CAPÍTULO IV

¿Es posible batir a los mercados financieros usando herramientas de *big data* y de *machine learning*?

Francisco J. Nogales*

Una de las premisas de la teoría económica financiera es la hipótesis del mercado eficiente: los precios de los activos financieros incorporan toda la información pública disponible. El consenso actual es que los mercados financieros no son 100 % eficientes en todo momento, por lo que siempre existen pequeñas anomalías que pueden ser explotadas para generar rentabilidades por encima del mercado.

En este trabajo se revisarán las herramientas de *machine learning* propuestas en los últimos años, en el contexto del *big data* financiero, para explotar las mencionadas anomalías, y desarrollar estrategias automáticas de inversión para tratar de batir a los mercados.

Palabras clave: aprendizaje automático, estrategias de inversión, grandes volúmenes de datos, interacciones y no linealidades.

* El autor agradece el apoyo financiero del Gobierno de España a través del proyecto de investigación PID2020-116694GB-I00.

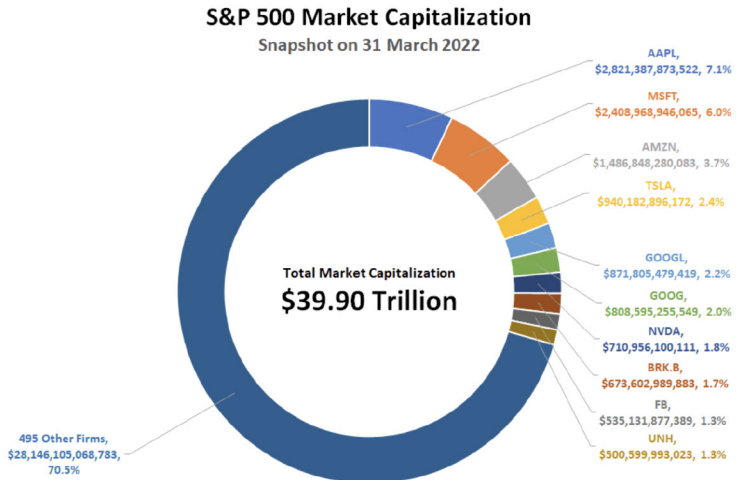
1. INTRODUCCIÓN Y MOTIVACIÓN

Empecemos definiendo en primer lugar qué se entiende por un mercado, y qué significa batir al mercado. La definición estándar de mercado viene dada por un índice financiero, que es una media ponderada por la capitalización bursátil de cada una de las compañías que componen el índice. El índice financiero más conocido en el mundo es el Standard & Poor's S&P 500.

A continuación, en la figura 1, se muestra la composición del S&P 500 a fecha 31 de marzo de 2022. Se puede observar cómo el índice está fuertemente concentrado en pocas empresas. En particular, diez empresas representan el 30 % del índice.

Figura 1.

Porcentaje de capitalización de las empresas componentes del S&P 500



Fuente: Standard & Poor's, Slickcharts.

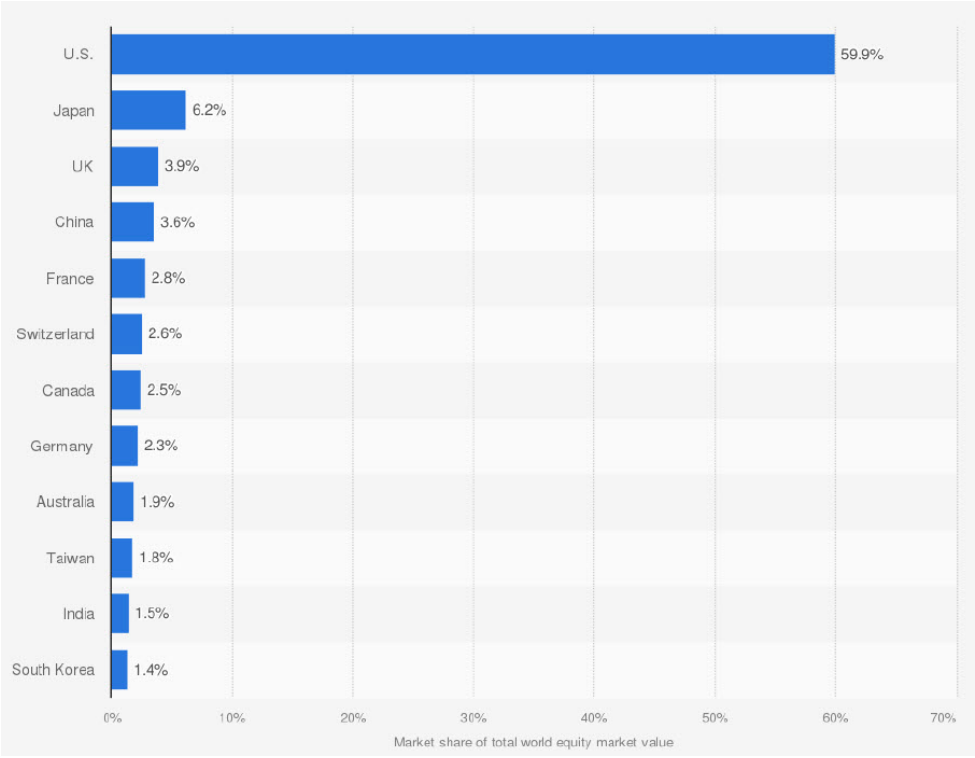
En la figura 2 se presenta la composición de un índice internacional a fecha enero de 2022. También se observa cómo unos pocos países están sobreponderados en el índice, en particular EE. UU. representa el 60 % de la capitalización bursátil mundial.

En el mercado español, el índice más conocido es el Ibex 35, representando a las 35 compañías con mayor volumen negociado. De nuevo, solamente cuatro compañías representan más del 40 % de la capitalización total. En resumen, un mercado es un índice financiero habitualmente concentrado en unas pocas componentes.

Una vez fijado un índice de mercado concreto, podemos definir el término *batir al mercado* como la habilidad de conseguir mejores rentabilidades por unidad de riesgo, de forma consistente y prolongada, que el propio índice.

Figura 2.

Porcentaje de capitalización de los países con mayor capitalización bursátil en el mundo



Fuentes: Credit Suise, FTSE.

Desde los años sesenta, una de las premisas populares de la teoría económica financiera es la hipótesis del mercado eficiente: en todo momento, los precios de los activos financieros incorporan toda la información pública disponible. Por tanto, si dicha hipótesis es cierta, no hay forma de batir al mercado: esto es, no sería posible desarrollar estrategias de inversión que obtengan rentabilidades, de forma consistente y prolongada, por encima de la media del mercado (representado por un índice, por ejemplo, el S&P 500).

Por otra parte, son bien conocidos inversores (Buffet, Lynch), compañías (AQR, Renaissance), y estilos de inversión (*momentum, value*) que han obtenido rentabilidades superiores al mercado a lo largo de los años. Además, existe evidencia de que ciertas estrategias y estilos de inversión (por ejemplo, *momentum* y *value investing*) que también son capaces de batir al mercado de forma consistente en el largo plazo.

Por estos motivos, actualmente existe consenso al pensar que los mercados financieros no son 100 % eficientes en todo momento, por lo que siempre pueden existir pequeñas

anomalías (desviaciones de la eficiencia) que pueden ser explotadas por los inversores para generar rentabilidades por encima del mercado.

Y es en este contexto donde las herramientas de *big data* y de *machine learning* resultan prometedoras para explotar anomalías, ya que actualmente existen grandes volúmenes de datos de activos financieros capaces de procesar y extraer información valiosa. Esta información es el punto de partida para que modelos recientes de ciencia de datos (estadística avanzada y herramientas de *machine learning*) puedan encontrar relaciones complejas entre dichos activos, y conseguir así predicciones suficientemente precisas que logren batir al mercado.

Pero la principal cuestión permanece: ¿hasta qué punto estas nuevas herramientas pueden conseguir mejores rentabilidades del mercado? Las anomalías que aparecen en los mercados son pequeñas, complejas, y dinámicas en el tiempo (pueden permanecer, pero también desaparecer), y por tanto la predictibilidad futura de las rentabilidades (en base a datos pasados) también será pequeña. En lenguaje estadístico, significa que la ratio señal-ruido en los datos financieros es muy próxima a cero.

En este artículo se hará un repaso a las herramientas de estadística avanzada y de *machine learning* que han ido apareciendo en los últimos años. Estas herramientas, en el actual contexto de *big data* financiero, tratan de explotar al máximo la baja predictibilidad existente, para desarrollar estrategias de inversión que intentan batir al mercado: el conocido valor del *big data* financiero. Con este repaso, se analizará si es posible batir al mercado, en qué condiciones, y hasta qué punto.

En particular, en la sección 2 se introducirán los modelos avanzados de estadística para reducir la variabilidad explicada en el valor esperado de las rentabilidades. En la sección 3 se presentarán los últimos avances en *machine learning* para explotar la escasa predictibilidad existente, y desarrollar así estrategias automáticas de inversión basadas en las predicciones de las herramientas de *machine learning*. En la sección 4 se mostrará una aplicación real donde se comparan todas las herramientas presentadas, y se mide la ganancia económica de las mismas a través de un *backtesting* exhaustivo en el mercado de acciones de EE. UU. La sección 5 trata de analizar las relaciones entre los predictores y las rentabilidades a predecir. Finalmente, la sección 6 presenta las principales conclusiones del trabajo.

2. METODOLOGÍA ESTADÍSTICA: MODELOS FACTORIALES

El modelo *CAPM (Capital Asset Pricing Model)* permite describir la relación entre el riesgo asociado al mercado y la rentabilidad asociada a cada activo (o compañía que cotiza en bolsa). Por tanto, es un modelo econométrico que establece una relación lineal entre la rentabilidad esperada para una compañía y el riesgo. En concreto, si definimos la rentabilidad en t para cada activo financiero i como $R_{t,i} = \frac{p_{t,i} - p_{t-1,i}}{p_{t-1,i}}$, donde el precio p está ajustado por dividendos, el modelo CAPM (Sharpe, 1964; Lintner, 1965) se define como:

$$E(R_i) - r_f = \beta_i \, (E(R_M) - r_f), \qquad [1]$$

donde $E(R_i)$ representa la rentabilidad esperada del activo *i*, r_f es la rentabilidad del activo libre de riesgo, y $E(R_M)$ es la rentabilidad esperada del índice de mercado.

La ecuación [1] representa un modelo estadístico que reduce la dimensionalidad en los mercados (por ejemplo, miles de *stocks*) a través de un factor común. El parámetro β permite medir la sensibilidad de las rentabilidades de los activos frente al mercado. El factor común (rentabilidad del mercado, R_M) es capaz de explicar el 70 % de variabilidad en la sección cruzada del panel de rentabilidades. Este hecho marca el inicio de los *ETF* (*Exchange Traded Funds*) en los noventa. Pero, por otra parte, el modelo deja un 30 % de variabilidad no explicada, que puede ser explotada para tratar de batir al mercado.

El modelo de Fama-French de tres factores, Fama y French (1993, 1996), expande el modelo CAPM para tratar de reducir su variabilidad no explicada. Para ello, el modelo de Fama-French añade dos factores de riesgo nuevos: el factor valor y el factor tamaño. Esto es, este modelo incorpora el hecho de que, a lo largo del tiempo, las compañías pequeñas (baja capitalización bursátil) han batido al mercado de forma consistente (en comparación con las compañías grandes). Y de igual forma, las compañías baratas (alto valor contable respecto a su valor bursátil) han obtenido mejores rentabilidades que las compañías caras.

En concreto, el modelo calcula dos nuevos factores de la siguiente forma. El factor tamaño, *SMB* (*small minus big*), ordena para un periodo concreto, todas las acciones de empresas en el índice en función de su capitalización bursátil. Y crea una cartera que compra las acciones con mayor capitalización bursátil y vende aquellas con menor capital, calculando finalmente la rentabilidad de esa cartera en ese periodo. De igual forma, el factor valor, *HML* (*high minus low*), ordena para un periodo concreto, las empresas en función de la ratio valor contable/valor de mercado (*book to market*). A continuación se compran las acciones más baratas (mayor *book-to-market*) y se venden las más caras (menor *book-to-market*), y finalmente se calcula la rentabilidad de esa cartera en ese periodo.

El modelo de Fama-French de tres factores se define como:

$$R_{i,t} - r_{f,t} = \alpha_i + \beta_{M,i} \, (R_{M,t} - r_{f,t}) + \beta_{S,i} \, \mathrm{SMB}_t + \beta_{H,1} \, \mathrm{HML}_t + \varepsilon_{i,t,} \qquad [2]$$

donde $R_{i,t}$ representa la rentabilidad de una acción/fondo/estrategia *i* en *t*, $r_{f,t}$ es la rentabilidad libre de riesgo en *t*, $R_{M,t}$ es la rentabilidad del mercado en *t*, SMB_t representa la rentabilidad del factor tamaño en *t*, y finalmente HML_t representa la rentabilidad del factor valor en *t*.

El modelo [2] es capaz de explicar más de un 90 % de variabilidad en la sección cruzada para carteras diversificadas. Los principales factores que explican la variabilidad entre rentabilidades son sensibles al mercado, al tamaño de la empresa, y a su valor (medido como la ratio *book-to-market*). Cualquier rentabilidad por encima de la esperada se atribuye, por

tanto a un riesgo no sistémico. La exposición al riesgo de cada empresa a cada uno de los tres factores se mide con el correspondiente parámetro β. Finalmente, el parámetro α mide el comportamiento de una empresa o estrategia (rentabilidad ajustada por riesgo): si es significativo, la rentabilidad obtenida no se explica solamente por estos factores comunes, sino que pueden existir otros (quizás específicos) que habría que considerar. Por tanto, el parámetro α resulta útil para saber si se bate al mercado, o para evaluar habilidades en gestión de fondos.

Por tanto, el modelo [2] explica una gran variabilidad en carteras diversificadas y permite calcular las rentabilidades ajustadas por riesgo (α) de una cartera o estrategia para evaluar su comportamiento respecto al mercado. Pero todavía presenta una variabilidad no explicada (alrededor del 10 %) que puede ser reducida introduciendo más factores de riesgo. Desde los años noventa, se han propuesto más y más factores de riesgo para reducir dicha variabilidad y además poder medir de forma más precisa las habilidades de un inversor (α).

Entre los factores de riesgo más conocidos, destacan (usando terminología inglesa, habitual en mercados financieros):

Cuadro 1.

Factores de riesgo más conocidos desde los años noventa

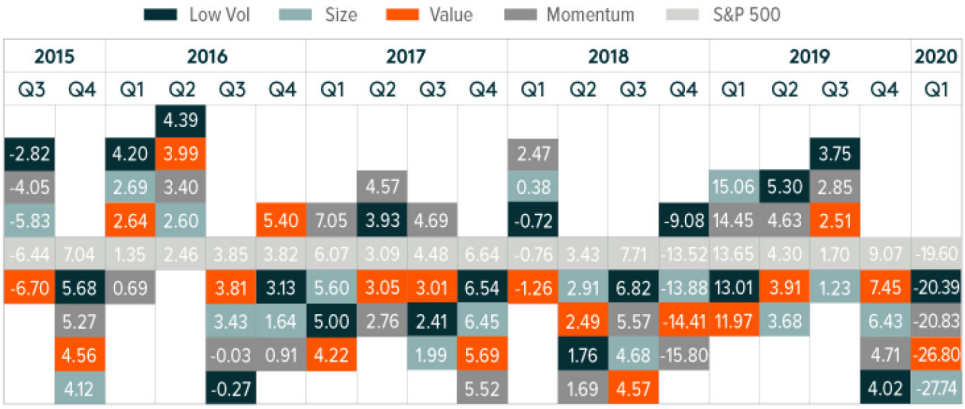
Firm size	Small stocks outperform large stocks
Value	High book-to-market ratio stocks outperform low book-to-market ratio stocks
Momentum	Winner stocks over the last 12 months outperform loser stocks
Investment	Conservative investment firms outperform aggressive investment firms
Profitatility	Profitable firms outperform less profitable firms
Low volatility	Low volatility firms outperform high volatility firms

Los factores de riesgo del cuadro 1 han sido capaces de batir al mercado (α significativa en largos periodos del pasado), pero tienen un comportamiento cíclico. Esto es, un factor de riesgo puede haber batido al mercado en un periodo de tiempo donde otro factor no ha sido capaz de hacerlo. La figura 3 ilustra este comportamiento cíclico. Se puede observar cómo los factores de riesgo que baten al mercado en determinados años no son capaces de batirlo en años posteriores.

Desde los años noventa, cientos de artículos y factores de riesgo se han ido proponiendo, tratando de explicar aún más la variabilidad en la sección cruzada de las rentabilidades esperadas. En concreto, Harvey, Liu y Zhu (2016), analizan al menos 316 factores (anomalías) de la literatura que han batido al mercado de forma consistente, y concluyen que las anomalías que aparecen en los mercados son pequeñas, complejas, y dinámicas en el tiempo. Y muchas de ellas son consecuencia de p-valores significativos pero espurios, algo habitual cuando se realizan cientos de test de hipótesis para validar si una estrategia ha logrado batir al mercado en el pasado. La figura 4 muestra la evolución del número de artículos y anomalías que han ido apareciendo en publicaciones científicas desde los años sesenta.

Figura 3.

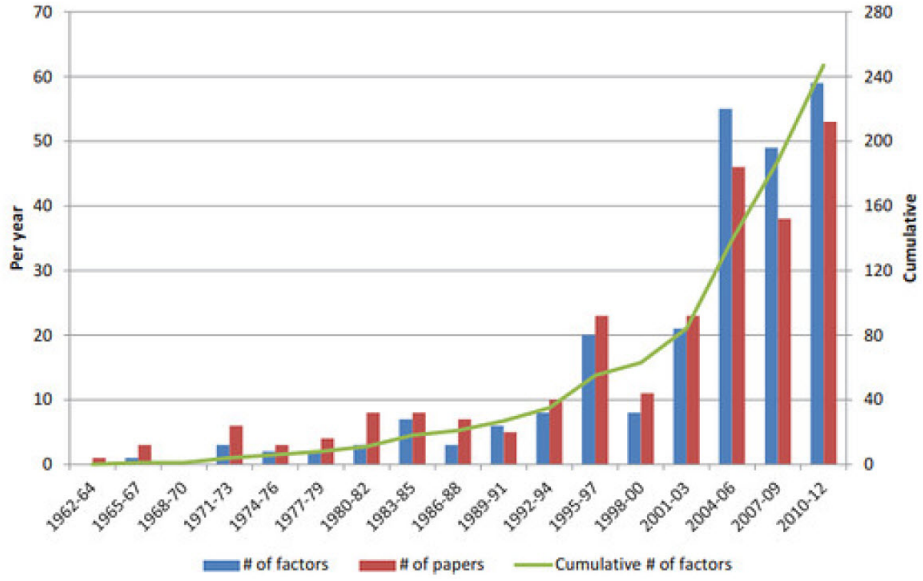
Naturaleza cíclica de los factores de riesgo



Fuente: ERI Scientific Beta.

Figura 4.

Evolución del número de anomalías y factores encontrados en la literatura académica



Fuente: Harvey, Liu y Zhu (2016).

Como se puede observar en la figura 4, la propuesta de nuevas anomalías o factores crece considerablemente a partir de los años 2000. De hecho, el presidente de la American Finance Association en 2011, John Cochrane, llegó a decir en su discurso presidencial (Johtn Cochrane, 2011): "financial academics and practitioners have created a zoo of factors. Most of the supposed market anomalies academics have identified don't exist, or are too small to matter". Por tanto, de los cientos de factores en el zoo, ¿cuáles son fruto del azar o de sesgos de selección?

Sí que hay ciertos factores de riesgo que parecen persistir a lo largo del tiempo, siendo los más conocidos: *beta, size, value, momentum, profitability, quality, term, carry* y *low volatility.* En todo caso, debido a que la ratio señal-ruido en los datos financieros es muy próximo a cero, cabe esperar que la predictibilidad futura de las rentabilidades (en base a datos pasados) sea escasa.

Es en este punto donde puede resultar muy conveniente el uso de herramientas avanzadas de estadística y de *machine learning,* con el objetivo de explotar al máximo la escasa predictibilidad observada en los mercados financieros. En concreto, estas herramientas tratan de identificar *a priori* los factores que batirán al mercado en el futuro en base a su comportamiento cíclico. Por tanto, es importante considerar una gran diversidad de factores para diversificar su exposición a los distintos periodos de tiempo y minimizar así el riesgo asociado. Además, estas herramientas permiten explotar no linealidades e interacciones entre los factores (predictores) y las rentabilidades futuras de las compañías. Finalmente, las herramientas avanzadas de estadística y de *machine learning* pueden presentar problemas de sobreajuste, es decir, son capaces de predecir bien el pasado pero mal el futuro. Es, por tanto, necesario saber manejar bien dichas herramientas para mitigar este problema.

En la siguiente sección se repasarán las herramientas más avanzadas capaces de explotar la baja predictibilidad en los mercados financieros.

3. METODOLOGÍA DE *MACHINE LEARNING*

Las herramientas de *machine learning* ponen el énfasis en reducir el error de predicción de rentabilidades futuras, mientras que los modelos estadísticos tratan de explicar las diferencias entre rentabilidades pasadas. En concreto, el siguiente modelo estadístico es una extensión del modelo factorial [2]:

$$R_{t,i} = \alpha_i + \beta_i^T f_t + \varepsilon_{t,i} \qquad [3]$$

donde $R_{t,i}$ representa las rentabilidades de (miles) de compañías, y f_i son (cientos) de factores (predictores) que explican las rentabilidades de forma lineal. Es un enfoque explicativo porque las relaciones se analizan en el mismo tiempo t.

En cambio, el enfoque de *machine learning* es predictivo, y consiste en desplazar una posición el tiempo en el modelo anterior:

$$R_{t+1,i} = \alpha_i + \beta_i^T f_t + \varepsilon_{t+1,i} \quad [4]$$

Este desplazamiento temporal implica que el enfoque de *machine learning* no pueda conseguir una predictibilidad superior al 2 %, mientras que el enfoque estadístico puede explicar más de un 90 % de variabilidad.

La cuestión importante que surge ahora es: ¿será posible batir al mercado con menos de un 2 % de predictibilidad?

En primer lugar repasemos qué herramientas analíticas son capaces de conseguir una predictibilidad del 2 %. Posteriormente veremos cómo se pueden diseñar estrategias de inversión a partir de predicciones poco precisas.

Las herramientas de Ciencia de Datos engloban tanto modelos avanzados de estadística como algoritmos de *machine learning,* y son aptas para tratar de predecir variables en grandes conjuntos de datos (*big data*). En concreto, las herramientas avanzadas de estadística tratan de estimar el modelo [4] asumiendo una relación lineal entre los factores (predictores) y la respuesta (rentabilidades). Como la cantidad de rentabilidades a predecir en cada instante t es muy elevada (orden de miles) y la cantidad de factores también (orden de cientos), se requieren herramientas avanzadas capaces de estimar bien dicho modelo tratando de evitar problemas de sobreajuste. Las mejores herramientas de estadística son extensiones de mínimos cuadrados, donde bien se añaden términos de regularización para reducir el sobreajuste (Lasso, Elastic Net), o bien se trata de reducir la dimensionalidad de los factores (PCR, PLS).

Por otro lado, las herramientas de *machine learning* tratan de relajar la hipótesis de linealidad en el modelo [4], y capturar así posibles no-linealidades e interacciones entre factores. Por tanto, en lugar de estimar el modelo [4] de forma explícita, estas herramientas entrenan una función no explícita en su lugar: $R_{t+1} = \text{map}(f_t) + \varepsilon_{t+1}$. Entre las herramientas de *machine learning* para este tipo de datos destacan: combinaciones de árboles de decisión (*random forests* y *gradient boosting*), y las redes neuronales.

Finalmente, hay que mencionar que para capturar la naturaleza dinámica en los mercados financieros, los modelos anteriores se reentrenan cada cierto tiempo (cada mes o cada año). En la siguiente sección veremos cómo estas herramientas son capaces de capturar parte de la escasa predictibilidad presente en los mercados financieros. Pero una vez capturada esta predictibilidad, ¿cómo se pueden diseñar estrategias de inversión que exploten bien las predicciones para tratar de batir al mercado?

En principio se pueden diseñar muchas estrategias de inversión automáticas, con cierto sentido de optimalidad. Markowitz (1952) propuso una estrategia de diversificación óptima teniendo en cuenta las rentabilidades esperadas de cada compañía, y la matriz de covarianzas asociada a ellas. En nuestro contexto, parece inviable calcular la matriz de correlaciones entre miles de rentabilidades de distintas compañías. Una estrategia más factible en la práctica consiste en diversificar entre las distintas compañías usando solamente la información

de las rentabilidades obtenidas mediante las herramientas de Ciencia de Datos anteriores. En concreto, y dado un mes determinado, se pueden ordenar todas las compañías en base a sus predicciones para el próximo mes. Una vez ordenadas, se pueden comprar las compañías con predicciones más altas (aquellas por encima del percentil 90 %) y vender (o descartar) aquellas con predicciones más bajas (aquellas por debajo del percentil 10 %). Para comprar o vender compañías se pueden usar los mismos pesos para todas ellas y así simplificar aún más la estrategia.

En la siguiente sección se muestra la aplicación de estas ideas en el mercado de EE. UU. donde se realiza un *backtesting* de distintas estrategias desde 1960 a 2016, considerando un total de 30.000 compañías en ese periodo, con una media de unas 6.000 rentabilidades a predecir cada mes. Además, se consideran alrededor de 100 características de cada compañía (factores), decenas de variables macroeconómicas, más sus posibles interacciones: en total, alrededor de 1000 predictores cada mes.

4. APLICACIÓN: *EMPIRICAL ASSET PRICING* VÍA *MACHINE LEARNING, REVIEW OF FINANCIAL STUDIES,* 2020

La siguiente aplicación está extraída de Gu, Kelly y Xiu (2020). Es este artículo, los autores realizan una comparación exhaustiva de diferentes herramientas de estadística avanzada y de *machine learning* para medir la ganancia económica de las mismas. A continuación se muestra un esquema del *backtesting* realizado en el artículo.

Para cada mes *t*, desde el año 1987 a 2016:

- Se estiman distintos modelos estadísticos y de ML, explicados en la sección anterior, usando el conjunto de predictores (factores). Estos modelos utilizan información de los 30 años inmediatamente anteriores (más de 18 años para entrenamiento, y 12 años para su validación).

- Una vez los modelos han sido entrenados con información previa, se usan para predecir la rentabilidad en el $t + 1$ para todas las compañías disponibles en ese mes.

- Se ordenan, de mayor a menor, todas las predicciones del paso anterior, y se desarrolla la siguiente estrategia de inversión: comprar las compañías en el percentil 10 % superior de las predicciones (con el mismo peso) y vender las compañías en el percentil 10 % inferior.

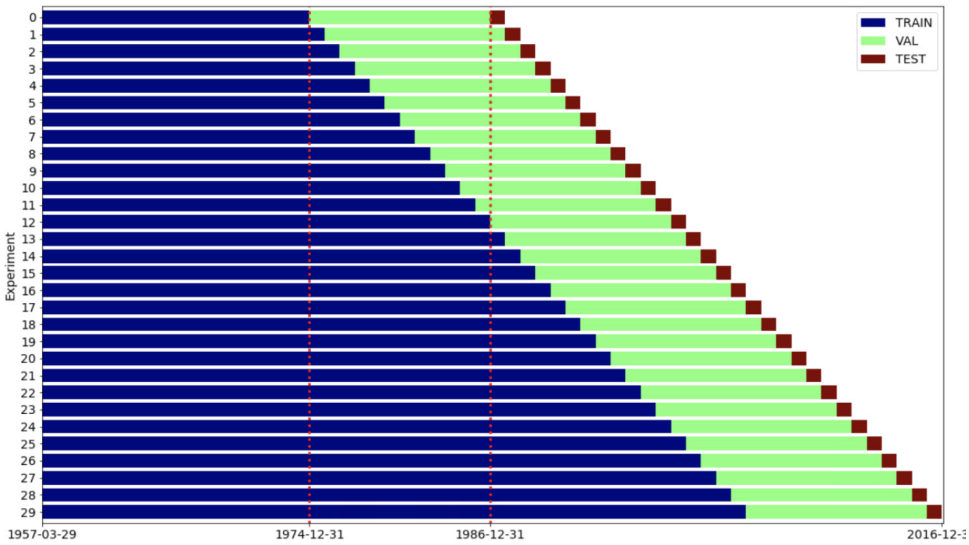
- Finalmente, se evalúa el comportamiento estadístico y económico en $t + 1$ de todas las herramientas consideradas.

Cabe destacar que, debido a que las herramientas de ML tienen un alto coste computacional, dichas herramientas se reentrenan (en el Paso 1) cada doce meses.

En la figura 5 se ilustra el *backtesting* propuesto en Gu, Kelly y Xiu (2020).

Figura 5.

Esquema de *backtesting*



Fuente: Gu, Kelly y Xiu (2020).

En color azul se muestran los meses utilizados para entrenar los modelos. Se comienza con 18 meses, y luego se va añadiendo un mes según avanza el *backtesting*. En color verde se muestra la ventana de validación (donde se optimizan hiperparámetros presentes en las herramientas de ML), que consta siempre de 12 meses. Finalmente, en color rojo aparece el mes a predecir en el Paso 2 del esquema anterior. En total, el *backtesting* consta de 30 meses donde se realizan predicciones (muestra test).

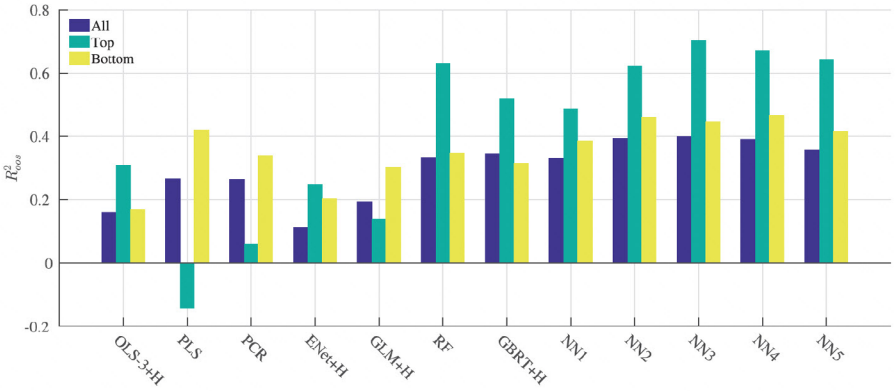
La figura 6 muestra la comparación de las distintas herramientas consideradas (13 en total) en términos del R^2 obtenido en el periodo de test (*out-of-sample*), que es una medida de precisión de las predicciones.

En concreto, el cuadro muestra el comportamiento de distintas herramientas analíticas: *OLS (ordinary least squares)* es una regresión lineal considerando todos los factores disponibles, OLS-3 es la misma regresión, pero considerando solamente los 2 factores de Fama-French (tamaño y valor) y el factor *momentum, PLS (partial least squares)* es una variante de regresión lineal que proyecta conjuntamente los predictores y la variable respuesta en un nuevo espacio, *PCR (principal component analysis)* es una regresión lineal considerando las primeras componentes principales de los predictores, *ENet (elastic net)* representa una versión regularizada de regresión lineal, *GLM (generalized linear model)* es una extensión no

Figura 6.

Comportamiento estadístico de las distintas herramientas

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	−3.46	0.16	0.27	0.26	0.11	0.19	0.33	0.34	0.33	0.39	0.40	0.39	0.36
Top 1,000	−11.28	0.31	−0.14	0.06	0.25	0.14	0.63	0.52	0.49	0.62	0.70	0.67	0.64
Bottom 1,000	−1.30	0.17	0.42	0.34	0.20	0.30	0.35	0.32	0.38	0.46	0.45	0.47	0.42



Fuente: Gu, Kelly y Xiu (2020).

lineal de la regresión múltiple, *RF (random forest)* representa una combinación de árboles de decisión, *GRBT (gradient boosted regression trees)* representa otra combinación de árboles pero de forma secuencial, y finalmente *NN (neural network)* son distintas arquitecturas de redes neuronales considerando desde una capa hasta cinco (NN1, . . . , NN5).

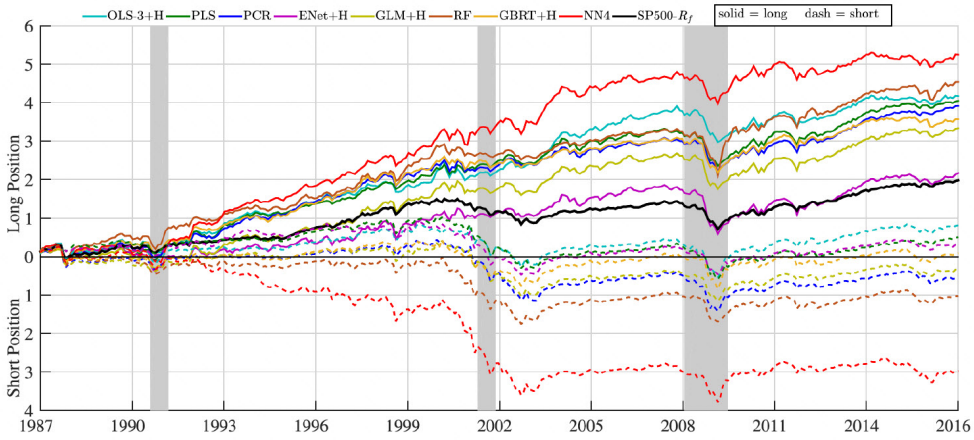
La primera fila del cuadro en la figura 6 muestra cómo herramientas tradicionales (OLS) obtienen una muy baja precisión (R^2 negativo o muy próximo a 0). Las herramientas basadas en estadística avanzada (PLS, PCR, ENet y GLM) consiguen mejorar la precisión, llegando a obtener un R^2 alrededor de 0,27 %. Pero las herramientas de ML (RF, GBRT, NN) consiguen aumentar más la precisión, llegando a obtener un R^2 = 0,4 en el caso de las redes neuronales (NN). Estos resultados indican el valor de incorporar interacciones complejas entre los predictores, que son capturadas por las herramientas de ML, y en menor medida por las de estadística.

En resumen, el R^2 (*out of sample*) es menor del 1 % para las herramientas de ML más avanzadas (muy baja predictibilidad). Sin embargo, la figura 7 muestra cómo la estrategia de inversión (Paso 3 en el *backtesting*) es capaz de explotar esta baja predictibilidad para conseguir un buen rendimiento económico.

En concreto, la figura 7 muestra la rentabilidad acumulada de las distintas estrategias de inversión (con las posiciones cortas y largas), e incluye como *benchmark* (en color negro)

Figura 7.

Comportamiento económico de las distintas herramientas



Fuente: Extraído de Gu, Kelly y Xiu (2020).

la rentabilidad acumulada del índice de mercado (S&P500). Se observa cómo las estrategias basadas en redes neuronales (especialmente NN4, en color rojo) dominan claramente al resto de estrategias. Más concretamente, las herramientas avanzadas de estadística baten al mercado ligeramente, mientras que las basadas en ML logran batirlo de forma más clara, siendo los modelos basados en redes neuronales los más prometedores: obtienen *Sharpe ratio* anualizado (rentabilidad anual entre volatilidad) de 2.45, y *alphas* significativamente positivos.

En resumen, las herramientas avanzadas de ML consiguen explotar la escasa predictibilidad existente para obtener una alta ganancia en términos económicos.

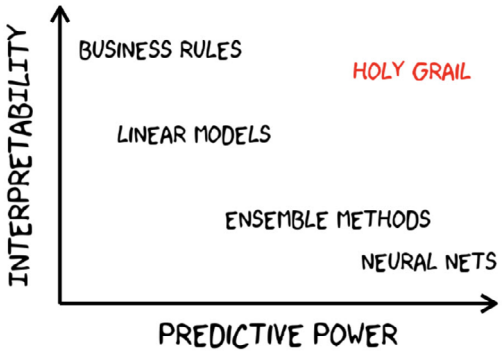
5. INTERPRETABILIDAD DE LAS HERRAMIENTAS DE *MACHINE LEARNING*

Las herramientas de *machine learning,* debido a que estiman el modelo $R_{t+1} = \text{map}(f_t) + \varepsilon_{t+1}$ de forma no explícita, no resultan adecuadas para entender de qué forma los factores (predictores) ayudan a explotar mejor la predictibilidad. Pero en los últimos años han ido apareciendo herramientas que ayudan a entender mejor estas relaciones. En particular, las herramientas de interpretabilidad en *machine learning* analizan la importancia de cada variable en las predicciones obtenidas, así como ayudan a entender mejor las relaciones entre los predictores, y cómo dichas relaciones evolucionan con el tiempo.

En general, cuanto más poder predictivo tiene un modelo, menor es su capacidad interpretativa, como muestra la figura 8.

Figura 8.

Capacidad predictiva de modelos vs. interpretabilidad

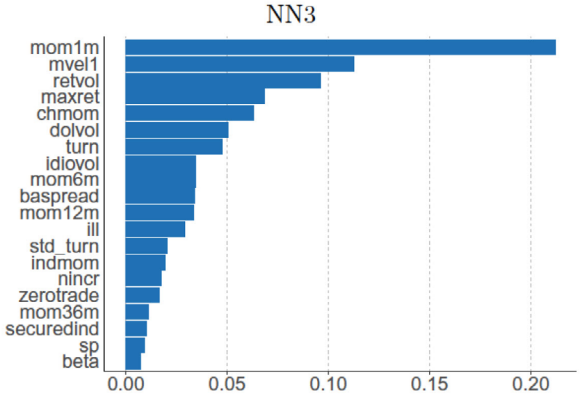


Fuente: @ManoMano.

En la aplicación financiera de la sección anterior, las herramientas de interpretabilidad analizan la importancia de cada factor considerado para cada modelo. La importancia se mide en términos de mejora del error de predicción para cada factor considerado. Además, estas herramientas ayudan a entender la relación marginal (no lineal) entre un factor y las predicciones de las rentabilidades. La figura 9 muestra la importancia de los 20 predictores más importantes para uno de los mejores modelos predictivos (red neuronal de tres capas). La importancia de las variables se ha normalizado para que sume 1 y sea más fácil su interpretación.

Figura 9.

Importancia de los predictores en la respuesta

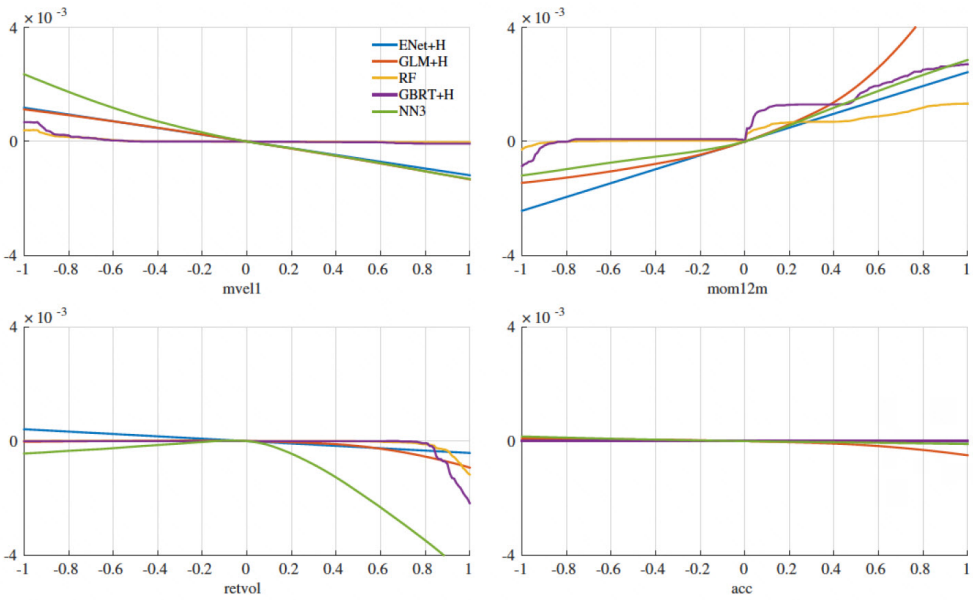


Fuente: Gu, Kelly y Xiu (2020).

De la figura 9 se deduce que los predictores dominantes para ese modelo están relacionados con *momentum,* liquidez y volatilidad. La figura 10 muestra el impacto marginal (para cada modelo) de un factor sobre las rentabilidades. Este impacto se consigue medir (para un predictor) asumiendo que el resto de predictores permanece fijo. La figura muestra, para distintos modelos, el impacto marginal de 4 de los factores más importantes. En modelos lineales, este impacto sería lineal.

Figura 10.

Efecto marginal de un predictor en la respuesta



Fuente: Gu, Kelly y Xiu (2020).

En concreto, la figura 10 identifica patrones no lineales en términos de predictibilidad, que los modelos lineales no son capaces de capturar, o los identifican con un efecto nulo. La figura superior izquierda muestra que los retornos esperados decrecen con el tamaño para los modelos basados en *machine learning,* mientras que los modelos estadísticos no encuentran relación. La figura superior derecha muestra que los retornos actuales son crecientes con los retornos encontrados en los últimos doce meses, pero las herramientas de *machine learning* logran modelar dicho crecimiento de forma no lineal. Algo similar ocurre en la figura inferior izquierda, donde las herramientas de *machine learning* logran capturar una relación decreciente no lineal entre los retornos y su volatilidad. Finalmente, la figura inferior derecha muestra que todos los modelos apenas encuentran relación entre los retornos y los devengos.

Estas relaciones no lineales explican parcialmente por qué las herramientas de *machine learning* son capaces de predecir algo mejor que las herramientas avanzadas de estadística.

6. CONCLUSIONES

En este trabajo se ha analizado si es posible diseñar estrategias de inversión basadas en *big data* y en *machine learning* para tratar de conseguir mejores rentabilidades (por unidad de riesgo) que los índices representativos de los mercados financieros. La principal conclusión es que sí es posible aunque no es sencillo, debido a que la predictibilidad futura de las rentabilidades (en base a datos pasados) es muy pequeña. Esto es, en los mercados financieros la ratio señal-ruido de los datos es muy baja. Pero los resultados mostrados en la aplicación de este trabajo muestran que pequeñas mejoras en las predicciones pueden dar lugar a grandes mejoras en las rentabilidades, explotando interacciones y no linealidades presentes en los datos para incrementar el valor económico. En concreto, las herramientas de *machine learning,* usando como *input* un gran volumen de datos sobre miles de compañías y variables macroeconómicas) son capaces de explotar un escaso 1 % de predictibilidad existente, y conseguir así un alto rendimiento económico (*Sharpe ratio* anualizado superior a 2 a lo largo de 30 meses).

Por tanto, las herramientas de *machine learning* resultan ser muy prometedoras para predecir miles de rentabilidades futuras en los mercados financieros. Además, las recientes herramientas de interpretabilidad de estas herramientas tratan de intuir el impacto (no lineal) de los predictores en la respuesta a lo largo del tiempo.

La evidencia empírica de este trabajo está basada en el análisis de Gu, Kelly y Xiu (2020) en el mercado de acciones de EE. UU. En DeMiguel *et al.* (2022) se presenta una evidencia empírica similar en fondos de inversión de EE. UU., mostrando la robustez de las herramientas de *machine learning* en distintos mercados.

Referencias

Cochrane, J. (2011). Presidential address: Discount rates. *Journal of Finance,* 66, pp. 1047–1108.

DeMiguel, V., Gil-Bazo, J., Nogales, F. J. y Santos, André A. P. (2022). Machine Learning and Fund Characteristics Help to Select Mutual Funds with Positive Alpha. *WP.*

Fama, E. F. y French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics,* 33, pp. 3–56.

Fama, E. F. y French, K. R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *Journal of Finance,* 51, pp. 55–84.

Gu, S., Kelly, B. y Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies,* 33, pp. 2223–2273.

Harvey, C. R., Liu, Y. y Zhu, H. (2015)... and the cross-section of expected returns. *Review of Financial Studies,* 29, pp. 5–68.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics, 47,* pp. 13–37.

MacLean, D. y Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? *Journal of Finance,* 71, pp. 5–32.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance, 7*(1), pp. 77– 91.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance,* 19, pp. 425– 42.