Coles Advanced Analytics presentation

Corporacion Favorita Grocery Sales Forecasting

Brendan Houng, 26 Oct 2021

Overview – Summary of Problem

The Kaggle competition is to forecast grocery sales for Corporacion Favorita, a large grocery retailer in Ecuador.

- Accurately forecasting product sales will assist with supply management and the logistics of much inventory to keep or ship
- this is particularly important for perishable products (fruit & veg)

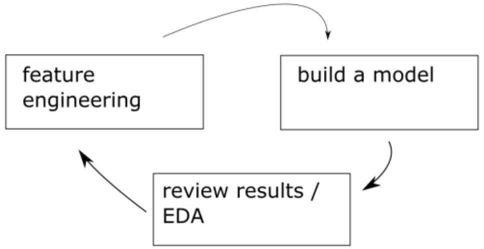
- Structure of data is : Date, Item number, Store number, Unit Sales
- Challenging to predict unit of sales for item, by store and date – not much detail on items (broad categories)

Step-by-Step Approach*

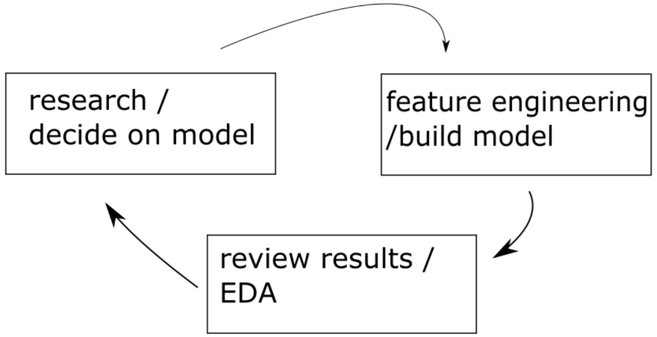
1. Preliminary Research
2. Exploratory Data Analysis
3. Problem Interpretation/Decide on Model
4. Feature Engineering
5. Build a Model
6. Review Results

* preliminary research was outside of Kaggle discussion – not in the spirit of the competition. Refer to notes.

iterative approach in model variations from steps (4) to (6), building and evaluating a model



iterative approach in model selection – moving back to steps (1,2,3) from (6)



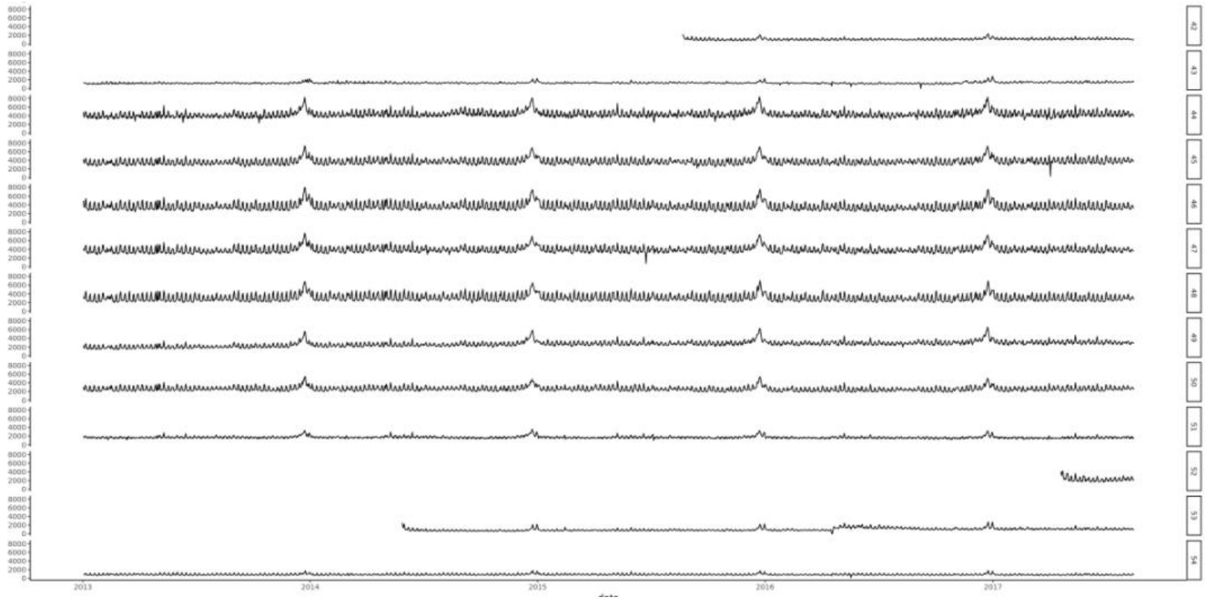
2. Exploratory Data Analysis

- Training and test data, with 1688 and 16 days of data respectively

```
"number of stores: 54"
"no of items in train dataset: 4036"
"no of items in test dataset: 3901"
```

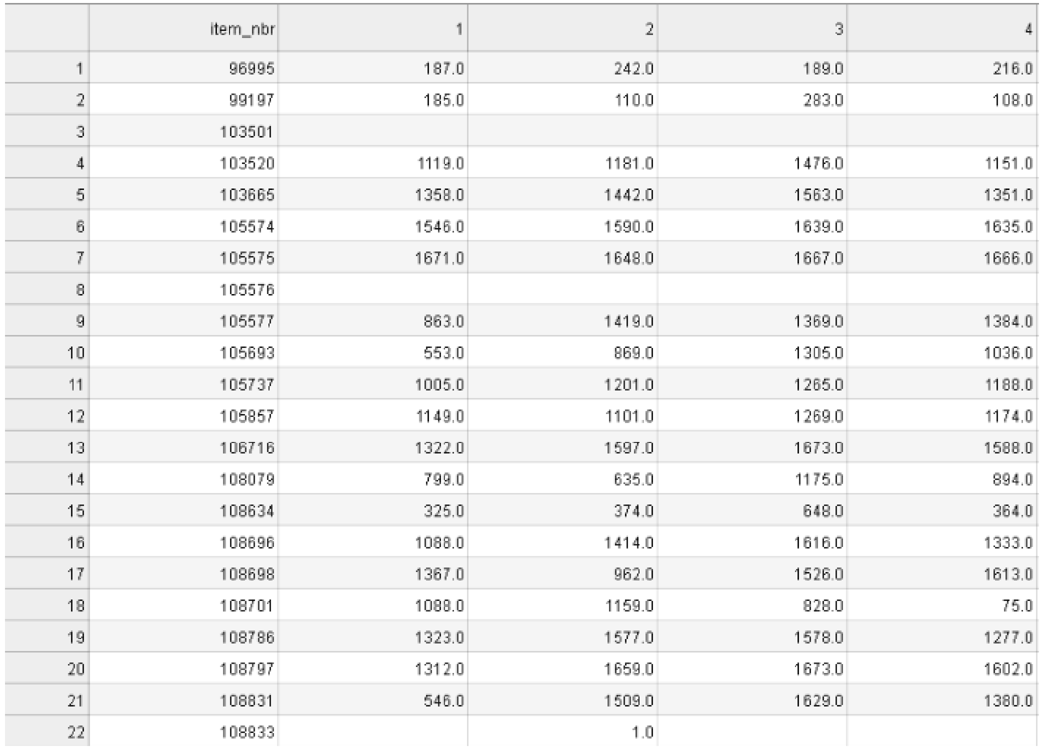
- A plot of the number of transactions by store over time (Fig 1) shows that some stores were opened throughout the training dataset time period (2014-2017). The volume of transactions also varies by geography and time.

Figure 1 – plot of no. transactions by store over time



- A count of the number of transactions by item number and by store number shows that some items are sparsely populated by store.

Table 1 – count of dates by item and by store

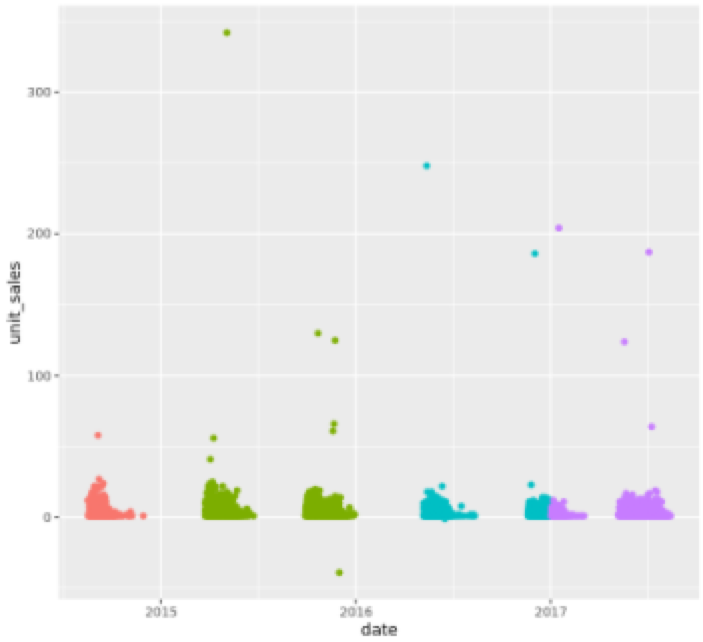


	item_nbr	1	2	3	4
1	96995	187.0	242.0	189.0	216.0
2	99197	185.0	110.0	283.0	108.0
3	103501				
4	103520	1119.0	1181.0	1476.0	1151.0
5	103665	1358.0	1442.0	1563.0	1351.0
6	105574	1546.0	1590.0	1639.0	1635.0
7	105575	1671.0	1648.0	1667.0	1666.0
8	105576				
9	105577	863.0	1419.0	1369.0	1384.0
10	105693	553.0	869.0	1305.0	1036.0
11	105737	1005.0	1201.0	1265.0	1188.0
12	105857	1149.0	1101.0	1269.0	1174.0
13	106716	1322.0	1597.0	1673.0	1588.0
14	108079	799.0	635.0	1175.0	894.0
15	108634	325.0	374.0	648.0	364.0
16	108696	1088.0	1414.0	1616.0	1333.0
17	108698	1367.0	962.0	1526.0	1613.0
18	108701	1088.0	1159.0	828.0	75.0
19	108786	1323.0	1577.0	1578.0	1277.0
20	108797	1312.0	1659.0	1673.0	1602.0
21	108831	546.0	1509.0	1629.0	1380.0
22	108833		1.0		

2. Exploratory Data Analysis

- A visual inspection of unit sales over time by item show that models that rely predictable seasonal (calendar) sales patterns would not be a good fit. See for example Figure 3.
- Further plots and tables are in the appendix and notebooks

Figure 2 – unit sales for item 99197*

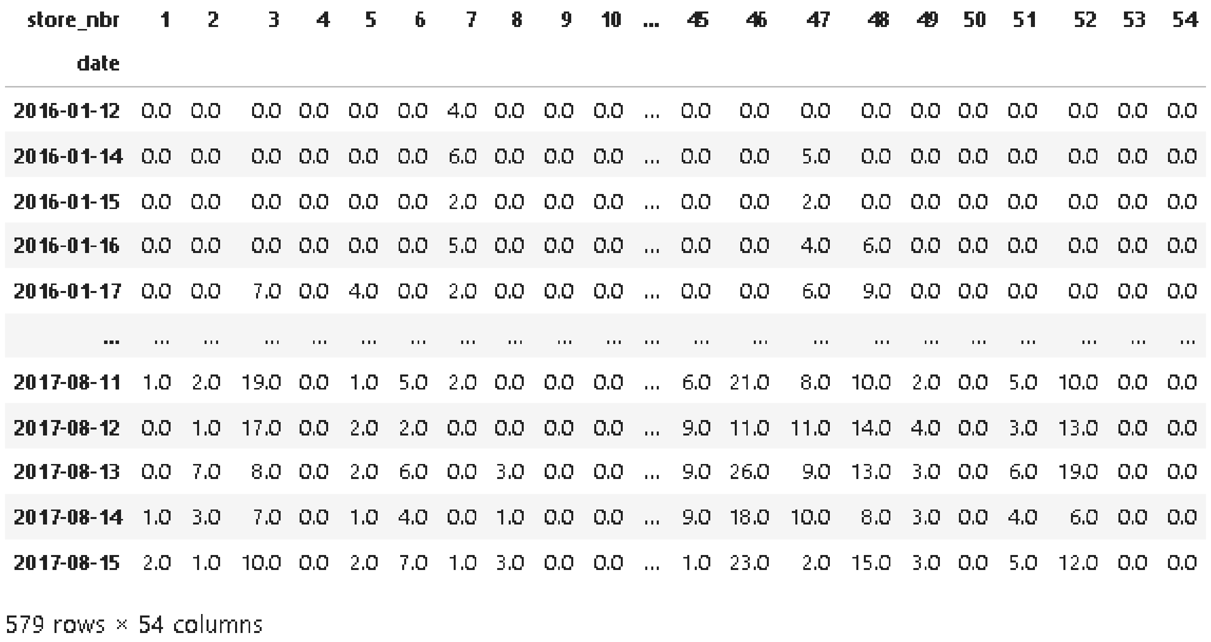


*coloured by year

3. Problem Interpretation

- The **sparsity of unit sales for certain items across stores** means that it will not be possible (or very difficult) to forecast the unit sales for items by specific store
- This means that to address the **large percentage of missing items across stores**, it is best to **model the item unit sales by pooling the stores**, such that the models can produce a result for the stores that have no or low volumes

Table 2 – unit sales for item 1956004 by store and date^



store_nbr	1	2	3	4	5	6	7	8	9	10	...	45	46	47	48	49	50	51	52	53	54
date																					
2016-01-12	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2016-01-14	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	...	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2016-01-15	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	...	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2016-01-16	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	...	0.0	0.0	4.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0
2016-01-17	0.0	0.0	7.0	0.0	4.0	0.0	2.0	0.0	0.0	0.0	...	0.0	0.0	6.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2017-08-11	1.0	2.0	19.0	0.0	1.0	5.0	2.0	0.0	0.0	0.0	...	6.0	21.0	8.0	10.0	2.0	0.0	5.0	10.0	0.0	0.0
2017-08-12	0.0	1.0	17.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	...	9.0	11.0	11.0	14.0	4.0	0.0	3.0	13.0	0.0	0.0
2017-08-13	0.0	7.0	8.0	0.0	2.0	6.0	0.0	3.0	0.0	0.0	...	9.0	26.0	9.0	13.0	3.0	0.0	6.0	19.0	0.0	0.0
2017-08-14	1.0	3.0	7.0	0.0	1.0	4.0	0.0	1.0	0.0	0.0	...	9.0	18.0	10.0	8.0	3.0	0.0	4.0	6.0	0.0	0.0
2017-08-15	2.0	1.0	10.0	0.0	2.0	7.0	1.0	3.0	0.0	0.0	...	1.0	23.0	2.0	15.0	3.0	0.0	5.0	12.0	0.0	0.0

579 rows × 54 columns

^ train dataset

3. Model Decision

models considered:

1. For each item, a **multi-output** (54 outputs, one for each store) and **multi-step** time series forecast with **regression** and **decision tree** models and unit sales from previous time periods as lags

$$Y_{j(t+1,t+15)} = Y_{jt} + Y_{jt-1} + \cdots + Y_{jt-49}$$

2. neural network LSTM forecast model with date and holiday features and one unit sale lag for each store, either 1 model for each time step to be forecast (16 days), or multi-step model predicating all dates together

$$Y_{j=1}, Y_{j=2}, Y_{j=3}, \ldots Y_{j=54}$$
$$= Y_{j=1t} + Y_{j=2t} + \cdots + Y_{j=54t} + date + holiday\ features$$

3. Neural network LSTM forecast model with the same structure as model 1, but with unit sales estimated jointly for all stores. unit sales with **N** lags for 54 stores.

$$Y_{j=1} \ldots Y_{j=54}$$

NN are good for representing complex r/ships*

*multi-output w/ geographic correlation

solution & performance

Performance of **1st model** – best of regression and decision tree

1. Private score of 0.76696 and public score of 0.73181 for store pooled items with 50 lags, best of regression and decision tree forecasts based on a training/validation 80/20 split of the training data

2. Ran a mean and median forecast based on code from discussion board – which provides results of 0.9086 and of 0.79525, respectively.

In the competition, the top 50 scores were between 0.50918 and 0.51887.

Limitations

One of the main limitations of the regression/decision tree based model is that:

- the model accounts for the weighting of past periods based on all the stores' unit sales
- but the model does not explicitly weight the previous sales from other stores of the same item
- this is something that a neural network model would account for, have previous sales from other stores as features for the model of a item for a given store

Evaluation metric

Normalized Weighted Root Mean Squared Logarithmic Error

$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i(\ln(\hat{y}_i + 1) - \ln(y_i + 1))^{2}}{\sum_{i=1}^{n} w_i}}$$

Intuition is that this is roughly:

log(predicted unit sales) – log(actual unit sales), where:

- n is the number of rows
- w is weighting of item. 1.25 for perishable

The metric was deemed suitable for when predicting across a large range of orders of magnitude.
i.e. less weighting is given to the error (difference between prediction and actual) when numbers are large

Discussion of alternatives

Remove logs

- weighting difference in units equally regardless of volume for item
- i.e. better to stock closer to demand, rather than customers preferring low volume items
- Also potentially approximates transport costs due to volume, depending on size of item.

$$RMSE = \sqrt{}\Sigma(P_i – O_i)^2 / n$$

Change weights

- Rather than only perishable items having greater weights could place greater weight on transport costs, size of item. Available inventory is another. Stockout?
- Mean Absolute Percentage Error would be even more sensitive to low volumes than the competition metric

$$\mathrm{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

3. Model Decision

Time constraints

- was able to prototype model 2 – LSTM neural network of 54 outputs with 1 lag and week, month and holidays features, with a view to also building out model 3

- my initial exploration of the results a validation dataset for one item led me to think that model 3 is a better solution.

Notes and Oversights

- Note: 3841 items were in both train and test datasets while there were additional 60 items in the test set but not in the train dataset.

- Due to time constraints – set the 60 items which were not in the training set to zero values, but would have liked to estimate on the basis of item family (only half of items had this data).

- Due to hastiness – I missed the inclusion of the feature capturing whether a product was perishable (though I wrote it down in the feature engineering notebook), which would have impacted the overall metric.

Improvements given more time

- Hyperparameter tuning for random forest decision-tree modelling for model 3 (and for model 1, which I didn't do). Gradient boosting alternatives

- A review (Bojer and Melgaard^) of Kaggle time series forecasting competitions indicates that combinations of LightGBM (Gradient Boosting Model), with one model per forecast horizon, and feed forward neural network models were adopted by the winner. Models were fitted over either 1, 3 or 5 months of data.

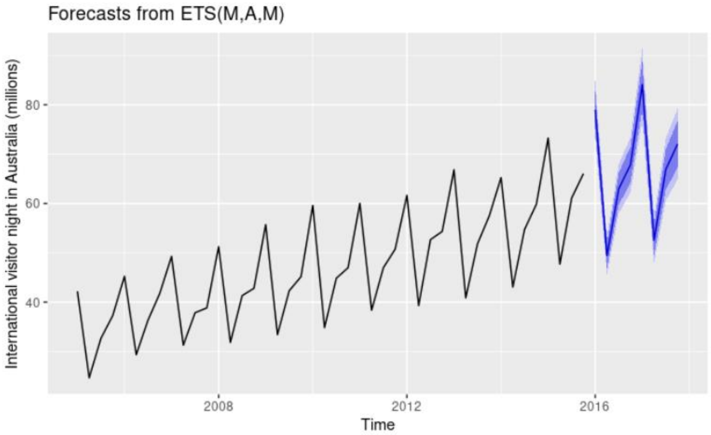
- Given more time I would have liked to have built a feed-forward multi-step multi-output neural network model with one output for each store.

- Would like to investigate LightGBM. In practice, through entering the competition and following the discussion, it is likely that a competitor would adopt more better performing models that are shared.

^Learnings from Kaggle's Forecasting Competitions, International Journal of Forecasting (2021)

- In practice – **models can be custom tuned** to an extent, for high volume items in larger stores more traditional seasonality regression models could be applied with expected greater degree of accuracy. These seasonality regression models probably perform better for longer time horizons

Forecast models suited to predictable patterns such as ETS and FB's prophet ruled out due to sparsity of items across stores. Further discussion later on suggested improvements.



how to extend the solution to production

Set up a scheduler* to the do the following:

- Update the items and unit sales time series data
- Create updated train and test sets for new day
  - the test set would then be the next 16 days
  - which can be extended to the desired forecast horizon
  - the accuracy of the forecast could be re-evaluated
- Re-run the data processing steps (and write new data to database/csv)
- Run the forecasting models
- Collate the results and write to desired location

Options: cron job, google cloud scheduler /aws cloudwatch, gitlab pipelines

how to extend the solution to production

Set up a scheduler* to the do the following:

- Update the items and unit sales time series data

- Create updated train and test sets for new day
  - the test set would then be the next 16 days
  - which can be extended to the desired forecast horizon
  - the accuracy of the forecast could be re-evaluated

- Re-run the data processing steps (and write new data to database/csv)

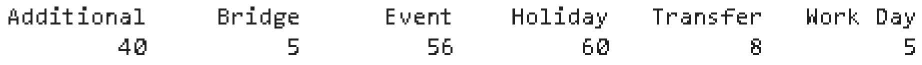
- Run the forecasting models

- Collate the results and write to desired location

Options: cron job, google cloud scheduler /aws cloudwatch, gitlab pipelines

Appendix. Exploratory Data Analysis

Table: A.1. Counts of holidays by type



```
Additional     Bridge      Event    Holiday   Transfer   Work Day 
        40          5         56         60          8          5 
```

Figure: A.1. Total Transactions over time

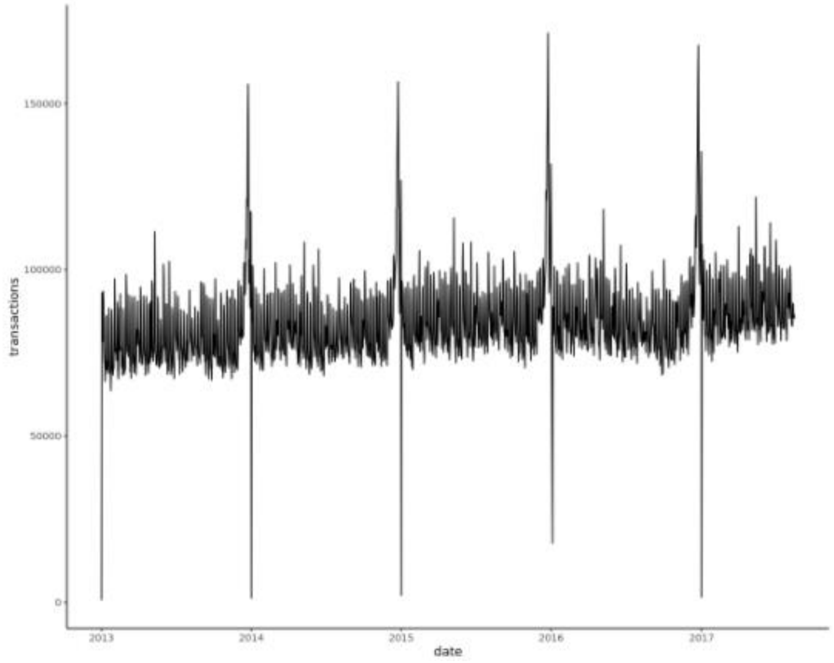
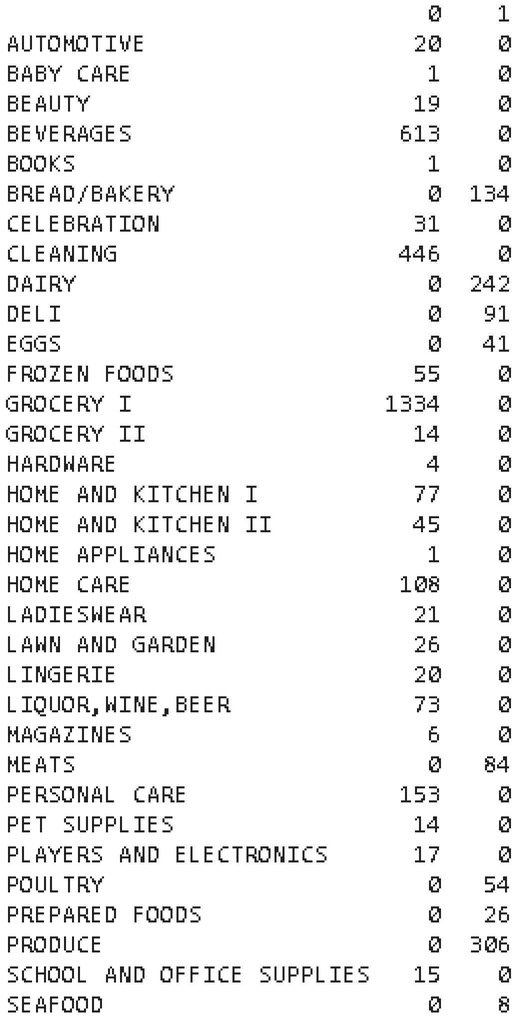


Table: A.2. Counts of item family by perishable

Perishable item



```
                               0    1
AUTOMOTIVE                    20    0
BABY CARE                      1    0
BEAUTY                        19    0
BEVERAGES                    613    0
BOOKS                          1    0
BREAD/BAKERY                   0  134
CELEBRATION                   31    0
CLEANING                     446    0
DAIRY                          0  242
DELI                           0   91
EGGS                           0   41
FROZEN FOODS                  55    0
GROCERY I                   1334    0
GROCERY II                    14    0
HARDWARE                       4    0
HOME AND KITCHEN I            77    0
HOME AND KITCHEN II           45    0
HOME APPLIANCES                1    0
HOME CARE                    108    0
LADIESWEAR                    21    0
LAWN AND GARDEN               26    0
LINGERIE                      20    0
LIQUOR,WINE,BEER              73    0
MAGAZINES                      6    0
MEATS                          0   84
PERSONAL CARE                153    0
PET SUPPLIES                  14    0
PLAYERS AND ELECTRONICS       17    0
POULTRY                        0   54
PREPARED FOODS                 0   26
PRODUCE                        0  306
SCHOOL AND OFFICE SUPPLIES    15    0
SEAFOOD                        0    8
```