# Corporación Favorita Grocery Sales Forecasting

Guayas province, January to March 2014

# Objective

Customer demand in Guayas changes quickly, influenced by promotions, seasonality, and local events.

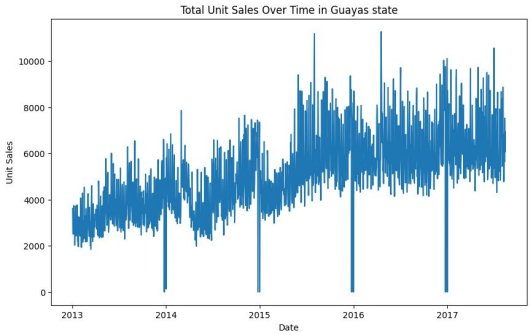
Our goal is to model these patterns and forecast daily sales for every product in every store for the first quarter of 2014, giving planners a reliable basis for operational decisions.

# Understanding the data and defining the approach

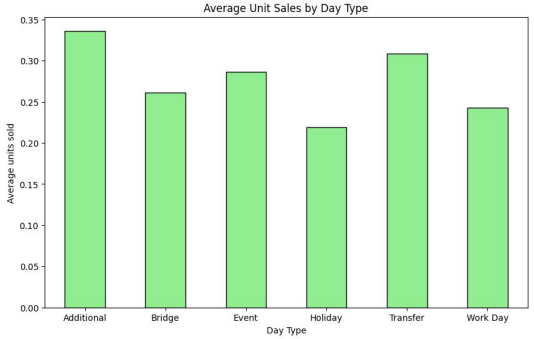
- We worked with several complementary datasets that describe how people shop across Ecuador
- Daily sales show what each store sold, while store and item metadata explain who sells what
- Transactions reveal footfall, holidays shift buying behavior, and oil prices capture the broader economic mood
- We started with an exploratory analysis based on the Pichincha example and adapted it to Guayas
- To make the data more manageable, we focused on the three largest product families; we avoided row sampling because it reduced precision noticeably
- We then tested SARIMAX, XGBoost, and Prophet

# Exploratory Data Analysis

# Insights


Total Unit Sales Over Time in Guayas state
Unit Sales
0
2000
4000
6000
8000
10000
2013
2014
2015
2016
2017
Date


Monthly Sales Trends Over Years
Year
2013
2014
2015
2016
2017
1
2
3
4
5
6
7
8
9
10
11
12
Month
Sales Volume
220000
200000
180000
160000
140000
120000
100000
80000


Average Unit Sales by Day Type
Average units sold
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Additional
Bridge
Event
Holiday
Transfer
Work Day
Day Type

# Prediction Models

# Models we tested

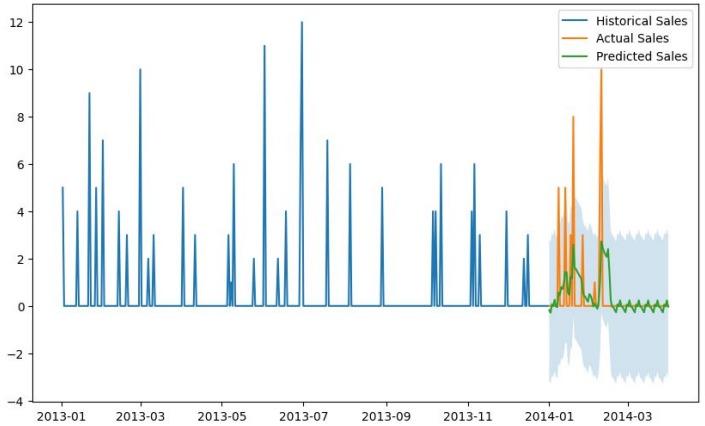
1. Started with SARIMAX on a single store–item pair as a proof of concept, then tested an aggregated version
2. Used XGBoost as the main model, with hyperparameter tuning to improve performance
3. Added Prophet as a simple baseline to compare against the more advanced methods

# How we evaluated the models

- Wee used RMSE, MSE, MAE,  and $R^2$, which are standard metrics for retail demand forecasting
- **RMSE** and **MSE** capture large errors that often occur during promotional spikes
- **MAE** provides an intuitive measure of the average daily deviation from actual sales
- **$R^2$** helps assess how much of the overall variability in demand the model can explain
- We avoided **MAPE**, which is commonly unreliable in retail because sales can be low or zero, causing distorted percentages

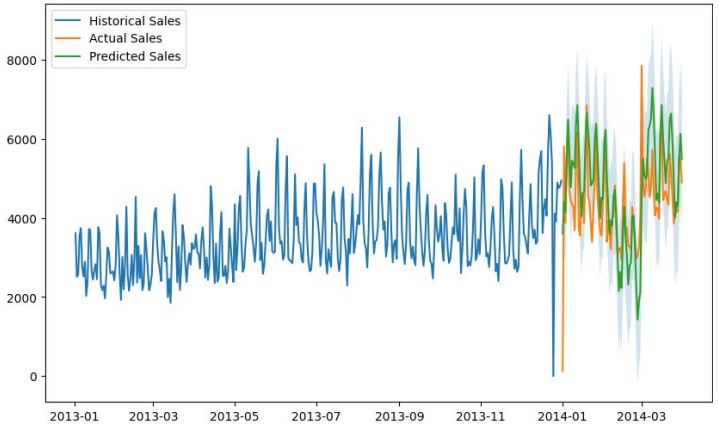
# SARIMAX

RMSE: 1.53
MSE: 2.34
MAE: 0.98
$R^2$: 0.16



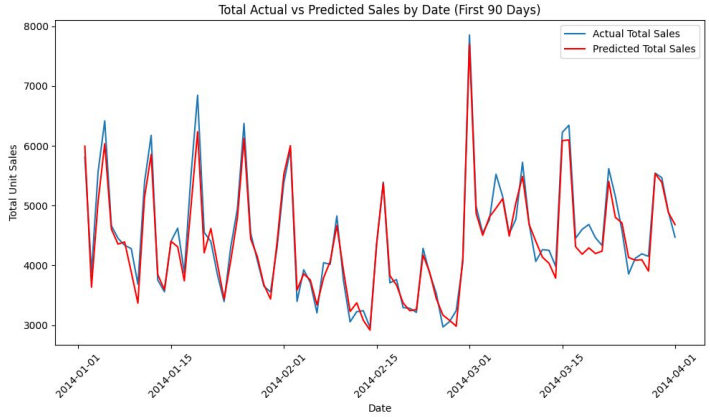
# (Aggregated) SARIMAX

RMSE: 1054.01
MSE: 1110933.65
MAE: 817.76
$R^2$: 0.01



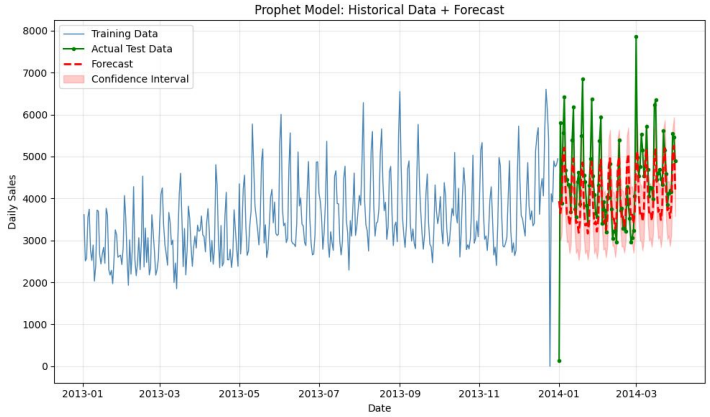
# XGBoost

RMSE: 1.33
MSE: 1.78
MAE: 0.10
$R^2$: 0.75



# Prophet

RMSE: 904.31
MSE: 817772.78
MAE: 673.12
$R^2$: 0.27



# Comparing model performance

| Model | RMSE | MSE | MAE | $R^2$ |
|---|---|---|---|---|
| SARIMAX | 1.53 | 2.34 | 0.98 | 0.16 |
| SARIMAX (agg.) | 1054.01 | 1110933.65 | 817.76 | 0.01 |
| XGBoost | 1.33 | 1.78 | 0.10 | 0.75 |
| Prophet | 904.31 | 817772.78 | 673.12 | 0.27 |

# Conclusions

# Insights, challenges, and next steps

- Preparing the data was the most demanding part of the project: every preprocessing choice had a noticeable effect on model performance
- Even small EDA decisions (like sampling) had a surprisingly large impact on accuracy, which shows how sensitive ML models are to upstream data work
- XGBoost emerged as the most robust model once the data was correctly engineered, showing consistent gains over the classical baselines
- A natural next step would be to expand the feature set beyond the three largest product families and evaluate whether model gains generalize across the full assortment