International Journal of Innovative Research and Scientific Studies, 8(6) 2025, pages: 2099-2120

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Credit risk prediction using behavioral features in an AI chatbot interface

Nidhal Ziadi Ellouze

Faculty of sciences of Tunis University of Tunis El Manar LARIMRAF Laboratory, Africa.

(*Email: nidhal.ziadiellouze@fst.utm.tn*)

Abstract

The following article presents an innovative credit risk assessment system that combines artificial intelligence with real-time interaction through a Streamlit-deployed chatbot. The aims of this article are demonstrated using a dataset of 3,080 clients from Tunisian banks in 2024, including financial and behavioral variables. Random Forest and XGBoost models were trained to predict loan defaults with up to 90% accuracy. The focus was on detecting high-risk profiles, achieving perfect recall (100%) and 76% precision for this class. The SHAP method ensures decision transparency by identifying key predictive variables. This system enables bank advisors to instantly obtain a risk score with personalized explanations, enhancing the speed, efficiency, and trustworthiness of credit evaluations, while outperforming traditional approaches.

Keywords: AI Chatbot, Behavioral Scoring, Credit Risk, Explainable AI, Machine Learning, Random Forest, SHAP, Streamlit, XGBoost.

DOI: 10.53894/ijirss.v8i6.10083
Funding: This study received no specific financial support.
History: Received: 17 June 2025 / **Revised:** 21 July 2025 / **Accepted:** 23 July 2025 / **Published:** 19 September 2025

Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.
Transparency: The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

In a rapidly changing financial landscape, managing credit risk is central to sustainable and responsible banking. Financial institutions are under pressure to reach more customers and offer a wider range of loans, making access to effective credit risk assessment methodologies increasingly crucial [1]. Conventional methods often rely on rigid scoring systems and manual processes, leading to systematic weaknesses in handling the volume and complexity of available credit data, especially concerning behavioral data, emerging market dynamics, and digital interactions with customers [2].

In parallel to this transition, Artificial Intelligence (AI) and Machine Learning (ML) methodologies are now integrated into credit risk models, representing a transformative development in the field [3]. These advancements allow lenders to learn from complex and multi-dimensional behavioral and financial data, uncovering hidden patterns and generating predictive decisions with a high degree of accuracy [4]. This progress enhances risk management while improving customer experience by providing quicker, fairer, and more personalized credit decisions.

Our research question for this article is: How can we develop an AI-enabled chatbot that utilizes behavioral credit scoring models to fully automate and optimize real-time predictions of retail clients' risk classes, while providing personalized and explainable responses to enhance understanding and trust in credit decisions? The goals include developing a rigorous and interpretable credit scoring model using machine learning techniques like Random Forest and XGBoost and integrating this model into an interactive chatbot that guides users through the credit assessment process [5].

The approach followed in this research is well-structured. First, customer data will be thoroughly explored, followed by feature engineering, model training, and evaluation using industry-relevant metrics such as success rate, recall, and AUC.

Emphasis will also be placed on robustness, generalizability, and the integration of the final model into a web-based user interface developed with Streamlit [5].

By combining advanced data science techniques with an intuitive conversational AI system, this research aims to support the digital transformation strategy of the Tunisian Bank while modernizing credit risk management and aligning with broader goals of automation and transparency [6]. This framework ensures a logical progression from understanding the institutional background to methodological design, implementation, and evaluation, demonstrating how AI and machine learning can enhance credit risk assessment in a practical banking context.

2. Literature Review

The examination of loan default and forecast risks reported by Provenzano et al. [7] displays a complicated interrelation of several economic variables, consumer behaviors, and credit policies that have immense implications on repayment levels of credit [7]. Over the past decades, levels of defaults have varied depending on general economic conditions such as employment levels, interest rates, and economic cycles. For instance, in economic boom conditions, levels of defaults fall as people have higher incomes and more secure employment, making it easier for them to meet their repayment schedules. Te-Cheng et al. [8] found that during economic recessions like that of 2007-2008, however, levels of defaults increase as people face layoffs, reduced earnings, and more economic strain. This pattern illustrates why it is crucial to consider economic conditions when assessing repayment risks of credit. Default can be defined as non-payment by a debtor on an agreed date of a sum due. In other words, a default is the total or partial non-payment of one or more monthly credit installments as a result of insolvency.

2.1. Traditional vs Machine Learning-Based Loan Risk Assessment

According to Sizan Chiamaka [9], the traditional approaches to credit measurement of loan credit have largely relied on conventional credit scoring approaches such as the FICO score that consolidates a borrower's credit history, repayment profile, and outstanding balances into a numerical credit score that captures creditworthiness in all its facets. Other measures of debt-to-income (DTI) ratio have similarly acted as pillar measures in the creditworthiness of a borrower by providing insight into their ability to make debt payments every month in proportion to their revenues [10]. As much as these conventional approaches have provided credit measurement of credit risk with a basic platform to work from, these have several disadvantages that make their performance in predicting loan defaults questionable at times. One significant disadvantage of traditional credit scoring methods is that these methods make assumptions based on records that may not reflect current economic conditions or the specific situations of borrowers.

In contrast, credit risk assessment through machine learning has several advantages that offset these disadvantages [1]. Employing algorithms that possess the capability to sift through large amounts of data from disparate sources, machine learning models have the potential to discern sophisticated interlinkages and patterns that traditional methods may not detect. They can consider a vast array of variables, from borrower details to economic signals to behavioral signals, to provide more inclusive credit risk measurement [11]. Machine learning models possess the potential to improve over time by constantly learning from fresh details to make their forecasts more accurate and their estimations more precise Possessing the potential to discern sophisticated risk patterns makes machine learning a forceful tool for credit risk improvement that has the potential to make more accurate estimations for lenders to make more prudent judgments and ultimately reduce default rates Luo et al. [12].

Ive et al. [11] stated that the use of credit risk analysis through machine learning within finance represents an example of how these methods may improve efficiency and reliability in lending. Supervised algorithms have attracted considerable attention as a vehicle for credit risk prediction performance. They are trained on past experiences, whereby they have learned from the past results and applied associated learning to review new credit applications. Items like Random Forest, XGBoost, and neural networks have been used broadly to develop prediction models for default probabilities that draw on the characteristics of the borrowers. According to Eustaquio-Jiménez et al. [13], a really great thing about supervised learning models is that they can incorporate a wide variety of input features to give deeper insight into borrower risk [14]. For example, traditional credit scores often focus solely on credit history, but machine learning models may include other variables or even other alternative data points (such as utility payments) that illuminate additional facets of a potential borrower's overall financial soundness. With this multi-faceted approach, lenders have a better chance of identifying high-risk applicants with finer precision, which allows them to tailor their lending strategy towards the specific clientele they serve. Success stories of AI-enabled loan approval processes further establish the potential of machine learning in credit decision-making. Financial organizations have implemented these technologies to automate their credit processes, enhance their capabilities in credit risk evaluation, and provide more sophisticated experiences to their customers. They also reduce default rates and improve the performance of credits. Through utilizing the capabilities of machine learning, these organizations take center stage in redefining credit risk analysis and management [15].

Through predictive modelling, machine learning has transformed the landscape of credit risk assessments and provided institutions with systems that can analyze very large and diverse amounts of financial and behavioral data with a high level of accuracy. Moving forward, AI chatbots have the potential to offer an effective combination of machine learning and natural language processing (NLP) that can automate customer interactions and enable credit decisions to be made in real time.

Al Shiam et al. [5] state that AI chatbots allow for effective user interaction and operational efficiencies by providing real-time, personalized, and contextualized responses that are based on predictive models. In the context of credit scoring, studies like [8] describe how AI chatbots are two-way interactive interfaces that facilitate the collection of behavioral data and provide explanatory credit decisions through explainable AI (XAI) methods such as SHAP and LIME. AI chatbots can also be integrated with behavioral scoring systems to allow financial institutions to predict risk classes in real time, based on assured transparency to the consumer to create better trust [16]. This integration of machine learning, conversational AI, and behavior analytics represents an innovative threshold in the world of retail credit risk management.

2.2. Problem Statement

2.2.1. Context

In the financial industry, credit risk evaluation is a strategic priority for banks. Traditionally, banks rely on historical financial data within credit scoring models with established customers' behaviors. Recently, as digital banking has grown, there is an increased demand to incorporate real-time behavioral data into credit risk assessment.

In addition, banks are seeking to use intelligent tools to improve customer engagement, such as an AI chatbot, while also creating a tool that is capable of collecting up-to-date information about the consumer. Using a combination of technology creates new opportunities in dynamic and personalized credit risk estimation.

2.2.2. Problem Definition

This article aims to work on a system that will simultaneously:

Assess a client's credit risk based on behavioral scoring models based on machine learning, and also Interact with clients to collect and process relevant data through an intelligent chatbot in real time.

Thus, our research question for this article is:

How to develop an AI-enabled chatbot that uses behavioral credit scoring models to fully automate and optimize real-time predictions of retail clients' risk classes, with personalized, explainable responses to bolster understanding and trust in clients' credit decisions?

2.2.3. Hypotheses and objectives

This article proposes a machine learning model, such as XGBoost and Random Forest, to predict credit risks of prospective trainees by implementing a predictive system. The institution can reduce financial risks. Additionally, the model serves as a decision-report tool for managing loan assistance more efficiently. To address the identified problem, this study proposes the following hypotheses:

$H_{1:}$ An AI chatbot using behavioral scoring models is better at predicting credit risk class than conventional scoring.

$H_{2:}$ The ability of AI chatbots to predict risk classes in real-time increases the speed and efficiency of credit risk evaluations for retail clients.

$H_{3:}$ The personalized and explainable responses of AI chatbots build client trust and acceptance of credit decisions.

$H_{4:}$ Models that incorporate behavioral data outperform models that rely solely on financial ratios to predict credit risk.

$H_{5:}$ The ability to automate credit risk assessments through chatbots creates cost savings and reduced processing times when assessing retail credit.

To validate the proposed hypotheses, a structured methodological approach is adopted.

2.3. Data Understanding

To address the problem at hand, I chose to take a supervised machine learning approach. I focused on two powerful ensemble algorithms, Random Forest and XGBoost, which are well known for their high performance in classification tasks and their ability to handle complex nonlinear relationships in data. I aimed to select models applicable within the financial sector, particularly when data is imbalanced or noisy.

2.3.1. Data Description

The dataset used in this study is extracted from Tunisian bank's files, specifically from client files contained in the Tunisian bank's credit portfolio for the year 2024. These files are handled by the Tunisian bank's Individual and Professional Division (IPD). The dataset comprises 3,080 observations and 18 predictive variables/examples.

The predictive variables are structured and categorical variables that characterize the actions of an applicant and the nature of a loan, as well as assessing the creditworthiness of the applicant.

The final decision outcome is denoted as AVISFINAL and is categorical. Using AVISFINAL, there are two unique values: reject for declined applications and accorded for accepted applications. For modeling, AVISFINAL will be considered a binary classification task where reject is coded as class 1 (high-risk or default-prone) and accorded is coded as class 0 (low-risk/accepted). The binary target serves as the dependent variable for training machine learning models to predict credit acceptance decisions with high discrimination ability.

The data was anonymized to protect privacy, pursuant to data policies.

Table 1.
Feature set with types and descriptions.

Feature Name	Type	Description
NUM-CANVAS-CNV	Categorical	Unique identifier for the customer file.
TYPE-CAN-CNV	Categorical	Type or category of the customer file.
DATECREATION	Categorical	Account creation date.
AVISFINL	Categorical	Final opinion or decision regarding the customer.
AVISPC	Categorical	Preliminary opinion or credit score check result.
NUM-DEC-DEC	Categorical	Unique identifier of the credit decision.
INTERDITCHEQUIER	Categorical	Indicator of whether the customer is banned from issuing checks.
NUM-CPTCLT-CNV	Categorical	Account number linked to the customer file.
LIB-GARPR-CNV	Categorical	Type of guarantee or collateral provided.
AGE	Numeric	Age of the customer.
GENRE	Categorical	Gender of the customer.
PROFESSION	Categorical	Professional occupation of the customer.
ANCIENNETE	Numeric	Account seniority in months.
CLASSE	Numeric	Risk classification score of the customer.
MONT-PRIN-DEC	Numeric	Principal loan amount requested.
NUM-NBECH-DEC	Numeric	Number of scheduled payments.
NUM-DUR-DEC	Numeric	Loan duration, adjusted based on the number of payments.
TAUX	Numeric	Interest rate applied to the loan.

2.3.2. Software Used

This article utilized Python 3.x within a Jupyter Notebook (Anaconda) environment for development. Data processing was facilitated by NumPy for numerical operations and Pandas for data manipulation, while predictive modeling employed scikit-learn's Random Forest algorithm and XGBoost. Model persistence was achieved using Joblib. Typical visualizations for the analysis were created with Matplotlib and Seaborn, and SHAP provided explanations for the classification of label classes. Finally, the Streamlit chatbot interface enabled interactive conversation management, real-time prediction displays, and dynamic processing of user inputs.

2.4. Data Preparation

2.4.1. Data Processing

Missing data accounted for approximately 8.7% of the entire dataset. A thorough data cleaning procedure was executed to fill these missing values and improve data quality. Categorical features with missing data were imputed using the mode (most frequent category) or another placeholder such as 'unknown'. Missing data in numeric features were imputed using medians (midpoint) to increase robustness to outliers. This process resulted in a fully populated dataset and improved the completeness and reliability of the dataset for modeling 0.8. See (Appendix 1)

Mistakes in categorical features such as PROFESSION were given the value of 'Unknown' for any missing entries, and GENRE was assigned its most frequent category (mode).

Numeric features such as MONT-PRIN-DEC and NUM-NBECH-DEC were replaced with their median values to enhance robustness against outliers.

NUM-DUR-DEC was recalculated as NUM-NBECH-DEC + 1 to correct missing or invalid entries.

For the AGE variable, any outliers greater than 74 years were replaced with the median ages to maintain the consistency of the data. 0.9 See (Appendix 2)

At this stage, I also performed a feature selection step to remove features identified as irrelevant or redundant. As a precaution, I removed the features LIB-GARPR-CNV, NUM-CPTCLT-CNV, TAUX, and AVISPC because they were either unnecessary for the scoring purpose or produced poor predictive power. This cleaning step helped to ensure that the final dataset was homogenized, noise-free, and ready for feature engineering and modeling stages. See (Appendix 4).

2.4.1.1. Feature Engineering

Feature engineering was critical in boosting our models' performance through the ability to transform raw data into features relevant to the task. The following steps were followed in this paper:

2.4.1.2. Feature Transformation and Creation

Table 2.
Feature Transformation and Creation.

Feature	Formula	Explanation
AGE_x_ANCIENNETE	AGE × ANCIENNETE	Age multiplied by job seniority (in months).
MONT_x_DUREE	MONT_PRIN_DEC × NUM_DUR_DEC	Loan principal multiplied by the loan duration (months).
PRESSION_REM	$\frac{\text{MONT\_PRIN\_DEC}}{\text{ANCIENNETE} \times 1000} / \text{NUM\_DUR\_DEC}$	Monthly installment pressure compared to estimated income.

0.10. See (Appendix 3).

It is important to highlight that the variable income is not a variable in the data set for this article. This is because the bank does not consider income as a fixed or standard variable in the database.

Also, income is different than most socio-demographics in that it is not dependent on any verification period, if at all. Income can change from year to year, and there is little feedback on how the income is declared. The client's income is often self-declared, with sometimes little follow-up from the bank.

When it comes to credit scoring models, credit managers have traditionally regarded income as a static value. However, this fixed income measure does not accurately reflect a client's true creditworthiness or financial behavior. To address the lack of a defined income variable and to improve the behavioral component of the model, three synergistic variables were introduced:

AGE-x-ANCIENNETÉ: Age * work defined as seniority (in months) - these variable captures both age and professional stability.

MONT-x-DUREE: The loan principal value * the duration of the loan in months - this reflects the client's engagement with credit.

PRESSION-REM: The MONT (loan principal) / the DUREE (loan duration) - estimating the monthly financial pressure on the client.

Although I consider these variables to be inspired by income, the premise of the model takes a more realistic and practical approach to model risk without relying on outdated and unreliable income data.

2.4.1.3. Encoding and Scaling Methods

Before using the machine learning algorithms, we preprocessed the data in two major ways: using Label Encoding for categorical variables and Standard Scaling for numerical features.

2.4.1.3.1. Label Encoding

We transformed all categorical features (GENRE, PROFESSION, TYPE-CAN-CNV) into numeric format using scikit-learn's LabelEncoder. LabelEncoding works by assigning each category an integer label like zero for Female 1, Male 2 or Other, or simply with no mathematical transformation, it can be thought of as:

$\text{Category}_i \rightarrow \text{Integer}_i$

It does not change categories; it only makes them more suitable for machine learning models.

2.4.1.3.2. Standard Scaling

The numerical features (AGE, ANCIENNETE, MONT-PRIN-DEC) were standardized using methods from StandardScaler. This means we rescaled these attributes to have a mean of zero and a standard deviation of one. This is particularly important since many models depend on the magnitude of a feature to learn or make predictions. The formula used for standard scaling is:

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (1)$$

where:

x: is the original feature value μ: is the mean of the feature, σ: is the standard deviation.

When we normalize all features, we can be sure that all features will contribute equally during the learning process, thus adhering to the principle of fair, improved model performance.

It does not change categories; it only makes them more suitable for machine learning models.

2.4.2. Feature Selection and Statistical Validation

2.4.2.1. Information Value (IV) and P-Value Test

To assess the predictive power and statistical significance of each feature, we calculated two numbers:

Information Value: This is an in-depth measure of the ability of the feature to distinguish between good and bad credit.

P-Value: This tests the statistical significance of each feature in relation to the target variable. The Information Value (IV) is calculated as follows:

$$\mathrm{WOE}_i = \ln\left(\frac{P_{good,i}}{P_{bad,i}}\right) (2)$$

where:

- WOEi is the Weight of Evidence for category i:

$$\mathrm{IV} = \sum_{i=1}^{n}(\mathrm{WOE_i} \times (\mathrm{P_{good,i}} - \mathrm{P_{bad,i}})) (3)$$

- $P_{\mathrm{good},i}$ and $P_{\mathrm{bad},i}$ are the proportions of "good" and "bad" clients in category *i*:

$$p_{good,i} = \frac{Number\ of\ good\ clients_i}{total\ good\ clients} \quad P_{bad,i} = \frac{Number\ of\ bad\ clients_i}{total\ \ bad\ clients}$$

2.4.2.1.1. Findings

Table 3.
Feature Importance Analysis by Information Value (IV) and Statistical Significance.

Feature	IV	p-Value
PRESSION_REM	1.683	0.015
MONT_x_DUREE	1.661	<0.001
MONT_PRIN_DEC	1.549	<0.001
JOURS_DEPUIS_CREATION	0.420	0.414
AGE_x_ANCIENNETE	0.400	0.688
PROFESSION	0.308	0.012
TYPE_CAN_CNV	0.122	<0.001
AGE	0.086	0.037
ANCIENNETE	0.072	0.716
NUM_NBECH_DEC	0.059	<0.001
NUM_DUR_DEC	0.059	<0.001
GENRE	0.011	0.009
CLASSE	0.006	<0.001

- The features PRESSION-REM, MONT-x-DUREE, MONT-PRIN-DEC reveal a strong predictive value (IV > 1).
- The features PROFESSION and TYPE-CAN-CNV have statistical significance (P-Value < 0.05).
- The features which had an IV = 0 and P-Value > 0.05, such as INTERDITCHEQUIER, were discarded.

Information Value (IV) analysis of variables showed that behavioural variables relating to household financial management – PRESSION-REM and MONT-x-DUREE – were the strongest outcome contributors and the strongest predictors of model performance.

The inference gained from these results is unequivocal regarding the value of a behavioral model in examining actual financial behaviors, rather than relying solely on a sociodemographic model. Information value analysis (IV) has already demonstrated that focusing exclusively on variables that measure actual financial activities will produce the best outcomes. Therefore, variables such as PRESSION-REM and MONT-x-DUREE are classified as behavioral because they exemplify how clients conduct their borrowings and repayments. Behavioral variables operate in dynamic conditions and relate to actual behaviors, whereas sociodemographic variables are static and pertain to fixed characteristics like age and gender. Identifying behavioral variables is crucial because they tend to be more informative about credit risk.

Table 4.
Classification of Features: Behavioral, Non-Behavioral, and Mixed.

Behavioral Features	Non-Behavioral Features	Mixed (Semi-Behavioral) Features
PRESSION_REM MONT_x_DUREE GENRE JOURS_DEPUIS_CREATIO N ANCIENNETE	MONT_PRIN_DE C PROFESSION TYPE_CAN_CNV AGE NUM_NBECH_DE C NUM_DUR_DEC	CLASSE AGE_x_ANCIENNETE

2.4.2.2. Correlation Matrix Examination and Non-Linearity

A correlation matrix was calculated to measure multicollinearity and evaluate the linearity of features. The Pearson Correlation Coefficient between variables X and Y is given by:

$$\Gamma_{X,Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\times\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (4)$$

where:

- Xi, Yi are individual sample points
- $\bar{X}$, $\bar{Y}$ are the sample means
- n is the number of observations.

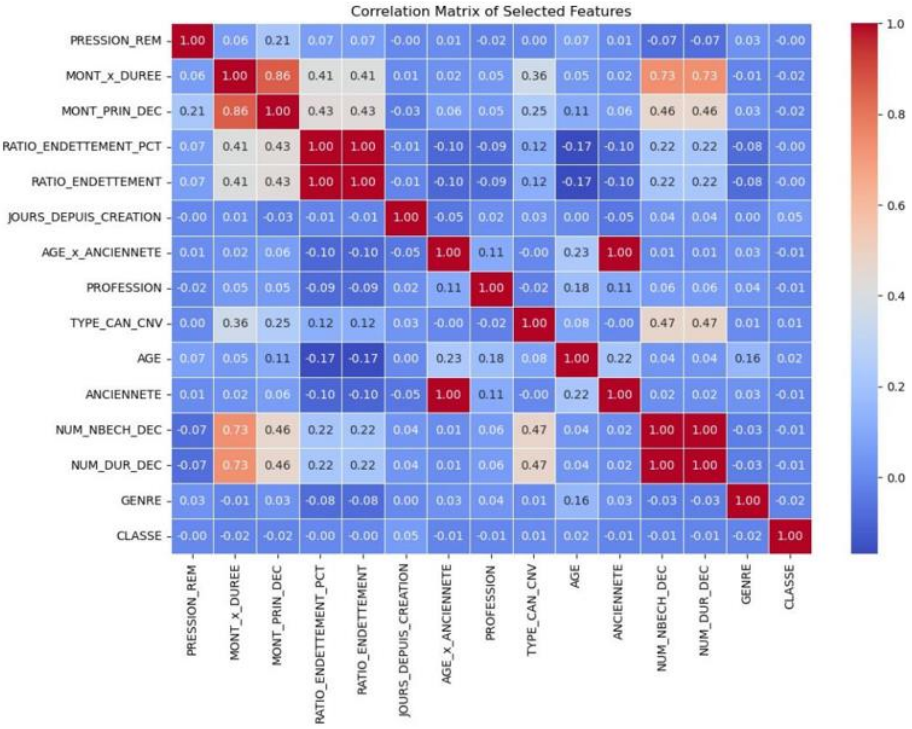

Figure 1.
Correlation Matrix of Features.

2.4.2.2.1. Findings

MONT-PRIN-DEC and MONT-x-DUREE are highly correlated (r = 0.86), indicating redundancy.

The other features are moderately to weakly correlated, meaning they are contributing uniquely. Non-linearity observations:

Although some features were correlated, the overall correlation values indicate that there are non-linear relationships among the features. Tree-based models such as XGBoost and Random Forest are better suited to capture these types of interactions compared to linear models, which assume additive effects. The non-linearity (or overall non-additive nature) of the relationships within the sample suggests that tree methods are more appropriate.

2.5. Modeling

2.5.1. Data Splitting Strategy

2.5.1.1. Target Variable Selection

In the modeling process, the selected target variable is AVISFINL, which indicates the final credit decision (i.e., whether a customer's request was approved (good risk) or rejected (bad risk)). This variable reflects the institution's real-world credit decision process; therefore, it is relevant to the predictive modeling task.

By being a binary variable, AVISFINL sets the task for supervised classification, which utilizes powerful algorithms such as Random Forest and XGBoost. These models enable us to capture complex and non-linear relationships between features such as income, behavior, and debt levels that inform credit risks.

Once AVISFINL was selected, the data was then split into training and testing sets. This ensures that models are built on one part of the data while they are evaluated using previously unseen data, which helps avoid biases and increases the ability to detect overfitting. This step enhances the models' reliability when implementing them in real-world credit processes.

2.5.1.2. Data Split: Training and Testing Sets

In this research, the dataset consisting of 3,080 records was divided into 80% training (2,464 samples) and 20% testing (616 samples). This partition allows models to have enough data to learn from while reserving a portion of the dataset to

evaluate the model's performance reliably. A stratified split of the data was used to maintain the balance of accepted and rejected credit decisions based on the target variable AVISFINL.

The training set is used to fit and tune the models, while the testing set is used to evaluate model performance on previously unseen cases. This separation of training and testing datasets is essential to avoid overfitting and to assess whether models generalize well to real-world applications in credit. (Appendix 2)

2.5.2. Machine Learning Algorithms

Machine learning (ML) is a subfield of artificial intelligence that assists systems in learning from data and improving performance over time without being explicitly programmed. In credit scoring applications, ML algorithms provide an effective methodology for analyzing large, complex datasets. This process involves identifying patterns and relationships, especially in cases where traditional statistical methods may not detect these outcomes, such as correlations among variables in multi-collinear datasets. Supervised learning, a key component of machine learning, enables the model to learn the mapping from input features to the outcome variable by training on labeled data, where the outcomes are known. This approach is used for classification tasks, such as predicting default or non-default outcomes to assess default likelihoods.

I have chosen Random Forest and XGBoost models for this article because they are both good candidates for supervised learning. They have the added benefit of gleaning useful correlations and relationships between features, regardless of the non-linear nature of these relationships, which is common in financial and behavioral data. Both RF and XGBoost can be used for the same classification tasks mentioned above.

The following sections will present the mathematical formulations for these two models.

2.5.2.1. Random Forest

Random Forest is an ensemble learning algorithm that trains many decision trees during the training process and takes the mode of their predictions (classification). The use of randomness in its tree construction reduces variance and overfitting.

Important Concepts: Bootstrap Aggregation (Bagging): Each tree is trained on a bootstrap sample (sample with replacement from the original dataset).

Feature Randomness: Each time a tree is split, only a random subset of features is considered. This reduces correlation among the trees and improves generalization. For classification, the Random Forest prediction $\hat{\mathbf{y}}$ is given by majority vote:

$$\hat{y} = \text{mode}\ T_1(X), T_2(X), \ldots, T_m(X) \quad (5)$$

- $T_i(\mathbf{X})$ is the prediction from the i-th decision tree
- m is the total number of trees in the forest
- $\mathbf{X}$ represents the input feature vector
- mode selects the most frequent class prediction

How it works (step by step): 1. Take multiple bootstrapped samples from the training data. 2. For each sample, create a decision tree using a random subset of features at each split. 3. Aggregate the predictions of all trees using either majority voting (when classification) or averaging (when regression).

2.5.2.2. XGBoost (Extreme Gradient Boosting)

XGBoost is a boosting algorithm that builds trees through a sequential process. Each new tree attempts to reduce the errors (residuals) of the combined ensemble of trees by optimizing the objective function, which includes regularization.

Key Concepts:

- Boosting: Essentially, trees are added, one at a time, and each tree takes into account the errors of the previous trees.
- Regularization, XGBoost considers a number of terms that penalize model complexity in order to reduce the amount of overfitting that will occur in the model.
- Gradient descent: The process of optimization is carried out using the gradients of the loss function.

The objective function at iteration t is:

$$\text{obj}(t) = \sum_{i=1}^{n}[L(l_{y_i}, y_i^{(t-1)} + f_t(X_i))^2] + \Omega(f_t) \quad (6)$$

The term (1/2) * y_i * ŷ_i^(t-1) represents a second-order approximation used in gradient boosting.

The model prediction after K trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i) \quad (7)$$

$\hat{y}_i$: The predicted value for the i-th data point.

K: The total number of trees (or models) in the ensemble.

$f_k(x_i)$: The prediction from the k-th tree (or model) for the input feature vector x_i.

x_i: The input feature vector for the i-th data point.

The gradient boosting algorithm proceeds as follows:
1. Initial Prediction:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (8)$$

where p is the class proportion.

2. Compute Residuals:

$$r_i^{(t-1)} = -\frac{\partial L(y_i - \hat{y}_i)}{\partial \hat{y}_i}\Big|_{\hat{y}_i = \hat{y}_i{}_i^{(t-1)}} \quad (9)$$

for $i = 1, \ldots, n$ and loss function
3. Fit Tree to Residuals

$$f_t = arg\min_f \sum_{i=1}^{n}[r_i^{(t-1)} - f(x_i)]^2 \quad (10)$$

4. Update Predictions:

$$y_i(t) = y_i(t-1) + f_t(X_i) \quad (11)$$

where η is the learning rate ($0 < \eta \leq 1$).
5Repeat steps 2-4 for $t = 1, \ldots, K$ trees.

$$F_k(X) = y(0) + \sum_{t=1}^{T} f_t(x) \quad (12)$$

Final Model:

2.5.3. Overfitting Mitigation Techniques

Overfitting occurs when a model has learned the underlying patterns in the training data along with the noise, resulting in a model that has no generalization and poor performance with new data.

Overfitting usually results in accuracy being high on the training data but performance is low on unseen data.

Here are some common techniques to reduce or prevent overfitting in machine learning

2.5.3.1. Regularization Strategy Across Models

To guarantee strong performance and to avoid overfitting, regularization procedures were consistently applied in both the classifier models. Both Random Forest and XGBoost offer similar regularization mechanisms; however, the models were tuned separately for their respective prediction tasks (binary classification).

For Regularization for Random Forest Classifier, over-fitting was avoided through:

- Max-depth: Restricting the depth of each tree to prevent the model from becoming too complex.
- N-estimators: The number of trees in the ensemble can be controlled, providing a balance of bias and variance.
- Max-features: Restricting the number of features considered at each split adds randomness and reduces three correlations.
- Bootstrapping sampling: The concept of bagging applies to training each tree on a random subset of the data reducing variance.
- For XGBoost Classifier, the regularization included the following: L1 (alpha) and L2 (lambda) penalties: The penalties provide a direct penalty of model complexity in the objective function, helping the creation of simpler model.
- Max-depth and min-child-weight: Control growth of trees to limit fitting small patterns (over-fitting).
- Subsample and colsample-bytree: Add randomness in both the samples of data and all the sampled features to enhance generalization.
- Early stopping: Stop training when validation performance stops improving and model is saturating on the training dataset adding limited additional value.

While the same regularization techniques were applied, the hyperparameters were tuned separately for classification to ensure that each model was as optimized as possible, with respect to its intended purpose

This approach to regularization was unified, yet task-appropriate and allowed us to build models that were fully accurate, generalizable, and that contributed to sound credit decision-making and accurate assessments of risk.

2.6. Evaluation Metrics

In this section, the results of the classification model will be presented, along with suitable metrics used to evaluate the accuracy and reliability of the models employed to predict credit decisions and risk scores.

Classification Metrics:

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision:

$$\text{Precision} = \frac{TP}{FP + TP}$$

- Recall:

$$\text{Recall} = \frac{TP}{FN + TP}$$

- F1-score:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC AUC: Measures the area under the ROC curve, reflecting the trade-off between true positive rate and false positive rate.

Definitions:

- TP (True Positives): Correctly predicted approvals.
- TN (True Negatives): Correctly predicted rejections.
- FP (False Positives): Incorrectly predicted approvals.
- FN (False Negatives): Incorrectly predicted rejections.
- Classification Results: Random Forest and XGBoost classifiers were submitted to a comparative analysis using training and test datasets. Recall and ROC AUC measures were particularly useful in their detection of compliant applicants and signified when a classifier accurately detected high-risk applicants. A summary of the comparative results achieved by the models is illustrated.
- Cross-Validation Scores We performed k-fold cross-validation. The mean and standard deviation of the scores of all folds provided some sense that the models are stable and reasonable for generalizability.
- Overfitting analysis: Training and test scores are complete after the analysis, and an overfitting process to the training set is explored. Employing regularization, k-fold cross-validation, and early stopping/dropping regularization are examples of potential options to maintain normal performance on unseen data.

2.6. System Deployment and Architecture of the Credit Scoring Chatbot

2.6.1. AI Chatbot for Credit Risk prediction

This part elaborates on an AI-enabled chatbot that performs credit risk assessment in banking, integrating machine learning-modeled predictions of credit approval/rejection decisions, as well as risk scores as part of that decision-making process. The chatbot makes it easier for users to complete the task while providing information that banks and financial services could leverage.

2.6.1.1. Chatbots in Banking

When it comes to credit scoring and banking, defining the scope of chatbot deployment is an important strategic decision. There are typically two deployment options in this context: internal (i.e., by bank staff) or external (i.e., by clients). After consideration, the internal deployment is the scope chosen for this article, for the reasons outlined:
Internal Chatbot (Recommended Option) The chatbot is designed to support personnel such as credit agents, customer advisors, and risk analysts at the bank. This internal deployment allows the chatbot to:

- Allow access to sensitive and confidential information, such as:
  - Cheque book interdiction
  - Historical banking incidents
  - Internal risk classifications
- Data from central credit registries/sources (such as BCT reports)
- Assist the credit decision-making process by providing real-time scoring and risk estimating based on full client information.
- Enhance operating efficiencies by assisting the staff with data input and automated evaluations and thus fostering faster and more consistent credit decisions.

2.6.2. System Architecture and Design

The internal AI chatbot is created for use by bank staff, such as credit agents, to assist with credit risk evaluations in a more efficient and accurate manner. The bank staff will enter customer information into the chatbot interface; the chatbot can take care of the rest. It will use feature engineering and predictive models to process customer information and arrive at a credit risk score.

The chatbot will also provide explanations for each prediction to allow bank staff to understand the variables impacting their risk evaluations. Specific checks for data issues or anomalous predictions are included to promote reliable results.

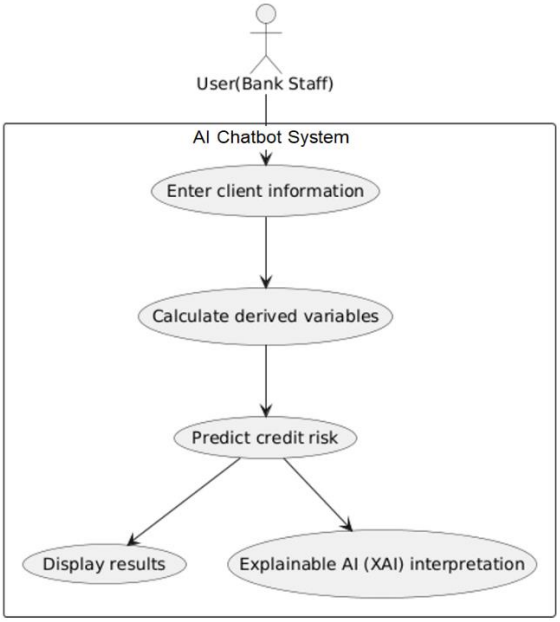

Figure 2.
Use Case Diagram of the AI Chatbot System.

As shown in Figure 2, the main actors interact with the chatbot to perform credit risk assessments.

The Use Case Diagram below demonstrates how bank staff interact with the chatbot system to conduct credit evaluations and review more in-depth explanations:

This internal tool will help bank staff make informed, transparent, data-driven lending decisions.

2.6.3. Process Flow of Credit Risk Evaluation

When the chatbot receives a prompt from either the bank staff or users, it follows a systematic set of steps to generate credit risk predictions. Specifically, the chatbot interface collects user inputs related to their financial histories and behaviors on the bank's platform. The user inputs are subsequently processed by a feature engineering module to create the features needed for the machine learning predictive model. The relevant machine learning model (e.g., Random Forest, XGBoost) processes the features describing the user to make an assessment of the credit risk. When the predictions are made, Explainable AI with Gemini models generates interpretable and understandable explanations about the model's decision-making process.

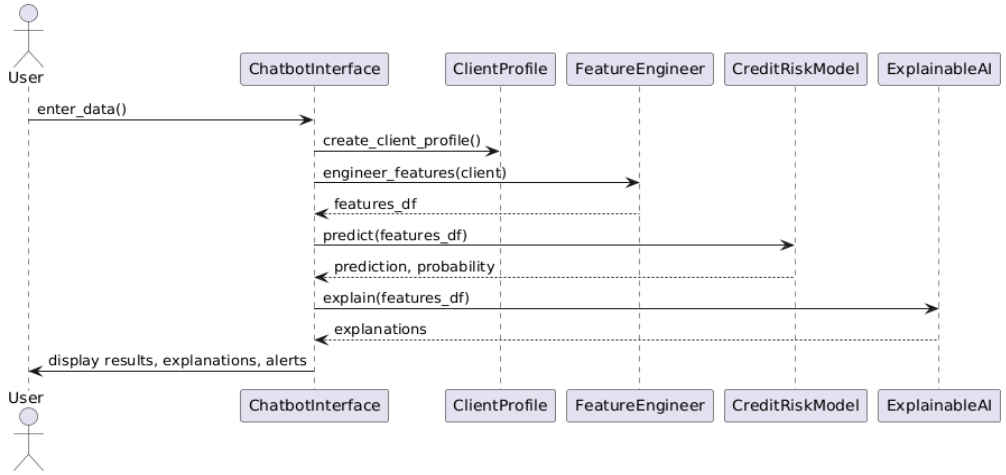

Figure 3.
Sequence Diagram of the AI Chatbot System.

The bot summarizes the risk class and explanation models again to the user through the chatbot interface, enabling empowered and transparent decision-making.

2.6.4. Internal Components and Structure

The chatbot is made up of various modules each playing a key role in returning accurate and explainable credit risk scores.

ClientProfile: This module collects the user's input, which includes financial and behavioral attributes such as age, MONT-PRIN-DEC, ANCIENNETÉ, and so on. It also validates and stores the input data for further processing.

FeatureEngineer: This module processes the raw client information in order to engineer features, which are generally

better suited for modeled prediction. This module performs any calculations, encodings, and normalization for all features to homogenize the specification of inputs.

CreditRiskModel: This is the machine learning model module (Random Forest, XGBoost) used to predict default based on the engineered features.

Explainable AI (Gemini): This module provides explainability by returning simple, clear explanations for the factors the model predicted, transparently showing which features best explain the default.

ChatbotInterface: This is the service facade layer used to connect to the user, manage user interaction, collect user inputs, display results, and present an explanation of what the results were.

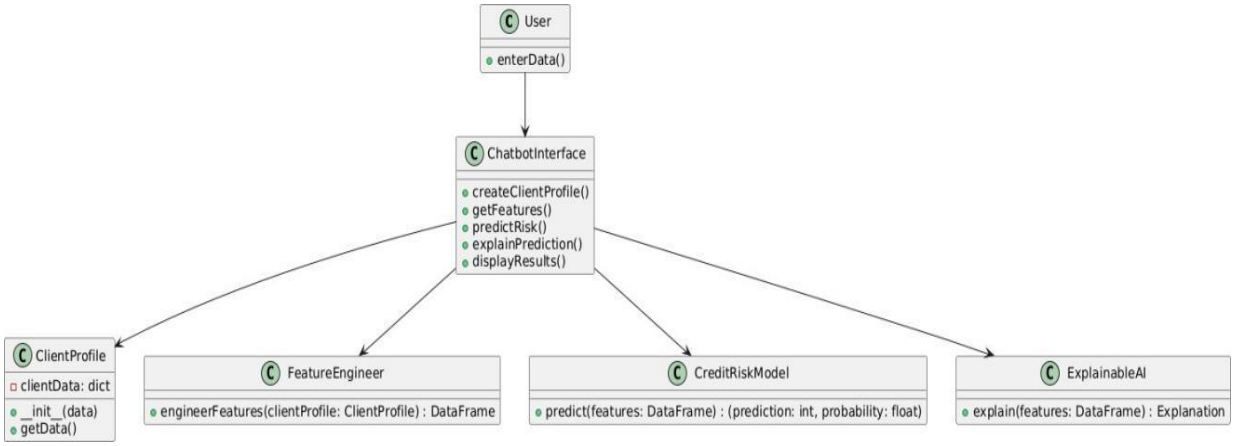

Figure 4.
Class Diagram of the AI Chatbot System.

The Class Diagram below shows the described modules along with their attributes, methods, and relationships:

2.6.4.1. Implementation Details and Tools

The chatbot has been built using Streamlit, a Python-based framework that facilitates the quick development of interactive web apps with minimal effort. Streamlit provides a clean and powerful interface for inputting data, obtaining predictions, and viewing results in real time.

Key libraries and tools were used, such as:

Python: The core programming language supporting the backend logic and model integration.

Scikit-learn and XGBoost: Machine learning libraries used for training and deploying the credit risk models.

Gemini Explainable AI API: Integrated to offer users transparent explanations of AI models' predictions, improving the user's trust and allowing for informed decisions based on the AI results.

Streamlit Python API: Facilitated the design of an intuitive chatbot interface to manage users, input forms, and dynamic display of results.

The interface is designed to be clean and intuitive for bank staff, guiding them through a 'step-by-step' process for data entry and displaying risk scores along with explanation reports. Streamlit creates a reactive interface that allows users to receive immediate feedback, making it practical for real-time decision-making in credit evaluations.

2.7. Conclusion

In this article, I have detailed the entire process of building the credit risk prediction system. Beginning with raw banking data, a number of steps were taken including cleaning, feature engineering and variable selection to prepare the data for modeling to ensure that the models could learn effectively from the data.

The machine learning phase showed us that both Random Forest and XGBoost algorithms were effective in predicting credit risk, along with careful tuning and evaluation in order to improve their performance and reduce errors.

Finally, the deployment of an AI-powered chatbot designed to support credit risk evaluation for bank staff. The system integrates data collection, feature engineering, predictive modeling, and Explainable AI through Gemini within a streamlined, interactive interface built on Streamlit.

Key advantages of this deployment include real-time, interactive risk predictions, transparent explanations that enhance user trust, and a user-friendly design that simplifies complex credit assessments

The following chapter will reveal the results of the models and provide detailed analysis of their performance.

3. Results and Performance Analysis

I have walked through the entire implementation process of building the credit risk prediction system. From raw banking data, several steps were taken, including data-cleanup, feature engineering, and variable selection, so that the data was in the most usable format intended for modelling and ensuring that the models could learn from the data effectively.

The machine learning portion demonstrated that both Random Forest and XGBoost algorithms were capable of predicting credit risk with careful tuning and evaluation to enhance performance and minimize errors.

Lastly, the solutions were embedded into an AI chatbot to utilize them live and openly with non-technical users as a viable tool to assess users' credit risk. The chatbot is better at simplifying and automating the decision-making process than human-based methods and improves the transparency of analyses by showing the key drivers behind each score.

The next chapter will present the results from the models and provide a detailed assessment of their performance.

3.1. Data Exploration

Having completed the data preparation in Chapter 2, we now present an exploration data analysis (EDA) to understand the characteristics of the data. The EDA involves examining descriptive statistics, the distribution of interest variables, and correlation analysis. This will provide an initial understanding of how the data is structured, in addition to confirming that the decisions made during the preprocessing steps are valid before modeling.

3.1.1. Descriptive Statistics

Provides the descriptive statistics of the dataset. As presented in the table, there are 3,079 observations. The average age of clients is approximately 47 years, with a standard deviation of 10 years, ranging from a minimum of 21 to a maximum of 74 years. The dataset is slightly more populated by men, as indicated by the mean of the GENRE variable (0.73).

The variable ANCIENNETE demonstrates very high variability in the sense that it had an average value of 23 months, indicating that there were clients with much longer tenures to credit.

Table 5.
Descriptive Statistics of the Dataset.

Variable	Count	Mean	Std. Dev.	Min.	Max.
NUM_CANVAS_CNV	3079	1539.00	888.98	0.00	3078.00
TYPE_CAN_CNV	3079	1.17	0.41	0.00	2.00
AVISFINL	3079	0.30	0.46	0.00	1.00
INTERDITCHEQUIER	3079	0.00	0.00	0.00	0.00
AGE	3079	46.95	10.15	21.00	74.00
GENRE	3079	0.73	0.44	0.00	1.00
PROFESSION	3079	125.59	80.75	0.00	260.00
ANCIENNETE	3079	23.17	108.78	0.00	35.00
CLASSE	3079	0.02	0.20	0.00	4.00
MONT_PRIN_DEC	3079	25,592,450	24,451,940	1,500,000	150,000,000
NUM_NBECH_DEC	3079	62.03	24.23	1.00	283.00
NUM_DUR_DEC	3079	63.03	24.23	2.00	284.00
PRESSION_REM	3079	470,629	982,555	33,898	100,000
AGE_x_ANCIENNETE	3079	1331.77	6795.10	0.00	65736.00
MONT_x_DUREE	3079	1,886,380,000	326,8205,000	4,800,000	594,000,000
JOURS_DEPUIS_CREATION	3079	301.73	104.00	112.00	476.00

The financial variables such as MONT-PRIN-DEC or MONT-x-DUREE are large-scale variables with a reasonable breadth of values and dispersions, indicative of the variety of loan amounts or the duration of maintained loan accounts that exist.

There is a maximum of 100 million for "PRESSION-REM," which indicates that some clients experience extreme financial pressure. In conclusion, there appears to be significant heterogeneity in the demographic and financial variables, which should be considered when modeling and conducting risk assessments.

3.1.2. Key Visuals for Credit Risk Analysis

Two plots were used to explore the relationship between key client financial and behavioral indicators and credit decision outcomes. Each visualization represents a specific hypothesis aligned with established best practices in credit scoring analysis.

3.1.2.1. Account Age vs Credit Decision (Grouped Histogram)

Purpose: Investigate whether newly opened accounts are more prone to rejection. Hypothesis: New clients with limited credit history face a higher risk of rejection.

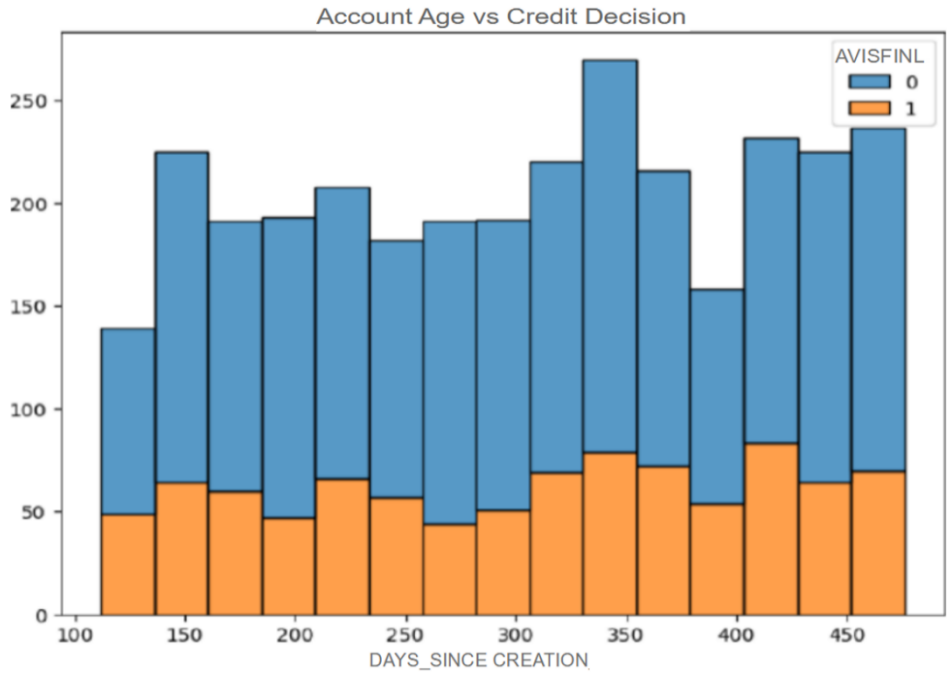

Figure 5.
Account Age vs Credit Decision.

Observation:

The histogram shows that accounts with fewer days since creation (younger accounts) have a higher proportion of rejections (AVISFINAL = 1). Specifically, the leftmost bars (representing newer accounts) are dominated by rejections, while approvals (AVISFINAL = 0) become more frequent as account age increase

The hypothesis is validated. Younger accounts are more likely to be rejected, likely due to insufficient credit history. This underscores the importance of client tenure in credit decision models.

3.1.2.2. Repayment Pressure vs Credit Decision (Boxplot)

Purpose: Validate whether high repayment pressure is associated with rejections. Hypothesis: Applications with repayment pressure (PRESSION-REM) beyond acceptable limits are likely to be rejected. Reason for choice: This metric reflects financial strain, a key indicator of default risk.

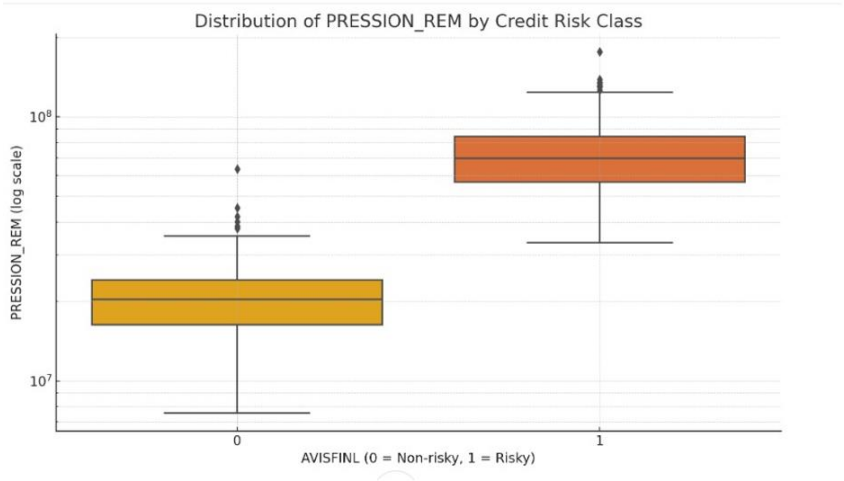

Figure 6.
Repayment Pressure vs Credit Decision.

Observation:

The boxplot reveals that rejected applications (class 1) tend to have significantly higher PRESSION-REM values, with a median around 70 million, compared to accepted applications (class 0), centered around 20 million. This confirms the hypothesis that excessive repayment pressure is associated with credit rejections, validating PRESSION-REM as a discriminant feature and a strong indicator of default risk.

The hypothesis is supported. Excessive repayment pressure negatively impacts approval chances, confirming its utility as a risk metric. Recommendations:

3.2. Model Training

In this section, we will demonstrate the training of Random Forest and XGBoost models for two complementary tasks of binary classification. The completion of both modeling tasks will allow a fully evaluated assessment of credit risk, providing insight into both the binary acceptance/rejection decision.

3.2.1. Random Forest Classifier

The Random Forest Classifier was trained to predict the binary response (to credit applications). The target variable AVISFINL, in our example, means credit decision with 0 by denoting Rejected and 1 denoting Accepted.

Training: The model was trained on a subset of the features selected for the analysis, which include cornerstone financials (like MONT-PRIN-DEC) and behavioral (like PRESSION-REM) information about applicants. Random Forest has a mechanism for creating multiple decision trees based on random subsets of the data samples and features! Random Forest takes a majority vote among all of the trees when predicting. Aggregating large results of multiple decision trees is a powerful technique to minimize overfitting.

Classification:
High Risk: The model classifies all applicants as high-risk class at rejection (Class 1).
Low Risk: The model classifies all applicants as low risk except (Class 0).

3.2.2. XGBoost Classifier

The XGBoost Classifier was also fitted to predict the outcomes of credit applications using the same target variable (AVISFINL). The XGBoost model uses gradient boosting to train a strong model through boosting that improves on the errors made by its predecessor.

Training Procedure: Generally, trees within the XGBoost ensemble are built sequentially, where each model attempts to minimize/table the error of the previous (often under-the-reather) tree. The learning process is accepted as a gradient descent algorithm, which minimizes the prediction error.

The benefit of this approach is that XGBoost can fit very complex patterns in the data while being a parametric model. The optimal values for hyperparameters are learning rate, maximum depth of tree, and a number of trees which are sought to provide the best accuracy without overfitting the training data.

Classification:

High Risk (reject): Clients identified as high risk clients are classified as rejected (class 1).

Low Risk (accept): Clients classed as low risk client are classified as accepted (class 0).

3.3. Model Performance Analysis

The performance of the developed credit risk prediction models, XGBoost and Random Forest, demonstrates strong predictive capabilities and excellent generalization.

Table 6.
Performance Metrics of XGBoost and Random Forest Models on Credit Risk Dataset

Metric	XGBoost	Random Forest
Train Accuracy	0.918	0.904
Test Accuracy	0.900	0.897
Overall Accuracy	0.900	0.897
Precision (Class 0)	1.00	1.00
Recall (Class 0)	0.85	0.85
F1-score (Class 0)	0.92	0.92
Precision (Class 1)	0.77	0.76
Recall (Class 1)	1.00	1.00
F1-score (Class 1)	0.87	0.87
Macro Average F1-score	0.89	0.89
Weighted Average F1-score	0.90	0.90
ROC AUC Score	0.934	0.936
Confusion Matrix (Test)	TN=353, FP=61	TN=350, FP=64
	FN=1, TP=205	FN=0, TP=206

Both the XGBoost and Random Forest models exhibit excellent predictive performance on credit risk, with test accuracies approaching 90%. Although the XGBoost model has slightly higher training accuracy and ROC AUC than Random Forest, this indicates a better fit to the data. The similar performance across test set results suggests that no overfitting has occurred. Both models also demonstrate perfect recall on the high-risk class (Class 1), effectively identifying the majority of defaulters.

Precision and recall values for each class show a balance which demonstrates strong and reliable predictions. Overall, while both models are good, the XGBoost model provides slightly better predictiveness and generalization without sacrificing

the stability of the model.

3.3.1. XGBoost Comparison and Explanation

Our XGBoost model achieves much better overall results than those of Al Shiam et al. [5], with an accuracy of 90% compared to 74.55% and a precision of about 77% compared to 22.03%. Our model recall is nearly 100% and an AUC of more than 93%, which shows better discriminatory power as well. These significant improvements are chiefly the result of using a much more comprehensive set of behavioral features in addition to financial data, which are in effect better capturing the complex and dynamic activities associated with credit risk. We also applied careful feature engineering, performed hyperparameter optimization, and leveraged explainability options offered through some advanced features of the Gemini framework, all of which add to the ruggedness and explainability of the XGBoost model. These factors working together help to render the model robust for generalization despite the challenges caused by it being on imbalanced classes.

Model Interpretation (with SHAP Analysis) Feature Importance via SHAP To interpret the internal decision logic of the model, we used SHAP (SHapley Additive exPlanations) values. The SHAPsummary plot (see Figure 7) reveals the top features influencing the model's predictions, both in direction and magnitude. we created a more trustworthy credit risk profile, effectively supporting decision-making in a banking context.

Like the XGBoost model, the SHAP summary plot for the Random Forest model also confirms that behavioral variables strongly dominate credit risk predictions. Most notably, PRESSION-REM, MONT-x-DUREE, and MONT-PRIN-DEC were the three main drivers. All these features produced negative impacts on the model output, meaning that a positive increase in their values decreases the probability of acceptance. Each of these behavioral variables is an important reflection of the applicant's weight and ability to repay, which is a core risk insight. It is undeniable that additional behavioral features like NUM-DUR-DEC, NUM-NBECH-DEC, and JOURS-DEPUIS-CREATION further add weight to our understanding of borrowing behavior and emphasize recent activity. In contrast, demographic variables like GENRE, PROFESSION, and TYPE-CAN-CNV had little to no impact on the Random Forest model predictions.

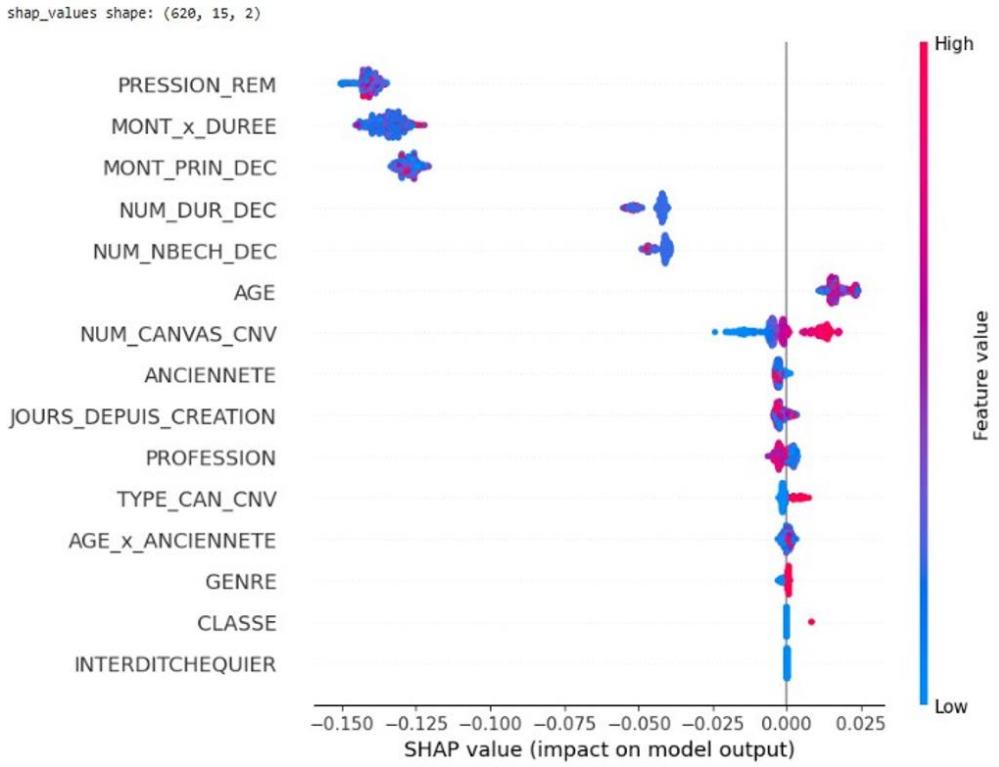

Figure 7.
SHAP summary plot Random Forest model.

The results from both models show that the best predictive model includes and ultimately identifies the strongest data-driven financial behaviors rather than personal characteristics, which allows for a more defined and explainable credit scoring decision.

3.3.2. Robustness and Overfitting Diagnostics

Both the XGBoost and Random Forest models show a solid and consistent fit to the training and test data without signs of significant overfitting. The Random Forest model achieved a training accuracy score of 90.35% and a testing accuracy score of 89.67%. Both the training and testing AUC were also exactly the same for Random Forest, with a score of 0.9363, which indicates outstanding generalizability and a low risk of overfitting in this model.

Similarly, the XGBoost model achieved a training accuracy of 91.76% and a testing accuracy of 90.00%, while the training AUC was 0.9425 and the testing AUC was 0.9338. Although the slight differences in the AUC indicate a small amount of

overfitting, it does not appear to be a major concern and does not affect the predictive accuracy of the diagnostic performance of this model.
Overall, both models have consistent fits, and Random Forest fits almost perfectly, but XGBoost has slightly greater predictive accuracy with very little overfitting.

3.3.3. Final Model Selection and Added Value
During the evaluation phase, both Random Forest and XGBoost model learning algorithms exhibited reliable performance in predicting credit risk. Both models achieved an overall accuracy of nearly 90%, demonstrated strong recall for the high-risk class (class 1), and showed balanced F1-scores. However, upon further inspection and comparison, there were instances in which XGBoost outperformed the Random Forest model.

- Better AUC (0.933 vs 0.93), which ensures the model's classification discrimination ability for positive and negative classes.
- Slightly better F1-score and precision, especially on the minority.
- Better generalization: smaller difference between training and test score.

For these reasons, as well as the interpretability, consistency, and availability of additional details, XGBoost was selected as the final model to serve as the predictive engine in the chatbot applying a credit risk prediction.
Business relevance and added value.
Selecting XGBoost as the predictive engine has substantial advantages for practical business utility and applicability to the credit risk model, including:

- Better risk identification: Almost perfect recall ensures that almost all high-risk clients will be properly classified as high risk, exposing the institution to negligible risk.
- Quicker and more informed decisions: Integrating XGBoost into the chatbot ensures that with each input, the system can continuously and immediately evaluate the clients' risk profile and exposure in real-time, creating operational efficiency.

3.4. Design, Integration, and Validation of the Credit Risk Chatbot
3.4.1. Design Description

This credit risk evaluation chatbot has a user interface that presents a clean and utilitarian form-like layout with clear separation of loan details, client details, and risk assessment details. The top of the interface includes key parameters of the loan, it then lists the client details, followed by financial summaries that show total repayment and social contributions. The layout is clean and has a modern design that prioritizes functionality over aesthetics, relying on bolded text and dividers to represent structure. The indicators for "Risk" could be more usable if they included an icon that indicated the levels of risk (e.g., red=high risk) or clearly defined options for the terms used.

Overall, the interface provides a good presentation of many types of important data, but even novice users would likely benefit from using interactive tooltips to explain technical terms.

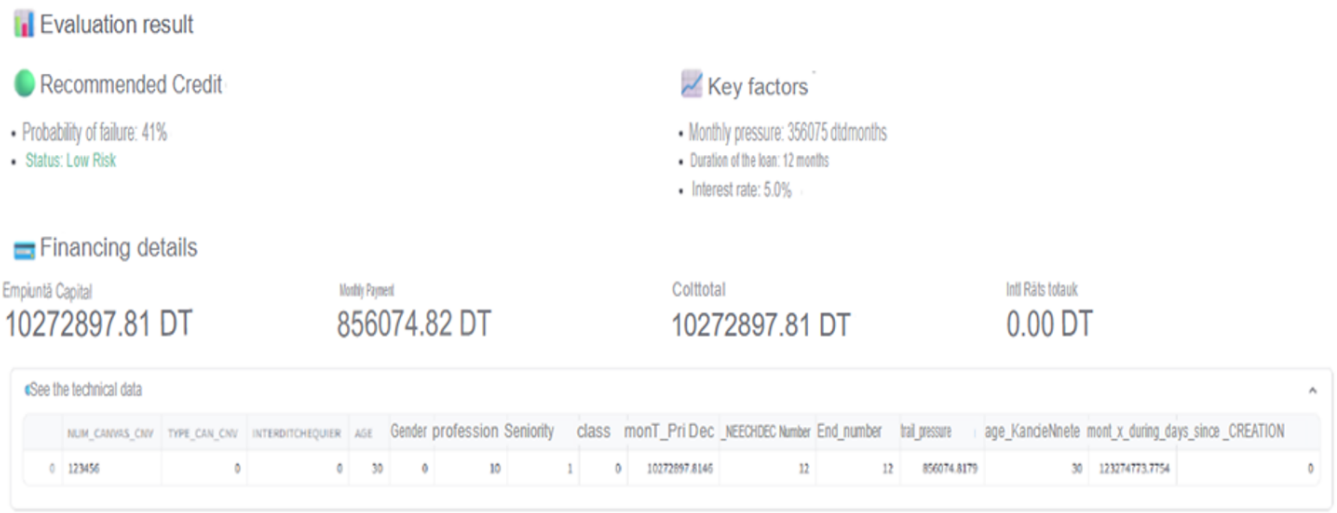

Figure 8.
Chatbot interface.

3.4.2. Integration of XGBoost Model into the Internal AI Chatbot

This section explains the incorporation of a credit risk prediction model within an AI chatbot intended for internal use by banks. Using the Streamlit platform, the chatbot features a user-friendly interface to assist with loan evaluations and decisions.

When a credit agent inputs the customer data, the data is automatically routed through a function engineering module in the background. This module re-organizes and enriches the raw inputs into the appropriate format for the model. The processed information is then passed into our pre-trained machine learning models, such as XGBoost or Random Forest,

which return a credit risk prediction (i.e., low or high risk).

The system is designed to be responsive and user-friendly, allowing credit officers to receive reliable results that are interpretable quickly. Importantly, as an internal-use system, it improves their day-to-day practice and confidence while keeping their risk assessment process data-driven and well-documented.

3.4.3. Functional Testing and Validation

This section provides a summary of the procedures and test observations using the AI chatbot integrated with the credit risk model.

3.4.3.1. Test Scenario Coverage

A range of test profiles was defined to imitate the reality of credit applications, including low, borderline, and high-risk applicants, in order to test the chatbot's robustness and adaptability through various conditions.

3.4.3.2. Prediction Dependability and Consistency

The model was found to be consistent in its outputs, which was validated against a wide array of inputs. Each test noted whether duplicate forms or slight modifications in the inputs produced the same result for both components of the chatbot as output.

3.4.3.3. Edge Cases and Anomaly Identification

Certain edge cases (extreme income values, conflicting financial behaviors, omission of optional fields) were used to test the system, including documenting the model's behavior when faced with uncertainty or overlapping risk patterns (e.g., borderline cases). In some instances, borderline decisions were made, which demonstrated lower confidence in the model.

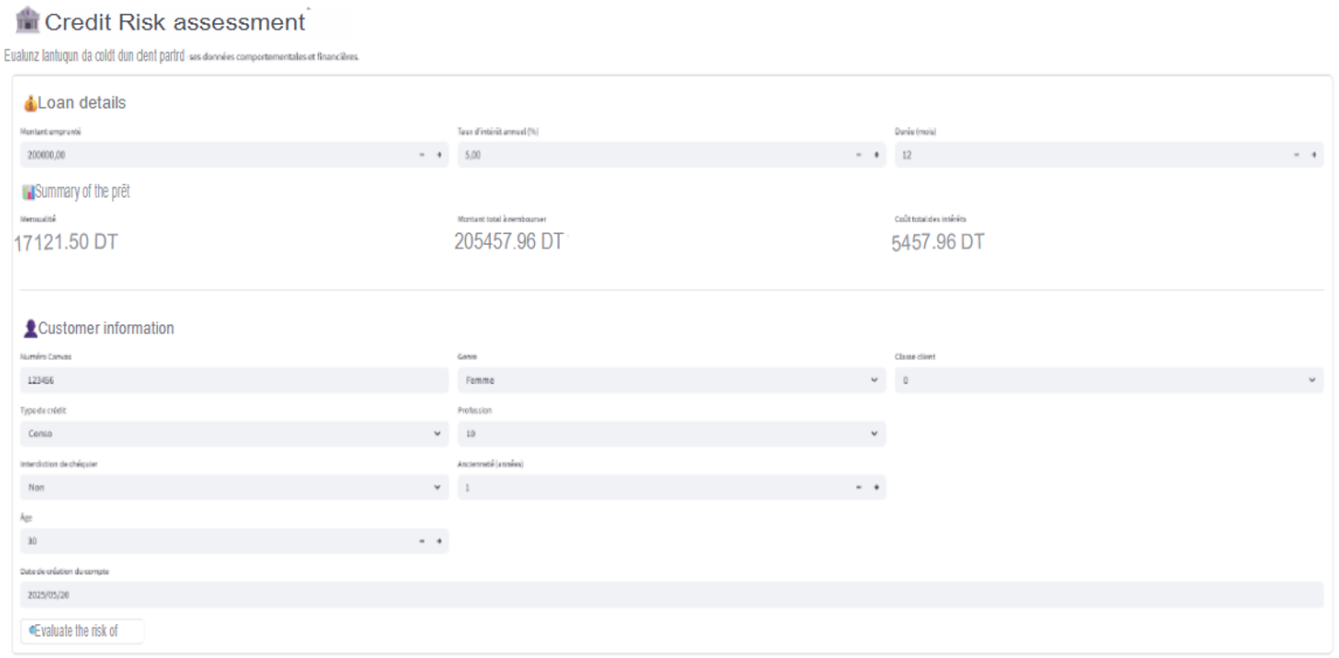

Figure 9.
Chatbot output.

3.5. Challenges and Limitations

This article faced several challenges and limitations, both technical and data-related, which are important to acknowledge for a fair evaluation of its scope and outcomes.

3.5.1. Data Limitations

One of the main limitations was the limited size of the dataset, which included 3,080 records, and this restricts the generalizability and stability of the machine learning models. Additionally, variables that are highly relevant for credit scoring, such as income and historical relationships with the bank (banking transactions), were missing, thereby limiting behavioral analysis and financial profiling. The target variable (loan decision: accepted or rejected) was also imbalanced, with a strong majority of accepted cases. This constituted a classification problem focused on identifying high-risk clients.

3.5.2. Technical Challenges

During model development, numerous technical concerns were considered. One of the most important concerns was overfitting, especially in early iterations of the model. While later iterations of the model avoided overfitting through hyperparameter tuning and cross-validation, the first few models overfitted substantially. Also, the dataset contained a mix

of numerical and categorical features, which meant sufficient preprocessing was needed. The feature "PROFESSION" posed an especially problematic issue as it included 250+ categories. To resolve the heterogeneity of the feature without inflating the dimensionality and introducing biases through encoding, an encoding strategy was developed.

3.5.3. Tool Limitations

The system uses Google's Gemini API to generate natural language explanations for credit decisions.

While effective, the API is currently limited in free usage, which may affect scalability or cost efficiency in a full deployment setting.

Despite these limitations, the system achieved strong performance and practical applicability, forming a solid base for further improvement and operational integration.

3.6. Perspectives and Future Enhancements

To extend the value of this work and integrate it into Tunisian bank's real-world credit risk assessment process, several future development perspectives are proposed:

- Deployment at Scale: The chatbot-based credit risk prediction tool can be deployed within Tunisian bank's internal credit evaluation platforms, allowing loan officers to instantly assess borrower risk with AI support. This not only enhances operational efficiency and decision-making speed but also ensures greater reliability and consistency in credit assessments.
- Wider Data Integration: Incorporating additional data sources such as customer transaction history, income records, and credit bureau scores (e.g., from national or private credit registries) will improve predictive power and make risk assessments more comprehensive.
- Open Banking APIs and Real-Time Insights: Leveraging Open Banking APIs could allow the chatbot to securely access real-time financial behavior, such as spending patterns, account balances, or payment history. This would enable a more dynamic and up-to-date evaluation of credit risk profiles.
- Explainability for Customers: The user interface could be enhanced to deliver more intuitive, personalized, and educational feedback to clients, helping them understand their credit rating and how to improve it, thereby promoting financial literacy.
- Information System Integration: Integrating and linking the chatbot with internal banking systems (e.g., customer relationship managers, external databases from credit bureaus, core banking software) would improve the chatbot's ability to retrieve corporate knowledge and store data as part of the credit assessment process. Explainability: The output of SHAP could be extended to include narrative summaries or simpler visualizations that target an audience of non-technical users.
- Multilingual Support: The chatbot could be modified to provide Arabic and French (in addition to English) language interface formats, which would enhance the usability for the diverse community of Tunisian banking professionals

3.7. Conclusion

In this article, we developed an empirical framework for credit risk analysis through systematic data exploration and key visualizations to understand the characteristics of data and relationships between key features. This foundation enabled modeling algorithm training in order of increasing performance levels, primarily XGBoost and Random Forest.

Accuracy training and test scores indicated a good, consistent fit to both training and test data with no notable signs of overfitting evident. The Random Forest model, in particular, achieved a training score accuracy of 90.35% and a test score accuracy of 89.67%, with training and test AUC scores at identical levels (0.9363), and neither model was made to develop overfitting. In this regard, Random Forest demonstrated excellent performance, and no other fitting models displayed better training and test scores. Random Forest presented a solid level of generalizability and suggested a quite low potential level of overfitted models.

Our formal Model Performance Analysis, Comparison, and Explanation, and Robustness and Overfitted Diagnostics confirmed good stability and predictability for the models we chose. Finally, the empirical findings being design, integration, and validation of the Credit Risk Chatbot mean we can translate them with any learning loss into a live, user-friendly credit risk assessment tool, as we approached credit risk analysis as a live, real-time, and data-driven activity.

4. General Conclusion

This report outlines substantial research and a proof-of-concept of a credit risk evaluation system with behavioral features and incorporating machine learning technologies in an AI-enabled chatbot. This article progressed from a thorough analysis of Tunisian bank BANK's context of operations, the types of credit products, and the challenges of credit risk management. This knowledge provided a baseline for many things, such as stating the problem clearly and defining a pragmatic, articleable approach to data.

The empirical framework utilized rigorous data exploration, thoughtful feature engineering, and reliable predictive modeling with the Random Forest and XGBoost models. Both models performed relatively well, achieving satisfactory accuracy over 89% and an AUC of 0.9363 from the Random Forest model; this indicates it is reliable and generalizes well. These results suggest that it is plausible to build upon, using machine learning, rather than traditional credit risk scoring processes, that can consider predictors of behavior and group behavior.

The last step of this article achieved its objective by delivering analytical insights in the form of a usable and interactive AI-enabled chatbot to assist credit analysts and decision-makers with risk assessments in real time. The chatbot enabled a drone review of credit applications that intensified intersection of data while enhancing interpretability and user experience, which is a significant change toward intelligent digital transformation in banking services.

This work represents a bridge between leading AI techniques and their applied use in the financial system. It illustrates the tremendous opportunity that exists in mining behavioral data and machine learning in a way that improves credit decisions, lowers exposure to risk, and supports financial inclusion. The tools and techniques advanced in this paper open the door to future improvement, such as ongoing learning, broader data integration, and automation.

This paper highlights that the use of AI-enabled methods to accomplish credit risk assessment, provided as an engaging chatbot, enables banks like Tunisian Bank to make more informed, accurate, and efficient lending decisions and also foster resilient consumer banking.

5. Future Research Directions

This study contributes to the increasingly large body of evidence on AI-based financial risk assessment. Recent publications by Gafsi and Louhichi [16] and Others have demonstrated the use of AI in the management of financial risk during the renewable energy transition, Gafsi and Louhichi [16] and the interplay among renewable energy, CO_2 emissions, and economic growth, Gafsi [17]. Other studies on agricultural finance, Gafsi [18] and Gafsi and Bakari [19], and digitalization of African development, Hlali and Gafsi [20], highlight the incorporation of behavioral, environmental, and macroeconomic factors into credit risk modeling approaches. These multi-factor data offer critical avenues for credit risk assessment enhancement using AI, particularly in emerging economies.

References

[1] N. Suhadolnik, J. Ueyama, and S. Da Silva, "Machine learning for enhanced credit risk assessment: An empirical approach," *Journal of Risk and Financial Management*, vol. 16, no. 12, p. 496, 2023. https://doi.org/10.3390/jrfm16120496

[2] Business News, "Ben moulehem: Tunisian bank has been resilient despite the economic situation," Business News, 2022.

[3] Anchala *et al.*, "Machine learning-based risk prediction model for loan applications: Enhancing decision-making and default prevention," *Journal of Business and Management Studies*, vol. 5, no. 6, pp. 160-176, 2023. https://doi.org/10.32996/jbms.2023.5.6.13

[4] D. Li, "Evaluating various machine learning techniques in credit risk area," *BCP Business & Management*, vol. 38, pp. 2836–2844, 2023. https://doi.org/10.54691/bcpbm.v38i.4198

[5] S. A. Al Shiam *et al.*, "Credit risk prediction using explainable AI," *Journal of Business and Management Studies*, vol. 6, no. 2, pp. 61–66, 2024. https://doi.org/10.32996/jbms.2024.6.2.6

[6] S. Mir Mohtasam *et al.*, "AI-enhanced stock market prediction: Evaluating machine learning models for financial forecasting in the USA," *Journal of Business and Management Studies*, vol. 5, no. 4, pp. 152-166, 2023. https://doi.org/10.32996/jbms.2023.5.4.16

[7] A. R. Provenzano *et al.*, "Machine learning approach for credit scoring," *arXiv preprint arXiv:2008.01687*, 2020.

[8] H. Te-Cheng, L. Shing-Tzuo, W. Yun-Ping, H. Yung-Shun, and Che-Lin, "Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction," presented at the In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[9] A. Chiamaka, "Determinants of credit default in banking in Nigeria," *International Journal of Modern Risk Management*, vol. 1, no. 2, pp. 33–44, 2023. https://doi.org/10.47604/ijmrm.2220

[10] J. Afrin Hoque *et al.*, "Understanding negative equity trends in us housing markets: A machine learning approach to predictive analysis," *Journal of Economics, Finance and Accounting Studies*, vol. 5, no. 6, pp. 99-120, 2023.

[11] B. Ive, P. Natalija, and B. Jurica, "Opportunities of gen AI in the banking industry with regards to the AI Act, GDPR, Data Act and DORA," presented at the 13th Mediterranean Conference on Embedded Computing (MECO), 2024.

[12] X. Luo, S. Tong, Z. Fang, and Z. Qu, "Frontiers: Machines vs. Humans: The impact of artificial intelligence chatbot disclosure on customer purchases," *Marketing Science*, vol. 38, no. 6, pp. 937–947, 2019. https://doi.org/10.1287/mksc.2019.1192

[13] R. Eustaquio-Jiménez *et al.*, "Chatbots for customer service in financial entities—A comprehensive systematic review," *Journal of Infrastructure, Policy and Development*, vol. 8, no. 16, p. 10122, 2024. https://doi.org/10.24294/jipd10122

[14] Google DeepMind, "Gemini – multimodal AI by google," 2024. https://deepmind.google/technologies/gemini

[15] Streamlit, "Streamlit – The fastest way to build and share data apps," 2023. https://streamlit.io/

[16] N. Gafsi and L. Louhichi, "Role of green finance in sustainable energy transition and economic growth in MENA region," *Edelweiss Applied Science and Technology*, vol. 9, no. 6, pp. 1704–1717, 2025. https://doi.org/10.55214/25768484.v9i6.8216

[17] N. Gafsi, "Perspective chapter: Financial risk management in the age of digital transformation challenges and opportunities in Africa," *IntechOpen*, 2025. https://doi.org/10.5772/intechopen.1010010

[18] N. Gafsi, "Analysing the impact of renewable energy use, CO_2 emissions, oil production, and oil prices on sustainable economic growth: Evidence from Saudi Arabia using the ARDL approach," *Edelweiss Applied Science and Technology*, vol. 9, no. 5, pp. 2317–2326, 2025. https://doi.org/10.55214/25768484.v9i5.7465

[19] N. Gafsi and S. Bakari, "Impacts of agricultural CO_2 emissions, agricultural exports and financial development on economic growth: Insights from East Asia and pacific countries," *International Journal of Energy Economics and Policy*, vol. 14, no. 6, pp. 136–153, 2024. https://doi.org/10.32479/ijeep.16960

[20] A. Hlali and N. Gafsi, "Analysis of digitalization and sustainable development in Africa," *Perspectives on Global Development and Technology*, vol. 22, no. 5-6, pp. 415–428, 2024. https://doi.org/10.1163/15691497-12341668

Appendix

```
Entrée [1]:
# Import basic libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("CANEVAS_2024.csv",encoding='Windows-1252', sep=';')
df.head()
```

Appendix 1.
Code to load the dataset from CSV and display initial rows using pandas.

```
Entrée [47]:
# Handle missing values (example)
df['PROFESSION'] = df['PROFESSION'].fillna('Unknown')
df['MONT_PRIN_DEC'] = df['MONT_PRIN_DEC'].fillna(df['MONT_PRIN_DEC'].median())
df['NUM_NBECH_DEC'] = df['NUM_NBECH_DEC'].fillna(df['NUM_NBECH_DEC'].median())
df['NUM_DUR_DEC'] = df['NUM_NBECH_DEC']+1
df['GENRE'] = df['GENRE'].fillna(df['GENRE'].mode()[0])

#remplacer les âges > 74 par la médiane des âges valides
median_age = df[df['AGE'] <= 74]['AGE'].median()
df.loc[df['AGE'] > 74, 'AGE'] = median_age
```

Figure 0.9.
Appendix 2– Handling missing values using median/mode imputation and replacing outlier ages.

```
# Feature engineering
df['PRESSION_REM'] = df['MONT_PRIN_DEC'] / df['NUM_NBECH_DEC']
df['AGE_x_ANCIENNETE'] = df['AGE'] * df['ANCIENNETE']
df['MONT_x_DUREE'] = df['MONT_PRIN_DEC'] * df['NUM_DUR_DEC']
df['JOURS_DEPUIS_CREATION'] = (
    pd.to_datetime('today') - pd.to_datetime(df['DATECREATION'], dayfirst=True)
).dt.days
```

Figure 0.10.
Appendix 3 – Feature engineering: computing derived columns relevant for credit prediction.

```
Entree [59]:
df = df.drop('LIB_GARPR_CNV', axis=1)
df = df.drop('NUM_CPTCLT_CNV', axis=1)
df = df.drop('TAUX', axis=1)
df = df.drop('AVISPC', axis=1)
df
```

Figure 0.11.
Appendix 4 – Dropping irrelevant or redundant columns to reduce noise in model training.