DMC 2020: Prediction of Product Stocks

Data Mining II (FSS2020)

Rebecca Armbruster Wei-Yi Chen Anna Fuchs Sang Hyu Hahn
Christian Klaus Jonas Klenk Jana Pfeffer Yen-Ting Wang

Team abraca-data

May 5, 2020

Data

	order	sales Price	simulation Price	brand	manufacturer	customer Rating	category1	category2	category3	recommended Retail Price
mean	1.239	37.92	34.5326	43.620	111.53	2.412	4.164	23.18	4.159	32.60
std	0.658	132.3	131	56.723	64.41	2.268	1.933	12.68	2.124	99.75
min	1	0	0.38	0	1	0	1	1	1	2.46
25%	1	7.43	6.1	0	66	0	2	10	2	10.83
50%	1	17.45	14.81	0	100	3	4	23	4	16.29
75%	1	34.99	30.96	92	164	5	5	32	6	26.26
max	100	9387.	9055.07	265	253	5	8	51	8	6955

First Approach

Methods employed

- Holt Winter
- Random Forest
- LSTM

General data preparation for time series analysis

- Aggregate to items sold per day per item.
- Replace missing days with zeros.

Holt Winter

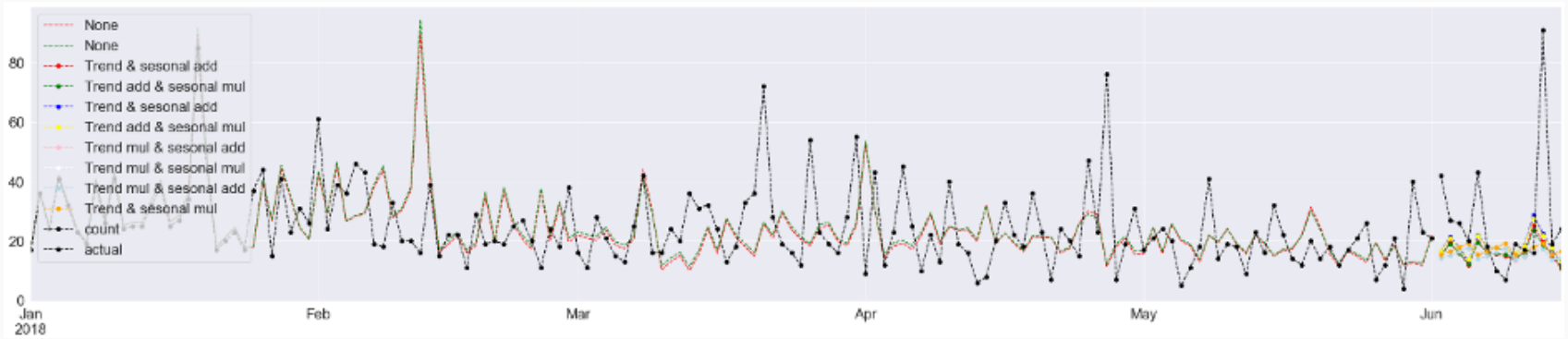


Figure 1: Forecasting sales of item 7798 with both additive and multiplicative seasonality.

Random Forest

- Training data is lagged to create features.
- Actual time series as target.

	count_0	count_1	count_2	count_3	count_4	count_5	count_6	count_7
2018-01-01	17	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-02	36	17.0	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-03	24	36.0	17.0	NaN	NaN	NaN	NaN	NaN
2018-01-04	41	24.0	36.0	17.0	NaN	NaN	NaN	NaN
2018-01-05	32	41.0	24.0	36.0	17.0	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...
2018-05-28	21	12.0	7.0	26.0	21.0	17.0	12.0	18.0
2018-05-29	4	21.0	12.0	7.0	26.0	21.0	17.0	12.0
2018-05-30	40	4.0	21.0	12.0	7.0	26.0	21.0	17.0
2018-05-31	23	40.0	4.0	21.0	12.0	7.0	26.0	21.0
2018-06-01	21	23.0	40.0	4.0	21.0	12.0	7.0	26.0

Table 1: Example training data.

Random Forest

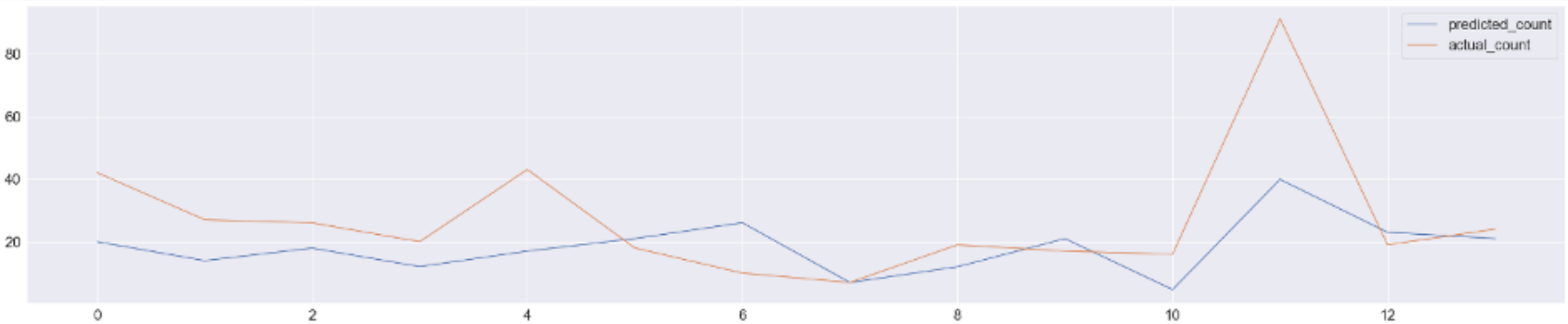


Figure 2: Random Forest prediction and actual for item 7798.

Performance compared to baseline:

Perfect Result: 7895975.87

Random Forest: -1,359,513.29

Baseline 2: -1,672,504.21

Baseline 1: -3,727,365.60

LSTM

- Generate time series data with 14 lags.
- Parameters: activation='relu', dropout(0.15), dense(1), optimizer='adam', loss='mse', shuffle = False, verbose = 1

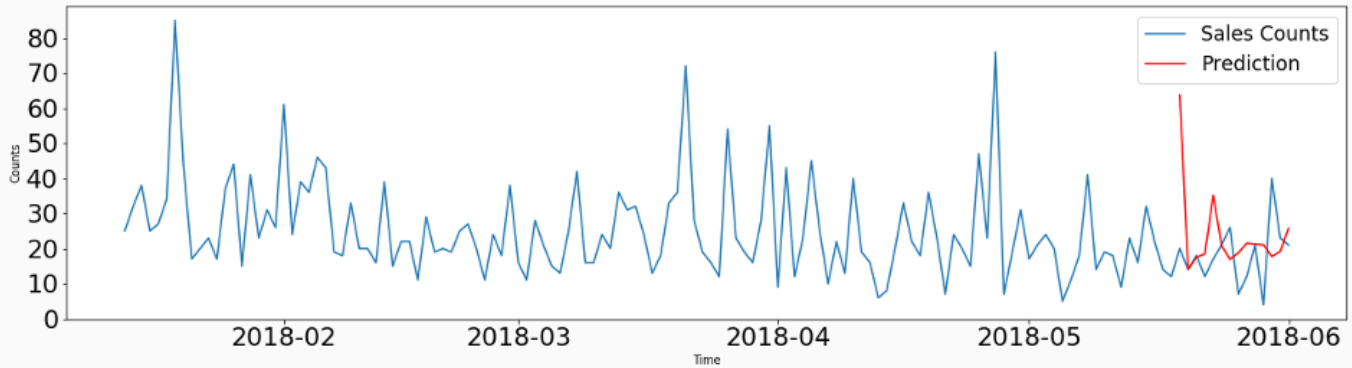


Figure 3: Prediction for item 7798 for the last 14 days in training dataset with 500 epochs. Loss: 0.033

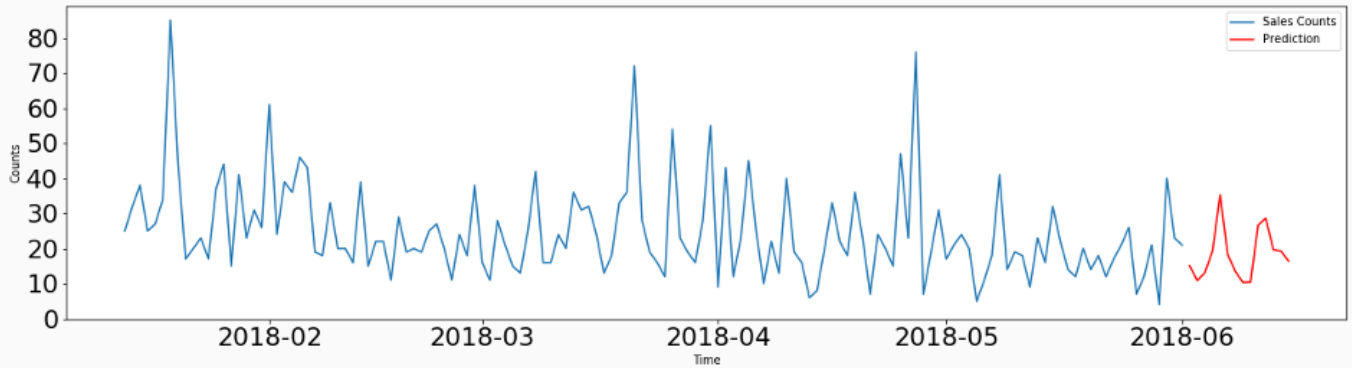


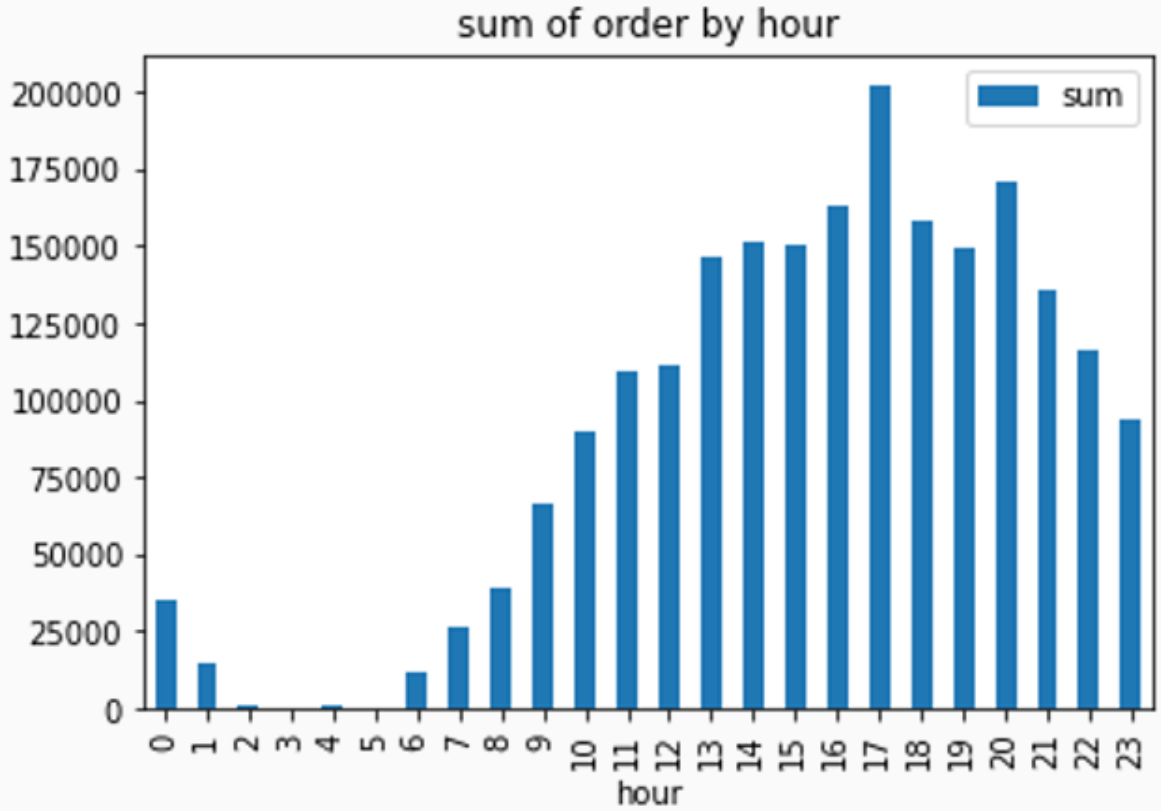
Figure 4: Prediction for item 7798 for the last 14 days in training dataset with 150 epochs. Loss: 0020

Next Steps

- Generate good features
- Check for seasonality

Feature Generation

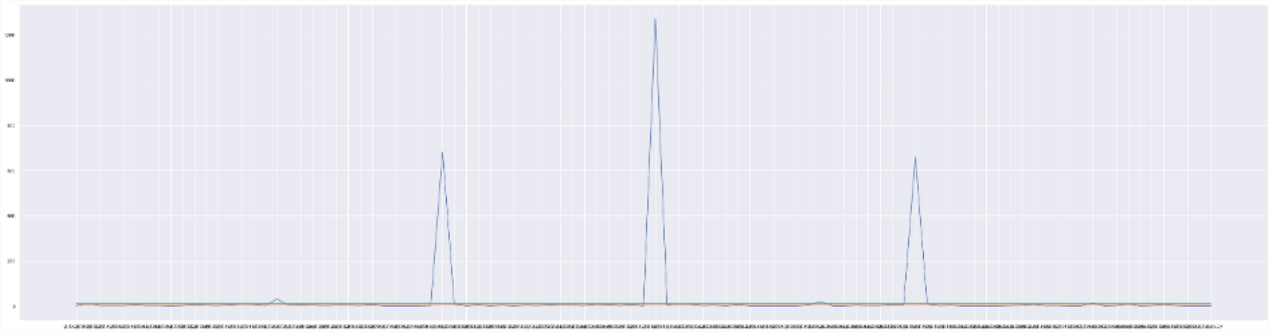
- Purchase time in various dimensions (weekday/day of the month/month/hour/calendar week)
- Discount from Recommended Retail Price to Sales Price (absolute/relative)

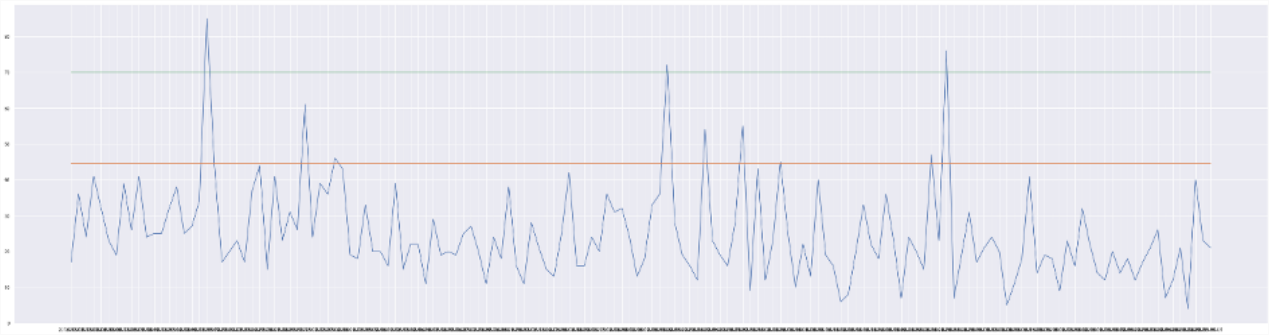


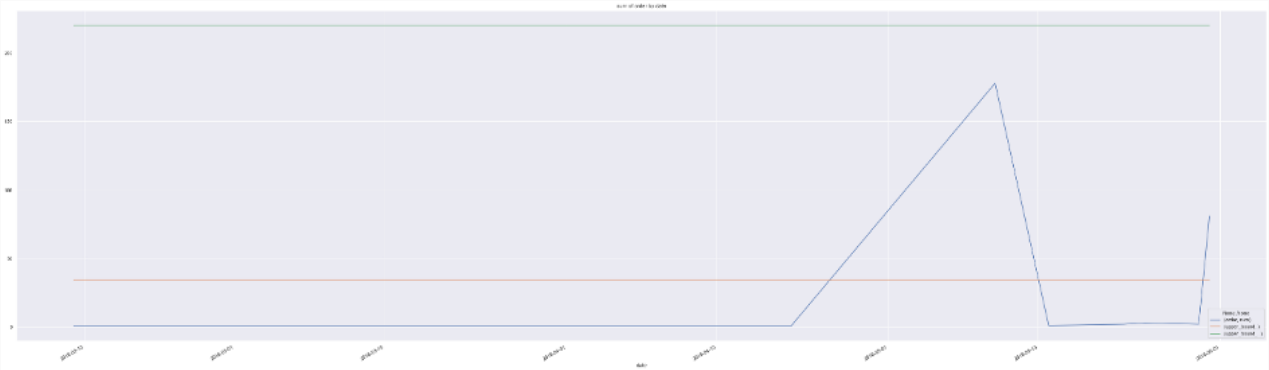
Promotion Flag

- Add a promotion flag to orders if an item was promoted that day.
  - $\rightarrow$ Promotions in Info.csv are only available for the future
- Explore ways to identify promotions in the orders data as outliers
  - $\rightarrow$ Use IQR as in the lectures
  - $\rightarrow$ Use modified IQR $(Median + 2 * IQR[25, 90])$

Promotion Flag







Seasonality Check

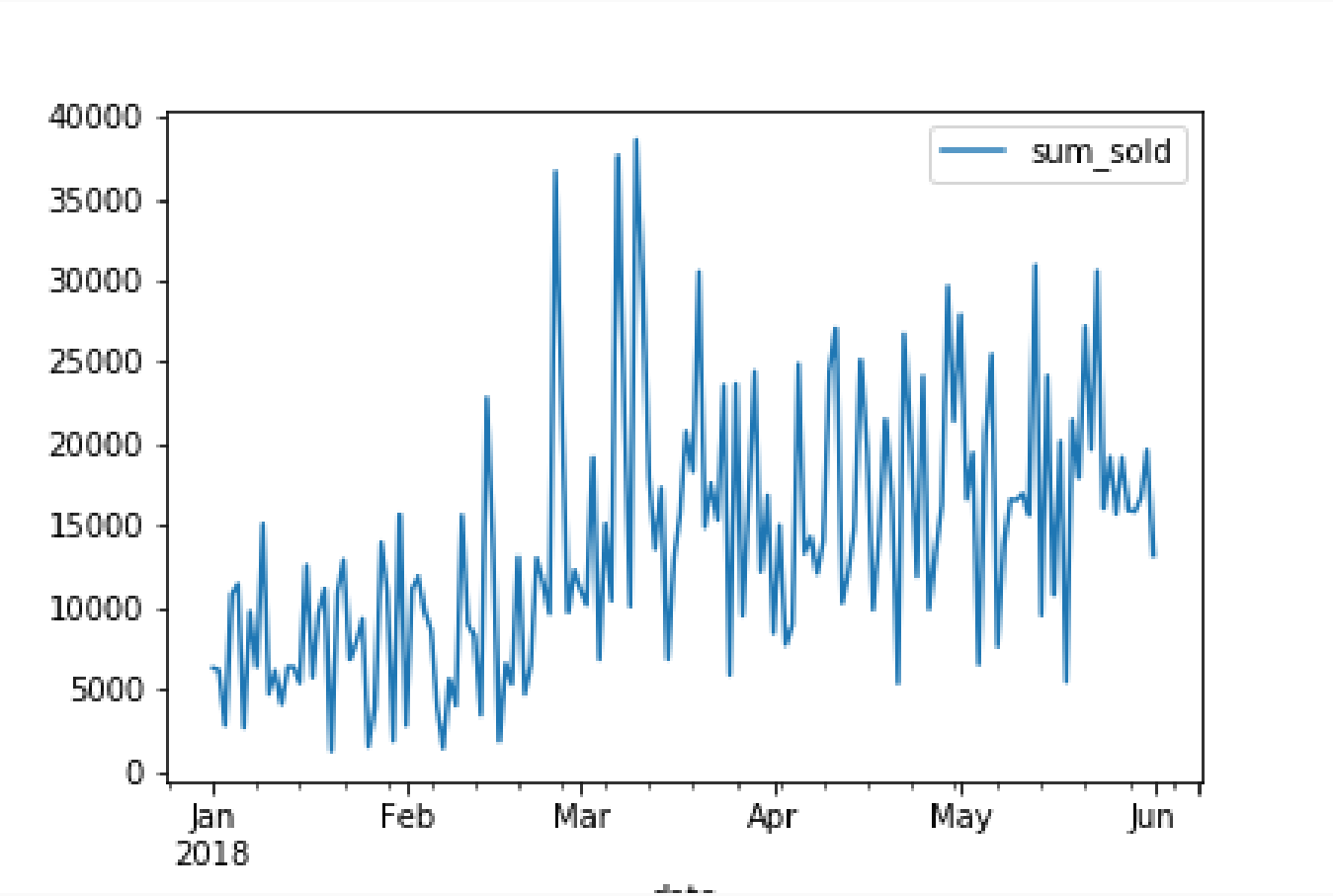


Figure 5: Seasonality.

Questions

-