T&O 567: Data Ming

Team Project

Benchmark the Actual Bond Prices

Inspired By a Kaggle Data Mining Competition

Team Member

Zhenshuo Zhang

E-mail: zhenshuo@umich.edu UMID: 54224588

Qiao Chen

E-mail: chenqiao@umich.edu UMID: 96072806

Ran Zhang

E-mail: zhran@umich.edu UMID: 53482321

Ruiwen Peng

E-mail: ruiwen@umich.edu UMID: 81041056

1. Executive Summary

There are a number of models trying to price bonds, and many of them focus primarily on the information available to the general public. Also more attention is paid to the bond issuer, rather than the features of bond itself. These models are good enough at benchmarking bond prices, however, in reality the bond trading is dynamic, and is connected with trade-specific factors like the type of bond trade and the size of the trade. Besides, some important information in deciding bond prices are not available to the public, or only available with time delay. The characteristics of bond buyers and sellers might also have an impact on the actual bond prices.

In this project we try to decide would an inclusion of bond's trade-specific features improve the explanation aptitude of the original model. The price that is calculated by the "original model" (the model is provided by Benchmark Solutions, the host of this Kaggle challenge is provided as curve_based_price in our data. By doing a univariate regression of actual bond price on this variable we can get a very high R^2 and a small RMSE, which indicates the model is itself a good prediction of actual bond prices. Starting from this variable, we include variables like **whether the bond is callable, the reporting delay of the transaction, the actual bond prices at which previous trades happened, trade size, and trade type** (sold by customer, bought by customer or traded between dealers), etc. into our model and try to decide whether the prediction (measured by RMSE) can be improved from the original model this way.

We began with a description of the major variables in our data, and since there are high correlations among a number of variables in our data, we then ran a factor analysis for lagged trade prices, lagged curve based prices and time differences between each of consecutive trade occurrences. As a result we got 2 factors which we think should respectively be name "Lagged Prices" and "Trade Frequency". Including the resulted factors and other several variables into our Partition Tree model and Neural Network model, we get models that can better predict the actual bond prices than the original curve based price model, although the magnitude of improvement is largely limited, which constitutes the most serious problem of our project. Of our two models, the Partition Tree model has a smaller RMSE, and therefore is more accurate in the bond price prediction based on our standard.

Since our data is bulky and observations are trade occurrences rather than various bonds listing, we predicted a cluster analysis might not be fruitful, which is then proved by the results of cluster analysis.

The improvement of our Partition Tree model over the curve based price model is very small, so we think a more rigorous model is needed to better predict the actual bond price.

2. Introduction

As indicated before, our project is about benchmarking the actual bond trade prices. The problem itself is a competition run by Benchmark Solutions, which describes itself as a "provider of real-time corporate bond prices", on the famous data mining competition website Kaggle.

Pricing bond prices has always been an interesting topic in finance research, since a bond's price changes on a daily basis, just like that of any other publicly-traded security. It depends not only on the factors that it is highly related with, but also on the interaction among those factors. Due to the evasiveness and volatility of bond price, and the resulting profitability that attracts profit-seekers in to the market of bond trading, pricing bond price, especially its actual trade price, becomes a major concern for bond traders, and the bond market as a whole.

There are many models trying to benchmark bond prices, but the accurate prediction of actual trade price of bonds is difficult. The problem arises partly as a result of the lack of transparency in the bond market.

A good number of models, as mentioned before, focus more on the information that is available to general public and the characteristics of the issuing companies, while failing to pay enough attention to "trading dynamics and microstructure of individual bonds", as stated by Benchmark Solution. In reality there is much information not available to the general public or available only with a time delay, therefore, the actual market price can deviate from the price that is calculated by various pricing models.

Therefore an urgent issue in pricing bond prices is to decipher the variation of actual prices from prices that are calculated from basic bond pricing models. In this project we assume that the "trading dynamics and microstructure of individual bonds" such as the type of bond trade, the size of trade and their lagged values can be of use in explaining the variation.

A final step is to compare the result of our model with the result of the original pricing model provided by Benchmark Solutions, and decide whether the "trading dynamics and microstructure of individual bonds" is truly meaningful here.

3. Data

The data is provided by Benchmark Solutions and downloaded from the competition homepage http://www.kaggle.com/c/benchmark-bond-trade-price-challenge/data.

The dataset used in this project is the training.csv file downloadable from the website. The original contains 762678 bond price observations from 3736 different bonds, which are too many for our analysis. Therefore, after getting the data, I deleted observations with missing values, and then kept all 148620 observations from the first 741 unique bonds.

Mean	103.44055
Std Dev	9.8242891
Std Err Mean	0.0112494
Upper 95% Mean	103.46259
Lower 95% Mean	103.4185
N	762678

Table 1. Trade Price Summary before Deleting Data

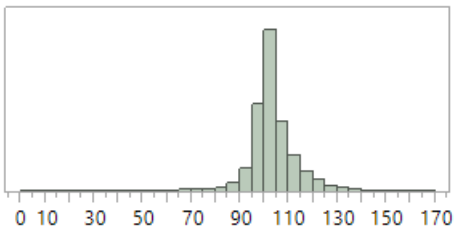


Figure 1. Trade Price Distribution before Deleting Data

Mean	105.39674
Std Dev	9.5509878
Std Err Mean	0.0247748
Upper 95% Mean	105.44529
Lower 95% Mean	105.34818
N	148620

Table 2. Trade Price Summary after Deleting Data

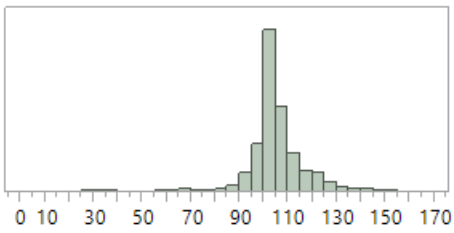


Figure 2. Trade Price Distribution after Deleting Data

The upper graph is the distribution of actual trade price before deleting observations, while the lower graph is the one after deleting observations. The two graphs are not very different from each other, each showing a roughly normal distribution, although the former one does have a bigger spread than the later. The summary statistics for the two data are also not very dissimilar, although the mean

price after deleting data is about $2 higher than before the data deletion. Judging from the summary statistics here, the deleting of data doesn't change the underlying characteristics of bond price by much.

The weight column for competition's evaluation purpose was also deleted.

The trimmed dataset was named bondtrim.csv, imported into JMP 11, and transformed into a JMP file named bondtrim.jmp.

Here is a variable dictionary.

Variables	Descriptions
id	The row id.
bond_id	The unique id of a bond to aid in time series reconstruction.
trade_price	The price at which the trade occurred.
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 occurred 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 (in seconds) 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.

Table 3. Variable Dictionary

The variable curve_based_price is the price that is calculated and provided by Benchmark Solutions. Below is the reason given by the company why the variable is included in the data.

> *Pricing bonds accurately requires an exacting knowledge of payment schedules, trading calendars and reference data for each bond. This, as well as synthesizing all of the bonds and CDS quotes and trades of a given issuer into implied hazard and funding curves, is something that we feel is beyond the scope of this challenge. Rather, we provide you with a reference price which is an intermediate result of our calculations and is labeled 'curve_based_price' in the dataset.*

4. Descriptive Analysis

1). Summary Statistics

As indicated before, there are 148620 trade price observations in the data, with no missing value.

After doing a summary for all the variables in the data, we get the following table. Summary statistics for variable trade_price is not shown here, since we've already provided tables in the previous part.

Column	N	Mean	Std Dev	Minimum	Maximum
current_coupon	148620	5.999	1.5315	0	13.5
is_callable	148620	0.0669	0.2498	0	1
reporting_delay	148620	33196.8	1252518	-57.373	99900000
time_to_maturity	148620	5.9665	7.0849	0.0989	84.1595
trade_price_last1	148620	105.395	9.5528	27	153.059
trade_price_last2	148620	105.392	9.5535	27	153.059
trade_price_last3	148620	105.389	9.5541	27	153.311
trade_price_last4	148620	105.386	9.5544	27	153.311
trade_price_last5	148620	105.383	9.5537	27	153.832
trade_price_last6	148620	105.381	9.5536	27	153.832
trade_price_last7	148620	105.379	9.5533	27	153.832
trade_price_last8	148620	105.377	9.5535	27	153.832
trade_price_last9	148620	105.375	9.5537	27	153.832
trade_price_last10	148620	105.373	9.553	27	153.832
curve_based_price	148620	105.241	9.6874	28.6191	152.078
curve_based_price_last1	148620	105.239	9.6892	28.6191	153.022
curve_based_price_last2	148620	105.236	9.6901	28.6731	153.072
curve_based_price_last3	148620	105.234	9.6908	28.6731	153.24
curve_based_price_last4	148620	105.231	9.6918	28.6731	153.24
curve_based_price_last5	148620	105.229	9.6923	28.6731	153.24
curve_based_price_last6	148620	105.228	9.6926	28.6731	153.24
curve_based_price_last7	148620	105.226	9.6932	28.6731	153.24
curve_based_price_last8	148620	105.224	9.6936	28.6731	153.24
curve_based_price_last9	148620	105.223	9.6946	28.6731	153.24
curve_based_price_last10	148620	105.223	9.6953	28.7	153.24
received_time_diff_last1	148620	32879.1	146532	0	7162758
received_time_diff_last2	148620	65760.4	213303	0	8382689
received_time_diff_last3	148620	98605.6	274300	0	8383730
received_time_diff_last4	148620	131395	331069	0	9072611
received_time_diff_last5	148620	163966	384380	0	9161295
received_time_diff_last6	148620	196504	436743	0	9596043
received_time_diff_last7	148620	229144	489405	0	9596043
received_time_diff_last8	148620	261691	540836	0	10300000
received_time_diff_last9	148620	294406	592275	0	10400000
received_time_diff_last10	148620	326921	641898	0	11000000
trade_size	148620	234260	735714	1000	5000001
trade_size_last1	148620	234006	734841	1000	5000001
trade_size_last2	148620	234274	735714	1000	5000001
trade_size_last3	148620	234350	735585	1000	5000001
trade_size_last4	148620	234462	735792	1000	5000001
trade_size_last5	148620	234482	736172	1000	5000001
trade_size_last6	148620	234427	736330	1000	5000001
trade_size_last7	148620	234648	737160	1000	5000001
trade_size_last8	148620	234971	737937	1000	5000001
trade_size_last9	148620	234950	737676	1000	5000001
trade_size_last10	148620	235206	738439	1000	5000001
trade_type	148620	3.2181	0.787	2	4

trade_type_last1	148620	3.2184	0.7869	2	4
trade_type_last2	148620	3.2188	0.787	2	4
trade_type_last3	148620	3.2184	0.7871	2	4
trade_type_last4	148620	3.2184	0.7873	2	4
trade_type_last5	148620	3.2187	0.7872	2	4
trade_type_last6	148620	3.2186	0.7873	2	4
trade_type_last7	148620	3.2185	0.7873	2	4
trade_type_last8	148620	3.2186	0.7874	2	4
trade_type_last9	148620	3.2185	0.7874	2	4
trade_type_last10	148620	3.2179	0.7877	2	4

Table 4. Summary Statistics

As stated before, there is no missing data here, and it's obvious that the summary statistics for lagged trade prices and curved based prices are very similar with each other, which indicates there might be strong correlations among them.

Since there are too many variables, we cannot not show all the histograms here. Histograms for a selected group of variables including **current_coupon, is_callable, reporting_delay, time-to-maturity, curve_based_price, received_time_diff_last1, trade_size, trade_size_last1, trade_type, trade_type_last1** are provided below. The histogram for **trade_price** is already provided before, so it will not be shown here. **Curve_based_price_last1** through **curve_based_price_last10**, **trade_price_last1** through **trade_price_last10** and **received_time_diff_last2** through **received_time_diff_last10** are not shown because their distribution is very similar to the distribution of **curve_based_price, trade_price** and **received_time_diff_last1,** respectively**. Trade_size_last2** through **trade_size_last10** and **trade_type_last2** through **trade_type_last10** are not included because they will not show up in our analysis model.

a). Current_coupon

There is nothing special about the current coupon of the bond. It roughly obeys a normal distribution.

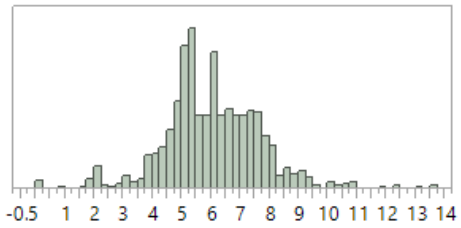


Figure 3. Histogram of current_coupon

b). Is_callable

0 indicates the bond is not callable, whereas 1 indicates the bond is callable. In most (more than 90%) of the trades the bonds are not callable.

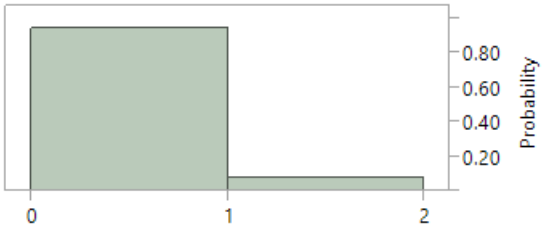


Figure 4. Histogram of is_callable

c). Reporting_delay

The distribution is highly skewed, and most of the observations are clustered to the leftmost side. Many outliers exist.

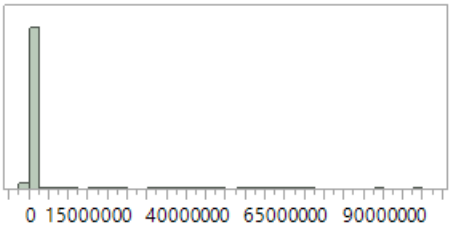


Figure 5. Histogram of reporting_delay

d). Time_to_maturity

The majority (99.5%) of bonds have time to maturity lower than 30 years, and a noticeable thing is that there are outliers around 80, which could possibly be measurement error.

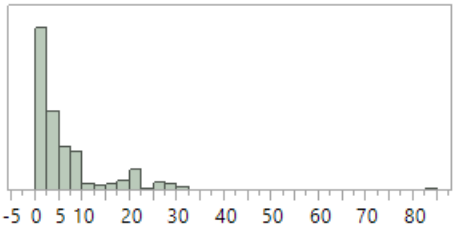


Figure 6. Histogram of time_to_maturity

e). Curve_based_price

The distribution is very similar to that of trade_price, which fact is consistent with argument that the curved based price is a good estimation of the real bond trade price.

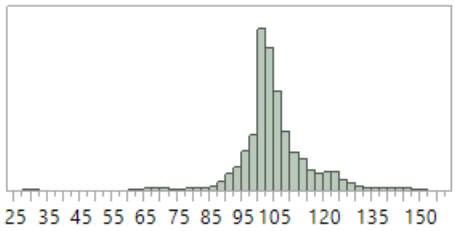


Figure 7. Histogram of curve_based_price

f). Received_time_diff_last1

Most (99.5%) of the trades happened within 852,114 seconds of the previous transaction.

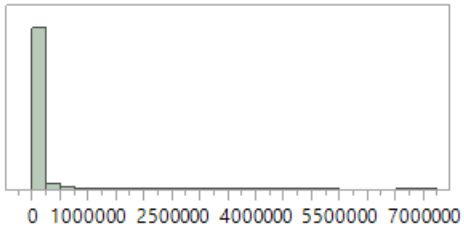


Figure 8. Histogram of received_time_diff_last1

g). Trade_size

Most (90%) of the trades have size smaller than $500,000, despite the fact that there exists some extremely large trade sizes.

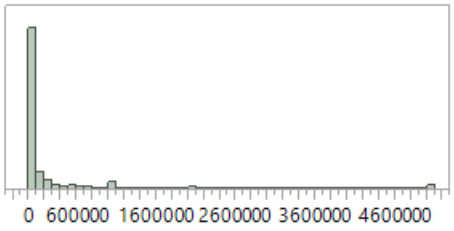


Figure 9. Histogram of trade_size

h). Trade_size_last1

The distribution is almost identical to trade_size. Most (90%) of the trades have size smaller than $500,000.

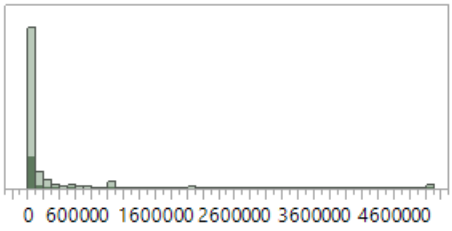


Figure 10. Histogram of trade_size_last1

i). Trade_type

22% of the trade are customer sells (trade_type=2), 33% of the trades are customer buys (trade_type=3), and 44% of the trades happen between dealers (trade_type=4).

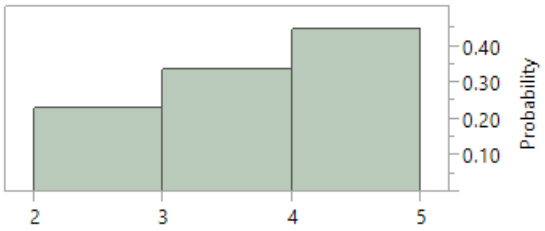
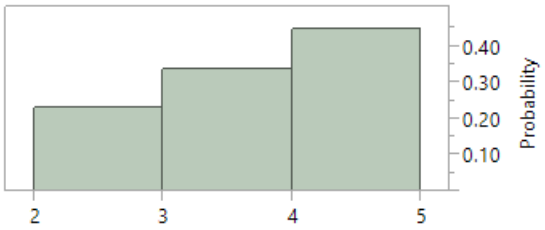


Figure 11. Histogram of trade_type

j). Trade_type_last1

The distribution is almost identical to that of trade_type. 22% of the trade are customer sells (trade_type=2), 33% of the trades are customer buys (trade_type=3), and 44% of the trades happen between dealers (trade_type=4).



b). Correlations

After a multivariate analysis for all the variables and a glance at the color maps, we found there are two major areas that are painted with light red colors (indicating high positive correlations), which respectively corresponds to variables **trade_price** through **curve_based_price_last10** and **received_time_diff_last1** through **received_time_diff_last10**, as demonstrated by Figure 13.

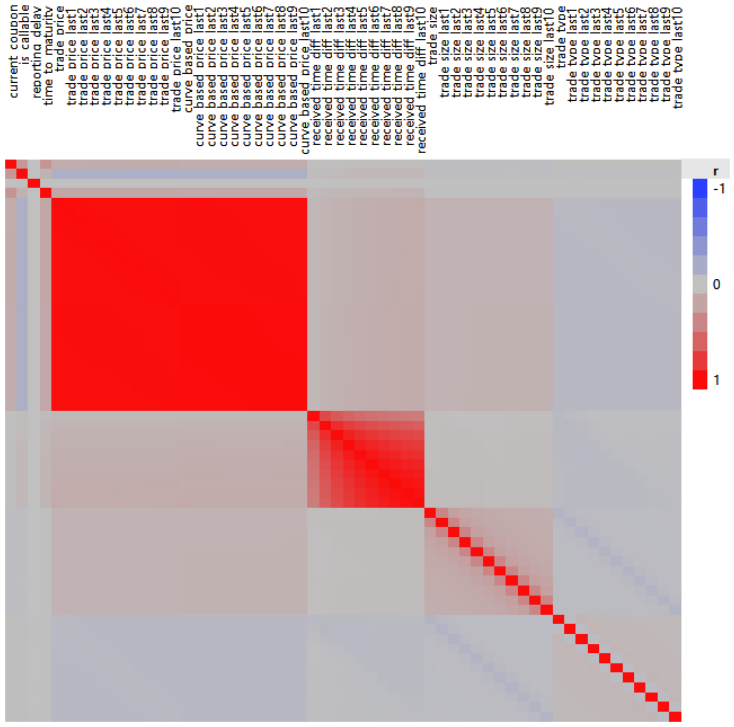
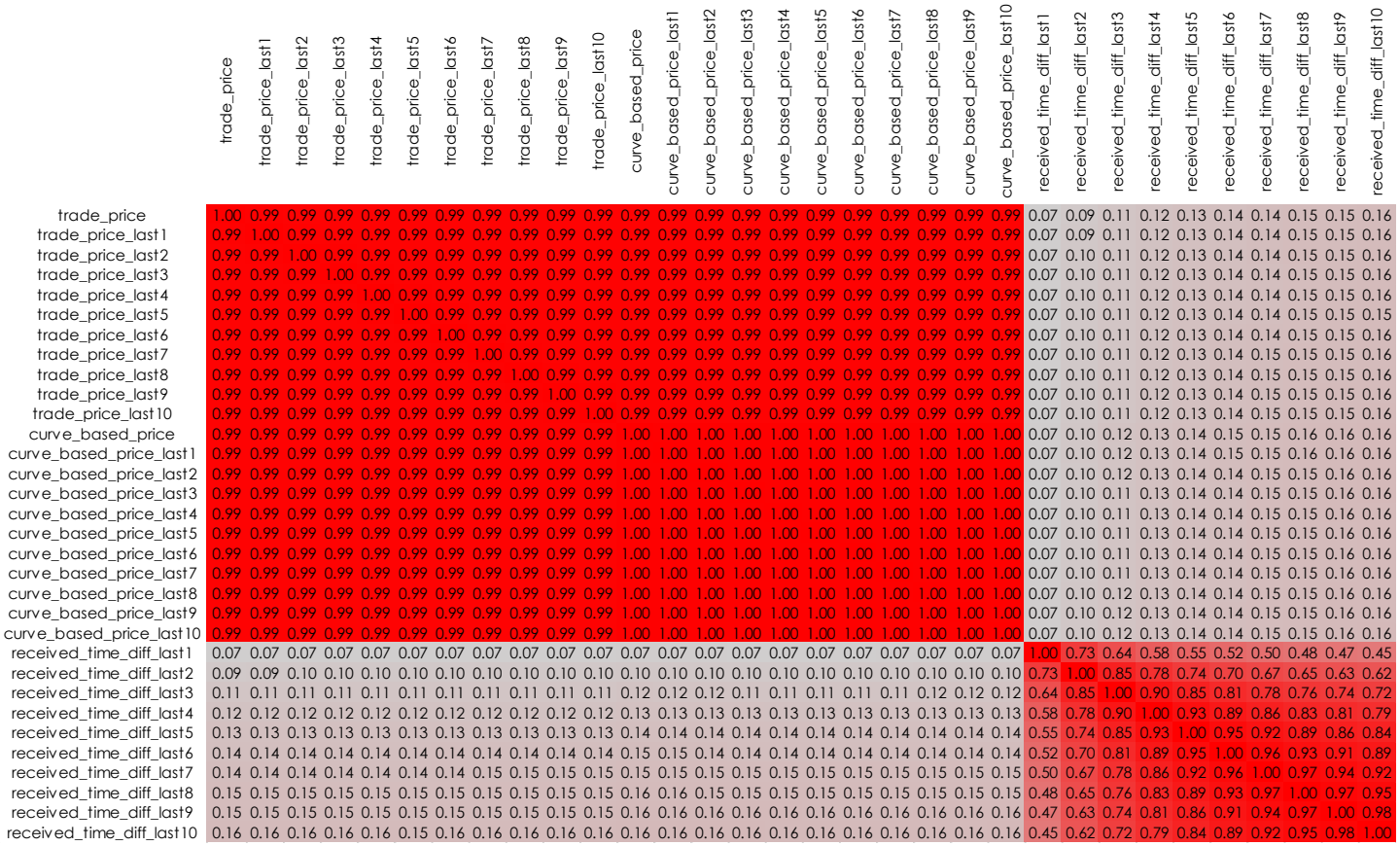


Figure 13. Correlation Color Map for all Variables

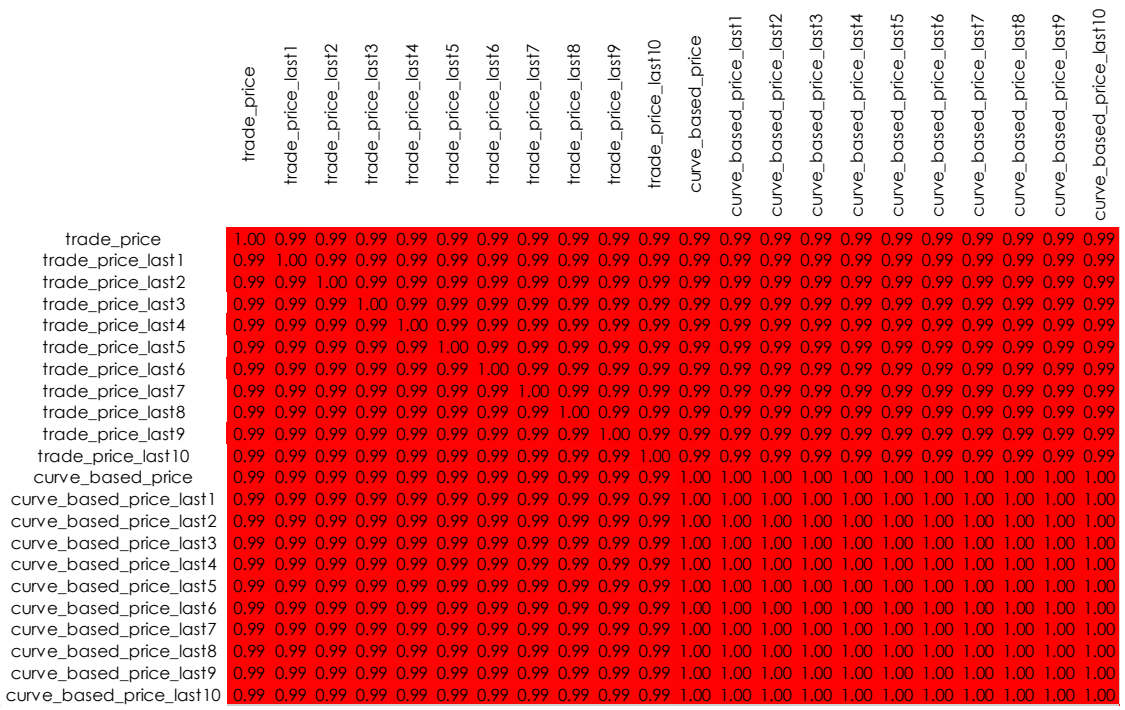
Now include only variables that are strongly correlated in our multivariate analysis. The result if shown below.



	trade_price	trade_price_last1	trade_price_last2	trade_price_last3	trade_price_last4	trade_price_last5	trade_price_last6	trade_price_last7	trade_price_last8	trade_price_last9	trade_price_last10	curve_based_price	curve_based_price_last1	curve_based_price_last2	curve_based_price_last3	curve_based_price_last4	curve_based_price_last5	curve_based_price_last6	curve_based_price_last7	curve_based_price_last8	curve_based_price_last9	curve_based_price_last10	received_time_diff_last1	received_time_diff_last2	received_time_diff_last3	received_time_diff_last4	received_time_diff_last5	received_time_diff_last6	received_time_diff_last7	received_time_diff_last8	received_time_diff_last9	received_time_diff_last10
trade_price	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.09	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.16
trade_price_last1	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.09	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.16
trade_price_last2	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.16
trade_price_last3	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.16
trade_price_last4	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.16
trade_price_last5	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.15
trade_price_last6	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.14	0.15	0.15	0.16
trade_price_last7	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.15	0.15	0.15	0.16
trade_price_last8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.15	0.15	0.15	0.16
trade_price_last9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.15	0.15	0.15	0.16
trade_price_last10	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.07	0.10	0.11	0.12	0.13	0.14	0.15	0.15	0.15	0.16
curve_based_price	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.12	0.13	0.14	0.15	0.15	0.16	0.16	0.16
curve_based_price_last1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.12	0.13	0.14	0.15	0.15	0.16	0.16	0.16
curve_based_price_last2	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.12	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last3	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.11	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last4	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.11	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last5	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.11	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last6	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.11	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last7	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.11	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.12	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.12	0.13	0.14	0.14	0.15	0.15	0.16	0.16
curve_based_price_last10	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10	0.12	0.13	0.14	0.14	0.15	0.15	0.16	0.16
received_time_diff_last1	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	1.00	0.73	0.64	0.58	0.55	0.52	0.50	0.48	0.47	0.45
received_time_diff_last2	0.09	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.73	1.00	0.85	0.78	0.74	0.70	0.67	0.65	0.63	0.62
received_time_diff_last3	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.64	0.85	1.00	0.90	0.85	0.81	0.78	0.76	0.74	0.72
received_time_diff_last4	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.58	0.78	0.90	1.00	0.93	0.89	0.86	0.83	0.81	0.79
received_time_diff_last5	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.55	0.74	0.85	0.93	1.00	0.95	0.92	0.89	0.86	0.84
received_time_diff_last6	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.15	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.52	0.70	0.81	0.89	0.95	1.00	0.96	0.93	0.91	0.89
received_time_diff_last7	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.50	0.67	0.78	0.86	0.92	0.96	1.00	0.97	0.94	0.92
received_time_diff_last8	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.48	0.65	0.76	0.83	0.89	0.93	0.97	1.00	0.97	0.95
received_time_diff_last9	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.47	0.63	0.74	0.81	0.86	0.91	0.94	0.97	1.00	0.98
received_time_diff_last10	0.16	0.16	0.16	0.16	0.16	0.15	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.45	0.62	0.72	0.79	0.84	0.89	0.92	0.95	0.98	1.00

Figure 14. Correlation Color Map for Selected Variables

Since the readability of table is not very good here, below we demonstrate the two areas separately. The first one is the correlation among variables **trade_price** through **curve_based_price_last10;** the second one is correlation among variables **received_time_diff_last1-10**.



	trade_price	trade_price_last1	trade_price_last2	trade_price_last3	trade_price_last4	trade_price_last5	trade_price_last6	trade_price_last7	trade_price_last8	trade_price_last9	trade_price_last10	curve_based_price	curve_based_price_last1	curve_based_price_last2	curve_based_price_last3	curve_based_price_last4	curve_based_price_last5	curve_based_price_last6	curve_based_price_last7	curve_based_price_last8	curve_based_price_last9	curve_based_price_last10
trade_price	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last1	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last2	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last3	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last4	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last5	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last6	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last7	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
trade_price_last10	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
curve_based_price	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last2	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last3	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last4	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last5	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last6	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last7	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
curve_based_price_last10	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 15. Correlation Color Map for Selected Variables A

	received_time_diff_last1	received_time_diff_last2	received_time_diff_last3	received_time_diff_last4	received_time_diff_last5	received_time_diff_last6	received_time_diff_last7	received_time_diff_last8	received_time_diff_last9	received_time_diff_last10
received_time_diff_last1	1.00	0.73	0.64	0.58	0.55	0.52	0.50	0.48	0.47	0.45
received_time_diff_last2	0.73	1.00	0.85	0.78	0.74	0.70	0.67	0.65	0.63	0.62
received_time_diff_last3	0.64	0.85	1.00	0.90	0.85	0.81	0.78	0.76	0.74	0.72
received_time_diff_last4	0.58	0.78	0.90	1.00	0.93	0.89	0.86	0.83	0.81	0.79
received_time_diff_last5	0.55	0.74	0.85	0.93	1.00	0.95	0.92	0.89	0.86	0.84
received_time_diff_last6	0.52	0.70	0.81	0.89	0.95	1.00	0.96	0.93	0.91	0.89
received_time_diff_last7	0.50	0.67	0.78	0.86	0.92	0.96	1.00	0.97	0.94	0.92
received_time_diff_last8	0.48	0.65	0.76	0.83	0.89	0.93	0.97	1.00	0.97	0.95
received_time_diff_last9	0.47	0.63	0.74	0.81	0.86	0.91	0.94	0.97	1.00	0.98
received_time_diff_last10	0.45	0.62	0.72	0.79	0.84	0.89	0.92	0.95	0.98	1.00

Figure 16. Correlation Color Map for Selected Variables A

Since there are too many variables here, we decide not to show scatterplot matrix.

According to the correlation analysis, there are strong correlations among variables **trade_price** through **curve_based_price_last10** and variables **received_time_diff_last1** through **received_time_diff_last10**, so it would be reasonable to do a factor analysis for these variables.

5. Analysis

1). Factor Analysis

According to the correlation color maps, there are strong correlation (marked by light red color) among variables **trade_price** through **curve_based_price_last10** and variables **received_time_diff_last1** through **received_time_diff_last10,** while among all the other variables the correlations are quite weak (marked by grey color). Therefore, it might be wise to conduct a factor analysis only for those highly-correlated variables. However, it would not hurt to do a factor analysis for all the variables first (curve_based_price with no lag is not included here, because we want compare our result with the prediction made by curve_based_price along).

As indicated by Figure 17, there are 13 factors having eigenvalue over 1. Do a factor analysis with these 13 factors, and we have the following variables with communality estimates over 0.5 (trade_size_last2 is included because it's the variable with the largest communality below 0.5).

As we can see, time_to_maturity, is_callable, reporting_delay, all the trade_size variables, and all the trade_type variables are not in Table 5, which means they have communality lower than 0.5. This is consistent with our correlation analysis.

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	20.4467	36.512		36.512
2	7.8590	14.034		50.546
3	3.2284	5.765		56.311
4	1.4602	2.608		58.918
5	1.3916	2.485		61.403
6	1.2922	2.308		63.711
7	1.2014	2.145		65.856
8	1.1758	2.100		67.956
9	1.1473	2.049		70.005
10	1.0910	1.948		71.953
11	1.0390	1.855		73.808
12	1.0252	1.831		75.639
13	1.0013	1.788		77.427

Figure 17. Eigenvalues of Factor Analysis

Variable	Communality	Variable	Communality
curve_based_price_last2	0.99993	trade_price_last7	0.9911
curve_based_price_last9	0.99988	trade_price_last5	0.99028
curve_based_price_last4	0.99973	trade_price_last6	0.99018
curve_based_price_last7	0.99966	trade_price_last1	0.98939
curve_based_price_last6	0.99954	trade_price_last10	0.9891
curve_based_price_last3	0.99952	received_time_diff_last7	0.98077
curve_based_price_last5	0.99952	received_time_diff_last8	0.97218
curve_based_price_last8	0.99949	received_time_diff_last6	0.96036
curve_based_price_last1	0.99909	received_time_diff_last10	0.95868
curve_based_price_last10	0.99904	received_time_diff_last5	0.95638
received_time_diff_last9	0.99465	received_time_diff_last4	0.94012
trade_price_last3	0.9924	received_time_diff_last2	0.9022
trade_price_last8	0.99235	received_time_diff_last3	0.89523
trade_price_last2	0.99166	current_coupon	0.79443
trade_price_last9	0.99157	received_time_diff_last1	0.58537
trade_price_last4	0.99123	trade_size_last2	0.38704

Table 5. Final Communality Estimates

Now include only variables that have communality over 0.50 in the factor analysis. We found loadings on almost all the variables are big enough, with the exception of curren_coupon. So we excluded the current_coupon and do the factor analysis again.

Variable	Factor 1	Factor 2	Variable	Factor 1	Factor 2
curve_based_price_last5	0.9967558	0.0755662	trade_price_last3	0.988106	0.0709214
curve_based_price_last6	0.9967174	0.0756097	trade_price_last9	0.9879556	0.0727749
curve_based_price_last4	0.9966311	0.075825	trade_price_last2	0.9879377	0.0711449
curve_based_price_last7	0.9965435	0.0758054	trade_price_last1	0.9877272	0.07134
curve_based_price_last3	0.9963856	0.0764016	trade_price_last10	0.9877253	0.0730398
curve_based_price_last8	0.996278	0.0761715	received_time_diff_last7	0.0749974	0.9786968
curve_based_price_last2	0.9960705	0.0771177	received_time_diff_last8	0.0798179	0.9777972
curve_based_price_last9	0.9959457	0.0766309	received_time_diff_last9	0.0846099	0.9641723
curve_based_price_last1	0.9957177	0.0777378	received_time_diff_last6	0.0707133	0.9605844
curve_based_price_last10	0.9955813	0.0770425	received_time_diff_last10	0.089827	0.9467116
trade_price_last5	0.9883508	0.0705643	received_time_diff_last5	0.0658023	0.9277274
trade_price_last6	0.9883347	0.0712355	received_time_diff_last4	0.0601794	0.8787479
trade_price_last7	0.9882685	0.0718468	received_time_diff_last3	0.0534734	0.8063649
trade_price_last4	0.9882372	0.0706384	received_time_diff_last2	0.0453459	0.6994366
trade_price_last8	0.9881275	0.0723692	received_time_diff_last1	0.0318191	0.5214407

Table 6. Factor Loadings on Variables

Every variable has at least 1 factor loading bigger than 0.5. Save the two factors and we will use them instead of original variables in our following analysis. A closer look at the loading suggests that factor 1 has high loadings on variables **curve_based_price_last1** through **trade_price_last10,** and factor 2 has high loadings on variables **received_time_diff_last1** through **received_time_diff_last10**. Therefore, it might be appropriate to describe factor 1 as an indication of "Lagged Prices" and factor 2 a measurement of "Trade Frequency".

As for other variables, we did a factor analysis again, but the results are not satisfying.

2). Partition Tree.

We include factor 1, factor 2 from our factor analysis, and all the other variables into our neural network model first, and let's call it Model 1 here. Also we did a neural network analysis without variables **trade_size_last1** through **trade_size_last10** and **trade_type_last1** through **trade_type_last10,** because we think lagged values of trade size and can only have minimal effect on present bond price**.** Let's call this model Model 2.

We found that the resulting RMSE of Model 2 is smaller than Model 1, so we decide to exclude variables **trade_size_last1** through **trade_size_last10** and **trade_type_last1** through **trade_type_last10** from our analysis. In both models validation proportion is set to 0.3.

Below are the results for Model 2.

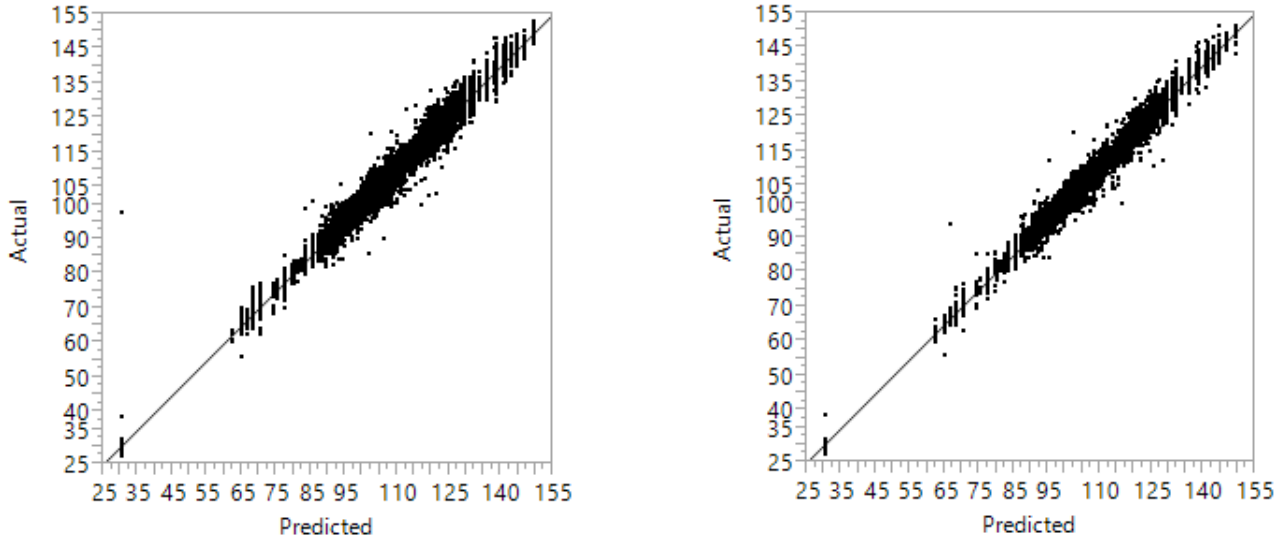


Figure 18. Actual vs. Predicted for Training and Validation Data

	RSquare	RMSE	N	Number of Splits	AICc
Training	0.988	1.0628332	104006	686	309217
Validation	0.985	1.1581579	44614		

Table 7. Neural Network Result

Term	Number of Splits	SS	SS	Portion
curve_based_price	159	9068443.36		0.9724
Factor1	69	202988.704		0.0218
trade_type	93	27218.9898		0.0029
time_to_maturity	189	13407.5236		0.0014
current_coupon	83	7307.14181		0.0008
is_callable	16	3137.84712		0.0003
trade_size	38	1944.80409		0.0002
Factor2	25	1157.34917		0.0001
reporting_delay	14	466.276043		0.0000

Table 8. Column Contribution of Partition Tree

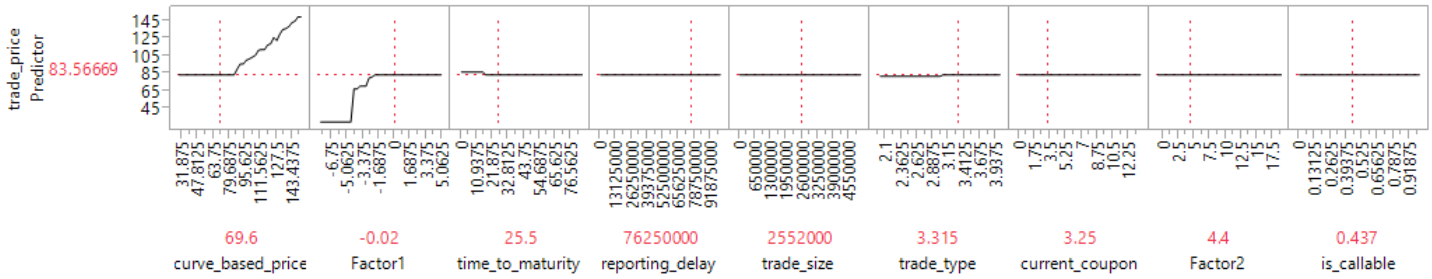


Figure 19. Profiler of Partition Tree Analysis

The R^2 is very high here, but much of the contribution is made by curved_based_price, which once again demonstrates that curved_based_price is itself a good estimation of the actual price. However, as we will show in our Summary part, we increase the accuracy of the original calculation model by improving RMSE, although not by much.

An analysis with bootstrap forest or booted trees might generate better parameters, but we did this project with the ordinary version of JMP 11, which doesn't provide those options, so we will just make do with the normal partition tree here.

3). Neural Network.

Based on the same thinking as in our Partition Tree analysis, we include **factor 1, factor 2** from our factor analysis, **current_coupon, time_to_maturity, is_callable, reporting_delay, trade_size** and **trade_type** into our Neural Network analysis.

The holdback proportion is set to 0.33, and hidden notes are set to 5.

Measures	Value
RSquare	0.9853575
RMSE	1.1601133
Mean Abs Dev	0.7240052
-LogLikelihood	155309.83
SSE	133353.47
Sum Freq	99084

Table 9. Result for Training Data

Measures	Value
RSquare	0.9854305
RMSE	1.1440344
Mean Abs Dev	0.7188651
-LogLikelihood	76954.153
SSE	64833.451
Sum Freq	49536

Table 10. Result for Validation Data

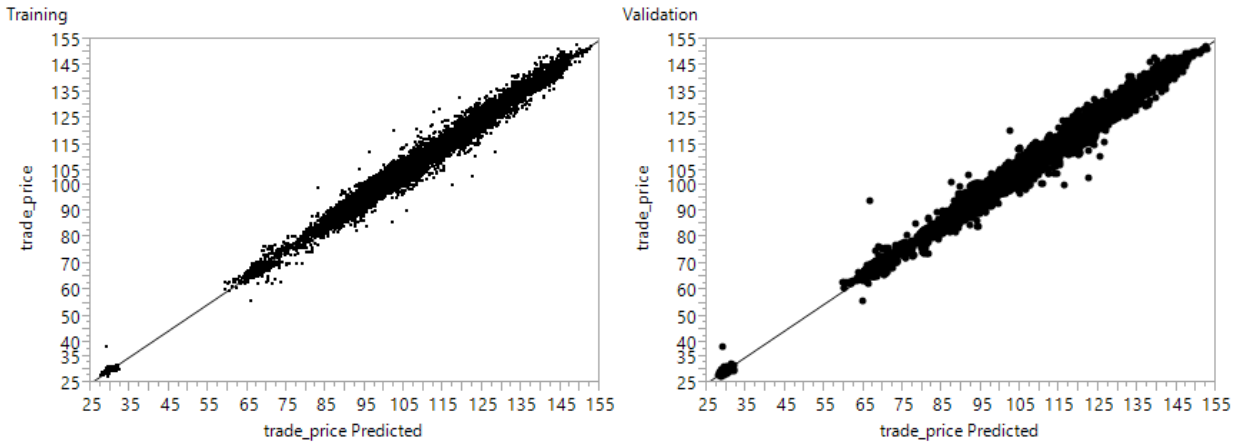


Figure 20. Actual vs. Predicted for Training and Validation Data

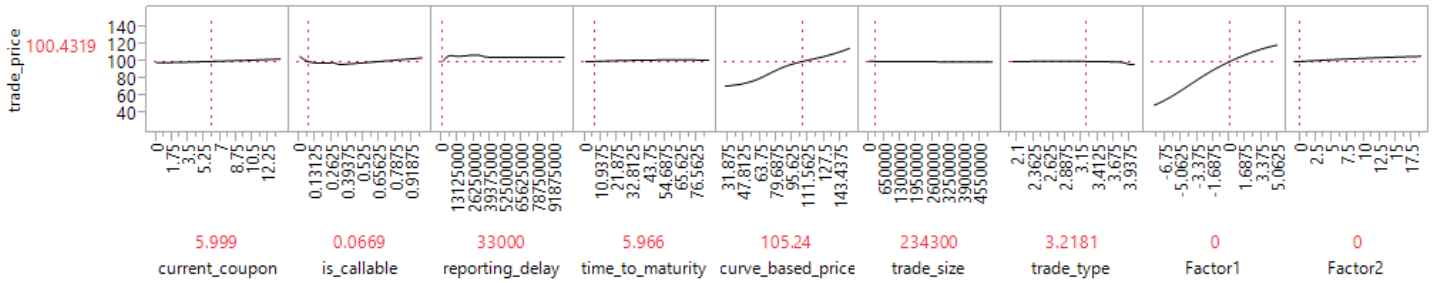


Figure 21. Profiler for Neural Network

4). Comparison

In general neither of the two models is dominating over the other, which is reasonable, because with the existence of variable curve_based_price in our model, there is no much room for improvement.

The Partition Tree model has a R^2 of 0.988 in training part of data, while the for Neural Network model the number is 0.9853, which is not very different from 0.987.

In terms of RMSE, the Partition Tree model performs a little better than the Neural Network model. Since lower RMSE in some way means a more accurate prediction, so in our analysis I would decide that Partition Tree model is the better one of them.

5). Cluster Analysis

There are 148620 observations of bond trade occurrence in our data, and it is not easy to distinguish one trade from another, so I think cluster analysis would not be very meaningful here.

Below is a tentative exercise of cluster analysis. Due to the bulky essence of our data, hierarchical clustering would be disastrous here, so we go with the k-means methods.

Set trade_price, trade_type and is_callable as clustering principles, and number of clusters range from 2 to 10, the Cubic Clustering Criterion decides 5 clusters is the best one.

Method	NCluster	CCC	Best
K-Means Clustering	2	85.1117	
K-Means Clustering	3	47.6461	
K-Means Clustering	4	-94.825	
K-Means Clustering	5	242.545	Largest CCC
K-Means Clustering	6	190.567	
K-Means Clustering	7	110.763	
K-Means Clustering	8	101.818	
K-Means Clustering	9	160.341	
K-Means Clustering	10	56.5142	

Table 11. Cluster Analysis Result

From table 11 we can see that the frequency distribution is very uneven, as large as 42.96% and as small as 1.00%.

Cluster	Count	Percentage
1	18690	12.58%
2	1537	1.03%
3	54609	36.74%
4	63848	42.96%
5	9936	6.69%

Table 12. Cluster Summary

Have a look at the biplot graph, and we can see that with large area of overlapping between cluster 2 and 5, 2 principles cannot distinguish clusters well,

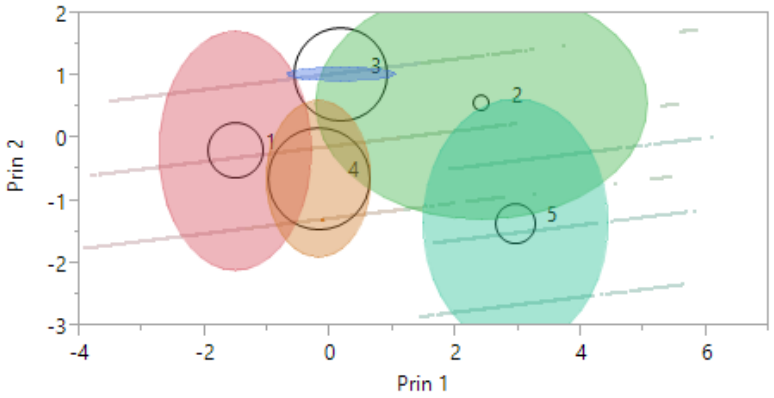


Figure 22. Biplot Graph for 5 Clusters Solution

Now have a look at biplot 3D graph, and we found 3 principles is also not good enough at distinguishing clusters.

We could roughly see the cluster that is marked by olive and green colors are better distinguished from the others, while the remaining 3 clusters are not well separated from each other. Therefore, it is not wise to use cluster analysis here.

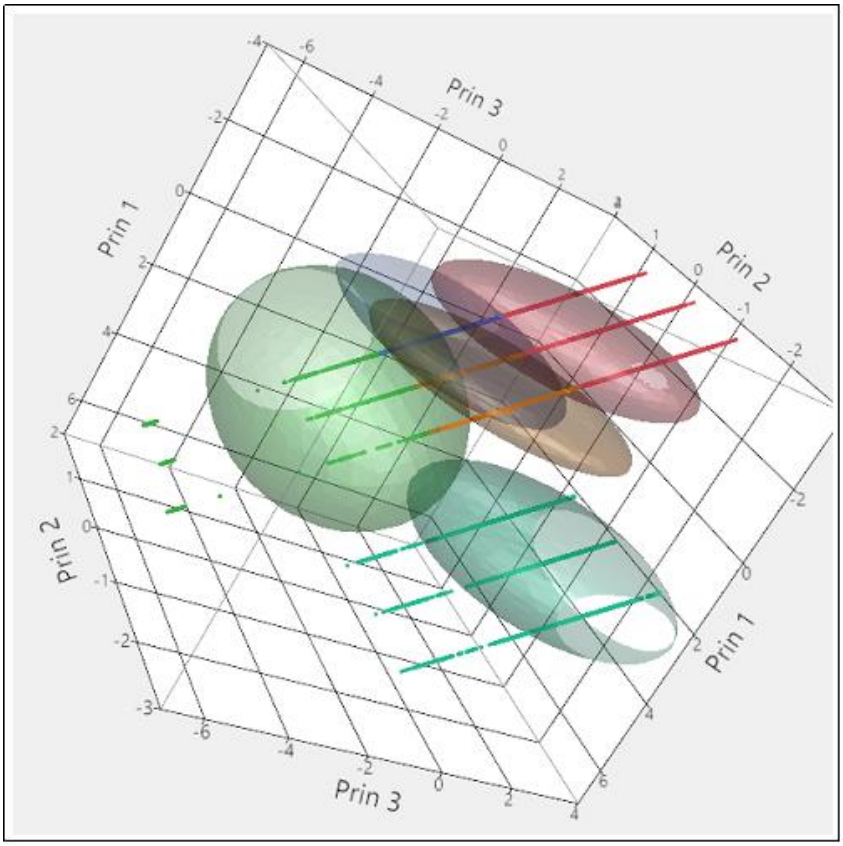


Figure 23. Biplot 3D Graph for 5 Clusters Solution

7. Summary

Now let's go back to our problem: would paying attention to "trading dynamics and microstructure of individual bonds" increase the accuracy of benchmarking actual prices? To answer this question, we must first know how the curve based price performs in predicting actual price.

Plot trade_price over curve_based_price and fit linear line on it, we get a simple univariate linear regression of trade_price on curve_based_price.

Results are shown below.

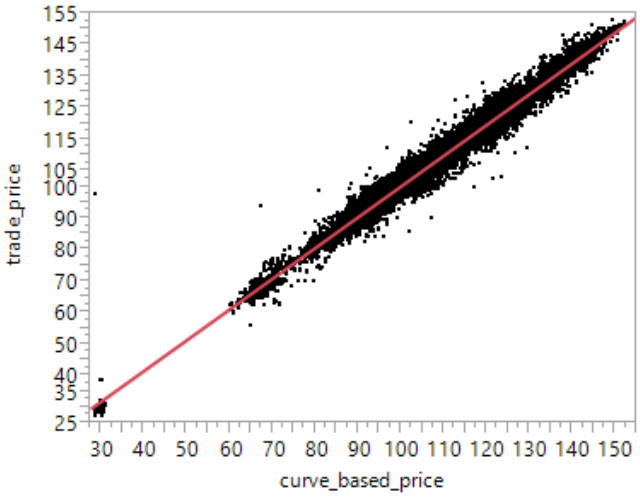


Figure 24. Fitted Line of trade_price vs. curve_based_price

RSquare	0.980494
RSquare Adj	0.980494
Root Mean Square Error	1.333943
Mean of Response	105.3967
Observations (or Sum Wgts)	148620

Table 13. Regression Result

We can see that curve_based_price itself is already good enough at predicting the actual bond price, with R^2 high as 0.98, and RMSE as 1.334, which is not very different from the results we get from our Partition Tree and Neural Network analyses.

However, our model does improve the prediction accuracy of the original model. By including other relevant variables into the model, in our Partition Tree analysis, we get a RMSE that is around 0.27 smaller than the one that results from the univariate model.

Although the improvement is not very significant, it does proves that "trading dynamics and microstructure of individual bonds" have effect on the variation of actual bond price from the curve based price, and this effect can be benchmarked by including variables such as bond type and trade size into the model.

However, the biggest problem with our model is that too few of the variation is explained: the improvement on RMSE is only 0.27. Therefore, more work need to be done to more accurately benchmark the actual bond price. Generally speaking more relevant variables need to be identified to explain the variation of actual bond price from model calculated rice, and maybe redoing the partition tree analysis with bootstrap forest or boosted trees might be a good choice.

Also, in our model the time gap between lagged observations are not the same, since they are values for the last transaction, not the transaction that happened a certain time ago. Therefore a more rigorous analysis might want to weight lagged values according to the time differences with the actual trading date, which are provided in the data as **received_time_diff_last1-10**.