Assignment 8: Exploratory Data Analysis

Saurav Mawandia

Sougandh Kohli

Harrisburg University of Science and Technology-Harrisburg, PA

ANLY 525-90- O-2022/Summer - Quantitative Decision Making

Professor E. André L'Huillier M

July 10, 2022

For this week's assignment, you have to submit a brief report with the data you will be using for your project, including

1. Mention the source. If you collected it, describe how you did it.

   The data is collected from Kaggle .
   https://www.kaggle.com/competitions/benchmark-bond-trade-price-challenge/data .

2. Present a simple Exploratory Data Analysis of the dataset: a) general statistics, b) visualization of main variables.

US corporate bond trade data is provided. Each row includes trade details, some basic information about the traded bond, and information about the previous 10 trades. Contestants are asked to predict trade price.

Column details:

- **id: The row id.**
- **bond_id: The unique id of a bond to aid in timeseries reconstruction. (This column is only present in the train data)**
- **trade_price: The price at which the trade occured. (This is the column to predict in the test data)**
- **weight: The weight of the row for evaluation purposes. This is calculated as the square root of the time since the last trade and then scaled so the mean is 1.**
- **current_coupon: The coupon of the bond at the time of the trade.**
- **time_to_maturity: The number of years until the bond matures at the time of the trade.**
- **is_callable: A binary value indicating whether or not the bond is callable by the issuer.**
- **reporting_delay: The number of seconds after the trade occured that it was reported.**
- **trade_size: The notional amount of the trade.**
- **trade_type: 2=customer sell, 3=customer buy, 4=trade between dealers. We would expect customers to get worse prices on average than dealers.**
- **curve_based_price: A fair price estimate based on implied hazard and funding curves of the issuer of the bond.**
- **received_time_diff_last{1-10}: The time difference between the trade and that of the previous {1-10}.**
- **trade_price_last{1-10}: The trade price of the last {1-10} trades.**
- **trade_size_last{1-10}: The notional amount of the last {1-10} trades.**
- **trade_type_last{1-10}: The trade type of the last {1-10} trades.**
- **curve_based_price_last{1-10}: The curve based price of the last {1-10} trades.**

We've created visualisation of some main variable

Attached the EDA PDF

3. Define your independent and dependent variables.

   Dependent variables: trade-price, current_coupon, time_to_maturity, is_callable, trade_size, trade_type, curve_based_price, trade_price_last(1-10), trade_type_last(1-10), curve_based_price

   Independent variable : id , bond_id, reporting_delay

4. Evaluate the quality of the data.

   Analyzed the data in attached ipynb. The overall data quality is good . There are some na's in trade_price_last(1-10) which can be easily cleaned by replacing them with a negative number and keeping count of those NA's for modelling

5. Evaluate the utility of the data (explain how it directly or indirectly could answer the research question or provide insights).

The last price , last trade type and curve based price last for last 10 trades can help us to determine the future trade for similar bond. Based on how the bond has traded in past for same bond type and size we can come up with a predicted price of the bond. Based on my experience with bond pricing for my company usually these variables are very helpful and can give us a very accurate prediction of bond price .