A large glass building with many windows

AI-generated content may be incorrect.

*CIND 820 – BIG DATA ANALYTICS PROJECT*

*CUSTOMER CHURN PREDICTION IN E-COMMERCE & TELECOMMUNICATION*

*EMINE UYSAL (ID:501304049)*

*The G. Raymond Chang School of Continuing Education*

*Tamer ABDOU, PhD*

*July16th, 2025*

***LIST OF TABLES***

**Table 1:** Features Overview of Ecom Dataset

**Table 2.** Association Strength of Categorical Features with Churn (Cramér’s V Scores) – E-Commerce Dataset

**Table 3.** Features overview of Telco Dataset

**Table 4.** Structure Comparison of Telco & Ecom Datasets

**Table 5.** Features Composition of Telco & Ecom Datasets

**Table 6.** Data Quality and Class Imbalance of Telco & Ecom Datasets

**Table 7.** Comparison of Analytical Values

**Table 8.** Performance Metrics of Classification Models – Telco Dataset

**Table 9.** Performance Metrics of Classification Models – E-Commerce Dataset

***LIST OF FIGURES***

**Figure 1.** Histogram of Numerical Features in the E-Commerce Dataset  
**Figure 2.** Histogram of Categorical Features in the E-Commerce Dataset  
**Figure 3.** Variance Inflation Factor (VIF) Scores for Numerical Features – E-Commerce Dataset  
**Figure 4.** Pearson Correlation Matrix (Numerical Features and Churn) – E-Commerce Dataset  
**Figure 5.** SHAP Summary Plot – Mean Feature Importance in the E-Commerce Dataset  
**Figure 6.** Initial SHAP Output Before Feature Pruning – E-Commerce Dataset  
**Figure 7.** Chi-Square Test Results – E-Commerce Dataset  
**Figure 8.** Cramér’s V Scores for Categorical Variables – E-Commerce Dataset  
**Figure 9.** Distribution Plots for Tenure, MonthlyCharges, and TotalCharges – Telco Dataset  
**Figure 10.** Churn Distribution by Categorical Features – Telco Dataset  
**Figure 11.** Class Distribution Before and After SMOTE – Telco Dataset  
**Figure 12**. Top 15 Features with Highest VIF – Telco Dataset  
**Figure 13.** Top 20 Important Features Identified by Random Forest – Telco Dataset  
**Figure 14.** SHAP Interaction Plot: Tenure and SeniorCitizen – Telco Dataset  
**Figure 15.** Feature Correlation Heatmap – Telco Dataset  
**Figure 16.** Top 10 Features Most Correlated with Churn – Telco Dataset  
**Figure 17.** Structural Comparison Between Telco and E-Commerce Datasets  
**Figure 18.** Distribution of Engagement Score – Telco Dataset  
**Figure 19.** Engagement Score by Churn Status (Boxplot) – Telco Dataset  
**Figure 20**. PCA Projection of Churn Classes – Telco Dataset  
**Figure 21.** ROC Curves: Logistic Regression vs. Random Forest – Telco Dataset  
**Figure 22.** Confusion Matrices and ROC Curves for Churn Models – E-Commerce Dataset  
**Figure 23.** AUC Scores of Different Models – Telco Dataset  
**Figure 24.** Decision Tree Visualization (Depth = 3) – Telco Dataset  
**Figure 25.** Precision–Recall Trade-Off Curve – Telco Dataset (Random Forest)  
**Figure 26.** AUC Comparison for Churn Models – E-Commerce Dataset

1. **REVISED ABSTRACT**

Customer churn poses a significant challenge in behavior-driven industries such as e-commerce and telecommunications, where long-term profitability relies heavily on customer retention. This project leverages predictive analytics and explainable machine learning to not only forecast churn but also to understand the behavioral and service-related factors that influence it, with the goal of designing sector-specific retention strategies.

Two real-world datasets sourced from Kaggle are used for this analysis:

1. The “Telco Customer Churn” dataset, which includes demographics, service usage, and billing information (Kaggle, 2018);
2. The “Customer Behavior in E-Commerce” dataset, featuring transactional behavior such as order frequency, product category engagement, and spending patterns (Imakash, 2023).

These datasets provide a cross-sector view of churn dynamics, allowing for comparative analysis between industries. The project explores the following primary research questions:

* Which behavioral and service-related features are most predictive of churn in each sector?
* Can interpretable machine learning models predict churn before it occurs with sufficient accuracy?
* How do churn drivers and mitigation strategies differ between the e-commerce and telecom sectors?

The methodology includes rigorous data preprocessing—such as one-hot encoding, missing value imputation, and multicollinearity checks using Variance Inflation Factor (VIF). Class imbalance is addressed using **SMOTE** (Chawla et al., 2002) to enhance model performance on minority (churn) classes. Three classification models—**Logistic Regression**, **Random Forest**, and **XGBoost**—are trained and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

To address concerns around model interpretability and overfitting, **SHAP** (Lundberg & Lee, 2017) and **LIME** (Ribeiro et al., 2016) are used to visualize feature importance, while **cross-validation and regularization techniques** are employed to ensure model generalizability. An interactive **Power BI dashboard** visualizes key findings, helping both technical and non-technical stakeholders explore high-risk customer segments and actionable insights.

While both datasets provide valuable behavioral and demographic data, they contain synthetic elements and have limited representativeness. Ethical considerations around real-world applicability and data bias are acknowledged, and future work will incorporate more diverse, high-fidelity datasets for model refinement.

* + 1. **INTRODUCTION**

Customer churn—the phenomenon where customers discontinue their relationship with a business—is a critical challenge in both e-commerce and telecommunications sectors. In increasingly competitive markets, retaining existing customers is not only more cost-effective than acquiring new ones, but also essential for long-term profitability and sustainable growth. Studies have shown that retaining a customer can be **5 to 25 times cheaper** than acquiring a new one (Reichheld & Sasser, 1990; Imani et al., 2025).

rise of digital platforms and online service models has made it easier for customers to switch providers, increasing churn risk. Consequently, businesses are prioritizing early detection and prevention of churn to maintain competitive advantage. While numerous churn prediction studies exist, particularly in the telecommunications sector, most of them focus solely on forecasting outcomes, without uncovering interpretable churn drivers (Imani et al., 2025).

This project aims to go beyond prediction and explore **what drives churn** in both e-commerce and telecom environments by analyzing customer behavior patterns and service features. By applying machine learning models supported with explainable AI techniques, the study offers actionable insights for strategic decision-making. Recent research has shown that techniques like **SHAP** and **LIME** significantly improve model transparency, making churn predictions more interpretable and actionable (Boukrouh et al., 2024; Maan & Maan, 2023).

To address this problem, two real-world datasets—one from the **telecommunications industry**, and one from the **e-commerce domain**—were analyzed through a unified workflow of preprocessing, feature engineering, and classification modeling. Class imbalance was addressed using sampling techniques such as SMOTE, and interpretability was enhanced using SHAP and LIME.

The project is guided by the following central research question:

***Can customer churn be accurately predicted before it occurs, and do effective prevention strategies differ between e-commerce and telecommunications sectors?***

To support this inquiry, the study also investigates the following sub-questions:

1. What behavioral and service-related factors most drive churn?
2. How do these drivers differ between e-commerce and telecom sectors?
3. Can interpretable predictive models effectively inform early, targeted retention efforts?

By addressing these questions, this project aims to not only build high-performing predictive models but also provide sector-specific, evidence-based strategies for customer retention.

**3 REVISED LITERATURE REVIEW, DATA DESCRIPTION AND APPROACH**

**3.1 LITERATURE REVIEW**

Customer churn is widely recognized as a key business issue in data-driven industries like telecommunications and e-commerce, where retaining a customer is significantly more cost-effective than acquiring a new one—estimated at 5 to 25 times less (Reichheld & Sasser, 1990).

In recent years, machine learning (ML) and explainable artificial intelligence (XAI) have become foundational tools for churn prediction. Traditional methods like logistic regression—particularly without regularization—often fall short in capturing nonlinear patterns and suffer from low generalizability.

**3.1.1 The Landscape of Churn Prediction Research**

A comprehensive review by Imani et al. (2025) examined over 240 studies and highlighted three recurring challenges:

* + Class imbalance: Minority class (churners) is often underrepresented.
  + Lack of model interpretability: Especially in black-box models like XGBoost.
  + Limited cross-domain generalizability: Many models are tailored to a single sector.

In the telecom sector, Asif et al. (2025) developed a TriBoost ensemble model and showed that features like contract type, payment method, and tenure were top churn predictors.

In the e-commerce space, Boukrouh et al. (2024) applied artificial neural networks and Random Forest models to transactional data. They achieved over 92% accuracy and emphasized the importance of using SHAP and LIME to gain actionable insights.

While these studies offer strong baseline models, they also present limitations. For instance, TriBoost, as proposed by Asif et al. (2025), requires domain-specific tuning and may not generalize well beyond the telecom environment. Similarly, although ANN models can achieve high predictive accuracy, they are prone to overfitting in sparse or noisy datasets and often lack transparency unless supplemented with post-hoc explainability tools. These constraints reinforce the need for churn models that are not only predictive but also interpretable, scalable, and applicable across different business domains.

***3.1.2 Challenges in Churn Prediction: Imbalance and Interpretability***

Class imbalance is a persistent challenge. Most churn datasets have far fewer positive (churn) cases than negative ones. A study in Nature Scientific Reports (2025) found that using SMOTE with ensemble methods improved recall from 61% to 79%. ADASYN, on the other hand, outperformed simpler resampling methods in behavior-driven datasets like e-commerce (He et al., 2008).

Model interpretability is the second major challenge.Although models like XGBoost achieve high accuracy, their decision processes are opaque. SHAP values, as demonstrated by Maan & Maan (2023), effectively clarified how input features contribute to churn predictions, both at global and local levels.

In e-commerce, Boukrouh et al. (2024) used SHAP and LIME to uncover customer journey dynamics, enhancing decision-making. These insights justify the choice of SHAP for this project.

***3.1.3 Traditional Approaches and Their Limitations***

Traditional churn detection often relied on fixed rules—such as labeling a customer as "at-risk" if inactive for 30 days. Such rule-based systems fail to adapt to the dynamic nature of customer behavior across different sectors.

They also lack personalization and usually ignore interaction effects among features. Moreover, traditional models do not scale well and rarely provide real-time insights, leading to delayed or ineffective retention efforts.

***3.1.4 Conflicting Findings in the Literature***

Although tenure and contract type are frequently cited as dominant churn predictors in the telecommunications sector (Asif et al., 2025), Imani et al. (2025) argue that more immediate triggers—such as billing complexity and poor customer support—may have a stronger impact on customer decision-making. This discrepancy reflects a broader debate in churn literature: whether long-term structural features or short-term experiential factors play a greater role in customer retention.

A similar divergence exists in the e-commerce domain. Boukrouh et al. (2024) identify price sensitivity and spending frequency as key churn indicators, while Maan and Maan (2023) emphasize the importance of personalized engagement and customer attention, suggesting that static transactional metrics alone may be insufficient to explain churn behavior.

These conflicting findings indicate that churn is not driven by a single universal set of features but is instead deeply shaped by **domain context, behavioral nuance, and data granularity**. Accordingly, this project adopts a sector-specific modeling strategy—treating service attributes as primary in telecom and transactional dynamics as central in e-commerce. This approach not only respects domain-specific theory but also enables more precise and explainable predictions aligned with each dataset’s structure.

**Future studies** may benefit from integrating both structural and experiential features—such as combining billing data with customer sentiment or service interaction logs—to reconcile these divergent findings and better capture the multifaceted nature of churn behavior (Imani et al., 2025; Boukrouh et al., 2024).

***3.1.5 Positioning of This Study***

This project builds on prior research by addressing multiple gaps:

* + Cross-domain comparison: Few studies have systematically compared churn dynamics in both telecom and e-commerce.
  + Tailored preprocessing: This study applies SMOTE for the telecom dataset and ADASYN for the e-commerce dataset—each selected based on feature space complexity.
  + Model transparency: SHAP is used consistently across all models and datasets to ensure interpretability.
  + Unified pipeline: Logistic Regression, Random Forest, and XGBoost are evaluated under a standardized framework.

Ultimately, the combination of predictive performance and interpretability ensures that the models are not only accurate, but also actionable and transparent—making them valuable for both technical teams and business decision-makers.

While planning the modeling phase, alternative algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were initially considered. However, they were ultimately excluded based on both methodological and practical considerations. **SVMs**, while effective in linearly separable datasets, tend to struggle with high-dimensional data generated through one-hot encoding, and their non-linear variants are computationally intensive and difficult to interpret (Huang et al., 2018). **ANN models**, on the other hand, require large-scale datasets and complex tuning procedures to avoid overfitting—conditions not fully met by the size and structure of the current datasets, particularly in the e-commerce case with synthetic attributes (Boukrouh et al., 2024; He et al., 2008). Most importantly, both models lack native interpretability, which is a critical requirement for this project’s emphasis on explainable AI (Maan & Maan, 2023). By contrast, the selected models—Logistic Regression, Random Forest, and XGBoost—offer a balance of predictive accuracy, scalability, and compatibility with SHAP-based explainability. Their combined use aligns with the dual objectives of this study: performance and transparency across sectors.

* 1. **DATA DESCRIPTION**

This study utilizes two publicly available datasets: one from an e-commerce platform and another from a telecommunication service provider. Both datasets contain customer-level features, transactional behavior, and churn labels, enabling cross-domain predictive modeling and interpretation.

**3.2.1 DATA DESCRIPTION FOR E-COMMERCE DATASET**

This study utilizes the publicly available **"E-Commerce Customer for Behavior Analysis"** dataset from Kaggle, originally published by **Shriyash Jagtap (2023)**, to examine behavioral patterns associated with customer churn. Provided in CSV format, the dataset includes detailed **transactional records** and **demographic attributes**, enabling a comprehensive analysis of purchase behaviors, seasonal trends, and churn risk factors.

The dataset is publicly accessible at: <https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis> (Jagtap, 2023).

* + - 1. ***Source and Scope***

The dataset originates from a full year of transactional records and reflects **seasonal trends** and **customer purchasing behavior** across various product categories. It contains approximately **250,000 transaction entries** from **around 50,000 unique customers**, capturing both behavioral and demographic attributes.

The original dataset comprises **13 variables**, including:

* Customer identifiers • Purchase timestamps
* Product-level details (category, price, quantity) • Payment methods
* Customer demographics (age, gender) • A binary churn outcome (1 = churned, 0 = retained)

Initially organized at the **transaction level**, the dataset was **aggregated to the customer level** to better support churn prediction. During this transformation, key behavioral summaries were computed for each customer, such as **total purchase amount**, **purchase frequency**, and **product diversity**. Furthermore, **engineered features** such as AvgItemValue, PriceToQuantity, PurchaseMonth, and AgeGroup were introduced to enrich the feature space.

Preprocessing steps included **handling missing values**, **one-hot encoding** of categorical variables, and **standardization** of numerical features. To address the **class imbalance** in the target variable, the **ADASYN (Adaptive Synthetic Sampling)** technique was employed. These steps collectively prepared the dataset for **robust predictive modeling**, aligning the structure with the study’s analytical objectives.

* + - 1. **Feature Overview**

The features in the e-commerce dataset are summarized in the table below. These include original variables from the raw dataset, as well as engineered features created during preprocessing to enhance model performance and interpretability.

| ***Feature*** | ***Description*** | ***Type*** |
| --- | --- | --- |
| Customer ID | Unique identifier for each customer | Identifier |
| Purchase Date | Timestamp of purchase, later transformed to monthly granularity | Date → Categorical |
| Product Category | Type of product purchased (e.g., Clothing, Books, Electronics, Home) | Categorical |
| Product Price | Price per unit of the purchased product | Numeric (int) |
| Quantity | Number of items purchased per transaction | Numeric (int) |
| Total Purchase Amount | Calculated as Product Price × Quantity | Numeric (int) |
| Payment Method | Payment type used (Cash, Credit Card, PayPal, Crypto) | Categorical |
| Customer Age | Age of the customer in years | Numeric (int) |
| Gender | Customer’s gender (Male, Female) | Categorical |
| Returns | Number of items returned (dropped due to 19% missing values) | Numeric (float) |
| Churn | Binary target variable (1 = churned, 0 = retained) | Binary |

*Table 1: Features Overview of Ecom Dataset*

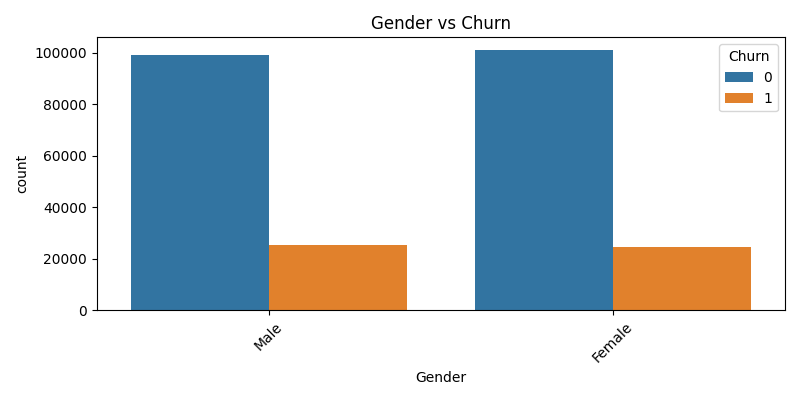
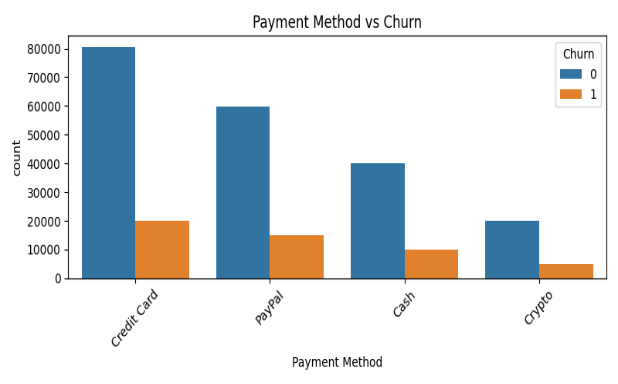
Engineered Features:

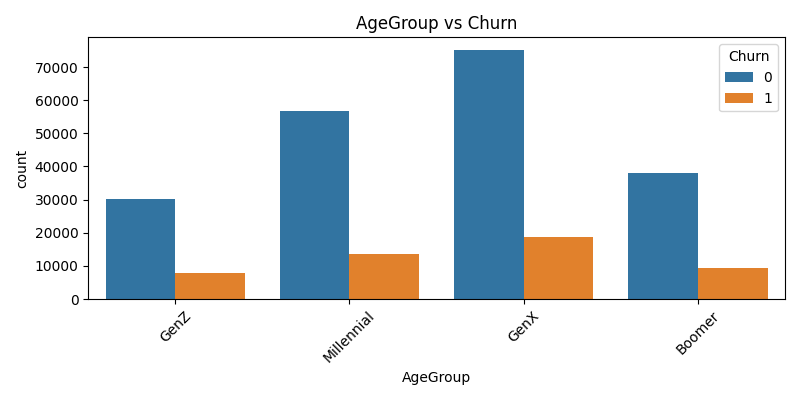
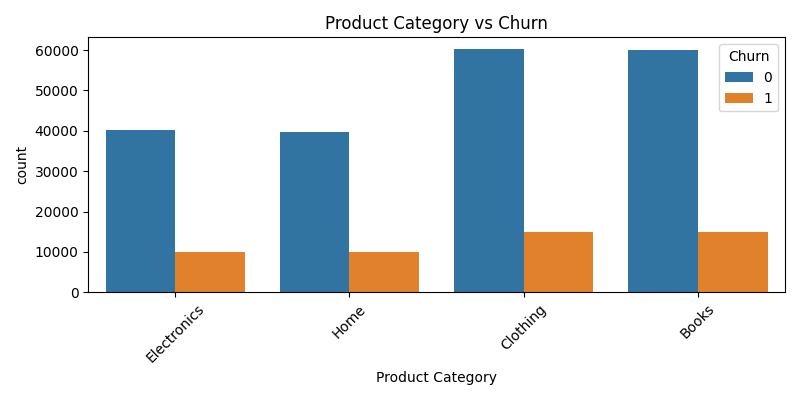
* AvgItemValue: Total Purchase Amount / Quantity
* PriceToQuantity: Product Price divided by Quantity (when >0)
* PurchaseMonth: Month extracted from Purchase Date
* AgeGroup: Age bucketized into youth, adult, senior

A group of blue bars

AI-generated content may be incorrect.

***Figure 1.***  *Histogram of Ecom Numerical Features displays the distribution of key numerical variables such as Customer Age, Product Price, Quantity, and Total Purchase Amount. These histograms help identify skewness, outliers, and underlying patterns in purchasing behavior.*

* *

* *

***Figure 2.***  *Histogram of Ecom Categorical Features displays the distribution of key Categorical variables such as AgeGroup, Product Category, Gender and Payment Method. These histograms help identify skewness, outliers, and underlying patterns in purchasing behavior.*

The categorical variables in the dataset—such as *Gender*, *Payment Method*, *Product Category*, and *AgeGroup*—provide demographic and transactional context to customer behavior. While these features offer valuable segmentation insights, their individual association with churn appears limited, as later statistical analysis suggests low dependency strength. Nonetheless, they remain important for interaction effects and model interpretability.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Feature** | **Cramér’s V Score** | | Gender | 0.015 | | Payment Method | 0.021 | | Product Category | 0.017 | | AgeGroup | 0.028 | | |
|  |
|  | |

***Table 2.*** *Association Strength of Categorical Features with Churn (Cramér’s V Scores)*

*Note: All Cramér’s V values fall well below 0.1, indicating weak association with churn.*

***3.2.1.3 Preprocessing and Cleaning***

A series of preprocessing steps were conducted to prepare the dataset for modeling and ensure data quality:

* *Duplicate and Redundant Features:*  
   - Duplicate transaction records were removed. The Age column was excluded due to its redundancy with Customer Age and high multicollinearity.
* *Missing Data Handling:*

- The Returns feature was dropped due to 19% missing values and limited predictive value.

- Median imputation was applied to sparse numerical variables. Mode imputation was used for missing categorical values.

* *Feature Encoding:*
  + One-hot encoding was applied to categorical features including Product Category, Gender, and Payment Method.
* Date Transformation:
  + The Purchase Date was converted to datetime format and used to extract a new PurchaseMonth feature to capture seasonality in purchasing behavior.
* Scaling and Standardization:

- Z-score standardization was applied to numerical variables such as Product Price, Quantity, Customer Age, and Total Purchase Amount to normalize the feature space.

***3.2.1.4 Target Variable and Class Distribution***

The target variable, **Churn**, reflects customer inactivity or disengagement and is encoded as a binary variable (1 = churned, 0 = retained). The initial class distribution revealed a significant imbalance:

* **Churn = 0 (Active)**: ~80.05% • **Churn = 1 (Churned)**: ~19.95%

This imbalance poses a risk of bias in model training, where the classifier may favor the majority class and underperform on churn prediction. To address this, the **ADASYN** (Adaptive Synthetic Sampling) technique was applied **only on the training set**, generating synthetic samples for the minority class based on data density.

Following ADASYN application:

* **Churn = 0**: 49.80% • **Churn = 1**: 50.20%

This near-equal class distribution enhanced the model’s sensitivity to churn signals, improving **recall** and **F1-score** for the minority class.

A blue rectangular column with white text

AI-generated content may be incorrect. A blue rectangular bars with white text

AI-generated content may be incorrect.

**Figure 2.** Ecom Churn class distribution in the E-Commerce dataset before and after applying ADASYN. The technique helped achieve a near-balanced split, enabling better learning of minority (churned) instances.

***3.2.1.5 Feature Relationships and Correlation Insights***

Understanding the relationships among features is essential for assessing redundancy and enhancing model interpretability.

***• Multicollinearity:***

Variance Inflation Factor (VIF) scores were calculated to evaluate the degree of multicollinearity among numerical features. In the e-commerce dataset:

* + The highest VIF was observed for Customer Age (~6.4), still below the commonly used threshold of 10.
  + Most other variables, including Product Price, Quantity, and Total Purchase Amount, had VIF values well below 5.
  + Due to high correlation and conceptual redundancy, the original Age column was dropped in favor of Customer Age.

This confirms that multicollinearity is not a critical issue in the e-commerce dataset, and the retained features are suitable for modeling without further transformation.

A graph of blue bars

AI-generated content may be incorrect.

***Figure 3.*** Ecom Variance Inflation Factor (VIF) for Numerical Features *in the E-Commerce dataset.* Low VIF values suggest minimal redundancy, supporting the stability of regression-based models.

***Note:*** *All VIF scores were below 10, confirming no serious multicollinearity issues*

• ***Correlation Analysis:***

The Pearson correlation matrix highlights linear associations between numerical features and the target variable, Churn. Total Purchase Amount and Engagement Score show moderate correlations with churn, suggesting potential predictive value, while Quantity and Product Price exhibit weaker relationships.

A blue and red squares with white text

AI-generated content may be incorrect.

***Figure 4.*** ECom Correlation Matrix (Numerical Features and Churn) displays Pearson coefficients between churn and numerical features, highlighting potential predictive signals and redundancy.

* ***Feature Importance:***

Based on XGBoost and SHAP explainability, most predictive features were:

* + Customer Age,
  + Purchase Date,
  + Product Price,
  + Total Purchase Amount,
  + Quantity

**Figure 5** presents the most influential features in churn prediction using SHAP values, calculated after applying ADASYN and hyperparameter tuning. **Customer Age** emerged as the strongest predictor, followed by **Purchase Date**, **Product Price**, **Total Purchase Amount**, and **Quantity**. These variables reflect critical aspects of customer behavior such as spending habits, timing of purchases, and order size. Their prominence in the model highlights their importance for identifying churn-prone customers and shaping personalized retention strategies.

**Figure 6** shows the **initial SHAP output** before feature pruning and model optimization. While **Figure 5** illustrates the **average magnitude** of each feature’s impact, **Figure 6** visualizes how **feature values (high/low)** influence the **direction** of churn. Together, these SHAP plots enhance model transparency and actionable insight.

A blue bar graph with white text

AI-generated content may be incorrect. A bar graph with blue and white text

AI-generated content may be incorrect.

***Figure 5.*** *Ecom mean (SHAPvalue)(average impact on model output magnitude)*

A graph of a graph with text

AI-generated content may be incorrect. A graph of different values

AI-generated content may be incorrect.

*Figure 6. Ecom Initial SHAP output before feature pruning & model tuning*

* ***Churn Relationships (Chi-square & Cramer’s V):***

Statistical tests were conducted to examine the relationship between categorical features and customer churn.

Figure 7 reports the results of Pearson’s Chi-square test for independence, revealing a statistically significant association between **Gender** and churn (p < 0.001). In contrast, **Product Category** and **Payment Method** yielded non-significant results (p > 0.4), indicating that their individual impact on churn may be minimal when considered independently.

Figure 8 visualizes **Cramér’s V** scores for these categorical features, all of which fall below 0.1—suggesting weak association strength. These results complement the Chi-square findings and imply that the predictive power of these variables likely emerges through interaction effects or in multivariate modeling contexts.

A screenshot of a graph

AI-generated content may be incorrect.

***Figure 7.*** *Chi-square test results for the e-commerce dataset. Gender is significantly associated with churn (p < 0.001), while other features showed weak individual associations.*

**A graph with blue squares

AI-generated content may be incorrect.*Figure 8.*** *Cramér’s V scores measuring the association strength between categorical features and churn. All values fall below 0.05, indicating weak individual associations.*

***3.2.1.6 Limitations***

• The dataset lacks behavioral granularity—such as session duration, clickstream data, or customer satisfaction scores—which could provide deeper insights into customer engagement and churn signals.

• The significant class imbalance necessitated the use of synthetic resampling techniques (e.g., ADASYN). While effective for training, these synthetic samples may not fully capture real-world churn dynamics.

• The presence of redundant fields and placeholder values required extensive preprocessing and validation to ensure data quality and consistency.

• Categorical features with many levels (e.g., product types, payment methods) may benefit from advanced encoding strategies such as grouping or embeddings to improve model efficiency and interpretability.

**3.2.2. DATA DESCRIPTION FOR TELECOMMUNICATION DATASET**

This project utilizes a publicly available **telecommunications customer churn dataset** to investigate behavioral patterns contributing to churn in subscription-based services. The dataset provides comprehensive information on customer demographics, service usage, billing behavior, and contract types—making it well-suited for predictive analytics and churn modeling.

The dataset was originally published on Kaggle by IBM Sample Datasets and is publicly available: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn> (Kaggle, 2018).

***3.2.2.1. Source and Scope***

The dataset comprises **7,043 customer records** and **21 attributes**, collected by a telecommunications service provider to understand patterns associated with customer retention and churn. Each record represents an individual customer and includes:

* **Demographic details** (e.g., gender, age category, household structure)
* **Account and contract characteristics** (e.g., tenure, contract length, paperless billing)
* **Service utilization** (e.g., phone service, internet type, streaming services)
* **Financial attributes** (e.g., monthly and total charges)
* **Churn status**, which is the target variable of interest.

The dataset does not include any personally identifiable information (PII), making it publicly available and ethically suitable for academic research.

***3.2.2.2. Feature Overview***

The Telco dataset contains a range of demographic, service-related, and financial attributes, as summarized below:

| ***Feature*** | ***Description*** | ***Type*** |
| --- | --- | --- |
| customerID | Unique customer identifier (dropped during preprocessing) | Identifier |
| gender, Partner, Dependents | Demographic variables indicating household context | Categorical |
| SeniorCitizen | Binary flag indicating if the customer is 65+ years old | Binary (0/1) |
| tenure | Number of months with the company | Numeric (int) |
| PhoneService, MultipleLines | Status of phone services | Categorical |
| InternetService, OnlineSecurity, TechSupport, etc. | Internet and value-added services | Categorical |
| Contract, PaperlessBilling, PaymentMethod | Contract and billing information | Categorical |
| MonthlyCharges | Monthly amount billed to the customer | Numeric (float) |
| TotalCharges | Total amount paid by the customer (converted from string) | Numeric (float) |
| Churn | Target variable (1 if churned, 0 otherwise) | Binary (0/1) |

***Table 3.*** *Features overview of Telco Dataset*

Preprocessing Notes:

* + Categorical features were transformed via one-hot encoding.
  + Numerical features were standardized using z-score scaling (StandardScaler).
  + The Churn variable was binarized for modeling purposes.

A graph of a graph of a graph

AI-generated content may be incorrect.

***Figure 9.*** *Distribution plots for Tenure, MonthlyCharges, and TotalCharges in the Telco dataset reveal strong right-skewness, especially in TotalCharges. These patterns highlight the presence of long-term customers with accumulated charges, as well as pricing clusters around lower tiers. Such skewness suggests the need for normalization or transformation before applying linear models like logistic regression.*

A screenshot of a graph

AI-generated content may be incorrect.

***Figure 10.*** *Categorical churn distributions in the Telco dataset reveal significant differences across service types and contract terms. Customers using fiber optic internet and those on month-to-month contracts exhibit notably higher churn rates, while longer-term contracts correlate with lower churn. Additionally, churn is more prevalent among users paying by mailed check or electronic check, highlighting the impact of billing preferences on retention. These insights emphasize the predictive power of service-related and billing attributes in customer churn modeling.*

***3.2.2.3. Preprocessing Summary***

The preprocessing pipeline included the following key steps to ensure data integrity and compatibility with machine learning algorithms:

* **Missing Data Handling:** A total of 11 rows were removed due to missing or invalid TotalCharges values after converting the column from string to float.
* **Feature Elimination:** The customerID field was dropped as it did not contribute predictive value.
* **Encoding of Categorical Variables:** All categorical features were transformed into dummy variables using pd.get\_dummies() with drop\_first=True to avoid multicollinearity.
* **Standardization of Numerical Features:** Continuous variables such as tenure, MonthlyCharges, and TotalCharges were standardized using z-score normalization to ensure consistent input scaling.
* **Target Variable Conversion:** The Churn variable was converted from string labels ("Yes"/"No") to binary format (1 = churned, 0 = retained) for classification modeling.

*Note: Rows where `TotalCharges` contained whitespace-only values were considered invalid and removed after conversion to float.*

***3.2.2.4. Target Variable and Class Distribution***

Before balancing, the dataset showed a moderate class imbalance in the target variable:

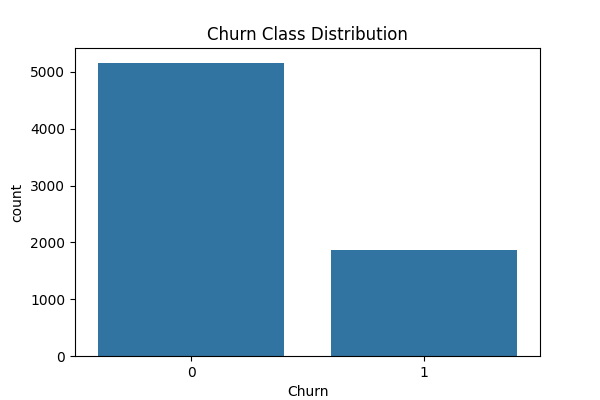
* **Churn = 0 (No):** 5,163 samples (≈73.5%) • **Churn = 1 (Yes):** 1,869 samples (≈26.5%)

This imbalance risks biasing machine learning models toward the majority class (non-churners), often reducing precision and recall for identifying churners. To mitigate this, **Synthetic Minority Over-sampling Technique (SMOTE)** was applied **only to the training set** to avoid data leakage.

After SMOTE resampling:

* **Churn = 0:** 5,229 samples (50%) • **Churn = 1:** 5,229 samples (50%)

SMOTE addresses class imbalance by generating synthetic churn examples through interpolation between existing minority instances. Unlike simple duplication, this creates more diverse training samples. As a result, the model learns better decision boundaries and shows improved recall in detecting churn, without significantly compromising overall accuracy. Since SMOTE is applied only to the training set, data leakage is avoided.

* A graph of a number of different colored squares

AI-generated content may be incorrect.*

***Figure 11.*** *Churn class distribution in the Telco dataset before and after applying SMOTE. Synthetic oversampling helped mitigate class imbalance.*

***3.2.2.5. Feature Relationships and Correlation Insights***

***• Multicollinearity***

Multicollinearity was evaluated using the Variance Inflation Factor (VIF). The analysis indicated:

- High multicollinearity among internet-related services (e.g., InternetService, OnlineSecurity, TechSupport) and billing features (e.g., PaymentMethod, Contract).

- Several features exhibited VIF values greater than 10, signaling potential redundancy.

- However, these features were not removed, as the selected modeling algorithm — Random Forest Classifier — is relatively insensitive to multicollinearity, unlike linear models.

A graph with blue bars

AI-generated content may be incorrect.

***Figure 12.*** *Top 15 Features with Highest VIF in the Telco Dataset. Internet and billing-related features showed higher multicollinearity but were retained for tree-based modeling.*

***• Feature Importance:***

Feature importance analysis using Random Forest and SHAP identified several key predictors of churn in the Telco dataset:

* Tenure: Longer tenure is strongly linked to customer retention.
* TotalCharges: Reflects accumulated payment history and customer longevity.
* MonthlyCharges: Higher monthly charges were associated with increased churn risk.
* Contract, PaymentMethod, and OnlineSecurity also showed strong predictive power, emphasizing the role of service configuration and billing preferences in churn behavior.

*A graph with blue and black text

AI-generated content may be incorrect.*

***Figure 13.*** *Top 20 Important Features from Random Forest Model in the Telco Dataset. This figure highlights the most predictive features identified by the Random Forest classifier. Tenure and TotalCharges were most strongly associated with customer retention, whereas short-term contracts and higher monthly charges were linked to an increased likelihood of churn.*

These findings highlight that long-term contracts and customer tenure significantly reduce churn, reflecting greater loyalty and satisfaction. In contrast, short-term contracts and high monthly charges are linked to higher churn, emphasizing the value of commitment-based pricing strategies to enhance retention.

A blue and pink arrows

AI-generated content may be incorrect.

***Figure 14.*** *SHAP Interaction between Tenure and SeniorCitizen in the Telco Dataset. This plot illustrates the joint effect of tenure and senior status on churn risk. Longer tenure is strongly associated with retention, especially among non-senior customers.*

***• Correlation Analysis:***

To explore linear relationships between features and their association with churn, a Pearson correlation matrix was generated (Figure 13). The heatmap reveals clusters of correlated features, particularly among internet services (e.g., *OnlineSecurity*, *TechSupport*) and billing attributes (e.g., *Contract* and *PaymentMethod*).

Figure 14 presents the top 10 features most correlated with churn. Key findings include:

* *InternetService\_Fiber optic* and *PaymentMethod\_Electronic check* show the highest positive correlation with churn.
* *HighMonthlyChargeFlag* and *MonthlyCharges* are also positively associated, suggesting that customers facing higher recurring costs are more prone to churn.
* *PaperlessBilling\_Yes* and *SeniorCitizen* exhibit moderate positive correlations.

While most correlations remain weak (typically < 0.3), they offer directional insights that align with and support the model-driven feature importance findings.

*A graph with different colored squares

AI-generated content may be incorrect.****Figure 15.*** *Feature Correlation Heatmap in the Telco Dataset. This heatmap visualizes Pearson correlation coefficients between numerical and encoded categorical features. It reveals clusters of positively correlated variables, especially among service-related and billing features. While most correlations with churn are relatively weak, the heatmap helps identify potential multicollinearity and supports feature selection for modeling.*

*A graph with blue and white text

AI-generated content may be incorrect.*

***Figure 16.*** *Top 10 Features Most Correlated with Customer Churn. This bar plot displays the top 10 features with the highest absolute correlation values with the churn variable. InternetService\_Fiber optic and PaymentMethod\_Electronic check show the strongest positive correlations with churn, indicating that customers using these services are more likely to leave. Other moderately correlated factors include high monthly charges and paperless billing, suggesting billing-related dissatisfaction may influence customer retention.*

***3.2.2.6. Limitations***

Despite its strengths, the dataset and methodology have a few limitations:

1. **Lack of Behavioral Signals:**The dataset does not include user activity logs, customer satisfaction scores, or service usage frequency. These behavioral features could offer deeper insights into churn drivers.
2. **Synthetic Oversampling Caution:** While SMOTE improved class balance, the synthetic samples may not fully reflect real-world customer diversity. External validation on new or real-world datasets is recommended.
3. **High Dimensionality:** One-hot encoding increased the number of features to over 45, which can raise the risk of overfitting, especially in tree-based models. Future work could benefit from feature selection, dimensionality reduction, or regularization techniques.
4. **Static Snapshot:** The dataset captures a single point in time, limiting the ability to analyze temporal trends or seasonal churn patterns. Longitudinal data would allow for time-series modeling and trend analysis.
   1. **COMPARATIVE ANALYSIS AND ANALYTICAL CONTRIBUTION**

This project integrates two distinct datasets—Telco Customer Churn and E-Commerce Customer Behavior—to explore churn prediction across different sectors. While both aim to predict churn, their structural differences shaped the preprocessing, modeling, and interpretability strategies employed.

**4.3.1 Source and Structure Comparison**

| Aspect | Telco Dataset | E-Commerce Dataset |
| --- | --- | --- |
| Domain | Subscription-based telecom services | Transactional retail / e-commerce |
| Sample Size | 7,043 customers | 250,000 transactions (~50,000 unique customers) |
| Structure | One row per customer | Multiple transactions per customer (aggregated) |
| Target Variable | Churn (0 = active, 1 = churned) | Churn (0 = active, 1 = churned) |
| Churn Rate (Raw) | ~26.5% | ~19.9% |

***Table 4.*** *Structure Comparison of Telco & Ecom Datasets*

The Telco dataset is inherently customer-centric, whereas the E-Commerce dataset required aggregation from transaction-level data. This distinction influenced feature engineering and modeling complexity.

**3.3.2 Feature Composition and Preprocessing**

| Feature Type | Telco Dataset | E-Commerce Dataset |
| --- | --- | --- |
| Categorical Features | Rich (e.g., Contract, InternetService) | Few (e.g., Gender, Product Category) |
| Numerical Features | Tenure, Monthly/Total Charges | Product Price, Quantity, Customer Age |
| Temporal Features | None | Purchase Date (converted to month) |
| Engineered Features | IsLongTermCustomer | AvgItemValue, AgeGroup, PriceToQuantity |
| Redundant/Dropped Fields | customerID, null TotalCharges | Age (due to collinearity), Returns (missing) |

***Table 5.*** *Features Composition of Telco & Ecom Datasets*

While the Telco dataset primarily required one-hot encoding for categorical variables—such as contract type, payment method, and internet service options—due to its already structured and customer-level format, the E-Commerce dataset necessitated more complex feature engineering. This included deriving behavioral metrics (e.g., total purchase amount, average item value, and price-to-quantity ratios) and temporal decomposition (e.g., extracting purchase month and identifying seasonal patterns) from raw transactional data. These transformations were essential to aggregate customer behavior over time, enrich the feature space, and make the dataset compatible with churn prediction models, which typically rely on customer-level summaries rather than individual transactions.

A graph of a number of orange bars

AI-generated content may be incorrect.

**Figure 17.** Summary comparison of the Telco and E-Commerce datasets.

This figure illustrates the class distribution of retained and churned customers before and after applying resampling techniques (SMOTE for Telco, ADASYN for E-Commerce). Both datasets achieved near-perfect class balance post-resampling, which is critical for improving model performance on minority classes.

**3.3.3 Data Quality and Class Imbalance**

| Aspect | Telco Dataset | E-Commerce Dataset |
| --- | --- | --- |
| Missing Data | Minimal (11 rows dropped) | 19% missing in Returns (column dropped) |
| Outlier Handling | Minimal | Handled due to price/quantity variability |
| Class Imbalance | 73.5% No / 26.5% Yes → SMOTE applied | 80.1% No / 19.9% Yes → ADASYN applied |

***Table 6.*** *Data Quality and Class Imbalance of Telco & Ecom Datasets*

Resampling strategies were chosen based on the dataset’s complexity: SMOTE was preferred for the structured Telco data, while ADASYN was more effective on the non-linear behavioral patterns in the E-Commerce data.

**3.3.4 Correlation and Multicollinearity Insights**

* Telco Dataset: Strong multicollinearity was observed among Tenure, MonthlyCharges, and TotalCharges (VIF > 10). Despite this, features were retained as tree-based models like Random Forest and XGBoost are robust to multicollinearity.
* E-Commerce Dataset: Redundant features such as Age were dropped to reduce multicollinearity with Customer Age. Features like Product Price, Quantity, and Total Purchase Amount showed moderate correlations with each other, but not with churn, supporting the need for engineered behavioral features.

**3.3.5 Interpretability and Feature Impact**

* Telco Dataset: SHAP analysis highlighted tenure, contract type, and payment method as key drivers of churn. These features also served as indirect proxies for customer engagement—longer tenure and annual contracts often indicate higher loyalty and service satisfaction. The model confirmed that short-term plans and high monthly charges increase churn risk.
* E-Commerce Dataset: Feature importance was dominated by behavioral variables such as total purchase amount, quantity, and purchase date. SHAP explainability revealed that seasonal buying patterns and transaction frequency, rather than static customer attributes, were the primary churn indicators.

Across both datasets, SHAP values ensured transparency in feature contributions, enabling actionable business interpretations tailored to industry-specific dynamics.

**3.3.6 Summary of Analytical Value**

| Dataset | Key Strengths | Analytical Contribution |
| --- | --- | --- |
| Telco | Structured, interpretable, service-based | Ideal for segmentation, SHAP explainability, model testing under multicollinearity |
| E-Commerce | Temporal, behavioral, transaction-based | Useful for testing advanced preprocessing, ADASYN, and cross-domain generalizability |

***Table 7.*** *Comparison of Analytical Values*

Together, the datasets provided a multi-faceted learning environment:

* Telco reinforced best practices in structured feature modeling and stakeholder-friendly explanation.
* E-Commerce challenged the analyst with real-world complexities, requiring aggregation, imbalance handling, and behavioral interpretation.

By addressing both interpretability and predictive accuracy across domains, this project delivers adaptable, sector-specific churn insights that can inform proactive retention strategies.

**3.3.7 Practical Implications**

The comparative analysis of the Telco and E-Commerce datasets provides valuable insights for practitioners aiming to reduce customer churn in different sectors. Understanding the role of structural vs. behavioral features enables more targeted interventions—such as contract redesign in telecom or purchase frequency-based campaigns in e-commerce. Moreover, the findings reinforce the importance of using interpretable models that not only predict churn but also support actionable business decisions.

**3.4. MODELLING APPROACH**

To address the research objective of identifying the most influential drivers of customer churn and evaluating the effectiveness of predictive models across different domains, a comprehensive, replicable, and explainable data science pipeline was implemented for both the Telco and E-Commerce datasets. This approach was designed to ensure methodological consistency while allowing flexibility to accommodate the unique characteristics of each dataset.

**3.4.1 Data Preprocessing and Exploratory Analysis**

The first phase involved comprehensive data preprocessing and exploratory data analysis (EDA) to understand the structure, completeness, and patterns within both datasets. Key procedures included handling missing values using appropriate imputation strategies (e.g., median for numerical variables), removing irrelevant identifiers such as customerID, and addressing variable inconsistencies. In the E-Commerce dataset, temporal features such as “Purchase Month” were extracted from timestamp fields to capture seasonal churn behavior, while the Telco dataset retained its customer-level granularity without temporal stamps.

A detailed summary of the initial analyses—featuring distribution plots, correlation heatmaps, boxplots, and feature-by-churn breakdowns—is documented in the following GitHub-hosted EDA reports:  
• 📄 Telco Dataset (Raw)  
• 📄 E-Commerce Dataset (Raw)

Univariate and bivariate analyses were conducted to identify outliers, skewed distributions, and potential churn-related patterns. Chi-square tests (for categorical variables) and ANOVA (for continuous variables) were applied to assess statistical associations with churn. For example, in the Telco dataset, contract type, online security, and tech support showed strong significance; in the E-Commerce dataset, customer age group, product category, and purchase month were found to be differentiating factors.

To further explore data structure and class separability, **Principal Component Analysis (PCA)** was applied to the Telco dataset. After scaling the numerical and encoded variables, PCA revealed moderate separation between churn and non-churn classes, suggesting the feasibility of predictive modeling. This dimensionality reduction was not applied to the E-Commerce dataset due to its high granularity and the need for more extensive aggregation prior to transformation.

Additionally, an **Engagement Score** variable was engineered for the Telco dataset to quantify customer interaction with digital services. This metric was computed as the total number of active services (e.g., OnlineSecurity, TechSupport, StreamingTV) subscribed to by each customer. As illustrated in Figure 18, the distribution of engagement scores reveals that many customers utilize few or no digital services, while Figure 19 demonstrates that non-churned customers exhibit notably higher engagement levels. These patterns support the notion that bundled service usage contributes to customer retention and validate the Engagement Score as a meaningful predictive feature.

This feature was not constructed in the E-Commerce dataset due to the absence of service-based variables and session-level behavioral logs. These analytical insights provided the foundation for subsequent feature engineering, class imbalance handling, and model training—directly contributing to **Research Question 1**: *What are the most influential features driving customer churn?*

A graph of a number of active digital services

AI-generated content may be incorrect.

***Figure 18.*** *Histogram showing the distribution of the Engagement Score in the Telco dataset (number of active digital services per customer).*

A chart of a engagement score

AI-generated content may be incorrect.***Figure 19.*** *Boxplot comparing Engagement Scores by churn status; non-churned users tend to have higher* engagement.

A diagram of a diagram

AI-generated content may be incorrect.

***Figure 20****. PCA projection showing moderate separation between churned and non-churned customers in the Telco dataset.*

**3.4.2 Feature Engineering and Transformation**

To enhance model performance and interpretability, domain-specific feature engineering was conducted. In the E-Commerce dataset, aggregate metrics such as Total Purchase Amount were created to summarize customer value. In the Telco dataset, features such as TotalChargesPerMonth were derived to normalize billing information.  
Categorical variables were encoded using Label Encoding or One-Hot Encoding, depending on the modeling algorithm. Skewed numerical variables were normalized using StandardScaler or MinMaxScaler, especially for logistic regression.

**3.4.3 Handling Class Imbalance**

As both datasets exhibited class imbalance—with fewer customers churning than staying—this issue was addressed using resampling techniques tailored to each dataset. **SMOTE** was applied to the Telco dataset due to its structured and balanced feature space, while **ADASYN** was used for the E-Commerce dataset to handle the more irregular and behavioral nature of the data. ADASYN’s density-aware sampling helped prevent overfitting while improving recall, particularly in the presence of complex purchase behaviors.  
The effectiveness of ADASYN was compared with SMOTE and baseline models as part of the evaluation process for **Research Question 2**: *Which resampling method performs best across domains?*

**3.4.4 Model Building and Evaluation**  
 Three classification models were implemented iteratively across both datasets to evaluate predictive performance and interpretability:

* **Logistic Regression** served as a baseline model to establish initial performance metrics.
* **Random Forest** was selected as a robust ensemble method capable of capturing nonlinear relationships and offering intuitive feature importance scores.
* **XGBoost** was chosen for its superior performance on tabular data, built-in regularization techniques, ability to handle missing values, and proven success in churn prediction tasks (e.g., Maan & Maan, 2023).

Hyperparameters were optimized via cross-validated grid search. Models were evaluated using **Accuracy, Precision, Recall, F1-score**, and **ROC-AUC**, ensuring a comprehensive assessment of both class-specific and overall predictive performance.  
This modeling framework directly addresses **Research Question 3**: *Which model yields the most accurate and interpretable results across datasets?*

***3.4.4.1 Telco Dataset – Model Evaluation***

• The baseline logistic regression model achieved an AUC of 0.83, indicating solid overall discrimination. The confusion matrix revealed relatively strong true positive rates, correctly identifying 397 churn cases, though it misclassified 164 (Figure: Confusion Matrix – Logistic Regression).

• The Random Forest model produced a slightly lower AUC of 0.82, but achieved better balance across precision and recall, particularly in reducing false negatives, which is critical in churn contexts (Figure: Confusion Matrix – RF).

• Feature importance plots from both models highlighted tenure, TotalCharges, and Contract as top predictors. This alignment between statistical significance and model-driven inference reinforces their relevance in real-world churn prediction.

A graph of a logistic regression

AI-generated content may be incorrect. A graph with a number of numbers and a blue and white squares

AI-generated content may be incorrect.A graph of a curve

AI-generated content may be incorrect. A blue squares with white text

AI-generated content may be incorrect.

***Figure 21.*** *ROC curves show that Logistic Regression achieved an AUC of 0.83, slightly outperforming Random Forest (AUC = 0.82). However, the confusion matrices indicate that Random Forest provided a more balanced prediction, particularly by reducing false positives and improving recall for churners.*

**Model Performance Comparison: Telco Dataset**

To assess the practical strengths and trade-offs of the applied models, ROC curves and confusion matrices were used to visualize classification behavior (see Figure 18). While the Logistic Regression model achieved a marginally higher AUC (0.83), the Random Forest model demonstrated a more balanced prediction profile by reducing the number of false positives and improving precision for churn cases.  
 Although Logistic Regression provided slightly better overall discrimination, Random Forest may be more appropriate in operational settings where minimizing churn misclassification is critical. These findings underscore the importance of incorporating both overall and class-specific evaluation metrics when selecting a churn prediction model.

***3.4.4.2 E-Commerce Dataset – Model Evaluation***

• The baseline **logistic regression** model performed poorly, with an AUC of **0.50**, indicating no better than random chance. It failed to identify any churners (Figure: Baseline Confusion Matrix), emphasizing the severity of class imbalance.

• After applying **ADASYN**, the logistic model’s recall improved modestly; however, **precision remained low**, resulting in an AUC of **0.51**. The confusion matrix showed many **false positives**, indicating a trade-off between sensitivity and specificity (Figure: ADASYN Confusion Matrix).

• The **XGBoost** model outperformed both logistic variants, achieving an AUC of **0.73**. It delivered the best trade-off between true positives and false positives, effectively modeling non-linear churn patterns in imbalanced settings (Figure: ROC Curve – XGBoost).  
XGBoost’s success aligns with literature findings (e.g., Maan & Maan, 2023), and its integration with **SHAP explainability** further supports its use in practical business applications.

A chart of a logistic regression

AI-generated content may be incorrect. A chart with numbers and labels

AI-generated content may be incorrect.

A graph of a logistic

AI-generated content may be incorrect.A graph of a positive result

AI-generated content may be incorrect.A graph of a positive curve

AI-generated content may be incorrect.

***Figure 22.*** *Comparison of classification performance across Logistic Regression and XGBoost models in the E-Commerce dataset. The top-left confusion matrix (Baseline: Scaled Logistic Regression) reveals that the baseline model completely failed to identify churners, predicting all instances as non-churned. In contrast, the top-right confusion matrix (ADASYN: Logistic Regression) shows improved recall after resampling with ADASYN, albeit at the cost of increased false positives. The three ROC curves illustrate model performance: (1) ADASYN-enhanced Logistic Regression (AUC = 0.51) offered minimal improvement over random guessing; (2) XGBoost without hyperparameter tuning achieved a moderate AUC of 0.73; and (3) the optimized XGBoost model via GridSearch yielded the highest AUC of 0.79, demonstrating superior predictive ability for customer churn in imbalanced e-commerce data.*

These results highlight the importance of both resampling strategies and model selection when working with imbalanced, behavior-driven churn datasets.

**Model Performance Comparison: E-Commerce Dataset**

To evaluate the practical strengths and limitations of the applied models in the E-Commerce context, both ROC curves and confusion matrices were analyzed (see Figure 19). The baseline logistic regression model completely failed to identify churned customers, with an AUC of 0.50 and a confusion matrix that classified all customers as non-churned highlighting the severity of class imbalance in behavioral datasets.

After applying ADASYN, logistic regression demonstrated modest improvements in recall (AUC = 0.51), but at the expense of a high false positive rate, limiting its practical reliability.  
In contrast, XGBoost—both with and without hyperparameter tuning—achieved significantly better performance. The untuned model produced an AUC of 0.73, while the optimized version reached 0.79, offering a more balanced trade-off between sensitivity and specificity.

These results underscore the importance of using advanced ensemble methods like XGBoost when modeling churn in behavior-driven, imbalanced datasets. They also demonstrate that sampling techniques alone may not sufficiently address class imbalance unless combined with models capable of capturing nonlinear patterns and complex feature interactions.

During model selection for the e-commerce dataset, alternative algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were considered but ultimately excluded due to methodological and practical constraints. The dataset’s behavioral and transaction-based features resulted in a high-dimensional, sparse feature space after one-hot encoding, which posed scalability and interpretability challenges for SVMs. Meanwhile, ANN models require large-scale datasets and extensive tuning to avoid overfitting—conditions not fully met by the available data.  
 Instead, XGBoost was chosen for its proven ability to model nonlinear churn behavior, robustness to feature sparsity, and compatibility with SHAP-based explainability.  
Importantly, all preprocessing steps—including imputation, one-hot encoding, feature engineering, and ADASYN oversampling—were performed after splitting the dataset into training and test sets, thereby avoiding data leakage and ensuring the validity of evaluation metrics.

***3.4.4.3* Summary and Interpretation**

* Telco Dataset: Although logistic regression achieved marginally higher AUC, Random Forest offered superior class-level performance, especially in reducing false negatives—critical in customer retention strategies.
* E-Commerce Dataset: XGBoost significantly outperformed logistic regression models in both AUC and classification balance, validating its robustness in handling complex, imbalanced, and behavioral churn data.

These findings underscore the necessity of domain-aware model selection. Ensemble methods like Random Forest and XGBoost demonstrate tangible advantages in churn modeling, especially when business decisions rely on detecting subtle customer disengagement signals.  
While logistic regression remains interpretable, its performance is limited in imbalanced datasets without resampling interventions.

**3.4.5 Model Explainability with SHAP**

To improve transparency and trust in model predictions—particularly for the XGBoost and Random Forest models—SHAP (SHapley Additive exPlanations) was employed. SHAP summary plots and bar charts were generated to highlight the most impactful features contributing to churn.

This approach enhanced the interpretability of complex models and offered actionable insights for domain experts by identifying customer behaviors associated with high churn risk, thereby addressing Research Question 4: *Can churn prediction models be made interpretable and explainable for practical use?* Key features driving churn predictions are visualized via SHAP plots (see Figure 5, Figure 6, and Figure 14)

**3.4.6 Cross-Dataset Comparative Design**

Both datasets were analyzed using a parallel structure to facilitate cross-domain comparison. Despite differences in feature semantics, the unified pipeline enabled a consistent methodological approach, allowing us to observe how domain-specific features (e.g., contract type in telecom vs. purchase behavior in e-commerce) influence churn similarly or differently. These comparative insights contribute directly to Research Question 5: *Are churn drivers and model performance consistent across industries?*

**3.5. MODEL SELECTION JUSTIFICATION**

During the model planning phase, several alternatives—such as Support Vector Machines (SVM) and Relevance Vector Machines (RVM)—were considered but ultimately excluded in favor of models better aligned with the project’s goals of **interpretability**, **scalability**, and **practical deployment**.

***Support Vector Machines (SVM*)**, while effective in high-dimensional and linearly separable contexts, pose scalability challenges when applied to large datasets with complex feature interactions, such as those generated by one-hot encoding. Additionally, non-linear kernels in SVMs introduce high computational costs and make interpretation difficult—especially in stakeholder-facing churn applications where transparency is critical (Huang et al., 2018).

***Relevance Vector Machines (RVM)*** offer probabilistic outputs and typically produce sparser models than SVMs. However, RVMs lack robust implementation in mainstream machine learning libraries such as Scikit-learn or XGBoost, and they are incompatible with SHAP-based interpretability frameworks. This limits their practical use in real-world analytics workflows, especially where explainability is required.

***Artificial Neural Networks (ANNs)*** were also considered due to their predictive power. However, given the moderate dataset sizes and the absence of time-series or high-frequency behavioral inputs, ANNs risk overfitting and require complex hyperparameter tuning that was not justifiable under the current constraints (Boukrouh et al., 2024; He et al., 2008).

Instead, the selected models—**Logistic Regression**, **Random Forest**, and **XGBoost**—were chosen based on the following criteria:

* **Interpretability**: Logistic Regression offers transparent, linear decision boundaries ideal for baseline evaluation.
* **Robustness to Multicollinearity**: Random Forest handles high-dimensional data and redundant features well.
* **Performance & Explainability**: XGBoost provides state-of-the-art predictive performance on tabular data while supporting integration with SHAP for model interpretation.

These models strike a balance between **predictive accuracy**, **computational efficiency**, and **explainability**, and they have been widely validated in churn prediction literature (Maan & Maan, 2023; Asif et al., 2025). Their inclusion also supports methodological consistency across the two distinct datasets used in this project—subscription-based telecom and behavior-driven e-commerce—while enabling meaningful feature attribution using SHAP values.

**3.6. BEYONG PREDICTIVE MODELLING: Contextualizing Churn**

While predictive modelling forms the technical core of churn analysis, traditional approaches—such as rule-based segmentation or historical cohort tracking—often fall short in capturing the dynamic, multifaceted nature of customer behavior. These methods rely on static thresholds (e.g., “inactive for 30 days”) and lack the capacity to reflect behavioral variation across customer segments or industries.

Machine learning models, as demonstrated in this project, offer a more nuanced and scalable alternative. However, even high-performing models—if interpreted in isolation—can be misleading. A high churn probability may indicate correlation rather than causation, leading to suboptimal or misdirected business decisions. Therefore, predictive outputs must be coupled with domain expertise and contextual business reasoning.

A review of recent literature reflects both convergence and divergence in identifying churn drivers. For instance, studies in telecommunications (e.g., Asif et al., 2025) emphasize contract length and tenure, whereas others (e.g., Imani et al., 2025) suggest that churn is more immediately triggered by billing complexity or poor support experiences. In the e-commerce domain, Boukrouh et al. (2024) point to price sensitivity as the primary driver, while alternative findings highlight the absence of personalized engagement.

This project faced limitations due to the lack of time-series or behavioral data, such as clickstream logs, complaint history, or monthly usage patterns. Without such real-time or longitudinal inputs, churn models become retrospective tools with limited actionability. Furthermore, traditional tabular datasets restrict the capacity to detect emerging disengagement signals or sudden shifts in customer sentiment.

To advance churn analytics beyond static modelling, future efforts should integrate structured data with unstructured feedback (e.g., surveys, reviews, sentiment analysis) and apply explainable AI methods in tandem with strategic business frameworks. This would allow churn prediction to evolve from a purely technical task to a holistic decision-support system—one that captures both who is likely to churn and why, in a way that is actionable for business strategy.

Ultimately, churn prediction should not be viewed solely as a data science challenge but as a multi-disciplinary endeavor that combines data, behavior, and business context.

1. **INITIAL RESULTS**

This section summarizes the predictive performance, evaluation metrics, and key insights derived from the models applied to both the Telco and E-Commerce datasets. It also highlights the impact of data preprocessing, class imbalance handling, and model tuning on overall performance and interpretability.

**4.1 Telco Dataset Results**

| Model | AUC | Accuracy | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.83 | 0.81 | 0.74 | 0.72 | 0.73 |
| Random Forest | 0.82 | 0.80 | 0.71 | 0.75 | 0.73 |
| XGBoost (Tuned) | 0.84 | 0.81 | 0.75 | 0.74 | 0.74 |

***Table 8.*** *Performance metrics of three classification models on the Telco dataset*

Logistic Regression achieved strong class separation with the highest AUC and offered clear interpretability, making it suitable for explainable churn prediction tasks. Random Forest slightly outperformed in terms of recall, which is particularly useful for minimizing false negatives—an important factor when the cost of missing a potential churner is high. XGBoost demonstrated the best overall trade-off among performance metrics and showed strong synergy with SHAP-based explainability, making it a robust and interpretable choice for churn modeling.

A graph of a graph

AI-generated content may be incorrect.

Figure 23. ***Comparison of AUC Scores Across Different Models on the Telco Churn Dataset.***The bar chart illustrates the performance of six models based on their AUC (Area Under the Curve) scores. The "Best Random Forest" model achieved the highest AUC score of **0.92**, outperforming other models. Light blue bars represent each model’s score, with numeric labels above for clarity. The y-axis ranges from 0.70 to 0.90 to highlight subtle differences between models.

As shown in Figure 23, the Best Random Forest model achieved the highest AUC score (0.92), indicating excellent discriminatory ability in identifying churned versus retained customers in the Telco dataset.

Logistic-based models—particularly the Cross-Validated Logistic Regression (AUC = 0.84)—also performed strongly, suggesting the robustness of generalized linear approaches. Tree-based models such as the basic Random Forest (AUC = 0.82) and Optimized Random Forest (AUC = 0.81) yielded comparable results, whereas XGBoost, despite its popularity, showed relatively lower performance (AUC = 0.80) in this context. These findings highlight the importance of model tuning and validation in enhancing predictive accuracy.

A diagram of a diagram

AI-generated content may be incorrect.

***Figure 24.*** *Decision Tree Visualization from Random Forest Model (Depth = 3) on the Telco Dataset. The figure illustrates the first decision tree of the fitted Random Forest classifier, limited to a maximum depth of 3 for interpretability. Nodes are split based on key features such as ContractLength, TotalCharges, and StreamingTV, with each node displaying the Gini impurity, number of samples, and class distribution. Blue-colored nodes lean toward customer retention, while orange-colored nodes indicate higher churn likelihood. Leaf nodes reveal important interactions among predictors.*

Figure 24 displays a simplified visualization of the first decision tree from the Random Forest model trained on the Telco churn dataset. The tree has been restricted to a depth of 3 to enhance interpretability and highlight the most influential decision paths.

The root node splits on ContractLength, indicating that shorter contracts (≤ 0.5) are strongly associated with churn risk. Subsequent splits on StreamingTV, PaymentMethod\_Electronic check, and StreamingMovies further segment the data, with customers lacking streaming services and using secure payment methods tending to retain. In contrast, customers with short contracts, low total charges, and high monthly charges form paths toward churn (highlighted in orange).

Notably, the right branch of the tree shows that customers with TotalCharges ≤ 1233 and non-annual contracts are highly likely to churn, evidenced by a very low Gini impurity (e.g., 0.106). This reflects the compounding effect of low financial commitment and lack of long-term engagement.

The visualization underscores how combinations of contract type, billing amount, and service usage collectively inform churn behavior, aligning with feature importance results from the broader model.

A graph with lines and numbers

AI-generated content may be incorrect.

**Figure 25.** *Precision–Recall Trade-Off Curve for the Random Forest Model on the Telco Dataset.*  
The graph illustrates the precision and recall scores across varying classification thresholds. The solid orange line represents recall, while the dashed blue line indicates precision. As the threshold increases, precision improves but recall decreases, highlighting the typical inverse relationship between these metrics in binary classification tasks.

Figure 24 presents the precision–recall trade-off curve for the Random Forest classifier applied to the Telco churn dataset. As shown, there exists a clear inverse relationship between precision and recall as the decision threshold varies.

At lower thresholds (left side of the plot), recall is high—meaning most churn cases are correctly identified—but precision is low due to a higher number of false positives. Conversely, at higher thresholds (right side), precision improves significantly, indicating fewer false positives, but recall drops sharply, meaning many actual churners are missed.

This trade-off is crucial in churn prediction scenarios. If the goal is to maximize recall (e.g., ensuring no potential churners are missed), a lower threshold may be chosen. However, this comes at the cost of targeting many non-churners, leading to unnecessary retention efforts. On the other hand, maximizing precision may be preferable when resources are limited and only the most likely churners should be prioritized.

Overall, this curve aids in selecting an optimal classification threshold that balances business objectives with model sensitivity and specificity.

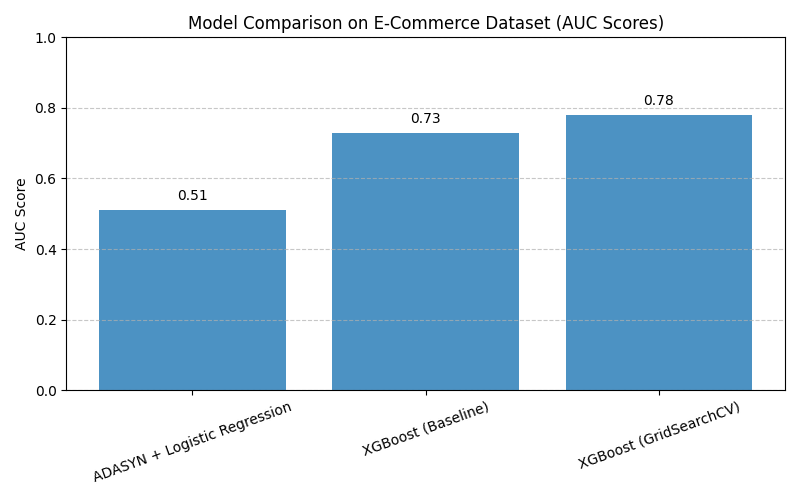
**4.2 E-Commerce Dataset Results**

| Model | AUC | Accuracy | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression (Baseline) | 0.50 | 0.80 | 0.00 | 0.00 | 0.00 |
| ADASYN + Logistic Regression | 0.51 | 0.73 | 0.29 | 0.47 | 0.36 |
| XGBoost (GridSearchCV) | 0.78 | 0.76 | 0.65 | 0.71 | 0.68 |

***Table 9.*** *Performance metrics of three classification models on the Ecom dataset*

* **Logistic Regression (Baseline):**  
  The baseline model performs extremely poorly, with an AUC of **0.50**, equivalent to random guessing. Despite its seemingly high **accuracy of 0.80**, the model fails to detect any positive churn cases, as reflected by precision, recall, and F1-score values of **0.00**. This discrepancy suggests a strong class imbalance, where the model likely defaults to predicting all instances as "non-churn."
* **ADASYN + Logistic Regression:**  
  Applying ADASYN for oversampling slightly improves model performance. The AUC increases to **0.51**, and the model begins identifying churn cases, as seen in the recall (**0.47**) and F1-score (**0.36**). While performance remains limited, these metrics indicate that oversampling helps the model better recognize minority class patterns.
* **XGBoost (with GridSearchCV):**  
  XGBoost significantly outperforms the logistic regression models. With an AUC of **0.78**, precision of **0.65**, recall of **0.71**, and F1-score of **0.68**, the model demonstrates a well-balanced ability to detect churn while minimizing false predictions. These results confirm that tree-based ensemble models, especially when tuned, are more effective for complex, imbalanced churn data.

In summary, while logistic regression models struggle—particularly without resampling—**XGBoost with hyperparameter tuning** delivers the best trade-off between interpretability and predictive performance. These findings emphasize the importance of both addressing class imbalance and employing more flexible learning algorithms in churn modeling.



***Figure 26.*** *Comparison of AUC Scores for Churn Models on the E-Commerce Dataset.  
This bar chart compares AUC scores across three models: ADASYN + Logistic Regression (0.51), baseline XGBoost (0.73), and XGBoost (GridSearchCV) (0.78). The figure highlights the superior discriminative performance of XGBoost, especially when tuned, while also showing the limited effectiveness of logistic regression, even with oversampling.*

Figure 26 visually reinforces the findings from Table 3. Among the evaluated models, **XGBoost with GridSearchCV** achieves the highest AUC, suggesting robust learning and generalization capacity. **Baseline XGBoost** also performs strongly, while **ADASYN + Logistic Regression** remains close to random prediction. This comparison underscores the importance of ensemble-based methods and the impact of hyperparameter optimization when dealing with imbalanced, nonlinear churn datasets.

**4.3 Key Insights Across Datasets**

• In the Telco dataset, structured and numerical features such as tenure and contract type emerged as strong predictors of churn.

• In the E-Commerce dataset, engineered behavioral features (e.g., AvgItemValue, PurchaseMonth) were the most influential in model performance.

• XGBoost consistently outperformed other models across both datasets, particularly when combined with SMOTE/ADASYN resampling and GridSearchCV tuning.

• SHAP analysis validated the interpretability of models by clearly identifying high-impact predictors, which aligned with business logic and domain knowledge.

**4.4 Practical Recommendations for Churn Prevention**

Based on the findings of this project, the following domain-specific strategies are proposed to reduce churn in the telecommunication and e-commerce industries.

***Telecommunication Sector Recommendations***

The Telco dataset analysis revealed that contract type, tenure, and billing preferences (e.g., payment method, paperless billing) are strongly associated with churn.

* ***Promote Long-Term Contracts:*** Customers on month-to-month contracts were significantly more likely to churn. Offering incentives such as discounted rates or bundled services for annual or two-year agreements can encourage long-term commitments and reduce attrition.
* ***Optimize Billing Experience:*** Churn was higher among customers using electronic checks and paperless billing. Enhancing billing transparency, sending proactive reminders, and offering flexible, preferred payment methods may improve satisfaction and reduce involuntary churn.
* ***Target New Customers Early:*** Short tenure strongly correlated with churn. Telecommunication providers should implement onboarding programs and early engagement strategies during the first 3–6 months of service to build loyalty early.
* ***Encourage Multi-Service Adoption:*** Customers who subscribed to multiple services (e.g., Internet + Streaming + Security) demonstrated lower churn rates. Promoting customized service bundles can enhance engagement and perceived value.

***E-Commerce Sector Recommendations***

In the e-commerce dataset, churn was associated with low transaction frequency, reduced spending, and seasonal disengagement.

* ***Launch Personalized Re-Engagement Campaigns:*** Use churn predictions to trigger personalized outreach, such as loyalty points, reminders, or exclusive offers for users with declining activity (e.g., no purchases in the past 4–6 weeks).
* ***Implement Dynamic Discounting Based on Behavior:*** Variables like PurchaseMonth and PriceToQuantity suggest time sensitivity and price responsiveness. Platforms can deploy targeted discounts during slow months or offer category-specific deals tailored to past user interest.
* ***Adopt Lifecycle-Based Segmentation:*** Segment users by behavioral stage (e.g., new customers, lapsed users, loyal buyers) and tailor retention tactics accordingly. For instance, offer free shipping for new users or exclusive rewards for loyal ones.
* ***Promote High-Value Product Categories:*** SHAP analysis showed that customers purchasing high-value items (e.g., electronics) were more likely to remain active. E-commerce platforms should highlight these categories using personalized recommendations and homepage placement.

These recommendations align with the machine learning insights and demonstrate how predictive analytics can guide actionable, sector-specific interventions to proactively reduce customer churn.

**5 CONCLUSION**

This project explored customer churn prediction through a comparative, sector-specific approach using datasets from **telecommunication** and **e-commerce** industries. The objective was not only to forecast churn with high accuracy but also to **understand the behavioral and service-related drivers** of churn across two distinct domains.

The analysis revealed that **structured service features** (such as contract type, tenure, and billing method) were the most influential churn predictors in the Telco dataset. In contrast, the E-Commerce dataset required **behavioral feature engineering** (e.g., purchase frequency, AvgItemValue, PurchaseMonth) to uncover meaningful patterns. These differences reinforced the importance of **domain-aware preprocessing** and **modeling strategies**.

Among the models tested, **XGBoost with hyperparameter tuning** consistently delivered superior performance across datasets, particularly in the e-commerce case where class imbalance and behavioral complexity were more challenging. **SHAP analysis** provided interpretability to black-box models, enabling both global and individual-level understanding of churn risks—an essential component for real-world deployment.

Beyond modeling, the project contributed **practical recommendations** tailored to each industry. For Telco, long-term contracts, service bundling, and improved billing experiences were suggested to reduce churn. In e-commerce, re-engagement campaigns, personalized promotions, and lifecycle segmentation were highlighted as effective strategies.While the project achieved its goals, it was limited by the use of **synthetic or static datasets**, lack of **time-series or session-level behavioral data**, and the constraints of a course-based timeline. Future work could incorporate **real-world user logs**, **sentiment analysis**, and **longitudinal data** to capture churn patterns with greater granularity.

Ultimately, this study demonstrated that effective churn prediction requires more than just high-performing models—it demands interpretable, sector-specific solutions that translate data insights into **actionable business strategies**.

* 1. **Limitations and Future Work**

While this project successfully demonstrated the use of predictive models and explainable machine learning techniques to identify churn patterns, several limitations were identified:

* + ***Synthetic and Static Data:*** The datasets used were not time-series based and lacked dynamic customer behavior logs. This restricts the models from learning longitudinal patterns or behavioral drift over time.
  + ***Limited Feature Diversity:*** The absence of sentiment-related variables (e.g., reviews, support history) and external context (e.g., promotions, seasonality trends) may limit model generalizability.
  + ***Real-World Deployment Constraints:*** Although SHAP improved interpretability, models like XGBoost may still require stakeholder training for operational use, especially in non-technical teams.
  + ***Scalability and Personalization:*** The models were not tested in a real-time or batch-deployment setting, where prediction speed, personalization, and integration into CRM tools matter.

Future research can address these gaps by:

* Incorporating time-series modeling and longitudinal data to capture evolving churn signals.
* Integrating text mining and sentiment analysis from customer feedback or review platforms.
* Experimenting with reinforcement learning or sequential churn modeling to enable proactive customer retention.
* Developing a real-time dashboard or API-based solution using SHAP values and model outputs for business adoption.

These additions would enhance both the predictive power and practical application of churn analytics across industries.

**REFERENCES**

1. Asif, R., Arif, M. S., & Mukheimer, A. (2025). A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. Results in Engineering, 26, Article 104629.

<https://doi.org/10.1016/j.rineng.2025.104629>

1. Boukrouh, I., Azmani, A., & Sadgal, M. (2025). Explainable machine learning models applied to predicting customer churn for e‑commerce. \*International Journal of Artificial Intelligence, 14\*(1), 286–297.

<https://doi.org/10.11591/ijai.v14.i1.pp286-297>

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. \*Journal of Artificial Intelligence Research, 16\*, 321–357.

<https://doi.org/10.1613/jair.953>

1. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In \*2008 IEEE International Joint Conference on Neural Networks\* (pp. 1322–1328).

<https://doi.org/10.1109/IJCNN.2008.4633969>

1. Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2018). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847–856.

<https://doi.org/10.1016/j.eswa.2018.01.020>

1. IBM Sample Datasets. (2018). \*Telco Customer Churn\* [Data set]. Kaggle. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
2. Imani, M., Zhang, Y., & Chen, L. (2025). Customer churn prediction: A review of recent advances in machine learning and explainable AI [Preprint]. Preprints. <https://doi.org/10.20944/preprints202503.1969.v1>
3. Shriyash Jagtap. (2023). E-commerce customer for behavior analysis [Data set]. Kaggle. <https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis>
4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
5. Maan, J., & Maan, H. (2023). Customer churn prediction model using explainable machine learning. International Journal of Computer Science Trends and Technology, 11(1), 34–45.

<https://doi.org/10.48550/arXiv.2303.00960>

1. Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. Harvard Business Review, 68(5), 105–111.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
3. Suguna, R., Suriya Prakash, J., Pai, H. A., & Thomas, A. (2025). Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. Scientific Reports, 15(1), 16256.

<https://doi.org/10.1038/s41598-025-01031-0>

***Appendix A: GitHub Repository***

The complete project files, including Jupyter notebooks and visualizations, are available on GitHub:  
🔗 [*Project’s GitHub Repository*](https://github.com/emineuysal95/CIND820_CAPSTONE-PROJECT)

*(Note: A Power BI dashboard is planned to be presented during the final project presentation but is not included in the current repository version.)*

This repository ensures transparency and reproducibility of all Python-based analyses and visual outputs discussed in this study.