# РОЗДІЛ 10. МАТЕМАТИЧНІ МЕТОДИ, МОДЕЛІ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЕКОНОМІЦІ

## FOZZY GROUP HACK4RETAIL COMPETITION OVERVIEW: RESULTS, FINDINGS, AND CONCLUSIONS

## ОГЛЯД ЗМАГАНЬ HACK4RETAIL ВІД FOZZY GROUP: РЕЗУЛЬТАТИ, ЗНАХІДКИ ТА ВИСНОВКИ

УДК 339.3:519.7
DOI: https://doi.org/10.32843/infrastruct67-42

**Kosovan Oleksandr**
Ph.D. student
Ivan Franko National University of Lviv

*Competitions are a common approach to finding state-of-the-art solutions in different fields. This also applies to forecasting sales in retail. Hack4Retail is a competition to determine the best projects or technological solutions, and the aim of this work was to review the results and describe the most accurate methods. The competition extended empirical findings for retail sales forecasting, especially for the smaller firms, and provided the following conclusions: (1) prediction methods such as Light Gradient Boosting Machine and Feed-Forward Neural Networks were found to be effective for community competition as well; (2) blending and feature engineering based on explanatory variables improved performance of forecasting models; (3) the competition confirmed the importance of high-quality data for the final evaluation sample. Also, this paper described competition organization and explored key characteristics of data sets.*
**Key words:** *Time series, forecasting competitions, machine learning, retail, sales forecasting.*

*Змагання є поширеною практикою для знаходження новітніх підходів у різних галузях. Це також стосується прогнозування продаж у роздрібній торгівлі. Hack4Retail – це конкурс на визначення найкращих проектів чи технологічних рішень. Конкурс розширив емпіричні висновки для прогнозування роздрібних продажів, особливо для невеликих фірм. Метою даної роботи був огляд організації та результатів змагань. Також було зроблено аналіз структури даних, які використовувалися для моделювання та їх розвідувальний аналіз. Було підбито підсумки стосовно того, що методи прогнозування, такі як Light Gradient Boosting Machine і Feed-Forward нейронні мережі, виявилися ефективними у прогнозування продаж для великої кількості товарних одиниць. Також показали свою ефективність у покращенні продуктивності моделей прогнозування такі підходи: ротаційне оцінювання для оцінки генералізації моделі; змішування моделей, що є одним із видів ансамблевого навчання; збільшення кількості вхідних даних на основі пояснювальних змінних, таких як зміна динаміки цін на товари. Змагання показало важливість високоякісних даних, особливо для тестової та валідаційної вибірок. Крім того, витік даних підтвердив, що іноді існує розрив між вимогами до даних для моделювання та фактичним потоком збору даних. Було розглянуто задачу прогнозування враховуючи два аспекти – теоретичний та практичний. Теоретичною цінністю дослідження є розширення емпіричних даних у сфері прогнозування продаж. З практичної точки зору описано складність прийняття рішень за умов використання моделей, які важко інтерпретуються, та у випадку не врахування всіх чинників системи, таких як складські витрати. Одним із актуальних напрямків майбутніх досліджень виділено проблему прогнозування продаж під час кризових періодів, серед яких пандемія коронавірусної хвороби та повномасштабне вторгнення Росії в Україну.*
**Ключові слова:** *Часові ряди, змагання по прогнозуванню, машинне навчання, роздрібна торгівля, прогнозування продаж.*

**Problem Statement.** Hack4Retail is a community competition organized by LLC McKinsey and Company Ukraine – an international consulting company specializing in solving tasks related to strategic management and LLC Silpo-Food – one of the bigger retail companies in Ukraine.

Silpo is the leading supermarket chain in Fozzy Group's sales structure. The chain consists of 241 supermarkets in 60 cities in Ukraine. Silpo supermarkets are self-service stores with product ranges consisting of up to 20,000 items of food and related products, depending on the sales area of each store [1].

The Hack4Retail hackathon was designed to raise and expand its objectives in several directions, as follows:

– A data set of 1961 SKUs daily series approximately for a 5-year time range was used along with a benchmark. All forecasting methods were evaluated by a 2-weeks forecasting horizon.

– The competition's main goal was to create a forecasting application that predicts daily SKU sales of retail stores for different locations and product groups with high quality.

– The submissions were evaluated by objective metrics. The competition used mean absolute error (MAE) which is calculated as the sum of absolute errors (deviation of point sales forecasts) divided by the sample size. The forecast horizon was equal to 2 weeks.

The hackathon started on October 29, at 07:00 p.m. (Kyiv time), when the initial training set became available, and ended on October 31, 2021, at 07:00 p.m. (Kyiv time), when Analytics Vidhya announced the final leaderboard. Context rules, prizes, and more information were available

on Analytics Vidhya and the hackathon contest website [2].

Also, the Hack4Retail was a completely open competition, encouraging the participation of both academics and practitioners in the field and ensuring fairness and objectivity, and emphasizing that each team was free to use its own method. It was also important to disseminate information about the approaches used and their results, to share information, and improve further developments. The result of the public discussion was 18 discussion topics and 42 public notebooks with a description of the methods.

**Literature review.** Forecasting competitions are a viable solution for evaluating current methods and finding state-of-the-art approaches. Hyndman, R. J. described that competitions provide empirical evidence and help improve forecasting theory and practice [3]. The Makridakis Competitions (*aka the M Competitions*) are a good example of a high impact on forecasting theory. Makridakis S., Spiliotis E., and Assimakopoulos V. researched M competitions and concluded that they shared new data and tasks from different domains [4].

M5 competition focused on a retail sales forecasting application with a few evaluation stages based on Walmart data. Makridakis et al. believe that more research is needed to generalize the findings of M5 like research for smaller retail firms, companies that operate outside the USA, retail e-commerce firms, etc. [4]. This argument is a motivation to review the community competition for the local retail company.

One more example of retail competition is "Corporacíon Favorita Grocery Sales Forecasting" which was researched and described by Valĺes-Ṕerez I., Soria-Olivas E., Martınez-Sober M., Serrano-Lopez A. J., Gomez-Sanchı́s J., and Mateo, F. Corporacíon Favorita, an Ecuadorian company owner of multiple supermarkets across Latin America, released this data set around 2017 as a Kaggle competition to challenge the community to forecast their sales [5].

**Objectives of the article.** The aim of the article was to review the organization of the competition and the achieved results from a theoretical and practical point of view. Conduct exploratory data analysis. Also, analyze challenges, limitations, and potential research areas in retail sales forecasting.

**Presentation of the main material of the study.** Hack4Retail is a competition to determine the best projects or technological solutions. The hackathon is aimed at professionals who study or work in the field of state-of-the-art IT / data science solutions [6]. All information about the Hackathon, as well as any changes to the terms of the Hackathon, was posted on the official website [2]. An online platform managed hackathon – Analytics Vidhya and organizers – LLC McKinsey and Company Ukraine and LLC Silpo-Food. The rules and information regarding the competition were posted, on both Analytics Vidhya and the Hack4Retail official competition's website.

The rules allowed you to participate if you reached 18 (eighteen) years after successfully registering on the Analytics Vidhya platform. Participants were allowed to create their teams with a limit of 4 (four) persons.

Organizers presented a list of requirements for ideas/solutions. It must meet the following criteria: (1) distinction; (2) novelty; (3) copyright belonging to the participant and/or the team of participants; (4) potential for use and further development of the project in the activities of the Customer.

Competition launched on October 29, at 07:00 p.m. (Kyiv time) and lasted until October, 31 2021, 07:00 p.m. (Kyiv time). On October 29, 2021, at 07:00 p.m. (Kyiv time) the registered participants were given access to the Hackathon online platform where they found the description of the task for the first stage, data sets, and the additional information they may want to use while working on the task. The teams were allowed to use the Hackathon online platform to upload their technological solutions and the updates of these solutions. The platform evaluated submission automatically [6].

**Submission.** All forecasts were submitted to the Analytics Vidhya platform using the template provided by the organizers. This template is required to forecast the 1,666,028 series for a 14-day forecasting horizon (from 2021-07-20 to 2021-08-02). The series contains information about the 1961 unique SKUs. Also, the submission data is normalized – each day has an equal number of series and SKUs.

As in the M5 accuracy competition, the submission template did not affect how the forecasts were produced, and teams were completely free to use their preferred forecasting method to forecast the individual series. However, the submission template ensured that the forecasts were coherent and in an appropriate form for direct evaluation [4].

The participants were allowed to submit up to 4 (four) entries per team per day on the Analytics Vidhya platform. Each team selected only one submission to be close to real-life when they choose a single set of forecasts that possibly will represent future sales. Typically, if no particular submission was selected, that with the highest performance during the "validation" phase were automatically selected by the system.

**Evaluation.** Various metrics were used for evaluating the accuracy of unit sales forecasting. The M5 Accuracy competition utilized a variant of the MASE originally proposed by Hyndman and Koehler (2006) [7] called the root mean squared scaled error (RMSSE). After estimating RMSSE in the M5 competition, the overall accuracy of the

forecasting method was computed by averaging the RMSSE scores across all series in the data set using appropriate weights. This measure is called the weighted RMSSE (WRMSSE) [4].

The Corporacíon Favorita Grocery Sales Forecasting competition used the Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE). This metric avoids penalizing large differences in prediction when both the predicted and the true number are large. Also, organizers defined custom weights [5].

The Hack4Retail competition chose mean absolute error (MAE). MAE is calculated as the sum of absolute errors divided by the sample size (1).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}, \quad (1)$$

where $y_i$ is the prediction, $x_i$ – the true value, and n – sample size.

Unfortunately, no measure is perfect because all have advantages and disadvantages, but MAE is the most natural measure of average error magnitude, and that (unlike RMSE) is an unambiguous measure of average error magnitude [8].

**Data.** The competition data was provided by Fozzy Group, consisting of the unit sales of various products sold in Ukraine. The data involves three data sets – time series, geo, and SKU files. The time-series data set contains the history of 1961 unique SKUs from 2011-01-29 to 2016-06-19. The SKU data set involves the metadata for each stock-keeping unit. Each unit is classified into 5 commodity groups, 208 categories, and 182 types. Also, there is information about 90 product brands in three languages – EN, UK, and RU. Products are sold in most regions of Ukraine, but this information is encoded in 515 geo clusters. So, all SKU information can be grouped based on either location (store and geo cluster) or product metadata (group, category, type, brand), as shown in Fig. 1.

Fozzy Group provided a limited commodity group number. The data set contains information about 5 groups, as shown in Fig. 2 – bakery, yogurts, cheese, mineral water, and tropical fruits. Also, the data set does not contain records for items on days when there were zero-unit sales. As a result, it lacks information about prices. These two factors make modeling and forecasting more complicated. Zero sales of some SKU at a given date can be affected by price, demand, or both reasons.

The level of sales of one product may differ significantly from the level of sales in other regions. It is also possible to observe certain differences in price dynamics in the different areas, as shown in Fig. 3. This is why the geo cluster is an important feature for modeling.

**Results, winning submissions, and key findings.** As a result, 194 teams registered in the competition, and 90 of them made submissions – it is less than 50 percent. Table 1 shows the aggregated score (MAE) reached by the top 10 teams. When you submitted with zero forecasts, MAE would be equal to one. It means that 1 is the score of naive forecasting. More than 35 percent of teams got better MAE than 1 MAE.

We can also observe that there is no linear relationship between submissions number, average

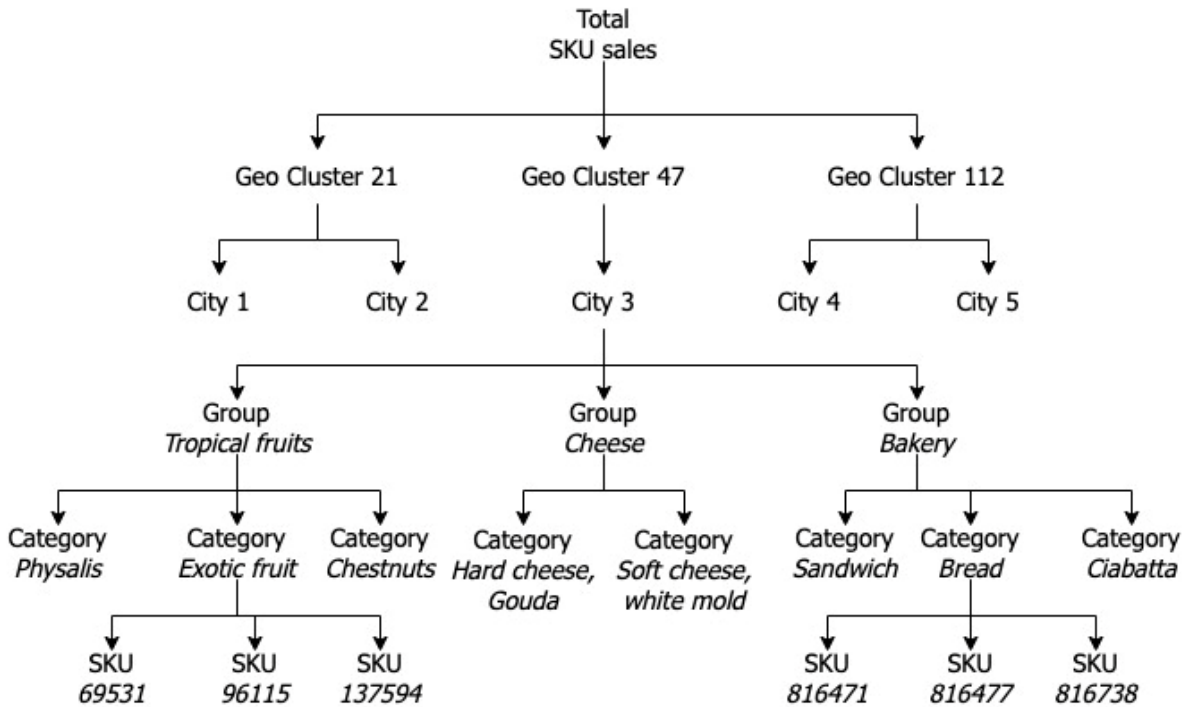


**Fig. 1. Grouped time series of the competition. The data can be aggregated in different levels by SKU-related information (*commodity group, category*) or location (*geo cluster, city*)**

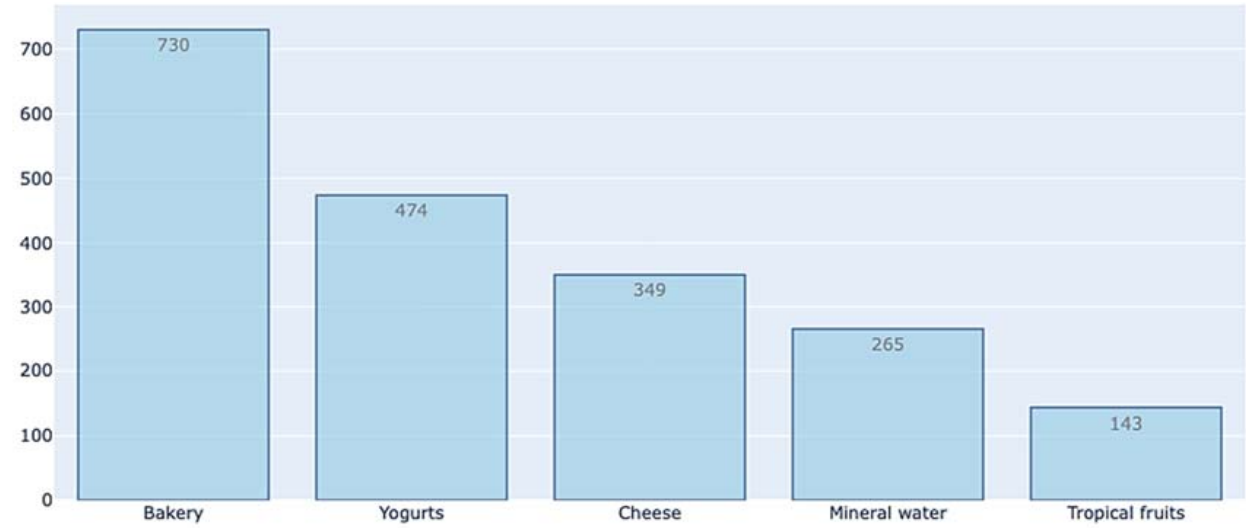


**Fig. 2. SKU count per commodity group**

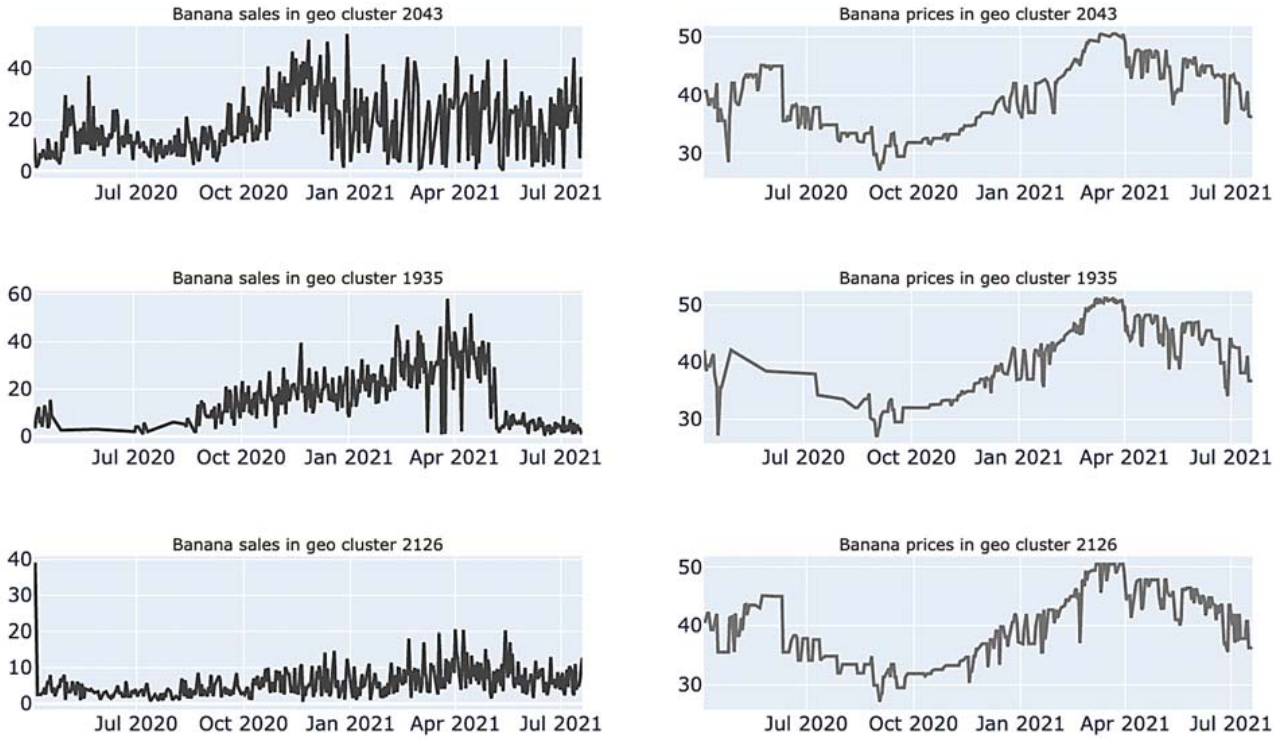


**Fig. 3. Time-series of daily sales and prices for bananas in different geo clusters**

submission score, and final score (MAE) (as shown in Table 1). This could mean that each of the top 10 teams had different strategies during the competition.

Platform "Vidhya Analytics" enables us to collect data about the history of solution submissions. This provides an opportunity to view the progress of the metric (MAE) during the two days of the competition. The analysis showed that there is no linear progression of the metric. In contrast, the metric value has many outliers in most teams (as shown in Fig. 4). Unfortunately, this may indicate an ineffectively selected metric or data leakage that will be described below.

Unfortunately, a limited number of teams that participated in the Hack4Retail competition shared their methods with descriptions. Nevertheless, available public methods can be useful as more effective approaches than dozens of other teams.

The forecasting methods used by the three winning teams with public methods can be summarized as follows.

– *First place (AfterParty; Fred Navruzov)*: the solution was validated by the K-fold strategy, where the window (input data) is one month and the forecasting horizon (output data) – is two weeks (each week was evaluated separately for public/private evaluation

Table 1

**Performance of the top 10 teams in the competition in terms of MAE**

| Team Name | Score (MAE) | Submissions Number | Average Submission Score |
|---|---|---|---|
| AfterParty | 0.748 | 21 | 1,91 |
| Prada.ai | 0.814 | 22 | 2,3 |
| Profesiyni Shatuni | 0.895 | 24 | 0,97 |
| Hack4reMONT | 0.896 | 15 | 1,24 |
| One More Mistake | 0.9 | 13 | 0,99 |
| Tiger Analytics | 0.902 | 32 | 1,04 |
| julia110995 | 0.902 | 61 | 1,32 |
| Final Submission | 0.902 | 16 | 1,0 |
| Raptus | 0.905 | 22 | 5,15 |
| sorochilco | 0.906 | 19 | 1,61 |

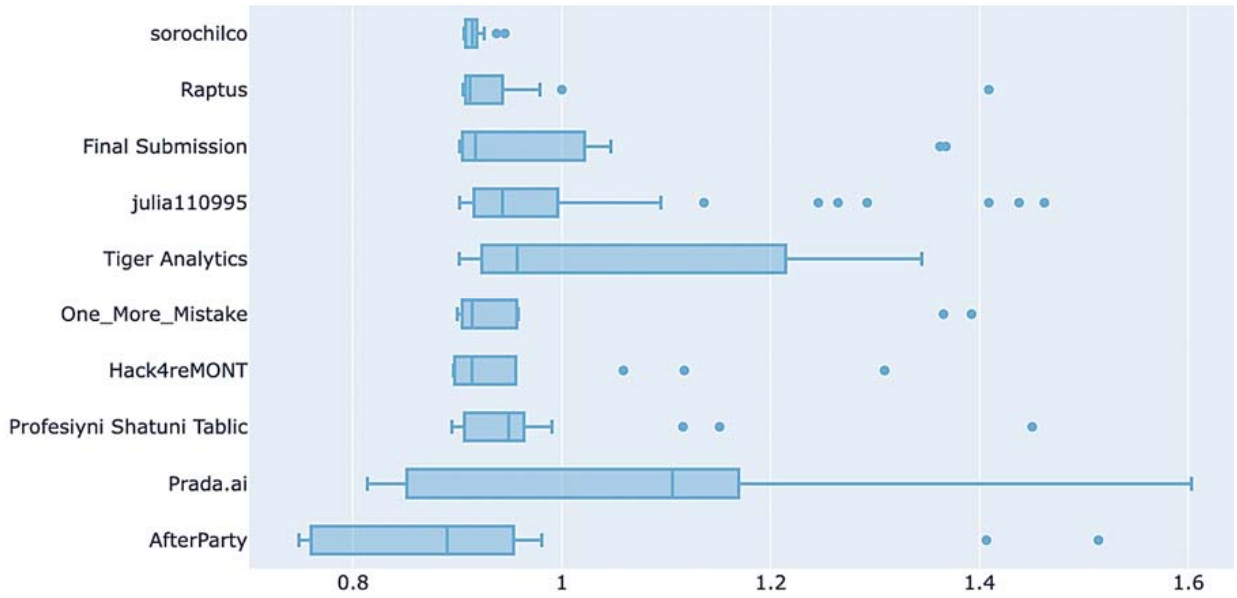


**Fig. 4. Submissions score per team. Filtered metrics, where MAE is higher than 2**

simulation). The final method was an ensemble of LightGBM Regressor (*0.7518 MAE*) and feed-forward neural network with MAE loss (*0.7598 MAE*) that was evaluated with 0.7482 MAE.

– *Second place (Prada.ai; eddiekro)*: LightGBM Regressor was used as final solution and gave 0.81 MAE. Most columns of data sets were prepared as input data. Also, the team splited train data into 5 folds for model validation.

– *Third place (Profesiyni Shatuni Tablic; Andrii Shalimov)*: the team used LGBM as a baseline model with all features from the train table, some features from the SKU table, and custom sales features – the median for the past 14 and past 21 days, and sales lag for the last 14 days. This baseline solution gave 0.97 MAE on the leaderboard. The final approach gained 0.89 MAE. It was a blending of Auto ARIMA results with the most popular 800 SKUs, which had the best sales and had zero sales within 3 days, and ARIMA for the least sales SKUs.

Among the best solutions were presented LightGBM and Feed-Forward Neural Network, which showed their effectiveness in other competitions for sales forecasting. LightGBM has become the standard choice for such tasks [4; 12]. This model has some advantages compared to others – the ability to process a large number of features, including categorical data (*this became an advantage when processing geo clusters and commodity groups*); also, it is faster than other GBM models and doesn't depend on data prepossessing. Feed-forward neural networks have also shown their effectiveness in previous competitions [5]. This model uses a sequence-to-sequence architecture, which receives product features, its sales history, and price as an input, and a vector (*forecast horizon*) as an output. Such a model is more difficult to implement and optimize than LightGBM (*which only requires hyperparameter selection*), but experienced developers can obtain quality predictions using Feed-Forward neural networks.

**Discussion, limitations, advantages, and directions for future research.** The main discussed topic in the competition was data leakage. When given SKU in a given location and date has zero sales, this date doesn't have information about price and sales. As a result, some teams revealed the logic of price information by filling in the test data set. For example, the winner used the feature price change and realized that it improved forecasting accuracy. Unfortunately, features related to price change are data leakage and cannot be used in production mode, since in production we will not know future prices (*or we will know, but they will be formed according to a completely different logic, and cannot be used in this way*).

An additional topic for discussion was the metric (MAE) and its effectiveness in cases where the SKUs had a lot of gaps in sales (it represents zero sales that day). For example, a product had sales every other day, and the next day it didn't. It was difficult to estimate the model that predicted the average number of sales, or whether there would be sales at all.

The first limitation of the competition is forecasting based on data about the past. Like other empiric studies [4], this competition provides useful information about the accuracy of different approaches for researchers. Also, the results of such competitions provide recommendations on how to improve the decision-making process for practitioners. However, the effectiveness of the developed models depends on the extent to which the data used correspond to the real data flow. There are cases when training data sets can be different on average from those data that are used in the process of forecasting and decision-making [9]. Also, it is not possible to evaluate the developed models for crisis situations, such as the coronavirus pandemic and the war in Ukraine. We can only assume that the quality of these models will be many times worse.

Another limitation of the competition was that the modeling focused on point forecasting accuracy without the context of business operations. Some empirical studies show that accurate forecasting can lead to higher costs for the company, such as increased inventory costs or service costs [4]. Also, it was not known which forecast could be more effective – the forecast for the guaranteed sale of all units of the product for a certain period, or the forecast for a stable balance in the warehouse.

Also, a significant limitation was the limited number of product groups used for the modeling. As shown in Fig. 2, there were only five of them – bakery, yogurts, cheese, mineral water, and tropical fruits. It is clear that a larger number of different types of goods would improve the quality of empirical studies.

The advantage of the competition is that Silpo is a local supermarket. The importance of research for smaller retail firms as described in other studies [10]. It allows evaluating different forecasting methods, not only based on large companies like Walmart.

The results obtained during the competition show that machine and deep learning methods are effective not only for global companies but also for regional companies. Therefore, the theoretical value and potential for practical use require additional research in the context of smaller firms. It is also important to take into account how the described approaches work during crisis periods, such as the COVID-19 and the war in Ukraine since smaller companies are more vulnerable to such changes, which forces us to reevaluate the performance of some methods.

It is also important to consider that machine learning and deep learning approaches are additional costs for the company. Therefore, it is important to investigate whether the improvement of forecasts and decision-making justifies the additional costs for companies at the regional level [4; 11].

An important direction for research is the interpretation of models for forecasting. This is important because managers are usually reluctant to make decisions when they cannot understand the logic of the methods they plan to use. Perhaps, from a practical point of view, a method that is more interpretable will be more effective than one that is more accurate from a metric perspective.

Also, it is worth reviewing the results obtained for a larger number of product groups and checking the effectiveness of the described models in the context of the war in Ukraine. Because the Russian aggression not only worsened the supply chains for Silpo but also caused the loss of many of the chain's supermarkets.

**Conclusions.** Hack4Retail competition extended the empirical data that we have about sales forecasting. Unlike Walmart, Fozzy Group is a smaller local retailer in Ukraine, which provides an opportunity to test the previously described approaches in other conditions. This paper described competition organization and analysis of data sets (explored key characteristics). The aim of this work was to review the results of the best teams and describe the most accurate solutions. In addition, it aimed to provide information to practitioners interested in applying the findings of the competition to improve the business operation's performance by the forecasting methods.

As in previous M5 "Accuracy", Corporacion Favorita competitions, the Hack4Retail competition focused on retail sales forecasting with an empirical evaluation of methods performance. To achieve this goal, the competition provided the history of 1961 unique SKUs from 2011-01-29 to 2016-06-19 and input data for two weeks forecasting horizon.

In summary, the Hack4Retail competition provided the forecasting researchers and practitioners with key findings and reaffirmation.

– Prediction methods such as LightGBM and Feed-Forward NN were found to be effective for community competition as well as for bigger competition (M5 "Accuracy", Corporaci´on Favorita).

– The following approaches have once again shown their effectiveness in improving the performance of forecasting models: cross-validation, blending, cross-learning, and feature engineering based on explanatory variables.

– The competition confirmed the importance of high-quality data, especially for the final evaluation sample. Also, the data leak confirmed that there is sometimes a gap between the data requirements for modeling and the actual data collection flow.

However, we believe that more research is needed to improve the findings of the Hack4Retail competition: re-evaluate used models on more commodity groups; find a way to avoid data leakage; reevaluate methods when used in real-time; review the benefits of accurate forecasting for various aspects of a retailer's operations; to investigate the performance of the described solutions in the context of the COVID-19 and the war in Ukraine.

**REFERENCES:**

1. Fozzy Group Homepage. Available at: https://www.fozzy.ua/en/ (accessed 19 July 2022).

2. "Hack4Retail" hackathon Homepage. Available at: https://hack4retail.fozzy.ua/ (accessed 19 July 2022).

3. Hyndman, R. J. (2020). A brief history of forecasting competitions. In *International Journal of Forecasting* (Vol. 36, Issue 1, pp. 7–14). Elsevier BV. DOI: https://doi.org/10.1016/j.ijforecast.2019.03.015.

4. Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. In *International Journal of Forecasting.* Elsevier BV. DOI: https://doi.org/10.1016/j.ijforecast.2021.11.013.

5. Vaĺes-Ṕerez, I., Soria-Olivas, E., Martınez-Sober, M., Serrano-Lopez, A. J., Gomez-Sanch́ıs, J., Mateo, F. (2022). Approaching sales forecasting using recurrent neural networks and transformers. In *Expert Systems with Applications* (Vol. 201, p. 116993). Elsevier BV. DOI: https://doi.org/10.1016/j.eswa.2022.116993.

6. "Hack4Retail" hackathon Homepage and rules page on Analytics Vidhya. Available at: https://datahack.analyticsvidhya.com/contest/hack4retail (accessed 19 July 2022).

7. Hyndman, R. J., Koehler, A. B. (2006). Another look at measures of forecast accuracy. In *International Journal of Forecasting* (Vol. 22, Issue 4, pp. 679–688). Elsevier BV. DOI: https://doi.org/10.1016/j.ijforecast.2006.03.001.

8. Willmott, C., Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. In *Climate Research* (Vol. 30, pp. 79–82). Inter-Research Science Center. DOI: https://doi.org/10.3354/cr030079.

9. Spiliotis, E., Kouloumos, A., Assimakopoulos, V., Makridakis, S. (2020). Are forecasting competitions data representative of the reality? In *International Journal of Forecasting* (Vol. 36, Issue 1, pp. 37–53). Elsevier BV. DOI: https://doi.org/10.1016/j.ijforecast.2018.12.007.

10. Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. In *International Journal of Forecasting*. Elsevier BV. DOI: https://doi.org/10.1016/j.ijforecast.2021.07.007.

11. Nikolopoulos, K., Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? In *Computers and; Operations Research* (Vol. 98, pp. 322–329). Elsevier BV. DOI: https://doi.org/10.1016/j.cor.2017.05.007.

12. Bojer, C. S., Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. In *International Journal of Forecasting* (Vol. 37, Issue 2, pp. 587–603). Elsevier BV. DOI: https://doi.org/10.1016/j.ijforecast.2020.07.007.