7th European Conference on Industrial Engineering and Operations Management
Augsburg, Germany, July 16-18, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/EU07.20240079
Published: July 16, 2024

Loan Eligibility System Using Machine Learning

Alpha Z Mnkandla, Belinda Ndlovu, Sibusisiwe Dube and Phillip Nyoni
Department of Information and Analytics
National University of Science and Technology
Bulawayo, Zimbabwe
mnkandlaalphazimkhulu@gmail.com; belinda.ndlovu@nust.ac.zw;
sibusisiwe.dube@nust.ac.zw, phillip.nyoni@nust.ac.zw

Fungai Jacqueline Kiwa
ICT and Electronics Department
School of Engineering and Technology
Chinhoyi, Zimbabwe
fungaijacquelinekiwa@gmail.com

Abstract

As people's demands grow, so does the need for loans to meet some of them. Every day, financial institutions receive numerous loan applications. The financial industry faces significant challenges in managing loan defaults, leading to increased fraud, bad debt, and financial losses. Financial institutions frequently lack reliable methods for determining a loan applicant's creditworthiness, particularly in Africa. The main objective of the study is to develop a machine learning-based loan eligibility system that accurately predicts the creditworthiness of loan applicants in the African financial sector. This study addresses a critical issue in the African financial industry by implementing a reliable predictive model that reduces risks and informs lending decisions. The methodology combines Design Science Research (DSR) and the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. The data used includes parameters such as the applicant's net salary, loan amount, loan term, credit history, and employment status. Cleaning, standardization, and dividing the dataset into training and testing sets were all part of the data preparation process. The Logistic Regression algorithm was utilized for model training, implemented in Python using the Scikit-learn library. The Logistic Regression model's performance was evaluated using the accuracy metric, achieving an accuracy rate of 80%. The developed machine learning model demonstrated high accuracy in predicting loan eligibility, indicating its potential effectiveness in real-world applications. The system was deployed as a web-based application using the Streamlit framework, providing an accessible tool for financial institutions. Future research and improvements, such as data enrichment, advanced analytics, and adherence to financial regulatory compliance, are necessary to further enhance the system's accuracy and reliability.

Keywords
Machine Learning, Loan Eligibility, Creditworthiness, Predictive Modelling.

1. Introduction
A loan eligibility system is a web-based program designed to automate the process of determining a borrower's eligibility for a loan and evaluating their creditworthiness(Wells Fargo 2023). A loan is a sum of money borrowed from a financial institution that is expected to be repaid with interest. People seek loans for various reasons, including purchasing homes, funding education, starting or expanding businesses, or managing unexpected expenses. To become eligible for a loan, applicants typically need to meet certain criteria, such as having a stable income, a good credit history, and sufficient collateral. The traditional loan approval process, which involves extensive manual evaluation of documents and financial histories, is both time-consuming and prone to human error. This inefficiency can lead to inaccuracies in assessing a borrower's eligibility and increase the risk of loan defaults. Despite advancements in financial technology, there is a significant gap in current research regarding the

full automation of loan eligibility assessments using machine learning (ML). Existing systems may partially automate the process but often lack the comprehensive analysis required for reliable and consistent predictions. Machine learning is a subset of artificial intelligence that involves training algorithms to recognize patterns and make decisions based on data(Orji 2022). ML technology is proposed to streamline and enhance the accuracy of the loan approval process due to its ability to process large volumes of data efficiently and generate predictive insights based on various criteria. By automating the loan eligibility assessment, an ML-based system reduces the workload of loan officers, increases the reliability of eligibility assessments, and mitigates the risks of loan defaults. This technology maximizes revenues from loan repayments by streamlining data analysis and providing consistent, data-driven predictions that guarantee loans are only given to trustworthy applicants. In this study, a machine learning model has been developed to predict eligibility by analyzing data from prior customers who applied for loans based on a set of established criteria. These criteria include net salary, employment status, property ownership, gender, level of education, and loan tenure (Kailey and Hagen 2023). By predicting a borrower's chance of repaying the loan, the web application helps loan officers decide whether to approve or deny a loan request by providing them with relevant information. The system also seeks to enable the viewing of data submitted for eligibility checks, facilitate the uploading of loan application data, and provide visualizations of the uploaded data.

2. Literature Review

The literature review delves into the fundamental concept of financial loans, elucidating how they involve the transfer of money from lenders to borrowers at specified interest rates and terms, as highlighted by (Shekhar Jha and Awodele 2022). With loans constituting a significant revenue stream for banks, ensuring the repayment of borrowed funds becomes paramount. This necessitates meticulous assessment processes, prompting the widespread adoption of machine learning techniques in loan eligibility prediction, as evidenced by the studies of (Anannya 2023) and (Hamayel 2021). These researchers employed various machine learning algorithms to analyze past loan recipient data, employing systematic methodologies to ensure prediction accuracy.

Additionally, emerging technologies such as artificial intelligence (AI) and blockchain offer promising avenues for enhancing loan eligibility systems. (Ann Schanare 2022) emphasize the potential of AI in providing more accurate and consistent decisions, potentially expanding credit access for underserved populations. However, concerns regarding historical discrimination and bias necessitate scrutiny of AI model inputs. Meanwhile, blockchain technology, as outlined by (Yaga 2019) and Guo & Liang (2016), bolsters security and efficiency in loan approval processes through tamper-proof records and smart contracts, streamlining underwriting procedures. Moreover, decision support systems, exemplified by the Weighted Product method, offer valuable insights into cooperative credit provision, highlighting the importance of robust decision-making frameworks in financial lending practices.

In a comparative analysis conducted by (Wang 2023) to develop a highly accurate supervised machine learning model for forecasting borrowers' loan repayment within a specified term, various algorithms including LightGBM, XGBoost, random forest, and decision tree were evaluated. The study revealed that the Random Forest algorithm demonstrated superior mean accuracy, surpassing 89.94%, in contrast to the Decision Tree method which achieved 86.69%. The difference was statistically significant ($p = 0.024$, where $p < 0.05$), demonstrating the higher stability and accuracy of the Random Forest in predicting loan eligibility.. This research underscores the efficacy of the Random Forest algorithm, providing valuable insights for financial institutions seeking robust forecasting models (Wang 2023)

2.1 Technologies Used in Existing Loan Eligibility Systems

2.1.1 Artificial Intelligence (AI)

Artificial intelligence (AI) has been applied in various domains such as marketing, customer relationship management, fraud detection, and loan servicing activities. AI models are praised for their ability to make decisions that are more accurate, consistent, and efficient than human models, and they are considered less susceptible to human biases and errors. This capability can potentially increase credit availability for underserved groups like low-income, black, and Hispanic consumers. However, AI models rely on historical data, raising concerns that they could perpetuate historical injustices and discrimination. The complexity of AI models also makes it difficult for non-developers to monitor and evaluate their inputs effectively (Ann Schanare 2022).

2.1.2 Blockchain

Blockchain technology provides a secure and tamper-proof record of a borrower's financial data, enhancing security in the loan approval system. Blockchains are decentralized, distributed digital ledgers that allow users to record transactions in a shared, immutable ledger(Jhamba P 2024). This technology addresses inefficiencies in bank credit information systems, such as limited and poor-quality data, data transfer difficulties between institutions, and unclear data ownership. Blockchain can also implement smart contracts, which are self-executing contracts used to automate the loan underwriting process, verifying borrower eligibility, and approving loans without human intervention (Guo & Liang 2016).

2.1.3 Decision Support Systems

Research on decision support systems for cooperative credit using the Weighted Product (WP) method has shown that these systems can effectively aid in decision-making. The WP method involves connecting attribute ratings through multiplication, with each attribute's rating raised to the power of the relevant attribute. (Dong X 2019) The Simple Additive Weighting (SAW) method is another approach, combining profit and cost criteria as a crucial basis for decision-making. These methods help in making more informed and balanced credit decisions.

2.1.4 Limitations of Existing Systems

Reliance on historical data has the potential to discriminate against historically oppressed groups by serving to confirm preexisting preconceptions. AI models are less visible and more difficult to assess for accuracy and fairness due to their complexity and opacity. Additionally, smaller financial institutions may find it difficult to develop and operate AI systems due to the large computational resources and expertise required.

The implementation of blockchain technology can be complex and costly, requiring substantial investment in infrastructure and expertise (Guo and Liang 2016). Additionally, it necessitates significant changes to existing systems and processes, which may face resistance from stakeholders accustomed to traditional methods. Blockchain's reliance on digital data might exclude applicants with limited digital footprints, such as those from lower-income backgrounds or developing regions(Yaga 2019). Moreover, the scalability of blockchain solutions can be a concern, as increasing transaction volumes may lead to slower processing times and higher costs.

Decision support systems like WP and SAW methods rely heavily on the correct weighting and assessment of attributes, which can be subjective and prone to human error. These systems may also struggle with the complexity of real-world data and the dynamic nature of loan applicants' financial situations, limiting their predictive accuracy and adaptability(Eweoya 2019). Furthermore, the manual calibration of these systems can be time-consuming and requires significant expertise. In rapidly changing financial environments, these static models may not adapt quickly enough to new trends or emerging risks, reducing their effectiveness over time.

3. Research Gap

All of the studies reviewed were conducted across the continents of Asia, Europe, and South America and contributed greatly towards predicting the creditworthiness of an individual on the continents of study. Little efforts have been made towards implementing and developing similar technologies in Africa; hence, the researcher seeks to expand the use of machine learning loan eligibility systems in loan approval in Africa. In Africa, most financial institutions have no system that accurately assesses the creditworthiness of loan applicants. This has resulted in institutions approving loans to borrowers who are unlikely to repay (Rajeev Dhir 2023). This has led to an increase in loan defaults, accumulation of bad debt, and incidents of fraud, negatively impacting the financial health of the institution. Furthermore, the research gap identified in the reviewed literature is the challenge of limited data availability. African financial institutions frequently struggle with obtaining comprehensive and standardized financial data repositories, which impedes the development and evaluation of machine learning models for assessing loan eligibility..

4. Method

The project follows the Design Science Research Methodology to construct a machine learning-based loan eligibility system and CRISP-DM (Cross-Industry Standards Process for Data Mining) architecture for data modeling.

Design Science Research Methodology

Design Science Research Methodology is a paradigm for solving problems that focuses on producing novel objects to further human understanding and solve problems in science and technology. The DSR process involves six key steps.

1. Problem Identification and Motivation

The primary focus of this project was to address the inefficiencies and inaccuracies inherent in manual loan eligibility assessments. Financial institutions often face high rates of loan defaults, which can be attributed to the subjective nature and potential errors in manual assessments. By automating this process through the machine learning model, aim was to enhance the accuracy and efficiency of loan eligibility predictions, thereby reducing the incidence of loan defaults. To achieve this, a literature review was conducted to understand the current challenges and limitations in manual loan assessments. Historical loan data was analyzed to identify patterns and factors contributing to loan defaults.

2. Define Objectives for a Solution
The main objectives were centered around developing a machine learning model to predict loan eligibility and assess borrower creditworthiness. Additionally, the solution aimed to provide functionalities for viewing, uploading, and visualizing loan data. The specific objectives included training a robust machine learning model for loan eligibility prediction, developing a user-friendly interface for financial institutions to upload and view loan data, and implementing data visualization tools to help stakeholders interpret the uploaded data effectively.

3. Design and Development
This phase involved the creation of a regression model to forecast borrowers' creditworthiness, along with the development of the necessary infrastructure for data handling and visualization. The steps included data collection, where a comprehensive dataset of historical loan applications was gathered, encompassing features such as borrower demographics, loan details, credit scores, and repayment histories. Data preprocessing involved cleaning and processing the dataset by addressing missing values, normalizing numerical features, and encoding categorical variables. For model selection, regression models such as Linear Regression, Decision Trees, and Random Forest were considered based on preliminary analysis and model performance metrics.

4. Demonstration
The developed artifact was demonstrated to Wandi Venture Capital Zimbabwe to showcase its application in solving the identified problem of inefficient loan assessments. A presentation was prepared to highlight the problem, solution, and benefits of the automated system. A live demo of the machine learning model was conducted, showing how it predicts loan eligibility based on input data. The interface functionalities, including data upload, viewing, and visualization, were also demonstrated to illustrate the practical application of the solution.

5. Evaluation
The artifact's effectiveness was rigorously evaluated through feedback from users and stakeholders. This phase aimed to determine if the solution met its objectives and provided value in a real-world setting. Feedback was collected from financial analysts and loan officers using structured questionnaires and interviews. The feedback was analyzed to identify areas of improvement and measure user satisfaction. A comparative analysis of loan default rates before and after the implementation of the model was conducted to assess its impact.

6. Communication
Effective communication strategies were employed to keep stakeholders informed and engaged throughout the project. Questionnaires and surveys were distributed to gather detailed feedback and insights from users and stakeholders. Regular meetings were held with key stakeholders to discuss progress, challenges, and obtain buy-in. Comprehensive documentation was prepared, outlining the problem, methodology, model performance, and user guides for the developed system. Workshops and training sessions were organized to train financial institution staff on using the new system and interpreting its outputs. These communication efforts ensured that pertinent stakeholders were well-informed and actively participated in the project's success

Cross-Industry Standards Process for Data Mining, or CRISP-DM
The data modeling methodology that was selected was CRISP-DM. The stages encompass understanding the business context, comprehending the dataset, preparing the data, modeling, evaluating the results, and deploying the solution.

1. Business Understanding
   The creation of a loan eligibility system with the potential to reduce the number of loan defaults is the primary objective of this project. Using machine learning approaches to forecast creditworthiness and visualizing and reporting the model's outcomes are among the particular criteria.

2. Data Understanding
   Data understanding was done to develop an understanding of the features of the dataset. This gave the researcher room to explore and derive some insights from the data that could not be understood by ordinary inspection of the data. Data exploration involves data profiling, data quality evaluation, and finding any problems with the data or values that are missing. Figure 1 shows the head and the tail of the data set which was used.

In [108]:
```
df.tail(5)
```

Out[108]:

	Loan_ID	Gender	Marital_Status	Dependents	Education	Status_of_Employment	Net_Salary	LoanAmount	Loan_Tenure	Pending_Installments	KYC_Docu
609	LP001609	Female	Not-Married	0	Graduate	Contract	2900	71.0	360.0	No	Not-Su
610	LP001610	Male	Married	3+	Graduate	Contract	4106	40.0	180.0	Yes	Not-Su
611	LP001611	Male	Married	1	Graduate	Contract	8072	253.0	360.0	Yes	Not-Su
612	LP001612	Male	Married	2	Graduate	Contract	7583	187.0	360.0	Yes	Not-Su
613	LP001613	Female	Not-Married	0	Graduate	Permanent	4583	133.0	360.0	No	Su

In [109]:
```
df.tail(5)
```

Out[109]:

	Loan_ID	Gender	Marital_Status	Dependents	Education	Status_of_Employment	Net_Salary	LoanAmount	Loan_Tenure	Pending_Installments	KYC_Docu
609	LP001609	Female	Not-Married	0	Graduate	Contract	2900	71.0	360.0	No	Not-Su
610	LP001610	Male	Married	3+	Graduate	Contract	4106	40.0	180.0	Yes	Not-Su
611	LP001611	Male	Married	1	Graduate	Contract	8072	253.0	360.0	Yes	Not-Su
612	LP001612	Male	Married	2	Graduate	Contract	7583	187.0	360.0	Yes	Not-Su
613	LP001613	Female	Not-Married	0	Graduate	Permanent	4583	133.0	360.0	No	Su

Figure 1. Data Summary

3. Data Preparation
   Data preparation is a pivotal step that entails cleaning and preprocessing the dataset to ensure it is suitable for modeling. As a researcher, I engaged in data cleaning, which involved resolving outliers, addressing inconsistencies, and handling missing values. Additionally, feature engineering was conducted to select relevant features and create new ones, thereby enhancing the model's predictive capability. This step ensures that the data is in optimal condition for the subsequent modeling phase.

4. Modelling
   After cleaning and preparing the data, the next step is the modeling stage. Here, machine learning techniques were utilized to build the prediction model. Various supervised machine learning methods, including random forests, logistic regression, decision trees, and support vector machines, were evaluated. The available data was split into training and testing sets, with the models being trained on the training set and evaluated on the testing set. Performance metrics such as accuracy, precision, recall, and F1 score were used to assess the models' capabilities in predicting creditworthiness.

5. Evaluation
   The model's performance, including metrics such as accuracy and recall, was evaluated using the testing dataset. Based on the review results, it was determined if further model improvements were necessary or if alternative strategies should be considered to enhance performance. The models' advantages, disadvantages, and constraints were examined before making any required modifications or improvements. The evaluation's findings were instrumental in determining which type of medical assistance fraud detection was most effective.

6. Deployment
   Using the Streamlit framework, the logistic regression model was integrated into a web-based application for loan processing and eligibility prediction during the deployment phase. This procedure comprises evaluating the model's functionality, keeping an eye on its efficacy, and updating or refining it as needed. For machine learning and data science teams, Streamlit is an open-source app framework that facilitates the quick building and distribution of data apps. When using Streamlit, the deployment procedure usually entails staging the application first and following successful staging, activating it.

5. Data Collecting

This study utilized a dataset downloaded from Kaggle, a data-sharing platform, containing just 613 records on loan applicants. While convenient, Kaggle data may have limitations in quality and origin. The data included both financial information (and potentially some personal details about the applicants. The research claimed the data was anonymized to protect privacy, but there's always a risk of re-identification. Additionally, even if anonymized, it's important to consider how the data was originally collected and whether informed consent was obtained, especially if the research aims to develop models that could impact loan decisions. Table 1 shows the description of the variables which were used in the model development

Table 1. Variable Description

Variables	**Description**
Gender	Applicant's gender, Female, Male
Marital Status	A binary feature indicating whether the applicant is married or not.
Number of dependents	The number of dependents the applicant supports, which may include children or other individuals financially reliant on them.
Education	educational background, possibly categorized as Graduate, Not Graduate
Employment Status	Indicating whether the applicant is a permanent worker or is in contract
Net Salary	income of the loan applicant after all deductions have been made
Co-applicant Income	income of a co-applicant
Loan Amount	the total amount of money the applicant is requesting in the loan application.
Loan Tenure	the duration of the loan, typically represented in months
KYC Documents (Know Your Customer)	Documents needed when applying for a loan including proof of employment, national identity, pay slip, etc
Property Ownership Status	the ownership of the property pledged by the applicant

7. Results and Discussions

The associations between the several numerical variables in the dataset are shown visually in Figure 2's heatmap. While the color itself denotes the direction, the intensity of the color in each box represents the strength of the link. A positive correlation is shown by warm hues, which indicate that an increase in one variable is usually accompanied by an increase in the other. Conversely, cool colors represent a negative correlation, suggesting that when one variable rises, the other tends to decrease. In essence, this heatmap serves as a valuable tool for understanding how changes in one aspect of the data might be linked to changes in another.

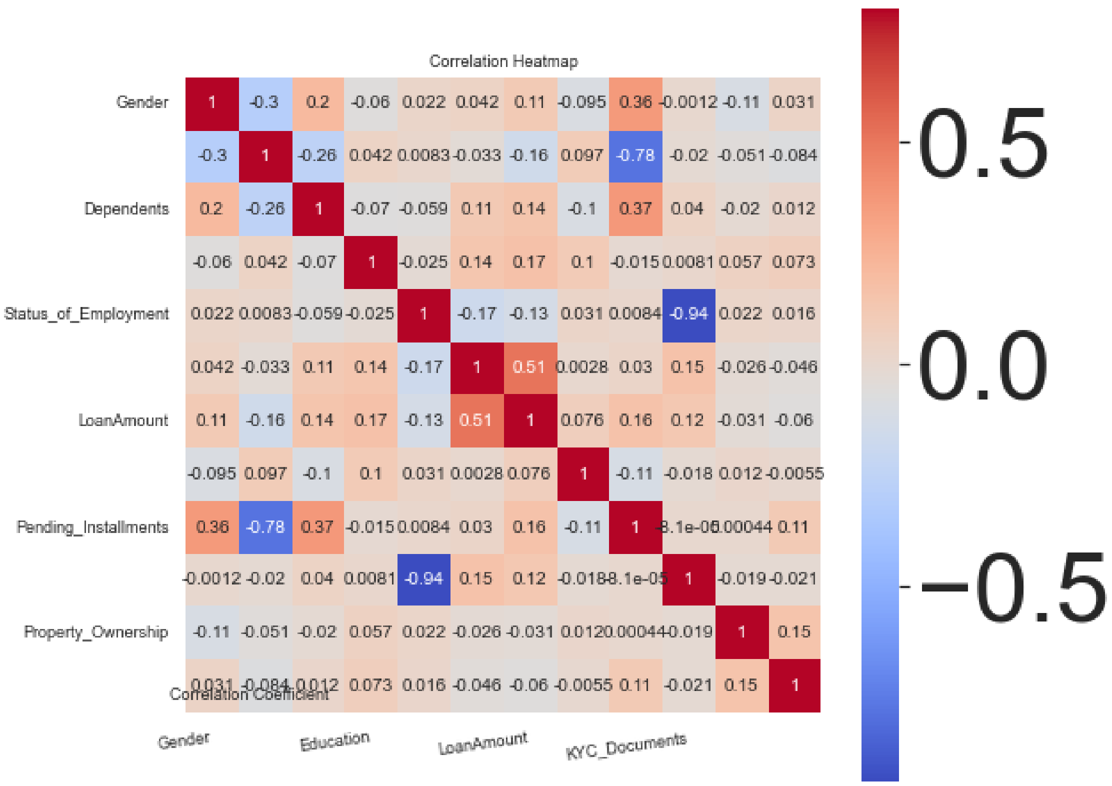


Figure 2. Heat Map

Evaluation of the machine learning algorithms used
The results of the machine learning models used in the loan eligibility system highlight significant differences in their performance metrics. Several processes were involved in developing, evaluating, and selecting the final model.

Data Collection
The data used in this research comprised historical records of loan applications, including various features such as net salary, employment status, property ownership, gender, level of education, and loan tenure. This dataset provided a robust foundation for training and evaluating the machine-learning models.

Data Processing

- Handle Missing Values: The missing data was handled by using appropriate imputation techniques tailored to each feature. Numerical values were imputed using the mean, while categorical values were filled using the mode. This approach ensured that the dataset was fully populated and ready for model training.
- Feature Scaling: Continuous variables underwent standardization to ensure uniform contribution to the model, enhancing convergence and overall performance.
- Feature Encoding: Categorical variables were transformed using methods such as one-hot encoding, enabling their conversion into a numerical format that is compatible with machine learning models. This transformation was essential to effectively incorporate categorical data into the models.

Feature Selection
Features were selected on their correlation with the target variable, using techniques such as correlation matrices and feature importance scores from initial model runs. This step ensured that only the most relevant features were included in the model, enhancing its predictive power and efficiency.

Model Selection and Evaluation
Four machine learning models were evaluated: Random Forest Classifier, K Nearest Neighbors (KNN), Gaussian Naive Bayes (NB), and Logistic Regression. Each model was assessed based on its accuracy, precision, recall, and F1 score.

1. Random Forest Classifier: Achieved an accuracy of 85%, with a precision and recall both at 72%, and an F1 score of 89%. This model demonstrated balanced precision and recall, indicating it is reliable in identifying both eligible and ineligible applicants.

2. K Nearest Neighbors (KNN): Had an accuracy of 79%, with a precision of 76%, but a notably low recall of 40%, leading to an F1 score of 87%. This suggests that while the model is fairly accurate, it struggles with recall, particularly for eligible applicants.

3. Gaussian Naive Bayes (NB): Achieved a high accuracy of 89%, with a precision of 79% and a recall of 34%, resulting in an F1 score of 88%. This indicates that the model is more precise but has difficulty correctly identifying all eligible applicants, as evidenced by the low recall.

4. Logistic Regression: Achieved an accuracy of 89%, with a precision of 82% and a recall of 72%, culminating in an F1 score of 92%. This model stands out with the highest F1 score, indicating a strong balance between precision and recall.

Final Model Training
After extensive evaluation, the Logistic Regression model was selected due to its highest F1 score and balanced performance metrics. The final model was trained on the complete training dataset using the optimal hyperparameters identified through tuning. This ensured the model's robustness and prepared it for deployment.

Model Evaluation
The Logistic Regression model was evaluated on a separate test set to ensure its performance was consistent. The model maintained high accuracy, precision, recall, and F1 score, confirming its robustness and reliability. This evaluation provided confidence in the model's ability to perform well in real-world scenarios.

6. Deployment

The Logistic Regression model was integrated into a web service using Streamlit for deployment. This interface allowed users to input new data and receive immediate predictions from the Logistic Regression model. This deployment marks a significant milestone in the research, allowing the developed model to be applied in real-world scenarios. It showcases the model's practical value by accurately identifying loan applicants who are at risk of defaulting. The interface, built with Python and the Streamlit framework, is directly connected to the model, ensuring seamless interaction between the user input and the predictive engine. Users can enter applicant information through the Streamlit interface, which then processes the input data and passes it to the Logistic Regression model for prediction. The immediate feedback provided by the system helps in making quick and informed financial decisions. This integration showcases the model's tangible utility, offering valuable insights for financial decision-making. The ability to provide real-time predictions and the user-friendly nature of the Streamlit interface enhance the model's accessibility and effectiveness for loan officers and financial institutions. Figure 3 shows the use interface where the applicant's information is entered to predict creditworthiness

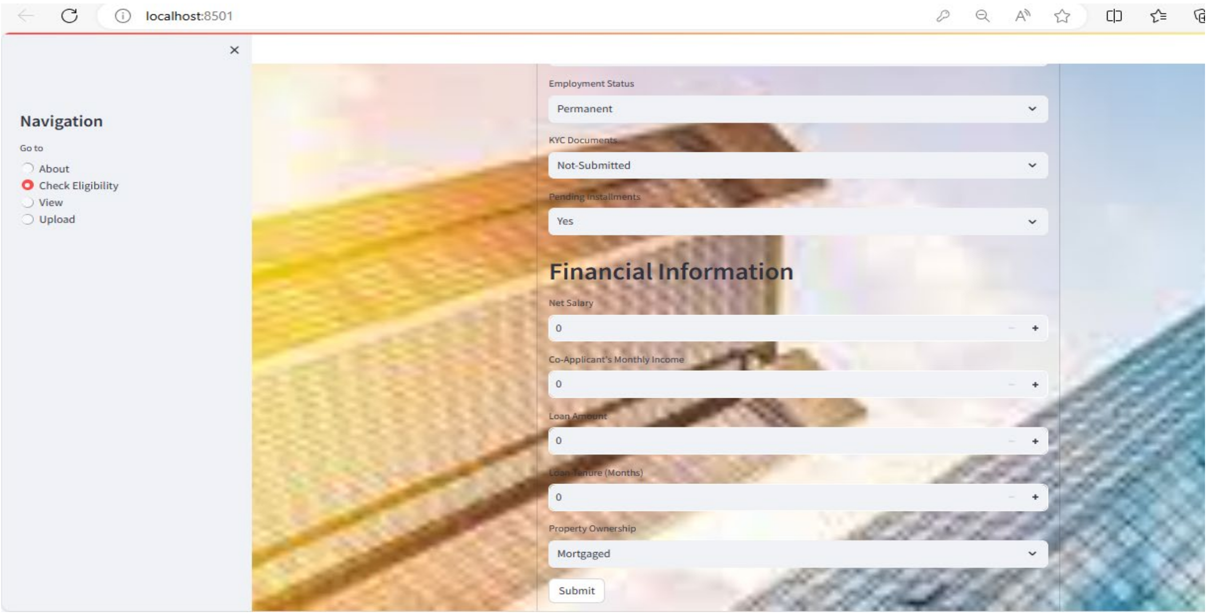
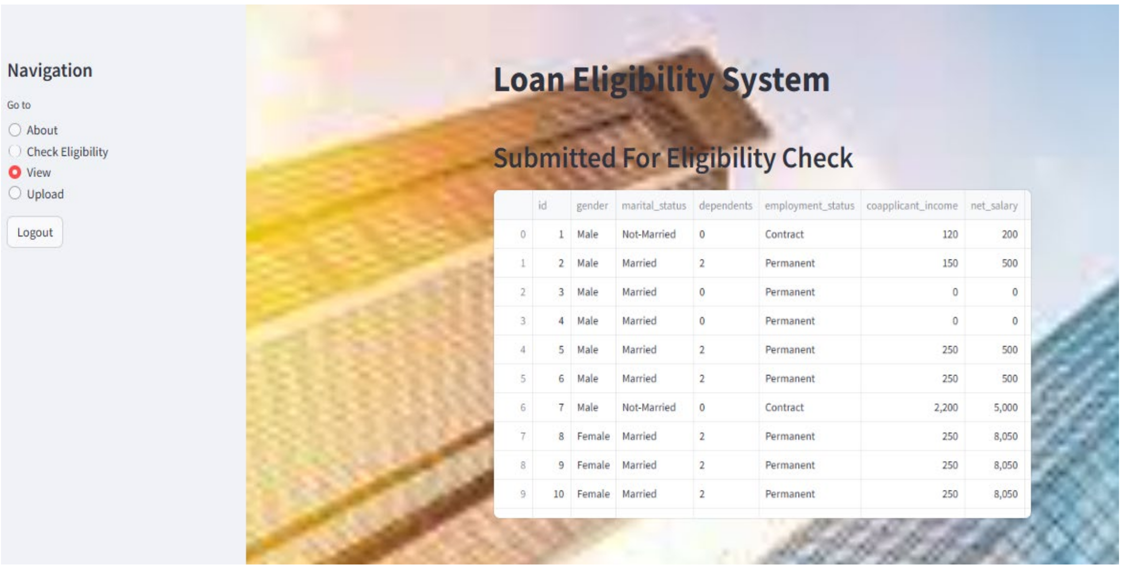


Figure 3. User-Interface

Viewing Submitted Loan Data

The submitted data for checking for eligibility can be viewed by the admin. This functionality allows the admin to access and review all the loan application details submitted by users. Figure 4 shows the page for viewing the submitted data



	id	gender	marital_status	dependents	employment_status	coapplicant_income	net_salary
0	1	Male	Not-Married	0	Contract	120	200
1	2	Male	Married	2	Permanent	150	500
2	3	Male	Married	0	Permanent	0	0
3	4	Male	Married	0	Permanent	0	0
4	5	Male	Married	2	Permanent	250	500
5	6	Male	Married	2	Permanent	250	500
6	7	Male	Not-Married	0	Contract	2,200	5,000
7	8	Female	Married	2	Permanent	250	8,050
8	9	Female	Married	2	Permanent	250	8,050
9	10	Female	Married	2	Permanent	250	8,050

Figure 4. Viewing Submitted Data

Uploading Data

The submitted data for checking for eligibility can be downloaded as a CSV file and uploaded by the admin. This functionality allows the admin to visualize the uploaded data. Figure 5 shows the uploading of loan CSV data set.

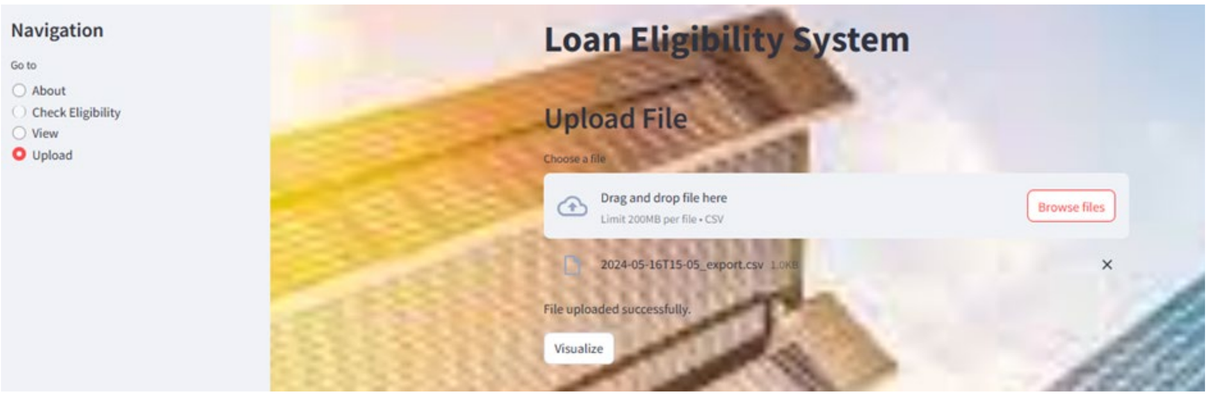


Figure 5. File upload

Visualizing Data

The uploaded file is visualised to gain insights into various aspects of the applications, such as the trends in loan distribution amounts, loan tenure distribution, net salary, and eligibility percentages. This analysis helps identify patterns that can be used to refine loan eligibility criteria, improve the application process, and make data-driven decisions to improve the loan approval system's general efficacy and efficiency. By examining these visualisations, administrators can better understand applicant profiles, identify common factors among applicant, and address any recurring issues that lead to ineligibility. Figure 6 shows the visualized data.

Figure 6. File upload

7. Proposed Improvements

To enhance the effectiveness of the loan prediction model, rigorous attention to data collection and quality is paramount. Effective data cleaning and preprocessing techniques are crucial for handling missing values, outliers, and inconsistencies. These steps are essential to ensure the model's performance and reliability. Moreover, integrating diverse external data sources, such as social media profiles and utility bills, alongside traditional metrics like credit history, enriches the borrower's financial profile and enhances predictive accuracy. Looking ahead, integration with regulatory compliance measures is crucial to adapt to evolving frameworks, necessitating continual adjustments to data practices and machine learning models. Proactive monitoring ensures adherence to regulations, mitigating compliance risks and maintaining trust with regulators and customers, thus ensuring the system's compliance while providing invaluable insights for lending decisions (Magar et al., 2022).

8. Conclusion

The study has demonstrated the effectiveness of employing machine learning techniques, particularly logistic regression, in developing a loan eligibility system capable of accurately predicting borrowers' creditworthiness. Through rigorous application of Design Science Research and the CRISP-DM framework, the study successfully addressed the prevalent issue of high loan defaults faced by financial institutions. The research evaluated and compared model performance using a dataset sourced from Kaggle, employing a range of machine learning algorithms including logistic regression, K Nearest Neighbors, Random Forest Classifier, and Gaussian NB. Ultimately, logistic regression was chosen for deployment due to its higher accuracy. The integration of the model into a user-friendly web service using Streamlit represents a significant milestone, enabling real-world application and showcasing practical utility in identifying loan applicants likely to default.

References

Anannya, M., Khatun, Most. S., Hosen, Md. B., Ahmed, S., Hossain, Md. F., & Kaiser, M. S., Eligible Personal Loan Applicant Selection using Federated Machine Learning Algorithm. *International Journal of Advanced Computer Science and Applications*, *14*(8), 2023. https://doi.org/10.14569/ijacsa.2023.01408110

Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., & Amosu, O., Fraud prediction in loan default using support vector machine. *Journal of Physics: Conference Series*, *1299*(1). 2019. https://doi.org/10.1088/1742-6596/1299/1/012039

Guo, Y., & Liang, C., Blockchain application and outlook in the banking industry. In *Financial Innovation* (Vol. 2, Issue 1), 2016. SpringerOpen. https://doi.org/10.1186/s40854-016-0034-9

Wells Fargo, Manzan, S., & Tosetti, E., Forecasting Loan Default in Europe with Machine Learning. *Journal of Financial Econometrics*, *21*(2), 569–596, 2023. https://doi.org/10.1093/jjfinec/nbab010

Hamayel, M. J., Mohsen, M. A. A., & Moreb, M. , Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine. *2021 International Conference on Information Technology, ICIT, - Proceedings*, 33–37, 2021. https://doi.org/10.1109/ICIT52682.2021.9491636

Jhamba, P., Ndlovu, B., Dube, S., Muduva, M., Jacqueline, F., & Maguraushe, K. , *A Blockchain-based Patient Portal for Mental Health Management*. N.d. https://doi.org/10.46254/AF05.20240251

Magar, S., Nikam, N. S., Taksale, N., & Hajare, S., Issue 8 www.jetir.org (ISSN-2349-5162). In *JETIRFP06097 Journal of Emerging Technologies and Innovative Research* (Vol. 9).2022. www.jetir.org

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N. , Machine Learning Models for Predicting Bank Loan Eligibility. *Proceedings of the 2022 IEEE Nigeria 4th International Conference on*

Disruptive Technologies for Sustainable Development, NIGERCON 2022. https://doi.org/10.1109/NIGERCON54645.2022.9803172

Wang, S., You, S., & Zhou, S., Loan Prediction Using Machine Learning Methods. *Advances in Economics, Management and Political Sciences*, *5*(1), 210–215, 2023. https://doi.org/10.54254/2754-1169/5/20220081

Yaga, D., Mell, P., Roby, N., & Scarfone, K. , *Blockchain Technology Overview, 2019*. https://doi.org/10.6028/NIST.IR.8202

Dong, X., Loan Default Prediction based on Machine Learning (LightGBM Model). In *BCP Business & Management EMEHSS* (Vol. 2022).2022.

Awodele, O., Alimi, S., Ogunyolu, O., Solanke, O., Iyawe, S., & Adegbie, F., Cascade of Deep Neural Network and Support Vector Machine for Credit Risk Prediction. *Proceedings of the 5th International Conference on Information Technology for Education and Development: Changing the Narratives Through Building a Secure Society with Disruptive Technologies, ITED 2022*. https://doi.org/10.1109/ITED56637.2022.10051312

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P., CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. 2021. https://doi.org/10.1109/TKDE.2019.2962680

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N., Machine Learning Models for Predicting Bank Loan Eligibility. *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development, NIGERCON 2022*.

Joshi #1, P., #2, V. T., Belakar, G., #4, S. M., & Wakankar, A. LOAN PREDICTION SYSTEM USING MACHINE LEARNING. *International Journal of Research and Analytical Reviews*. 2022.www.ijrar.org

Hamayel, M. J., Mohsen, M. A. A., & Moreb, M.. Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine. *2021*.

Biographies

Alpha Mnkandla Alpha Mnkandla is a researcher specializing in Informatics and Analytics at the National University of Science and Technology. During his attachment, he gained experience at Steward Bank Zimbabwe as a POS (Point of Sales) Systems Technician and Analyst within the technology services department. His research focuses on the application of machine learning technology in the banking and finance sectors.

Belinda Mutunhu Ndlovu is a Ph.D. in Information Systems student at UNISA. She holds an MSc in Information Systems and a BSc in Computer Science. She is a seasoned software developer and academic. She has published several papers in the fields of Data Analytics, Health Informatics, ICT4D, and 4IR.

Sibusisiwe Dube is an experienced lecturer of Information Systems and Computer Science courses. She holds a PhD in Information Systems, an MSc in Computer Science, and a BSc in Information Systems. She has been lecturing since 2004. She is also an active researcher and supervisor of Postgraduate dissertations and undergraduate student projects.

Fungai Jacqueline Kiwa holds a Doctor of Philosophy Degree in Cultural Heritage and Information Technology from Chinhoyi University of Technology, complemented by a Post Graduate Diploma in Higher Education. Additionally, she holds a Master of Science degree in Information Systems, a Bachelor of Technology (Honors) degree in Computing and Information Technology, and an Advanced Diploma in Computing and Information Technology. With a wealth of academic achievements, Dr. Kiwa has authored 17 publications, encompassing articles, conferences, thesis, and dissertations. Currently, she is a candidate for a Master of Mechatronics and AI at the University of Zimbabwe. Her expertise extends to Artificial Intelligence, creative IoT framework designing, and intensive programming skills, particularly in Java, Python, and C++.

Phillip Nyoni is a lecturer at the National University of Science and Technology, Zimbabwe. He has a BSc (Hons.) from the National University of Science and Technology, an MCom and a PhD in Information Systems from North-West University in South Africa. His research interests are cyber security and data protection.