Entrega guía 4 semana 6: Anteproyecto grupo Stanley

Integrantes: Alexis Abreu, Edgar Balaguera**,** María Paula Cortés Riaño, Ricardo Pretelt Villadiego

Objetivo

Identificar las técnicas para el preprocesamiento de datos requeridos para el problema de Analytics identificado.

1. Parte 1. Características de la calidad de los datos

Para el proyecto se usa el data set de valores de cierre para las acciones de las empresas que hacen parte del S&P500. En total se tiene información de 503 empresas. La información se da por fecha en un rango iniciando el 3 de enero de 2022 y terminando el 18 de noviembre de 2022. Los días no hábiles no opera la bolsa de valores por lo que no existen registros para estos días.

Se desarrolló un microservicio de almacenamiento de datos que alberga la información de los precios de cierre de los 503 símbolos del índice SP&500 y los actualiza todos los días a las 7:00 PM. De esta forma se asegura tener un insumo que puede ser llamado desde diversas plataformas de inteligencia de negocio y desde Python para el análisis y procesamiento al que haya a lugar.

Origen de los datos

Se deja una relación de los orígenes de datos y de las herramientas necesarias para la implementación de este workflow:

- yfinance librería de Python, puede encontrarla en el siguiente link (https://pypi.org/project/yfinance/)
- Cuenta de Google
- Configuración de una cuenta de servicios en GCP y habilitación de las APIs de Google Drive y Spreadsheets
- Lenguaje computacional Python

Se puede obtener acceso a la base de datos en cualquier momento mediante el siguiente enlace https://docs.google.com/spreadsheets/d/e/2PACX-1vQJhwKp9kXl0ta8NsK9oKwNOPw8pBNldvR2_am8SOuJidAuywapixZ03kL__IqAOZZUsX1ctCGd1lfp/pub?output=csv

Inicialmente los datos provienen de una única fuente, y es Yahoo! Finance. En esta página se encuentran los valores de las acciones para las empresas del S&P500.

Nivel de granularidad

El nivel de detalle que se tiene para los datos es el valor de cierre para la acción en bolsa de valores de cada día hábil. Estos datos son provenientes de una única fuente, se obtienen los valores de acciones para cada día de las empresas en estudio.

Fidelidad y exactitud

Los valores que se obtienen de la fuente son las ventas sin transformaciones. Estos están en su mayoría completos por lo no hay riesgos de sesgos por falta de información, o por errores en medición.

Nivel de integridad dentro y a través de los almacenes de datos.

Se utiliza sólo una fuente de datos y se integra usando un microservicio para siempre tener los datos actualizados. Esta información permite derivar hechos porque se puede agregar y estudiar las distintas empresas.

Tiempo de accesibilidad

Se utiliza el servicio en línea de Yahoo! Finance, que ha sido confiable, además se usa un microservicio para mantener actualizado los datos. El tiempo de accesibilidad es mínimo y no se encuentran “teóricamente” retrasos por estos casos.

Entendimiento de su ubicación

Dentro de la cadena de suministro el proyecto y sus desarrolladores reciben los datos tal cuál son generados por la fuente externa, son procesados y serán presentados en un tablero de control o visualizaciones para que los interesados que toman decisiones los vean y generen valor.

Edad

Los datos obtenidos son muy recientes ya que son de este año, por lo que son relevantes y tienen mucho potencial para crear valor entendiendo su dinámica.

Habilidad de su entendimiento

Los datos son muy fáciles de comprender ya que son básicamente valores de acciones a través del tiempo. Su entendimiento puede ser fácilmente revelado a través de su uso.

Características de calidad

Totalidad de los datos en la base: De acuerdo con el análisis exploratorio se observa que los datos tienen una completitud del **99.6%**, esto debido a que no se tiene toda la información del valor de las acciones para todos los días.

Consistencia: Hay coherencia entre el nombre de los campos y los datos, ya que cada variable es el símbolo de una empresa y su valor es el precio de la acción para cada día.

Claridad: Los datos son entendibles y se le pueden aplicar todas las técnicas estadísticas necesarias.

Formato: Los datos poseen un formato coherente, en su mayoría son el valor de las acciones de las empresas en dólares, y las fechas que corresponden, como dd/mm/AAAA.

2. Parte 2. Técnicas de limpieza de datos

Para emplear las técnicas de limpieza de datos, se parte del análisis e identificación de los *Datos no válidos* y de *Valores nulos* a fin de determinar si son candidatos para eliminarse o si se puede aplicar alguna técnica de imputación que permita alcanzar la máxima proporción posible de validez de los datos para los análisis consecuentes.

Identificación de Datos no válidos y Valores nulos

Dada la naturaleza de los datos, no se debería dar la nulidad o invalidez de estos. Sin embargo, el 1.8% de los datos de la base corresponden a valores nulos y el 0.5% son valores no válidos.

Dependiendo de cada uno de los casos, se dirá que los datos serán eliminados o imputados mediante alguna técnica, como se describe a continuación.

Eliminación de datos

Para los datos de los símbolos como **BRK.B** y **BF.B** se considera la eliminación de datos, de primera mano, dado que no contienen ningún dato a lo largo de las dimensiones de la base de datos lo que inhabilita la imputación de datos.

A partir de esta situación, definimos un umbral mínimo de completitud de datos para garantizar la máxima cantidad posible de información. El umbral es del 90% y la base tiene una tasa de completitud del **99.6%** por lo que se detiene el proceso de eliminación de datos.

Técnicas de imputación de datos

El 77.8% de los datos nulos (1.8% de la totalidad de la base) se consideran imputables mediante la técnica de KNN configurado con 2 datos vecinos. Esto garantiza que los datos nulos son calculados e imputados a partir de los 2 datos más cercanos que permiten guardar consistencia en el comportamiento de los datos considerando que estamos trabajando con datos financieros que difícilmente pueden res representados a partir de las medias y comportamiento de toda la serie de datos.

Una vez eliminados los datos mencionados y ejecutada la técnica de imputación, obtenemos un porcentaje de validez de la base del **100%**

3. Parte 3. Entendimiento de los datos

Para tener un conocimiento más profundo de los datos y de la forma en que pueden generar valor de cara al análisis que se pretende, consideramos que las técnicas apropiadas para su entendimiento son Filtración (o eliminación de valores atípicos), Imputación de datos, Descripción de estadísticas básicas, Determinación de variables relevantes y Agrupación de los datos así:

Filtración

La naturaleza de la información que usaremos para el proyecto no considera la eliminación de datos por atipicidad o según su valor. Sin embargo, para poder llevar a cabo los análisis que nos ocupan, es necesario contar con los valores de los datos a lo largo de toda la serie de forma que efectivamente podamos ejecutar los modelos y garantizar su alto nivel de exactitud en la entrega de resultados.

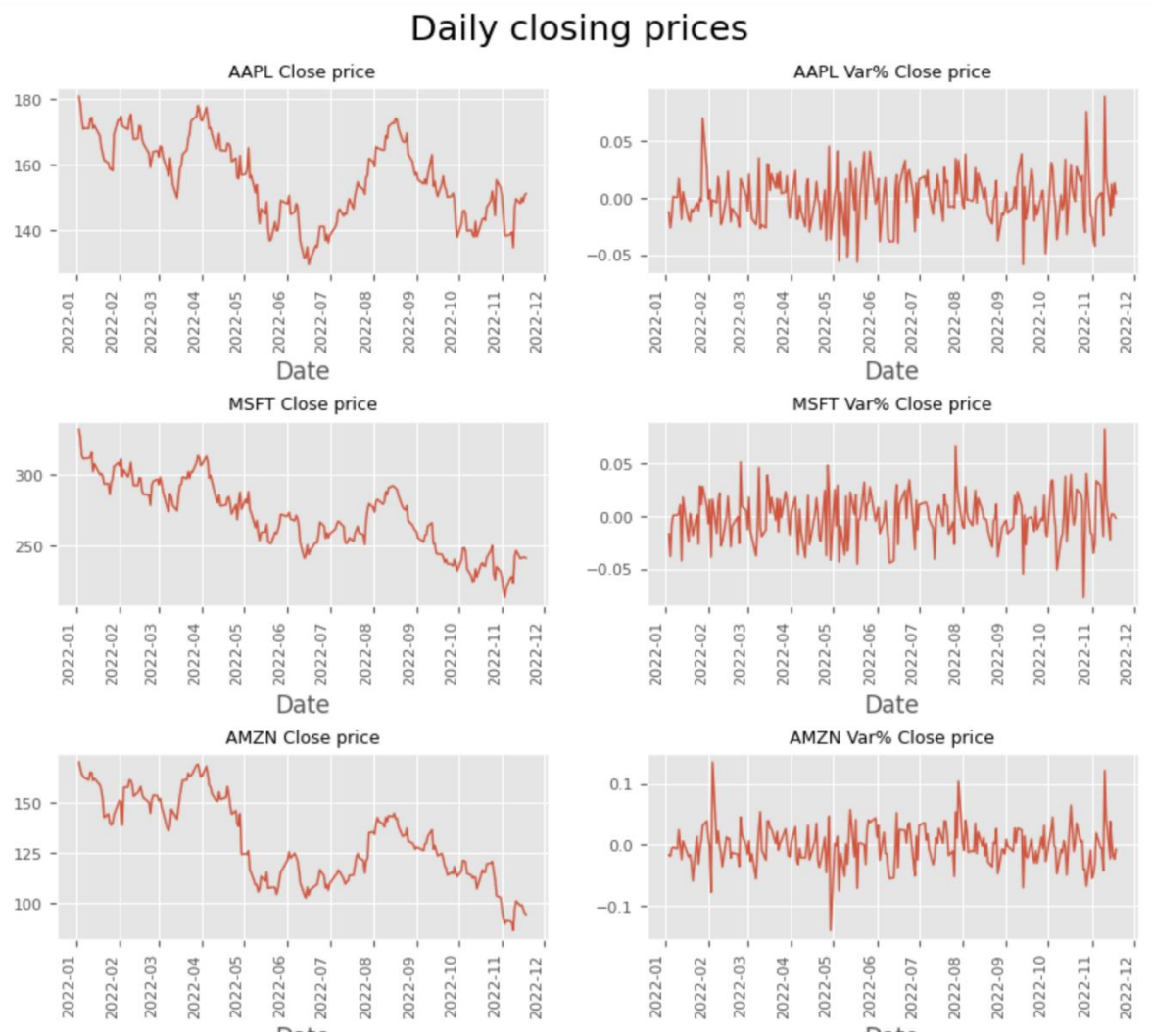
Imputación

Tal y como se explicó antes y aunado al proceso de filtración, un dato no existente no es candidato para eliminarse de primera mano pues podría tener altísimo impacto en el resultado que se entregue. Por el contrario, aplicamos la técnica de KNN con 2 vecinos de forma que la información no altere la tendencia y comportamiento de los datos ya que, en promedio, el valor de cada uno de los símbolos que se imputen estará acorde con los resultados de los 2 días más cercanos de operación bursátil.

Descripción de Estadísticas Básicas

Independientemente del tipo de análisis que se lleve a cabo y el dato (siempre estructurado) con el que se esté trabajando, es importante contar con estadísticas básicas descriptivas. Esto permite definir supuestos relevantes (por ejemplo, distribución normal) para la aplicación de diferentes técnicas de modelado de datos.

Para las etapas siguientes del proyecto, trabajaremos con 2 métricas: El precio de cierre y la variación porcentual de estos precios, como se ve a continuación



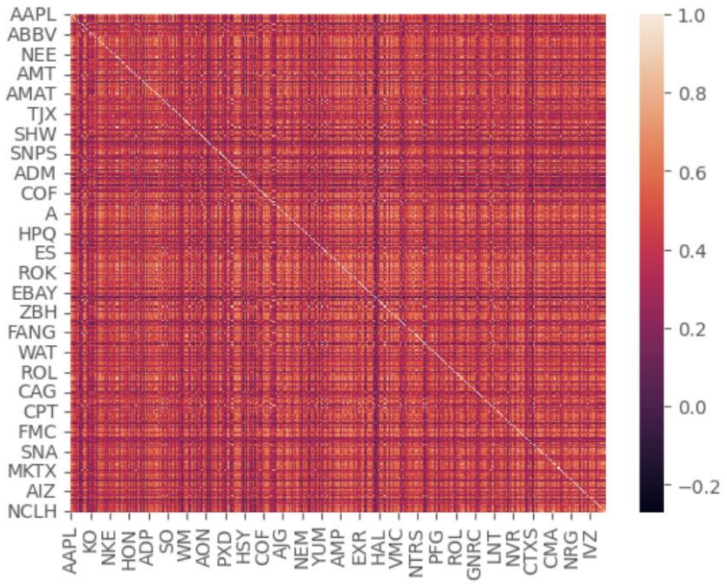
Una de las estadísticas más interesantes, de cara a los objetivos del proyecto, es la correlación entre la variación de los precios de las diferentes acciones.

Sin embargo, considerando la dimensión de la base con la que trabajamos (503 acciones), no es posible hacer un análisis 1:1 entre todas las combinaciones posibles para llegar a los resultados que se esperan. En ese sentido, se trabajarán 2 enfoques para el análisis

Determinación de variables relevantes

En el primer enfoque se propone partir de la fecha como una de las variables con mayor relevancia en el análisis de forma que se estudie la variación de los precios con respecto a la variable fecha.

Este enfoque permite hacer un análisis integral del índice usando toda la información de las acciones del S&P con mayor demanda para la generación de estrategias y portafolios de alto rendimiento.



Agrupación

Por otra parte, el segundo enfoque parte de hacer un análisis basado en un subconjunto de las acciones en las que el Grupo Stanley tiene inversiones.

Este enfoque restringe el análisis a las acciones que le competen al Grupo Stanley aumentando la efectividad en la generación de portafolios y estrategias de inversión de alto rendimiento

Anexos

Revisar notebook para comprobar soportes técnicos. En este se encuentran fuentes, preprocesamiento y visualizaciones realizadas.