

# Using Machine Learning to Predict Food Prices in Kazakhstan

Saya Sapakova<sup>1</sup>, Askar Sapakov<sup>2</sup>, Nurgul Madinesh<sup>3</sup>, Ali Almisreb<sup>4</sup> and Meruyert Dauletbek<sup>1</sup>

<sup>1</sup> International Information Technology University, Manas St. 34/1, Almaty, 050000, Kazakhstan

<sup>2</sup> Kazakh National Agrarian Research University, Valikhanov St. 137, Almaty, 050000, Kazakhstan

<sup>3</sup> Information System Department, al-Farabi Kazakh National University, Almaty, Kazakhstan

<sup>4</sup> Faculty of Natural Sciences and Engineering, International University of Sarajevo, Bosnia and Herzegovina

## Abstract

The relevance of using machine learning to predict consumer prices for food in Kazakhstan is due to the need to effectively manage prices and ensure the availability of food products for the population. Machine learning can analyze multiple factors, such as inflation, seasonal fluctuations, and changes in supply and demand, and create accurate forecasts, which is critical in times of economic uncertainty and food price volatility. In this work, we used a machine learning method to predict food prices in Kazakhstan. The main focus was on the application of machine learning methods in this context. We used monthly rural, urban, and mixed-price data from 2005 to 2020 to develop and test the models. The study examines the use of machine learning models such as Decision Tree, Random Forest and Gradient Boosting using a dataset obtained from the World Food Program price database. This helps accurately predict future food prices. We also assessed the accuracy of these models using various metrics such as root mean square error (RMSE), root mean square error (MSE), as a mean absolute error (MAE) and coefficient of determination ( $R^2$ ). In our work, we came to the conclusion that the Random Forest model performed best in all measurement metrics. Experimental results confirm that the high accuracy produced by the Random Forest model has (0.99) in predicting future food price values.

## Keywords

Machine learning, forecasting, consumer price index, random forest, decision tree, gradient boosting.

## 1. Introduction

This article explores the utilization of machine learning methods to predict food prices in Kazakhstan. In the dynamic and variable economic environment associated with agriculture and global markets, accurate food price forecasts become a key factor for both consumers and producers. The article provides an overview of modern machine learning methods that can be successfully applied to analyze food price data, including time series, regression, and neural networks. Next, we discuss the methodology for collecting and processing data necessary for training and evaluating models. The study results show that machine learning can significantly improve the accuracy of food price forecasts in Kazakhstan, which can be useful for making strategic decisions at both the consumer, business and government levels. The methods and results presented in the article can be used in practical applications to optimize the processes of managing food prices and mitigating the impact of economic fluctuations on the population of Kazakhstan. artificial intelligence (AI) and particularly Machine learning play a substantial role in the prediction of prices, offering precise and highly reliable tools for scrutinizing and prognosticating the costs of commodities and services. Below are several ways in which machine learning and AI are exerting an impact on this procedure:

---

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

✉ s.sapakova@iitu.edu.kz (S. Sapakova); Sapakovaskar@gmail.com (A. Sapakov); nurcagana84@gmail.com (N. Madinesh); a.almisreb@iitu.edu.kz (A. Almisreb); m.dauletbek@iitu.edu.kz (M. Dauletbek)

ORCID 0000-0001-6541-6806 (S. Sapakova); 0000-0001-5708-7862 (A. Sapakov); 0000-0001-9376-8489 (N. Madinesh); 0000-0001-7581-5747 (A. Almisreb); 0009-0005-5569-4980 (M. Dauletbek)



© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

**Examination of Extensive Data:** Machine learning proficiently assesses substantial quantities of pricing data, encompassing historical records, supply and demand statistics, macroeconomic indicators, and more. This aids in the identification of concealed patterns and trends that can exert an influence on price dynamics.

**Prognostication of Temporal Trends:** Machine learning methodologies, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), are adept at constructing models for predicting prices grounded on temporal sequences. This is especially crucial for the anticipation of price fluctuations in financial markets and commodity sectors.

**Sentiment Analysis:** AI can scrutinize social media, news outlets, and alternative information sources to gauge public opinion and sentiments capable of influencing price movements. For instance, sentiment analysis of social media posts can furnish insights into market reactions in response to news and occurrences.

**Enhanced Pricing Strategy:** Machine learning aids organizations in enhancing pricing strategies by taking into account a diverse array of considerations, including competitive pricing, cost structures, seasonal variations, and even individual consumer preferences.

**Risk Mitigation:** Machine learning models can be deployed for the evaluation of financial risks linked to commodity and asset prices. This enables businesses and investors to formulate effective risk management tactics.

**Augmented Accuracy and Velocity of Forecasts:** Through the utilization of machine learning and AI, it is possible to automate the forecasting process and handle data in real-time, culminating in forecasts that are not only more precise but also more swiftly generated.

In summation, machine learning and AI have a substantial impact on augmenting the capacity to foretell prices, rendering the process more precise and adaptable to the swiftly evolving landscape of market conditions. This holds profound significance for enterprises, financial institutions, and governmental entities that rely on dependable price forecasts to inform their strategic decision-making.

## 2. Literature review

Nowadays, where agricultural and food markets are becoming increasingly dynamic and subject to various influences, food price forecasting has become an important tool for farmers, consumers and governments. The application of machine learning approaches to this task has attracted increasing attention and research. For instance, in Refs. [1, 2], researchers use Vector Autoregression (VAR) and Nonlinear Autoregressive Distributed Lag (NARDL) techniques to explore the impact of economic policy uncertainty on food prices. They find that rising economic policy uncertainty leads to a substantial reduction in prices. Therefore, in line with existing studies, we anticipate a positive connection between economic policy uncertainty and global food prices. The US Economic Policy Uncertainty Index is employed as an indicator for economic policy uncertainty in their empirical analysis. The rapid rise in food prices, particularly burdensome for the poor in developing countries who allocate for foods approximately half of their income, is primarily attributed to increased biofuel production in the US and the EU. In [3] identifies this as the major factor behind rising internationally traded food prices since 2002. Other contributing factors include a weakened US dollar, higher energy costs in food production, and occasional droughts. This research aligns with many other studies recognizing biofuels as a significant driver of food price increases. It raises important policy concerns, as government policies in the EU and the US incentivized biofuel production, which had a substantial impact on food prices. Reconsidering these biofuel policies in light of their influence on food prices is essential.

The impacts of agricultural output investigated in [4], production material prices, and production prices on China's food prices using panel data from 26 provinces between 2004 and 2015. The research reveals that the primary driver of food prices is price inertia shock, with limited influence from agricultural output or price transmission shocks within the vertical chain, emphasizing the importance of price expectations in shaping food prices. The factors influencing price changes in international food commodity markets considered in [5]. By analyzing various

models, it differentiates between fundamental, conditional, and internal drivers of these changes. The research shows a growing connection between food, energy, and financial markets, contributing significantly to observed food price spikes and volatility. Short-term price spikes are amplified by financial speculation, while medium-term volatility is intensified by fluctuations in oil prices. In [6] showed that statistical analyses linking agricultural yields to weather fluctuations indicate that high temperatures are a key factor in predicting outcomes. Climate change is projected to raise these temperatures, leading to yield reductions. These models have demonstrated their ability to effectively forecast future results. For example, a model based on historic US corn and soybean yield data from 1950 to 2011 accurately predicted yields for 2012–2015, a period marked by higher-than-average temperatures due to climate change. The implications of declining yields are discussed, with a particular focus on adaptation strategies and their impact on commodity prices.

Numerous scholars examining food price dynamics in the existing body of research incorporate a diverse array of factors into their investigations. In recent years, advanced methodologies such as Autoregressive Distributed Lag (ARDL), causality and cointegration analysis (as proposed by Granger and Toda-Yamamoto), Dynamic Conditional Correlation (DCC), Generalized Method of Moments (GMM), Nonlinear Autoregressive Distributed Lag (NARDL), Generalized Autoregressive Conditional Heteroscedasticity (GARCH), Panel Data Analysis, and various regression techniques (including Ordinary Least Squares (OLS) and pooled-OLS), Vector Autoregression (VAR), and Vector Error Correction Model (VECM) have been employed.

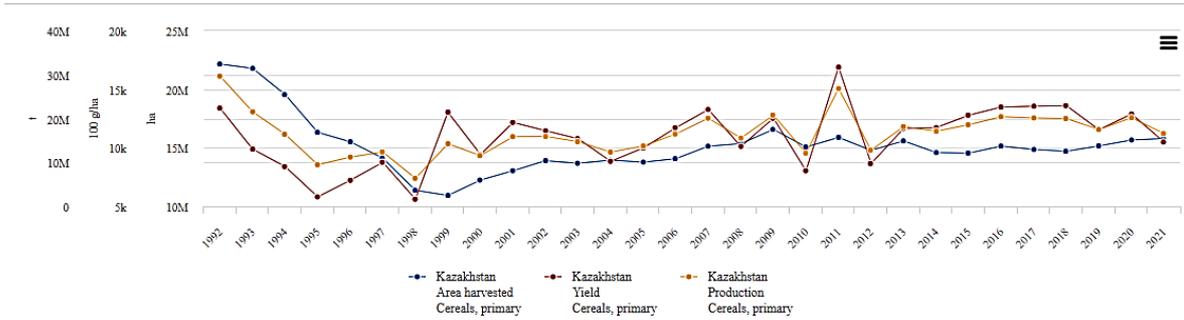
Recent research findings demonstrate the compelling outcomes achieved by employing ensemble learning algorithms. Ensemble methods endeavor to formulate a collection of hypotheses and amalgamate them, offering practical solutions to various real-world challenges, including those pertaining to yield prediction [7, 8, 9, 10]. However, there is a notable scarcity of research cases that directly compare the predictive performance of econometric models with machine learning algorithms in this domain. Our research addresses this gap, making a distinctive contribution to the existing literature by focusing on methodological advancements. As a result, the outcomes of our study carry significant implications for nations seeking to effectively manage and stabilize food prices.

### **3. Methods**

#### **3.1. Data**

In this research, we leverage a dataset pertaining to food prices in Kazakhstan, sourced from the World Food Programme's price database. This extensive database covers a wide array of food items, including maize, rice, beans, fish, and sugar, across 98 countries and approximately 3000 markets. It undergoes weekly updates, primarily consisting of monthly data spanning from 2005 to 2020.

As previously discussed, we have employed machine learning methods to achieve precise forecasting. The selection of appropriate models represents a crucial step in attaining high prediction accuracy. Within this article, we have opted to utilize the following regression models: Decision Trees, Random Forest, and Gradient Boosting. Following the implementation of these models, we assess their accuracy by partitioning the dataset into two segments: a training set and a testing set.



**Figure 1:** Cereals, total production [11]

Based on empirical analysis, monthly data frequency aligns with the monthly announcements of global food prices. The study spans from January 1991 to May 2021, encompassing a substantial timeframe.

	date	market	latitude	longitude	category	commodity	unit	priceflag	price
0	#date	#loc+market+name	#geo+lat	#geo+lon	#item+type	#item+name	#item+unit	#item+price+flag	#value
1	2005-11-15	Nur-Sultan	51.17	71.43	cereals and tubers	Potatoes	KG	actual	40.0
2	2005-11-15	Nur-Sultan	51.17	71.43	cereals and tubers	Wheat flour (first grade)	KG	actual	61.0
3	2005-11-15	Nur-Sultan	51.17	71.43	meat, fish and eggs	Meat (beef)	KG	actual	403.0
4	2005-11-15	Nur-Sultan	51.17	71.43	milk and dairy	Milk	L	actual	58.0
...	...	...	...	...	...	...	...	...	...
3361	2020-11-15	Kostanay	53.2	63.63	oil and fats	Oil (sunflower)	L	actual	514.0
3362	2020-11-15	Aktau	43.65	51.2	cereals and tubers	Potatoes	KG	actual	147.0
3363	2020-11-15	Aktau	43.65	51.2	cereals and tubers	Wheat flour (first grade)	KG	actual	241.0
3364	2020-11-15	Aktau	43.65	51.2	meat, fish and eggs	Meat (beef)	KG	actual	1875.0
3365	2020-11-15	Aktau	43.65	51.2	oil and fats	Oil (sunflower)	L	actual	586.0

**Figure 2:** Data used in the study

The dataset encompasses 3365 rows and 11 columns, encompassing information on various aspects such as commodities, pricing, and market dynamics in Figure 2. This dataset was meticulously gathered from diverse regions within Kazakhstan. To import this dataset, we harnessed the power of Pandas along with an array of libraries tailored for both visual and statistical analysis. Subsequently, we subjected the input data, which includes market, category, product, and unit of measure, to a transformation process. Utilizing Scikit-learn's OneHotEncoder function, we effectively converted these categorical variables into numerical representations, thereby assigning numeric values to each of these pertinent factors. Following necessary data preprocessing, diverse models are deployed and assessed in this study. Model evaluation employs a data splitting technique where 67% of the data serves as the training set, while the remaining 33% is allocated for testing. The initial phase of the research involves data training, which also serves as the model-fitting process. Subsequently, the second phase assesses the model's accuracy.

### 3.2. Methods

This section elaborates more on the proposed methods such as Decision Tree, Random forest, and Gradient boosting.

#### Decision tree

Decision trees (DT), serve as versatile tools applicable to both classification and regression tasks. These algorithms organize datasets into a tree-like structure consisting of a root node,

intermediate nodes, and leaf nodes. Within this framework, each internal node corresponds to one of the available input variables, determined through criteria such as information gain for classification challenges and standard deviation reduction for regression scenarios. The leaves, on the other hand, signify the ultimate predictions or labels. It's worth noting that random forest and gradient boosting, two prominent algorithms, are fundamentally rooted in decision tree principles.

In the context of this study, we specifically employ the decision tree method for regression problems. Our approach involves the reduction of variance to guide variable selection at internal nodes. Initially, we calculate the variance of the root node using a specific equation, and subsequently, we employ a similar equation to compute the variance of the features, a process instrumental in constructing the decision tree (1).

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}, \quad (1)$$

where in (1),  $n$  is the total number of samples, and  $\mu$  is the average of the samples in the training set. After calculating the variance of the root node, the variance of the input variables is calculated as follows:

$$\sigma_X^2 = \sum_{c \in X} P(c) \sigma_c^2, \quad (2)$$

where in (2),  $X$  is the input variable, and  $c$  is the different values of this attribute.  $P(c)$  is the probability that  $c$  is in attribute  $X$  and  $\sigma_c^2$  is the change in the value of  $c$ . The input variable that has the smallest variance or the largest reduction in variance is selected as the best node, as shown in the equation below:

$$vr_X = \sigma^2 - \sigma_X^2. \quad (3)$$

This iterative procedure persists until the divergence at the leaf node falls below a specified threshold, or alternatively, until all input variables have been exhausted. After the construction of the tree is completed, the evaluation of a new instance is facilitated through an inquiry process directed at the nodes within the tree structure. Upon arriving at a leaf node, the prediction is derived from the value associated with that particular leaf.

### **Random forest**

Random Forest (RF) is a sophisticated form of meta-learning, strategically employing a multitude of decision trees to tackle classification and regression tasks. In the Random Forest paradigm, a key feature is the random selection of objects and samples for each tree within the forest. These individual trees are trained autonomously, contributing to the collective wisdom of the forest. This widely adopted machine learning algorithm, characterized by its assembly of multiple decision trees, diverges from conventional methods that simply average the predictions of these trees. Instead, it harnesses the power of randomness through the selection of training data points and subsets of objects to create node splits.

#### **Advantages of RF:**

- Proficiency in handling datasets replete with a profusion of features and categories.
- Remarkable resilience to variations in the scaling of feature values.
- Availability of techniques for gauging the significance of individual features within the model.
- Exceptional precision in the classification of novel data points.
- Versatility in conducting measurements across divergent scales, encompassing numerical, ordinal, and nominal realms.

#### **Disadvantages:**

- Requisition of substantial memory resources for model storage, which can be quite prodigious.
- Languid computational speed when dealing with missing data values, a drawback that warrants consideration.

### Gradient boosting

Gradient boosting represents the process of elevating the proficiency of less capable learners into more proficient ones. Typically, the less capable learner takes the form of a decision tree that exhibits subpar performance in data classification. Each additional tree incorporated into the boosting scheme is essentially a modified version of the original dataset. Boosting can be conceptualized as a numerical optimization challenge with the objective of minimizing the loss function inherent to the model by employing a gradient descent approach when introducing new, less capable learners.

In essence, gradient enhancement follows a stepwise, cumulative model-building approach. It systematically trains multiple models in a sequential and incremental manner. When a new less capable learner is introduced, the previously incorporated learners remain unaltered within the model. Importantly, gradient boosting offers the capacity to optimize diverse loss functions that are characterized by their differentiability.

$$F_0(x) = \operatorname{argmin} \sum_{i=1}^n L(y_i, \gamma), \quad (4)$$

in the realm of our discourse,  $\gamma$  alludes to the anticipated outcomes, while the term "argmin" signifies the imperative criterion of assigning to  $F_0(x)$  the predicted value that minimizes the summation denoted by  $\Sigma$ .

The computation of  $F_0(x)$  entails the differentiation of the summation related to the loss function, followed by its equating to zero. This operation, in turn, furnishes us with the mean of the observed values of the dependent variable. This inaugural predictive value assumes the role of the primary leaf node within the framework of the gradient-boosting algorithm.

For  $m=1$  to  $M$

$$r_{im} = - \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \quad \text{for } i=1, \dots, n$$

For  $j=1 \dots J_m$   $\gamma_{jm} = \operatorname{argmin} \sum_{x \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \quad (5)$$

In this study, the data set is divided using the specified 67%/33% ratio. Once preprocessing and data division are complete, models like Decision Tree, Random Forest, and Gradient Boosting are implemented, and their respective accuracies are evaluated.

The assessment of algorithm performance was conducted by measuring key metrics, including mean absolute error (MAE), root mean square error (RMSE), mean square error (MSE), and coefficient of determination ( $R^2$ ). The algorithms demonstrating minimal errors are indicative of superior accuracy and are thus deemed preferable. MAE quantifies the cumulative absolute disparities between actual and predicted variables, while MSE represents the mean of squared errors. The optimization of the loss function, defined as the sum of squared differences between observed values ( $y_i$ ) and predicted values ( $\hat{y}_i$ ), is employed to calculate the best-fitting linear regression line, as depicted in the following equation:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

where,  $n$  is the number of observations or data points.  $y_i$  is the actual value (observed value) for the  $i$ -th observation,  $\hat{y}_i$  is the predicted value generated by the model for the  $i$ -th observation. A smaller MSE value indicates that the model is better at approximating the data. A high MSE value suggests that the model is making significant errors in its predictions. MSE is also widely

used in machine learning during model training because it can be minimized during the parameter optimization process.

### **RMSE**

The Root Mean Square Error (RMSE) stands as a prevalent metric within the realms of statistics and machine learning, serving as a pivotal tool for assessing the precision of predictive models, especially in the domain of regression tasks. RMSE, an offshoot of the Mean Square Error (MSE), furnishes a quantification of the extent to which a model's predictions align with the observed empirical data.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

where:

$n$  is the number of observations or data points;

$y_i$  is the actual value (observed value) for the  $i$ -th observation;

$\hat{y}_i$  is the predicted value generated by the model for the  $i$ -th observation;

$\Sigma$  represents the summation of squared differences across all data points.

RMSE serves as a singular metric encapsulating the typical scale of discrepancies between observed and forecasted values. Diminished RMSE values signify a model with predictions that, on average, closely align with the real data, implying a superior model fit. Conversely, elevated RMSE values indicate less precise predictions, signifying an inferior fit to the dataset.

### **R<sup>2</sup> (the coefficient of determination)**

The coefficient of determination, commonly represented as R-squared ( $R^2$ ), quantifies the fraction of variance in the dependent variable elucidated by the independent variables within a statistical model. Its scale spans from 0 to 1, with higher values signifying a more substantial proportion of the dependent variable's variability being elucidated by the independent variables, thus suggesting an improved model fit.

$$R^2 = 1 - \frac{SSE}{SST}$$

where:

**SSR** is the sum of squared residuals, which represents the sum of the squared differences between the observed values and the predicted values by the model;

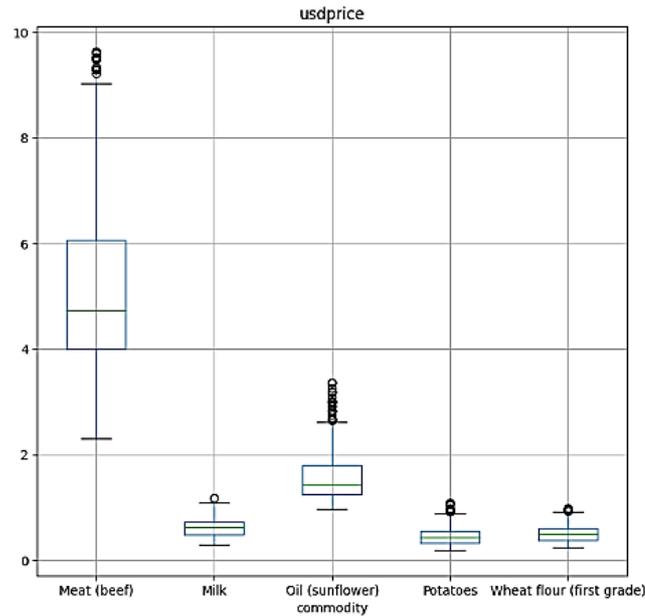
**SST** is the total sum of squares, which represents the sum of the squared differences between the observed values and the mean of the dependent variable.

$$\sum (y^i - \hat{y})^2,$$

where  $y_i$  is the observed value,  $y^{\wedge}$  is the predicted value.

## **3.3. Analyses of results**

In this study, we have employed a machine learning methodology to predict consumer food prices in Kazakhstan. Our primary emphasis has been on assessing the effectiveness of machine learning techniques in this context. Forecasting consumer demand plays a pivotal role in the retail sector, as it aids in cost reduction, maintaining appropriate inventory levels, optimizing warehouse utilization, and mitigating out-of-stock issues. Precisely predicting future demand poses a significant challenge for retailers and wholesalers, owing to abrupt shifts in demand patterns, limited historical data, emerging trends, and seasonal spikes in demand.



**Figure 3:** Price chart for food products

**Table 1**  
**Comparison of the results**

Indicator name	Decision Tree	Gradient Boosting	Random Forest
MAE	0.13	0.15	0.042
RMSE	0.29	0.27	0.085
MSE	0.08	0.074	0.007
R2	0.97	0.98	0.99

In our research, we analyzed the performance of three different models on consumer food products. The results indicate that the Root Mean Square Error (RMSE) values were 0.29 for the Decision Tree model, 0.27 for Gradient Boosting, and 0.085 for Random Forest. Notably, Random Forest outperformed the other models, while Decision Trees and Gradient Boosting exhibited similar performance levels.

Conclusions on the performance evaluation of the Decision Tree algorithm:

- Mean Absolute Error (MAE) indicates that the difference between actual and predicted food prices is \$0.13;
- Root mean square deviation (RMSE) is 0.29;
- Mean square error (MSE) 0.08;
- The coefficient of determination (R2) is 0.97 – this indicates that the agreement between real data and the data obtained as a result of using the model is high and the model can be considered good.

The following results were obtained for the Random Forest algorithm:

- Mean absolute error (MAE) indicates that the difference between actual and predicted food prices is \$0.04;
- Standard deviation is 0.08;
- The coefficient of determination (R2) is 0.99 – that is, the correspondence of real data with the data obtained as a result of using the model is high and the model can be considered excellent.

Additionally, the R-Square values for all models approached 1, indicating a high degree of accuracy, with 98% and 99% correspondence between real data and model predictions. This suggests that our models are quite effective. Further analysis revealed that the Random Forest model had a mean squared error of 0.085, with a minimal difference of 0.042 between actual and predicted food prices.

## 4. Conclusion

In this study, we investigated established approaches for predicting sales volumes and opted for machine learning techniques to construct our model. We identified three machine learning algorithms, namely random forest, gradient boosting, and decision tree, as the most effective for sales volume prediction. Our primary criterion for these models was achieving the highest level of accuracy. Using these selected algorithms, we constructed three distinct models. Employing key quality assessment metrics such as mean absolute error, standard deviation, and coefficient of determination, we determined that the boosting algorithm yielded the most precise forecasts with an impressive accuracy rate of 99%. Furthermore, our research focused on evaluating the performance of these models in the context of consumer food products. Intriguingly, our findings and subsequent discussion indicated that the Random Forest model outperformed the others in terms of accuracy. This discovery holds particular significance for busy decision-makers. In today's volatile economic landscape, characterized by heightened competition in the trade sector, organizations are increasingly in need of tools that enable seamless and uninterrupted sales processes. This, in turn, facilitates more precise allocation of financial resources within the organization. Anticipating the challenges associated with demand forecasting and how to effectively address them remains a complex analytical trend in the upcoming year. This complexity arises from the diverse and ever-evolving desires of consumers. Even top-tier manufacturers may struggle if they fail to align their production with consumer preferences. As such, the relevance of machine learning models, featuring a range of algorithms for assessing and responding to consumer demand, continues to persist. This is especially true for fast-moving consumer products, such as food or fashion, where real-time forecasting is essential.

## 5. References

- [1] Xiao, X.; Tian, Q.; Hou, S.; Li, C. (2019). Economic Policy Uncertainty and Grain Futures Price Volatility: Evidence from China. *China Agric. Econ. Rev.* 11, 642–654.
- [2] P. S. Abril, R. Plant, The patent holder's dilemma Wen, J.; Khalid, S.; Mahmood, H.; Zakaria, M. (2021). Symmetric and Asymmetric Impact of Economic Policy Uncertainty on Food Prices in China: A New Evidence. *Resour. Policy.* 74, 102247.
- [3] Mitchell, D. A Note on Rising Food Prices; World Bank Policy Research Working Paper, No. 4682; World Bank-Development Economics Group: Washington, DC, USA, 2008.
- [4] Wu, X.; Xu, J. (2021). Drivers of Food Price in China: A Heterogeneous Panel SVAR Approach. *Agric. Econ.* 52: 67–79.
- [5] Tadasse, G.; Algieri, B.; Kalkuhl, M.; Von Braun, J. (2014). Drivers and Triggers of International Food Price Spikes and Volatility. *Food Policy.* 47: 117–128.
- [6] D'Agostino, A.L.; Schlenker, W. (2016). Recent Weather Fluctuations and Agricultural Yields: Implications for Climate Change. *Agric. Econ.* 47: 159–171.
- [7] Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Comput. Electron. Agric.* 121: 57–65.
- [8] Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors* 18, p.2674. <https://doi.org/10.3390/s18082674>.
- [9] Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sens. Environ.* 164: 324–333. <https://doi.org/10.1016/j.rse.2015.05.020>.

- [10] Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J. (2016). Analysis of big data technologies for use in agro-environmental science. *Environ. Model Softw.* 84: 494–504. <https://doi.org/10.1016/j.envsoft.2016.07.017>.
- [11] FAOSTAT. Kazakhstan, 2023. URL: <https://www.fao.org/faostat/en/#country/108>.
- [12] Huang, J., Chai, J. & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Front. Bus. Res. China* 14 (13). <https://doi.org/10.1186/s11782-020-00082-6>.
- [13] Haselbeck, F., & Grimm, D. G. (2021). Evars-GPR: Event-triggered augmented refitting of Gaussian process regression for seasonal data.
- [14] In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Lecture Notes in Computer Science: vol. 12873, Ki 2021: 44th German Conference on AI, Virtual Event, September 27 - October 1, 2021, Proceedings (Vol. 12873) (pp. 135–157)*. Springer, [http://dx.doi.org/10.1007/978-3-030-87626-5\\_11](http://dx.doi.org/10.1007/978-3-030-87626-5_11).
- [15] Liu, B., Qi, Y., & Chen, K.-J. (2020). Sequential online prediction in the presence of outliers and change points: An instant temporal structure learning approach. *Neurocomputing*, 413: 240–258. <http://dx.doi.org/10.1016/j.neucom.2020.07.011>.
- [16] Salini Suresh; Suneetha V; Niharika Sinha; Sabyasachi Prusty; Sriranga H.A. 92020). *Machine Learning: An Intuitive Approach In Healthcare*". *International Research Journal on Advanced Science Hub*, 2, (7): 67-74. doi: 10.47392/irjash.2020.67.
- [17] Singh N, Singh P, Singh KK, Singh A. (2020) Machine learning based classification and segmentation techniques for CRM: a customer analytics. *J. Bus Forecast Mark Intell, Int.* <https://doi.org/10.1504/IJBFMI.2020.10031824> (InPress).
- [18] Asteriou, D., & Hall, S. g. (2021). *Applied Econometrics*, 4rd edition. Bloomsbury.
- [19] Borenstein, S. (2018). Trying to unpack California’s Mystery Gasoline Surcharge. “Research that Informs Business and Public Policy”.
- [20] EIA (2022). U.S. Energy Information Administration released on retail gas price in California.
- [21] Frank, S. (2019). Reasons for California high gas prices. “California Political Review”.
- [22] Mugabe, D., Elbakidze, L., & Carr, T. (2021). All the DUCs in a row: Natural gas production in U.S. *The Energy Journal*, 42(3).
- [23] Trupti S. Gaikwad; Snehal A. Jadhav; Ruta R. Vaidya; Snehal H. Kulkarni. (2020). *Machine learning amalgamation of Mathematics, Statistics and Electronics*". *International Research Journal on Advanced Science Hub*, 2, (7): 100-108. doi: 10.47392/irjash.2020.72.
- [24] Singla, C., & Sahoo, A.K, (2019). Modelling Consumer Price Index: An Empirical Analysis Using Expert Modeler, *Journal of Technology Management for Growing Economies*, 10(1): 43-50. <http://dx.doi.org/10.15415/jtmge.2019.101004>.
- [25] Sarangi, P.K., Chawla, M., Ghosh, P., Singh, S., Singh, P.K. (2021). FOREX trend analysis using machine learning techniques: INR vs USD currency exchange rate using ANN-GA hybrid approach, *Materials Today: Proceedings*, <https://doi.org/10.1016/j.matpr.2020.10.960>.
- [26] Sarker, I.H. (2021). *Machine Learning: Algorithms, Real-World Applications and Research Directions*. *SN COMPUT. SCI.* 2, 160. <https://doi.org/10.1007/s42979-021-00592-x>.
- [27] Schmidt, J., Marques, M.R.G., Botti, S. et al. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* 5, 83. <https://doi.org/10.1038/s41524-019-0221-0>.