PROJECT PROPOSAL (DATS – 6103)

OPTIMIZING SUPPLY CHAIN AND SALES PERFORMANCE

SUBJECT - DATA MINING

SUBMITTED BY - NIKHILESH DHAVALE & ADITYA KARANDE

Objective

The objective of this project is to leverage data mining and predictive modeling techniques to extract actionable insights from the DataCo Supply Chain dataset. The analysis focuses on enhancing operational efficiency, predicting delivery risks, understanding customer behavior, and optimizing product inventory. By using machine learning and data analytics, the goal is to develop data-driven recommendations that improve delivery reliability, customer segmentation, and inventory planning in supply chain operations.

Impact

The project aims to demonstrate how data mining can significantly improve decision-making within supply chain management and sales analytics. Key expected impacts include:

- Improved delivery time predictions to mitigate late shipments and enhance customer satisfaction.
- Enhanced customer segmentation through RFM (Recency, Frequency, Monetary) analysis for targeted marketing.
- Optimized inventory through ABC product classification, ensuring focus on high-value items.
- Better demand forecasting and cost efficiency in logistics planning.
- Increased operational transparency through data visualization and predictive insights.

Dataset(s)

The dataset used for this analysis is the 'DataCo Smart Supply Chain for Big Data Analysis' obtained from Kaggle (https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis). It contains over 180,000 records detailing customer orders, shipment details, product information, sales, and delivery performance. The data covers multiple dimensions, including order region, market, shipping mode, product category, and customer segments. It is a structured dataset with both categorical and numerical features, making it suitable for statistical analysis, machine learning, and clustering techniques.

Approach

The project follows a comprehensive data mining pipeline comprising the following steps:

1. Data Understanding and Cleaning: Imported and examined the dataset, handled missing values, removed irrelevant columns, and standardized data formats. Feature engineering was applied to derive new variables such as year, month, day, and time of order and shipment.

2. Exploratory Data Analysis (EDA):Visualized data distributions, correlations, and relationships between shipping mode, customer segment, and delivery performance. Used heatmaps, scatter plots, and boxplots to uncover patterns in sales and shipping behavior.

3. Model Building: Applied multiple machine learning models including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting to predict late delivery risks and classify order outcomes. Hyperparameter tuning using RandomizedSearchCV improved model generalization.

4. Customer and Product Analytics: Conducted RFM analysis for customer segmentation to identify loyal, high-value, and inactive customers. ABC analysis was performed to classify products based on their contribution to overall revenue, supporting inventory prioritization.

5. Evaluation and Visualization: Model accuracy, precision, recall, and ROC-AUC were used for performance evaluation. Visualizations were created to interpret model results and provide clear business insights.

Timeline

• Week 1 – Data acquisition, understanding dataset structure, and initial cleaning.
• Week 2 – Exploratory Data Analysis and feature engineering.
• Week 3 – Model development (classification and predictive modeling).
• Week 4 – Model tuning, RFM and ABC segmentation, and insight generation.
• Week 5 – Final report writing and result presentation.

Possible Issues

Several challenges may arise during the project, including:
• Data Quality: Missing or inconsistent entries may affect model accuracy, requiring robust cleaning and imputation.
• Imbalanced Classes: Delivery risk prediction may suffer from imbalance, necessitating resampling or weighting strategies.
• Multicollinearity: Highly correlated numerical variables could distort regression results, addressed through VIF analysis and feature reduction.
• Computational Complexity: Large datasets and model tuning may demand significant processing time, mitigated using optimized algorithms.
• Interpretability: Ensuring that complex models provide explainable insights remains crucial for practical decision-making.

Despite these challenges, careful validation and systematic workflow design will ensure meaningful and reliable results.

SOURCE

The dataset for this project is obtained from Kaggle.
https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis

---