Predicción Univariante y Multivariante de los Principales Índices Financieros del Mundo

Introducción

A lo largo de esta mentoría se intentó implementar distintos modelos univariantes y multivariantes de series temporales y aprendizaje automático que permitiera anticipar los valores de los principales índices financieros mundiales y detectar relaciones entre ellos.

Además, se trabajó también en la detección de los distintos comportamientos de cada una de las series para identificar la presencia de mercados “líderes” y “seguidores”. Donde los líderes hacen referencia a los mercados más influyentes y los seguidores a aquellos mercados que por su menor influencia a nivel mundial “siguen” el comportamiento de los líderes.

En particular, se trabajó con series de datos financieras. Una serie de tiempo financiera, puede ser considerada como una caminata aleatoria (“random walk”). Es decir, la mejor predicción para la serie de tiempo en el momento t+1, es su valor al tiempo t más un shock puramente aleatorio.

La hipótesis de caminata aleatoria fue popularizada por B.G. Malkiel en su libro “A Random Walk Down Wall Street” en el cual expresa que no es posible superar constantemente los promedios del mercado (los precios del mercado reflejan toda la información disponible).

Análisis y Curación de datos

¿Qué datos utilizamos?

Se trabajó con un dataset que contenía información sobre los valores de cierre diario de 11 índices bursátiles correspondientes a Argentina, Brasil, Estados Unidos, Reino Unido, Alemania, Francia, Japón, China y la India para los años 1997 a 2020.

Cada índice se encontraba expresado en su propia moneda por lo que se trabajó también con un dataset que proveía las series de tipo de cambio para cada uno de estos países, posibilitando así el análisis en una moneda común. En particular para la Argentina, fue necesario incluir una serie de tipo de cambio extra que mostraba los valores del dólar “blue”.

¿Qué modificaciones realizamos sobre el dataset?

Dado que los mercados financieros sólo cotizan los días hábiles, no se contaba con información para los días correspondientes a los fines de semana por lo que se decidió eliminar las observaciones correspondientes a los días sábados y domingos.

Además, para algunas series se encontró información faltante para los días que correspondían a feriados en algún país. Para imputar estos datos, se probaron dos métodos de interpolación: interpolación lineal e interpolación cúbica. Finalmente se comprobó que la interpolación cúbica generaba datos más “realistas” por lo cual se eligió esta opción.

Como se mencionó anteriormente, los datos se encontraban expresados en distintas monedas. Para generar una base de datos comparable, se expresaron todos los índices en una moneda común: el dólar. Para la Argentina fue necesario antes construir una serie de tipo de cambio corregida sobreescribiendo los valores oficiales con la serie de dólar blue en los años que correspondía.

¿Qué análisis preliminares realizamos?

Para identificar que efectivamente existían mercados “líderes” y mercados “seguidores” efectuamos un análisis de correlación entre los distintos índices. Con este estudio se pudo identificar la presencia de crisis globales -índices altamente correlacionados- así como también la presencia de crisis locales -baja correlación entre índices y gran diferencia entre el comportamiento de una serie con el resto-.

Se analizó también el comportamiento de las distribuciones de las series. Se encontró que las distribuciones se modifican a lo largo del tiempo, lo que llevó a concluir que las series no eran estacionarias. Además, se pudo identificar que las trayectorias de las series no eran determinísticas sino que se trataba de procesos con raíces unitarias (autocorrelación).

Para superar estos obstáculos, se trabajó con las series transformadas a logaritmo y se calcularon también sus tasas de rendimiento.

Se intentó identificar también outliers en las series pero no se encontraron valores significativamente diferentes.

Por último, se calcularon una serie de medidas descriptivas clásicas para evaluar las series -media, varianza, desviación estándar- y se utilizaron varias herramientas gráficas -histogramas, boxplots, gráficos de autocorrelación y autocorrelación parcial-.

Modelos

¿Qué modelos utilizamos?

En una primera etapa se entrenaron modelos univariantes sencillos que luego fueron utilizados para comparar su desempeño contra modelos relativamente más complejos.

Particularidades de las series de tiempo a la hora de entrenar modelos

Para entrenar modelos, es necesario particionar los datos en conjuntos de entrenamiento, validación y test. Generalmente esta partición se realiza con una selección aleatoria lo que garantiza que los datos provengan de una misma distribución y además evita el sobreajuste.

En el caso de series temporales no es posible utilizar el método anterior dado que el *orden* de los datos importa (se pierde la autocorrelación). Esto llevó a que fuera necesario utilizar otras metodologías. Se utilizaron las siguientes dos:

- Train-Test-Split: se particiona la serie de tiempo en un punto fijo utilizando los datos más antiguos para entrenar el modelo y los más nuevos para validar.

- Walk-Forward Validation: se realizan cortes recursivos . A partir de una ventana temporal inicial, se toman los primeros *s* datos para predecir el valor de la variable en *s+1,* luego se toman los primeros *s+1* datos para predecir el valor de la variable en *s+2,* etc.

  Esta modalidad tiene como consecuencia directa que el conjunto de datos utilizados para entrenamiento crezca con cada recursión, generando un mayor uso de recursos (y tiempo) a la hora de entrenar los modelos.

Modelos Baseline

Se entrenaron dos modelos: Average Forecast y Naive Forecast.

El Average Forecast es un proceso formado por una constante más un término aleatorio y consiste en calcular el valor promedio de la serie.
Por otra parte, el Naive Forecast es una caminata aleatoria donde el valor de la variable en un momento *t* está determinado por su valor en *t-1* más una perturbación aleatoria.

Las métricas utilizadas fueron mean squared error, mean absolute error y median absolute error.

Resultados obtenidos para la serie en logaritmo:

	naive	average
mean squared error	0.00067	0.28174
mean absolute error	0.01637	0.42556
median absolute error	0.01121	0.36331

Resultados obtenidos para las tasas de retorno:

	naive	average
mean squared error	0.00117	0.00062
mean absolute error	0.02223	0.0.01629
median absolute error	0.01542	0.01124

Modelo ARIMA

Los modelos ARIMA son modelos Autorregresivos Integrados de Medias Móviles y emplean la autocorrelación de las series para mejorar los pronósticos. En este trabajo se entrenó un modelo ARIMA con parámetros (1,1,0) donde los parámetros corresponden a:

p = cantidad de retardos de la serie que se incluyen
d = cantidad de veces que se diferencia la serie para lograr estacionariedad
q = cantidad de retardos del término de perturbación que se incluye.

Resultados obtenidos para la serie en logaritmo:

Mean squared error: 0.001

Modelo VAR

Los modelos Vectoriales Autorregresivos (VAR) son la extensión natural de los modelos ARMA para series multivariantes. Estos modelos permiten modelar las tasas de rendimientos de varios índices de manera simultánea.

Resultados obtenidos para las tasas de retorno:

Mean absolute error: 0.018192
Mean squared error: 0.000861
Real mean squared error: 0.029338

Modelo VECM

Los modelos Vector Error Correction se utilizan con series no estacionales por lo que en este caso pudo ser aplicado a la base de datos transformada a logaritmo. Para conocer la cantidad de relaciones de cointegración que existían, se realizó el Test de Johansen.

Resultados obtenidos para la serie en logaritmo:

Mean squared error: 0.00107

Modelo LSTM

Las redes Long Short Term Memory son un caso particular de redes neuronales recurrentes que buscan resolver el problema del decaimiento del gradiente.

Con estos modelos se busca predecir los retornos del tiempo siguiente utilizando una red de una única capa oculta con función de costo “mae”.

En este trabajo se entrenó un modelo LSTM Univariante para la serie MERV y los resultados obtenidos fueron los siguientes:

Resultados obtenidos para la serie en logaritmo:

mae: 0,029880
mse: 0,001765
rmse: 0,042008

Conclusiones

Sobre los modelos trabajados

Se entrenaron modelos sencillos y algunos más complejos tanto univariantes como multivariantes. Sin embargo, no fue posible superar el modelo base Naive. La métrica que se utilizó para comparar fue el Mean Squared Error (MSE) y este modelo fue el que menor valor obtuvo.

Conclusiones del trabajo

A la hora de predecir los valores futuros de los principales índices financieros nos encontramos con distintas dificultades, algunas de las principales fueron:

- Hay muchas variables que afectan al precio de los mismos que no fueron tenidas en cuenta. Sólo se utilizó el precio de los índices en momentos anteriores.
  Si bien entendemos que la información que se deja de usar por no incorporar otras variables es menos de la que uno imagina (debido a que los valores pasados incorporan la información pasada de las series relacionadas), creemos que los modelos podrían sofisticarse con mayor información.
- Se contó con poco manejo de dominio para el análisis de series financieras. Cada modelo predictivo cuenta con hiperparámetros a estimar y diferentes formas de tratamiento de los datos.
- Se entrenaron modelos de distintas formas. A algunos modelos se los entrenó una sola vez para predecir los datos mientras que a otros se los reentrenó a medida que se realizaban las nuevas predicciones.