Modeling the Optimal Grocery Store Trading Area Using Machine Learning Methods

Olena Liashenko and Bohdan Yakymchuk

Taras Shevchenko National University of Kyiv, 90-A, Vasylkivska st., Kyiv, 03022, Ukraine

Abstract

Over the past few years, the COVID-19 pandemic has significantly transformed consumer behavior, which has undoubtedly affected a large number of industries. Food retail was among the sectors where the effect was significant and led to the transformation of the approach to customer interaction. A large part of consumers began to use online delivery services more, and key players were able to provide delivery of products with their own delivery services or third-party on-demand courier service companies. Undoubtedly, in addition to operational changes in retailers' business model, this also affected their investment activities. Some key players began to reduce their trading floor areas to increase financial efficiency and look for options to work in a convenience store format. In our research, we offer an approach for making the right investment decisions when opening a new store to balance financial metrics and customer satisfaction indicators, which is a key sales driver for the segment of customers who substitute delivery service for brick-and-mortar store visits. Using Machine Learning methods, we solve the task of scenario modeling of revenue and operational efficiency metrics for different areas of the store's trading floor, which allows us to identify the optimal choice for the retailer. Using traffic metrics during peak operation hours, we determine the minimal density of the trading area that will not lead to a decrease in the activity of guests inside the store. Such an approach allows us to evaluate the best format of the store, forecast the object's revenue, and recommend investment project parameters.

Keywords [1]

Grocery retail, consumer behavior, machine learning, investment project, Data Science, regression, time series clustering.

1. Introduction

The central part of the investment activity of grocery supermarket chains includes organic growth by opening new stores, which allows expanding the audience by covering new regions or increasing the coverage of the population with the company's services. Traditionally, the format of opening a store depends entirely on the external characteristics of the location: such as population density, residential real estate, and the presence of competitors in the radius of the store's geolocation. However, in recent years, with the rapid growth and increasing penetration of e-commerce, traditional brick-and-mortar stores are losing their profitability and efficiency due to significant changes in consumer behavior. Furthermore, due to the COVID-19 pandemic, previously conservative Ukrainian consumers became accustomed to e-commerce channels, which led to considerable improvement in the retail industry towards digitalization and, as a result, active development of delivery services [4].

Retailers now have to reinvent the brick-and-mortar store format, enhance their digital capabilities, improve their loyalty programs, and consider new ways to engage with the audience to strengthen their business model and operational efficiency [3]. In Central and Eastern Europe, traffic and frequency of retail purchases have significantly decreased over the past two years (2019-2021). Still, the average consumer's basket has increased, which could offset the sales drop. At the same time, the penetration

Information Technology and Implementation (IT&I-2022), November 30 - December 02, 2022, Kyiv, Ukraine
EMAIL: olenalyashenko@knu.ua (O. Liashenko); bogdan.yakymchuk3@gmail.com (B. Yakymchuk)
ORCID: 0000-0002-0197-4179 (O. Liashenko); 0000-0003-2574-3815 (B. Yakymchuk)

of online trade has increased significantly (3.6% of offline sales in the Czech Republic and 1.5% in Poland). For European countries, such as Germany, the UK, the Netherlands, France, Sweden, Spain, Italy, Portugal, the Czech Republic, and Poland, e-commerce penetration grew to 6.6% of offline sales (weighted average) [10]. Such changes in consumer behavior are leading to the following trends for traditional brick-and-mortar supermarket chains:

1. Store strategy is shifting to providing a distinctive experience that brings customers to physical locations. The role of a place for issuing orders formed online (click and collect format) should be added for brick-and-mortar stores to create a synergy with online channels and reduce delivery costs. In general, grocery stores may need less physical space and may need to reduce costs as offline formats lose sales.
2. Retail companies should be flexible to gain market share and improve margins through three primary levers: branding and marketing, sustainable value proposition, and differentiation.
3. Now the main investment direction for retailers as expansion through new store openings should be eliminated with digitalization and deep direct interaction with consumers.

All the above trends indicate that retailers must effectively evaluate the initial store trading area and qualitatively evaluate external factors describing the target audience when searching for new locations. Opening a new point of sale is a rather complicated and expensive process. Therefore, the problem of external factors evaluation that may affect the potential of the frequency of visits to the new store and the purchasing power of the target audience is quite relevant among researchers.

Most often, attention is paid to the approach precisely from the position of physical rotation of points of sale, which is a rather complex analysis that must consider many external factors and be sensitive to even minor environmental changes [13]. Location analysis also includes such scientific developments as retail location theory, theory of land value or central place theory, or minimum differentiation principle as proposed by Baviera-Puig [1]. However, it is worth considering that companies often take geomarketing factors into account to reduce the risks of a negative consumer experience and, as a result, loss of consumer loyalty and financial and reputational risks. According to Cheng et al. [2], geographical analysis is a mandatory tool for representatives of the retail market. Therefore, most scientific works are devoted to the geographic choice of location as a critical variable.

Scientists also studied the descriptive-deterministic approach, which describes the consumer as a person looking for the nearest location [12]. Among these are gravity models developed by Yrigoyen and Otero, which evaluate the relationship between the attractiveness of a store for the consumer and the distance to it [16]. The thesis "size-distance to the store" [7] is also popular for research, which over time was supplemented by the multifactorial Multiplicative Competitive Interaction Model proposed by Nakanishi and Cooper [11]. Finally, relatively many works are also devoted to the analysis based on direct utility assessments [9]. However, as previously stated, the pandemic has forced consumers to go out less often and spawned a boom in e-commerce and mobile commerce. As a result, even industry retail market giants, such as Walmart or Sephora [5], are reducing their retail space in parallel with the boost in online sales. Therefore, for the survival of offline retailers, in addition to changing the format and diversification, it is necessary to optimize the operational processes and the store trading area. That is why there is a need to review the approach to choosing the location format, considering the depreciation of the location trading area, which may lead to a range of business effects to the brick-and-mortar retail model:

- Reduction of the rental burden and staff structure, and accordingly reducing the economic buying quantity (EPQ), the sales level that minimizes the total holding costs and ordering costs in inventory management;
- Decrease of technological equipment and, accordingly, utility costs;
- Capital Expenditures and Cash Flow optimization;
- Increase of Internal Rate of Return and Net Present Value of the new store opening as a project.

Our research will be devoted to evaluating the optimal area of the trading floor at the time of opening of the store to optimize investment budgets and improve the profitability of the square meter of the trading floor area without potential losses for customer service during peak hours.

2. Methodology

Our study considered the decision-making process of opening a new store. The management of the retail company (Growth Office) at the time of signing a long-term lease or construction contract to open the new store has information only about geographical factors of location. At that moment, the scenarios of the store format can be simulated to decide the optimal trading area and potential traffic. Of course, taking into account the trends of the modern retail market, the most effective solution is maximizing income from a square meter of the trading floor area. However, it should also be taken into account that there is a certain level of traffic per square meter ratio, which will create an uncomfortable environment and, on the contrary, will push away the consumer in peak hours, when there can be expected possible utilization of high-volume goods, high level of "Out-of-Shelf" and queues at the counter area. That is why the following approach was considered in our research:

1. Development of a Machine Learning model to forecast sales during the store's peak hours based on external factors and store trading area.
2. Determination of the income from a square meter ratio, at which the level of guest service satisfaction does not decrease.
3. Simulation of scenarios at which the ratio will be maximized with a high customer satisfaction level constraint.

As a first step, a dataset was collected with a set of metrics that can describe the activity of consumers within each individual store during the day. A total of 236 stores of one of the leading Ukrainian supermarket chains of the premium format were selected as the sample for our research. The dataset by day, time, and store includes the following metrics: number of unique SKUs sold, number of SKUs sold, number of cheques (non-unique visitors who made purchases), and total sales. Furthermore, in order to study the level of store loading with consumer traffic, the dynamics at the level of each store were grouped by month, day of the week, and store hours.

The first noticeable trend, which can be easily distinguished from the dynamics (Fig. 1):

- the weekly and monthly seasonalities of sales;
- stores trade more on Fridays and in December;
- the effect of weekly seasonality varies significantly throughout the year.

The intraday seasonality, though, has a tendency over months and weeks: peak hours are from 11:00 to 13:00 and from 17:00 to 19:00. However, on weekends, the behavior of consumers changes significantly, and the evening peak of traffic loads cannot be observed. If you single out the dynamics of four representative months (January, May, July, and December), you can see how much the dynamics change during the year. For example, peak hours are more pronounced on January and December weekends, which is explained by the abnormal load on holidays (Fig. 2)..

This approach allows us to highlight the main trends that indicate the inconsistency of the traffic distribution during the day from seasonality factors during the week and year by month. In addition, of course, the level of traffic is affected by holidays and weather conditions. Furthermore, with the impact of the COVID-19 pandemic and military externalities, consumer behavior is becoming less and less predictable. Still, the suggested approach is to identify the store format that best matches the comfortable visit to the store and allows us to optimize the traffic flow for higher sales and service.

According to the proposed approach, factors of temporal influence were identified, but there is also a high dependence on consumer behavior within the framework of geographical distribution and the properties of the store itself. Therefore, machine learning methods for clustering will be used to identify factors or groups of stores with similar properties and explain the distribution of traffic throughout the day. The approach chosen to identify time series with a high level of similarity is Dynamic Time Warping (DTW). This algorithm is used to measure the similarity between two time sequences. The DTW distance is calculated using a dynamic programming algorithm, which allows you to construct an optimal transformation path under boundary, monotonicity, and continuity constraints [14].

The formula is used to initialize the initial state of the algorithm:

$$dtw(i,j)=\begin{cases}\infty, \text{if } i=0 \text{ or } j=0\\ 0, \text{if } i=j=0\end{cases} \quad (1)$$

A recursive relationship is described by the following dependency:

$$dtw(i,j) = c(x_i, y_j) + \min\{dtw(i-1,j), dtw(i,j-1), dtw(i-1,j-1)\} \quad (2)$$

where $i = \overline{1,n}, j = \overline{1,m}, c(x_i, y_j)$ - costs for matching observations of two time series x_i, y_j, respectively, calculated according to the Euclidean distance formula.

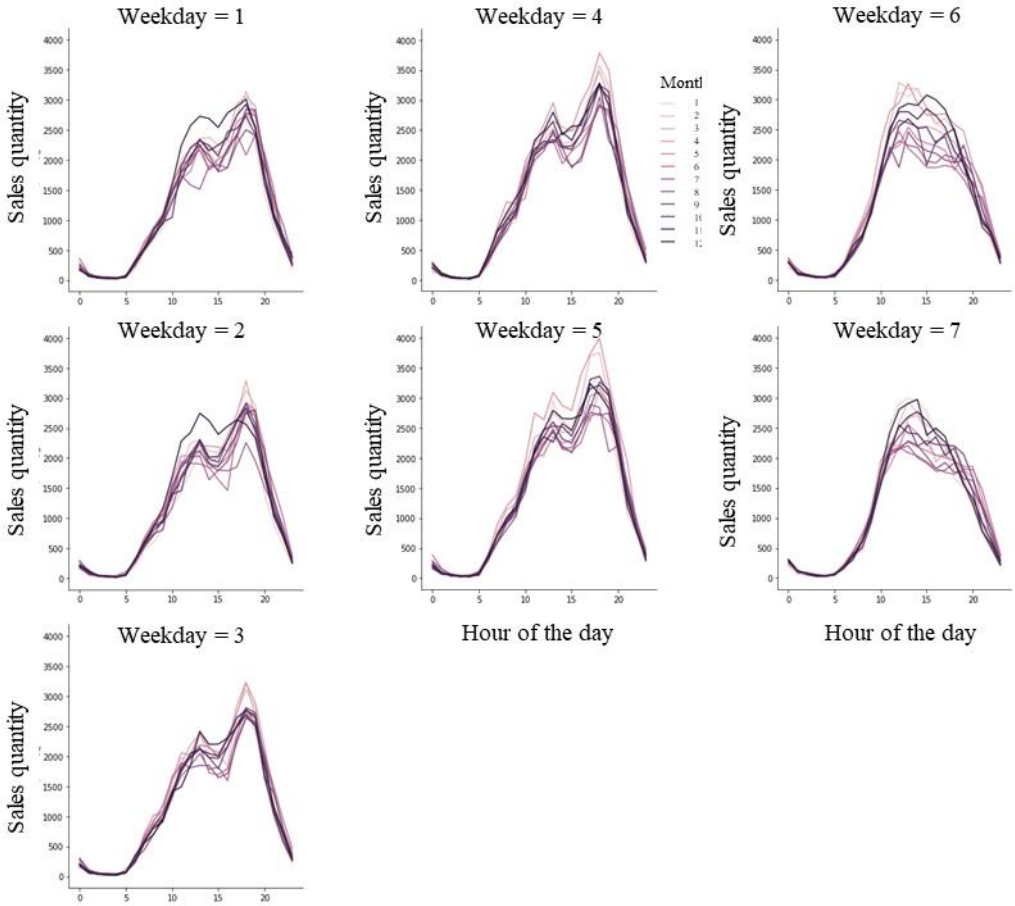

Figure 1: Distribution of sales during the day in terms of days of the week and months.

Using this approach, you can build a cost matrix (Fig. 3) on which you can display the optimal path of transformation (red line). This approach allows for a pairwise assessment of the relationship between time series representing the entire network of supermarkets. At the same time, this approach can be applied to series reflecting different dynamics patterns: by average sales dynamics during the day, week, and months. The obtained pairwise metrics of the density of time series dynamics can be combined with factors that can explain the similarity of store sales curves. To extract the factors' coherence, the density-based spatial clustering of applications with noise (DBCAN) approach is used. First, the neighbor search radius ε and the minimum number of neighbors in the cluster are input to the algorithm. The algorithm first finds neighbors around each point and determines core points that satisfy the given minimum. In the next step, the algorithm identifies connectivity components for core points, excluding non-core points. Finally, the method connects each non-core point to the nearest cluster, provided that the cluster is in the ε neighborhood; otherwise, it marks it as noisy [6].

In this way, our stores can be split into cluster groups, and then cluster binary features can be used in the modeling of the optimal area. The indicator that describes the optimal area for maximizing income without loss of service level is such traffic in terms of the number of sold SKUs per square meter of the store's trading floor that satisfies the desired level of NPS (Net Promoter Score). NPS is such an indicator that allows you to assess consumer loyalty to the service provided by a specific network store.

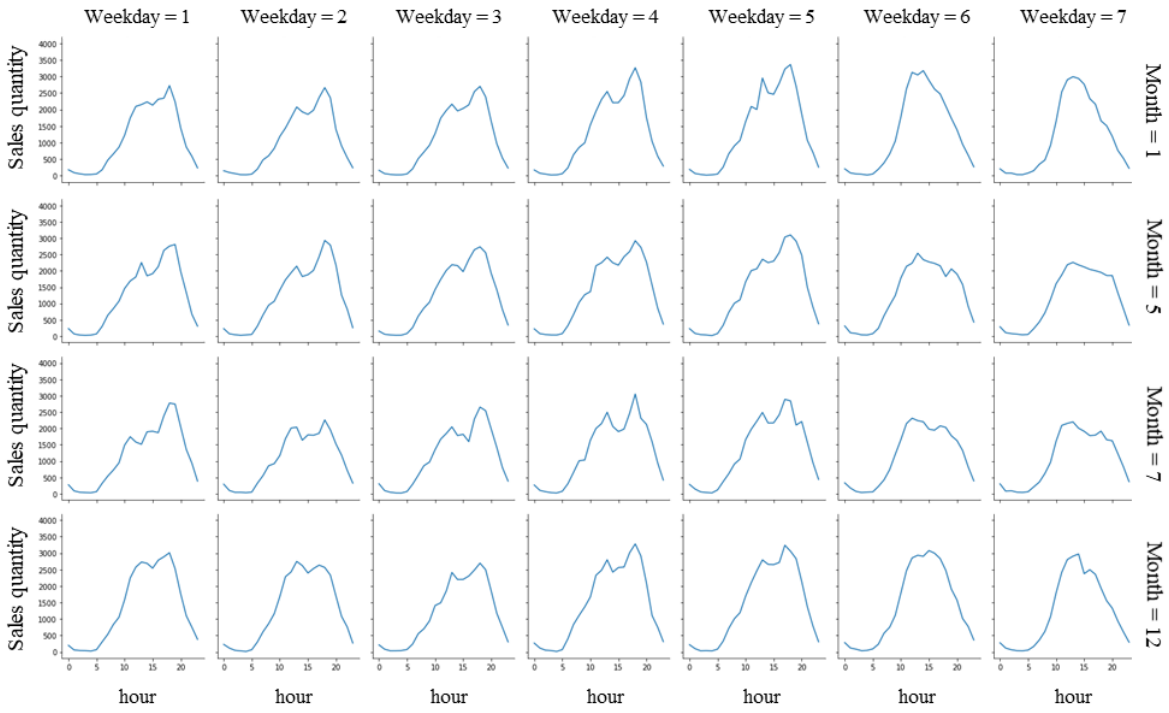

Figure 2: Distribution of sales during the day by days of the week and months (without grouping).

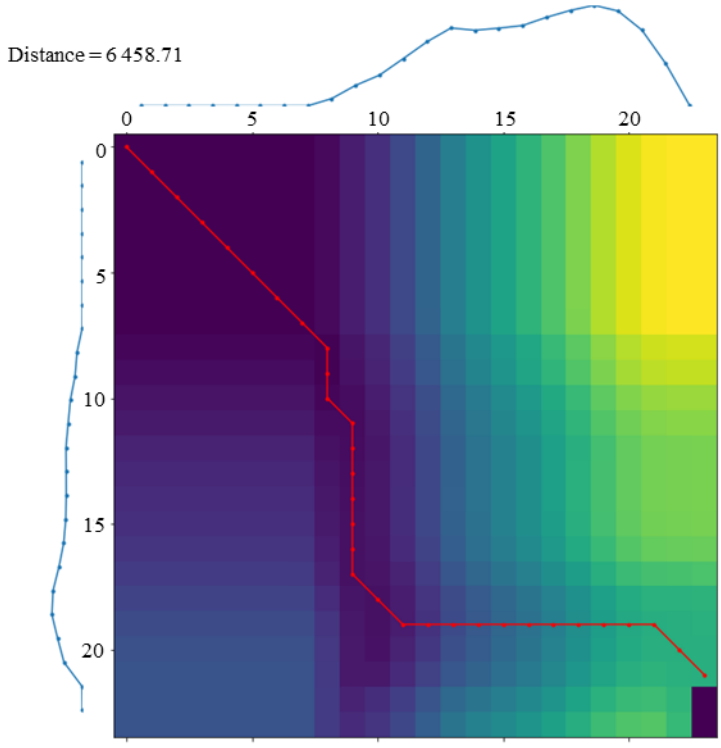

Figure 3: DTW matrix between two time series showing daily dynamics of two stores.

NPS is calculated according to the formula:

$$NPS = \frac{1}{n}\left(\left| \{ X : x_i \in [9,10] \} \right| - \left| \{ X : x_i \in [0,6] \} \right| \right) \quad (3)$$

where $x_i \in [0,10], i = \overline{1,n}$ - the assessment with which the consumer is ready to recommend the company's goods and services.

Usually, three groups of consumers are distinguished through the survey:

- "promoters" - those who rated the service above 9, the most loyal group of consumers,
- "passives" - those who rated the visit less than 9, but above 6;
- "detractors" - guests who rated the service less than 6 and are the least loyal audience.

The NPS is the difference in survey structure between promoters and detractors and can range from -1 (all rated negatively, i.e., below 6) to 1 (all rated positively, i.e., above 9).

Therefore, based on the NPS, which was evaluated during the period of the highest traffic during peak hours for the store, the top 25 supermarkets were determined by the level of NPS. As a result of research, the median ratio of traffic in the number of SKUs sold to the area of the sales floor was chosen as a reference (at the level of 5.27 SKU per square meter of trading floor area). Based on this indicator, there is a need to simulate scenarios of the store opening format based on the area of the sales hall. A considerable number of factors affect store sales, including the size of the store. Accordingly, an approach was chosen which uses machine learning regression methods to build a revenue forecast model during peak hours (the upper percentile of the sample to eliminate anomalies, Fig. 4).

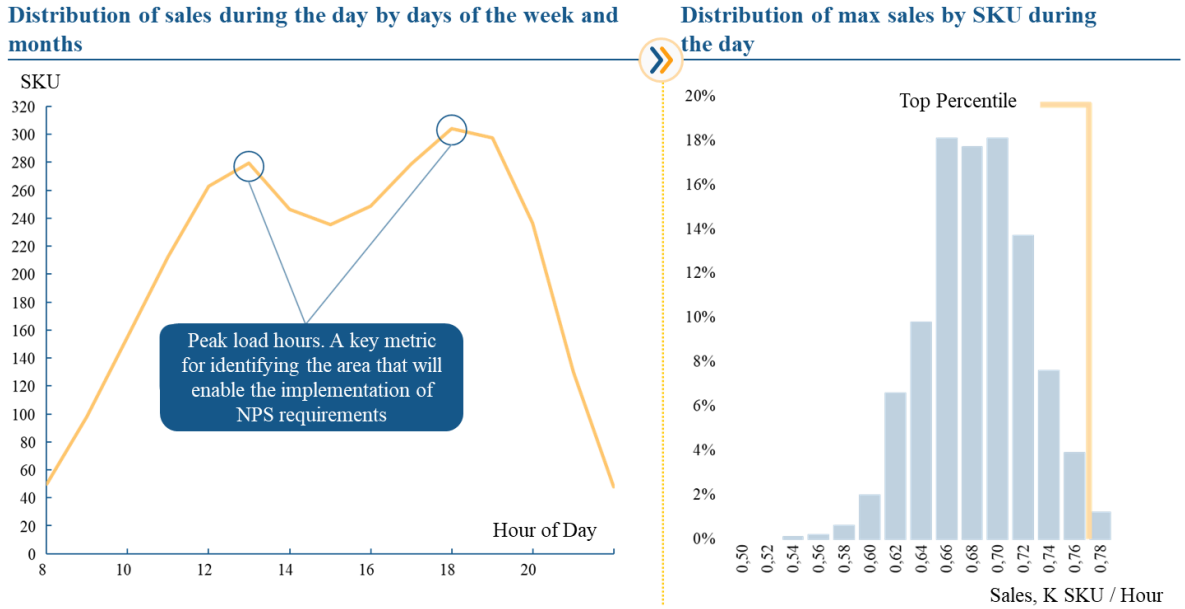

Figure 4: Anomaly rejection approach

3. Empirical Results

To develop the model, a sample of 236 stores of the supermarket chain was formed, and the following factors were selected to predict the peak load of SKU sales:

1. **Features describing the location**: region, city, population, and area of the city, the population in the radius of the site, distance from the city center, availability of parking, binary factor (located in a shopping center, shopping center), number of competitors in the radius of the location.
2. **Features describing the store format**: the total area of the store, the trading area, the number of cash registers, the number of self-service cash registers, the number of operation hours of the store, the assortment cluster, the total area of shelves, the year the store was opened.
3. **Clustering results**: cluster by the distribution of sales during the day, cluster by the distribution of sales during the day and week, and cluster by the distribution of sales during the day and year.

The following transformations were applied to the dataset:

1. Combining or discarding factors with a unique number of levels
2. Elimination of multicollinearity
3. Normalization of quantitative variables using the z-score transformation: $(x-\mu)/\sigma$

4. The Yeo-Johnson transformation [15], which allows you to reduce asymmetry and approach a normal distribution:

$$\psi(\lambda, y) = \begin{cases} \left((y+1)^{\lambda} - 1\right)/\lambda & , if\ \lambda \neq 0, y \geq 0 \\ \log(y+1) & , if\ \lambda = 0, y \geq 0 \\ -\left((-y+1)^{2-\lambda} - 1\right)/(2-\lambda) & , if\ \lambda \neq 2, y < 0 \\ -\log(-y+1) & , if\ \lambda = 2, y < 0 \end{cases} \quad (4)$$

Among the models that had the best metrics of peak load prediction accuracy, the following can be distinguished:

1. Random Forrest and Extra Trees Regressor that combine decision tree framework and ensemble learning to randomly simulate decision trees and by mixing their results improve model accuracy metrics.

2. Gradient boosting, which is based on a step-by-step search for the optimal model. It starts with differential loss function initialization $F_0 = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F(x))$ and after each step improves model accuracy metrics by determination of the optimal multiplier to conduct the appropriate descent $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x_i)$ [8].

The highest result on the test sample was achieved when using the CatBoost model according to the RMSE, determination coefficient, RMSLE, and MAPE (Table 1). On the other hand, according to the MAE and MAPE metrics, the highest result was recorded when using the Extreme Gradient Boosting model.

Table 1

Comparison of modeling performance

Model	MAE	RMSE	R2	RMSLE	MAPE
CatBoost Regressor	81 438	116 088	0.8366	0.2134	0.1640
Extreme Gradient Boosting	79 869	128 285	0.7971	0.2151	0.1517
Light Gradient Boosting Machine	89 781	134 636	0.7832	0.2270	0.1749
Random Forest Regressor	89 971	136 217	0.7718	0.2251	0.1745
Extra Trees Regressor	92 648	143 527	0.7500	0.2387	0.1765

As the final model, the CatBoost algorithm will be used. Using the feature importance, the following metrics turned out to be the most influential (Fig. 5).

The most important factor turned out to be the year the store was opened. This variable actually demonstrates the strategy of grocery retail company development, which was investigated in this paper. Every year, in new openings, the chain is inclined to diversify the assortment, increase internal departments, and saturate the assortment with products of its own production and craft goods. This variable qualitatively explains the difference between stores with similar characteristics in the same region. At the opening, excitement is created, which stimulates consumers to overcome even greater distances in search of a better selection of assortment.

It can be seen that the model relies heavily on the clusters that were obtained using the DWT approach. Of course, clustering allows you to identify potential peak load hours, but it is also worth considering that in the case when a store opens in a new region or with a new unique format, there is a high risk of error when assigning such a store to one of the clusters. The mistake of associating a store with a cluster can cause a sufficiently high error in the prediction. When looking at significant deviations in the folds of the test sample, the most significant deviations were characteristic of stores that correspond to poorly representative regions of the eastern regions of Ukraine. Therefore, this approach can be effective with a qualitatively formed sample within the given attributes.

Among the metrics describing the format of the location, the following were chosen as the most significant: city population, distance from the city center, population within a radius of 1 km from the location (approximated through open data of building density from OSM), number of competitors and others. Finally, the metrics that we will use to build the opening scenarios are the trading area, the number of cash registers and self-service cash registers, the store's work schedule, and the assortment cluster.

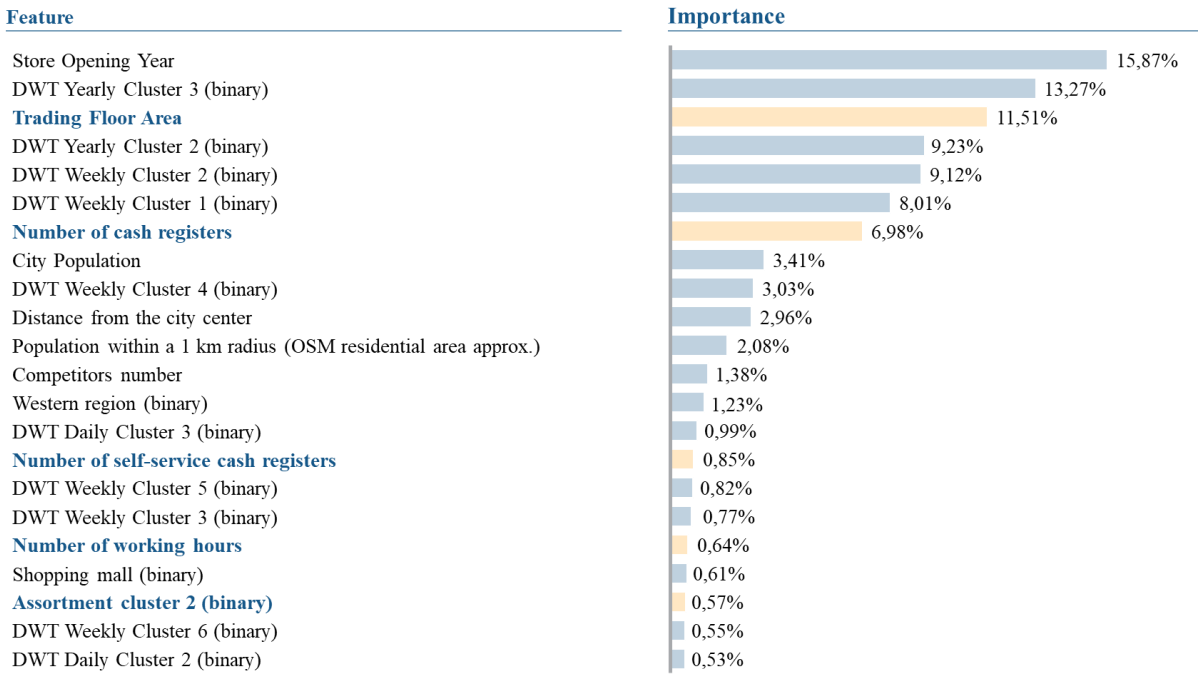

Figure 5: Feature Importance Comparison

Usually, the assortment cluster is the metric that is determined as a result of researching the store's target audience and is set according to fixed rules by the company's commercial office. Store scheduling is a metric that is usually sought to be maximized when opening a store in order to cover all traffic flows regardless of the time of day, but it is usually limited either by the operation of the shopping center or by the requirements of the legislation. Therefore, the main metrics that can be managed to estimate the peak load are the trading area and the number of cash registers (including self-service).

In the ideal decision-making process regarding the object, the fundamental parameter during negotiations is the area of the entity for construction or lease. Therefore, the developed tool can be used for simulation scenarios of the trading area. When transferring the parameters to the object, you can get a set of load forecasts in peak hours and the corresponding ratio of the number of sold SKUs per square meter of the trading floor. The optimal area vorticity rule can be described by the following rule:

$$\max\left\{sales_i : sales_i / area_i \le \gamma\left(NPS\right)\right\} \tag{5}$$

where $sales_i / area_i$ is the scenario of the revenue per square meter of the trading floor during peak hours; $\gamma\left(NPS\right)$ is the critical level from the study, estimated at 5.27 SKUs per square meter per hour, $i = \overline{1,n}$ - allowable for consideration of scenarios is the square of the hall.

It is worth noting that the more effective considering sales the square meter of the trading area is, the more marginal the business model of the brick-and-mortar store.

4. Conclusions

The described approach creates an opportunity for the business to enrich decisions with additional forecasts based on public and internal data sources. Such an algorithm may propose the optimal trade area that may lead to several improvements in the business brick-and-mortar grocery retailer model:

- Maximize the sales per square meter of trading area ratio;
- Reduce utility costs, the rental burden, and labor costs, leading to a lower level of the economic buying quantity (EPQ) and faster break-even point achievement;
- Increase EBITDA and profitability margins in a long-term period;
- Decrease the level of technological equipment and, as a result, Capital Expenditures Budgets
- Optimize Cash Flow structure leading to Internal Rate of Return and Net Present Value growth of the projects portfolio of the company.

It should be mentioned that the resulting machine learning model can be improved with a broader range of input parameters: behavioral characteristics of buyers, assessment of the population's purchasing power, the loyalty of the target audience to the brand, and others. However, the model effectively copes with simulating scenarios within the given metrics and can be used in the decision-making process.

An implicit result of the model that can be used by grocery retail business is the possibility of optimizing the existing network based on benchmark metrics of the cluster. If there is an opportunity to reduce store space with a clear improvement in operational efficiency, then the tool can be applied to improve the company's current financial performance. Also, the model makes it possible to evaluate the potential of opening in certain types of locations based on basic metrics and allows the development office to optimize the process of finding new locations, taking into account critical limits to ensure the desired level of service and profitability.

5. References

[1] A. Baviera-Puig, N. Roig-Tierno, J. Buitrago-Vera and F. Mas-Verdu. «Comparing trade areas of technology centres using geographical information systems.» The Service Industries Journal 33 (2013) 789–801. http://dx.doi.org/10.1080/02642069.2013.740467

[2] E. W. L. Cheng, H. Li and L. Yu. «A GIS approach to shopping mall location selection.» Building and Environment 42 (2007) 884–892. http://dx.doi.org/10.1016/j.buildenv.2005.10.010

[3] O. Chernyak, B. Yakymchuk, Assessment of the Impact of COVID-19 on Grocery Retail in Ukraine, KnE Social Sciences 5 (2021) 202–214. DOI: https://doi.org/10.18502/kss.v5i9.9894.

[4] Deloitte. Consumer sentiment of Ukrainians in 2020. Industry group for retail and wholesale distribution, 2021. URL: https://www2.deloitte.com/content/dam/Deloitte/ua/Documents/Press-release/RWD 2020 UA.pdf

[5] FastCompany.org, C. Dillow, Wal-Mart Plans to Grow by Shrinking, 2009. URL: https://www.fastcompany.com/1417301/wal-mart-plans-grow-shrinking

[6] J. Jokinen, T. Räty and T. Lintonen. "Clustering structure analysis in time-series data with density-based clusterability measure." IEEE/CAA Journal of Automatica Sinica 6 (2019) 1332-1343, doi: 10.1109/JAS.2019.1911744.

[7] D. Huff, B. M. McCallum. Calibrating the Huff Model Using ArcGIS Business Analyst. Redlands, CA, Esri White Paper, 2008

[8] J. Lambers, The Method of Steepest Descent, 2011. URL: https://www.math.usm.edu/lambers/mat419/lecture10.pdf

[9] J. Louviere, G. Woodworth. Design and Analysis of Simulated Consumer Choice of Allocation Experiments: an Approach based on Aggregated Data Journal of marketing research 20 (1983) 350-367

[10] McKinsey & Company. The State of Grocery Retail, 2022 URL: https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/state%20of%20grocery%20europe%202022/navigating-the-market-headwinds-the-state-of-grocery-retail-2022-europe.pdf

[11] M. Nakanishi, L. G. Cooper. Parameter Estimation for a Multiplicative Competitive Interaction Model: Least Squares Approach. Journal of Marketing Research 11 (1974) 303-311.

[12] W. J. Reilly, The Laws of Retail Gravitation, Knickerbocker Press, New York, 1931.

[13] N. Roig-Tierno, A. Baviera-Puig, J. Buitrago-Vera and F. Mas-Verdú «The retail site location decision process using GIS and the analytical hierarchy process.» Applied Geography 40 (2013) 191–198. http://dx.doi.org/10.1016/j.apgeog.2013.03.005

[14] D. F. Silva, G. E. A. P. A. Batista, Speeding Up All-Pairwise Dynamic Time Warping Matrix Calculation, SDM (2016) DOI:10.1137/1.9781611974348.94

[15] I. K. Yeo, R. A. Johnson. "A New Family of Power Transformations to Improve Normality or Symmetry." Biometrika 87 (4) (2000): 954–959. doi:10.1093/biomet/87.4.954. JSTOR 2673623.

[16] C.C. Yrigoyen, J.V. Otero, Spatial interaction models applied to the design of retail trade areas, European Regional Science Association conference papers, 1998