**Analysis of Superstore’s Product Profitability and Customer Purchase Prediction**

**Problem Statement:**

Superstore wants to improve its marketing strategies and services by using machine learning (ML) algorithm to gather insights into consumer purchasing trends, discover important factors influencing product profitability, and predict which product categories customers are likely to buy. This proposal presents an overall approach to resolving these business difficulties using ML technologies.

**Business Understanding:**

The Superstore is a mid-sized retail company that operates across many regions in the United States. Sector competitors are increasingly implementing ML to optimize pricing methods, improve customer segmentation, and refine marketing approaches, shipping expenses, inventory management procedures and learning about customer demographics and purchasing trends. There is an increasing demand to keep ahead of the trend in utilizing new analytics to improve decision-making, and understanding market analysis, operational efficiency, competitive analysis.

**Data Understanding:**

We have taken the dataset from the data world and here is the link for the dataset <https://data.world/vizwiz/superstore-2023>. It consists of 21 features and 9994 instances.

**Data Preparation:**

Converted the 'Order Date' and 'Ship Date' columns from object to datetime format to enable temporal analysis. Checked for null values we found some rows with missing values so dropped the rows with null values to prepare clean dataset for modeling. This reduced the number of rows from 9994 to 9980.Performed one-hot encoding to deal with categorical columns like Category, Sub-Category, Ship Mode, Segment and Region.

**Exploratory Data Analysis:**

We have conducted EDA to analyze variability in profits and sales. Profit exhibits moderate correlation with sales, weak with quantity, and negative with discounts. Product category, sub-category, customer segment and region impact average profit significantly. These insights present optimization of product, pricing, promotions, and geographical focus to boost profitability. We have also analyzed the target variable Category with respect to other continuous variables to improve customer segmentation.

**Data Modelling:**

Our aim is to determine which factors had a major influence on the Superstore’s products’ profitability. We have used advanced regression analysis methods, such as Linear Regression, Random Forest Regression, and Gradient Boosting Regression. Linear Regression is used to predicts profitability based on sales, quantity, discount, category, sub-category, ship mode, segment, and region, Random Forest Regression is used to predicts profitability by capturing non-linear correlations and Gradient Boosting Regression is used to correct errors sequentially for better prediction of complicated patterns.

For the prediction of the product categories customers were most likely to purchase based on their demographics and past purchases. We have used classification models such as Support vector machines, Random Forest Classifier, and Logistic Regression. Support Vector Machines (SVM) is used to Predicts product categories based on past purchases and demographic information, Random Forest Classifier is used to classifies products into categories with high accuracy whereas Logistic Regression predicts product categories based on historical data and customer demographics.

**Conclusion/Evaluation:**

Linear regression performed poorly in predicting profitability, explaining only 28% variance, indicating a complex non-linear relationship. Random forest regression significantly improved performance to 74% by modeling non-linearities and interactions. Gradient boosting was most effective with 78% R-squared, as it sequentially corrects errors of previous decision trees to capture complex data patterns driving profit**.**

Random Forest classifier performs best with 76% accuracy in predicting product categories purchased based on customer data. Logistic Regression is least accurate at 64% while SVM has moderate 66% accuracy. Random Forest balances performance across categories unlike the other models.