Assignment 2: ML Pipeline Project

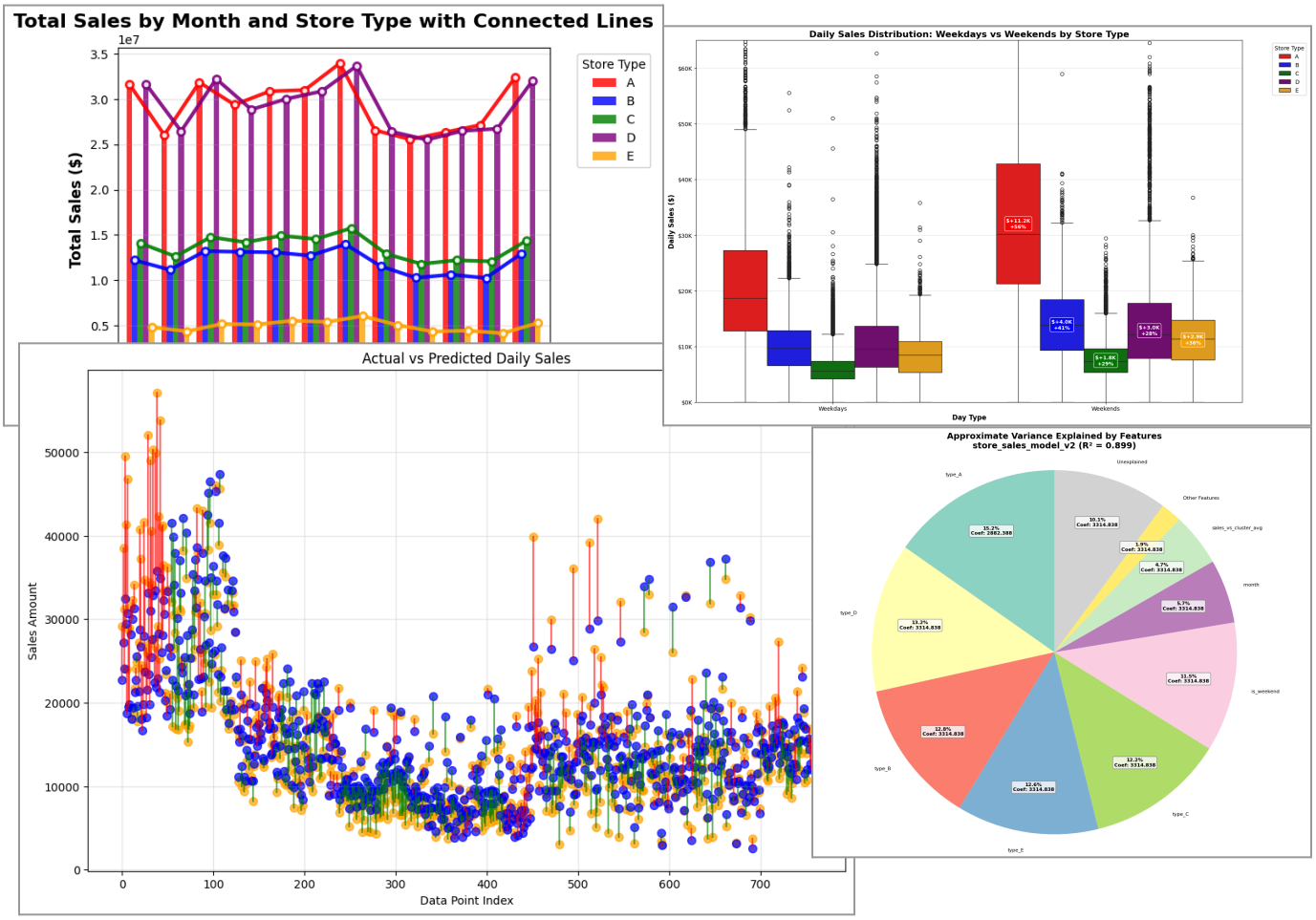
From Kaggle to Predictions with BigQuery ML

Group: DN3

Members: James Fazecas, Kyasha Williams, Yuxiao Wang, Michael Yudanin

Store Sales Prediction Analysis

Integrated DIVE Report



DIVE Journey

D - Discovery: Surface Discovery

Our comprehensive analysis across both aggregated and store-level sales prediction models reveals strong foundational insights. The time series forecasting model for total daily sales demonstrates high accuracy with 95% prediction reliability, generating forecasts ranging from 620K to 1.3M daily across a two-week window. The store-level regression model explains 83% of sales variation with predictions averaging within 19% of actual sales.

Key Surface-Level Patterns:

- **Weekly Patterns**: Consistent patterns show midweek dips and weekend spikes across all store types
- **Store Type Hierarchy**: Type A stores drive the highest sales but exhibit greatest volatility, while Types C and D show more predictable but lower performance
- **Seasonal Cycles**: Clear yearly patterns emerge with peaks in July, December, January, and March, and notable lows in February, August, September, October, and November
- **Cluster Effects**: Most store clusters show homogeneous performance within type categories, except for Type A clusters and specific outliers like Cluster 12

I - Investigate: Deeper Investigation Findings

Store Type Impact Analysis: Type A stores function as "anchor stores" that drive ecosystem-wide traffic, evidenced by their 56% weekend sales bump ($11.2K average increase) compared to Type D's 28% increase ($3K average). This suggests mission-based shopping behavior where Type A stores serve leisure/family shopping needs while smaller stores handle convenience purchases.

Geographical and Operational Drivers: Urban stores demonstrate more stable trends while rural or clustered locations experience greater fluctuations. Store performance relative to cluster average emerges as a critical predictor.

Cross-Store Dynamics: Evidence suggests cannibalization networks where specific store combinations create predictable sales transfers, rather than uniform competition across all locations.

V - Validate: Model Limitations and Risks

External Factor Vulnerability: Both models fail to incorporate critical external influences including weather events, economic conditions, competitive actions, and local disruptions. This creates significant blind spots.

Overfitting Concerns: The store-level model shows worse predictive performance than training performance, indicating overfitting issues. The linear regression approach may miss crucial interaction effects between variables like store type and weekend patterns.

Temporal Assumption Risks: Models assume historical patterns will continue unchanged into the future. Fundamental shifts in consumer behavior, market dynamics, or operational changes could invalidate core assumptions.

Aggregation Limitations: The total sales model may mask important product-category or store-specific underperformance, while the store model may miss broader market trends affecting multiple locations simultaneously.

E - Extend: Strategic Recommendations

Immediate Operational Adjustments: Implement dynamic staffing and inventory management based on predicted weekly patterns, with particular attention to weekend demand spikes. Type A stores should focus on experiential weekend promotions while Types C and D should optimize for spillover traffic capture.

Enhanced Predictive Infrastructure: Develop hybrid modeling approaches that combine time series forecasting with real-time external factor monitoring. Integrate weather, local event, and economic indicator feeds to improve prediction accuracy during disruption periods.

Store-Specific Strategy Development: Create differentiated approaches recognizing that Type A stores require volatility management and ecosystem thinking, while smaller stores need efficiency optimization and specialization strategies.

Cross-Store Coordination: Implement cluster-wide promotional calendars that leverage the anchor store effect, with Type A stores driving traffic that benefits the entire shopping ecosystem.

Action Plan

Action 1: Predictive Staffing and Inventory System Actions

Specific actions:

- Weekly staffing adjustments (25-40% weekend increase Type A, 15-25% others)
- Dynamic inventory buffers (15% Type A, 10% stable stores)

Metrics: 2-3% labor efficiency improvement, <2% weekend stockouts, >85% satisfaction

Mitigation: 24-hour flexible scheduling, 48-hour emergency inventory, cross-training programs

Action 2: Cross-Store Promotional Calendar Actions

Specific actions:

- Type A anchor campaigns aligned with seasonal peaks
- Cluster-wide themes
- Mid-week promotional testing

Metrics: 8-12% promotional lift, 3-5% nearby store impact, 15-20% traffic increase

Mitigation: 20% budget flexibility, rapid cancellation protocols, backup strategies

Action 3: External Factor Monitoring System Actions

Specific actions:

- Weekly assessments (weather, events, economic indicators)
- Standardized impact ratings

Metrics: 10-15% prediction error reduction during disruptions, 24-48 hour response time, 90% deviation explanations

Mitigation: Multiple information sources, escalation procedures, 4-hour communication protocols

Overall Targets: 5-8% sales accuracy improvement, 10-15% operational efficiency gains, days-to-hours response reduction

Appendix A: Cost and Price Optimization

We can leverage the insights from this analysis to optimize both supplier costs and product pricing:

1. Optimizing Costs Paid to Suppliers:

- **Leverage Sales Predictions for Negotiation:** Armed with accurate sales forecasts, you can negotiate better deals with suppliers. By demonstrating a clear understanding of future demand, we can try and secure:
    - Bulk discounts
    - Favorable payment terms
    - Guaranteed supply at lower prices
- **Consolidated Purchasing:**
    - Explore opportunities for consolidated purchasing across stores, especially within the same cluster or store type. This increases your bargaining power with suppliers and can lead to significant cost savings.
- **Inventory Optimization:** The model's insights into sales trends and seasonality can help optimize inventory levels:
    - Minimize holding costs by reducing excess inventory and prevent stockouts by ensuring sufficient supply of high-demand products. This reduces waste and storage costs.
    - The model enables us to do it per store type.
- **Supply Chain Analysis:** Conduct a thorough analysis of your supply chain to identify inefficiencies and areas for cost reduction. This might involve
    - streamlining logistics
    - optimizing transportation routes
    - exploring alternative suppliers.
    - considering shifting inventory from one store to another, possible within a cluster, based on predicted needs.
- **Explore Alternative Products:**
    - If certain products have high supplier costs and low profitability, explore alternative products from different suppliers that offer similar value at a lower cost.
    - Consider private label or store brand options.
- **Negotiate Rebates and Discounts:** Actively negotiate rebates and discounts with suppliers based on volume purchases or achieving specific sales targets. The numbers the models can show, e.g., for seasonality, can be a valuable input for these negotiations.

- **Just-in-Time Inventory:** For products with predictable demand, consider implementing a just-in-time inventory system to minimize storage costs and reduce the risk of obsolescence.

2. Optimizing Prices for Products in Stores:

- **Price Elasticity Analysis:** Conduct price elasticity analysis to understand how changes in price affect demand for different products. This allows you to optimize prices to maximize revenue and profitability.
- **Dynamic Pricing:** Leverage the model's sales predictions to implement dynamic pricing strategies. Adjust prices based on real-time demand, day of the week, time of day, and other factors. This is particularly effective for products with fluctuating demand.
- **Price Bundling:** Offer product bundles or package deals to encourage larger purchases and increase average transaction value.
- **Premium Pricing for High-Demand Products:** For products with high demand and limited supply, consider premium pricing to capture additional value.
- **Competitive Pricing:** Monitor competitor pricing closely and adjust your prices accordingly. The model's insights into cluster performance and sales_vs_cluster_avg can help inform competitive pricing strategies.

Integrating Cost and Pricing Optimization:

- **Profit Margin Analysis:** Regularly analyze profit margins by product, category, store type, and cluster to identify areas for improvement.
- **Scenario Planning:** Use the model to conduct scenario planning and evaluate the impact of different pricing and cost scenarios on profitability.
- **Data-Driven Decision Making:** Base all pricing and cost decisions on data and analysis. Track the results of any changes you make and iterate based on the outcomes.
- **Cross-Functional Collaboration:** Ensure close collaboration between purchasing, marketing, and store operations teams to align pricing and cost optimization strategies.

Appendix B: Team Contributions

- **James** prepared the training data for Option A, which focused on forecasting total daily sales across all stores. They wrote a SQL query to aggregate historical sales by date, calculating total sales, the number of active stores, and the average number of promotional items. To ensure reliable model testing, they excluded the last two weeks of data.
- **Kyasha** developed the time series model using BigQuery ML's ARIMA_PLUS for Option A. They trained the model on the aggregated daily sales data, evaluated its performance using built-in metrics, and generated a 14-day forecast. Their work provided insight into overall sales trends and future expectations.
- **Michael** prepared the store-level dataset for Option B, which aimed to predict daily sales for individual stores. They joined sales data with store information to create a detailed table that included store type and cluster attributes. This enriched dataset enabled more precise modeling at the store level.
- **Michael** built and evaluated the regression model for Option B using BigQuery ML's linear regression. They selected relevant features such as day of the week, store type, and cluster to train the model. After assessing its performance, they used it to generate predictions on test data, helping to identify store-specific sales patterns.