DATA ANALYSIS

--

Studies and interpretation of the bond market data.

Purpose:

This is a personal initiative to support my application for the data scientist role.

Exercise:

Using these two datasets (datasets are described below), build a model that predicts whether the request for quote is going to be won or not by the bank, given the price quoted and other information provided in the datasets or derived from them (feature engineering).

Context:

The bond sales team requests its data scientists that they would like to have a model that estimates whether prices quoted to clients in bonds are going to be accepted by them, therefore resulting in a trade. In this business line, clients can request at any time during the trading session both buy and sell prices of bonds, for any volume in €. The client might request quotes from multiple banks at the same time, and therefore will close with the one that answers the best price (if any of the prices is considered fair by the client). Banks have no access to the prices quoted by their competitors; they only know the number of competitors for each request for quote (RfQ) from a client.

The objective of this data analysis part is to explore and interpret the historical data to enhance understanding of the bond market dynamics, enabling informed decision-making for model selection. This step is critical to ensure the success of our analysis.

Data source:

These datasets are collected from an open-source online project. They are two datasets. Dataset A which contains the information regarding the request for quote (timestamp, client, volume requested, price quoted by the bank, number of competitors, …), and dataset B contains intraday historical market mid-prices of the bonds for which request for quotes are provided in dataset A.

Description:

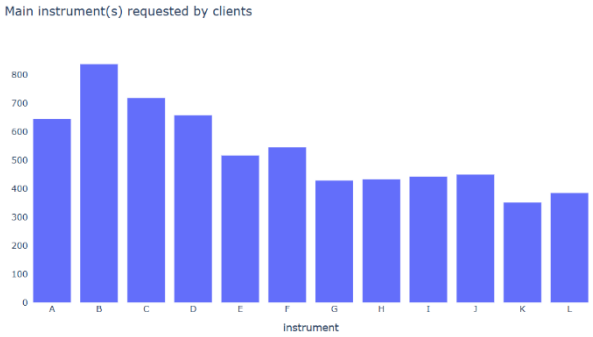
Dataset A: rfqs.csv:
- date_time: date and time at which the quote is requested.
- Instrument: the bond for which the customer has requested a price
- client: client code (anonymized)
- price: price quoted to the client by the bank
- mid: market mid-price of the bond captured by the bank's system at the time of the operation
- vol_MM: amount requested by the client (in millions of euros).
- dv01: sensitivity of the bond to variations in its yield (a measure of risk of the bond)
- num_dealers: number of banks from whom the client has requested a quote
- side: 1 if it is buy -1 if it is sell (from the point of view of the bank, not the client)
- won: 0 if the bank did not close the operation, 1 if it did.

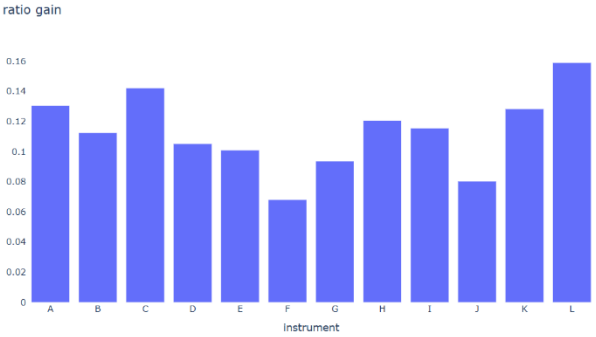
Dataset B: mids.csv:
- date_time: day and time of the market mid-price of the bond
- mid: market mid-price of the bond provided by a data provider with greater reliability than our platform, sampled with a frequency of 5 minutes
- instrument: the bond for which the market price is provided

Data quality check:

There are no data quality issues on the collected datasets.

Data Analysis





NB: Please refer to the script named "*DataAnalysis"* for more details on the plots.

Key Observations and Feature Engineering

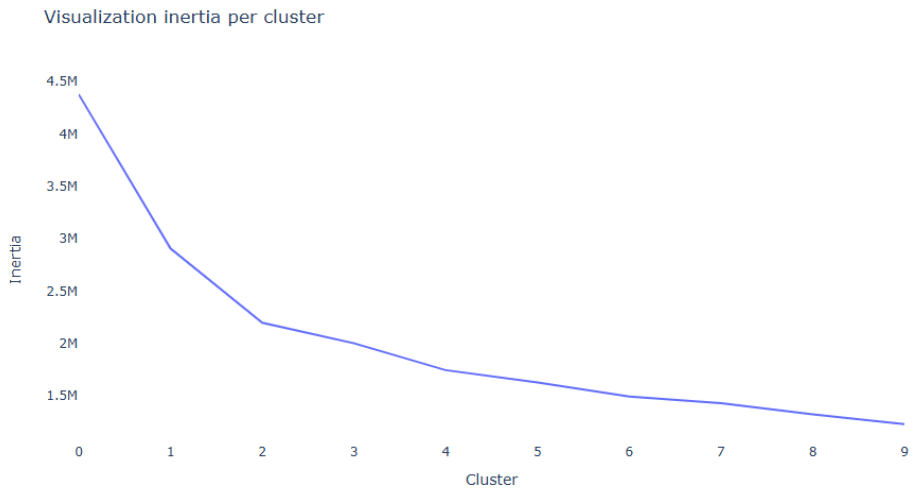
1. **Financial Data Characteristics**
    - The **date column** plays a pivotal role as it may capture temporal patterns such as trends and seasonality. To leverage this, we created additional features derived from the date, like month and quarter indicators.
    - We also engineered the **quoted spread** as a key feature. This metric assesses the credibility of quotes provided by the bank, which is a crucial determinant for clients' decisions.
2. **Market Insights from Historical Data**
    - **Liquidity and Market Share**: Instrument B emerges as the most liquid in the bond market. However, the bank's market share for this instrument remains low at just 11%.
    - **Pricing Challenges**: The underperformance of Instrument B may stem from its high average quoted spread of 25%, significantly exceeding the real market price. This not only hampers the bank's competitiveness but could also impact its reputation and business efficiency.

- **Competitive Pricing**: Instruments A, C, E, and L exhibit very competitive pricing, with quoted spreads near zero. This correlates with higher win rates, particularly for Instrument L, which achieves a win ratio exceeding 15%.
- **Underpricing Concern**: Instrument G appears to be underpriced, indicating the need for a review of its pricer settings.
- **Seasonality**: Market activity peaks between **May and September**, indicating a seasonal trend.

Clustering Analysis

To delve deeper, we implemented a clustering model to segment the bond market and extract actionable insights for the sales team.

1. **Methodology**
   - We employed the **k-means clustering algorithm**, a well-suited method for segmenting financial datasets.
   - To determine the optimal number of clusters, we utilized the **Elbow Technique**, which balances cluster count and variation across clusters.
   - The principle involves minimizing inertia (variance within clusters). Beyond a certain point (the "elbow"), the reduction in inertia becomes negligible, indicating stable segmentation.
2. **Elbow Analysis**



   - The **inertia vs. cluster count graph** highlights three major breaks, suggesting an optimal segmentation with **four clusters**. This configuration balances analytical precision and business interpretability.

Cluster Characterization.

Note: for more

After applying the four-cluster model to our historical data, we derived the following profiles:

Group 1

- **Key Clients**: A, B
- **Primary Instruments**: D, F, I
- **Seasonality**: Peak volumes from June to August; quoted spreads for Instruments B, D, F peak in the same period.
- **Win Ratio**: 10%

- **Average Quoted Spread**: 5%

Group 2

- **Key Clients**: A, B
- **Primary Instruments**: A, C, D
- **Seasonality**: Peak volumes from May to August; quoted spreads for Instrument B peak in September.
- **Win Ratio**: 12%
- **Average Quoted Spread**: 8%

Group 3

- **Key Clients**: A, B
- **Primary Instruments**: B
- **Seasonality**: Peak volumes from June to August; quoted spreads for Instruments B and F peak in August.
- **Win Ratio**: 11.5%
- **Average Quoted Spread**: 9%

Group 4

- **Key Clients**: A, B
- **Primary Instruments**: A, C, E, G, L
- **Seasonality**: Peak volumes in March; quoted spreads for Instrument K peak in September.
- **Win Ratio**: 11%
- **Average Quoted Spread**: 5.5%

General Comments and Recommendations

1. **Client Focus**
    - Our primary clients, **A and B**, frequently request Instrument B. Despite its liquidity, its poor win ratio highlights inefficiencies in pricing strategies.
    - **Group 1**: Instruments D, F, I are well-priced but underperform in terms of win ratio compared to our market share. This anomaly warrants further investigation.
    - **Group 2**: Instrument B's inefficiencies in both quoting and win ratios are evident. A review of its pricing model is imperative.
2. **Improving Competitiveness**
    - Review and adjust the pricer for **Instrument B** to align it more closely with market realities.
    - Monitor and refine pricing strategies for underperforming instruments within each cluster to improve the win ratio.
    - Capitalize on the success of instruments with low spreads, such as **A, C, E, and L**, while maintaining their pricing efficiency.
3. **Seasonality Considerations**

- Focus sales efforts and promotional strategies during peak market periods (May-September) to maximize market share.

This comprehensive analysis and clustering approach enable actionable insights with regards to the pricing strategies, address inefficiencies, and enhance overall market competitiveness.