

MYSQL CAPSTONE PROJECT

E-Commerce Customer Churn Analysis

Introduction:

Customer churn is a significant concern for businesses in the highly competitive e-commerce industry. Churn refers to the phenomenon where customers stop engaging with a business, leading to reduced revenue and profitability. Understanding the reasons behind customer attrition and proactively addressing them is vital for maintaining a loyal customer base and ensuring sustainable growth.

In this project, "E-Commerce Customer Churn Analysis in MySQL," we aim to address the problem of customer churn by utilizing historical customer data to uncover its root causes. The problem statement highlights the challenge of identifying churn patterns, which can arise due to factors like dissatisfaction with services, lack of engagement, or better offers from competitors. By analyzing key customer attributes such as tenure, purchase behavior, preferred payment methods, and satisfaction levels, the project provides a structured approach to understand and mitigate churn.

The process involves several steps to ensure a comprehensive analysis:

1. Data Cleaning and Preprocessing: Handling missing values, outliers, and inconsistencies in the dataset to prepare it for analysis.
2. Feature Engineering: Creating derived metrics, such as churn status and complaint indicators, to enhance the analysis.
3. Exploratory Data Analysis: Using MySQL queries to investigate patterns and trends in customer behavior and churn.
4. Insights and Recommendations: Identifying actionable insights that help businesses design targeted retention strategies.
5. Reporting: Developing visual dashboards and detailed reports for clear communication of findings.

This systematic approach ensures the identification of churn patterns and provides e-commerce businesses with the tools needed to proactively address customer attrition, foster loyalty, and maintain competitiveness in a dynamic market.

Project Title: E-Commerce Customer Churn Analysis

Problem Statement:

In the realm of e-commerce, businesses face the challenge of understanding customer churn patterns to ensure customer satisfaction and sustained profitability. This project aims to delve into the dynamics of customer churn within an e-commerce domain, utilizing historical transactional data to uncover underlying patterns and drivers of churn. By analyzing customer attributes such as tenure, preferred payment modes, satisfaction scores, and purchase behavior, the project seeks to investigate and understand the dynamics of customer attrition and their propensity to churn. The ultimate objective is to equip e-commerce enterprises with actionable insights to implement targeted retention strategies and mitigate churn, thereby fostering long-term customer relationships and ensuring business viability in a competitive landscape.

Objective:

The primary objective of this project is to analyze customer churn patterns using MySQL to:

- Identify key factors influencing customer churn and their impact on business outcomes.
- Explore the correlation between customer satisfaction metrics and churn rates.
- Develop data-driven strategies to enhance customer loyalty and reduce attrition.
- Provide actionable insights for segmenting customers based on churn propensity.
- Evaluate the effectiveness of different payment modes and order categories in retaining customers.
- Analyze customer tenure and its relationship to purchase behavior and retention.
- Enable e-commerce businesses to proactively address churn through predictive modeling and targeted interventions.
- Create visual dashboards and reports for intuitive understanding and strategic decision-making.

Key Features:

- Comprehensive data cleaning and preprocessing to handle missing values, outliers, and inconsistencies.
- Standardized and transformed dataset for accurate analysis and reporting.
- Use of MySQL for efficient data storage, querying, and manipulation.
- Creation of derived metrics and indicators such as churn status and complaint status.
- Segmentation of customers based on behavior, satisfaction, and purchase patterns.
- Detailed churn analysis to identify high-risk customer groups.
- Insights into the effectiveness of payment modes, order categories, and other attributes on churn.
- User-friendly visualizations and reports to support strategic decision-making.
- Emphasis on actionable recommendations to reduce churn and boost customer retention.

Methodology:

The methodology adopted for this project is designed to systematically analyze customer churn using MySQL as the primary tool for data processing and analysis. The process includes the following key steps:

1. Data Acquisition:

- The dataset, containing customer details and transactional records, is imported into a MySQL database for analysis.

2. Data Cleaning:

- Missing values in numerical columns (e.g., WarehouseToHome, HoursSpentOnApp) are imputed with their respective means.
- Missing values in categorical columns (e.g., Tenure, CouponUsed) are imputed with the mode.

- Outliers are identified and removed from relevant columns (e.g., WarehouseToHome values greater than 100).
- Inconsistencies in categorical values (e.g., "Phone" replaced with "Mobile Phone") are resolved to ensure uniformity.

3. Data Transformation:

- Renamed columns for better readability and consistency (e.g., PreferredOrderCat to PreferredOrderCat).
- New derived columns are created:
 - ComplaintReceived based on the Complain column.
 - ChurnStatus based on the Churn column.
- Dropped unnecessary columns (Churn and Complain) to focus on actionable metrics.

4. Exploratory Data Analysis (Eda):

- SQL queries are used to explore relationships and trends in the dataset.
- Key metrics such as churn rates, average tenure, and total cashback for churned customers are calculated.
- Gender distribution, satisfaction scores, and order categories are analyzed for their impact on churn.

5. Insights Generation:

- Patterns and correlations between churn and customer attributes are identified.
- High-risk customer groups are segmented for targeted interventions.

6. Reporting And Visualization:

- The insights are presented through intuitive reports and dashboards.
- Key findings and recommendations are summarized for stakeholders.

By following this structured methodology, the project ensures a thorough analysis of customer churn, providing actionable insights to e-commerce businesses for effective decision-making.

2. Data Cleaning :

Data cleaning is a crucial step to ensure that the dataset is consistent, complete, and ready for analysis. Below are the specific steps taken for handling missing values, outliers, and inconsistencies in the dataset.

2.1. Handling Missing Values And Outliers:

To address missing data and ensure consistency, the following imputation strategies were used:

2.1.1 Imputation Of Missing Values (Mean And Mode)

The **mean** was used to impute missing values for the following columns:

- WarehouseToHome
- HourSpendOnApp
- OrderAmountHikeFromlastYear
- DaySinceLastOrder

The **mode** (most frequent value) was used to fill missing values for the following columns:

- Tenure
- CouponUsed
- OrderCount

Rationale: Imputing the mean for numerical columns helps maintain data consistency, while using the mode for categorical columns ensures the most common value is used to fill gaps.

2.1.2 Outliers Handling

Rows with WarehouseToHome values greater than 100 were deleted, as these were considered outliers.

Rationale: Removing outliers ensures that extreme values do not skew the analysis, especially for columns related to delivery times and distances.

2.2 Dealing With Inconsistencies:

Data inconsistencies were corrected to maintain uniformity across the dataset:

2.2.1 Uniformity in Categorical Columns:

- Replaced occurrences of "**Phone**" in the PreferredLoginDevice column with "**Mobile Phone**".
- Replaced occurrences of "**Mobile**" in the PreferredOrderCat column with "**Mobile Phone**".

Rationale: Standardizing these terms ensures consistency and prevents confusion in downstream analysis.

2.2.2 Standardization of Payment Mode Values:

- Replaced "**COD**" with "**Cash on Delivery**" in the PreferredPaymentMode column.
- Replaced "**CC**" with "**Credit Card**" in the PreferredPaymentMode column.

Rationale: Consistent labeling of payment methods helps with accurate analysis, particularly when grouping customers based on their preferred payment modes.

3. Data Transformation

Data transformation steps were applied to refine the dataset for analysis, ensuring it is more meaningful and aligned with the project's objectives. The key transformations include renaming columns for clarity, creating new calculated columns to better represent key information, and removing unnecessary columns.

3.1. Column Renaming

To enhance clarity and ensure consistency, the following columns were renamed:

1. "**PreferedOrderCat**" to "**PreferredOrderCat**":
 - The column "**PreferedOrderCat**" was renamed to "**PreferredOrderCat**" to correct the spelling error and standardize the naming convention across the dataset.
2. "**HourSpendOnApp**" to "**HoursSpentOnApp**":
 - The column "**HourSpendOnApp**" was renamed to "**HoursSpentOnApp**" to better reflect the data it represents (i.e., the number of hours spent on the app) and ensure consistency in naming.

3.2. Creating New Columns

To enhance the analytical capabilities of the dataset, the following new columns were added:

1. **ComplaintReceived:**

A new column "**ComplaintReceived**" was created, which assigns the value:

- "**Yes**" if the corresponding value in the "**Complain**" column is 1 (indicating that a complaint was received).
- "**No**" otherwise.

Purpose: This new column enables us to easily analyze customer complaints and identify their relationship with churn.

SQL Query:

```
UPDATE customer_churn  
  
SET ComplaintReceived = IF(Complain = 1, 'Yes', 'No');
```

2. ChurnStatus:

- A new column "**ChurnStatus**" was created to flag whether a customer has churned. The values are:
 - "**Churned**" if the corresponding value in the "**Churn**" column is 1.
 - "**Active**" otherwise.

Purpose: This new column simplifies the analysis of churn patterns and allows for easy categorization of customers into "Churned" and "Active" groups.

SQL Query:

```
UPDATE customer_churn  
  
SET ChurnStatus = IF(Churn = 1, 'Churned', 'Active');
```

3.3. Column Dropping

To streamline the dataset and remove redundant information, the following columns were dropped:

1. Churn and Complain:

- These columns were dropped from the table because the newly created columns "**ChurnStatus**" and "**ComplaintReceived**" now represent the same information in a more user-friendly and efficient format.

SQL Query:

```
ALTER TABLE customer_churn  
  
DROP COLUMN Churn,  
  
DROP COLUMN Complain;
```

4. Data Exploration And Analysis :

The data exploration phase is critical for uncovering trends and patterns within the dataset. This section provides insights into customer behavior, churn patterns, and various other metrics that are essential for understanding the factors influencing customer retention. Below are the key analysis queries and their results.

4.1. Churn and Customer Activity Analysis

1. Count of Churned and Active Customers:

- The number of churned and active customers was determined based on the **ChurnStatus** column.

- **SQL Query:**

```
SELECT ChurnStatus, COUNT(*) Count FROM customer_churn  
GROUP BY ChurnStatus;
```

- **Result:**

	ChurnStatus	Count
▶	Churned	948
	Active	4680

2. Average Tenure and Total Cashback for Churned Customers:

- The average tenure and total cashback for customers who have churned were calculated.

- **SQL Query:**

```
SELECT ROUND(AVG(tenure)) Avg_tenure ,  
SUM(CashbackAmount) TotalCashbackAmount  
FROM customer_churn WHERE ChurnStatus = 'Churned';
```

- **Result:**

	Avg_tenure	TotalCashbackAmount
▶	3	152030

3. Percentage of Churned Customers Who Complained:

- The percentage of churned customers who have complained was determined by analyzing the **ComplaintReceived** column.

- **SQL Query:**

```
SELECT CONCAT(ROUND(SUM(IF(ComplaintReceived = 'Yes', 1,0))/  
COUNT(*)*100,2),'%') AS ComplainedPercentage FROM customer_churn  
WHERE ChurnStatus = 'Churned';
```

- **Result:**

	ComplainedPercentage
▶	53.59%

4. Gender Distribution of Customers Who Complained:

- The gender distribution of customers who have lodged complaints was retrieved.

- **SQL Query:**

```
SELECT Gender, COUNT(*) AS Count FROM customer_churn  
WHERE ComplaintReceived = 'Yes' GROUP BY Gender;
```

- **Result:**

	Gender	Count
▶	Female	690
	Male	914

4.2. Customer Segmentation and Behavior Analysis

5. City Tier with Highest Number of Churned Customers (Preferred Order Category: Laptop & Accessory):

- Identified the city tier with the highest number of churned customers who prefer laptops and accessories.

- **SQL Query:**

```
SELECT CityTier, COUNT(*) ChurnedCustomers FROM customer_churn  
WHERE ChurnStatus = 'Churned' AND PreferredOrderCat = 'Laptop &  
Accessory'  
GROUP BY CityTier ORDER BY ChurnedCustomers DESC LIMIT 1;
```


- **Result:**

	CityTier	ChurnedCustomers
▶	3	150

6. Most Preferred Payment Mode Among Active Customers:

- The preferred payment mode among active customers was analyzed.
- **SQL Query:**

```
SELECT PreferredPaymentMode , COUNT(*) ActiveCustomers
FROM customer_churn WHERE ChurnStatus = 'Active'
GROUP BY PreferredPaymentMode
ORDER BY ActiveCustomers DESC LIMIT 1;
```

- **Result:**

	PreferredPaymentMode	ActiveCustomers
▶	Debit Card	1956

7. Total Order Amount Hike from Last Year for Customers Who Are Single and Prefer Mobile Phones:

- Calculated the total order amount hike from last year for customers who are single and prefer mobile phones.

- **SQL Query:**

```
SELECT SUM(OrderAmountHikeFromlastYear) TotalHike
FROM customer_churn
WHERE MaritalStatus ='Single' AND PreferredOrderCat = 'Mobile Phone' ;
```

- **Result:**

	TotalHike
▶	12177

8. Average Number of Devices Registered Among Customers Who Used UPI:

- The average number of devices registered by customers who used UPI as their preferred payment method was determined.

- **SQL Query:**

```
SELECT FLOOR(AVG(NumberOfDeviceRegistered)) AvgDevices
FROM customer_churn
WHERE PreferredPaymentMode = 'UPI';
```

- **Result:**

	AvgDevices
▶	3

9. City Tier with the Highest Number of Customers:

- The city tier with the highest number of customers was identified.

- **SQL Query:**

```
SELECT CityTier , COUNT(*) CustomerCount
FROM customer_churn
GROUP BY CityTier
ORDER BY CustomerCount DESC
LIMIT 1;
```

- **Result:**

	CityTier	CustomerCount
▶	1	3666

10. Gender with the Highest Number of Coupons Used:

- The gender that utilized the highest number of coupons was found.

- **SQL Query:**

```
SELECT Gender, SUM(CouponUsed) TotalCoupons
FROM customer_churn
GROUP BY Gender
ORDER BY TotalCoupons DESC
LIMIT 1;
```

- **Result:**

	Gender	TotalCoupons
▶	Male	5629

4.3. Customer Purchase and Satisfaction Analysis

11. Number of Customers and Maximum Hours Spent on the App in Each Preferred Order Category:

- A breakdown of the number of customers and maximum hours spent on the app in each preferred order category was provided.

- **SQL Query:**

```
SELECT PreferredOrderCat, COUNT(*) CustomerCount,  
MAX(HoursSpendOnApp) maximum_hours_spent  
FROM customer_churn  
GROUP BY PreferredOrderCat ;
```

- **Result:**

	PreferredOrderCat	CustomerCount	maximum_hours_spent
▶	Laptop & Accessory	2050	5
	Mobile Phone	2078	5
	Others	264	4
	Fashion	826	5
	Grocery	410	4

12. Total Order Count for Customers Who Prefer Credit Cards and Have the Maximum Satisfaction Score:

- The total order count was calculated for customers who prefer credit cards and have the highest satisfaction score.

- **SQL Query:**

```
SELECT SUM(OrderCount) TotalOrders  
FROM customer_churn  
WHERE PreferredPaymentMode = 'Credit Card' AND SatisfactionScore =  
(SELECT MAX(SatisfactionScore) FROM customer_churn) ;
```

- **Result:**

	TotalOrders
▶	1122

13. Number of Customers Who Spent Only One Hour on the App and Had More Than 5 Days Since Their Last Order:

- The number of customers who spent only one hour on the app and had more than 5 days since their last order was retrieved.

- **SQL Query:**

```
SELECT COUNT(*) CustomerCount
FROM customer_churn
WHERE HoursSpendOnApp = 1 AND DaySinceLastOrder > 5;
```

- **Result:**

	CustomerCount
▶	8

14. Average Satisfaction Score of Customers Who Have Complained:

- The average satisfaction score of customers who have complained was calculated.

- **SQL Query:**

```
SELECT FLOOR(AVG(SatisfactionScore)) CustomerCount
FROM customer_churn
WHERE ComplaintReceived = 'Yes';
```

- **Result:**

	CustomerCount
▶	2

15. Preferred Order Category Among Customers Who Used More Than 5 Coupons:

- The preferred order category for customers who used more than 5 coupons was determined.

- **SQL Query:**

```
SELECT PreferredOrderCat , COUNT(*) CustomerCount
FROM customer_churn
WHERE CouponUsed > 5 GROUP BY PreferredOrderCat
ORDER BY CustomerCount DESC;
```

- **Result:**

	PreferredOrderCat	CustomerCount
▶	Laptop & Accessory	99
	Fashion	89
	Mobile Phone	45
	Grocery	42
	Others	28

16. Top 3 Preferred Order Categories with the Highest Average Cashback:

- The top 3 preferred order categories with the highest average cashback amount were identified.

- **SQL Query:**

```
SELECT PreferredOrderCat ,  
FLOOR(AVG(CashbackAmount)) AvgCashback  
FROM customer_churn  
GROUP BY PreferredOrderCat  
ORDER BY AvgCashback DESC LIMIT 3 ;
```

- **Result:**

	PreferredOrderCat	AvgCashback
▶	Others	304
	Grocery	266
	Fashion	210

17. Preferred Payment Modes of Customers with Average Tenure of 10 Months and More Than 500 Orders:

- Identified the preferred payment modes of customers who have an average tenure of 10 months and have placed more than 500 orders.

- **SQL Query:**

```
SELECT PreferredPaymentMode, COUNT(*) CustomerCount  
FROM customer_churn  
WHERE Tenure = 10 AND OrderCount >  
((SELECT SUM(OrderCount) FROM customer_churn) > 500)  
GROUP BY PreferredPaymentMode  
ORDER BY CustomerCount DESC ;
```

- **Result:**

	PreferredPaymentMode	CustomerCount
▶	Debit Card	72
	Credit Card	43
	E wallet	16
	UPI	14
	Cash on Delivery	14

18. Categorization of Customers Based on Distance from Warehouse:

- Customers were categorized based on their distance from the warehouse into the following groups:
 - Very Close Distance:** ≤ 5 km
 - Close Distance:** ≤ 10 km
 - Moderate Distance:** ≤ 15 km
 - Far Distance:** > 15 km
- The churn status breakdown for each category was displayed.
- SQL Query:**

```
ALTER TABLE customer_churn
ADD COLUMN DistanceCategory VARCHAR(20);
UPDATE customer_churn
SET DistanceCategory =
CASE
    WHEN WarehouseToHome <= 5 THEN 'Very Close Distance'
    WHEN WarehouseToHome <= 10 THEN 'Close Distance'
    WHEN WarehouseToHome <= 15 THEN 'Moderate Distance'
    ELSE 'Far Distance'
SELECT DistanceCategory, ChurnStatus, COUNT(*) AS Count
FROM customer_churn
GROUP BY DistanceCategory, ChurnStatus;
```

- Result:**

	DistanceCategory	ChurnStatus	Count
►	Close Distance	Churned	265
	Far Distance	Churned	498
	Moderate Distance	Churned	184
	Close Distance	Active	1696
	Moderate Distance	Active	1106
	Far Distance	Active	1871
	Very Close Distance	Active	7
	Very Close Distance	Churned	1

4.4. Customer Demographics and Purchase Behavior

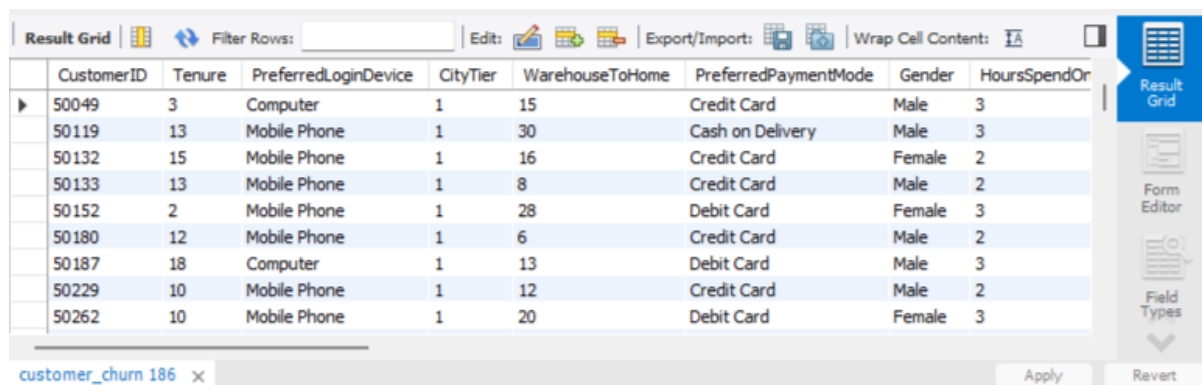
19. Customer Order Details Who Are Married, Live in City Tier-1, and Have More Than the Average Number of Orders:

- Retrieved the order details of customers who are married, live in City Tier-1, and have placed more orders than the average customer.

- SQL Query:**

```
SELECT * FROM customer_churn
WHERE MaritalStatus = 'Married' AND
CityTier =1 AND OrderCount >
(SELECT AVG(OrderCount) FROM customer_churn );
```

- Result:**



	CustomerID	Tenure	PreferredLoginDevice	CityTier	WarehouseToHome	PreferredPaymentMode	Gender	HoursSpendOn
▶	50049	3	Computer	1	15	Credit Card	Male	3
	50119	13	Mobile Phone	1	30	Cash on Delivery	Male	3
	50132	15	Mobile Phone	1	16	Credit Card	Female	2
	50133	13	Mobile Phone	1	8	Credit Card	Male	2
	50152	2	Mobile Phone	1	28	Debit Card	Female	3
	50180	12	Mobile Phone	1	6	Credit Card	Male	2
	50187	18	Computer	1	13	Debit Card	Male	3
	50229	10	Mobile Phone	1	12	Credit Card	Male	2
	50262	10	Mobile Phone	1	20	Debit Card	Female	3

4.5. Customer Returns Analysis

20. a) Customer Returns Table Creation and Data Insertion:

- A **customer_returns** table was created in the 'ecomm' database and the following data was inserted:

ReturnID	CustomerID	ReturnDate	RefundAmount
1001	50022	2023-01-01	2130
1002	50316	2023-01-23	2000
1003	51099	2023-02-14	2290
1004	52321	2023-03-08	2510
1005	52928	2023-03-20	3000
1006	53749	2023-04-17	1740
1007	54206	2023-04-21	3250
1008	54838	2023-04-30	1990

- **SQL Query:**

```
CREATE TABLE customer_returns (  
  ReturnID INT PRIMARY KEY,  
  CustomerID INT,  
  ReturnDate DATE ,  
  RefundAmount DECIMAL(10, 2));  
INSERT INTO customer_returns  
VALUES  
  
  (1001, 50022, '2023-01-01', 2130),  
  (1002, 50316, '2023-01-23', 2000),  
  (1003, 51099, '2023-02-14', 2290),  
  (1004, 52321, '2023-03-08', 2510),  
  (1005, 52928, '2023-03-20', 3000),  
  (1006, 53749, '2023-04-17', 1740),  
  (1007, 54206, '2023-04-21', 3250),  
  (1008, 54838, '2023-04-30', 1990);  
  
SELECT * FROM customer_returns;
```

- **Result:**

	ReturnID	CustomerID	ReturnDate	RefundAmount
▶	1001	50022	2023-01-01	2130.00
	1002	50316	2023-01-23	2000.00
	1003	51099	2023-02-14	2290.00
	1004	52321	2023-03-08	2510.00
	1005	52928	2023-03-20	3000.00
	1006	53749	2023-04-17	1740.00
	1007	54206	2023-04-21	3250.00
	1008	54838	2023-04-30	1990.00
✱	NULL	NULL	NULL	NULL

b) Display Return Details for Churned Customers Who Complained:

- Retrieved the return details along with the customer details for customers who have churned and made complaints.

- **SQL Query:**

```
SELECT cr.*, cc.* FROM customer_returns AS cr
JOIN customer_churn AS cc ON cr.CustomerID = cc.CustomerID
WHERE cc.ChurnStatus = 'Churned' AND cc.ComplaintReceived = 'Yes';
```

- **Result:**

	ReturnID	CustomerID	ReturnDate	RefundAmount	CustomerID	Tenure	PreferredLoginDevice	CityTier	WarehouseToHon
▶	1002	50316	2023-01-23	2000.00	50316	0	Computer	2	29
	1004	52321	2023-03-08	2510.00	52321	18	Mobile Phone	3	19
	1006	53749	2023-04-17	1740.00	53749	1	Mobile Phone	3	31

Conclusion:

This analysis highlights the key factors that contribute to customer churn within the e-commerce platform. By focusing on improving customer support, addressing complaints promptly, and offering personalized retention programs, the business can reduce churn rates, improve customer satisfaction, and increase long-term profitability.

Moreover, understanding customer demographics, such as gender distribution among complaints, provides valuable opportunities to design targeted strategies that resonate with different customer segments. The insights on average tenure and cashback received by churned customers highlight the financial impact of losing long-term, high-value customers, reinforcing the importance of loyalty programs.

Effective implementation of these strategies will contribute to creating a loyal customer base and securing sustainable growth in a competitive market.