Original features

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Description | Values | Decision |
| "TRADEDATE" | Trade session date | datetime | Keep |
| "SECID" | Security ID | str | Keep |
| "NUMTRADES" | Amount of trades during the day | float | keep |
| "VALUE" | Total aggregate value of all trades during the session | float | keep |
| "OPEN" | Opening price | Float | Highly correlated with LOW HIGH and CLOSE, more likely to drop |
| "LOW" | Minimum price reached during trading session | float | Highly correlated with OPEN, HIGH, CLOSE, more likely to drop |
| "HIGH" | Maximum price reached during trading session | float | Highly correlated with OPEN, LOW, CLOSE, more likely to drop |
| "LEGALCLOSEPRICE" | Closing price with legal restrictions | float | Almost identical to CLOSE, more likely to drop |
| "WAPRICE" | Weighted Average Price | float | More likely to drop |
| "CLOSE" | Closing price | float | For me best option out of OPEN, HIGH, LOW, more likely to keep |
| "VOLUME" | Total volume of security bought or sold during trading session | float | keep |
| "MARKETPRICE2" and "MARKETPRICE3" | Don’t know for sure what is it | float | Highly correlated with ADMITTEDQUOTE, more likely to drop |
| "ADMITTEDQUOTE" | Admitted quotation (price) of the instrument | float | I think keep this |
| "MP2VALTRD" | Ratio of market price to the total value traded | float | drop |
| "MARKETPRICE3TRADESVALUE" | Don’t know what is this for sure | float | drop |
| "ADMITTEDVALUE" | Admitted value of the instrument | float | keep |
| "TRENDCLSPR" | Closing price associated with trends or price changes | float | Keep |
| "currencyid" | Currency ID, like RUB | str | Only for plots |

After some investigation, I’ve reduced the number of features to 8:

* 'NUMTRADES',
* 'VALUE',
* 'CLOSE',
* 'VOLUME',
* 'ADMITTEDQUOTE',
* 'MARKETPRICE3TRADESVALUE',
* 'ADMITTEDVALUE',
* 'TRENDCLSPR'

What about derived features?

There are several features that can be extracted from this data.

Rolling Statistics: In pump cases, there is almost always a preparation phase characterized by volume growth before the actual pump. Therefore, we can calculate various rolling statistics such as sum, standard deviation, mean value, kurtosis, and skewness.

Choosing the Window Size: How do we determine the window size for these statistics? The answer is straightforward. By examining the data, we can identify patterns. One of the most obvious patterns is the weekly cycle: there are typically more deals made from Tuesday to Thursday and fewer on Mondays and Fridays. I’ve tried some and stopped at 3, 7, 10, 14, and 30 days windows.

Date Features Extraction: Another option is extracting date-related features. However, it's unclear how to effectively use them in classification problems. As mentioned earlier, there is a weekly pattern in the data. Theoretically, we can weight values based on the weekday, but this approach seems unusual. Its primary purpose would be to smooth out the weekly pattern to aid in more effective anomaly detection, but it's a questionable approach and can be skipped for now.

Lag Components: Lag components can also serve as valuable features. We can use simple percentage change or shift operations, and potentially more complex transformations, although these would need thorough testing to prove their effectiveness.

News Data: Currently, we have news data from the Moex news page, which provides technical information about securities. Initially, our interest was in changes in risk parameters. Unfortunately, the Moex news portal lacks substantial information about companies related to securities. Consequently, we can effectively use this data only for detecting TRUE POSITIVE cases.

Potential for Official Company News: If we could access official company news, it would be a powerful feature. Positive news about a new product or a significant event can act as a green light for investors. Conversely, if there's no news, and we observe sudden rapid growth, it might indicate a pump and dump process. Since our goal is to create a classifier, having growth cases with a valid reason could be highly valuable.

Forum Data: Active users often discuss noteworthy events on forums. When something good or unexpected happens, it tends to generate discussions. Considering we are dealing with low-liquidity securities, the volume of messages in a forum thread can be an informative feature.

In Conclusion: To build a precise classifier, it's advisable to consider not just two classes but three: 0 for normal activity, 1 for expected events, and 2 for pump and dump cases. This approach can enhance precision compared to a model with only normal and pump cases.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Description | Potential? | Access |
| Original features | Historical securities data | Good predictor | Yes |
| Date column features | Derived from date features | Questionable | Yes |
| Rolling window features | Derived from original features rolling statistics | Might be a good predictor | Yes |
| Lag features | Shifted values of original data | Might be good | Yes |
| Official news data | Official news from company news page | Might be good | Hard to get automatically, Not every company has one |
| Forum data | Security forum thread data | Might be good | Can be acquired, |