Springboard – Data Science Career Track

Capstone Project 2

Grocery Store Sales – Time Series Forecasting

By Zhiling Xie

Capstone 2 Final Report

Grocery Store Sales Forecasting

1. Introduction

1.1 Problem statement

Corporación Favorita, a large Ecuadorian-based grocery retailer, would like to increase its revenue and please its customers by having just enough of the right products at the right time. Therefore, we need to forecast the sales quantity of the products in a future period.

Forecasting over would cause overstocked, perishable goods, while forecasting under would let popular items quickly sell out, leading to lost revenue and upset customers. More accurate forecasting could help the retailer ensure the supply of the right, enough products.

1.2 Goal

This project aims to develop time series forecasting models for multiple products sold at multiple Favorita stores. The predictive models are used to provide guidance for Favorita's product stocking strategy. The models with lower mean absolute percentage error would be selected for future prediction.

The implementation details can be found in the Notebooks in this GitHub repository (https://github.com/zxie9/Capstone-Project-Two).

2. Approach

2.1 Data Acquisition and Wrangling

a) Dataset

The raw data can be downloaded from the webpage:
https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data

- Sales: number of sales for each product family sold at each store during Jan. 2013 to Aug. 2017 – The **sales** column is our target for predictions.
- On-promotion: number of items being promoted.
- Store metadata: city, state, type, and cluster.
- Ecuador's daily oil price.
- Ecuador's holidays and events.

b) Data wrangling

1. Target

- The raw dataset contains sales information from 33 product families and 54 Favorita stores in Ecuador. Each product family in each store has a time series to be forecasted.
- The resolution of the time series is daily, and the forecasting horizon is 16 days.

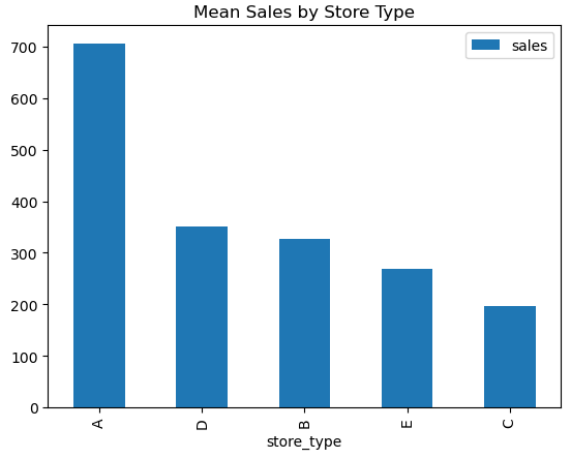
2. Exogenous variables:
- Categorical features: store's city, state, type, and cluster, whether the date is a Holiday/Weekend or not.
- Numerical Features: daily oil price, daily transactions per store, quantity of on-promotion products.

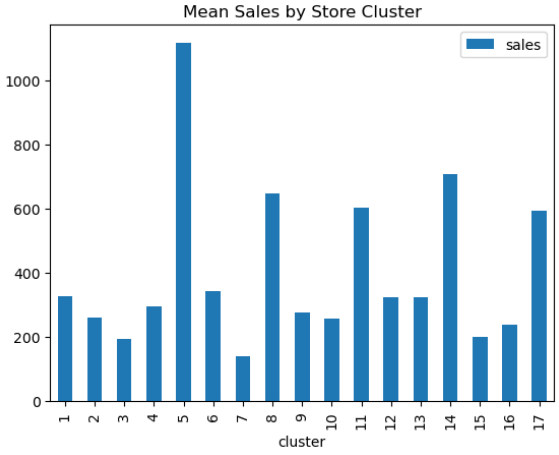
3. Missing values and time gaps
- The sales do not have missing values but have time gaps on every Christmas day.
- The daily transactions, daily oil price, and holiday columns have missing values and time gaps.

The missing values and time gaps in this dataset were filled using forward fill method.

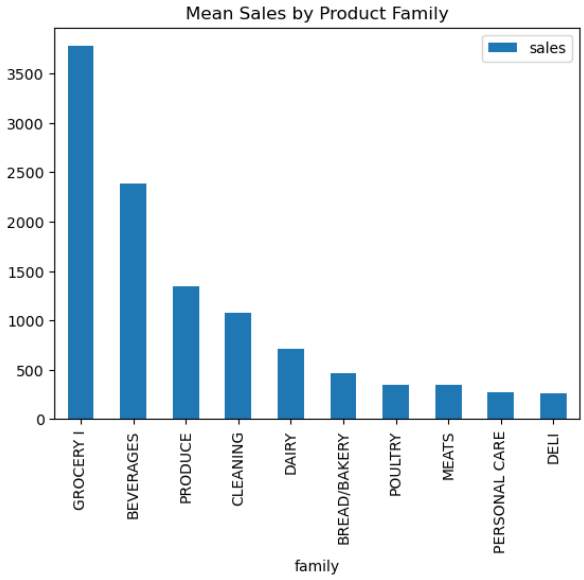
2.2 Exploratory Data Analysis

1. Mean sales by store: Type-A and Cluster-5 stores have the highest mean sales.

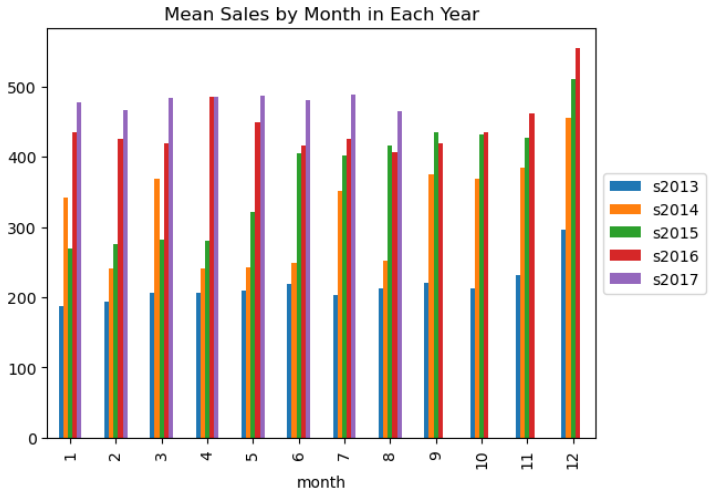




2. Mean sales by product family: Top three mean sales are from Grocery I, Beverages, and Produce product families.



3. Mean Sales by Year and Month: The sales were increased by year and the sales had an annual cycle.



2.3 Baseline Modeling

The baseline models were built for one time series forecasting. The one with store number of 44 and product family of "Grocery I" was selected for baseline modeling because it had highest mean sales.

1. ARIMA model: Used Mean Squared Error (MSE) as a criterion for choosing p, d, q values. Results showed that the best p, d, q, parameters for the ARIMA model were 0, 0, 14 respectively - an MA(14) model. The MA(14) forecasts had a Mean Absolute Percentage Error (MAPE) of 10.5%.
2. Linear Regression model: Used above mentioned exogenous variables. Besides, the lagged features were 1-day, 7-day, 14-day lagged values and the average of last 7- or 14-days' values. We found k=15 is the best k (number of used features) for this linear regression model, and the MAPE has been improved from 12.5% (k=10) to 11.0% (k=15).
3. Random Forest model: Like linear regression model, and it had an MAPE of 12.0%.

The most important features identified from both linear regression and random forest models were weekend, weekday, sales_lag1 (previous day's sales), sales_7days_avg (average sales for the previous 7 days).

2.4 Extended Modeling

The baseline models were built for multiple time series forecasting. Three product families ('BREAD/BAKERY', 'DAIRY', 'GROCERY I') in stores numbers 1 to 6 were selected, because the sales in these columns had very few zero values and the sales had relatively high values which were better for predictions.

1. pmdarima.auto_arima: Automatically selected best p, d, q values for each time series. The Best MAPE is about 7%, and the Worst MAPE is about 26%.
2. Prophet: Got reasonable forecasts on multiple time series with no manual effort. The Best MAPE is about 8%, and the Worst MAPE is about 22%.

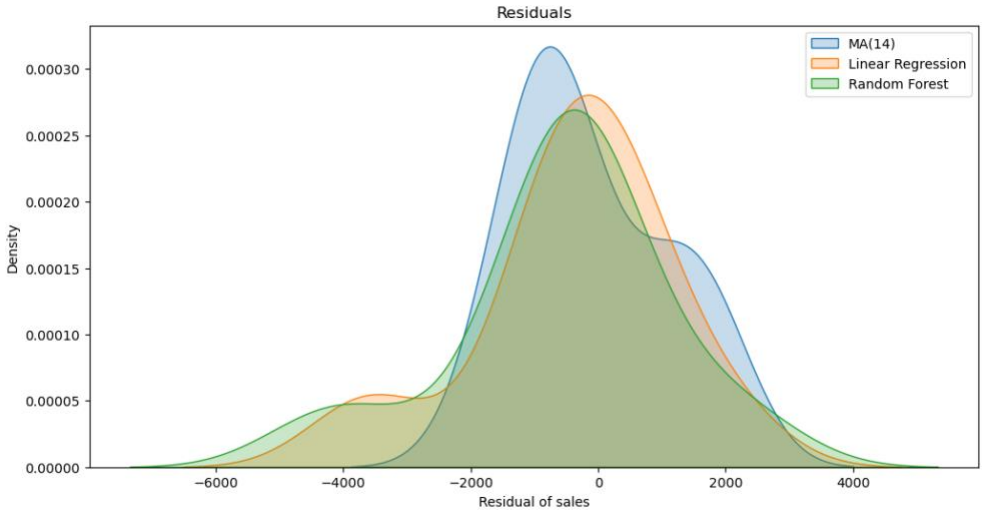
3. mljar-supervised AutoML: Applied multiple model algorithms for each time series and got an ensemble result. The Best MAPE is about 6%, and the Worst MAPE is about 21%.

3. Findings

3.1 Baseline Modeling Comparison

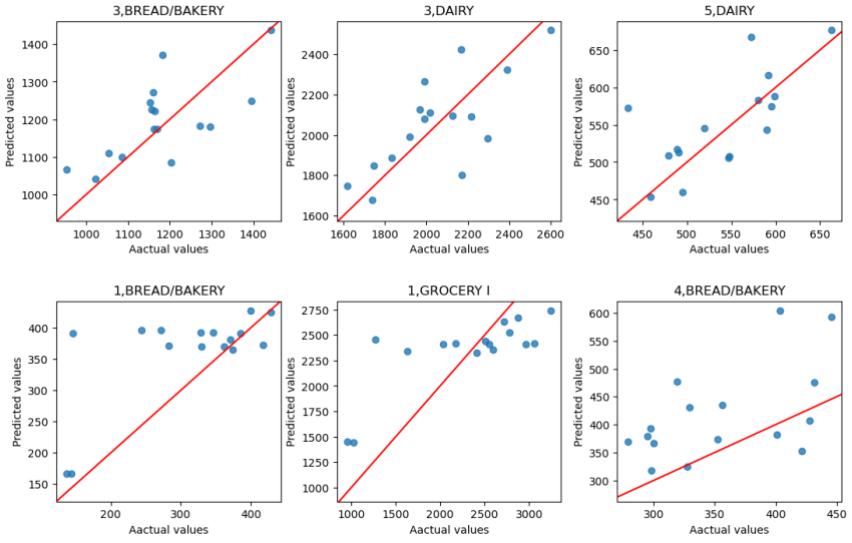
The MAPE for MA(14), linear regression, and random forest were 10.5%, 11%, and 12.0%, respectively.

From the distribution of residual plots, we can see the MA(14) model had a relatively concentrated distribution, but its center was offset from zero. The linear regression and random forest had similar distributions, and the linear regression was slightly better.

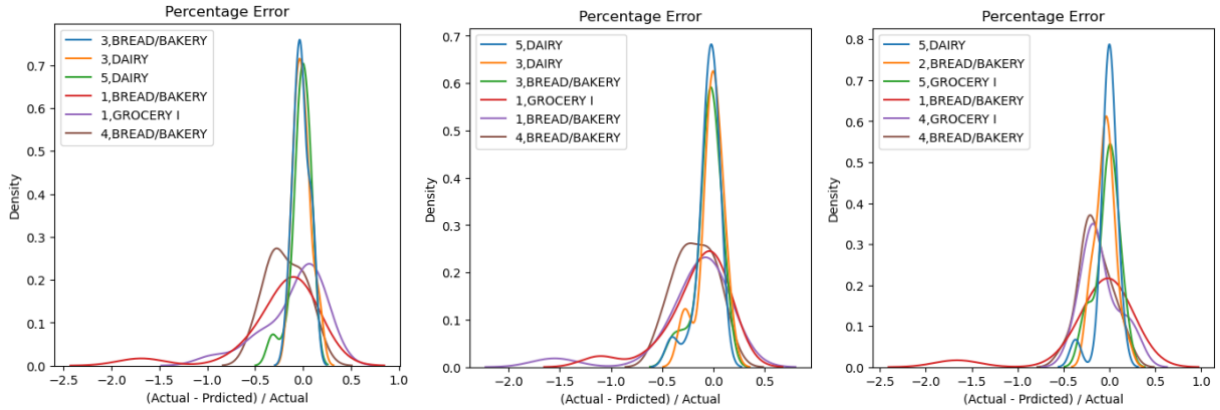


3.2 Extended Modeling Comparison

For each model, Best 3 MAPE and Worst 3 MAPE were identified. From the scatter plot of the pmdarima result, we cans see the scatter points in the upper Best 3 plots are closer to the y=x line than those in the lower Worst 3 plots. Other two models' results are similar.



From the distribution of Percentage Error plots (for three models respectively), we can clear see that the center lines of the "Best 3" are closer to zero and their distributions are also more concentrated.



4.Conclusions and Future Work

We applied several models for the sales time series forecasting and found that different models had similar results with slight differences. It means these time series forecasting models can be used for the sales predictions. Besides, we found that large sales amounts were easier to predict, and the predictions of small sales products always had large errors.

Due to some technical issues, we forecasted only 18 time series in the extended modeling section instead of the thousands provided by the original data. The next step is to compute the forecast of thousands of time series.

We may need larger computing sources for large amount of data (such as High-performance computing). We can use hierarchical forecasting approach. We need a separate forecast for each separate time series, but there are similarities between products that might help with the forecasting. Grouping the product together along a product hierarchy and then using hierarchical forecasting to improve accuracy.

Additionally, we can further conduct forecasting in a Cloud-based environment that would allow us to simulate a realistic scenario for which new data is regularly ingested. The models are computed regularly (e.g., weekly), and results are also stored in the Cloud.

5. Consulted Resources

[1] Forecasting: Principles and Practice (3rd Ed) – Chapter 9. *Rob J Hyndman and George Athanasopoulos*. https://otexts.com/fpp3/

[2] ARIMA:
https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima.model.ARIMA.html

[3] pmdarima:
https://pypi.org/project/pmdarima/
https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html

[4] Prophet:
https://facebook.github.io/prophet/

[5] AutoML mljar-supervised:
https://github.com/mljar/mljar-supervised
https://supervised.mljar.com/