

Forecasting Demand for Food Delivery Services

Yurii Kryvenchuk, Nazarii Hryhorash

Lviv Polytechnic National University, 12 Stepan Bandera Street, Lviv, 79013, Ukraine

Abstract

This research article examines the relevance of forecasting demand for food delivery services and explores various artificial intelligence methods to optimize this process. The growing popularity of food delivery services creates the need for improved forecasting systems to ensure efficiency and meet the growing needs of consumers.

Keywords

Demand forecasting, food delivery, artificial intelligence, model, optimization, machine learning analysis.

1. Introduction

In today's world, where technology has already penetrated all spheres of life, the food delivery industry is becoming increasingly important. It affects the lives of everyone who devotes most of their time to work, study, and family and does not have enough time or desire to cook. In this context, companies offering food delivery services need to reliably forecast demand to prevent their customers from switching to competitors. In addition, correct forecasts help companies to use their resources rationally, reducing costs and contributing to sustainable development.

This research article is devoted to the study of the use of various artificial intelligence methods to predict the demand for food delivery services. It is worth noting that artificial intelligence has already proven to be effective in solving similar problems, including the tasks of modeling and forecasting demand in various fields.

The artificial intelligence methods that will be analyzed for their application in this study include regression, random forest, gradient boosting, lasso, and decision tree. All of these methods have their strengths and limitations, and they have already been used to solve a wide range of problems.

Regression, for example, is one of the most widely used forecasting methods. It allows to estimate the relationship between two or more characteristics. Random Forest creates a 'forest' of algorithms that allows for more accurate forecasts. Gradient Boosting is a method that uses a sequential improvement of predictive models. It allows you to work with different types of data and is extremely effective for recognizing complex patterns. Lasso (Least Absolute Shrinkage and Selection Operator) helps to reduce model complexity. Decision Tree allows you to analyze data using a binary algorithm.

Thus, the purpose of this article is to explore ways to use these artificial intelligence methods to predict the demand for food delivery services to help companies in the industry become more efficient and competitive.


2. Related Works

The integration of innovative artificial intelligence technologies into logistics management has recently gained particular relevance in the context of research on the functioning of social enterprises. Existing scientific developments reveal key aspects of this topic, in particular, related to the efficiency of logistics in social enterprises.

Thus, it is worth noting the study of T. O. Shmatkovska, who argues that social enterprises, due to their specific purpose of solving social problems, often face complicated logistics tasks. At the same

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

 yurii.p.kryvenchuk@lpnu.ua (Yu. Kryvenchuk); nazarii.hryhorash.knm.2020@lpnu.ua (N. Hryhorash);

 0000-0002-2504-5833 (Yu. Kryvenchuk); 0009-0005-7162-1490 (N. Hryhorash)



© 2024 Copyright for this paper by its authors.

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

time, the author identifies the importance of efficient logistics in social enterprises and adds that optimization of logistics processes will improve their ability to solve social problems [7, 8].

The study of N. Y. Kyrlyk is aimed at studying the impact of innovative artificial intelligence technologies on the optimization of logistics processes. It identifies the positive impact of artificial intelligence on efficiency, cost reduction, and improved forecasting accuracy, which are key benefits for social enterprises [2].

It is also worth paying attention to the work of M. I. Dziamulych, who notes the growing interest in the social responsibility of enterprises and argues that the use of artificial intelligence to improve logistics processes can help increase the transparency and responsibility of the production supply chain and reduce the negative impact on the environment [1, 10].

In addition, it is worth noting the study of Y. Chalyuk in the field of social entrepreneurship and innovation, which proves that social enterprises have sufficient potential to innovate and have a positive impact on society. The author determines that in this aspect, the integration of artificial intelligence can be a key factor in achieving the strategic goals of social enterprises [3, 4, 5, 6].

The use of linear regression models to process the results of the study is presented in [12, 14]. The author also provides a brief description of the process of forming the input image for further processing. Naomi Altman in [12] analyzes linear regression algorithms in detail and presents the results of linear regression using visual tools.

Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization method used in regression analysis and machine learning. Its main goal is to provide a simpler, less variable analysis using fewer predictive variables (features). It does this by introducing a penalty that drives some model coefficients to zero, thereby reducing model complexity. [14]

Random Forest Classifier [15] is an ensemble machine learning method that takes into account several learning algorithms together when generating a prediction result. Random Forest combines bootstrap aggregation [16] and random feature selection [13] to build a set of decision trees that exhibit manageable variation. Thus, a random forest generates multiple decision trees rather than a single decision tree.

Decision tree is a well-known and often discussed method for classification and further use for forecasting. The algorithm builds a decision tree using a top-down approach, in which a greedy search over a given training data set is used to test each attribute or context at each of the nodes. It calculates entropy and information gain, which is a statistical property used to select which attribute to test at each node of the tree [12, 13, 15].

Gradient Boosting is a machine learning technique for regression and classification tasks that is based on the principle of constructing a new, improved model based on the weaknesses of the previous one. This technique is based on the concept of boosting, which allows combining many weak prediction models to create one powerful model. [11]

In general, current research identifies the potential benefits and challenges of implementing innovative artificial intelligence technologies in the logistics management of social enterprises. However, there is also an objective need to deepen existing developments to ensure the effective development of social enterprises.

3. Methods analysis

3.1. Linear regression

Linear regression is the simplest and most commonly used statistical approach for forecasting. This approach is used when there is a linear relationship between the dependent and independent variables, which has been empirically proven for many different data sets, including food delivery demand.

In the context of predicting the demand for food delivery services, the dependent variable could be the number of orders, and the independent variables could be the price of the service, the time it takes to place an order, the distance from the restaurant to the customer, the restaurant's rating, and so on. Linear regression will use the data to establish linear relationships between the dependent and independent variables. Using these relationships, it predicts the value of the dependent variable. However, this method has its limitations. It assumes that changes in the dependent variable can be

fully explained by changes in the independent variables, which may not be true in the real world. For example, there are many other factors that can affect the demand for food delivery services that were not included in the model, such as weather conditions, seasonal fluctuations, the influence of social media, etc. Therefore, you need to use linear regression carefully and understand its limitations.

However, despite these limitations, linear regression provides insight into what factors may be influencing demand and provides an easy way to start forecasting. With the same caveat that it needs to be used in combination with other methods to produce more accurate forecasts.

3.2. LASSO

The Least Absolute Shrinkage and Selection Operator (Lasso) is a regression modification that not only helps prevent overfitting, but also performs automatic feature selection.

One of the key aspects of the Lasso method is its ability to reduce the influence of less important features. By introducing the L1 penalty, Lasso reduces the weights of some features to zero, thereby eliminating them from the model. This means that when using the Lasso method, the model can make predictions using only important features, and those that are less important are simply excluded. This distinguishes Lasso from classical linear regression, which includes all available features and tries to minimize forecast errors regardless of the degree of influence of individual features. In the case of linear regression, this can lead to overly complex models and overfitting, especially when we have a large number of features and a reduced number of observations.

However, both Lasso and linear regression are models that require linearity between the dependent and independent variables. Both methods can be limited in cases where such linearity is not appropriate and may not perform well if there are high levels of noise or outliers in the data.

3.3. Gradient Boosting

Gradient Boosting is an exciting machine learning technique that uses the concepts of boosting and gradient descent. It creates a strong prediction model by combining simpler models. The basic idea is to use new models to correct the errors of previous models. The process involves creating new simple models (e.g., decision trees) that correct the errors of the previous ones and then combining them into a total.

In the context of predicting demand for food delivery services, these simple models can be built on the basis of various attributes, such as distance from a restaurant, restaurant rating, price, order time, and so on. Each new model improves the existing whole model by minimizing prediction errors. The Gradient Boosting method is extremely flexible: it can be applied to both binary and continuous dependent variables, and can be used for various types of forecasting, from simple regression to classification. However, despite its power, Gradient Boosting can be sensitive to overfitting, especially if the data has a lot of noise or outliers. It can also be relatively computationally intensive, especially when we have very large datasets, due to its iterative nature.

3.4. Decision tree

A decision tree is another powerful tool for predicting demand for food delivery services. It is a machine learning method that makes predictions using a model of decisions made based on the values of input features.

The decision tree works by splitting the dataset into smaller and smaller subsets until it reaches subsets with the same or similar outcome. Each level of the decision tree represents a decision that is made based on one or more attributes. In the case of food delivery service demand, the attributes that can be included in a decision tree can include delivery cost, delivery distance, restaurant rating, and many others. One of the advantages of a decision tree is its interpretability - you can visualize the decision tree and interpret which features are most influential in making a prediction. However, a decision tree can become overtrained if it is not sufficiently constrained, and some values in the data may be too well fitted to the tree, leading to problems with the overall performance of the model on new data.

This method can be particularly useful if the data contain a few important features that have a strong influence on demand, and these features are reflected at the top of the decision tree.

3.5. Random forest

The Random Forest method is one of the most popular ensemble machine learning methods used for classification and regression.

Random Forest aggregates the results of a large number of independently selected decision trees that operate on random subsets of the features and trials that form the model. This method performs predictions by aggregating the results of all trees, which reduces variance and helps avoid overfitting. In the context of predicting demand for a food delivery service, a random forest can be used to identify key factors that influence demand and predict the number of orders using data on previous order trends, prices, distance to the restaurant, delivery time, and other features. However, a "random forest" can be computationally intensive, especially with a large number of features and trees. In addition, it can be difficult to interpret because it includes many decision trees.

Despite these limitations, the random forest is known for its high accuracy and ability to handle complex relationships between features, making it an important tool in forecasting demand for food delivery services

4. Dataset

To test the methods, we chose the dataset used on 'Analytics Vidhya' in the 'Food Demand Forecasting Challenge' (<https://datahack.analyticsvidhya.com/contest/genpact-machine-learning-hackathon-1>).

The dataset simulates the following situation: your client is a food delivery company that operates in several cities. In these cities, they have different fulfilment centers to send food orders to their customers. The client wants you to help these fulfilment centers predict the demand for the following weeks so that these centers can plan their raw material inventory accordingly. The dataset consists of three tables: train.csv - contains historical order data for all centers see table 1, fulfilment_center_info.csv - contains information about each delivery center see table 2, meal_info.csv - contains information about each meal see table 3.

Table 1

Table train.csv

Field	Meaning
Id	A unique identifier
Week	Number of the week
Center_id	Unique identifier of the delivery centr
Meal-id	Unique identifier for food
Checkout_price	The final price includes discount, taxes and delivery costs
Base_price	Base price of the meal
Emailer_for_promotion	Whether the email discount was used
Hamepage_featured	Whether the food is available on the homepage
Num_orders	Number of orders

After merging the tables, I got a table consisting of 15 columns and 521565 records. I gave it to the data analysis. First, I will identify the delivery centers with the largest number of orders see Figure 1.

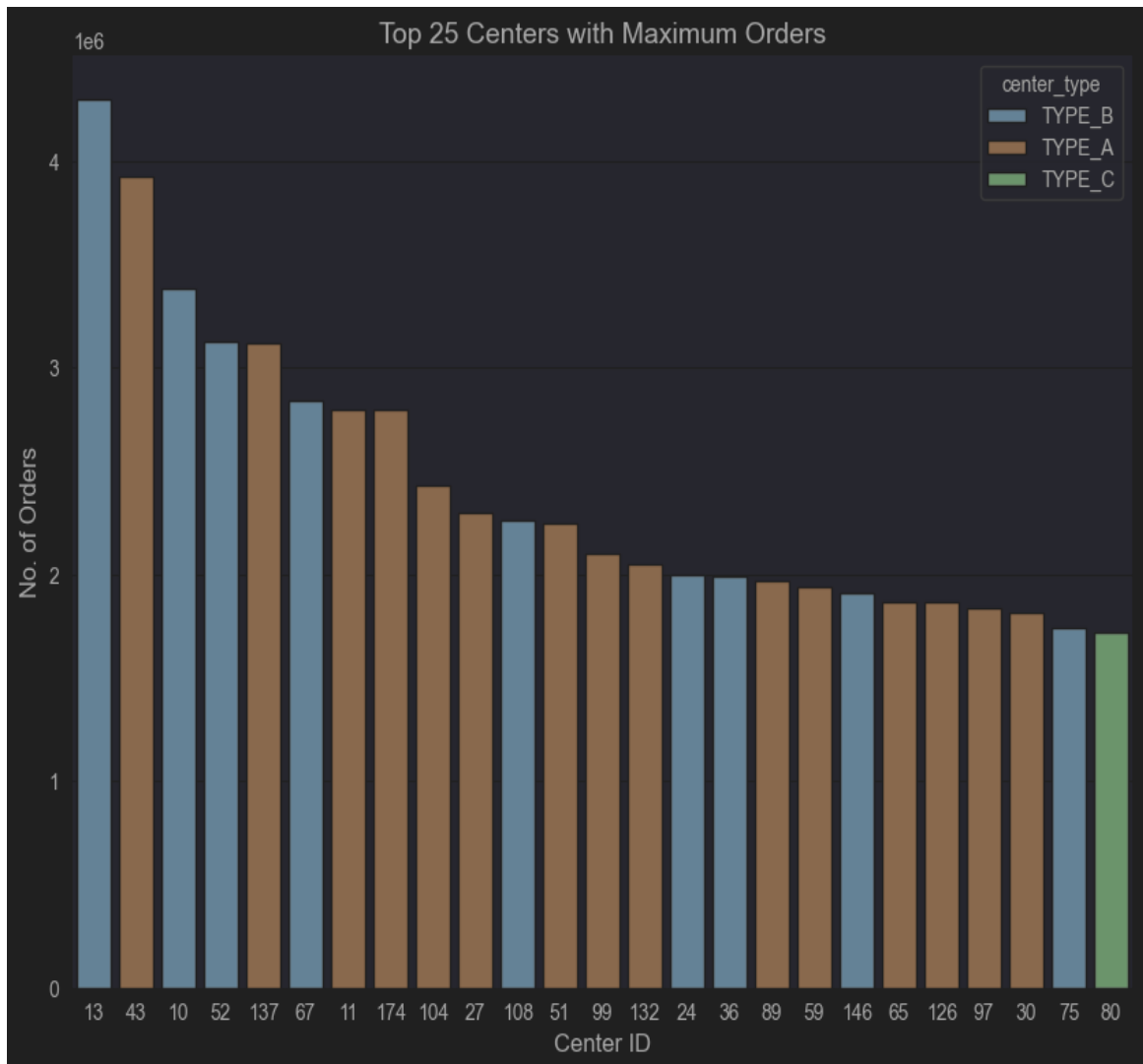


Figure 1: Graphics of delivery centres by the number of orders

I will calculate the number of orders for each type of delivery center see Figure 2.

Table 2

Table fulfilment_center_info.csv

Field	Meaning
Center_id	Unique identifier of the delivery centr
City_code	Unique city identifier
Region_code	Unique region indexer
Center_type	Type of delivery centr
Op_area	The scope of the delivery centr

Table 3

Table meal_info.csv

Field	Meaning
Meal_id	A unique indexer for food
Category	category
Cuisine	Kitchen types

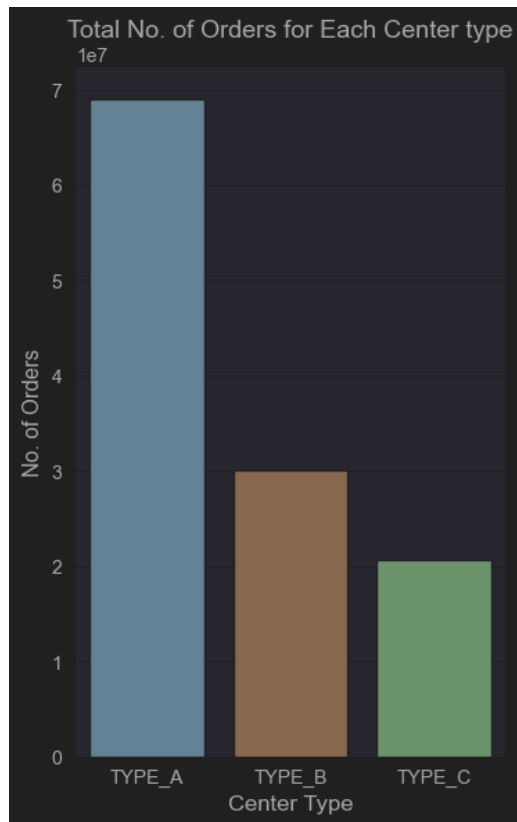


Figure 2: Graphics of delivery centres by number of orders and type

Now I will determine the number of delivery centers for each type see Figure 3.

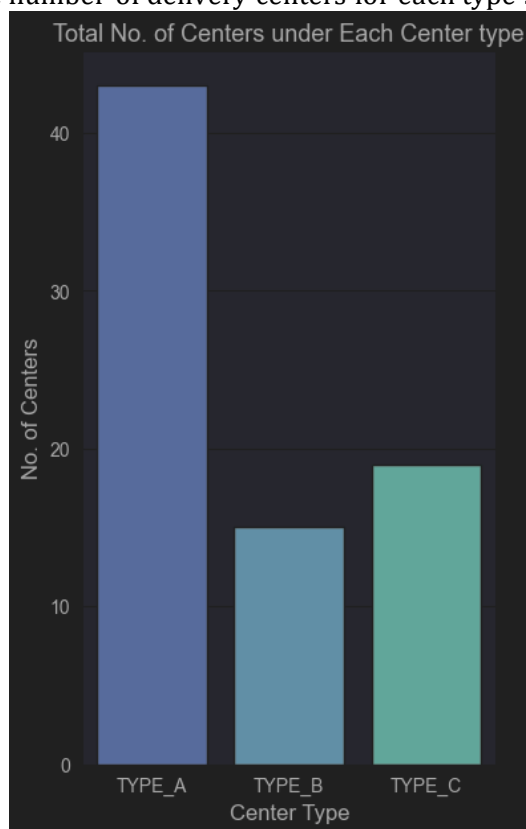


Figure 3: Graphics of delivery centres by number and type of centres

From the graphs above, it can be concluded that centre type A has a large number of centres and orders - it is very popular. Centre type B has a large number of orders relative to centre C, but the number of centres is small.

Now I will evaluate how discounts affect the number of orders see Figure 4.

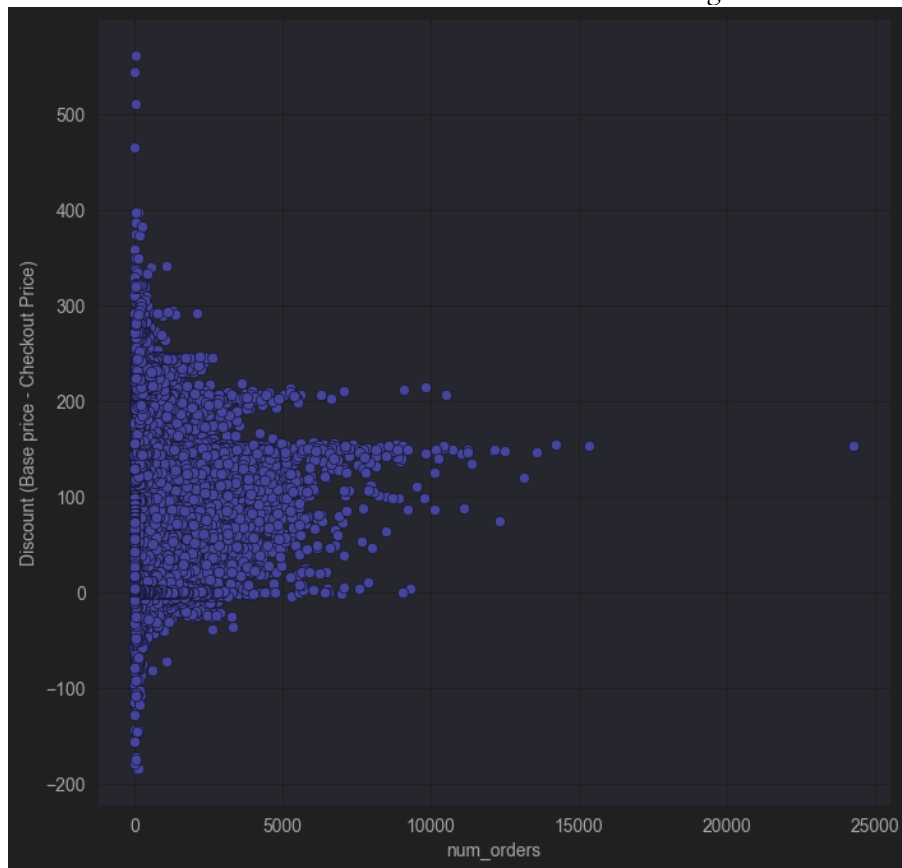


Figure 4: Chart of the ratio of discount to the number of orders

The graph shows that the range from 100 to 200 has the highest number of orders, which may indicate that this is a typical discount for food and people use it. Discounts over 200 are relatively rare, so the number of orders drops.

It is also necessary to estimate the number of orders in relation to the kitchen see Figure 5.

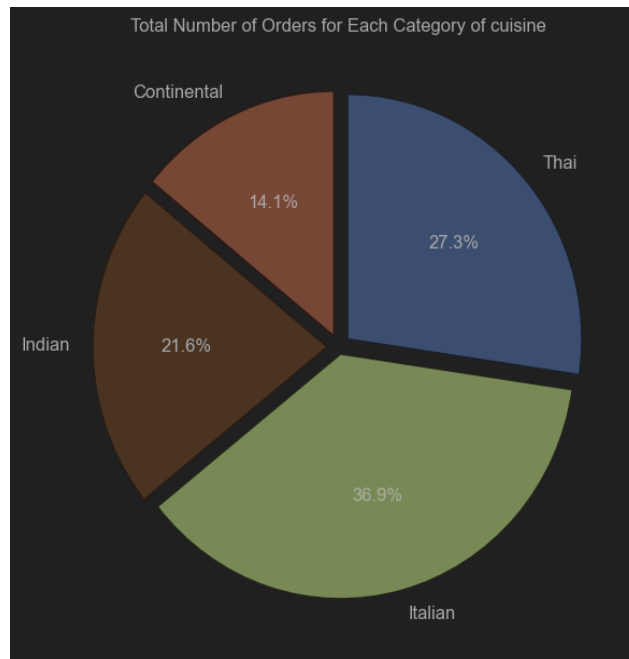


Figure 5: Diagram of the ratio of the number of orders to the kitchen

The diagram shows that Italian cuisine dominates the rest in terms of the number of orders. Thai and Indian cuisines have similar popularity, and people rarely choose content cuisine.

Analysis of the results

I will evaluate each of the methods using two metrics to choose the best one for predicting demand for food delivery.

I will use the Mean Squared Error (MSE) and R-squared. MSE is one of the most common metrics for evaluating the quality of machine learning models. It is a measure of the difference between model predictions and actual values. MSE is calculated as the root mean square of the differences between the true and predicted values. For each predicted value, you subtract the actual value, square the result, and then calculate the average of all the squared differences. The R-squared, also known as the coefficient of determination, is a statistical measure that shows how much of the variance in the response (dependent) variables a machine learning model can explain.

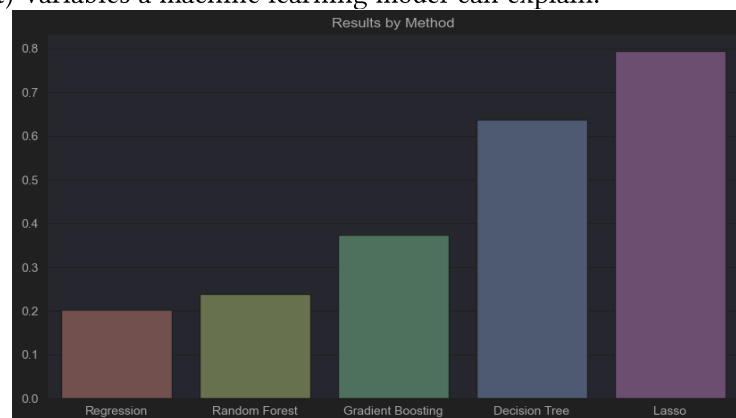


Figure 6: MSE results

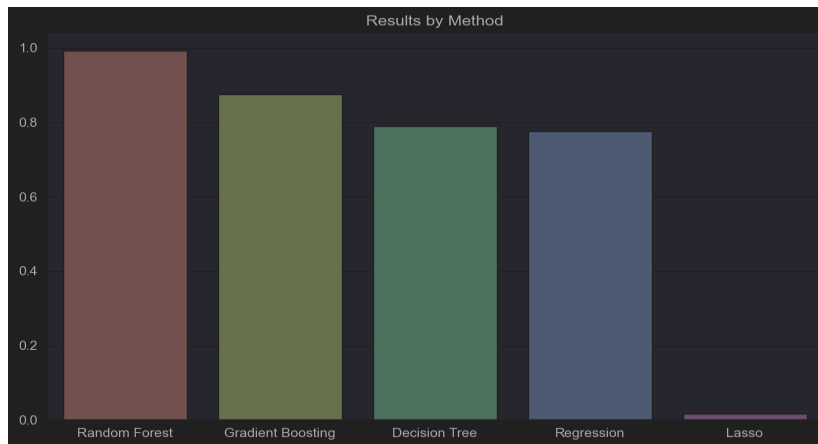


Figure 7: R-squared results

After conducting the study, it is clear that for the mean squared error (MSE) metric, regression showed the best results, and for R-squared, the random forest gave the best result by a large margin.

Linear regression is a fairly simple model that takes into account linear interactions between variables. If there is not a lot of data or it is highly linear, linear regression may be the best model for predicting a variable. This explains why linear regression has the best MSE value in this case. Random Forest, on the other hand, is an improved decision tree algorithm that uses an ensemble (or 'forest') of many decision trees for training and prediction. Random Forest makes its predictions by voting among all the trees in the forest, which helps it avoid overfitting and improves stability. It also works well with non-linear and interdependent data. This may explain why the random forest has a high R-squared value - the model is good at explaining the existing variability in the data.

References

- [1] M.I. Dziamulych, T.O. Shmatkovska, Influence of modern information systems and technologies on the formation of the digital economy, *Economic Forum* (2022): 3-8.
- [2] N.Y. Kirlik, Artificial intelligence and its use in logistics processes, *Actual Problems of the Economy* (2021): 243-244.
- [3] Y. Chalyuk, Determinants of digitalisation of the economy and society, *Intellect XXI* (2020): 138-143.
- [4] Y. Chaliuk, Scenarios of socio-economic development of the EU after BREXIT and COVID, *Scientific Notes of Vernadsky TSU, Series: Economics and Management*, 31(70) (2020): 25-32.
- [5] Y. Chalyuk, Digital competitiveness of countries, *Market Infrastructure* 50, (2020): 23-30.
- [6] T.O. Shmatkovska, M.I. Dzyamulych, Modern information and communication technologies in professional activity in the system of new trends in the digitalisation of the economy, *Economic Sciences* 18 (2021): 248-255.
- [7] T.O. Shmatkovska, O.V. Stashchuk, Big data and business modelling of economic systems, *Effective Economy* 5 (2021): 125-133.
- [8] M. Dziamulych, I. Sadovska, The study of the relationship between rural population spending on peasant households with the main socio-economic indicators: a case study of Volyn region, Ukraine, *Management, Economic Engineering in Agriculture and Rural Development*, volume 20, 2020.
- [9] O. Stashchuk, T. Shmatkovska, Model for efficiency evaluation of financial security management of joint stock companies operating in the agricultural sector: a case study of Ukraine, *Management, Economic Engineering in Agriculture and Rural Development* (2021): 715-728.
- [10] Q. Abdulqader, Applying the Binary Logistic Regression Analysis on The Medical Data, *Science Journal of University of Zakho* (2017): 330-334.
- [11] N. Altman, M. Krzywinski, Simple linear regression, *Nat Methods* (2015): 999-1000.

- [12] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics* 35 (2002): 352-359.
- [13] Y. Amit, D. Geman, Shape quantization and recognition with randomized trees, *Neural Comput* (2019): 1545-1588.
- [14] B. Lanz, *Machine Learning in R: Expert Techniques for Predictive Analysis*, Peter, 2020.
- [15] M. Kuhn, K. Johnson, *Applied predictive modeling*, Springer, NY, 2013.
- [16] L. Breiman, Random forests, *Mach Learn* (2021): 5-32.