Credit Card Customer Churn Prediction and Retention Strategy

Instructor: Jing Yi
Student: Caroline Kuo; Kailing Li
Date: November 2, 2025

1.Field or topic of interest

This project focuses on predictive modeling and customer analytics in the field of financial economics. It applies machine learning techniques to identify which credit card customers are most likely to close their accounts. By combining methods such as logistic regression, ensemble models, and clustering, the project connects the concepts learned in AAE 722, including data preprocessing, model selection, cross-validation, and model interpretability, to a real-world business problem. This topic also aligns with data analytics career paths and demonstrates how applied machine learning can generate actionable insights for customer retention and business decision-making.

2. Data Source and Structure

The dataset used is the publicly available Credit Card Customers Dataset from Kaggle ("BankChurners.csv", created by Sakshi Goyal, 2020).
Dataset link: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers/data
It contains approximately 10,000 observations, each representing an individual credit card customer from a U.S. bank. The unit of analysis is the individual customer. The dataset is moderately imbalanced, with about 16% of customers labeled as "Attrited." This makes it suitable for applying class-imbalance handling and model evaluation techniques that were covered in AAE 722.

Key variables include:

- Target variable: Attrition_Flag (whether the customer is "Existing" or "Attrited").
- Demographics: Customer_Age, Gender, Education_Level, Income_Category, Marital_Status.
- Account behavior: Total_Trans_Ct, Total_Trans_Amt, Months_Inactive_12_mon, Avg_Utilization_Ratio, Credit_Limit, Contacts_Count_12_mon.
- Derived features (to be created): average transaction amount, engagement score, and utilization volatility.

3. Objective of the Project

The main objective is to predict customer churn and identify the key behavioral factors that explain it. The project will follow the end-to-end machine learning pipeline discussed in AAE 722.

- EDA and Data Cleaning: Explore variable distributions, detect class imbalance, and handle missing values.
- Feature Engineering: Create new behavior-based variables such as average transaction amount and engagement score.
- Model Building: Begin with a logistic regression baseline, then compare Random Forest and XGBoost models to improve predictive accuracy.
- Model Evaluation: Assess model performance using ROC-AUC, PR-AUC, and F1-score to

ensure robust results.

- Explainability: Apply SHAP values to interpret which variables have the strongest influence on customer churn.
- Customer Segmentation: Use K-Means clustering to identify high-value but inactive customers and propose data-driven retention strategies.

Expected contribution:
The project will produce a well-performing predictive model, an interpretable explanation of churn factors, and a concise business summary that translates machine learning results into practical retention strategies and potential return on investment (ROI).

References

Kaggle Dataset: Credit Card Customers (BankChurners.csv), Sakshi Goyal (2020).
Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions (SHAP).
AAE 722 Course Lectures, Fall 2025.