Trabajo de fin de Máster Data Science

Estudio sobre la aplicación de técnicas de Machine Learning en el análisis fundamental de valores cotizados y en la selección de carteras

José Felipe Cuesta Cabot

Introducción

En este trabajo de fin de master se estudia la aplicación de técnicas de machine learning en la selección de una cartera de inversión, y su rendimiento frente a un índice de referencia en el medio plazo. Para ello se pretende averiguar la posible capacidad de predicción del comportamiento relativo de los valores frente a la media del conjunto analizado.

La predicción de los mercados financieros es un objetivo ampliamente estudiado desde hace ya mucho tiempo. En este sentido, las estrategias de análisis más extendidas son las conocidas como “análisis fundamental” y “análisis técnico”. Muy resumidamente, se puede definar la primera de ella como el intento de estimar el valor “real” de una compañía, en función de su información financiera y determinadas estimaciones de mercado, para compararlo con el valor “de mercado” ímplicito en el precio de cotización y asumir que a la larga, este precio “de mercado” tenderá a aproximarse a nuestra estimación “real”.

Por su parte, el análisis técnico, intenta predecir el comportamiento del precio futuro de un valor al asumir que en determinadas circunstancias, los precios siguen determinados patrones fijos y repetitivos.

Ambas estrategias son susceptibles de ser analizadas mediante técnicas de machine learning, y en el presente trabajo intentaremos aplicar, tentativamente y siempre sujetos a la disponibilidad de información, los principios del análisis fundamental.

Para ello, utilizamos ratios financieros y de mercados como variables explicativas, mientras que la variable objetivo es una etiqueta a la que hemos denominado ‘mejor’ y ‘peor’, y que refleja si el el periodo de 180 días, un determinado valor mejora o empeora el rendimiento del índice de referencia.

Una vez preparados los datos, probamos a entrenar diferentes modelos de machine learning. El objetivo es utilizar el modelo que mejor precisión obtenga para seleccionar los valores con mayor probabilidad de comportarse ‘mejor’ que el índice. Para probar el rendimiento del método, generamos pruebas sucesivas en las que utilizamos como población de test los valores en un periodo determinado, y le aplicamos el modelo entrenado con los datos anteriores a dicho periodo, para simular una situación real de decisión de inversión basada en la información disponible en ese momento. Por último, comparamos la rentabilidad de esta cartera con respecto al índice de referencia.

Descripción de los datos

Los ficheros que vamos a tratar se han obtenido de la web simfin https://simfin.com/data/bulk, que proporciona gratuitamente datos de compañías cotizadas en EEUU desde 2007. Esta web ofrece información más completa mediante una modalidad de pago que podría ser muy interesente utilizar para profundizar en este análisis. Para mi trabajo, los ficheros utilizados han sido los siguientes:

- **Companies:** Información general de todas las compañías cotizadas, incluyendo el código de su subsector económico.
- **Share Prices:** Cotizaciones diarias de los valores disponibles. El tamaño de este fichero excede la capacidad de GitHub, por lo que en este repositorio se encuentra filtrado con los valores que nos interesan.
- **Sector/Industry:** Detalle de cada sector y subsector.
- **Balance sheet:** Información contable del balance, por trimestres.
- **Income sheet:** Información contable de la cuenta de pérdidas y ganancias, por trimestres.
- **Cash Flow:** Información contable del flujo de caja, por trimestres.

Con respecto a los tres últimos ficheros, SimFin divide los datos en tres grupos de empresas: general, bancos y compañías de seguros. Esto se debe a que la normativa contable, y por tanto, los Estados Financieros de estos dos últimos sectores difieren de los del resto.

Al ejecutar el notebook 'prep_datos.ipynb' se procesan estos ficheros y se genera una única tabla con un registro por valor y trimestre, en el que se incluyen los datos relevantes contable-financieros del periodo.

Metodología

Con el fin de evitar la influencia de factores exógenos específicos sobre compañías del mismo sector, agrupamos nuestro análisis por industrías homogéneas. De este modo, es más fácil que el comportamiento bursátil de un valor concreto frente al resto se pueda predecir por valores financieros propios y no por otros factores. En nuestro trabajo hemos elegido los sectores energético, sanitario, industrial y tecnológico.

Por otro lado, como la disponibilidad de datos, tanto de precios como información financiera es limitada, el estudio lo realizamos sobre una muestra acotada de valores para los que sí disponemos información en un periodo dado. Por ello, en cada industria analizada seleccionamos un grupo de valores con información constante durante el mayor tiempo posible. De este modo nos quedamos con el periodo comprendido entre el 1 de octubre de 2013 y el 30 de junio de 2019.

En el notebook "etiquetado" es donde asignamos a cada registro la etiqueta correspondiente ("mejor" o "peor"). Para ello, primero normalizamos los precios fijando todos los precios a 1 de octubre de 2013 a base 100, y calculando las variaciones porcentuales a partir de dicha referencia.

Asimismo, creamos 4 índices sectoriales calculados como la media diaria del precio en base 100 de cada grupo de valores.

Por último, para cada registro, calculamos la variación de precio con respecto al precio futuro en 180 días, y lo comparamos con la variación del índice correspondiente en el mismo periodo temporal. Si el comportamiento del valor mejora al índice lo etiquetamos como "mejor", en caso contrario como "peor".

Las ariables explicativas que alimentan al modelo de machine learning se han calculado en el notebook "ratios", en donde se utiliza la información financiera trimestral por valor descargada de la web Simfin. Con la información disponible calculamos los siguientes ratios básicos de análisis financiero y contable: PER, Price to Book Value, ROE, ROA y Deuda corto plazo / FFPP

Además, para cada uno de estos ratios hemos calculado la variación con respecto al periodo anterior, el percentil del ratio con respecto al resto de valores del sector, y la variación de este percentil con respecto al periodo anterior.

El tratamiento y preparación de los datos explicado hasta el momento se ha realizado utilizando la librería Pandas de Python. Por su parte, los modelos de machine learning se han aplicado con la librería Scikit-learning en el notebook "modelos". En concreto, los algoritmos de clasificación utilizados son:

- Regresión Logarítmica
- Nearest Neighbors Classification
- Decission Tree Classifier

Resultados de los modelos

Tras aplicar seleccionar las variables para alimentar el modelo teniendo en cuenta la correlación entre ellas, y utilizando técnicas de cross validation, obtenemos los siguientes resultados en la precisión de los modelos.

Accuracy	LR	KN	DT
Energy	0.5031	0.5088	0.5119
Industrial	0.5347	0.5347	0.5267
Healthcare	0.5378	0.5316	0.5526
Technology	0.5311	0.5412	0.5377

Simulación de carteras

El resultado de la aplicación de los modelos no ha sido buena, no superando en ningún caso significativamente los resultados esperados de intentar predecir el comportamiento completamente al azar.

Sin embargo, todavía podemos intentar testear la posibilidad de automatizar la selección de una pequeña cartera de valores que mejore los resultados del índice de referencia de cada sector.

Es razonable asumir que para poder obtener un buen resultado de inversión no necesitamos predecir el comportamiento de todos los valores de la población, por lo que vamos a simular la creación de carteras utilizando las funciones predict_proba de cada modelo, y seleccionaremos los valores a los que cada modelo otorgue mayor probabilidad de mejorar su índice correspondiente, con la esperanza de que a la larga sí obtengamos mejores resultados que el índice.

Hemos preparado un notebook “test_cartera” en el que aplicamos en bucle el modelo de Logistic Regression (dado que ningún modelo a priori parece mejorar a los demás, optamos por el más rápido), de forma que se genera automáticamente una cartera para cada periodo de los 10 últimos disponibles, utilizando en cada periodo como set de entrenamiento la información de los periodos anteriores. De este modo intentamos simular una situación real en la que en cada periodo únicamente disponemos de información histórica.

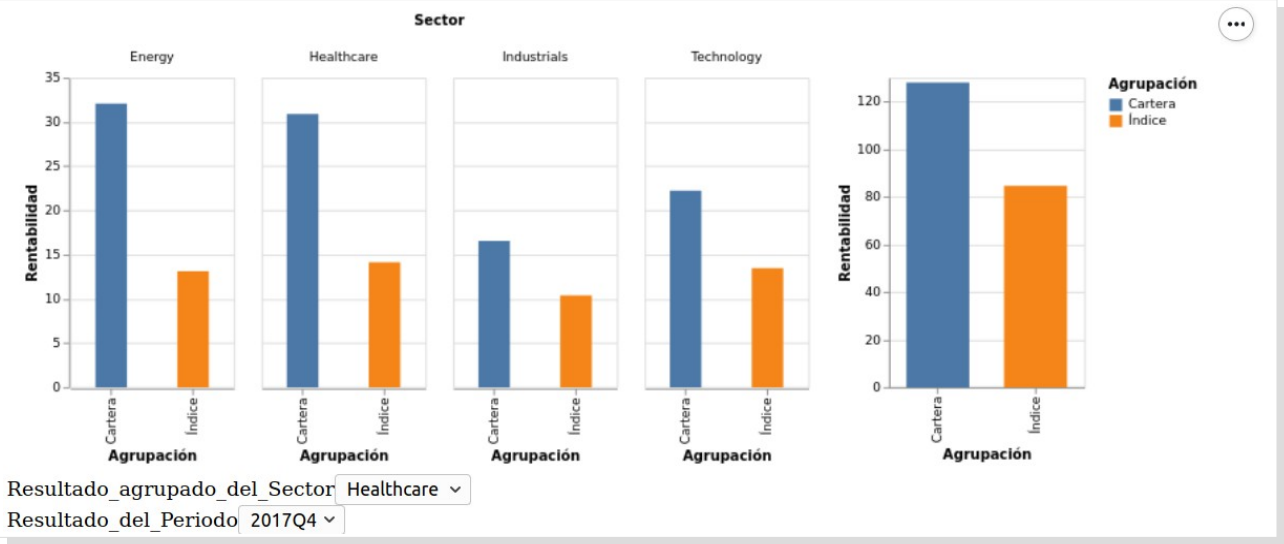
Resultados de la simulación de carteras

En el siguiente cuadro se resume el resultado agregado obtenido en las simulaciones:

Accuracy	N.º de periodos que la cartera mejora al índice	Rentabilidad acumulada de la cartera	Rentabilidad acumulada del índice
Energy	9	47.44 %	-18.33 %
Industrial	8	111.14 %	81.99 %
Healthcare	5	97.37 %	84.65 %
Technology	6	134.94 %	84.72 %

Visualización de la simulación de carteras

Al ejecutar cada uno de los notebooks de “test_cartera” se vuelcan los resultados en un fichero csv con el objetivo de procesarlo y generar una representación gráfica interactiva del resultado obtenido. Esta representación gŕafica la generamos en el notebook “visualización” en el que utilizamos el paquete *altair* para obtener 3 ficheros html con el siguiente aspecto:



En el gráfico es posible seleccionar el periodo deseado para comprobar como se comporta la cartera generada por nuestro modelo frente al índice del sector.

También es posible modificar el sector para verificar la rentabilidad acumulada de los 10 periodos en el gráfico de barras de la derecha.

Conclusiones

En ninguno de los modelos desarrollados en este trabajo ha sido posible obtener resultados significativamente mejores que el modelo base. Por lo tanto, al menos en cuanto al alcance aquí planteado, los resultados refuerzan la teoría de los mercados eficientes.

No obstante, se presentan indicios de que sí es posible utilizar los modelos de machine learning para seleccionar una cartera reducida de valores que mejore la rentabilidad media de los valores utilizados como Benchmark.

Instrucciones para replicar el análisis y generar la visualización

Al clonar el repo en GitHub se habrán obtenido los ficheros de datos descargados de SimFin y los notebooks para cada fase del trabajo. A continuación deberán ejecutarse sucesivamente, utilizando jupyter-notebook, los siguientes notebooks:

1. prep_datos.ipynb
2. etiquetado.ipynb
3. ratios.ipynb
4. test_carteras_LR.ipynb
5. visualización.ipynb

Tras ello, se habrán generado los ficheros Rentabilidad_LR.html, Rentabilidad_KN.html y Rentabilidad_DT.html, que podrán abrirse con cualquier navegador para ver los resultados de las carteras generadas con cada uno de los tres modelos.