Tipología y Ciclo de Vida de Datos

PRA1 - Scrap Yahoo Finance

Antonio Caparrini López - 10 de noviembre de 2019

Contexto

Desde la aparición de las acciones y las bolsas de valores mucho esfuerzo e investigación se ha dedicado a estudiar el comportamiento y la predicción del valor de una acción a futuro. Yahoo finance ofrece datos financieros de las empresas y noticias que pueden ser utilizados para tomar decisiones de inversión. En este proyecto planteamos la extracción de los datos financieros y las noticias asociadas a las empresas pertenecientes al IBEX35 (https://en.wikipedia.org/wiki/IBEX_35), que son las 35 empresas de mayor capitalización bursátil en la bolsa de Madrid.

Título para el dataset

El título para el dataset es **IBEX35_Dataset** ya que el índice del que está tomando el listado de empresas es lo suficientemente descriptivo para comprender que contiene y además es un título breve.

Descripción del dataset

El dataset consta de 3 ficheros distintos que pasamos a describir con su contenido:

- **yahoo_news.csv**
    1. **ticker**: Identificador de la empresa.
    2. **date**: Fecha de la publicación de la noticia.
    3. **url**: URL de la noticia.
    4. **content**: Contenido de la noticia preprocesado.
- **yahoo_statistics.csv**
    1. **ticker**: Identificador de la empresa.
    2. **Market Cap (intraday)**: Valor de la empresa en bolsa.
    3. **Enterprise Value**: Valor de la empresa.
    4. **Trailing P/E**: Valor de la acción entre las ganancias por acción de los últimos 12 meses.

5. **Forward P/E**: Predicción de Thomson Reuters.
6. **PEG Ratio (5 yr expected)**: Indicador que relación el precio con las ganancias y el crecimiento de la empresa.
7. **Price/Sales**: Precio/Ventas.
8. **Price/Book**: Valor en acciones / Valor en libros.
9. **Enterprise Value/Revenue**: Valor de la empresa / Ganancias.
10. **Enterprise Value/EBITDA**: Valor de la empresa / EBITDA.
11. **Beta (3Y Monthly)**: Métrica de volatilidad.
12. **52-Week Change**: Porcentaje de cambio en 52 semanas.
13. **S&P500-Week Change**: Cambio semanal en el S&P.
14. **52 Week High**: Valor más alto en 52 semanas.
15. **52 Week Low**: Valor más bajo en 52 semanas.
16. **50-Day Moving Average**: Valor medio de los últimos 50 días.
17. **200-Day Moving Average**: Valor medio de los últimos 200 días.
18. **Avg Vol (3 month)**: Volumen medio 3 meses.
19. **Avg Vol (10 day)**: Volumen medio 10 días.
20. **Shares Outstanding**: Cantidad de acciones de la empresa.
21. **Float**: Reajuste.
22. **% Held by Insiders**: Porcentaje de acciones propiedad de la empresa.
23. **% Held by Institutions**: Porcentaje de acciones propiedad de instituciones.
24. **Shares Short**: Acciones en corto.

25. **Short Ratio**: Acciones en corto / Volumen.

26. **Short % of Float**: Porcentaje de corto respecto a la empresa total.

27. **Short % of Shares Outstanding**: Porcentaje de corto respecto a la empresa total.

28. **Shares Short (prior month)**: Acciones en corto desde el mes pasado.

29. **Forward Annual Dividend Rate**: Estimación del dividendo.

30. **Forward Annual Dividend Yield**: Estimación del dividendo.

31. **Trailing Annual Dividend Rate**: Dividendo de los últimos 12 meses.

32. **Trailling Annual Dividend Yield**: Dividendo de los últimos 12 meses.

33. **5 Year Average Dividend Yield**: Dividendo de los últimos 5 años.

34. **Payout Ratio**: Porcentaje de ganancias repartidas como dividendo.

35. **Dividend Date**: Proxima fecha de pago de dividendo.

36. **Ex-Dividend Date**: Proxima fecha en la que las acciones se transmiten sin derecho al dividendo.

37. **Last Split Factor (new per old)**: Nuevas por antiguas.

38. **Last Split Date**: Última fecha de partición de acciones.

39. **Fiscal Year Ends**: Fin del año fiscal

40. **Most Recent Quarter**: Último cuatrimestre.

41. **Profit Margin**: Margen de beneficios.

42. **Operating Margin**: Margen operativo.

43. **Return on Assets**: Beneficios por activos.
44. **Return on Equity**: Beneficios por patrimonio neto.
45. **Revenue**: Beneficios.
46. **Revenue Per Share**: Beneficios por acción.
47. **Quarterly Revenue Growth**: Crecimiento cuatrimestral.
48. **Gross Profit**: Beneficio bruto.
49. **EBITDA**: Beneficio después de impuestos y amortizaciones.
50. **Net Income Avi to Common**: Beneficio neto.
51. **Diluted EPS**: Métrica sobre acciones convertibles.
52. **Quarterly Earnings Growth**: Crecimiento en beneficios cuatrimestrales.
53. **Total Cash**: Cash líquido en la empresa.
54. **Total Cash Per Share**: Cash por acción.
55. **Total Debt**: Total de deuda.
56. **Total Debt/Equity**: Deuda / Patrimonio neto.
57. **Current Ratio**: Ratio de líquidez.
58. **Book Value Per Share**: Valor en libro por acción.
59. **Operating Cash Flow**: Medida de la cantidad de flujos de caja generados.
60. **Levered Free Cash Flow**: Cash sobrante de hacer frente a todas las obligaciones.

- **yahoo_summaries.csv**

1. **ticker**: Identificador de la empresa.
2. **Previous Close**: Valor de la acción en último cierre.

3. **Open**: Valor de la acción en la última apertura.
4. **Bid**: La cantidad de acciones que se ofertan al precio más bajo.
5. **Ask**: La cantina de acciones que se demandan al precio más alto.
6. **Day's Range**: Valor mínimo y máximo del día.
7. **52 Week Change**: Valor mínimo y máximo de las últimas 52 semanas.
8. **Volume**: Cantidad de acciones que se ofertan o demandan.
9. **Avg. Volume**: Volumen medio.
10. **Market Cap**: El valor de la empresa contando el valor de las acciones.
11. **Beta (3Y Monthly)**: Métrica de volatilidad.
12. **PE Ratio (TTM)**: Precio por ganancias.
13. **EPS (TTM)**: Ganancias por acción.
14. **Earnings Date**: Fecha de anuncio de las cuentas de la empresa.
15. **Forward Dividend & Yield**: Estimación de los dividendos del año expresados como porcentaje del valor de la acción.
16. **Ex-Dividend Date**: Fecha a partir de la cuál no se tiene derecho a recibir el dividendo.
17. **1y Target Est**: Estimación del valor de la acción a un año.

Inspiración

Este conjunto de datos consta de 3 ficheros de los cuáles el de noticias se diferenciaría mucho de los otros dos por constar de texto. Podría utilizarse para realizar **tareas de clustering** entre las empresas tanto en función de sus estadísticas y sumarios financieros o por las noticias. También podrían utilizarse junto con los datos de precios de las acciones

para crear **modelos predictivos** que pronostiquen el precio futuro. Los datos de precios son más fáciles de conseguir en una API que los sumarios y estadísticas que ofrece Yahoo.

Licencia

La licencia elegida es **Released Under BY_NC-SA 4.0 License** (https://creativecommons.org/licenses/by-nc-sa/4.0/).

Esto es debido a que los datos que cogemos de Yahoo no pueden utilizarse con fines comerciales, por lo que no podemos elegir permitir uso comercial además obligamos que las copias/modificaciones sobre el dataset cumplan con estas mismas condiciones.