Optimising Hierarchical Demand Forecasting with Explainable AI: Insights into Key Drivers

Mátyás Kuti-Kreszács

Babeş-Bolyai University, Cluj-Napoca

Abstract

Demand forecasting is a prediction problem that aims to estimate future needs based on historical data. It serves as the basis for optimal decision making in multiple areas of value chains such as manufacturing, logistics, and retail. It is particularly important in demand forecasting models where demand drivers like price, promotions, and resource planning can help companies optimise pricing, promotional activities, resource planning, and inventory planning. Our goal is to identify applicable feature importance techniques to hierarchical forecasting problems by providing insights into feature importance and the underlying decision-making process and helping to understand the model's reasoning. We propose applying SHAP values to a forecasting model while using part of a real-world dataset. The results will provide insight into the key drivers of the forecast and help to understand the impact of the features on the decisions made by the model.

1. Introduction

Demand forecasting became really important for businesses and serves as the basis for optimal decision making in multiple areas in value chains such as manufacturing, logistics, and retail. By having multiple products, manufacturing locations, sales channels, and geographical regions, demand forecasting can be complex and hierarchical in nature.

However, the problem can be formulated as a regression problem, with the aim of predicting the future demand based on historical data. This regression problem can be solved using machine learning models such as random forests, gradient boosting, and neural networks. Unfortunately, these models are considered black boxes and their predictions are hard to interpret. This is where explainable AI (XAI) techniques come into play, providing insights into the model's decision-making process and helping to understand the underlying rules and reasoning behind the predictions.

One of the most fundamental methods for understanding a model's reasoning is feature importance or attribution, which allows identifying key contributor factors to the model's predictions. This is especially important in demand forecasting models, where demand drivers, such as price, promotions, weather, holidays, and economic indicators, can influence demand. Understanding these drivers can help companies optimise pricing, promotional activities, resource planning, and inventory management.

Our goal is to identify applicable feature importance techniques to demand forecasting models, aiming to discover key features contributing to the decisions and explain the model's reasoning at different levels. The significance of our research is to improve the reasoning and transparency of multiseries and hierarchical demand forecasting models by providing insights into feature importance and the underlying rules at various levels. The methods employed are expected to be used not only in demand forecasting, but also in other grouped and hierarchical forecasting problems in different domains.

1.1. Research Questions

The gap identified in the literature is the lack of studies that apply feature importance techniques to multiseries models for hierarchical demand forecasting problems and analyse the underlying decision

RuleML+RR'24: Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning, September 16–22, 2024, Bucharest, Romania

matyas.kuti@ubbcluj.ro (M. Kuti-Kreszács)
https://www.linkedin.com/in/kkmatyas/ (M. Kuti-Kreszács)
0009-0004-4997-2000 (M. Kuti-Kreszács)

drivers at different levels. Other studies focused on the representation of the explanation for sales forecasting models, but not on the explanation methods themselves[1]. Our research questions are:

- **RQ1:** Can existing feature importance techniques be applied to multi-series and hierarchical models to identify key features and explain the underlying decision factors?
- **RQ2:** How feature importance can be translated to different hierarchical levels?
- **RQ3:** How do these methods perform when applied to real-world datasets?
- **RQ4:** How can the results be visualised and interpreted?
- **RQ5:** What methods are most effective in this context?

In our current work, we partially address RQ1 and RQ2 by proposing a method to apply SHAP values to a LightGBM model used for forecasting hierarchical time-series data. Furthermore, we make progress on RQ3 using part of a real-world dataset; however, evaluation is still pending. Last but not least, we address RQ4 by visualising the results in a way that can be interpreted by the user. RQ5 is still open and will be addressed in future work.

2. Literature review

2.1. Demand Forecasting with machine learning

Demand forecasting is a prediction problem that aims to estimate future needs based on historical data. Statistical forecasting methods such as ARIMA[2, 3] and exponential smoothing [3] have been widely used in demand forecasting. However, they have limitations in intermittent multi-series and hierarchical forecasting, where machine learning models have shown better performance[4]. An important aspect also is that there may be multiple exogenous variables so-called demand drivers[5] that can influence the demand. Internal factors such as price, promotions, and external factors like weather, holidays, and economic indicators can be considered as demand drivers. These can be used as features in machine learning models to improve forecast accuracy.

Machine learning models such as tree ensembles and neural networks have been successfully applied to demand forecasting tasks[4]. Ensemble models in general can be homogeneous with individual models of the same type or heterogeneous with models of different types. We considered only homogeneous ensemble tree models because of the applicability of some model-specific explanation methods. To build tree ensembles, bagging methods such as random forest[6] can be used, which trains multiple decision trees on different subsets of the data, and the final prediction is the average of the predictions of the individual models. In addition, boosting methods such as Gradient Boosting Machines (GBM) [7], XGBoost [8], and LightGBM [9], which train models sequentially on the residuals of the previous model, in this case using the sum of individual predictions. In a notable forecasting competition [10], a LightGBM model was the winner and secured four of the top five positions.

2.2. Forecasting techniques

Forecasting techniques can be divided into single-series or multi-series forecasting from the perspective of the model's input. Single-series forecasting refers to the prediction of a single time series, while multiseries forecasting involves the prediction of multiple time series, with the same global model[11]. These series can be related to each other, such as sales of different products, or they can be independent, such as sales in different regions; therefore, it is important to consider the hierarchical structure of the data.

Hierarchical forecasting refers to the prediction of multiple time series that are related to each other in a hierarchical structure[12]. It can be tackled with different single-level approaches, such as bottom-up, top-down, or middle-out[12]. The top-down approach would involve a single series model for the total demand and then disaggregating it to the lower levels. The middle-out and bottom-up approach would involve a multiseries model. Grouped time-series forecasting is a special case of hierarchical forecasting, where the series are aggregated based on attributes such as product type, region, or sales channel.

[5] suggests three major hierarchies in demand forecasting: product hierarchy, geographical hierarchy, and time hierarchy. The product hierarchy refers to the categorisation of products according to their attributes, such as product type, brand, or category. The geographic hierarchy involves the division of sales regions based on geographic attributes, such as country, state, or city down to the point of sale. Time hierarchy refers to the temporal structure of the data, such as year, month, week, day, and hour.

2.3. Feature importance

Feature importance (FI) or feature attribution is considered an interpretation method resulting in a summary statistic that assigns a score to each input feature [13]. Depending on their scope, the FI methods can be global or local [14, 13]. The global feature importance (GFI) or model feature attribution methods explain the contribution of features to overall predictions, while the local FI quantifies feature contributions to specific predictions [13]. Although related, GFI methods differ from feature selection, which identifies irrelevant features before training. GFI methods can be model-specific, which are limited to specific model types, while model-agnostic ones are applicable independent of the model type[13]. Another categorisation of FI methods is given by how it is calculated, in which case the importance can be based on the model's structure, while the other approach relies on a dataset.

Among the model-agnostic methods, one of the most common is permutation feature importance (PFI) which was proposed to measure FI in random forests[15]. It is a model-agnostic, data-dependent method that measures the decrease in the model's performance when the features are permuted. The PFI can be calculated using different metrics such as the mean squared error (MSE), the mean absolute error (MAE), or the coefficient of determination (R^2). PFI also has limitations, as it is sensitive to over- and underfitting[16], in which case the FI differs on training and test data, so the use of both datasets can be beneficial. In addition, another flaw of the PFI method is that it can generate cases in which the model does not have training data[17, 18], but other methods were proposed to overcome this[19, 20].

SHAP(SHapley Additive exPlanation)[21] values contribute local explanation for individual predictions, but aggregates of it are useful to assess the importance of global features. For example, the mean absolute SHAP values quantify the importance of the feature regardless of the direction of the impact on the prediction. There are different algorithms for approximation from which Kernel SHAP[21] is one that is model-agnostic. TShap [22] is a method for estimating SHAP values for time series data, but it uses a surrogate model, so it gives the FI of the surrogate. Another related method is SAGE *(Shapley additive global importance)* [19], which estimates the contribution of each feature to the model's performance.

Tree specific GFI methods are gain-based importance values which were already introduced with decision trees [23] It measures of the reduction in mean average error(MAE) made by the decisions based on the respective feature. Another measure is the split-based importance[8] refers to the number of decisions made by the model based on a feature. The previously presented SHAP also has a tree model-specific solution for approximation, called TreeSHAP [24]

2.4. Explainability in forecasting

The number of publications on forecasting explainability is limited. [1] tackled the presentation of explanations for sales forecasting models, but not the explanation methods themselves. [25] used SHAP values to explain the prediction of a time series model but on local level and not global level. Skforecast [11] library extracts model specific global feature importance from tree ensemble models. The work is focused on either global feature importance or local feature contribution without considering the multi-series and hierarchical structure of the data.

2.5. Feature importance as a basis for model reasoning

Feature importance methods can provide insight into the model's decision-making process and help to understand the underlying rules and reasoning behind the predictions. By including demand drivers as features in the model, the feature importance methods can help to identify the key drivers of

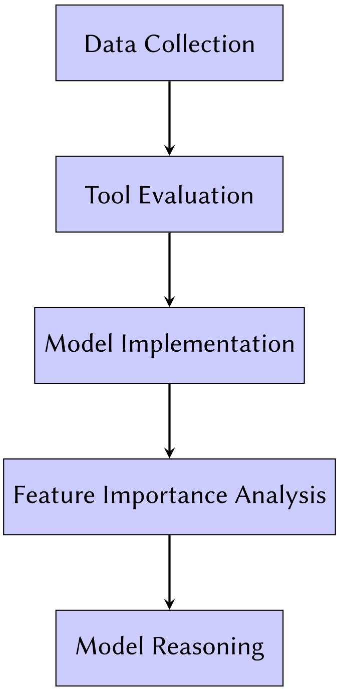


Figure 1: Research design

demand. For external factors such as weather, holidays, and economic indicators, the importance of the characteristics can help to understand their impact on demand. Through internal factors like price, promotions, the feature importance can help to understand post-promotion effects and the impact of price changes on the demand[5]. Knowing the influence of internal factors can help to optimize pricing strategies and promotional activities. However, causation and correlation are different concepts, and the feature importance methods can only provide correlation; therefore, the identified key features should be further analyzed to understand the causation[15]

3. Methodology

Our preliminary research focusses on the methodological aspects of applying feature importance techniques to hierarchical forecasting models. This includes adaptation of existing methods, but also tool development to support the analysis of hierarchical forecasting models. Later we plan to conduct an empirical study to evaluate the methods on real-world datasets.

Our initial research design1 includes the following steps:

- Data collection: identify datasets with hierarchical time series data describing sales/demand for multiple product categories and regions with exogenous variables.
- Tool evaluation: assess the applicability of existing libraries for hierarchical forecasting and XAI techniques.
- Model implementation: we build global models that consider multiple series and exogenous variables.
- Feature importance analysis: We apply model attribution methods and aggregation and decomposition techniques to identify key features and analyse their impact on the forecast.
- Model reasoning: analyse the feature contributions to forecast and identify underlying rules on different levels of the hierarchy.

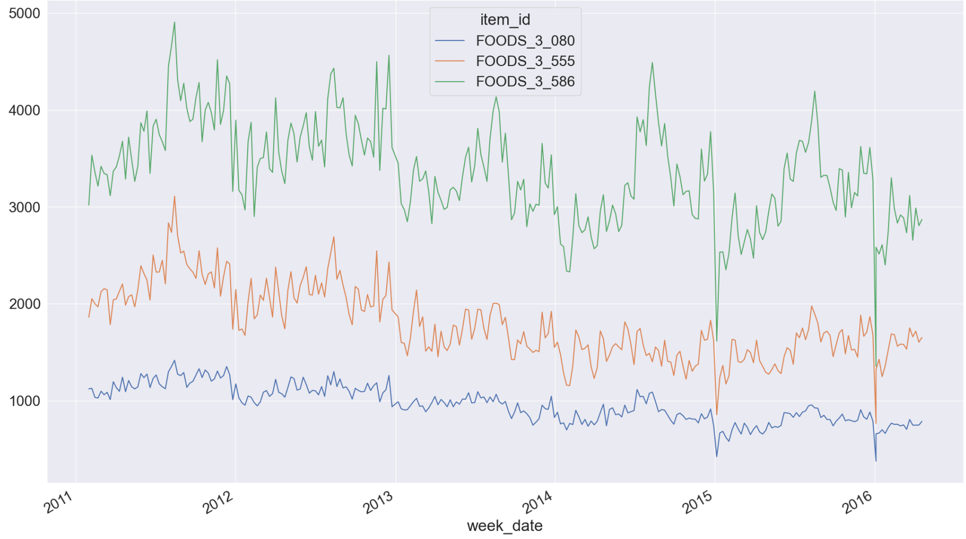


Figure 2: Total weekly sales for the chosen products

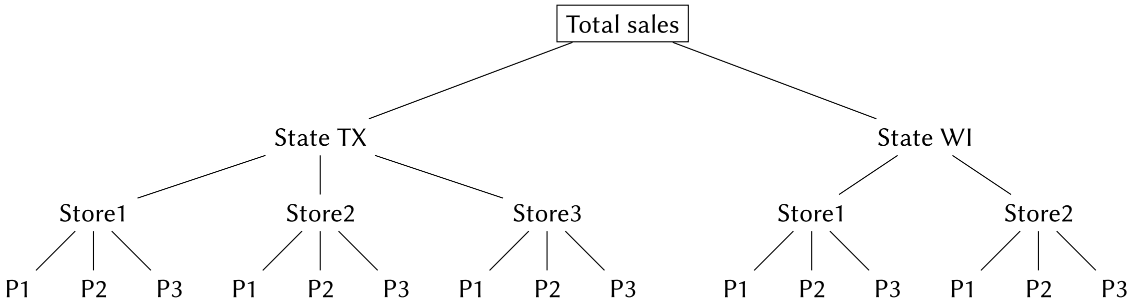


Figure 3: Hierarchical structure of product sales data (P1-3 = Product1-3)

3.1. Data collection and preprocessing

To model the hierarchical impact of features on forecasting, we must use datasets with multiple series and exogenous variables that represent demand drivers. There are multiple open sales data sets available; however, there are just a few, such as the M5 competition[10] and the Kaggle datasets[26]. For our initial exploration, we sampled M5 competition[10] dataset, which includes sales data for multiple product categories and regions. The dataset contains daily sales information for 3049 products in 10 stores over 5 years. For our analysis, we identified three products that have similar sales patterns and are sold in two states and five stores. As products are from the same category and department, the hierarchy at the product level was not considered. The reason for this filtering is to reduce the complexity of the model and to focus on the feature importance analysis. The selected products are FOODS_3_586, FOODS_3_080, and FOODS_3_555 and are sold in three states of Texas (TX), Wisconsin (WI). The total sales data for these products are shown in Figure 2. Our hierarchical structure is shown in Figure 3. It should be mentioned that the hierarchical structure can be inverted, meaning that the products can be at the top level and the stores at the bottom level, so technically our data set is grouped time series data.

Data preprocessing two main parts: preparing the sales data and the exogenous variables. Sales data were aggregated at the weekly level. The weeks at the beginning and end of the data set were removed to have a consistent time period. As features, lagged sales data was included to capture the temporal dependencies. The exogenous variables were related to pricing and calendar events. The selling price was already aggregated at the weekly level for each store and product. Calendar events included whether a day was a holiday, had special events, and if it was a SNAP (Supplemental Nutrition Assistance Programme) day in a respective state. To include these variables in some way, they were

Feature name	Type	Description
week_date	date	Starting date of week used for aggregation
week	int	Week number of the year
series_id	string	Unique identifier for the series made of state+store+product combination, representing the hierarchy
sales	float	Total sales for the week
lag_*n*	float	Sales from the previous *n* weeks

Table 1
Sales data structure

Feature name	Type	Description
week_date	date	Starting date of week used for aggregation
week_of_year	int	Week number of the year
series_id	string	Unique identifier for the series made of state+store+product combination, representing the hierarchy
sell_price	float	Selling price for the week for
num_of_events	int	The number of special events and holidays in the week
snap_days	int	Number of SNAP days in the week

Table 2
Exogenous variables

Parameter	Search space	Description
n_estimators	50-1000	Number of boosting iterations
max_depth	5-50	Maximum depth of the tree
min_samples_leaf	1-10	Minimum number of samples required to be at a leaf node
num_lagged_sales	4-52	Number of lagged sales records used as features

Table 3
Model hyperparameters search space

counted for each week and state. In addition, the week of the year was included as a feature to capture seasonality. The data were split into training and test sets, and the last complete year(2015) was used to test the model. The structure of the data sales data is shown in Table 1 and the exogenous variables in Table 2.

3.2. Model implementation

The modelling approach is to build a single global on all series and exogenous variables for bottom-up aggregation. For creating forecast models, the skforecast[27] library was used. The base model for hierarchical forecasting was LightGBM[9] which we chose because of its efficiency and also because of its widespread usage in the M5 competition in this data set[10]. Other ensemble models such as Random Forest or Gradient Boosting Machines could be used as well. Other reasons for choosing LightGBM are that it can handle categorical variables without the need for one-hot encoding, and that it supports model-specific split and gain-based global feature importance methods.

Hyperparameter tuning was performed using the Optuna library[28], by Bayesian optimisation. The search space3 was defined for the parameters of the LightGBM model, including the number of predictors, the minimum number of samples in the leaf, and the maximum depth of the tree. In addition, the number of lagged sales records used as features was included in the search space. For the search, the data was split into training and validation sets, the last year being the validation set used for backtesting. The performance of the model was evaluated as a mean square error (MSE) in the validation set for each configuration. The best configuration found was with 239 estimators and a maximum depth of 26 with a backtesting MSE 4263.01 The lagged sales records used as features were 1, 4, 5, 13, and 52 weeks.

The feature input for the final model is a table with the following columns:

- week_of_year represented as numerical values (1-52)
- sell_price for the week for the product in the store
- num_of_events for the week
- snap_days for the week in the state
- lag_*n* for n in [1, 4, 5, 13, 52] representing the sales from the previous weeks
- series_id noted as (_level_skforecast) encoded as a numerical value representing the series hierarchy

Series_id could have been encoded as a one-hot encoded vector or as a categorical variable given it is supported by LightGBM. One-hot encoded vector would have increased the number of features and the complexity of the model, while with the categorical variable

3.3. Feature importance analysis and model reasoning

Two initial ideas were considered to analyse the importance of characteristics. The first involves using the mean SHAP values for cohorts representing different levels of the hierarchy, providing information on the contribution of features throughout the structure. The second approach is based on conditional permutation importance, which evaluates the importance of features while accounting for the hierarchical structure on the idea of subgroup-based permutation importance[29]. The first method was prioritised for implementation due to the availability of support in the SHAP library[30]. Given an instance x for prediction, the SHAP value of the feature i is $\phi_i(x)$ Each x is part of a cohort C_k based on the series hierarchy k. The contribution value or importance of feature i for a C_k cohort is calculated as

$$\phi_i(C_k) = \frac{1}{|C_k|} \sum_{x \in C_k} |\phi_i(x)| \tag{1}$$

where $|C_k|$ is the cardinality of C_k and $|\phi_i(x)|$ is the absolute SHAP value of feature, i for instance x.

Steps for the feature importance analysis:

- For each prediction instance x and feature i calculate SHAP value $\phi_i(x)$.
- Split the instances into cohorts according to the hierarchy levels.
- Calculate the mean SHAP values for each cohort C
- Visualize the mean SHAP values for C and summary plots

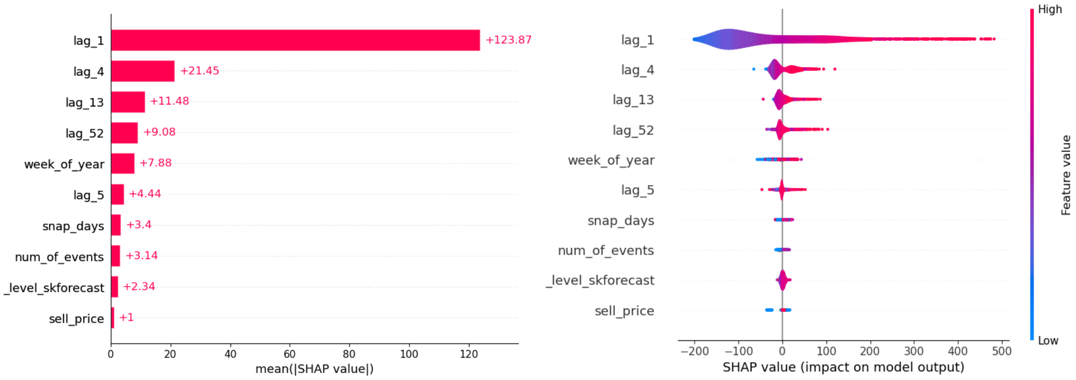
The reasoning of the model is based on the analysis of the contributions of the features to the forecast. The aim is to identify the underlying rules and patterns that the model uses to make predictions. The SHAP values provide a way to understand the impact of the features on the forecast. The analysis can be done at different levels of the hierarchy, from the global model to the state and store levels.

4. Preliminary Results

The preliminary results focus mainly on the practical application of SHAP values in hierarchical forecasting models rather than on the theoretical aspects of feature importance. As preliminary results, we present mean average SHAP values at different aggregation levels. These provide an overview of the main contributors to the forecast at different levels of the hierarchy.

Furthermore, we visualise the distribution of SHAP values at different aggregation levels using violin plots, which provide a representation of variability and density of SHAP values for each feature across the hierarchy. This double representation allows for a more detailed analysis of the contribution of features to the forecast, for example, if a feature contributes positively or negatively to the forecast. In the following, we present several cases of different aggregation levels, starting from the global level to the store and product level, but it is not meant to be exhaustive.

At the global level, the SHAP values4a show that the most important features are the lagged sales values, especially the sales value of lag 1, which has the highest SHAP value. This is expected as prior sales are the most important factor in predicting future sales. The violin plot in Figure 4b shows the distribution of SHAP values for each feature. It shows that the actual impact of the lag value is most of the time negative, as after a week with higher sales, demand the following week can drop.



(a) Global SHAP values

(b) Global SHAP violin plot

Figure 4: Global summary

In case of state-level grouping, the number of samples differs for the two groups, one of them having only two stores included. This can be observed in the wider distribution on the violin plot of the Texas(TX) state.

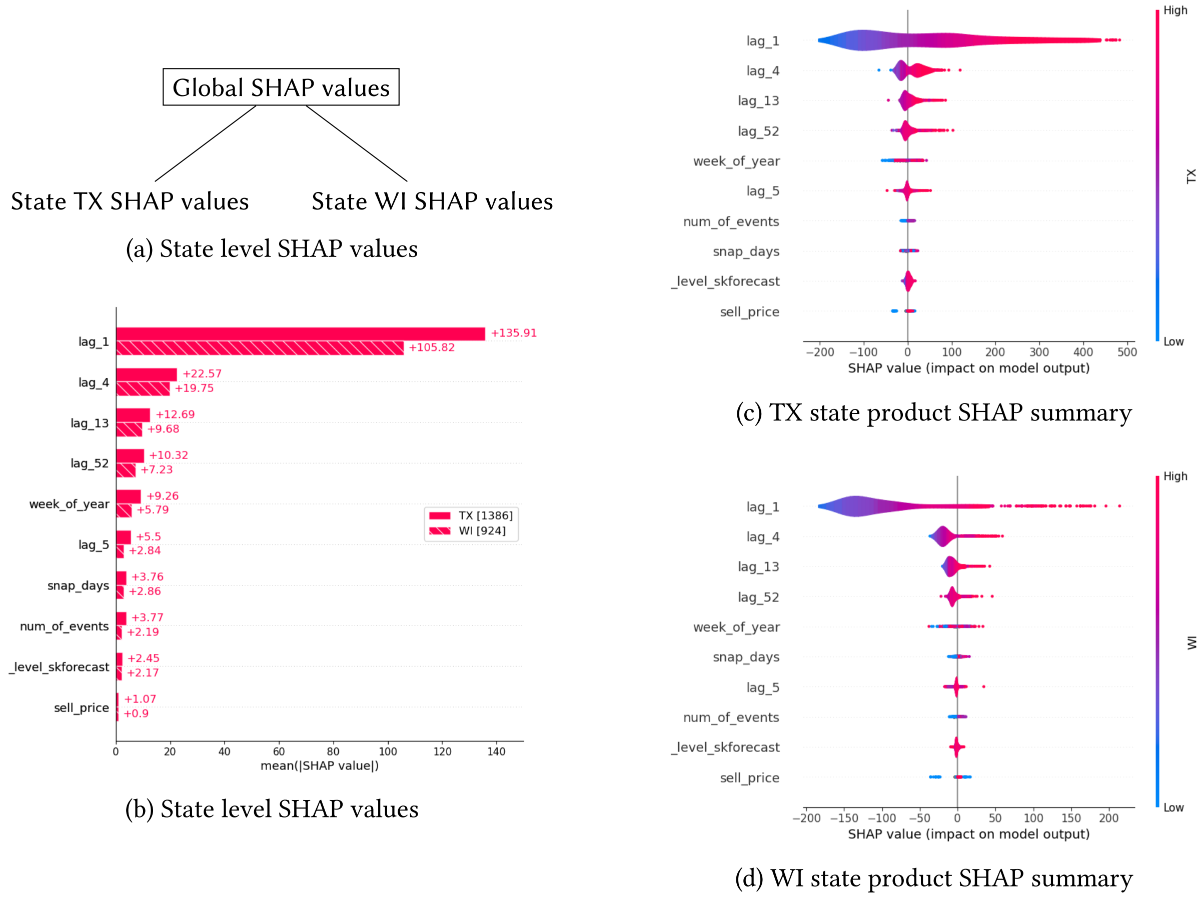
On the lowest level of the hierarchy presented in Figure 7, deviations can be revealed in the order of importance of the features. For example, in Figure 7d the week-of-year feature has a greater impact on the forecast than some of the lag values that occurred in other cases in Figure 7c. This can be due to the fact that the store TX_2 has a different seasonality pattern or might have recurring special events, since the number of events is also a feature with higher impact on this store. What is problematic in this case is that, due to the large number of series, representation of the mean absolute SHAP value is hardly comprehensible in the previous form of the bar plot. As a workaround, the grouping of feature contribution of each group is presented in Figure 7b. What can be misleading in this case is the lack of order by impact and a different scale of the x axis for each feature.

5. Discussion

In this work, we managed to calculate the Shapley values for the predictions of a hierarchical forecasting model with some limitations, while we also aggregated these values to different levels of the hierarchy. By this we addressed the first two research questions. We used a sample of a real-world dataset to evaluate the proposed method working towards the third research question. We visualised the SHAP values at different levels of the hierarchy and provided some interpretation of the results. To respond to the fourth research question, we plan to expand the literature review to include a wider range of XAI techniques.

This research is expected to contribute in several key areas. First, it will provide an evaluation of feature importance methods in the context of hierarchical forecasting models. This will help to identify what methods are most effective and how they can be applied to improve model interpretability. Second, it aims to provide guidelines, best practices, and limitations of effectively explaining these models. Finally, the research will support the development of tools that improve the understanding of hierarchical forecasting models and their underlying rules and reasoning.

The limitations of our study include the handling of categorical variables in the SHAP library. The effect of ordinal encoding that induces an order on the categorical variables may not be appropriate for all cases. Low feature importance for the categorical variables may be due to the encoding method. Recent research [20] proposes a method to handle categorical variables for conditional feature importance. With regard to data limitations, the dataset used is simplified in multiple dimensions. First, with aggregation



(a) State level SHAP values

(b) State level SHAP values

(c) TX state product SHAP summary

(d) WI state product SHAP summary

of sales data at the weekly level, multiple exogenous variables such as special events could not be included. Second, the dataset is limited to a single product category, which may not be representative of all hierarchical forecast scenarios. Lastly, the input data was limited to the sales lag of the product without considering other products in the same category. The independence of products in the same category may not be a realistic assumption.

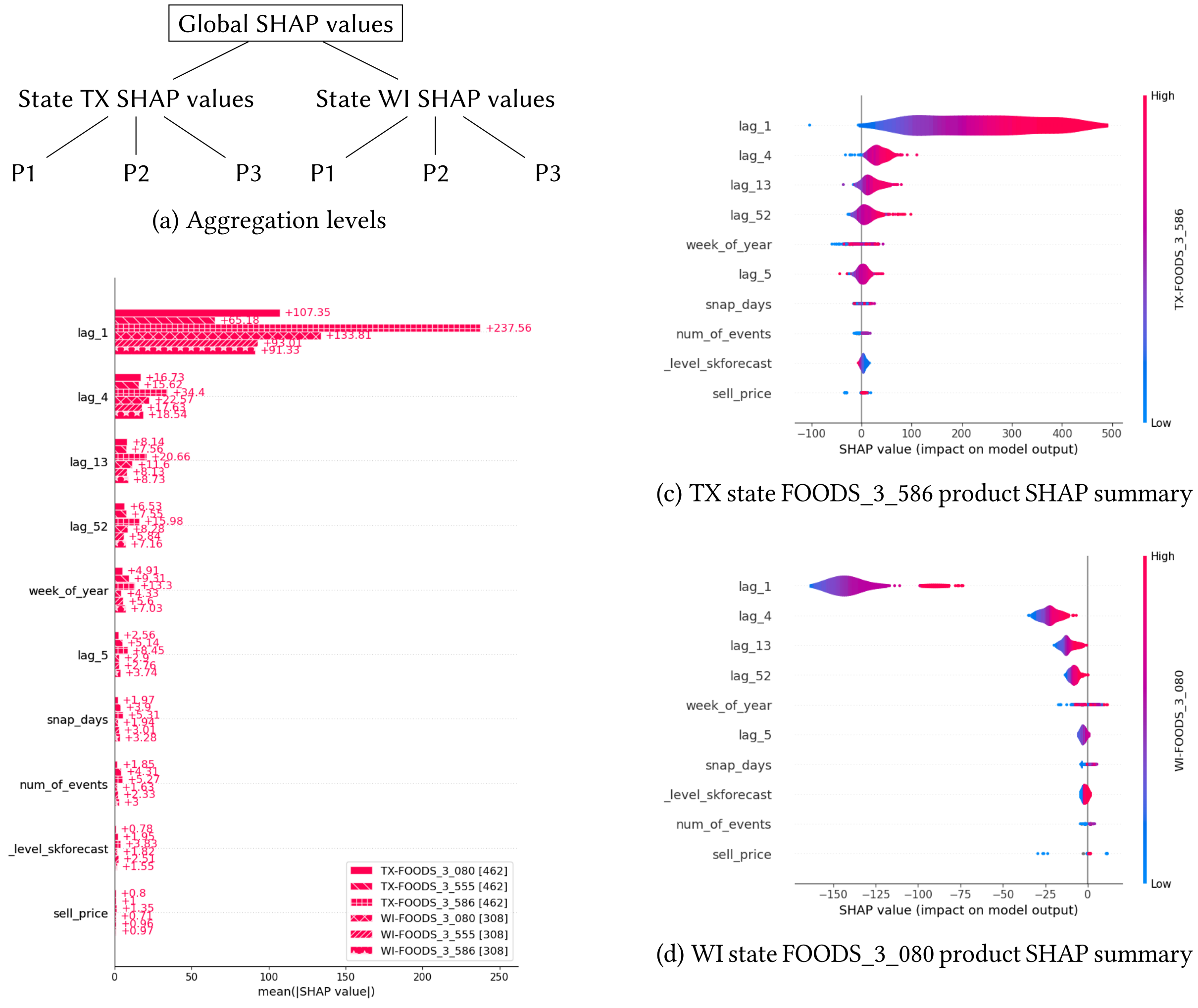
During the implementation of our initial approach, we encountered several challenges. One of the main issues was the lack of appropriate tools. For example, the SHAP library does not support categorical features in the current version. In addition,we faced difficulties in visualising the results; although the SHAP library offers integrated graphing functions, these have not been effectively used to deal with a large number of cohors, leading to errors and incomplete plots. Although these issues are not straightforward to solve, they are a sign of unexplored areas in the field of XAI and hierarchical forecasting.

There are also potential risks that could impact research in addition to challenges. One of the main risks is the availability of data, especially real-world datasets that include exogenous variables or demand drivers. Synthetic datasets can be used as an alternative, but they may not capture the complexity of real-world scenarios. The evaluation of methods is another potential risk, as it may be difficult to assess the performance of the explanation methods. In case of application grounded evaluation, it may be difficult to find experts in the field who can provide meaningful feedback given that each product category may require different domain knowledge[31].

5.1. Future work

To address the challenges and limitations of the current research, several next steps are proposed for each part of the research.

- First, the literature review will be extended to include a broader range of XAI techniques. Given



(a) Aggregation levels

(b) State and product level SHAP values

(c) TX state FOODS_3_586 product SHAP summary

(d) WI state FOODS_3_080 product SHAP summary

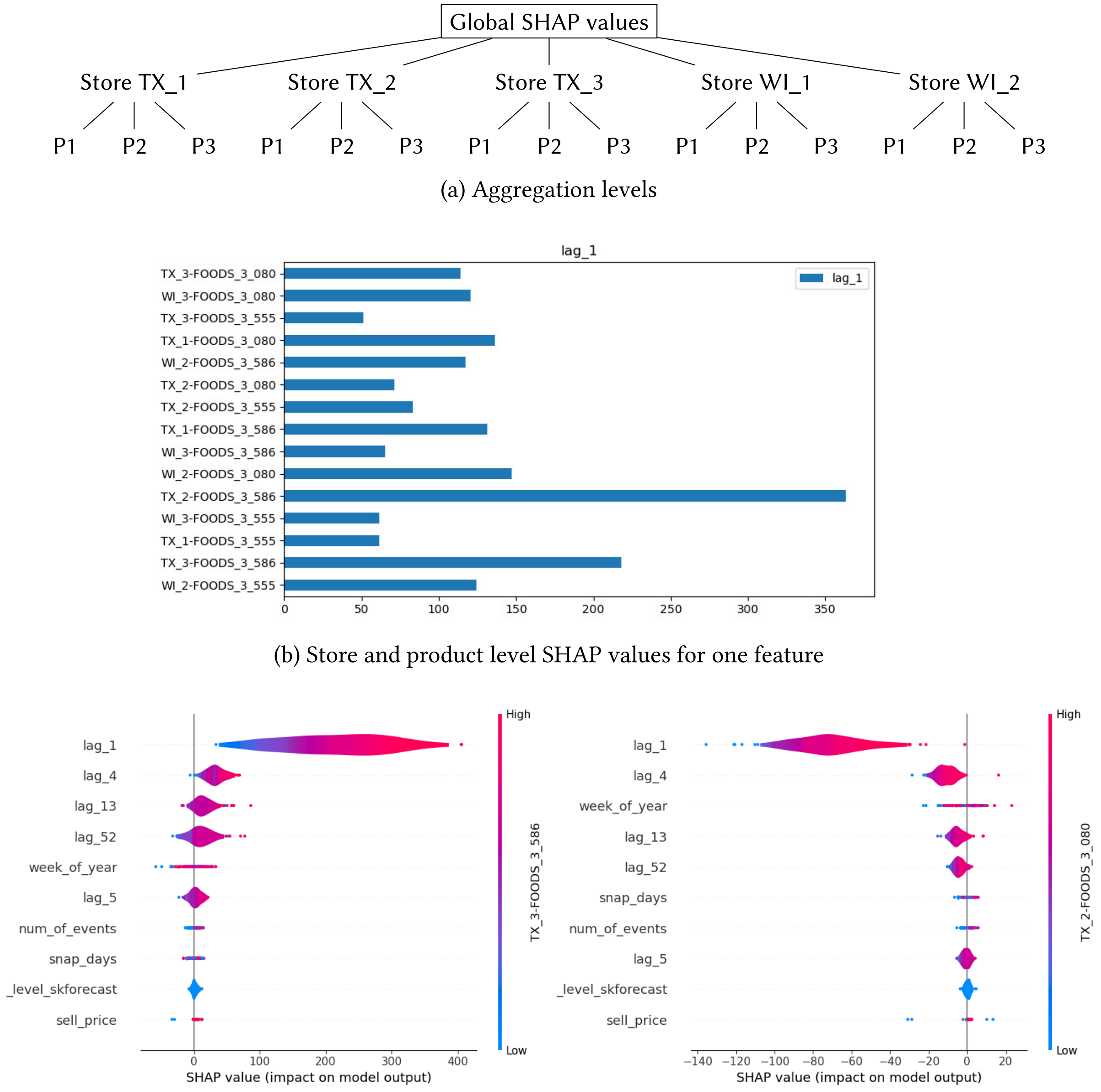
Figure 6: State and product level summary

the current context, we focus on feature importance-based evaluation, partial dependence plots, feature interaction, and other XAI techniques that should be included in the review.

- Data collection will be expanded to include synthetic datasets and additional real-world datasets. In addition, including more data from the actual dataset and forecasting on the day level can be a future direction.
- Evaluation and implementation of the tool for other methods will be needed. Conditional permutation importance can be also evaluated after implementing the method.
- The model implementation could be extended to include additional ML models for hierarchical forecasting. An additional enhancement to that would be dependent multiseries forecasting, as usually product sales are not independent of each other, especially in the same product category.
- Rule extraction based on feature importance and interaction can be a future direction.
- After covering the methodological aspects, an empirical study with evaluation in terms of accuracy and computational efficiency is planned.

5.2. Conclusion

Nowadays every organisation thrives in the direction of becoming data driven. In this context, data-driven decision making is crucial for optimising business processes to remain competitive. This effort is supported by the use of data mining, machine learning, and AI techniques. To avoid blindly trusting



(a) Aggregation levels

(b) Store and product level SHAP values for one feature

(c) Store TX_3 FOODS_3_586 product SHAP summary (d) Store TX_2 FOODS_3_080 product SHAP summary

Figure 7: Store and product level summary

ML models,it is crucial to understand the reasoning behind their decisions. Our goal is to demystify hierarchical forecasting models by applying XAI techniques.

This study explores the usage of SHAP values to explain the importance of features in hierarchical forecasting models. Our preliminary results focused on the practical aspects of aggregating SHAP values at different levels of hierarchy. This approach provides insights into the model's reasoning. We plan to extend this work by evaluating other XAI techniques to enhance the explainability of hierarchical forecasting models.

Acknowledgement

This work was done in collaboration with my Ph.D. supervisor, Laura Diosan, from Babes-Bolyai University. I am grateful for her continued support and encouragement throughout this research.

References

[1] T. Fahse, I. Blohm, R. Hruby, B. van Giffen, Explanation interfaces for sales forecasting, in: ECIS 2022 Research-in-Progress Papers, 2022.

[2] J. Fattah, L. Ezzine, Z. Aman, H. E. Moussami, A. Lachhab, Forecasting of demand using arima model, International Journal of Engineering Business Management 10 (2018) 1847979018808673. URL: https://doi.org/10.1177/1847979018808673. doi:10.1177/1847979018808673. arXiv:https://doi.org/10.1177/1847979018808673.

[3] C. Ingle, D. Bakliwal, J. Jain, P. Singh, P. Kale, V. Chhajed, Demand forecasting: Literature review on various methodologies, in: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2021, pp. 1–7.

[4] E. Spiliotis, S. Makridakis, A.-A. Semenoglou, V. Assimakopoulos, Comparison of statistical and machine learning methods for daily sku demand forecasting, Operational Research 22 (2022) 3037–3061.

[5] N. Vandeput, Demand forecasting best practices, Manning, 2023. URL: https://books.google.ro/books?id=C_u8EAAAQBAJ.

[6] Leo Breiman, L. Breiman, Random Forests, Machine-mediated learning 45 (2001) 5–32. doi:10.1023/a:1010933404324, mAG ID: 2911964244 S2ID: 8e0be569ea77b8cb29bb0e8b031887630fe7a96c.

[7] Jerome H. Friedman, J. H. Friedman, Jerome H. Friedman, Greedy function approximation: A gradient boosting machine., Annals of Statistics 29 (2001) 1189–1232. doi:10.1214/aos/1013203451, mAG ID: 1678356000 S2ID: 1679beddda3a183714d380e944fe6bf586c083cd.

[8] Tianqi Chen, T. Chen, Carlos Guestrin, C. Guestrin, XGBoost: A Scalable Tree Boosting System, arXiv: Learning (2016). doi:10.1145/2939672.2939785, mAG ID: 3102476541.

[9] Guolin Ke, Guolin Ke, G. Ke, Qi Meng, Q. Meng, Taifeng Wang, T. Wang, W. Chen, Wei Chen, Wei Chen, Wei Chen, W. Chen, W. Chen, Weidong Ma, W. Ma, Tie-Yan Liu, T.-Y. Liu, A Highly Efficient Gradient Boosting Decision Tree, Neural Information Processing Systems (2017) 3108–3116. MAG ID: 2753094203 S2ID: 497e4b08279d69513e4d2313a7fd9a55dfb73273.

[10] S. Makridakis, E. Spiliotis, V. Assimakopoulos, M5 accuracy competition: Results, findings, and conclusions, International Journal of Forecasting 38 (2022) 1346–1364. URL: https://www.sciencedirect.com/science/article/pii/S0169207021001874. doi:https://doi.org/10.1016/j.ijforecast.2021.11.013.

[11] J. A. Rodrigo, J. E. Ortiz, Global forecasting models: Dependent multi-series forecasting (multivariate forecasting), 2024. URL: https://skforecast.org/0.12.1/user_guides/dependent-multi-series-multivariate-forecasting.html.

[12] R. J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, 3 ed., OTexts, 2021. URL: https://otexts.com/fpp3/index.html.

[13] C. Molnar, Interpretable Machine Learning, 2 ed., online, 2022. URL: https://christophm.github.io/interpretable-ml-book.

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

[15] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. URL: https://doi.org/10.1023/A:1010933404324.

[16] C. Molnar, S. Gruber, P. Kopper, Limitations of interpretable machine learning methods, 2020. URL: https://slds-lmu.github.io/iml_methods_limitations/, mAG ID: 3041627266.

[17] C. Molnar, C. Molnar, G. König, G. König, J. Herbinger, J. Herbinger, T. Freiesleben, T. Freiesleben, S. Dandl, S. Dandl, C. A. Scholbeck, C. A. Scholbeck, G. Casalicchio, G. Casalicchio, M. Grosse-Wentrup, M. Grosse-Wentrup, B. Bischl, B. Bischl, Pitfalls to avoid when interpreting machine learning models, arXiv.org (2020). doi:null.

[18] Giles Hooker, G. Hooker, Giles Hooker, Lucas Mentch, L. Mentch, Siyu Zhou, S. Zhou, Unrestricted permutation forces extrapolation: variable importance requires at least one

more model, or there is no free variable importance, Statistics and Computing 31 (2021) 1–16. doi:10.1007/s11222-021-10057-z, aRXIV_ID: 1905.03151 MAG ID: 3209434414 S2ID: 29a4d6988a14ac694f4a73017705fe1506bcca92.

[19] Ian Covert, I. Covert, Scott Lundberg, S. Lundberg, Su-In Lee, S.-I. Lee, Understanding Global Feature Contributions With Additive Importance Measures, Neural Information Processing Systems 33 (2020) 17212–17223.

[20] Kristin Blesch, David S. Watson, Marvin N. Wright, Conditional feature importance for mixed data, AStA Advances in Statistical Analysis (2023). doi:10.1007/s10182-023-00477-9, aRXIV_ID: 2210.03047 MAG ID: 4367393752 S2ID: decb3af989963b43ce552f68e1481738ec6ed55d.

[21] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[22] Vikas C. Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi K. Sarpatwar, Giridhar Ganapavarapu, Roman Vaculín, TsSHAP: Robust model agnostic feature-based explainability for time series forecasting, arXiv.org (2023). doi:10.48550/arxiv.2303.12316, aRXIV_ID: 2303.12316.

[23] A. Gordon, L. Breiman, Jerome H. Friedman, R. A. Olshen, Charles J. Stone, classification and regression trees, Biometrics (1984). doi:10.2307/2530946.

[24] Scott Lundberg, S. Lundberg, Gabriel Erion, G. Erion, Hugh Chen, H. Chen, Alex J. DeGrave, A. J. DeGrave, Jordan M. Prutkin, J. M. Prutkin, Bala G. Nair, B. G. Nair, Ronit Katz, R. Katz, Jonathan Himmelfarb, J. Himmelfarb, J. T. Flynn, Nisha Bansal, N. Bansal, Su-In Lee, S.-I. Lee, From Local Explanations to Global Understanding with Explainable AI for Trees., Nature Machine Intelligence 2 (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[25] R. Saluja, A. Malhi, S. Knapič, K. Främling, C. Cavdar, Towards a rigorous evaluation of explainability for multivariate time series (2021) –. doi:10.2139/ssrn.4627337.

[26] M. M. Corporación Favorita inversion, Julia Elliott, Corporación favorita grocery sales forecasting, 2017. URL: https://kaggle.com/competitions/favorita-grocery-sales-forecasting.

[27] J. Amat Rodrigo, J. Escobar Ortiz, skforecast, 2024. URL: https://skforecast.org/. doi:10.5281/zenodo.8382788.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[29] C. Molnar, C. Molnar, G. König, G. König, B. Bischl, B. Bischl, G. Casalicchio, G. Casalicchio, Model-agnostic feature importance and effects with dependent features - a conditional subgroup approach., Data mining and knowledge discovery (2020). doi:10.1007/s10618-022-00901-9.

[30] Scott Lundberg, S. Lundberg, Gabriel Erion, G. Erion, Su-In Lee, S.-I. Lee, Consistent Individualized Feature Attribution for Tree Ensembles., arXiv: Learning (2018).

[31] F. Doshi-Velez, B. Kim, Considerations for evaluation and generalization in interpretable machine learning, The Springer Series on Challenges in Machine Learning (2018) 3–17. doi:10.1007/978-3-319-98131-4_1.