Price trend prediction using the high-frequence Limited Order book data

Nikita Bogomazov, Vyacheslav Blinov, Roman Vetrin

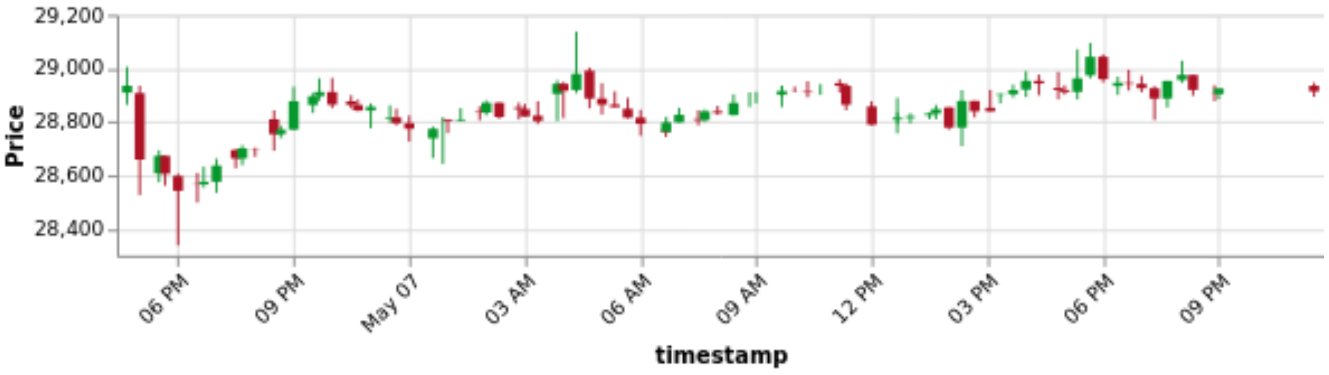
This section will present the results of Explorational Data Analysis and Predictive Data Analysis.

We will illustrate key insights from the data aquired as well as models considers and their results.

Exploratory Data Analysis

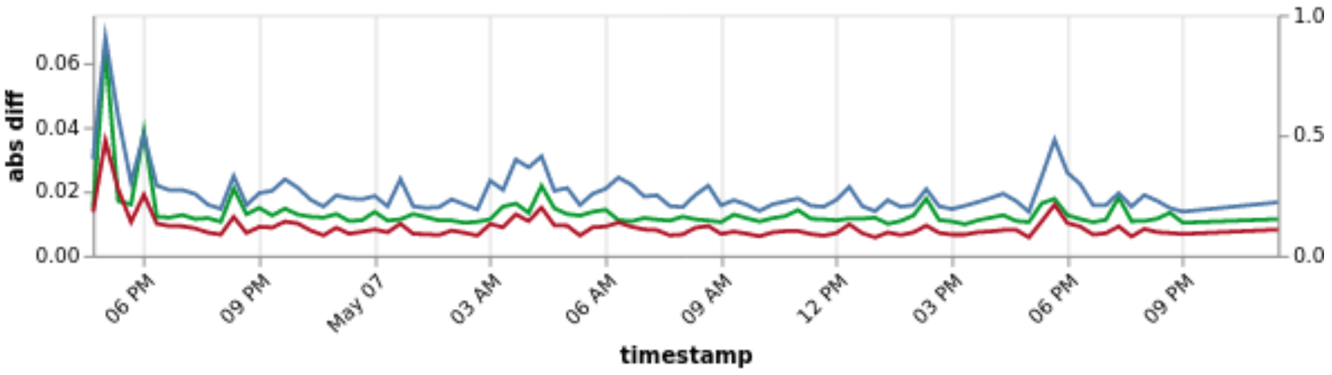
Our data represent snapshots from a bitcoin order book downloaded from the Binance exchange for one day. The order bool contains a price and volume for each of a 5 levels of all bids and asks at the given timestamp. In order to have overall view on the data, we reconstructed the popular "candlestick" chart below:

Plot 1: Reconstructed candlestick chart



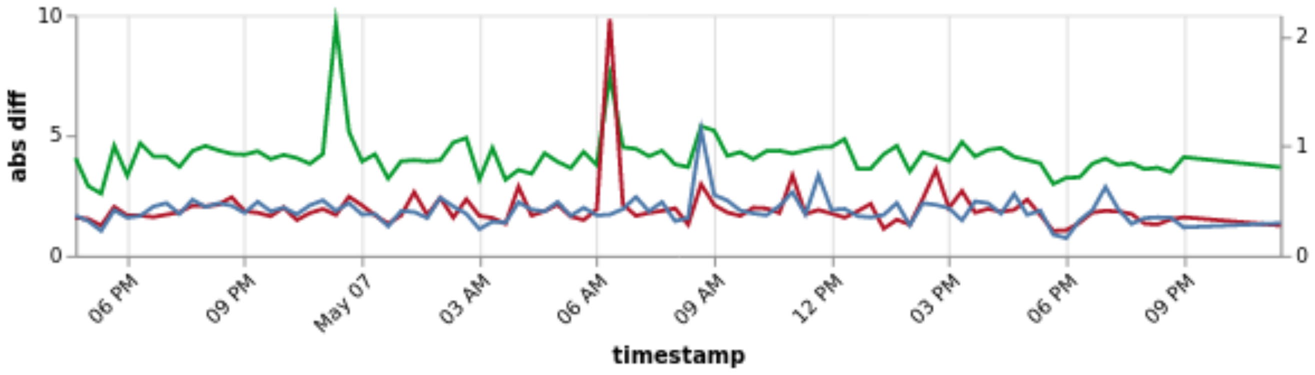
As we may observe, the price changes wichin appx. 5% margin within the day which is quite high comparing to less volitile assets. Further we will take a closer look into the order book microstructure.
First, we may look into differencies between bid and ask prices on various levels. Potentially, huge difference between prices should indicate unpredictability in future mid price which we could use as a valuable feature.

Plot 2: Difference between price levels 1(green, secondary axis), 2(red) and 3(blue)



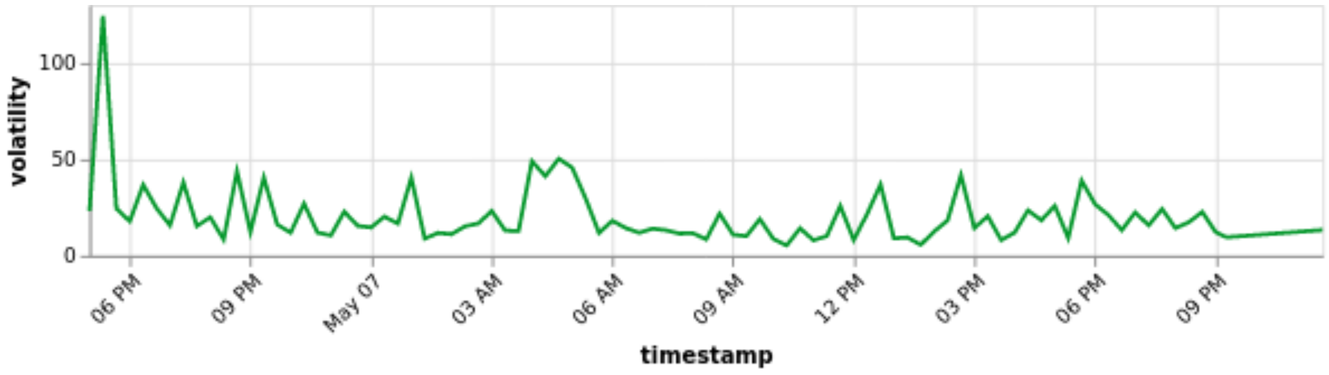
The plot above demonstrates that few sudden changes in candlestick graph corresponds to jumps in price difference between bids and asks. Such feature can be useful in the final model.
Additionally, we can take a look into difference between amount of coin available at each level. Shortage of the asset on the first levels should signify a price change in the following periods.

Plot 3: Difference between quantity levels 1(green, secondary axis), 2(red) and 3(blue)



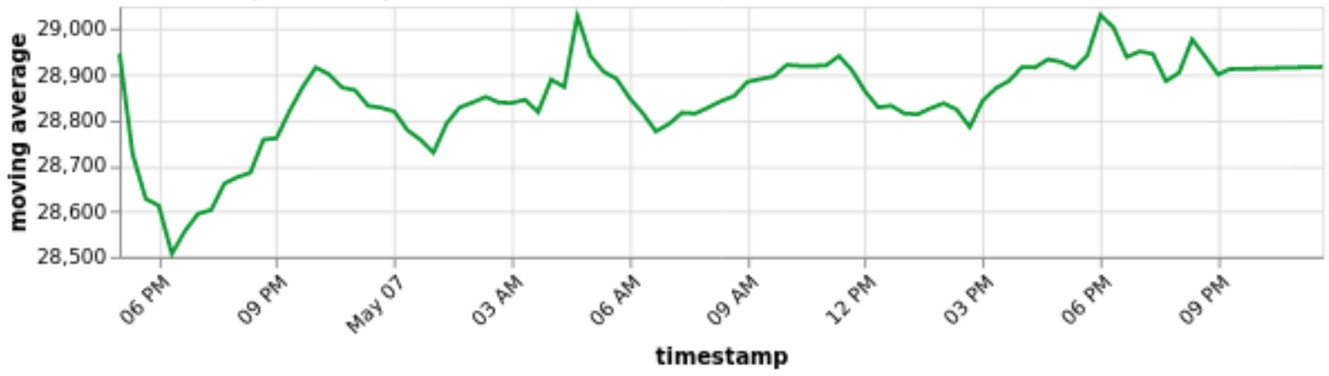
As is shown on the plot, some pikes in quantity difference correspond with future highly volitile displayed on the Plot 1. Further we used such feature during the model training.
Next we will take a look into volatility estimation of a given trend. This is one of the parameters to which should define the current trend and ensure that it is feasable to divide the timeseries into train and test dataset saving the overall trend properties.

Plot 4: Volatility value



According to the volatility values, it is feasable to divide the trend into the two parts as the levels seems overall constant throught the day.
Our task will be generation of a feature, signifying the ovarall trend direction. In order to achieve this, we will predict the moving average value 1s ahead of the given tick and output the corresponding binary signal. The overall moving average value is displayed below:

Plot 5: Moving average

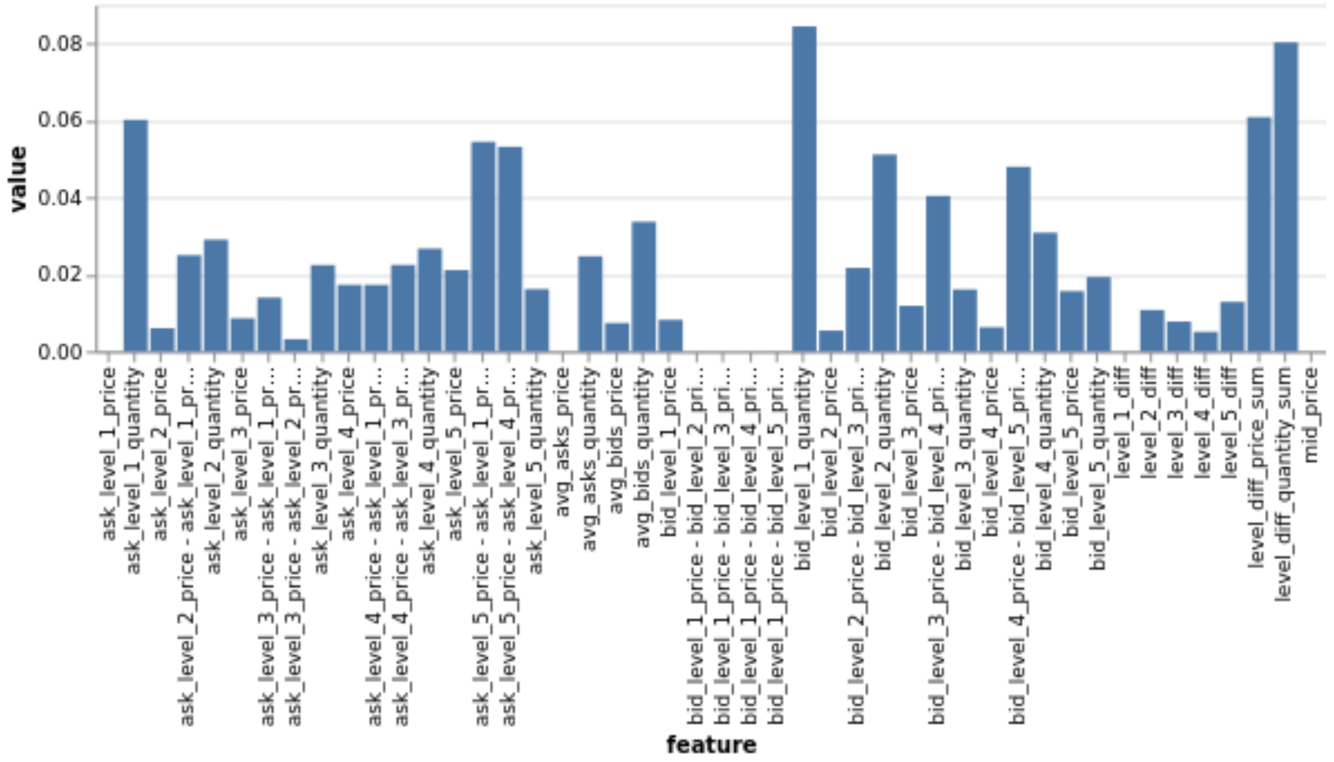


Prediction of the moving average allow us to output useful feature for overall trend direction.

PDA analysis

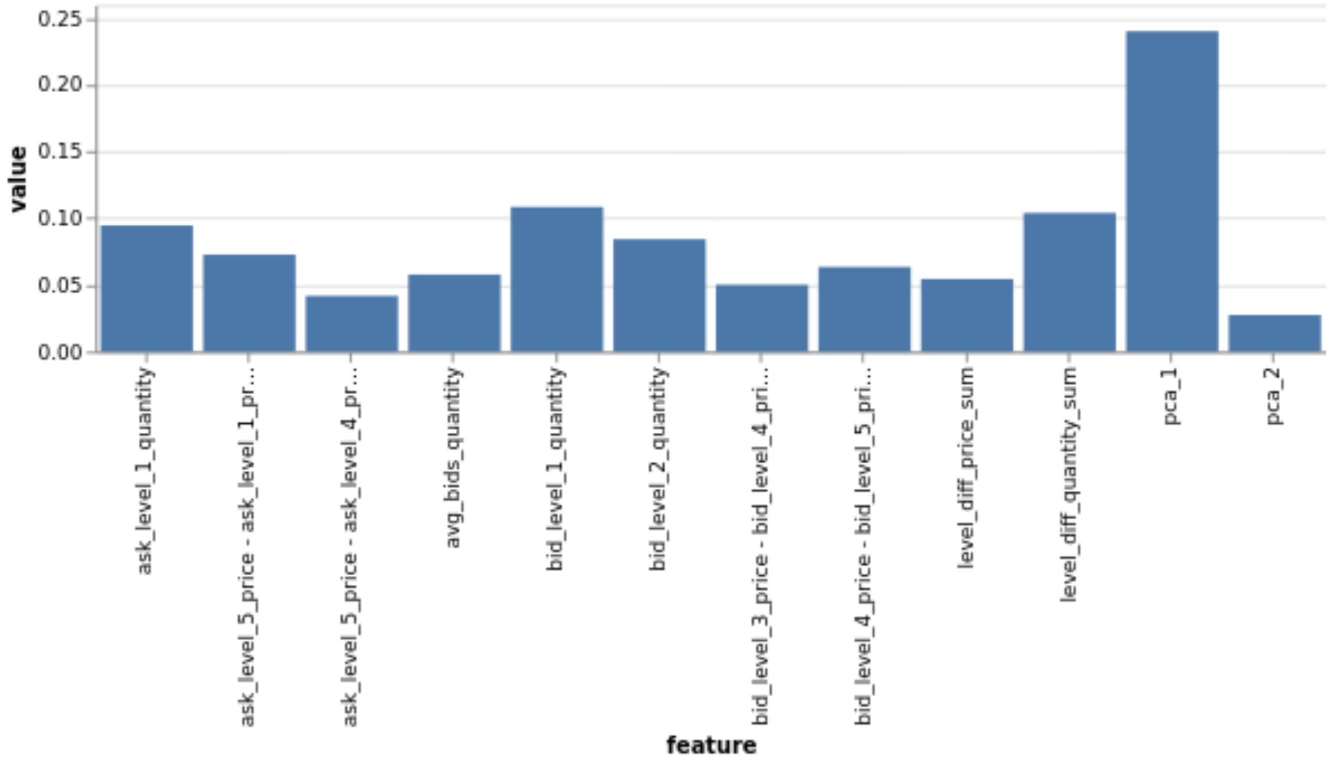
In this section we will take a look into features and aspects related to the model training and evaluation.
In order to evaluate existing and generated features we trained a Gradient Boosting Classifier Tree and extrated information on feature importance. The result is presented below:

Plot 6: Feature importance



We may observe that a number of proposed features indeed evulated as quite important. However, a lot of features seems not to be important for our goals. In order to reduce dimensionality, we will extract key vectors from those features via PCA algorithm. We selected 10 top features for direct usage and PCAed the rest of features transforming them into two dimensions. Then we reevaluated importance of the new features and ended up with the following imoprtance distribution:

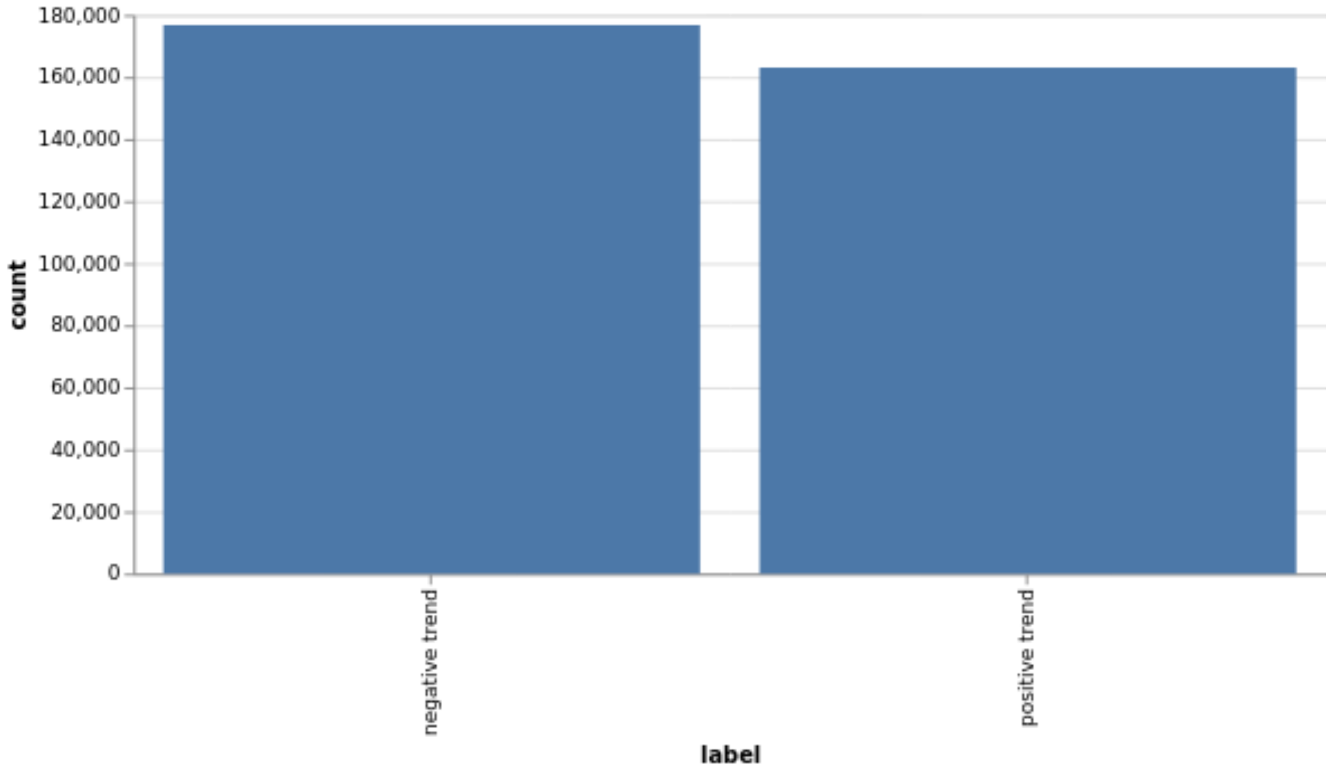
Plot 7: Feature importance after PCA



Thus, we kept the mos important features and combined the rest in the PCA vectors. As we may observe from the plot, the first PCA vector should represent an important feature.

As we transformed our task into the classification problem, it is important to asses label imbalance as it may affect the further training process. As demonstrated on the plot below, the labels are balanced thus there is limited threat of our metrics' validity.

Plot 8: Distribution of labels (test data)



Final model results

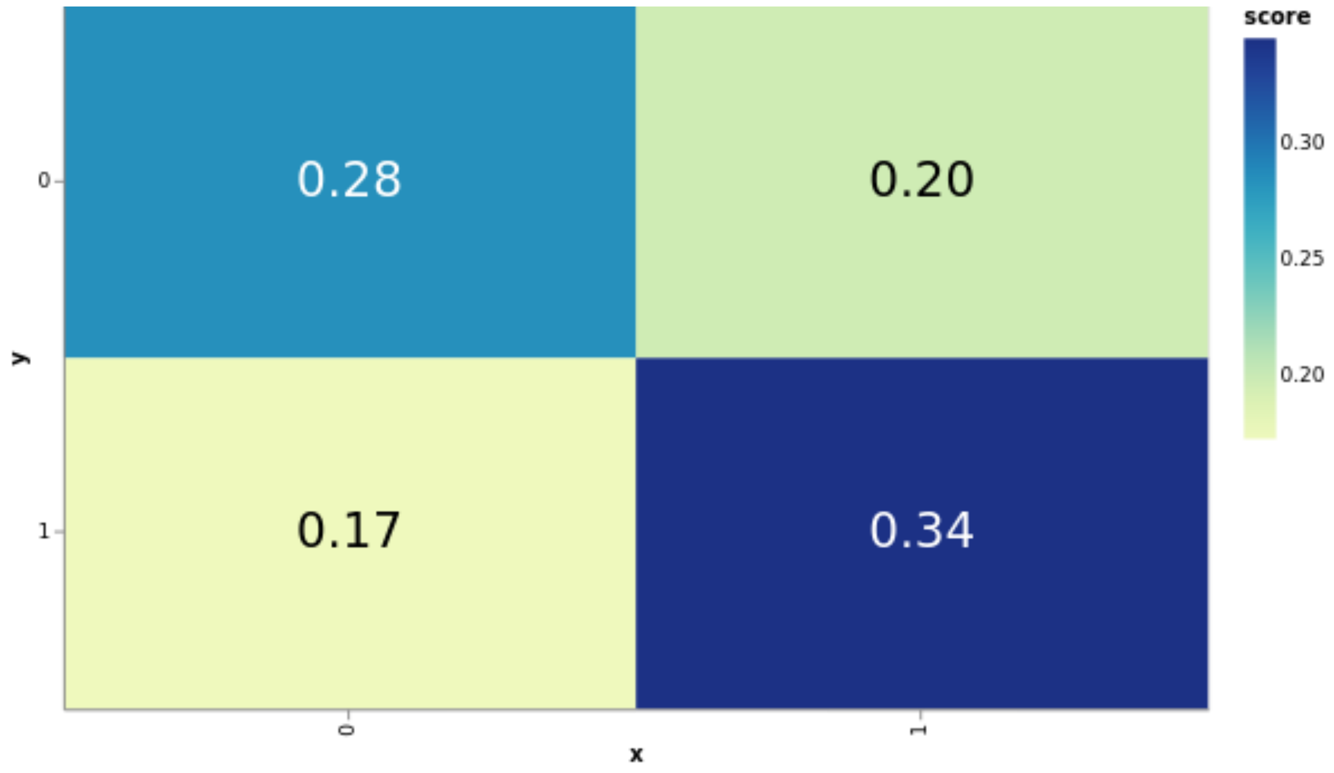
We evaluated performance of the three models: Gradient Bosting Classifier Tree, Supported Vector Machine and Multilayer Perceptron Classification model. According to our analysis, the Supported Vector machine performed better than the other and thus we use this model as the primary one. We used the following model parametrs during the training stage:

Final model key parametrs

	setting	value
0	maxIter	30
1	aggregationDepth	4
2	tol	0.0001

Overall we achieved 64 percent of accracy which exceeds the result of a random coin toss. As we may observe from the confusion matrix below, we achieved stable result without imbalance in recall or precision metrics.

Plot 9: Confusion matrix



☰