Automated Credit Scoring: AI-Driven Predictive Analytics for Smarter Lending

Final Year Project

PROPOSAL

Supervisor

Dr. Muhammad Saqib Sohail

Co-Supervisor

Mr. Saad Abrar Khan - CEO, Aletheia-AI Pvt. Ltd.

Submitted by (BSCS-S25-012)

03-134221-031 Muhammad Wahaaj Tauqir

03-134212-080 Muzammil Amjad

Bachelor of Science in Computer Science

Department of Computer Sciences

Bahria University Lahore Campus

February 2025

Version History

Version	Date	Remarks
1.0	Feb 24, 2025	First Complete Submission to FYP Office

Abstract

Commercial clients (business owners, retailers) of wholesale distributors often rely on credit lines to purchase goods, especially during economic crises when banks hesitate to lend to retailers. This affects the sales of distributors, prompting them to offer products on credit. However, when distributors extend credit, they face the risk of customer default. They need a reliable method to evaluate the creditworthiness of their customers to minimize financial loss.

To address these challenges, we propose an automated AI driven credit scoring system that automates the evaluation process by analysing several features including customer purchase behaviour, transaction history, and repayment patterns. The system will include machine learning algorithms such as Logistic Regression, Decision Trees, and XG-Boost, performing automated processes, to generate accurate and scalable credit scores. We will integrate this system with a cloud environment and a frontend dashboard. The expected impact is to enhance distributor's decision-making processes, allowing them to make informed lending decisions quickly, reduce financial risks, and improve overall business stability.

Table of Contents

1. Introduction

We are collaborating with Aletheia-AI Pvt. Ltd. to develop an AI-driven credit scoring system for wholesale distributors. Small business owners and retailers often depend on credit lines to obtain goods. Especially during economic crisis, retailers face difficulty in obtaining loans and they themselves don't have enough cash in the bank to buy goods for sale. This affects the sales of distributors. To keep the business running, the wholesale distributor directly sells goods on credit. But this activity involves a high risk of customer default, which results in the retailer not being able to return the credit. So, the distributor needs to evaluate the retailer's creditworthiness to reduce financial losses. (Niti Saxena, 2024)

Traditional businesses take months and require many documents to make credit decisions. This makes it challenging for distributors to assess a retailer's creditworthiness. Our goal is to do it in real time without asking for additional documents, only looking at the behavioral data. Currently, giants like SCHUFA and Experian offer creditworthiness systems, but there are two major challenges. Firstly, their cost is too high and secondly their data is not updated. To solve these problems, we propose an automated AI-driven credit scoring system that automates the evaluation process by analyzing customer purchase behavior, transaction history, repayment patterns and several other factors. (Alessandro Bitetto, 2024)

Our system will automate processes like data processing, feature selection, and machine learning modeling to generate precise and scalable credit scores. This will help the distributors in identifying potential bad retailers, to lend smartly.

2. Problem Description

Problem Description of our final year project is divided into the following subsections.

2.1. Primary Scope

The company (wholesale distributor) is extending credit lines for high-value purchases, as banks are reluctant to lend to retailers. There remains a risk of customer default, which can result in significant financial loss.

Traditional credit evaluation methods are manual, time-consuming, and prone to human error. The existing solutions are costly, and their model is not adaptive. As the data is continuously updated, the model must be extremely adaptive and respond according to the updated data.

To address this, an automated, AI-driven credit scoring system is needed to assess customer creditworthiness based on updated transactional behavior, financial history, and repayment patterns. Our project involves the following major phases. Organizing data, i.e. bringing it into structured form. Analyzing variables (features), identifying the most important ones. Applying univariate, multivariate analysis and finding the most effective combination. Compiling a score card. We will use machine learning techniques to automate these processes. Then it will be deployed in a cloud environment.

2.2. Final Deliverable of the Project and Beneficiaries

The deliverables will include an automated credit scoring system deployed in a cloud environment. The beneficiary is the wholesale distributor who will use it to make faster and accurate decisions regarding the creditworthiness of the retailer.

2.3. Objectives

We aim to develop an automated, credit risk assessment system that enhances the accuracy of credit decision making.

This can be divided into sub objectives:

- Develop a workflow in which a supervised machine learning model is deployed and managed.
- Implement explainable ML models at key stages for creditworthiness scoring.
- Automate data preprocessing, feature engineering, and model training with machine learning.
- Build an interactive interface for visualizing credit decisions and managing scoring processes.

2.4. Novelty

There is no similar FYP already approved in BULC. Our project's key differentiator is the automation of the overall process of credit scoring. This will result in a generalized model that will be highly effective in faster and accurate decision making regarding the creditworthiness of retailers. This will result in saving the distributor's time and cost and will minimize financial loss.

Unlike existing solutions such as Experian and SCHUFA, our model is adaptive, leveraging real-time transactional data and automated retraining to ensure up-to-date credit scoring.

3. Methodology

1. Data Collection and Ingestion Engine

Customer data, such as purchase behavior, transaction history, and repayment patterns, will be collected from the Google drive storage spaces of the distributors. An ingestion engine will be needed to inject the data from google drive into our application. This will give access to the data to our system. Then it will be further processed.

2. Statistical Analysis

Customer data will be categorized into behavioral scorecards:

- Master Data – Discrete variables for customer identification and classification.
- Movement Data – Continuous variables reflecting buying and payment behavior.

With 150+ variables, users will classify them as master, movement, or target. This flexible categorization will allow the model to analyze multiple behaviors beyond credit default prediction. Univariate and multivariate analysis will refine feature selection. (Jomark Pablo Noriega, 2023)

3. Machine Learning & Predictive Modeling

Supervised learning models, including Logistic Regression, Decision Trees, KS Statistics and XG Boost, will be trained to predict repayment probability. Hyperparameter tuning will optimize model performance.

4. Model Validation

The models will be validated using accuracy, precision, recall, F1-score, and AUC-ROC metrics. To ensure interpretability, final output will be mapped into three risk levels:

- Red: High risk of default
- Yellow: Moderate risk
- Green: Low risk

Each score will include a probability of default (PD) for precise decision-making. Techniques such as Weight of Evidence (WoE) will be employed to enhance risk assessment.

5. Future Enhancements

To maintain and improve model accuracy over time:

- External Datasets – Incorporate macroeconomic variables (e.g., inflation, interest rates) to enhance predictive performance.
- Population Stability Models – Monitor model effectiveness, considering the typical two-year lifespan of credit models, with adjustments for market volatility.
- Effectiveness Formula – Develop a metric to determine when model retraining is required, ensuring adaptability and reliability.

These refinements will enhance the system's scalability and long-term performance.

6. Deployment & System Integration

The trained models will be deployed using Docker for portability and Kubernetes for scalable cloud-based orchestration. A backend will integrate the credit scoring system with distributor financial applications via APIs. A React.js dashboard will provide real-time access to credit scores and risk assessments.

7. Automation & Security

Data ingestion pipelines will be automated to update credit scores in real-time. Security measures, including data encryption for storage and OAuth for API authentication, will ensure compliance with data protection regulations.

4. Feasibility Plan

We will achieve our objectives as we fulfill all the requirements. Our team members already have experience in the industry and have completed and delivered several projects.

4.1. Resource Requirement

Required Resources includes some cloud-based infrastructure and computing resources to ensure efficient data processing, model training, and deployment.

Cloud Services & Deployment: Docker-based deployment on AWS/GCP for scalability and real-time processing.

Dataset: The company (Aletheia-AI Pvt. Ltd.) will provide a dataset related to training the credit scoring model.

Hardware: High-performance machine with GPU (if required for model training) to handle complex computations.

4.1.1. Expertise of the Team

To ensure the successful completion of the Automated Credit Scoring project, our team is equipped with the necessary knowledge and skills. Our team members have strengths in React Js and Python in particular modules such as (Django, Pandas) and training machine learning models. We have also taken university courses in AI. Additionally, we have certifications from Google in Data Analytics.

The extensive industrial experience of our team as a software engineer will benefit us in project execution and management. As both team members have the relevant knowledge, have taken the required courses, and have equal interest, our team is fully capable of delivering the full project.

4.1.2. Tools / Technology

This project is completely software based so software tools and technologies are needed despite the model training hardware.

Category	Tools	Purpose	Availability
Data Processing	Pandas, NumPy, Big-Query	Data cleaning, transformation, and large-scale financial analysis.	Available
Machine Learning & Predictive Modeling	Logistic Regression, Decision Trees, XG-Boost	Supervised learning models and improving accuracy for credit risk classification.	Available
Database Management	PostgreSQL, Big-Query	Secure data storage, real-time querying, and scalable data management.	Available
Backend Development	Python (Django)	API management, business logic implementation, and database communication.	Available
Frontend Development	React.js	Build a dynamic and interactive dashboard for credit risk analysis.	Available

Containerization & Orchestration	Docker, Kubernetes	Application deployment, scalability, and environment consistency.	Available
Cloud Deployment	AWS/GCP	Hosting, real-time data processing, and high availability.	Available
Version Control	Git, GitHub/GitLab	Code management, cloning, automated testing, and continuous integration/deployment.	Available

4.1.3. Budget

No Financial Requirements. The project has no major financial requirements, as most tools and services are open source or available through free tier cloud options. Costs for the deployment of the system will be covered by the company.

4.2. Risks Involved

There are two major risks involved in our project.

1. **AI Model Explainability Risk:**

It is very crucial for our model to be explainable and logical. The customer will only implement such a system if there is a logical explanation behind its score assessment. Complex ML models may operate as "black boxes," making it difficult to explain decisions

Mitigation:

- Use interpretable ML models.
- Develop clear visualization of the purpose of every AI model we will use.
- Rigorous testing.

2. **Automation of the process**

We will have to automate the complete process by keeping the model accuracy on top.

Mitigation:

- Implement robust monitoring and logging to track performance.
- Regularly update and retrain models with up-to-date data.
- Use cross-validation and hyperparameter tuning to optimize performance.

5. Key Milestones and Schedule

5.1. Key Milestones

The project has been divided into eleven major milestones.

Table 5-1: Breakdown of work in the form of milestones

S. No.	S. No. of Predecessor Milestone	Key Milestone Name / Description	Duration (person-hours)
1	-	Initial Workflow Design	20
2	1	Data Collection & Preprocessing	40
3	2	Ingestion Engine	55
4	3	Feature Engineering & Variable Selection	70
5	4	Creditworthiness Scoring ML Model Development	100
6	5	Model Enhancement & Optimization	70
7	6	Automation of Processes through ML	90
8	-	Web Application Development	40
9	8	Backend API Development	40
10	9	Deploying on GCP	30
11	10	System Integration & Testing	21

The project follows a structured workflow, starting with Initial Workflow Design (20 hours) to define system architecture and data flow. Next, Data Collection & Preprocessing (40 hours) gathers and structures data, followed by Ingestion Engine (55 hours) to automate data retrieval. Feature Engineering & Variable Selection (70 hours) identifies key variables for model training, leading to Creditworthiness Scoring ML Model Development (100 hours) for risk assessment. Model Enhancement & Optimization (70 hours) fine-tunes models, while Automation of Processes through ML (90 hours) streamlines data handling. Parallelly, Web Application Development (40 hours) builds the frontend, supported by Backend API Development (40 hours) for system communication. Deploying GCP (30 hours) ensures scalable deployment, and System Integration & Testing (21 hours) finalizes and validates the complete system.

5.2.Gantt Chart

Activity	Weeks																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Initial Workflow Design									Midterm Exam Week									Exam Week
Data Collection & Preprocessing																		
Ingestion Engine																		
Feature Engineering & Variable Selection																		
Creditworthiness Scoring																		

Figure 5-1: SEMESTER 1 (Spring 2025)

In Semester 1, the project begins with Initial Workflow Design (20 hours) to establish system architecture and data flow. This is followed by Data Collection & Preprocessing (40 hours) to gather, clean, and structure customer data. Next, the Ingestion Engine (55 hours) is developed to automate data retrieval from external sources. Once the data pipeline is established, Feature Engineering & Variable Selection (70 hours) is performed to identify critical variables for model training. Finally, Creditworthiness Scoring ML Model Development (100 hours) is conducted, where machine learning models are trained to assess customer credit risk effectively.

Activity	Weeks																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Model Enhancement & Optimization									Mid Exam Week									Exam Week
Automation of Processes through ML																		
Web Application Development																		
Backend API Development																		
Deploying on GCP																		
System Integration & Testing																		

Figure 5-2: SEMESTER 2 (Fall 2025)

In Semester 2, the focus shifts to optimization and deployment. Model Enhancement & Optimization (70 hours) fine-tunes ML models for better accuracy, followed by Automation of Processes through ML (90 hours) to streamline data handling. Meanwhile, Web Application Development (40 hours) is carried out to build an interactive frontend, supported by Backend API Development (40 hours) to enable seamless system communication. Deploying on GCP (30 hours) ensures scalable cloud deployment, and the project concludes with System Integration & Testing (21 hours) to validate and finalize the complete system.

REFERENCES

Journals:

[1] Alessandro Bitetto, P. C. (2024). Can we trust machine learning to predict the credit risk of small businesses? *Review of Quantitative Finance and Accounting*, 925–954.

[2] Jomark Pablo Noriega, L. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data, Article 169 in Volume 8*.

Conference Papers:

[3] Niti Saxena, T. S. (2024). Machine Learning Techniques for Credit Scoring in Banking with Management, HR, and Organizational Key Components. *024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (p. 6). India: Institute of Electrical and Electronics Engineers (IEEE).