18/01/2025

RETAIL TRANSACTION ANALYSIS AND TOTAL PRICE PREDICTION

PREPARED BY,
SHIBILA KP

TABLE OF CONTENTS

1 Introduction

2 Dataset Overview

3 Data Preprocessing

4 Feature Engineering

5 Data Visualisation

6 Model building and Validation

7 Model Evaluation

8 Challenges faced

9 Applications

10 Conclusion

INTRODUCTION

This dataset focuses on understanding and predicting the Total Price of customer transactions by analyzing product details, customer demographics, operational metrics, and sales performance. The primary goal is to extract insights into purchasing behavior, profitability, and business strategies in a retail context.The dataset provides a comprehensive view of retail operations, offering details such as customer satisfaction, discounts, and annual income, which are crucial for understanding purchasing behavior.

DATASET OVERVIEW

Dataset Details:

- Rows: 50 k
- Columns: 20
- Features :

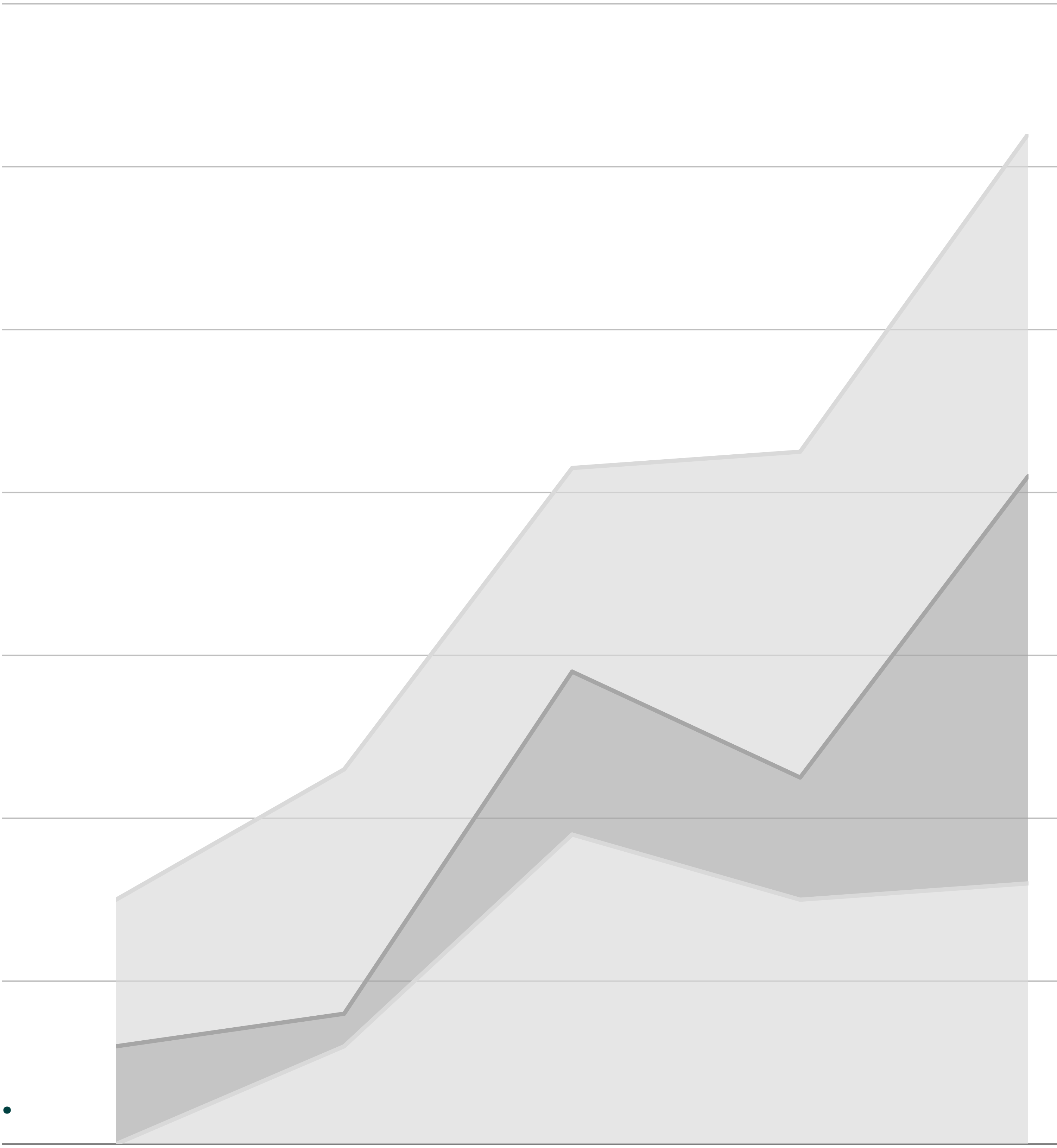
  It includes 20 features such as transaction identifiers, customer age and income, product categories, applied discounts, profit margins, and payment methods.

- Purpose :

  The dataset is designed to provide comprehensive insights into retail business operations by analyzing transactions, customer demographics, product details, and sales performance. Its primary purpose is to enable predictive modeling, such as forecasting the total price of transactions, optimizing pricing strategies, and understanding customer behavior. By leveraging this dataset, businesses can make data-driven decisions to enhance revenue, improve customer satisfaction, and streamline operational efficiency.

DATA PREPROCESSING

- **Steps :**
    - Handling Missing Values:
        - Filled missing dates with mode.
        - Imputed numerical features with mean.
- **Outlier Treatment:**
    - Identified using box plot
    - Removed extreme outliers from Total price
- **Categorical Encoding:**
    - Converted categorical values into numerical values
- **Scaling:**
    - Standardized numerical features using StandardScaler.



FEATURE ENGINEERING

- **Created features**
  - **Effective_Price** : Adjusted price after applying the discount.
    Formula : Product_Price * (1 - Discount (%) / 100)
  - **Total_Product_Value** : Pre-discount total price base quantity.
    Formula : Product_Price * Quantity
  - **Discounted_Total_Price** : Total price after discount for all quantities.
    Formula : Total_Product_Value * (1 - Discount (%) / 100)
  - **Profit** : Contribution to profit after applying discounts.
    Formula : Discounted_Total_Price * (Profit_Margin (%) / 100)

- **Impact of Features:**
  - Helped identify product-level profitability.
  - Improved model performance by capturing pricing dynamics.

DATA VISUALISATION

Histograms: Analyze the distribution of numerical features like Total_Price to identify skewness or outliers.

- The presence of peaks reflects frequent pricing tiers, possibly due to discounts or popular product categories.
- The distribution of Total_Price is slightly right-skewed, indicating some high-value transactions.

Correlation Heatmap: Identify relationships between numerical features and their influence on Total_Price

- Moderate correlation observed between Discount (%) and Profit, showing discounts influence profitability.
- Strong positive correlation between Total_Price and Quantity, indicating that higher quantities drive total sales.
- Weak correlations between demographic attributes like Customer_Age and sales metrics, suggesting limited impact.

Box Plot: Examine the relationship between Quantity and Total_Price, and identify outliers in transactions

- A positive relationship is visible between Quantity and Total_Price, with higher quantities generally leading to higher total prices.
- Some extreme outliers in higher Total_Price indicate transactions involving expensive products or bulk purchases.
- The variability in Total_Price is wider for larger quantities, suggesting a diverse product mix.

Pair Plot: Explore pairwise relationships among key numerical features to uncover interactions.

- Product_Price vs. Total_Price: A direct positive relationship is observed, as higher product prices increase the total transaction value.
- Quantity vs. Total_Price: Larger quantities result in a significant increase in the total price, as expected.
- Discount (%) vs. Total_Price: Discounts show varying impacts, with reduced prices sometimes correlating with increased quantities and total price.
- Feature Interactions: These pairwise plots highlight potential multi-collinearity and non-linear relationships to consider during modeling.

MODEL BUILDING AND VALIDATION

Regression Models Used :

- Linear Regression
- Random Forest Regression
- XGB Regression
- KNN
- Decision tree regression

Why these models?

- The dataset contains a mix of numerical and categorical data, making ensemble methods (Random Forest, XGBoost) effective.
- Non-linear relationships between variables are well-suited for tree-based models.
- Linear Regression provides a simple baseline for comparison.

Feature Engineering and Scaling Considerations:

- Categorical features were encoded for compatibility with all models.
- Numerical features were scaled for models like SVR and KNN that are sensitive to feature magnitudes.

HYPERPARAMETER TUNING

METHODOLOGY:

- Used GridSearchCV for hyperparameter tuning
- Cross-validated with 5-fold or 3-fold CV.

LINEAR REGRESSION:

- Simpler model; minimal hyperparameter tuning
- Regularization can be added (e.g., Lasso or Ridge) with tuning for alpha.

RANDOM FOREST REGRESSION:

KEY PARAMETERS USED:

- n_estimators: Number of decision trees in the forest.
- max_depth: Maximum depth of each tree (prevents overfitting).
- min_samples_split,min_samples_leaf: control the size of splits and leaf.
- **Impact of Tuning: Balances overfitting (deep trees) and underfitting (shallow trees**).

XGBOOST REGRESSION:

- learning_rate: Controls step size for each update.
- max_depth: Depth of the trees for capturing complex patterns.
- n_estimators: Number of boosting rounds.
- **Impact of Tuning: Fine-tuning these parameters improves convergence and generalization.**

K-Nearest Neighbors (KNN):

KEY PARAMETERS USED:

- n_neighbors: Number of neighbors considered for prediction.
- weights: Weighting scheme for neighbors (uniform or distance-based).
- metric: Distance metric used for finding neighbors (e.g., Euclidean, Manhattan).

- **Impact of Tuning: Helps balance bias (low k) and variance (high k).**

Decision Tree Regression

- max_depth:Maximum depth of the tree. Controls overfitting (deep trees) and underfitting (shallow trees).
- min_samples_split,min_samples_leaf: control the size of splits and leaf.

- **Impact of Tuning: ignificantly influences the model's performance, interpretability, and generalization capabilities**.

MODEL EVALUATION

Metrics Used:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE).
- R² Score.

Comparison:

	RMSE	MAE	R2
Linear Regression	36.965680	2.375115	0.999487
Random Forest	34.848228	0.878707	0.999544
KNN	76.951362	50.170391	0.997776
XGBoost	0.997776	7.048143	0.999400
Decision tree	45.024114	1.075409	0.999239

Best Performing Model : Random Forest .

CHALLENGES FACED

Data Quality Issues:

- Missing values in critical columns like Date.
- Presence of outliers in Total_Price and Profit_Margin (%).

Model Performance:

- SVR struggled with high-dimensional data
- Random Forest required significant computational time for tuning.

Feature Interactions:

- Non-linear relationships among features were difficult to capture.

APPLICATIONS

Real-World Use Cases:

- Revenue Forecasting: Predict total revenue based on historical data.
- Price Optimization: Optimize pricing strategies to maximize profit
- Customer Insights: Segment customers for targeted marketing.
- Inventory Management: Predict demand to reduce overstock or stockouts.
- Sales Channel Analysis: Evaluate the performance of different sales channels
- Customer Satisfaction Analysis: Explore the impact of factors like discounts, product categories, and prices on customer satisfaction.

CONCLUSION

This retail dataset provides valuable insights into customer behavior, product performance, and sales trends. By analyzing key attributes like pricing, discounts, and customer demographics, we developed predictive models to forecast total price accurately. These insights support strategic decisions in revenue optimization, customer segmentation, and inventory management, highlighting the dataset’s importance in driving data-driven growth and efficiency in the retail sector.

This dataset not only aids in solving immediate business challenges but also provides a framework for long-term planning and growth, showcasing the importance of leveraging data for achieving a competitive advantage in the retail sector.