E-commerce

Sales Prediction

Springboard Data Science Capstone Project
by Abhinaya Gyawali

Business Problem

Challenges in retail inventory management

Overstock Costs

Excess inventory leads to increased holding costs, tying up capital and resources that could be better utilized elsewhere, ultimately affecting profitability and operational efficiency in retail.

Stockouts Impact

Insufficient stock leads to missed sales opportunities, resulting in reduced revenue and customer dissatisfaction. Retailers must prioritize accurate demand forecasting to avoid these costly scenarios.

Project Objective

Defining the prediction target

Target Variable

The primary target variable in this project is **Units Sold**, which quantifies product demand and serves as the basis for all predictive modeling efforts conducted throughout the analysis.

Supervised Learning

This analysis employs a **supervised regression problem** approach, utilizing historical sales data to train models aimed at minimizing prediction error and enhancing forecasting accuracy.

Dataset Overview

Understanding the E-commerce dataset

Data Source

The dataset utilized for this project originates from Kaggle, providing a rich collection of e-commerce transactions that can support predictive modeling efforts effectively.

Observations Count

The dataset comprises approximately 1,000 observations, offering a substantial sample size for analysis while ensuring a manageable scope for model training and validation processes.

Feature Types

It includes both numeric and categorical features, such as product details, pricing information, and marketing variables, which are essential for understanding sales patterns and trends.

Exploratory Data Analysis

Understanding patterns, trends, and key relationships

Variability in Units

Units sold show moderate variability across different product categories in the dataset.

Linear Relationships

Numeric features exhibit weak to moderate linear relationships with overall sales performance.

Demand Factors

Results suggest that demand is influenced by multiple factors together rather than isolated ones.

Outlier Inspection

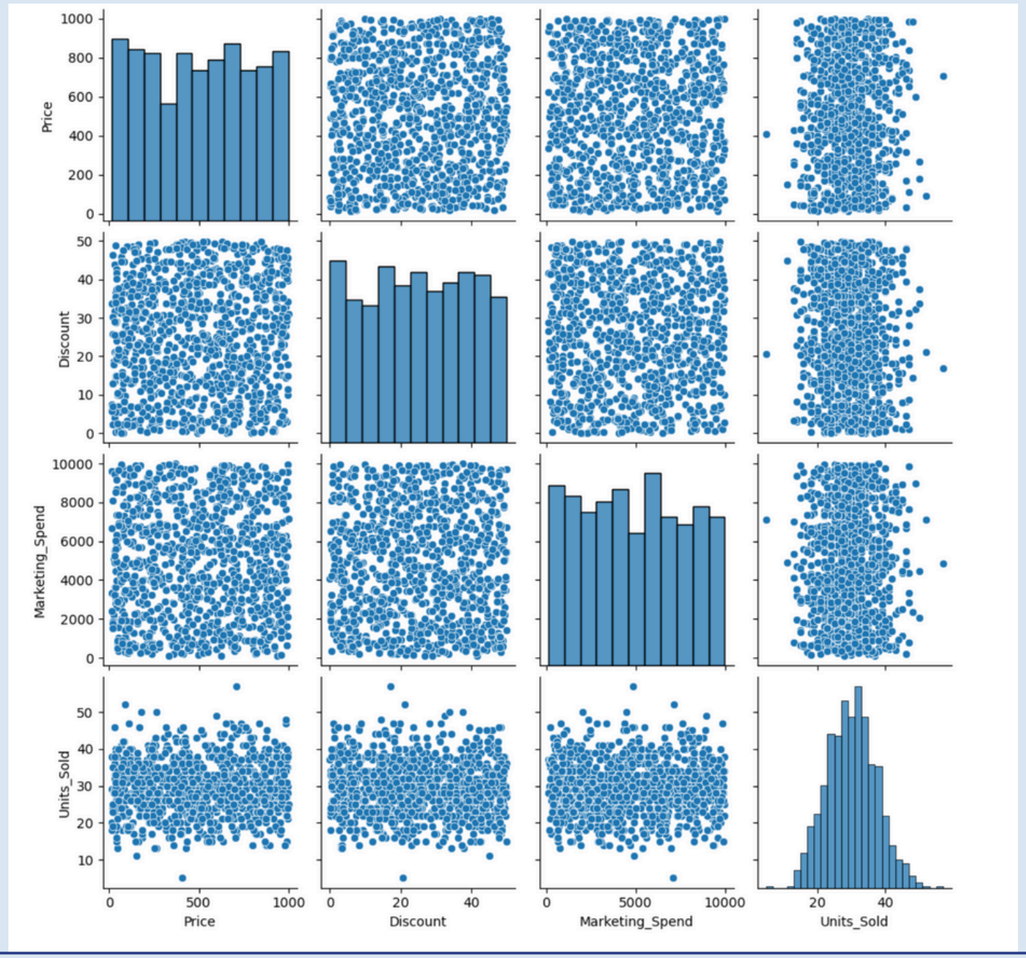
No extreme outliers were found after conducting an initial data inspection process.

Feature Influence

No single feature strongly dominates the sales behavior observed in the analysis.

Exploratory Data Analysis

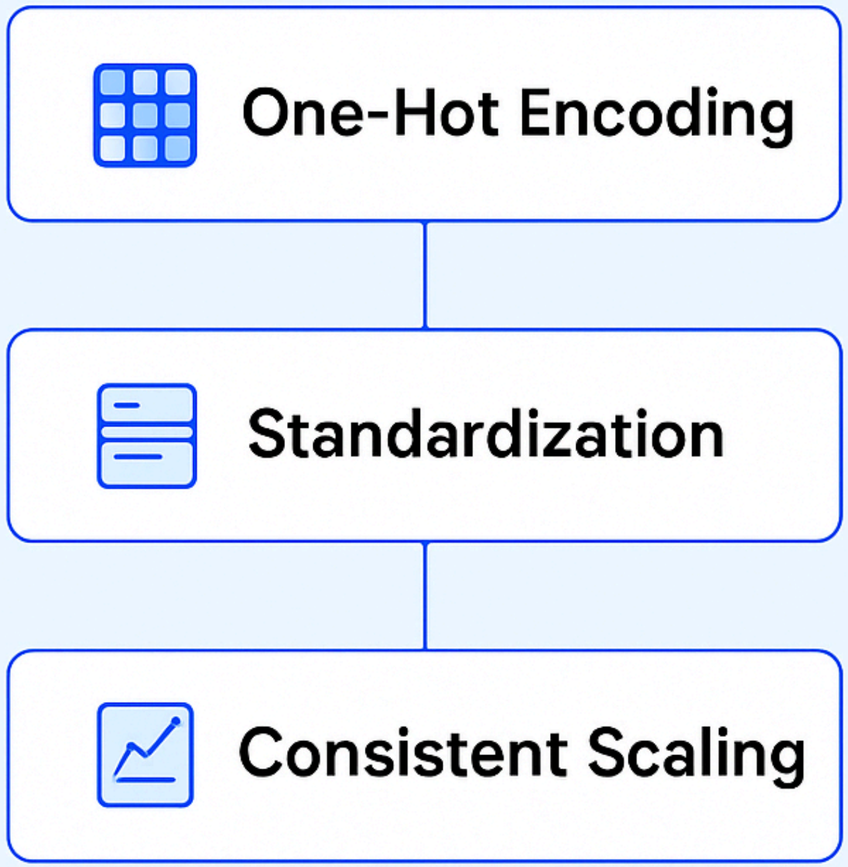
Understanding patterns, trends, and key relationships



- No single feature shows a strong linear relationship with Units Sold, indicating that sales depend on multiple interacting factors rather than one dominant driver.

Data Preparation Process

This step involved one-hot encoding for categorical variables, standardizing numeric features with a pre-fitted scaler, and ensuring consistent scaling across data splits without explicit imputation for missing values.



Modeling Strategy

Exploring different regression models

Baseline Approach

The baseline model employs a DummyRegressor that predicts the mean value of units sold. This provides a reference point for assessing the effectiveness of other models.

Linear Regression

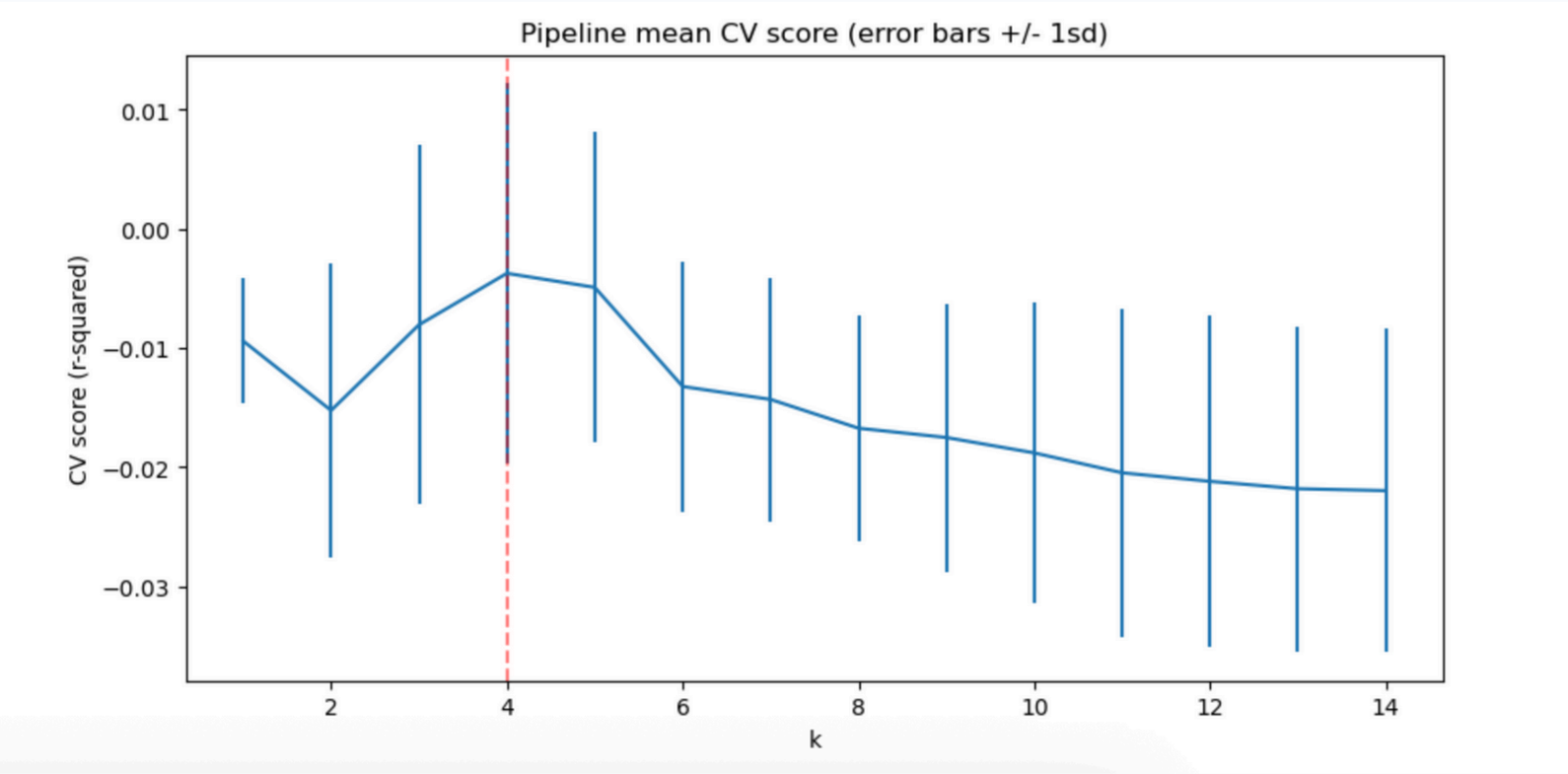
Linear Regression is utilized with hyperparameter tuning to optimize performance. This model is favored for its simplicity and interpretability, making it a strong candidate for this analysis.

Random Forest

The Random Forest Regressor, also tuned, is included for comparison. It captures complex relationships within the data, though it may introduce variability in predictions due to its ensemble nature.

Optimal Model Complexity Selection

Cross-validation identifies k = 4 as best



Model performance peaks at k=4, with higher values adding variance and no improvement in cross-validated R^2, so k=4 was selected as the optimal model complexity.

Training Size vs. Cross-Validation Score

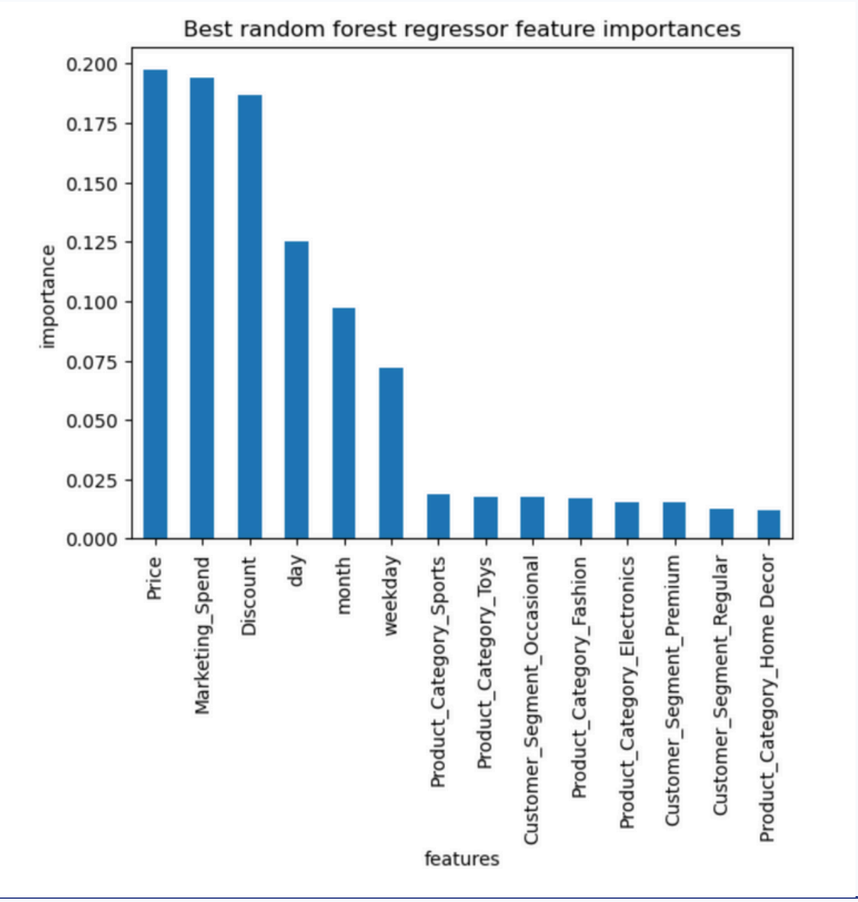
A look at where the model stabilizes

Model performance improves steadily with more training data, and begins to stabilize at larger sample sizes, indicating additional data would yield incremental but diminishing gains.

Key Drivers of Sales Predictions

Pricing and promotions dominate model importance

Sales are primarily driven by pricing and marketing decisions, with timing effects playing a secondary role and product/customer categories having limited influence



Evaluation Methodology

Metrics for Model Assessment

Primary Metric

The primary evaluation metric used is the **5-fold cross-validated Mean Absolute Error (MAE)**, which provides a robust assessment of average prediction error across different data subsets.

Error Measurement

MAE measures the average absolute difference between predicted and actual values, providing clear insight into the model's performance and its effectiveness in real-world applications.

Outlier Robustness

This metric is particularly **robust to outliers**, making it a reliable choice for evaluating models in e-commerce datasets where extreme values can skew results and inflate error measurements.

Model Performance

Comparing Cross-Validated Results

Linear Regression

The Linear Regression model achieved a CV MAE of approximately **5.90**, indicating its reliability and accuracy in predicting product sales, contributing to effective inventory management.

Random Forest

The Random Forest model recorded a CV MAE of about **6.12**, demonstrating its predictive capability; however, it showed slightly more variability in its performance compared to Linear Regression.

Stability Insights

Overall, the Linear Regression model exhibited greater stability and generalization ability across different data segments, making it a preferred choice for consistent e-commerce sales predictions.

Test Set Performance

Comparing model accuracy on unseen data

Random Forest Results

The Random Forest model achieved a **mean absolute error** (MAE) of approximately 5.52, indicating reliable predictive performance on the test dataset with minimal overfitting.

Linear Regression Results

The Linear Regression model recorded a **mean absolute error** (MAE) of around 5.55, showcasing its effectiveness, albeit slightly higher than the Random Forest in this specific scenario.

Model Performance Comparison

Cross-validation stability versus test-set accuracy

Cross-validated MAE (average performance across folds)

- Random Forest (RF):
  - Mean MAE = 6.12
  - Std = 0.18
- Linear Regression (LR):
  - Mean MAE = 5.90
  - Std = 0.17

→ LR has better cross-validated MAE than RF.

Final Test MAE (performance on unseen test data)

- RF Test MAE: 5.516
- LR Test MAE: 5.553

→ RF has slightly better test-set MAE. (But the difference is tiny: only ~0.037)

Final Model

Choosing the best approach

Selected Model

The **Linear Regression** model was selected for its robustness and simplicity, making it easy to interpret and implement in real-world applications for accurate sales predictions.

Best Performance

Through cross-validation, Linear Regression achieved the **lowest CV MAE**, demonstrating its effectiveness in predicting sales while minimizing error, which is crucial for inventory management.

Lower Variance

Unlike more complex models, Linear Regression exhibited **lower variance**, ensuring consistent performance across different datasets and providing greater reliability in sales forecasting for decision-making processes.

Business Insights

Data-driven insights to guide pricing and marketing decisions

Multiple Factors

Sales patterns are driven by multiple interacting factors affecting overall demand dynamics.

Meaningful Variation

Linear relationships capture a meaningful portion of demand variation across different products.

Complex Models

More complex models do not provide significant performance gains over simpler alternatives.

Reliable Forecasts

Reliable forecasts can still support inventory and planning decisions effectively for business.

Simple Models

Simple, interpretable models perform competitively on this dataset, offering useful insights.

Limitations

Recognizing the study's constraints

Dataset Size

The limited dataset size of approximately 1,000 observations may restrict the model's ability to generalize, affecting its reliability in predicting sales for a broader range of products.

Seasonal Effects

The absence of seasonal effects in the dataset means that important patterns related to seasonal demand fluctuations are not captured, potentially leading to inaccurate forecasts during peak sales periods.

Future Work

Enhancing predictive modeling capabilities

Time-Series Features

Incorporating **time-series features** can improve accuracy by capturing trends and seasonality, allowing more nuanced predictions and better alignment with actual sales patterns across time periods.

Gradient Boosting

Exploring **gradient boosting** methods may enhance performance by addressing complex nonlinear relationships in the data, leading to better adaptability and improved predictive results compared to traditional models.

Thank you for your attention!

Let's discuss any questions you may have.