Tata Data Analytics Job Simulation (Forage)

This document compiles the key deliverables from the Tata Data Analytics Virtual Experience Program (Forage). It includes the business summary report, imputation guide, predictive model plan template, and dataset description.

# Dataset Description

The dataset, titled 'Delinquency\_prediction\_dataset.xlsx', contains customer-level financial information used for predictive analysis. It includes attributes such as credit utilization, missed payments, debt-to-income ratio, and demographic details like age. The dataset was the foundation for data cleaning, imputation, exploratory analysis, and predictive modeling.

Key Variables:  
- Credit Utilization (% of available credit used)  
- Payment History (number of missed payments)  
- Debt-to-Income Ratio (DTI)  
- Age (years)  
- Delinquency Status (target variable)

# Business Summary Report

Business Summary Report: Predictive Insights for Collections Strategy

1. Summary of Predictive Insights

The predictive model identified several customer segments at elevated risk of credit card delinquency. Key risk indicators include high credit utilization, missed payments, and elevated debt-to-income ratios. These insights can help prioritize which customers may benefit most from early outreach or financial support strategies.

Key Insights Summary Table:

2. Recommendation Framework

Restated Insight:

Customers under 30 with two or more missed payments have a significantly higher likelihood of delinquency.

Proposed Recommendation:

Launch a 6-week pilot outreach campaign targeting this segment with proactive SMS and email messaging. The goal is to offer tailored payment plans or financial counselling support before accounts reach 30+ days delinquent.

Justification and Business Rationale:

Specific: Focused on a clearly defined, high-risk group.

Measurable: Target a 10–15% reduction in delinquency within the pilot group.

Actionable: Uses existing communication infrastructure.

Relevant: Aligns with Geldium’s goals to reduce credit risk and improve customer outcomes.

Time-bound: Designed as a time-limited pilot with measurable outcomes.

3. Ethical and Responsible AI Considerations

The model was evaluated for fairness using multiple performance metrics across age and income groups. No disproportionate flagging of protected segments was observed.

Bias: The analyst tested for overrepresentation in delinquency predictions and found a balanced outcome across customer demographics.

Explainability: The model uses logistic regression, which allows clear explanation of how key variables influence predictions.

4.Responsible use: The recommendation is focused on early, supportive interventions rather than punitive action, reinforcing fairness and customer care principles.

# Imputation Guide

Understanding Imputation:   
A CHEAT SHEET

What is Imputation?

Imputation is the process of filling in missing data in a dataset using reasonable or estimated values. Instead of deleting rows with missing values, imputation helps retain valuable information by replacing the missing entries with typical or representative data.

What are Mean, Median, and Mode?

Mean (Average)

The mean is calculated by adding all the values in a dataset and dividing by the number of values.  
Example: For values $2,000, $3,000, $4,000, $5,000, $6,000:  
Mean = (2000 + 3000 + 4000 + 5000 + 6000) ÷ 5 = $4,000

Median (Middle Value)

The median is the middle number when the data is ordered from smallest to largest.  
Example: For values $2,000, $3,000, $4,000, $5,000, $6,000:  
Median = $4,000 (the third number in the sorted list)

Mode (Most Frequent)

The mode is the value that appears most frequently in a dataset.  
Example: For values $2,000, $2,000, $3,000, $4,000, $5,000:  
Mode = $2,000 (it appears twice)

How is Imputation Used?

When data is missing, analysts use the mean, median, or mode to fill in those blanks depending on the nature of the data:  
- Use the MEAN if the data is well-balanced without extreme values.  
- Use the MEDIAN if the data contains outliers that could skew the average.  
- Use the MODE if a single value is very common and likely represents others.