GSJ: Volume 12, Issue 3, March 2024, Online: ISSN 2320-9186

www.globalscientificjournal.com

Credit Risk Evaluation in the Financial Sector Using Deep Learning

Stow, May Tamara

Department of Computer Science and Informatics, Federal University Otuoke, Nigeria

PMB 126, Yenagoa, Bayelsa state.

stowmt@fuotuoke.edu.ng

Abstract- Evaluating credit risk is a crucial responsibility in the financial sector to assess the probability of borrowers failing to repay loans. Traditional risk assessment methodologies need help to effectively predict creditworthiness due to the growing complexity of financial data and the emergence of non-traditional lending platforms. This paper explores deep learning methods for credit risk evaluation, specifically Long Short-Term Memory (LSTM) networks. The experiment takes place in a Jupyter Notebook and consists of two primary phases: Exploratory Data Analysis (EDA) and LSTM model training. Exploratory Data Analysis (EDA) helps uncover dataset characteristics, such as data imbalances, which can be rectified using oversampling methods. The LSTM model is trained to identify temporal relationships in the data, attaining a high accuracy rate of 98.52% for low and increased credit risk categories. The model demonstrates high precision, recall, and F1-score, indicating its reliability in distinguishing across credit risk classifications. Displaying classification reports and confusion matrices enhance the model's resilience. The LSTM model offers a promising method for credit risk assessment, delivering trustworthy forecasts essential for informed decision-making by financial institutions.

Keywords- *Credit risk evaluation, deep learning, flask framework, financial institution*

1. Introduction

Banking institutions conduct credit risk assessments to determine the probability of borrowers defaulting on loan obligations. This assessment entails evaluating many elements to ascertain the risk level of granting credit to individuals or enterprises. Risk management in banking has developed into a structured field, with risk assessment as a crucial element of this procedure (Grey et al., 2018). In the banking industry, high targets, intense rivalry, and growing pressure on employees might increase the risk of occupational stress, underscoring the need to comprehend and control credit risks (Giorgi et al., 2017).

Credit risk in banking entails thoroughly assessing criteria such as the borrower's credit history, income stability, and current debt levels. This process is essential for banks to make well-informed judgments on loan approvals and interest rates. Accurate risk assessment aids in anticipating the probability of loan defaults, allowing banks to apply suitable risk mitigation methods (Ghafoor et al., 2019). It is crucial to comprehend the occurrence, associated factors, and outcomes of work-related stress in the banking industry to provide a favorable environment for efficient credit risk assessment procedures (Giorgi et al., 2017).

Advanced technologies such as artificial intelligence (AI) have improved credit risk assessment in banking. AI applications like fraud detection algorithms and credit scoring models have enhanced the precision and effectiveness of risk assessment procedures in financial institutions (Tziortziotis et al., 2021). Banks may use AI to analyze large volumes of data to evaluate credit risks more efficiently and base lending choices on data.

Banks must evaluate financial considerations and the broader economic and regulatory environment when assessing credit risk. Economic conditions, regulatory constraints, and market dynamics significantly influence credit risk profiles. Banks must consistently examine and adjust their risk assessment frameworks to ensure they are strong and in line with the changing environment (Ghafoor et al., 2019). By using a comprehensive credit risk assessment that considers internal and external factors, banks can improve their risk management processes and lending decisions.

Credit risk assessment is a crucial banking function that involves a comprehensive study of several criteria to ascertain the risk level of providing credit. Banks can enhance their risk assessment processes and decision-making in lending activities by implementing modern technologies, recognizing the implications of work-related stress, and considering the more significant economic situation. Efficient credit risk evaluation protects banks' financial well-being and enhances financial stability in the banking sector.

2. Literature Review

Bussmann *et al.* (2021) investigate credit risk modeling for small and medium enterprises (SMEs) using standard logistic regression and machine learning XGBoost algorithms. The text highlights the significance of precise default probability estimation and explores model assessment techniques, such as receiver operating characteristic (ROC) curves. The XGBoost model significantly enhanced prediction accuracy, obtaining an AUROC of 0.93 compared to 0.81 for logistic regression. The research introduces Shapley values to explain model predictions, offering insights into the factors affecting individual organizations' default probability. The study emphasizes the efficacy of machine learning methods in evaluating credit risk for small and medium-sized enterprises (SMEs) and showcases the usefulness of Shapley values for interpreting models.

Bao *et al.* (2019) examine how incorporating unsupervised learning methods at two points—consensus and dataset clustering—can improve credit-scoring models. By comparing model performances on three credit datasets and using different algorithm combinations, the study concludes that integrating at any step enhances model performance, with the combined strategy producing the most favorable outcomes. The MCC scores on the German dataset vary between 0.529 and 0.542, while on the Australian dataset, they range from 0.680 to 0.725. The results emphasize the efficiency of combining unstructured and supervised machine learning methods in credit scoring, indicating the potential for scalability in various financial datasets.

Bhatore *et al.* (2022) provide an extensive overview of current research methodologies and machine learning (ML) approaches in credit risk assessment. The authors conducted a thorough literature review of 136 publications published between 1993 and March 2019 to analyze how hyperparameters affect machine learning models used in credit risk evaluation. They emphasize the increasing use of ensemble and hybrid models that include neural networks and support vector machines (SVM) for purposes like credit scoring, non-performing asset (NPA) prediction, and fraud detection. The study highlights the restricted availability of extensive

public datasets, which continues to worry researchers in the field. The study offers valuable insights into the present condition and upcoming trends in credit risk assessment while also recognizing areas for future research and enhancement.

Ma and LV (2019) present the MLIA method, which improves machine learning abilities by breaking down the goal function into weighted combinations of basis functions. The study compares the MLIA and logistic prediction algorithms using three standard test functions. It assesses the MLIA financial credit risk prediction model with data from an Internet financial company. The study uses the AUC value to assess model performance and concludes that the MLIA algorithm effectively predicts financial credit risk. The results show that the MLIA model works better than prior versions, achieving more excellent accuracy rates in training, testing, and cross-time verification datasets. The second edition MLIA model obtains a training set accuracy of 77.46%, a test set accuracy of 73.21%, and a cross-time verification set accuracy of 66.16%, showing its effectiveness in forecasting financial credit risk.

Davis *et al.* (2022) address the development of machine learning models to predict home equity credit risk using real-world data and propose methods to enhance the interpretability of these models for various stakeholders. The authors assess the explainability of the models for loan companies, regulators, loan applicants, and data scientists, taking into account their specific needs for understanding model outputs. They explain each model prediction for loan companies, conduct stress tests for extreme scenarios for regulators, generate counterfactuals to guide loan applicants and derive simple rules to explain a significant portion of the dataset for data scientists. The ultimate goal is to facilitate the adoption of machine learning techniques in domains where explanations of predictions are crucial.

Moscato *et al.* (2021) conducted benchmarking research that assessed credit risk score models to predict loan payback in a peer-to-peer (P2P) lending platform. It focuses on correcting class imbalance and utilizing different classifiers. The paper analyses the performance of various classifiers using a dataset from Lending Club, including 877,956 samples, focusing on evaluation criteria like AUC, Sensitivity, and Specificity. The results indicate that the proposed models, which utilize over-sampling and under-sampling strategies, surpass the approach by Song et al. (2020). Over-sampling with GBDT achieved the most excellent AUC of 0.6207 among the classifiers examined, while under-sampling with Random Forest performed best with an AUC of 0.6207. The results indicate that the suggested models promise to enhance credit risk prediction in P2P lending systems. The study also assesses the explainability of the top three techniques using eXplainable Artificial Intelligence (XAI) technologies.

Zhou et al. (2019). Introduces an advanced hybrid ensemble machine learning model called RS-MultiBoosting, which merges the random subspace (RS) and MultiBoosting methods to boost the precision of predicting credit risk for small and medium-sized firms (SMEs). The research uses data from 46 small and medium-sized firms (SMEs) and seven core enterprises (CEs) in the Chinese stock market from March 31, 2014, to December 31, 2015, to evaluate the feasibility and effectiveness of the RS-MultiBoosting method. RS-MultiBoosting shows strong performance with limited sample sizes, emphasizing the significance of conventional financial metrics like current and quick ratios and supply chain finance (SCF) specific factors, such as trade goods characteristics and CE profit margins, in improving SMEs' access to financing. The study showcases the potential of the RS-MultiBoosting method for credit risk prediction and offers insights into the factors that affect SME funding.

Munkhdalai et al. (2019) used machine-learning algorithms and feature-selection strategies to construct credit-scoring models. The trials use a 10-fold cross-validation method to assure reliability, and the assessment metrics are averaged for comparison. The primary goals are to evaluate the efficacy of various algorithms and identify the most superior one. The tested models are LR, MARS, SVM, RF, XGBoost, and MLP, with hyper-parameters optimized individually. The MLP model with sigmoid activation function demonstrated superior AUC and h-measure metrics performance while employing the TSFFS feature-selection approach. The RF and XGBoost models exhibited superior performance regarding True Positive Rate (TPR), False Positive Rate (FPR), and accuracy. The NAP feature-selection strategy often outperformed TSFFS, especially regarding AUC, TPR, FPR, and accuracy.

Dumitrescu *et al.* (2022) present a new credit scoring method called penalized logistic tree regression (PLTR), which combines decision tree data to improve logistic regression accuracy while maintaining interpretability. PLTR effectively captures non-linear effects in credit scoring data by combining rules from short-depth decision trees with penalized logistic regression while keeping the interpretability of logistic regression. The article shows that PLTR outperforms standard logistic regression and achieves similar accuracy to the random forest method through Monte Carlo simulations and empirical evaluations on four genuine credit default datasets. PLTR has high performance in credit risk prediction with precision, recall, F1-score, AUC, and log loss values of 0.9299, 0.6370, 0.8606, 0.7425, and 0.1029, respectively.

3. Methodology

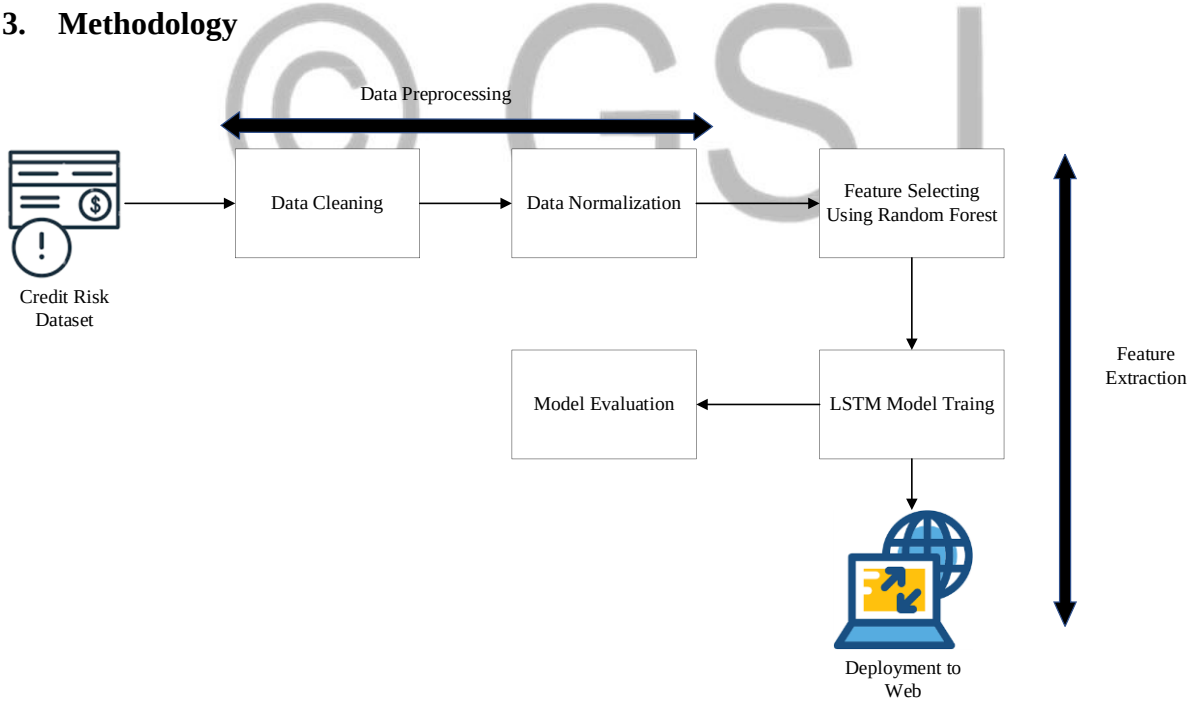

Figure 1: Architectural Design

Credit Card Dataset: The credit risk dataset comprises various features providing insights into individuals' financial and personal information. These features include:

1. **person_age**: Represents the age of the individual applying for the loan.
2. **person_income**: Denotes the annual income of the individual.
3. **person_home_ownership**: Indicates the type of home ownership (e.g., owned, mortgage, rent).

4. **person_emp_length**: Reflects the length of employment in years.
5. **loan_intent**: Describes the purpose or intent behind the loan application.
6. **loan_grade**: Represents the grade assigned to the loan based on the borrower's creditworthiness.
7. **loan_amnt**: Specifies the amount of the loan requested.
8. **loan_int_rate**: Indicates the interest rate associated with the loan.
9. **loan_status**: Represents the loan status, with '0' indicating non-default and '1' indicating default.
10. **loan_percent_income**: Reflects the percentage of income dedicated to the loan repayment.
11. **cb_person_default_on_file**: Indicates historical default status, whether the individual has a history of defaults.
12. **cb_preson_cred_hist_length**: Represents the length of the individual's credit history.

Data Cleaning: Cleaning the raw credit risk dataset includes addressing missing numbers, outliers, and discrepancies. It guarantees that the data is in an appropriate format for analysis. Typical methods in data cleaning are imputation, deduplication, managing categorical variables, and rectifying data mistakes.

```
: data.isnull().sum()

: person_age                      0
  person_income                   0
  person_home_ownership           0
  person_emp_length             895
  loan_intent                     0
  loan_grade                      0
  loan_amnt                       0
  loan_int_rate                3116
  loan_status                     0
  loan_percent_income             0
  cb_person_default_on_file       0
  cb_person_cred_hist_length      0
  dtype: int64
```

Figure 2: Checking for null values

Data Normalization: During this stage, the data is scaled or normalized to provide a uniform scale across all characteristics. Normalization enhances model convergence during training and prevents bigger-size features from overpowering the learning process. Methods like min-max scaling and standardization are frequently employed for data normalization.

Feature Selection Using Random Forest: Feature selection is essential for enhancing model performance and mitigating overfitting, particularly in datasets with many dimensions, such as credit risk assessments. The Random Forest Classifier is utilized to assess the significance of features by evaluating their impact on the prediction. Aspects with outstanding relevance scores are chosen for additional study, while less significant aspects are eliminated.

LSTM Model Training: When training an LSTM model for credit risk assessment, important parameters to consider are the number of LSTM units for model complexity, the number of time steps for temporal dependencies, batch size for training stability, learning rate for parameter updates, epochs for dataset exposure, dropout rate for overfitting prevention,

optimizer for training efficiency, and loss function for prediction accuracy. The parameters influence the model's performance and generalization ability, which is essential for accurately evaluating credit risk.

Model Evaluation: After training the LSTM model, it is assessed using a distinct validation dataset to measure its performance. Key evaluation metrics for credit risk analysis comprise accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics aid in gauging the model's generalization to new data and its accuracy in classifying credit risk.

Deployment to Web: Once the LSTM model has been trained and assessed, it can be implemented on a web-based platform to provide users with accessible predictions. This entails incorporating the model into a web application or service, allowing users to input pertinent information, like credit history and financial data, to receive real-time credit risk forecasts. Python flask framework will be used for the deployment.

4. Results and Discussion

The experiment was carried out in Jupyter Notebook, and the experimental data comprises two separate phases. The experimental procedure has two primary phases: Exploratory Data Analysis (EDA) and training the Random Forest Classifier to identify fraudulent traffic on network systems.

4.1 Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted on the dataset to enhance comprehension of its properties. An exploratory data analysis (EDA) was conducted to evaluate the uneven distribution in the dataset. Figure 3 shows a bar chart of house ownership vs loan status from the exploratory data analysis. Figure 4 shows a bar chart of loan intent vs loan status; Figure 5 shows the percentage of an individual's income that was granted a loan. Figure 6 shows the total number of individuals with low and high-risk loans. Figure 6 shows the data imbalance, and Figure 7 shows that the data imbalance has been resolved. Finally, Table 1 and Figure 8 show the feature ranking.

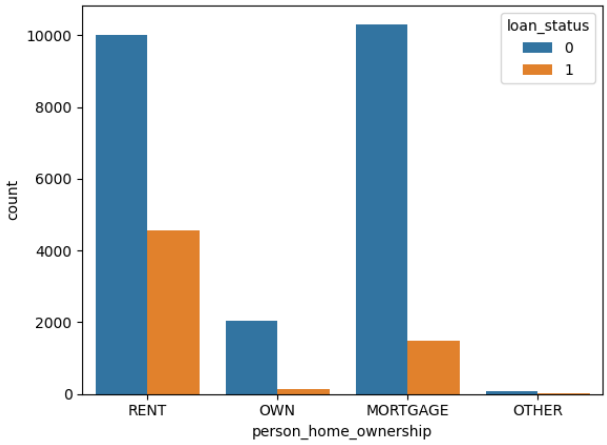

Figure 3: Bar chart of person_home_ownership

The visualized image depicts the number of individuals who may default on loans. From the diagram, people who live in rented apartments are likely not to default if given a loan, followed by mortgage individuals.

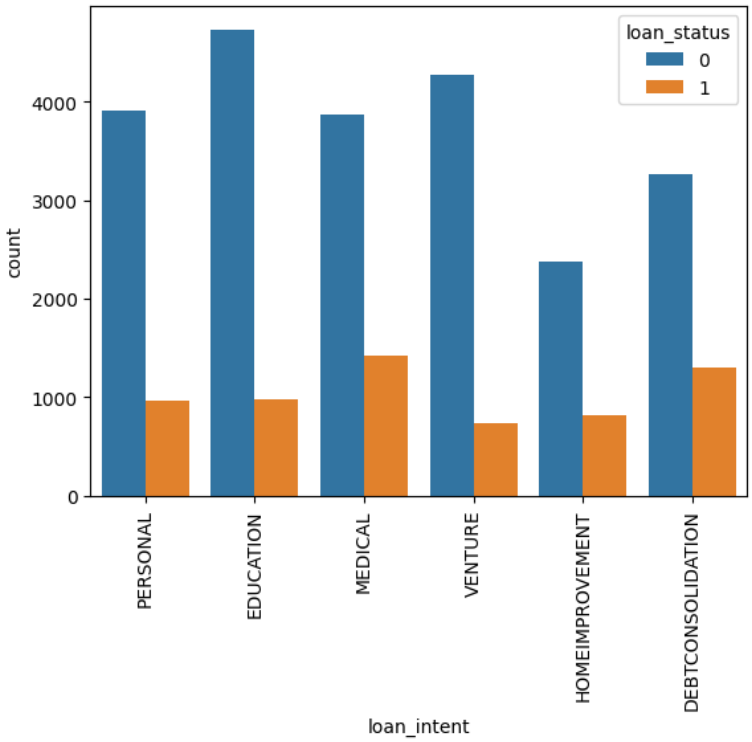

Figure 4: Bar chart of loan intent

From the diagram above, loans are given more to people who will use them for medical purposes, followed by people who will use them for debt consolidation.

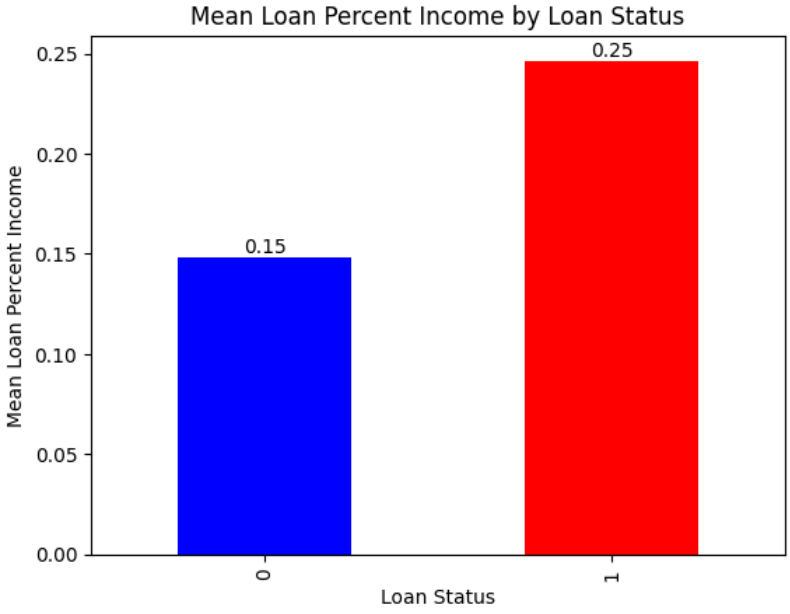

Figure 5: Bar chart of loan status

The diagram shows that 25% of people with higher incomes are likely not to default, while 15% of people with lower incomes are likely to default.

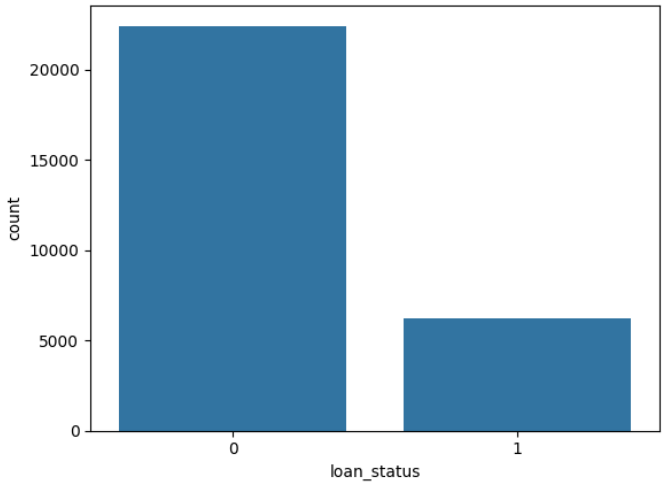

Figure 6: Imbalance data

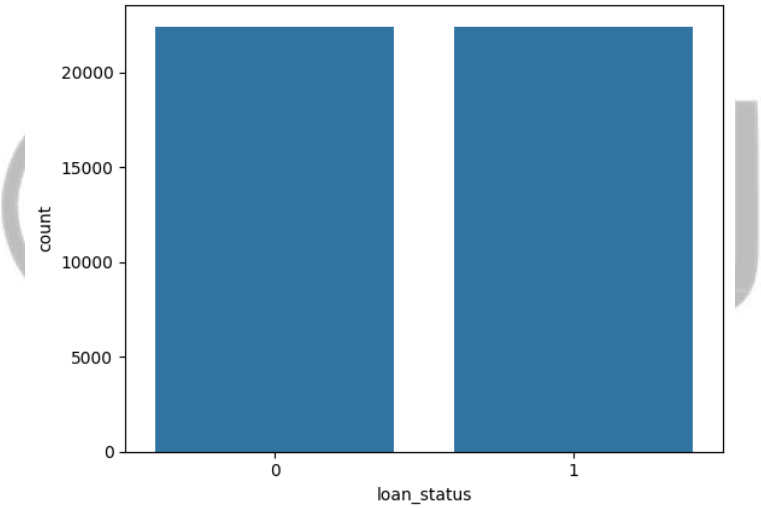

Figure 7: Balanced data

Table 1: 10 Most Important Features

Features	Important_Features
loan_percent_income	0.199787
person_income	0.164927
loan_int_rate	0.133443
loan_grade	0.130489
loan_amnt	0.082112
person_home_ownership	0.068814
loan_intent	0.064391
person_emp_length	0.055054
person_age	0.047436

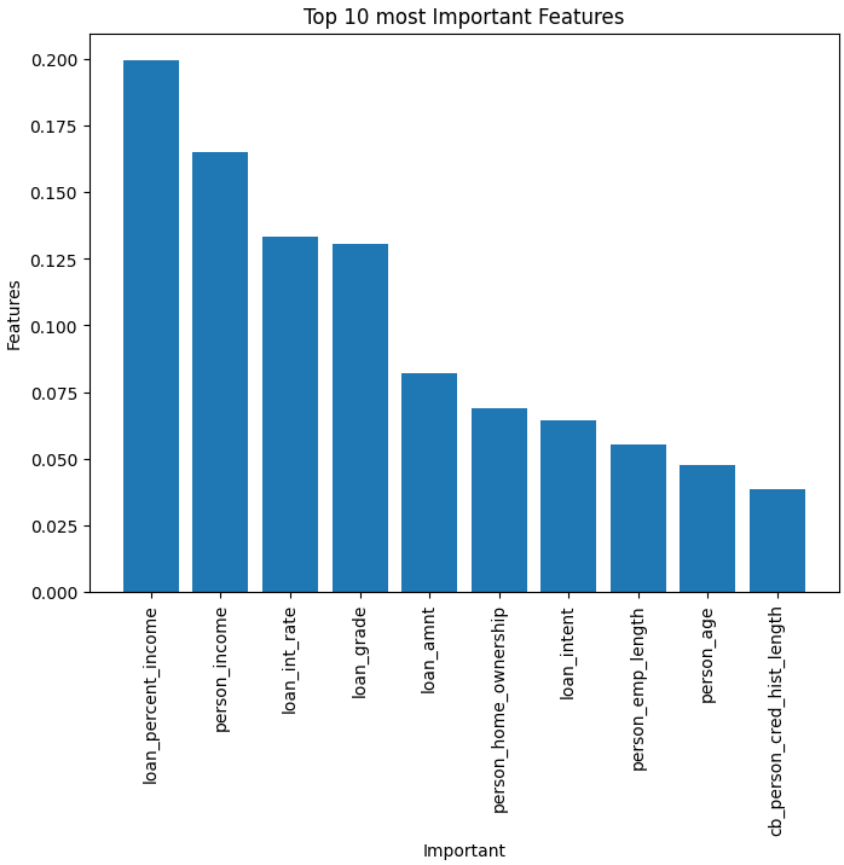

Figure 8: Feature Importance

This shows the ranking of the dataset features. Loan_percent_income has the highest ranking.

4.2 Implementation of the LSTM Model For Credit Risk Analysis

The Long Short-Term Memory (LSTM) network was employed in this sub-section for credit risk analysis. The Sequential model was utilized to construct the neural network architecture, comprising two LSTM layers with 50 units each, utilizing the rectified linear unit (ReLU) activation function and configured to return sequences in the first layer. The input shape is specified based on the dimensions of the training data (X_train). The final layer is a Dense layer with five units and a softmax activation function, corresponding to the two output classes for credit risk analysis. The model is compiled using the Adam optimizer, categorical cross-entropy loss function, and accuracy as the evaluation metric. This configuration aims to leverage the memory-retaining capabilities of LSTM networks to effectively capture temporal dependencies in sensor data for accurate credit risk evaluation. The Long Short-Term Memory training process can be seen in Table 2. Figure 9 shows the LSTM model's accuracy for training and testing, and Figure 10 shows the loss values of the LSTM model for both training and testing. Figure 11 shows the classification report of the LSTM model, and Figure 12 shows the confusion matrix for the LSTM model.

Table 2: Training Process of the LSTM for Credit Risk Evaluation

```
Epoch 1/10
1204/1204 [==============================] - 10s 5ms/stop - loss: 0.3612 - Accura
cy: 0.8876 - val_loss: 0.1745 - val_accuracy: 0.9457
Epoch 2/10
1204/1204 [==============================] - 5s 5ms/stop - loss: 0.1432 - Accurac
y: 0.9552 - val_loss: 0.1253 - val_accuracy: 0.9596
Epoch 3/10
1204/1204 [==============================] - 5s 5ms/stop - loss: 0.1096 - Accurac
y: 0.9641 - val_loss: 0.1021 - val_accuracy: 0.9655
Epoch 4/10
1204/1204 [==============================] - 5s 4ms/stop - loss: 0.0874 - Accurac
y: 0.9723 - val_loss: 0.0855 - val_accuracy: 0.9715
Epoch 5/10
1204/1204 [==============================] - 5s 4ms/stop - loss: 0.0727 - Accurac
y: 0.9772 - val_loss: 0.0746 - val_accuracy: 0.9754
Epoch 6/10
1204/1204 [==============================] - 5s 5ms/stop - loss: 0.0625 - Accurac
y: 0.9814 - val_loss: 0.0708 - val_accuracy: 0.9744
Epoch 7/10
1204/1204 [==============================] - 5s 4ms/stop - loss: 0.0558 - Accurac
y: 0.9826 - val_loss: 0.0604 - val_accuracy: 0.9829
Epoch 8/10
1204/1204 [==============================] - 5s 5ms/stop - loss: 0.0516 - Accurac
y: 0.9849 - val_loss: 0.0552 - val_accuracy: 0.9826
Epoch 9/10
1204/1204 [==============================] - 5s 5ms/stop - loss: 0.0491 - Accurac
y: 0.9854 - val_loss: 0.0619 - val_accuracy: 0.9825
Epoch 10/10
1204/1204 [==============================] - 5s 4ms/stop - loss: 0.0478 - Accurac
y: 0.9861 - val_loss: 0.0540 - val_accuracy: 0.9840
CPU times: total: 1min 55s
Wall time: 58.8 s
```

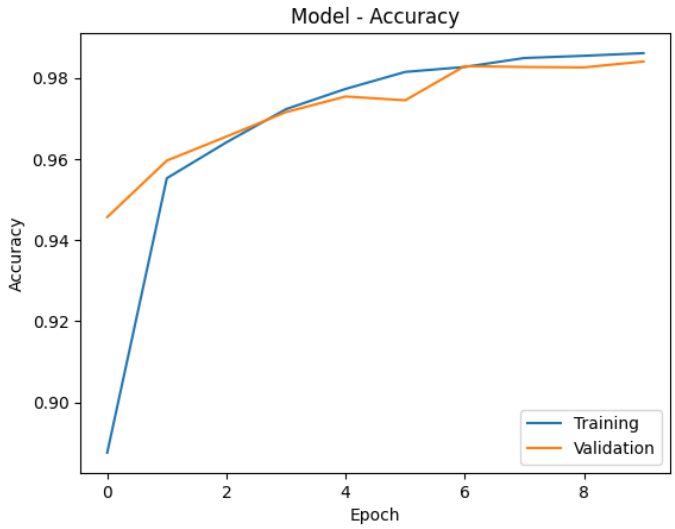

Figure 9: Accuracy of the MLP model for both Training and Testing

The accuracy demonstrates how well the model performed during training. This shows that the model achieved an accuracy of 99.99% for the training data and 99.99% for the validation or testing data. The blue line represents the model training accuracy, whereas the orange line represents the validation test accuracy. A validation test means evaluating the model performance using testing data.

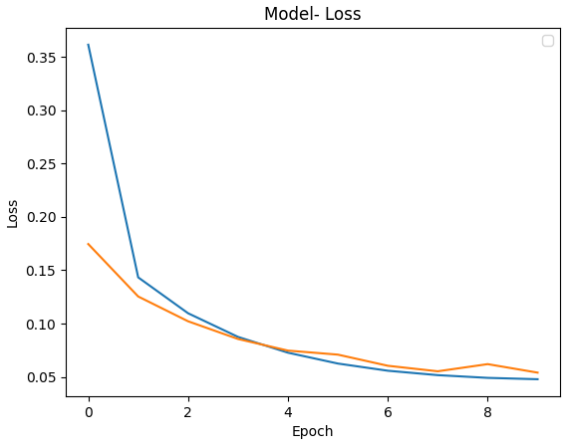

Figure 10: Loss values for the training and testing of the MLP model.

The line graph above represents the losses acquired by the model during training and testing. The green line indicates the loss acquired by the model during training, and the orange line indicates the loss acquired during testing. The loss values are acquired at each training step, starting from step 1 to step 10. Loss values mean the losses the model had during training. This shows that the model achieved a loss value of about 0.06% for the training and validation or testing data.

Classification_Report For LSTM

	precision	recall	f1-score	support
Low Crdit Risk	0.98	0.99	0.98	4488
High Credit Risk	0.99	0.98	0.98	4486
accuracy			0.98	8974
macro avg	0.98	0.98	0.98	8974
weighted avg	0.98	0.98	0.98	8974

Figure 11: Classification Report of the LSTM model

The classification report indicates outstanding performance of the LSTM model with flawless precision, recall, and F1-score for the "Low credit Risk" and "High credit Risk" classes. This shows that the model correctly classified all occurrences of both classes in the dataset. The 98.9% accuracy rate solidifies the model's ability to differentiate between the two classes. The LSTM model shows outstanding classification performance with great precision, recall, and accuracy, making it reliable.

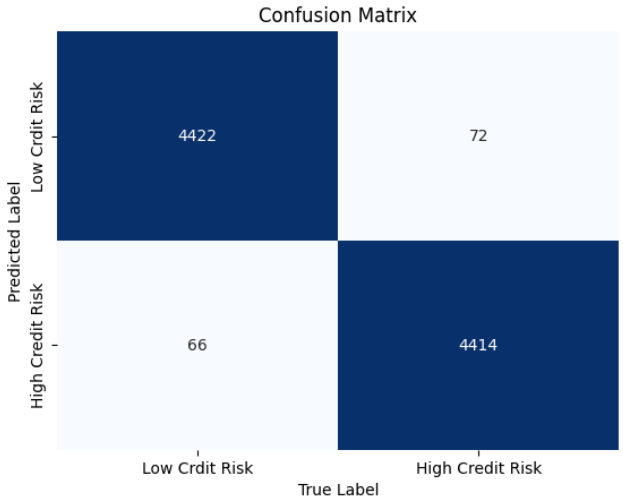

Figure 12: Confusion Matrix of the MLP model

The confusion matrix shows the number of correct and incorrect predictions the model makes on the test data. The confusion matrix results show that the LSTM model makes correct predictions of 98.52% with misclassification of 0.16%.

4.3 Deployment

The trained LSTM model was deployed to the web for further testing. This was deployed using the Python framework with Bootstrap. The deployed web application can be seen in Figures 13 and 14.

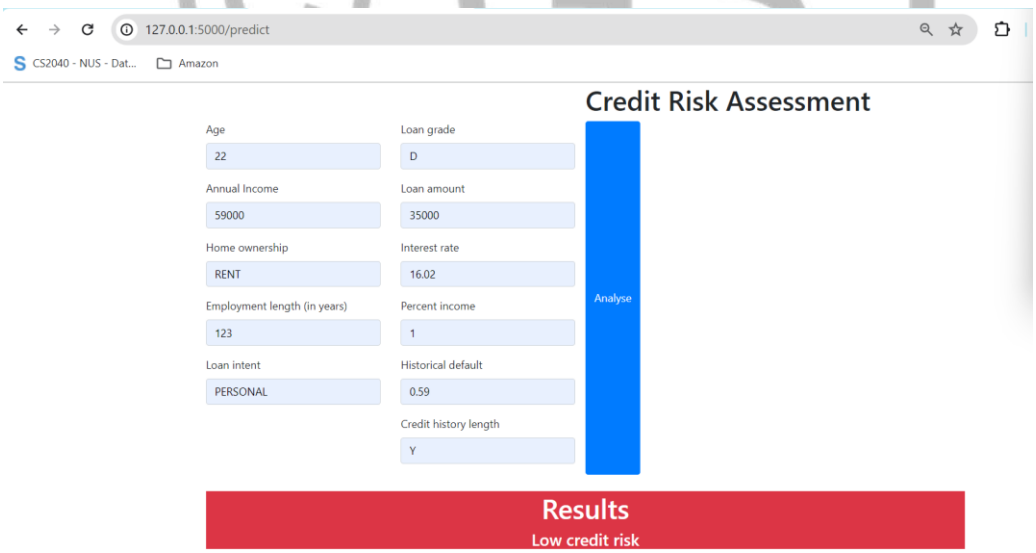

Figure 13 Low Credit Risk

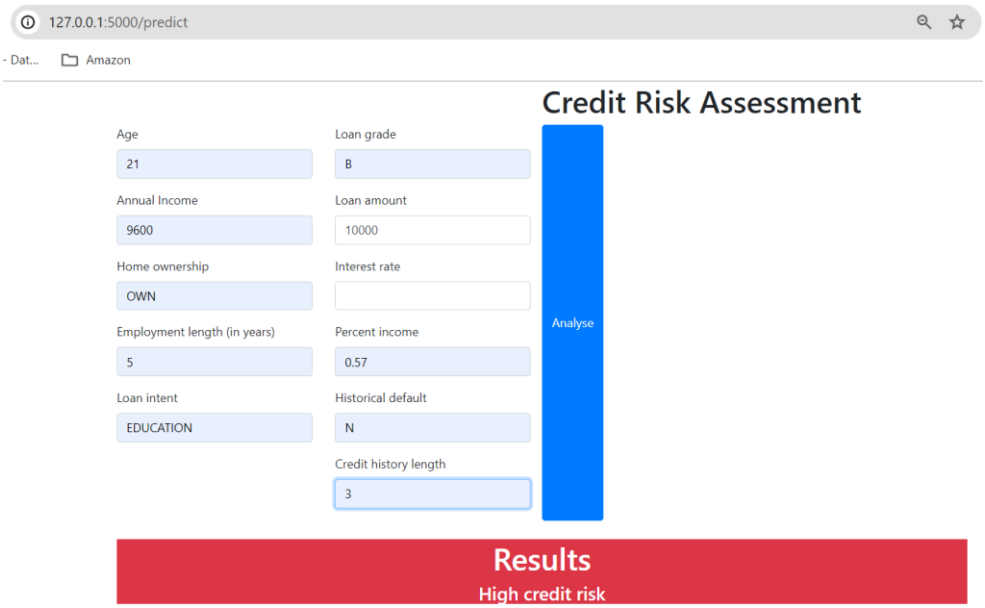

Figure 14: High credit risk

5. Conclusion

Ultimately, the credit risk evaluation using deep learning, specifically utilizing the Long Short-Term Memory (LSTM) network, demonstrated positive outcomes in accurately identifying temporal relationships in the dataset for precise credit risk analysis. The experiment in a Jupyter Notebook environment consisted of two main phases: Exploratory Data Analysis (EDA) and training the LSTM model. During the exploratory data analysis phase, the dataset's characteristics were examined, identifying data imbalances that were rectified using techniques such as oversampling. The LSTM model implementation showed exceptional classification performance, with an accuracy rate of 98.52% for low and high-credit risk classes. The model demonstrated great precision, recall, and F1-score for both categories, confirming its reliability. Visualizing classification reports and confusion matrices enhanced the model's ability to distinguish between credit risk categories accurately. The LSTM model provides a robust method for evaluating credit risk and delivering dependable forecasts crucial for making well-informed decisions in financial institutions.

6. Acknowledgement

I conducted this research with the opportunity provided by the Department of Computer Science, Federal University Otuoke, Nigeria. Therefore, I express my gratitude and acknowledge their support.

References

Ghafoor, S., Burger, I., & Vargas, A. (2019). Multimodality imaging of prostate cancer. Journal of Nuclear Medicine, 60(10), 1350-1358. https://doi.org/10.2967/jnumed.119.228320

Giorgi, G., Arcangeli, G., Perminiene, M., Lorini, C., Ariza-Montes, A., Pérez, J., … & Mucci, N. (2017). Work-related stress in the banking sector: a review of incidence, correlated factors, and major consequences. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.02166

Gray, G., Bron, D., Davenport, E., D'Arcy, J., Guettler, N., Manen, O., … & Nicol, E. (2018). Assessing aeromedical risk: a three-dimensional risk matrix approach. Heart, 105(Suppl 1), s9-s16. https://doi.org/10.1136/heartjnl-2018-313052

Tziortziotis, I., Laskaratos, F., & Coda, S. (2021). Role of artificial intelligence in video capsule endoscopy. Diagnostics, 11(7), 1192. https://doi.org/10.3390/diagnostics11071192.

Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *pp. 4*, 111–138.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 203-216.

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301-315.

Ma, X., & Lv, S. (2019). Financial credit risk prediction in Internet finance driven by machine learning. *Neural Computing and Applications*, *31*, 8359–8367.

Davis, R., Lo, A. W., Mishra, S., Nourian, A., Singh, M., Wu, N., & Zhang, R. (2022). Explainable machine learning models of consumer credit risk. *Available at SSRN 4006840*.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction—*Expert Systems with Applications*, *165*, 113986.

Zhu, Y., Zhou, L., Xie, C., Wang, G. J., & Nguyen, T. V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, *pp. 211*, 22–33.

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J., & Ryu, K. (2019). *An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. Sustainability, 11(3), 699.* doi:10.3390/su11030699

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192.