

Data-Driven Credit Risk Assessment: Predicting Customer Default for Smarter Lending Decisions

A PROPOSAL REPORT FOR THE BDM CAPSTONE PROJECT

JATIN PURI

23f3000916



INDIAN INSTITUTE OF TECHNOLOGY, MADRAS, CHENNAI TAMIL NADU,
INDIA, 600036

(BS) DEGREE IN DATA SCIENCE AND APPLICATIONS

Table of Contents

1. Executive Summary	3
2. Organization Background	4
3. Problem Statement	5
4. Background of the Problem	5
5. Problem Solving Approach	6
5.1 Methods	6
5.2 Data Source	6
5.3 Tools	7
6. Expected Timeline (WBS and Gantt Chart)	7-8
7. Expected Outcome	8

Declaration

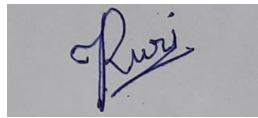
I am working on a project titled “Data-Driven Credit Risk Assessment: Predicting Customer Default for Smarter Lending Decisions.” I hereby declare that the project submitted for the BDM Capstone is my original work and has not been copied from any other student, individual, or entity.

The data presented and analyzed in this project has been sourced from the UCI Machine Learning Repository, specifically the dataset titled “Default of Credit Card Clients.” I affirm that all procedures followed for data analysis have been clearly documented in the report and that the results presented are accurate and based on analytical evaluation.

I understand the importance of academic honesty and integrity. This is an independent project undertaken by me, and I have not collaborated with others. In the event that plagiarism is detected at any stage, I am fully aware of the disciplinary actions that may be imposed by the institution.

I confirm that the recommendations provided are specific to this project and based solely on the context defined herein. I understand that IIT Madras does not endorse or validate the findings or suggestions in this report.

Signature of the candidate:



Name: Jatin Puri

Date: 11 July 2025

1. Executive Summary

This capstone project focuses on using data-driven methods to enhance credit risk assessment for financial institutions. The initiative is set within the context of a fictional company, FinSecure Corp, a mid-sized financial services firm specializing in consumer lending products. With rising default rates impacting operational profitability, the company is seeking innovative ways to refine its credit approval process and mitigate financial risk.

The primary business challenge involves identifying high-risk credit card applicants prior to issuance, with the aim of reducing default rates. This proposal employs a publicly available secondary dataset titled 'Default of Credit Card Clients' from the UCI Machine Learning Repository to address this challenge. Using predictive analytics and machine learning classification models such as logistic regression and decision trees, the project aims to uncover patterns and insights within historical repayment data.

The goal of this project is to build a reliable model that can help spot customers who are likely to default on their credit card payments. This can support banks or financial institutions in making smarter, more cautious lending decisions. Along the way, the project will also highlight which customer behaviors like late payments or high credit usage which are the biggest red flags. The final outcome is meant to be simple, useful, and easy to apply in real-world decision-making.

2. Organization Background

FinSecure Corp is a fictional financial services company headquartered in a major metropolitan city. The company operates in the B2C segment, offering credit card products tailored to middle-income urban consumers. FinSecure's product portfolio includes standard and premium credit cards designed for everyday purchases, travel, and lifestyle spending.

Over the past few years, FinSecure has experienced a gradual increase of customer defaults, specially the new applicants. This trend has led internal reviews of the company's underwriting policies and risk assessment mechanisms. Despite having access to repayment and demographic data, the company has not fully utilized these insights to predict and prevent defaults.

FinSecure's leadership has committed to adopting a data-centric approach to credit evaluation. With the objective of improving customer segmentation and creditworthiness analysis, the company is exploring predictive analytics tools to identify risky applicants at the pre-approval stage. This project simulates that transformation using publicly available data as a stand-in for the company's internal database.

3. Problem Statement

- **To develop a predictive model that identifies customers that are likely to default on their credit card payments.**

This will help banks flag high-risk accounts before even their issuing credit, hence reducing loan loss exposure and strengthening risk control measures.

- **To understand what affects the risk of defaulting on payments, we need to look at important factors like a person's payment history, how much of their credit limit they're using, and their demographic details.**

By analyzing this information, we can better segment our customers, tailor credit limits to individual needs, and make informed policy adjustments for those who may be at a higher risk.

- **Aim to create a data-driven approach that helps us make better decisions when it comes to approving credit and assessing customer risk.**

This will allow the business to refine its lending strategy and improve the overall quality of loan portfolio.

4. Background of the Problem

Credit card companies take a significant financial risk every time they approve a new customer. If someone fails to repay their dues consistently, it can lead to direct losses and affect the company's overall credit health. Traditionally, financial institutions have relied on basic scoring methods or historical intuition to approve or reject applications. But with more people using credit and more data available, there's a strong need to move toward smarter, data-driven decision-making.

The dataset used for the project contains data of about 30,000 customers in Taiwan, including their age, credit limits, payment history, bill amounts, and also whether they defaulted. By analyzing this data we can look for patterns and understand that there might be signals that show higher risk of defaulting. This includes things like maxing out credit limits, missing payments for consecutive months, or showing inconsistent repayment behavior.

Understanding these can help banks make better choices not just in approving or rejecting applications, but also in customizing credit limits or offering early interventions. With better prediction models, companies can lower default rates, reduce financial losses, and create safer lending ecosystems. The goal here isn't just to build a model but to

understand what drives customer defaults and how that knowledge can support smarter business decisions.

5. Problem Solving Approach

5.1 Methods:

The project begins with a thorough exploration of the dataset to understand its structure, check for missing values, and analyze the distribution of key features like credit limit, payment delays, and bill amounts. After proper cleaning and preparing the data, we'll perform exploratory data analysis (EDA) to identify trends, patterns, and potential outliers.

Next feature selection techniques will be used to identify variables that most strongly influence default risk. Then build classification models starting with logistic regression for interpreting and then experimenting with decision trees or random forests for improved performance. Then the models will be evaluated using metrics like accuracy, precision, recall, and AUC-ROC to find the best balance between identifying defaulters and avoiding false positives.

5.2 Data Source:

This dataset is sourced from UCI Machine Learning Repository.

Link: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

It contains detailed data on 30,000 credit card users in Taiwan, including demographic information, repayment behavior, credit limits, and bill/payment amounts across six months.

5.3 Tools:

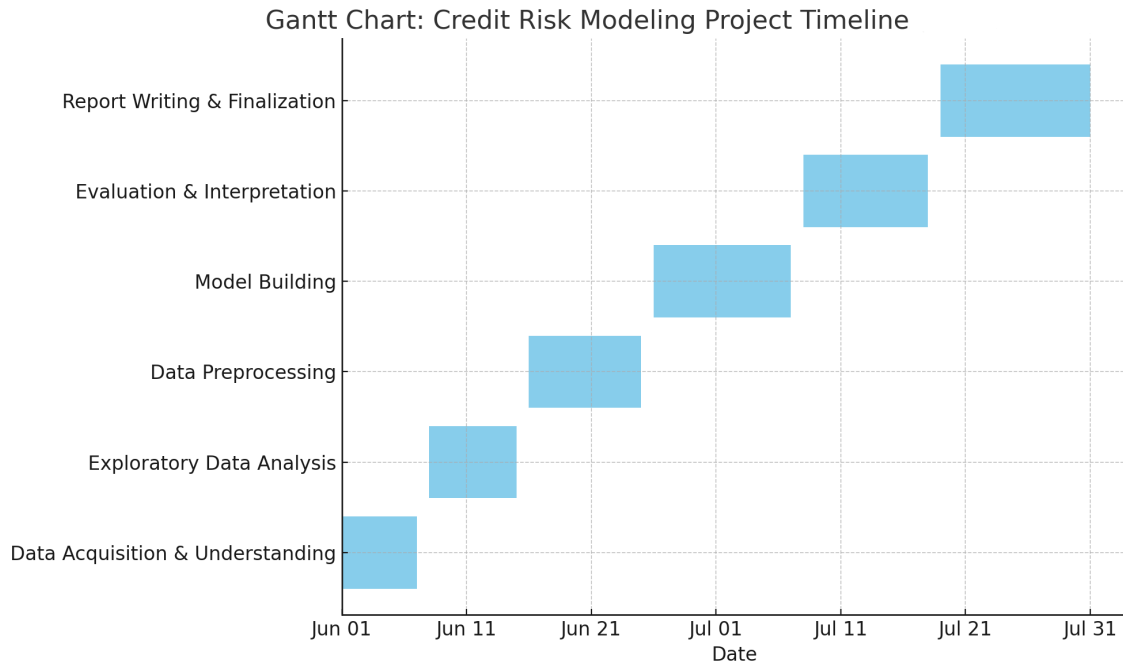
Python will be used for data cleaning, visualization, and modeling. Libraries such as **pandas**, **matplotlib**, **seaborn** and **scikit-learn** are the required tools for analysis.

Jupyter notebook will be the main working environment, allowing both analysis and visualization.

The methods: EDA, logistic regression, and decision trees are all suited for binary classification tasks like default prediction and allow for both interpretability and performance comparison. The dataset is available on the internet, structured, and directly relevant to the problem. Tools like Python and Jupyter are appropriate due to their wide use in business analytics and their capability to handle data preprocessing, modeling, and result visualization efficiently.

6. Expected Timeline (WBS and Gantt Chart)

S.No.	Task	Start Date	End Date
1	Data Acquisition & Understanding	June 01, 2025	June 07, 2025
2	Exploratory Data Analysis	June 08, 2025	June 15, 2025
3	Data Preprocessing	June 16, 2025	June 25, 2025
4	Model Building	June 26, 2025	July 07, 2025
5	Evaluation & Interpretation	July 08, 2025	July 18, 2025
6	Report Writing & Finalization	July 19, 2025	July 31, 2025



7. Expected Outcome

The main outcome of this project will be a working classification model that can predict whether a customer is likely to default on their credit card payments. This model will help financial institutions assess credit risk more accurately before approving or adjusting credit limits.

The project will provide details into the factors such as late payments, high credit usage, or inconsistent billing behavior as these are most closely linked with default risk. These findings can help businesses design targeted policies for high-risk customers or introduce preventive strategies.

The final report will also include solutions for how the institution can use these details to make better decisions during credit approval and customer profiling. All outcomes will be driven by data and backed by statistical evaluation of the model's performance. No assumptions or intuition-based reasoning will be used.