Modelo predictivo valor EUR/USD

Soluciones del reto 3 del Hackathon de Nuwe "Reto Enseña x Oracle España"

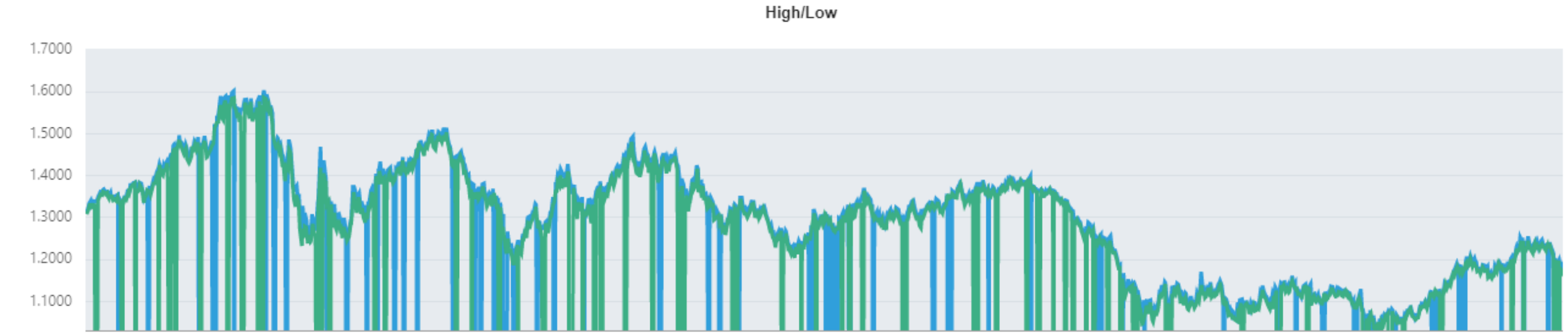
Análisis exploratorio

- En primer lugar, calculamos algunas **estadísticas descriptivas básicas** de las variables numéricas (Open, High, Low, Close, Volume) utilizando consultas SQL. Nos permitirá hacernos una idea de la distribución y variabilidad de los datos.

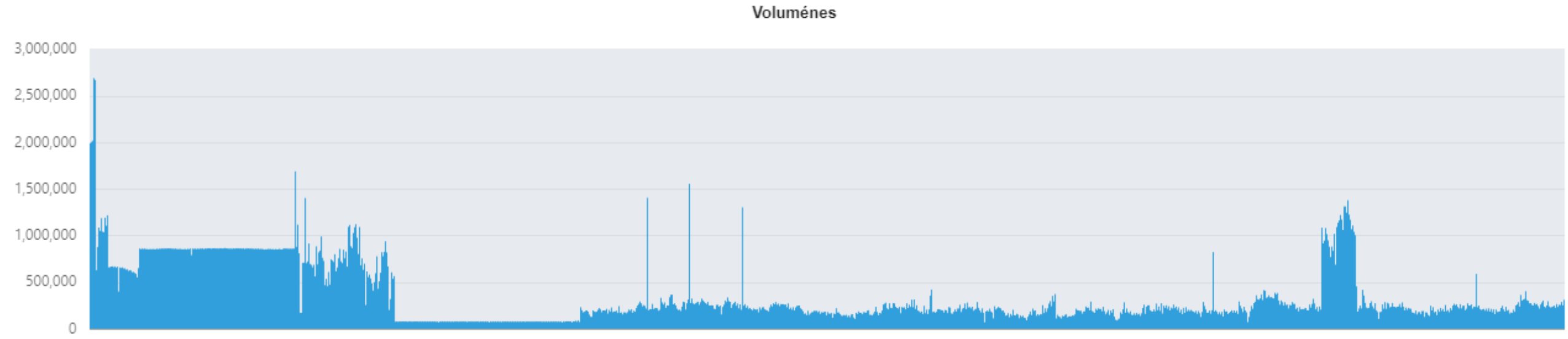
	Open	High	Low	Close	Volume
Media	1.3159	1.3083	1.2985	1.3094	279384.30
Mediana	1.3131	1.3187	1.3072	1.3129	191209
Desviación estándar	.4636	.3618	.3671	.4068	299512.52
Mínimo	.1363	.1331	.1339	.1327	497
Máximo	12.6045	11.5227	10.8712	10.8770	2693602

Vemos desviaciones altas y gran disparidad entre máximos y mínimos, es probable que haya **valores atípicos**. Al hacer las gráficas y análisis posteriores no los tendremos en cuenta

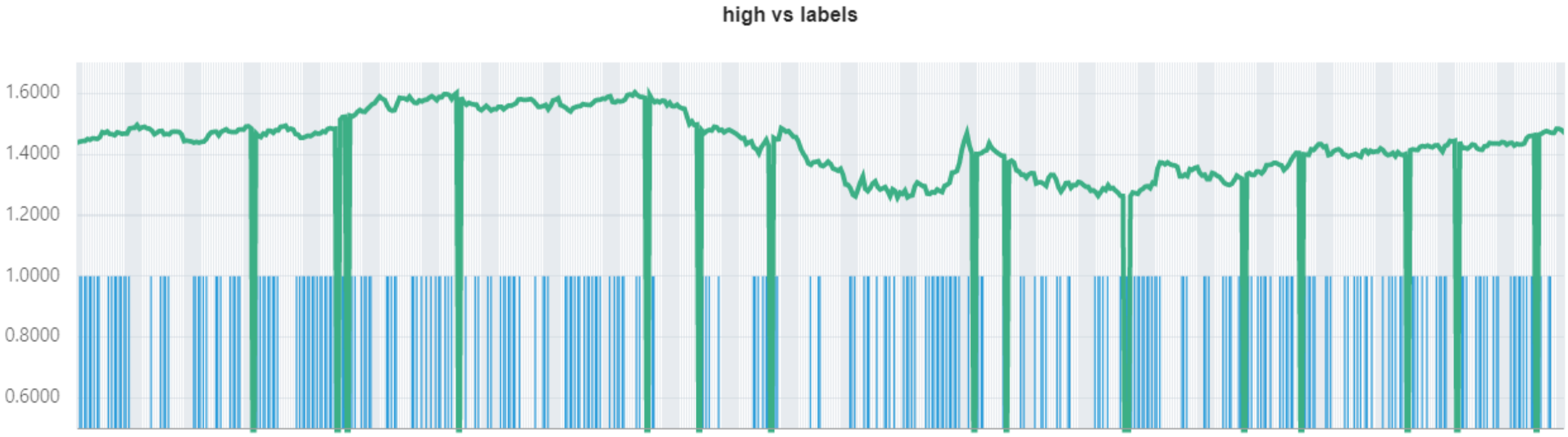
En este caso, la mediana es más útil que la media como medida central.



Se muestra la variación de precios a lo largo de los años a través de los HIGH (azul) y LOW (verde). Se observa lo mismo con CLOSE y HIGH. Las líneas verticales son valores atípicos o nulos, se pueden ignorar.



Volúmenes de operaciones a lo largo del tiempo.



Tratamos de visualizar los labels (columnas azules) junto a los precios, con el objetivo de encontrar algún patrón que nos indique qué es lo que hace que el label tome el valor 0 o 1. Haciendo varias gráficas como esta en distintos periodos de tiempo, y también respecto a las otras métricas de precios, no llegamos a una conclusión clara. A excepción, quizás, de que parece haber más labels 0 en épocas de bajada de precios y viceversa, pero no siempre es así. Trataremos de que el modelo de predicción encuentre una relación más clara.
Por otro lado, vemos que los labels suelen ir por bloques, varios unos o varios ceros seguidos.

Preprocesamiento de los datos

- Pasos principales y criterios utilizados:
  - Convertir los csv de datos en un DataFrame, para poder manipular fácilmente los datos.
  - Eliminamos las filas a las que les faltan valores o tienen valores atípicos (alrededor del 7.5%). Dejarlos generaría ruido en los datos.
  - Feature engineering: Como el análisis exploratorio con las variables que teníamos nos nos revelaba demasiado, añadimos nuevas variables útiles usando la librería "Technical Analysis Library in Python".
    - Investigando cuáles son las métricas que se utilizan en el análisis técnico que podrían ser usadas en nuestro caso, añadimos las siguientes: RSI, MACD, ADX, CCI, ATR, Bollinger Bands. Iremos probando si los resultados mejoran cuando se quita alguna o se añade otra nueva.
  - Eliminar valores NaN
  - Dividir el training dataset en training (80%) y validation (20%), lo que nos permitirá evaluar los resultados de cada modelo que probemos con datos que no han sido usados para entrenarlo.

Modelos y técnicas de predicción

- Probaremos varios modelos que podrían ser útiles en este contexto. Mediremos sus resultados al ir ajustando sus parámetros y evaluaremos cuál es el óptimo para nuestro caso particular. Al ser un problema de clasificación, probaremos los siguientes modelos:
  - Logistic regresion
  - Random Forest
  - Decision Tree
  - Support vector machine (SVM)
  - Red neuronal

Entrenamiento, validación y resultados

- Tras entrenar cada modelo, calculamos con el set de validación los resultados de las métricas de evaluación más populares en problemas de clasificación. Nos centraremos en maximizar el F1-score, al ser lo que pide el reto.
- Del que mejores resultados de en un principio, ajustaremos sus parámetros para optimizar su rendimiento.

	Accuracy	Recall	F1-Score
Random Forest	0.6969	0.6987	0.6955
Logistic Regression	0.4953	1.0	0.6625
Decision Tree	0.5907	0.5838	0.5857
SVM	0.5077	0.0745	0.1304
Red Neuronal	0.4954	0.9876	0.6597

Conclusiones

- Aun ajustando un poco más el modelo y probando distintas variables de análisis técnico, no logramos superar el 70% de precisión en el set de validación. Probablemente hay algún patrón más claro en los datos que no hemos logrado encontrar. Además, el rendimiento empeora con los datos de test.
- Aunque no hayamos conseguido un gran rendimiento para el modelo, este proyecto nos ha servido mucho para aprender sobre exploración de datos, preprocesamiento, modelos predictivos y técnicas de evaluación.