Retail Sales and Customer Behavior Analysis Using Machine Learning

Piyush Harchandani

Abstract :

This project aims to predict customer churn in a retail environment using machine learning techniques. A real-world dataset consisting of demographic, behavioral, and transactional data was used to train several classification models. Through comprehensive data preprocessing, feature engineering, and model evaluation, XGBoost was identified as the most effective classifier. The results demonstrate the potential of predictive analytics in helping retail businesses proactively manage customer retention strategies.

Table of contents :

List of terms :

- **Churn**: When a customer stops purchasing from the business.
- **EDA (Exploratory Data Analysis)**: The process of analyzing datasets to summarize main characteristics.
- **SMOTE (Synthetic Minority Oversampling Technique)**: A technique to handle class imbalance by creating synthetic data.
- **ROC-AUC**: A performance measurement for classification problems at various thresholds.
- **SHAP**: A tool for explaining the output of machine learning models.
- **ANN**: Artificial Neural Network.
- **XGBoost**: A scalable and highly efficient machine learning algorithm for classification and regression.

Acknowledgments :

I would like to thank all those who provided guidance and support during this project. Special thanks to open-source communities and contributors whose tools and libraries such as Scikit-learn, Keras, XGBoost, and SHAP made this analysis possible.

Problem Statement :

Customer churn has a direct impact on a retail business's revenue and growth. Identifying customers who are likely to stop purchasing allows businesses to take timely action. This project addresses the problem of churn prediction using machine learning techniques and aims to build a reliable model that can classify customers as likely to churn or not based on historical data.

Retail Sales and Customer Behavior Analysis Using Machine Learning

Piyush Harchandani

Introduction :

In the competitive retail landscape, retaining customers is more cost-effective than acquiring new ones. With access to large datasets capturing customer behavior, machine learning can offer predictive insights to support retention strategies. This project builds a predictive model using real customer data and analyzes key features that influence churn.

Background :

Customer churn prediction is a critical application in the retail and marketing domains. Various statistical and machine learning models have been successfully applied to such problems in telecom, banking, and e-commerce sectors. With the evolution of tools like SHAP and SMOTE, models can now not only predict with high accuracy but also provide interpretable insights.

Materials and apparatus:

- **Software & Libraries**:
  - Python 3.x
  - Pandas, Numpy
  - Scikit-learn, XGBoost, Keras
  - Matplotlib, Seaborn
  - SHAP, SMOTE
- **Hardware Requirements**:
  - Any modern system with at least 8GB RAM and multi-core CPU
  - Jupyter Notebook environment or any IDE supporting Python
- **Dataset**:
  - Retail customer behavior and transaction dataset (CSV format)

Procedure :

---

1. Data Acquisition and Loading

- Imported the dataset into a Jupyter Notebook using Pandas.
- Performed initial inspection (head(), info(), describe()) to understand the structure, data types, and volume.
- Verified the number of records and identified missing or anomalous values.

---

2. Data Cleaning

- **Missing/Unknown Values**:
  - Features like job, education, contact, and poutcome had many "unknown" values. These were treated as missing.
  - Categorical columns with manageable missing values were filled using **mode** (most frequent value).
  - Numerical columns were filled using the **median** to preserve distribution without being skewed by outliers.
  - Features with over 50% missing values were removed due to low predictive power.
- **Duplicates**:
  - Checked and confirmed there were no duplicate rows in the dataset.

---

3. Outlier Detection and Treatment

Retail Sales and Customer Behavior Analysis Using Machine Learning

Piyush Harchandani

- Visualized distributions using **boxplots** and **histograms**.
- Applied the **Interquartile Range (IQR)** method to cap extreme values in numeric features like balance, duration, campaign, pdays, and previous.

---

4. Exploratory Data Analysis (EDA)

- Performed **univariate** analysis to understand individual feature distributions.
- Conducted **bivariate analysis** to identify relationships between features and the target variable (churned).
- Used **heatmaps** to observe correlations between numerical variables.
- Created insightful visualizations using:
  - Countplots (for categorical features)
  - Boxplots (to observe feature distributions with respect to churn)
  - Barplots (for churn frequency across features like job, education, month, etc.)
  - Pairplots (to analyze interaction between multiple variables)

---

5. Feature Engineering

- Converted all **categorical variables** into numerical representations:
  - **Label Encoding** for binary or ordinal variables (marital, default, housing, etc.)
  - **One-Hot Encoding** for nominal variables with multiple categories (job, month, etc.)
- Scaled numerical features using **MinMaxScaler** to bring all values into the [0,1] range for algorithms sensitive to scale.

---

6. Handling Class Imbalance

- Found significant imbalance in the target variable (churned):
  - Only ~12% of customers churned, while ~88% did not.
- Applied **SMOTE (Synthetic Minority Oversampling Technique)** to synthetically generate new churned customer samples to balance the dataset.
- Verified the class distribution post-SMOTE to ensure balance.

---

7. Feature Selection and Importance

- Computed **mutual information** between each feature and the target.
- Removed features with near-zero importance scores.
- Used **model-based feature importance** (Random Forest and XGBoost) to refine selection further.
- Visualized top contributing features.

---

8. Model Building and Evaluation

- Split the dataset into **training and testing** sets (e.g., 80/20 ratio).
- Trained multiple classifiers:
  - Logistic Regression

  - Decision Tree
  - Random Forest
  - Gradient Boosting Machine (GBM)
  - XGBoost
  - K-Nearest Neighbors
  - Naive Bayes
  - Support Vector Machine (SVM)
  - Artificial Neural Network (ANN with Keras)
- Used **GridSearchCV** or manual parameter tuning for hyperparameter optimization where applicable.
- For each model:
  - Computed performance metrics: **accuracy**, **precision**, **recall**, **F1 score**, **ROC-AUC**
  - Evaluated using both train-test split and **cross-validation** to ensure robustness

---

9. Model Interpretation

- Applied **SHAP (SHapley Additive Explanations)** on the best-performing model (XGBoost):
  - Plotted SHAP summary plots to interpret feature contributions.
  - Analyzed which features positively or negatively impacted predictions.
  - Explained model decisions to make the solution business-friendly and interpretable.

---

10. Final Model Selection

- Based on evaluation, **XGBoost** was selected as the final model due to:
  - Highest accuracy (93%)
  - Balanced precision and recall
  - Good AUC and generalization
  - Interpretability via SHAP

Conclusion :

This project effectively built a predictive model to identify churn behavior using real-world retail data. Among all models, XGBoost achieved the best performance with an accuracy of 93%. The use of SHAP enabled interpretable predictions, allowing for actionable business decisions. The approach demonstrated here can be integrated into customer relationship management systems to reduce churn, increase customer loyalty, and enhance long-term profitability.