Predicting Price of Daily Commodities using Machine Learning

Md Nur Amin
Faculty of Sciences and Technologies
University Jean Monnet Saint Etienne
Saint Etienne, France
nuramin.aiub@gmail.com

***Abstract*—Daily commodities are necessities that take up a large part of the product market. The fluctuation of the price in daily commodities is an apparent phenomenon that has a high bearing on the cost of living. Early prediction of price can help to control prices by monitoring and adjusting the price in the market beforehand so that the commodity market runs in a stable way. Suppliers and manufacturers can choose to produce or supply commodities accordingly. It will help to balance inventory and profitability as well as improve availability to consumers. Thus, it is directly related to the interest of consumers and producers. The selection of suitable and favorable algorithms is one of the most practiced studies in the field of data forecasting. It takes the advantage of the empirical evidence available at hand to choose the most appropriate model since no model can be contemplated as the best. In this study, We performed an extensive and comprehensible experimental evaluation to predict commodity price using state-of-the-art Machine Learning algorithms.**

***Index Terms*—Daily Commodities, Price Prediction, Regression, Machine Learning, Ensemble Methods**

I. Introduction

The price of daily commodities often becomes unstable due to various reasons. Many aspects contribute to this problem depending on production and supply management chain, type of food and farming location. Change in the weather has a direct impact on the fluctuations in the price of daily commodities. Yielding of crops is directly affected by the exposure to climatic change. The weather has a high impact on the production of the crops due to the diversity of seasons and extreme climatic conditions can lead to a decrease in crop yield, surging the price and creating anarchy in the commodity market. As demand shifts upwards and is not satisfied by the local market, the price of daily commodities go up. Being able to forecast the price of commodities due to adverse climatic situations, preventive measures like changing the crop varieties, determining planting dates for better adapting to changing weather patterns may increase the yield to control the price. Predicting the price of daily commodities can help the appropriate governing body and policy makers to monitor beforehand the price surge takes place. It can play a significant role to control the price hike and fluctuations in the commodity market.

The advancement in computing domain has strengthen the conversion of data into knowledge and useful information to aid decision making. It has bolstered the evolution of computer system that enables efficient storing, analysis and organization of ever increasing data.Forecasting contributes to ease the decision-making process in financial, commercial, social and economic domain.

There are few works in the literature of predicting models that illustrate the seemingly random movement of price of the commodity market addressing variety of daily commodities. Achieving high accuracy requires the appropriate selection of variables with the right transformation and tuning of the parameter. This study aims to predict the price of 6 different daily commodities such as wheat, avocado, and dairy foods(40 pound block cheddar cheese, 500 pound Barrel Cheddar cheese, Butter, Dry Whey) using sate-of-the-art Machine Learning algorithms such as Support Vector Machine, Random Forest, Bagging, AdaBoost, GradientBoost, XGBoost and LightGBM.

II. Related Works

In this section, we dive deep into the Machine Learning methods that were applied by the practitioner to perform regression task to solve real word problem.

Statistical methods such as Moving Average (MA) and Autoregression (AR) dominated the data analysis field for over half a century. Despite some studies took place in the early eighties, they were not perfectly carried out to real world application. Over the last decades due to dawn of data mining techniques, Machine Learning has seen a growing popularity and interest to aid time dependent analysis. Non-parametric models have risen for thoroughness to the classical models. Therefore, researchers have engaged themselves to improve these techniques as well as developing new solution.

Deep learning has been widely used by experts to make an accurate prediction in various domains including agriculture and economics. With the increasingly huge amount of data being recorded every day about consumer and purchase behavior, forecasting has become crucial where it can support the government to determine suitable policy. Inflation is the tendency of price to rise of products and goods. High and unstable inflation can have a negative impact on the economy. Consumer Price Index (CPI) value is a broadly used indicator of inflation. S Zahara et al [1] proposed a Consumer Price Index (CPI) prediction model with non-linear parameter input of daily staple food prices using Long Short Term Memory

978-1-7281-9673-2/20/$31.00 ©2020 IEEE

(LSTM) to predict the next inflation move. Adaptive Gradient (AdaGrad), Adaptive moment (Adam), Stochastic Gradient Descent (sgd), Root Mean Square Propagation (RMSProp), and Nesterov Adam optimization algorithms were used to improve the accuracy where Nesterov Adam achieved the best result with RMSE of 4.08 to predict the CPI value.

Firms that deal with supply chain, despite exchanging information and maintaining integration, does not reduce the forecast error completely. Real et al [2] applied nonlinear Machine Learning algorithm in an extended supply chain to forecast the distorted demand signal at the upstream end of the supply chain. RNN, SVM, NN, and MLR found to be significantly better than naïve, moving average and trend methods to predict the distorted signal.

Zuriani et al[3] proposed an Enhanced Artificial Bee Colony(eABC) to predict the price of daily commodities of a time series data in three different arrangements. A Swarm Intelligence approach, Artificial Bee Colony (ABC) has been used to optimize the parameter of the Least Square Support Vector Machine considering the critical issue of overfitting and later compared with Back Propagation Neural Network and Genetic Algorithm. Empirical results demonstrate a higher PA due to the ability of the model to learn the pattern effectively in the time series data.

ANN are non-parametric statistical estimators and are good for predictive modeling having the ability to capture nonlinear characteristics and time series patterns. Backpropagation Neural Network (BPNN) is the training algorithm to update the wights in RNN to minimize the error of network outputs. NARX is a recurrent dynamic network that has a feedback architectures with feedback connections that stitches several layers of the network which come only from the output neuron instead of hidden neurons. Azme Khamis et. al [4] used BPNN and nonlinear autoregressive models with exogenous inputs (NARX) networks to predict the price of wheat from historical wheat data. In addition, the NARX model with 8 nodes in the hidden layer and 4 tapped delay lines is adopted for the prediction as an alternative model. NARX outperformed the BPNN with an MSE and R of 0.0047 and 0.9728.

Avocado has been seeing growing demand and is one of the common fruits that people love to have in their breakfast. This growing popularity has caused the price of avocado to be inconsistent. Juan et al [5] proposed an approach to estimate the sales of avocado using historical sales records and weather data in the United States. Linear Regression, Multilayer Perceptron, Support Vector Machine for Regression and Multivariate Regression Prediction Model were applied where the last two achieved the best result having a correlation coefficient of 0.995 and 0.996, and an RMSE of 7.971 and 7.812 respectively.

Sidra et al[6] demonstrate a trade-oriented forecasting framework consisting of machine learning and deep learning models to predict the price of the stock and short term movement in the stock. They built eight regression models to predict the stock price and classification models for predicting the movement of the price using Logistic Regression, Decision Tree, K-Nearest Neighbor, Artificial Neural Networks, Random Forest, Bagging, Boosting, Support Vector Machines, Long and Short-Term Memory Networks. The best result for regression model was achieved by LSTM with an RMSE and correlation of 2.36 and 0.99 respectively.

Yuehjen et al [7] proposed an integrated model using ARIMA-ANN, ARIMA-SVR, and ARIMA-MARS to predict the price of three major crops such as rice, wheat, and corn. The main contribution of the proposed model is that the model can predict without requiring extensive effort to obtain the future values of explanatory variables. Integrated ARIMA-SVR outperformed in a well over margin upon single ARIMA and SVR model in terms of MSE, RMSE and MAPE . However, the limitation of the model is it may be intuitive and computationally expensive to identify the correct ARIMA model from the variety of possible models.

Forecasting price trends of products have become a priority in the consumer industry. It can also help the buyers to find suitable prices for their desired product. Huy Voung et al [8] developed a system for short term prediction of seasonal products using Auto ARIMA and a comparison with Moving Average(MA). Auto ARIMA was preferred instead of ARIMA as it is hard to set the parameter and correct order of AR and MA. Using MA they found that forecast trends tend to be flat, while Auto ARIMA is not suitable for long term prediction. Nevertheless, the model advances through a hierarchical process that is computationally costly.

With the increase in the world population, the demand for food is growing. Policymakers should estimate the demand of food so that the food market does not become chaotic. Marziye et al [9] proposed a predictive model based on the energy inputs applied from different sources and output energy during the production season of rice. Polynomial and Radial Basis Function (RBF) kernel of Support Vector Machine (SVR) was implemented to minimize the generalization error bound instead of minimizing the training error. Selecting the right parameter of the kernels affects the accuracy of the model. The model achieved the best results with (10,0.02,0.06) for parameter selection of RBF kernel and with (10,1,0.3) for the polynomial kernel. SVR with RBF kernel outperforms polynomial kernel with an RMSE of 8 and a coefficient determination of 94.

From this section we found that there are very few works on forecasting of daily commodities addressing variety of daily commodities with comparative analysis of Machine Learning Models.

III. DESCRIPTION OF DATASET

In this section, we will discuss about the data sets that are used in this study. To conduct the experiment, seven different data sets have been collected from various sources. Description of each data set are as follows-

US daily wheat stock price: This data set [16] is collected from Kaggle containing the daily wheat price in the US having a sample size of 2273 and 4 features. The features include, ‘open’: opening price of wheat stock on a particular day, 'high': highest price of wheat stock on a particular day, 'low': lowest price of wheat stock on a particular day, 'close': closing price of wheat stock on a particular day.

Weekly butter price: This data set [17] is collected from Kaggle containing weekly dairy price of Butter having a sample size of 1408 and 5 features. Features are 'week ending date': Last day of the week, 'report date': Last day of the reporting period, 'date': Day of record, 'weighted prices': average price, 'Sales': total sales in a week.

Weekly 40 pound block cheddar cheese price: This data set [17] is collected from Kaggle containing weekly dairy price of 40 pound Block Cheddar Cheese having a sample size of 1408 and 5 features. Features are 'week ending Date': Last day of the week, 'report date': Last day of the reporting period, 'date': day of record, 'weighted prices': average price, 'Sales': total sales in a week.

Weekly dry whey price: This data set [17] is collected from Kaggle containing weekly dairy price of Dry Whey having a sample size of 1408 and 5 features. Features are 'week ending Date': last day of the week, 'report date': last day of the reporting period, 'date': day of record, 'weighted prices': average price, 'Sales': total sales in a week.

Weekly 500 pound barrel cheddar cheese price: This data set [17] is collected from Kaggle containing weekly dairy price of 500 pound of Barrel Cheddar Cheese having a sample size of 1408 and 5 features. Features are 'week ending Date': Last day of the week, 'report date': Last day of the reporting period, 'date': day of record, 'weighted prices': average price, 'Sales': total sales in a week, 'moisture content': moisture content in the cheese, 'weighted price adjusted to 38% moisture': adjusted price due to moisture in the cheese.

US daily avocado price: This data set [18] is collected from Kaggle containing the daily price of Avocado in the US having 18248 instances and 12 features. Features include 'week ending': last day of the week, 'average price': average price of a single Avocado, 'type': conventional or organic Avocado. 'year': year of record, 'region': city or region of the observation, 'total volume': total number of avocado sold,'4046': total number of Avocados of PLU 4046 type, '4225': total number of Avocados of PLU 4225 type, '4770': total number of Avocados of PLU 4046 type, '4770': total number of Avocados of PLU 4770 type, 'small bags': sold in small bags, 'large bags': sold in large bags, 'xlarge bags': sold in xlarge bags.

Wheat price of Bangladesh: Two data sets of weather and wheat price have been merged to create a single data set. The weather data was collected from the Bangladesh Agricultural Research Council (BARC) and the price of wheat was collected from Humanitarian Data Exchange. Together it contains the price of wheat in Bangladesh in different regions along with weather data. It has 550 instances and 25 features. Two different data sets of wheat price and weather data were aggregated. Some of the main features are 'price': price of wheat, 'max_temp': max temperature in a region, min_temp': minimum temperature in a region, 'max_temp_avg': average of max temperature in a region, min_temp_avg': average of min temperature in a region, 'rainfall_avg': average rainfall in a region,'humidity_avg': average humidity in a region, 'sunshine_avg': average sunshine in a region, 'windspeed_avg': average wind speed in a region, 'cloud_avg': average cloud over a region.

IV. METHODOLOGY

In this section, the methodology for the experiment is described. The process is carried out into several consecutive phases. It includes missing value imputation, feature scaling, building test and train model and evaluation of the results. The methodology can be illustrated in the following diagram.

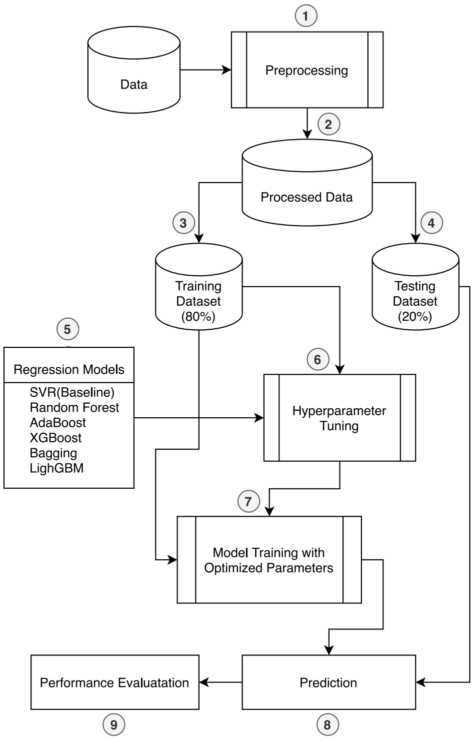

Fig. 1. Detailed Workflow Diagram

A. Data Preprocessing

Preprocessing is an important task before feeding the data into the model. It includes missing value imputation and feature scaling. Missing value imputation is an important task in Machine Learning particularly when the data is small to ensure utilizing all possible samples. Although there are many missing value imputation techniques, mean and mode were used in this study. Mean is used to handle the missing numerical feature where mode, on the other hand, is used to handle the categorical feature. Feature scaling is a method to standardize the data when the values in the input features varies highly. *StandardScaler* from *Scikit-learn* preprocessing library has been used for scaling the input features. Feature engineering involves adjusting and reworking the raw predictors to represent a model to better uncover predictor-outcome relationships. The objective of the feature engineering is to provide a strong and simple relationship between new input features and output. Supervised learning algorithms should have enough input features as low dimensional input features are a bottleneck for the learning algorithm.

B. Data Segregation

The data sets were partitioned into train and test sets. The train set includes the 80% of the actual data set and the test includes the 20% of the actual data set.

C. Regression

In this study, we have used 7 different Machine Learning algorithms. Support Vector Regressor was used as a Baseline model to set a baseline performance.Random Forest and five different ensemble methods such as Adaptive Boosting (AdaBoost), GredientBoost,eXtreme Gradient Boosting (XGBoost), Bagging, LighGBM (Light Gradient Boosted Machine) were used to improve upon the baseline model. Ensemble methods are chosen for this experiment because they are proven to perform well from small to medium seized data set. The data sets are used in this experiment has low sample size except one data set.*Scikit-learn* library has been used to deploy the models.

1) Support Vector Regressor (SVR): Support Vector Regressor [10] is a part of Support Vector Machine that is mainly used for classification, but the algorithm applied in SVM is also useful as a regressor. It is a non-parametric technique that supports that both linear and nonlinear regression. SVM projects the original training data to higher dimensional plane and looks for the optimal hyperplane.

2) Random Forest (RF): Ensemble methods are a combination of several models in order to produce optimal predictive model. Random Forest [11] is an ensemble method that operates by constructing multiple Decision Trees during the train phase. Using slightly different data to build each tree adds diversity to the models. Averaging the results of multiple trees together reduces the risk of overfitting. Multiple trees also give fine- grain predictions than a single tree.

3) Bagging (BG): Bagging [15] is a homogeneous ensemble method which works on by applying the same algorithm on all the estimators and the algorithm must be a weak model. The Bagging algorithm trains individual models using a random subsample for each which is known as bootstrapping . Bagging uses the same weak model for all the algorithms but the dataset for each is a different subsample which provides diversity. After the individual models are trained with their respective samples, they are aggregated using voting or averaging. The advantage of the bagging is it helps to reduce variance as the sampling is truly random.

4) AdaBoost (AB): AdaBoost [15] is an ensemble method having two distinctive properties of AdaBoost compares to other boosting algorithms. First, the instances are drawn using a sample distribution of the training data into each subsequent data set. This sample distribution makes sure that instances which were harder to predict for the previous estimator, have a higher chance to be included in the training set for the next estimator by giving them higher weights. Secondly, estimators are merged with weighted majority voting. Estimators that showed better performance are awarded with higher weights for voting.

5) Gradient Boosting (GB): Gradient Boosting [13] is an ensemble method that builds up a model by incrementally improving the existing one. It initiates a model by fitting a single , usually , shallow tree to the data afterwards fits a tree to the residuals of the model and finds the weighted sum of that tree with the first one that gives the best fit. Because Gradient Boosting optimizes error on the training data, it is very easy to overfit the model. So it is better to estimate out of sample error via cross- validation for each incremental model then retroactively deciding how many trees to use.

6) LightGBM (LGB): LightGBM [12] is an efficient Gradient Boosting framework that uses tree based learning. It uses a leaf- wise tree growth algorithm. LightGBM is used for its faster training time when the train data is big. It has become popular for its higher training speed and it also supports parallel computing with GPU learning.

7) XGBoost (XGB): XGBoost [12] is an advanced version of Gradient Boosting that uses a gradient boosting framework focusing on the computational speed and model efficiency. Its supports parallelization and has a custom objective function to create decision trees parallelly that requires gradient and hessian. In XGBoost, complex models are penalized by using both Lasso and Ridge regularization to prevent overfitting. It comes with a built in cross- validation methods that eliminates the need to specify the exact number of boosting iterations needed in a single run.

D. Hyperparamter Tuning

Hyperparamter is a crucial part to obtain the best performance of the model. Parameters of all the seven models for each data sets were tuned using *GridSeacrhCV* from *Scikit-learn* library. Grid search builds models for each combination of parameters that ultimately results in a better performance.

E. Metrics for Performance Evaluation

Three evaluation metrics have been used in this experiment to evaluate the performance of the model such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and R^2. MSE measures the average of the squares of the errors. The average magnitude of the errors are measured using MAE where R^2 on the other hand, demonstrates how close is the regression line from the original data point.

V. EXPERIMENTAL RESULTS

In this section, we will analyze the findings in depth for all the datasets that have been used.

1) US daily wheat price: From table 1 we can see that, overall performance of all the models have increased after hyperparameter tuning. Random Forest and ensemble methods could not outperform the the base model(SVR) and all ensembles methods had higher MSE even after hyperparamter tuning. Support Vector Regressor had the best MSE score. AdaBoost, GradientBoost and XGBoost had the same MAE score where Support Vector Regressor and LightGBM have the highest score in term of R^2.

TABLE I
RESULTS OF THE US DAILY WHEAT PRICE PREDICTION

	Before Hyperparameter Tuning			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	20.725	3.246	0.999	**3.246**	0.999	**0.999**
RF	37.717	4.193	0.998	38.713	4.557	0.998
AB	42.464	4.273	0.998	37.103	**0.998**	0.998
GB	40.355	4.573	0.998	32.518	**0.998**	0.998
XGB	42.353	4.673	0.998	32.995	**0.998**	0.998
BG	41.716	4.437	0.998	38.862	4.549	0.998
LGB	39.069	4.302	0.999	35.468	4.386	**0.999**

2) Weekly butter price: From table 2 we can see that, overall performance of the models have increased after the hyperparamter tuning. All the ensemble methods have performed well in this dataset which indicates that the features of this dataset are good for prediction. Support Vector Regressor, Random Forest, GradientBoost, XGBoost, Bagging had the same MSE score and XGBoost had the best MAE score for MAE where GradientBoost had the highest score for R^2.

TABLE II
RESULTS OF WEEKLY BUTTER PRICE PREDICTION

	Before Hyperparameter Tuning			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	0.001	0.105	0.903	**0.001**	0.105	0.903
RF	0.002	0.018	0.982	**0.001**	0.010	0.994
AB	0.002	0.008	0.990	0.009	0.010	0.994
GB	0.025	0.110	0.827	**0.001**	0.150	**0.995**
XGB	0.025	0.111	0.824	**0.001**	**0.007**	0.987
BG	0.004	0.022	0.971	**0.001**	0.014	0.992
LGB	0.003	0.025	0.981	0.006	0.035	0.987

3) Weekly 40 pound block cheddar cheese: From table 3 we can see that, performance of all the models have increased after hyperparameter tuning. Random Forest, AdaBoost and Bagging had the lowest score of MSE. AdaBoost had the lowest MAE score where Random Forest, AdaBoost and LightGBM had the highest R^2 score. Overall, all the models have performed well for this dataset.

TABLE III
RESULTS OF WEEKLY 40 POUND BLOCK CHEDDAR CHEESE PRICE PREDICTION

	Before Hyperparameter Tuning			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	0.011	0.087	0.809	0.002	0.004	0.053
RF	0.000	0.006	0.994	**0.000**	0.001	**1.000**
AB	0.000	0.003	0.996	**0.000**	**0.000**	**1.000**
GB	0.003	0.042	0.949	0.001	0.039	0.965
XGB	0.003	0.003	0.944	0.003	0.003	0.984
BG	0.000	0.009	0.992	**0.000**	0.006	0.996
LGB	0.000	0.015	0.986	0.002	0.002	**1.000**

4) Weekly dry Whey price: From table 4 we can see that, the performance of the models have increased after hyperparameter tuning. Random Forest, GradientBoost, XGBoost, Bagging and LightGBM had the lowest MSE score. GradientBoost, Bagging and LightGBM had the lowest MAE where Bagging have the highest R^2 score. All the ensemble methods have a good R^2 score except the AdaBoost.

TABLE IV
RESULTS OF WEEKLY DRY WHEY PRICE PREDICTION

	Before Hyperparameter Tuning			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	0.005	0.065	0.754	0.005	0.065	0.754
RF	0.000	0.002	0.999	**0.000**	0.005	0.996
AB	0.000	0.000	0.999	0.006	0.061	0.899
GB	0.001	0.002	0.951	**0.000**	**0.002**	0.998
XGB	0.001	0.025	0.947	**0.000**	0.004	0.994
BG	0.000	0.003	0.998	**0.000**	**0.002**	**0.999**
LGB	0.000	0.007	0.985	**0.000**	**0.002**	0.984

5) Weekly 500 barrel cheddar cheese price: From table 5 we can see that, there is a slight improvement after hyperparameter tuning. The base model, Support Vector Regressor obtained a good score in terms of all performance measures. All the ensemble methods have similar lowest MSE, MAE and highest R^2. This depicts that the features of this dataset are suitable for prediction.

TABLE V
RESULTS OF WEEKLY 500 BARREL CHEDDAR CHEESE PRICE PREDICTION

	Cross Validation Result			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	0.057	0.004	0.950	0.057	0.004	0.950
RF	0.000	0.001	1.000	**0.000**	0.001	**1.000**
AB	0.000	0.000	1.000	**0.000**	**0.000**	**1.000**
GB	0.000	1.000	1.000	**0.000**	**0.000**	**1.000**
XGB	0.000	0.002	1.000	**0.000**	**0.000**	**1.000**
BG	0.000	0.001	1.000	**0.000**	**0.000**	**1.000**
LGB	0.001	0.011	0.985	**0.000**	0.001	0.992

6) US daily Avocado price: From table 6 we can see that, performance of all the models have improved after hyperparameter tuning except the base model Support Vector Regressor. Notably, the ensemble methods and Random Forest could not obtain a better score than the base Support Vector Regressor model in terms of MSE score. AdaBoost had the lowest MAE and XGBoost had the highest R^2 score.

TABLE VI
RESULTS OF US DAILY AVOCADO PRICE PRICE PREDICTION

	Before Hyperparameter Tuning			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	0.058	0.004	0.950	**0.055**	0.004	0.945
RF	0.954	0.709	0.977	1.023	0.434	0.993
AB	0.674	0.623	0.984	0.663	**0.001**	0.990
GB	0.044	1.072	0.957	0.674	0.975	0.967
XGB	0.950	1.181	0.949	0.953	1.003	**0.999**
BG	1.104	0.759	0.621	1.105	0.790	0.647
LGB	2.721	1.188	0.935	1.496	0.984	0.978

7) Wheat Price Bangladesh: From table 7 we can see that, after hyperparameter tuning the results have improved for the models. Random Forest had the worst MAE and R^2 score.

AdaBoost, GradientBoost and Bagging had the similar lowest MSE, MAE and highest R^2 score.

TABLE VII
RESULTS OF WHEAT PRICE PREDICTION OF BANGLADESH

	Before Hyperparameter Tuning			After Hyperparameter Tuning		
	MSE	MAE	R^2	MSE	MAE	R^2
SVR	0.057	0.004	0.950	0.057	0.004	0.950
RF	0.018	0.093	0.891	0.040	0.152	0.740
AB	0.015	0.827	0.907	**0.000**	**0.000**	**1.000**
GB	0.044	0.160	0.731	**0.000**	**0.000**	**1.000**
XGB	0.044	0.016	0.730	0.011	0.073	0.932
BG	0.021	0.099	0.874	**0.000**	**0.000**	**1.000**
LGB	0.020	0.105	0.878	0.014	0.085	0.978

From the result analysis we can state that the ensemble models have performed well in all datasets in overall to make a better prediction. The datasets used in this study, do not have a large sample size except US daily Avocado price. Despite having a small sample size, ensemble models were able to make a better prediction and outperformed the base model.

VI. CONCLUSION

In this study, we aimed to predict the price of seven different daily commodities using supervised Machine Learning models. From the analysis, we found that ensemble methods have a very good performance in medium to large dataset and outperformed the base Support Vector Machine model. The future endeavor of this study is to work with time series and regional data of daily commodities. Since the price of commodities varies locally, inclusion of regional data will further help to predict price precisely. Apart from time series analysis, we would also like to investigate the state-of-the-art wide & deep learning [14] models.

REFERENCES

[1] Zahara, S., and M. B. Ilmiddaviq. ”Consumer price index prediction using Long Short Term Memory (LSTM) based cloud computing.” Journal of Physics: Conference Series. Vol. 1456. No. 1. IOP Publishing, 2020.

[2] Carbonneau, Real, Kevin Laframboise, and Rustam Vahidov. ”Application of machine learning techniques for supply chain demand forecasting.” European Journal of Operational Research 184.3 (2008): 1140-1154.

[3] Mustaffa, Zuriani, Yuhanis Yusof, and Siti Sakira Kamaruddin. ”Enhanced artificial bee colony for training least squares support vector machines in commodity price forecasting.” Journal of Computational Science 5.2 (2014): 196-205.

[4] Khamis, Azme, and S. N. S. B. Abdullah. ”Forecasting wheat price using backpropagation and NARX neural network.” The International Journal of Engineering and Science 3.11 (2014): 19-26.

[5] Rincon-Patino, Juan, Emmanuel Lasso, and Juan Carlos Corrales. ”Estimating avocado sales using machine learning algorithms and weather data.” Sustainability 10.10 (2018): 3498.

[6] Mehtab, Sidra, and Jaydip Sen. ”A Time Series Analysis-Based Stock Price Prediction Using Machine Learning and Deep Learning Models.” arXiv preprint arXiv:2004.11697 (2020).

[7] Shao, Yuehjen E., and Jun-Ting Dai. ”Integrated feature selection of ARIMA with computational intelligence approaches for food crop price prediction.” Complexity 2018 (2018).

[8] Nguyen, Huy Vuong, et al. ”A smart system for short-term price prediction using time series models.” Computers & Electrical Engineering 76 (2019): 339-352.

[9] Yousefi, Marziye, et al. ”Support vector regression methodology for prediction of output energy in rice production.” Stochastic environmental research and risk assessment 29.8 (2015): 2115-2126.

[10] Al Imran, Abdullah, Md Rifatul Islam Rifat, and Rafeed Mohammad. ”Enhancing the Classification Performance of Lower Back Pain Symptoms Using Genetic Algorithm-Based Feature Selection.” Proceedings of International Joint Conference on Computational Intelligence. Springer, Singapore, 2020.

[11] Al Imran, Abdullah, et al. ”The Impact of Feature Selection Techniques on the Performance of Predicting Parkinson’s Disease.” (2018).

[12] Al Imran, Abdullah, and Md Nur Amin. ”Predicting the Return of Orders in the E-Tail Industry Accompanying with Model Interpretation.” Procedia Computer Science 176 (2020): 1170-1179.

[13] Rafsunjani, Siam, et al. ”An Empirical Comparison of Missing Value Imputation Techniques on APS Failure Prediction.” (2019).

[14] Al Imran, Abdullah, Md Nur Amin, and Fatema Tuj Johora. ”Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide & Deep Learning.” 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE, 2018.

[15] Bauer, Eric, and Ron Kohavi. ”An empirical comparison of voting classification algorithms: Bagging, boosting, and variants.” Machine learning 36.1-2 (1999): 105-139.

[16] Daily Wheat Price — Kaggle. https://www.kaggle.com/nickwong64/daily-wheat-price. Accessed 30 Oct. 2020.

[17] Weekly Dairy Product Prices — Kaggle. https://www.kaggle.com/sohier/weekly-dairy-product-prices. Accessed 30 Oct. 2020.

[18] Avocado Prices — Kaggle. https://www.kaggle.com/neuromusic/avocado-prices. Accessed 30 Oct. 2020.