Predicting the Jaén Olive Oil Price

Diego Hueltes
diego@hueltes.com

Introduction

Jaén is a Spanish region producing 20% of the world olive oil. In this region, it is quite common for families to own a piece of olive tree plantation. The whole family works in the harvest, and that will provide the oil used during the year or a financial bonus. These bonuses have benefits on the regional economy which has high unemployment rates.

The olive oil price depends on demand and production. The production depends mostly on the weather. When the production is low, the price rises. Sometimes, this price raising is not enough to compensate for bad production years.

For this reason, the farmers know in advance what to expect from the harvest by looking at the weather, so they can prepare for it.

This paper will show some prediction models able to do this price prediction using the weather, production and price data.

Background

Jaén contains more than 550.000 olive trees plantation hectares, more than 66 million olive trees. These olive trees produce around 20% of the global olive oil, in the last years a mean of 600.000 tons, more production than the second country, which is Italy.

The economic impact in the province is high. Each Jaén's olive oil campaign requires about 8 million work days for the harvest plus 150.000 work days for the olive oil processing, which has been estimated in a 300 million Euros impact in the province ("Aceite de Oliva de Jaén | Esencia de Olivo - Aceite de Oliva," n.d.).

The olive oil price is negotiated in advance by cooperatives. These cooperatives usually negotiate it using the mean olive market price. This market can suffer from manipulation, so these negotiated prices may not be fair enough to cooperatives, especially the small ones.

A good price indicator, or price prediction, would help those co-ops to negotiate better prices, increasing the profit, which means more benefits for the farmers, which will impact positively in the province.

For building the prediction model, we should understand which are the main factors that affect the price, in order to include data about them in the model. The price depends mainly on the production, so past production data will be useful for the model.
The future production can be forecasted with weather data such as the rain, the temperature or the existence of extreme winds. There are other factors like the crop or the type of plantations. Weather data can be easily included but not crop or plantation types because it is not publicly accessible.

The price is also affected by manipulation, but there is no way to have the certainty of this market manipulation so we cannot include data for it.

There are previous research about the Olive Oil price prediction. In (Pérez-Godoy, Pérez, et al., 2010), they tested several forecast methods based just on the prices, for a one week forecast. The best result they got was a MAPE (Mean Absolute Percentage Error) of 2,257% for CO2RBFN (an evolutionary cooperative-competitive hybrid algorithm for the design of Radial Basis Function Networks). They compared it with methods such as ARIMA which had a 2,659% MAPE.

Besides, in (Pérez-Godoy, Pérez-Recuerda, et al., 2010) the authors improved the previous results to 1,914% MAPE with the same CO2RBFN algorithm (one week forecast). They also introduced a four-weeks forecast with results such us 2,970% for MLP ConjGrad or 3,003% for NU-SVR.

Using these results as a benchmark, in this paper, we will test multiple modern machine learning techniques to forecast the Extra Virgin Olive Oil.

Data extraction and preprocessing

One of the goals of this research is to provide results using free and public available data in order to make the model practically available for any interested producer. The required data is needed price, production, and weather data. We will detail the way in which each data series is acquired as follows.

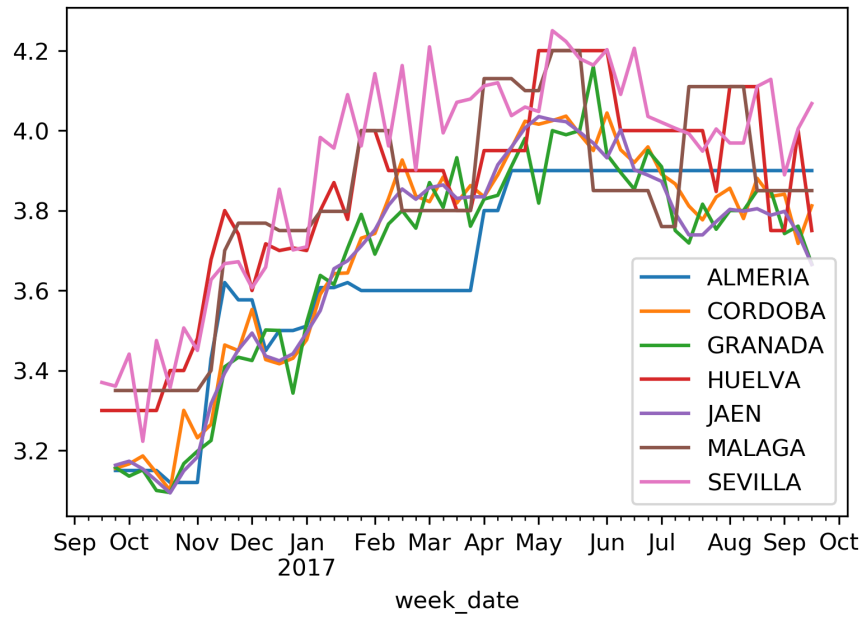
Data extraction, cleaning, and feature engineering

Price data

The price data was extracted in September 2018 from the Junta de Andalucía's (Andalucía regional government, to which the Jaén province belongs) agriculture open data ("Observatorio de Precios y Mercados. Consejería de Agricultura, Pesca y Desarrollo Rural. Junta de Andalucía," n.d.). It was filtered by extra-virgin olive oil, aggregated per province. This is the mean price data at the end of the week.

The acquisition process is simple, the open data tools provide the set of filters and granularity selectors. The data can be visualized or downloaded in excel format.

The data was clean, but with a large number of missing values in provinces that do not sell extra virgin olive oil on a regular basis. Those missing values were instantiated with a very basic strategy, taking directly the last week price data.



The dataset was filtered for keeping the Jaén province data only. After that, the dataset was enriched by included some additional features allowing us to create new and useful variables.

- **Price for the last week**: That is expected to be one of the most important features, this is the last price data we can use in the time series modeling.
- **Price for the previous 2, 4, 6 and 8 weeks**: These features can help the model with past price data and also will be used for calculating means.
- **Mean price in the last 4 and 8 weeks**: This mean price can help to know how the current price is in comparison with the last month / 2 months.
- **Price in three weeks**: That is a future price, it will not be used in the modeling but will be used as the base for the 4 weeks prediction.
- **Price percentages**: All the previous prices but calculated as the percentage change with respect to the last week. The price percentages help the model to check the price difference without looking at external factors like inflation.

Weather data

The weather data was extracted from the Spanish meteorology agency (Agencia Estatal de Meteorología, AEMET) using the available open data REST API ("AEMET OpenData," n.d.)

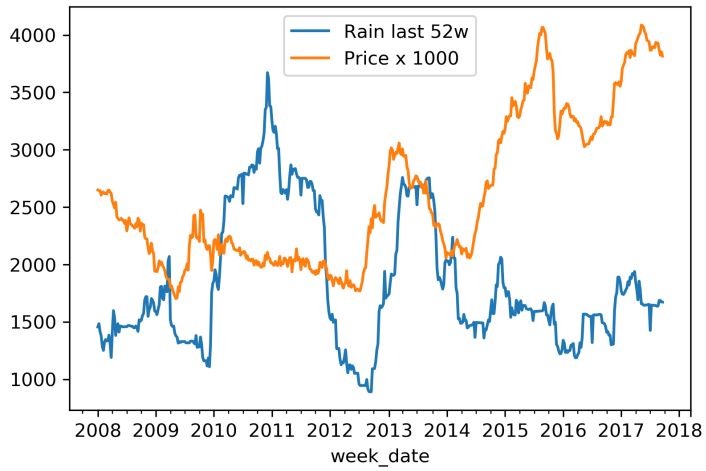
The API provides multiple data from the meteorology stations located close to or in the cities. This data contains, among others, rain amount, wind speed, and temperatures. After aggregating per province, the following variables were extracted on a weekly basis:

- **Temperature**: Max, min and mean temperature
- **Rain**: Sum of rain amount per week
- **Wind**: Max wind speed

Besides, after some computations with the features, five variables were added to the model:

- **Rain sum** for the last 4, 13, 25 and 52 weeks
- **Mean temperature** for the last 4 weeks
- **Percentages**: For all the previous variables, it was calculated the percentage change compared with the mean value for every variable.

The 52 weeks rain sum seems to be correlated with the price, here a comparison between this created variable and the price per ton.



We can observe when the rain is in minimums, the price starts growing fast.

We also took a research study by the Junta de Andalucía *(Efecto de las heladas en el olivar andaluz: identificación y evaluación, análisis térmico y técnicas de teledetección,* 2007) as a base, where the authors state how low temperatures can damage the olive tree. For this reason, a new feature was created for the model, so it can easily account for those cases where there is some damage in the tree or the olive.
Under -5°C the olive is damaged and under -10°C the olive tree is also damaged. The variable **freeze damage** will be an indicator of under zero temperatures, from 0 to 1 indicating the olive/olive tree damage probability.

Production data

There are some available production data that can be extracted from the Spanish department of food control and information ("InformacionMercados (AICA)," n.d.).
While the data is publicly available, the institution does not provide an API, but the data should be extracted doing some AJAX calls.

The data extracted is the produced oil per month, per province. It contains not only the Extra Virgin Olive Oil but also another olive oil like the Virgin Olive Oil.

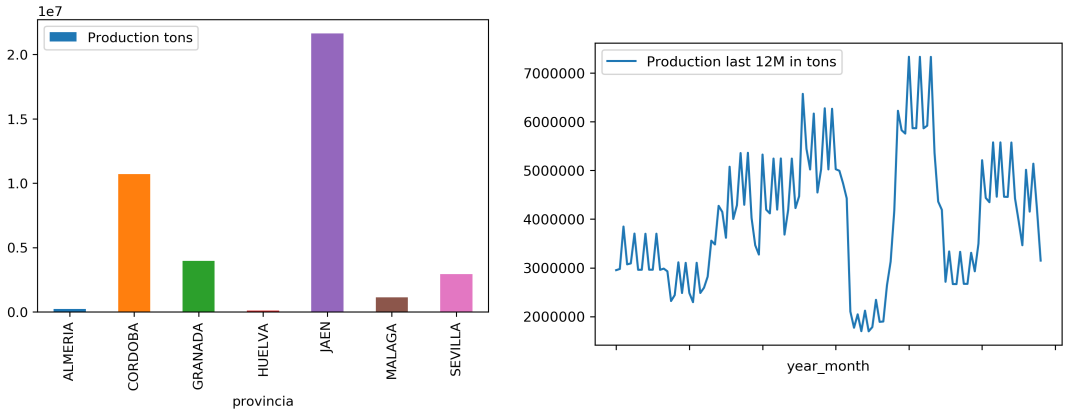
Because of that, this data will not be extremely helpful for the model but it can help to discriminate between good and bad months.
The extracted variable was:

- **Production tons**: Tons of produced olive oil in the last month. (Excluding the current month)

Also, from this variable there were created:

- **Production tons in the last 3, 4, 9 and 12 months**: Sum of productions in these months.

- **Percentages**: For all the production variables, the percentage change compared with the mean value was calculated.

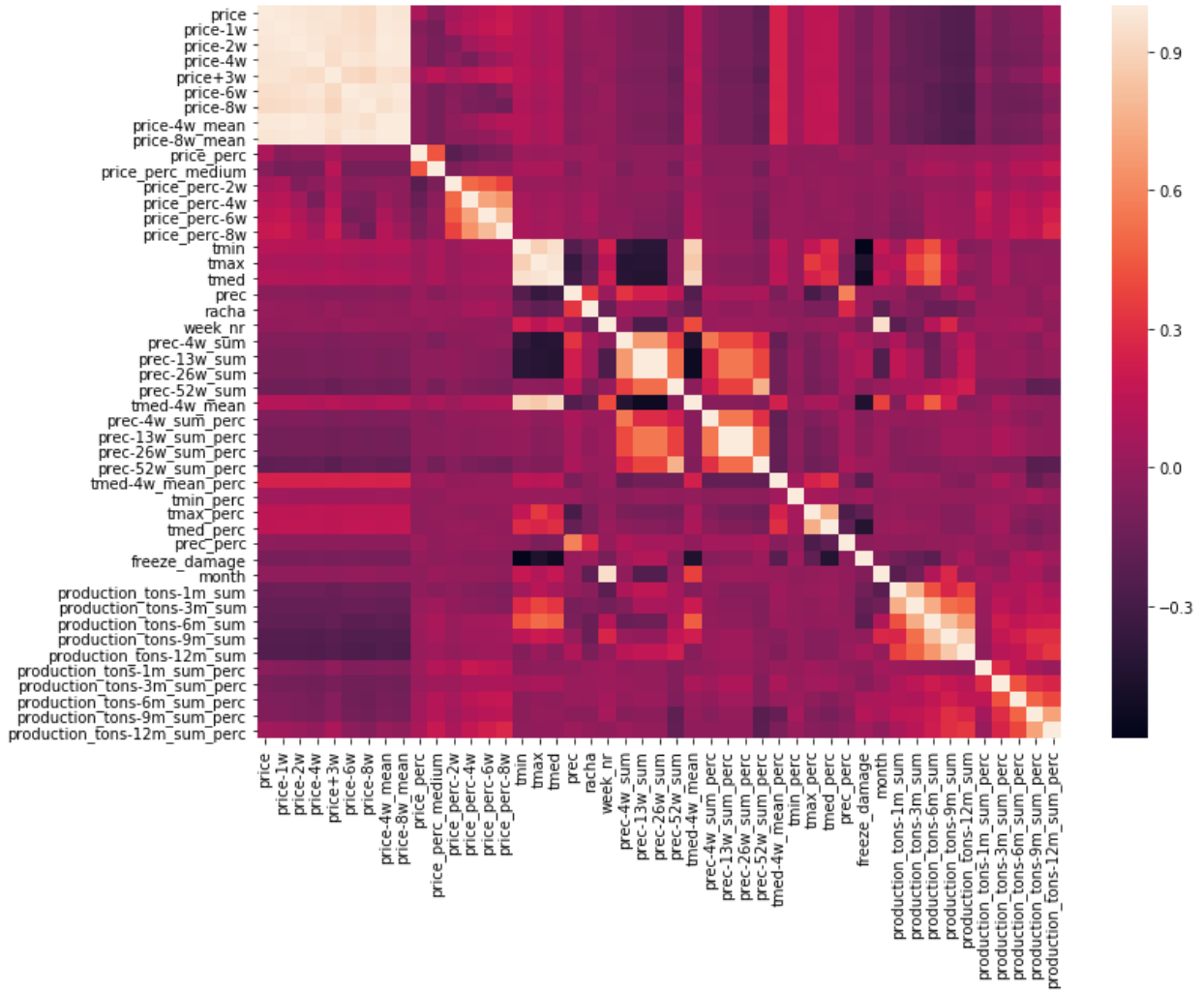


As can be seen, Jaén produced more than the double of the second Andalusian province, i.e. Córdoba. In addition, looking at the aggregation per month, we can see the production movements show a seasonal behavior but also match with the rain aggregation and price.

Feature selection

After joining the extracted data, with the designed features, our feature set is composed of more than 45 features. Most of these features contain redundant information, so a feature selection is required to ensure we consider all the possible information, but with as little repetition as possible.

The following chart is a heatmap based on the Pearson pairwise correlation where the correlation between variables can be observed.



There are two different kinds of variables: percentage based and absolute amounts. In a visual way, we can check the impact of those variables in the price and price change, but firstly there is a need to choose the target variable. It can be either the price or the percentage change.

We have several years of price data, and there are external variables that influence the price, such as inflation. In order to avoid considering those external variables, the price percentage change was selected as target variable instead of the absolute price.

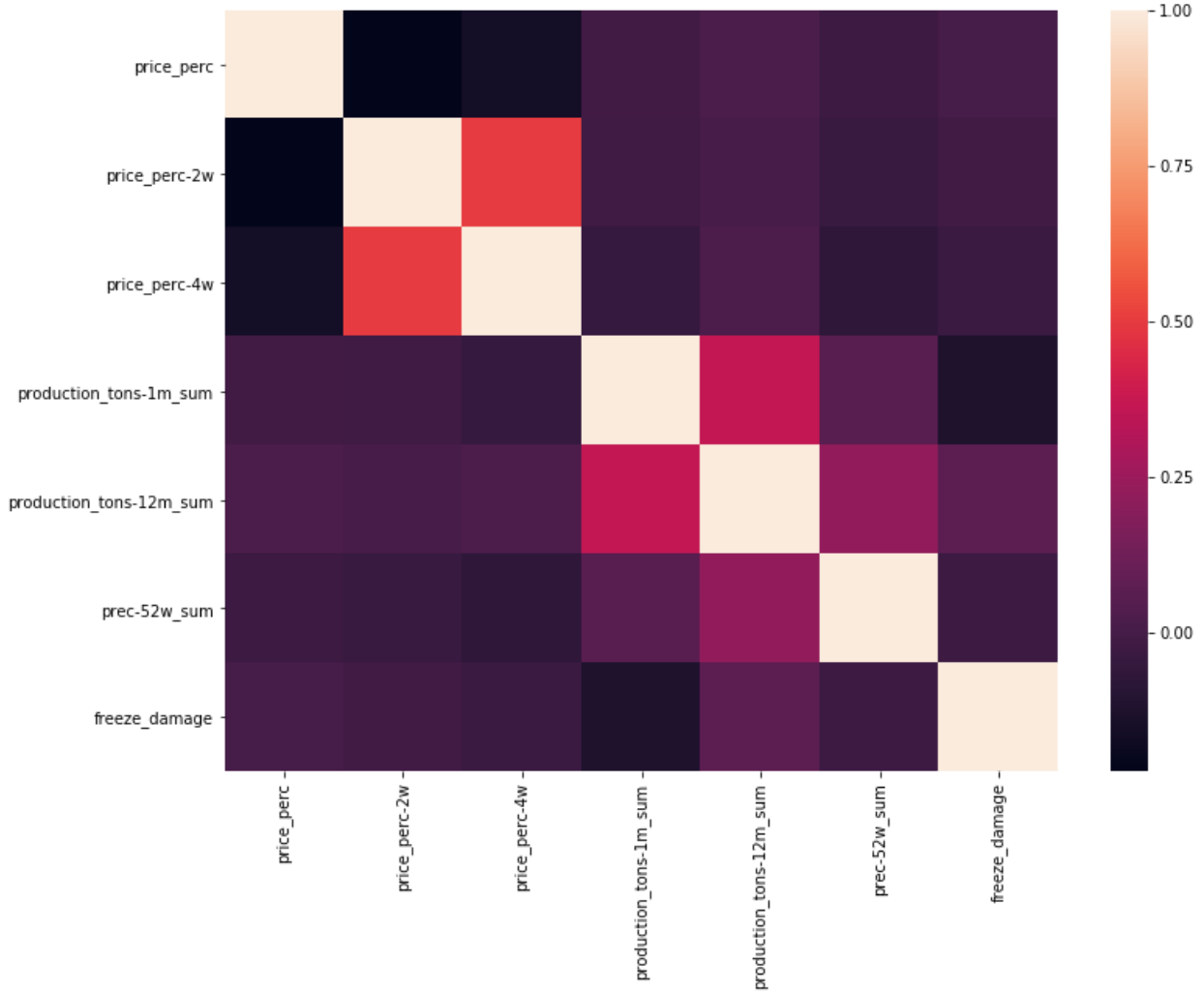
The name of this variable is price_perc and it is calculated as follows:

$$price\_perc = \frac{current\_price - last\_week\_price}{last\_week\_price}$$

After looking at the correlations, and after doing some basic linear regressions for checking the feature importance, the selected variables were the following:

- **price_perc**: Target variable, the price percentage change of the current week compared with the last week.

- **price_perc_2w**: the price percentage change between two weeks ago and the last week
- **price_perc_4w**: the price percentage change between four weeks ago and the last week
- **production_tons-1m_sum**: The last month production change compared with the mean production
- **production_tons-12m_sum**: The last twelve months of production change compared with the mean production
- **prec-52w_sum**: The last 52 weeks rain sum
- **freeze_damage**: Variable that indicates if there is some potential damage due to the low temperatures



Modeling

After developing all the preprocessing tasks, it is time to start with the modeling. To properly validate the model, we will perform a training-test partitioning based on a date split. The whole data set includes data from October 2007 to July 2018. The last year will be considered for model validation, i.e. hold-out data, while the remaining of the data will be taken for training. Hence, the training set contains the data from October 2007 to July 2017 while the test set includes data from July 2017 to July 2018.

Price forecast

The first goal is to forecast the price. That price prediction can be used to trade oil with a favorable price, at the right moment. Olive oil trading is daily, but this research is using open data, which is weekly.

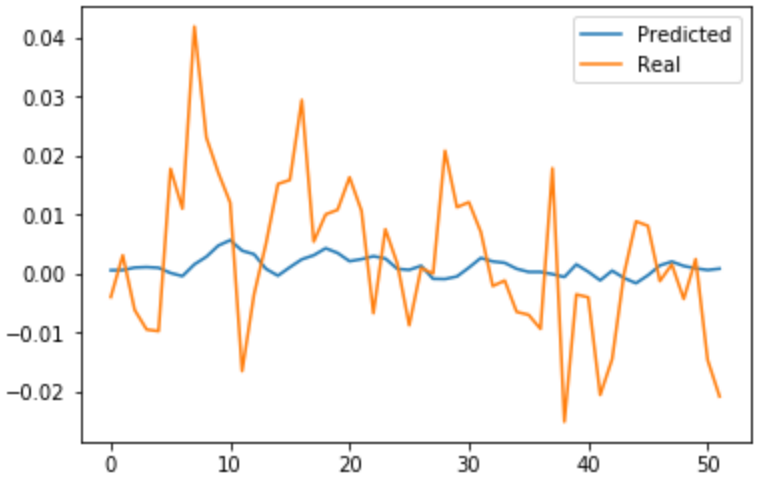
As explained before, the target variable is going to be the price percentage change, but the error variable used for comparing results is the Mean Absolute Percentage Error (MAPE) over the predicted price and not the price percentage. That is a more fair error metric, because the real target for the end user is the price, and it also will allow us to benchmark the models' results compared with the (Pérez-Godoy et al. 2010) and (Pérez-Godoy, Pérez-Recuerda, et al., 2010) research results.

One week price forecast

The first model will try to forecast the price for the next week using the current week data. For that learning task, 3 main model types will be tested: Linear models, Tree-based models, and Deep Learning models.

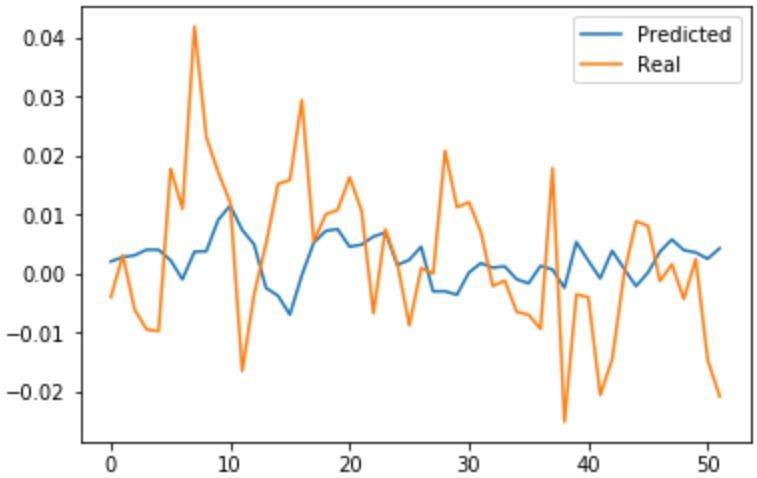
Linear models

The first tested model was the Lasso - Lars model. Lasso-Lars works better with high dimensional data and maybe the feature selection was too selective, but anyways is a model that can provide interesting results.
The Mean Absolute Percentage Error (MAPE) is **1.01%** for the test set, which is better than the results in (Pérez-Godoy et al. 2010) and (Pérez-Godoy, Pérez-Recuerda, et al., 2010), but looking at the data we see that the model is pretty conservative.



The orange line is the real price percentage change and the blue line the predicted percentage. The prediction is almost a flat line, so we can discard this model.

The next tested one is a simple Linear Model, and the MAPE obtained is **1.07%.**

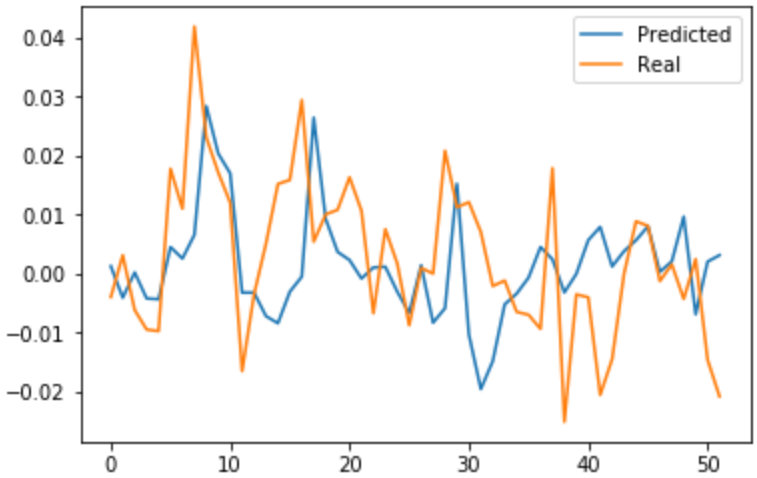


As we can see in the plot, the model is less conservative but still not good enough.

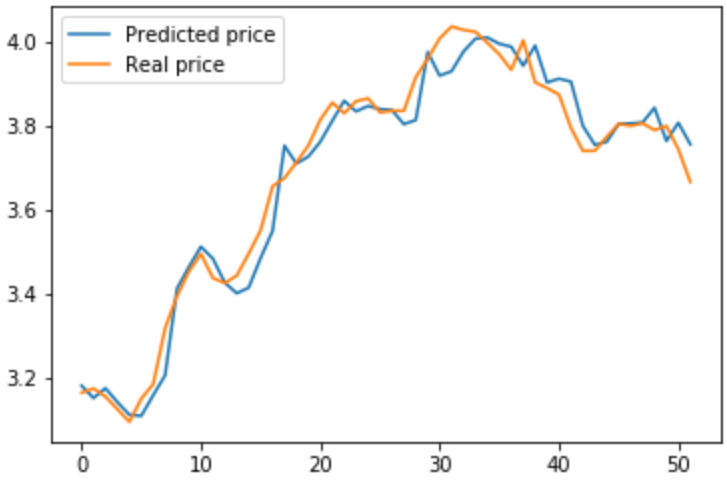
Tree-based models

Those models are better to discover hidden relationships between variables, which will work well for this kind of price prediction.

The first tested one from this model family is the Random Forest Regressor and the MAPE obtained is **1.09%**

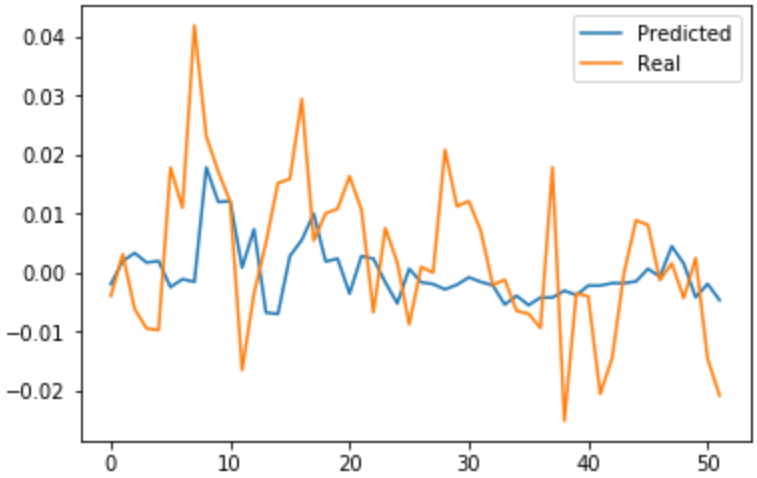


Looking at the results, we see that is not a conservative model, and it has some good predictions especially on big changes, but there are multiple weeks where the prediction is totally different from the real one.
Even if the MAPE is worst than the linear models, the results are interesting.
The next figure shows the predicted price vs the real price in an absolute manner.

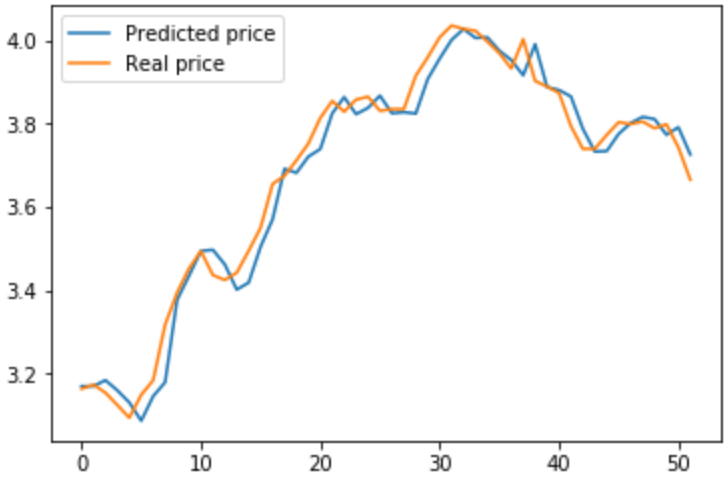


It is interesting because we can see that the predicted price, some times is ahead of the real price. That means that it detected a tendency that is fulfilled in the next weeks, even if it fails in other cases.

The next tested one was the XGBoost Regression model, with a MAPE of **0.99%.**



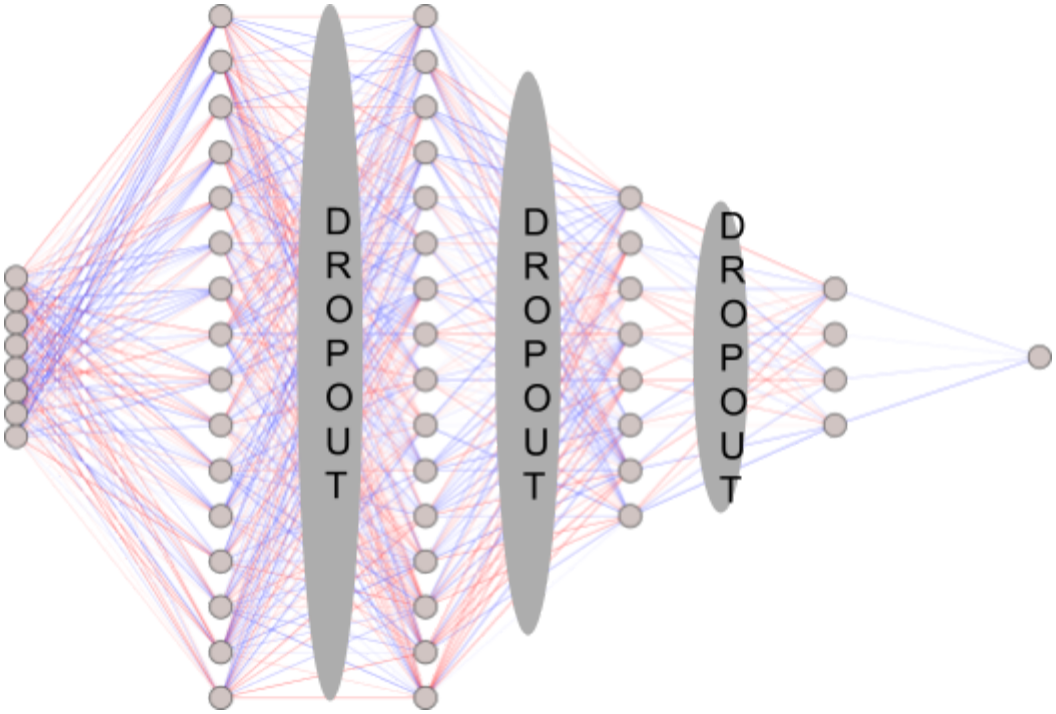
These predictions seem more balanced, predicting some strong price movements, but being conservative on uncertainty. That is why the MAPE obtained is the best one, **0.99%**



Looking at the predicted price figure, it is possible to observe that the price tendency mostly tries to imitate last week's price behavior and that is why the prediction line is in a big number of cases behind the real. This is a valid approach which obtains good results because in some key weeks with big changes, the forecast is the right one.

Deep learning models

For the deep learning models, I used the same data but normalized between 0 and 1 with the MaxMin Scaler.
Finding the right architecture is difficult, but after some tests, I ended with a good one: A neural network with 4 hidden layers and dropout between the four hidden layers, represented in the next figure:



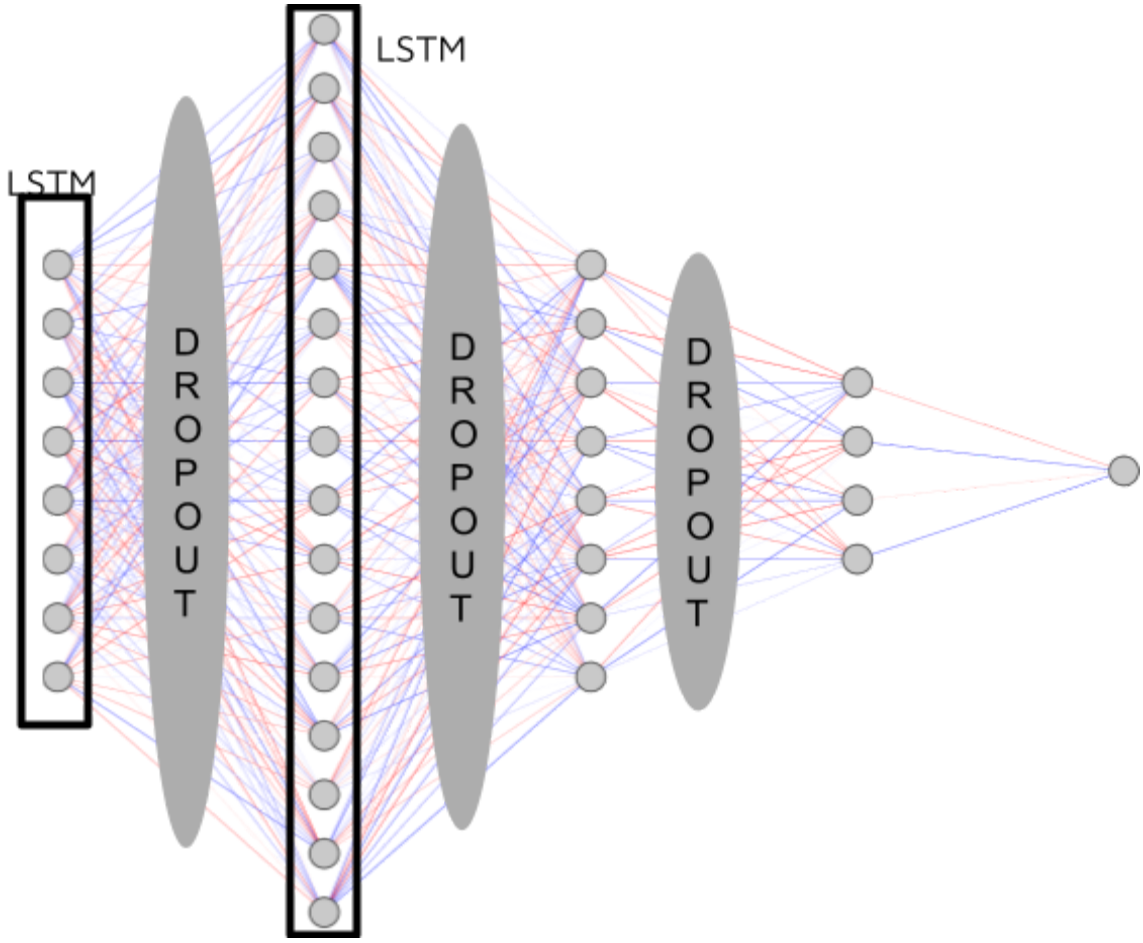
The cells activation function is RELU and the training uses the optimizer adam. The loss function was the mean squared error.

After a 100 epochs training the MAPE obtained for the test set was **0.98%.**

The results are similar to the XGBoost ones, guessing some good changes but conservative in the rest.

After this model, the next one was another neural network but changing the cell type. In this case, Long-Short Term Memory (LSTM) neurons combined with standard neurons.
The LTSM neurons have the ability to "remember" important facts from the past and multiple from recent times. That behavior works really well with data series.

The chosen architecture was a 3 hidden layers model with LSTM cells in the two first layers, standard cells in the next two layers and dropout between the hidden layers as shown in the figure:



After a 100 epoch training, the MAPE in the test set is **1.1%** and looking at the predictions we can observe that are not good enough.

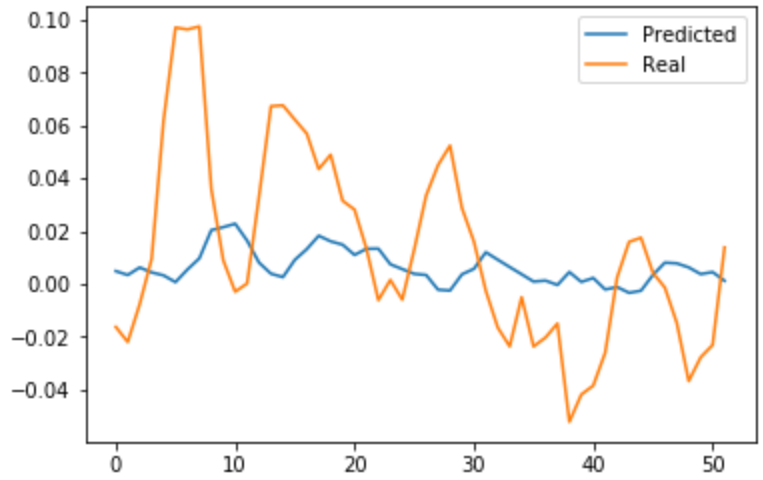
The deep learning models surely would work better with more data points.

Four weeks price forecast

In the (Pérez-Godoy, Pérez-Recuerda, et al., 2010) research, they also try to predict the Olive Oil price for the next 4 weeks, to give a middle term forecast, so the dataset was prepared to get the current week data and the price in 3 weeks as the target variable.
Again, instead of predicting the price, the model tries to predict the price percentage change.

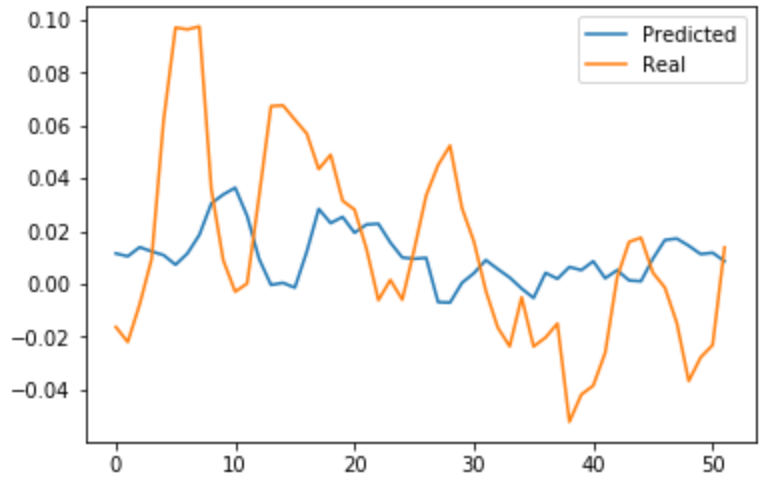
Linear models

For this 4 week forecast, the first tested model is Lasso Lars, with a MAPE of **2.84%** which is slightly better than the error obtained in the best model of (Pérez-Godoy, Pérez-Recuerda, et al., 2010) (**2.9%** MAPE).



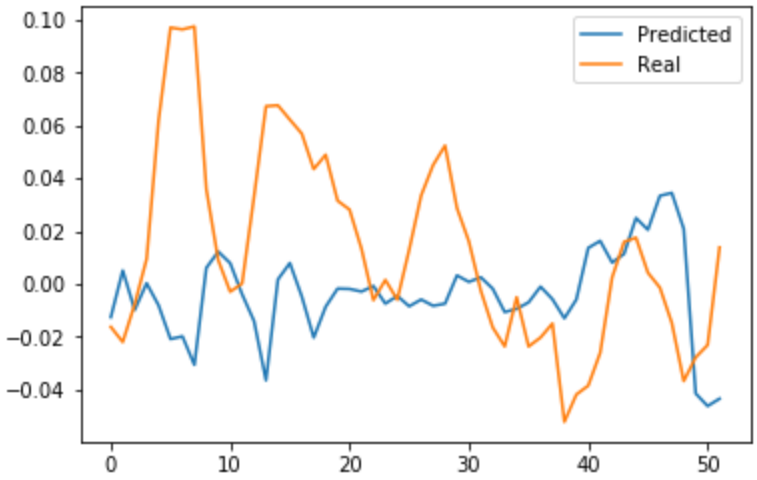
Again, the model is pretty conservative, so unuseful.

The simple linear regression has a **2.96%** MAPE, and looking at the predictions we can see that are almost the same:



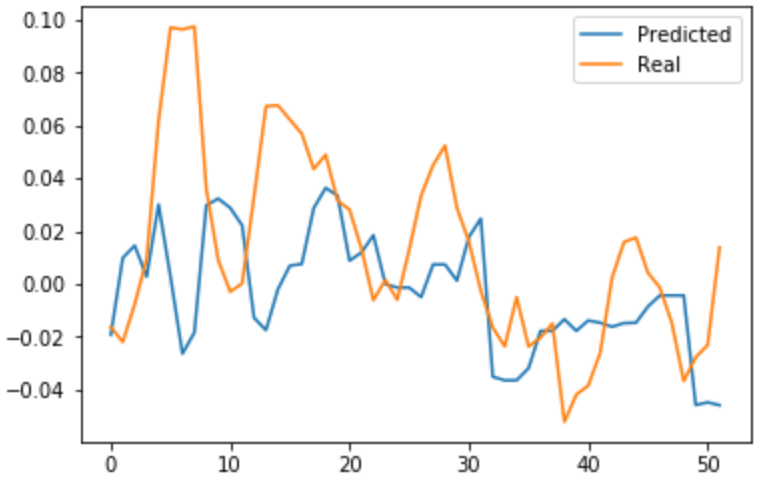
Tree-based models

The first tree-based model for the mid-term forecast is Random Forest Regressor with **3.35%** MAPE



The results are not good at all.

The second tested for this family was the XGBoost, with a **2.79%** MAPE.



That looks better, it detects some of the important variations, and the overall result is good enough.

Deep learning model

The architecture for these 4 weeks prediction is the same as for the 1-week prediction. 4 hidden layers with dropout in the middle cells.
The error is **3.32%** MAPE and in the next figure, we can see what the predictions look like.

Is interesting that the prediction gets the general trend of future movements in a very conservative way.

That is interesting because even if the error obtained in this model is greater than others, like XGBoost regressor, that deep learning model is able to "predict" the direction of the price.

Price direction forecast

The price prediction models obtained better results than previous researches, but one of the most interesting results was the "price" direction sensibility of the last deep learning model.

For that reason, the next step is to try to do exactly that: predict the price direction using classifiers instead of regressors.

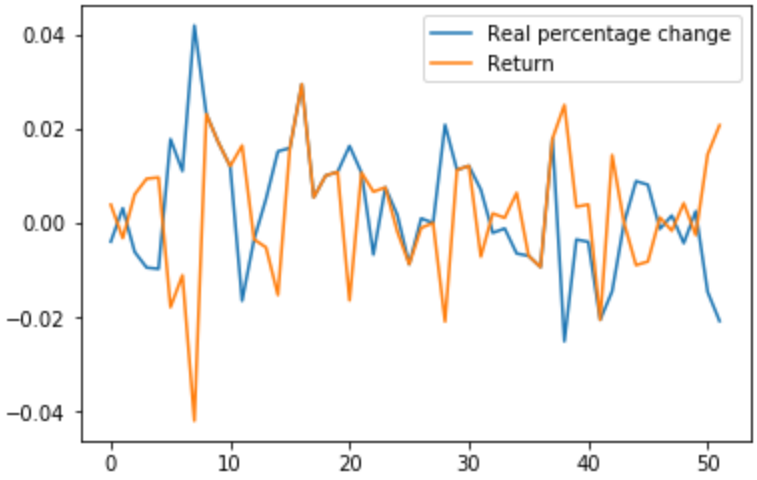
The first step is to create a new target variable, named **increase**, that is 1 if the percentage change is positive and 0 if it is not.

For this price direction prediction, linear models are not expected to perform well based on the regression results, so it has been tested with Tree-based models and Deep Learning.

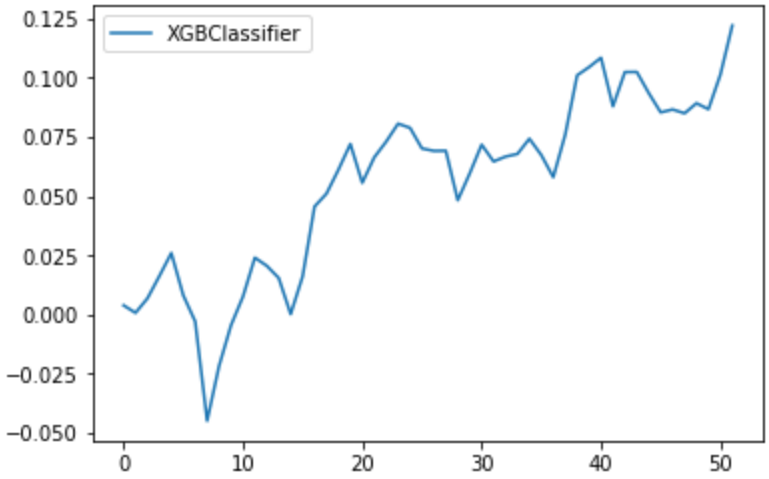
This price direction prediction is for the short-term, so one week.

Tree-based models

The first tested model was XGBoost. That model got an accuracy of **61%**. The following chart shows how the classifier performs. If the model matches the real direction, then the orange line shows the return as an investment gained with that price direction match.

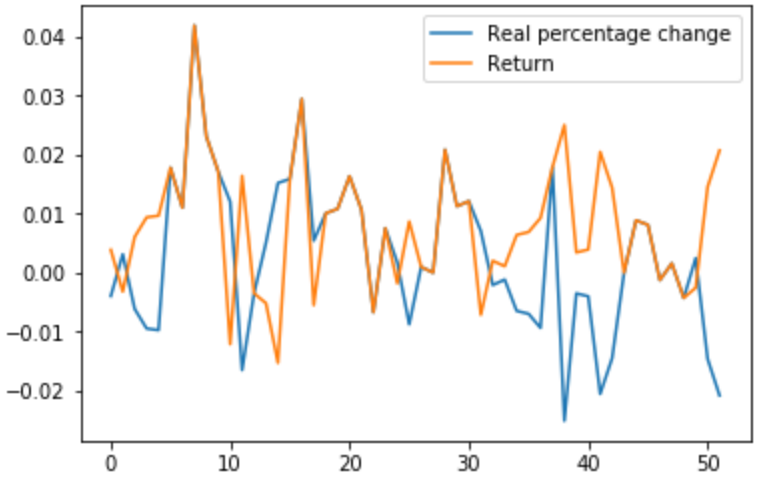


As we can see, it matches in a good number of cases but it failed some important peaks like the weeks 7-8. We can see the cumulative return in the following figure:

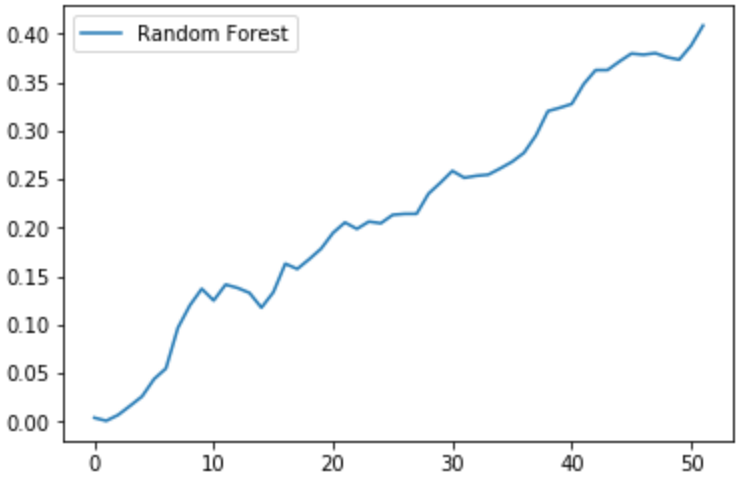


The total is a **12.5%** of annualized return.

After that model, the next one is the Random Forest, getting an accuracy of **77%**.

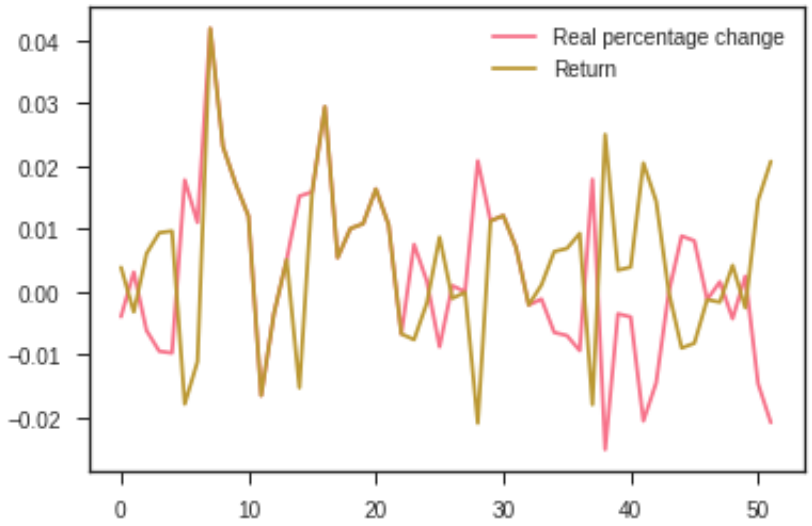


Another good thing about the model is that it matched the bigger price directions and it failed mostly on small ones. Because of that, the cumulative return is a **40%** in that 52 test set weeks, or what is the same a **40% of annualized return** as we can check in the following figure:

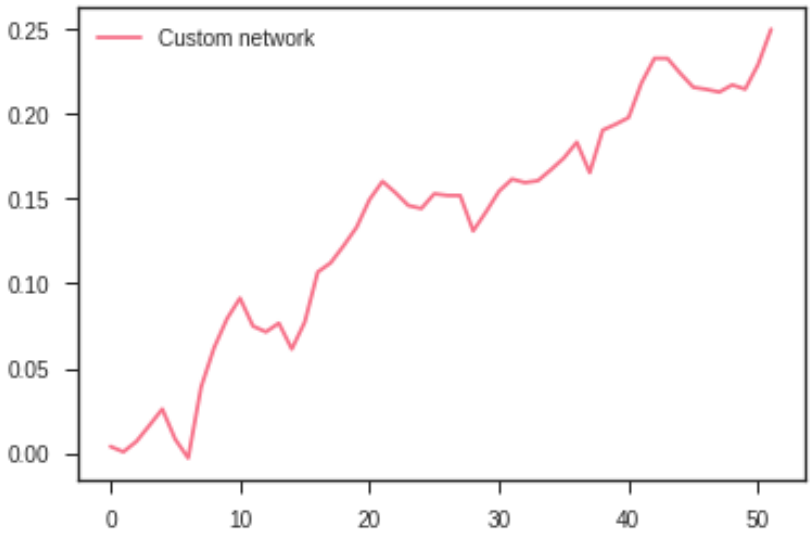


Deep learning models

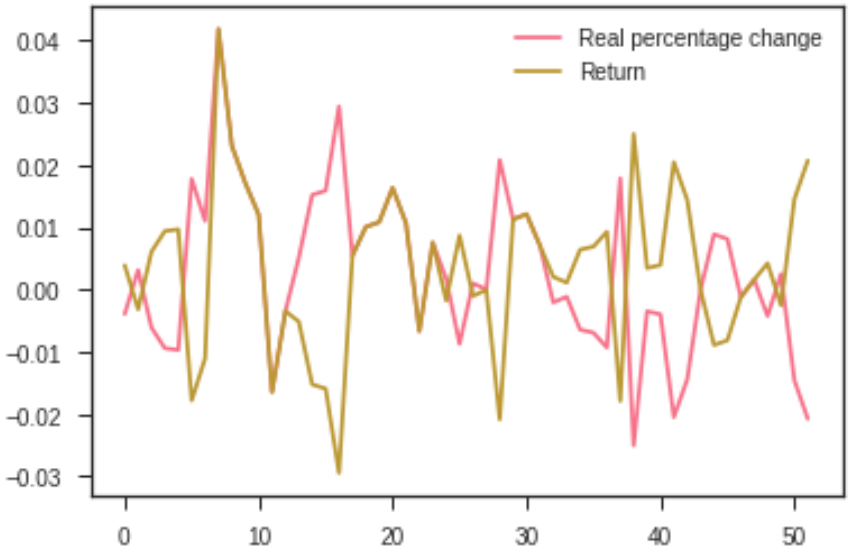
After having these results, the next goal was to try to develop the same classification task using neural networks. In that case, after testing several architectures, the most performant one is a simple neural network with one hidden layer.
The results are **61%** accuracy with an annual return of **25%**

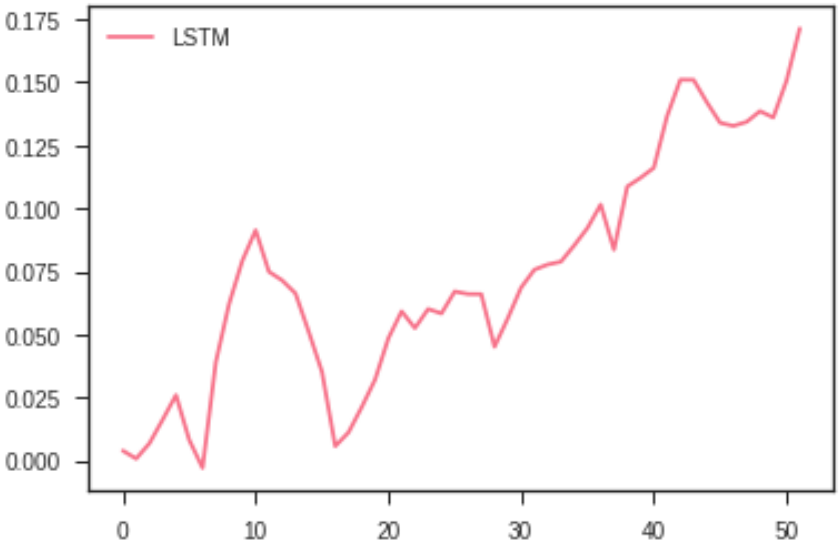


In this chart we can see that is a balanced classifier, with an acceptable annualized return as shown in the next chart:



Then, the next tested one is a Deep Learning network using LSTM cells. The best architecture is 2 hidden layers with LSTM cells and dropout between the hidden layers. The result is a **62%** of accuracy with a **17%** annual return as we can see in the following charts:





Those results are acceptable but not better than the random forest models. The reason, probably, is the number of points to train.

Conclusion

The price prediction based on production, weather and price open data is possible. The price prediction results shown in this research are good, but it can be improved.

There are some unknowns that affect the model such us price manipulation which is difficult to detect and will make the models fail.

This research prediction has been done on a weekly basis but it would be better to do it on a daily basis because the market trading is daily. For improving that model, there a few kinds of variables that could be used such as:

- International market data
- Harvest types of data
- Plantation types of data
- Olive tree diseases data

Another interesting analytic that would help the model would be a price manipulation indicator. That could be achieved by using anomalies detection models, but it requires intensive research.

Anyways, even if the price prediction is not as good as it could be, especially in the mid-term forecast, the price direction prediction is good enough to help investors in their decisions. It has been demonstrated with a real annualized return in the test set (last year of data) of **40%**.

References

Aceite de oliva de Jaén | Esencia de Olivo - Aceite de Oliva. (n.d.). Retrieved February 21, 2019, from http://www.esenciadeolivo.es/aceite-de-oliva/aceite-de-jaen/

AEMET OpenData. (n.d.). Retrieved May 3, 2019, from https://opendata.aemet.es/

Consejería de Innovación, Ciencia y Empresa. Efecto de las heladas en el olivar andaluz: identificación y evaluación, análisis térmico y
técnicas de teledetección. (2007).

InformacionMercados (AICA). (n.d.). Retrieved May 3, 2019, from https://servicio.mapama.gob.es/InformacionMercado_Aica/index.jsp?aplica=IMA
Observatorio de Precios y Mercados. Consejería de Agricultura, Pesca y Desarrollo Rural.

Junta de Andalucía. (n.d.). Retrieved February 21, 2019, from http://www.juntadeandalucia.es/agriculturaypesca/observatorio/servlet/FrontController?action=Static&subsector=33&producto=33000&url=generadorInformesOR.jsp

Pérez-Godoy, M. D., Pérez, P., Rivera, A. J., del Jesus, M. J., Carmona, C. J., Frías, M. P., & Parras, M. (2010). CO2RBFN for short-term forecasting of the extra virgin olive oil price in the Spanish market. International Journal of Hybrid Intelligent Systems, 7(1), 75–87.

Pérez-Godoy, M. D., Pérez-Recuerda, P., Frías, M. P., Rivera, A. J., Carmona, C. J., & Parras, M. (2010). CO2RBFN for Short and Medium Term Forecasting of the Extra-Virgin Olive Oil Price. In Studies in Computational Intelligence (pp. 113–125).