Predicting Store Sales Using Time Series Forecasting

Springboard Data Science Track Capstone Project
Ngoc Tran: Data Scientist Student, PharmD.

What is Food Waste?

Food is wasted at 3 levels:

1. **Production**- food is damaged or spoiled
2. **Retail**- food thrown out due to:
    a. Overbuying
    b. Deviations from looking “optimal” i.e. bruising, discoloration, abnormal shape, etc.
3. **Consumer**- customers buy more than needed & throw out unused or spoiled food

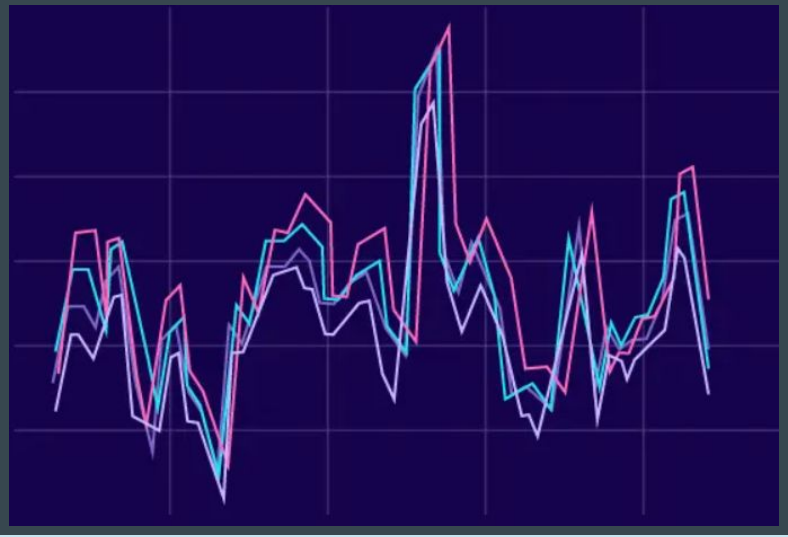
Why Should You Care?

- According to the World Food Program USA, ~$1 **trillion** dollars worth of food is wasted **yearly**
- **95%** of food waste ends up at landfills
- Rotten food emits methane, carbon dioxide, and other greenhouse gases
- Contributes to **global warming** and **climate change**

How Can We Fix This?

Reduce food at the **Retail** level

- **Create time series forecast model to predict sales**
- Help grocery stores find balance between having just enough inventory while avoid product shortages
- Accurate forecasting results in happy customers and less food waste



The Data

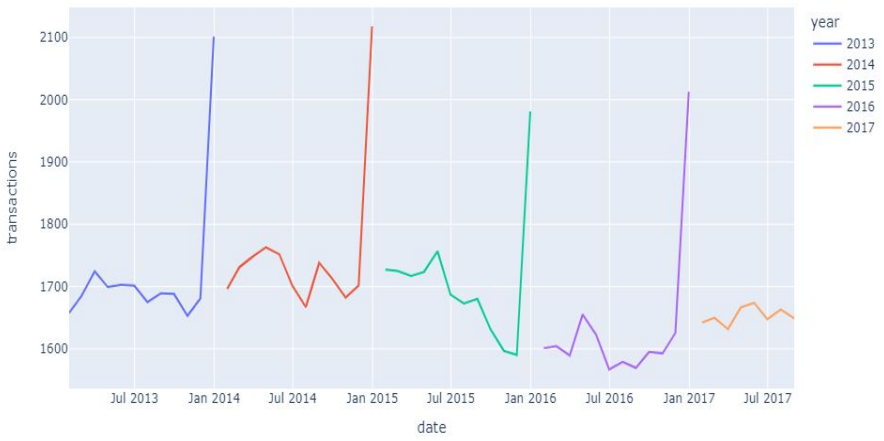
- Dataset from Kaggle
- Corporación Favorita, Ecuadorian-based grocery retailer
- Training data: 2013-01-01 to 2017-08-13
- Test data: 2017-08-16 to 2017-08-31
    - 15 days after training data

The Data

- Train and test data features:
  - **store_nbr**: store number
  - **family**: type of product sold
  - **onpromotion**: total number of items in a product family that were being promoted
  - ➢ **sales**: target feature; total sales for a product family
- Transactions features
  - store_nbr
  - transactions: how many invoices were created in a data

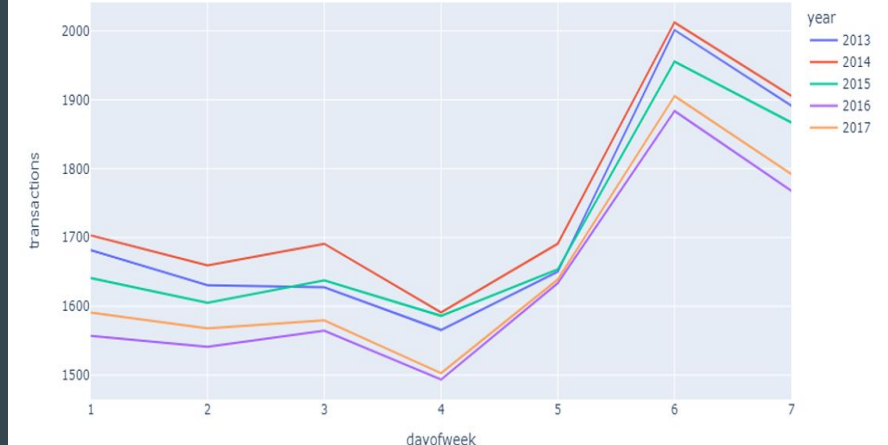
Exploratory Data Analysis: Transaction Analyses

Figure 1



The monthly average transactions data from 2013 to 2017 all seem to follow the same pattern. Sales increase around March-May, and then decrease around July before increasing to a huge spike in December. From Figure 1, It would be a good idea to reduce inventory during the summer and increase inventory towards the end of the year to accommodate for the sales pattern.

Figure 2



Narrowing the transactions down to a weekly basis, all the years follow the same pattern with day 6 and 7 (Saturday and Sunday, respectively) having the highest amount of transactions. There is also a dip on day 4 (Thursday) with the least amount of transactions. From Figure 2, the stores should consider stocking enough inventory to meet the needs of the weekend shoppers.

Exploratory Data Analysis: Product Families Analysis

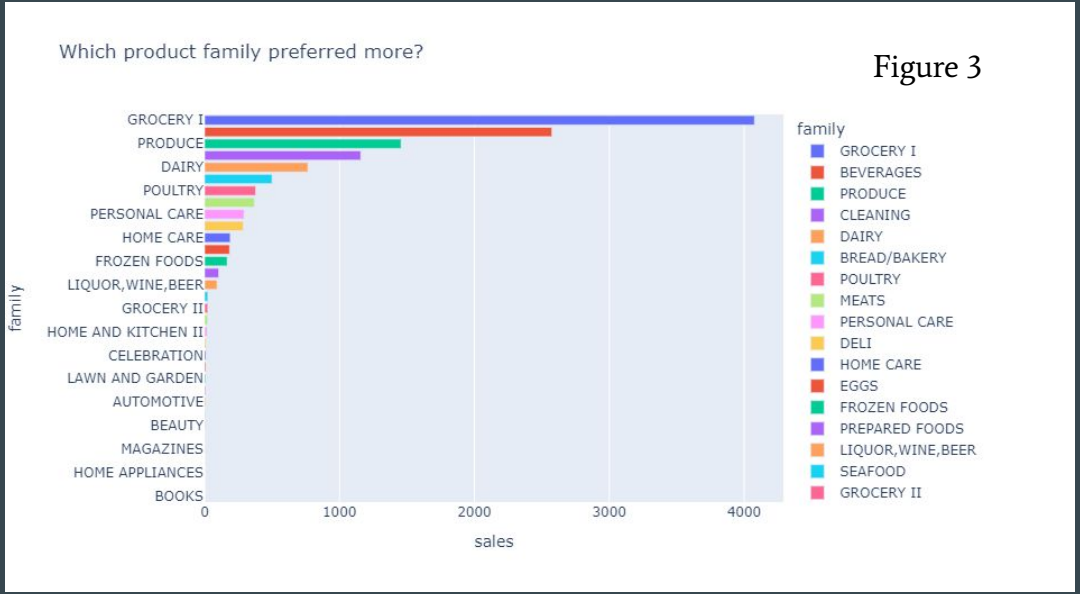


Figure 3

Figure 3 depicts which product families require a higher number of inventory compared to other product families. Product families like Grocery I, Produce, and Dairy are in higher demand and need to be restocked more often than Frozen Foods, Alcohol, and Grocery II. This information can also help the stores determine how much shelf space each product family can take up within the store. A popular product family should take up more space than a product family with lower sales.

Exploratory Data Analysis: Products On Promotion Analysis

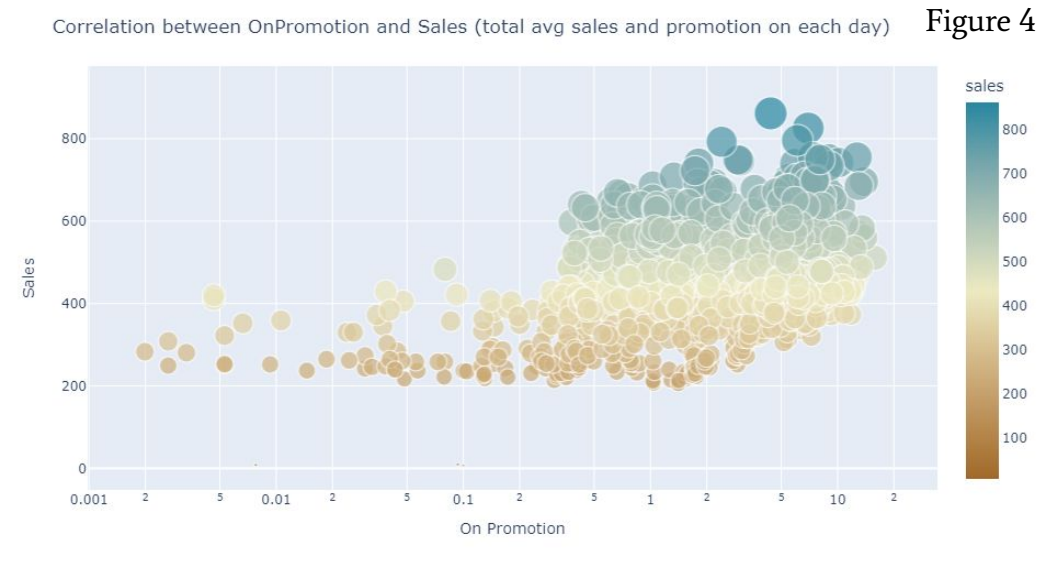


Figure 4

Figure 4 shows there is a positive correlation between on promotion items and sales units sold. If a store is looking to clear a certain type of inventory quickly, putting the item on promotion can influence customers to want to buy the product.

Modeling: Defining the Forecasting Task

- **Features**- info readily available to build model
  - Training data: 2013-01-01 to 2017-08-15
    - **Forecast origin**
- **Target**- time period being forecasted
  - Testing data: 2017-08-16 to 2017
  - Forecast 15 days after end of training data
    - **Forecast horizon**
    - One step between origin and horizon = lead time of 1 day
- ***16-step forecast with a 1-step lead time***
  - Using lags starting with lag 1 and making the entire 16-step forecast with features from 2017-08-15

Modeling: Forecasting with DirRec Strategy

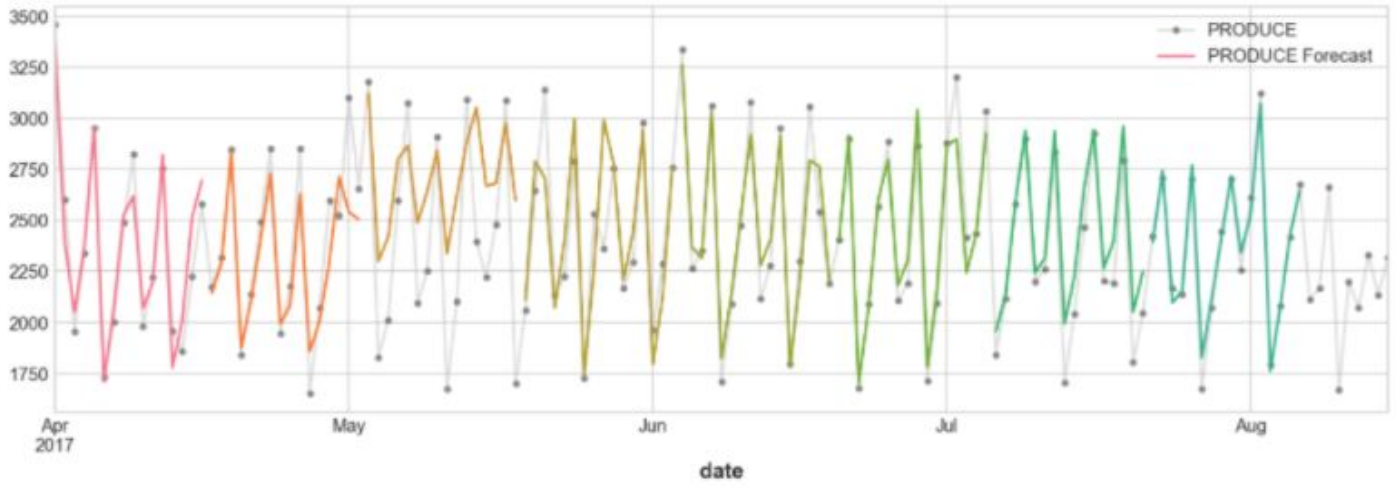


Figure 5

The DirRec strategy trains a model for each step and uses forecasts from previous steps as new lag features. Step by step, each model gets an additional lag input.The forecast and the actual values seem to overlap with each other quite nicely.

Comparing Predicted vs. Actual Values

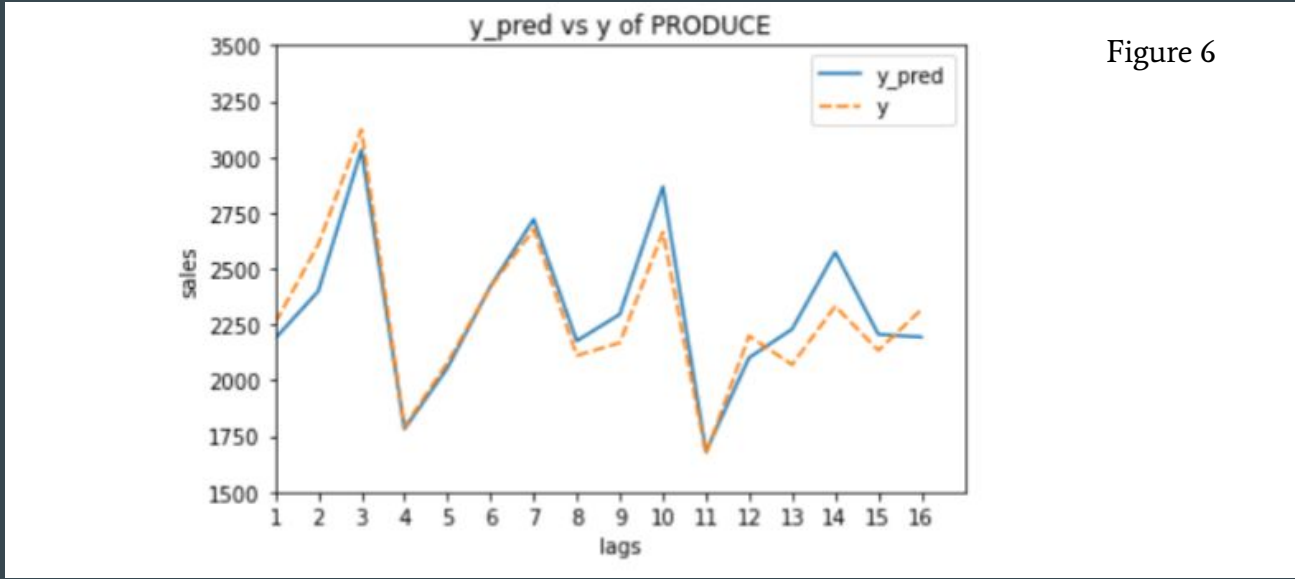


Figure 6

The forecasted (y_pred) and the actual (y) values for the lags have very similar Produce sales values, validating the accuracy of the time series. Towards the beginning, the model tends to slightly underestimate, but towards the later lags, the model tends to overestimate. However, it does not look like there is a huge disparity.

Evaluation Metric: Root Mean Squared Logarithmic Error (RMSLE)

$$\text{RMSLE}\,(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(1+\hat{y}_i) - \log(1+y_i))^2}$$

Figure 7

```
RMSLE for lag 1: 0.031152
RMSLE for lag 2: 0.083929
RMSLE for lag 3: 0.029653
RMSLE for lag 4: 0.005451
RMSLE for lag 5: 0.011249
RMSLE for lag 6: 0.001015
RMSLE for lag 7: 0.017099
RMSLE for lag 8: 0.030481
RMSLE for lag 9: 0.056990
RMSLE for lag 10: 0.073972
RMSLE for lag 11: 0.007164
RMSLE for lag 12: 0.045324
RMSLE for lag 13: 0.074060
RMSLE for lag 14: 0.098723
RMSLE for lag 15: 0.033067
RMSLE for lag 16: 0.054542
```

Figure 7 shows the calculated RMSLE values for the 16 lags of the Produce product family. The errors are low, with the highest error at lag 14 (RMSLE = 0.098723), further verifying the accuracy of the time series forecast using the DirRec strategy.

Takeaways

- Analyzing transactions trends can help retailers identify the shopping pattern
- Promoting a product leads to higher sales
- Time series forecasting with the DirRec strategy helps retailers accurately predict sales
- **Accuracy is crucial**:
  - Underestimating leads to understocking, resulting in customer dissatisfaction
  - Overestimating leads to overstocking, resulting in food waste
- Accurate forecasting can combat against food waste, emission of greenhouse gases, and climate change!