A Data-Driven Machine Learning based Forecasting Approach for Posterior Product Demand Explained with Explainable AI

Adria Binte Habib
Dept. of CSE
BRAC University
Dhaka, Bangladesh
binte.adria708@gmail.com

Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@bracu.ac.bd.com

***Abstract*—Demand forecasting is critical for conducting business successfully. Supply chain management systems are heavily under the influence of demand forecasting. With time, the rapidly growing, volatile, uncertainty poses a major complex and ambiguous situation in the market. And to tackle this level of complexity, forecasting and the dependency on its accuracy has become a must to meet the demand and build as efficient as possible of a supply chain. To address this, we conducted some experiments using various regression analysis including Logistic Regression, Random Forest, KNN and XGBoost and utilizing Shapley Additive explanations (SHAP) algorithm we are proposing that "monthly average sales" is the best indicator for demand forecasting. The results suggests that XGBoost performed the best achieving RMSE value of 10.63 and R^2 score of 0.89. In the future works, we will experiment with Multivariate Transformer for demand forecasting and explain the outputs with multiple explainable AI algorithms to extract out the best predicting features.**

***Index Terms*—Time-Series Datasets, XGBoost, Random Forest, K-Nearest Neighbors, SHAP, Explainable AI**

I. Introduction

Supply chain management systems are heavily under the influence of demand forecasting. The reason for this is that it helps the supply chain management system to make vital decisions and planning, like, building capacity, allocation of resources, expanding the size of the business, etc. to run the business with the utmost efficiency. With time, the rapidly growing, volatile, uncertainty poses a major complex and ambiguous situation in the market. And to tackle this level of complexity, forecasting and the dependency on its accuracy has become a must to meet the demand and build as efficient as possible of a supply chain. There are multiple ways to understand the demand using forecasting. It may also differ based on the model used to perform the forecasting. For example, passive demand forecasting is used in stable and low-changing businesses, like local or small businesses or enterprises. This is because, the complexity of such businesses do not require much assumptions and hence can have simpler models. On the other hand, accurate demand forecasting is much more critical for the business. Demand forecasting helps in delivering the demanded items to the customers with utmost satisfaction. It helps in planning by feeding us with reports and data on which the businesses would act upon. The volume, time, period of the products being produced with minimal waste. To make sure nothing is under-delivered, to make sure that the business isn't making the wrong product in the wrong time. As a result of this, business enterprises can act accordingly to produce the best quality products without cutting down it by hurrying to meet the customer demand. Moving on, forecasting doesn't only depend on the algorithm used. It widely depends on the selecting the appropriate statistical and survey methods. The questions need to be well defined, meaningful and concise. Survey methods that involve just asking them about their preference and their feedback for the product can be improved in the future, forecasts are done on shorter horizons while only including the opinions in a poll-basis method. By contrast, statistical methods are used for forecasts done over a longer period of time. These forecasts are much more reliable in terms of surveys because there's no subjectivity involved. Rather, it's fully dependent on the actual behaviour of the consumers through raw data. Nonetheless, survey method is much cheaper to execute in comparison to the statistical methods.

The statistical demand forecasting can be divided into three types. The first is 'Trend Projection' approach. It considers the trend of the product and accordingly predicts the future of the product assuming the trend of the product is constant throughout the time. Hence, the data fed to the model in this approach is in the time-series format. The next is "Econometric Method". This method takes input from both the statistical methods and theories of economics to forecast. A regression model, based on either one or simultaneous equations, is used in the econometric models. Although, the single equation method is mostly used due to it satisfying the needs of the forecasting much more easily. Finally, there's the "Barometric forecasting", however, is not of our concern since it's not relevant for the study we are currently in.

However, in this work, we have implemented three type of algorithms to predict posterior demand of any particular product. The applied algorithms are Logistic Regression, K-Nearest Neighbour, Random Forest & XGBoost. K-Nearest

Neighbour works based on finding closest neighbor approach on the other hand, XGBoost is a tree based algorithm. We used these three type of algorithm to find the best model. However, to explain the result, we used explainable AI.

II. Demand Forecasting Models

In the early days, demand forecasting was performed by statistical methodologies. There were a number of approaches and the most effective one was Auto-Regressive Integrated Moving Average (ARIMA). ARIMA models are defined with three parameters, p denotes auto-regressive nomenclature,q denotes the number of moving average nomenclature n & d denotes the number of times the series has to be differentiated before it becomes stationary. Adding to it, there were other statistical methodologies such as Exponential Smoothing, Independent & Identically Distributed (IID) methods which were used to be widely used for handling various dynamic demand forecasting problems [1] [2] [3] [4] [5]. Consequently, a quantitative study of the immediate resource requirements and their temporal patterns after Hurricane Katrina in 2005 was conducted with ARIMA. Moreover, in another work, ARIMA was used to find the forecasting model for consumer demand and the respective lead time. Moving on to Exponential Smoothing, in 2016, a forecasting model was developed for demand time series which could handle additive and multiplicative seasonality using the concept of exponential smoothing. This method provided better short-term forecasting results than other classical methods. However, in another work, exponential smoothing was compared with potential demand forecasting. In this work, it was illustrated that exponential smoothing does not work as accurately as maximum likelihood techniques [5] [6] [7] [8] [9] [10]. However, due to the arousal of multivariate and multi-period time-series problems, fuzzy network theories were vastly spread. However, a large number of instances happened with the collaboration of Machine Learning, Back Propagation, Genetic Algorithms etc. In 2019, the exponential functions of multiple linear regression for forecasting were modified with the help of the data collected from the Wenchuan earthquake (2008) in China, Chichi earthquake (1999) in Taiwan and Kobe earthquake (1995) in Japan. In another work, Bayesian Decision Framework was used to create prediction model. Consequently, a two stage stochastic program was introduced for Humanitarian Organizations (HO) to procure and distribute humanitarian goods with a limited budget. Besides, a preemptive multi-objective programming model was introduced to address issues in supply chain management with a trade off stockpile and shortage cost [11] [12] [13] [14]. In a work of 2010, multiple methodologies like multi-source data fusion, fuzzy clustering, TOPSIS, hybrid fuzzy clustering–optimization approach etc. were introduced. Moreover, a fuzzy rough set model was introduced in 2012. Additionally, a rough non-deterministic information analysis (RNIA) framework was introduced back in 2016 [2] [15] [16].

III. Implemented Algorithms

A. Logistic Regression

Logistic regression forecasts binary consequences based on previous surveillance with the help of statistical analysis. In other words, the dependent variable is predicted by this model through exploring the relationship between divers one or multiple existing independent variables [20].

B. K-Nearest Neighbour (KNN)

K-Nearest Neighbour (KNN) is a supervised machine learning algorithm and this algorithm can be used for classification and regression. The goal of this algorithm is to learn Y = f(X) where the input is X. KNN summarizes the output variable for those cases while looking through all the data for the K most comparable neighbors. The mean output variable is employed in regression, and it serves as the most typical class value in classification [19].

C. Random Forest

Random Forest consists of a number of decision trees. While constructing a single tree, this algorithm utilizes bagging and feature randomness. The reason behind this is to generate an uncorrelated forest of trees. This type of forest helps to predict more perfect results than any individual tree [18].

D. XGBoost

In the paradigm of approximate tree learning, XGBoost is used as a sparsity-awarded algorithm for the data that are sparse and quantile sketches that are weighted. Furthermore, this algorithm used parameters from cache access patterns, compression of data and a boosting system to build a tree that is scalable. The XGBoost gets an advantage by combining all the above which helps it be more efficient and more accurate using less computational power. [17].

E. Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a new area of study in artificial intelligence. XAI can describe how AI arrived to a given answer (for example, classification or regression). This level of explanation is not attainable in older systems. Explainability is vital in applications such as e-commerce, health care, banking and finance, law and order, and so on. So far, a variety of XAI approaches have been developed for such applications. In our study we have attempted to find the explanation of our models outcome by one of the XAI algorithms: Shaply additive explanations (SHAP).

1) SHAP: SHapley Additive exPlanations: For understanding a particular model's mechanism a lot of methods have been invented. However, it is still imprecise how these techniques relate to one another and when one technique is better than the other. This issue was addressed by the SHAP framework, which unifies prediction interpretation (SHapley Additive exPlanations). Each feature is given a relevance value by SHAP for a specific prediction. The discovery of a new class of additive feature significance measures and theoretical findings

demonstrating the existence of a singular solution in this class with a set of desirable qualities are its novel components. Because some more current methods in the class don't have the suggested desirable features, the new class unifies six existing methods [21].

IV. Data Analaysis

A. Dataset Description

In this study, a realworld sales dataset is undertaken from an e-commerce business company. In this time series dataset, transactions are stored from 2013–01–01 to 2017–12–31 timestamp. There are total 913,000 Sales Transactions recorded in the whole dataset with 50 unique SKU in 10 Stores. There are total four attributes in the dataset which are named as 'date', 'store', 'item', 'sales'.

B. Data visualization

A sales per year graph is created using our dataset to weed out the fundamental trend (Fig 1). The graph clearly shows that a dramatic increase in sales occurs in the middle of each year.

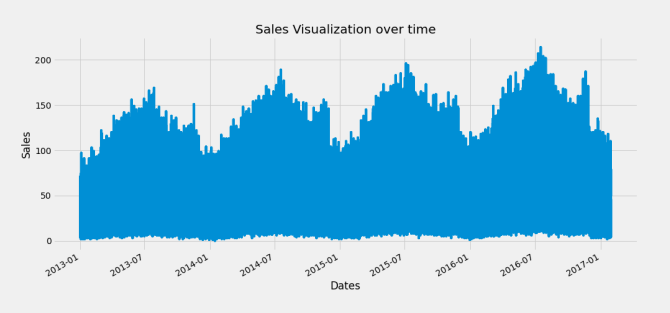

Fig. 1. Sales (yearly) data visualization.

To make the trend more evident, we generated a boxplot of sales by month throughout the whole dataset (Fig 2). According to the plot, there is a steady increase in sales, with a peak in sales occurring around the seventh month of each year.

C. Feature Extraction

To extract features from a time-series dataset, at first a coversion is needed to make it a series data in datetime dataframe. Later, some extraneous features can be derived from these existing features to find a set of robust features for the dataset. In this study, after extracting date, time, month and year attributes, day^{year} is calculated. Daily average sales and monthly average sales are also derived from the existing attributes. Later, from the correlation matrix of all the features (Fig 3), the most correlated features with sales data are daily average sales and monthly average sales (Fig 3). Based on this fact, these two features are selected as our main features to train our model.

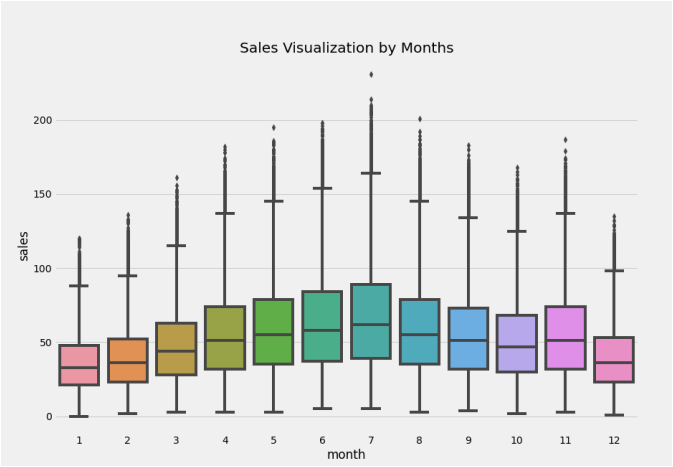

Fig. 2. Sales (monthly) data visualization.

Fig. 3. Correlation matrix of the features

V. Methodology

In this study, our first step was to load the 'Store Item Demand Forecasting' dataset. As it was a time-series dataset, we had splitted the date attribute in to separate date, month and year attributes using pandas datetime module. After getting our desired attributes we calculate daily average and monthly average attribute with the help of mean and merge these values with the main dataframe. A correlation matrix among features helped us to determine the most effective features for this set. Basically, correlation matrix is just another form of principal component analysis (PCA) (Fig 4).

From the heatmap of correlation matrix (Fig 3), it is safe to say that the daily average and monthly average are highly correlated with our output attribute i.e. sales data. After preprocessing, cleaning and extracting features for all the data, our dataset is splitted into train sets (2013-01-01 to 2016-12-31) and test sets (2017-01-01 to 2017-12-31). A XGBoost (Extreme Gradient Boosting) model was hypertuned with n-estimators=1000 and learning-rate=0.01 for the training phase and evaluated the performance metric with K-Nearest Neighbor, Random Forest Regressor and Logistic Regressor model. For understanding each model's mechanism, an game theoretic approach (SHAP) is used in each of the ML model and evaluated the performance parameters.

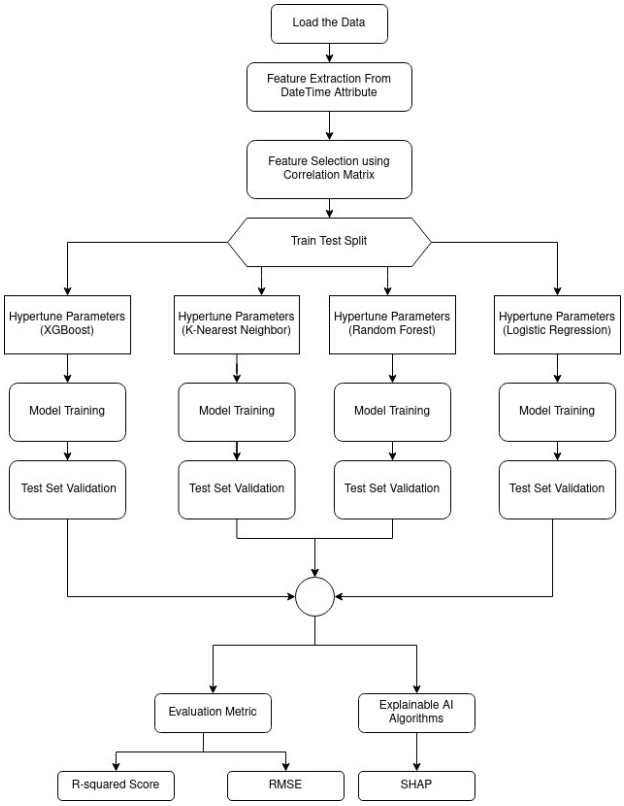

Fig. 4. Proposed Pipeline

VI. Result Analysis

A. Performance Measurement Metrics

There are several metrics that may be used to assess the performance of regression algorithms. We have picked the measures for evaluating ML performance with caution as they will determine how the performance of ML algorithms is assessed and compared. In addition, the metrics will totally affect how we weight the relevance of various variables in the outcome. In this work, we have used two assessment indicators to validate our models: R2 coefficient and root mean square error (RMSE)

1) R^2 Score: R-squared (R^2) is a statistical measure that shows how much of a dependent variable's variance is explained by one or more independent variables in a regression model. It measures how well the variation of one variable accounts for the variance of the second, as opposed to correlation, which describes the strength of the relationship between independent and dependent variables.

2) Root Mean Square Error (RMSE): Root Mean Square Error (RMSE) is the residuals' standard deviation (prediction errors). The distance between the data points and the regression line is measured by residuals, and the spread of these residuals is measured by RMSE. It clarifies how concentrated the data is around the line of greatest fit, to put it another way. In addition, root mean square error is frequently utilized in regression analysis, forecasting, and climatology to validate experimental results.

B. Model Evaluation

For the dataset, four models were used: Logistic Regression, Random Forest, KNN, and XGBoost. Table-I shows that the XGBoost model outperformed the others. Among the other models we tested, it had the lowest rmse value (10.63). The model also has the greatest R^2 value (0.89).

TABLE I
Performance Analysis of the Algorithms

Model	R^2	RMSE
Logistic Regression	0.82	13.42
Random Forest	0.85	12.3
K-Nearest Neighbor (KNN)	0.87	11.32
XGBoost	**0.89**	**10.63**

TABLE II
Hyperparameters

Algorithms	Hyperparameters
Logistic Regression	default settings
Random Forest	n_estimators = 45, max_features = 'sqrt', max_depth = 5
K-Nearest Neighbor	n_neighbors=20
XGBoost	n_estimators=1000, early_stopping_rounds=50, learning_rate=0.01

1) Logistic Regression: First, we used logistic regression to build a baseline model. For the hyperparameters, we utilized the default values. Logistic regression produced an R^2 score of 0.82 and an rmse value of 13.42.

2) Random Forest Regressor: Following that, we used a decision tree-based random forest regressor model to improve the metrics. We ran the random forest regressor model on the dataset with 5 to 50 trees and calculated the rmse value. Based on this, we chose 45 trees to be utilized in the final model along with additional hyperparameters (Table-II). The trained random forest model increased the rmse by 1.12 and the R^2 score by 3% from logistic regression.

3) K Nearest Neighbors Regressor: Furthermore, we used the data-set to test the K-nearest neighbors (KNN) regressor model in order to improve the assessment metrics. We trained the KNN regressor model for different neighbors ranging from 1 to 50, just as the random forest regressor model, and then created an elbow plot. From the elbow plot, We decided to train the final model with 20 neighbors using the default hyperparameters. We were able to enhance the R^2 score by 2% while decreasing the rmse value to 11.32.

4) XGBoost Regressor: Finally, we wanted to go further into the tree-based algorithms and see if we could enhance the metrics even more. As a result, we trained the XGBoost Regressor algorithm using 1000 trees, 50 early stopping rounds, and a 0.01 learning rate. This model produced the

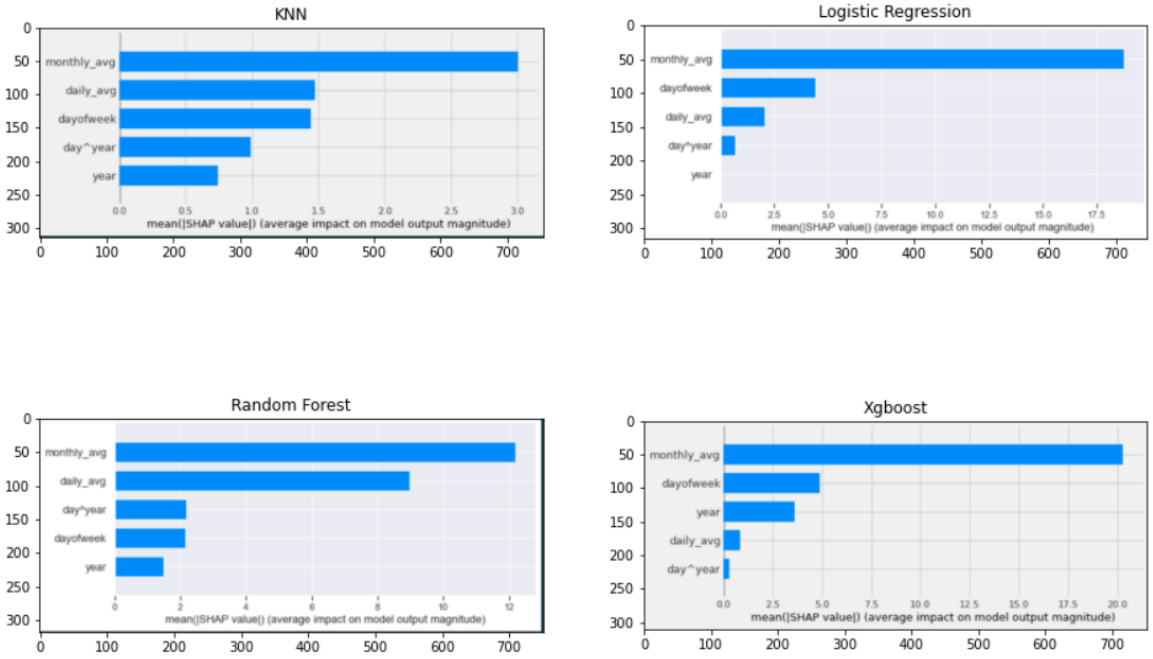

Fig. 5. Performance Analysis of the models with explainable AI (XAI)

greatest results thus far, increasing R^2 value by 7% over the baseline model and decreasing rmse value by 2.79. (Table-I).

C. Explanation of the models outputs

For each model, we determined the validation losses. The assessment measures, however, do not disclose how the model arrived at a given answer. We needed to know which features were important for the model to use to determine the regressor output. On this occasion, we used one of the XAI models: Shaply additive explanations (SHAP).

1) SHAP: We chose five highly correlated date time features with regard to sales amount earlier: monthly average sales, day of the week, daily average sales, year, and day^{year}. Following training, we fed each model, together with the test-set and training-set, through the SHAP algorithm. The SHAP model then created a summary plot for each model, displaying the model's relevance of characteristics based on the SHAP values. According to the results (Figure 5), all of the models used the monthly average sales (monthly_avg) as the most relevant variable to forecast sales. Furthermore, with the exception of XGBoost, every model ranked the "year" characteristic of each sale as the least important. In addition, KNN and Random Forest Regressor have selected daily average sales (daily_avg) as their second most essential attribute. However, Logistic Regression and the XGBoost model rank the day of the week (dayofweek) as the second most relevant variable in predicting sales volume for a given date. This contrast also suggests that forecasting product sales does not necessarily depend on the year the items are sold.

VII. Conclusion

Demand forecasting is very much essential for successfully undertaking business. To address this, we did some experiments with the help of several regression analysis. After that, we explained the results of those regression analysis tasks using two approaches, one of them is traditional approach which includes R squared and Root Mean Square Error metrics and another of them is Explainable AI. In all aspect, we found XGBoost working in the best way. However, the target of this experiment was not to find the best model rather to find why and for which features that particular model performed in the best way. We used R squared and RMSE as evaluation metrics to evaluate the results. The explainable AI suggests us that forecasting product sales does not necessarily depend on the year the items are sold rather depends on monthly average sales of any product mostly. Our next goal is to use transformer models for demand forecasting and use more XAI models and compare between the results of XAI models.

References

[1] Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis: forecasting and control, 3rd edn. Prentice Hall, Englewood Cliffs
[2] Sheu J-B (2010) Dynamic relief-demand management for emergency logistics operations under largescale disasters. Transp Res Part E 46:1–17
[3] Gujarati D (2003) Basic econometrics. Mc-Graw Hill, Boston
[4] Wei WWS (1990) Time series analysis: univariate and multivariate methods. Addison-Wesley Publishing Company, New York
[5] Aviv Y (2003) A time-series framework for supply chain inventory management. Oper Res 51(2):210–227
[6] Holguin-Veras J, Jaller M (2012) Immediate resource requirements after hurricane Katrina. Nat Hazards Rev 13(2):117–131
[7] Gilbert K (2005) An ARIMA supply chain model. Manag Sci 51(2):305–310

[8] Wu SL (2012) A research of dynamic demand forecasting model for large earthquake emergency supplies. Harbin Institute of Technology, Harbin
[9] Tratar LF, Mojškerc B, Toman A (2016) Demand forecasting with four-parameter exponential smoothing. Int J Prod Econ 181(Part A):162–173
[10] Zehna PW (1972) Some alternatives to exponential smoothing in demand forecasting. Technical Report Collection
[11] Wu X, Gu J, Wu H (2009) A modified exponential model for reported casualties during earthquakes. Acta Seismol Sin 31(4):457–463
[12] Taskin S, Lodree EJ (2011) A Bayesian decision model with hurricane forecast updates for emergency supplies inventory management. J Oper Res Soc 62:1098–1108. https://doi.org/10.1057/ jors.2010.14
[13] Park J, Kazaz B, Webster S (2018) Surface versus air shipment of humanitarian goods under demand uncertainty. Prod Oper Manag 27(5):928–948
[14] Wyk EV, Yadavalli VSS, Carstens H (2013) Decision support in supply chain management for disaster relief in Somalia. Springer, Berlin, pp 13–22
[15] Bingzhen S, Weimin Ma, Haiyan Z (2012) A fuzzy rough set approach to emergency material demand prediction over two universes. Appl Math Modell 31:7062–7070
[16] Zhu X, Sun B, Jin Z (2016) A new approach on seismic mortality estimations based on average population density. Earthq Sci 29(6):337–344. https://doi.org/10.1007/s11589-016-0170-3
[17] Chen, T., Guestrin, C. (n.d.). XGBoost: A scalable tree boosting system . https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf. Retrieved August 31, 2022, from https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
[18] Breiman, L. (n.d.). random forests - University of California, Berkeley. https://www.stat.berkeley.edu/ breiman/randomforest2001.pdf. Retrieved September 12, 2022, from https://www.stat.berkeley.edu/ breiman/randomforest2001.pdf
[19] Maudoux, C., Boumerdassi, S. (n.d.). Smart and Sustainable Agriculture. Springer. Retrieved September 12, 2022, from https://link.springer.com/chapter/10.1007/978-3-030-88259-48
[20] An introduction to logistic regression: From basic concepts ... - koreamed. (n.d.). Retrieved September 12, 2022, from https://synapse.koreamed.org/upload/SynapseData/PDFData/0006jkan/jkan-43-154.pdf
[21] Lundberg, S., Lee, S.-I. (2017, November 25). A unified approach to interpreting model predictions. arXiv.org. Retrieved September 12, 2022, from https://arxiv.org/abs/1705.07874