Algoritmos del comercio de acciones: Pronósticos aplicando técnicas de ML

Ma. Luisa Argáez Salcido

14 de marzo de 2023

Resumen

El objetivo de este ejercicio es pronosticar series temporales financieras utilizando y comparando diferentes algoritmos de aprendizaje automático supervisados de clasificaón para predecir la dirección del mercado de las series temporales financieras, en este documento se intentará pronosticar la dirección diaria de la acción llamada S&P500

1. Introduction

El comercio de acciones o también llamado en inglés "Trading.[es] la compra y venta de valores, como acciones, bonos, divisas y materias primas. El éxito comercial depende de la capacidad del comerciante para ser rentable con el tiempo.Halls-Moore

Ahora bien, el trading algoritmico es el uso de un sistema automatizado para la realización de operaciones, que se ejecutan de forma predeterminada a través de un algoritmo específicamente sin intervención humana.Halls-Moore

Bien se conoce que el aprendizaje automático es un subcampo de la inteligencia artificial que permite a las computadoras realizar tareas sin ser explicitamente programadas, si no que a través de un entrenamiento sean capaces de realizar cálculos por ellas mismas.

En el trading, se utiliza el "backtesting", la cual es una forma de estudiar el rendimiento de una estrategia potencial en la que se aplica a muestras de datos históricos de la vida real.

El trading algortimico presenta ventajas y desventajas. Entre las ventajas es posible mencionar:

Ventajas

- Valoración Histórica. El backtesting, que es el proceso de creación de una estrategia automatizada, es que su rendimiento se puede determinar en los datos históricos del mercado, que (con suerte) son representativos de los datos del mercado futuro. El backtesting permite determinar las propiedades estadísticas (anteriores) de la estrategia, lo que proporciona información sobre si es probable que una estrategia sea rentable en el futuro.
- Eficiencia. El comercio algorítmico es sustancialmente más eficiente que un enfoque discrecional. Con un sistema completamente automatizado, no hay necesidad de que un individuo o equipo esté constantemente monitoreando los mercados para la acción del precio o la entrada de noticias.
- Comparación. Las estrategias sistemáticas brindan información estadística sobre el desempeño histórico y actual.
- Frecuencias más altas. Las estrategias que operan a frecuencias más altas en muchos mercados se vuelven posibles en un entorno automatizado. De hecho, algunas de las estrategias comerciales más rentables operan en el dominio de frecuencia ultra alta en los datos del libro de órdenes limitadas.

Desventajas

- Requerimientos de Capital. El comercio algorítmico generalmente requiere una base de capital mucho más grande que la que se utilizaría para el comercio minorista discrecional, esto se debe simplemente al hecho de que hay pocos corredores que admiten la ejecución automatizada de operaciones que no requieren grandes cuentas mínimas.

Los feeds intradía minoristas comunes a menudo tienen un precio de entre $300 y 500$ por mes, mientras que los feeds comerciales tienen un orden de magnitud superior.

- Buenas equipo de computo.
- Programación/Experiencia científica. Es un requisito para el comerciante algorítmico ser relativamente competente tanto en programación como en modelado científico.Halls-Moore

El comercio algorítmico se distingue de otros tipos de clases de inversión porque es posible proporcionar expectativas sobre el rendimiento futuro de forma más fiable a partir del rendimiento pasado, como consecuencia de la abundante disponibilidad de datos, a esto se le conoce como backtesting.

Razones clave para el backtesting

- Filtración. Se eliminan estrategias que no satisfagan nuestras necesidades de desempeño
- Modelado. El backtesting nos permite de una manera segura, probar nuevos modelos de ciertos fenómenos del mercado, como los costos de transacción,
- Optimización El backtesting permite aumentar el rendimiento de una estrategia modificando la cantidad o valores de los parámetros asociados a esa estrategia y recalculando su rendimiento
- Verificación Las estrategias a menudo se obtienen de forma externa, a través de nuestro pipeline de estrategias

Desventajas del backtesting

- Sesgo de optimización Implica ajustar o introducir parámetros comerciales adicionales hasta que el rendimiento de la estrategia en el conjunto de datos de backtest sea muy atractivo. Sin embargo, una vez en vivo, el rendimiento de la estrategia puede ser notablemente diferente. Otro nombre para este sesgo es .ªjuste de curvas".
- Sesgo de anticipación Se introduce en un sistema de backtesting cuando se incluyen accidentalmente datos futuros en un punto de la simulación en el que esos datos no habrían estado realmente disponibles. Algunos ejemplos son.
  - Errores técnicos
  - Cálculo de parámetros
  - Máximos/mínimos
  - Sesgo de sobreviviente
- El sesgo de supervivencia. Es un fenómeno particularmente peligroso y puede conducir a un rendimiento significativamente ïnflado"para ciertos tipos de estrategias. Ocurre cuando las estrategias se prueban en conjuntos de datos que no incluyen el universo completo de activos anteriores que pueden haber sido elegidos en un momento determinado, sino que solo consideran aquellos que han "sobrevivido"hasta el momento actual.
- Sesgo cognitivo
- Problemas de cambio
  - Consolidación de precios. Datos diarios en forma de cifras Apertura-Alto-Bajo-Cierre o en inglés .ºpen- High- Low -Close"(OHLC), especialmente para acciones. Si la estrategia comercial hace un uso extensivo de cualquiera de los datos .ºHLC.ªspecíficamente, el rendimiento de la prueba retrospectiva puede diferir del rendimiento en vivo, ya que las órdenes pueden enrutarse a diferentes intercambios según su corredor y su acceso disponible a la liquidez. La única forma de resolver estos problemas es hacer uso de datos de mayor frecuencia u obtener datos directamente de un intercambio individual, en lugar de una fuente compuesta más barata.
  - Comercio de divisas y ECN

- Costos de Transacciones Uno de los errores más frecuentes de los principiantes al implementar modelos comerciales es descuidar (o subestimar enormemente) los efectos de los costos de transacción en una estrategia.

    - Comisión. Todas las estrategias requieren alguna forma de acceso a un intercambio, ya sea directamente oa través de un intermediario de corretaje (.[el] corredor"). Estos servicios incurren en un costo incremental con cada operación, conocido como comisión.
    - Deslizamiento (Slippage). El Slippage es la diferencia de precio que se logra entre el momento en que un sistema de negociación decide realizar una transacción y el momento en que se realiza realmente una transacción en una bolsa. El deslizamiento es un componente considerable de los costos de transacción y puede marcar la diferencia entre una estrategia muy rentable y una que funciona mal.
    - Impacto en el mercado. El impacto en el mercado es el costo en el que incurren los comerciantes debido a la dinámica de oferta/demanda del intercambio (y el activo) a través del cual intentan operar. Halls-Moore

La ejecución automatizada es el proceso de permitir que la estrategia genere automáticamente señales de ejecución que se envían al corredor sin intervención humana. Esta es la forma más pura de estrategia comercial algorítmica, ya que minimiza los problemas debido a la intervención humana.

Es importante para una ejecución automátizada definir los siguientes conceptos:

- Latencia. Se define como el intervalo de tiempo entre una simulación y una respuesta
- Elección del lenguaje de programación. Puede ser $C++$, C y Java o bien MATLAB, R y Python
- Tipo de colocación. Un comerciante minorista probablemente ejecutará su estrategia desde casa durante el horario de mercado, encenderá su PC, se conectará a la correduría, actualizará su software de mercado y luego permitirá que el algoritmo se ejecute automáticamente durante el día. Por el contrario, un fondo cuantitativo profesional con importantes activos bajo administración (AUM) tendrá una infraestructura de servidor ubicada en el intercambio dedicado para reducir la latencia en la medida de lo posible para ejecutar sus estrategias de alta velocidad.

La estrategia de abastecimiento.

La identificación de estrategias tiene tanto que ver con las preferencias personales como con el rendimiento de la estrategia, cómo determinar el tipo y la cantidad de datos históricos para la prueba, cómo evaluar desapasionadamente una estrategia comercial y, finalmente, cómo proceder hacia la fase de backtesting y la implementación de la estrategia. Entre las ideas principales a evalucar para una estrategia de abastecimiento se encuentran:

- Identificar sus propias preferencias personales para operar
- Abastecimiento de ideas comerciales algorítmicas
- Evaluación de estrategias comerciales

    - Metodología
    - Relación de Sharpe El ratio de Sharpe caracteriza heurísticamente el ratio riesgo/beneficio de la estrategia. Cuantifica cuánto retorno puede lograr para el nivel de volatilidad soportado por la curva de acciones. Naturalmente, necesitamos determinar el período y la frecuencia en que estos rendimientos y la volatilidad (es decir, la desviación estándar) se miden durante
    - Apalancamiento Responde a la pregunta ¿La estrategia requiere un apalancamiento significativo para ser rentable?
    - Frecuencia La frecuencia de la estrategia está íntimamente relacionada con la tecnología con la que cuenta (y, por lo tanto, la experiencia tecnológica), el índice de Sharpe y el nivel general de los costos de transacción.
    - Volatilidad La volatilidad está fuertemente relacionada con el riesgo"de la estrategia. La relación de Sharpe caracteriza esto. Una mayor volatilidad de las clases de activos subyacentes, si no están cubiertas, a menudo conduce a una mayor volatilidad en la curva de acciones y, por lo tanto, índices de Sharpe más bajos

- Ganancia/pérdida, ganancia/pérdida promedio Las estrategias diferirán en sus características de ganancia/pérdida y ganancia/pérdida promedio. Uno puede tener una estrategia muy rentable, incluso si el número de operaciones perdedoras supera el número de operaciones ganadoras. Las estrategias de impulso tienden a tener este patrón, ya que se basan en una pequeña cantidad de "grandes éxitos"para ser rentables. Las estrategias de reversión a la media tienden a tener perfiles opuestos donde la mayoría de las operaciones son "ganadoras", pero las operaciones perdedoras pueden ser bastante graves.
- Reducción máxima La reducción máxima es la mayor caída porcentual general de pico a valle en la curva de capital de la estrategia. Es bien sabido que las estrategias de impulso sufren períodos de caídas prolongadas (debido a una serie de muchas operaciones perdedoras incrementales). Muchos comerciantes se darán por vencidos en períodos de reducción prolongada, incluso si las pruebas históricas han sugerido que esto es "negocios como de costumbre"para la estrategia. Deberá determinar qué porcentaje de reducción (y durante qué período de tiempo) puede aceptar antes de dejar de operar con su estrategia.
- Capacidad/Liquidez La capacidad determina la escalabilidad de la estrategia para aumentar el capital. Muchos de los fondos de cobertura más grandes sufren importantes problemas de capacidad a medida que sus estrategias aumentan la asignación de capital.
- Parámetros Debe probar y orientar estrategias con la menor cantidad de parámetros posible o asegurarse de tener suficientes cantidades de datos con los que probar sus estrategias.
- Punto de referencia Comparar su trabajo con otros

La Obtención de Datos Históricos

- Datos Fundamentales Esto incluye datos sobre tendencias macroeconómicas, como tasas de interés, cifras de inflación, acciones corporativas (dividendos, división de acciones), presentaciones ante la SEC, cuentas corporativas, cifras de ganancias, informes de cosechas, datos meteorológicos, etc. Estos datos a menudo se usan para valorar empresas. u otros activos sobre una base fundamental
- Datos de noticias –
- Datos de precios de activos –
- Instrumentos Financieros
- Frecuencia Cuanto mayor sea la frecuencia de los datos, mayores serán los costos y los requisitos de almacenamiento
- Puntos de referencia –
- Tecnología

El almacenamiento de datos financieros puede ser complejo y llegar a requerir demasiado esfuerzo. A continuación se muestran conceptos claves al respecto.

- Conjuntos de datos financieros Para el comerciante minorista algorítmico o el pequeño fondo cuantitativo, los conjuntos de datos más comunes son los precios históricos al final del día (EOD) e intradía para acciones, índices, futuros (principalmente materias primas o renta fija) y divisas (forex). Para simplificar esta discusión, nos concentraremos únicamente en los datos de fin de día (EOD) para acciones, ETF e índices de acciones.
- Es posible obtener datos financieros de: Yahoo Finanzas,Finanzas de Google,Cuantificauion.
- Formatos de almacenamiento Hay tres formas principales de almacenar datos financieros. Todos ellos poseen diversos grados de acceso, rendimiento y capacidades estructurales. Entre ellos se encuentran: almacenamiento de archivos planos, almacenamiento de documentos/NoSQL y sistemas de gestión de bases de datos relacionales
- Estructura de datos históricos La primera tarea es definir nuestras entidades, que son elementos de los datos financieros que eventualmente se mapearán en las tablas de la base de datos. Para una base de datos maestra de acciones, preveo las siguientes entidades:

- Intercambios: ¿cuál es la última fuente original de los datos?
- Proveedor: ¿de dónde se obtiene un punto de datos en particular?
- Instrumento/ticker: el ticker/símbolo de la acción o el índice, junto con la información corporativa de la empresa o el fondo subyacente.
- Precio: el precio real de un valor en particular en un día en particular.
- Acciones corporativas: la lista de todas las divisiones de acciones o ajustes de dividendos (esto puede conducir a una o más tablas), necesaria para ajustar los datos de precios.
- Días festivos nacionales: para evitar clasificar erróneamente los días festivos comerciales como errores de datos faltantes, puede ser útil almacenar los días festivos nacionales y las referencias cruzadas.

Evaluación de la precisión de los datos

- Los datos de precios históricos de los proveedores son propensos a muchas formas de error:

  Acciones Corporativas - Incorrecto manejo de splits de acciones y ajustes de dividendos. Uno debe estar absolutamente seguro de que las fórmulas se han implementado correctamente.

  Picos: puntos de precios que superan en gran medida ciertos niveles históricos de volatilidad. Uno debe tener cuidado aquí, ya que estos picos ocurren: vea May Flash Crash para ver un ejemplo aterrador. Los picos también pueden ser causados por no tener en cuenta las divisiones de acciones cuando ocurren. Los scripts de filtro de picos se utilizan para notificar a los comerciantes de tales situaciones.

- Agregación de OHLC: los datos gratuitos de OHLC, como los de Yahoo/Google, son particularmente propensos a situaciones de 'agregación de ticks incorrecta' en las que los intercambios más pequeños procesan transacciones pequeñas muy por encima de los precios de intercambio 'principales' del día, lo que conduce a máximos/sobreinflados. mínimos una vez agregados. Esto es menos un .[er]rorçomo tal, pero más un problema del que hay que tener cuidado.
- Datos faltantes: los datos faltantes pueden deberse a la falta de transacciones en un período de tiempo particular (común en datos de resolución de segundo/minuto de empresas de pequeña capitalización sin liquidez), por feriados comerciales o simplemente un error en el sistema de intercambio. Los datos faltantes se pueden rellenar (es decir, rellenar con el valor anterior), interpolar (linealmente o de otro modo) o ignorar, según el sistema de comercio.
- Clasificación de Mercados e Instrumentos

  Mercados

  Instrumentos

  Datos Fundamentales

  Datos no estructurados: Datos de texto completo y Datos de redes sociales

- Frecuencia de datos La frecuencia de los datos es una de las consideraciones más importantes al diseñar un sistema de comercio algorítmico. Impactará cada decisión de diseño con respecto al almacenamiento de datos, la prueba retrospectiva de una estrategia y la ejecución de un algoritmo. Es probable que las estrategias de mayor frecuencia conduzcan a un análisis estadísticamente más sólido, simplemente debido a la mayor cantidad de puntos de datos y de operaciones que se utilizarán. Las estrategias "High Frequency Trading"(HFT) a menudo requieren una importante inversión de tiempo y capital para el desarrollo del software necesario para llevarlas a cabo. Las estrategias de menor frecuencia son más fáciles de desarrollar e implementar, ya que requieren menos automatización. Sin embargo, a menudo generarán muchas menos transacciones que una estrategia de mayor frecuencia, lo que lleva a un análisis estadísticamente menos sólido.

  Datos Semanales y Mensuales

  Datos diarios Los datos .[En]d Of Day"(EOD) no implican requisitos de almacenamiento particularmente grandes. Hay 252 días de negociación en un año para las bolsas estadounidenses y,

por lo tanto, durante una década habrá 2.520 barras por valor. Incluso con un universo de 10 000 símbolos, son 25 200 000 barras, que se pueden manejar fácilmente dentro de un entorno de base de datos relacional.

Barras Intradía Las estrategias intradía a menudo hacen uso de barras OHLCV por hora, quince, cinco, un minuto o segundo. Los proveedores de feeds intradía, como QuantQuote y DTN IQFeed, a menudo proporcionan barras por minuto o por segundo en función de sus datos de ticks. Los datos en tales frecuencias tendrán muchas barras "faltantes"simplemente porque no se realizaron transacciones en ese período de tiempo

Datos del libro de órdenes y ticks Cuando se completa una operación en una bolsa u otro lugar, se genera un tick. Los feeds de ticks consisten en todas esas transacciones por intercambio. Los feeds de ticks minoristas se almacenan y cada dato tiene una marca de tiempo precisa al nivel de milisegundos. Los datos de ticks a menudo también incluyen el mejor precio de oferta/demanda actualizado

- Fuentes de datos Los datos del mercado financiero proporcionados con una frecuencia diaria diferida o más larga están disponibles libremente, aunque con una calidad general dudosa y el potencial de sesgo de supervivencia. Para obtener datos intradiarios suele ser necesario adquirir un feed de datos comercial. Los proveedores de tales alimentaciones varían enormemente en su capacidad de servicio al cliente, calidad general de alimentación y amplitud de instrumentos.

  Fuentes gratuitas Los datos de barra (bar data) gratuitos al final del día, que consisten en precios de instrumentos de Open-High-Low-Close-Volume (OHLCV), están disponibles para una amplia gama de acciones y futuros estadounidenses e internacionales de Yahoo Finance, Google Finance y Quandl. .

  Fuentes Comerciales Para llevar a cabo operaciones algorítmicas intradía, normalmente es necesario comprar un feed comercial. El precio puede oscilar entre 30 dolares por mes y alrededor de 500 dolares por mes para feeds de "nivel minorista".

- Obtención de datos. Es posible obtener datos de: Yahoo Finanzas, Pandas, Quandl , DTN IQFeed.

- Limpieza de datos financieros Con posterioridad a la entrega de datos financieros de los proveedores, es necesario realizar una limpieza de datos. Desafortunadamente, este puede ser un proceso laborioso, pero muy necesario. Hay varios problemas que requieren resolución: datos incorrectos, consideración de agregación de datos y reposición

- Calidad de los datos Quitar o reemplazar datos contradictorios e incorrectos

- Contratos de Futuros Continuos Consiste en un breve resumen de los contratos de futuros y Formación de un contrato de futuros continuo Halls-Moore

2. Diseño de Experimentos

Desarrollada por Ronald Fisher, el Diseño de Experimentos (DoE , por sus siglas en inglés), es una técnica estadística que estudia el efecto de uno o más factores sobre la media de una variable continua.

Las suposiciones que se plantean es:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_1 : \mu_i \neq \mu_j$$

Es decir se plantea la hipótesis nula es el supuesto en que la media de los grupo es igual mientras que la hipótesis alternativa es que al menos una media de los grupos es diferente.

Es importante tener en cuenta las condiciones iniciales para aplicar este tipo de ANOVA, los cuales son:

- Independencia. Es decir que las muestras son independientes y aleatorias.
- Distribución normal. Esto indica que los datos de cada grupo deben seguir una distribución normal. Este supuesto en ocasiones puede ser ignorado ya que la técnica del ANOVA es robusta.
- Homoscedasticidad, Misma varianza entre grupos. La varianza entre los grupos debe de asemejarse. Este supuesto puede no ser tan estricto si existe el mismo número de observaciones por grupo Amar.

En mi ejercicio se compararon los resultados de predicción clasificación del algoritmo de bosques aleatorios, XG boost y los datos reales. Es importante destacar que los datos no siguen una distribución normal, más bien es binomial, sin embargo este supuesto no será tan extricto para efectos del ejercicio. Por otro lado, el supuesto de Homocedasticidad puede ser laxo en el aspecto de que existe el mismo número de muestras en los gurpos.

2.1. Procedimiento

1. Datos. Se redistribuyeron los datos partiendo de la tabla 1 a la tabla 2 para llevar la técnica estadística adecuadamente.

Bosques Aleatorios	XG Boost	y_real
1	0	0
1	1	0
1	1	0
...	...	...

Cuadro 1: Tabla 1 . Datos Originales

Variable	Valor
xgboost	0
bosque aleatorio	1
y_{real}	1
...	...

Cuadro 2: Tabla 2 . Datos transformados

2.2. Descriptivos generales

Los datos en cada grupo se observan de la siguiente manera.

Variable	Número de muestra	Media	Desviación estándar
y_real	100	0.41	0.4943
Bosque Aleatorio	100	0.91	0.2876
XG Boost	100	0.85	0.3588

Cuadro 3: Tabla 3. Descriptivos por grupo.

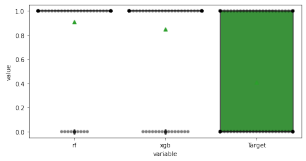

Figura 1: Gráfico de caja y bigotes para cada grupo

2.3. Supuestos

Es importante validar los supuestos de normalidad y homocedasticidad.

Para el supuesto de normalidad se utilizó la prueba de normalidad de Shapiro- Wilk y los resultados fueron que para bosques aleatorios, XG Boos y y_{real} los valores p son 2.17e-19, 5.72 e-18 y 1.21 e-14 respectivamente, lo cual indica que se rechaza la hipótesis nula de que los valores provienen de una distribución normal.

Para la prueba de homocedasticidad presento un valor p de 0.65 e-08, lo cual indica que la varianza entre las grupos no es igual.

Dicha conclusiones ya se habian discutido al inicio del análisis sin embargo es importante llevar la metodología adecuadamente.

2.4. Resultados ANOVA del diseño de experimentos.

Fuente	SS	DF	MS	F	p-unc	np2
variable	14.90	2	7.4533	49.05	4.9195e-19	0.2448
Within	45.13	297	0.1519			

Cuadro 4: Tabla 4. Resultados

2.5. Conclusiones

Es importante destacar que los datos no siguen una distribución normal y no cumplen con el supuesto de homocedasticidad. Sin embargo se continuo con el Diseño de Experimentos y se concluyo que como el valor p es muy cercano a cero y el alfa de 0.05. Se conoce que se rechaza H_0 si pvalor es menor a α . Entonces, se rechaza H_0 , los grupos que se compararon no tienen la misma media.

2.6. Código

Para ver el código se inserta el link del google colaboratory Código

Referencias

Michael L. Halls-Moore. *Successfull Algorithmic Trading.* URL https://www.quantstart.com/successful-algorithmic-trading-ebook/.

Joaquin Amar. Analisis de varianza (anova) con python. URL https://www.cienciadedatos.net/documentos/pystats09-analisis-de-varianza-anova-python.html.