WESTMINSTER UNIVERSITY

Data Capstone Modeling Plan

SAP Backorder Prediction, Regression & Classification Models

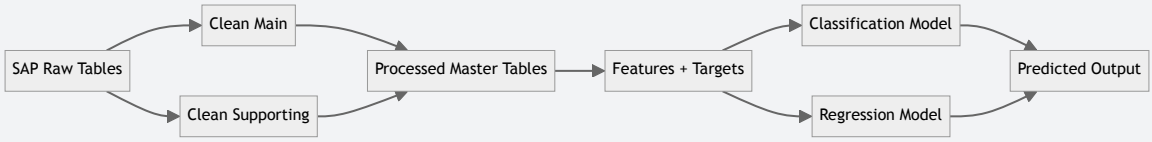
Purpose: Model selection for backorder prediction, regression (magnitude) and classification (risk). Primary + Plan B for each.

Scope: Predict backorder magnitude from historic sales trends and inventory; classify backorder risk.

Status: Plan, to be refined after EDA and baseline results

Data Flow: From Raw to Prediction

How data moves from the SAP repository through ETL to final predictions.



Stages

- **SAP Raw:** Sales, inventory, delivery, billing, material master, etc.
- **Clean Main / Supporting:** Normalized, deduplicated, joined by keys
- **Processed Master:** master_order_fulfillment_brd, master_inventory_material, master_woc
- **Features + Targets:** Sales trends, inventory, WOC, backorder flags
- **Models:** Classification (yes/no) and regression (magnitude)

Final Prediction Output

Each row is a material/plant with predicted backorder risk and inventory context.

Material	Plant	Predicted Backorder	WOC	Saleable Inventory
000000000000002733	1000	Yes	2.1	45
000000000000002480	1000	No	8.3	120
000000000000001107	1000	Yes	0.5	12

Regression: Magnitude of Backorder

Predict numeric backorder (such as `backorder_units`, shortfall, demand) from historic sales trends and inventory.

Primary: Ridge / ElasticNet Regression

> **Rationale:** Linear, interpretable, and robust to multicollinearity (common in sales + inventory features). Use layered feature groups (sales trends, inventory, lead time) for interpretability.

- **Pros:** Fast, stable, easy to explain to stakeholders
- **Cons:** Assumes mostly linear relationships

Plan B: Gradient Boosting Regressor (XGBoost / LightGBM)

> **Rationale:** Captures non-linear effects and interactions without manual feature engineering. Often best performance on tabular supply-chain data.

- **Pros:** Handles missing values and outliers; strong on complex relationships

- **Cons:** Less interpretable; needs tuning

Classification: Backorder Risk (Yes/No)

Predict backorder risk, binary or risk level classification.

Primary: XGBoost / LightGBM

> **Rationale:** Strong on tabular data; handles ~10% backorder class imbalance well. Built-in feature importance and interaction capture.

- **Pros:** Often best performance; handles imbalance; feature importance for reporting
- **Cons:** Less interpretable than logistic regression

Plan B: Logistic Regression

> **Rationale:** Interpretable baseline; coefficients show direction and relative importance of inventory, lead time, order/delivery timing.

- **Pros:** Simple, fast, easy to explain
- **Cons:** Assumes linear decision boundary; weaker if relationships are non-linear

Model Selection Summary

Task	Primary	Plan B
Regression (magnitude)	Ridge / ElasticNet	XGBoost / LightGBM Regressor
Classification (risk)	XGBoost / LightGBM	Logistic Regression

> **Workflow:** Start with Ridge/ElasticNet for regression and XGBoost/LightGBM for classification. If Ridge underperforms, switch to gradient boosting for regression. If XGBoost is hard to explain or overfits, fall back to logistic regression for classification.

Glossary

Term	Definition
AWD	Average Weekly Demand, rolling 24-week average of units shipped
BRD	Business Requirements Document
CSV	Comma-Separated Values, flat file format
EDA	Exploratory Data Analysis
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
PO	Purchase Order
SAP	Systems, Applications, and Products (ERP software)
SI	Saleable Inventory, unrestricted stock available to fulfill orders
SO	Sales Order
agg	Aggregated, summed/grouped by key columns
sales_org	Sales Organization, SAP organizational unit
WOC	Weeks of Coverage, net available inventory ÷ AWD