Original Research Article

Indian Commodity Market Price Comparative Study of Forecasting Methods - A Case Study on onion, potato and tomato

ABSTRACT

In this study several characteristics are taken into account so that the crop price forecast is accurate. Forecasting the price of agriculture commodities based on Volume, diesel price helps the agriculturist and also the agriculture mandi's in India. We look at onion, tomato, and potato trading in India and present the evaluation of a price forecasting model, and anomaly detection and compared differently Supervised, Unsupervised and Forecasting prediction models. We prefer to use wholesale prices, retail prices, arrival volumes of the agricultural commodities and Diesel prices in India. We also provide an in-depth forecasting analysis of the effect on these retail prices. Our results are encouraging and point towards the likelihood of building pricing models for agricultural commodities and to detect anomalies. These data can then be stored and analyzed. The empirical comparison of the chosen methods on the various data showed that some methods are more suitable than others for this type of problem. In this research, we did a comparative study of Auto ARIMA (Autoregressive Integrated Moving Average), RNN (Recurrent Neural Network), LSTM, VAR (vector autoregressive model), and Random Forest Regression, XGBoost in their ability to predict Retail prices of potatoes, onions and tomatoes.

KEYWORDS

Auto ARIMA, RNN, LSTM, VAR, Random forest Regression and XG Boost Regression

1. INTRODUCTION

In this project we study several mandi's across the India for onion, tomato and potato based on arrival volumes and diesel prices between 2015 to July 2020. Identified top mandi for onion, tomato and potato for the prediction of retail price.

The work focuses on analyzing such agricultural data using techniques of machine learning and gaining the knowledge obtained from the result. The comparison of results obtained from several algorithms will help in selecting the more appropriate algorithm for agricultural data to predict retail prices of onion, tomato and potato from 2015 to 2020 [4]

Agriculture commodity Wholesale price, retail price and volume for onion, tomato and potato has been a collected from Horticulture website from 2015 – 2020.

Diesel Price also collected from 2015-2020 and it contributes in deriving retail price of the

agricultural commodity. During this study of the data we observed several anomalies which are explained with more details [5]

The prices of agricultural commodities are forecasted using various time-series, machine learning, and deep learning models. onion, tomato and potato are an important commercial tree crop mainly traded in Mumbai, Bengaluru and Delhi.

Although there exist numerous online platforms such as commodity online, that provide details of prices of onion, tomato and potato from various cities predicted using VAR, Auto ARIMA, LSTM, Random forest and XG Boost. The performance of these models was evaluated and compared to identify the best model [3]

India ranks second worldwide in farm outputs. As of 2018, agriculture employed more than 50% of the Indian workforces and contributed 17–18% to the country's GDP. According to the latest report, agriculture is the primary source of livelihood for 58% population in India [10]

onion (Allium cepa) is one of the most important commercial crops of the India.

In India, the onion crop is grown in about 1.20-million-hectare area with an annual production of 19.40 million tons with a productivity of 16.12 tons per hectare.

The quantity of onion 2415.75 thousand tons is exported from India which outputs a value of 3,10,650.09 Rs. Lakhs (in 2017) [10]

Tomato is one of the most important vegetable crops cultivated for its fleshy fruits. tomato is considered an important commercial and dietary vegetable crop. The botanical name of tomato is Lycopersicon esculemtun and belongs to the family Lycopersicae. As it is short duration crop and gives a high yield, it is important from the economic point of view and hence area under its cultivation is increasing day by day. tomato is used in preserved products like ketchup, sauce, chutney, soup, paste, puree, etc. The estimated area and production of tomato for India are about 3,50,000 hectares and 53,00,000 tons respectively [13]

potato (Solanum tuberosum) is the most important food crop in the world. potato is a temperate crop grown under subtropical conditions in India. India’s potato production is 52.5 million tons in 21.8 lakh hectares [14]

I explained about different models used in this study as follows:

Auto ARIMA is essentially ARIMA with a grid search part to find the best metric parameters such as the Akaike Knowledge Criterion (AIC).

ARIMA stands for Autoregressive Integrated Moving Average. ARIMA can also be broken down into part Autoregressive (AR), part Moving Average (MA) and part Integrated (I). Models AR and MA can be used as separate models to produce

predictions of the time series. It can however produce better predictions when combined as an ARMA model.

“VAR - Vector autoregression (VAR) is a

the statistical model used to capture multiple quantity relationships as they change over time. VAR is a model of the stochastic process.

The VAR models generalize the autoregressive

single variable (univariate) model by

allowing multivariate time series” [8]

“The model of vector autoregression (VAR) adds the idea of univariate autoregression to kk

time-series regressions, in which all kk

series lagged values appear as regressors. Put differently,
a vector of time series variables on lagged vectors of
those variables are regressed in a VAR model.
As for AR (pp) models, the lag order is denoted by pp so the VAR(pp) model of two variables
Xt and Yt (k=2) is given by the equations" [8]
"Yt=β10+β11Yt−1+⋯+β1pYt−p+γ11Xt−1+⋯+γ1pXt−p+u1t,

Xt=ß20+ß21Yt-1+⋯+ß2pYt-p+γ21Xt-1+⋯+γ2pXt-p+u2t" [8]
"The ββs and γγs can be estimated using OLS on each equation" [8].

"Randomforest The "forest" it builds, is an ensemble of decision trees, usually trained by the method of "bagging." The general idea of the bagging method is that the overall result increases with a combination of learning models. Advantages of Random Forest algorithm – it will avoid the overfitting problem, handle missing values identify important features and can be used in both classification and regression problems" [3]
"XGBoost is a Machine Learning algorithm based on a decisiontree ensemble that uses a gradient boosting system" [9]
"LSTM - Long Short-Term Memory networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory" [1]
"LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. It trains the model by using back-propagation. In an LSTM network, three gates are present: Input gate, output gate and forget gate" [1]

Diesel price also play a major role in deciding retail prices of the agricultural commodity, Diesel price change every day due to changes in international crude oil price and policies. [2]
Research work done to identify anomalies in the prediction of onion, tomato and potato retail price. In our project we identified the anomalies based on the events which results in sudden rise in price trend. [6] [1]

It is a great opportunity to implement and compare different machine learning algorithms with the assumption of variables are not dependent, with normal distribution and with real time data collected from different websites. [7]

Work done in comparing different performance result RMSE for all the forecasting prediction models will help to choose the right models for the prediction of onion, tomato and potato retail price.

In this study we found the better models as follows:
for the onion Retail Price prediction: LSTM, tomato Retail Price prediction: LSTM and potato Retail Price prediction: XG Boost

2. Objectives of the Study

The problem statement of the "Indian commodity market price comparative study
of forecasting methods - A case study on onion,
potato and tomato" are to develop a machine learning models that can accurately predict the retail price of onion, tomato and potato from the collected data. The "National Horticulture Board runs a portal which provides retail prices from across 30 district centers across the country". Web crawling done to collect data from January 2015 till July 2020.
I also collected diesel prices in India between January 2015 till July 2020.
The collected data from different websites is defined as the set of features selected for the agricultural commodity in India – onion, tomato and potato. The idea behind the problem is that an

accurate forecasting model will be able to reduce the variation in forecasting of retail prices successfully.

During the study of different forecasting prediction models, also need to identify different anomalies which are outside the range of 2 sigma levels. Need to identify key events based on anomalies which resulted in rise or fall or volume and resulted in changes in retail prices of the agricultural commodities – onion, tomato and potato. This problem is also a supervised task because the targets for the training data are known future of time and the model will learn based on tabular data.

3. DATA ACQUISITION AND UNDERSTANDING

The National horticultural board website run by the Government of India makes publicly available the monthly data on mandi wholesale and retail prices in Rupees and arrival volume in quintals of onion, tomato and potato from different mandi are in India. We scraped all the data for onions, tomato & potatoes from all mandi's for 5 years from January 1st 2015 to July 2020"

Diesel price in rupees per liter also collected which is also contributing to the prediction of retail prices, we collected for 5 years from January 1st, 2015 to July 2020.

There were missing values in all these commodities which will be treated and explained in the experimental section.

These are the variables considered from the collected data as follows:

- Retail Price of onion, tomato and potato are in Rupees
- Retail Price of onion, tomato and potato are in Rupees
- Wholesale price of onion, tomato and potato are in Rupees
- Arrival volume in quintals (100kg/quintal) for onion, tomato and potato
- Diesel Price per month are in Rupees/litre

- Downloaded onion, tomato and potato data from January 2015 till July 2020 from National horticultural and Diesel price from Gatti website
- Combined data for commodity prices, volume and diesel prices into onion, tomato and potato using panda's data frames, respectively.
- Data is having missing values <15%, total 67 rows of data and for different mandi's for onion, tomato and potato respectively.

Table 1, shows different parameters of commodity onion. Similar data collected for tomato and potato dataset

Center / Mandi	Year	Month	Retail Price in rupee	Volume in quintal	Diesel Price in rupee
bengaluru	2015	January	2791	74978	53.74
bengaluru	2015	February	2750	42164	51.96
bengaluru	2015	March	2396	41339	55.41
bengaluru	2015	April	2260	43534	52.56
bengaluru	2015	May	2458	44227	58.19
bengaluru	2015	June	3145	43403	56.69
bengaluru	2015	July	3314	45957	53.37
bengaluru	2015	August	4029	83437	48.23

Table1:datasample

Figure 1 shows different top 10 mandi's/center for onion commodity-based Arrival volume in quintals, considered data from Jan 2015 to Jul 2020. bengaluru is the top mandi for onion

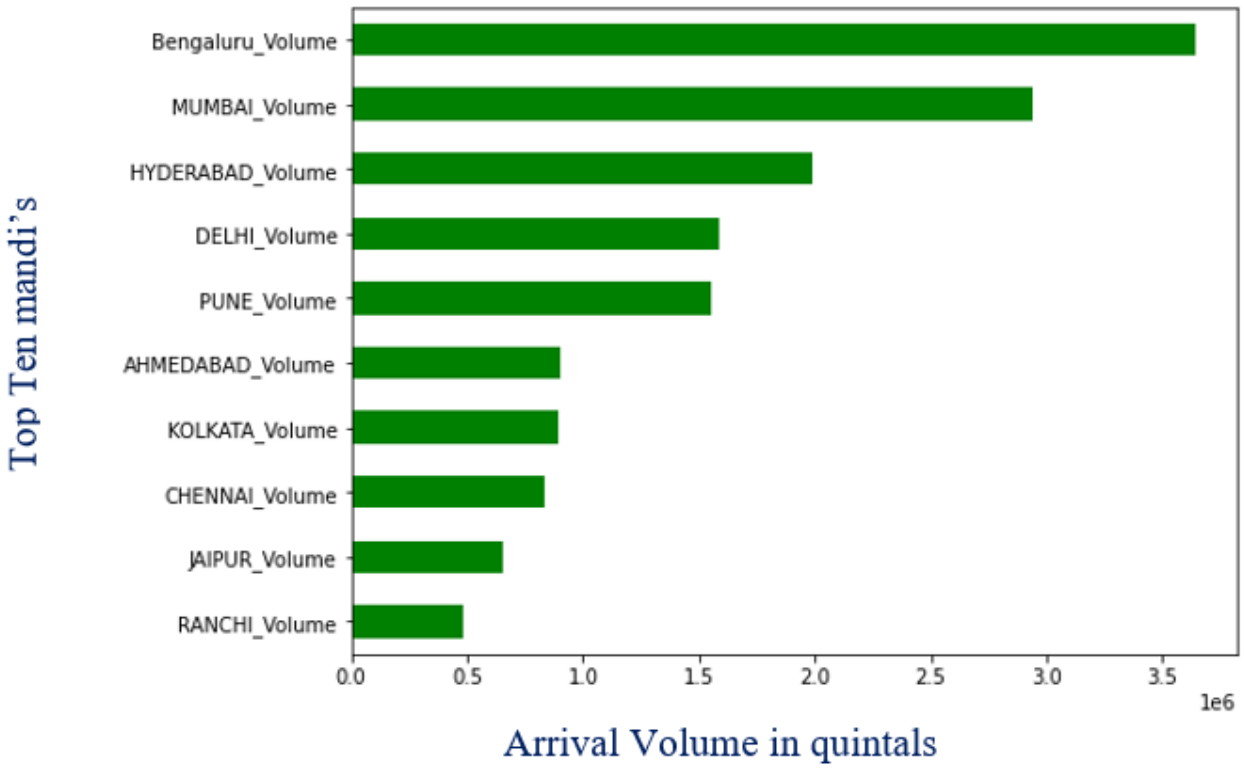

Figure 2 shows different top 10 mandi's/center for onion commodity-based Arrival volume in quintals

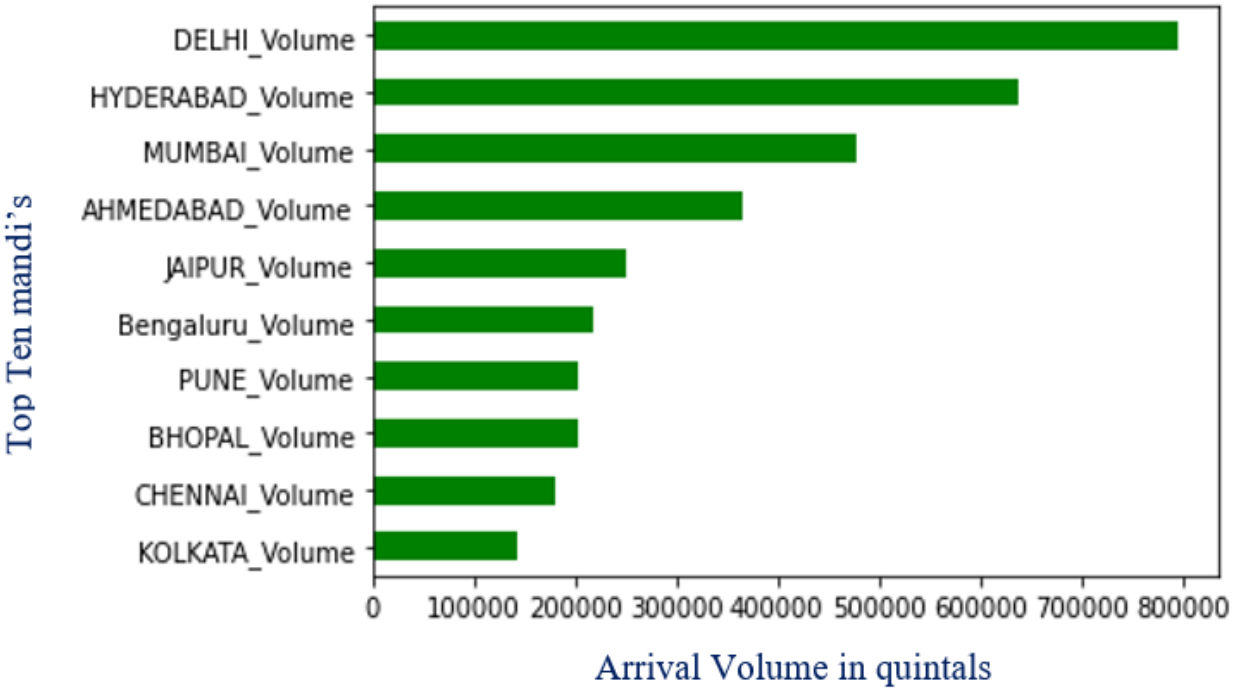

Figure 3 shows different different top 10 mandi's/center for tomato commodity-based Arrival volume in quintals, considered data from Jan 2015 to Jul 2020. delhi is the top mandi for tomato

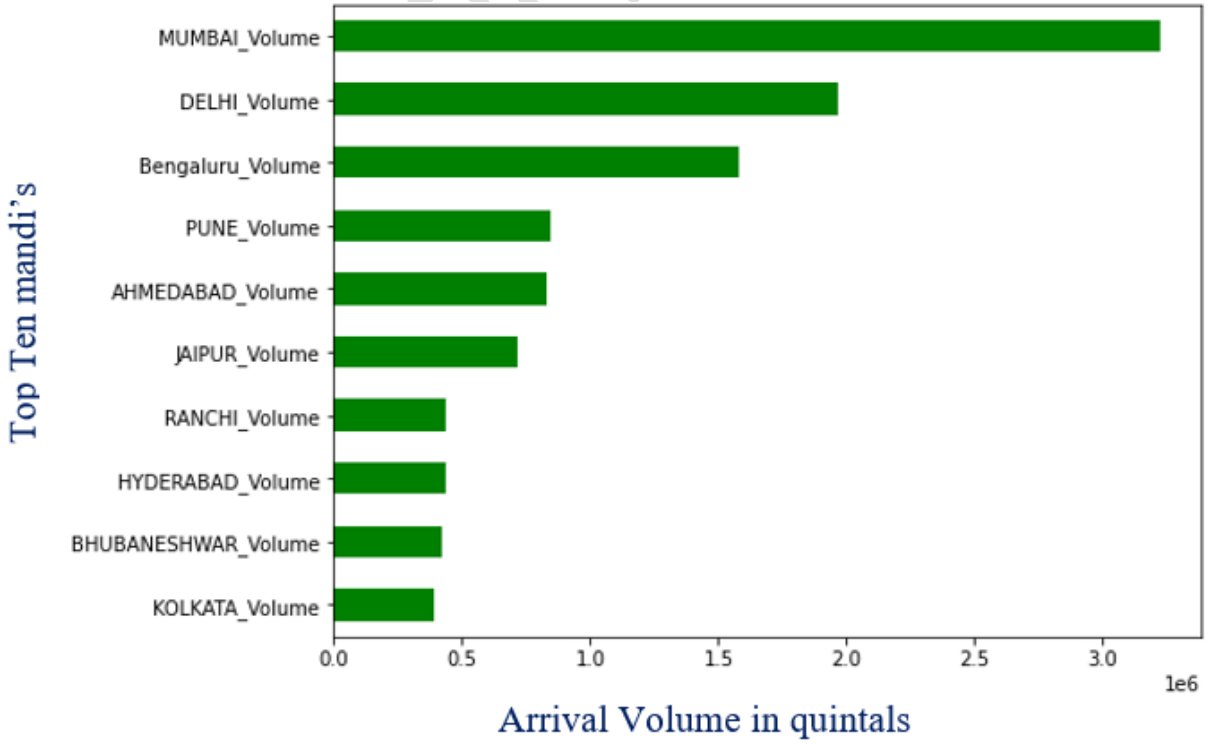

Figure 4. shows different top 10 mandi's/center for potato commodity-based Arrival volume in quintals, considered data from Jan 2015 to Jul 2020. mumbai is the top mandi for potato

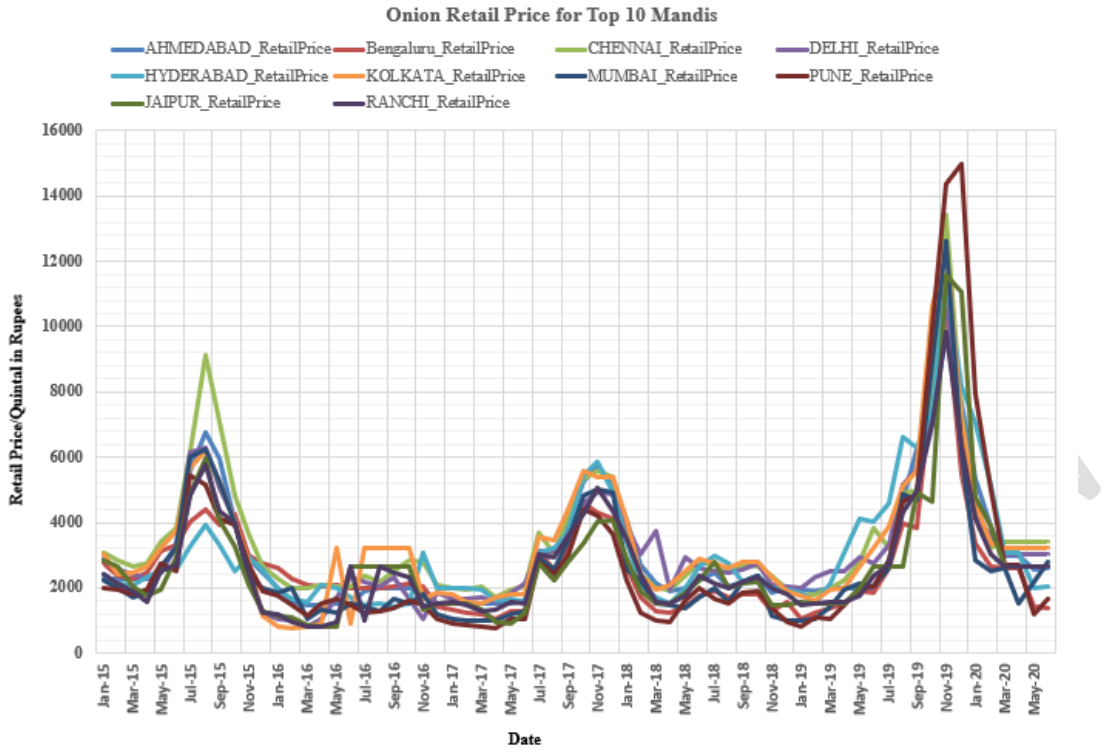

Figure 5 The figure shows the retail price of onion commodity in different mandi's between Jan 2015 till July 2020.

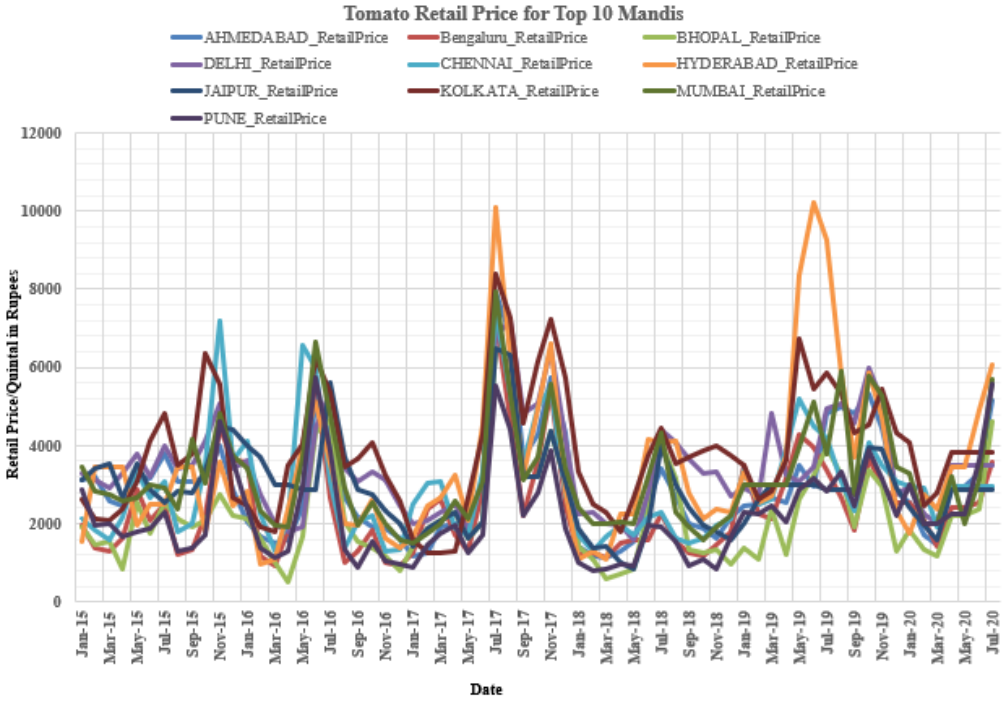

Figure 6 shows the retail price of tomato commodity in different mandi's between Jan 2015 till July 2020.

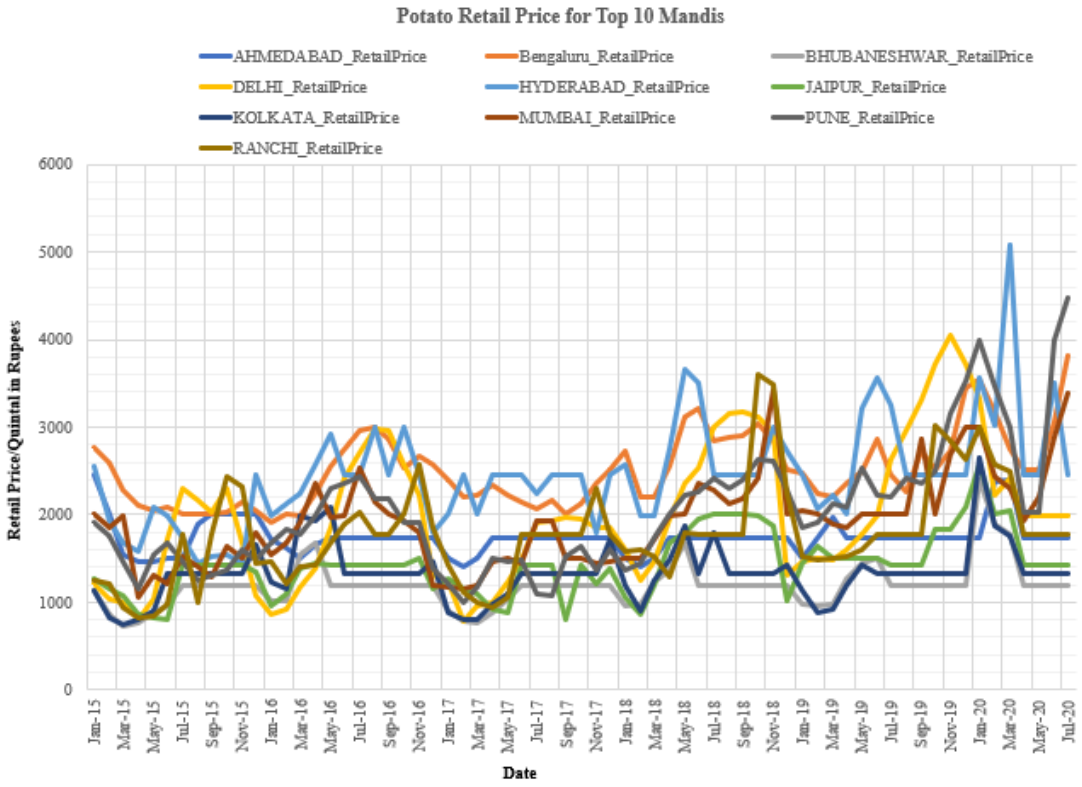

Figure 7 shows the retail price of potato commodities in different mandi's between Jan 2015 till July 2020

Following are the activities during data preparation:

Missing Value Imputation
Identify Top 10 mandi's of India for onion, tomato and potato based on volume
Data Distribution, Scaling
Exploratory data analysis

Missing value Imputation: Observed missing value in onion, tomato, potato and diesel data. Missing value imputation using replacing missing value by mean.
Top mandi's for onion is bengaluru, tomato is delhi and for potato is mumbai.

Data Scaling:
Variables calculated at various scales do not contribute equally to the model fitting & model role learn ed and which ultimately generates a bias.
Thus, standardized feature wise ($\mu=0$, $\sigma=1$) is typically used before model fitting to solve this potential issue.
StandardScaler does away with the mean and measures the variance of any features. StandardScaler eliminates the mean and scales to the unit variance for each function/variable. This scaling is achieved independently of feature-wise.
Used Scikit python's basic scaler function to scale variables for onion, tomato and potato.

"Anomalies: Unexpected values often surface in a distribution of values, especially when working with data from unknown sources which lack poor data validation controls" [15]. A detailed explanation of these anomalies.

onion, tomato and potato are three vegetables that drive food inflation in India.
Regional supply and demand, changes in weather conditions, increased production rates, increased wages, higher cost for transportation, lack of availability of manpower, lack of sufficient cold storage are some of the common factors that significantly contribute to onion, tomato and potato price's inflation. We identified key events from the data which are anomalies in the data of onion, tomato and potato, from the collected data between January 2015 till July 2020.

Figure, 9 and 10 represent anomalies in the retail prices of onion, tomato and potato respectively.

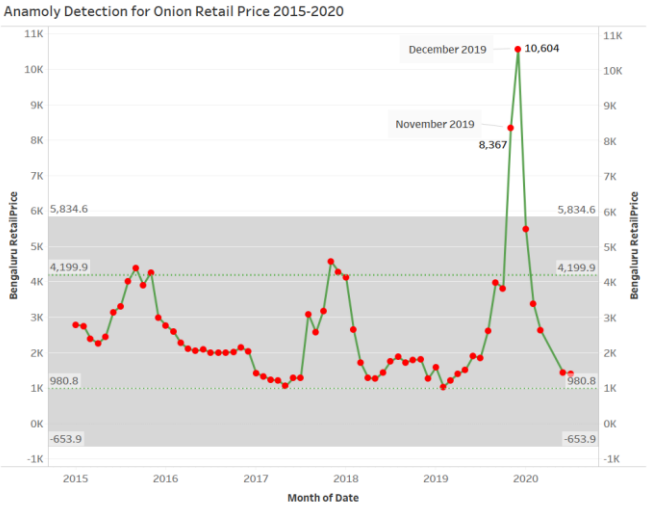

Figure 8 shows Anomalies identified in November 2019 and December 2019 in onion commodity retail price in bengaluru Mandi using ± 2σ level from the data

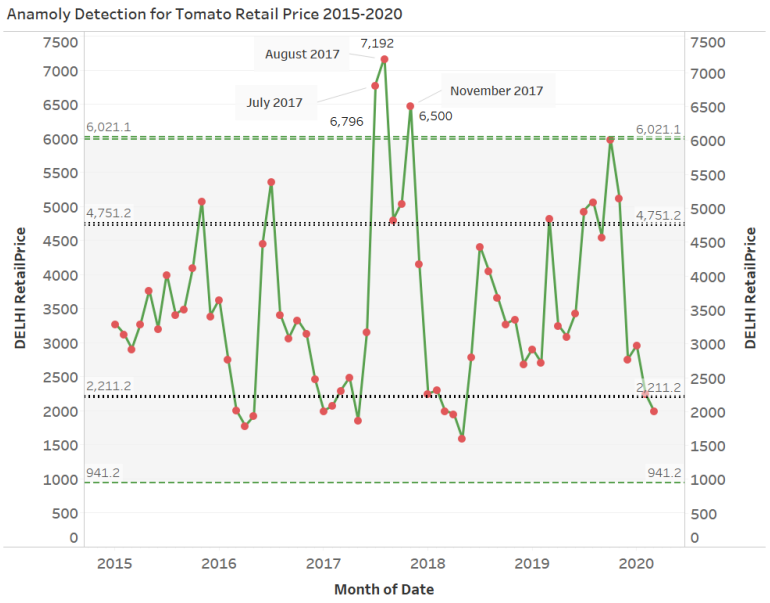

Figure 9 shows Anomalies identified in July 2017, August 2017 and November 2017 in tomato commodity retail price in delhi Mandi using ± 2σ level from the data

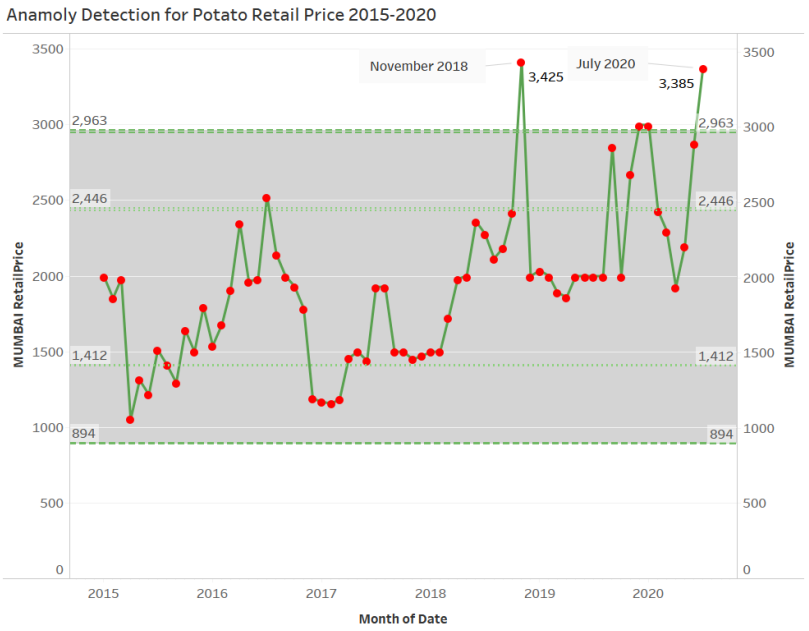

Figure 10 shows Anomalies identified in November 2018 and July 2020 in potato commodity retail price in mumbai Mandi using ± 2σ level from the data.

Following statistical test performed on onion, tomato and potato for bengaluru, delhi and mumbai mandi's are as follows:

- Testing Causation using Granger's Causality Test
- Cointegration Test
- Augmented Dickey-Fuller Test (ADF Test)
- Durbin Watson test
- Autocorrelation and functional analysis

Granger's causality test help to understand causality with the help of P Value <0.05

"Cointegration Test: It helps to establish the presence of a statistically significant connection between two or more time series. This we have performed using the Johansen test. The results show retail prices of onion, potato and tomato is having a significant connection with time series"[16].
"Augmented Dickey-Fuller Test (ADF Test):The Augmented Dickey-Fuller Test (ADF) is the Stationary Unit Root Test" [17]. Unit roots in your analyses of the time series will trigger unpredictable results. These tests are run for bengaluru, mumbai and delhi mandi data for

onion, tomato and potato to understand the data is stationary or not. The test is run twice before and after differencing to ensure data is stationary and rejecting the null hypothesis.
Durbin Watson test: This is a test of autocorrelation (also known as serial correlation) / tests for homoscedasticity in regression residuals.
This test also can be performed using OLS (ordinary least square) method.
Autocorrelation is a time series similitude over consecutive periods. It can cause the standard error to

be underestimated and can cause predictors to be large
when not. This test searches for a particular form of
serial correlation, the AR (1) method. Thus, the test
statistic equals 2 for r = 0, suggesting no serial
correlation. This number often ranges from
0 to 4. The closer the numbers are to 0, the greater the proof for a strong serial correlation.
The closer to the four,
the greater the evidence of negative serial correlation.
Autocorrelation f u n c t i o n (ACF) and seasonality analysis: ACF is a form of dependency on the sequence.
Autocorrelation, in particular, is when a time series is linearly connected to a lagged version of itself.
Partial autocorrelation solves this problem by measuring the correlation between variables and it's timeseries when the influence of the intermediate variables has been removed.
In comparison, correlation is simply when there are linear ties between two independent variables, the trend in time series data, deleted the seasonality

Figure 11 shows seasonal plot for onion commodity retail price in bengaluru Mandi

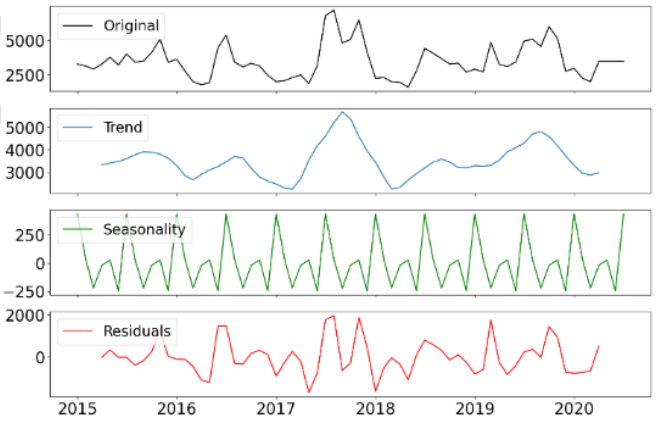

Figure 12 shows the seasonal plot for tomato commodity retail price in delhi Mandi

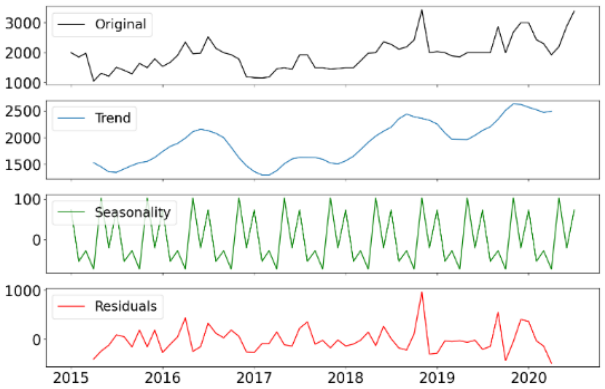

Figure 13 shows the seasonal plot for potato commodity retail price in delhi Mandi

4. Data Modeling

Modeling is performed by splitting train and test for onion, tomato and potato data from bengaluru, delhi and mumbai Mandi's respectively. In our datasets splitting of train and test is done in the ratio of 75%, 25%.

RMSE – Root Mean Square Error is a standard method for calculating a model's error when predicting quantitative results.

RMSE is formally defined as follows:

$$\mathbf{RMSE} = \sqrt{[\ \Sigma(Pi - Oi)2 / n\ ]}$$

RMSE formula where Σ is a symbol for "sum", Pi is the predicted value for the ith observation in the dataset, Oi is the observed value for the ith observation in the dataset, and n is the number of observations

RMSE has a twofold aim: To act as a heuristic model for the training, to assess the utility / accuracy of trained models. I used RMSE as a measure to identify and to choose good prediction models for further analysis.

I chose LSTM for onion , LSTM for tomato and XGBOOST for potato as best prediction models

I checked whether by using Auto ARIMA, VAR, Random forest, XG Boost, Multivariate deep learning model and LSTM for multivariate models, did evaluation of right models use its RMSE results. Best 3 Models for onion, tomato and potato retail price predictions are as follows:

onion Retail Price prediction: LSTM, Deep learning and Random forest

tomato Retail Price prediction: LSTM, Deep learning and Random forest

potato Retail Price prediction: XG Boost, Random forest and Auto ARIMA (Univariate)

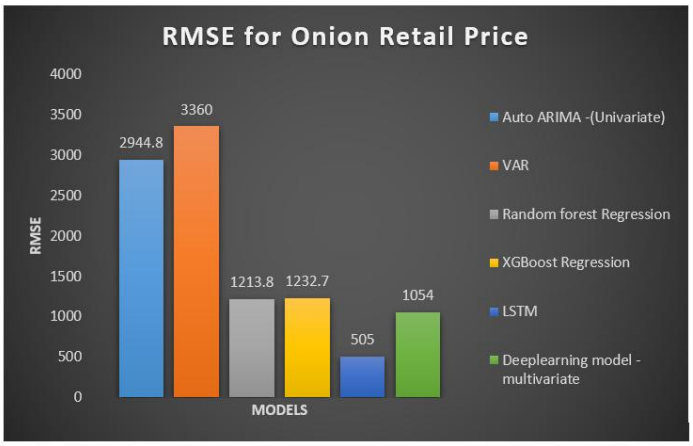

Figure 14 RMSE evaluation from prediction models for onion retail price

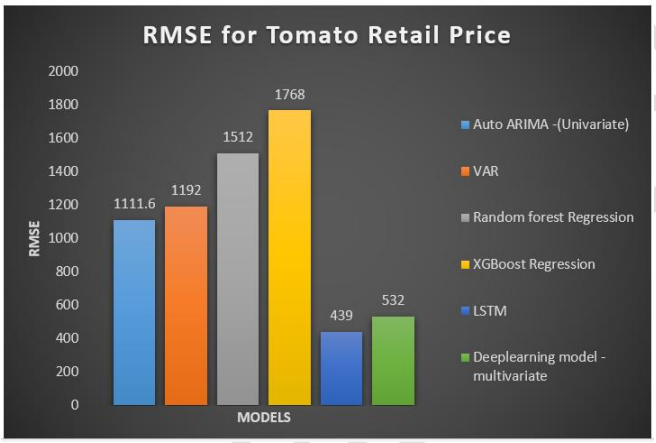

Figure 15 RMSE evaluation from prediction models for tomato retail price

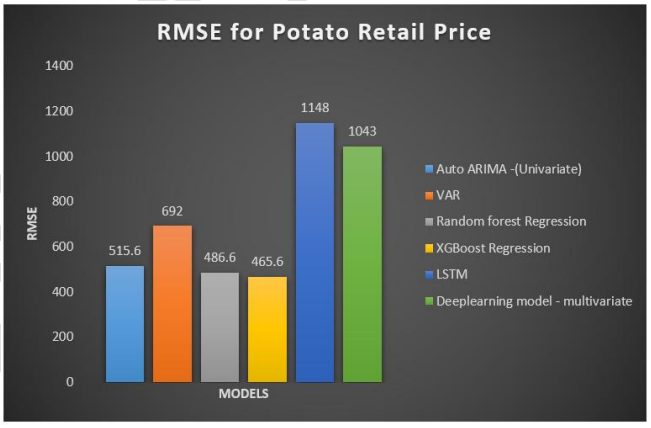

Figure 16 RMSE evaluation from prediction models for potato retail price

5. Deployment

Finally, we built different models and tested all the different metrics and are now ready to incorporate the model in
output. Proposal for the deployment of models for the prediction of onion, tomato and potato are as follows:

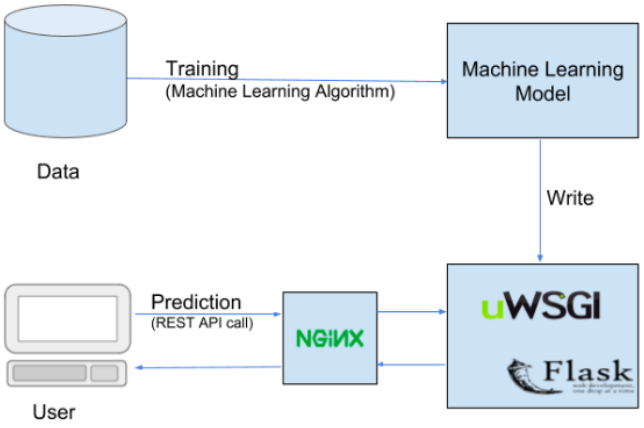

Figure 17 Data pipe line deployment architecture proposal

6. Analysis and Results

In this project, I collected 5 years of data for commodity price for onion, tomato and potato.
In this research, we identified key models to predict multivariate prediction considering volume in quintal and diesel price in rupees for onion, tomato and potato from January 2015 till July 2020. In this research, we also understood the relationship of retail price with wholesale price, volume and diesel price of onion, tomato and potato. During data understanding also observed a linear relationship between the wholesale and retail price of the onion, tomato and potato. I took a decision to drop the wholesale price variable from the study to avoid overfitting.

7. Conclusions and Recommendations for future work

In this research, we also understood the relationship of retail price with wholesale price, volume and diesel price of onion, tomato and potato. During data understanding also observed a linear relationship between the wholesale and retail price of the onion, tomato and potato. I took a decision to drop the wholesale price variable from the study to avoid overfitting.
There are few major anomalies identified from the data for onion, tomato and potato and identified due to key events are as follows:
onion Anomalies were identified from the data during this project which happened during November 2019 and December 2019.
“Heavy unseasonal rainfall has locked latest onion produce in wet fields throughout onion growing states of Maharashtra, Karnataka and Telangana. As a ripple effect, onion prices increased everywhere in the country. The politically sensitive bulb’s prices however are set to ease in coming months, sparking fears farmers will not get fair prices for their crop” [18]. “Between July 2017 to November 2017 there was a huge reduction in the arrival of volume in different mandi’s in India” [19]. tomato Anomalies were identified from the data during this project which happened during July 2017, August 2017 and November 2017. potato Anomalies were identified from the data during this

project which happened during Nov 2018 and July 2020. "potato price increased continuously from July till December 2018" [20]

"Supply shortage, rise in demand push potato prices in India up by 40% in major cities" [21]. The prediction models selected also able to understand the data and pattern, we identified different models like Auto ARIMA, VAR, LSTM, Deep learning model, Random forest and XGBoost. Based on different tests and data understanding we considered necessary variables for further prediction of Retail price which is the key output we are predicting in this project.

Recommendation for future work: Our study in this research helps to understand the geological location of different mandi's based on the overall volume of onion, tomato and potato from January 2020 till July 2020. We are also having potential opportunity to continue this work to further do the deployment, rest API, web apps and also to include other variables which are affecting Retail prices of onion, tomato and potato. Key contribution in this research is to emphasize the role of multivariate Machine learning prediction and other forecasting methods which will help in the future .

REFERENCES

[1] L. Madaan, A. Sharma, P. Khandelwal, S. Goel, P. Singla, and A. Seth, "Price forecasting & anomaly detection for agricultural commodities in India," 2019. doi: 10.1145/3314344.3332488.

[2] S. Shome, U. Khatri, D. Joshi, and S. Mehndiratta, "Dynamic Fuel Pricing in India: An Event Study Methodology," Int. J. Manag. Stud., vol. V, no. 4(6), p. 32, 2018, doi: 10.18843/ijms/v5i4(6)/05.

[3] D. Zhang, S. Chen, L. Liwen, and Q. Xia, "Forecasting Agricultural Commodity Prices Using Model Selection Framework with Time Series Features and Forecast Horizons," IEEE Access, vol. 8, pp. 28197–28209, 2020, doi: 10.1109/ACCESS.2020.2971591.

[4] H. Ouyang, X. Wei, and Q. Wu, "Agricultural commodity futures prices prediction via long- and short-term time series network," J. Appl. Econ., vol. 22, no. 1, pp. 468–483, 2019, doi: 10.1080/15140326.2019.1668664.

[5] M. R. Istambul et al., "Performance Analysis of the Regression and Time Series Predictive Models using Parallel Implementation for Agricultural Data," Procedia Comput. Sci., vol. 132, no. 6, pp. 52–64, 2019, doi: 10.1016/j.procs.2018.05.187.

[6] C. O. F. onion, "Supplementary Material : Price Forecasting & Anomaly Detection for Agricultural Commodities in India," 2016.

[7] C. Bratsas, K. Koupidis, J. M. Salanova, K. Giannakopoulos, A. Kaloudis, and G. Aifadopoulou, "A comparison of machine learning methods for the prediction of traffic speed in Urban places," Sustain., vol. 12, no. 1, pp. 1–15, 2020, doi: 10.3390/SU12010142.

[8] Wikipedia, "VAR," 2020. https://en.wikipedia.org/wiki/Vector_autoregression#:~:text=Vector autoregression (VAR) is a,type of stochastic process model.&text=This equation includes the variable's,model%2C and an error term.

[9] Timothy Susanto, "XG Boost," 2020. https://mc.ai/xgboost-time-series-for-forecasting-stocks-price/.

[10] Wikipedia, "onion In india," WIKIPEDIA, 2020. https://en.wikipedia.org/wiki/Agriculture_in_India.

[13] Wikipedia, "tomato," WIKIPEDIA, 2020. https://en.wikipedia.org/wiki/tomato.

[14] Wikipedia, "potato," wikipedia, 2020. https://en.wikipedia.org/wiki/potato.

[15] A. Bhattacharya, "Anomalies," towardsdatascience, 2020. https://towardsdatascience.com/effective-approaches-for-time-series-anomaly-detection-9485b40077f1#:~:text=What is Time Series Anomaly,from rest of the data.

[16] Corporatefinanceinstitute, "Cointegration test," corporatefinanceinstitute, 2020. https://corporatefinanceinstitute.com/resources/knowledge/other/cointegration/.

[17] wiki, "Augmented_Dickey test," WIKIPEDIA, 2020. https://en.wikipedia.org/wiki/Augmented_Dickey–Fuller_test.

[18] E. Times, "Anomaly news1," Times, 2019. https://economictimes.indiatimes.com/news/economy/agriculture/flooding-in-major-onion-producing-states-has-led-to-spike-in-rices/articleshow/71975921.cms?from=mdr.

[19] Hindu, "Anomaly news2," Hindu, 2019.

https://www.thehindubusinessline.com/economy/agri-business/onion-tomato-price-spike-season-not-the-only-reason/article9957255.ece.
[20] F. Express, “Anomaly 3,” Financial express, 2019. https://www.financialexpress.com/economy/potato-prices-are-up-nearly-50-in-december-but-here-is-a-good-news/1442954/.
[21] P. News, “Anomaly 4,” potato news, 2019. https://www.potatonewstoday.com/2020/07/28/supply-shortage-rise-in-demand-push-potato-prices-in-india-up-by-40-in-major-cities/
“Global EV Outlook 2020,” Glob. EV Outlook 2020, 2020, doi: 10.1787/d394399e-en.