# Churn Prediction Project Report

## Overview

This project aimed to identify key drivers of credit card customer churn using advanced analytics and machine learning techniques. Leveraging the IBM BankChurners dataset, we cleaned, engineered, and modelled features to accurately predict attrition risk and support actionable business decisions.

## Top Insights

**Transaction Volume is King**
The number of transactions ( `Total_Trans_Ct` ) emerged as the most powerful indicator of churn across all models. Customers with fewer transactions were significantly more likely to churn, highlighting inactivity as a major red flag.

**Low Activity Signals Risk**
Low activity in general—evidenced by lower `Total_Trans_Amt` , `Avg_Transaction_Value` , and `Avg_Utilization_Ratio` —was strongly associated with churn. This supports the understanding that dormant or low-engagement customers are at high risk of attrition.

**Relationship Depth Helps Retention**
Depth of customer relationship played a protective role. Customers with more products or services ( `Total_Relationship_Count` ) showed higher retention. This indicates that offering a broader service portfolio can improve loyalty.

**Marital Status & Income Matter**
Demographic and socioeconomic factors also proved relevant. Single customers and those with incomes below $40K were more likely to churn. This suggests financial strain or lack of perceived value may contribute to attrition.

**Feature Engineering Paid Off**
Feature engineering substantially enhanced model interpretability and performance. Derived variables such as `Avg_Transaction_Value` helped clarify behavioural patterns and improved predictive accuracy.

**XGBoost Outperformed All Models**
Among all models tested, **XGBoost** performed best. It achieved:

- **Accuracy:** 97%
- **F1-score (Attrited):** 0.89
- **AUC-ROC:** 0.99

## Recommendations

**Prioritise At-Risk Segments**
Churn-prone customers should be proactively identified based on transactional behaviour. Specifically, low and declining transaction counts or amounts should trigger early alerts for intervention.

**Design Retention Offers**
Retention offers should be designed and personalised for dormant or low-activity users, particularly those in lower income brackets or without strong relationship depth. Campaigns targeting single and economically vulnerable users could be more effective in reducing attrition.

**Track Feature Trends**
Feature usage metrics such as transaction count and relationship depth should be closely monitored on a monthly basis. These indicators can serve as operational KPIs for customer success teams.

**Simplify or Consolidate Low Impact Features**
To reduce multicollinearity and simplify future models, highly correlated features should be evaluated for consolidation. For example, `Credit_Limit` and `Avg_Open_To_Buy` show nearly perfect correlation and could be merged into one representative variable.

**Deploy XGBoost Model**
The XGBoost model is recommended for production deployment. It should be embedded within CRM systems to perform real-time churn scoring and support automated engagement strategies.

## Limitations

The dataset is limited to a snapshot of customer behaviour at a single point in time. As such, time series patterns and seasonality effects are not captured.

Some features, such as marital status or education, may not be updated frequently in the bank's system and could lose relevance over time.

Although class imbalance was addressed using SMOTE, real-world class distributions might require ongoing calibration to ensure predictive consistency.

Feature importance from tree-based models reflects relative influence, not causality. Therefore, interventions based on feature impact should be validated through A/B testing or user surveys.

## Justifications for Modelling Choices

XGBoost was selected as the final model based on its high predictive performance and ability to handle non-linear interactions effectively. It is also well-suited for tabular data and has strong community and production support.

Feature engineering decisions were driven by both domain knowledge and statistical analysis, such as the introduction of `Avg_Transaction_Value` and consolidation of highly correlated variables based on correlation heatmaps.

Cross-validation and AUC metrics were used to ensure robustness and generalisability of model performance, reducing the risk of overfitting on a single train-test split.

In [ ]:

```

```