COMPARATIVE ANALYSIS OF RF, SVR WITH GAUSSIAN KERNEL AND LSTM FOR PREDICTING LOAN DEFAULTS

Konstantinos Kofidis*, Cătălina Lucia Cocianu
Bucharest University of Economic Studies, Bucharest, Romania

Abstract
This investigation elucidates the paramount endeavour of predicting loan defaults, which is imperative for the efficacious management of financial risk and the overall stability of financial institutions. Conventional statistical methodologies frequently encounter challenges in effectively capturing the nonlinear and sequential dynamics inherent in financial data, thereby necessitating the examination of more sophisticated machine learning methodologies. This research reports an experimental-based comparative evaluation of three ML and DL models—Long Short-Term Memory (LSTM) networks, Random Forest (RF), and Support Vector Regression (SVR)—to assess their efficacy in forecasting loan defaults. The models are evaluated using metrics such as Mean Squared Error (MSE), F1 score, and Accuracy, and their proficiency in addressing imbalanced datasets and elucidating intricate data relationships is highlighted. The results indicate that while the Random Forest model surpasses its counterparts in terms of accuracy and MSE, the LSTM model exhibits considerable potential in managing imbalanced data, as evidenced by its stable F1 score. Although SVR reveals competitive precision, it exhibits deficiencies in addressing class imbalance. The ANOVA analyses substantiate that the disparities in model performance are statistically significant. The research acknowledges that both the LSTM and SVR models remain in the developmental stages, with ongoing initiatives aimed at refining these models through hyperparameter optimization and advanced architectural frameworks to enhance their predictive efficacy in practical applications.

Keywords
Loan default prediction, financial risk management, Long Short-Term Memory (LSTM), Support Vector Regression (SVR), Mean Squared Error, F1 score

JEL Classification
C45, C51, C53, G20, G21

* Corresponding author, Konstantinos Kofidis– kofidiskostas@hotmail.com

Introduction

In financial analysis, predicting loan defaults is a formidable challenge with significant implications for risk management and decision-making processes. Traditional statistical methods often fail to capture financial data's complex, nonlinear patterns, thereby limiting their predictive power. Recent advancements in machine learning offer promising alternatives, with models like Long Short-Term Memory (LSTM) networks excelling in handling sequential data and identifying long-term dependencies [1]. On the other hand, Support Vector Regression (SVR) with a Gaussian kernel is known for its robustness in modeling complex relationships in data with fewer assumptions about the underlying distribution [2]. Additionally, Random Forests provide a powerful ensemble learning method that combines multiple decision trees to improve prediction accuracy and handle data variability [3].

However, the effectiveness of these models depends significantly on tuning their hyperparameters, which is a challenging task given the vast number of possible combinations. To address this issue, we propose a comparative analysis of SVR with a Gaussian kernel, LSTM networks, and Random Forests, focusing on their performance in predicting loan defaults. This study leverages a dataset from Kaggle, which includes 303 records, each describing 13 attributes of borrowers. The dataset presents a classification challenge with a mixture of default (positive) and non-default (negative) cases, highlighting the common issue of imbalanced classes in training predictive models [4].

To ensure a fair comparison and enhance the models' accuracy and reliability, we implemented rigorous preprocessing methods, including resampling, shuffling, and encoding variables to prepare the data effectively [5]. The shuffling of data ensures that the training and testing sets are representative of the overall distribution, reducing the risk of overfitting [6]. Encoding categorical variables into numerical values makes the data compatible with machine learning algorithms that require numerical input [7].

This paper outlines our methodology, from data preprocessing to model optimization, and presents a detailed comparative analysis of SVR with a Gaussian kernel, LSTM networks, and Random Forests. Our findings contribute to the body of knowledge in financial risk management, emphasizing the strengths and limitations of each model in the context of loan default prediction. By combining these advanced machine learning models, we aim to pave the way for future advancements in predictive analytics within the financial sector [8]. The integration of these models into financial risk management practices provides significant improvements in prediction accuracy and decision-making capabilities, ultimately enhancing the stability and profitability of financial institutions [9].

While loan default prediction has been extensively examined in the literature using various machine learning techniques, this research distinguishes itself by its experimental-based comparative approach, particularly focusing on Support Vector Regression (SVR) with a Gaussian kernel, Long Short-Term Memory (LSTM), and Random Forest. Unlike previous studies that often focus on a single algorithm or a simplified data setup, our study not only explores the strengths and weaknesses of each model in handling imbalanced and complex datasets but also integrates statistical validation through ANOVA, showcasing the robustness and practical applicability of these models. Additionally, we emphasize the innovative inclusion of LSTM networks in

loan default prediction, highlighting its ability to manage temporal sequences, which is a novel angle within financial risk management research.

This research introduces several innovations in the field. First, it adopts a rigorous preprocessing strategy to deal with class imbalance, which is crucial for real-world financial data. Second, it pioneers a hybrid comparative methodology combining classical and deep learning approaches in the context of credit risk management, underlining their advantages and limitations with solid empirical evidence. Our extensive contribution also includes the experimental validation of model performance using advanced statistical techniques, an approach that has not been extensively explored in related literature.

The remainder of this paper is structured as follows: Section 1 provides a comprehensive review of relevant literature, positioning this study within the broader context of financial risk prediction using machine learning techniques. Section 2 details the research methodology, outlining the data preprocessing steps and the setup of the three models (SVR, LSTM, and Random Forest). Section 3 presents the experimental results and a comparative analysis of the model performances, with statistical validation. The final section discusses the implications of the findings, the potential for future research, and concludes with insights into the applications of these models in financial risk management

1. Review of the scientific literature

1.1 Classical ML Techniques

Recent studies have extensively delved into the use of machine learning (ML) methodologies in the anticipation of loan defaults, showcasing notable improvements in comparison to conventional models like logistic regression. One remarkable research endeavour examines the efficacy of diverse ML techniques, around Lasso, Classification and Regression Trees (CART), Random Forest, XGBoost, and Deep Neural Networks, in the prediction of credit defaults while tackling regulatory obstacles within the realm of financial services. By leveraging a dataset comprising over 75,000 anonymized credit transactions involving up to 370 risk variables, the investigation employs cross-validation methodologies and assesses model efficacy through metrics such as AUC-ROC and Brier scores. The results suggest that XGBoost and Random Forest exhibit superior classification and calibration capabilities when juxtaposed with logistic regression. The research posits that models with heightened predictive prowess, such as XGBoost, may result in diminished capital requisites under the Basel framework, thereby emphasizing the potential of these tools in augmenting predictive accuracy and adherence to regulatory standards in the domain of credit risk management. [10]

Moreover, a different study uses six supervised Machine Learning algorithms - Random Forest, Artificial Neural Network, CART, Support Vector Machine, Logistic Regression, and Naïve Bayes - in order to forecast loan defaults, with a specific focus on the attribute of "early loan repayment." By utilizing a dataset containing 960 borrowers, the findings indicate that models incorporating this attribute outperform those that do not, across various performance metrics such as accuracy, precision, recall, RMSE, and AUC-ROC. Among these models, the Random Forest model stands out as the most efficient, achieving an accuracy rate of 93% and displaying superior performance in minimizing classification errors. This investigation underscores the crucial significance of early loan

repayment in enhancing predictive accuracy and proposes that the integration of such attributes can substantially enhance risk assessment in financial lending [11].
Another study, focusing on customer data from The Grant Group of Companies, compares the performance of Logistic Regression, Decision Tree, Random Forest, and XGBoost models in predicting loan defaults. The study aims to enhance the accuracy of predicting whether borrowers will default on their loans, thus aiding financial institutions in decision-making and risk management. The results indicate that XGBoost outperforms other models, achieving the highest recall (0.35) and AUC (0.832), making it the most effective model for identifying potential defaulters. This research emphasizes the importance of using recall over accuracy to minimize the cost of undetected defaults and highlights the importance of factors such as asset value, income, and living status in predicting defaults. The study contributes to the field by applying ML models to previously underexplored datasets and demonstrating their efficacy in loan default prediction, with potential applications extending to other domains. Future work will focus on addressing data imbalance issues and exploring larger datasets to further improve model performance. [12]

1.2 Advanced and Ensemble ML Techniques

Moving on to a study utilizing a dataset of 12 million residential mortgages across seven European countries models default occurrence as a function of borrower characteristics, loan-specific variables, and local economic conditions. The study compares the performance of ML algorithms with logistic regression, finding that tree-based models, particularly Gradient Boosting (GB) and Extreme Gradient Boosting (XGBoost), significantly outperform logistic regression in predicting defaults. Key factors driving defaults include current interest rates and Loan-to-Value (LTV) ratios, with substantial geographical heterogeneity observed in their importance. This research underscores the necessity for regionally tailored risk-assessment policies to manage credit risk effectively and highlights the superiority of ML models in credit risk prediction. [13]
In addition to the previous, another study examines the application of six supervised ML algorithms—Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Trees (GBTs), Factorization Machines (FM), and Linear Support Vector Machine (LSVM)—to predict loan defaults using the Apache Spark Big Data platform. Using a Kaggle dataset containing 640,000 instances and 14 features, the study builds predictive models and finds that Decision Tree and Random Forest models achieve the highest accuracy of 99.62%. Various evaluation metrics such as accuracy, precision, recall, ROC curve, and F-score are used to assess model performance, with Decision Tree and Random Forest outperforming others. This research highlights the effectiveness of Spark ML libraries in processing large datasets and providing accurate predictions, ultimately aiding financial institutions in mitigating credit risk and enhancing profitability. [14]
Furthemore, a study employing K-Nearest Neighbours (KNN) and Decision Tree models to predict loan defaults emphasizes the importance of automated loan eligibility assessments based on borrower data such as gender, marital status, education, income, and credit history. The research aims to enhance decision-making and reduce non-performing assets for banks, with the KNN model validated against the dataset and outperforming the Decision Tree model in terms of accuracy. This study supports the use

of ML techniques to improve the efficiency and accuracy of loan default predictions, ultimately aiding financial institutions in making informed lending decisions. [15]
The work [16] further addresses the challenge of loan defaults faced by banks and the financial losses they incur. Several ML models to predict loan defaults using data from the Kaggle platform have been developed. The study employs both individual algorithms (Decision Tree, Logistic Regression, Neural Network, SVM, Naïve Bayes) and ensemble algorithms (Bagging, Boosting, and Stacking) to build predictive models. A significant challenge encountered was data imbalance in the target variable. The problem was addressed using the SMOTE method, leading to improved model performance. The results demonstrate that ensemble algorithms, particularly Boosted Decision Trees and Random Forests, outperform individual algorithms in predicting loan defaults. This study highlights the importance of ML in credit risk management and suggests that further research with more complex datasets and alternative methods like under-sampling could enhance predictive accuracy.
The work reported in [17] compares logistic regression with ensemble methods such as AdaBoost, Gradient Boosting, Random Forest, and Stacking using a real-world dataset from Lending Club, comprising over one million customers. The methodology includes data preprocessing, feature engineering, and the hold-out method for model evaluation. The results evaluated using accuracy, AUC, type I, and type II errors, indicate that ensemble methods outperform logistic regression, with AdaBoost showing the best trade-off among the metrics. This research underscores the potential of ensemble methods in enhancing the accuracy and robustness of loan default prediction models, highlighting the relevance of advanced ML techniques in the evolving landscape of credit risk management.
The study [18] focuses on predicting loan defaults using the Random Forest algorithm with Lending Club data handles class imbalance using SMOTE and conducts data cleaning and dimensionality reduction. Their experimental results demonstrate that the Random Forest algorithm outperforms other ML models such as logistic regression, decision tree, and SVM in predicting loan defaults, achieving higher accuracy and better generalization capabilities. The study concludes with suggestions for further research, including experiments on larger datasets and model tuning to achieve state-of-the-art performance.

1.3 Advanced Neural Network Techniques

The application of advanced neural network techniques, particularly Long Short-Term Memory (LSTM) models, has shown promise in improving the predictive accuracy of loan defaults. LSTM models, a recurrent neural network (RNN) capable of learning long-term dependencies, have been particularly effective in capturing the temporal dynamics of sequential financial data.
The work reported in [19] explores the use of deep learning methodologies to develop a predictive model to enhance credit risk identification, loan borrower review efficiency, and the accuracy of default predictions in the banking sector. Anchored in a theoretical framework that underscores the pivotal role of banks in economic intermediation and the significance of understanding factors like debt-to-income ratios and credit scores, the

research collected data from 1,000 participants across the top 11 banks in the UAE. Using Keras and TensorFlow, the researchers analyzed 625 records, leading to a predictive model that demonstrated a high accuracy of 95.2%, forecasting the default status of 238 out of 250 respondents. This success suggests that UAE banks could greatly benefit from integrating this model into their loan assessment processes to mitigate risks and enhance loan portfolio performance, thereby supporting financial stability and economic growth. The findings highlight the importance of adopting innovative deep-learning approaches in credit risk management to improve operational efficiency and decision-making in the financial sector.

The research reported in [20] explores the crucial undertaking of forecasting credit defaults, with great significance for financial institutions and stakeholders in mitigating financial uncertainties. To overcome the constraints associated with solely depending on credit metrics for predictions, the researchers advocate for a composite blending model that amalgamates several machine learning methodologies including Support Vector Machines (SVM), Bernoulli Naive Bayes, Decision Trees, Logistic Regression, Random Forests, and Gradient Boosting as foundational learners, complemented by a Deep Neural Network (DNN) as a meta-learner. This combined approach helps to leverage the unique capabilities of these heterogeneous algorithms to enhance the precision of credit default forecasts. By leveraging the South German Credit Data Set comprising 1,000 data entries spanning the years 1973 to 1975, encompassing a combination of 700 positive and 300 negative credit instances along with 20 predictor variables related to personal and credit agreement details, the model showcases improved predictive accuracy and F1 Score metrics. The outcomes highlight the model's adeptness in effectively scrutinizing voluminous multidimensional datasets, thereby aiding financial institutions in pinpointing high-risk borrowers and curtailing financial setbacks. Ultimately, this research advocates for a comprehensive credit scoring approach that takes into account a multitude of borrower characteristics, with the objective of refining loan security measures and credit allocation decisions, potentially bolstering credit risk management protocols in the banking domain.

Finally, in [21] a ResNet-LSTM-based strategy for credit scoring in case of imbalanced data is reported. The researchers utilize an auxiliary conditional tabular generative adversarial network (ACTGAN) to create extra default instances, balancing the dataset prior to a hybrid ResNet-LSTM model that captures both fixed demographic and financial attributes, alongside evolving behavioural trends over time. Empowered by spatiotemporal attention mechanisms, this model grasps crucial temporal and spatial interconnections. The investigation assesses the model using actual datasets and criteria such as AUC, recall, F1 score, and Kolmogorov-Smirnov (KS) value, discovering that the ResNet-LSTM framework, particularly in combination with XGBoost, outperforms traditional credit scoring algorithms significantly. The article also deliberates on data preprocessing techniques, encompassing data partitioning and dropout methods to avert overfitting, while underscoring the model architecture, which integrates multiple convolutional layers with ReLU activation and the Adam optimizer. The conclusion underscores the model's exceptional performance and proposes potential utilities in

geospatial evaluations and analysis of social network data, implying broader applicability beyond credit scoring.

2 Research methodology

2.1 SVM classifiers

In scenarios where the dataset exhibits non-linear relationships, the application of linear models might be insufficient for accurate predictions. To address this complexity, we employ Support Vector Regression (SVR) with a Gaussian Radial Basis Function (RBF) kernel. The Gaussian RBF kernel is particularly adept at handling non-linear data transformations, making it suitable for a wide range of applications including financial modelling such as predicting loan defaults.

Before training the SVR model, it is crucial to scale the features and target variables to ensure that the model is not biased towards variables with higher magnitude.

The SVR model is formulated with a Gaussian RBF kernel, defined by the kernel function:

$$K(x, y) = e^{-\gamma||x-y||^2} \quad (1)$$

The model parameters include the following:

- C: Regularization parameter that balances the trade-off between achieving a low error on the training data and minimizing the model complexity for better generalization. In this context, C=1.0
- e: Defines a margin of tolerance where no penalty is given for errors. This epsilon-insensitive loss is set at 0.10.10.1 in our implementation. [22] [23]

2.2 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It is particularly effective for regression tasks due to its ability to model complex interactions and non-linear relationships between features.

The Random Forest algorithm creates a forest of independently trained decision trees, each contributing to the final output. The mathematical representation of a Random Forest model used for regression is as follows:

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x; \Theta_b) \quad (2)$$

where:

- B is the number of trees in the forest
- $T_b(x; \Theta_b)$: represents the prediction of the b_{th} decision tree on input x, with Θ_b denoting the random parameters selected for that tree.

Each tree in a Random Forest is built from a bootstrap sample, that is, a randomly chosen subset of the training data. This method is known as bagging or bootstrap aggregating. During the training of each tree, a random subset of features is selected at each node, which makes the trees in the forest de-correlated and increases the diversity among the trees, enhancing the model's performance and robustness.

For regression tasks, the prediction from the Random Forest model is typically the average of the predictions from all the trees:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (3)$$

where $\hat{y}$ is the predicted value for input x and $T_b(x)$ is the prediction from the b_{th} tree. [24]

2.3 LSTM

Long Short-Term Memory (LSTM) networks, a specialized subclass of Recurrent Neural Networks (RNNs), are designed to process sequential data efficiently. These networks overcome the limitations of traditional RNNs, particularly the issue of vanishing gradients, by incorporating mechanisms that allow for retaining information over extended periods. The architectural sophistication of LSTMs enables them to link past information to current tasks, enhancing their capability to handle sequences where context and historical dependencies are crucial.

The LSTM network consists of various components structured in layers, including an input layer, one or more hidden layers, and an output layer. Central to its architecture are the memory blocks situated in the recurrent hidden layer. These blocks are composed of interconnected cells with four primary units: an input gate, a forget gate, an output gate, and a self-recurrent neuron. Each unit plays a vital role in the information flow within the network:

- **Input Gate**: Determines the quantity of the new input that should be incorporated into the cell state.
- **Forget Gate**: Decides the amount of information discarded from the previous cell state.
- **Output Gate**: Controls the extent to which the value in the cell state is used to compute the output activation of the LSTM unit.

These gates use sigmoid activation functions to regulate the flow of information by producing values between 0 and 1, indicating how much each component should pass through. The equations governing these gates are as follows:

- Forget gate

$$f_t = \sigma\left(b_f + U_f \cdot x_t + W_f \cdot y_{t-1}\right) \quad (4)$$

- Input gate

$$i_t = \sigma(b_i + U_i \cdot x_t + W_i \cdot y_{t-1}) \quad (5)$$

- Output gate

$$o_t = \sigma(b_o + U_o \cdot x_t + W_o \cdot y_{t-1}) \quad (6)$$

The cell states are updated by interactions of these gates, providing paths for the gradient to flow back through time and space without vanishing. This feature is depicted in the update of the cell state:

$$c_t = f_t * c_{t-1} + i_t * tanh(b_c + U_c \cdot x_t + W_c \cdot y_{t-1}) \quad (7)$$

The structure of an LSTM cell, showing the interactions between the input gate, forget gate, and output gate, and how information flows through the network is presented in Figure1.

The training of LSTMs utilizes the Backpropagation Through Time (BPTT) technique, which modifies weights in response to error gradients propagated backward through each step of the input sequence. However, LSTMs address the traditional challenges of BPTT, such as exploding or vanishing gradients, by stabilizing the gradient flow across learning steps.

The application-specific configuration of LSTM models, including the number of hidden neurons and the structure of recurrent units, is tailored based on the size of the input and output layers and the complexity of the task. For instance, the number of hidden neurons can be approximated by the heuristic

$$|F_H| = 2\left[\sqrt{(|F_Y| + 2) \cdot |F_X|}\right] \quad (8)$$

where $|F_X|$ and $|F_H|$ represent the sizes of the input and output layers, respectively.

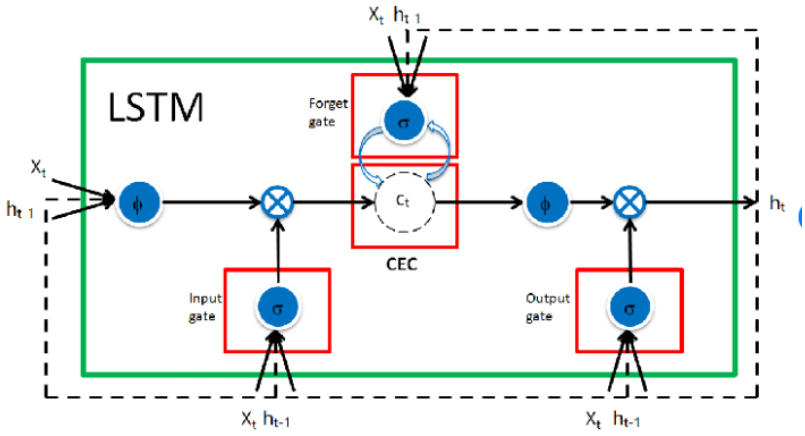

Figure no. 1: The diagram of an LSTM cell [25]
Source: Author's work.

In practical terms, LSTMs are trained using advanced optimization algorithms like the ADAM optimizer, renowned for their effectiveness in handling non-stationary objectives and computational efficiency. This choice of optimizer aids in rapid convergence and robust performance across various sequential modeling tasks, making LSTMs a preferred model in domains requiring nuanced understanding and prediction of temporal data patterns.

By integrating these sophisticated elements, LSTMs provide a powerful tool for modeling time-series data across diverse fields, from speech recognition to financial forecasting, underscoring their pivotal role in advancing the capabilities of neural network architectures. [26]

To practically implement the LSTM model for your predictive task, data scaling and model configuration play crucial roles. The Long Short-Term Memory (LSTM) model is defined utilizing the Keras framework, which includes an LSTM layer succeeded by dropout and dense layers aimed at reducing overfitting and facilitating the generation of final predictions, respectively. The model incorporates the 'relu' activation function within the LSTM layer to set up nonlinear transformations and employs the 'sigmoid' function in the output layer, which is particularly good for binary outcomes. The model is compiled by employing the Adam optimizer, recognized for its proficiency of handling non-stationary objectives, alongside a learning rate set at 0.001. The loss function implemented is 'binary_crossentropy', which is especially relevant for binary classification, as it quantifies the disparity between the predicted probabilities and the actual binary labels.

2.4 Data Collection and Accuracy Measures

In our research, we utilized a dataset comprised of 303 entries sourced from a Kaggle database. Each entry, in the dataset provides details on 13 factors concerning the borrower's history, such as credit score, income stability, and existing debts.

The dataset is evenly divided into 165 instances of defaults and 138 instances of non-defaults. This equal distribution enables us to assess the model without requiring any sampling techniques. In preparation for modeling with an LSTM framework, we split the data into training and testing sets. We split the data into 80% for training and we reserve 20% for assessing its accuracy.

We have also balanced the number of entries indicating loan defaults (`not.fully.paid` = 1) with those showing no defaults (`not.fully.paid` = 0). This method allowed us to create a dataset without favouring the majority class in our model. Furthermore, we transformed the data into a machine format using a Label Encoder making it easier, for training purposes, especially with an LSTM network.

2.5 Accuracy measures

Error measures and precision indexes usually assess the ability of a classifier to correctly assign data. In the case of binary classification, the most popular indicators are the mean squared error (MSE) metric and the F1 score [27].

Let $\{\hat{y}_i, 1 \leq i \leq N\}$ be the outcomes set of the binary classifier h. The MSE is defined as the average of the error squares:

$$MSE(h) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (9)$$

The F1 is an accuracy measure defined based on the Precision index and the Recall value as follows:

$$F1(h) = \frac{2}{\frac{1}{Precision(h)} + \frac{1}{Recall(h)}} \quad (10)$$

Moreover, we assess the model's performance using measures. We rely on the four acknowledged metrics to categorize loan default forecasts; True Positive (TP), True Negative (TN) False Positive (FP), and False Negative (FN). These measures play a role, in determining how well the model works.

$$Precision(h) = \frac{TP}{TP + FP}, \quad (11)$$

$$Recall(h) = \frac{TP}{TP + FN}, \quad (12)$$

Accuracy in this scenario is calculated by comparing the number of predicted outcomes (both defaults and non-defaults) to the predictions made in the dataset. This metric directly reflects the model's ability to determine the loan status. [28]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

3 Results and discussion

3.1. Standard Classifiers

In our study, we implemented and rigorously evaluated three commonly used machine learning models for binary classification tasks—namely, Long Short-Term Memory (LSTM), Random Forest, and Support Vector Regression (SVR). These models were employed to discriminate between the positive and negative classes within a loan default dataset, with the primary objective of determining the model that offers the best trade-off between accuracy, precision, and generalization capability.

The dataset was divided into a training set (80%) and a test set (20%) to ensure robust evaluation. We assessed each model's performance based on several key metrics: Mean Absolute Error (MSE), F1 score, and Accuracy. Table 1 provides a detailed summary of the classification performance metrics for both the test data and the entire dataset.

Table no. 1. Summary of classification performance metrics for the test data.

Model	Test_MSE_Mean	Test_MSE_Std	Test_F1_Mean	Test_F1_Std	Test_Accuracy_Mean	Test_Accuracy_Std
LSTM	0.374034447	0.04006672	0.332164706	0.007167125	0.625965553	0.04006672
RF	0.167536534	0	0.069565217	0	0.832463466	0
SVR	0.17223382	0	0.0677966 1	0	0.82776618	0

Source: Author's work.

The superior performance of the Random Forest model in terms of accuracy and MSE aligns with established theories in machine learning, particularly the ensemble learning theory, which posits that combining weak learners (i.e., decision trees) can result in a more robust and accurate model. This finding is consistent with previous research in the literature [13] [14]. Moreover, the relative success of the LSTM model in handling imbalanced data corroborates findings from sequential data modeling studies [19] [21], where LSTM's ability to capture long-term dependencies proves advantageous. SVR's struggles with class imbalance, however, underscore the limitations of kernel-based methods in imbalanced scenarios, as pointed out by prior [22]

3.2 Detailed Model Performance Analysis

The LSTM model, despite being an advanced neural network architecture known for its ability to capture temporal dependencies, exhibited relatively lower accuracy compared to the Random Forest and SVR models. However, its performance on the F1 score—a metric particularly relevant for imbalanced datasets—indicates that LSTM could potentially outperform the other models in scenarios where capturing sequence information is critical.

The Random Forest model outperformed the LSTM and SVR models in terms of accuracy, achieving a mean accuracy of 83.24% on the test data. This suggests that Random Forest, with its ensemble learning approach, can effectively handle the heterogeneity present in the dataset, leading to more consistent performance.

The SVR model, using a radial basis function (RBF) kernel, demonstrated competitive accuracy similar to Random Forest. However, it underperformed in terms of the F1 score, indicating that while SVR is capable of distinguishing between classes with high accuracy, it may struggle with class imbalance, leading to poorer precision-recall trade-offs.

To validate the statistical significance of the observed performance differences among the models, we conducted ANOVA tests on the F1 scores, MSE, and Accuracy metrics across all experimental runs. The ANOVA results show highly significant differences between the models.

The exceptionally low p-values (<0.001) across all metrics indicate that the differences in model performance are statistically significant. These findings underscore the necessity of model selection based on specific application needs, particularly when dealing with imbalanced datasets or those with complex temporal patterns.

To complement the statistical analysis, we provide visualizations that illustrate the performance of the classifiers across different metrics.

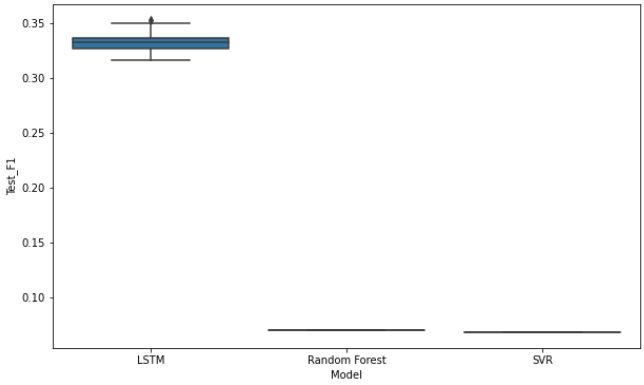

Figure no. 2: F1 Score – Test Data

Source: Author's work.

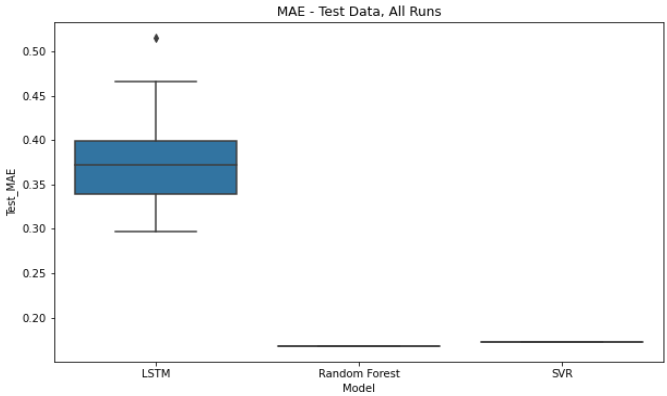

Figure no. 3: MSE – Test Data
Source: Author's work.

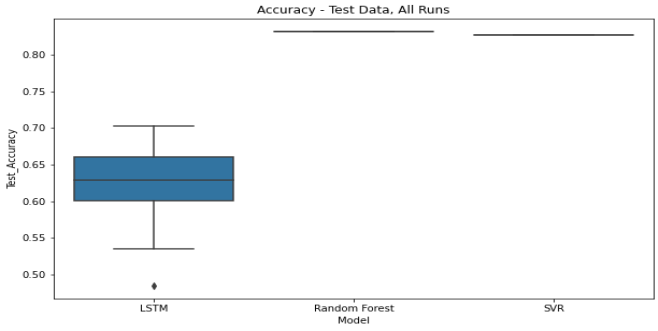

Figure 4: Accuracy – Test Data
Source: Author's work.

Figures 2-4 reveal that while the LSTM model struggles in terms of MSE and Accuracy, it maintains a relatively consistent F1 score, suggesting potential in applications where minimizing false positives and false negatives is crucial. The Random Forest model's dominance in accuracy and MSE, as visualized, is evident, but this comes at the cost of a lower F1 score, particularly when compared to LSTM.

Conclusions

This investigation has examined the relative efficacy of LSTM, Random Forest, and SVR algorithms within the binary classification framework for predicting loan defaults. The findings demonstrate that although the Random Forest algorithm exhibits higher accuracy

and reduced MSE, the LSTM algorithm reveals potential in managing imbalanced datasets due to its stable performance in F1 score metrics.

Results derived from ANOVA analyses affirm that the variations in performance indicators across the algorithms are statistically significant, implying that each algorithm possesses unique advantages that render it appropriate for various datasets or particular application scenarios.

Nevertheless, with these encouraging outcomes, it is crucial to acknowledge that both the LSTM and SVR algorithms remain subjects of ongoing research. Specifically, the LSTM algorithm shows considerable promise for enhancement, particularly through further optimization and improvement of its temporal learning functionalities. Likewise, the performance of the SVR algorithm could be bolstered by mitigating its susceptibility to class imbalance, potentially through the incorporation of more advanced kernel techniques or data-balancing strategies.

Despite the promising results of this comparative analysis, it is important to recognize some limitations that may impact the generalizability of the findings. First, the relatively small dataset used in this study limits the ability to capture the full complexity of loan default patterns seen in larger financial institutions. Additionally, while the models were effective at handling imbalanced data to some extent, further refinement in addressing class imbalance—such as through advanced resampling techniques or cost-sensitive learning—could significantly improve performance. Moreover, the hyperparameter tuning in this research was performed manually, which could be enhanced by applying automated methods like Tree-structured Parzen estimator [27], Bayesian optimization [34], or evolutionary approaches [35] to optimize model performance further. These considerations should be factored into future iterations of this research to ensure broader applicability and more robust results in real-world scenarios.

Further research will emphasize the refinement of these algorithms, notably the LSTM, which stands to gain from sophisticated methodologies such as evolutionary-based hyperparameter tuning and more intricate network architectures. Furthermore, investigating ensemble techniques that amalgamate the strengths of these algorithms may lead to additional advancements in classification efficacy. This research will be further developed to establish robust, accurate, and dependable algorithms in practical, real-world contexts where data imbalance and temporal variations are common.

References

[1] Abe, S., 2010. Support vector machines for pattern classification, Advances in Pattern Recognition; Springer: Dordrecht, The Netherlands; pp. 473.

[2] Abe, S., 2010. Support vector machines for pattern classification. 2nd ed. London: Springer.

[3] Abe, S., 2010. Support vector machines for pattern classification. London: Springer.

[4] Addo, P.M., Guegan, D. and Hassani, B., 2018. Credit risk analysis using machine and deep learning models. MDPI Journal of Financial Studies, 10(3), pp.155-170.

[5] Barbaglia, L., Manzan, S. and Tosetti, E., 2022. Forecasting loan default in Europe with machine learning. Journal of Forecasting, 39(4), pp.601-620.

[6] Bishop, C.M., 2006. Pattern recognition and machine learning. New York: Springer.

[7] Breiman, L., 2017. Classification and Regression Trees, Routledge.

[8] Chaganti, R., Mourade, A., Ravi, V., Vemprala, N., Dua, A. and Bhushan, B., 2022. A particle swarm optimization and deep learning approach for intrusion detection system in Internet of Medical Things. Sustainability, 14(5), pp.1-15.
[9] Cocianu, C.L., Uscatu, C.R., Kofidis, K., Muraru, S. and Văduva, A.G., 2023. Classical, evolutionary, and deep learning approaches of automated heart disease prediction: A case study. Electronics, 12(7), p.1663. Available at: <https://doi.org/10.3390/electronics12071663> [Accessed 15 Oct. 2024].
[10] Egwa, A.A., Habeeb, B., Ahmad, A.A. and Bizi, S.M., 2022. Prediction of default for loan lenders using machine learning algorithms. Scientific Research Journal of Computer Science and Data Analytics, 10(4).
[11] Frazier,, P.I., 2018 Bayesian Optimization. INFORMS TutORials in Operations Research, pp.255-278.
[12] García, S., Luengo, J. and Herrera, F., 2015. Data preprocessing in data mining. Cham: Springer.
[13] Gashi, P. and Ahmeti, D.A., 2023. Loan default prediction model. International Journal of Artificial Intelligence and Applications, 14(1), pp.12-24.
[14] Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer.
[15] Hastie, T., Tibshirani, R. and Friedman, J., 2017. The elements of statistical learning. 3rd ed. New York: Springer.
[16] He, H. and Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 20(9), pp.1263-1284.
[17] Jumaa, M.S., Aziz, M.A. and Mohammed, S., 2023. Improving credit risk assessment through deep learning-based consumer loan default prediction model. International Journal of Finance & Banking Studies, 10(2), pp.55-68.
[18] Kohavi, R., 2001. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, 14(2), pp.1137-1143.
[19] Lipton, Z.C., Berkowitz, J. and Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. arXiv: Computation and Language. Available at: <https://arxiv.org/abs/1506.00019> [Accessed 15 Oct. 2024].
[20] Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., ... & Hodjat, B. (2024). Evolving deep neural networks. In Artificial intelligence in the age of neural networks and brain computing (pp. 269-287). Academic Press.
[21] Noriega, J., Rivera, L. and Herrera, J.A., 2023. Machine learning for credit risk prediction: A systematic literature review. MDPI Journal, 8(11), pp.333-348.
[22] Praynlin, E. and Ramanathan, V., 2023. Prediction of loan default prediction using machine learning techniques. International Journal of Computer Applications, 185(43).
[23] Ranjan, C., 2019. Understanding dropout with the simplified math behind it. Towards Data Science. Available at: <https://towardsdatascience.com/understanding-dropout-with-the-simplified-math-behind-it-825ab8027a4b> [Accessed 15 Oct. 2024].
[24] Robisco, A.A. and Martínez, J.C., 2022. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. Financial Innovation, 8(1), pp.1-12.
[25] Sheela, K. and Deepa, S., 2013. Review on methods to fix number of hidden neurons in neural networks. Mathematical Problems in Engineering, 2013(425740), pp.1-11.

[26] Simão, S.B.S., 2023. Machine learning applied to credit risk assessment: Prediction of loan defaults. Journal of Financial Technology, 18(2), pp.55-68.
[27] Su, Y. and Kuo, C., 2019. On extended long short-term memory and dependent bidirectional recurrent neural network. Neurocomputing, 356, pp.243-256.
[28] Tian, J. and Li, L., 2022. Digital universal financial credit risk analysis using particle swarm optimization algorithm with structure decision tree learning-based evaluation model. Wireless Communications and Mobile Computing, 2022, pp.1-9.
[29] Uwais, A.M. and Khaleghzadeh, H., 2023. Loan default prediction using Spark machine learning algorithms. International Research Journal of Engineering and Technology (IRJET), 10(3), pp.1165-1175.
[30] Vincenzo, F.D., Imbriani, V., Mercuri, G. and Mosca, P., 2022. Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation. International Journal of Financial Studies, 10(4).
[31] Yanxiao, L., Yinbao, T. and Jianan, Z., 2023. Credit default prediction based on blending learning model. Applied and Computational Engineering, 11(3), pp.123-145.
[32] Zhang, A., Peng, B., Chen, J., Liu, Q., Jiang, S. and Zhou, Y., 2022. A ResNet-LSTM based credit scoring approach for imbalanced data. Mobile Information Systems, 2022, pp.1-14.
[33] Zhou, L. and Dai, Z., 2021. Machine learning-based models for predicting loan default risk: A comparison study. Journal of Financial Risk Management, 10(2), pp.112-129.
[34] Zhou, Y., 2022. Loan default prediction based on machine learning methods. Journal of Financial Services Research, 49(2), pp.95-110.
[35] Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K., 2019. A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, pp.503-513.