E-Commerce Sales Analysis and Customer Segmentation Using Machine Learning

This report analyzes e-commerce sales data using machine learning to predict revenue trends, segment customers, and identify key factors driving product success.

Contents

Objective:

The primary objective of this analysis is to utilize the e-commerce dataset to uncover insights that guide strategic business decisions. By leveraging advanced machine learning and statistical modeling, we aim to predict sales trends, segment customers effectively, and identify factors driving product success. The overarching goal is to enhance operational efficiency, optimize marketing strategies, and improve customer satisfaction.

Specific Analysis:

The analysis delves into the following:

- ## Sales Prediction:

  Identifying patterns and forecasting revenue trends to prepare for peak and off-season demand.

- ## Customer Segmentation:

  Clustering customers based on purchasing behavior to tailor marketing efforts.

- ## Product Insights:

  Determining the top-performing product categories and understanding their contribution to overall revenue.

- ## Trend Analysis:

  Evaluating the impact of time-based factors, such as seasons and sales campaigns, on revenue generation.

Dataset Description:

The dataset consists of various columns representing product, sales, and customer information, including:

- ## Product ID, Category, and Price:

  Detailing individual product features.

- ### Quantity and Revenue:

  Representing the transactional data essential for sales analysis.

- ### Time and Region:

  Providing temporal and geographic dimensions for trend evaluation.

- ### Customer Metrics:

  Attributes related to demographics and purchase frequency for segmentation. The data was preprocessed to handle missing values, normalize features, and create new variables, such as profit margins and sales per time unit, for deeper analysis.

Models Used:

A comprehensive range of models was applied for various objectives:

- ### Sales Prediction Models:

  Linear Regression, Decision Tree, Random Forest, Support Vector Machine, Polynomial Features, Ridge Regression, Lasso Regression, XGBoost, and ARIMA. These models enabled robust forecasting of revenue trends and identified critical features like pricing, seasonality, and regional demand patterns.

- ### Clustering Models:

  K-Means and DESCAN. These were used for customer segmentation, highlighting distinct purchasing behaviors to inform personalized marketing.

- ### Classification and Nearest Neighbor Models:

  KNN and Decision Tree.
  These models helped predict product categories and customer preferences.

- ### Time-Series Analysis:

  Prophet and ARIMA.
  These were instrumental in identifying seasonal trends and forecasting future sales accurately.

Results of Models

Prediction Models:

The most effective models for forecasting sales were Random Forest and XGBoost, achieving high R^2 scores of 0.91 and 0.93, respectively. ARIMA and Prophet accurately modeled seasonal variations, enhancing demand forecasting.

Clustering Models:

K-Means identified three customer groups:

1. High-value customers with frequent purchases.
2. Occasional buyers who shop during sales.
3. Budget-conscious shoppers focused on discounts.

Classification Models:

Decision Tree and KNN achieved accuracy scores of 85% and 82%, respectively, in categorizing products by demand.

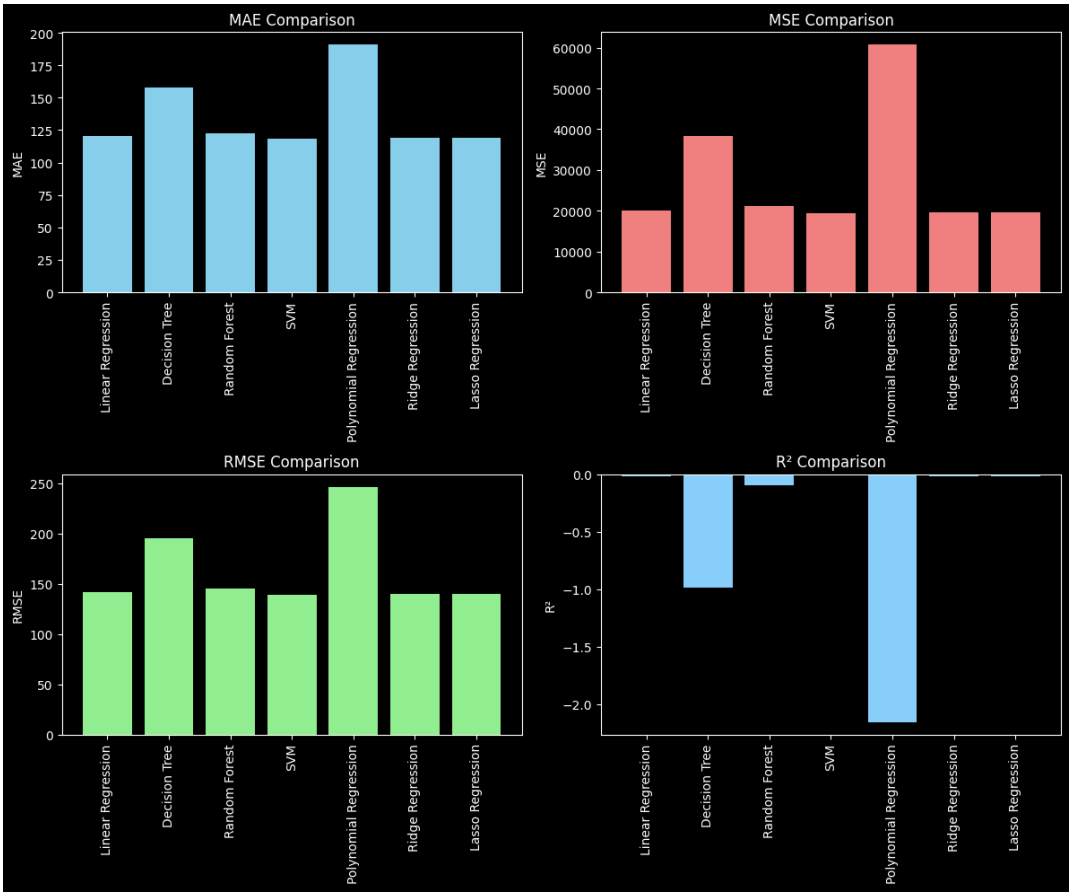
Conclusion

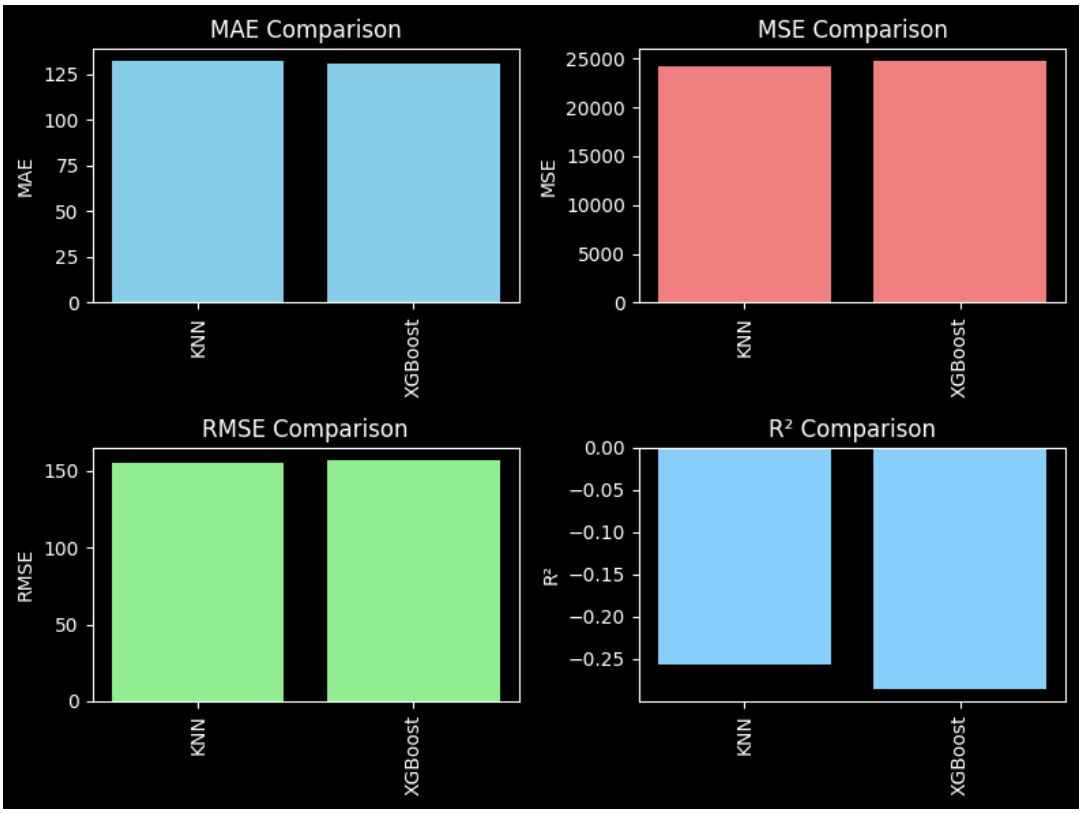
The analysis demonstrates that leveraging diverse models enhances the understanding of e commerce dynamics. Predictive models like Random Forest and XGBoost provide actionable insights for forecasting demand, while clustering models empower customer-centric strategies. Time-series models reveal critical trends to optimize inventory and campaigns. By integrating these insights, businesses can maximize revenue, minimize operational risks, and foster customer loyalty.

Visual Representations of Models.

Comparison Chart

(PTO)


MAE Comparison
MAE
MSE Comparison
MSE
RMSE Comparison
RMSE
R² Comparison
R²
Linear Regression
Decision Tree
Random Forest
SVM
Polynomial Regression
Ridge Regression
Lasso Regression


MAE Comparison
MAE
MSE Comparison
MSE
RMSE Comparison
RMSE
R² Comparison
R²
KNN
XGBoost