Corporación Favorita Insight Extraction & Time Series Analysis

Diego Losada
3rd of January 2021

CONTEXT

Corporación Favorita is a conglomerate that operates in many industries. The oldest of their lines of businesses is a chain of supermarkets and convenience stores. They are present in many Latin American countries with Ecuador being the place where the headquarters are located and where they conduct most of their activity.

Disclaimer: the content in this analysis is for educational purposes, use at your own discretion.

TABLE OF CONTENTS

01 The Goal

02 The Data

03 The Insights

04 The Forecast

05 The Conclusions

01

THE GOAL

THE GOAL

Extract insights from the sales of a particular store of Corporación Favorita

Create a time series model to forecast sales 15 days into the future

02

THE DATA

Search Foods
Search:
Search
Source: All
Category: All
Language: All
Description
Source
Rank
Cancel

Original Datasets

Train

Rows = 125,497,040
Columns = 6

Transactions

Rows = 83,488
Columns = 3

Oil

Rows = 1,218
Columns = 2

Holidays

Rows = 350
Columns = 6

Items

Rows = 4100
Columns = 4

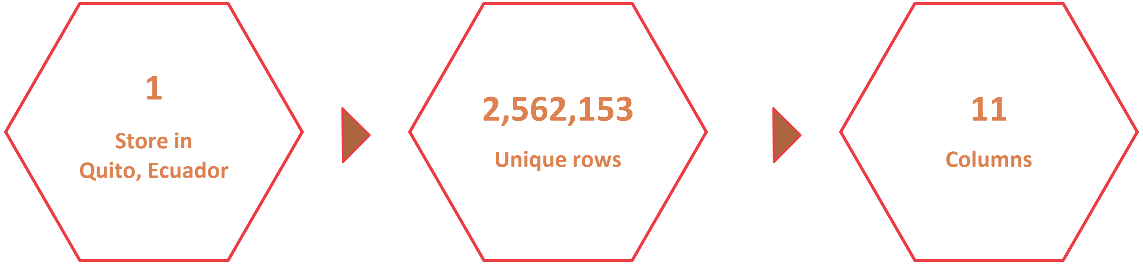
Stores

Rows = 54
Columns = 5

Data Wrangling

I decided to analyse the sales of 1 of the stores instead of the 54 in the dataset, this makes our analysis much more manageable for 1 personal computer. Having said this, the same process can be applied for the other 53 stores.

Final Dataset



1

Store in Quito, Ecuador

2,562,153

Unique rows

11

Columns

03

THE INSIGHTS

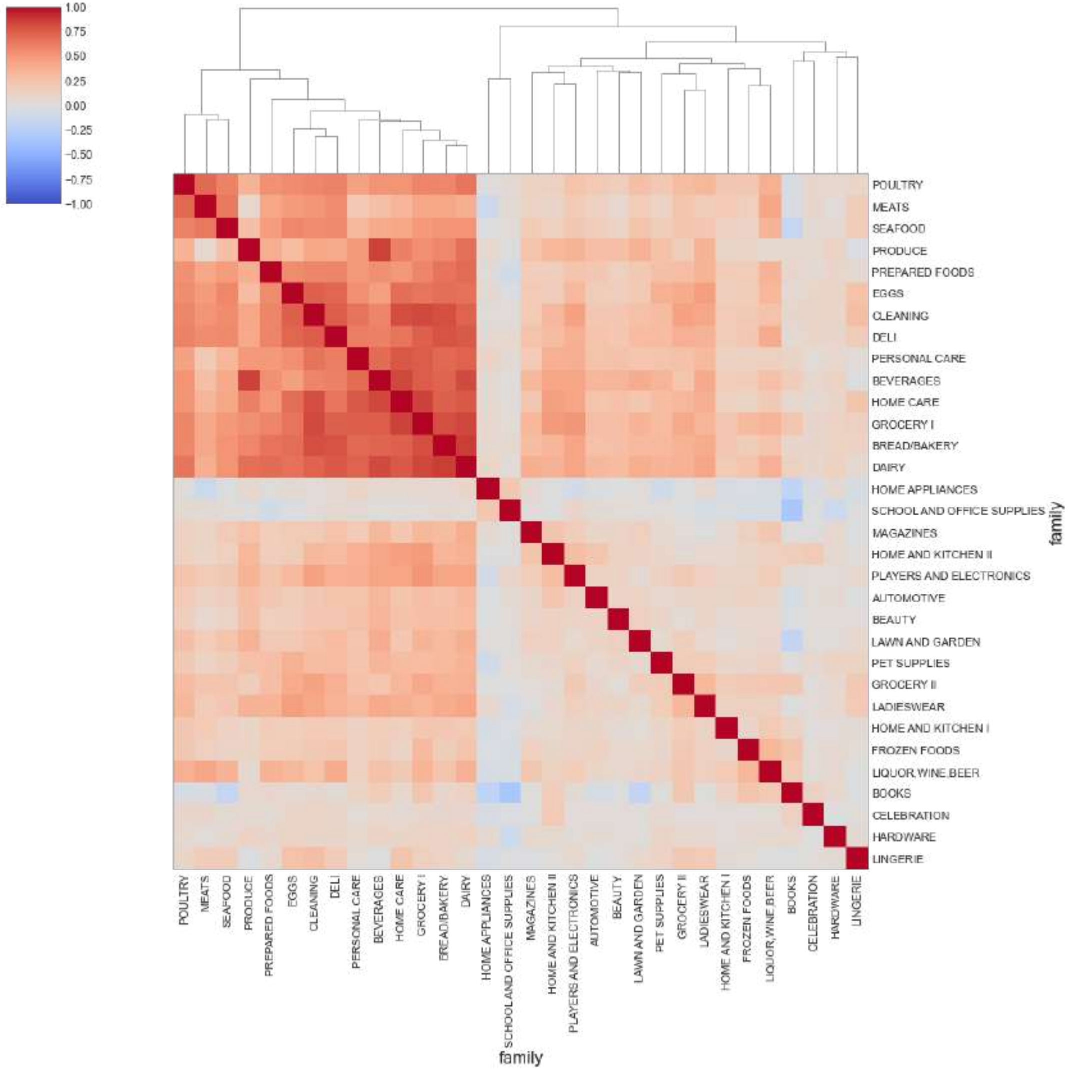
Are sales of families of products correlated between them?

Clustermap of unit sales by family of products

The clustermap shows there is a strong correlation between the unit sales of the first 14 product families of our dataset. As one might expect, these products are mostly perishable and food related products.

However, this information can still be useful. Some supermarkets use the strategy of putting far apart products that sell well together. This, in turn, makes clients walk more and thus, seeing other products they might not have considered buying on the first hand.

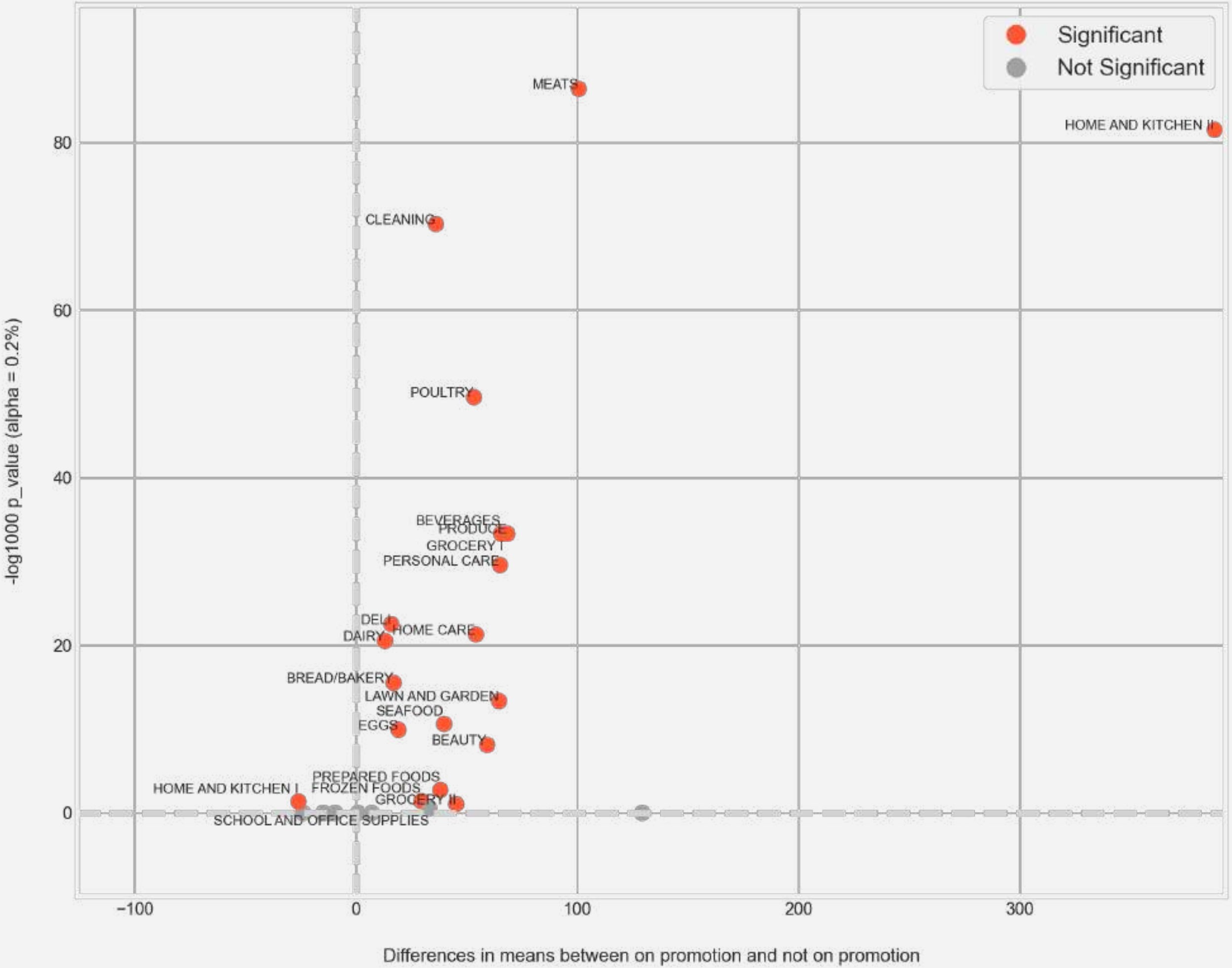
Conversely, if you want to make it easier for clients to find the products that sell well together, this clustermap can also be useful for that.



What family of products are most influenced by promotions?

Families of products that show a significant statistical difference in the number of units sold with and without promotion

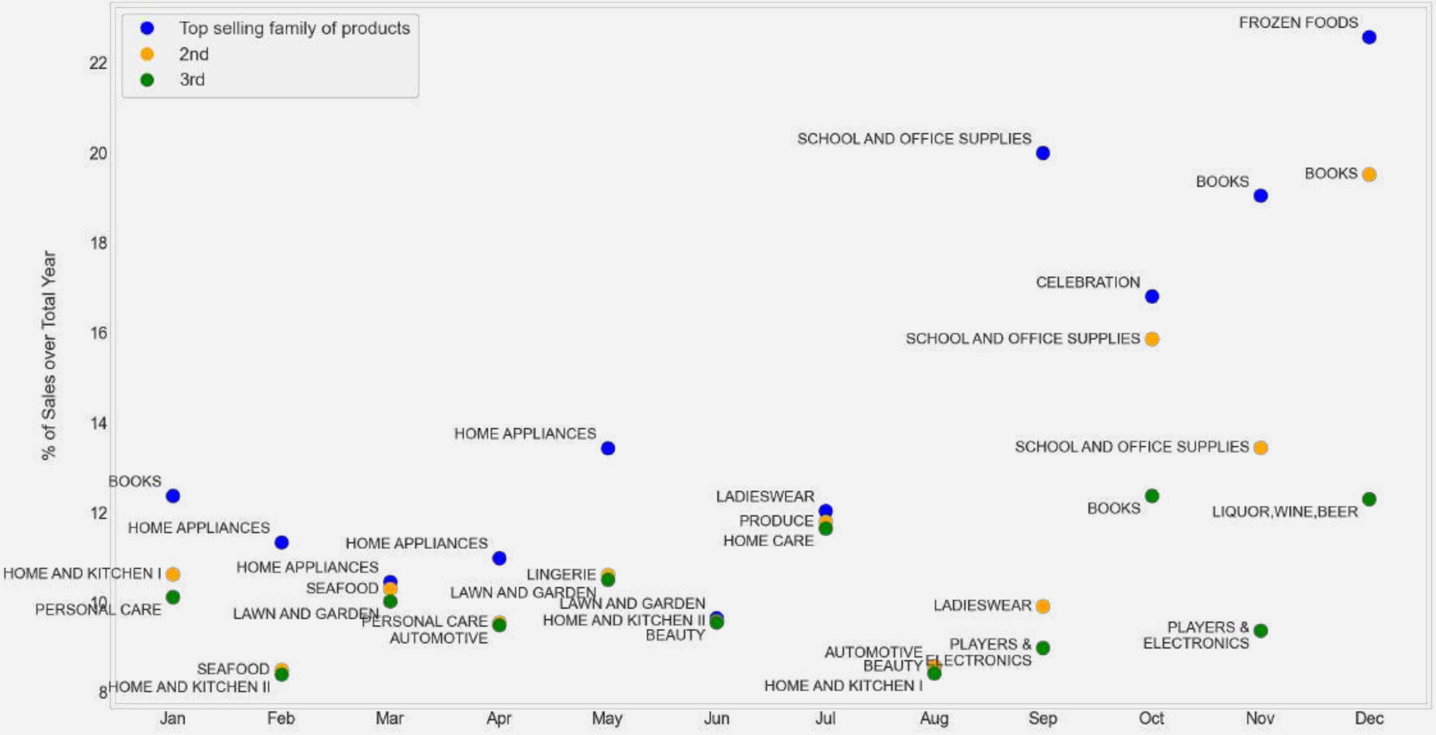
This volcano plot shows that some families of products are very responsive to promotions. Product in the categories of “home and kitchen II” and “meats” sell much more on average than when they are not on promotion. In other families of products like ”pet supplies” and “celebration” there is no effect with promotions. Interestingly, products inside the “Home and Kitchen I” category seem to have the opposite effect, when products are in promotion they sell less.



What family of products are sold more seasonably?

Top 3 family of products that sell the most each month a percentage of whole year sales

This plot is another way of understanding seasonality. We see that home appliances sell well from February to April, while School and Office Supplies sell a lot from September, to November. Interestingly, almost 25% of the frozen food sales are made in December. This type of information can be useful to decide when to make promotions, when to advertise certain products, and when not to do it.



04

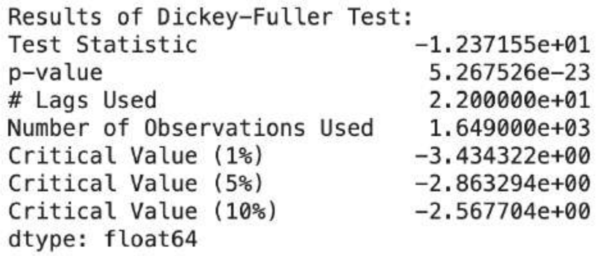
THE FORECAST

Stationarity

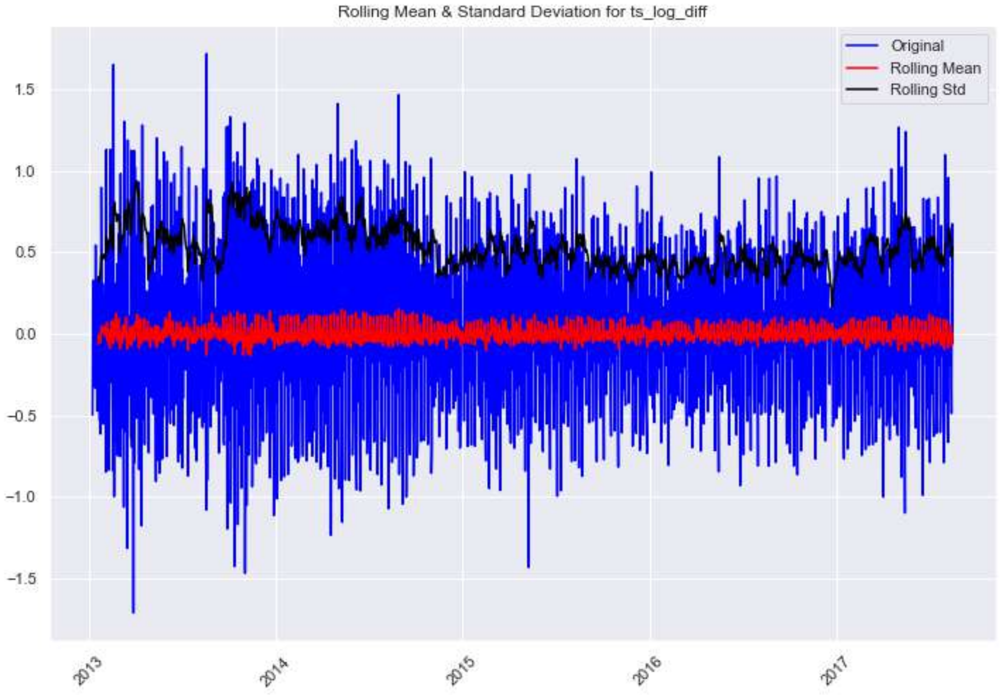
To forecast time series data we need to make our data stationary since it is an assumption of many of the time series models. For our data to be stationary, it should have constant mean, constant variance, and the covariance of the i th term and the (i + m) th term should not be a function of time. Of our product families only Poultry and Books were not stationary, also, our oil data (which we will use as an additional variable in our models) was not stationary.

Example

On the image we see the transformations we needed for out “poultry” time series data to become stationary. We did this in two steps 1) taking the log values and 2) doing the first-difference ((t+1) – t). The Dickey-Fuller test tells us if we can reject the null hypothesis that the series in non-stationary.



```
Results of Dickey-Fuller Test:
Test Statistic                 -1.237155e+01
p-value                         5.267526e-23
# Lags Used                     2.200000e+01
Number of Observations Used     1.649000e+03
Critical Value (1%)            -3.434322e+00
Critical Value (5%)            -2.863294e+00
Critical Value (10%)           -2.567704e+00
dtype: float64
```



Models

We used three models:

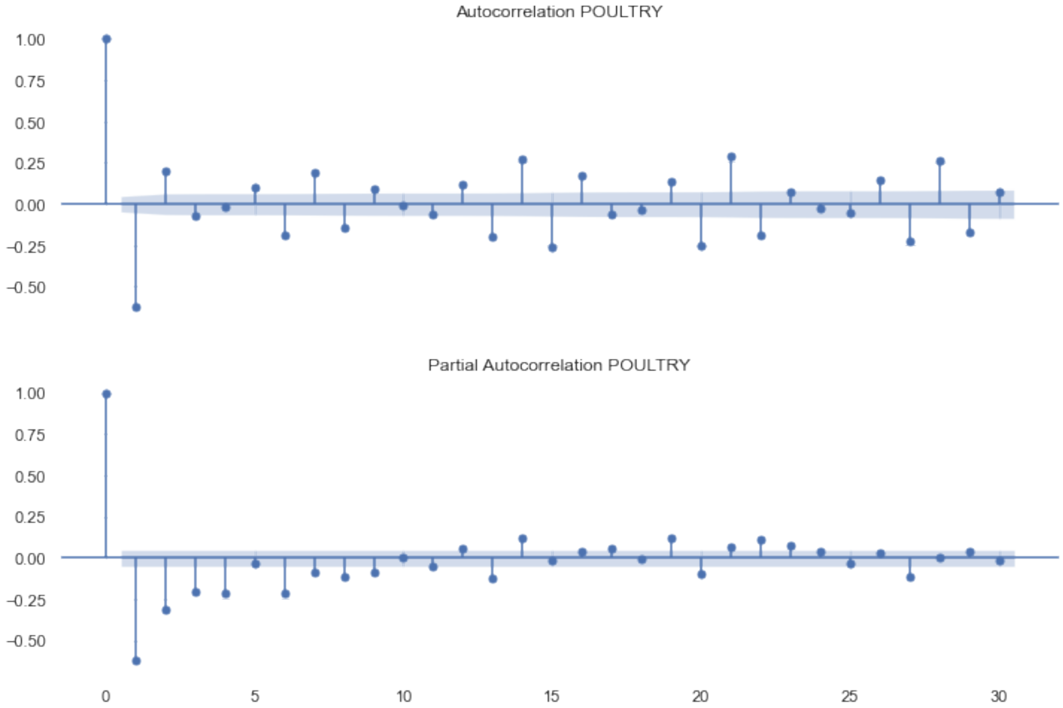
- ARIMA (Autoregressive Integrated Moving Average) Model
- Facebook´s Prophet Model
- LSTM (Long Short Term Memory) Model

Predictions

We predicted the sales of August 1st to August 15th. Unfortunately for the families of products “Home Appliances” and “School and Office Supplies” we did not have data to test our models on (they were thus dropped).

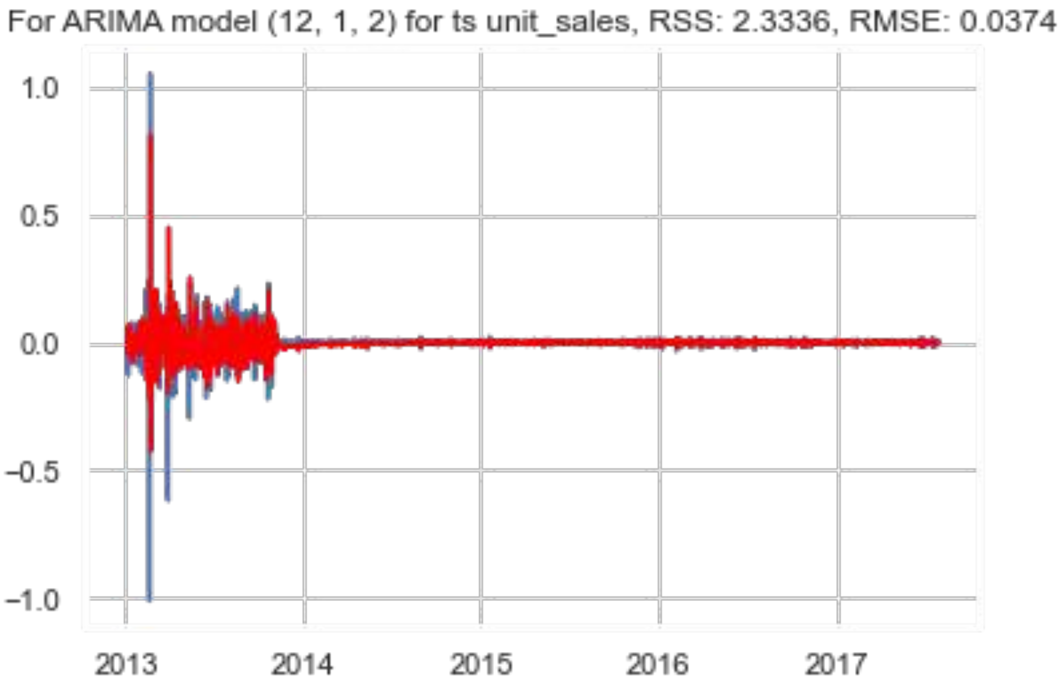
ARIMA Model Example

To do the ARIMA model, we need to first calculate the ACF (Autocorrelation Function) and the PACF (Partial Autocorrelation Function). These two functions will provide us relevant information to then create our model. Namely, we will get the autoregressive parameter (p) and the moving average parameter (q), more info here.



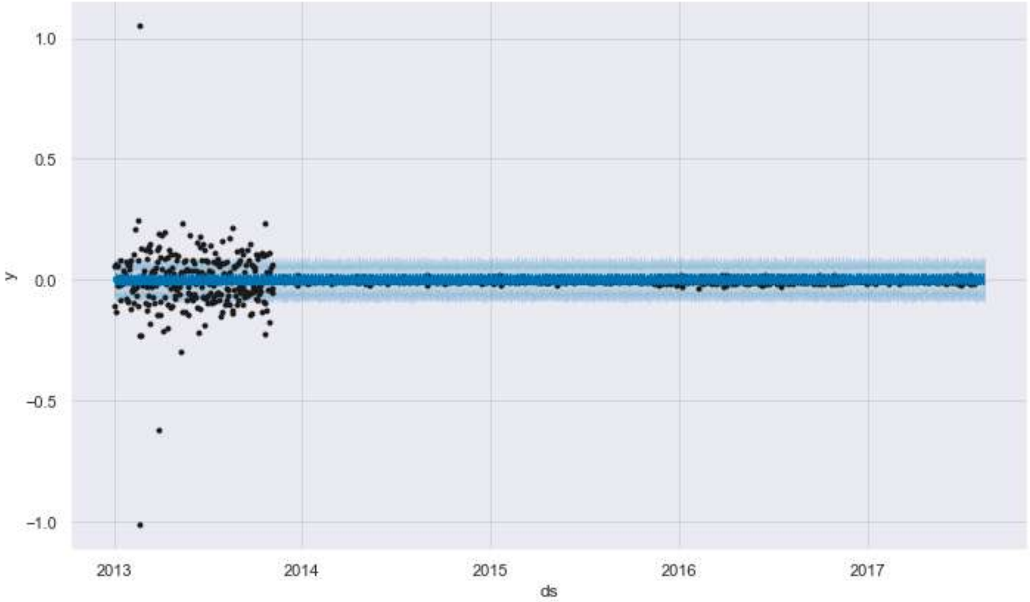
ARIMA Model Continuation

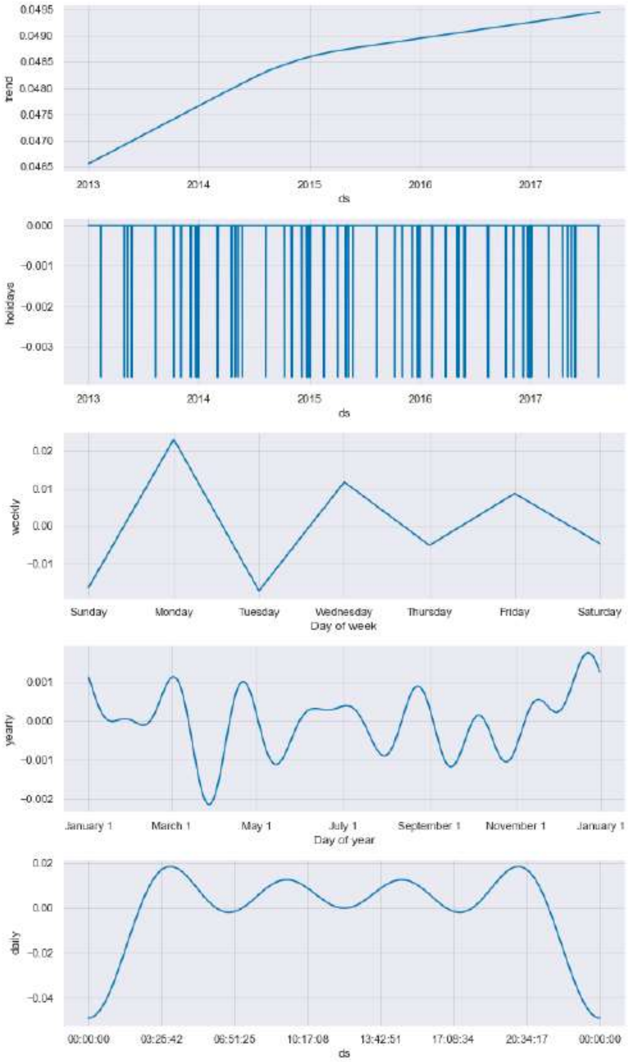
We used the arima model with the p and q values we got from our ACF and PACF. Here is the result for the poultry sales.



FB's Prophet Model Example

Facebook´s Prophet is a great solution for a quick and easy forecast. We were able to include information about holidays in our analysis and, as we will see in the final result slide, this model worked out the best for most of our families of products.



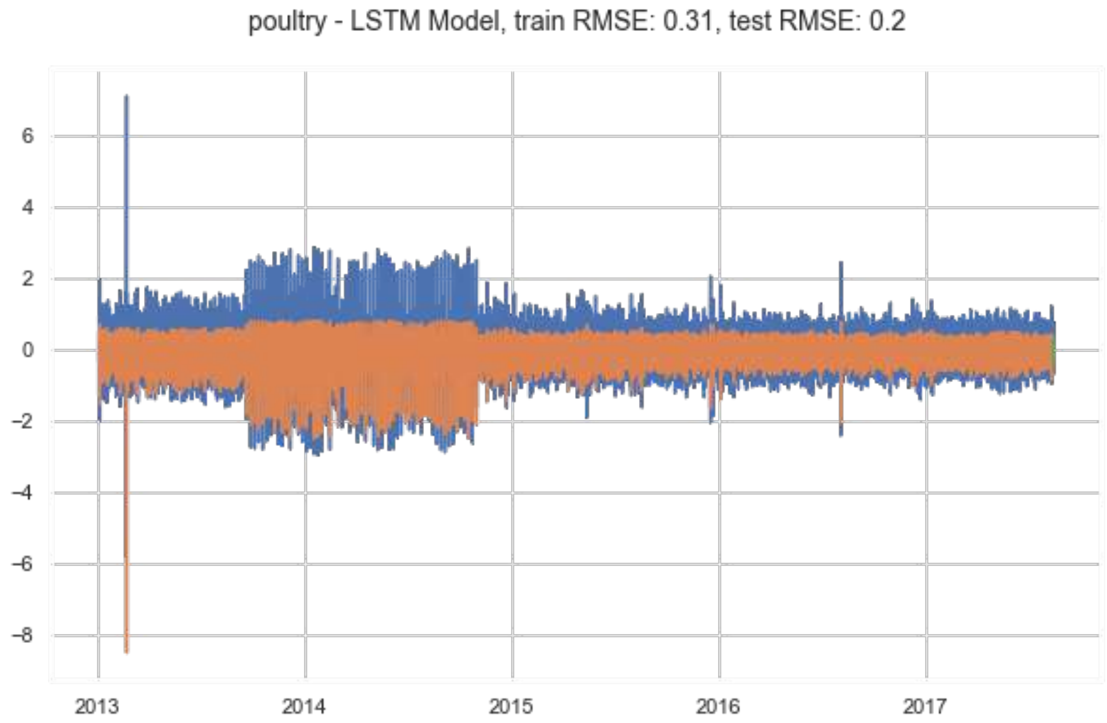


LSTM Model Example

The LSTM we build was a simple one, we added 4 units and we the tanh activation function and the sigmoid recurrent_activation function.

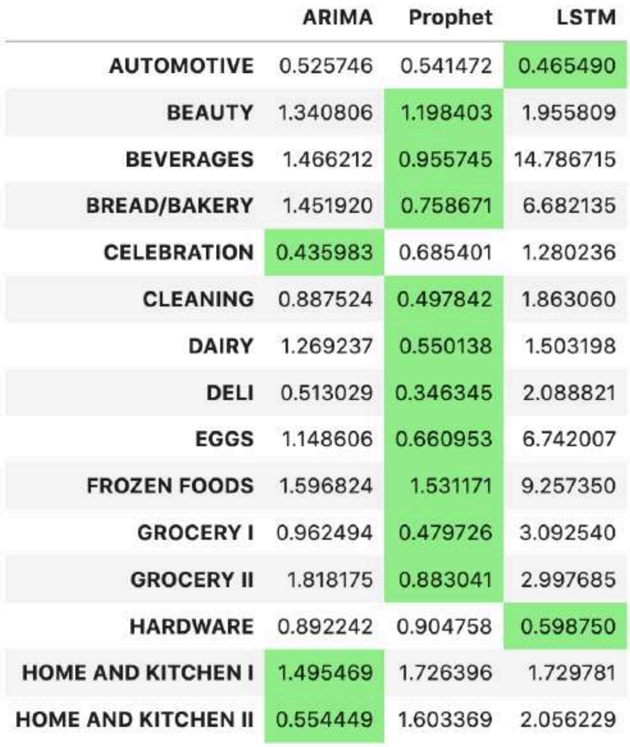
The results of the LSTM show a lot of variability. For a few product families it worked out well, whilst for others they were far from optimal.

Our LSTM model was very simple, adding more layers and tweaking the other parameters will surely provide better results.

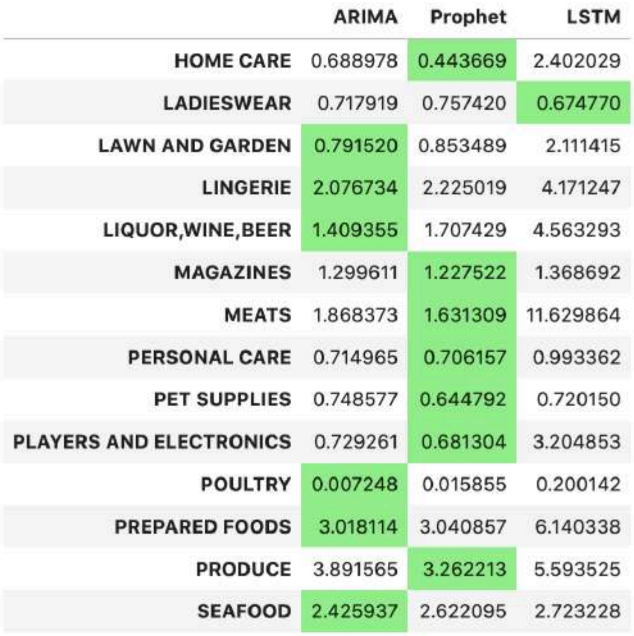

poultry - LSTM Model, train RMSE: 0.31, test RMSE: 0.2

The Final Results

As we mentioned before, the results are quite clear, FB’s Prophet provided the best results in most of the categories. The results we are seeing here are root mean squared errors (RMSE) for our test dataset. In color green are the models that performed the best for each family of products.



	ARIMA	Prophet	LSTM
AUTOMOTIVE	0.525746	0.541472	0.465490
BEAUTY	1.340806	1.198403	1.955809
BEVERAGES	1.466212	0.955745	14.786715
BREAD/BAKERY	1.451920	0.758671	6.682135
CELEBRATION	0.435983	0.685401	1.280236
CLEANING	0.887524	0.497842	1.863060
DAIRY	1.269237	0.550138	1.503198
DELI	0.513029	0.346345	2.088821
EGGS	1.148606	0.660953	6.742007
FROZEN FOODS	1.596824	1.531171	9.257350
GROCERY I	0.962494	0.479726	3.092540
GROCERY II	1.818175	0.883041	2.997685
HARDWARE	0.892242	0.904758	0.598750
HOME AND KITCHEN I	1.495469	1.726396	1.729781
HOME AND KITCHEN II	0.554449	1.603369	2.056229



	ARIMA	Prophet	LSTM
HOME CARE	0.688978	0.443669	2.402029
LADIESWEAR	0.717919	0.757420	0.674770
LAWN AND GARDEN	0.791520	0.853489	2.111415
LINGERIE	2.076734	2.225019	4.171247
LIQUOR,WINE,BEER	1.409355	1.707429	4.563293
MAGAZINES	1.299611	1.227522	1.368692
MEATS	1.868373	1.631309	11.629864
PERSONAL CARE	0.714965	0.706157	0.993362
PET SUPPLIES	0.748577	0.644792	0.720150
PLAYERS AND ELECTRONICS	0.729261	0.681304	3.204853
POULTRY	0.007248	0.015855	0.200142
PREPARED FOODS	3.018114	3.040857	6.140338
PRODUCE	3.891565	3.262213	5.593525
SEAFOOD	2.425937	2.622095	2.723228

05

THE CONCLUSIONS

Summary of Conclusions:

- The timeseries data from Corporacion Favorita offered us good insights:
  - We saw the families of products that sell better together and which don’t. This can become a good starting point to try different placement strategies in each store.
  - We also saw which families of products respond better to promotions and for which there is no apparent difference in unit sales when they are in promotion versus when they are not.
  - Finally, we saw how some families of products have a clear seasonality. Corporación Favorita could take this information to do targeted advertisements on certain times of the year or use the insights to support or rethink their supply strategy.
- In relation to the forecast, we saw that our FB Prophet models, followed by the ARIMA models are good at forecasting sales with a 15-day outlook. These models could also be used for supply strategies.

Improvements and Next Steps:

- One of the first improvements would be to use all the exisiting data. We used the data from 1 store but managing it together might shed some interesting insights for the whole company. One way to do this would be to use external servers with more computing power, leverage the power of PySpark or Dask, and be more efficient on my code.

- There are more insights to be extracted from the dataset. I focused on analysing the promotions and product families but there are also ”class” and “individual products”. These, however, have no description attached to it, so product 590 we know is from the ”poultry” family but nothing else. Getting this information would be ideal.

- Test other forecasting models. We tested three models (ARIMA, FB Prophet, and LSTM) but we have many other tools at our disposal. A first possibility would be to try a multivariate time series forecasting. Only with FB´s Prophet model we added the variable “holidays” but we can also consider using the daily oil prices, the day of the week, etc.

More info

CREDITS

- Content by Diego Losada
- Presentation template by Slidesgo
- Icons by Flaticon and Pixel Perfect
- Infographics by Freepik
- Images created by Freepik - Freepik
- Author introduction slide photo created by Freepik
- Text & Image slide photo created by Freepik.com