Forecasting Grocery Store Product Sales

Austin Lackey, Tomy Sabalo Farias, and Sam Herold

DSCI 478, Colorado State University

February 20, 2024

Abstract

This paper presents a methodology for forecasting grocery store sales across many stores and product subsets. The data used in this project is from Favorita Grocery Stores in Ecuador from 2013 to 2017 and is provided by Kaggle. A link to the data can be found in the references section at the end of this paper. Our methodology is split into three main sections: data preprocessing, model engineering, and model evaluation. In this project, we use a variety of models including linear and polynomial regression and gradient boosting.

1 Introduction

1.1 Background

The 'Store Sales' time series forecasting dataset contains sales data from Favorita Grocery Stores in Ecuador from 2013 to 2017[1]. Sales data is broken down by store and product category for each day (1,782 pair-wise combinations = $stores \times products$).

1.2 Problem Statement

The goal of this project is to forecast sales for each store and product category for the next day while reducing $RMSE$. $RMSE$ is defined as the square root of the average of the squared differences between the forecasted and actual sales.

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

Variable	Type	Unique Counts	Description
date	datetime64	1684	Date-stamp of the sales data
store_nbr	int64	54	Store number
family	categorical	33	Product family (Grocery, Beverages, Deli, etc.)
city	categorical	22	City of the store
state	categorical	16	State of the store
type	categorical	5	Type of store (A, B, C, D, E)
cluster	int64	17	Cluster of the store (Similar stores grouped together)

Table 1: Variable Descriptions

For each day, we are making 1,782 pair-wise predictions of store and product category sales. This is a challenging problem since sales are influenced by more than just the seasonality. Below is an example that shows the sales over the course of a three-month period for only the top 5 product categories for the store with the most sales.

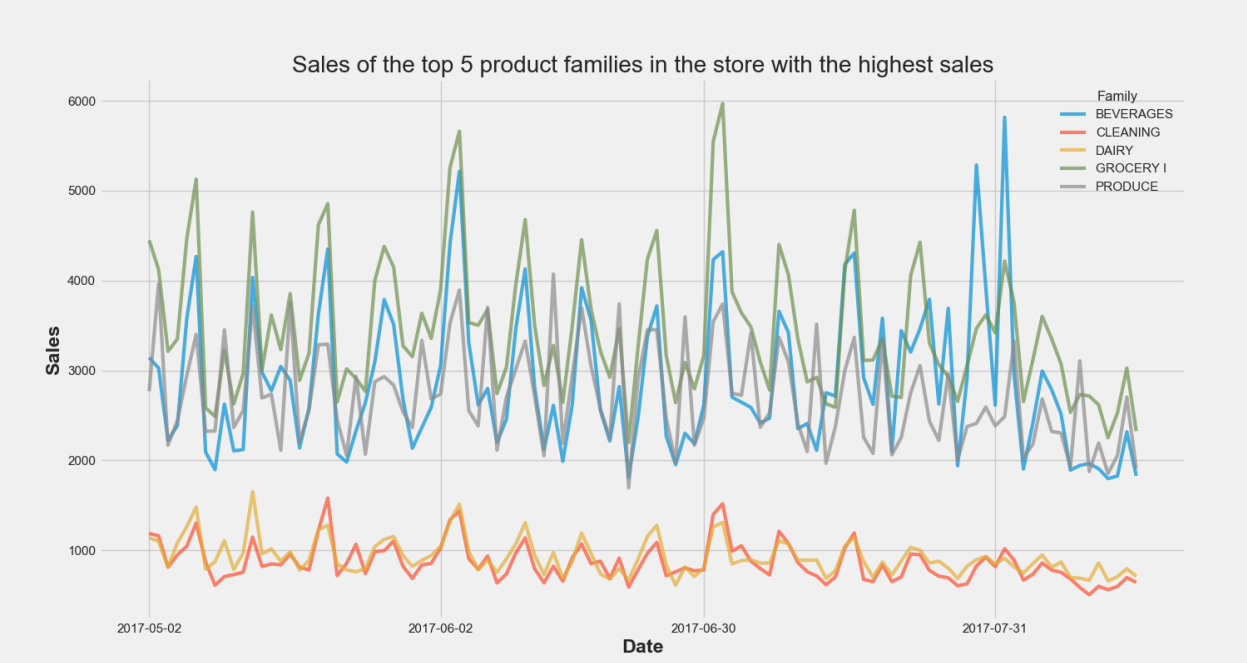

Figure 1: Top 5 Product Categories for Store 44

Fix plot and expand on it later...

Date	Store Number	Product Family
2013-01-01	1	Grocery
2013-01-01	1	Beverages
2013-01-01	1	Deli
⋮	⋮	⋮
2013-01-01	54	Grocery
2013-01-01	54	Beverages
2013-01-01	54	Deli
⋮	⋮	⋮
2013-01-02	1	Grocery
⋮	⋮	⋮

Table 2: Example of Pair-Wise Combinations

2 Solution

2.1 Feature Engineering

The biggest challenge in this project was to create a feature set that would allow us to forecast sales for each store and product category. We found that the most important part in reducing $RMSE$ was to create a feature set that captured as much information regarding seasonality and trends as possible. We created a feature set that included the following extra variables:

- **Datetime Features:** We created features that captured what day of the week it was, the day of the month, the month, the quarter, the week of the year and the year.
- **Lag Features:** We implemented a 7-day lag feature to capture the previous week's sales that led up to the current day. This feature really helped the model fine-tune the forecast.
- **Store-Specific Features:** We joined the store data with the sales data to give the model more information about the store.

Expand more here...

2.2 Model Engineering

Add intro here...

References

[1] Favorita Grocery Store Sales - Time Series Forecasting. Kaggle. Retrieved from `https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data`

[2] XGBoost Documentation. dmlc. Retrieved from `https://xgboost.readthedocs.io/en/stable/python/python_api.html`

[3] Insert more here...