Modelo Predictivo Valor EUR/USD

Reto 3 Enseña Oracle

Grupo INFORADE

Iago Barreiro Río
Santiago Pérez Acuña
Victor Figueroa Maceira

Análisis exploratorio de los datos

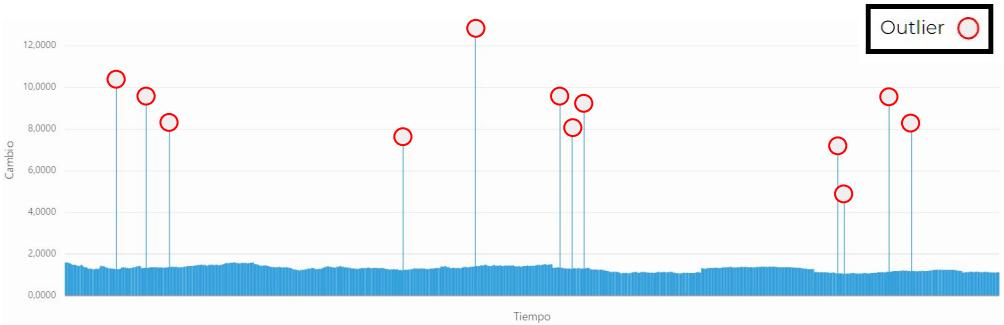


Gráfico de los valores “Open” con sus respectivos outliers detectados.

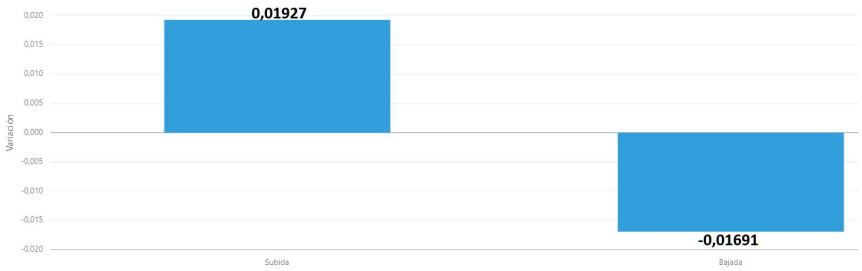
Existe una **amplia diferencia** entre máximos y mínimos, así como considerable desviación, lo que nos lleva a pensar que se presentan ***outliers****.

Además, se observan **valores imposibles**, e.g. **“Open” superior** a **“High”**.

	Open	High	Low	Close
Desv.Tipica	0.4638	0.3619	0.3672	0.4069
Max	12.6045	11.5228	10.8713	10.8770
Min	0.1363	0.1331	0.1339	0.1327
Q1	1.182	1.1846	1.1768	1.181278
Q2	1.3131	1.31872	1.30726	1.313005
Q3	1.3769	1.38156	1.37211	1.376593

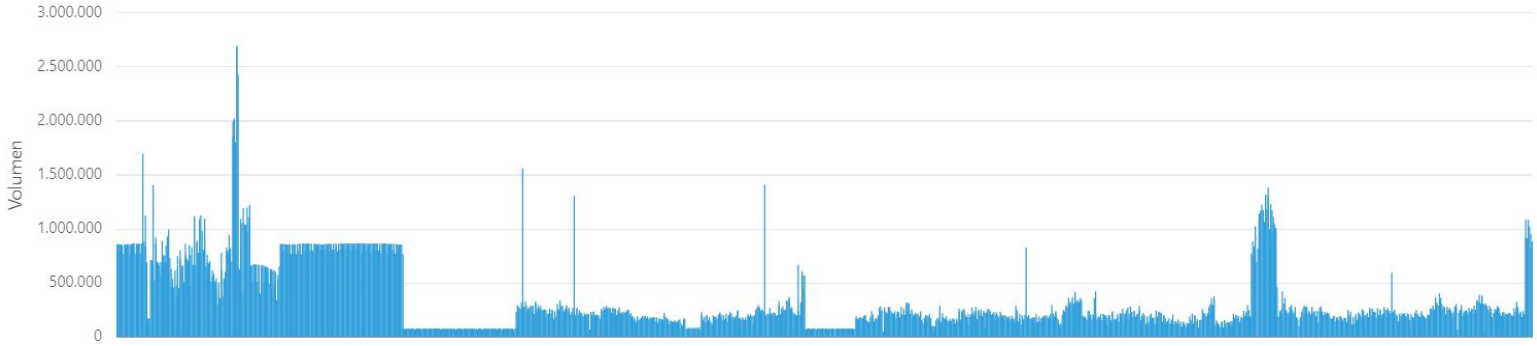
* En estadística, un ***outlier*** **(o valor atípico)** es un **punto** de datos que **difiere significativamente** del resto de los datos en un conjunto. Cabe aclarar que **se detectaron** outliers en **todos los campos**, no exclusivamente en el representado en la figura.

Análisis exploratorio de los datos



Si calculamos la **variación máxima** de precio, tanto positiva como negativa, para los valores de **apertura** de un día con respecto a la de su **cierre anterior** (sin *outliers*; ignorando casos para tipos de cambio superiores a 2), observamos que son **valores muy pequeños**.

Tras analizar los valores de **volumen**, hemos detectado, además de ***outliers***, períodos en los que se producen **"series extrañas"**, como si se hubiesen **"recortado"** los datos.



Preprocesamiento de los datos

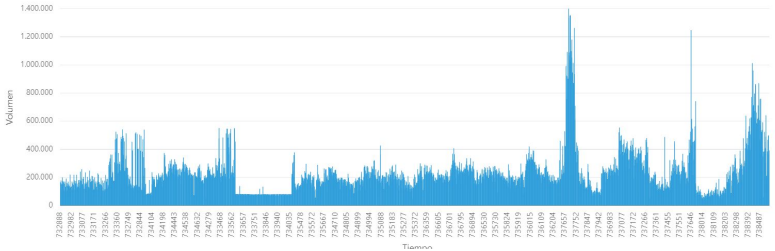
Se empleará el método **IQR** para la detección y eliminación de **outliers**, i.e.:

1. Ordenar por cuartiles los datos.
2. Siendo **IQR = Q3 - Q1**, los valores que se sitúen por debajo de **Q1 - 1.5 * IQR** o por encima de **Q3 + 1.5 * IQR** deben marcarse como outliers.
3. Eliminar los valores atípicos.

Al **eliminar** los **valores atípicos**, debemos eliminar esa tupla o rellenarla con información relevante:

- Los **open/close** se rellenan con los **homólogos del día anterior**, dada la poca variación observada en el análisis exploratorio con APEX.
- Para **volume** y **high/low** se aproximan con una **Bayesian Ridge**; modelo de regresión para estimar los coeficientes caracterizado por ayudar a evitar el sobreajuste y controlar la complejidad del modelo al **combinar información previa y los datos observados**.
- Se verifica que no se produzcan **máximos (mínimos)** por **debajo (encima)** de la **apertura** y del **cierre**. En caso de haberlos, se **reasigna** el valor del máximo/mínimo de acuerdo a la apertura/cierre según corresponda.

Una vez tratados los outliers, hemos **verificado con APEX el resultado**. Si bien para la **mayoría** el tratamiento ha **resultado satisfactorio**, **exclusivamente** para el campo **“Volume”** no ha sido así, dado que el IQR detectó falsos outliers. Hemos optado por **conservar** sus valores **originales**.



Preprocesamiento de los datos

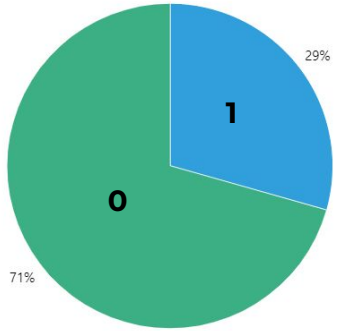
Tras un análisis exhaustivo de los datasets de ***training*** y ***testing*** proporcionados, mediante el uso de **APEX**, y relacionando los datos presentes en el conjunto de entrenamiento con el **F1-Score** para todas las etiquetas 0 y 1, respectivamente, se ha determinado que el **criterio seguido para determinar el valor de "label"** en el dataset de testing (que no en el dataset de entrenamiento) es el cumplimiento conjunto de las dos condiciones siguientes:

1. El precio de **cierre** dentro de 3 días sea **superior**.
2. El **volumen** de mercado dentro de 3 días sea **mayor** ó la **diferencia entre** el **máximo** y el **mínimo** dentro de 3 días sea **mayor** ó ambas condiciones.

Conociendo el criterio de etiquetado, **resulta trivial obtener el ground-truth** (i.e. la solución perfecta para el conjunto de prueba), así como reetiquetar de manera correcta el dataset de entrenamiento.

Es por esto que aquellos modelos que expliquen patrones complejos jugarán un gran papel.

Cabe resaltar que con las indagaciones anteriores sobre el criterio de etiquetado, se concluye que **el conjunto de prueba presenta un sesgo** hacia la label 0, según se aprecia en la figura.



Modelos y técnicas de predicción

Modelos con **elevado grado de aleatoriedad** (poco explicativos): pese a poder llegar producir buenos en iteraciones concretas para una ventana de tiempo delimitada, no son consistentes a largo plazo.

⇩ Se descartan en favor de:

A) Modelos basados en **análisis estadístico**:

SVM (Support Vector Machines): ideal para **minimizar** el impacto de **outliers** y otros valores anómalos.
→ Los datos preprocesados ya no presentan grandes cantidades de datos anómalos (descartado).

ARIMA (Auto Regressive Integrated Moving Average): destaca en series de tiempo **estacionales**.
→ El análisis de la ventana de datos propuesta (training y testing) no concluye estacionalidad (descartado).

B) Modelos basados en **redes neuronales**:

RNN (Recurrent Neural Networks): es, a priori, el enfoque más intuitivo para **series temporales** genéricas.
→ La realidad demuestra que el enfoque con MLPs produce mejores resultados* (descartado).

MLP (Multilayer Perceptrons): ideal para identificar **patrones complejos** en los datos, tanto en problemas de **clasificación** como de **regresión**.
→ **Elegido**.

* Honchar, A. (2019, 17 noviembre). Neural networks for algorithmic trading. Simple time series forecasting. *Medium*. https://alexrachnog.medium.com/neural-networks-for-algorithmic-trading-part-one-simple-time-series-forecasting-f992daa1045a

MLP: Topología de la red y técnicas empleadas

1) Entrada de datos:

El modelo recibe como entrada **N 6-tuplas de datos**, correspondientes a las últimas N* muestras hasta la muestra de inferencia, no incurriendo en **look-ahead bias**.

* Los mejores resultados se han conseguido para entradas de N=21 días.

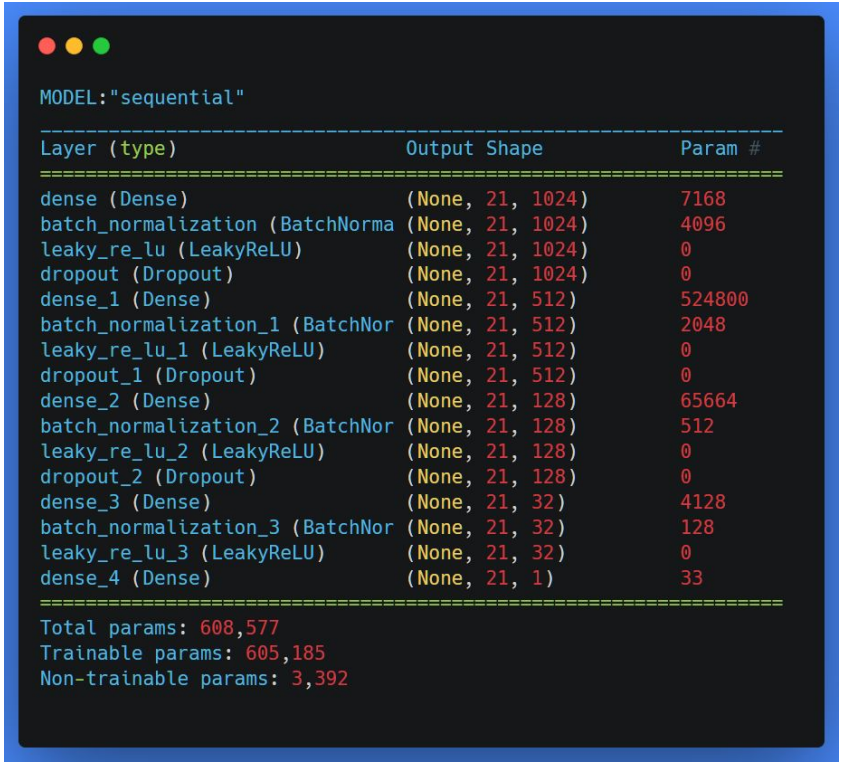
2) Topología del modelo:

Se opta por una topología en **4 macro-capas** de tamaño decreciente (por obtenerse los mejores resultados de precisión).

Cada macro-capa implementa **LeakyReLU** como función de activación y una **capa de normalización**. A esto se le considera state-of-the-art.

Se incluyen una capa de **dropout*** en cada macro-capa para evitar **overfitting** (muy difícil de detectar de modo fehaciente en algoritmos de trading).

* Los mejores resultados se han conseguido para dropout-rate=0.25 (severo).



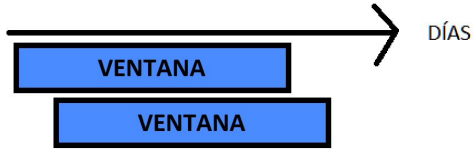
```
MODEL:"sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 21, 1024)          7168
batch_normalization (BatchNorma (None, 21, 1024)       4096
leaky_re_lu (LeakyReLU)      (None, 21, 1024)          0
dropout (Dropout)            (None, 21, 1024)          0
dense_1 (Dense)              (None, 21, 512)           524800
batch_normalization_1 (BatchNor (None, 21, 512)        2048
leaky_re_lu_1 (LeakyReLU)    (None, 21, 512)           0
dropout_1 (Dropout)          (None, 21, 512)           0
dense_2 (Dense)              (None, 21, 128)           65664
batch_normalization_2 (BatchNor (None, 21, 128)        512
leaky_re_lu_2 (LeakyReLU)    (None, 21, 128)           0
dropout_2 (Dropout)          (None, 21, 128)           0
dense_3 (Dense)              (None, 21, 32)            4128
batch_normalization_3 (BatchNor (None, 21, 32)         128
leaky_re_lu_3 (LeakyReLU)    (None, 21, 32)            0
dense_4 (Dense)              (None, 21, 1)             33
=================================================================
Total params: 608,577
Trainable params: 605,185
Non-trainable params: 3,392
```

Entrenamiento, validación y resultados

Inicialmente, se entrena el modelo con **todos los datos** (excepto los tres días) **anteriores** a la primera muestra de prueba, a fin de evitar incurrir en **look-ahead bias**.

Para cada **muestra de prueba** subsiguiente, se emplea el modelo actual para **inferir su etiqueta**, dando al modelo tanto esa misma muestra a etiquetar como un cierto número de muestras previas contiguas también sin etiquetar, según lo descrito en el apartado anterior.

A continuación, se **prosigue con el entrenamiento** del modelo durante un menor número de épocas y con una menor cantidad de datos (correspondientes a una **ventana deslizante de tamaño fijo**, que abarca hasta la muestra N+1-3, siendo N la última muestra inferida).



Análisis de los resultados en función del Ground-Truth

Precisión de las Predicciones	
65% Correctas	
Errores de Predicción	Porcentaje
Falsos Negativos	70%
Falsos Positivos	30%

Observamos que el modelo presenta una cantidad **sobredimensionada de falsos negativos**, lo cual, pese a que merme el F1-Score (que se encuentra entorno al 50-60%), puede resultar **positivo** a la hora de elegir una política de **fijación de precios** como es el caso de que nos ocupa.

Conclusiones

Un **proyecto útil** y, sobre todo, con **explotabilidad económica**

Se han detectado ciertas iteraciones del modelo que presentan una **cantidad sobredimensionada de falsos negativos**, lo que implica que en una operativa como el trading, en la que hay 3 decisiones posibles (short, long, y no action) el criterio de la red desarrollada se pueda llegar a tornar **muy efectivo** en uno de los sentidos (pese a resultar en valores bajos para el indicador F1-macro), siendo similar al problema del restaurante NUWE EVA de este reto, en donde el precio de un plato podrá bajar, aumentar o mantenerse.

Diversidad de **enfoques**

Existe una **gran cantidad de modelos** que cubren una amplia variedad de necesidades, aunque la principal diferencia podría decirse que es si basan su criterio de acierto en la aleatoriedad o la inferencia de **criterios de análisis fundamental** a partir de estadísticos provenientes del análisis técnico. Además, hemos aprendido que ciertos modelos, pese a que en momentos iniciales otorguen peores resultados, una vez configurados y **personalizados** en profundidad, pueden aportar los **mejores resultados**.

APEX, una herramienta **útil**

APEX ha resultado extremadamente útil no solo a la hora de realizar el **análisis exploratorio** de los datos iniciales, sino también a lo largo de **todo el proyecto**. Ha sido una herramienta **versátil** que ha aportado multitud de soluciones de manera inmediata y con apenas código.