Interpretable ML-driven Strategy for Automated Trading Pattern Extraction

A Case Study of CME Futures

Artur Sokolovsky, Jaume Bacardit, Thomas Groß

Newcastle University, School of Computing
Newcastle Upon Tyne, UK

Luca Arnaboldi

University of Edinburgh, School of Informatics
Edinburgh, UK

Abstract

Financial markets are a source of non-stationary multidimensional time series which has been drawing attention for decades. Each financial instrument has its specific changing over time properties, making their analysis a complex task. Improvement of understanding and development of methods for financial time series analysis is essential for successful operation on financial markets. In this study we propose a volume-based data pre-processing method for making financial time series more suitable for machine learning pipelines. We use a statistical approach for assessing the performance of the method. Namely, we formally state the hypotheses, set up associated classification tasks, compute effect sizes with confidence intervals, and run statistical tests to validate the hypotheses. We additionally assess the trading performance of the proposed method on historical data and compare it to a previously published approach. Our analysis shows that the proposed volume-based method allows successful classification of the financial time series patterns, and also leads to better classification performance than a price action-based method, excelling specifically on more liquid financial instruments. Finally, we propose an approach for obtaining feature interactions directly from tree-based models on example of CatBoost estimator, as well as formally assess the relatedness of the proposed approach and SHAP feature interactions with a positive outcome.

Keywords: Futures Trading, Volume Profiles, Tree Boosting, Explainable ML

Email addresses: artur.sokolovsky@gmail.com,{jaume.bacardit,thomas.gross}@ncl.ac.uk (Artur Sokolovsky, Jaume Bacardit, Thomas Groß), luca.arnaboldi@ed.ac.uk (Luca Arnaboldi)

1. Introduction

Market events are key means to evaluate changes in a market [1]. The study of these particular events allows traders to predict potential changes of market dynamics, and permit the study of factors which may impact the profitability of trades. The concept of event here is very broad and include natural disasters, economic news, political speeches, tweets, or even certain financial time series patterns. These event-based techniques permit the human to make and justify their investment or trading decisions. In this study we focus on financial time series patterns, namely volume profiles by making them suitable for machine learning pipelines.

Using two different types of time series pattern extraction, namely, volume-centred range bars (VCRB) and price levels, we showcase how we are able to classify the patterns using explainable machine learning models. An explainable model, is one which is non-inherently understandable by design, but whose results can be explained by means of post-hoc methods. As previous research has extensively discussed price level patterns [2], the core analysis of this paper is the volume-based pattern extraction. We adjust manually traded volume profiles to the ML applications by proposing VCRB. See previous work from Sokolovsky & Arnaboldi [2] for details on the price level work. For any terms related to finance or machine learning that may be unknown, we introduce a glossary in the supplementary material for the readers convenience.

As automated machine learning (ML) algorithms take over several aspects of trading, accountability becomes a huge factor at play [3, 4, 5, 6, 7]. Due to the complexity, and unintuitive nature of most machine learning models, it is unreasonable to expect that users, who may or may not be an expert in the field, to understand how a model works and comes to decisions. Furthermore, even if a user may understand exactly how the model works, following a decisions through a very complex and large set of mathematical steps, is unreasonable and often unfeasible for any real-world application [8]. ML has historically been treated as a black box, data is input in the model, and a prediction is made. However, whilst this kind of approach may very well trade with high levels of profitability, what happens when it does not? The downside of the black box approach is that we are unable to understand why certain decisions are made - consequently when things go wrong, it becomes difficult to ascertain why [7, 9]. What is instead ideal is to have explainable machine learning. Through more explainable AI; with careful choice of algorithms and data curation we are able to ascertain why decisions are made, why things go wrong and consequently adjust our strategies to maintain profitability.

Explainable machine learning is an active research area across many different disciplines [7]; however

we have yet to reach a consensus on how to achieve perfect understanding as several challenges arise [10, 11, 12]. It is generally understood, that focusing on more understandable machine learning algorithms, such as logistic regressions, and with careful feature selection we can greatly improve understanding. Whilst this is a very active area in certain domains, such as medicine [13], comparatively little research has been applied to finance and trading. In this work we showcase a combination of novel state of the art machine learning techniques and statistics methods to create effective automated trading means that can both potentially garner profits whilst still being potentially understandable by a human trader to monitor and adjust as needed.

Generally speaking, one can achieve explainability in AI in three ways [9]: 1) using more understandable algorithms, 2) reverse engineering the estimator to understand how it came to a decision, and/or 3) domain-specific adjustment of the input entries and feature design. Whilst the first approach is the more desirable of the two, as explanability comes inbuilt, there is often a trade-off between using simpler, more understandable models that may be less accurate, and more complex (less understandable) models that may well be highly accurate [14]. The second approach is gaining traction in recent years, focusing on using: i. visualisations, ii. natural language explanations and iii. explanations by example [9]. The third approach leads to optimal input entries and feature space. Applied to financial markets, this means careful selection of the events (time series patterns in our case) as well as use of the most relevant features, leading to less convoluted model explanations. Which approach is most suitable depends on the prediction tasks. In this work we use domain knowledge to design the time series patterns of interest. Additionally, due to the financial implications of decisions, we choose a state-of-art boosting trees algorithm - CatBoost [15]. It is explainable by post-hoc explanations i.e. SHAP [16], efficient and easy to tune.

We specifically make use of several statistical techniques to evaluate the approach, namely: effect sizes and hypothesis testing to measure the effects and assess significance of the results, backtesting [1] to simulate trading performance, as well as SHAP [16] and Monoforest-based approach [17] to allow explanation of the decisions. The backtesting is performed using a bespoke trading strategy and assumptions presented in a previous work [2]. Additionally, we discuss a more conservative approach accounting for a bid-ask spread.

[1] Python Library for testing trading strategies available at: https://www.backtrader.com

1.1. Contributions

The contributions of the current study are two-fold. Firstly, we propose a trading volumes-based method for extracting potentially tradeable patterns from financial time series, which we also call time series subsets. Aimed at comparability across entries, our method is specifically designed to be used in machine learning setting. As a way of assessing its usefulness, we compare the proposed method to a conventional price levels pattern, adapted for automatic extraction from time series as demonstrated by Sokolovsky & Arnaboldi [2]. Both these methods are analysed across two different financial instrument - CME Globex British Pound futures (B6) and S&P E-mini Futures.

Secondly, we propose a way of obtaining any order feature interactions directly from tree-boosting models using CatBoost as an example. We feel that interpretable machine learning is an essential part of modern finance. The proposed way can be used for interpreting the models and potentially mining new trading ideas. We qualitatively compare the obtained results with the SHAP feature interactions.

1.2. Research Questions and Paper Outline

In this study we are answering four research questions. First, we want to evaluate whether the proposed method allows better classification performance than no-information model for the considered feature space. The second question is whether the developed volume-based extraction method is better than a previously proposed approach based on price action. We compare two types of time series pattern extraction methods from the classification performance point of view. Namely, we consider price action-based patters - price levels [2] and volume-based patters, which we propose in the current study. Then, we investigate whether volume-based patterns lead to better pattern classification on liquid markets. The answer suggests which type of market the proposed method works best on. Finally, we investigate if the commonly used SHAP values feature interactions are associated with the ones exacted from the model decision paths and quatify this relatedness.

In order to make use of the extracted patterns, we classify them into scenarios as a way of predicting the future. We consider two the most general price behaviours which can be traded - price crossing the target and price reversing from the target. The target can be arbitrary. In the current study it is either a price extremum or PoC (Point of Control - the price with the largest volume in the volume profile) as an attempt of bringing stationarity in the non-stationary financial time series. More complex scenarios can be considered for non-linearly valued instruments like options. The classification depends a lot on the feature space. Hence, answering the first research question, we investigate whether the chosen market

microstructure-based feature space and model are appropriate for the proposed method and further experiments.

RQ1: Given our proposed volume-based pattern extraction method baseline performance (as always-positive), can we further increase it with a domain-led feature engineering and ML model?

RQ2: Are the proposed volume-based patterns potentially better suitable for trading than price level-based? To answer this question generally, we do not limit ourselves to a particular trading strategy. Instead, we set up a classification task and compare the classification performance of the methods.

For the third research question we hypothesise that volume-based patterns will perform statistically better on a more liquid market (S&P E-mini in our case). The reasoning is that if there are more large players on the market, the price is less price-action driven or "noisy". We use volume-based approach for pattern extraction and expect it to be more suitable for liquid markets.

RQ3: Does classification of volume-based patterns show better results on a liquid market?

RQ4: Are feature interactions discovered with SHAP associated with the ones obtained directly from the decision paths?

This RQ arises from the field of explainable machine learning. SHAP values can be applied to interpret various models [16]. Among other information, they provide a ranking of features based on contribution to the prediction per instance, and feature interactions. Since SHAPs provide an intuitive understanding and are easy to interpret, they are an ideal candidate to evaluate diverse real-world models. Nevertheless, since SHAP values are an approximation of the original model, it would be useful to compare SHAPs to an explicit model interpretation method for completeness. To do so, we propose a way of obtaining any order feature interactions explicitly from a tree-based boosting model by leveraging the Monoforest approach [17].

The remainder of the paper is structured as follows: Section 2 communicates the research protocol and evaluation criteria used for this paper; Section 3 presents the results of our evaluation; Section 4 discusses the results and the extent to which our analysis is successful as well as future work; finally, Section 5 concludes the paper. Further analysis and extra details are provided in the supplementary materials after the bibliography.

2. Materials and Methods

In this section we provide details of all the experiments performed, datasets used, model training and evaluation methodology. Most of the discussion around the market structures will focus on the currently proposed volume bars, details of price levels approach are discussed in the previous work [2].

2.1. Datasets

In the current study we use a sample of data from S&P E-mini (ES) and British Pound (B6) futures instruments, traded on CME Globex. Namely, we consider a time range of 39 months - from March 2017 until June 2020. When discussing the results, the data outside of this time range is considered as statistical population.

Since order book data is less commonly available and requires extra assumptions for pre-processing, we use tick data with only Time&Sales per-tick statistics. Concretely: we obtain numbers of trades and volumes performed by aggressive sellers and buyers, at bid and ask, respectively. Additionally, we collect millisecond-resolution time stamps on tick starts and ends as well as prices of the ticks. In order to make the instruments comparable, we use price-adjusted volume-based rollover contracts data. Detailed data pre-processing is explained later on.

2.2. Volume-centred range bars (VCRB)

Volume profile is a representation of trading activity over a specified time and price range, its variations are also called market profile. They are commonly used for market characterisation (when considering daily volume profiles) and trading. Volume profiles are usually built from the temporal or price range bars, once every n bars. These settings make the obtained profiles highly non-stationary and applying ML to them has the same drawbacks as feeding raw financial time series into a model. Namely, in such a setting models are hard to fit and require large datasets due to constantly changing properties of the data.

In the proposed method we aim to increase the stationarity of the volume profiles, as well as obtain as many entries per time interval as possible. We illustrate a high-level diagram of the VCRB extraction pipeline in Figure 1. The core of the proposed method is a set of tick buffers; a new buffer is started at a given price if there is currently no active buffer that has been initialised at this price - ensuring that each price is covered. Ticks are fed into all active buffers simultaneously. When the desired price range is reached for a specific buffer, it becomes complete. The price range is measured as a minimum price

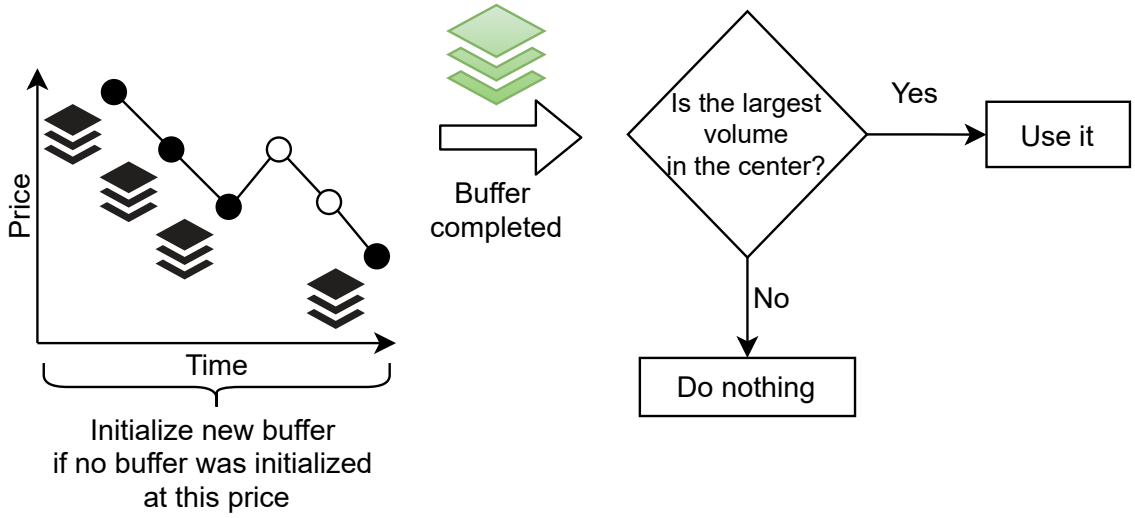

Figure 1: Visualisation of the volume-based pattern extraction approach. Entries filled with black indicate initializations of new buffers.

subtracted from the maximum price in the tick buffer. The price range is referred from now on as *range configuration*.

After the buffer is complete, we build its volume profile and check if the largest volume is in its centre. If so, we proceed with computing its features and labelling it. Otherwise, we ignore it. As we show later in Results, the tick buffer fulfilling the condition can be visualised as a volume-centred range bar (VCRB). The largest volume in the volume profile of the buffer is called Point of Control (PoC). We use the PoC as a target for labelling where price reverses from or crosses it.

2.3. Experiment Design

In the current subsection we describe all the components of the experiment design. We ensure that the design is as uniform as feasible across the experiments. When highlighting the differences, we refer to the experiments by the associated research questions - from RQ1 to RQ4. To support the readers, we summarise the experiment design for all the experiments in Table 1.

2.3.1. Label Design and Classification Setting

For the sake of comparability, in the current study we reuse labelling procedure from the study by Sokolovsky & Arnaboldi [2]. The labelling process starts when the VCRB is formed. After the price reaches

Table 1: Experiment design across the experiments. Ft_{sel} & M_{tune} column accounts for feature selection and model tuning. "Comparison H_0/H_1" column indicates the settings used to obtain null and alternative hypotheses data. "Other" indicates whether the test is conducted on both instruments (ES, B6) and whether only VCRB pattern extraction method was used. All hypotheses are evaluated using Wilcoxon test.

Experiment	Ft_{sel} & M_{tune}	Metric	Comparison H_0/H_1	Other
RQ1	✓	Precision	No-info. / CatBoost	VCRB; ES, B6
RQ2	✓	PR-AUC	Price levels / VCRB	ES, B6
RQ3	✓	PR-AUC	B6 / ES	VCRB
RQ4	-	Footrule distance	Bootstrap / SHAP, Monoforest-based	ES, B6

the target (POC or extremum of the pattern for VCRB and price level, respectively), there are two scenarios: it reverses or continues its movement. For the reversal we require the price to move for at least 15 ticks from the target. For the crossings we take the 3 ticks beyond the target. If the price crosses the target for less than 3 ticks and then reverses, we exclude this entry from training and label as negative for testing.

Since price reversals require 15 ticks price movement, they are directly applicable for trading and are considered positives in our binary classification tasks. Target price crossings are labelled based on 3 ticks price movements, hence their direct use in trading is limited and we consider them negatives. Consequently, price reversals are labelled as positives and crossings - as negatives.

We perform binary classification of the extracted VCRB and price levels patterns. Namely: we classify whether the entries are followed by price rebounds (reversals) from the target or target crossings. In the study we use one of the novel tree boosting algorithms - CatBoost [15], which is claimed to require little-to-none parameter tuning and dealing well with large feature spaces.

2.3.2. Feature Space and Model Parameter Tuning

We conduct two types of experiments - with and without feature selection and model parameter tuning. We have done so to study different aspects of the matter:

- Experiments with feature selection and model tuning. Since the models are free to choose the feature subset and model configurations, these experiments are not limited by a particular feature set, but rather by an overall feature space. We use this setting for RQ1-3.

- Experiments with a fixed feature space and model parameters. These are aimed at studying feature interactions across datasets. Fixed feature space and model parameters ensure that any differences in the feature interactions are caused by the data or the model analysis approach and not by varying model configuration or feature space. These are used specifically in RQ4.

When designing the feature space we follow the same approach as suggested in [2] - we extract a set of features from the volume-based pattern or a price level - called Pattern features, and the second one from the most recent ticks before the target is approached (being 2 ticks away from it) - called Market Shift (MS) features. In order to decrease the number of uncontrolled factors and their impact on the experiment design, we limit the feature space to only market microstructure-based features. We list the features and the associated equations in Table 2. The provided equations are valid for volume-based patterns. Since the data is available only below or above the extrema in case of the price level patterns, the features are computed slightly differently. Negative t values are considered as odd number of ticks distances from extrema and positive - as even numbers. We believe that this alteration is the closest possible to the original while preserving the domain knowledge at the same time.

To perform the feature selection, we run a Recursive Feature Elimination with Cross-Validation (RFECV) process with a step of 1 and using the internal CatBoost feature importance measure for the feature ranking. Model parameter tuning is done for the following model parameters, which are recommended for optimisation: i) number of iterations; ii) maximum tree depth; iii) whether temporal component of the data is considered; and iv) L2-regularisation. We considered it infeasible to optimise over a larger number of parameters, as with the current setting the experiments take in total >7k (single core) CPU hours. When optimising the feature space and tuning the model parameters, precision metric is used - the one directly related to the success of the model in a trading setting, as we explain later.

2.3.3. Performance Metrics

For the RQ1 experiments, the performance is evaluated using precision metric as it directly defines the potential profitability of the model. Namely, as fraction of true positives over a sum of true positives and false positives, which defines the fraction of times the participant enters the market with a positive outcome. We do not account for negatives has they have no impact on the outcome, due to our approach not trading these. We elaborate more on this in the Discussion section.

We use binary Precision-Recall Area Under Curve (PR-AUC) score as the metric for RQ2 and RQ3 as there we are comparing classification performance of differently imbalanced datasets and this measure

Table 2: Features (referred to by code '[code]') used in the study in two stages, Stage 1 - Pattern and Stage 2 - Market Shift (MS)

	Equation	Description
Pattern features	$\frac{\sum_{t\in[1;5]}^{p=X+t}(V_b)}{\sum_{t\in[-5;-1]}^{p=X+t}(V_b)}$	Sum of upper (above PoC) bid volumes divided by lower ones [P0]
	$\frac{\sum_{t\in[1;5]}^{p=X+t}(V_a)}{\sum_{t\in[-5;-1]}^{p=X+t}(V_a)}$	Sum of upper ask volumes divided by lower ones [P1]
	$\frac{\sum_{t\in[1;5]}^{p=X+t}(T_a)}{\sum_{t\in[-5;-1]}^{p=X+t}(T_a)}$	Number of upper bid trades divided by lower ones [P2]
	$\frac{\sum_{t\in[1;5]}^{p=X+t}(T_b)}{\sum_{t\in[-5;-1]}^{p=X+t}(T_b)}$	Number of upper ask trades divided by lower ones [P3]
	$\frac{\sum_{t\in[1;5]}^{p=X+t}(V_b)}{\sum_{t\in[1;5]}^{p=X+t}1}$	Average upper bid trade size [P4]
	$\frac{\sum_{t\in[1;5]}^{p=X+t}(V_a)}{\sum_{t\in[1;5]}^{p=X+t}1}$	Average upper ask trade size [P5]
	$\frac{\sum_{t\in[-5;-1]}^{p=X+t}(V_b)}{\sum_{t\in[-5;-1]}^{p=X+t}1}$	Average lower bid trade size [P6]
	$\frac{\sum_{t\in[-5;-1]}^{p=X+t}(V_a)}{\sum_{t\in[-5;-1]}^{p=X+t}1}$,	Average lower ask trade size [P7]
	$\frac{\sum^{p=X}V_b}{\sum_{t\in[1;5]}^{p=X+t}V_b}$	Sum of PoC bid volumes divided by sum of upper bid volumes [P8]
	$\frac{\sum^{p=X}V_a}{\sum_{t\in[1;5]}^{p=X+t}V_a}$	Sum of PoC ask volumes divided by sum of upper ask volumes [P9]
	$\frac{\sum^{p=X}V_b}{\sum_{t\in[-5;-1]}^{p=X+t}V_b}$	Sum of PoC bid volumes divided by sum of lower bid volumes [P10]
	$\frac{\sum^{p=X}V_a}{\sum_{t\in[-5;-1]}^{p=X+t}V_a}$	Sum of PoC ask volumes divided by sum of lower ask volumes [P11]
	$\frac{\sum_t^{p=X+t}V_b}{\sum_t^{p=X+t}V_a}$	Sum of bid volumes divided by sum of ask volumes as price X+t; $t\in[-1;1]$ as different features [P12]
	$\frac{\sum_t^{p=X+t}T_b}{\sum_t^{p=X+t}T_a}$	Number of bid trades divided by number of ask trades as price X+t; $t\in[-1;1]$ as different features [P13]
	-	Side - below or above the price when the pattern is formed [P14]
MS features	$\frac{\sum_{t,b}^{w=237}(V_b)}{\sum_{t,a}^{w=237}(V_a)}$	Fraction of bid over ask volume for last 237 ticks [MS0]
	$\frac{\sum_{t,b}^{w=237}(T_b)}{\sum_{t,a}^{w=237}(T_a)}$	Fraction of bid over ask trades for last 237 ticks [MS1]
	$\frac{\sum_t^{w=237}V_b}{\sum_t^{w=237}V_a}-\frac{\sum_t^{w=21}V_b}{\sum_t^{w=21}V_a}$	Fraction of bid/ask volumes for long minus short periods [MS2]
	$\frac{\sum_t^{w=237}T_b}{\sum_t^{w=237}T_a}-\frac{\sum_t^{w=21}T_b}{\sum_t^{w=21}T_a}$	Fraction of bid/ask trades for long minus short periods [MS3]
Key	t - number of ticks; a - ask; p - price; w - tick window; P_N - neighbours until distance N	
	V - volume; b - bid; T - trades; X - PoC or extremum	

is advised for use in this setting [18]. For the sake of completeness and easier interpretation of the results, we provide the mentioned metrics together with ROC-AUC and f1-score for all the experiments. As there is no notion of performance in the experiments associated with RQ4, we describe its metric in Section: Relatedness of feature interactions from SHAP and decision paths (RQ4).

2.3.4. Cross-validation

Price-adjusted volume-based rollover pre-processing of the data allows us to split the data into 3-month chunks without any extra care of contracts rollover. We do so since length of contracts is different for ES and B6 and contracts cannot be compared directly. When training the models, we apply a temporal sliding window approach, using batch N for training, $N+1$ for testing, for $N \in [1, B-1]$, where B is the total number of 3-month batches available. For the feature selection and model parameter tuning, we use 3-fold time series cross-validation within the training batch with the final re-training on the whole batch.

2.3.5. Effect Sizes

Before evaluating the hypotheses, we compute Hedge's g_{av} effect sizes for paired data together with the .95 confidence intervals (CIs) in order to generalise the results to the population (unseen data) as well as make them easier to interpret and compare across studies [19]. We consider the effect sizes significant if the CI does not overlap with the 0.0 effect size threshold. We expect the significant to be also present in the unseen data.

2.3.6. Statistical Tests

In order to formally validate the hypotheses, we have to choose a suitable paired difference statistical test. Since the number of data batched is small, we take a conservative approach and require the selected test to be applicable to non-normally distributed data. We choose Wilcoxon signed-rank test as the best suitable candidate [20] and apply its single-sided version to validate the hypotheses of the study. We refer to null and alternative hypotheses as H_{0X} and H_X, respectively, with X being the number of the research question.

Throughout the study, we set the significance level for the statistical tests of the study to $\alpha = .05$. Finally, we apply Bonferroni corrections to all the statistical tests within each experiment family [21]. The experiment families are defined based on the data and objectives - each research question forms a separate experiment family.

2.3.7. Model Interpretation

The challenge of interpreting the modern ML models roots in their complexity. Even considering decision tree-based models, interpretation becomes problematic when the number of features grows. When it comes to interpretations, number of decision paths of CatBoost model is usually well beyond the limit of manual analysis. Fortunately, Monoforest [17] approach makes the decision paths uniform and its implementation gives access to the machine-readable decision paths. For each decision path we retrieve information about the subset of the involved features, their thresholds, as well as support (w) and contribution (c) to the output of the model.

We assume that interactions between features take place if they are found in the same decision path. Please note that the way SHAP defines feature interactions is principally different, as we elaborate in Discussion section. In order to get interactions for the whole model, we average across the decision paths in the following way:

$$I_{(F1,F2)} = \frac{\sum_{i=0}^{N} c \times w}{N}, \tag{1}$$

where N is the total number of decision paths containing features 1 and 2 ($F1$, $F2$), c is the contribution of the decision path to the model output and w is the support, computed as number of times the path was activated in the training set divided by the total number of training entries.

If there is a single feature in the decision path, we consider it as a main-effect. We include main effects into the interaction matrix the same fashion as it is done for SHAP [16] - as diagonal elements of the matrix. When there are more than two features in the decision path, we treat them as multiple pairwise interactions to be able to represent them in a single 2d matrix and compare to SHAP interactions. By considering decision paths with a fixed number of features, one can get interactions of a particular order. Moreover, these interactions are directly comparable across orders, models and datasets.

2.3.8. Backtesting

We perform backtesting of the proposed method in Python Backtrader platform. We use the same strategy, assumptions and trading fees as used in the price levels study [2]. In the study we focus on performance comparison based on classification tasks over the backtesting simulations for two reasons: i) different trading intensities lead to varying impacts of the modelling assumptions; ii) limiting our comparison to a particular strategy significantly decreases the generality of the finding. Elaborating on the first point: influences of order queues, slippages and bid-ask spreads are partially taken into account, however it is obvious that these effects have increasing impacts with an increase of the trading intensity.

Quantification of these is out of the scope of this study, however, might be very useful. We describe how the modelling assumptions might affect the backtesting results in the Discussion section. As a complementary analysis, in the current study we report annual rolling Sharpe ratios for all the range configurations, and for the price level method.

2.4. Experiments

In the current subsection we list the conducted experiments associated with the research questions, as well as highlight any experiments-specific choices.

2.4.1. VCRB method, CatBoost versus no-information estimator (RQ1)

Firstly we assess the performance of the no-information and CatBoost classifiers, and compare them. Prior to evaluating the hypotheses, we compute the effect sizes. Then, we run the statistical test with the following hypotheses:

H_{01}: CatBoost estimator performs performs equally or worse than the no-information model.

H_1: CatBoost estimator performs performs better than the no-information model.

In the results section we report statistical test outcomes for the configuration with the largest effect sizes, the other configurations are reported in the supplementary materials.

2.4.2. VCRB vs price levels approach (RQ2)

To answer the second research question we need to compare the performance of the two methods - price levels and VCRB. We obtain the PR-AUC classification performance for both methods and both instruments. Then, we compute the effect sizes, and, finally, we run the statistical test with the following hypotheses:

H_{02}: Price level patterns are classified with performance equally good or better than volume-based patterns.

H_2: Volume-based patterns are classified with statistically better performance.

2.4.3. VCRB method, ES versus B6 datasets (RQ3)

Answering the third research question, we compare CatBoost classification performance on the VCRB-extracted data over the two datasets - B6 and ES, where the latter is far more liquid. After the PR-AUC classification performance is obtained, we compute the effect sizes and run the statistical tests with the following hypotheses:

H_{03}: Both instruments perform comparably or performance on B6 is statistically better.

H_3: Volume-based patterns are classified with statistically better performance for the instrument with higher liquidity (ES).

2.4.4. Relatedness of feature interactions from SHAP and decision paths (RQ4)

To answer the last research question we need to get the data representing the null hypothesis - having no relation (potentially in contrast to SHAP and decision paths), and propose a method for assessing the relatedness. As the first step, we obtain feature interactions in a form of a square matrix using SHAP and the Monoforest-based method. To perform the comparison in a statistical fashion, we need data representing the null hypothesis. To generate it, we bootstrap (randomly sample with replacement) the interaction matrices' elements separately for both approaches. We choose to generate 500 bootstrapped entries for each method. As a result, we get two sets of matrices with the same value distributions as the original feature interaction matrices (SHAP and Monoforest-based). Since we performed bootstrapping, by definition there is no association between any two matrices from the two sets.

Since both methods output interactions scaled differently, we need to make the matrices comparable. To do so, we rank the interaction strengths within each matrix. After that, we compare the ranks across the two methods by computing their Footrule distances [22]. This measure is designed specifically for ranks data. Later the distances are used as a proxy to assess the relatedness of the methods - the larger the distance, the weaker the relationship.

In order to get a reliable null hypothesis distance, we compute mean of the distances between the bootstrapped matrices. At this point the relatedness of SHAP and decision paths methods can be compared against the bootstrapped data. To quantify the differences, we obtain the effect sizes using the Footrule distance (instead of the classification performance using in the other RQs). Finally, we run the statistical test with the following hypotheses:

H_{04}: There is no difference between Footrule distances on ranks of SHAP-decision paths and bootstrapped feature interactions matrices or SHAP-decision paths are larger.

H_4: There is a difference between Footrule distances on ranks of SHAP-decision paths and bootstrapped feature interactions matrices with SHAP-decision paths being smaller than bootstrapped.

3. Results

This section presents all the results from our methodology described in the previous section, We reserve detailed analysis of how the results match our hypothesis for Sec. 4, and this section only contains outcomes of the experiments.

3.1. Pattern extraction from the datasets

In this subsection we report statistics on the original datasets as well as numbers of entries obtained from every dataset batch, for both extraction methods and financial instruments.

In Table 3 we show numbers of ticks and total volumes per-batch for both instruments.

Batch	**ES**		**B6**	
	Volume	Ticks	Volume	Ticks
3/17 to 6/17	86123932	151965	6663094	118098
6/17 to 9/17	82384394	132964	6575559	115576
9/17 to 12/17	73925568	98600	7971963	142548
12/17 to 3/18	88050918	517451	7588515	159565
3/18 to 6/18	96653879	536280	7267680	131215
6/18 to 9/18	71968775	235841	6956995	116275
9/18 to 12/18	109410969	581345	7128825	155895
12/18 to 3/19	95948559	673838	6078533	136554
3/19 to 6/19	92201997	378788	6643282	132441
6/19 to 9/19	93229922	458141	5677468	94248
9/19 to 12/19	75613694	343666	7193235	153081
12/19 to 3/20	101547199	587658	6122643	100381
3/20 to 6/20	126756329	3482845	5593815	270096

Table 3: Original datasets statistics. Volume columns correspond for the total volume traded per the stated time interval. The Ticks columns show the numbers of ticks per the time interval.

We see in Table 4 that in year 2017 there are more entries for VCRB B6 than for ES, later the situation reverses. Overall, there are around 5-10 times less entries for price level-based method in comparison to the volume-based. We address potential consequences of these differences in the Discussion section.

Batch	**VCRB**		**PL**	
	ES	B6	ES	B6
3/17 to 6/17	2224	2764	447	278
6/17 to 9/17	1950	2781	360	267
9/17 to 12/17	1405	3432	268	378
12/17 to 3/18	10398	3992	1643	418
3/18 to 6/18	10199	3121	1695	335
6/18 to 9/18	4142	2633	725	287
9/18 to 12/18	11809	3750	1926	370
12/18 to 3/19	13894	3330	2234	348
3/19 to 6/19	6908	3463	1186	358
6/19 to 9/19	9078	2290	1413	235
9/19 to 12/19	6156	3844	1060	410
12/19 to 3/20	12829	2483	1856	227
3/20 to 6/20	87987	8340	12085	738

Table 4: Numbers of extracted patterns for volume-based (VCRB) range 7, and price level-based (PL) methods, reported for the analysed data sets, both instruments.

3.2. Volume-centred range bars

We generate VCRBs for range sizes of 5,7,9,11. For the comparison purposes we extract price-based patterns using a configuration suggested in [2]. We visualise the VCRBs in Fig 2. As we stated in the methods, the volume in the centre is the largest one. The volume distributions differ a lot for the provided entries. If there is a price with zero volume (second volume profile in Figure 2), we skip that price in the visualisation.

3.3. Prediction of the reversals and crossings

The initial experimental stages are feature selection and model parameter tuning. We do not report the optimised feature spaces and model parameters in the manuscript, however we make this data available in the shared reproducibility package [23].

For the classification task we report all the stated performance metrics for the configuration range 7, which is chosen based on RQ1 effect sizes, together with PR-AUC metric for the price levels method

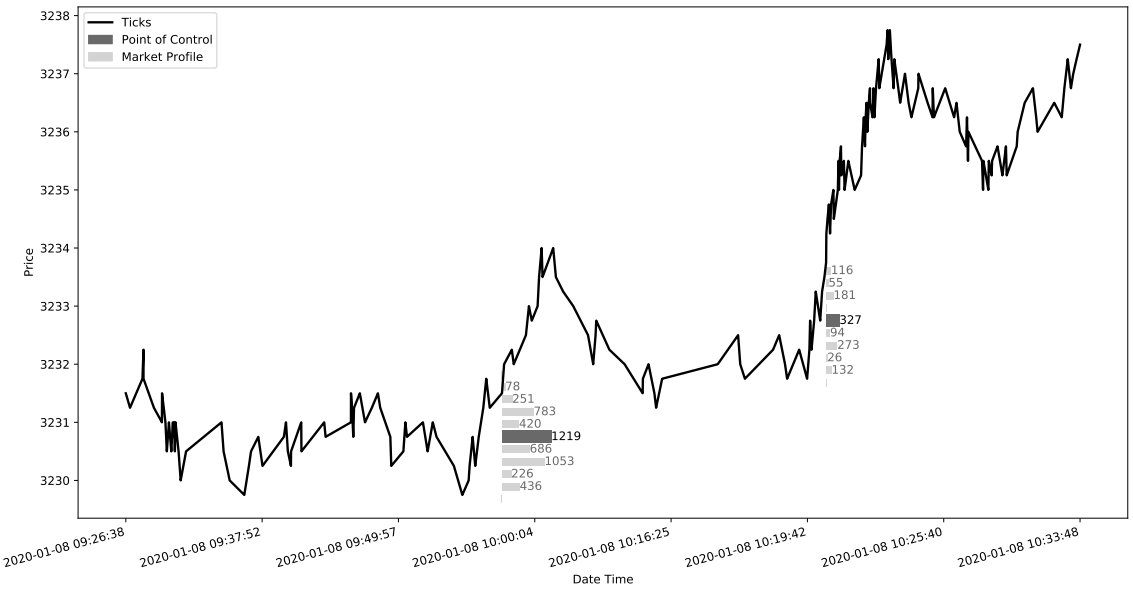

Figure 2: Example of volume-centred range bars generated for ES instrument. Histograms indicate traded volumes within the buffer. Points of control are in the centre of the volume profiles, marked with dark grey. The profiles are formed when the price buffer is complete (9 ticks in this case). Zero-volume entries are not shown.

in Tables 5 and 6. The rest of the metrics for price levels are reported in the supplementary materials, in Tables 12 & 13. Additionally, in the supplementary materials we plot the data representing null and alternative hypotheses throughout the study - in Figures 9, 10, 11 and 12.

In all the experiment families we report effect sizes for all the VCRB configurations (ranges 5, 7, 9 and 11). The statistical test results are reported only for the largest effect size configuration from RQ1 experiment family, on S&P E-mini (ES) instrument. We report the rest of the statistical tests in the supplementary materials. When evaluating the statistical tests, we correct for multiple comparisons - the corrected significance levels are provided separately for each experiment group.

3.3.1. VCRB method, CatBoost versus No-information estimator (RQ1)

Here we provide outcomes of classification performance comparison between the CatBoost estimator and the no-information estimator. First, we plot the effect sizes with .95 confidence intervals in Fig 3. Then, we provide performance statistics and the statistical test results in Table 7. Additionally, to allow easier comprehension of the results, we visualise precision of the models from Tables 5 and 6 in supplementary materials, Fig 9.

Batch	**VCRB**					**Price levels**
	PR-AUC	ROC-AUC	F1-score	Precision	Null_precision	PR-AUC
3/17 to 6/17	0.46	0.57	0.49	0.46	0.41	0.25
6/17 to 9/17	0.45	0.56	0.45	0.43	0.39	0.34
9/17 to 12/17	0.44	0.55	0.37	0.45	0.40	0.26
12/17 to 3/18	0.44	0.57	0.43	0.45	0.38	0.27
3/18 to 6/18	0.45	0.57	0.44	0.46	0.40	0.28
6/18 to 9/18	0.45	0.57	0.47	0.45	0.39	0.26
9/18 to 12/18	0.45	0.57	0.43	0.45	0.39	0.24
12/18 to 3/19	0.45	0.58	0.51	0.43	0.38	0.30
3/19 to 6/19	0.44	0.55	0.44	0.43	0.39	0.29
6/19 to 9/19	0.48	0.59	0.52	0.47	0.40	0.28
9/19 to 12/19	0.42	0.55	0.35	0.43	0.38	0.22
12/19 to 3/20	0.40	0.54	0.35	0.41	0.37	0.25
3/20 to 6/20	0.44	0.58	0.48	0.43	0.38	0.28

Table 5: Performance metrics for ES, volume-based pattern extraction configuration range 7. The dates are reported in the form YY/MM.

From the Figure 3 we see that effect sizes for ES are generally larger. The effect size pattern across configurations varies between the two instruments, one might even say that it is flipped. For ES the maximum effect size is observed at range 7 and minimum at range 11, while for B6 the maximum is at range 5 and the minimum at range 9. Finally, judging about the significance of the effect sizes by the confidence intervals (CIs) crossing the significance threshold line, we see that there is only one configuration with a statistically significant effect size at range 5 for B6 (and marginally significant range 7), while range 11 is the only insignificant configuration for ES.

Performance statistics in Table 7 indicates no skew in the data between CatBoost and no-information models, at the same time there is up to 2 times difference in the variance of the performance. The results for other VCRB configurations are provided in the supplementary materials, Table 14. Since 8 statistical tests were conducted (4 configurations $\times$ 2 instruments), we apply Bonferroni corrections to the significance level threshold for null hypothesis rejection. The corrected significance level is $\alpha = .05/8 = .00625$. For the rest of the experiment families, the statistical tests are reported for the range 7 configuration, as

	VCRB					Price levels
	PR-AUC	ROC-AUC	F1-score	Precision	Null_precision	PR-AUC
3/17 to 6/17	0.38	0.50	0.37	0.38	0.37	0.26
6/17 to 9/17	0.36	0.51	0.34	0.38	0.36	0.26
9/17 to 12/17	0.37	0.51	0.37	0.37	0.36	0.27
12/17 to 3/18	0.38	0.52	0.34	0.39	0.36	0.24
3/18 to 6/18	0.37	0.51	0.42	0.36	0.36	0.21
6/18 to 9/18	0.40	0.54	0.34	0.41	0.37	0.25
9/18 to 12/18	0.35	0.48	0.37	0.35	0.36	0.25
12/18 to 3/19	0.39	0.54	0.28	0.40	0.36	0.26
3/19 to 6/19	0.33	0.50	0.31	0.34	0.34	0.36
6/19 to 9/19	0.36	0.52	0.26	0.37	0.36	0.31
9/19 to 12/19	0.39	0.54	0.37	0.39	0.36	0.23
12/19 to 3/20	0.35	0.51	0.37	0.35	0.34	0.24
3/20 to 6/20	0.36	0.53	0.27	0.37	0.33	0.28

Table 6: Performance metrics for B6, volume-based pattern extraction configuration range 7. The dates are reported in the form YY/MM.

Statistics	**Dataset**			
	ES		**B6**	
One-tailed Wilcoxon test p-value	< .001		0.0017	
Test Statistics	91.0		85.0	
	CatBoost	No-information	CatBoost	No-information
Mean (precision)	0.44	0.39	0.37	0.36
Median (precision)	0.45	0.39	0.37	0.36
Standard Deviation (precision)	0.018	0.01	0.02	0.011

Table 7: Statistics supporting the outcomes of the Wilcoxon test. The test is aimed to validate whether on the VCRB data and the considered feature space, CatBoost performs significantly better than the no-information estimator. The provided result is for the range 7 configuration.

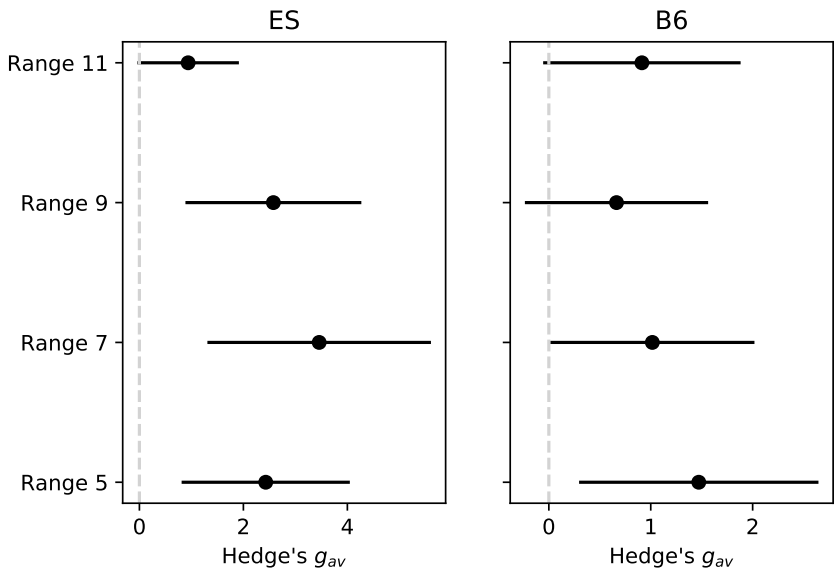

Figure 3: Hedge's g_{av} effect sizes quantifying the improvement of the precision from using the CatBoost over the no-information estimator. The error bars illustrate the .95 confidence intervals, corrected for multiple comparisons. The dashed line corresponds to the significance threshold.

it demonstrated the largest effect size for RQ1 experiments.

3.3.2. VCRB versus price levels approach (RQ2)

Here the results of the investigation on whether volume-based centred bars lead to better classification performance than the price level approach are presented. We report Hedge's g_{av} effect sizes on paired data with .95 confidence intervals in Fig 4, and the outcomes of the statistical test with the supporting statistics in Table 8. Complementing the results, we visualise PR-AUC values from Tables 5 and 6 in supplementary materials, Fig 10).

One can see that all the effect sizes in Fig 4 are significant as the confidence intervals do not overlap with the significance threshold line. Generally, confidence intervals for ES are larger in comparison to B6. The largest effect size is observed for range 5 configuration across instruments.

Performance statistics in Table 8 shows no skew in the data, however data variances differ up to 2 times between the two methods with the difference is preserved across the instruments. The rest of the VCRB configurations is reported in Table 15. In the current experiment family we run 8 tests in total, hence the corrected significance level is $\alpha = .05/8 = .00625$.

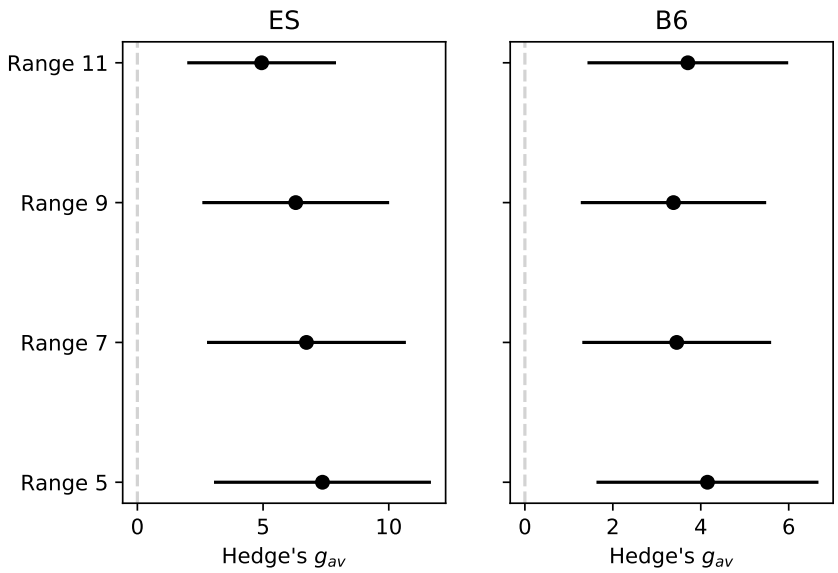

Figure 4: Hedge's g_{av} effect sizes, quantifying the supremacy of the VCRB over the price levels approaches on the basis of the PR-AUC metric. The error bars illustrate the .95 confidence intervals, corrected for multiple comparisons. The dashed line accounts for the significance threshold.

Statistics	**Dataset**			
	ES		**B6**	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	91.0		90.0	
	VCRB	Price levels	VCRB	Price levels
Mean (PR-AUC)	0.44	0.27	0.37	0.26
Median (PR-AUC)	0.45	0.27	0.37	0.26
Standard Deviation (PR-AUC)	0.018	0.029	0.018	0.036

Table 8: Statistics supporting the outcomes of the Wilcoxon test which validates whether Volume-based pattern extraction method leads to better classification performance than the price level pattern extraction. The result is reported for the range 7 configuration.

3.3.3. VCRB method, ES versus B6 datasets (RQ3)

Here we detail results of comparing classification performance of the VCRB entries extracted from ES and B6 datasets. We report effect sizes with .95 confidence intervals in Fig 5, and the statistical test

with the supporting statistics from range 7 configuration in Table 9. The test outcomes for the rest of the configurations are provided in supplementary materials, in Table 16. Additionally, we visualise PR-AUC performance from Tables 5 and 6 in supplementary materials, Fig 11.

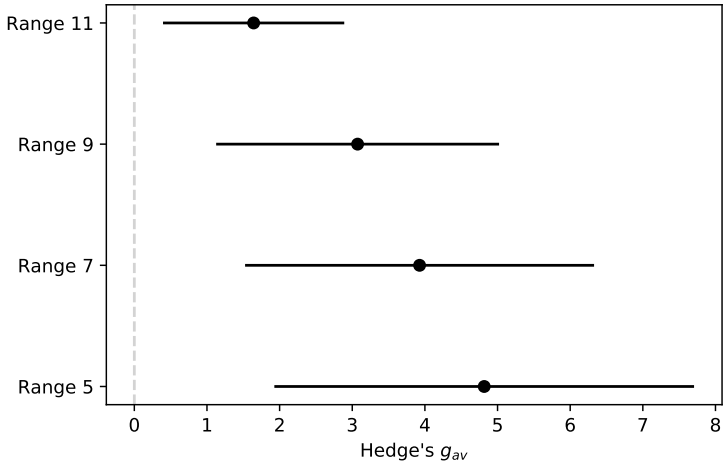

Figure 5: Hedge's g_{av} effect sizes for Volume-based method PR-AUC performance improvement on ES over B6 datasets. Error bars illustrate the .95 confidence intervals corrected for multiple comparisons.

Statistics	**Datasets**	
One-tailed Paired Wilcoxon test p-value	< .001	
Test Statistics	91.0	
	ES	B6
Mean (PR-AUC)	0.44	0.37
Median (PR-AUC)	0.45	0.37
Standard Deviation (PR-AUC)	0.018	0.018

Table 9: Statistics supporting the outcomes of the Wilcoxon test which assesses whether VCRB pattern extraction method leads to better classification performance on the more liquid market (ES in comparison to B6). The result is reported for the range 7 configuration.

In Figure 5 increasing range of the VCRB leads to a smaller effect size. At the same time, the confidence intervals shrink with the range increase, indicating that the performance difference for the larger ranges is smaller but more stable.

The statistics on the results in Table 9 shows that variances are the same for both cases and there is no skew in the distribution. In the current experiment family we run 4 tests in total, hence the corrected significance level is $\alpha = .05/4 = .0125$.

3.4. Backtesting

In Figures 6,7 we show annual rolling Sharpe ratios with 5% risk-free rate and cumulative profits in ticks for all the configurations of the VCRB and price level methods. For easier interpretation of the figures, we plot Sharpe ratios averaged over 90-day periods. The lag of Sharpe ratio plots with respect to the cumulative profits is caused by the requirement of the Sharpe ratio metric to have a year of data available for obtaining the initial value.

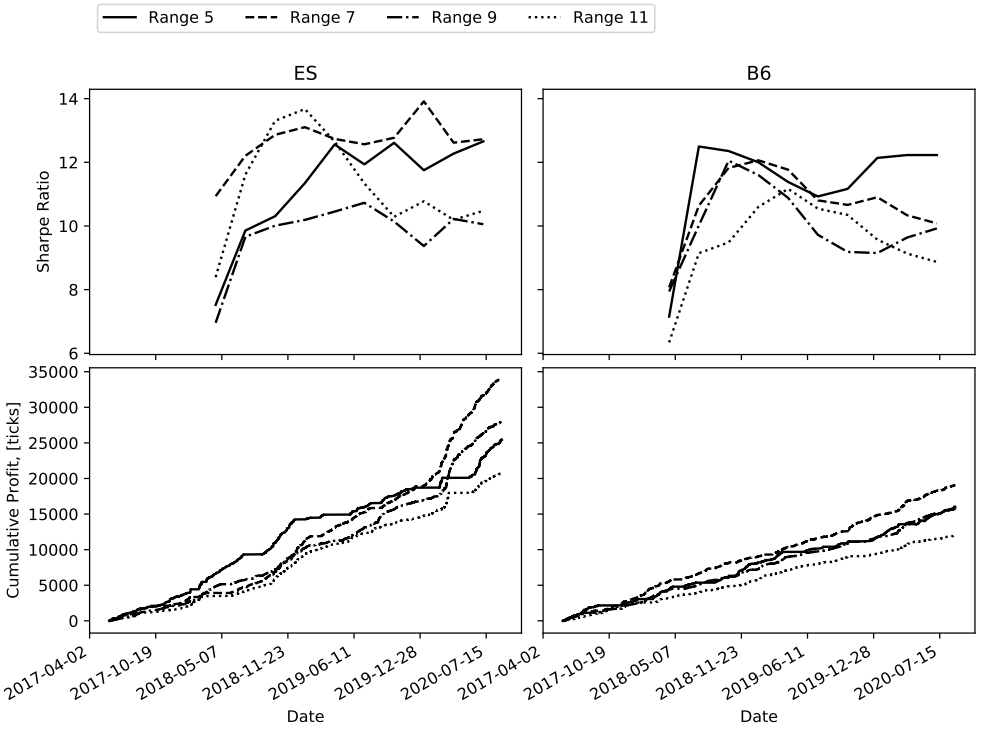

Figure 6: Sharpe ratios and cumulative profits of the volume-based method configurations. Profits are provided in ticks.

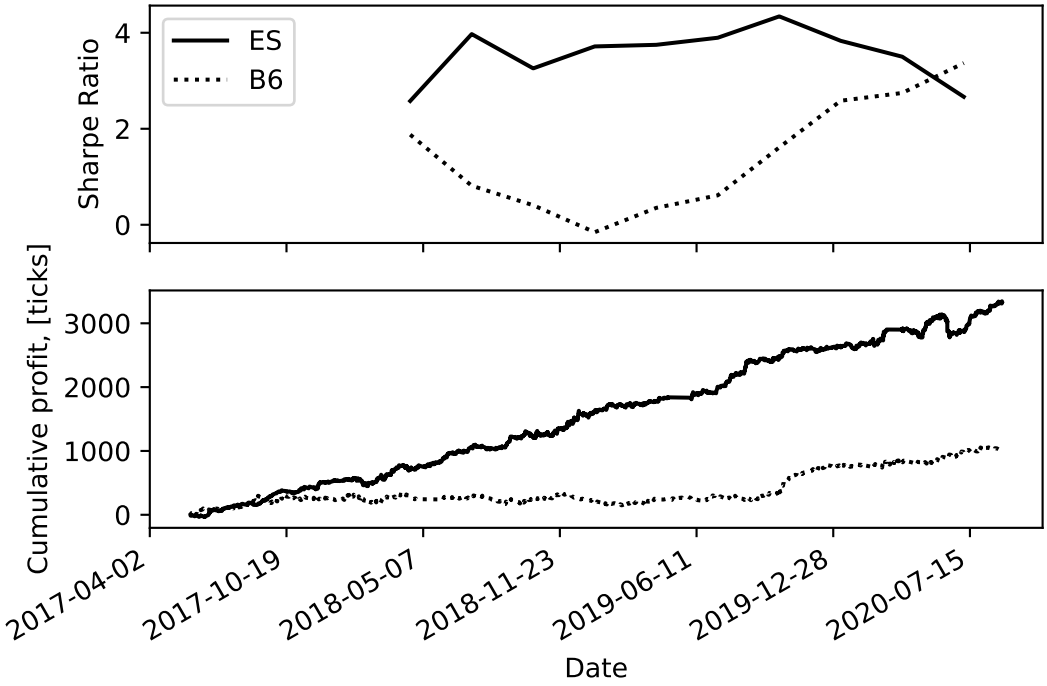

Figure 7: Sharpe ratios and cumulative profits of the price level-based method. Profits are provided in ticks.

3.5. Relatedness of feature interactions from SHAP and decision paths (RQ4)

The following results assess the relatedness of the feature interactions data extracted using SHAP and the proposed decision paths methods by comparing their relatedness to the bootstrapped data. Following the format of the previous experiments, we report effect sizes in Figure 8, and provide results of the statistical test as well as supporting statistics in Table 10. Additionally, we report the statistical test outcomes for the rest of the configurations in the supplementary materials, Table 17. To allow easier assessment of the results, we plot the differences between data representing both hypotheses in supplementary materials, Figure 12.

From the Figure 8 we see that the smallest effect sizes are observed for configurations 7 and 9 in ES and for configuration 5 in B6. Interestingly, the behaviour across the configurations is flipped for ES and B6. In Table 10 one can see that there is no skew in the data as mean and median values are very similar. At the same time we see significant differences in the data variances, which we address in the Discussion

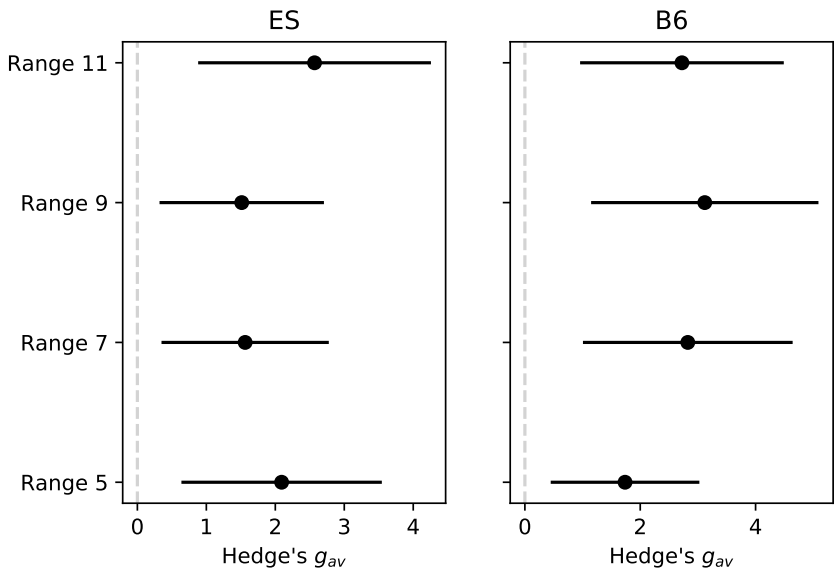

Figure 8: Hedge's g_{av} effect sizes quantifying the relatedness strength of the SHAP and decision paths methods for extracting feature interactions with respect to the relatedness of the bootstrapped data. The relatedness of the feature interactions is assessed through the Footrule distances of the ranked interaction strengths. Error bars illustrate the .95 confidence intervals corrected for multiple comparisons.

Statistics	**Dataset**			
	ES		**B6**	
Wilcoxon	< .001		< .001	
Test Statistics	2.0		0.0	
	Actual distance	Bootstrapped	Actual distance	Bootstrapped
Mean	88281	93281	86552	93280
Median	88104	93279	86296	93281
SD	4208.0	5.3	3134	6.6

Table 10: Outcomes of the one-tailed Wilcoxon test which validate whether SHAP and decision paths feature interactions extraction methods are related significantly stronger than the bootstrapped data. Mean, Median and Standard Deviation (SD) are produced for footrule distance. Footrule distance is inversely proportional to the relatedness. The result is reported for the range 7 configuration.

section. The test statistics values are small in comparison to the previous tests - in the current experiment smaller Footrule distances represent our alternative hypothesis, hence the statistical test is computed for the opposite difference sign with respect to the previous cases. In the current experiment family we run 8 tests in total, hence the corrected significance level is $\alpha = .05/8 = .00625$.

4. Discussion

In the current section we reflect on the obtained results in the same order as the experiments were conducted. Namely, performance comparison between: i) CatBoost and no-information models; ii) volume-based (VCRB) and price levels extraction methods; iii) ES and B6 financial instrument; and iv) relatedness of feature interactions obtained from SHAP and the explicit model decision paths. Furthermore, we reflect on the limitations of the study as well as its broader implications and future work.

When discussing effect sizes, we note that in different fields interpretation of the effect sizes varies depending on the commonly observed differences in the effects between the test groups [24]. For instance, in social and medical sciences it is common to consider Hedge's g effect sizes above 0.2, 0.5 and 0.8 as small, medium and large, respectively [19]. To our knowledge, there are no established effect size thresholds in the field of financial time series data analysis. Hence, we mainly use the 0-threshold indicating absence of the effect size, and contribute to the establishment of the domain-specific thresholds by reporting the effect sizes. Interpreting the data, we are guided by the confidence intervals as they represent 95% chance of finding the population effect size within the reported range.

4.1. RQ1 - Classification Performance of VCRB Bars

Larger effect sizes for the ES instrument in Figure 3 mean that there is a larger improvement from using the CatBoost model for ES rather than for B6 instrument. We reported the p-value of the statistical test in Tables 7 and 14. For the configuration range 7 the null hypothesis is rejected for both instruments. Considering the corrected significance level for the current experiment family, the null hypothesis is rejected for the rest of the configurations with one exception of range 9, B6 instrument 14. The rejects of the null hypothesis mean that CatBoost estimator performs significantly better than the no-information model. From these results we conclude that the feature space and the model work acceptably well across the considered configurations.

No-information models, whose precision represents the fraction of the positively labelled patterns, are the highest for range 5 configuration (Table 14). An obvious interpretation is that range 5 configuration is the most suitable for the studied markets. Potential underlying reasons include lack of interest in the implied trading frequency from the larger market participants, whose capitals exceed liquidity offered at this time scale. Alternatively, it might mean that the observed performance supremacy is purely theoretical and in reality is levelled off by higher risks associated with more frequent market exposure (more trades).

Looking at the precision of the models in Figure 9, there is an evident descending trend in time for both models and instruments. This can be interpreted as a gradual increase of market efficiency in the aspect of the wider use of market microstructure data for investment and trading decision making. Overall, CatBoost model with the considered feature space gives a larger performance increase with respect to the no-information model for ES than for B6 dataset. This might be due to a larger trading volumes of the ES instrument, hence less price action not supported by volumes.

4.2. RQ2 - Comparison of VCRB vs Price Level Trading

The statistical test outcomes are reported in Tables 8 and 15. Considering the corrected significance level for the current experiment family, all the statistical tests in the current experiment family have a significant outcome, and the null hypothesis is rejected. Answering the research question, the results mean that VCRB patterns can be classified with a significantly better performance than price level-based patterns.

All the effect sizes in Figure 4 are significant based on the provided confidence intervals. In Figure 10 we see quite a stable gap of around 0.15 in the PR-AUC performance between VCRB and price levels for ES. The gap is smaller (around 0.1) and possibly converging for B6, with one entry where price levels method performs better than the volume-based.

Estimators might be worse at learning a reliable classification path from the price levels data for at least two hypothetical reasons: i) higher non-stationarity of the price level patterns in comparison to the VCRBs; ii) smaller price levels dataset sizes (Table 4). The latter can be potentially solved by increasing the considered time ranges of the training datasets. However, verification of any of these reasons is out of the scope of the current work and requires a separate research for validation. Finally we see a better backtesting performance for the volume-based method in comparison to the price level in Figs 6, 7.

4.3. RQ3 - Impact of Market Liquidity on VCRB

As per our initial hypothesis (H_3), VCRB trading performed significantly better in the more liquid market (ES). This was shown by rejecting the null hypothesis, as the VCRB bars were able to classify with significantly better performance on the ES dataset for all the considered configurations. Looking at the model performance for both instruments in Figure 11, one sees that the method performs consistently better for ES dataset with differences in a range between 0.02 and 0.08. This is something expected as with more liquidity comes more impact from the volume-based features. This is also reflected in the

number of Point of Controls, with ever increasing identified points correlating to an increase in liquidity (more recent years).

One of the reasons of the better performance for ES dataset is much larger number of extracted patterns in the ES market (Table 4). We can also observe from the results in Tab 9, that the PR-AUC of the prediction is significantly higher for the more liquid market (Seen graphically in the supplementary materials in Fig. 9. Finally we also see a better backtesting performance for the more liquid asset in Fig 6. Interestingly, we see that Sharpe ratios have the same character of changes across configurations and even instruments. There is less similarity between the two pattern extraction methods. While it is a known fact that most of financial instruments are related, making it hard to diversify risks, we see that these relations cause similar impacts on the trading performance. This might suggest that using the same trading approach across instruments might not contribute to risks diversification to the expected extent, and requires careful prior research.

4.4. RQ4 - Feature Interaction Associations

In this work we investigated the relatedness of two different feature interaction methods. SHAP values are a widely used measure which is easy to compute and approximates the predictor. However, due to this approximation our thinking was that it might be not suitable for every single setting, especially for the cases where the expected performance is far from ideal, like financial markets. Hence, we aimed at checking the relatedness of SHAP and explicitly extracted feature interactions. We constructed a null hypothesis dataset using a bootstrapping approach, if the relatedness between the null dataset and the other feature analysis dataset were similar, than the relatedness would be insignificant; conversely, if the relatedness was akin between the interaction analysis methods but not the null hypothesis then our hypothesis (H_4) would be confirmed. From the test outcomes in Tables 10 and 17 we conclude that for all the configurations and both instruments the the null hypothesis is rejected and relatedness between the two methods is significant. In Figure 12 we see that mean bootstrapped distances has a negligible variance in comparison to the actual feature interaction methods. Low variance of the null hypothesis data is a sign of the correct choice of the bootstrapped sample size. Also, distances of all the entries are smaller than the null hypothesis data for B6 and ES with one exception entry being slightly larger than the bootstrapped entry. We also have a strong relatedness between the two measures highlighting a strong interaction between the feature results of the two methods. Distances between the two methods are around 5% smaller on average than the null hypothesis datasets. One should remember that

the approaches have completely different underlying principles which probably causes these large differences. SHAP assigns a local explanation based on introducing a numeric measure of credit to each input feature. Then by combining many local explanations, it represents global structure while retaining local faithfulness to the original model [25]. This is in contrast to our explicit approach which instead computes global values, accounting for the large mean distance.

4.5. Limitations

We would like to highlight that the analysis provided in this paper is but one of the possible means to empirically answer our research questions. Whilst we choose to focus on strict statistical analysis to showcase the significance of the results, other valid approaches could be implemented. A limitation of our work is that whilst our analysis is extensive, it focuses on only two examples of trading instruments. Whilst these are chosen to represent different markets that can observe performance of our method in vastly different scenarios, one could envision that a more thorough systematic analysis of the same approach across varied instruments could have some benefits; although we argue this may be out of the scope of the current work. One potential limitation of our analysis resides in the relatively small sample size. To validate statistically the significance of the performance increase we made use of the Wilcoxon test. This decision was guided by the small sample size, which in turn made it not feasible to reliably establish whether the data was normally distributed, consequently we could not use parametric tests such a t-test. Whilst sometimes less precise, this decision is supported by literature to be the optimum choice in these circumstances [26].

When assessing the absolute model performance, we consider theoretical profitability threshold computed for the simplistic strategy proposed in [2]. Concretely, assuming that the take-profit is 15 ticks, and the stop-loss is 3 ticks, we account 0.5 ticks for the trading fees (which is above a typical trading fee at the beginning of 2021). Here we do not take into account slippage as our approach uses limit orders for executing trades. For each entry we have a maximum possible theoretical profit of 14.5 ticks and maximum loss of 3.5 ticks. Dividing one by another we get the theoretical profitability threshold at 24.1% precision. Being more conservative, we would want to account for the bid-ask spread, which is 1 tick most of the time for ES and B6 (less stable). Presence of the spread means that after we enter the market, our open P&L (profit & loss) is -1 tick - if the position is opened by bid, it will be liquidated by ask which is 1 tick away, an vise versa. Of course, it affects the fraction of the trades closed by the stop loss in live setting in comparison to the no-spread simulation. We don't have enough information to probabilistically model

this, but if we account for the spread by subtracting 1 tick from all trades, we end up with the profitability threshold of 33.3% precision. While the original take profit to stop loss sizes relate as 1/5, the actual picture (after taking into account all the mechanics and fees) is very different. The main limitation of the profitability threshold value comes from multiple entries observed within a short time range. With an already open position the assumed strategy does not make use of the following signals until the current position gets liquidated. Also, there is a limitation caused by order queues hampering the strategy performance on the live market - potentially profitable orders are more likely to not be executed, as we are expecting the price to reverse. Losing positions will be executed always as the price continues its movement. We note that the different instruments and datasets will involve differences in volatility and liquidity, which impacts the length of the trades and other factors which may influence backtesting results. Considering all these, we conclude that it is necessary to take the backtesting results with a certain grain of salt.

We note that in some other machine learning contexts the obtained performance may seem low if not horrendous, however in these very complex classification scenarios it is inline with expectations as shown in previous work [2].

It should be noted that by design our hypotheses are tested on market microstructure-based feature space. By rejecting the null hypotheses we cannot claim that these findings hold for an arbitrary feature space, but rather for the proposed one. By running the experiments with the feature selection and model optimisation steps, we make an informal effort to expand the findings to a flexible feature set and model configuration. Even though we have made the best effort to unify the experiment design, the feature spaces slightly vary between the two methods, which may have some impact on the results.

4.6. Implications for practitioners

The current study proposes an approach to classification of patterns in multidimensional non-stationary financial time series. This work advances the body of knowledge in the applied financial time series analysis by proposing a pattern extraction method and shaping the contexts in which it can be used. The proposed method can be directly applied as a part of an algorithmic trading pipeline. Moreover, the method might become an alternative way of sampling the market and stand in a row with other types of sampling, like volume and range bars.

There are other research areas dealing with the similar setting, like social networks and forums analysis, topic detection and tracking, fraud detection, etc. We believe that the idea of the proposed method is

applicable to some of these fields. Namely, identification of a pattern in one of the time series dimensions as an "anchor" for the multidimensional pattern extraction. For obtaining optimal results, the design of the anchor should involve domain knowledge.

4.7. Future work

We see a number of paths for the future work. Namely, extension of the experiments to other markets, more advanced backtesting, fundamental assessment of the stationarity, and extension of the feature space. Other financial markets, like Forex, Crypto and stocks would require certain adjustment of the proposed method, since there are multiple marketplaces and volumes are distributed across them. Moreover, prices might also differ between the marketplaces, making it harder to aggregate the data. Overall, each financial instrument has its own characteristic properties and it would be interesting to see how generalisable the proposed method is.

While the used backtesting engine is relatively simple, there are more advanced ones exist in the field. However, they are usually made available as parts of trading platforms. Since the trading platforms are usually proprietary, the mechanics of the backtesting engines are not always clear and transparent and cannot be replicated outside the platform. Hence, it would be beneficial to develop an open-source package for backtesting which includes bid-ask spreads and models the trading queues.

Our study is based on empirical methods and is considered application-oriented. We hypothesise that the proposed method allows extracting patterns which are more stationary that the market itself. However, we never formally measure the stationarity, to avoid further complication of the study design. Formal assessment of the stationarity would allow choosing the most promising pattern extraction methods which in theory might require less training entries for successful classification and trading.

In the current study we used volumes-based feature space. There is a different approach to feature design - technical indicators, like RSI, MACD, Parabolic SAR, etc. To fully incorporate the indicators into the pipeline, the feature space should be increased significantly. This might be done in parallel with more fine-grained configuration of the estimator and development of a more realistic trading strategy. Even though this point is rather implementation-focused, it might lead to very interesting results, which could be further supported by the statistical approach followed in the current study.

5. Conclusion

In this study we present a new automated trading pattern extraction method suitable for ML called Volume-Centred Range Bars (VCRB). The study presents a detailed statistical analysis of the presented approach to thoroughly assess its performance.

We firstly assess the volume-based pattern extraction validity by evaluating 1) significance of the classification performance, 2) improvement for the proposed feature space and 3) model configurations with respect to the baseline (RQ1). This expands beyond simply trading using VCRBs as we showcase how performance can be improved using state-of-the-art feature engineering approach and a machine learning estimator. We further investigate method's effectiveness by comparing it with another successful pattern extraction method based on price levels. The results showcase a net improvement in performance across two different financial instruments (RQ2). We also confirm our hypothesis (H3) that liquid markets improve the effectiveness of the proposed approach, and successfully validate this.

Additionally, contributing to the explainability, we compare two different feature interaction extraction approaches - popular ML approach SHAP, which approximates the model, and an extension of Monoforest, which uses explicit decision paths from the model. The analysis shows that in the considered setting, both methods are significantly related, hence holding some common findings. We conclude that SHAP is effective in providing explainability in the considered setting, something which had previously not been investigated.

To conclude, all our methodology is structured in a way that allows for comparability across studies by providing the effect sizes; and reproducibility, by detailing the method and sharing the reproducibility package. Our hope is that this will make it easier for the practitioners to test this same approach and evaluate it against other methods, hopefully helping improve the field for the better. Code to reproduce our analysis and match our results is available online [23].

6. Supplementary Materials

Glossary

Terms	Definitions
Futures contract	provide means to trade a commodity (instrument) at a predetermined price at a specific time in the future.
Tick	represent a single movement upward or downward by a specific increment in price for a specific instrument (e.g. 0.25$ for S&P futures).
Bar	used to identify a window of interest based on some heuristic, and then aggregate the features of that window. May contain several features and it is up to the individual to decide what features to select, common features include: *Bar start time*, *Bar end time*, *Sum of Volume*, *Open Price*, *Close Price*, *Min* and *Max* (usually called High and Low) prices, and any other features that might help characterise the trading performed within this window
Volume	refers to the number of traded contracts (or shares) for a particular instrument
Volume profile	refers to the volumes traded per price, visualised as a vertical histogram for a range of prices over a certain time range
Liquidity	how rapidly stocks may be traded without affecting market price. Has an impact on whether you are able to get the desired instrument at your choice of price (sell or buy)
Volatility	degree of variation for the price of a given instrument over a period of time.
Trading	the buying and selling of an instrument
Trading platform	is software that you use to conduct your trading. Allows for the centralised management of instruments and positions
Time & Sales	a set of features provided real time for each trade executed in an exchange. Features include: *volume*, *price*, *direction*, *date*, and *time*
Order Book	the list of orders used by a trading venue to keep track of offers and bids by buyers and sellers for a particular instrument. These are then matched in specific order to execute a trade

Flat Market	is a stable state in which the range for the broader market does not move either higher or lower, but instead trades within the boundaries of recent highs and lows.
Trending Market	shifts in the market towards a raise or decrease in price compared to expected highs or lows. Used to buy and sell at the point where ones is most likely to gain profit
Long positions	owning the asset for a time period with the expectation that the asset will go up in price
Short positions	if the expectation is that the price will decrease over time, you can short an asset to profit from is decreasing value.
Actionable ML	In the context of machine learning and more specifically algorithmic trading, actionability refers to the ability to act upon a prediction. This may directly relate to understanding the reason behind the prediction, in turn allowing you to make informed decisions on how to act upon it.
Take-profit	an order that specifies a price at which to trade at exactly. The order remains open until the price is reached
Stop-loss	an order placed at a specific price that gets closed if the price lowers beyond a certain amount. This is meant to reduce potential losses incurred if the desired price is not reached.

Glossary: This glossary contains some essential definitions used throughout the paper. Sometimes similar definitions are reintroduced in specific contexts to centre the discussion.

	PR-AUC	ROC-AUC	F1-score	precision	Null_precision
3/17/3 to 6/17/6	0.25	0.02	0.04	0.50	0.26
6/17 to 9/17	0.34	0.04	0.18	0.54	0.32
9/17 to 12/17	0.26	0.30	0.27	0.51	0.25
12/17 to 3/18	0.27	0.19	0.26	0.54	0.25
3/18 to 6/18	0.28	0.16	0.32	0.51	0.27
6/18 to 9/18	0.26	0.16	0.28	0.52	0.25
9/18 to 12/18	0.24	0.16	0.26	0.50	0.24
12/18 to 3/19	0.30	0.33	0.33	0.53	0.27
3/19 to 6/19	0.29	0.18	0.32	0.51	0.26
6/19 to 9/19	0.28	0.22	0.31	0.52	0.26
9/19 to 12/19	0.22	0.08	0.17	0.50	0.23
12/19 to 3/20	0.25	0.08	0.26	0.50	0.25
3/20 to 6/20	0.28	0.17	0.30	0.54	0.26

Table 12: Performance metrics for ES, price levels pattern extraction method.

	PR-AUC	ROC-AUC	F1-score	precision	Null_precision
3/17 to 6/17	0.26	0.10	0.19	0.46	0.28
6/17 to 9/17	0.26	0.20	0.24	0.53	0.24
9/17 to 12/17	0.27	0.30	0.30	0.52	0.25
12/17 to 3/18	0.24	0.11	0.22	0.48	0.25
3/18 to 6/18	0.21	0.17	0.16	0.44	0.21
6/18 to 9/18	0.25	0.21	0.24	0.50	0.25
9/18 to 12/18	0.25	0.33	0.26	0.58	0.21
12/18 to 3/19	0.26	0.33	0.23	0.53	0.23
3/19 to 6/19	0.36	0.24	0.48	0.57	0.26
6/19 to 9/19	0.31	0.29	0.28	0.49	0.31
9/19 to 12/19	0.23	0.19	0.21	0.50	0.25
12/19 to 3/20	0.24	0.33	0.24	0.50	0.24
3/20 to 6/20	0.28	0.25	0.27	0.54	0.26

Table 13: Performance metrics for B6, price levels pattern extraction method.

Statistics	Dataset			
	Range 5			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	91.0		90.0	
	CatBoost	No-information	CatBoost	No-information
Mean (precision)	0.47	0.42	0.38	0.37
Median (precision)	0.47	0.42	0.38	0.37
Standard Deviation (precision)	0.021	0.018	0.011	0.007
	Range 9			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		.029	
Test Statistics	91.0		73.0	
	CatBoost	No-information	CatBoost	No-information
Mean (precision)	0.43	0.39	0.37	0.36
Median (precision)	0.43	0.39	0.37	0.36
Standard Deviation (precision)	0.013	0.013	0.023	0.01
	Range 11			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		.004	
Test Statistics	88.0		82.0	
	CatBoost	No-information	CatBoost	No-information
Mean (precision)	0.43	0.40	0.38	0.36
Median (precision)	0.43	0.39	0.39	0.36
Standard Deviation (precision)	0.028	0.024	0.031	0.015

Table 14: Statistics supporting the outcomes of the Wilcoxon test. The test is aimed to validate whether on the VCRB data and the considered feature space, CatBoost performs significantly better than the no-information estimator. The result is reported for the range configurations of 5, 9 and 11.

Statistics	Dataset			
	Range 5			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	91.0		91.0	
	VCRB	Price levels	VCRB	Price levels
Mean (PR-AUC)	0.48	0.27	0.38	0.26
Median (PR-AUC)	0.48	0.27	0.39	0.26
Standard Deviation (PR-AUC)	0.025	0.029	0.012	0.036
	Range 9			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	91.0		90.0	
	VCRB	Price levels	VCRB	Price levels
Mean (PR-AUC)	0.42	0.27	0.37	0.26
Median (PR-AUC)	0.43	0.27	0.37	0.26
Standard Deviation (PR-AUC)	0.0140	0.029	0.02	0.036
	Range 11			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	91.0		91.0	
	VCRB	Price levels	VCRB	Price levels
Mean (PR-AUC)	0.42	0.27	0.38	0.26
Median (PR-AUC)	0.43	0.27	0.38	0.26
Standard Deviation (PR-AUC)	0.028	0.029	0.02	0.036

Table 15: Statistics supporting the outcomes of the Wilcoxon test which validates whether Volume-based pattern extraction method leads to better classification performance than the price level pattern extraction. The result is reported for the range configurations of 5, 9 and 11.

Statistics	Datasets	
	Range 5	
One-tailed Wilcoxon test p-value	< .001	
Test Statistics	91.0	
	ES	B6
Mean (PR-AUC)	0.48	0.38
Median (PR-AUC)	0.48	0.39
Standard Deviation (PR-AUC)	0.025	0.012
	Range 9	
One-tailed Wilcoxon test p-value	< .001	
Test Statistics	91.0	
	ES	B6
Mean (PR-AUC)	0.42	0.37
Median (PR-AUC)	0.43	0.37
Standard Deviation (PR-AUC)	0.014	0.020
	Range 11	
One-tailed Wilcoxon test p-value	.0012	
Test Statistics	86.0	
	ES	B6
Mean (PR-AUC)	0.42	0.38
Median (PR-AUC)	0.43	0.38
Standard Deviation (PR-AUC)	0.028	0.020

Table 16: Statistics supporting the outcomes of the Wilcoxon test which assesses whether VCRB pattern extraction method leads to better classification performance on the more liquid market (ES in comparison to B6). The result is reported for the range configurations of 5, 9 and 11.

Statistics	Dataset			
	Range 5			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	3.0		2.0	
	Actual distance	Bootstrapped	Actual distance	Bootstrapped
Mean (footstep distance)	88247	93280	87891	93279
Median (footstep distance)	87390	93281	88408	93278
Standard Deviation (footstep distance)	3166.0	5.0	4083.0	3.4
	Range 9			
	ES		B6	
One-tailed Wilcoxon test p-value	.00171		< .001	
Test Statistics	6.0		0.0	
	Actual distance	Bootstrapped	Actual distance	Bootstrapped
Mean (footstep distance)	88188	93279	86172	93278
Median (footstep distance)	88728	93279	86132	93279
Standard Deviation (footstep distance)	4429.0	6.1	2998.0	4.6
	Range 11			
	ES		B6	
One-tailed Wilcoxon test p-value	< .001		< .001	
Test Statistics	1.0		0.0	
	Actual distance	Bootstrapped	Actual distance	Bootstrapped
Mean (footstep distance)	87266	93281	87228	93280
Median (footstep distance)	86994	93281	86198	93279
Standard Deviation (footstep distance)	3081.0	5.5	2925.0	3.4

Table 17: Statistics supporting the outcomes of the Wilcoxon test which validates whether SHAP and decision paths feature interaction extraction methods are related significantly stronger than the bootstrapped data. Footstep distance is inversely proportional to the relatedness. The result is reported for the range configurations of 5, 9 and 11.

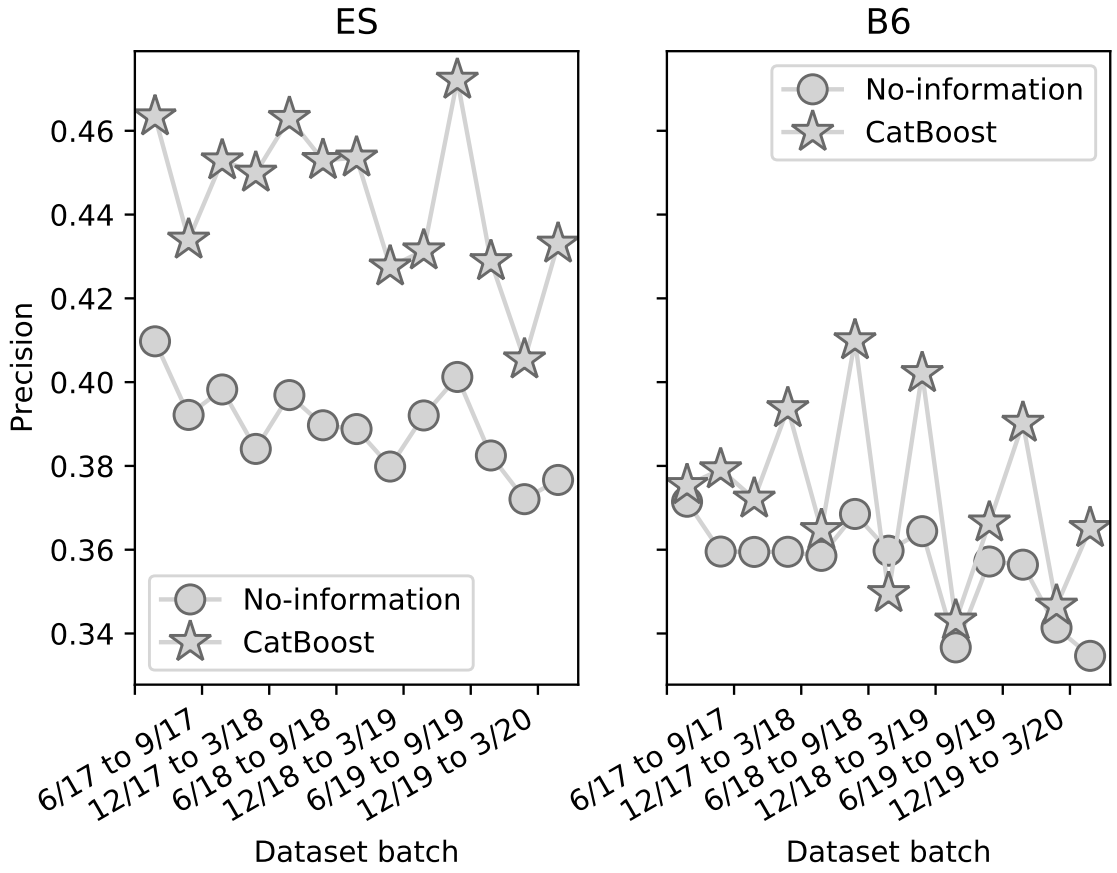

Figure 9: Precision performance metric for the no-information model and CatBoost plotted for both instruments - ES and B6. The Y-axis is mutual for both subplots.

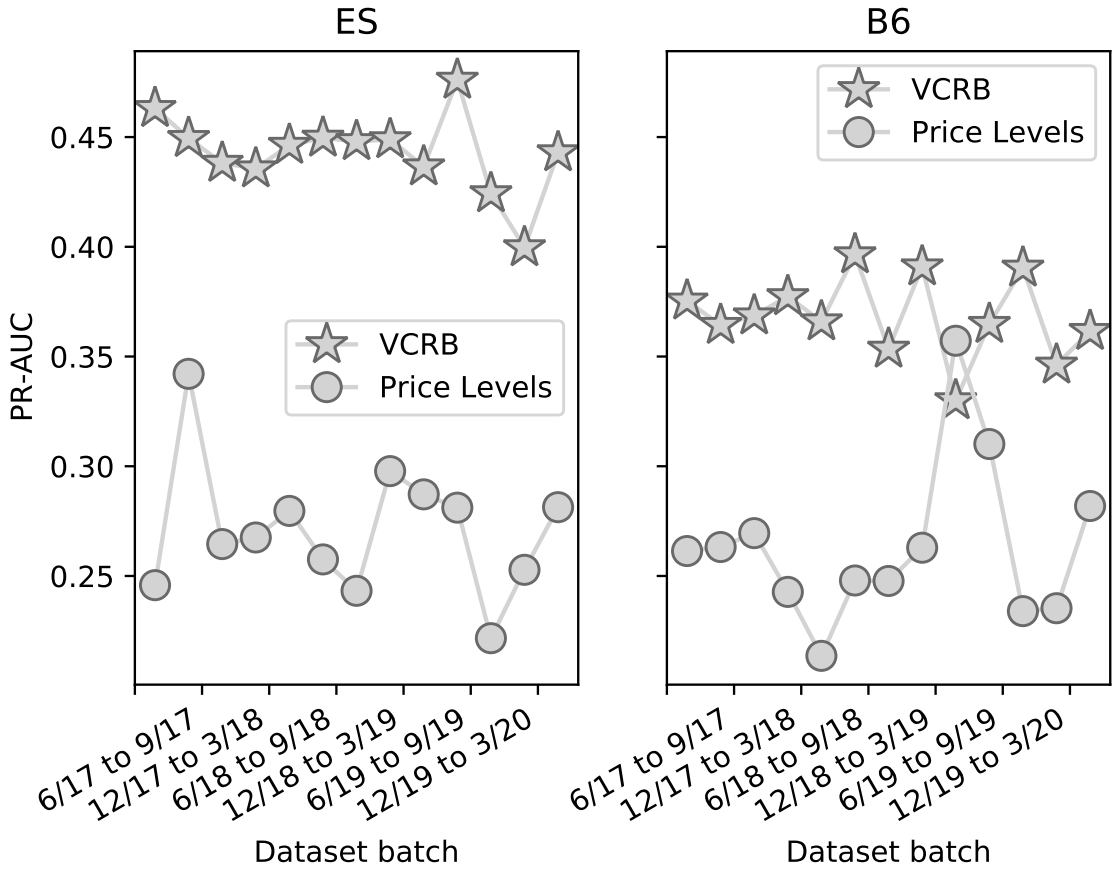

Figure 10: We plot PR-AUC for volume-based pattern extraction method and price level-based. The metric is reported for ES and B6 instruments. Volume-based method is reported for range 7 configuration.

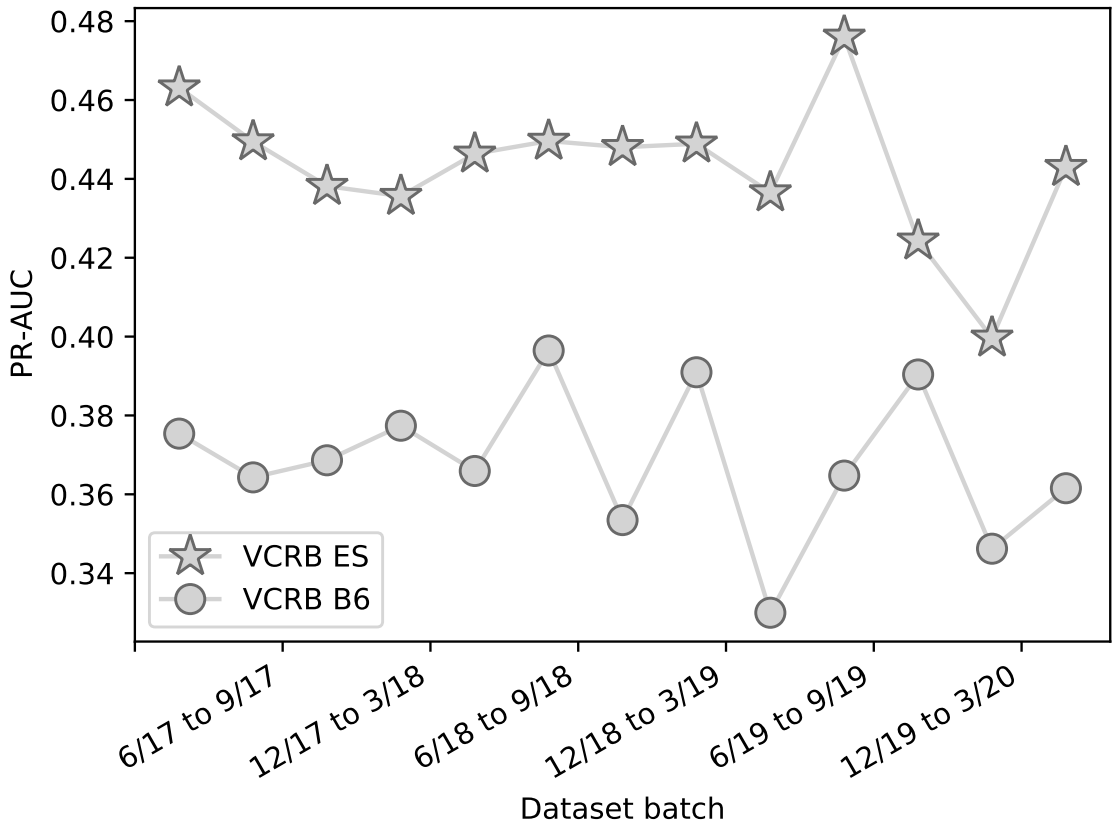

Figure 11: PR-AUC of CatBoost models obtained for ES and B6 futures instruments. Reported for range 7 configuration.

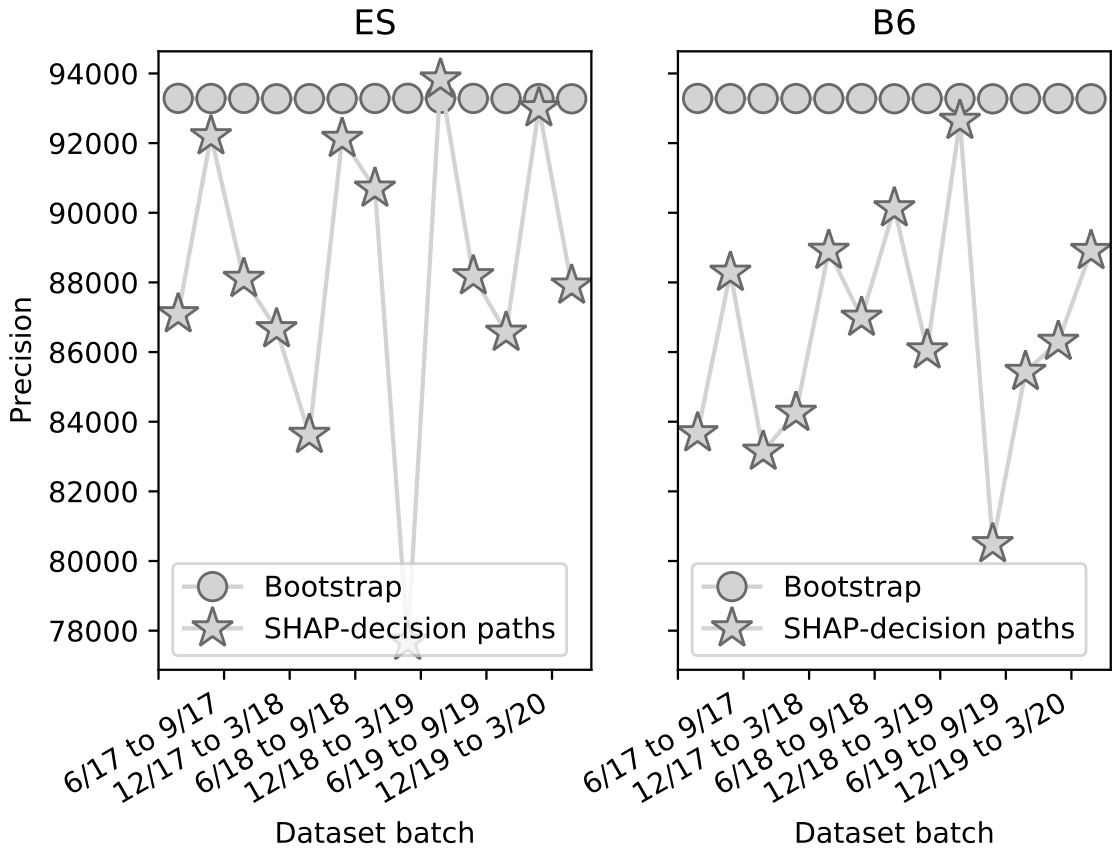

Figure 12: Footrule distances between ranks of the feature interactions for SHAP and decision path-based methods for S&P E-mini and British Pound futures instruments.

References

[1] M. L. De Prado, Advances in financial machine learning, John Wiley & Sons, 2018.

[2] A. Sokolovsky, L. Arnaboldi, Machine learning classification of price extrema based on market microstructure features: A case study of s&p500 e-mini futures, arXiv preprint arXiv:2009.09993 (2020).

[3] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 33–44.

[4] M. A. Ahmad, A. Teredesai, C. Eckert, Fairness, accountability, transparency in ai at scale: Lessons from national programs, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 690–690.

[5] K. Martin, Ethical implications and accountability of algorithms, Journal of Business Ethics 160 (2019) 835–850.

[6] H.-W. Liu, C.-F. Lin, Y.-J. Chen, Beyond state v loomis: artificial intelligence, government algorithmization and accountability, International Journal of Law and Information Technology 27 (2019) 122–141.

[7] M. Wieringa, What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 1–18.

[8] O. Biran, K. R. McKeown, Human-centric justification of machine learning predictions., in: IJCAI, volume 2017, 2017, pp. 1461–1467.

[9] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[10] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

[11] H. Hagras, Toward human-understandable, explainable ai, Computer 51 (2018) 28–36.

[12] D. Doran, S. Schulz, T. R. Besold, What does explainable ai really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794 (2017).

[13] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017).

[14] L. Breiman, et al., Statistical modeling: The two cultures (with comments and a rejoinder by the author), Statistical science 16 (2001) 199–231.

[15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: Unbiased boosting with categorical features, Advances in Neural Information Processing Systems 2018-Decem (2018) 6638–6648.

[16] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[17] I. Kuralenok, V. Ershov, I. Labutin, Monoforest framework for tree ensemble analysis, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019, pp. 13780–13789. URL: https://proceedings.neurips.cc/paper/2019/file/1b9a80606d74d3da6db2f1274557e644-Paper.pdf.

[18] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS ONE 10 (2015).

[19] D. Lakens, Calculating and Reporting Effect Sizes to Facilitate Cumulative Science, Frontiers in Psychology 4 (2013) 863.

[20] F. Wilcoxon, Individual Comparisons by Ranking Methods, Biometrics Bulletin 1 (1945) 80.

[21] B. Ce, Teoria statistica delle classi e calcolo delle probabilit, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936) 3–62.

[22] C. Spearman, The Proof and Measurement of Association between Two Things, The American Journal of Psychology 15 (1904) 72.

[23] A. Sokolovsky, L. Arnaboldi, J. Bacardit, T. Gross, Interpretable ML-driven Strategy for Automated Trading Pattern Extraction - Reproducibility Package, 2021. URL: https://doi.org/10.5281/zenodo.4629568. doi:10.5281/zenodo.4629568.

[24] J. A. Durlak, How to select, calculate, and interpret effect sizes, Journal of Pediatric Psychology 34 (2009) 917–928.

[25] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature machine intelligence 2 (2020) 56–67.

[26] E. Skovlund, G. U. Fenstad, Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?, Journal of clinical epidemiology 54 (2001) 86–92.