao Y, and Saffell M. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(56):441–470, 1998.

[10] Watkins C. Learning from delayed rewards, 1989.

[11] Mohan S. N-step bootstrapping in reinforcement learning, 2023.

[12] Donadio S and Ghosh S. *Learn Algorithmic Trading*. Packt Publishing, 2019.