Store Sales Forecasting Project

Welcome to the Store Sales Forecasting Project report. This presentation outlines our approach, findings, and the robust predictive model developed to optimize sales strategies.

NAME- Rajdeep Chakraborty

Project Overview

Unpacking the Challenge

Our primary objective was to build a predictive framework for weekly sales across stores. By leveraging historical data and store-specific details, we aimed to:

- Identify crucial factors influencing sales.
- Develop robust regression models.
- Optimize inventory, resource allocation, and promotional strategies.
- Maximize revenue and operational efficiency.

Data Foundation

The Datasets: A Closer Look

The project utilized three core datasets, each providing unique insights into store operations and sales performance.

Store_Details

- 45 rows, 6 columns
- Unique store attributes (ID, Type, Address, Area Code, Location, Size).

Business_Data

- 8,190 rows, 15 columns
- Weekly records: Temperature, Fuel Price, MarkDowns 1-5, CPI, Unemployment Rate, Holiday status, and engineered time features.

Sales_History

- 421,570 rows, 8 columns
- Transactional data: Store, Department, Date, Total Sales, Holiday, Year, Month, Weekday. Advanced features like lags and rolling means were engineered.

Data Preprocessing

Addressing Missing Values & Outliers

Ensuring data quality was paramount. We meticulously handled missing values and outliers to prevent model bias.

Missing Values

- **Store_Details:** No missing data.
- **Business_Data:**
  - **MarkDowns (1-5):** ~50% missing, imputed with zero (assuming no promotion).
  - **CPI & Unemployment Rate:** ~7% missing, filled with median values.
- **Sales_History:** No missing values post-cleaning.

Outliers

- Identified in Sales, Temperature, Fuel Price, CPI, and Unemployment Rate.
- **Method:** IQR (Interquartile Range) methodology used for capping values outside [Q1-1.5_IQR, Q3+1.5_IQR].
- **Sales Outliers:** Extreme sales during holidays/promotions were capped, not removed, to preserve valuable event data.

Store_Details	None	N/A	Not required
Business_Data	MarkDowns, CPI, Unemployment	Zero (MarkDown), Median (others)	IQR capping
Sales_History	None	N/A	IQR capping for sales

Exploratory Data Analysis

Key Insights from Descriptive Statistics

Understanding the raw data's characteristics provided a foundation for feature engineering and model selection.

Store_Details

- Majority store type: E-Commerce Fulfillment (22/45).
- Store sizes: 34,875 to 219,622 sq ft; median ~126,512 sq ft.

Business_Data

- Temperature: Mean 59.4°F (SD 18.7), Min ~4, Max ~102.
- MarkDowns: Mostly zero (no promotions).
- CPI: Mean ~173, range 126 to 229.
- Unemployment Rate: Mean 7.72%, range 4.3–11.0%.

Sales_History

- Total Sales: Mean ~$13,649 (SD ~$14,909), Min ~-$4,989, Max ~$47,395.
- Departments: 81 unique. Stores: 45.
- Most data points are non-holiday weeks.

Visual Discoveries

Sales Patterns & Relationships

Visualizations helped us uncover critical sales dynamics and variable correlations.

- **Total Sales Distribution:** Skewed right, indicating many low/medium sales weeks with sharp spikes (likely holidays/promotions).
- **Holiday Effect:** Sales during holiday weeks are significantly higher (statistically confirmed).
- **Correlation Heatmap:**
  - Sales moderately correlated with **store size** (0.26).
  - Weakly correlated with MarkDowns, CPI, and Unemployment Rate.
  - High correlation between MarkDown1 and MarkDown4 (0.84).
- **Scatter Plots:** Larger stores consistently show higher total sales, while economic/MarkDown variables show weak direct correlation with sales.

Feature Importance

The Predictive Power of Variables

Information Value (IV) analysis highlighted which variables are most crucial for predicting sales.

1

Holiday

HIGH IV (>0.4)

Significant impact: Holidays **increase** sales.

2

Store Size

MODERATE IV (0.25–0.3)

Bigger stores consistently **sell more**.

3

MarkDowns

LOW IV (<0.1 individually)

Only large, infrequent discounts show notable impact.

4

Economic Variables

LOW IV

(CPI, Unemployment): Effects present but **weak** compared to holiday/promotions.

Model Development

Benchmarking Regression Models

We evaluated several regression models, assessing their performance using RMSE (Root Mean Squared Error) and R^2 (Coefficient of Determination).

Linear Regression	~0.45	~8,900	Baseline, interpretable
Random Forest	~0.62	~7,550	Robust, captures non-linearities
Gradient Boosting	~0.64	~7,350	Slightly better, slower
XGBoost	~0.66	~7,200	**Best performing**, flexible

Feature importance from tree models consistently highlighted Holiday, Store Size, and certain MarkDowns as top contributors.

Final Selection

The Chosen Model: XGBoost Regression

After thorough evaluation, XGBoost emerged as the optimal choice for our sales forecasting framework.

Why XGBoost?

- **Superior Performance:** Achieved the highest test R^2 and lowest RMSE on validation data.
- **Robustness:** Handles interactions and non-linear relationships better than other models.
- **Flexibility:** Adapts well to feature scaling, collinearity, and works effectively with engineered features.

Summary & Outlook

Conclusion & Future Opportunities

This project successfully delivered a robust sales prediction model, offering valuable insights for Store's strategic planning.

Key Achievements

- Integrated diverse datasets for comprehensive analysis.
- Quantitatively modeled holiday and promotional effects.
- Developed a high-accuracy XGBoost model.
- Provided actionable insights for optimizing inventory, resources, and promotions.

Challenges Overcome

- Ensured proper data merging without leakage.
- Effectively managed significant missing data in MarkDowns.
- Handled computationally intensive feature engineering.
- Judiciously treated outliers to preserve event-driven sales spikes.
- Optimized hyperparameter tuning for boosting models.

THE END