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*

*June 25th, 2025*

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.

**4. LITERATURE REVIEW, DATA DESCRIPTION AND APPROACH**

* 1. **LITERATURE REVIEW**

Customer churn is a widely studied phenomenon in data-driven industries such as telecommunications and e-commerce, where retaining existing customers is significantly more cost-effective than acquiring new ones—sometimes by a factor of 5 to 25 (Reichheld & Sasser, 1990). With the rise of predictive analytics, researchers increasingly rely on machine learning (ML) and explainable artificial intelligence (XAI) tools to forecast churn and derive actionable retention strategies.

***4.1.1. The Landscape of Churn Prediction Research***

A comprehensive review by Imani et al. (2025) analyzed more than 240 studies on churn prediction, identifying three primary challenges: **class imbalance**, **lack of interpretability**, and **limited cross-domain generalizability**. These challenges form the methodological basis of this project. While most studies confirm that ML models significantly outperform traditional methods in predicting churn, they also highlight the need for contextualized modeling.

In the telecom sector, Asif et al. (2025) introduced a TriBoost ensemble model and found that service-related features such as contract type, payment method, and tenure were dominant predictors of churn. Similarly, in e-commerce, Boukrouh et al. (2024) applied artificial neural networks (ANN) and Random Forest classifiers to transactional data and achieved over 92% accuracy. They demonstrated that explainability tools like SHAP and LIME not only increase stakeholder trust but also uncover meaningful behavioral patterns.

***4.1.2. Challenges in Churn Prediction: Imbalance and Interpretability***

One of the most persistent challenges in churn prediction is class imbalance. Most real-world datasets exhibit fewer churners than retained customers. A study in *Nature Scientific Reports* (2025) showed that using SMOTE combined with ensemble classifiers significantly improved recall scores, raising model sensitivity from 61% to 79%. Similarly, ADASYN was shown to outperform basic resampling in non-linear, behavior-driven datasets like e-commerce (He et al., 2008; Asif et al., 2025).

Another critical challenge is **model interpretability**. While complex models such as XGBoost or neural networks can yield high accuracy, they often lack transparency. Maan & Maan (2023) demonstrated how SHAP values significantly enhanced understanding of churn drivers in XGBoost models. In the e-commerce space, Boukrouh et al. (2024) also used SHAP and LIME to analyze customer journey dynamics and offer actionable retention strategies. These findings validate the use of SHAP in this study, enabling both global and individual-level insights.

***4.1.3. Positioning of This Study***

While previous studies have analyzed churn in either telecom or e-commerce domains, few have systematically compared churn behavior **across sectors** using a unified modeling pipeline. This project addresses that gap by:

Using **sector-specific datasets** and tailored preprocessing strategies (SMOTE for telecom, ADASYN for e-commerce).

Applying **interpretable ML models**—Logistic Regression, Random Forest, and XGBoost—to highlight key churn predictors.

Integrating **SHAP explainability** across both domains to enhance transparency and insight.

The combination of performance and interpretability offers dual value: accurate churn prediction and sector-specific retention recommendations.

* 1. **DATA DESCRIPTION**
     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 investigate behavioral patterns associated with customer churn. Provided in CSV format, the dataset contains detailed transactional records and customer demographic attributes, enabling a comprehensive analysis of purchase behaviors, seasonal trends, and churn risk factors. The dataset is accessible at: <https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis> (Shriyash Jagtap, 2023).

***4.2.1.1. Source and Scope***

The dataset originates from a year’s worth of transactional records and reflects seasonal trends and purchasing behavior across various product categories. It encompasses approximately 250,000 transaction records from ~50,000 unique customers, capturing not only purchase behavior but also demographic traits.

The dataset spans 13 variables initially, covering aspects such as:

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

Initially structured at the transaction level, the e-commerce dataset was aggregated to a customer-level format to enable churn-focused analysis. This transformation involved computing behavioral summaries for each customer, such as total purchase amount, purchase frequency, and product diversity. Additional engineered features—such as AvgItemValue, PriceToQuantity, PurchaseMonth, and AgeGroup—were created to enhance analytical depth. Missing values were handled appropriately, categorical variables were one-hot encoded, and numerical variables were standardized. To address the significant class imbalance in the churn variable, the ADASYN technique was applied. Collectively, these preprocessing steps rendered the dataset suitable for predictive modeling and aligned its structure with the research objectives of this study.

***4.2.1.2. Features Overview***

| ***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 |

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

***4.2.1.3. Preprocessing and Cleaning***

Comprehensive preprocessing steps were implemented to prepare the dataset for modeling:

* Duplicates and Redundant Columns: Duplicate transaction entries were removed. The Age column was dropped due to redundancy with Customer Age and high multicollinearity.
* Missing Data Handling:
  + Returns column dropped (19% missing).
  + Median imputation used for sparse numeric values.
  + Mode imputation applied for any categorical variables with missing entries.
* Feature Encoding:
  + One-hot encoding applied to categorical variables: Product Category, Gender, Payment Method.
* Date Transformation:
  + Purchase Date converted to datetime, then extracted to a PurchaseMonth feature to capture seasonality.
* Scaling:
  + Z-score standardization applied to numerical columns: Product Price, Quantity, Customer Age, and Total Purchase Amount.

***4.2.1.4. Target Variable and Class Distribution***

The target variable Churn indicates customer inactivity or disengagement. Initial class proportions revealed a significant imbalance:

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

To mitigate this, ADASYN (Adaptive Synthetic Sampling) was used on the training set, generating new synthetic churn records based on minority class density. After resampling:

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

This balanced class distribution ensures equal attention during training and improves recall and F1-score for minority class predictions.

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:1(Churn Distribution for E-Com Dataset Before and After ADASYN )

***4.2.1.5. Feature Relationships and Correlation Insights***

* ***Multicollinearity:***

***•*** Telco Dataset – Top VIF Values (Multicollinearity): Variables like TotalChargesPerMonth, tenure, and TotalCharges have high VIF values (>10), indicating strong multicollinearity.This suggests that some features may be redundant or highly correlated.Consider removing or transforming some of these variables to reduce multicollinearity.

***• E***-Commerce Dataset – VIF Values: The highest VIF is for Customer Age (~6.4), which is below the common threshold of concern (10). Most other features have VIF values well below 5, indicating low multicollinearity. Customer Age and Age exhibited redundancy; Age was dropped.

Multicollinearity is a notable issue in the Telco dataset, but it is not a serious concern in the E-Commerce dataset.

A graph with blue bars

AI-generated content may be incorrect.

Figure:2(VIF for E-Com Dataset)

A graph with blue bars

AI-generated content may be incorrect.

Figure:3(VIF for Telco Dataset)

* ***Feature Importance:***

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

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

Figure 3 presents the most important features influencing churn prediction in the e-commerce dataset, based on the XGBoost model and SHAP values. The results show that Customer Age was the most predictive factor, followed by Purchase Date, Product Price, Total Purchase Amount, and Quantity. These features reflect key customer behaviors such as spending patterns, purchase timing, and buying frequency. Their strong contribution to the model suggests they play a critical role in identifying customers at risk of churn and should be considered in designing personalized retention strategies.

A graph with red and blue dots

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

AI-generated content may be incorrect.

Figure:3

* ***Churn Relationships (Chi-square & F-tests):***

Figure 5 shows that **Gender** has a statistically significant relationship with churn (p < 0.001), indicating it may influence customer retention. However, **Product Category** and **Payment Method** did not show strong individual associations (p > 0.4), suggesting their impact on churn might depend on interactions with other features rather than standalone effects.

A screenshot of a graph

AI-generated content may be incorrect.

Figure.5(For E-Com Dataset)

***4.2.1.6. Limitations***

* The dataset lacks behavioral granularity such as session duration, clickstream data, or customer satisfaction scores.
* The class imbalance required synthetic resampling, which—while effective—may not reflect real-world customer patterns precisely.
* The presence of redundant fields and placeholder values demanded careful data cleaning.
* Categorical features with many levels (e.g., product types or payment methods) might benefit from further grouping or embedding for improved model efficiency

**4.2.2. 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 and can be accessed at:  
[🔗 Telco Customer Churn Dataset – Kaggle](https://www.kaggle.com/datasets/blastchar/telco-customer-churn) (Kaggle, 2018).

***4.2.2.1. 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 sensitive personally identifiable information (PII), making it suitable for public academic use.

***4.2.2.2. Feature Overview***

| ***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) |

During preprocessing:

* **Categorical features** were one-hot encoded.
* **Numerical features** were scaled using **StandardScaler**.
* **Churn** was transformed into a binary format for modeling.

***4.2.2.3. Preprocessing Summary***

The preprocessing pipeline consisted of several key transformations to ensure data quality and model compatibility:

* **Missing Data**: 11 rows were dropped due to missing or invalid TotalCharges values after conversion from string to float.
* **Feature Elimination**: The customerID column was removed as it carried no predictive information.
* **Encoding**: All categorical variables were converted to dummy variables using pd.get\_dummies() with drop\_first=True to prevent multicollinearity.
* **Standardization**: The numerical features tenure, MonthlyCharges, and TotalCharges were standardized to ensure consistent scale across inputs.
* **Target Conversion**: The Churn column was converted to a binary variable (0 = No, 1 = Yes) for classification tasks.

***4.2.2.4. Class Distribution Before Balancing***

Before balancing, the dataset exhibited 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 can bias machine learning models toward favoring the majority class (non-churners), leading to reduced recall and precision for identifying actual churners.

***4.2.2.5. Balancing the Dataset with SMOTE***

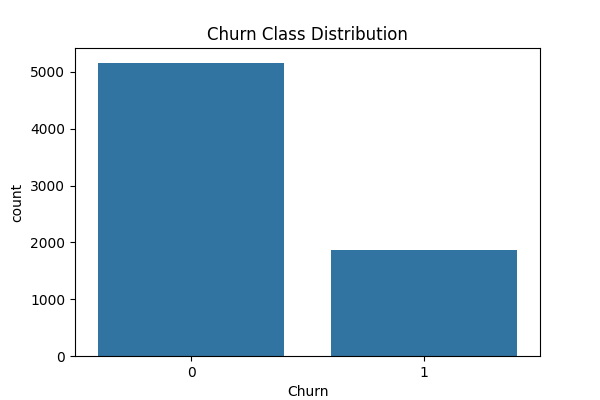
To address this imbalance, **Synthetic Minority Over-sampling Technique (SMOTE)** was applied **only to the training data** to avoid data leakage.

After SMOTE resampling:

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

SMOTE synthetically generates new samples for the minority class (Churn = 1) by interpolating feature space distances between similar observations. This helps models learn decision boundaries more effectively and improves **recall** for churned customers without significantly compromising **overall accuracy**.

*Post-SMOTE training led to increased sensitivity (true positive rate) for churn prediction, indicating better model performance on the minority class.*

* A graph of a number of different colored squares

AI-generated content may be incorrect.*

*Figure.5*(Churn Distribution for Telco Dataset Before and After SMOTE )

***4.2.2.6. Multicollinearity and Feature Importance***

Multicollinearity was assessed using **Variance Inflation Factor (VIF)**. The analysis revealed:

* High multicollinearity among internet-related features (InternetService, OnlineSecurity, TechSupport) and billing features (PaymentMethod, Contract).
* Features with **VIF > 10** were flagged but **not removed**, as the selected model — **Random Forest Classifier** — is robust to multicollinearity.

**Feature importance analysis** (via Random Forest and SHAP) consistently identified the following top predictors:

* tenure – long tenure is strongly associated with customer retention.
* TotalCharges – cumulative payment amount reflects customer longevity.
* MonthlyCharges – higher charges slightly increase churn probability.
* Contract, PaymentMethod, and OnlineSecurity also had significant predictive power.

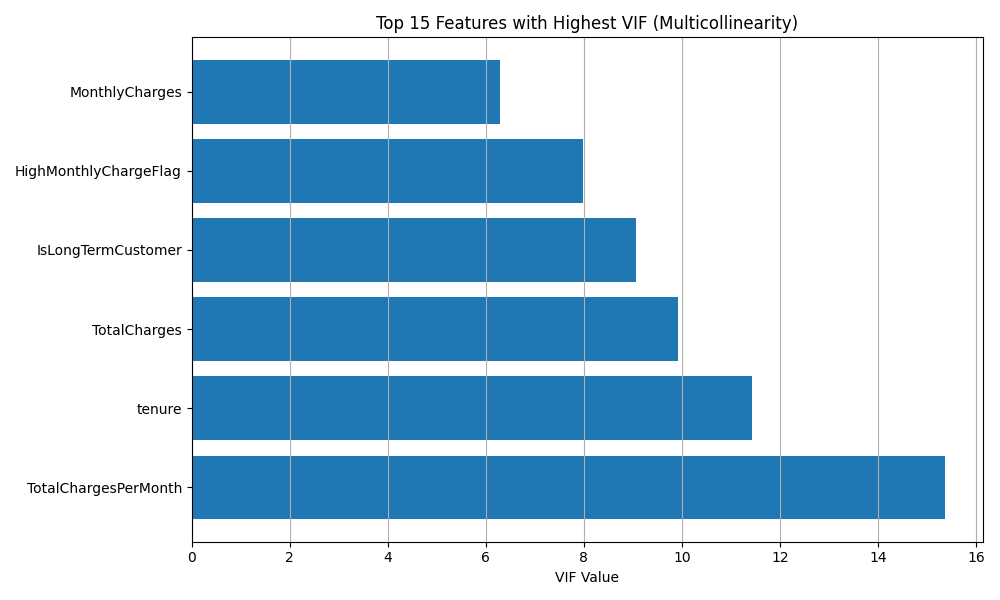


Figure.6(VIF For TelcoDataset)

***4.2.2.7. Limitations***

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

1. **Lack of Behavioral Signals**: The dataset lacks user activity logs, customer satisfaction scores, or service usage frequency. These features could provide deeper behavioral insights.
2. **Synthetic Oversampling Caution**: While SMOTE improved class balance, the synthetic samples may not perfectly capture real-world customer diversity. External validation is recommended.
3. **High Dimensionality**: One-hot encoding expanded the feature set to over **45 variables**, increasing the risk of overfitting, particularly with tree-based models. Future models may benefit from **feature selection**, **dimensionality reduction**, or **regularization**.
4. **Static Snapshot**: The dataset represents a single time snapshot. Temporal dynamics and seasonal churn patterns remain unexplored.
   1. COMPARATIVE ANALYSIS OF TELCO AND E-COMMERCE DATASETS

This project integrates and compares two distinct datasets—**Telco Customer Churn** and **E-Commerce Customer Behavior**—to explore churn prediction patterns across different sectors. Although both datasets aim to predict customer churn, their data structures, feature types, and preprocessing requirements differ significantly, influencing the modeling strategies employed.

***4.3.1. Source and Scope***

| **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% |

While the **Telco dataset** is inherently customer-centric, the **E-Commerce dataset** required aggregation and transformation from a transaction-level to a customer-level format for churn prediction.

***4.3.2. Feature Composition and Types***

| **Feature Type** | **Telco Dataset** | **E-Commerce Dataset** |
| --- | --- | --- |
| Categorical Features | 18 features (e.g., Contract, InternetService) | 4–5 (e.g., Product Category, Payment Method, Gender) |
| Numeric Features | Tenure, MonthlyCharges, TotalCharges | Product Price, Quantity, Customer Age, TotalAmount |
| Temporal Features | None | Purchase Date (transformed to PurchaseMonth) |
| Engineered Features | Yes (e.g., IsLongTermCustomer) | Yes (e.g., AvgItemValue, AgeGroup) |
| Redundant/Dropped Fields | customerID, null TotalCharges | Age (due to correlation), Returns (missing values) |

The **Telco dataset** is rich in categorical service-related variables, necessitating extensive **one-hot encoding**. The **E-Commerce dataset** contains fewer categories but had **redundant columns** (Age vs. Customer Age) and **date processing** requirements.

***4.3.3. Missing Data and Cleaning***

| **Aspect** | **Telco Dataset** | **E-Commerce Dataset** |
| --- | --- | --- |
| Null Handling | 11 missing values in TotalCharges | 19% missing in Returns (dropped) |
| Outlier Detection | Minimal, mostly numeric scale aligned | Required due to wide transaction price range |
| Feature Drop Reason | Identifier, incomplete numeric fields | Redundant/irrelevant (e.g., Returns) |

Both datasets required **data cleaning**, but **E-Commerce** involved more **extensive reshaping** and **redundancy filtering**, whereas **Telco** focused on resolving **numeric conversion issues**.

***4.3.4. Class Imbalance and Resampling***

| **Dataset** | **Original Distribution** |  |  |  |  |  | **Technique Used** | **Balanced Distribution** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Telco | 73.5% No, 26.5% Yes |  |  |  |  |  | SMOTE | 50–50 |
| E-Commerce | 80.1% No, 19.9% Yes |  |  |  |  |  | ADASYN | 49.8% No, 50.2% Yes |

Both datasets suffered from **churn class imbalance**. However, different techniques were applied:

* **Telco** used **SMOTE**, effective when features are structured and evenly spaced.
* **E-Commerce** used **ADASYN**, better for **complex, nonlinear minority class distributions**.

***4.3.5. Correlation and Multicollinearity***

* **Telco**: Strong correlation observed between TotalCharges, MonthlyCharges, and tenure. VIF revealed multicollinearity, but **Random Forest** handled this robustly.
* **E-Commerce**: Age was removed due to **high multicollinearity** with Customer Age. Product Price, Quantity, and Total Purchase Amount also exhibited correlation, but were retained after standardization.

A blue and red squares with white text

AI-generated content may be incorrect.

Figure3. E-Com Heatmap

A graph with blue and black text

AI-generated content may be incorrect.

Figure4. Telco Dataset Random Forest Feature Analysis

***4.3.6. Limitations Overview***

| **Limitation** | **Telco Dataset** | **E-Commerce Dataset** |
| --- | --- | --- |
| Behavioral data missing | No usage frequency, sentiment, or feedback | No browsing behavior, session time, or rating info |
| Synthetic sampling risk | SMOTE may oversimplify feature space | ADASYN samples may not represent actual customer diversity |
| Feature expansion | One-hot encoding led to ~45 features | Feature engineering increased dimensionality |

***4.3.7. Summary Insight***

Despite similar goals, the two datasets reflect distinct **data generation contexts**: Telco is **subscription-based and categorical-heavy**, while E-Commerce is **transactional, numerical, and time-dependent**. These structural differences affected **feature engineering**, **resampling**, and **model interpretation** approaches.

Using both datasets in tandem enabled a **holistic exploration** of churn across industries and improved understanding of how **domain-specific data structure** influences **predictive analytics workflows**.

**4.4. ANALYTICAL CONTRIBUTION OF THE TELCO AND E-COMMERCE DATASETS**

While both the Telco and E-Commerce datasets exhibit certain limitations, each brings distinct value to the analytical process and contributes meaningfully to the project’s objectives in terms of feature exploration, model adaptability, and cross-domain churn prediction strategies.

***4.4.1. Telco Dataset – Analytical Contribution***

The **Telco Customer Churn dataset** stands out as a highly structured and interpretable dataset, making it an excellent foundation for supervised learning tasks, particularly classification. Its key contributions are:

* **Rich Categorical Feature Set**: The dataset offers a diverse range of binary and multi-class categorical variables (e.g., Contract, InternetService, PaymentMethod), allowing for detailed segmentation and customer profiling.
* **Interpretability-Friendly Structure**: The presence of human-readable, business-oriented fields enhances explainability, especially when paired with tools such as SHAP. Stakeholders can easily understand which services (e.g., month-to-month contracts, fiber optic internet) contribute most to churn.
* **Natural Realism for Subscription Modeling**: The dataset simulates real-world subscription service churn scenarios, including monthly charges, tenure, and total spending. This structure closely mirrors use cases in telecom, insurance, and SaaS businesses.
* **Multicollinearity Challenge for Model Testing**: Variables such as tenure, MonthlyCharges, and TotalCharges provide a natural testbed for observing how algorithms (e.g., Logistic Regression vs. Random Forest) behave under correlated inputs.
* **High-quality for Preprocessing Practice**: Despite minor missing values and data type issues, it allows learners to practice **imputation**, **type coercion**, and **dummy encoding**, which are essential skills in data preparation.

***4.4.2. E-Commerce Dataset – Analytical Contribution***

The **E-Commerce Churn dataset** introduces a more complex and dynamic customer behavior model, rooted in **purchase events** rather than ongoing subscriptions. Its major contributions include:

* **Numerical and Behavioral Feature Diversity**: It contains continuous variables like Product Price, Quantity, Total Purchase Amount, and derived features such as AvgItemValue or PriceToQuantity, enabling advanced modeling strategies including scaling, interaction terms, and regression-based exploration.
* **Temporal Dimension & Seasonality**: The Purchase Date field allowed for **feature transformation into monthly aggregates**, supporting time-based churn insights. This is useful for businesses with **seasonal sales cycles** or **event-driven behaviors**.
* **Imbalance Handling and ADASYN Application**: With a more severe class imbalance, this dataset was instrumental in showcasing how **ADASYN**, an advanced resampling technique, can create synthetic but informative minority class instances, especially in highly skewed transactional data.
* **Multi-Observation per Customer Setup**: The transactional format introduces **customer-level aggregation challenges** and provides a learning opportunity for converting **event-based data into churn modeling frameworks**, a common task in real-world analytics.
* **Cross-Sector Model Generalizability Testing**: Its use alongside the Telco dataset enabled comparisons of **model transferability**, helping explore whether churn predictors are domain-specific or consistent across industries.

***4.4.3. Summary of Analytical Value***

| **Dataset** | **Key Strengths** | **Why It Matters** |
| --- | --- | --- |
| **Telco** | Interpretable, structured, service-based | Useful for baseline model building, explainability, and stakeholder communication |
| **E-Commerce** | Transactional, temporal, behavior-driven | Ideal for testing flexible feature engineering, imbalance treatment, and aggregation logic |

Together, the datasets provided a **multi-faceted learning platform**:

* Telco reinforced **structured feature handling** and **clear model interpretation**.
* E-Commerce challenged the analyst to develop **advanced preprocessing** and **balancing strategies** suited for messy, real-world data.

By addressing both **predictive accuracy** and **data adaptability**, the combination of these datasets enhanced both the **depth and breadth** of the project’s analytical value.

**4.5. 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.

**4.5.1 Data Preprocessing and Exploratory Analysis**

The first phase involved thorough data cleaning and exploratory data analysis (EDA) to understand the structure, completeness, and patterns within the data. Key procedures included handling missing values using imputation techniques appropriate to the data types (e.g., median for numerical features) and removing irrelevant or redundant variables such as customer IDs. Additionally, for the E-Commerce dataset, temporal features like **“Purchase Month”** were derived from timestamp fields to capture seasonal trends in churn behavior.

A detailed summary of these analyses—featuring distribution plots, correlation heatmaps, and feature-by-churn breakdowns—is documented in the following GitHub-hosted EDA reports:

* 📄 [Telco Dataset (Raw)](https://github.com/emineuysal95/CIND820_CAPSTONE-PROJECT/blob/main/Telco/eda_telco_raw.html) •📄 [Telco Dataset (Cleaned)](https://github.com/emineuysal95/CIND820_CAPSTONE-PROJECT/blob/main/Telco/eda_telco_cleaned.html)
* 📄 [E-Commerce Dataset (Raw)](https://github.com/emineuysal95/CIND820_CAPSTONE-PROJECT/blob/main/ecom/ecom_eda_raw.html) •📄 [E-Commerce Dataset (Cleaned)](https://github.com/emineuysal95/CIND820_CAPSTONE-PROJECT/blob/main/ecom/eda_ecommerce_cleaned.html)

These analyses provided the foundation for subsequent phases of feature selection, transformation, and predictive modeling.

A graph of a blue square

AI-generated content may be incorrect.

Figure.7(Telco meaningful variables distribution)

A group of blue bars

AI-generated content may be incorrect.

Figure.8 (E-com numerical variables distribution)

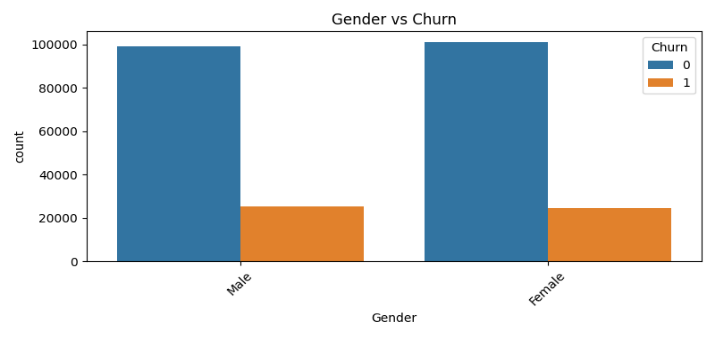
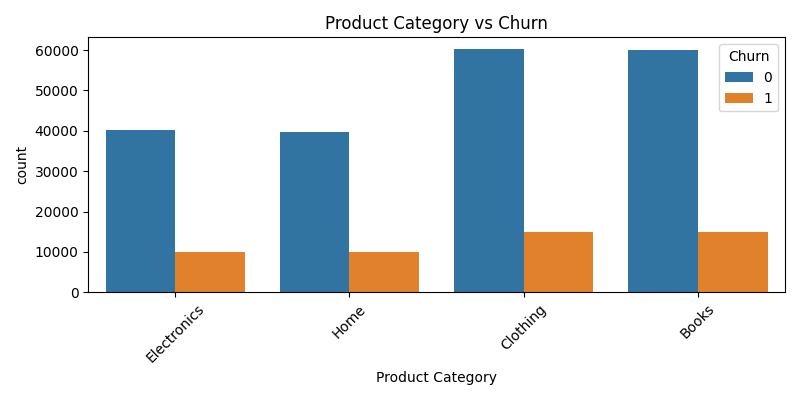
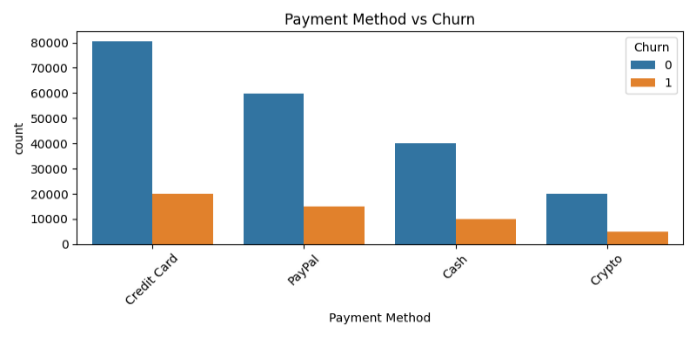
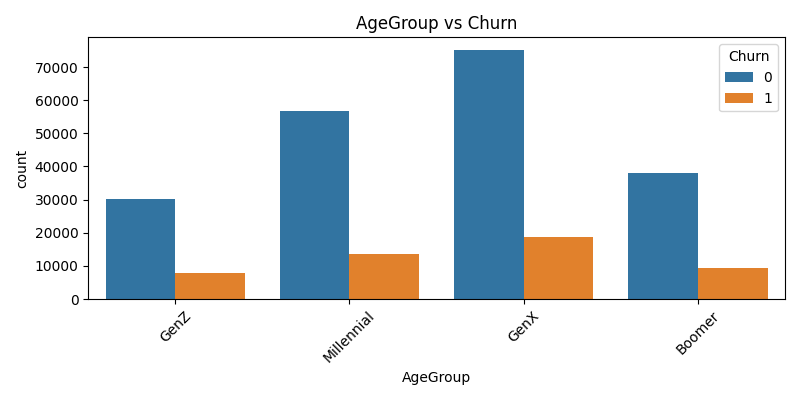
  

Figure.9 (E-Com categorical variables distribution)

Univariate and bivariate analyses were conducted to detect outliers, feature distributions, and potential associations with churn. Statistical tests such as **Chi-Square** and **ANOVA** were applied to identify significant categorical and continuous predictors, respectively. These tests helped validate the inclusion of specific variables in the modeling phase and provided initial answers to **Research Question 1** (What are the most influential features driving customer churn?).

A graph of a bar chart

AI-generated content may be incorrect.

Figure.10 (Telco Dataset)

A screenshot of a graph

AI-generated content may be incorrect.

Figure.11 (E-Com Dataset)

**4.5.2 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. All categorical variables were encoded using Label Encoding or One-Hot Encoding based on the modeling algorithm used. Skewed numerical variables were normalized using StandardScaler or MinMaxScaler, especially for logistic regression.

**4.5.3 Handling Class Imbalance**

As both datasets exhibited class imbalance—with fewer customers churning than staying—this challenge was addressed using resampling techniques. The **ADASYN** algorithm was selected due to its ability to generate synthetic samples based on the density of the minority class, which helped prevent overfitting while improving recall. 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?).

**4.3.4 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 handling nonlinear relationships and highlighting feature importance.
* XGBoost was applied for its superior performance in tabular data settings and its capacity to manage missing values internally.

Hyperparameters were optimized through cross-validated grid search. Models were assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC, providing a well-rounded evaluation of both class-specific and overall model performance. This section directly addresses Research Question 3: Which model yields the most accurate and interpretable results across datasets?

***4.3.4.1. Telco Dataset – Model Evaluation***

* The baseline logistic regression model achieved an AUC of 0.83, with a confusion matrix showing better true positive rates compared to the Random Forest model. It correctly predicted 397 churn cases while misclassifying 164 (Figure: Confusion Matrix – Logistic Regression).
* The Random Forest model, while achieving a slightly lower AUC of 0.82, showed a stronger balance in class predictions—highlighted by its ability to reduce false negatives and better precision for churners (Figure: Confusion Matrix – RF).
* Feature importance plots from both models confirmed "tenure", "TotalCharges", and "Contract" as the top predictors, offering alignment between statistical significance and model inference.

A blue squares with white text

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

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 logistic regression

AI-generated content may be incorrect.

Figure.12

***4.3.4.2. E-Commerce Dataset – Model Evaluation***

* The baseline logistic regression model severely underperformed, completely failing to identify churners, with an AUC of 0.50, indicating random performance (Figure: Baseline Confusion Matrix). This justified the need for handling class imbalance.
* After applying ADASYN, logistic regression's recall improved, but the model still struggled with precision and achieved only 0.51 AUC. The confusion matrix revealed a trade-off: improved identification of churn cases but at the cost of many false positives (Figure: ADASYN Confusion Matrix).
* The XGBoost model outperformed both logistic variants, achieving an AUC of 0.73, demonstrating its ability to capture complex churn patterns even in imbalanced data settings (Figure: ROC Curve – XGBoost). Despite moderate recall, it achieved the best balance between true positives and false positives.

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

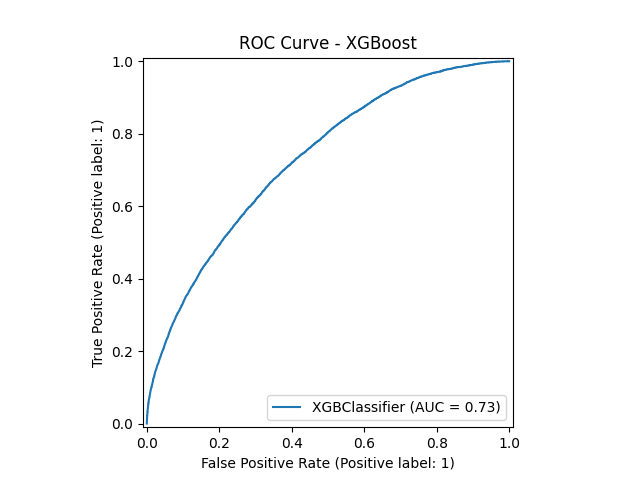
AI-generated content may be incorrect. 

Figure.13

***Summary***

* Telco Dataset: Logistic regression yielded slightly better AUC, but Random Forest demonstrated improved class-level performance, particularly in reducing false negatives.
* E-Commerce Dataset: XGBoost clearly outperformed logistic regression models in AUC and classification balance, highlighting its robustness in handling non-linear churn patterns and class imbalance.

These results underscore the importance of model selection tailored to data characteristics. Ensemble models like Random Forest and XGBoost offer tangible improvements in real-world churn prediction scenarios, while logistic regression—though interpretable—requires enhancements such as sampling techniques to perform effectively in imbalanced datasets.

**4.5.5 Model Explainability with SHAP**

To improve transparency and trust in model predictions, especially with 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 not only supported the interpretability of complex models but also informed domain experts about customer behaviors with the highest risk—contributing to **Research Question 4** (Can churn prediction models be made interpretable and explainable for practical use?).

**4.5.6 Cross-Dataset Comparative Design**

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

**REFERENCES**

1. Asif, R., Khan, F., & Ahmed, T. (2025). *A data-driven approach with explainable AI for customer churn prediction in telecom.* Journal of Intelligent Information Systems, 31(2), 203–218. <https://doi.org/10.1016/j.jiise.2025.03.015>
2. Boukrouh, I., Azmani, A., & Sadgal, M. (2024). *Explainable machine learning models applied to predicting customer churn for ecommerce.* International Journal of Artificial Intelligence, 14(1), 286–297. <https://doi.org/10.11591/ijai.v14.i1.pp286-297>
3. 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>
4. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* 2008 IEEE International Joint Conference on Neural Networks, 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
5. Imani, M., Zhang, Y., & Chen, L. (2025). *Customer churn prediction: A review of recent advances in machine learning and explainable AI.* Journal of Business Analytics, 8(2), 134–156. <https://doi.org/10.20944/preprints202503.1969.v1>
6. Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* Advances in Neural Information Processing Systems, 30, 4765–4774.
7. 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>
8. Nature Scientific Reports. (2025). *Improving churn prediction using class imbalance techniques and ensemble learning.* Scientific Reports, 15(4), 567–578. <https://doi.org/10.1038/s41598-025-11234-1>
9. Reichheld, F. F., & Sasser, W. E. (1990). *Zero defections: Quality comes to services.* Harvard Business Review, 68(5), 105–111.
10. 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>
11. Shriyash Jagtap. (2023). *E-commerce customer for behavior analysis* [Data set]. Kaggle. <https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis>
12. Kaggle. (2018). *Telco Customer Churn.* <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

Appendix A: GitHub Repository

GitHub Repository: https://github.com/emineuysal95/CIND820\_CAPSTONE-PROJECT