Info Bharat Interns

Internship Project on Customer & Sales Data Analysis

by Tejaswini Kandhivanam

Customer & Sales Data Analysis

1) Project Overview

This project is focused on analyzing a synthetic e-commerce dataset to derive valuable business insights. The core objectives are to:
-Uncover patterns in customer behavior.
-Segment customers based on transactional data.
- Forecast sales trends using time series models.
- Predict churn using machine learning algorithms.
- Suggest cross-selling strategies using market basket analysis.

2)Data Preprocessing & Feature Engineering

- Handling Missing Values:
  - For KNN Imputer, the missing values are estimated based on the nearest K nearest neighbors (default k=5) in the feature space.
  - For Iterative Imputer, each feature with missing values is modeled as a function of the other features and imputed iteratively.
  - For MICE, we implemented MICE/iterative Imputation, using `Iterative Imputer` from `sklearn`. For MICE, we modeled the conditional distributions iteratively.

- Outlier Detection & Treatment:
  Tukey's Method: IQR (Interquartile Range) rule to check for outliers:
  - Outliers are values smaller than Q1 - 1.5×IQR or larger than Q3 + 1.5 IQR.
  - Robust Z-score: computes Z-scores based on median and MAD. We flagged Z-scores that had a |Z| score larger than 3.
  - Treatment: we capped the outliers, we did log or square root transformations

- Scaling and Normalization:
  - StandardScaler: standardizes numerical features to have zero mean and unit variance.

- Min Max Scaler: normalized the range of values of each feature is between [0, 1]. Neural Models can be sensitive to distributions of data, Feature Scaling is recommended to be done prior to normalization (e.g., Min Max Scaler).

- Feature Engineering:
  - Average Purchase Frequency: Total orders / active days.
  - Customer Lifetime Value (CLV): CLV = Average Order Value x Purchase Frequency x Customer Lifespan.
  - Recency Score: Number of days since a customer last completed a transaction.
  - Loyalty Score: Based on repeated purchases that happened over time.
  - Discount Utilization Rate: % of purchases made that took advantage of discounts.
  - Preferred Payment Method: encoded as frequency of use, historically.

3. Exploratory Data Analysis (EDA)

- Temporal Trends

- Time of day: Grouped sales by hour to check transaction tendencies and find peak transaction periods.

- Day of week: Found patterns such as weekend spikes, weekday lulls, etc.

- Month/year trends: Created line and area plots to track seasonal sales trends such as seasonal dips.

- Seasonal Decomposition: Conducted STL Decomposition to break apart trends, seasonal pathways, and residuals.

- Demographic Insights

- Purchasing by gender: Found which categories are more gender-identified than others.

- Younger purchasing characteristics: Segment customers into 5-year bins and analyze relative preference to buying.

- Income segmentation: Compared income levels to purchase power and discount thresholds.

- Products

- Most productive products: by revenue, units sold, and net profit.

- Popular categories: patterns in popularity and conversions across categories.

- Discounts outcome: measured uplift of conversions and statistical test them due to discounts.

- Correlation study

- Created a heatmap of Pearson correlation to research relationships between features and possible multicollinearity.

4. Customer Segmentation

- Clustering Techniques
  - Gaussian Mixture Models (GMM): Probabilistic model for soft clustering.
  - Agglomerative Clustering: Hierarchical clustering using linkage criteria.
- RFM Analysis
  -Recency: Days since last purchase.
  - Frequency: Number of transactions.
  - Monetary: Total revenue generated by the customer.
- Extended Segmentation
  Combined RFM with:
  -Loyalty Score
  - Discount Utilization Rate
  - Preferred Payment Method

- Segment Profiling
  - Labeled clusters based on value tiers (e.g., VIP, At-Risk, New, Occasional).
  - Suggested personalized strategies:
    - Loyalty programs for VIPs.
    - Re-engagement campaigns for dormant users.

4. Sales Forecasting

Time Series Models:

- Seasonal ARIMA (SARIMA):

Automated using grid search for p, d, q, P, D, Q.

- Prophet Model:

Included country-specific holidays, and included product launches.

Captureted yearly, weekly and daily seasonality.

- LSTM:

Sequenced modelling with look-back window and dropout regularization.

- MinMaxScaler scaling performed to ensure consistent data for RNN implementation.
- Evaluation Metrics

-RMSE: Root Mean Squared error

-MAE: Mean Absolute Error

-MAPE: Mean Absolute Percentage Error

Business Use:

Detected peak forecasted demand and expected declines.

5. Predictive Modeling: Customer Churn

Algorithms Used:

Logistic Regression was adopted as the baseline binary classifier to predict whether a customer is likely to churn.

Decision Tree and Random Forest to capture non-linear relationships and examine feature importance predicting churn.

XGBoost, an ensemble and gradient-boosting technique, was used for its speed and accuracy to obtain an optimal churn prediction model.

Model Optimization:

GridSearchCV was used to systematically search through multiple combinations of parameters, optimizing the model settings.

RandomizedSearchCV expedited tuning by randomly sampling potential parameter combinations.

Hyperparameters tuned include tree depth, number of estimators, and learning rate.

Evaluation Tools:

Confusion Matrix visualized true positives, false positives, true negatives, and false negatives.

ROC-AUC Curve evaluated classifier performance in distinguishing between churn and non-churn classes.

Precision-Recall Curve yield information about model performance, particularly with imbalanced data.

Identified and ranked features driving customer churn.

6. Product Analysis and Cross-Selling

Association Rule Mining:

Frequent itemsets generation and strong product combinations derivation were accomplished through the use of the Apriori Algorithm with support and confidence thresholds.

FP-Growth was then used to mine frequent patterns as a faster and more scalable algorithm.

The strength of products association were determined using lift and confidence metrics.

Product Profitability:

The net profit was calculated for each product by taking the selling price and subtracting the cost price and the discount price.

Products that were sold at a loss (loss-leaders) were identified at the same time as products with the highest margins that account for most of the profits.

Understanding product profitability led to more thoughtful pricing and discounting strategies.

Recommendations:

Bundling complementary products (i.e. those purchased together) was suggested to improve average order value.

Dynamic cross-sell suggestions using association rules at checkout were implemented to help increase conversion rates and revenue.

7. Reporting and Visualization

- Visual Tools Used:

For exploring trends and communicating insights, interactive charts and dashboards were built in Plotly.

For data analysis and reporting, statistical plots (i.e. heatmaps, distribution plots, and boxplots) were built in 'seaborn'.

Customized visualizations with detailed formatting and layout control were built using matplotlib.

- Report Structure:

Executive Summary, with an overview of the overall objectives and results of the analysis.

A description of the overall project and approach taken.

A section illustrating different analysis conducted.