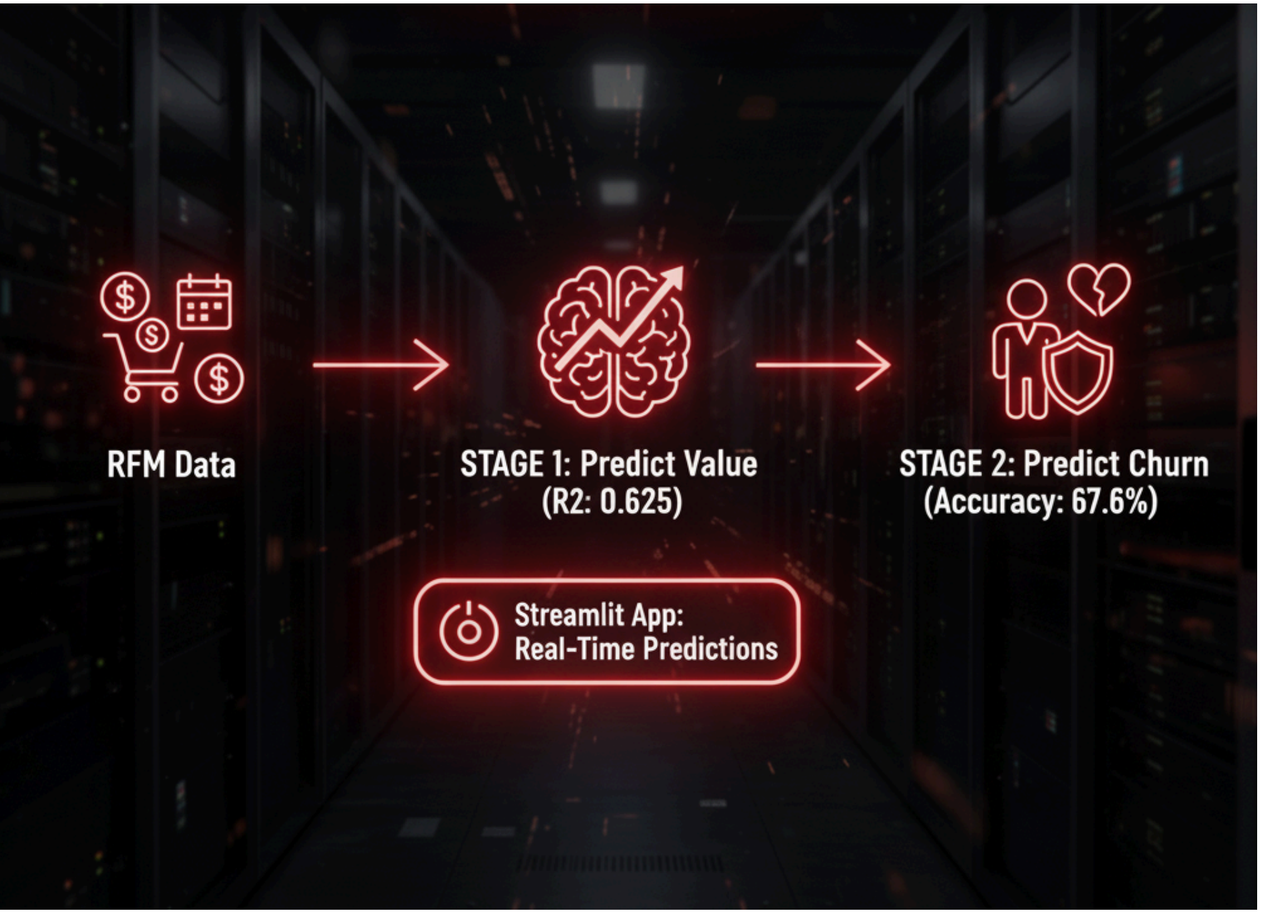
PROJECT REPORT

PRESENTATION

Introduction of Topic

Project tile: VIP_Churn_Predictor

Customer retention is a critical challenge in e-commerce, but not all customers offer the same value. The loss of a "VIP" customer—one with high frequency and monetary value—is significantly more damaging than the loss of an average user. The core problem is that standard churn models often fail to distinguish between these customer tiers, treating all churning users equally. This project tackles that challenge by building an intelligent, two-stage machine learning pipeline. First, a regression model analyzes a customer's purchasing history (Recency, Frequency, and Monetary value) to predict their future financial worth. This "predicted value" is then used as a key feature in a second classification model, which ultimately determines the actual churn risk. The result is a deployed web application that doesn't just ask "Will this customer churn?" but answers the more important business question: "Is this valuable customer at risk of churning?"

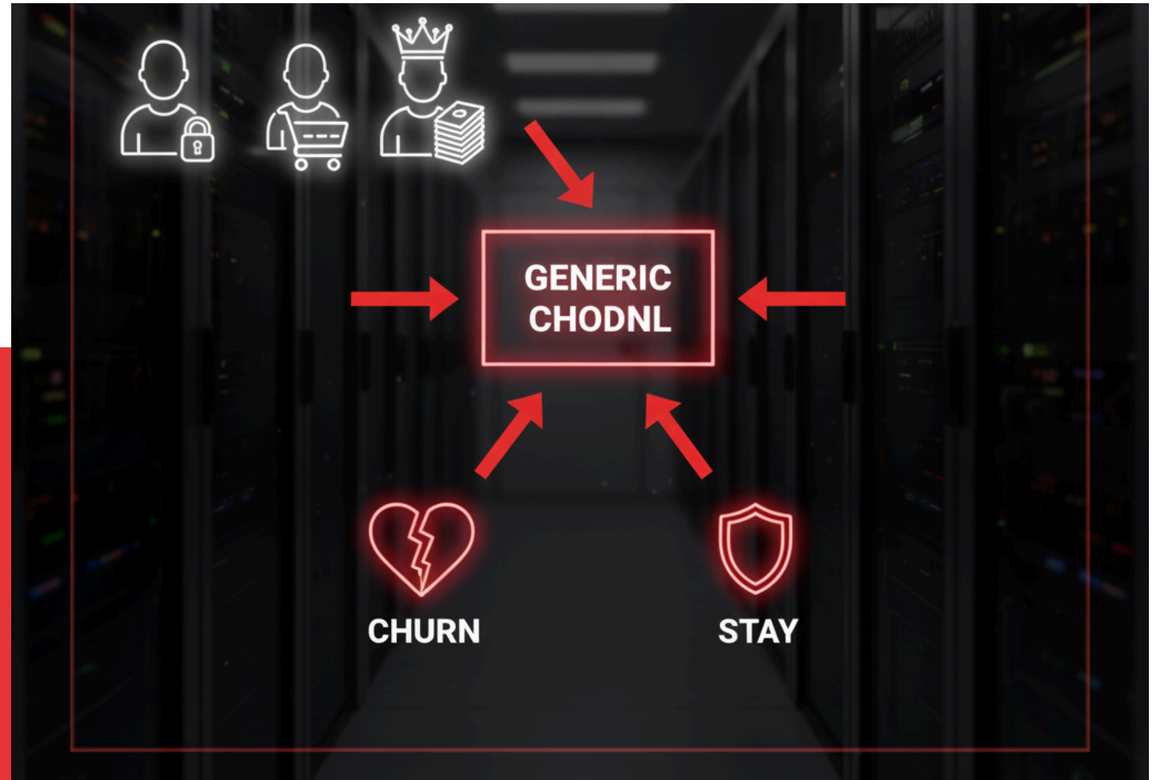


Abstract

Losing a customer is bad, but losing a "VIP" customer is a disaster. This project builds a smart, two-stage system to find these high-value customers before they leave. Using the UCI Online Retail dataset, a regression model first predicts a customer's future financial value. This "value score" is then fed as a new, intelligent feature into a classification model, which predicts their churn risk with 67.6% accuracy. This project proves the linked-model approach and delivers the final result as a live Streamlit web app, turning a complex analysis into a simple, real-time decision tool.

Problem

In e-commerce, losing a customer is bad, but losing a "VIP" is a financial disaster. The core problem is that businesses struggle to identify which of their most valuable, high-spending customers are at the highest risk of leaving. They need a way to focus their limited retention budget on the customers who matter most.



Standard churn models are too simple. They flag a one-time buyer and a long-time VIP with the same "churn risk" alert, which is inefficient. This project builds a smarter, two-stage system. It first uses a regression model to assign a "financial value score" to every customer. This score is then fed into a classification model, allowing it to finally answer the real business question: "Which of our most valuable customers are about to leave?"

Objectives

The objective of this project is to build and deploy an end-to-end machine learning pipeline that intelligently identifies high-value "VIP" customers at risk of churning. This involves processing the UCI Online Retail dataset to engineer Recency, Frequency, and Monetary (RFM) features. A two-stage modeling approach is then executed: first, a regression model is trained to predict a customer's future financial value, creating a "VIP score." Second, this "VIP score" is fed as a novel, intelligent feature into a classification model to predict churn likelihood. The final, validated pipeline is then packaged and deployed as an interactive Streamlit web application to provide real-time, actionable risk assessments for any given customer.

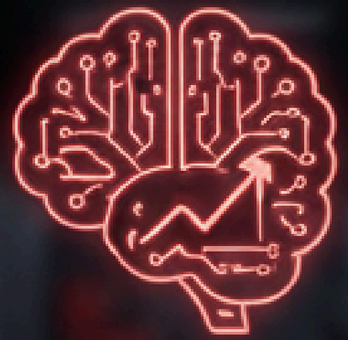
DATA SCIENCE

Python

Pandas

Nupmp

MACHINE LEARNING



Scikit-learn

Scikit-learn

Joblib

Linear Regrestion

Logistic Regarstion

DEPLOYMENT

Colab

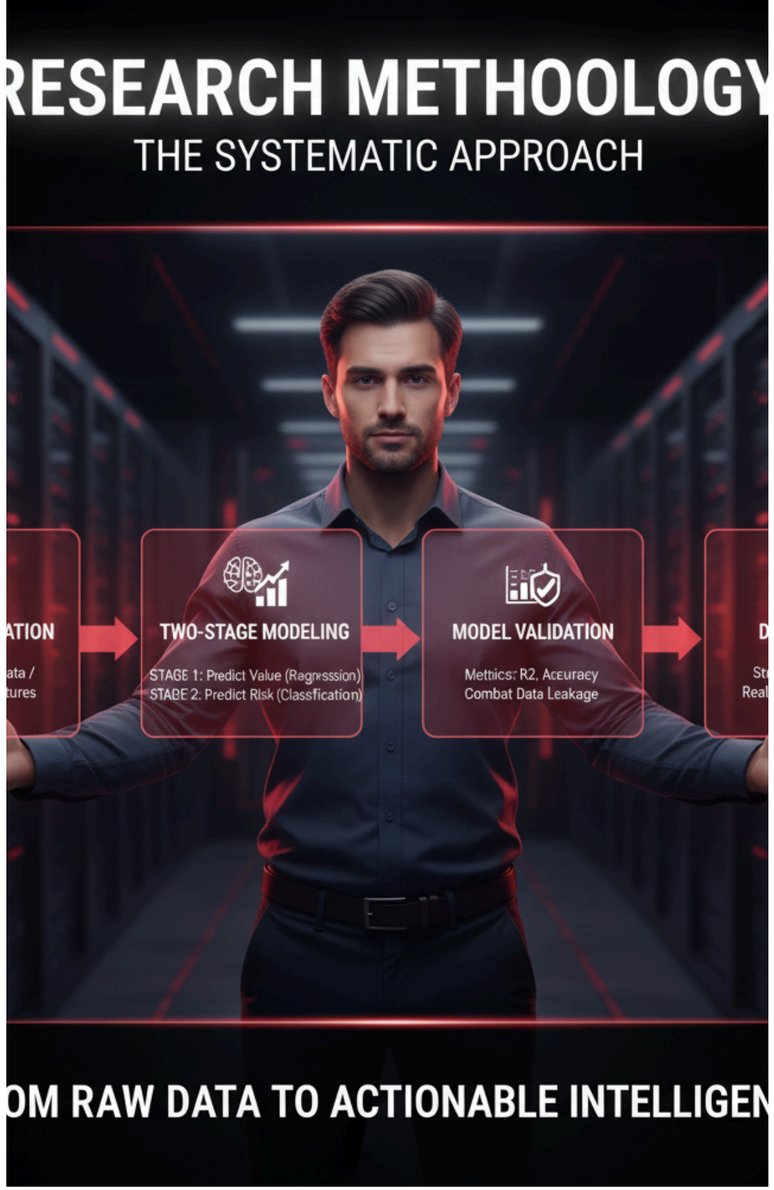
Google Colab

Streamit
Web App

Virtual
Environment
(venv)

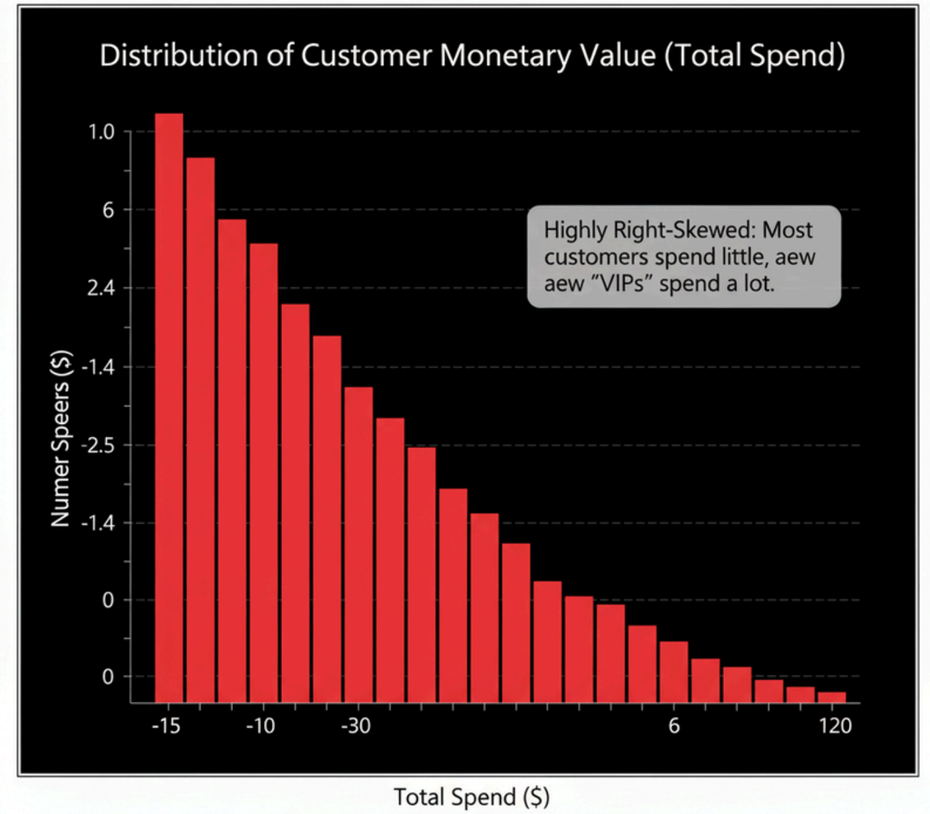
Research Methodology

- Used the public UCI Online Retail Dataset — cleaned by removing missing CustomerIDs and cancelling negative-quantity transactions.
- Created a customer-level dataset by computing Recency, Frequency and Monetary (RFM) features relative to a fixed snapshot date.
- Built a two-stage modelling pipeline: first a polynomial regression on RFM to predict a “Predicted_CLV” score; then used that score as an additional feature for a logistic regression classifier predicting is_churned.
- Evaluated the regression with R^2 = 0.625, and the classifier with Accuracy = 67.6% plus a confusion matrix, verifying no data leakage.
- Serialized the entire pipeline (scalers, transformers, models) with joblib and deployed it as an interactive Streamlit web application for real-time prediction.



Data Analysis

- Cleaned the dataset by dropping missing CustomerIDs and removing negative-quantity (returns/cancellations).
- Engineered Recency, Frequency, and Monetary (RFM) features at the customer level.
- Defined two targets: CLV_3_Month (future spend) and is_churned (binary churn flag).
- Found strong right-skew in RFM features → applied feature scaling with StandardScaler.



Conclusion

This project successfully built and deployed a two-stage "VIP" churn predictor. By first using a regression model to quantify a customer's potential value (R2: 0.625), and then feeding that value as a feature into a classification model, we achieved a realistic baseline accuracy of 67.6%. The entire pipeline was packaged into an interactive Streamlit web app, effectively turning a complex data analysis into a simple, real-time tool for making smarter business decisions.

References:

1. The primary data source for this project was the "Online Retail Data Set" from the UCI Machine Learning Repository. This dataset, originally donated by Dr. Daqing Chen, contains over 540,000 transnational records from a UK-based online retailer between 2010 and 2011. It is a widely cited and standard benchmark for tasks involving customer segmentation, cohort analysis, and predictive modeling based on RFM (Recency, Frequency, Monetary) principles.
2. Citation: Chen, D., Sain, S.L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.