



UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS ECONÓMICAS Y SOCIALES
ESCUELA DE ESTADÍSTICA Y CIENCIAS ACTUARIALES

EVALUACIÓN DE LA EFICACIA DE UNA
ESTRATEGIA DE TRADING BASADA EN MODELO
LINEAL GENERALIZADO

Trabajo Final de Pregrado

PARA OPTAR POR EL TÍTULO DE:
Licenciado en Ciencias Actuariales

Rodrigo Alejandro Serrano Morales

TUTOR:
Prof. Jonattan Ramos & Prof. Eloy Eligon

Caracas, Marzo 2019

*Dedicado a mis padres y a mi hermana, por ser los pilares de mi vida y ser mi motivo de
haber cumplido esta meta.*

Agradecimientos

Agradecimientos

Gracias

Índice general

Índice de figuras	4
Índice de tablas	5
Introduccion	6
1. El Problema	7
1.1. Justificación	7
1.2. Planteamiento del Problema	7
1.3. Objetivo General	7
1.4. Objetivos Específicos	7
2. Marco Teórico	8
2.1. Antecedentes	8
2.1.1. Trading de Cryptomonedas basado en Aprendizaje Automatico	8
2.1.2. Modelos predictivos para el mercado FOREX	8
2.1.3. Diseño e implementación de un sistema automatizado para operar en el mercado de divisas usando reglas de asociación	8
2.2. Bases Teóricas	8
2.2.1. Hipótesis del Mercado Eficiente	8
2.2.2. Análisis Técnico	9
2.2.3. Introducción al aprendizaje automático	10
2.2.4. Regresión Logística	11
2.2.5. Análisis de Componentes Principales como método de reducción de variables	12
2.2.6. Matriz de Confusión	13
2.3. Bases Legales	14
3. Marco Metódico	15
3.1. Análisis Exploratorio de los datos	15
3.1.1. Datos OHLC y Fuente de los datos	15
3.1.2. Series de Índices	15
3.2. Entrenamiento del Modelo	16
3.2.1. Variable dependiente	16
3.2.2. Indicadores Técnicos como variables predictoras	17
3.2.3. Validación Cruzada en Series de Tiempo	19
3.2.4. WalkForward Backtesting	20
3.2.5. Reducción de la dimensión con Análisis de Componentes Principales	21

4. Análisis de Resultados	25
4.1. Series de Rendimientos	25
Conclusiones y Recomendaciones	26
Lista de Referencias	27

Índice de figuras

2.1. Matriz de Confusión	13
3.1. Precios de Cierre de los índices en el período de estudio (26/10/2008 - 18/01/2019)	16
3.2. Correlación entre indicadores originales calculados con los precios del SP500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)	18
3.3. Correlación entre indicadores definitivos calculados con los precios del SP500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)	19
3.4. Metodología WalkForward	20
3.5. Observaciones que presentan ocurrencias con $sl = 2.5\%$, en el primer conjunto de datos de entrenamiento (26/10/2008 - 31/12/2010) para el índice SP500, para diferentes valores de tp y h	21
3.6. d	22
3.7. d	23
3.8. d	24

Índice de tablas

Introducción

Las economías de países latinoamericanos son reconocidas históricamente por depender en gran magnitud del comercio de sus materias primas, lo cual hace que el comercio con dichos commodities resulte de gran impacto para el gasto fiscal y para la balanza de pagos, esto deja como consecuencia, la necesidad de los actores económicos de estudiar el riesgo a profundidad para poder evitar resultados que reduzcan sus retornos positivos.

1.1. Justificación

1.2. Planteamiento del Problema

1.3. Objetivo General

Evaluar la eficacia de una estrategia de trading basada en técnicas de aprendizaje automático en diferentes instrumentos financieros

1.4. Objetivos Específicos

- Definir los parámetros de la estrategia, así como los indicadores técnicos a utilizar como variables predictoras
- Utilizar Análisis de Componentes Principales como técnica de reducción de la dimensión de variables predictoras
- Aplicar método de Regresión Logística con las componentes arrojadas por el ACP
- Desarrollar la metodología Walkforward para tomar en cuenta el dinamismo del mercado en la estrategia
- Evaluar la predicción del modelo como estrategia en los instrumentos seleccionados

2.1. Antecedentes

2.1.1. Trading de Cryptomonedas basado en Aprendizaje Automatico

2.1.2. Modelos predictivos para el mercado FOREX

2.1.3. Diseño e implementación de un sistema automatizado para operar en el mercado de divisas usando reglas de asociación

2.2. Bases Teóricas

2.2.1. Hipótesis del Mercado Eficiente

La Hipótesis del Mercado eficiente fue desarrollada por Eugene Fama en los años 60, en la misma argumenta que los precios de los activos reflejan toda la información disponible, es decir que siempre son transados a un valor adecuado para su riesgo, haciendo imposible para los inversores obtener retornos más elevados que los del mercado en general.

Fama sugiere tres suposiciones. Primero, el mercado eficiente requiere un gran número de competidores buscando maximizar ganancias. Segundo, la información que afecta al activo llega al mercado de manera aleatoria y cada anuncio es independiente de los demás. Tercero, todos los competidores intentarán ajustar sus posiciones lo más rapido posible conocida la información del mercado. Existen tres variantes de la hipótesis:

Eficiencia débil, en esta variante, los precios del pasado no sirven para predecir el precio futuro, es decir cualquier movimiento del activo es determinado por información no contenida en la serie de precios. *Eficiencia media*, en esta forma se asume que los precios se ajustan instantáneamente a la información pública, por lo que rechaza cualquier tipo de arbitraje intentando aprovechar nueva información. *Eficiencia fuerte*, esta última forma de la hipótesis plantea que los precios reflejan tanto información pública como privada, por lo cual incluso obteniendo información no conocida por todos los competidores, no se pueden obtener retornos anormales a los de los mercados.

Aunque esta hipótesis es la piedra angular de la teoría financiera moderna, es controversial entre la comunidad financiera y disputada frecuentemente. Gran parte de sus detractores argumentan que el precio del activo está influenciado por suposiciones sesgada de los individuos, formuladas por la manera en como estos responden ante nueva información. Algunas de las hipótesis que explican este razonamiento sesgado son:

Los inversores interpretan la información de manera distinta, por lo que generarán diferentes

valuaciones de un mismo activo, lo que sugiere que la reacción del inversor a la misma noticia será distinta. Day and Wangr (2002) argumentan que si los precios son continuamente influenciados por estas interpretaciones erróneas, los movimientos contrarios del precio pueden ser predecidos estudiando la data histórica. Sugieren también que mientras más extremo sea el movimiento inicial, mayor será el ajuste de precio.

Los inversores se dejan influenciar por la tendencia del mercado, este comportamiento se ha visto a lo largo de la historia en casos de colapso del mercado como en la caída del mercado bursátil en 1987 ó la burbuja del puntocom a finales de los 90. Froot (1992) muestra como estos comportamientos pueden resultar en ineficiencias del mercado.

Aglunos académicos como Hong y Stein's (1999) categorizan a los inversores en *Informados* y *Noinformados*. Los inversores que tienen acceso a la información solo operan al obtener nueva información, mientras que los no informados operan basados en el pasado reciente del activo. A medida que la información es conocida por todos los competidores, se forma el fenómeno de reversión a la media.

Es evidente la postura que se asume en la presente investigación con respecto a la hipótesis de mercado eficiente. Además de los aspectos del comportamiento de los competires, se ha evidenciado en la historia, casos de inversores que han logrado vencer el mercado por largos períodos de tiempo, como Warren Buffet, lo cual por definición de la hipótesis es imposible. Por otro lado, Los avances tecnológicos y la capacidad de procesamiento de las computadoras en la actualidad hacen pensar que cualquier anomalía presente en el mercado por muy pequeña que sea puede ser aprovechada por sofisticados softwares automatizados.

2.2.2. Análisis Técnico

Los inversionistas que rechazan la hipótesis del mercado eficiente buscan interpretar la situación del mercado, bien a través de noticias que afecten al activo o estudiando su movimiento intentando extrer patrones de conducta. A la primera técnica se le llama *AnlisisFundamental* y el segundo *AnlisisTcnico*. El Análisis Fundamental está mas asociado a estrategias de inversión pasivas a largo plazo aunque en la actualidad se han desarrollado algoritmos de compra y venta que buscan predecir la dirección del precio en función de noticas utilizando minería de texto.

El análisis técnico es aquel que busca patrones y tendencias de comportamiento en la cotización de los activos financieros, basándose en la serie de tiempo del activo, con esto intenta predecir el movimiento futuro mediante el uso de gráficos. Según J.J.Murphy (1999) existen tres fundamentos básicos en los que se basa el análisis técnico: Los movimientos del mercado lo descuentan todo, los precios se mueven por tendencias y la historia se repite.

Murphy establece que cualquier efecto que posiblemente pueda afectar al precio se ve reflejado en la cotización del mismo. Por lo que un estudio del desplazamiento del activo en un período de tiempo sería suficiente para lograr predecir su movimiento. Esto quiere decir que el análisis técnico no es mas que una manera indirecta de estudiar los fundamentos del activo, suponiendo que la cotización del mismo resume toda la información que lo afecta.

El analista técnico acepta la premisa de que los mercados tienen tendencias. Buscar tendencias en las primeras etapas de su desarrollo es la razón de toda la representación gráfica dentro del análisis, con el fin de que las transacciones vayan en dirección de esa tendencia. Por otro lado, la afirmación de que la historia se repite tiene que ver con el estudio de la psicología humana, la cual según murphy se repite. Ésta afirmación tiene también una estrecha relación en los ciclos económicos.

2.2.3. Introducción al aprendizaje automático

Aprendizaje automático refiere a una rama de la Inteligencia Artificial, que busca crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de un conjunto de datos. Supongamos que existe una variable respuesta Y y distintos predictores X_1, X_2, \dots, X_j . Se asume que existe una relación entre Y y $X = X_1, X_2, \dots, X_j$, la cual puede ser escrita de forma general como

$$Y = f(X) + \epsilon$$

donde f es una función desconocida de X y ϵ es un término de error aleatorio, independiente de X y de media 0. En esta formulación f representa información sistemática que X proporciona sobre Y .

En esencia, el aprendizaje automático refiere a un conjunto de enfoques para estimar f

Métodos Paramétricos vs No Paramétricos

La mayoría de los métodos de aprendizaje automático pueden ser caracterizados como paramétricos o no paramétricos. Los primeros, involucran un enfoque basado en dos pasos. Primero se asume que los datos toman una forma específica, una vez asumida la forma que debe tener la función f , el problema de la estimación se simplifica. Al seleccionar el modelo se procede a ajustar el modelo en la data de entrenamiento. Este es el caso de los Modelos Lineales Generalizados como la Regresión Logística. La desventaja de este enfoque paramétrico es que el modelo escogido puede no ser apropiado a la verdadera forma de f , por lo que la estimación puede ser pobre.

Por otro lado los modelos No Paramétricos no asumen ninguna forma para f , en cambio, estos modelos buscan estimar f acercándola lo más posible a los datos observados. Esto les permite evadir el problema de ajustarse a alguna forma en específico. Sin embargo, al no reducir el problema a estimar unos parámetros sino utilizar los datos directamente, se necesita un gran número de observaciones -muchas más que las necesarias por los métodos paramétricos- para obtener una estimación precisa. Además en general la interpretación del modelo se hace más difícil con estos métodos y son propensos a caer en sobreoptimización.

Aprendizaje Supervisado vs No Supervisado

Se le llama aprendizaje Supervisado, a los métodos en los cuales para cada observación de las variables predictoras x_i existe un valor asociado a la variable respuesta y_i . Por lo que se ajusta un modelo que relacione la respuesta con los predictores, con el fin de predecir acertivamente respuestas futuras. Este es el caso de los modelos Lineales así como los métodos de boosting, SVM, GAM, etc.

En contraste, los métodos No Supervisados describen una situación más complicada, en donde para cada observación, se cuenta con variables predictoras, pero no existe ninguna variable respuesta. Lo que se busca en este tipo de modelos es buscar entender la relación entre las variables o entre las observaciones. Para esto se utilizan métodos de agrupación o cluster y métodos de reglas de asociación. Los primeros intentan describir las agrupaciones subyacentes en los datos, como por ejemplo, el tipo de clientes dependiendo de su comportamiento de compra. Las reglas de asociación buscan descubrir patrones inherentes que describan el comportamiento de los datos u observaciones, por ejemplo, un grupo de clientes que compren un producto r conjunto con otro producto s .

Regresión vs Clasificación

Las variables pueden ser divididas entre cuantitativas ó cualitativas -también llamadas categóricas-. Las cuantitativas toman valores numericos mientras que las cualitativas son categorías o clases. Dependiendo del tipo de variable respuesta se realiza el enfoque del modelo. En el caso de que la variable respuesta sea cuantitativa se refiere a problemas de regresión, mientras que los que involucran una variable respuesta cualitativa, son referidos como problemas de clasificación.

Compensación entre sesgo y varianza

2.2.4. Regresión Logística

Los modelos Lineales Generalizados asumen que existe una aproximada relación lineal entre la variable respuesta Y y la variable predictora X . Matemáticamente se puede describir la relación como:

$$Y \approx \beta_0 + \beta_1 X_1 + \dots + \beta_j X_j$$

En donde X_j representa las variables predictoras y los coeficientes β_j cuantifican la asociación entre la variable predictora X_j y la variable respuesta Y . Por lo que se interpreta a β_j como el efecto promedio que tiene en Y un incremento de una unidad en X_j , bajo el supuesto de que todas las demás variables se mantienen constantes.

En problemas de clasificación, la variable predictora asume valores categóricos, por lo que al utilizar este enfoque se pueden obtener probabilidades fuera del intervalo $[0, 1]$, haciendo imposible su interpretación. Esto concluye en que se deba utilizar una función, tal que permita la generación de valores entre $[0, 1]$, en el caso de la regresión logística esta función es:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j}}$$

Despejando se obtiene

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j}$$

El lado izquierdo de la ecuación puede tomar valores entre 0 e ∞ , lo cual indicaría muy bajas o muy altas probabilidades, aplicando logaritmo en ambos miembros de la ecuación se obtiene la función logit

$$\log \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j$$

Se observa que la función logit es lineal en X , por lo que incrementar una unidad de X afecta el lado izquierdo de la ecuación en β . Sin embargo dado que la relación entre $p(X)$ y X no es una linea recta, β no corresponde a un cambio en $p(X)$ asociado a una unidad de incremento en X . Se debe hacer la respectiva transformación para interpretar el coeficiente β en relación a Y .

Máxima Verosimilitud

Los coeficientes β son desconocidos, por lo que deben estimarse en la data de entrenamiento. Para esto se utiliza el método de *Mxima Verosimilitud*, el cual consiste en estimar los coeficientes para los cuales la probabilidad de predicción para cada individuo, utilizando (formula arriba),

corresponda lo más cercano posible al valor observado del individuo. Se define la función de verosimilitud como

$$l(\beta) = \prod_{i=1}^j P(x_i/\beta)$$

Por conveniencia se trabaja con el logaritmo, dado que esto transforma una operación de productos de probabilidades en una sumatoria, por lo que se obtiene

$$l(\beta) = \sum_{i=1}^N \log P(y_i/x_i; \beta)$$

Al codificar las clases en 0 y 1, la función de verosimilitud para la regresión logarítmica puede ser escrita como

$$l(\beta) = \sum_{i=1}^N (y_i \beta^T x_i - \log 1 + e^{\beta^T x_i})$$

Para maximizar la función, se iguala la derivada a 0

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^N x_i (y_i - P(x_i; \beta)) = 0$$

Para resolver la ecuación (n arriba) se utiliza un algoritmo de optimización llamado *Newton–Raphson*

2.2.5. Análisis de Componentes Principales como método de reducción de variables

Los modelos lineales tienen distintas ventajas en cuanto a interpretación y muchas veces son sorprendentemente competitivos en relación con los métodos no lineales. Existen técnicas para relajar el supuesto de que la relación entre la respuesta y los predictores es lineal, arrojando mejores predicciones e interpretabilidad.

Una clase de métodos es el enfoque de *Reducción de la Dimensión*, el cual involucra proyectar los p predictores en M -dimensiones o componentes, donde $M < p$. Esto se logra transformando los predictores en combinaciones lineales que recogen parte de la información, estas M componentes o dimensiones son entonces utilizadas como nuevos predictores en el modelo de regresión. Esto es:

$$Z_m = \sum_{i=1}^p \phi_{im} X_i$$

Para cualquier constante $\phi_{1m}, \phi_{2m}, \dots, \phi_{pm}, m = 1, \dots, M$. Se ajusta el modelo

$$y_i = \theta_0 + \sum_{m=1}^M \theta_m z_{im} + \epsilon_i, \quad i = 1, \dots, n$$

En situaciones donde p es relativamente grande con relación a n , seleccionar un valor de $M \ll p$ puede reducir considerablemente la varianzade los coeficientes. Es de notar que si $M = p$, ajustar el modelo con las combinaciones lineales de los coeficientes originales es equivalente a ajustar el modelo original.

El Análisis de Componentes Principales (ACP) es una técnica que reduce la dimensión de una matriz de datos. La dirección del primer componente principal es aquella en la cual exista mayor variación entre las observaciones, es decir

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

donde, $\sum_{j=1}^p \phi_{j1}^2 = 1$. Los elementos de ϕ son llamados *loadings*, en donde el subíndice representa el número de componente. Juntos, los *loadings* forman el vector de loading de componente principal $\phi_1 = (\phi_{11}\phi_{21}\dots\phi_{p1})^T$.

Dado una matriz de datos X de $n \times p$, se asume que cada variable en X está normalizada -tiene media 0-, entonces se obtiene la combinación lineal de los valores de los predictores, llamadas *scores*

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip}$$

que contiene la mayor varianza. El primer vector de loadings de componente principal resuelve el problema de optimización

$$\max_{\phi_{11}, \dots, \phi_{p1}} \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \quad \text{sujeto a } \sum_{j=1}^p \phi_{j1}^2 = 1$$

El problema de maximización en (formula de arriba) se soluciona mediante la descomposición de los eigenvalores. Luego de determinar el primer componente Z_1 , se procede a encontrar el segundo componente Z_2 , el cual es una combinación lineal de X_1, \dots, X_p que tiene la maximiza varianza de todas las combinaciones lineales que no están correlacionadas con Z_1 . Así los scores del segundo componente principal toman la forma

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$$

donde ϕ_2 es el segundo vector loading del componente principal. Es de notar que restringir Z_2 a no ser correlacionada con Z_1 es equivalente a restringir la dirección de ϕ_2 a ser ortogonal a la dirección de ϕ_1 .

El utilizar la técnica de componentes principales en el modelo de regresión también soluciona el tema de la multicolinealidad entre las variables.

2.2.6. Matriz de Confusión

En los problemas de clasificación se utiliza la matriz de confusión para evaluar el desempeño del modelo. La misma es una tabla que categoriza las predicciones realizadas por el modelo de acuerdo a la coincidencia con los valores reales.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 2.1: Matriz de Confusión

La estrategia solo toma la señal cuando el modelo predice un incremento en el precio, la venta por el contrario no depende del modelo, sino de los parámetros predefinidos (porcentaje de Stop Loss y Horizonte de tiempo). Esta característica implica que el valor a maximizar es la predicción de los verdaderos positivos, conocido como *Precisin*.

$$Precisin = \frac{VP}{VP + FN}$$

2.3. Bases Legales

3.1. Análisis Exploratorio de los datos

3.1.1. Datos OHLC y Fuente de los datos

La estructura de los datos utilizados en el trabajo es de tipo OHLC por sus siglas en inglés Open, High, Low, Close. La misma, resume en 4 registros el comportamiento del precio del activo (Apertura, Cierre, Mínimo y Máximo) en un intervalo de tiempo. En el caso de la presente investigación, de un día. Este tipo de dato provee la información necesaria para cubrir las exigencia del modelo, tanto para la creación de la variable dependiente como para el cálculo de los indicadores técnicos.

Los datos fueron extraídos del portal www.investing.com, uno de los portales financieros con mayor prestigio en el mundo. Fue fundado en 2007 y es conocido por su prestigioso calendario económico y directorio de brokers.

3.1.2. Series de Índices

El universo de estudio está representado por los índices bursátiles de los mercados financieros existentes entre el período 26/10/2008 - 18/01/2019. Un índice bursátil es un promedio de los precios de los activos que representan un mercado o sector determinado. Los mismos sirven como 'benchmark' o referencia de la economía de un país, sector financiero, etc. En el ámbito de los 'hedge funds' son una referencia para medir la rentabilidad de una estrategia de inversión y el riesgo del mercado.

En la presente investigación se utilizan los índices como reflejo del comportamiento de varios activos, de esta manera, se mide la estrategia en un sector y no en un instrumento en específico. Otras de las ventajas de utilizar los índices es que al representar un promedio de varios activos, sus variaciones son menos drásticas. La muestra está constituida por 5 índices bursátiles que representan distintos mercados del mundo: NASDAQ, NIKKEI, FTSE 100, BOVESPA y SP500.

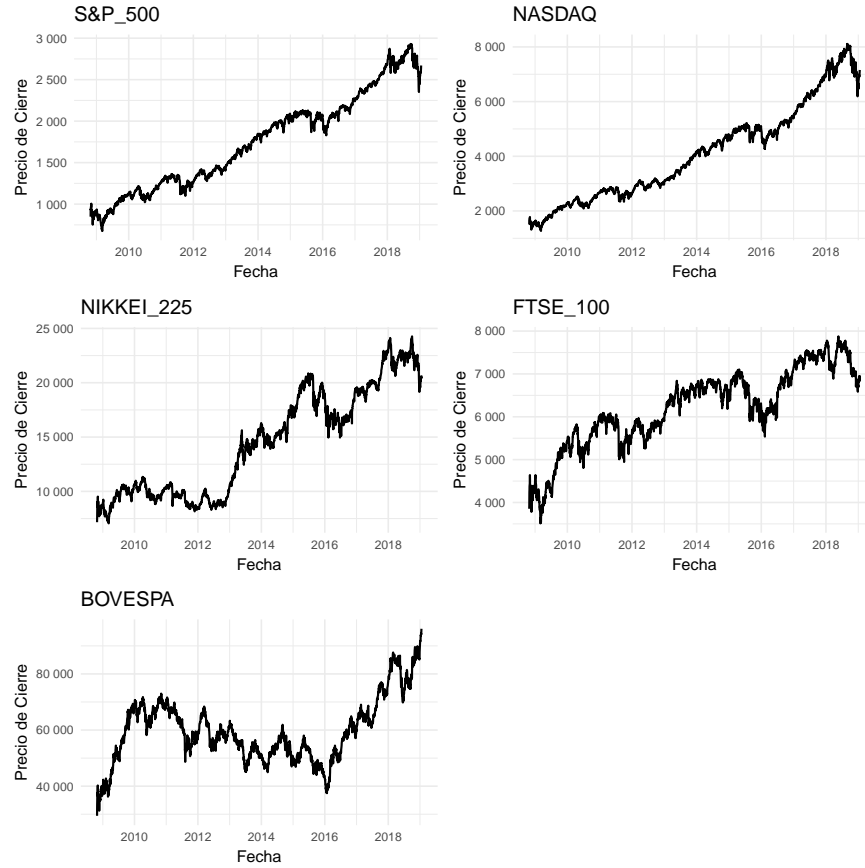


Figura 3.1: Precios de Cierre de los índices en el período de estudio (26/10/2008 - 18/01/2019)

3.2. Entrenamiento del Modelo

3.2.1. Variable dependiente

Las decisiones de entrada en el trading pueden ser producto de muchos factores, en la presente investigación se analiza el enfoque donde se define un porcentaje objetivo de ganancia y se intenta predecir si dicho objetivo se materializará en un futuro cercano, sin que se haya concretado una venta por Stop Loss. Este enfoque reduce la toma de decisión en una variable tal que:

$$P_X(x) = \begin{matrix} p ; & x = c \\ 1 - p ; & x = -d \end{matrix}$$

Dado los datos OHLC del activo es posible identificar los períodos en donde se materializa la variable dependiente. la identificación se realiza, comparando el precio de cierre con los precios máximos y mínimos de las siguientes h observaciones, donde h es el número de períodos, en este caso días en los cuales se desea evaluar la condición.

En la práctica se identifica los registros que cumplen con esta condición añadiendo una columna a la data donde incluimos 'buy' para identificar los registros donde se da la señal y

'stay' en caso de que no haya ocurrido o hubiese ocurrido primero el retroceso del precio.

3.2.2. Indicadores Técnicos como variables predictoras

Los indicadores a utilizar fueron seleccionados buscando recoger la mayor información posible sobre el precio del activo, se pueden resumir en tres categorías: tendencia, momentum y volatilidad.

No es de interés en la presente investigación describir como funciona cada indicador para la toma de decisiones en el trading basado en fundamentos técnicos. Cada indicador puede utilizarse de distintas maneras, calcularse con distintos parámetros y asociarse a discreción del trader, lo que conlleva a un sin fin de reglas de asociación.

Lo que busca la investigación es utilizar la relación entre estos indicadores como variables independientes que ayuden al modelo a predecir oportunidades de entradas. En este sentido se asume la existencia de una dinámica local del mercado que puede ser predecida con ayuda de estos indicadores.

A continuación se presentan los indicadores utilizados:

- Retornos con respecto al precio de Cierre.
- RSI (Relative Strength Index) de 14 períodos, el cual es un indicador de volatilidad.
- MACD (Moving Average Converge/Divergence) el cual es una diferencia de dos EMAs (Exponential Moving Average) de 12 y 26 períodos. Este es un indicador de tendencia que se complementa con un MA (Moving Average) de 9 períodos.
- ADX (Average Directional Index), este es un indicador que utiliza dos indicadores de dirección +Di y -Di, se calculó en base a 14 períodos y mide tendencia.
- Bandas de Bollinger, el cual es un indicador de tendencia y volatilidad, utiliza dos bandas calculadas a partir de una media móvil con desviaciones estándar. Se utilizó en base a 14 períodos y una desviación de 2.5.
- ATR (Average True Range), es un indicador de volatilidad calculado a partir de los máximos y mínimos de un período, en este caso 14.

Estos indicadores están fuertemente correlacionados por lo que se decidió, disminuir el número de variables dejando solo las más representativas de cada indicador.

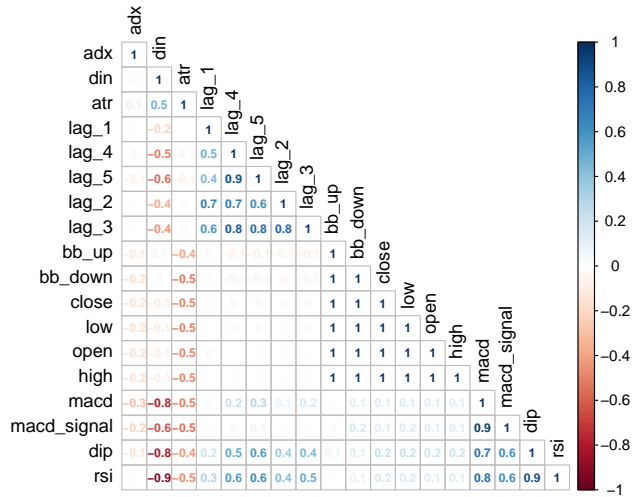


Figura 3.2: Correlación entre indicadores originales calculados con los precios del SP500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

Como se puede observar existe alta correlación entre las distintas variables del precio (Apertura, Cierre, Máximo y Mínimo) por lo que se decidió trabajar solo con los precios de cierre dado que ésta es la misma utilizada para determinar la variable dependiente. Así mismo se observa alta correlación entre los rezagos de los rendimientos, para esto se decidió trabajar solo con los rezagos de 1, 3 y 5 períodos. Por su parte se descarta la variable dip -elemento utilizado en el indicador ADX- por su fuerte correlación con el RSI. Se determina lo mismo para la banda inferior del indicador de Bollinger.

En la figura – se muestra las correlaciones de las variables definitivas

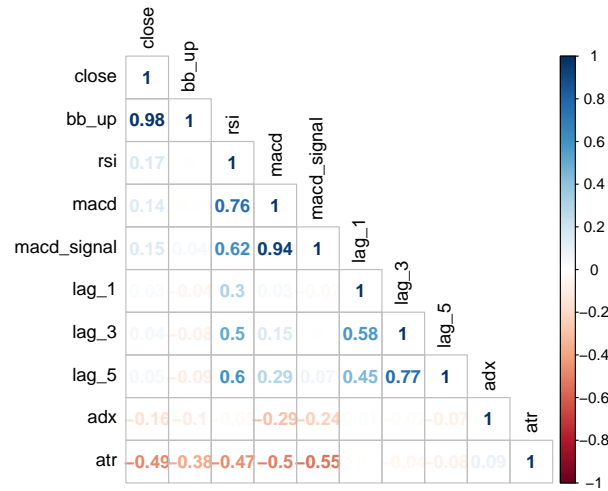


Figura 3.3: Correlación entre indicadores definitivos calculados con los precios del SP500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

Ahora bien, la idea base de la investigación era utilizar los valores de cada indicador como variables predictoras. Dado que el cálculo de todos los indicadores provienen de la misma variable -precio del activo, en la mayoría de los casos precio de cierre-, existe una alta colinealidad entre ellos, la idea de utilizar el ACP es precisamente para enfrentar este problema como se detallará más adelante. Sin embargo, es de notar que los valores de los indicadores por sí solos no proveen un poder predictivo, lo que realmente usa el trader son las asociaciones entre indicadores para encontrar patrones.

Se decidió entonces, utilizar como predictores no los indicadores por si solos, sino, las relaciones entre cada uno de ellos. Esto se abordó agregando al modelo las interacciones entre todos los indicadores, y removiendo los valores de los indicadores por sí solos. De esta manera el hecho de utilizar ACP, no solo es visto ahora como una manera de remover la colinealidad entre predictores sino como método de reducción de variables, ya que el modelo pasó de tener 10 predictores -incluyendo el precio de cierre- a 45.

A continuación se presentan una serie de gráficos para reflejar lo anteriormente expuesto en relación a la colinealidad entre los indicadores.

3.2.3. Validación Cruzada en Series de Tiempo

La validación Cruzada es un método de validación y prueba que consiste en dividir los registros aleatoriamente en grupos de similar tamaño. El primer grupo es utilizado como validación del modelo que ha sido entrenado en el resto de los datos, este proceso se realiza k veces, y el resultado final es el promedio arrojado por cada una de las k validaciones.

Ahora bien este método asume que no existe relación entre las observaciones, es decir que son independientes. Esto no es verdad en el caso de las series de tiempo debido a la condición

de autoregresión. Por lo tanto al dividir la data se debe respetar el orden temporal de cada observación.

3.2.4. WalkForward Backtesting

Al principio de la investigación se implementó el método de entrenamiento, validación y prueba comúnmente utilizado, en donde la mayor parte de la data es destinada a entrenamiento del modelo, otra sección es destinada a validación, para elegir los parámetros óptimos, y finalmente se testeaba el modelo en la data de prueba. Sin embargo este tipo de metodología en opinión del investigador no es el más óptimo para desarrollar el presente modelo, dado el dinamismo de los mercados bursátiles la estrategia no puede permanecer estática en el tiempo.

Para contrarrestar esta situación se optó por el método de *backtestingWalkforward*, el cual consiste en entrenar el modelo en un período base de data, en este caso los primeros 4 años de estudio, posteriormente se aplica la estrategia en el año siguiente y se obtiene los primeros resultados. Luego este año de aplicación es incluido en la data de entrenamiento -es decir, la data de entrenamiento pasa a ser de 5 años- y se evalúa el modelo en el siguiente año. De esta manera, contemplamos el dinamismo del mercado permitiéndole al modelo -y por ende a la estrategia- utilizar el período mas reciente con respecto al cual será implementado. En la presente figura se ilustra la metodología implementada.

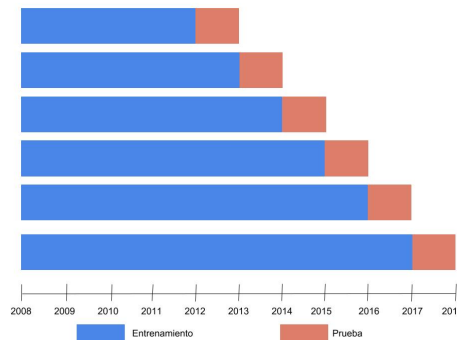


Figura 3.4: Metodología WalkForward

Otra de las características de la metodología que se modificó fue la elección de los parámetros óptimos. Previamente se utilizaba la data de validación para buscar la combinación de parámetros óptima. Ahora bien en la metodología de Walkforward se utilizan los mismo parámetros. A opinión del investigador al buscar los mejores parámetros se estaría incurriendo en un posible sesgo de sobreoptimización. El hecho de que en un año determinado unas configuraciones óptimas den los mejores resultados no asegura que se replique en el siguiente año.

La selección de los parámetros debería ser un estudio previo de la serie financiera a testear. Evidentemente al seleccionar un período de tiempo mas corto se obtendrán menos observaciones que cumplan con el patrón por lo tanto se estaría en presencia de un problema de data imbalanceada que debe tener un tratamiento distinto. Por otro lado el utilizar un horizonte mayor no representa gran cambio en el número de ocurrencias, pero sí en el caso de que la transacción quede abierta -hay recordar que h representa una condición de salida para la estrategia de no ocurrir el target ni el stop loss-. La estrategia implementada en la investigación toma un horizonte de 20 períodos ya que este valor representa un umbral para la ocurrencia del objetivo.

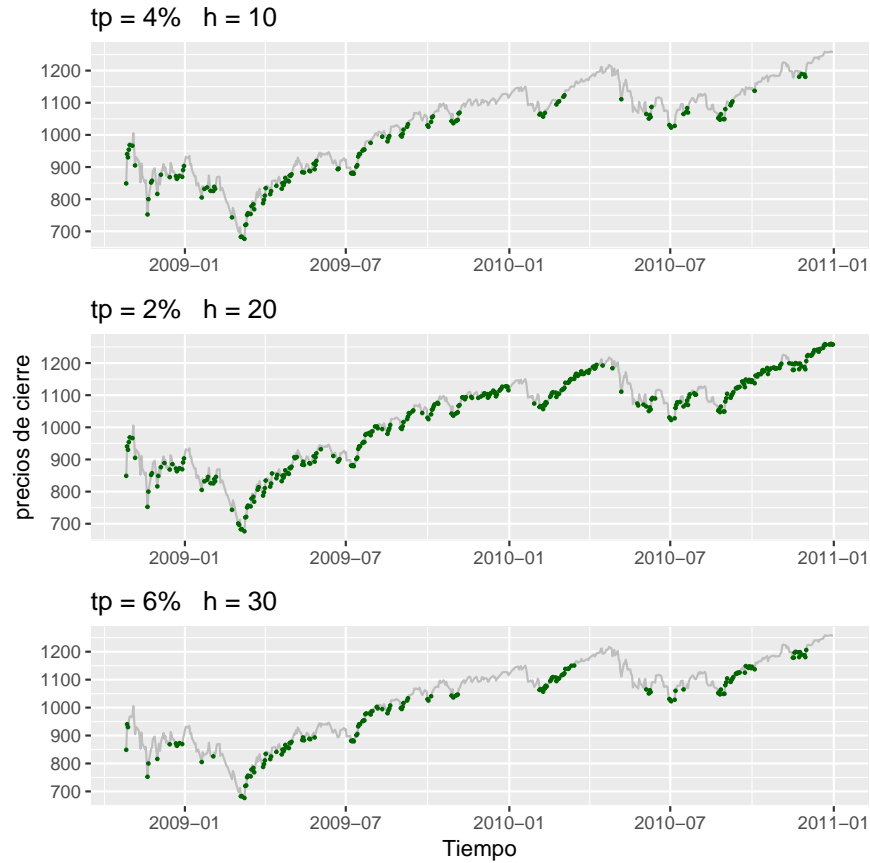


Figura 3.5: Observaciones que presentan ocurrencias con $sl = 2.5\%$, en el primer conjunto de datos de entrenamiento (26/10/2008 - 31/12/2010) para el índice SP500, para diferentes valores de tp y h

3.2.5. Reducción de la dimensión con Análisis de Componentes Principales

Como se señaló anteriormente, el modelo utiliza las interacciones entre las variables predictoras, esto aumenta el número de variables de 10 a 45, las cuales además en muchos casos están correlacionadas. Se decidió entonces utilizar el ACP como método de reducción de las variables.

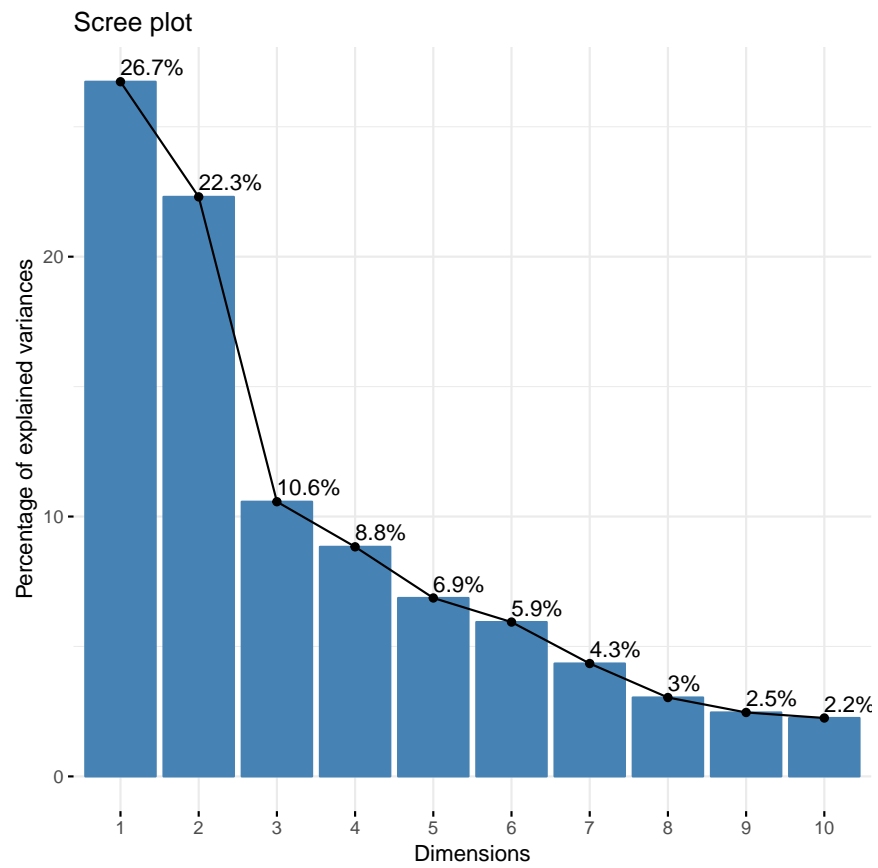


Figura 3.6: d

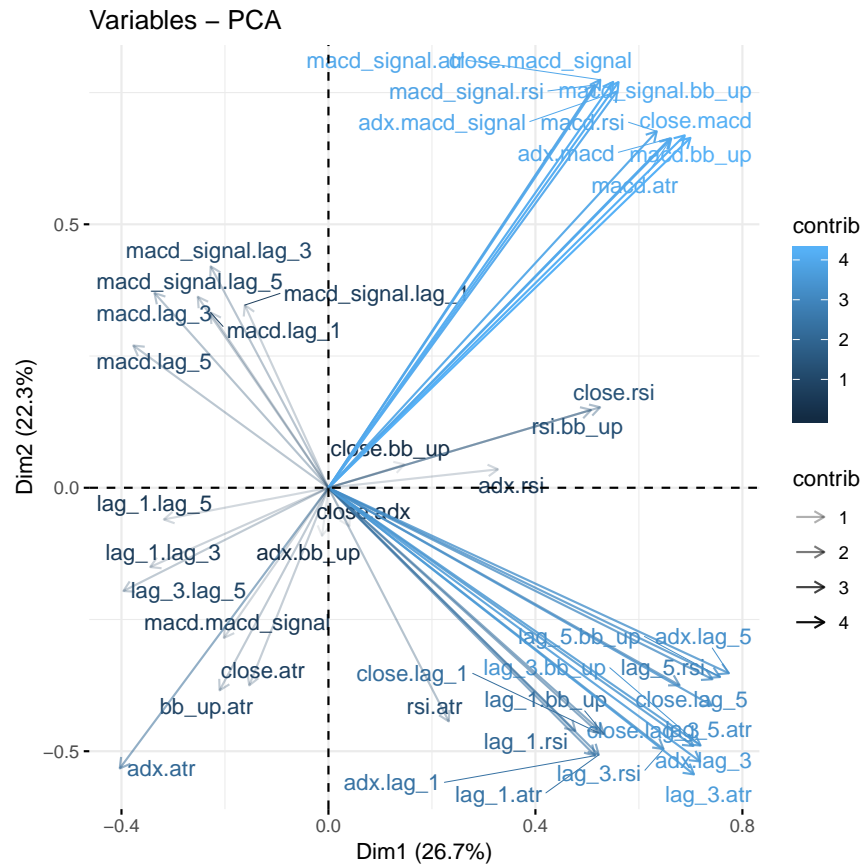


Figura 3.7: d

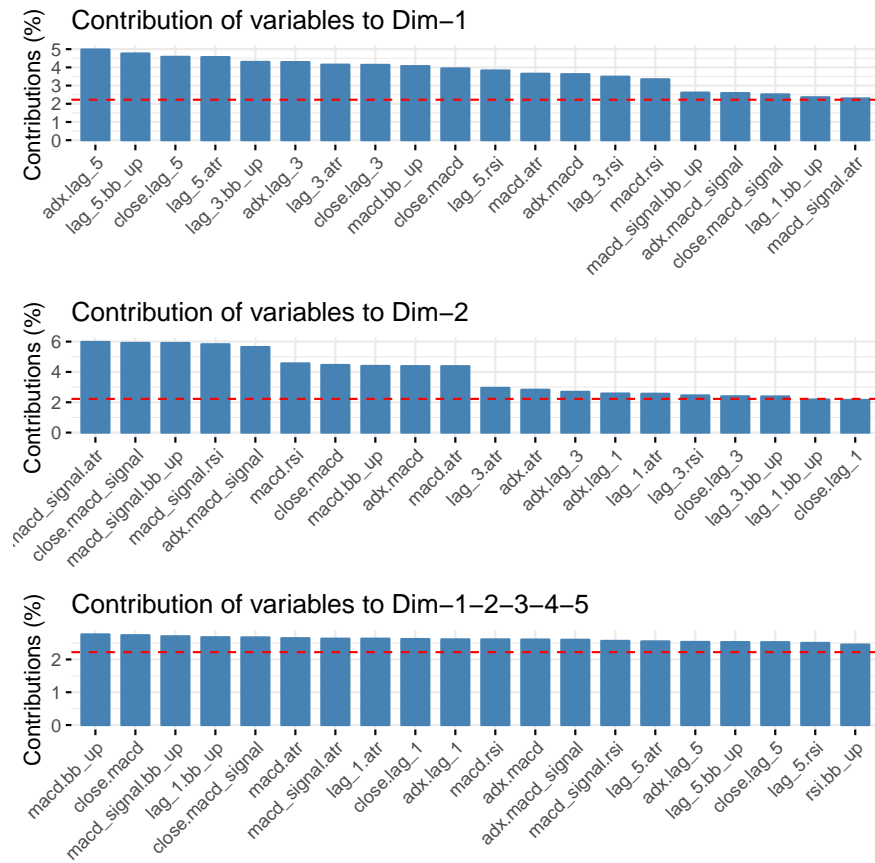


Figura 3.8: d

Análisis de Resultados

En el presente capítulo se realiza la descripción de los resultados obtenidos despues de la aplicación del método propuesto para la estrategia. De igual modo, se presentan los resultados arrojados por las pruebas de Backtesting simulando las entradas y salidas.

4.1. Series de Rendimientos

>

Conclusiones y Recomendaciones

El estudio presentado se centró en confirmar alternativas eficaces para el cálculo del VaR en mercados de materias primas. Los hechos estilizados de las series de tiempo financieras fueron considerados a lo largo del trabajo. Y frente a estos planteamientos, se desarrolló una investigación que permitiera detectar una aproximación con el método GARCH-EVT-COPULAS que tuviese la capacidad de estimar el riesgo del mercado de commodities.

Lista de Referencias

- Alexander, N. Y Zafer, D. (2016). *Gestión del riesgo con el uso de commodities energéticos utilizando Valor en riesgo y Teoría del Valor Extremo*. Suecia, Universidad de Lund.