

Phase 1 – Data Cleaning and Validation

Learners begin by exploring and cleaning RetailSmart's core datasets — customers, sales, products, marketing, and reviews. They use SQL and Python to:

- Handle missing values and outliers.
- Validate referential integrity.
- Standardize categorical data.
- Conduct univariate, bivariate, and time-series EDA.
- Derive RFM metrics and analyze churn patterns.

Output Files:

customers_cleaned.csv, sales_cleaned.csv, marketing_cleaned.csv, products_cleaned.csv

Phase 2 – Predictive Modeling

Using cleaned data, learners integrate multiple sources to create a unified modeling dataset. They then:

- Engineer features such as Recency, Frequency, and Monetary value.
- Train baseline and advanced models (Logistic Regression, Random Forest, Gradient Boosting).
- Perform hyperparameter tuning and evaluate using Accuracy, Precision, Recall, F1, and ROC-AUC.
- Interpret model outputs and identify top churn predictors.

Output Files:

model_input.csv, final_rf_model.pkl, scaler.pkl

Phase 3 – Advanced Analytics

Building upon the predictive phase, learners explore unsupervised and time-series analytics to derive strategic insights. They:

- Perform Customer Segmentation using clustering.
- Execute Demand Forecasting using time-series models.
- Analyze seasonal trends and cluster profiles for strategic planning.

Output Files:

cluster_summary.csv, customers_with_clusters.csv, forecast_results.csv

Phase 4 – Visualization and Storytelling

Learners synthesize all analytical outputs to build an interactive Power BI dashboard that bridges technical findings and business insights. They:

- Integrate Phase 1–3 outputs into Power BI.
- Create KPI measures with DAX.
- Design multi-page dashboards: Executive Summary, Customer Insights, Churn Prediction, and Forecasting Trends.
- Embed narrative storytelling for actionable insights.

Output Files:

RetailSmart_Dashboard.pbix, RetailSmart_Storytelling_Report.docx