Forecasting new product demand using machine learning

P S Smirnov¹ and V A Sudakov¹,²*

¹Institute number 8 “Information Technology and Applied Mathematics”, Moscow Aviation Institute (National Research University), 4 Volokolamskoe Hhighway, 125993, Moscow, Russia
²Department “Problems of mathematical modeling and high-performance computing”, Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), 4 Miusskaya Square, 125047, Moscow, Russia

*E-mail: sudakov@ws-dss.com

Abstract. The problem of predicting demand for a new product based on its characteristics and description is critical for various industrial enterprises, wholesale and retail trade and, especially, for modern highly competitive sector of air transportation, since solving this problem will optimize production, management and logistics in order to maximize profits and minimize costs. Classic demand forecasting methods assume the availability of sales data for a certain historical period, which is obviously not the case when concerning a new product. Most research papers are limited either to a specific category of goods or use sophisticated marketing methods. This paper proposes the use of machine learning methods. We used data about new product demand from the Ozon online store. The input data of the algorithm are characteristics such as the price, name, category and text description of the product. To solve the regression problem, various implementations of the gradient boosting algorithm were used, such as XGBoost, LightGBM, CatBoost. The forecast accuracy is now about 4.00. The proposed system can be used both independently and as part of another more complex system.

1. Introduction

In modern highly competitive economics accurate sales (demand) forecasting is necessary for success in various fields. Especially, in the field of air transportation. The proposed methodology can be especially useful in onboard trade and catering. Trading on board an aircraft is associated with great logistical difficulties and the range of products offered cannot be too wide. The introduction of new products into the assortment offered on board an airliner is especially costly and increased requirements are imposed on the forecasting algorithms. In the case of introducing into the assortment goods that will not be in demand, the costs and lost profits in the onboard retail trade would significant exceed those of traditional or e-commerce.

Companies must have the ability to adapt to a frequently changing environment. Accurate sales forecasting is critical for intelligent transportation [1] as far as it can optimize logistics, supply chain management and wholesale management. In retail trade (especially in E-commerce) one needs to manage assortment, so it is necessary to know which goods will be popular in the future. Accurate sales forecasting is certainly an inexpensive way to meet the goals, since this leads to improved customer service, reduced lost sales and product returns and more efficient production planning [2]. In production, forecasting is also needed to decide what types of goods to produce. In management, decision makers can optimize processes based on accurate forecasts. The problem of product demand forecasting is
especially critical in food retail where overestimated demand can lead to the spoilage of short shelf-life food products.

The proposed paper is based on data provided by Russian e-commerce company Ozon. The Ozon marketplace is a service with over 40 million unique visitors per month and more than 180,000 orders per day. So, even small improvement in forecast accuracy can lead to significant increase of company profit and customers can buy quicker and at lower price.

Literature research shows that there are several common types of papers related to the forecasting product demand. One of them uses marketing research methods like [3]. Marketing research needs a lot of human resources and it is highly unlikely they can be automatized. Other researchers [4] use classical statistic methods like Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH, etc. applied to historical data time series. Although such methods are well-developed and have high efficiency, we cannot use only them, because we work with new goods. In turn it means we do not have historical data for them. There are several types of new product introductions: new-to-the-world products, new-to-the-firm products, additions to existing product lines, improvements and revisions of existing products, etc. and none of them have historical data. Also, many researchers provide techniques which work well in special cases. For example, forecasting method for book sales proposed in [5], for 3D TV in [6], for oral care goods in [7].

The general objective of this work is to propose a forecasting method based on machine learning models to forecast the demand of new products satisfying the following conditions:

- does not depend on the type of goods;
- can work without history of sales;
- can work with large data set;
- no need for marketing research.

2. Methods

2.1. Machine learning

The machine learning discipline contains a lot of powerful methods to solve complex problems. There are several definitions of machine learning. One of the most reasonable is as follows: “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.” [8]. So, if our problem is too complex to be solved directly (with analytical or numeric methods) and we have enough data we can try to build a machine learning model able to extract knowledge from this data. During recent years, we can observe an increasing interest in this field due to several amazing results in different domains; for example, image processing, natural language processing, recommendation systems, games and even art. Such progress in developing machine learning algorithms can be explained by several reasons: the growing available amount of data, growing computational performance and developing effective algorithms.

Common machine learning project workflows consist of five main stages presented in figure 1 [9]. Usually, it is not enough to perform these procedures only one time. Several cycles are needed until the goal is achieved. Although there are automated machine learning (AutoML) techniques which try to automatically cover the complete pipeline, nowadays it does not work well when dealing with complex problems, so we need to release all the stages manually.
2.2. Data set description, feature engineering, validation and metric
At the stage of data exploration available data were examined. There were defined types, distributions, percentages of missing values and information was received from domain knowledge experts to understand which data is possibly useful for the problem solution. After that a data cleaning procedure was performed – outliers were removed, some rows with missing values were dropped. After all these procedures, the data set consisted of more than 4.5 billion samples of sales which were made from 2012 to 2020. There were data for about 89 000 items. An example of one item data is presented in figure 2. All data were divided by 24 categories like “Sport”, “Zoo”, “Furniture”, “Books” etc.

The next important procedures in the machine learning pipeline are feature processing and generation. All the data presented in this paper were taken from Ozon company database which stores the history of sales, pre-processed by means of Pandas library, and visualized using Matplotlib library. There are two commonly used approaches of feature generation – use already trained models for feature extraction or manually create features based on the domain knowledge. Figure 3 presents sales distributions by month and figure 4 presents sales by date. Obviously, time-dependent features contain useful information for our problem. Other important features are price, promotion and category. Although we could not use data about the history of each item’s sales, we used historical data grouped by categories. Examples of formulated features include such features as: “Average brand sales within a
group of products in the first week of products in stock”, “The ratio of price to average price of products from the same subtype” and so on. Our hypotheses about the importance of features will be checked after model learning. A summary about features is presented in table 1.

Table 1. Features types description.

| Category              | Features | Description                                                                 |
|-----------------------|----------|-----------------------------------------------------------------------------|
| Date                  | 8        | Features based on date: day of week, weekend etc.                           |
| Price and promotions  | 3        | Information about price and sales                                          |
| Identity              | 1        | Category                                                                    |
| Aggregates            | 2        | Averages over all available data on different aggregation levels            |

Figure 3. Sales distributions by month.  

Figure 4. Sales by date.  

For the correctness of training, validation and testing procedures, data were divided according to the rules of time series cross-validation as presented in figure 5. This avoided both types of data leaks from future and between items.

Figure 5. Time series and items cross validation.
In order to make reliable conclusions about the accuracy of the proposed model appropriate metric were chosen. Although there are a lot of different metrics, for example, Geometric Mean of the Relative Absolute Error (GMRAE) and Median Absolute Percentage Error (MdAPE) [10], standard Root-Mean-Square Error (1) (RMSE) was used for this research. The RMSE is one of the simplest metrics frequently used to draw conclusions about model accuracy both in time series forecasting and in standard regression models.

\[
RMSE = \left( \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N} \right)^{1/2}
\]

where \(x_i\) are actual sales values, \(\hat{x}_i\) are predicted sales values and \(N\) is the sample size.

2.3. Gradient Boosting

The problem of forecasting new product demand can be treated as regression task. In this type of task, the computer program is asked to predict a numerical value given some input. To solve this task, the learning algorithm is asked to output a function \(f : \mathbb{R}^n \rightarrow \mathbb{R}\) This type of task is similar to classification, except that the format of output is different [11]. One of the best methods to solve regression task is gradient boosting. Boosting is a powerful technique for combining multiple ‘base’ classifiers to produce a form of committee whose performance can be significantly better than that of any of the base classifiers [12]. The advantages of boosting trees: natural handling of data of “mixed” type, handling of missing values, robustness to outliers in input space, computational scalability and, of course, high accuracy. Due to computational efficiency gradient decent learning algorithm described in table 2 was used [13].

There are three well-developed packages belonging to the family of gradient boosting decision trees: XGBoost, LightGBM and CatBoost. Each of them has advantages and disadvantages. For example, CatBoost shows the best results when the data contains many categorical features. In our case, the best result was achieved with LightGBM package.

**Table 2. GTB Algorithm description.**

**Algorithm 1: Gradient Tree Boosting for Regression.**

1. Initialize \(f_0 = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma).\)
2. For \(m = 1\) to \(M:\)
   a. For \(m = 1, 2, ..., N\) compute
      \[
      r_{im} = -\left[ \frac{\partial L(y_i,f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}} .
      \]
   b. Fit a regression tree to the targets \(r_{im}\) giving terminal regions \(R_{jm}, j = 1, 2, ..., J_m.\)
   c. For \(j = 1, 2, ..., J_m,\) compute
      \[
      \gamma_{jm} = \arg\min_{\gamma} \sum_{x \in R_{jm}} L(y_i,f_{m-1}(x_i) + \gamma) .
      \]
   d. Update \(f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}).\)
3. Output \(\hat{f}(x) = f_m(x).\)

where \(L(y, f(x))\) is a loss function, \(M\) is a number of iterations and \(J_m\) are the sizes of each of the constituent trees.

The gradient tree boosting algorithm has several hyperparameters like learning rate, maximum depth and minimum data in leaf needed to be adjusted. Grid search and manual search are the most widely used strategies for hyper-parameter optimization, but random search strategy for hyper-parameter was
used because in [14] it was shown empirically and theoretically that this optimization strategy is more efficient.

3. Results

The resulting learning curves are presented in figure 6. There are two important results displayed in learning curves. Firstly, despite the complex data, the error decreases during the learning process, which indicates that the algorithm is being trained correctly. Secondly, no increase is observed in the error on the validation set, which means that the effect of overfitting was successfully avoided, and the model will work correctly on new data.

The resulting feature importances are presented in figure 7. Feature importances show how much, on average, the prediction changes if the feature value changes. The most important features are date, price, and discount, but days of the week and aggregated features also contribute to final accuracy. The values of feature importances play an important role in the selection of features and have been used for this purpose.

![Figure 6. Learning curves.](image)

The best accuracy achieved at testing data set is RMSE = 4.00129. There are no published articles with results on the same data set, so it is not possible to compare results directly. This result was achieved at hyper-parameters presented in table 3. This parameter can be used by other researchers to solve the same or similar problem. In order to clarify how this result was obtained optimization table fragment is shown in table 4.

| Hyper-parameter          | Value  |
|--------------------------|--------|
| Learning rate            | 0.051  |
| Number of leaves         | 64     |
| Minimum data in leaf     | 12     |
| Number of iterations     | 243    |
There are a lot of forecasting techniques. Most popular methods are described in a detailed review - “A review of forecasting models for new products” [15]. But none of them are suitable for direct comparison with the proposed method because they do not satisfy the previously mentioned requirements. Disadvantage of the proposed method is that it works only on a large amount of data, which is relatively common for algorithms based on machine learning.

Table 4. Optimization procedure (fragment). Optimal values marked in bold.

| Learning rate | Number of leaves | Iteration | RMSE    |
|---------------|------------------|-----------|---------|
| 0.100         | 32               | 100       | 5.04192 |
| 0.100         | 32               | 200       | 5.04038 |
| 0.100         | 32               | 300       | 5.03282 |
| 0.100         | 32               | 400       | 5.02162 |
|               |                  |           |         |
| 0.051         | 64               | 240       | 4.00314 |
| 0.051         | 64               | 241       | 4.00259 |
| 0.051         | 64               | 242       | 4.00186 |
| **0.051**     | **64**           | **243**   | **4.00129** |
|               |                  |           |         |
| 0.001         | 128              | 100       | 4.20755 |
| 0.001         | 128              | 200       | 4.19667 |
| 0.001         | 128              | 300       | 4.18807 |
| 0.001         | 128              | 400       | 4.18365 |

Figure 7. Feature importance.

4. Conclusion
The proposed model satisfies initial conditions because it meets all the requirements – forecasting demand without any marketing research and to work independent of the type of goods and historical
data, so the goal of the paper has been achieved. This means that community can use this model for predicting software development which could successfully work with very complex problems of predicting new good demand. This model can be used by companies’ analysts for optimizing sales assortment, planning and logistic optimization. Also, this model can be useful for increasing the accuracy of other prediction systems as part of committee. Practical value of the obtained results for the international community in avia-industry is ability to use the proposed method to increase efficiency of logistics in onboard trade and catering. It is important due to high competitiveness of commercial air transportation. Yet another application is usage as a part of marketplace recommendation system. In future studies it is planned to improve accuracy by using features extracted from other types of data like images and videos.

References
[1] Zhenchang X, Xue S, Wu L and Sun J 2020 ForeXGBoost: passenger car sales prediction based on XGBoost. Distrib. Parallel Dat. 38 713 https://doi.org/10.1007/s10619-020-07294-y
[2] Doganis P, Alexandridis A, Patrinos P and Sarimveis H 2006 Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing J. Food Eng. 75(2) 196 https://doi.org/10.1016/j.jfoodeng.2005.03.056
[3] Green K C and Armstrong J S 2007 Structured analogies for forecasting. Int. J. Forecasting 23 365 https://doi.org/10.1016/j.ijforecast.2007.05.005
[4] Pavlyshenko B 2019 Machine-learning models for sales time series forecasting. Data 4 15 https://doi.org/10.3390/data4010015
[5] Sharma S K, Chakraborti S and Jha T 2019 Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. Inf. Syst. E-Bus. Manag. 17 261 https://doi.org/10.1007/s10257-019-00438-3
[6] Lee H, Kim S G, Park H W and Kang P 2014 Pre-launch new product demand forecasting using the Bass model: A statistical and machine learning-based approach. Technol. Forecast. Soc. 86 49 https://doi.org/10.1016/j.techfore.2013.08.020
[7] Vhatkara S and Diasb J 2016 Oral-care goods sales forecasting using artificial neural network model. Procedia Comput. Sci. 79 238 https://doi.org/10.1016/j.procs.2016.03.031
[8] Géron A 2017 Hands-On Machine Learning with Scikit-Learn and TensorFlow (Sebastopol, CA: O’Reilly Media) p 21
[9] Nain N and Behera G 2019 A Comparative Study of Big Mart Sales Prediction 4th International Conference on Computer Vision and Image Processing (Jaipur: MNIT) p 4
[10] Armstrong J S and Collopy F 1992 Error measures for generalizing about forecasting methods: Empirical comparisons. Int. J. Forecasting 8(1) 69-80 https://doi.org/10.1016/0169-2070(92)90008-W
[11] Goodfellow I, Bengio Y and Courville A 2016 Deep Learning (Cambridge: The MIT Press) p 101
[12] Bishop C. M 2006 Pattern Recognition and Machine Learning (New York: Springer) p 657
[13] Hastie T, Tibshirani R and Friedman J 2001 The Elements of Statistical Learning: Data Mining, Inference, and Prediction (New York: Springer) p 361
[14] Bergstra J and Bengio Y 2012 Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 281-305 https://dl.acm.org/doi/10.5555/2188385.2188395
[15] Mas-Machuca M, Sainz M and Martinez-Costa C 2014 A review of forecasting models for new products. Intangible Capital 10(1) 1 https://doi.org/10.3926/ic.482