A. MODEL SUMMARY

A1. Background

- Competition Name: M5 Forecasting - Accuracy
- Team Name: Alan Lahoud
- Private Leaderboard Score: 0.53604
- Private Leaderboard Place: 5th

- Academic and professional background:

  2013 – 2018: Electrical Engineer – University of Sao Paulo, Brazil
  2017 – 2018: 1-year Master Exchange studies – Lund University, Sweden
  Jun/2018 – Nov/2019 - Data Scientist at Big Data Brasil
  Feb/2020 – Now – Data Analyst at BTG Pactual Bank

- Experiences in Data Science and Machine Learning:

  Machine Learning courses taken at Lund University
  Data Science projects at Big Data and BTG Pactual
  MOOC specific courses

- Motivation and time spent:

  I decided to enter this competition because I wanted to improve my knowledge in timeseries. I spent about 60 hours of work.

A2. Summary

Processed price, calendar and sales history data were used, as well as categorical variables provided by the competition.

The LGBM model was used, and the training and forecasting execution time is approximately 3 hours.

More details are provided in the next sessions

A3. Features Selection / Engineering

From the calendar, I created features from the remaining days for the event representing the "strength" of the proximity of the event. Also used some variables to indicate seasonality. For example, indicating how near the day is to the end of the month. Week of the year, day of the week etc.

From the prices, I focused on calculating the percentual difference of the prices from the weeks before to the current week of the training row.

From the historical series, thanks to @kkiller (@kneroma) team and @kxx team that shared the idea to insert the rolling mean and lag of the sales of 7 days and 28 days before.

It was not used external data

The most important variables varied from department to department.

Some of these most important variables are:

The item itself.
The mean sales of last month and mean sales of last week.
The month itself.
The day of the week.
The week of the year.
The store itself.
The percentual difference between the prices from the weeks before and the week of the sales.

All the important variables can be checked in the solution, for each department, in the third notebook.

A4. Training Method

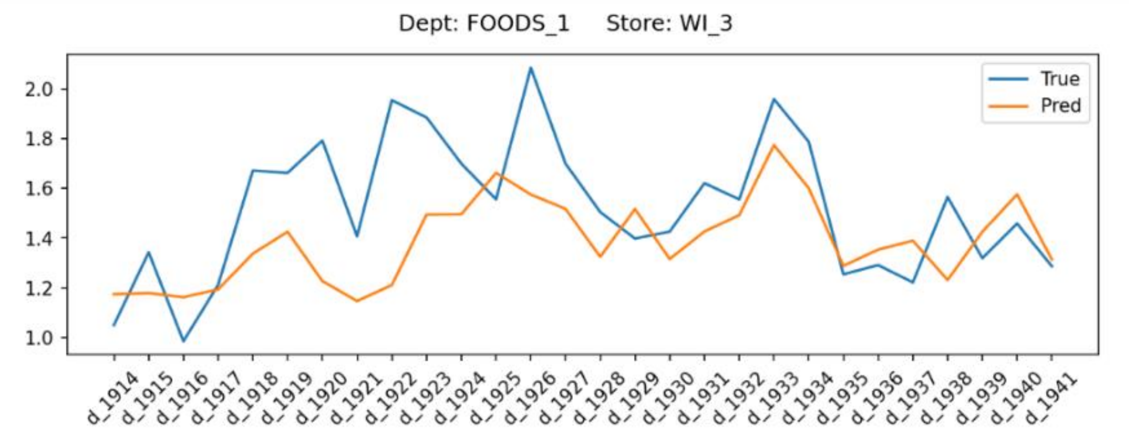
I used an LGBM model with early stopping, and with Poisson objective, for each department, resulting in a total of 7 models. A simple and random separation between training and validation was carried out so that the model was trained and created. It takes around 3 hours to train all the models and use them to forecast values.

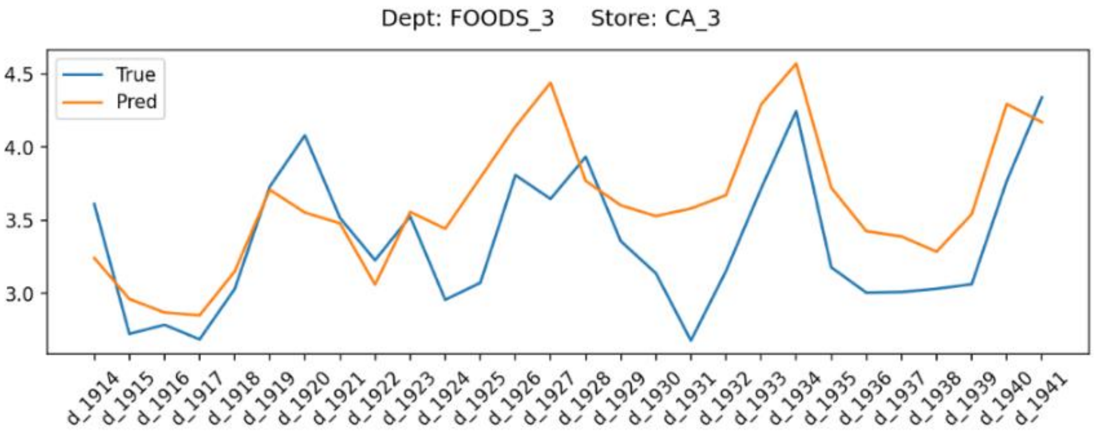
A5. Interesting findings

Post processing:

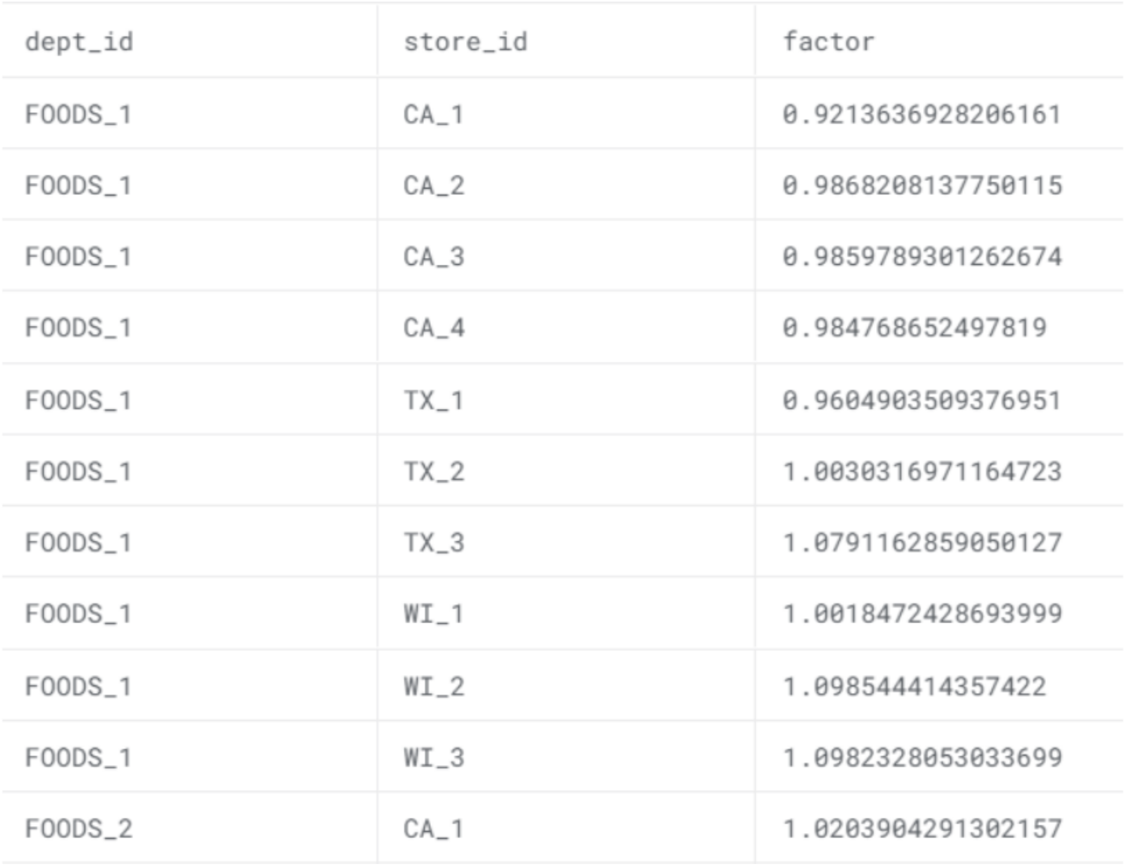
The model itself was resulting in a different offset than desired. That is, on average, the values were 3% to 4% below what they should be. Therefore, following the discussion of the participants to insert a correction multiplier, I chose to bring the average of each department and store of the validation data to the evaluation data. I believe that this was the main different point between my solution and the solution of the other participants, and perhaps the one that made the biggest difference for a good result.

This idea was created after verifying that some stores/departments were constantly below or constantly above what they should be, when we compared the predicted values with the true values, as shown in the graphs below:





Some of the factors for each store/department is shown below, but you can have the whole file in one of the outputs in the submission model



dept_id	store_id	factor
FOODS_1	CA_1	0.9213636928206161
FOODS_1	CA_2	0.9868208137750115
FOODS_1	CA_3	0.9859789301262674
FOODS_1	CA_4	0.984768652497819
FOODS_1	TX_1	0.9604903509376951
FOODS_1	TX_2	1.0030316971164723
FOODS_1	TX_3	1.0791162859050127
FOODS_1	WI_1	1.0018472428693999
FOODS_1	WI_2	1.098544414357422
FOODS_1	WI_3	1.0982328053033699
FOODS_2	CA_1	1.0203904291302157

A6. Simple Features and Methods

For a simplification of the model, I would suggest a few things:

- Reducing the list of variables by choosing the first 10 most important variables from the list in item A3.
- It could also shorten the training days, taking only the last 2 years.
- And if still necessary, do not carry out training for each department, but add the department in the variables and carry out only one training for all departments at the same time

Here there would be a tradeoff between runtime and scoring. However, I do not suggest removing the multipliers from the solution, which I believe was the differential of this specific solution. In addition, multipliers are calculated and applied in a very short time, so it is not a limitation on the execution time.

A7. References

All references used were internal from Kaggle discussions and kernels. All notebooks used were:

1) https://www.kaggle.com/kneroma/m5-first-public-notebook-under-0-50
Author: @kneroma
Main ideas: rolling mean and lags from historical sales

2) https://www.kaggle.com/kneroma/m5-forecast-v2-python
Author: @kneroma
Main ideas: rolling mean and lags from historical sales

3) https://www.kaggle.com/kyakovlev/m5-dark-magic
Author: @kyakovlev
Main ideas: single multiplier