High Training Loss

Analysis Report

David Sasaki

October 21, 2022

**Overview**

During training of FMV model, we’ve encountered a problem that the training loss is a lot higher than the validation and test set losses. This report shows analysis to determine the cause of high training loss.

**Scope**

This report based on FMV Model Data gathered from records for the past 3 and half years.

**Methodology**

There are 2 hypotheses to determine the crucial cause that impact on training loss.

1. Extreme Outliers

There may be some extreme loss outliers in the data set that affect the training and evaluation of the model. The hypothesis is that there might be some huge outliers in terms of loss in the training set.

To prove the hypothesis, I needed to see what outliers exist, what patterns there are that explain. First, I identified the outliers. The analysis of this hypothesis based on data of May 27th, 2022.

I selected candidates of outliers with top-100 loss trades of the day. There were 28542 trades, so I selected 0.35% of the whole trades as candidates. Then, I detected each candidate is real outlier or not.

To detect real outliers, I used 2 methods.

First one is to find the sudden change in the price.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| figi | report\_date | price | loss | diff\_price |
| 28744 | 1.65E+18 | 88.607 | 0.16031 |  |
| 28744 | 1.65E+18 | 88.872 | 0.242285 | 0.265 |
| 28744 | 1.65E+18 | 89.897 | 1.016563 | 1.025 |
| 28744 | 1.65E+18 | 88.931 | 0.028158 | 0.966 |
| 28744 | 1.65E+18 | 88.931 | 0.231312 | 0 |
| 28744 | 1.65E+18 | 88.931 | 0.212234 | 0 |
| 28744 | 1.65E+18 | 88.927 | 0.235313 | 0.004 |
| 28744 | 1.65E+18 | 88.931 | 0.219177 | 0.004 |
| **28744** | **1.65E+18** | **99.931** | **10.85562** | **11** |
| 28744 | 1.65E+18 | 88.931 | 3.337603 | 11 |
| 28744 | 1.65E+18 | 88.931 | 1.854297 | 0 |

Table-1

This table shows trades of same bond. As you can see in this table, once the price is changed from 88.931 to 99.931, then, it is returned 88.931, again. This is a typical example of sudden change, and obviously, this trade is an outlier. This method detected 5 real outliers among 100 candidates.

This method is an excellent greedy approach. But, there were some mistakes to detect real outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| figi | report\_date | price | loss |
| 8686 | 1.65E+18 | 114.461 | 5.192771 |
| **8686** | **1.65E+18** | **116.461** | **4.162824** |
| 8686 | 1.65E+18 | 114.461 | 5.246438 |

Table-2

Look at this table. The price is changed from 114.461 to 116.461, then, returned. According to this method, it is a sudden change. But, it’s not a real outlier. That’s because the loss of the trade is lower than the previous one and the next one.

|  |  |  |  |
| --- | --- | --- | --- |
| figi | report\_date | price | loss |
| 19769 | 1.65E+18 | 83.75 | 0.167824 |
| 19769 | 1.65E+18 | 83.75 | 0.020135 |
| 19769 | 1.65E+18 | 85.477 | 1.590473 |
| 19769 | 1.65E+18 | 85.509 | 1.141658 |
| 19769 | 1.65E+18 | 85.447 | 1.079658 |
| 19769 | 1.65E+18 | 86.447 | 1.432809 |
| 19769 | 1.65E+18 | 86.233 | 1.186745 |
| 19769 | 1.65E+18 | 87.417 | 2.043553 |
| 19769 | 1.65E+18 | 85.061 | 0.621253 |
| 19769 | 1.65E+18 | 85.096 | 0.586253 |
| 19769 | 1.65E+18 | 85.061 | 0.362 |
| **19769** | **1.65E+18** | **81.367** | **4.429708** |
| **19769** | **1.65E+18** | **81.267** | **4.272357** |
| **19769** | **1.65E+18** | **81.167** | **4.137801** |
| **19769** | **1.65E+18** | **81.267** | **4.272357** |
| **19769** | **1.65E+18** | **81.267** | **4.272357** |
| 19769 | 1.65E+18 | 85.296 | 2.933132 |
| 19769 | 1.65E+18 | 85.234 | 3.212119 |
| 19769 | 1.65E+18 | 85.906 | 2.073587 |
| 19769 | 1.65E+18 | 85.806 | 2.420412 |
| 19769 | 1.65E+18 | 85.806 | 2.727143 |
| 19769 | 1.65E+18 | 85.806 | 2.458101 |
| 19769 | 1.65E+18 | 85.286 | 0.699526 |
| 19769 | 1.65E+18 | 85.286 | 0.382465 |

Table-3

And, in this table, there are 5 trades with high loss values. And the price was changed from 85.061 to 81.367, then, returned to 85.296. This is a kind of sudden change. But, actually, first method couldn’t detect these outliers. That’s why I used the second method.

Second method is to detect by statistics. I calculated mean and std of loss, then detected outlier that wasn’t in the range [mean-3\*std, mean+3\*std] as a real one. This method is one of the general approaches to find outliers in statistics. I wanted to find the general characteristics of outliers, and used this method.

Finally, it showed outliers are not really the main cause.

1. Histogram with buckets of time

The purpose of the histogram is to know where the maximum loss and volatility exists.

Here, the loss is MSE (mean squared error) of trade losses in each bucket.



And, the volatility is the mean value of set of averages of difference in price of each bond.



To analyze loss and volatility, I did histogram.

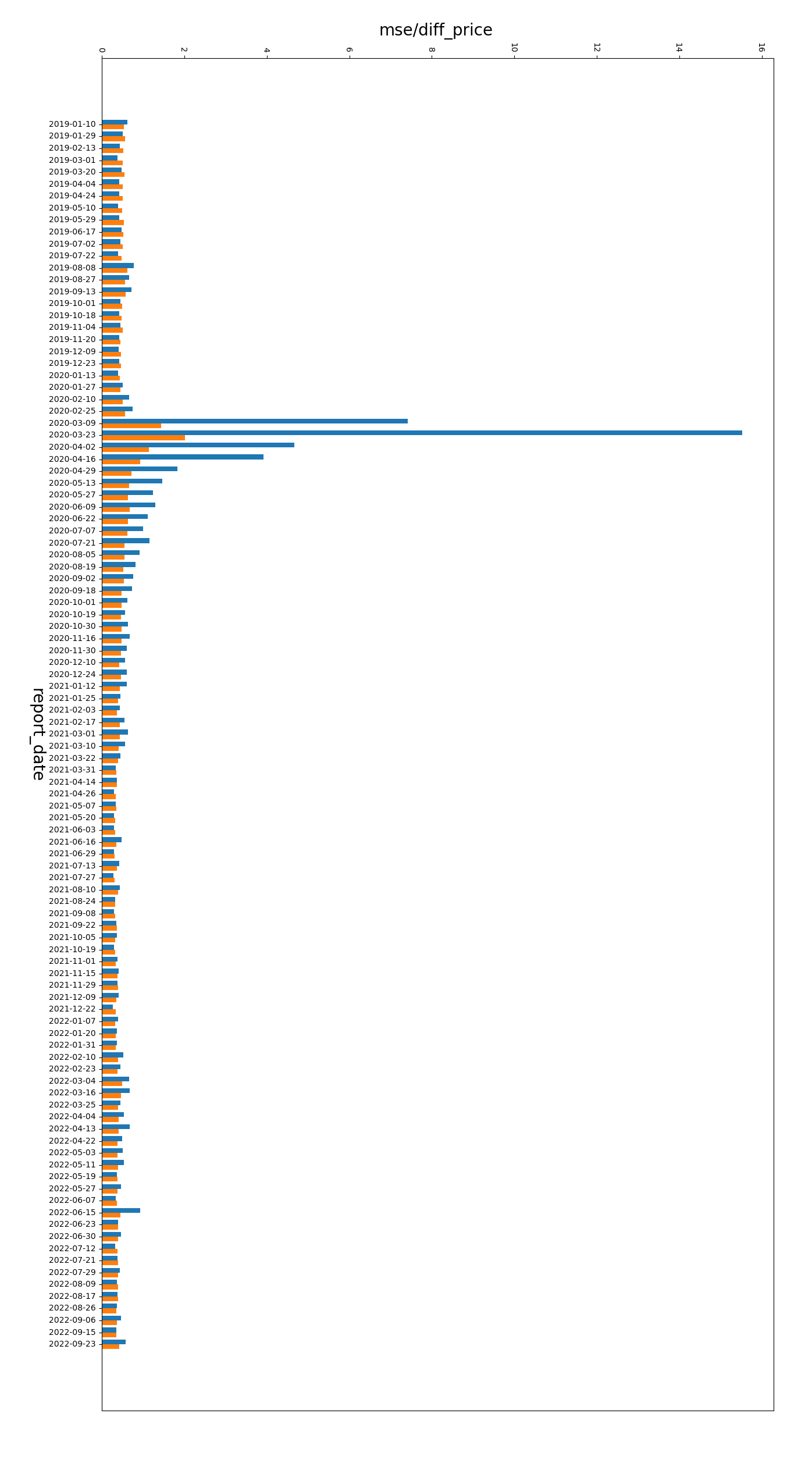
I divided the whole date range from 2019.1 ~ 2022.9 into 100 trade buckets. Each bucket consists of about 13~20 days - 300000 trades. I set the batch size to 300000. The whole data consists of about 30 million trades.

In each trade bucket, I calculated the mean squared error of trade losses. Then, I divided the bucket according to bond (figi), and sorted them by reported date. Then, for each bond, I got the difference of every consequent to calculate the difference average. The volatility is the mean value of set of difference averages.

As you can see in the below histogram, the blue rectangles represent loss (MSE) of each bucket, while the yellow rectangles represent volatility.

According to the histogram, the highest loss exists in 2020.3 ~ 2020.4, the beginning of pandemic. And there was also the highest volatility.

As a result, there are high market volatilities at the beginning of pandemic. And, that’s why the training data has so much higher loss than the validation and test sets.



**Conclusion**

When we train FMV model, the training loss is a lot higher than validation and test set losses. According to the above histogram, there are high losses at the beginning of pandemic with high volatilities. It shows that the main cause of high loss is high market volatility.

To improve the performance, first, it is needed to increase the number of training examples – trades at the beginning of pandemic. That helps the model to capture the features correctly to avoid underfitting.

Second, hyper-parameter tuning techniques are required. We can initialize weights differently, use different activation functions, and try different regularization methods. It is useful to find the possible best sets of hyper-parameters of the model to maximize its predictive accuracy.