**Project Problem and Hypothesis**

This project will seek to identify the drivers of bid-ask spread in the municipal marketplace and use those factors and the resulting model to predict spreads on bonds/positions going forward. This information will help municipal bond traders make better-informed decisions regarding execution, bond liquidity/selection, and position sizing. In addition, this model will help us market our services to potential clients. To understand how this will be the case, please see the Domain Knowledge section below. As a machine learning problem, the prediction outcome will be a continuous number representing the spread in basis points that one could expect to incur for a “round-trip” (buy and then selling) trade for a particular position. This model, if it were accepted on a trading desk, could help traders/clients /risk managers better assess risk. In this case, the risk is liquidity risk specifically. I believe that the biggest determinants of bid-ask spread in municipal bonds are (1) position size, (2) rating as a proxy for credit quality, and (3) coupon, ranked in terms of importance.

**Datasets**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Years to maturity | Years to call date | Years between call date and maturity | Coupon (%) | Quantity/Position size (000’s $ par) | CUSIP | Description | Deal Size | Rating (median of S&P, Fitch, & Moody’s) | Date | Municipal Rate Volatility (10-year MMD rate change on the day) | Treasury Rate Volatility (10-year Treasury rate change on the day) | Dealer-to-dealer trade or dealer-to-customer |

Domain Knowledge

I work as a portfolio manager on a municipal bond trading desk. We manage municipal bond SMAs (separately managed accounts) which are basically retail accounts for high-net-worth individuals that offer scale for these clients (think buying products from Costco instead of your local “mom-and-pop” grocer) while still providing customized solutions. So, I interact with/utilize this data every day in order to inform my own trading decisions. I believe that my experience in this domain will greatly benefit my analysis of the data as I’m aware of what results are expected and will be able to interpret whether the results of the model make sense empirically.

As far as existing liquidity-related research in the municipal bond space, it unfortunately doesn’t get as much interest as corporate or treasury/gov’t bonds do from the academic community or finance industry. While looking at corporate and gov’t bond research on the topic may be instructive from a methodological standpoint, those are asset classes are much simpler than municipals. Not only are municipal bonds less standardized than them (for some fun, try explaining a 30 year muni with a sinking fund feature and a par call in 10 years to a corporate trader/PM), but there are vastly more unique CUSIPs (by several orders of magnitude) due to the way that issuers typically issue bonds serially. Looking internally, I know we have a crude marketing slide that was produced years ago about the effect of trade size on execution. This slide helps us sell our services to clients (financial advisors) because we’re basically saying to them, “we can transact at tighter spreads and thus provide better outcomes for clients than you.” In addition, there was a similar analysis to the one I’m looking to do which was done by one of our competitors, Nuveen Asset Management. Theirs was proprietary so, while I’m not able to view the data underlying their analysis, below is a brief explanation of how they went about their research.

“Nuveen Asset Management Research conducted a study of daily municipal bond trades to calculate estimated spreads for fixed coupon municipal securities. Each file was sorted by CUSIP to find multiple trades on a given day. For each CUSIP that appeared multiple times in one day, the algorithm searched to see if both a sale to a customer and a purchase from a customer occurred on that day. If there were multiple sales and purchases on a given day, the average was taken for each category. The estimated spreads were calculated for each CUSIP by subtracting the average purchase price from the average sale price of the CUSIP. Transactions between dealers were excluded. Spreads for some security-days appeared too large or small to be anything but data errors, so we sorted security days by the size of the spread and deleted observations with the largest and smallest 0.5% of spreads. Institutional size trades were separated from the retail trades and the median spread was taken by category ($25K or Less, $25K - $100K, $100k - $1M, and Over $1M). For example, the $25,000 or less dataset contains all instances where there was at least one customer purchase of principal amount of $25,000 or less and at least one customer sale of principal amount of $25,000 or less on the same day in the same security. Negative spreads (average purchase price greater than the average sale price from the customer) and zero spreads were excluded from the calculation”

Source: <https://www.nuveen.com/Home/Documents/Default.aspx?fileId=56469>

Clearly, their analysis looks at a single factor, trade size, as a determinant of bid-ask spreads. While this factor intuitively makes sense as the leading driver in municipal bid-ask spreads, I believe that there are indeed other contributing factors to this measure. Other than that research, the only existing papers are either from >10 years ago when the municipal bond market was much more homogenized due to the prevalence of bond insurance. The decline of municipal bond insurance following the credit crisis was a definite regime change in that municipal bonds began to trade in much wider bands and as a result the market now resembles more of credit market than a rates-driven one. For this reason, the research prior to this regime change wouldn’t be all that applicable.

Project Concerns

The concerns I have are as follows:

* Pre-processing of the data: While I have definitely gained experience in Python/Pandas/NumPy since the course started, this will be the time I’m delving into a dataset on my own. I’m not sure that I know how to do a lot of the preprocessing that will be required for this project. For example, in order filter out the “noise” in this data, I would have to look at each day’s trades, and I really only want to include trades on CUSIPs where the buy and sale occur on that same day. Any single un-matched buy or sale would be ignored for this analysis. In addition, I’d have to make sure that trade pairs are of the same/similar quantity so that I’m not looking at the bid-ask spread on a pair where the buy was >$1 million and the sale was $10k lot sizes for example. This mixed quantity pair would “muddy” my analysis as I couldn’t necessarily say what size trade the resulting bid-ask spread was for.
* Adding the Treasury/municipal market rate volatility to the other data: While the trade-related/CUSIP data is all coming from one source (MSRB), the rate data would be from another source (Thomson Reuters TM3) and would involve some calculation and then matching against the respective dates on the MSRB data. This is a relatively minor concern because I know it can be done, just not how to actually perform this transcription.
* Model choice: I’m not entirely sure how I’m going to choose which model to fit to this data. While the existing models that I’ve looked at mostly resemble simple linear regression (“bid-ask spreads are a function of trade size only”), I wouldn’t want to make the wrong assumption here.
* Feature selection/identifying multi-collinearity if present: I’ve purposefully included many variables that may be correlated because I didn’t want to leave any of the pertinent features out. As a result, I’ll have to really emphasize ruling out variables that are correlated as well as those that offer little value to the model.

The main assumption/caveat of this problem is that a model CAN be fit to help predict bid-ask spreads in municipal bonds. The municipal bond market, due to its retail and fragmented nature, is sometimes very irrational/headline-driven so I have a feeling that whatever model is fit will have low predictive value practically. This inefficiency is partially driven by the audience for municipal bonds (not the most technologically advanced market by any means) and partially by its very nuanced character. There are tens of thousands of issuers with hundreds of thousands of financing streams responsible for over one million unique CUSIPs outstanding and 99% of them don’t trade on a given day! Understandably, drawing inference from any model on the market as a whole is quite an over-simplified view. This is something that multi-million dollar operations at the pricing service-providers still consistently gets wrong, so I don’t expect the resultant model to have a very high r-squared. Besides that, I’m looking at bid-ask spreads on a daily basis. This ignores the fact that many round-trip trades occur over multiple days. Again, a simplification but a livable one for explain-ability purposes. Fortunately, there’s no gap in the data in terms of trades not being captured. Every dealer/customer is required by law to report trades to MSRB within 15 minutes of execution, so there’s very little chance of any data inaccuracy or of any trade(s) not making this trade feed.

The cost of model being wrong would be very high in terms of trading risk and/or potential losses. If a trader/PM bases his risk management on a faulty model, he/she may be overlooking a potential source of risk and his P&L may suffer due to this. On the other hand, the benefit would be lower transaction costs and liquidity risk for our clients and thus better outcomes as we’d be able to quickly trade into/out of positions without incurring high slippage. This would allow us to be more responsive in terms of market developments and monitor our risks much more closely.

Outcomes

My expectations fall largely in line with those of my target audience. The target audience (our trading desk) doesn’t expect much in terms of the model’s applicability because of the nuanced nature of the market described above. A complex model in this space runs the risk not being explainable as well as having a high variance/not being applicable. The only municipal-specific research in the space is thus very crude (also described above) and looks at a single feature. In other words, the audience is most likely expecting a simple model with only a handful of features at most. Chief among these features, they expect bid-ask spreads to be largely driven by (1) position size, (2) rating as a proxy for credit quality, and (3) coupon. While I don’t know how important position size is alone, I would expect that these 3 features are responsible for the vast majority of the bid-ask spreads in the municipal marketplace.

The model doesn’t have to yield 100% accuracy to be of use to trades. It’s more important that its output is explainable. At the end of the day, traders and PMs will not rely solely on any model (at least I hope not) to make buy/sell decisions but will use it as part of the mosaic of market “color” (who knew trading was so art reference-heavy?) to inform their decisions. In that sense, the model will be successful if it yields new insight into the features that drive bid-ask spreads. For instance, variable coefficients could be useful in either determining relationships that we didn’t know existed or explaining long-held trader intuition. The project will be a bust if there is no discernable relationship amongst any of the independent variables and the predicted values. The model will thus yield little insight and probably cause many career traders to question their life choices.