CREDIT CARD DEFAULT PREDICTION

Submitted as a partial fulfilment of Bachelor of Technology in Computer Science & Engineering
(Data Science)
Of
MCKV Institute of Engineering
(An Autonomous Institute under UGC Act, 1956
Approved by AICTE
Affiliated to Maulana Abul Kalam Azad University of Technology, West Bengal)

Project Term Paper

Submitted by

Name of Students	*Examination Roll No.*
UDIT NARAYAN BAIRI	***11600322067***
SANKARASISH MISHRA	***11600322051***
SATANIK MANNA	***11600322053***
SAYAN BARIK	***11600322055***
SOMNATH SASMAL	***11600322058***
SOUMIK SEN	***11600322059***

Under the supervision of

Mr. Sumit Majumdar
Assistant Professor, Dept. of Computer Science & Engineering

Department of Computer Science & Engineering,
MCKV Institute of Engineering
243, G.T. Road(N)
Liluah, Howrah – 711204

April 2025

MCKV Institute of Engineering

(An Autonomous Institute under UGC Act, 1956
Approved by AICTE
Affiliated to Maulana Abul Kalam Azad University of Technology,
West Bengal)

CERTIFICATE OF APPROVAL

[B.Tech. Degree in Computer Science & Engineering (Data Science)]

This project term paper is hereby approved as a creditable study of an engineering subject carried out and presented in a satisfactory way to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is to be understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed, and conclusion drawn therein but approves the project term paper only for the purpose for which it has been submitted.

COMMITTEE ON FINAL
EXAMINATION FOR
EVALUATION OF
PROJECT TERM PAPER

1. _____
2. _____
3. _____

Department of Computer Science & Engineering
MCKV Institute of Engineering
243, G. T. Road (N),
Liluah, Howrah-711204

CERTIFICATE OF RECOMMENDATION

I hereby recommend that the term paper has been prepared under my supervision by the students listed below

Sl. No.	Name the Student	Signature
1	Udit Narayan Bairi	
2	Sankarasish Mishra	
3	Satanik Manna	
4	Sayan Barik	
5	Somnath Sasmal	
6	Soumik Sen	

for the project entitled **"Credit card default prediction"** to be accepted in partial fulfillment of the requirements for the degree of **Bachelor of Technology in Computer Science & Engineering (Data Science)**.

Mr. Avijit Bose
Assistant Professor & Head of the Department
Computer Science & Engineering
MCKV Institute of Engineering, Howrah

Project Guide
Mr. Sumit Majumdar
Assistant Professor
Computer Science & Engineering Dept.

Acknowledgement:

We would like to express our heartfelt gratitude to all those who have contributed to the successful completion of our 4th year even semester project during our B Tech course. First and foremost, we extend our deepest appreciation to our project guide Mr. Sumit Majumdar, whose guidance, expertise, and constant support have been invaluable throughout the project. He has provided us with valuable insights, suggestions, and encouragement, helping us navigate through the various challenges and ensuring the project's smooth progress.

We would like to thank the faculty members of the Computer Science and Data Science Department for their knowledge and expertise, which formed the foundation of our project. Their teachings and guidance have been instrumental in shaping our understanding of the subject matter and have contributed significantly to the successful execution of our project. We are grateful to the staff and authorities of MCKVIE for providing us with the necessary resources, infrastructure, and access to important materials. Their continuous support and encouragement have been vital in our project's completion.

We would also like to express our gratitude to our classmates and friends for their assistance and motivation throughout the project. Their valuable feedback and brainstorming sessions have enriched our project and made the journey more enjoyable.

Last but not least, we extend our heartfelt thanks to our families for their unwavering support and understanding during this entire journey. Their encouragement, patience, and belief in us have been crucial in overcoming challenges and achieving our goals.

We acknowledge that this project would not have been possible without the combined efforts and support of all the individuals mentioned above. We are sincerely grateful for their contributions, and we consider ourselves fortunate to have been surrounded by such a wonderful team.

Thank you all once again for your valuable support and contributions.

Content:

Abstract:

The banking sector is widely acknowledged for its intrinsic unpredictability and susceptibility to risk. Bank loans have emerged as one of the most recent services offered over the past several decades. Banks typically serve as intermediaries for loans, investments, short-term loans, and other types of credit. The usage of credit cards is experiencing a steady increase, thereby leading to a rise in the default rate that banks encounter.

The increasing use of credit cards has made it essential for financial institutions to assess the risk associated with lending. One of the key challenges is predicting whether a customer is likely to default on their credit card payment. This project
aims to develop a machine learning-based model to predict credit card defaults using historical customer data. The dataset used includes features such as credit limit, payment history, bill statements, and demographic information.

The dataset used 30000 of bank customer about their payment details, collected from the Kaggle data sets. These samples information about credit card holders, their credit usage, payment history, and demographic details, along with a binary target variable indicating whether they defaulted on their payment in the following month.

Data preprocessing techniques like handling missing values, normalization, and feature selection were applied to prepare the data. Multiple classification models—including Logistic Regression, Decision Trees, Random Forest, and other classification algorithm were trained and evaluated using metrics such as accuracy, precision, recall, and F1-score.
The best-performing model was selected based on its ability to correctly identify potential defaulters.
This project demonstrates how data science and machine learning can be effectively used to assist banks and financial institutions in making informed decisions, minimizing risk, and improving financial stability.

1. Introduction:

A Credit Card Default Prediction Project aims to develop a model that can accurately forecast whether a credit cardholder will default on their payments in the future. This is a critical task for financial institutions as it directly impacts their risk management strategies and profitability. By identifying potential defaulters early, these institutions can take proactive measures to mitigate losses.

Credit card default occurs when consumers fail to fulfil the minimal repayment for their credit card bill consecutively for two or more reporting cycles. These actions can result in adverse outcomes for the cardholder, such as incurring late fees, fines, detrimental impact on their
credit rating, and other legal actions from the credit card issuing authority. The banking sector currently utilizes ML techniques, particularly those that apply diverse categorization algorithms, to accurately categorize consumers into distinct categories for improved trend predictions.
One benefit of utilising ML for credit card default predictions, as opposed to traditional methods, is its ability to accurately detect probable default risks. The accuracy of credit card default prediction models is contingent upon the reliability and quality of the input data used for training them.

2. Scope of the work:

The scope of this project is to develop a machine learning model that can accurately predict whether a credit card holder will default on their payment in the next month. The project involves collecting and preprocessing relevant data, selecting appropriate features, applying classification algorithms, and evaluating model performance using metrics like accuracy, precision, recall, and F1-score.

This system aims to assist financial institutions in assessing customer risk and making better credit-related decisions. The prediction model can help reduce financial losses by identifying high-risk customers early. The scope also includes visualizing important trends and insights from the dataset, and potentially deploying the model for practical use.

The project is limited to historical data analysis and does not involve real-time data or integration with existing banking systems.

The scope of a credit card project can vary significantly depending on the specific goals and context. Here's a breakdown of potential aspects:
Possible Objectives:
Understanding Consumer Behaviour: Analysing how consumers use credit cards, their spending habits, and the factors influencing their choices.
Customer Satisfaction: Evaluating the satisfaction levels of credit card users with the services provided by different banks.
Identifying Influencing Factors: Determining the factors that prompt customers to use credit cards.

Problem Identification: Understanding the challenges and issues faced by credit card users.
Fraud Detection: Developing systems and algorithms to identify and prevent fraudulent credit card transactions.
Financial Inclusion: Examining the role of credit cards in promoting financial inclusion.
Improving Creditworthiness: Understanding how responsible credit card use can build a positive credit history.
Enhancing Security: Implementing measures to protect credit card users from unauthorized transactions and data breaches.
Potential Features to Explore:
Credit Limits: The maximum spending power offered to cardholders.

The main scope is this is also helpful for banking purpose.

3. Survey of the Previous Works:

3.1 Traditional Statistical Methods

Early methods relied heavily on statistical models:

Logistic Regression: One of the most popular and interpretable models for predicting default. It models the probability of default as a logistic function of predictors like income, debt, etc.
Discriminant Analysis: (e.g., Linear Discriminant Analysis - LDA) was used for classification between defaulters and non-defaulters.
Probit Models: Similar to logistic regression but assumes a normal cumulative distribution function.
Example:
Hand and Henley (1997) - Surveyed statistical models used in credit scoring, emphasizing logistic regression.

3.2 Machine Learning Approaches

With the availability of more data and computational power, machine learning models started outperforming traditional ones:
1.Decision Trees (e.g., CART, C4.5)
2.Random Forests (Ensemble of decision trees)
3.Support Vector Machines (SVM): Good for high-dimensional feature spaces.
4.Neural Networks: Used to capture complex non-linear relationships.
Gradient Boosting Machines (GBM, XGBoost, LightGBM): Popular for their strong performance.
Example:
Yeh and Lien (2009) ("The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients"): Compared SVM, decision trees, neural networks, and logistic regression. Found that SVM slightly outperformed others.

3.3 Commonly Used Datasets

Banking Institution Proprietary Datasets
Some studies use real-world datasets obtained from financial institutions.
Typically larger (millions of records) but not publicly available due to confidentiality.
Kaggle Datasets: Variants datasets are available. Example: "Default of Credit Card Clients Dataset" on Kaggle.

4. Hardware and Software Details:

4.1 No additional hardware require.

4.2 Only mention those Software (Packages/ Tools):

4.2.1 Name of the Package/ Tool:

1. Python
2. Pandas
3. Numpy
4. Matplotlib
5. Seaborn
6. Scikit-learn (Sklearn)
7. XGBoost
8. Jupyter Notebook / Google Colab

4.2.2 Purpose of use:

1. Python:
Purpose: Core programming language used to implement the project.
Why: Python is versatile, has vast libraries for data science, and is easy to use for machine learning workflows.

2. Pandas:
Purpose: Used for data manipulation and preprocessing.
Why: Efficient for reading datasets (CSV), handling missing values, filtering rows, creating few features, and summarizing statistics.

3. NumPy:
Purpose: Used for numerical operations and array handling.
Why: Faster mathematical computations and vectorized operations which speed up model training and preprocessing.

4. Matplotlib & Seaborn:
Purpose: Used for data visualization.
Why: Matplotlib: For plotting basic graphs like bar charts, histograms, and line plots.
Seaborn: For advanced visualizations like correlation heatmaps, box plots and distribution plots to identify feature relationships.

5. Scikit-learn :
Purpose: The main machine learning library used for:
Data splitting (train/test)

Model training (Logistic Regression, Decision Tree, etc.)
Evaluation (Accuracy, Precision, Recall, ROC)
Preprocessing (Standard Scaler, Label Encoder)
Why: Easy-to-use, comprehensive, and supports a wide range of ML algorithm and tools.

6. **XGBoost:**
Purpose: Used for training a high-performance gradient boosting model.
Why: Known for delivering high accuracy in classification tasks, handles missing Data well, and offers regularization to reduce overfitting.

7. Jupyter Notebook / Google Colab:
Purpose: Coding environment used for implementation, testing, Visualization
Why: Interactive interface makes it easy to write, test, and visualize results place with markdown support for documentation.

4.2.3. Licensed/ Free /Open Sourced (with source link):

VsCode v1.86
Type: Open-Sourced, free software
Available: https://code.visualstudio.com/

Jupyter Notebook
Type: Open-Sourced, free software
Available: https://jupyter.org

Google Colab
Type: Free online service (proprietary backend, free to use)
Available: https://colab.research.google.com

4.3 Dataset to be used (if any):

4.3.1 Source of the dataset:

The **"Default of Credit Card Clients"** dataset is originally sourced from the UCI Machine Learning Repository. It was introduced by Yeh and Lien in their 2009 study, which investigated the behaviour of Taiwanese credit card clients with respect to default payments. The dataset comprises 30,000 instances featuring 25 attributes—including demographic details, credit limits, payment history (recorded as ordinal values), billing statement amounts, and actual payment amounts.

Over time, this dataset has also been popularized on platforms such as Kaggle. The Kaggle dataset is essentially a replica of the original UCI version and provides an easy-to-access resource along with community notebooks and kernels demonstrating various analysis techniques.
If you are looking to cite the dataset in your report, you might include the following details based on the citation style you are using:
Original Source (UCI Repository):
Yeh, I., & Lien, C.-H. (2009). *Default of Credit Card Clients* [Data set]. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
Kaggle Hosted Version:
(n.d.). *Default of Credit Card Clients Dataset* [Data set]. Kaggle. Retrieved April 25, 2025, from https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

4.3.2 Type of the data:

This is a classification model-based data, with 25 attributes of different type.
Understanding these data types helps tailor the preprocessing steps:

Categorical variables (like SEX, EDUCATION, and MARRIAGE) often require encoding techniques (e.g., one-hot encoding) before being used in many machine learning algorithms.

Ordinal variables (such as PAY_0 to PAY_6) might be maintained as-is or transformed based on the modeling technique because their inherent order is meaningful.

Continuous numerical features (like LIMIT_BAL, BILL_AMT*, and PAY_AMT*) can be scaled or normalized depending on the needs of your model.
The **binary target variable** is ideal for classification tasks.

4.3.3. Number of total data and features:

Number of total data and features, you can report that the Default of Credit Card Clients dataset consists of:

30,000 instances (observations): Each row represents a distinct credit card client.

25 attributes (features):
24 explanatory variables: These include demographic information (e.g., sex, education, age), credit data (e.g., credit limit), payment history (e.g.,PAY_0 to PAY_6), and billing and payment amounts (e.g., BILL_AMT1–BILL_AMT6, PAY_AMT1–PAY_AMT6).

target variable: The column default.payment.next.month indicates whether the client defaulted (1) or not (0).

5. Roadmap of the Future work

5.1 Proposed works for the Seventh Semester.

In next upcoming semester (7th semester) we built a machine Learning model by using this dataset with good predictive performance using traditional machine learning models like Logistic Regression, Decision Tree, Random Forest, XG-Boost, there are several directions for improvement and further exploration:

By using following classification algorithm we test our model and understand how does our model behaviour on this algorithms ,Check the accuracy and other classification measure like(Precision Recall ,F1 score , ROC-AUC curve etc)

Applying techniques such as Grid Search, Random Search, or Bayesian Optimization to further optimize model performance.

5.2 Proposed works for the eighth semester.

Deploy the model as an API for real-time credit application assessment.
Create a dashboard for bank officers to visualize customer risk.
Explore deep learning models for potentially better performance on larger datasets.
Crafting an effective **UI/UX design** for a credit risk dashboard is crucial for enhancing user experience and facilitating informed decision-making. Here's a comprehensive guide to help you design an intuitive and user-friendly dashboard.

6.Reference:

1.Hand, D. J., & Henley, W. E. (1997). *Statistical classification methods in consumer credit scoring: a review*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.

2. Towards Data Science. (2020). *Credit Card Default Prediction: Data Science Project.* *https://towardsdatascience.com/credit-card-default-prediction-data-science-project-5cf5c7d3a7d6*

3. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. ***Expert Systems with Applications****, 36(2), 2473-2480.*
https://doi.org/10.1016/j.eswa.2007.12.020

4. Baesen's, B., Van Gastel, T., Viaene, S., Stepanova, M., suvken, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. ***Journal of the Operational Research Society****, 54(6), 627-635.*
https://doi.org/10.1057/palgrave.jors.2601545