IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Credit Score Prediction using Machine Learning

Rachit Deven Modi[1] and Dr. Pallavi Devendra Tawde[2]
Student, Department of Msc.IT[1]
Assistant Professor, Department of IT[2]
Nagindas Khandwala College, Mumbai, Maharashtra, India
rachitmodi12@gmail.com and pallavi.tawde09@gmail.com

Abstract: *Credit score prediction plays a crucial role of financial risk assessment and for making informed decisions. This paper gives the effectiveness of various machine learning (ML) algorithms in predicting credit scores by using a dataset obtained from Kaggle. The dataset has financial attributes such as income, credit history, outstanding debt, credit score. This research involves data preprocessing, feature selection, model training, and evaluation of multiple ML algorithms, including logistic regression, decision trees, random forests, and neural networks. Results provides that Support Vector Machine (SVM) gives the highest accuracy of 87%. Additionally, feature importance analysis helps to identify that income is the most significant factor influencing credit scores, while loan history has the least impact. The findings contribute to the optimization of credit score prediction models for financial institutions.*

Keywords: Credit Score, Machine Learning, Algorithms, Prediction.

I. INTRODUCTION

Credit rating is essential since it affects interest rates, loan approvals, and stability. An individual's creditworthiness is assessed by a credit score, which is a numerical representation that takes into account a number of variables, including prior credit behavior, loan repayment history, and outstanding debt. This score is used by lenders to evaluate the risk of making a loan to a person or company. Conventional credit scoring algorithms, such expert-based scoring systems and logistic regression, frequently have trouble processing large amounts of complicated data. These models may not adjust well to changing financial circumstances since they rely on manually created features and preset criteria.

By using advanced methods to find patterns, uncover hidden relationships, and provide data-driven forecasts, machine learning presents a possible substitute. In contrast to traditional techniques, machine learning (ML) models have the ability to handle massive datasets, automatically identify pertinent characteristics, and constantly increase prediction accuracy. More accurate risk assessment is made possible by machine learning in credit scoring, which also lessens the possibility of human bias in judgment. This strategy is very helpful for spotting possible defaulters, streamlining the loan approval procedure, and enhancing financial inclusion.

Credit scoring models are becoming even more capable because to recent developments in data science and artificial intelligence (AI). Big data from several financial sources is becoming more widely available, and as a result, ML-based credit scoring models are improving in accuracy and efficiency.

This study compares many machine learning models and evaluates how well they predict credit scores, building on earlier studies. This study attempts to illustrate the benefits and difficulties of machine learning applications in credit risk assessment by putting several ML algorithms into practice and assessing them.

II. RESEARCH OBJECTIVES

1: Predicting Credit Score using Regression Analysis
2: Identifying Key Factors Affecting Credit Score using Feature Importance

III. LITERATURE REVIEW

Previous research has focused on traditional statistical methods like linear regression and discriminant analysis for credit scoring. However, these approaches often struggle with high-dimensional data and complex relationships. Recent studies have explored ML-based techniques:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

- **Golbayani et al. (2020):** This study compared neural networks, support vector machines, and decision trees in forecasting corporate credit ratings. The findings indicated that decision tree-based models, particularly bagged decision trees and random forests, outperformed other techniques in predictive accuracy.
- **Biecek et al. (2021):** The authors explored the integration of Explainable Artificial Intelligence (XAI) methods to enhance the interpretability of complex ML models in credit scoring. They demonstrated that advanced tree-based models, when combined with XAI techniques, offer both high predictive performance and transparency, addressing the "black box" nature of ML models.
- **Perera (2022):** This research introduced a search-based fairness testing approach for regression-based ML systems, including those used in credit scoring. The study emphasized the importance of assessing and ensuring fairness in ML models to prevent biased credit evaluations.
- **Feng(2023):** The study proposed the use of Large Language Models (LLMs) for credit scoring tasks, highlighting their potential to generalize across multiple tasks and datasets. The authors introduced an open-source framework and benchmark for evaluating LLMs in credit assessment, demonstrating that LLMs can match or surpass traditional models in predictive performance.
- **Rida (2024):** The author explored the deployment of ML models, specifically Gradient Boosting Machines, in credit scoring within the constraints of Basel II and III regulations. The research demonstrated that these models significantly enhance performance and default capture rates, utilizing Shapley Values to interpret model outputs and ensure compliance with financial oversight requirements.

IV. METHODOLOGY

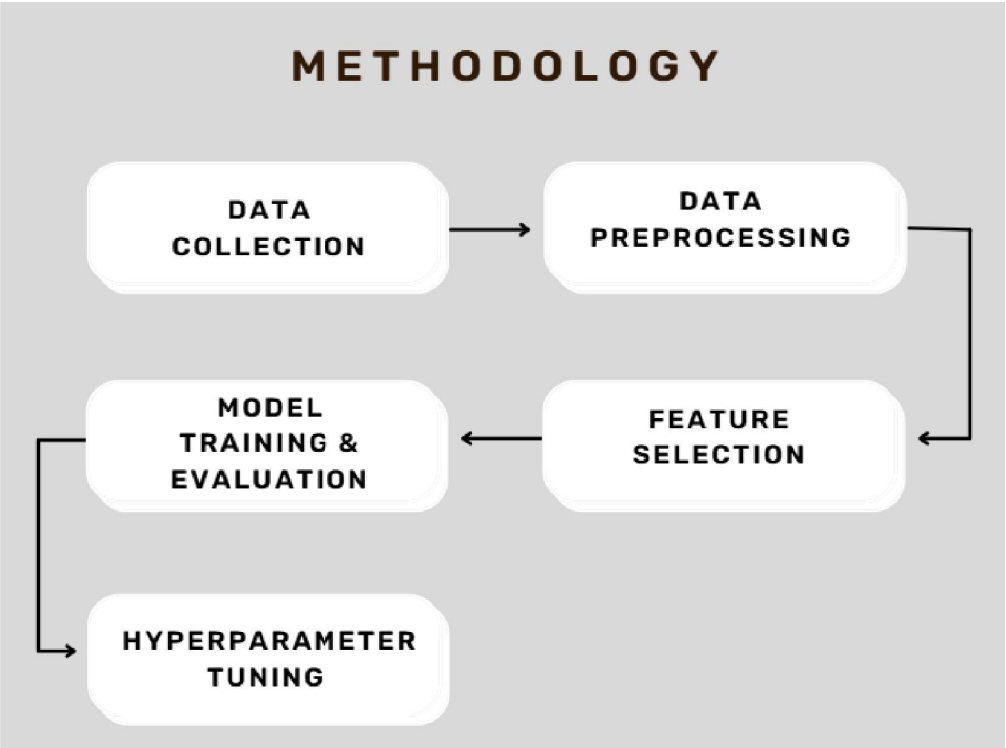

Fig 1: Methodology

This study employs a comparative analysis of ML algorithms for credit score prediction. The methodology involves the following steps:

1. **Data Collection**: Credit Score Dataset is used from Kaggle
   Column – Age, Employee, Income, Education, Debt to Income, Credit_utilization, Loan_Hitstory,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Credit_Score.

Rows of data – 1200 Rows

2. **Data Preprocessing**: Missing values are handled, categorical features are encoded, and data is normalized.
3. **Feature Selection**: Relevant features are selected using statistical tests and ML-based techniques.
4. **Model Training and Evaluation**: Various ML algorithms, including logistic regression, decision trees, random forests, and neural networks, are implemented and evaluated using accuracy, precision, recall, and F1-score.
5. **Hyperparameter Tuning**: Grid search and cross-validation techniques are applied to optimize model performance.

V. RESULTS

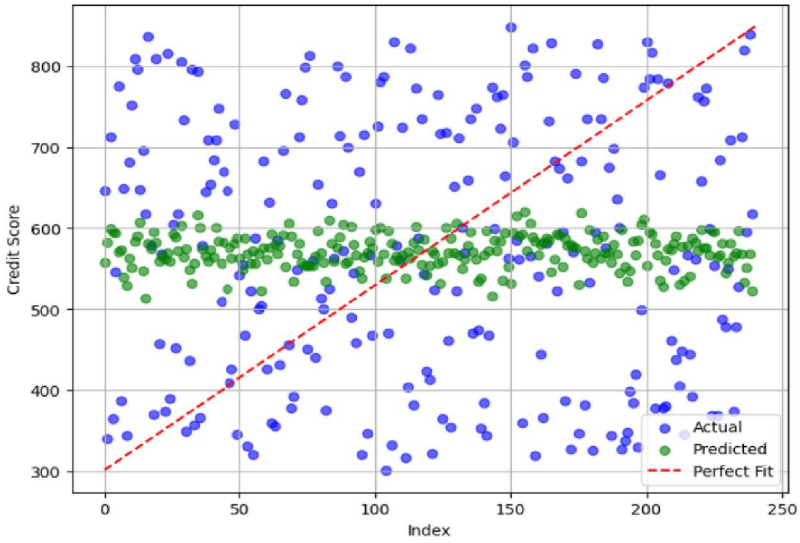

Fig 2: Actual vs Predicted Credit Score

Predicting Credit Score using regression analysis implemented Actual vs Predicted credit score using scatter plot and calculated mean square error and r2 scores. Where MSE score is 25303.713472465788 and R2 score is 0.000136. In diagram there are two color blue color shows actual credit score and green color shows predicted credit scores

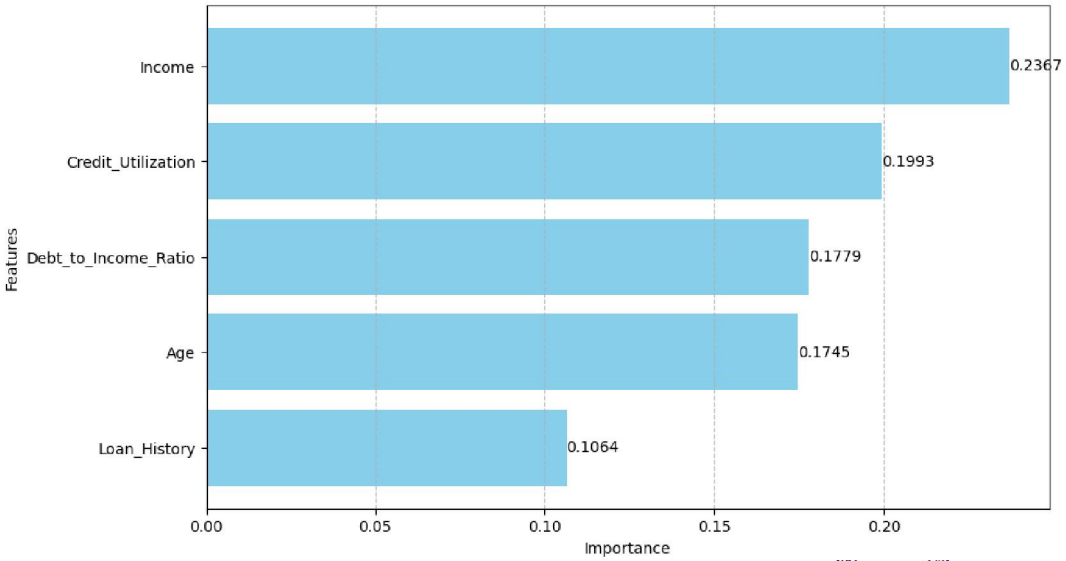

Fig 3: Top 5 Factors Affecting Credit Score

DOI: 10.48175/IJARSCT-23348

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Identified top 5 factors affecting credit scores using feature importance with the help of bar diagram and numbers we can visualize income is the highest affecting factor with 0.2367 to the credit score then the other factors as follows.

	Precision	Recall	F1 Score	Support
Poor	0.85	0.87	0.86	120
Average	0.89	0.89	0.89	160
Good	0.86	0.85	0.85	120
Accuracy	0.87	0.87	0.87	400
Macro avg	0.87	0.87	0.87	400
Weighted avg	0.87	0.87	0.87	400

Table 1: Classification Report

The above Classification report helps us to better understand about the models accuracy, precision, recall, f1 score and support it also helps for future reference for making decision based on credit scores. It show the overall accuracy of different algorithms is 87%.

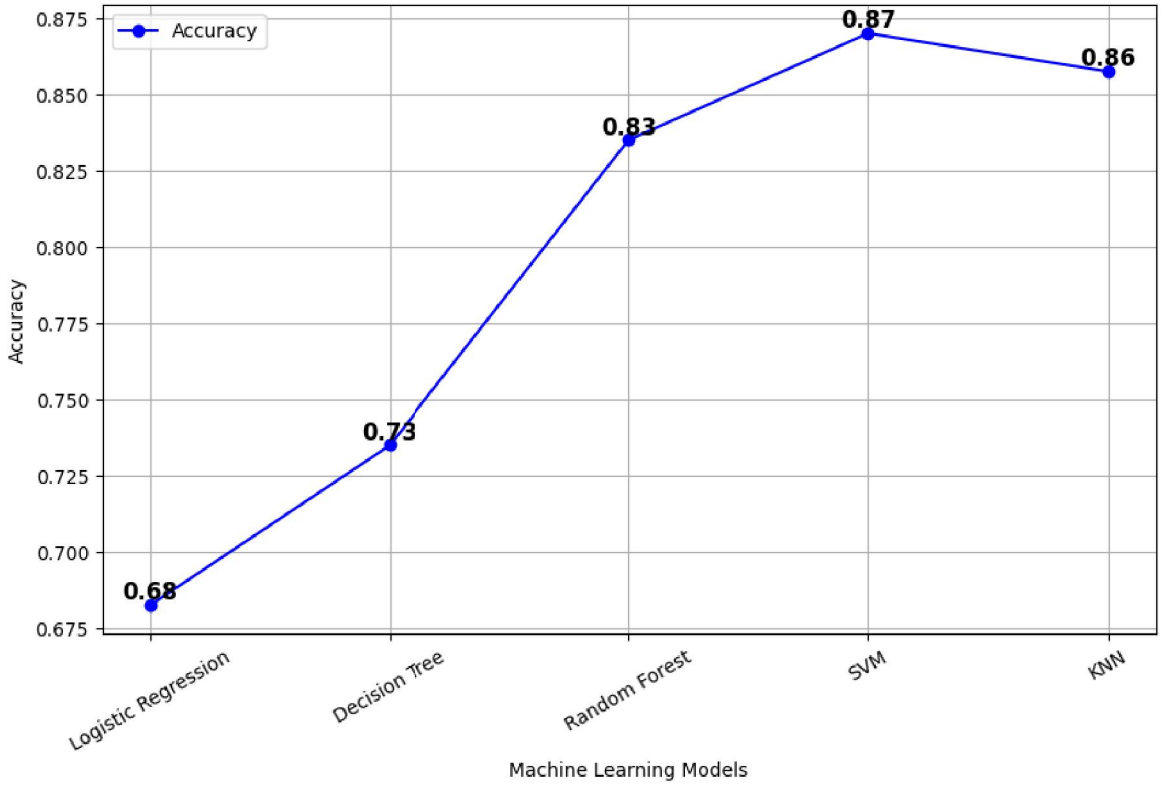

Fig 5: Line Graph Accuracy Comparison of Different Algorithms

The above line graph with shows the accuracy and percentage of different machine learning algorithms it indicates the Support vector Machine (SVM) has the highest accuracy of 87% followed by KNN which has 86% accuracy and as follows.

VI. CONCLUSION

This study evaluates the effectiveness of multiple machine learning algorithms in predicting credit scores based on financial and demographic attributes. The results indicate that SVM model achieves the highest accuracy of 87%, making it the most reliable model for credit score prediction. Feature importance analysis reveals that income significantly impacts credit scores, while loan history has the least influence.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

The research highlights the potential of ML techniques in improving credit risk assessment and aiding financial institutions in making data-driven decisions. Future work could explore larger datasets, incorporate deep learning approaches, and refine feature selection techniques to further enhance predictive accuracy.

The paper also emphasizes the benefits of machine learning for credit scoring, such as its capacity to manage big datasets, adjust to shifting financial trends, and lessen human bias in judgment. For broad implementation, nevertheless, issues like data privacy, model interpretability, and regulatory compliance has to be removed.

To increase the transparency and equity of credit scoring algorithms, future studies might investigate the combination of Explainable AI (XAI) with deep learning methodologies. Predictive accuracy and financial inclusion might also be improved by utilizing real-time data and alternate credit data sources, such as transaction history and social behavior.

REFERENCES

[1]. Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques, and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59-88.

[2]. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 66(6), 740-755.

[3]. Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. Omega, 31(2), 83-96.

[4]. Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit scoring and its applications. SIAM.

[5]. Zhang, D., Li, X., & Wang, S. (2019). Machine learning for credit risk prediction: A survey. IEEE Access, 7, 150199-150222.

[6]. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

[7]. Baesens, B., Setiono, R., Mues, C., &Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science, 49(3), 312-329.

[8]. Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36(2), 3302-3308.

[9]. Chen, M., &Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[10]. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480