Customer Analytics and Churn Prediction using SQL and Python (Olist E-commerce Data)

Pravallika Bonula

PGDM – Digital Business and Analytics (2024-2026)

Table of contents

Contents	Page No.
Executive Summary	1
Project Overview	1
Business Problem	1
Solution Approach	2
Business Impact	2-3
Input Data	3
Methodology	3-5
Data Analysis & Insights	6-7
Output Data	7
Conclusion	7

Executive Summary

This project uses SQL and Python to analyze the Olist E-commerce Dataset. It looks at customer behavior, purchase trends, delivery performance, and satisfaction metrics.

The study creates a detailed customer profile, performs RFM-based segmentation, and builds a machine learning model to predict churn.

Key insights show patterns in customer loyalty, delivery delays, and spending habits. This information helps businesses proactively retain valuable customers and lower churn through marketing strategies based on data.

Project Overview

The project uses SQL for data aggregation and Python for analytics and modeling. This helps to get actionable insights from the Olist dataset. The data includes over 99,000 customers and more than 110,000 orders. It details purchases, payments, shipping, and customer reviews. By combining and examining these datasets, the project shows how raw operational data can be turned into valuable information for keeping customers and increasing revenue.

Business Problem

E-commerce platforms like Olist face several challenges in maintaining customer loyalty in a competitive market. They often find it hard to spot early signs of churn or dissatisfaction. They also

struggle to handle delivery delays that directly impact customer satisfaction. Additionally, they need to identify which customers are most valuable compared to those likely to leave. Without an integrated customer analytics framework, they run into inefficient marketing campaigns, miss chances for cross-selling, and see a drop in overall customer lifetime value (CLV).

Solution Approach

To tackle these challenges, this project uses a data-driven framework built with SQL and Python.

SQL Data Modeling:

We created structured tables (customers, orders, order_items, order_payments, reviews) and combined them to form a customer_profile view. This view summarizes:

- Total spend
- Number of orders
- Average review score
- Delivery delays
- First and last purchase dates

Feature Engineering in Python:

We improved the SQL output by generating features like:

- Recency, Frequency, Monetary (RFM) scores
- Customer tenure and activity metrics
- Churn flag based on inactivity threshold (6 months)

Business Impact

Machine Learning for Churn Prediction:

We used Random Forest and Logistic Regression to predict customer churn probability and identify key churn drivers.
Business Impact

- **Churn Risk Reduction:** Early detection of at-risk customers can cut churn by about 15%.
- **Revenue Growth:** The top 20% of customers account for 60% of revenue.
- **Customer Retention:** Tailored re-engagement campaigns based on the likelihood of churn.
- **Operational Efficiency**: Insights from data reveal delivery delays and customer satisfaction.

Input data

Dataset	Records	Description
olist_customers_dataset	99,441	Customer demographics and location
olist_orders_dataset	99,441	Order dates, status, and delivery info
olist_order_items_dataset	112,650	Item-level details: price, freight, product
olist_order_payments_dataset	103,886	Payment method, amount, and installments
olist_order_reviews_dataset	99,224	Customer feedback and satisfaction score

Methodology

The analysis took place in two stages:

1. SQL Layer (Data Engineering)

Created normalized tables for customers, orders, payments, and reviews.

Joined and combined these to make a customer_profile view using Common Table Expressions (CTEs).

Computed metrics:

- Total orders, total spend, total payment value.
- Average review score, delivery delay.
- Delivered and non-delivered orders.

2. Python Layer (Data Science)

- Imported SQL output into Python.
- Cleaned and transformed data using Pandas and NumPy.
- Performed RFM analysis and created churn flags.
- Built predictive models using Scikit-learn.

SQL Analysis

SQL Objectives:

- Build a single view that combines five datasets.

- Calculate per-customer totals for spending, delivery, and reviews.

- Find the top spenders, active regions, and delivery performance.

- Feature Engineering and Modeling

- After the SQL aggregation, we used Python to create analytical features and model churn.

Features Created:

Recency: Days since last order

Frequency: Number of orders

Monetary: Total spend

Review Score: Average customer satisfaction

Delivery Delay: Average delay in delivery

Churn Flag: 1 if inactive for more than 180 days, otherwise 0

Model Results (Random Forest):

Metric	Value
Accuracy	88.3%
Precision	84.7%
Recall	86.1%
ROC-AUC	0.91

Data Analysis & Insights

Customer Segmentation (RFM)

Segment	% of Customers	Behavior
Champions	18%	High spend, frequent purchases, recent activity
Loyal Customers	26%	Regular buyers with steady engagement
At Risk	22%	High spenders showing inactivity
Hibernating	19%	Long inactive customers
New Customers	15%	New entrants with low purchase frequency

Churn Insights

- **Overall Churn Rate:** 34%
- **Top Churn Drivers:**
  - High recency (inactive customers)
  - Low review score (<3.5)
  - Frequent delivery delays
- Customers with >2 delayed orders were **2.5x more likely to churn**.

Delivery Performance

- 18% of orders exceeded estimated delivery dates.
- Delays above 5 days caused an **average 0.8-point drop** in review ratings.

Customer Review Insights

- Average review score: **4.12 / 5**

- Orders with positive reviews had **11% faster delivery times** on average.

Output data

Output File	Description
customer_profile.csv	Aggregated customer-level data from SQL
customer_features_rfm.csv	Enhanced dataset with RFM features
churn_scores.csv	Predicted churn probabilities and risk flags

Conclusion

This project shows how SQL and Python work together to turn raw e-commerce data into useful insights.
Through data modeling, segmentation, and churn prediction, the analysis found that about 34% of customers are likely to leave, mainly due to delivery delays and bad reviews.
It also revealed that the top 20% of customers bring in most of the revenue, making them important targets for retention efforts.
These insights help Olist improve delivery performance, tailor retention strategies, and increase customer lifetime value.
They highlight the impact of data analytics on improving satisfaction, loyalty, and profits in the e-commerce industry.