Capstone 3: Time Series Forecasting Report

Executive Summary

In this project, I worked with data from Corporación Favorita, a large Ecuadorian-based grocery retailer to predict store sales using time series forecasting. The data is readily accessible through Kaggle and includes dates, store and product information, whether the item was being promoted, as well as the sales numbers. The training data includes dates from 2013-01-01 to 2017-08-15 and the test set comprises dates from 2017-08-16 to 2017-08-31. Analyzing the training data revealed:

- There is a sharp increase in number of transactions towards the end of the year
- The busiest day of the week is Saturday
- Grocery I, Beverages, Produce, Cleaning, and Diary are the top 5 selling product families
- Products on promotion are correlated with a higher number of sales

Employing the DirRec strategy, a 16-step forecast with a 1-step lead time is used to build the time series. The predicted values were compared to the actual test values using Root Mean Squared Logarithmic Error (RMSLE) as the metric, and verified the actual values were very close to the predicted values. This information would be very useful to help the stores reduce food waste by knowing when to increase or decrease inventory according to the trends of customer demands.

Background

Food loss and food waste is a major global concern. Food is wasted at three levels:

- At the production level, where food is damaged or spoiled
- At the retail level, where food is thrown out due to overbuying or deviations from what is considered optimal (bruising, discoloration, abnormal shape, etc.)
- At the consumer level, when customers buy more than they need and throw out unused or spoiled food.

According to the World Food Program USA, nearly one third of all food produced each year is squandered or spoiled before it can be consumed and approximately $1 trillion dollars' worth of food is lost or wasted yearly. According to this article published in 2013 by the Central European Journal of Engineering, more than 95% of food waste ends up at landfills where they rot, emitting methane, carbon dioxide, and other greenhouse gasses, thus contributing to global warming and climate change.

A possible solution to decreasing food waste at the retail level is to help grocery stores find the delicate balance between providing just enough inventory for customers and restocking at the appropriate time to avoid shortages. Accurate forecasting can help ensure retailers please customers by having just enough of the right products.

Data Cleaning and Wrangling

The training data contained features such as: **store_nbr, family, onpromotion,** and **sales**:

- **store_nbr** refers to the store at which the products are sold
- **family** refers to the type of product sold
- **onpromotion** gives the total number of items in a product family that were being promoted at a store on a given date
- **sales**, the target feature, gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).

The training data comprises dates from 2013-01-01 to 2017-08-15. There was an additional transaction.csv sheet that included the store_nbr and number of transactions corresponding with the dates of the training data The test data have the same features as the training data and comprises dates from 2017-08-16 to 2017-08-31.

The training data had over 3 million entries and did not have any missing entries. As the data was clean and did not need further wrangling aside from converting the date column into datetime format, it was ready to be analyzed.

Exploratory Data Analysis

Transactions Analysis

Transactions means how many people came to the store or how many invoices were created in a day. Sales gives the total sales for a product family at a particular store on a given date.

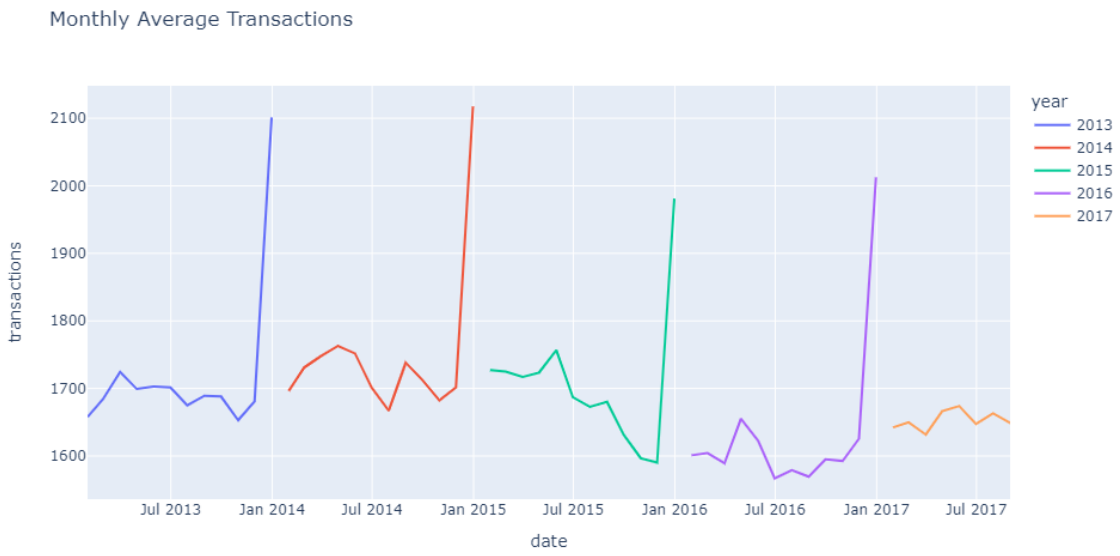

Figure 1: Sales have a tendency to skyrocket at the end of the year. May need to increase inventory around this time. July usually has the least amount of transactions, so can reduce inventory around the summertime.

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.

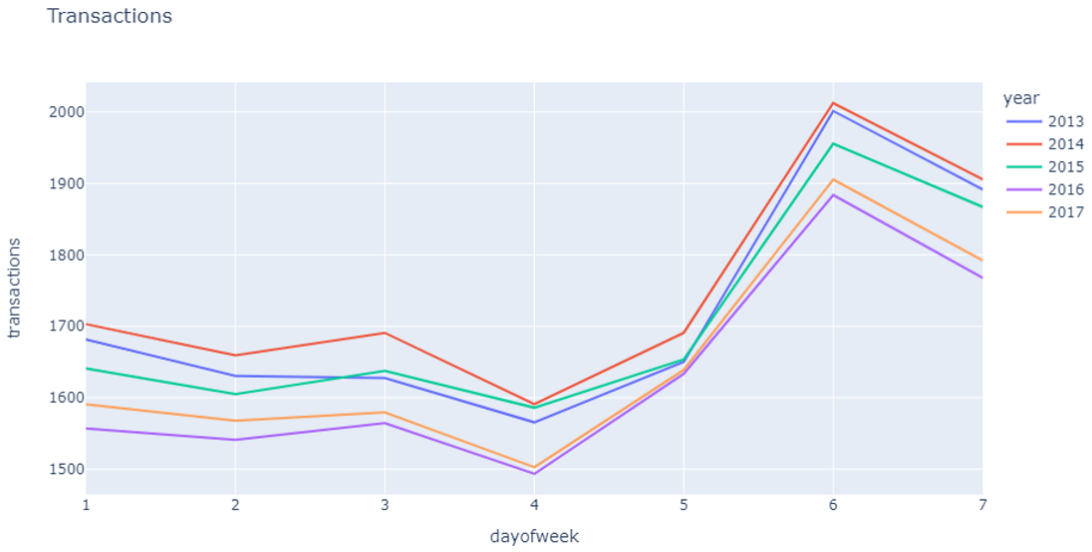

Figure 2: Most transactions are done on Saturdays (day 6). Important to have enough inventory during the weekends to accommodate for the rise in customer demand.

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.

Analysis of Product Families

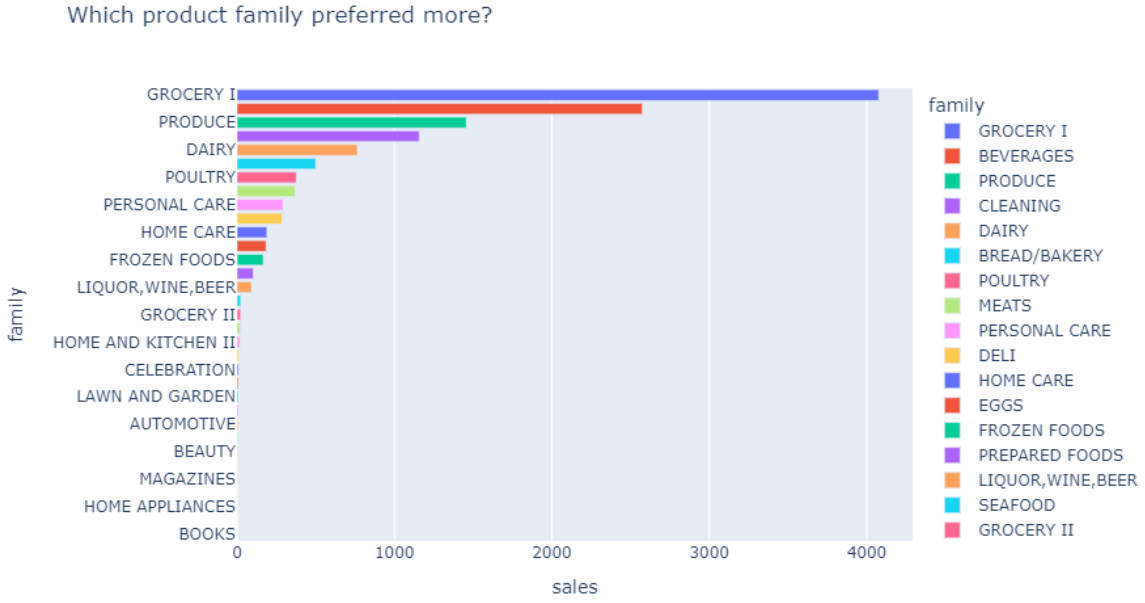

Figure 3: Calculating the average sale of each family, Grocery I, Beverages, Produce, Cleaning, and Diary are the top 5 selling product families.

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.

Analysis of Products On Promotion

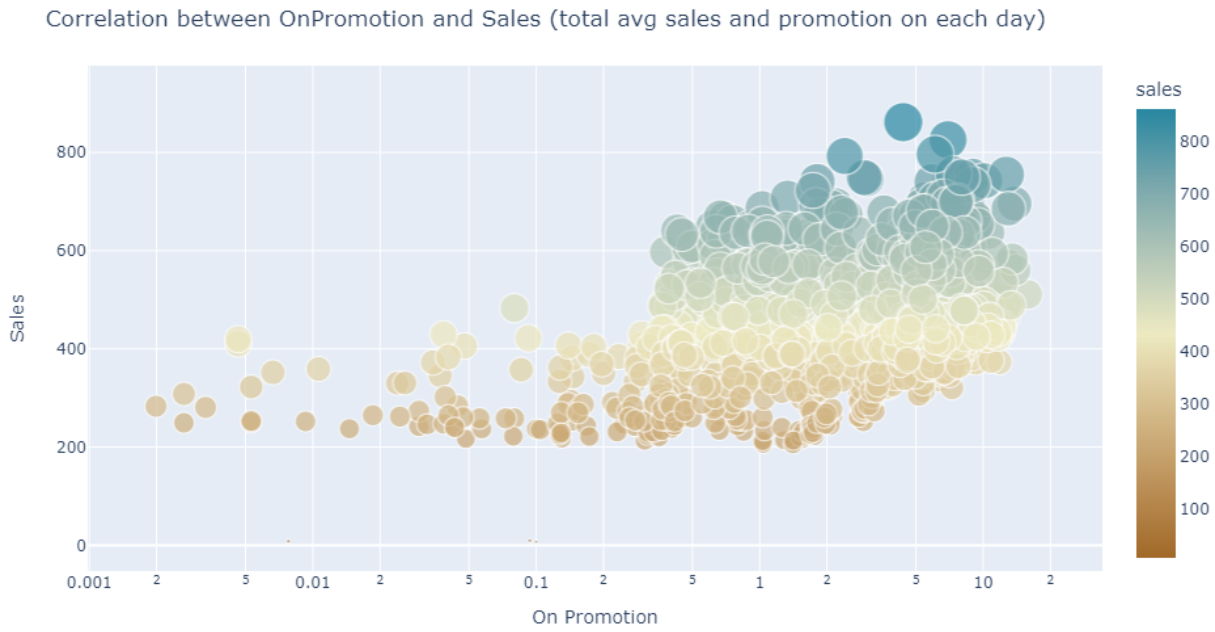

Figure 4: As expected, items on promotion will have a higher number of sales.

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 and Evaluation

Defining the Forecasting Task

To design a forecasting model, two things need to be established--the features and the target. The features refer to information readily available to create the model, and the target refers to the time period that is being forecasted. In this project, the training data contains the features needed to build the time series, with data from 2013-01-01 to 2017-08-15. The target is to forecast 15 days after the end of the training data, 2017-08-16 to 2017-08-31. The predictions made for the target can be compared to the test data to see how accurate the forecast is.

The training set ends on 2017-08-15, which gives us the forecast origin. The test set comprises the dates 2017-08-16 to 2017-08-31, and this gives us the forecast horizon. There is one step between the origin and horizon, so there is a lead time of one day. To prepare the dataset for modeling, a 16-step forecast with a 1-step lead time is needed, using lags starting with lag 1 and making the entire 16-step forecast with features from 2017-08-15.

Forecast with DirRec Strategy

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.

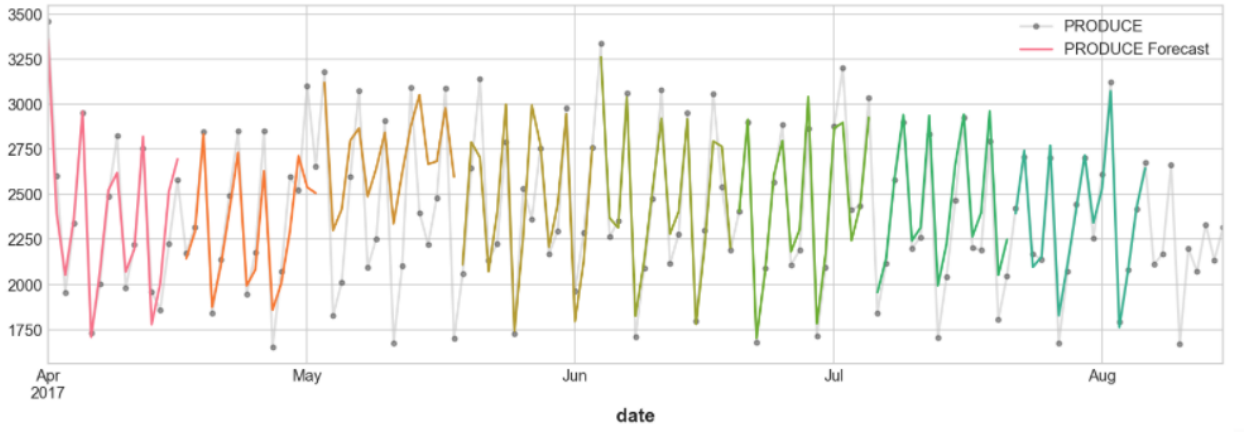

Figure 5: Time Series Forecast for the Produce product family.

The forecast and the actual values seem to overlap with each other quite nicely. Note: a forecast was made for every product family, but for brevity, only the results of the Produce product family will be shown in this report. Refer to this python notebook for further details.

Comparing the Predicted and Actual Values

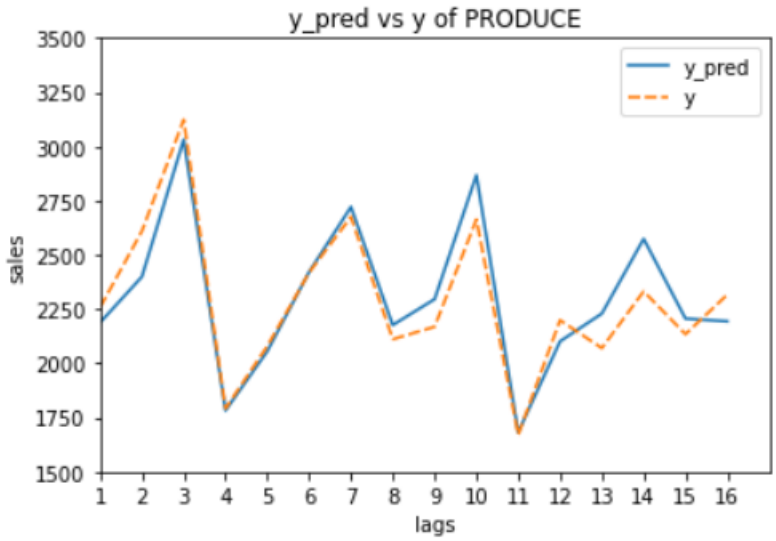

Figure 6: The forecasted (y_pred) and the actual (y) values for the lags have very similar Produce sales values, validating the accuracy of the time series.

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: Root Mean Squared Logarithmic Error (RMSLE)

The metric for evaluation is the root mean squared logarithmic error, which is given by:

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

,

where ŷ is the predicted value and y is the actual value of the target. Note that RMSLE is asymmetric and penalizes more on underestimated predictions than overestimated predictions.

```
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: Calculated RMSLE values for the 16 lags of the Produce product family.

The RMSLE for the Produce product family was calculated and showed very little error, with the highest error at lag 14 RMSLE = 0.098723), further verifying the accuracy of the time series forecast using the DirRec strategy.

Takeaways

Figure 1 shows that transactions at the Corporación Favorita stores steadily increase at the beginning of the year, then hit a nadir in summer and skyrocket towards December. This suggests that the stores may want to decrease their inventory during the summer to combat food waste, and prepare to have enough inventory towards the end of the year to accommodate for the influx of shoppers.

Figure 2 reveals that on a weekly basis, Saturdays and Sundays are when the highest numbers of transactions occur, and Figure 3 depicts the most popular product families. Curating this information, stores can get an idea of which days to expect a high number of shoppers, and which items are most in demand so that they can prepare for inventory and restocking accordingly. Figure 4 illustrates a positive correlation between products on promotion and number of sales, so if stores are looking for ways to improve the sales of a certain product, they can consider putting it on promotion.

Shown in Figure 5 is a time series forecast for the Produce product family using the DirRec strategy, in which the predicted sales numbers were very close to the actual values (Figure 6) with very low root mean squared logarithmic error (Figure 7). It is important that the time series model be as accurate as possible, as underestimating can lead to understocking of inventory resulting in customer dissatisfaction. On the contrary, overestimating the sales may lead to having a surplus of inventory and thus contribute to more food wastage.

Going Forward

Food waste amounts to a major squandering of water, land, energy, labor, and capital resources, producing greenhouse gas emissions, and contributing to catastrophic climate change. To combat this problem, a model can be built to predict the unit sales for items sold at groceries stores so that consumers can be provided with just the right amount of inventory without over or understocking. However, this comes with another challenge as current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing.

To reduce food waste at the consumer level, we can try to implement this forecasting strategy to every household where they input what they want to cook and how many servings of it. This information is then used to calculate the precise amount of groceries needed to make the dishes to prevent overbuying groceries which can lead to food spoilage and food waste. By employing a reliable forecasting model, food waste can be scaled down at both the retail and consumer level.

To improve this project, various other forecasting strategies can be explored to see if a more accurate model can be created. In addition, if the store's goal is to reduce food waste, a different evaluation metric can be applied, in which it would penalize more for making overestimations than underestimations.