

## Analytics Startup Plan

<b>Project</b>	Predicting loan recovery status for borrower
<b>Requestor</b>	Centennial College
<b>Date of Request</b>	14 <sup>th</sup> July 2025
<b>Target Quarter for Delivery</b>	13 <sup>th</sup> August 2025
<b>Epic Link(s)</b>	
<b>Business Impact</b>	This project will enhance loan recovery efficiency by enabling data-driven decisions that reduce costs, prioritize high-yield recovery actions, and improve overall financial performance.

# 1.0 Business Opportunity Brief

## The Ask:

Develop an AI-driven predictive model to accurately forecast the recovery status of delinquent loans (Fully Recovered, Partially Recovered, or Written Off) and recommend cost-effective recovery strategies, enabling our organization to maximize loan recovery rates and minimize costs by 10 % while adhering to Canadian regulatory standards (FCAC/PIPEDA).

## The Opportunity:

By leveraging predictive analytics on loan data (e.g., number of collection attempts, legal actions taken, loan amounts, collateral values), we can increase loan recovery rates and reduce recovery costs. This opportunity allows us to optimize resource allocation, prioritize high-impact recovery actions (e.g., legal notices for home loans, phone calls for personal loans), and improve financial outcomes amid rising delinquency rates.

## The Problem:

Currently, our loan recovery process lacks predictive insights, leading to inefficient resource allocation and inconsistent recovery outcomes. Without a data-driven approach, we struggle to identify which delinquent loans are recoverable, select the most effective recovery methods (e.g., calls, legal notices), and manage costs, particularly for high-value loans (e.g., home loans) or high-delinquency personal loans. This results in missed recovery opportunities, higher costs, and potential non-compliance with Canadian regulations, hindering our ability to meet financial and regulatory objective

# 1.1 Supporting Insights

## Trends and Research Findings:

- **Delinquency and Default Trends:** According to recent reports by Equifax Canada and the Canadian Bankers Association (CBA), consumer debt delinquency rates have risen post-pandemic, especially in unsecured credit lines and personal loans. This highlights the growing importance of effective loan recovery strategies. Canada's private debt-to-GDP ratio is 215.9%, with credit card (8.5%) and auto loan (7.7%) delinquencies rising, especially among Gen Z and Millennials (Q1 2025).
- **Shift Toward Digital Collections:** Leading financial institutions and FinTech's are transitioning to **digital-first collection approaches** using AI and predictive analytics to personalize recovery methods and increase repayment compliance. AI-driven software improves recovery via predictive analytics. Omnichannel communication (SMS, email) reduces costs and boosts engagement.

- **Cost of Recovery:** Traditional recovery methods such as in-person visits and legal action are **costly and often inefficient**. Research by McKinsey & Company shows that digital and data-driven collections can reduce cost-to-collect by up to **10%** while improving customer experience.
- **Regulatory Pressure:** Stricter consumer protection laws, like those enforced by the Financial Consumer Agency of Canada (FCAC), push institutions to ensure fair treatment during collections—strengthening the case for smarter, more personalized recovery methods.

### **Key competitors**

- The Canadian loan recovery industry is increasingly driven by data, digital automation, and consumer behavior modeling. Financial institutions are recognizing the value of predictive analytics to reduce delinquency and optimize collection efforts while maintaining customer trust. Equifax Canada leads in this space with offerings in credit scoring and debt recovery analytics, positioning itself with the message “Know your borrowers better – predict and act with data.” It operates nationally as a primary data provider.
- Borrowell, a growing fintech player, provides free credit monitoring and AI-powered financial tools, aiming to “help Canadians improve their financial health.” It enjoys a strong national presence, particularly in urban markets. Clearbanc (now Clearco) offers revenue-based financing and uses predictive models to track repayments, promoting non-dilutive capital solutions with intelligent recovery mechanisms. With a growing international footprint, Clearco remains a national leader in alternative lending.
- Symcor, known for its outsourced digital collection services, emphasizes compliance and automation, catering largely to enterprise clients across Canada. Traditional banks such as TD and RBC maintain dominant market share through in-house loan recovery solutions, increasingly enhanced by AI and digital engagement tools. They brand their services around “trusted recovery with respect and precision,” leveraging national infrastructure and customer relationships.

These competitors highlight an evolving landscape where the ability to combine digital efficiency with customer sensitivity is key

## **1.2 Project Gains**

The KPIs of a financial company is related to revenue growth, operational efficiency, customer satisfaction and compliance regulation

- **Revenue Gains:**  
By predicting loan recovery status more accurately, the company can focus recovery efforts on loans with high recovery potential, leading to increased recovered amounts and reduced revenue loss from write-offs.
- **Quality Improvements:**  
The predictive model will introduce consistency and objectivity in evaluating delinquent loans. This minimizes bias, removes guesswork, and ensures data-driven recovery decisions that align with performance goals.
- **Cost and Time Savings:** Targeting the right accounts reduces operational overhead by eliminating low-yield collection attempts. It also frees up staff time and call center capacity, enabling leaner and more focused collection campaigns.

### **What We will Do Differently:**

Instead of a one-size-fits-all collection strategy, the business will deploy tailored recovery actions—legal escalation, negotiated settlements, or payment plans—based on predicted recovery status, optimizing both efficiency and customer sensitivity.

### **Why Customers Will Care:**

Customers benefit from fairer, more personalized recovery approaches. High-risk customers may be offered early interventions or flexible repayment options, while low-risk borrowers avoid aggressive tactics—improving customer experience and reputation.

### **Implications of Doing Nothing**

Failing to adopt an AI-driven loan recovery model could have significant negative consequences for your organization, especially in Canada's competitive financial services market:

1. **Revenue Loss:**
  - Without AI, recovery rates remain suboptimal, potentially missing out on 25% additional recoveries (\$25 million for a \$100 million portfolio). Competitors like HES FinTech and Collect are already capitalizing on AI to maximize recoveries.
  - Inability to cross-sell effectively reduces revenue growth, as AI-driven competitors identify high-potential borrowers
2. **Operational Inefficiencies:**
  - Reliance on manual processes increases cost-to-collect by 10% compared to AI-driven methods, straining budgets. For a lender with \$1 million in collection costs, this could mean \$300,000 in avoidable expenses.
  - Slower recovery processes (hours vs. seconds) reduce staff productivity and delay cash flow, as manual document reviews bottleneck operations.

### **3. Competitive Disadvantage:**

- Canada's debt collection software market is the fastest-growing, and competitors like VQN, Collect, and HES FinTech are adopting AI rapidly. Inaction risks losing market share to these players.
- Traditional methods fail to address rising delinquencies (8.5% credit cards, 7.7% auto loans), particularly among younger borrowers, leading to higher write-offs.

### **4. Regulatory and Reputational Risks:**

- Non-compliance with FCAC and PIPEDA regulations due to outdated, non-transparent recovery methods could result in fines and legal challenges
- Aggressive or unfair recovery tactics erode borrower trust, damaging brand reputation, especially among Gen Z and Millennials who value ethical practices.

### **5. Missed Customer Expectations:**

- Borrowers, particularly younger Canadians, expect digital, personalized experiences. Without AI-driven omnichannel communication, satisfaction drops, reducing repayment compliance.
- Lenders miss opportunities to serve underserved segments, limiting financial inclusion and portfolio diversity.

## **2.0 Analytics Objective**

In order to develop a predictive model for loan recovery status, we have identified a set of analytical objectives designed to guide our exploration and modeling efforts. This section outlines the key business questions, underlying assumptions, and testable hypotheses that form the foundation of the analysis.

### **Key Analytical Questions**

- What borrower and loan characteristics are most predictive of loan recovery success?
- Can we accurately predict whether a delinquent loan will be recovered or written off?
- How do collection methods (e.g., legal action, phone contact, letters) affect recovery outcomes?
- What patterns in payment behavior (e.g., number of missed payments, days past due) are indicative of recovery likelihood?
- Can predictive modeling support more cost-effective and targeted recovery strategies?

## Hypotheses

The following hypotheses will be tested through data exploration and machine learning modeling:

Hypothesis	Type	Description
H1	Predictive	Borrowers with higher monthly income and fewer missed payments are more likely to be successfully recovered.
H2	Predictive	Loans secured with high-value collateral are more likely to be recovered.
H3	Associative	Legal action increases the probability of recovery for high-value loans.
H4	Comparative	A higher number of collection attempts correlates with reduced recovery likelihood, indicating diminishing returns.
H5	Exploratory	Recovery outcomes significantly vary depending on the collection method used.

## 2.1 Other related questions and Assumptions:

To conduct this analysis, the following assumptions have been made:

1. **Data Accuracy:** The “Recovery\_Status” field correctly represents the final outcome of each loan’s collection process.
2. **Feature Reliability:** Variables such as Monthly\_Income, Payment\_History, and Collateral\_Value are assumed to be accurate and reflect current borrower conditions.
3. **Operational Consistency:** Collection methods have been consistently applied and documented across the dataset.
4. **Representative Sample:** The dataset used in this analysis is representative of the broader loan portfolio.
5. **Stable External Conditions:** No significant external economic or policy changes have skewed recovery outcomes during the observed period.

## 2.2 Success measures/metrics

Goal	What Success Looks Like	What Makes It Happen	How We Measure Success	Target	How We Check
Recover More Loans	Get more loans paid back (fully or partially) compared to now (e.g., from 60% to 75% of loans).	- Accurate predictions of which loans can be recovered, using data like number of collection attempts and legal actions taken. - Smart recovery plans for high potential risk loans (e.g., legal notices for home loans, phone calls for personal loans).	<b>Recovery Success Rate:</b> Percentage of loans predicted to be paid back that actually get paid.	At least 75%	Compare how many loans we recover with the model vs. before.
			<b>Model Strength:</b> How well the model separates recoverable loans from those that won't be paid.	Strong (precision $\geq 0.85$ )	Use a test score to measure model performance.
			<b>Accuracy of Predicting Full Recovery:</b> How often the model correctly picks loans that are fully paid back.	ROC AUC $>= 0.85$ or better	Check if loans predicted as fully paid are actually fully paid.

## 2.3 Methodology and Approach

### Type of Analysis

The analysis will include a combination of statistical and machine learning techniques: Chi-square test, ANOVA, correlation analysis, Decision Trees, Logistic Regression, Random Forest, and Neural Networks.

The initial approach will involve conducting correlation analysis, Chi-square tests, ANOVA, and logistic regression to identify which variables are most significantly associated with the target variable—**Recovery Status**. These insights will inform effective feature engineering for the modeling phase

### Methodology

We will begin by exploring the dataset through descriptive statistics and data visualization. This will be followed by data cleaning to address issues such as missing values, duplicates, unbalance target and incorrect data types. After preprocessing, feature engineering will be performed to enhance model performance.

The dataset will be split into **training, validation, and test sets** to ensure reliable model training and evaluation. The target variable, *Recovery Status*, contains three outcomes. Therefore, we will build two binary classification models:

- **Model 1:** Recovered (1) vs. Written Off (0)
- **Model 2:** Fully Recovered (1) vs. Partially Recovered (0)

Supervised learning algorithms including Decision Trees, Logistic Regression and Random Forest will be trained on the training dataset to identify key predictors of recovery outcomes. Hyperparameter tuning will be conducted to optimize performance.

Following this, the models will be validated and tested using the respective datasets. Evaluation metrics such as Precision and ROC AUC, and **Interpretability** will be used to select the best-performing model.

### Expected Output

The final deliverable will include a set of data-driven insights and strategic recommendations to better assess borrower profiles, accurately predict their loan recovery status and recommend best strategies for loan collection from high risk loan. These outcomes will inform actionable decisions aimed at maximizing loan recovery while minimizing costs, thereby improving operational efficiency and financial performance.

## 3.0 Population, Variable Selection, considerations

The data set is about borrowers from a bank with their information on past recovery outcomes. it is a table made up of 21 variables and 501 records

Variables	Description
Gender	Gender of the borrower Male (0) or Female (1)
Employment type	Type of employment of the borrower 1-Salaried 2-Self-employed 3-Business Owner
Monthly Income	Salary obtained per month
Num_Dependents	Number of children below 17 years old financially depending
Loan_ID	unique identifier for each loan record
Loan_Amount	principal amount of the loan (in Canadian Dollars, CAD) borrowed by the individual.

Interest_rate	annual interest rate (expressed as a percentage) applied to the loan
Loan_type	type or category of the loan taken by the borrower 1-Auto 2-Business 3-Home 4-Personnal
Collateral_value	monetary value (in Canadian Dollars, CAD) of the asset(s) pledged as security for the loan
Monthly_EMI	The fixed monthly payment (in Canadian Dollars, CAD) that the borrower is required to make to repay the loan, including both principal and interest components
Payment_History	Track record of making loan payments 1-on time 2-Delayed 3-Missed
Num_Missed_payments	Number of times payments were missed
Days_past_due	Number of days a borrower's loan payment is overdue at the time of data collection
Collection_attempts	The number of times the bank has attempted to contact the borrower or take action to recover the outstanding loan amount
Collection_method	The type of method used by the lender to recover the outstanding loan amount from a delinquent borrower. 1-Calls 2-Debt collectors 3-Legal notice 4-Settlement Offer
Legal_Action	Indicating whether the banker has initiated formal legal action to recover the outstanding loan amount from a delinquent borrower. 0-No 1-Yes
Recovery_status	The outcome of the loan recovery process Written_off (0), fully recovered (1), partially recovered(2)

## 4.0 Dependencies and Risks

Risk	Description	Likelihood	Impact on Outcome
Data Quality and Completeness	The accuracy, completeness, and consistency of loan and borrower records, including key variables such as income, missed payments, and recovery status.	High	Inaccurate or missing data could lead to poor model performance and misleading insights.
Model Interpretability	The ability of stakeholders (e.g., recovery agents, finance teams) to understand and trust the outputs of the model.	Medium	A highly accurate model may not be adopted if its logic is opaque or overly complex.
Stakeholder Buy-in	Support from leadership, recovery operations teams, and IT for model development, deployment, and integration.	High	Strong buy-in improves alignment, facilitates data access, and increases model adoption.
Regulatory Compliance	Adherence to legal and ethical guidelines for the use of personal and financial data.	High	Non-compliance could result in fines or reputational damage, derailing the project.
Technological Infrastructure	Adequacy of systems for storing, processing, and deploying machine learning models (e.g., cloud platforms, databases).	Medium	Limitations in infrastructure may affect scalability or real-time use.
Change Management Readiness	The organization's readiness to integrate data-driven decision-making into existing recovery workflows.	Medium	Resistance to change may hinder implementation and reduce impact.

## 5.0 Deliverable Timelines

Item	Major Events / Milestones	Description	Scope	Days	Date
1.	Project initiation  -Presentation of data set and definition of the problem be addressed  - Initial analysis plan	-Acquisition of data from a reliable source and preliminary screening to understand the problem and potential solutions.  -Development of a high-level business document outlining the project objectives and scope. - Initial data collection and formulation of a high-level analytical approach.	-Data collection  -high level analysis plan	3  4	2025-07-08 to 2025-07-10  2025-07-11 to 2025-07-14
2.	<i>Data preparation</i>  -Data cleaning	- Checking missing values, duplicates, the columns and rows and data types to assess data readiness for analysis.	-Checking the quality of the dataset	3	2025-07-15 to 2025-07-17
3.	Data exploration  -Statistical analysis  -Visualization	-Descriptive statistics and exploratory data analysis. - Use of statistical tests such as Chi-square, ANOVA, and Pearson correlation to evaluate relationships with the target variable. - Visualizations including scatter plots and histograms to identify distribution patterns and outliers. - Initial insights into dataset trends and potential modeling features.	Data quality check and preprocessing	4	2025-07-18 to 2025-07-21
4.	Stakeholder review and approval  -One on one meeting with my advisor	-Review of initial technical work to gather feedback and secure stakeholder alignment and approval.	Approval of the initial findings and direction	1	2025-07-24
5.	Model development -Decision tree		Predictive model		

	<ul style="list-style-type: none"> <li>-Random forest</li> <li>-Logistic regression</li> <li>-Neural network</li> </ul>	<ul style="list-style-type: none"> <li>-Development and evaluation of multiple supervised machine learning models</li> <li>- Comparison of performance metrics to select the most effective model for prediction.</li> </ul>	building and evaluation	7	2025-07-26 to 2025-08-04
6.	Governance	<ul style="list-style-type: none"> <li>-Review of processes to ensure accountability, transparency, and adherence to ethical and data governance standards.</li> </ul>	Compliance and oversight	3	2025-08-05 to 2025-08-07
7.	Documentation -Final report	<ul style="list-style-type: none"> <li>-Compilation of a final written report detailing all technical work, analysis, and findings.</li> <li>- Preparation of deliverables for project closure</li> </ul>	Project completion	4	2025-08-08 to 2025-08-11
8.	Final presentation to stakeholder	<ul style="list-style-type: none"> <li>-Formal presentation of project results, insights, and recommendations to stakeholders.</li> </ul>	Project handoff	1	2025-08-13
9.	Delivery & sign-off	<ul style="list-style-type: none"> <li>-Final approval and sign-off from stakeholders for deployment and implementation of the model.</li> </ul>	Final stakeholder endorsement	1	2025-08-13