Retail Sales and Customer Behavior Analysis – Project Summary

This project focuses on utilizing machine learning techniques to analyze customer behavior and predict churn in a retail setting. The data used in this project comes from a retail sales and customer transaction database and includes input variables such as age, gender, location, product category, total spending, payment methods, and channel of purchase.

Objective

The primary goal of the project is to develop a classification model that can accurately predict whether a customer will churn (stop purchasing) or not. By applying various machine learning algorithms, the model aims to identify which customers are likely to leave, enabling the business to take proactive retention actions. This analysis also helps uncover key factors driving customer churn behavior, supporting data-driven marketing and engagement strategies.

Dataset Overview

The dataset consists of 77 features and several thousand observations related to customer purchases, product categories, and behavioral metrics.
Key features include:

- Demographic variables (e.g., age, location)
- Behavioral data (e.g., product preferences, frequency of purchases)
- Channel data (e.g., online vs offline purchase)
- Churn label (churned – the target variable)

Data Characteristics:

- Total records: [actual count inferred during EDA]
- No duplicate rows found
- Contains both categorical (nominal and ordinal) and numerical features
- Several columns with "unknown" or missing values, treated as null
- Features with excessive missing values were dropped
- Remaining missing values were handled using:
    - Mode for categorical features
    - Median for numerical features

Exploratory Data Analysis (EDA)

Descriptive statistics and visualization techniques were used to better understand the distribution and relationships among variables.

Tools used:

- Countplots, distplots, boxplots, and heatmaps for detecting trends and outliers
- Correlation matrix for numerical features
- Identified strong relations between churn and variables like channel, location, and product category

---

Feature Engineering

- Outliers in variables such as balance, duration, campaign, and previous were treated using the interquartile range (IQR) method
- Label encoding was used for binary and ordinal categorical features
- One-hot encoding was applied for nominal categorical features with multiple categories
- Feature scaling was applied using MinMaxScaler
- Feature selection performed using mutual information classifier and model-based importance scoring

---

Handling Class Imbalance

The dataset showed significant class imbalance, with only a small portion of customers labeled as churned. To address this:

- SMOTE (Synthetic Minority Oversampling Technique) was used to oversample the minority class and balance the dataset

---

Machine Learning Models Used

Multiple classification algorithms were trained and evaluated:

- Logistic Regression
- Decision Tree

- Random Forest
- Gradient Boosting Machine (GBM)
- XGBoost
- K-Nearest Neighbors
- Naive Bayes
- Support Vector Machine (SVM)
- Artificial Neural Network (ANN) using Keras

Each model was:

- Trained using train-test split and cross-validation
- Evaluated using:
    - Accuracy
    - Precision
    - Recall
    - F1 Score
    - ROC-AUC
    - Confusion Matrix (visualized)

Best Model: XGBoost

The XGBoost model performed best among all classifiers with the following metrics:

- Accuracy: 0.93
- Precision: 0.93
- Recall: 0.93
- F1 Score: 0.93
- ROC-AUC Score: 0.93

These metrics indicate that the XGBoost model is highly reliable for predicting customer churn with strong generalization on unseen data.

Model Interpretation with SHAP

- The XGBoost model was interpreted using SHAP (SHapley Additive Explanations)
- Top contributing features included:
    - Purchase channel
    - Month of last transaction

    - Product category
    - Total spend
    - Number of purchases
- SHAP summary plots revealed:
    - Lower values of certain features (e.g., no loans, less product returns) increase likelihood of retention
    - Higher values of other features (e.g., longer inactivity, high returns) increase churn probability

---

Insights from the Data

- Customers without personal or housing loans are less likely to churn
- Customers with higher purchase frequency and call durations are more likely to stay
- Customers who were contacted fewer than 3 times had a higher conversion rate
- Marketing channel and month of contact had a significant impact on retention
- Educated and debt-free customers with regular interaction are most loyal
- Only around 12% of customers churned — indicating a strong imbalance in behavior

---

Challenges Faced

- Dealing with missing and unknown values
- Class imbalance, which initially led to biased model predictions
- Model tuning and avoiding overfitting in complex models like ANN
- Ensuring interpretability of complex models for business use

---

Conclusion

The project successfully implemented a full pipeline for churn prediction using real-world retail customer data. By testing and tuning various models, XGBoost emerged as the most effective classifier. The use of SHAP enhanced transparency in prediction, helping stakeholders understand why a customer may churn. This analysis can help retailers optimize targeting, retain valuable customers, and design smarter loyalty programs.