Analytical p-value of feature synergy for machine learning explainability

Sohaib Choufani, Jean-Luc Dupire, Nicolas Fabrigoule, Aya Khazri, Fabien Labarre

May 2023

Abstract

This article investigate feature synergies in machine learning models using Shap values. An analytical p-value is developed to evaluate their significance. A benchmark analysis on several models reveals consistent results in detecting significant synergies. The methods are then applied to a machine learning algorithm used to predict prices of an order book, highlighting synergies between some features.

1 Introduction

Machine learning has revolutionized complex problems solving and making decisions in various fields such as healthcare, finance, transportation... However, the increasing use of Machine Learning algorithms and models has brought attention to the issue of interpretability and explainability. Machine Learning models are typically designed to optimize their performance, without necessarily considering the transparency of their decision-making process, it's often described as a black box. This lack of transparency raises important questions about how to understand and explain results generated by these models. As a result, the need for interpretability and explainability in machine learning has become a crucial topic in the field, and a growing number of researchers and practitioners are exploring different ways to address this challenge.*Been Kim*'s paper defines interpretability of machine learning and how it should be measured [2].The goal of this effort is to improve the trustworthiness, accountability, and safety of Machine Learning systems, as well as to facilitate better decision-making in various applications.

Explaining how and why a machine learning model produces its output is crucial for building a robust and reliable solution. There are several reasons why data scientists may want to unbox their models, including validating the relationships discovered by the model, ensuring that the evaluation protocol is not compromised, and detecting any biases or violations. However, some machine learning models, by design, offer limited insights into their decision-making process. Therefore, model explanation is essential to attribute importance to input features individually or by groups. This paper deals with one such approach that has gained popularity in recent years : Shap (SHapley Additive exPlanations) [6]. The studies related to this topic all stem from a paper by Shapley on game theory, titled "A value for n-person games" [7]. While Shap quantifies the local contributions of one or more features, it is not designed to explain global relationships among features from the perspective of a given model. The global relationships can reveal if the model combines information from groups of features, which features are redundant with respect to the target variable, and which ones can be substituted with little or no loss of model performance (features selection).

After the introduction of Shap values, which represent the contribution of individual features to the output of a model, the natural next step is to consider how features interact with each other to

collectively affect the output. SHAP Interaction values provide a solution to this question by quantifying the interactions between pairs of features. Concretely, the SHAP Interaction value of two features represents the average difference in the SHAP values of the model output when the two features are included together versus when they are excluded from the model.

By understanding the interactions between features, data scientists can gain insights into the complex relationships between inputs and outputs in their models, which is particularly useful for improving model interpretability and enhancing its performance. In this sense, SHAP Interaction values offer a valuable tool for understanding the complex inner workings of machine learning models, and can help to make better decision for several applications. '*Consistent individualized feature attribution for tree ensembles*' [5] authors propose SHAP values for understanding tree ensemble models, widely used models, such as as random forests or gradient boosting models. Tree explainer, which is a method to provide explanations for the predictions and features contributions of tree-based models, is consistent with SHAP values interpretation [4].

When it comes to complex systems as financial markets, the interactions between different variables are often more significant than their isolated effects [1]. This is where the concept of synergy must be introduced. Synergy defines the combined impact or interaction between features, which can often result in outcomes that cannot be attributed to any single factor alone. Synergy can help to gain a more comprehensive understanding of how different features collaborate and influence model predictions [3]. This deeper insight allows us to make more accurate and nuanced interpretations of the model's behavior, particularly in domains where dependencies among variables play a crucial role, such as order book analysis.

Synergy calculation plays an important role in improving the performance of machine learning models, particularly when applied to order books. By analyzing and understanding the inter-relationships and dependencies within an order book, we can highlight valuable insights that can assist in making informed trading decisions. Model's components can include bid and ask prices and order volumes, calculating synergy involves examining how these components influence each other and contribute to the global market's dynamics. This allows us to extract meaningful patterns, detect price anomalies, and predict potential price movements or market behavior more accurately. Then, synergy calculations enable the model to go beyond analyzing individual data points and instead focus on the holistic view of the order book, capturing the collective impact of various factors simultaneously.

In this context, particularly for algorithmic trading, interpreting machine learning models is of great importance for orderbook forecasting. In the order book, a set of limit orders are placed by buyers and sellers at different prices. The order book reflects the supply and demand for an asset and is an important input for trading decisions. Accurate forecasting of the order book can provide significant trading benefits, like minimizing transaction costs and maximizing profits.

The use of machine learning models has shown promising results in improving order book forecasting accuracy. The greatest challenge in this approach is to understand how the machine learning model makes its predictions, particularly in the context of the order book, where the relationship between features is complex and non-linear.

With synergy's computation, machine learning models for order book forecasting can be made more transparent, interpretable, and robust, facilitating more informed trading decisions.

1.1 Shap vectors

Shapley values are a concept of game theory which provide a unique distribution of a total surplus generated by a coalition of players in a cooperative game [7]. This distribution is *fair* in the sense that it is the only distribution with certain desirable properties. Formally, for a game with a set of players

N and outcomes $f_x(S)$ for a coalition $S \subseteq N$, Shapley value of player i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \nabla_i(S)$$

where $\nabla_i(S) = f_x(S \cup \{i\}) - f_x(S)$ represents the marginal contribution of player i in the coalition $S \cup \{i\}$.

Applying to a predictive model $f : \mathbb{R}^n \mapsto \mathbb{R}$ where players are the features and the outcome is f_x the model evaluated for a sample $x \in \mathbb{R}^n$ gives SHAP values. Thus $f_x(S)$ is the model evaluated for sample $x \in \mathbb{R}^n$ using only the subset of features S. With notation of [1] it can be written as:

$$f_x(S) = \mathbb{E}\left[f(x)|S\right]$$

and the SHAP value ϕ_i represents the marginal contribution in prediction given by the feature i.

Then SHAP interaction value can be defined as the difference between the marginal contribution of i when j is present and the marginal contribution of i when j is absent [5]. Formally:

$$\phi_{ij} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!\,(|N| - |S| - 2)!}{2\,(|N| - 1)!} \nabla_{ij}(S)$$

where $\nabla_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S)$

For m samples in the training dataset, it is possible to compute m SHAP values and m SHAP interaction values and group them in order to create a SHAP vector and a SHAP interaction vector [3]:

$$\mathbf{p}_i = \begin{pmatrix} \phi_i^{(1)} & \dots & \phi_i^{(m)} \end{pmatrix}$$

$$\mathbf{p}_{ij} = \begin{pmatrix} \phi_{ij}^{(1)} & \dots & \phi_{ij}^{(m)} \end{pmatrix}$$

Another advantage of using SHAP values is their capacity to explain not only individual predictions but also provide a global understanding of the model by aggregating local explanations. This feature enables us to gain insights into the overall importance of each feature and the interactions between them within the model.

1.2 Synergy

From these SHAP vectors, different types of relationships can be expressed and quantification of the strength of these relationships is possible through formulas. One of these relationships is called synergy. Synergy as well as the two other relationships, redundancy, and independence, are expressed as percentages of feature importance. For each pair of features, the synergy S_{ij} quantifies the degree to which predictive contributions of x_i rely on information from x_j. As an example, two features representing coordinates on a map need to be used synergistically to predict distances from arbitrary points on the map.

$$S_{ij} = \left\langle \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|}, \frac{\mathbf{p}_{ij}}{\|\mathbf{p}_{ij}\|} \right\rangle^2$$

Geometrically, feature vector $\mathbf{p}_i$ is projected on interaction vector $\mathbf{p}_{ij}$ to obtain synergy vector $\mathbf{s}_{ij}$. Then S_{ij} is expressed by :

$$S_{ij} = \left\langle \mathbf{s}_{ij}, \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|^2} \right\rangle$$

Synergy's computation enhance our understanding of the relationships among model features.

2 Significance test of the synergy

2.1 Bootstrap method for p-value computation

We propose a method to provide the synergy computation with a p-value using re-sampling. Let us define the signed synergy $\sigma_{i,j} = \left\langle \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|}, \frac{\mathbf{p}_{i,j}}{\|\mathbf{p}_{i,j}\|} \right\rangle$. The signed synergy is calculated on k new samples constructed by choosing N instances with replacement among the initial sample $\mathcal{S}$ of size N. This allows to obtain a distribution $\mathcal{D}$ of the signed synergies. Zero synergy is considered to be the null hypothesis, which is equivalent to a zero signed synergy. Therefore, the p-value of the test can be estimated by computing the proportion of negative realizations in the distribution $\mathcal{D}$. k must be sufficiently large to obtain a very small accuracy in front of the chosen threshold of significance s (typically $k \geq \frac{100}{s}$).

The choose of a one-sided p-value is based on synergy's positivity. As a reference for the analytical method, the test aims to evaluate whether the observed value is significantly different from zero in this direction. The research hypothesis is one-sided and an effect is expected in a positive direction, then a unilateral test, with a one-sided p-value, can be used to evaluate whether the observed value is significantly different from zero in this direction. In this case, the one-sided test may be more powerful to detect a specific effect in a given direction.

In order to simplify the computations and the following tests, the problem is reduced to the study of the scalar product $\rho_{i,j} = \langle \mathbf{p}_i, \mathbf{p}_{i,j} \rangle$ which has the same sign as the signed synergy.

The p-value obtained via bootstrapping gives a reference to compare the results of the analytical method below which is much less computationally demanding.

2.2 Analytical p-value

Let $\mathcal{S}$ be the initial sample with N instances on which we want to compute a synergy. For each instance l in $\mathcal{S}$, one can compute the SHAP value $\phi_i^{(l)}$ and the SHAP Interaction value $\phi_{i,j}^{(l)}$ associated with the model. For i and j fixed, these values are considered to be respectively independent identically distributed random variables. The random variables $\phi_i^{(l)}\phi_{i,j}^{(l)}$ have therefore the same properties and the scalar product $\rho_{i,j}$ can be written as $\sum_{l \in \mathcal{S}} \phi_i^{(l)}\phi_{i,j}^{(l)}$. Let assume that $\mathbb{E}\left[\phi_i^{(l)}\phi_{i,j}^{(l)}\right] < \infty$ and $\mathbb{E}\left[\left(\phi_i^{(l)}\phi_{i,j}^{(l)}\right)^2\right] < \infty$. For N sufficiently large, the Central Limit Theorem implies that the distribution of $\frac{\rho_{i,j} - N\mu}{\sqrt{N}}$ tends to a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ (Fig.1). This allows to estimate analytically the distribution and the p-value obtained via the bootstrap method when $\mathcal{S}$ is large and the variance of $\phi_i^{(l)}\phi_{i,j}^{(l)}$ is well defined.

This method provides a fast estimation of the p-value associated to the computation of the synergy between to variables. The next part gives an overview of its efficiency and its limitations.

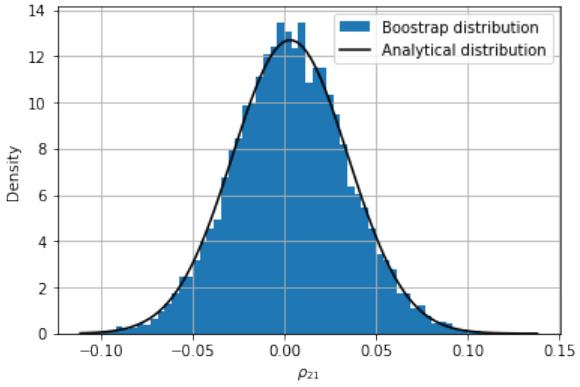

Figure 1: Density of ρ_{21} in the model of subsection 3.3 with $\alpha = 0$

3 Benchmark

3.1 Model presentation

We utilize an Extreme Gradient Boosting (XGBoost) model, an algorithm esteemed for its performance and efficiency, to scrutinize the synergy between multiple features. This choice is influenced by XGBoost's excellent capacity to handle a variety of data structures, the possibility of tuning various parameters for model optimization, and its capability to fit complex nonlinear relationships.

We employ SHapley Additive exPlanations (SHAP) Tree Explainer. This tool, specifically tailored for tree-based models like XGBoost, facilitates an efficient and accurate computation of SHAP values. We consider observations of $n = 5$ features, represented by n dimensional vectors $x_1, ..., x_n)$ drawn independently from a uniform distribution on $[0, 1]^5$. We use the following model introduced in [3] to build our datasets:

$$f(x) = \sin(2\pi x_1)\sin(\pi(x_2 + x_3)) + x_4 + x_5 \tag{1}$$

3.2 True null hypothesis

In order to test the performance of the previous methods in the case of a true null hypothesis H_0 (the absence of synergy), a pair of features without synergy is chosen from the model above, for example x_2 and x_5. The synergies S_{25} should be equal to 0 as the contribution of x_2 is independent to the value of x_5 assuming that the machine learning model used fits sufficiently well the function f. The first and second moments of the random variables $\phi_2^{(l)}\phi_{2,5}^{(l)}$ are well defined so that the computation of the analytical p-value is unbiased with respect to the bootstrap p-value (Fig.2).

Here, an XGBoost model is fitted on a training dataset of 100000 points randomly drawn from a standard normal distribution for each feature. Then, 100 test datasets with 1000 instances each are used to study the behavior of the estimation of the p-value when the null hypothesis is true.

For each test dataset, the p-value associated to the computation of the synergy S_{14} is estimated via both bootstrap and analytical methods. Under H_0, the distribution of the obtained p-values should be a uniform distribution by definition, *i.e.* the cumulative distribution should be along the first bisector. If it is above (resp.below) this bisector, the estimation of the p-value is too permissive (resp. too restrictive). In Fig.2, the cumulative distributions of bootstrap and analytical p-values do not display any tendency to be permissive or restrictive.

A special case of true null hypothesis for tree-based models is obtained when the maximum depth of the trees is set to 1. Each submodel distinguishes instances according to only one variable, therefore the effects of combinations of variables are not captured by the global model. In particular, no synergy

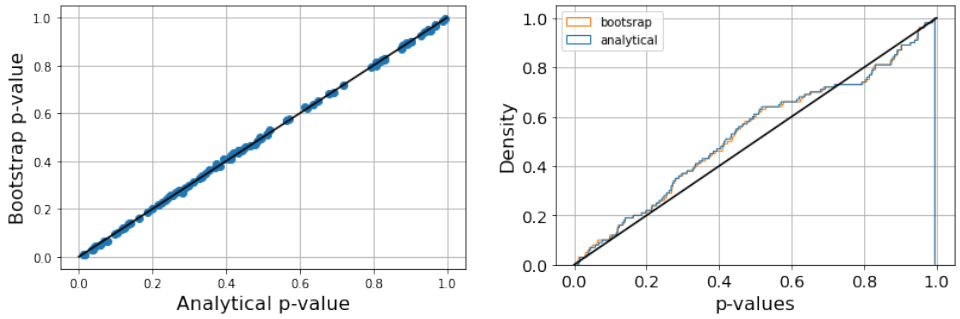

Figure 2: a) Comparison of the analytical and the bootstrap p-values obtained on the 100 test datasets. — b) Cumulative distributive distributions of the p-values obtained on the 100 test datasets.

can be observed in this case, as all the SHAP Interaction values are zero. Thus, the computation of the analytical p-value must be adapted to the case of a zero-valued variance of the random variable $\phi_i^{(l)}\phi_{i,j}^{(l)}$ by setting the p-value to 1 if $\mathbb{E}\left[\phi_i^{(l)}\phi_{i,j}^{(l)}\right] \leq 0$ and 0 otherwise.

3.3 Benchmark model

We now introduce the following benchmark model to control the variation of the synergy of a pair of features:

$$g_\alpha(x) = \sin(2\pi x_1)\sin(2\pi(\alpha x_2 + (1-\alpha)x_3)) + (1-\alpha)x_2 + \alpha x_3 \tag{2}$$

with $\alpha \in [0,1]$. The benchmark datasets are built by randomly drawing independent instances $x = (x_1, x_2, x_3)$ from a uniform distribution on $[0,1]^3$. The synergy between x_2 (resp.x_3) and x_1, denoted S_{21} (resp.S_{31}), depends on the coefficient α. At $\alpha = 0$ for instance, knowing x_1 without knowing x_3 does not change the expected value of $g_\alpha(x)$, *i.e.* the synergy S_{31} equals 1. On the contrary, x_2 does not provide information on the contribution of x_1, *i.e.* the synergy S_{21} equals 0. The variation of the synergies S_{21} and S_{31} with respect to the parameter α is displayed in Fig.3 for the fitted models.

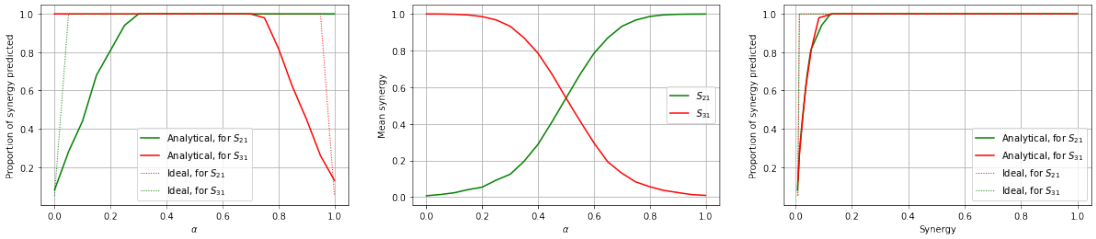

Figure 3: a) Proportion of rejected null hypothesis for a threshold of 0.05 — b) Average synergies S_{21} and S_{31} over 100 datasets. c) Proportion of rejected null hypothesis with respect to the average synergy

For α in $\{0, 0.5, 0.1, ..., 1\}$, a training dataset of 100000 instances is built using g_α and an XGBoost model is fitted to predict $g_\alpha(x)$. Then p-values of the feature synergies are computed for 100 tests datasets of size 1000. Fig.3 display the proportion of significant p-value among them for S_{21} and S_{31}. Ideally, as the threshold of significance is set to 0.05, this proportion should be 0.05 when there is no synergy and 1 otherwise. In practice, apart from the cases with too small synergies (typically smaller than 0.1), significant synergy are well detected. The discrepancy between theory and practice when

the synergy is zero may be due to the fact that g_α is not perfectly fitted (with phenomenon such as overfitting for instance) on a finite-size dataset.

3.4 Overfitting

Overfitting can lead to the emergence of non-zero synergies between features that do not exhibit true synergy in the underlying model. When a machine learning algorithm attempts to approximate the true model, it may erroneously hallucinate synergies due to the overfitting phenomenon. It is important to note that the synergy values computed using our proposed method correspond to the synergies present in the model built by the machine learning algorithm itself, rather than the actual synergies present in the true model. This distinction emphasizes that the synergy values obtained reflect the learned relationships within the specific model used, and caution should be exercised in interpreting these values as indicative of the true underlying synergies between features.

3.5 Fat-tailed distributions

The presence of some outliers and extreme events in the data may introduce fat-tailed distributions of the $\phi_i^{(l)}\phi_{i,j}^{(l)}$. In this case, the Central Limit Theorem as used in part 2.2 no longer applies.

In order to observe the impact of this phenomenon on the precision of the analytical p-value, the same calculations as in part 3.2 are reproduced but with a different way of creating the training dataset. This time, the observations are drawn independently from a t-distribution with 2 degrees of freedom on $\mathbb{R}^5$.

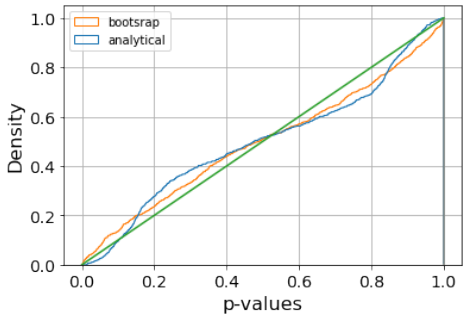

Figure 4: Cumulative distributions of the p-values obtained on the 100 test datasets.

The cumulative distributions of the analytical and bootstrap p-values displayed in Fig.4 suggests that extreme events generate a bias in the analytical p-value, which has no particular tendency (neither too restrictive nor permissive).

4 Application to order book models

4.1 Introduction to order book

The concept of the order book is fundamental in the world of financial markets. An order book is a list that records orders to buy and sell a specific security or financial instrument, classified by date. It provides an overview of the supply and demand for a security, including both the price and the volume at which market agents are prepared to buy or sell.

Analyzing an order book with Shapley values and synergy could bring significant insights. In the context of an order book, each order can be seen as a contribution to the overall market dynamics. Analyzing the 'contributions' of these orders can help understand their influence on price movements.

The idea of synergy, also has relevance in this context. Order synergy can materialize when combinations of orders collectively impact the market more significantly than their standalone influences. This could be due to factors such as timing, size, or price of orders.

By employing Shapley values and synergy in order analysis, we can quantify the contribution of each order and order combinations to market dynamics. This can illuminate the underlying mechanisms driving price formation and liquidity fluctuations, providing a more nuanced understanding of market behavior. This analysis can be critical for market participants, ranging from individual traders to regulatory bodies, aiding in decision-making, strategy formulation, and market oversight.

4.2 Model for prediction

In this study, the order books of Microsoft and Yahoo have been chosen for analysis, primarily due to their significant influence and benchmark status in the technology sector. As key players in the industry, their trading activities often mirror larger market trends and can provide valuable insights into broader market dynamics.

The selected parameters for this analysis are bid-price, bid-volume, ask-price, and ask-volume for both companies, yielding a total of 8 parameters. These parameters provide a comprehensive view of the trading landscape, encapsulating both sides of the market and enabling a detailed analysis of market behavior.

Data for training was gathered on the first week of January 2008 from the 7th to the 11th, between 1 pm and 9 pm and the test was made on the data the week after from the 14th to the 18th. The year 2008 was characterized by significant market volatility, thus showing a rich and informative dataset for this study.

The data in our set was not synchronised. We synchronised on the basis of Yahoo's bid price because our prediction will be based on this parameter. To accommodate the asynchronous changes of the remaining seven parameters, their latest value was retained each time the Yahoo bid-price changed. This approach ensures that we maintain a comprehensive dataset for each instance of Yahoo bid-price change. The log-difference of the parameters was calculated between two consecutive time steps. This was done in order to study the variation in the parameters on the date before the one we wish to predict. This also makes it possible to stabilise the variance and avoid outliers. Finally, we seek to predict whether Yahoo's bid price is increasing or decreasing.

This approach enables us to obtain a prediction of the variation in Yahoo's bid price, and our objective will be to study the relationships between the parameters involved in this study.

In order to predict whether the bid-price of Yahoo increases or decreases, the XGBoost method was employed. The XGBoost model was trained using the **'XGBClassifier'** function, which is specifically designed for classification problems. In this case, we have a binary classification problem, as we aim to predict whether the bid-price will increase or decrease. To ensure the model's performance and accuracy, parameter tuning was carried out through grid search.

Through this approach, we were able to train a highly accurate XGBoost model that eventually predicts whether the bid-price of Yahoo increases or decreases.

4.3 ROC curve and AUC

The performance of our predictive model was evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). These are metrics for binary classification problems and provide an evaluation of the model's performance.

In this case, the model achieved an AUC of 0.84. While not perfect, this score indicates a reasonable level of predictive power. It suggests that the model can distinguish between price increase and decrease scenarios with a fair degree of accuracy. The ROC curve for this model is shown on Fig.5.

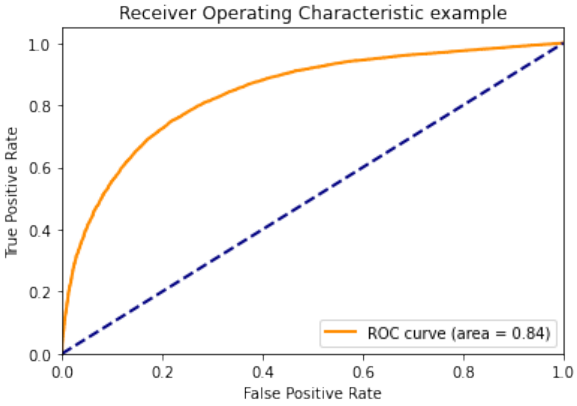

Figure 5: ROC curve and AUC

An AUC of 0.84 and the shape of the ROC curve indicate that the model can reasonably be used to predict whether Yahoo's bid-price will rise or fall. By setting the parameter of XGBoost `max_depth` to 1, all the synergies are removed, which results in a loss of predictive information. In this case, the AUC is lower (0.82).

4.4 Results

The synergies obtained on the test dataset are displayed in Fig.6. The analytical and bootstrap p-values allow to filter really significant ones, and give the same results. Here the threshold of significance is set to $s = 0.05$. Because of the large number of tests performed (56, one for each pair of different features), the probability to observe extreme events increases. In order to compensate this phenomenon, Bonferroni correction is applied by testing each individual hypothesis at $s = 0.05/56 \approx 8.9 \times 10^{-4}$. This allow to derive more consistent results.

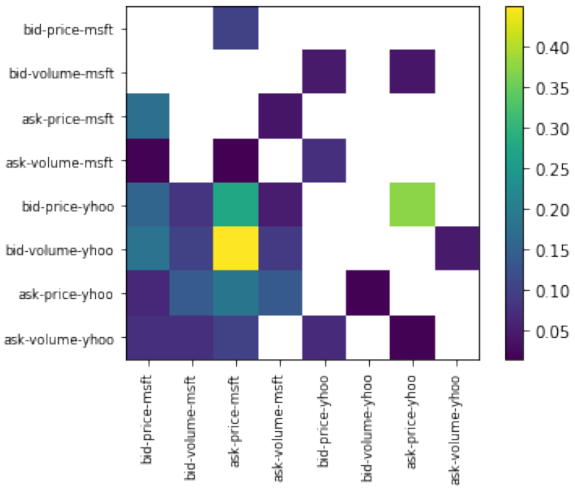

Figure 6: Synergy of each pair of features (a blank box indicates that there is no significant synergy). Each feature is logarithmically differenciated.

These synergies provide information on the extent to which the predictive power of some features depends on the presence of other features. A first global observation in this case is that the bottom

left-hand quarter contains almost exclusively significant synergies (Fig.6). In other words, the features associated to the Microsoft order book need information about the Yahoo order book for maximum impact on the prediction of the bid price increase of the Yahoo order book.

It should also be noted that a large part of the predictive power of the ask return of Yahoo lies in the presence of the bid return of Yahoo, which suggests that the evolution of the ask price should always be studied in the light of the evolution of the bid price to better predict the latter.

However, the bid volume of Yahoo which is the most important feature in the prediction (in terms of SHAP value) as it gives direction to the market does not need any further feature to increase its contribution to the prediction.

By repeating the process on the next week, we obtained similar significant synergy. Especially, by performing a Fisher test, the hypothesis of no contingency between both distributions of significant synergies among all the pair of features is rejected with a p-value of 0.0006. This allows us to be confident that the results obtained are not the result of chance.

5 Conclusion

This article shows the effectiveness of our method in investigating synergies between features in machine learning models. The analytical method presented provide reliable means to quantify and assess the significance of these synergies. The application of these methods to order book highlights the importance of some features in price predictions. Overall, this research contributes to advancing the explainability of machine learning algorithms and offers valuable insights in order books forecasting.

References

[1] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data?, 2020.

[2] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

[3] Jan Ittner, Lukasz Bolikowski, Konstantin Hemker, and Ricardo Kennedy. Feature synergy, redundancy, and independence in global model explanations using shap vector decomposition. *arXiv preprint arXiv:2107.12436*, 2021.

[4] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

[5] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[7] Lloyd S Shapley et al. A value for n-person games. 1953.