STORE SALES

Times Series Forecasting

Dongyu Fan
Maksim Kosmakov
Fernando Liu Lopez
Xin Su
Wenchuan Tian

TABLE OF CONTENTS

01

Project Overview

PROJECT DESCRIPTION

OBJECTIVES

- Develop predictive models to forecast sales
- Find trends in sales data
- Incorporate external data beyond sales
- Address real-world challenges in sales forecasting

- **Bonus**: Do well in Kaggle competition!

02

DATA Description

Datasets

SALES

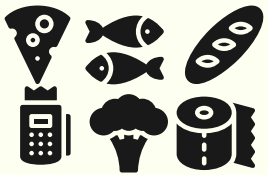
META DATA

OIL

HOLIDAY
S

STORES

FAMILIE
S



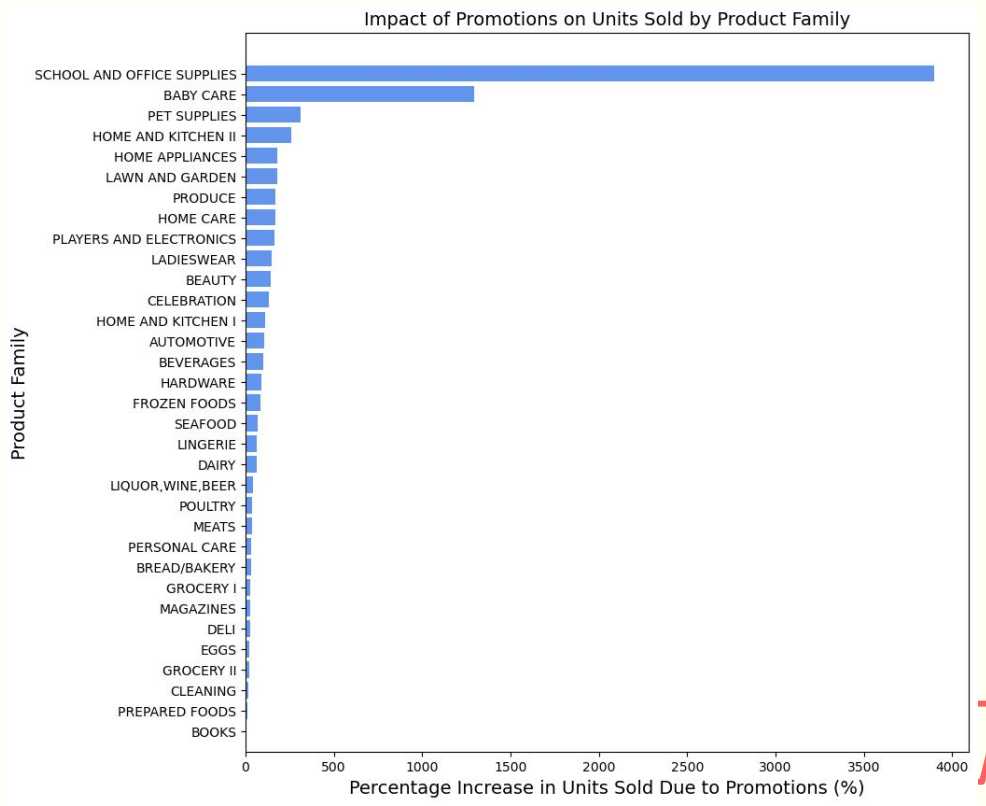
03

DATA Analysis

Impact of Promotions, Oil Prices, and Holidays on Sales

- **Largest Impact (3898.93%)**: *School & Office Supplies* – Sales jumped from 1.11 to 44.54 units with promotions.
- **Smallest Impact (11.5%)**: *Prepared Foods* – Sales increased modestly from 102.72 to 114.54 units with promotions.
- *Books*: No promotions recorded.

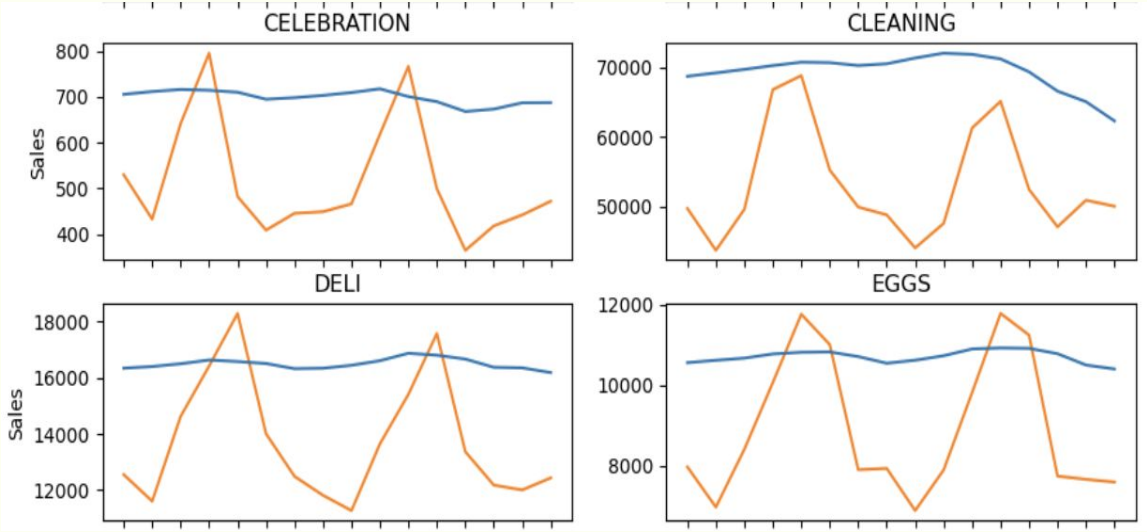
We analyzed the correlations between **oil prices**, **holidays/events**, and **sales**. While the impact is **less pronounced**, it **varies across families and stores**. Among holidays, **national holidays** have the **most significant effect**.

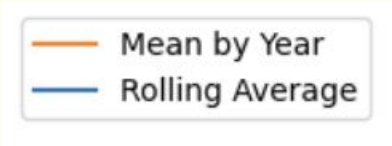


04

Model Selection

Baseline Models: Mean by Year, Rolling Average





Observation: The rolling average tends to overestimate the actual values.

Overestimation might lead to excess inventory, which is often less harmful than stockouts (lost sales!)

Scoring Metric: RMSLE (Root mean squared log error)
Mean by year score: 0.902
Rolling average score: 0.461

Monthly rolling average is a good option in terms of RMSLE.

RMSLE penalizes underestimation of actual values more heavily than overestimation.

Root Mean Squared Log Error(RMSLE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

Root Mean Square Error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

X is the predicted value
Y is the actual value

Example:

Overestimated Value =1400, RMSE: 400, RMSLE: 0.33

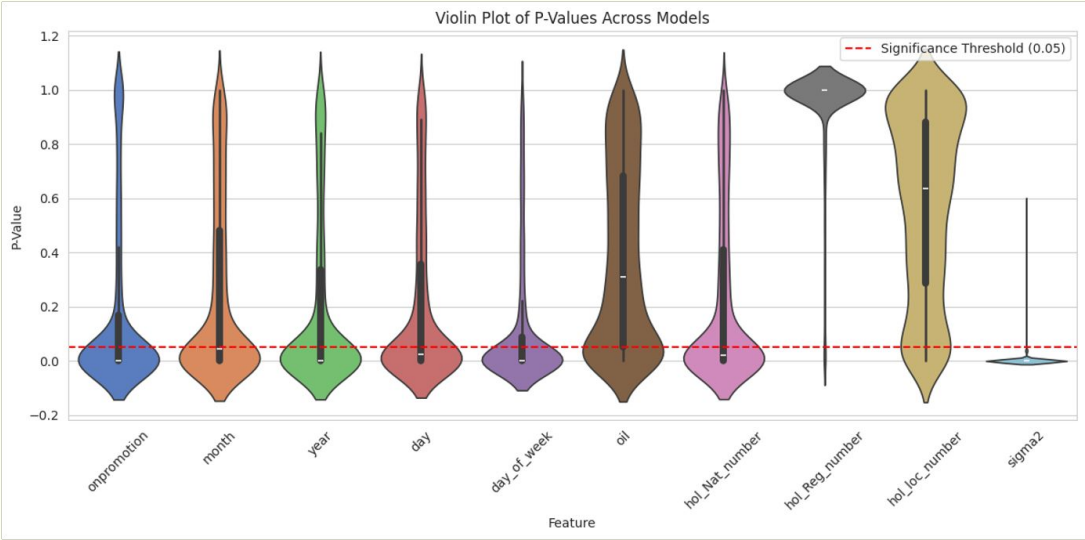
Actual Value =1000

Underestimated Value =600,RMSE: 400, RMSLE: 0.51

Advanced models: approach and results

- **Feature Selection:** We focused on the most important features for each model, reducing training time and improving efficiency.

- **Results:**
  - **Random Forest regression:** 0.514
  - **Prophet Model:** 0.484
  - **SARIMAX:** 0.476
- **Combined model:**
  - ⅓ Rolling average + ⅓ Prophet + ⅓ SARIMAX : 0.429



05

Summary

+Future Directions

Data Preprocessing

• Cleaned 2.7M rows, merged holidays, interpolated oil prices

Key Insights

• Promotions: Significant sales boost, varying by product family and store
• Oil Prices: Weak correlation with sales
• Holidays: National holidays have the strongest impact

Modeling Approach

• Baseline: Rolling Average performed well, hard to beat
• Advanced: Random Forest, Prophet, SARIMAX
• Final RMSLE: 0.42913 (Top 100 on Kaggle)

Model Improvement

• Experiment with Deep Learning for weight optimization
• Explore Long Short-Term Memory (LSTM) for better sequence forecasting