April 2020

Tomato price variation factors

By Paula Soto

***Abstract*—Exploratory Data Analysis, Time Series Forecasting, ensemble models, recurrent neural network, comparison of the results with CausaLens application. Tools used in this exploratory study to predict tomato price and variation factors.**

I. Introduction

The objectives of this project are to analyse the price variation of tomato, its seasonality, and to develop a model to forecast the monthly tomato prices in the UK. To avoid autoregression, to compare and discover the relation with the tomato price, the historical weather data in Bognor Regis, West Sussex, where the tomatoes are produced, was used as well as price data for other commodities such as vegetables and oil.

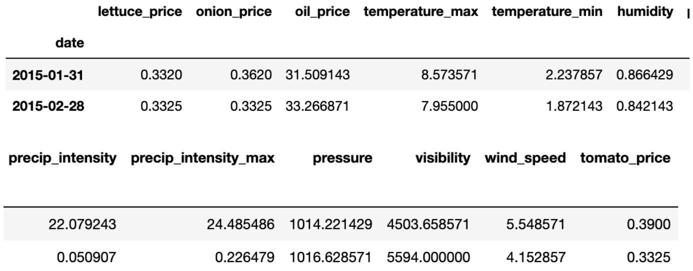
Forecasting tomato prices can provide critical and useful information to make marketing decisions.

To get historical weather data of Bognor Regis a REST API from Dark Sky was used.

The data is explored in the following parts:

1) Exploratory data analysis using Time Series.
2) Forecasting the Tomato price with Prophet.
3) Forecasting the Tomato price with ensemble models.
4) Forecasting the Tomato price with RNN.

II. Exploratory Data Analysis



date	lettuce_price	onion_price	oil_price	temperature_max	temperature_min	humidity	precip_intensity	precip_intensity_max	pressure	visibility	wind_speed	tomato_price
2015-01-31	0.3320	0.3620	31.509143	8.573571	2.237857	0.866429	22.079243	24.485486	1014.221429	4503.658571	5.548571	0.3900
2015-02-28	0.3325	0.3325	33.266871	7.955000	1.872143	0.842143	0.050907	0.226479	1016.628571	5594.000000	4.152857	0.3325

Fig. 1: Dataset

A. EDA: Data preparation

The datasets used were: vegetables price, weather data and oil price. It was necessary to clean the data, see some basic statistics variables about the distribution of each feature, replace the missing data with the mean value for that feature, resample the data monthly by mean, parse the column dates and make the time as the index of the dataset. After this, the resulting datasets were merged to one, and it was necessary to normalize the data to improve the data visualization in the charts. Figure 1 shows the dataset after data preprocessing.

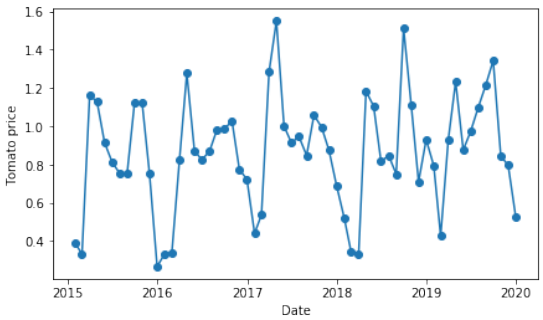


Fig. 2: Mean monthly of tomato price variation

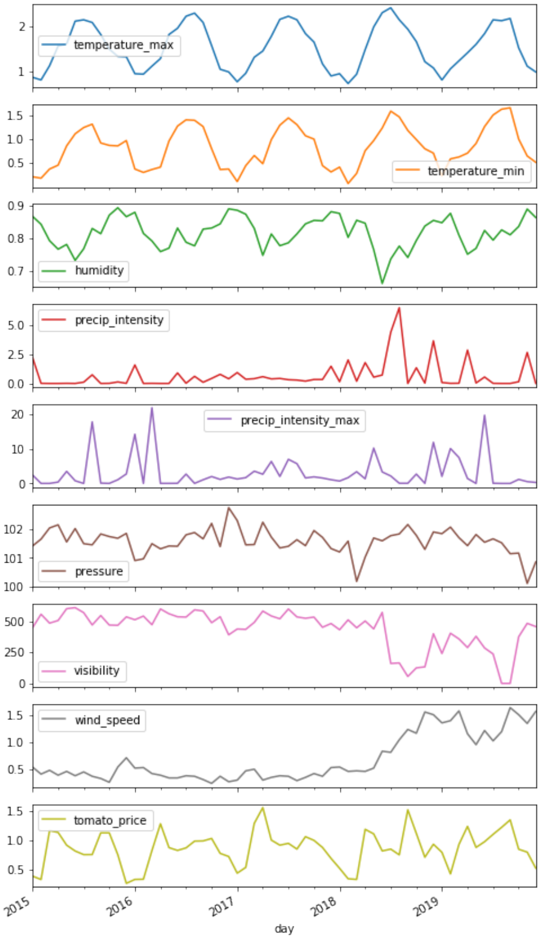


Fig. 3: Mean monthly weather variation

B. EDA: Description of analysis

Figure 2 shows the mean monthly tomato price variation. The features date and tomato price were used, and the data was

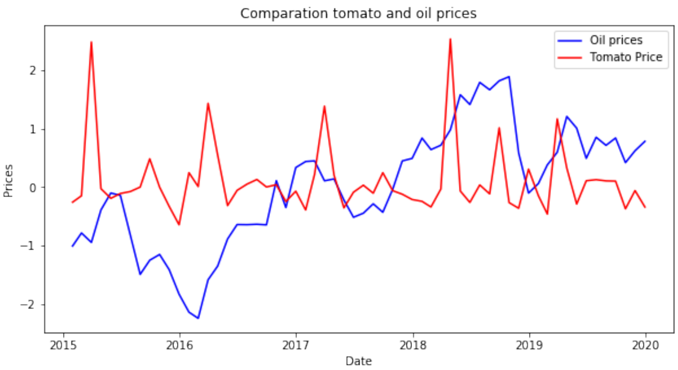


Fig. 4: Mean monthly oil price (GBP) variation

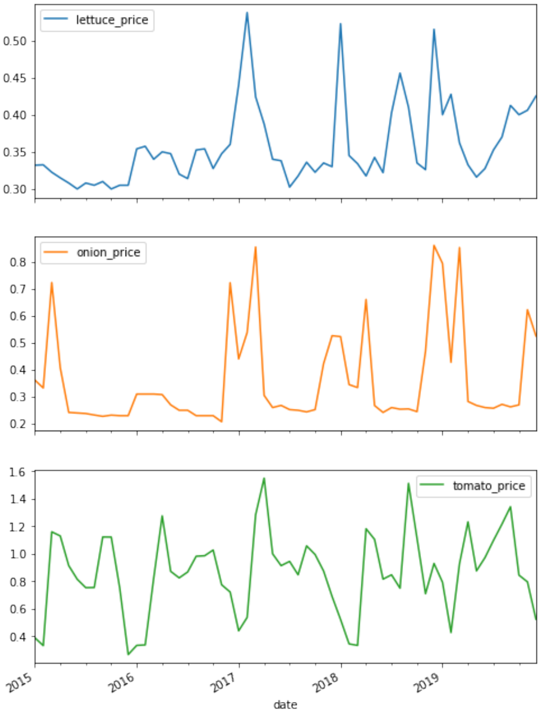


Fig. 5: Mean monthly vegetables price

resampled monthly by mean. We can see a clear seasonality: the lower tomato prices are at the beginning and the end of the years, there is a fall in the prices in the Summer, and the higher values are approximately in April, September and October.

Figure 3 shows a mean monthly weather variation. Weather historical data in Bognor Regis, West Sussex, was used, where the tomatoes are produced. There could be a relation between weather variables and the tomato price, due to its impact in the tomato production: when the maximum temperature fall the tomato price is lower.

Figure 4 shows a plot of mean monthly oil price variation. The oil price was included because the tomatoes need to be transported, the oil prices might have an influence on shipping cost and finally influence in the tomato price. In general the peak in tomato prices coincides with the rise in oil prices.

Figure 5 shows mean monthly vegetables price variation, these variables were normalized(mean 0 and standard deviation 1).

The lettuce and onion price were included because it might be a correlation between a dip price of lettuce or onion, and the tomato price, since if there is a very large harvest of these vegetables, then their price fall and the tomato price rises. The peak of tomato prices could be related to dropping lettuce prices.

III. Forecasting the tomato price

A. *Forecasting the tomato price with Prophet*

The goal is the tomato price forecast using Prophet. The Prophet package was chosen because fits an easily interpretable regression model, which can account for seasonality trends of the tomato price variation.

Figure 6 represents the tomato price prediction in the next 4 years. Weekly tomato price variation dataset was used, with date and tomato price features.

The figure shows: the black dots correspond to the observed tomato price each week, the dark blue line corresponds to the estimated tomato price in four years. Finally, the light blue lines correspond to the 80 per cent confidence interval for the model's predictions.

As shown in figure 7, the trend is the tomato price increases during the years because the inflation.

The yearly seasonality in figure 8 seems there is a dip in prices around the summer, it might be because the tomato harvest is in summer and there are more volume in the market.

Diagnostics: to measure forecast error it was used the Prophet functionality for cross-validation of time series, the result of mean square error was 0.1.

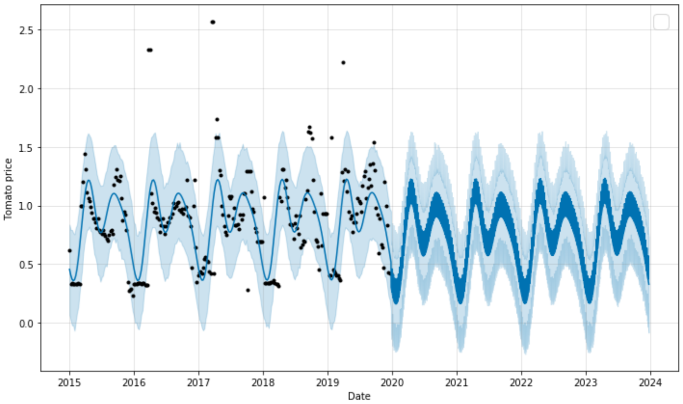


Fig. 6: A plot of the Tomato price forecast for the next 4 years

B. *Forecasting the tomato price with ensemble models*

The ensemble technique that was used in this section was bagging, Random Forest Regressor. All the features were used to forecast the tomato price. Decision Tree was used as the baseline model to compare with the Random Forest model, and Random Forest model with GridSearchCV.

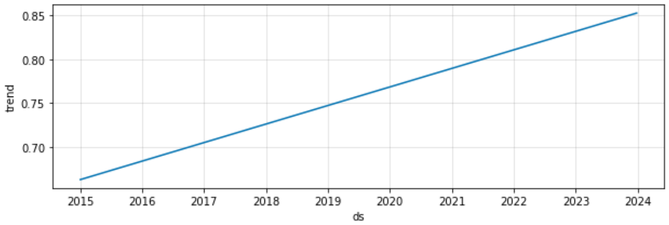


Fig. 7: Tomato price trend

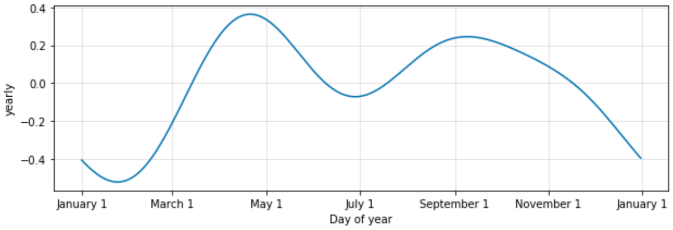


Fig. 8: Tomato price yearly trend

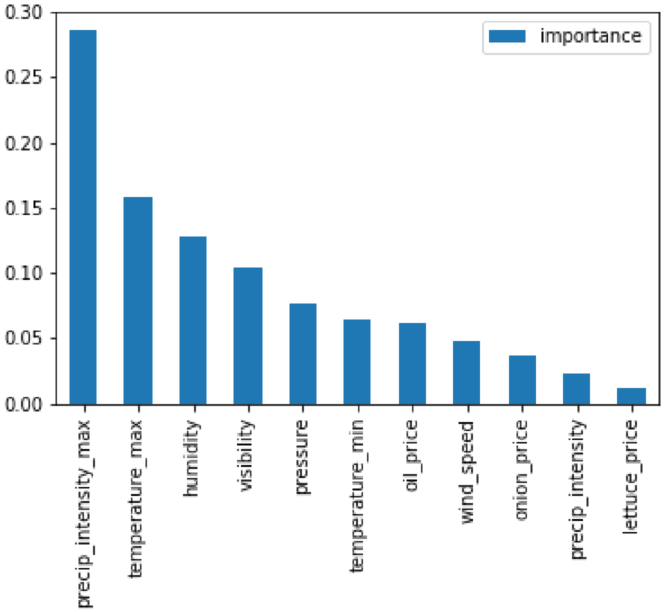


Fig. 9: Random Forest model with GridSearchCV: Feature importance

Decision Tree has the best result. The Random Forest regressor with the default hyper-parameters resulted in a 0.715 accuracy score. After hyper-parameter tuning with GridSearchCV, cross-validation to avoid overfitting, its accuracy score increased to 0.756.

In figure 9 we can see the importance of the features. The maximum precipitation intensity, the maximum temperature and humidity are the features with the greater impact on the tomato price variation.

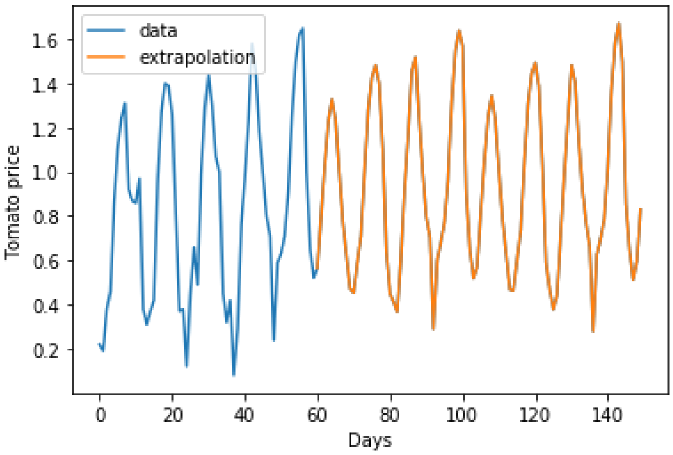


Fig. 10: Tomato price, its trend and extrapolation

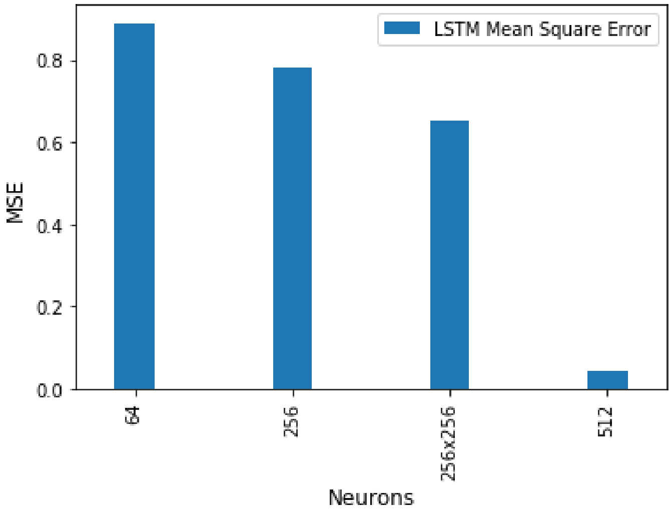


Fig. 11: LSTM: Mean square error with different models

C. *Forecasting the Tomato price with RNN*

The purpose of this section was forecasting the Tomato price using a recurrent neural network with Long Short-Term Memory (LSTM), a deep learning system that avoids the vanishing gradient problem.

All the features were used, checking before there were no missing data, and then a standard scaling to the data was applied.

The goal was to be able to predict the tomato price tomorrow, to the prediction was used daily data instead of monthly.

As it is a regression problem, in all the cases linear activation function was used.

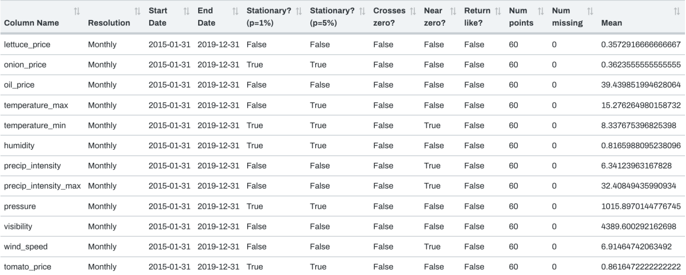
As the figure 11 shows, the diagnostic measure used was the loss of mean square error, with the differences detailed below:

- The first LSTM model used has a layer with 64 neurons, without regularization, epochs: 100, resulted in mean squared error: 0.89.

- Then the number of neurons was increased to 256, resulted in mean squared error: 0.78.
- Then two layers of 64 neurons, dropout, with regularisation, epochs: 500 were used, resulting in mean square error: 0.65
- The final result was LSTM with two layers, 512 neurons, without regularisation, epochs: 500, resulting in mean square error: 0.040 and an accuracy: 0.97

Figure 10 shows tomato price extrapolation compared to the general trend.

One of the main limitations of neural networks is that they typically need a ton of data for training, and the dataset used in this project was by no means enough.



Column Name	Resolution	Start Date	End Date	Stationary? (p=1%)	Stationary? (p=5%)	Crosses zero?	Near zero?	Return like?	Num points	Num missing	Mean
lettuce_price	Monthly	2015-01-31	2019-12-31	False	False	False	False	False	60	0	0.3572916666666667
onion_price	Monthly	2015-01-31	2019-12-31	True	True	False	False	False	60	0	0.3623555555555555
oil_price	Monthly	2015-01-31	2019-12-31	False	False	False	False	False	60	0	39.439851994628064
temperature_max	Monthly	2015-01-31	2019-12-31	False	True	False	False	False	60	0	15.276264980158732
temperature_min	Monthly	2015-01-31	2019-12-31	True	True	False	True	False	60	0	8.337675396825398
humidity	Monthly	2015-01-31	2019-12-31	True	True	False	False	False	60	0	0.8165988095238096
precip_intensity	Monthly	2015-01-31	2019-12-31	False	False	False	True	False	60	0	6.34123963167828
precip_intensity_max	Monthly	2015-01-31	2019-12-31	False	False	False	True	False	60	0	32.40849435990934
pressure	Monthly	2015-01-31	2019-12-31	True	True	False	False	False	60	0	1015.8970144776745
visibility	Monthly	2015-01-31	2019-12-31	False	False	False	False	False	60	0	4389.600292162698
wind_speed	Monthly	2015-01-31	2019-12-31	False	False	False	True	False	60	0	6.91464742063492
tomato_price	Monthly	2015-01-31	2019-12-31	True	True	False	False	False	60	0	0.8616472222222222

Fig. 12: CausaLens Application: features information

IV. Forecasting the Tomato price with CausaLens application and comparation of the results

The main objective was to predict the tomato price with the CausaLens application using the methods used before and then compare the results.

To work with the CausaLens web application it was necessary to adapt the data format to the CausaLens format. The data with all the features was uploaded into application. Fig 12 shows the information of each feature using the CausaLens application, this information, like resolution, period, stationarity, mean, missing values, is important to make modelling decision.

The following set up was done:

- The target transformation to return and difference.
- The splitting of the dataset was 40 per cent to training, 20 per cent to validation, 20 per cent to testing, and 20 per cent to holdout.
- The algorithms used to the forecast selected were Prophet, Neural Network and ensemble.

Figure 13 shows the tomato price forecast with Prophet using the CausaLens application where we can see the prediction in blue, and the truth values in red. As it shows the data splitting, the holdout data is used in CausaLens application, as an additional test to assess the predictive power of the selected models.

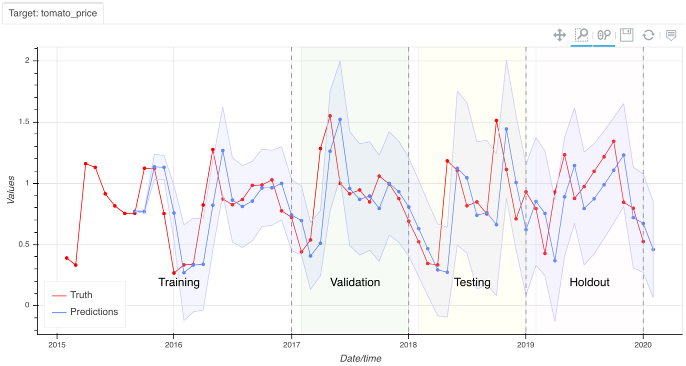
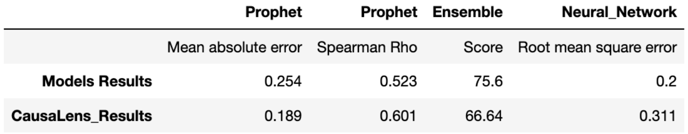


Fig. 13: CausaLens Application: Forecasting the tomato price with Prophet



	Prophet	Prophet	Ensemble	Neural_Network
	Mean absolute error	Spearman Rho	Score	Root mean square error
Models Results	0.254	0.523	75.6	0.2
CausaLens_Results	0.189	0.601	66.64	0.311

Fig. 14: Models results and CausaLens Application results

In figure 14 we can see the models results and a comparison with the CausaLens application results. The metrics used in Prophet were mean absolute error and Spearman (correlation coefficient), the results look similar, as the Prophet charts show. As ensemble techniques, Random Forest was used, as we saw before, and CausaLens application used XGB Regressor, the metric used was cross-validation score. In a neural network it was used LSTM and CausaLens used Multi-layer Perceptron Regressor, and the metric used to compare the result was the root mean square error.

V. Conclusion

With the exploratory data analysis, the conclusion is, that there is clear seasonality in the tomato price variation, as shown in the figure 2 and figure 6, the lower tomato prices are at the beginning and the end of the years, after a fall in prices in the summer, as shown in the figure 8 there is a dip in prices around the summer, it might be because the tomato harvest is in summer, and there is more volume in the market. The higher prices are approximately in April, September, and October. The trend is the increase in the tomato price during the years because the inflation.

In figure 9 we can see the importance of the features. The maximum precipitation intensity, the maximum temperature and humidity are the features that most impact have in the variation of the price of tomato. In CausaLens application oil price was one of the influences in the tomato price, because the tomatoes need to be transported and the oil prices have an influence on shipping cost and finally, influence in the tomato price.

With CausaLens, similar results were obtained, but the time spent was less.

Further work could be done to optimise LSTM model using a bigger volume data.