PACE UNIVERSITY

CS 667 Practical Data Science
Shopping Trends

Date: 4/1/25
Name: Priyansh Parmar
Email: pp91088n@pace.edu
Class Name: Practical Data Science
Program Name: M.S. in Data Science
University Name: Pace University

Agenda

- Data and EDA deck
- Methods: Technical Version
- Findings
- Recommendations
- GIT Repo: Link

Summary

Business Problem Summary:

The business problem at hand revolves around predicting customer purchasing behavior and understanding key factors that influence purchase decisions. Specifically, the goal is to develop a model that can predict whether a customer will make a purchase based on various features such as promo code usage, purchase amount, location, item category, and previous purchase history. By leveraging these insights, the business can optimize marketing strategies, personalize offers, and improve customer retention.

Objectives:

To tackle this challenge, we will develop a robust sales trend model designed to identify feature importance. Feature importance helps businesses focus on the most influential factors driving sales, enabling smarter decision-making. Feature importance helps businesses identify the key factors driving sales and customer behavior. By understanding which features (e.g., seasonality, discounts, customer demographics) influence sales the most, businesses can optimize pricing strategies, marketing efforts, and inventory planning. For example, if holiday season has high importance, companies can allocate more stock and promotions during that period. This data-driven approach enhances decision-making and improves overall profitability.

DATA

Recap

- Data and EDA
- Methods, Findings and Recommendations
- Final

Data

Data link: https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data
The dataset contains 3,900 rows. Each row represents a single shopping transaction made by a customer, capturing details such as:

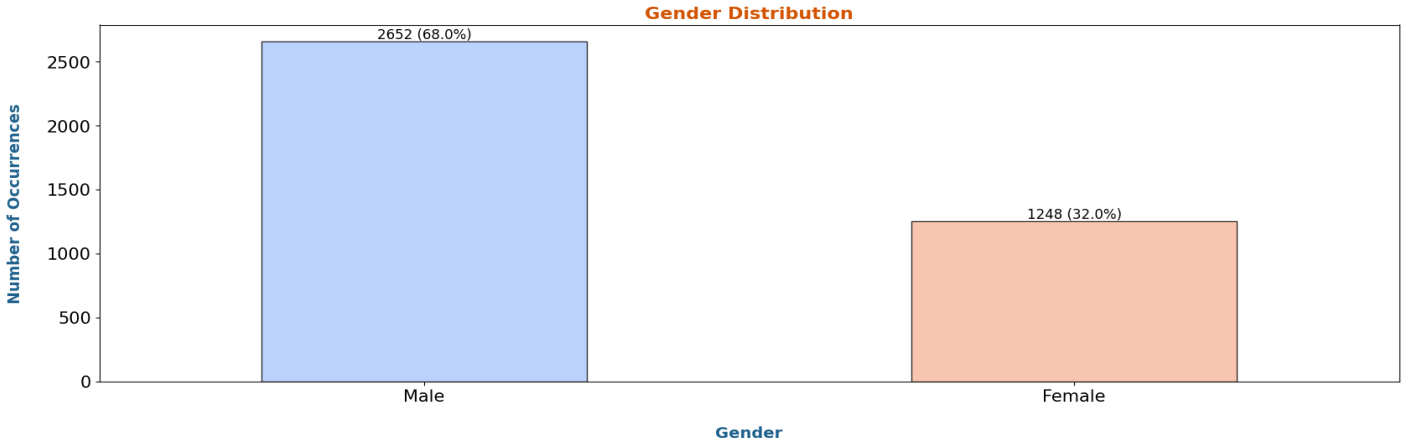
- Customer ID, Age, Gender
- Item Purchased, Category, Purchase Amount
- Location, Size, Color, Season
- Review Rating, Subscription Status
- Payment Method, Shipping Type, Discount Applied
- Promo Code Used, Previous Purchases, Preferred Payment Method
- Frequency of Purchases

Time Period: There's no column with time period and dates

Assumptions: The dataset assumes that it accurately represents customer shopping behavior, including purchase frequency, payment methods, and promo code usage. It is presumed that all recorded transactions reflect real purchases without missing or incorrect data. Seasonal shopping trends, if applicable, are assumed to be captured, ensuring a complete view of consumer behavior. Additionally, key factors influencing sales, such as location and previous purchases, are considered comprehensive and relevant. Lastly, it is assumed that external influences, like economic conditions, do not significantly distort purchasing patterns

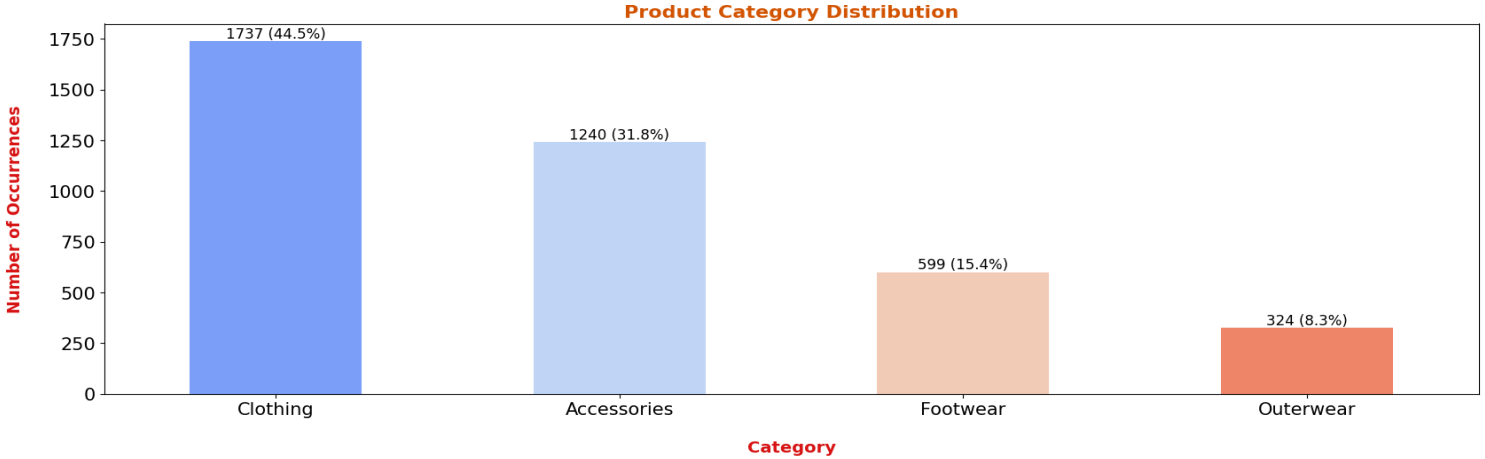
EDA

EDA



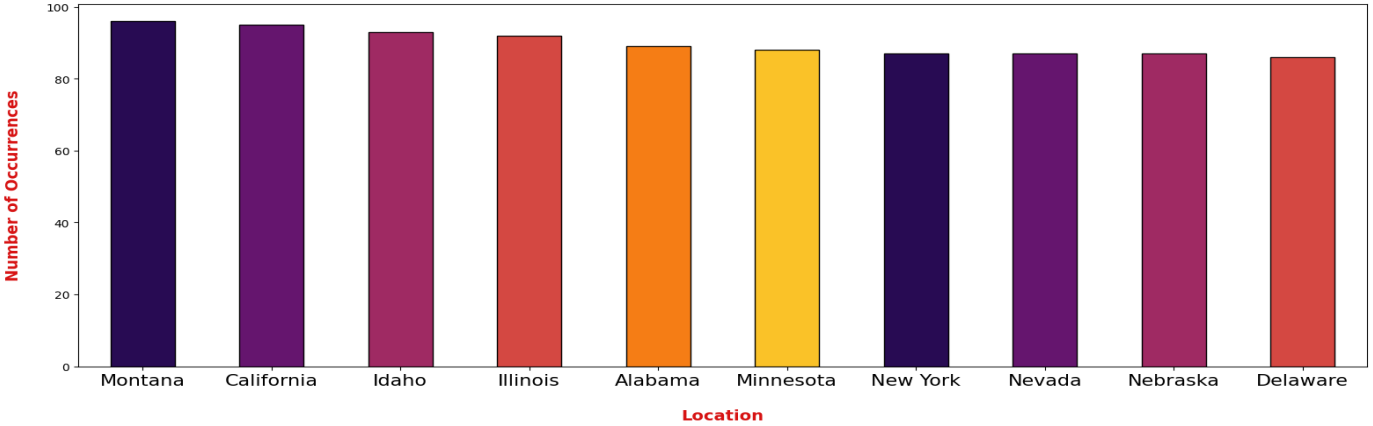
- The Male category has a significantly higher count (2652 occurrences) compared to Female (1248 occurrences).
- The Male segment covers 68% of the total sales data, while the Female segment accounts for 32%
- This reflects product preferences, so the business may need to consider gender-based product customization or marketing adjustments.

EDA



- The bar graphs illustrate the distribution of product sales across four categories: Clothing, Accessories, Footwear, and Outerwear.
- Clothing has the highest sales volume counting to 44.5% of the total sales, followed by Accessories, while Footwear and Outerwear have significantly lower sales figures.
- The data suggests that Clothing and Accessories dominate the market, whereas Footwear and Outerwear have relatively lower demand.

EDA

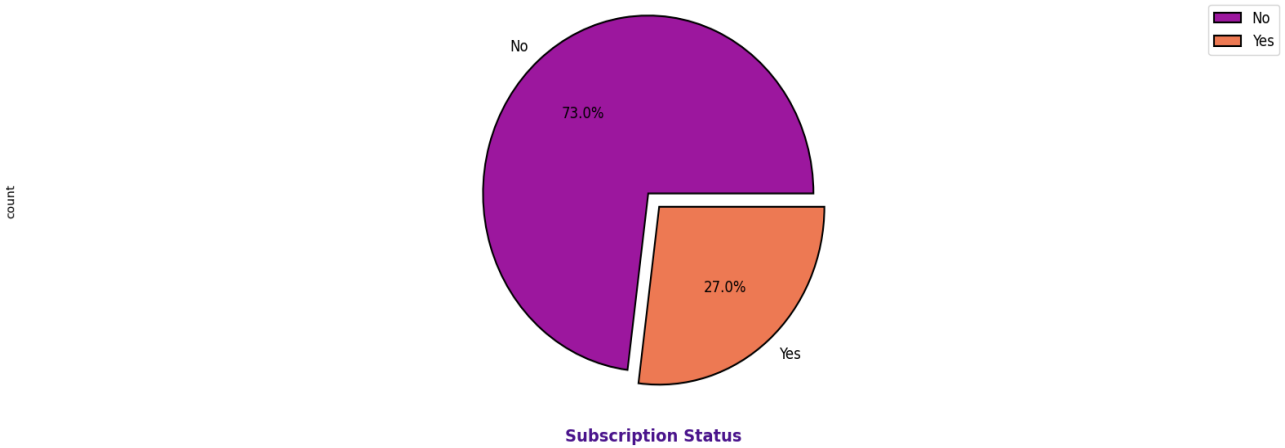


- The bar chart represents the number of occurrences across different U.S. states, including Montana, California, Idaho, Illinois, Alabama, Minnesota, New York, Nevada, Nebraska, and Delaware.

- The occurrences are relatively high and evenly distributed, with slight variations between states. Montana has the highest number of occurrences, while Delaware has the lowest.

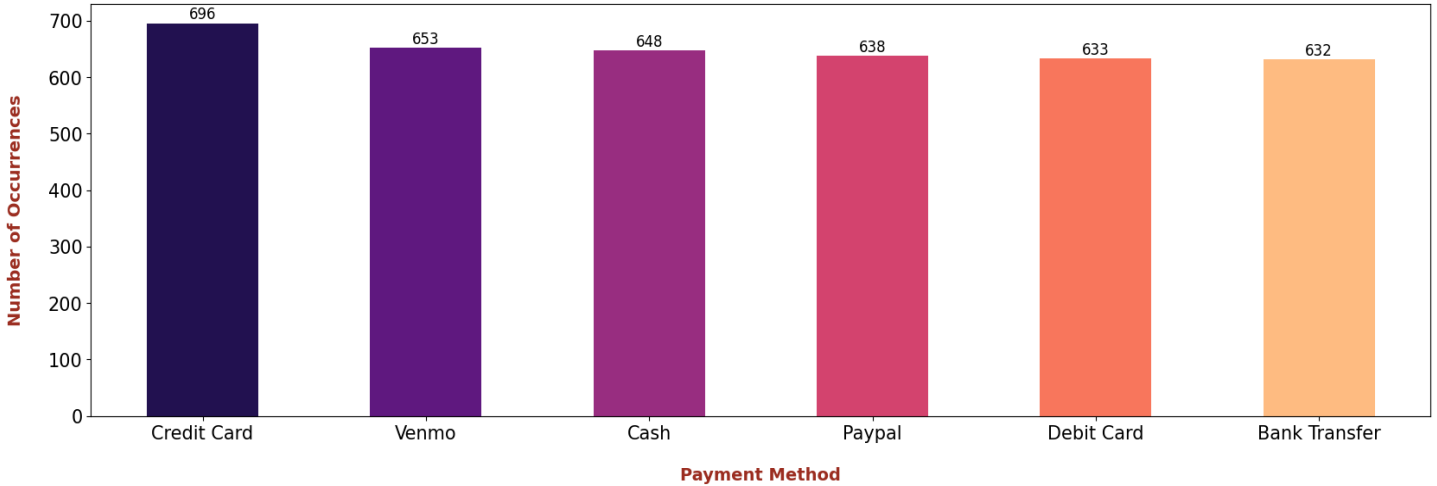
- The colors used in the bars provide a visually appealing gradient effect. The labels and axis titles are clearly marked in red for emphasis.

EDA



This pie chart illustrates the distribution of subscription status among customers. It shows that 73% of customers are non-subscribers, while only 27% have an active subscription. This indicates that a majority of shoppers prefer to make purchases without subscribing, suggesting that the subscription model might not be widely adopted. Businesses could explore ways to enhance subscription benefits, offer exclusive deals, or improve marketing strategies to increase subscription rates and boost customer retention.

EDA



This bar chart displays the distribution of payment methods used by customers. Credit cards are the most commonly used payment method, followed by Venmo and Cash, indicating a preference for digital and convenient transactions. PayPal, debit cards, and bank transfers also show significant usage, suggesting that customers prefer multiple payment options. Businesses can optimize their checkout process by ensuring seamless transactions for the most preferred payment methods.

MODELING

Modeling

The model used is a Decision Tree Classifier, trained on a dataset with features such as "Age," "Item Purchased," "Location," and "Promo Code Used." The model achieved an accuracy of 73%, with a higher precision and recall for class 1 (positive class). It also includes performance evaluation metrics like the F1 score, confusion matrix, and ROC curve, which help assess the model's effectiveness in distinguishing between classes. Additionally, the feature importance values highlight that "Promo Code Used" and "Purchase Amount (USD)" were the most influential factors in predictions.

The outcome variable in this model is predicting whether a customer will make a purchase (1) or not (0). By analyzing factors such as promo code usage, purchase amount, location, and other customer characteristics, the model aims to forecast purchase behavior. The decision tree model's feature importance highlights that variables like promo code usage and purchase amount are the most influential in predicting outcomes.

The features selected for the Decision Tree model focus on key factors like purchase amount, item purchased, and promo code usage, which directly influence customer buying decisions. The model highlights that promo code usage and purchase amount are the most significant predictors, supporting the hypothesis that discounts and spending levels drive purchasing behavior. Other features like "Location" and "Color" also contribute, though less prominently.

Modeling

Non Technical Version:

- In simple terms, this model looks at several factors to predict customer purchase behavior.
- For instance, if someone uses a promo code or spends a certain amount, they are more likely to make a purchase.
- Other factors, like the type of item they buy, where they are located, or even the color of the product, also help in predicting what they might choose to buy.
- For example, a customer who uses a discount code and spends a larger amount of money is more likely to make a purchase than someone who doesn't.
- Similarly, if a person is located in a region where certain products are more popular, that could influence their buying decision.
- For more details you can go on to this slide

Modeling

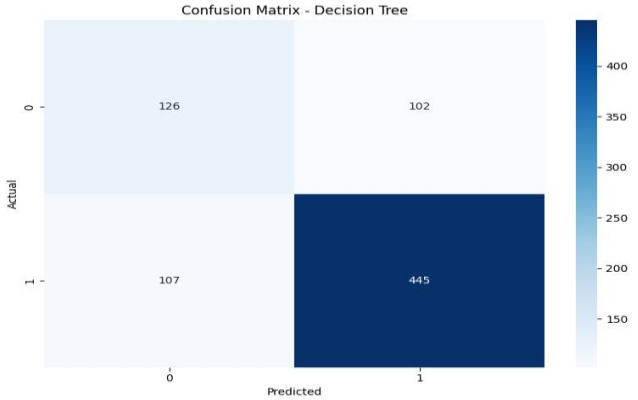
Technical Version:

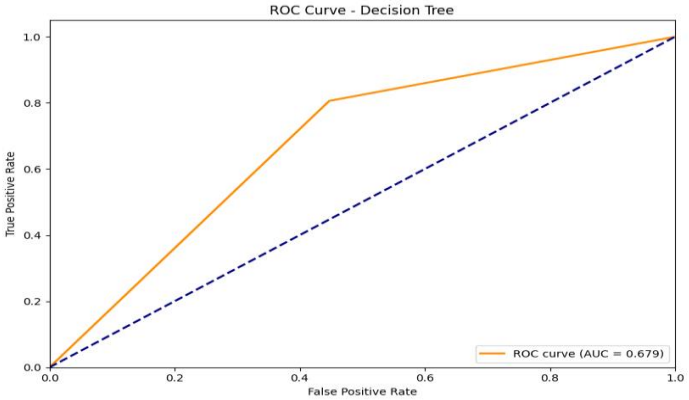
In this Decision Tree model, several features were chosen based on their potential influence on consumer purchasing behavior, derived from both domain knowledge and exploratory data analysis. Features such as "Purchase Amount (USD)" and "Promo Code Used" were hypothesized to be strong predictors, as higher spending and discount application are commonly linked to purchasing decisions. The model's feature importance ranking aligns with this assumption, with "Promo Code Used" showing the highest importance (0.36), followed by "Purchase Amount (USD)" (0.08).

The Decision Tree algorithm works by recursively partitioning the feature space to create splits that maximize the information gain at each node, resulting in the creation of a tree structure that can predict the target variable. Feature importance is calculated based on how often a feature is used to split the data at each node, weighted by the reduction in impurity (Gini Index or Entropy). One interesting finding is the relatively low importance of features like "Subscription Status" and "Discount Applied," which suggests that they may not directly affect the purchasing decision as strongly as initially anticipated. This could imply that customers either disregard these factors in their decision-making or that other, more influential features are overshadowing their impact. Overall, this decision tree model provides both theoretical insights into customer behavior and practical implications for targeted marketing strategies.

FINDINGS

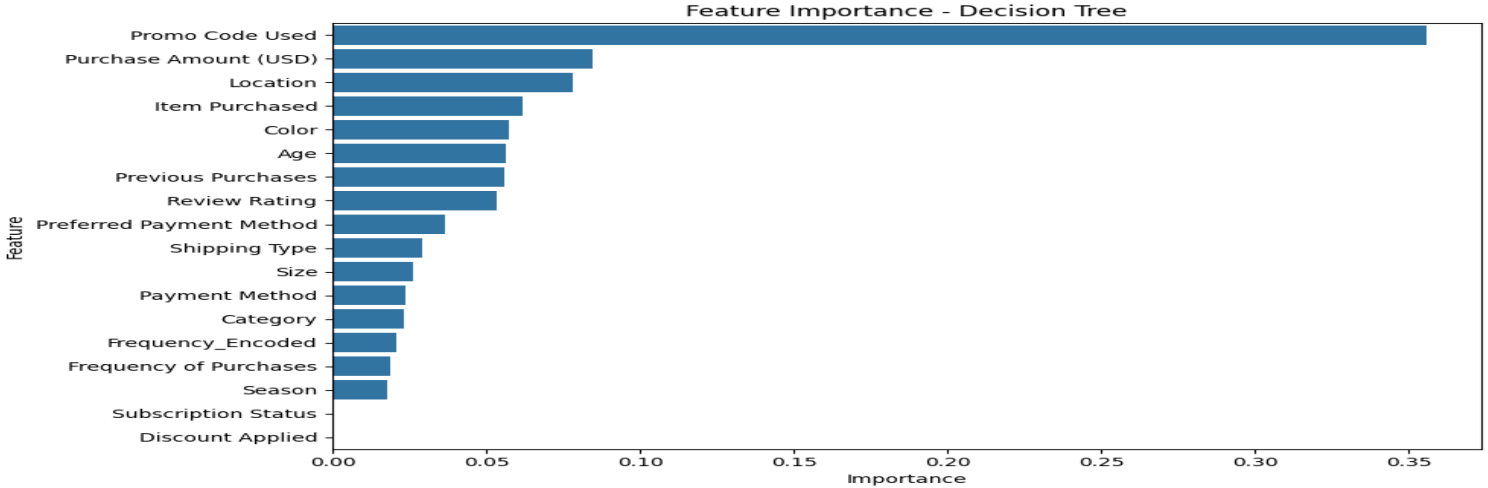
Findings





- The model has a good ability to correctly identify positives (1s), as shown by the 445 true positives.
- The ROC curve (Receiver Operating Characteristic curve) plots the True Positive Rate (Sensitivity/Recall) against the False Positive Rate (1 - Specificity) at different threshold values.
- The AUC (Area Under Curve) = 0.679, which indicates moderate performance. A perfect model would have an AUC of 1.0, meaning it perfectly separates the two classes.
- A random classifier would have an AUC of 0.5 (diagonal dashed line).

Findings



- The feature importance chart shows that "Promo Code Used" is the most influential factor in the Decision Tree model, significantly impacting predictions. Other key features include "Purchase Amount (USD)," "Location," and "Item Purchased," which also contribute notably. Less important features, such as "Subscription Status" and "Discount Applied," have minimal influence on the model's decisions. This suggests that promotional activity and purchasing behavior play a crucial role in classification outcomes.

BUSINESS RECOMMENDATION

Business Recommendation

Promo Code Used (36% Importance)

- Business Insight: Customers using promo codes have a significantly higher impact on purchase amounts.
- This suggests that discounts & promotions are a strong driver of sales.

Purchase Amount (8% Importance)

- Business Insight: Past purchase amounts influence future buying behavior.
- High-spending customers may continue to make significant purchases, showing strong repeat customer potential.

These findings highlight that promo codes drive conversions, while past purchase behavior helps predict future spending. Businesses should leverage targeted promotions and personalized marketing to boost sales and customer retention.

Business Recommendation

Actionable Recommendation: Based on the model's feature importance, Promo Code Usage emerged as a critical factor influencing customer purchases. To capitalize on this insight, the business should implement targeted marketing strategies that promote the use of promo codes to increase sales. Specifically, the business could use Personalize Promo Code Offers: Identify customers with a high likelihood of responding to promo codes (based on previous purchase behavior or demographics like location and age) and send personalized promo code offers via email, SMS, or app notifications.

Building a more advanced model could provide deeper insights into customer behavior, address limitations of the current model, and improve prediction accuracy by incorporating additional features, advanced algorithms (e.g., neural networks), or time-series analysis. This would help answer open questions and refine business strategies further.

The Decision Tree model has limitations, as shown by its moderate AUC score (0.679), indicating that a more sophisticated model might better capture complex relationships and reduce misclassification.