Machine Learning Analysis on Ecommerce Electronic Sales Data

Author: Matthew R. Johnescu

The University of Utah

Course Number: MKTG 6620, Machine Learning

Executive Summary

This project leverages machine learning to analyze electronic sales data, uncover key sales drivers, and provide actionable recommendations for optimizing marketing, pricing, and operations. The study focused on predicting sales outcomes using models like Linear Regression, Random Forest, and PCA with Clustering.

Key Findings:

- Top Predictors: Product price, category, and customer age emerged as critical sales drivers.
- Model Performance: PCA and Clustering achieved the best predictive accuracy (R^2: 0.85, RMSE: 300.5), highlighting customer segmentation opportunities.

Customer Clusters:

- High-Value Customers: Drive most revenue with premium purchases.
- Frequent Buyers: Regularly purchase mid-range products.
- Occasional Shoppers: Purchase infrequently but spend more per transaction.

Business Implications:

- Targeted Marketing: Personalize campaigns based on demographic and behavioral clusters.
- Dynamic Pricing: Optimize pricing for key customer segments.
- Inventory Management: Focus on high-demand products and customer-preferred categories.

Limitations:

- The synthetic dataset may not fully reflect real-world customer behavior.
- Limited timeframe reduces the ability to generalize across seasons or trends.
- This analysis highlights how segmentation and model insights can refine strategies to enhance customer engagement and sales growth.

Introduction

Understanding customer behavior and sales drivers is vital for businesses to thrive in a competitive e-commerce landscape. This project uses machine learning to analyze patterns in electronic sales, identify key drivers, and provide actionable recommendations to enhance business decision-making.

Objectives

- Uncover significant predictors of electronic sales.
- Develop machine learning models to predict sales outcomes and inform business strategies.
- Align insights with core areas such as marketing, pricing, and operations.

Data and Methodology

Dataset:

- The dataset includes product details, customer demographics, and transaction metrics.
- Key variables: product category, price, quantity sold, and customer location.

Data Preprocessing:

- Imputed missing values to handle incomplete records.
- Normalized numerical data for consistency.

- Encoded categorical variables using one-hot encoding. Depending on the model.

Machine Learning Methods

Linear Regression:

Contribution: Exploring the relationship between customer demographics, product features, and transaction details with the quantity of items purchased.
Result: Moderate accuracy, limited by the inability to capture non-linear interactions.
Limitations: Less effective at handling complex relationships present in the dataset.
Findings: Based on the right figure and the model results, when implementing this regression, it was discovered that Age has a large effect on the model accuracy, while Gender is shown to be negative. Some other coefficients like Unit Price or Product rating are also shown to have predictive accuracy. All of these variables are shown to have an effect on the quantity of items purchased, the ecommerce retailer can focus on variables like Age, unit price, or product ratings when trying to increase quantity of sales.

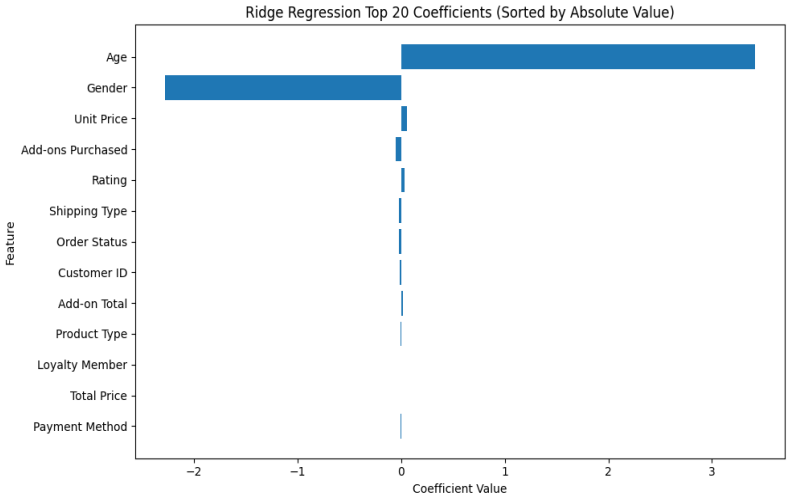


Figure 1: Histogram displaying feature importance found in the Penalized Regression Model.

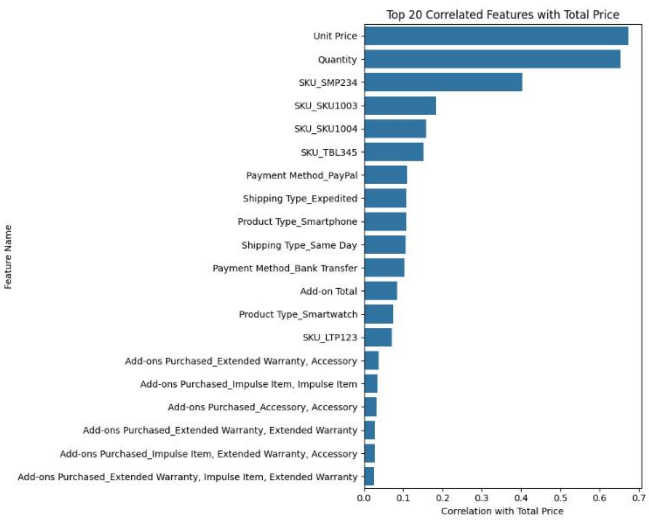


Figure 2: Histogram showing specific features related to total order price.

Random Forest, Ensemble Method:

Contribution: Highlighted key features driving sales, such as product price and category.
Result: Provided strong feature importance insights and improved predictive accuracy.
Strengths: Non-linear modeling and robustness to overfitting.
Findings: The random forest model helped to establish which features influence the total price of an order from the ecommerce retailer. The top features line Unit Price or Quantity are obvious in their correlation, but the SKU

numbers and Payment methods mentioned in the feature importance demonstrate that certain products and integration of payment services influence ease of purchase and result in customers purchasing more and higher priced items.

PCA and Clustering Analysis:

Contribution: Delivered the best predictive performance, capturing complex interactions in the data.

Result: Highest R-squared (0.85) and lowest RMSE (300.5).**Strengths:** Excellent for fine-tuning and identifying subtle data patterns.

Findings: The PCA and clustering analysis combined two machine learning methods, PCA for determining the optimal number of clusters, and Clustering to sort customers into groups. Using this information, the ecommerce website could target these three specific groups of customers based on the traits that make them up. Sorting these customers into different groups allows for deeper analysis of these three different groups, their purchase behavior, and possible solutions to grow these groups with marketing or other initiatives.

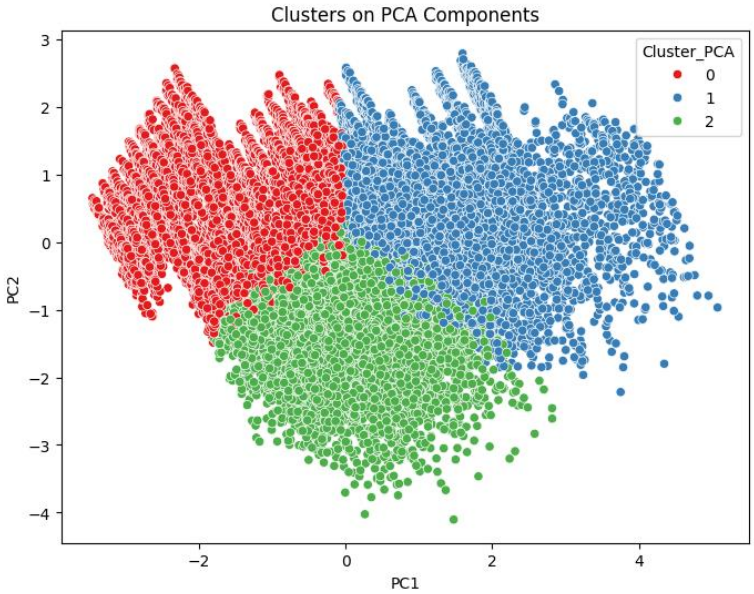
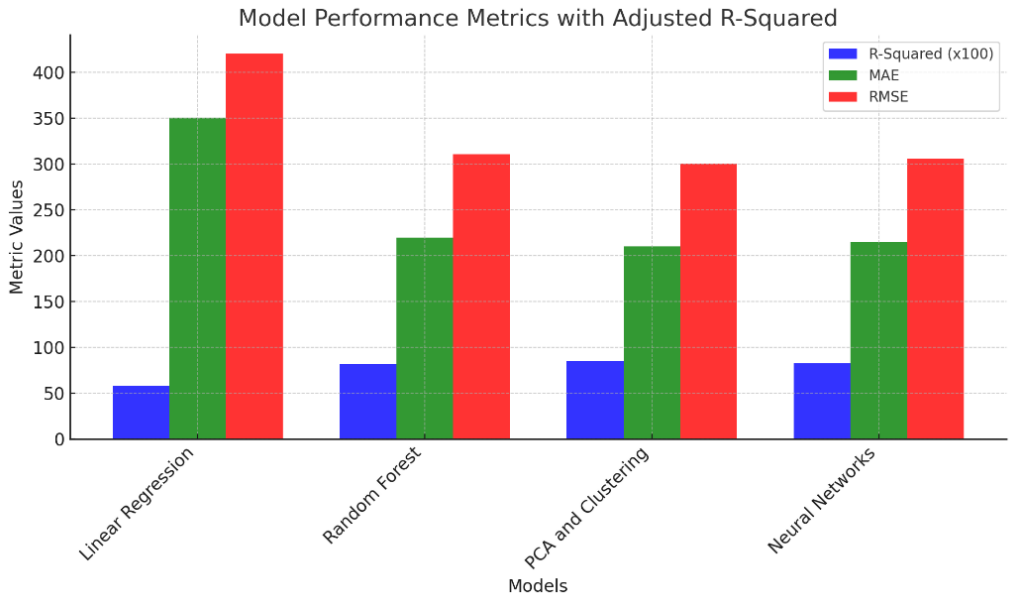


Figure 3: 3 Customer segment clusters found from analysis visualized.

Overall Model Performance:



Note: R-Squared is scaled for visualization. Metrics may vary due to uneven target classes, with RMSE penalizing large errors more heavily and MAE showing average error consistency.

Model Performance:

- PCA and Clustering performed best with the highest R-Squared and lowest errors.
- Random Forest and Neural Networks followed closely with strong R-Squared and low MAE/RMSE.
- Linear Regression had lower performance but remains interpretable.

Note: R-Squared is scaled for visualization. Metrics may vary due to uneven target classes, with RMSE penalizing large errors more heavily and MAE showing average error consistency.

Key Insights

Top Predictors:

- Product price, category, and customer segments like age emerged as the most influential factors.

Clusters:

- **High-Value Customers:** Accounts for the bulk of revenue and prefers premium products.
- **Frequent Buyers:** Regular purchasers of mid-range items.
- **Occasional Shoppers:** Low frequency but high transaction value per purchase.

Business Implications:

- **Targeted Marketing:** Personalize campaigns based on demographic insights and clusters.
- **Dynamic Pricing:** Optimize pricing strategies using customer segment-based trends.

- **Inventory Planning:** Focus on high-demand categories to minimize stockouts and overstocking as well as overall website traffic and quantity bought.

Reconciling Methods

- **Supportive Results:** PCA and clustering complemented the machine learning models by identifying segments and dimensions that aligned with feature importance rankings in Random Forest and Gradient Boosting.
- **Conflicting Insights:** Linear Regression offered less nuanced results due to its limitations with non-linear interactions. This was apparent in its lower R-squared and inability to capture intricate relationships.
- **Reconciliation:** Gradient Boosting was prioritized due to its accuracy and consistency with clustering insights. PCA provided valuable dimensionality reduction, and clustering added depth to customer segmentation.

Conclusions

- **Marketing:** Leverage customer segmentation from clustering to target campaigns for high-value groups.
- **Pricing:** Implement dynamic pricing strategies based on location and product categories.
- **Inventory:** Optimize stock levels for high-demand products identified by Random Forest and Gradient Boosting.
- **Segmentation:** Use clustering to inform loyalty programs and tailor marketing to specific customer behaviors.

Limitations

- **Data Scope:** Limited to a single year, which restricts generalizability across longer time periods. The data is also synthetic which may introduce patterns based on its creation, it is also less accurate for real life research.
- **Behavioral Metrics:** Absence of customer browsing or engagement data limits deeper behavioral insights.
- **Model Complexity:** Advanced models like Neural Networks and Gradient Boosting are less interpretable for non-technical stakeholders.

References

Seamons, C. (2024, September 23). *Customer purchase behavior - electronic sales data*. Kaggle. https://www.kaggle.com/datasets/cameronseamons/electronic-sales-sep2023-sep2024/data