Contexto

Un equipo de ventas de bonos de un banco comenta a sus científicos de datos que necesitan saber de antemano si el precio de un bono que dan a un cliente va a ser aceptado o no por este. El cliente les puede pedir precio tanto de compra como de venta del bono, además el cliente podrá pedir el precio a otros competidores y en función de todos los precios que recibe elegirá cerrar la operación con uno de ellos.
Para esta tarea se aportan dos datasets, en el que el dataset A recoge peticiones de precios con el precio que se ha cotizado, y el dataset B recoge el histórico del precio medio de los bonos de los que se han pedido precio en el dataset A.

Notas: Los datasets no corresponden con datos reales del negocio

Descripción de los dataset

Dataset A rfqs.csv:

- date_time: dia y hora en la que se pidió la operación
- instrument: el bono del que el cliente ha pedido precio
- client : código de cliente.
- price: precio dado al cliente.
- mid: precio de mercado del bono capturado por el sistema del banco en el momento de la operación
- vol_MM: cantidad pedida por el cliente (en millones de euros).
- dv01: sensibilidad del bono a variaciones de su yield
- num_dealers: número de vendedores a los que el cliente ha pedido precio.
- side: 1 si es compra -1 si es venta.
- won: 0 no se cerró y 1 si se cerró.

Dataset B mid_price.csv:

- date_time: dia y hora del precio medio del bono
- mid: precio de mercado del bono que nos facilita un proveedor de datos de mayor fiabilidad que nuestra plataforma pero con frecuencia de 5 minutos
- instrument: el bono

Ejercicio:
Utilizando estos dos datasets diseñar uno o varios modelos que a partir de los datos de entrada nos predigan si la operación se va a cerrar o no.
Desde negocio se transmite que una característica importante a tener en cuenta es el spread. Este valor se calcula como:
Si es compra mid - price.
Si es venta price - mid.
Comparar la calidad del modelo usando el mid capturado en el momento de la operación o interpolando usando la fuente externa, ¿impacta mucho en el resultado el no contar con datos de alta frecuencia?

Se valorará:

- Limpieza y tratamiento de los datos
- Análisis exploratorio de los datos
- Interpretación de los resultados del modelo
- Calidad de código (se podrá pedir explicación del mismo en posterior entrevista)
- Eficiencia del código

Restricciones:

- Tiempo : 1 semana
- Lenguaje: python o scala, se valorará positivamente si se usa Spark

Entregable:

- Jupyter Notebook (.ipynb) con el análisis