Utilizing Predictive Modeling for Retail Sales and Inventory Optimization

Purdue University, Daniels School of Business

Jose Augusto Heighes jheighes@purdue.edu | **Marcus Page** page108@purdue.edu

Swati Rajasekaran rajasek0@purdue.edu | **Colin Wellington** cwellin@purdue.edu

Mentor - Xing Wang wang5719@purdue.edu

BUSINESS PROBLEM

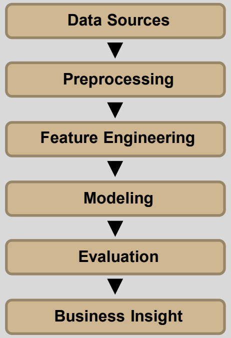
Business Problem: This project is a data analytics consulting initiative where our student team is helping an industrial supplies company, a major B2B distributor of shipping, industrial, and packaging materials, improve its demand forecasting process. Team members was assigned a particular product line and used historical sales and web traffic data to build predictive models that estimate future demand.

Importance / Motivation: Accurate demand forecasting is critical for reducing stockouts, limiting excess inventory, and improving operational efficiency. For a large B2B distributor managing thousands of products, even small improvements in forecast accuracy can generate significant cost savings. By incorporating web-traffic indicators like pageviews and add-to-cart activity, our project explores whether online behavior can act as an early signal of future demand. This approach supports more proactive decision-making in purchasing and inventory planning. Ultimately, the goal is to build a scalable forecasting framework the company can apply across product lines.

Audience/Stakeholder: Industrial Supplies Company

ANALYTICAL FRAMING

Analytical Goal: Develop predictive models that use historical sales and web-traffic behavior to generate more accurate short-term demand forecasts for a specific product line.



Data Sources

Preprocessing

Feature Engineering

Modeling

Evaluation

Business Insight

DATA

Data Source: The data comes from the Company and contains product ID's, Quantity Sold, Adds to Carts, Page Views, and other relevant metrics. Below is more specifications of the dataset.

- **Timeframe: 01/01/2022 to 6/30/2025**
- **Over 20,000+ Data Entries**
- **Daily totals of variables, split by product type and color.**

Data Preprocessing:

- ***Data Selection:*** Separated the dataset by product type to perform analysis.
- ***Data Type Standardization:*** Unified date formats, numeric values, and categorical variables.
- **Added Columns:** Added a day_of_week column to capture seasonality among weekday

DATA UNDERSTANDING

Desk Variables

Variable	*Color*	*Adds*	*Pageviews*	*Orders*	*Total Qty Ordered*	*Homepage*
Color	100%					
Adds	49%	100%				
Pageviews	45%	73%	100%			
Orders	53%	76%	73%	100%		
Total Qty Ordered	46%	66%	61%	86%	100%	
Homepage	0%	1%	-1%	2%	2%	100%

Pedestal Variables

Variables	*Color*	*adds*	*pageviews*	*orders*	*Quantity*
Color	100%				
adds	64%	100%			
pageviews	61%	92%	100%		
orders	65%	94%	89%	100%	
Quantity	62%	90%	86%	95%	100%

Ice Melt Variables

	Total order quantity			
Variable	**ID_12**	**ID_13**	**ID_14**	**ID_15**
adds	0.988	0.468	0.841	0.961
pageviews	0.802	0.456	0.795	0.923
orders	0.968	0.516	0.977	0.992
on_homepage	0.432	0.214	0.452	0.411
qty_ma7	0.755	0.492	0.753	0.751
qty_ma30	0.581	0.215	0.569	0.549
adds_ma7	0.681	0.363	0.621	0.734
pageviews_ma7	0.615	0.346	0.683	0.717
is_winter	0.396	0.155	0.394	0.357
is_holiday_season	0.241	0.107	0.166	0.143
is_weekend	-0.253	-0.147	-0.235	-0.223
dow	-0.038	-0.013	-0.039	-0.031
month	-0.087	-0.046	-0.154	-0.137

Statistical summaries and visualizations were used to examine seasonality, demand spikes, weekday effects and correlations between sales and web visibility metrics. It was identified that the Ice Melt product was extremely seasonal while the other products were not. It also helped identify variables with predictive potential, such as add to cart and pageviews.

PURDUE UNIVERSITY® | Mitch Daniels School of Business

METHODOLOGY

The process began with data collection and preprocessing. Following preprocessing, exploratory data analysis was conducted to understand historical patterns and behaviors across product lines. Multiple predictive models were developed to forecast demand and compare forecasting accuracy. Linear regression models served as initial baselines to quantify relationships between different predictors and total order quantity. These were followed by time-series forecasting approaches, including Holt-Winters exponential smoothing and ARIMA models to capture recurring seasonal cycles depending on the product. Model performance was evaluated using an 85/15 train-test split, using the data from 01/01/2022 to 12/31/2024 as the training set, and 01/01/2025 onwards for the testing set. Prediction accuracy was measured using RMSE, MAE, and R^2. Beyond statistical fit, business interpretability was considered critical, coefficients were analyzed to make sure that variables were statistically significant and not over training the models. When inconsistencies emerged, feature sets and model configurations were iteratively refined to reduce overfitting and increase generalization. Through iterative refinement, the models were progressively improved by testing alternative variable combinations. The final methodology allowed the forecasting framework to balance statistical accuracy with interpretability, creating models suitable not only for prediction but also for operational decision-making related tc inventory planning and promotional strategy.

MODEL BUILDING

Pedestal Modeling Results

	Training Data			Testing Data		
	MSE	**MAE**	**RSQ**	**MSE**	**MAE**	**RSQ**
Pedestal #1 Linear Regression	1.2	4.7	51%	1.5	6.3	45%
Pedestal #2 Decision Tree	4.2	47.3	74%	4.5	66.9	71%
Pedestal #3 Linear Regression	7.2	133.7	81%	9.1	193.3	84%

Ice Melt Modeling Results

	Training Data			Testing Data		
	RMSE	**MAE**	**RSQ**	**RMSE**	**MAE**	**RSQ**
Linear Regression (Variables provided)	152.9	58.3	92%	238.4	156.7	76%
Linear Regression (Moving Average Variables added)	89.61	51.89	93%	286.07	124.4	87%
Arima	39.6	10.4		43.4	19.4	

Desk Product Modeling Results

	Training Data			Testing Data		
	MSE	**MAE**	**RSQ**	**MSE**	**MAE**	**RSQ**
Linear Regression	343.3	13.1	77.9%	710.8	18.3	77.4%
Decision Tree (1% complexity, min. 10 splits)	295.8	12.4	80.8%	733.8	18.5	69.2%
Random Forest (50 Trees)	141.8	8.58	90.9%	624.5	16.6	75.3%

RESULTS

Business Validation: To ensure the models were not only statistically sound but also operationally useful, results were reviewed in the context of real-world business processes. Forecast outputs were compared against historical ordering patterns, lead times, and known seasonal events to confirm that predictions aligned with operational realities. Key stakeholders evaluated whether the model's drivers reflected observed customer behavior and could reliably inform purchasing decisions. This validation highlighted where models provided actionable insight, such as early demand signals from web traffic, and where adjustments were needed. The approach delivered forecasts that better support inventory planning, reduce uncertainty, and align with the company's data-driven decision-making goals. These results can then be utilized by the stakeholders to make decisions regarding when their products should be promoted, advertised, and placed in stock.

Model Comparison:

Top Model Comparison Across Products

	Training Data			Testing Data		
	MSE	**MAE**	**RSQ**	**MSE**	**MAE**	**RSQ**
PEDESTAL- Pedestal #3 Linear Regression	7.2	133.7	81%	9.1	193.3	84%
ICE MELT- Arima	1592 .01	10.4		1883. 56	19.4	
DESK - Random Forest (50 Trees)	141. 8	8.58	90.9%	624.5	16.6	75.3 %

CONCLUSION

This project demonstrated that combining historical sales data with web-traffic indicators can meaningfully enhance demand forecasting for a large B2B distributor. Linear and time-series models each contributed unique strengths, and iterative refinement allowed us to balance predictive accuracy with interpretability. The final models offer improved visibility into short-term demand trends and provide earlier signals for inventory planning. These insights can help the company reduce stockouts, optimize purchasing decisions, and support more proactive operational strategy. The forecasting framework developed here can be scaled across product lines to further strengthen data-driven decision-making.

FUTURE WORK

Future work on this project could include developing more advanced machine learning models, such as tree-based approaches like Random Forest, to better capture nonlinear relationships in the data. Additionally, incorporating seasonality and holiday effects would help improve predictive accuracy by accounting for recurring patterns and demand fluctuations throughout the year.

ACKNOWLEDGEMENTS

We sincerely appreciate Professor Xing Wang, our TA Keun Jun Park, and the Industrial Supply Company stakeholders we partnered with for providing us with this opportunity, and for their invaluable guidance and support throughout the course of this project.