Predict Credit Score with Four Models

BHAKTI PALKAR, AASHLESHA SHIROLE, DIPALI AVHAD

Abstract-

This study tackles the critical need for financial institutions to improve their credit scoring systems, which are essential to the lending sector. This study focuses on creating and enhancing credit scoring models using creative methods because it acknowledges the critical role that credit scoring plays in reducing risks and encouraging responsible lending. By utilizing machine learning techniques, particularly ensemble classification algorithms, the study seeks to offer a thorough assessment of creditworthiness. The inquiry begins by outlining how personal credit is changing and demonstrating how sophisticated and adaptable credit assessment systems are necessary. The research, dedicated to progressing the area, suggests a unique credit scoring algorithm designed for top financial organizations. Through the integration of machine learning algorithms, such as Random Forest and Extratree, the study seeks to improve predictive accuracy and identify creditworthy persons more accurately. This study's comparative analysis evaluates the performance of several classifiers using error measures like MAE and RMSE in addition to metrics like recall, precision, F-measure, and accuracy. As a solid basis for experimentation, the research uses the Australian credit dataset from the UCI Machine Learning repository, guaranteeing the applicability and generalizability of the suggested credit scoring model.

In the end, our study adds to the continuing conversation about credit scoring by providing financial institutions with useful information about implementing innovative approaches that keep up with the quick advancement of data science and technology. This study aims to set the stage for top institutions to adapt and prosper in this dynamic environment as the financial landscape continues to change, making it vital to improve credit scoring methods.

Literature Review-

1. **An ensemble classifier model to predict credit scoring -comparative analysis.**
   The supplied research paper's literature review examines the field of credit scoring research with an emphasis on base and ensemble classification methods. The paper discusses prior research on credit score prediction by He et al., Cheon J H, Cortes and Vapnik, Jensen et al., Cheng-An Li, Siami et al., and others. These studies used various algorithms, including logistic regression, decision trees, support vector machines, neural networks, and k-nearest neighbors. Notably, ensemble techniques are recognized for their capacity to improve prediction performance, as demonstrated by Ala'raj et al., Lessmann S et al., Whitehead M et al., Guo et al., and Tabik et al. The literature analysis highlights how credit scoring does not take into account specific algorithms such as Random Forest, Bagged Decision Tree, and Extra Tree Classifier. In order to close this gap, the current study compares these algorithms against pre-existing ones in an effort to determine which model is the most accurate at predicting credit scores.

2. **Credit risk scoring analysis based on machine learning models.**
   A notable change in modeling approaches is highlighted by the literature on credit scoring in the setting of big data. While credit scoring algorithms of the past relied on credit history, modern models make use of large datasets that include a variety of user data. This change indicates how financial environments are changing and how institutions are using additional personal information for a more thorough credit

evaluation, such as property characteristics and length of work. To handle these complex datasets, feature engineering and machine learning models—most notably LightGBM—have become indispensable tools. The research emphasizes how historical credit data and alternative credit-related information interact dynamically, highlighting the necessity for complex models that may extract expressive features. The literature indicates a desire for precise, understandable, and flexible credit scoring models to handle the difficulties presented by the integration of big data as academics experiment with different approaches. The Kaggle Home Credit Default Risk dataset is a useful tool for assessing these creative strategies since it shows how credit scoring techniques are always changing to meet the needs of modern financial environments.

3. **Variable Selection for Credit Risk Scoring on Loan Performance using regression analysis.**
   Within the framework of the Department of Science and Technology VII Small & Medium Enterprise Technology Upgrading Program (DOST VII-SETUP), the literature study explores the dynamic field of credit risk analysis. The article emphasizes the need for strong financial strategies against the backdrop of the global financial crisis and its implications for small and medium-sized firms (SMEs) in emerging nations like the Philippines. It draws attention to the way the government is responding with programs like DOST-SETUP, which are designed to improve the financial standing of SMEs. The report highlights the critical role that credit risk scoring plays in the painstaking credit appraisal processes that financial institutions undergo, while also acknowledging the information constraints they encounter. The article demonstrates its dedication to utilizing technology improvements for well-informed decision-making through the application of linear regression analysis to specific variables and the integration of data analytics tools, such as Tableau. In the area of SME financing and economic development, the suggested decision matrix for credit risk scoring becomes an essential component of a future Credit Risk Analysis and Recommendation System, indicating a progressive move toward more advanced and data-driven methods.

4. **A local binary social spider algorithm for feature selection in credit scoring model.**
   The research on credit risk assessment and scoring models emphasizes how crucial it is to manage the risk of borrower default, especially in the context of online lending. Previous studies have largely focused on improving evaluation techniques, possibly ignoring the careful scrutiny of credit data quality. This body of work highlights the importance of feature selection to reduce computational complexity and improve model performance, acknowledging the common problem of noisy and redundant features in online credit databases. The research presents the Local Binary Social Spider Algorithm (LBSA) as a novel solution to the issues raised by the widely used BinSSA. LBSA attempts to mitigate extreme dispersion at initialization and the risk of falling into local optima during iteration by combining opposition-based learning with an enhanced local search algorithm. The better performance of LBSA in lowering feature redundancy and enhancing the general effectiveness and accuracy of credit scoring models is demonstrated by evaluation against a variety of online credit datasets. The body of research highlights the crucial role that logical feature selection plays in improving the accuracy of credit scoring, especially in the ever-changing world of online lending with its intricate borrower-lender relationships.

5. **Technology credit scoring model with fuzzy logistic regression.**
   The body of research on credit scoring models, especially as it relates to small and medium-sized businesses (SMEs), consistently emphasizes the need to improve assessment techniques in order to reduce default risks. Conventional methods, which frequently utilize logistic regression models, concentrate on pre-established evaluation characteristics in order to evaluate creditworthiness. But the rapidly changing internet lending scene presents new difficulties as well, emphasizing how critical it is to address credit

data quality issues in addition to refining assessment methods. Interestingly, the majority of the research to date has focused on improving assessment techniques, frequently ignoring the influence of credit data quality. In this regard, combining data mining and analytics becomes an effective way to find information that is hidden in organizational databases. Data selection, preparation, transformation, mining, and result interpretation and evaluation are some of the steps involved in this process. Credit scoring algorithms are greatly aided by data analytics approaches, particularly those that have their roots in data science.

Introduction-

The group of machine learning models used in the credit score prediction project was carefully selected to handle different facets of creditworthiness evaluation. By combining the results from several decision trees, Random Forest—which is renowned for its proficiency with high-dimensional data and nonlinear relationships—provided reliable forecasts. In order to produce a more precise and trustworthy credit score prediction, voting, as an ensemble technique, combined the intelligence of several models, such as Decision Tree, KNN (K-Nearest Neighbors), XGBoost, Logistic Regression, and AdaBoost. Decision Trees made the decision-making process more transparent and made it easier to understand the important factors that affect credit scores. KNN helped to capture localized relationships in the data by using proximity-based patterns. With its gradient boosting architecture, XGBoost enhanced prediction performance step by step. AdaBoost concentrated on iteratively improving the model by highlighting cases of misclassification, while Logistic Regression provided a baseline for binary classification tasks. The project's goal was to develop a comprehensive credit scoring system that could handle intricate datasets, offer insightful analysis, and produce precise predictions for better financial decision-making by merging these many methods.

Related work-

The Random Forest algorithm, central to our credit score prediction project, entails constructing multiple decision trees and aggregating their predictions through an ensemble approach. Decision trees are built on random subsets of training data, and each tree contributes to the final prediction through voting. Entropy, a measure of impurity, guides decision tree splitting, with the formula $H(S) = -p1 \log2(p1) – p2 \log2(p2)$, where p1 and p2 represent the proportions of the two classes. K-Nearest Neighbors (KNN) classifies based on data point proximity, with predictions influenced by the majority class among the k-nearest neighbors. XGBoost, an ensemble algorithm utilizing boosting, balances model complexity and data fit in its objective function. Logistic Regression predicts class probabilities using the sigmoid function. AdaBoost combines weak learners, assigning weights to misclassified instances to enhance subsequent iterations. These concepts and formulas form the foundation for our comprehensive approach to credit score prediction, ensuring a robust and accurate model.

Preferential method –

1. **Random forest-**
   One effective ensemble learning method that is frequently used for both classification and regression problems is the Random Forest algorithm. Building several decision trees and combining their predictions is the fundamental notion underlying Random Forest, which aims to produce results that are more reliable and accurate. A random portion of the training data is used to build each decision tree, and they are all autonomous. Every tree "votes" for the ultimate result during prediction; the class

or average that the majority of trees forecast becomes the final prediction. By using an ensemble technique, overfitting is lessened and the model's capacity for generalization is enhanced. Random Forest is a well-liked option in many machine learning applications because of its adaptability, scalability, and capacity to manage complicated datasets. In order to optimize the information gain or decrease in impurity, each decision tree is created during the training phase by recursively partitioning the data based on the most informative characteristics. The ultimate forecast is then ascertained by combining the individual forecasts from every tree, either by means of an average for regression issues or a majority vote for classification problems.

When it comes to managing numerical and categorical information, processing high-dimensional datasets, and revealing the significance of individual aspects, Random Forest is quite effective. For machine learning tasks like feature selection, regression, and classification, it is a popular choice because of its scalability, resilience, and capacity to uncover intricate relationships in the data. Furthermore, Random Forest's low hyperparameter tuning requirements add to its use and adaptability across a range of fields.

2. **Entropy**

In the context of decision tree splitting, ntropy refers to a data set's degree of disorder or impurity. It is frequently employed to measure the degree of uncertainty surrounding a data point's classification. The following is the expression for the entropy formula, which comes from information theory:
$H = p_1 \log_s(1/p_1) + p_2 \log_s(1/p_2) + \cdots + p_k \log_s(1/p_k)$.

$$Entropy\,(p) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

The formula calculates the entropy by summing over all classes (i=0 to c) the product of the proportion of instances pi belonging to class i and the base-2 logarithm of pi . The negative sign is used to ensure that entropy is always non-negative.
When building a decision tree, the algorithm evaluates the entropy of different splits and aims to minimize it. A split with lower entropy indicates better purity, meaning that the resulting subsets are more homogenous in terms of class labels.
To put it briefly, entropy plays a key role in decision tree algorithms because it directs the process of choosing the appropriate features and thresholds for data splitting, resulting in the creation of a tree that best divides classes according to the available features.

3. **K-Nearest Neighbors (KNN)**

K-Nearest Neighbors (KNN) is a classification algorithm that makes predictions based on the proximity of data points in a feature space. The algorithm's decision for a new data point is determined by the majority class among its k-nearest neighbors, where k is a user-defined parameter.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

After calculating distances, the algorithm identifies the k training instances with the shortest distances to the new data point. The class labels of these k neighbors are considered, and the majority class among them is assigned to the new data point as its predicted class. This approach reflects the intuition that instances with similar feature values are likely to belong to the same class. The choice of k influences the algorithm's sensitivity to noise and its ability to capture local patterns in the data.

4. **XGBoost:**

XGBoost, also known as eXtreme Gradient Boosting, is an ensemble learning technique that builds a reliable and accurate predictive model by aggregating the predictions of several weak learners, typically decision trees. Based on boosting, the method trains weak learners in a sequential fashion, with each new learner correcting the mistakes caused by the ensemble of previous learners.

- **Loss Function** : The loss function quantifies how well the model predicts the target variable. It measures the difference between the predicted values and the actual values. Common choices for regression tasks include mean squared error, while for classification tasks, it may be log loss (cross-entropy).
  L = sum(y_i*log(P_i)+(1-y_i)*log(P_i))

- **Regularization Term (Ω) :**
  It penalizes the complexity of the model to prevent overfitting. It is a combination of L1 and L2 regularization terms and is defined as Ω(F) = λ * Ω1(F) + 0.5 * γ * Ω2(F), where Ω1(F) is the L1 norm of the leaf weights and Ω2(F) is the L2 norm of the leaf weights.
  The ultimate goal is to reduce the loss of the regularization term and the model together for all weak learners. The algorithm iteratively updates the model and enhances its forecasting capabilities using gradient descent techniques. XGBoost is a strong and popular method in machine learning competitions and real-world applications because it strikes a balance between fitting the data effectively and preventing overfitting.

5. **Logistic Regression:**

The specific method of Logistic Regression is as follows: find a suitable hypothesis *h*, which is a function that needs to be classified to predict the judgment result of the input data. A cost function (loss function) is then constructed, which represents the deviation between the predicted output *h* and the training data category *y*, which can be different between *h* and *y* ($h - y$) or other forms between the two. Considering the "loss" of all training data, the Cost is summed or averaged and recorded as a $J(\theta)$ function, indicating the deviation of the predicted values of all training data from the actual category. The smaller the value of the $J(\theta)$ function, the more accurate the prediction function (i.e. the more accurate the *h* function), and the minimum value of the $J(\theta)$ function can be found by the Gradient Descent method.

6. **AdaBoost:**

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that combines the predictions of weak learners (usually simple models called weak classifiers) to form a strong learner. The algorithm assigns different weights to each instance in the training data, and it focuses on improving the classification of instances that are misclassified by the current ensemble in subsequent iterations.

- **Initialize Weights:** Assign equal weights to all training instances so that each instance has an equal chance of being selected.

- **Train Weak Classifier:** Utilizing the training data, fit a weak classifier and calculate the error, which is the weighted sum of misclassifications. Usually, the weak classifier is a straightforward model, such as a decision stump.

- **Compute Classifier Weight:** Based on its accuracy, give the weak classifier a weight. Higher weights are assigned to classifiers with greater accuracy, reflecting their significance within the ensemble.
- **Update Instance Weight:** Increase the weights of misclassified instances, making them more likely to be chosen in the next iteration. This emphasizes instances that are challenging for the current ensemble.

Through the repeated adjustment of misclassified instance weights, AdaBoost aims to improve the model's performance and make the following weak classifiers more attentive to the difficult cases. An effective ensemble may be produced by AdaBoost even from weak base learners thanks in part to this adaptive weighting approach.

Experiment –

Data set –

Data Attributes: The dataset comprises essential attributes that play a crucial role in understanding credit scoring model. These attributes include: Demographic Details, Financial Information, Loan and Credit Details, Credit Profile, Payment and EMI Details, Monthly Balance and Credit Score

Data Dimensions:

1. This dataset contains a total of 100000 rows and 28 columns.
2. These dimensions reflect the size of the dataset and the number of data points and attributes available for our analysis.

3. ID: Unique identifier for each record.

4. Customer_ID**:** Identification for individual customers.

5. Month**:** Timeframe of data collection.

6. Demographic Details:
    a. Name, Age, SSN, Occupation**:** Personal and professional information.
7. Financial Information:
    a. Annual_Income, Monthly_Inhand_Salary: Key indicators of financial health.
    b. Num_Bank_Accounts, Num_Credit_Card**:** Count of banking and credit accounts.
8. Loan and Credit Details:
    a. Interest_Rate, Num_of_Loan, Type_of_Loan: Parameters related to loans.
    b. Delay_from_due_date, Num_of_Delayed_Payment**:** Instances of payment delays.
    c. Changed_Credit_Limit, Num_Credit_Inquiries**:** Credit limit changes and inquiries.

Credit Profile:

1. Credit_Mix, Outstanding_Debt**:** Variety of credit and current outstanding debt.
2. Credit_Utilization_Ratio, Credit_History_Age**:** Utilization ratio and credit history age.

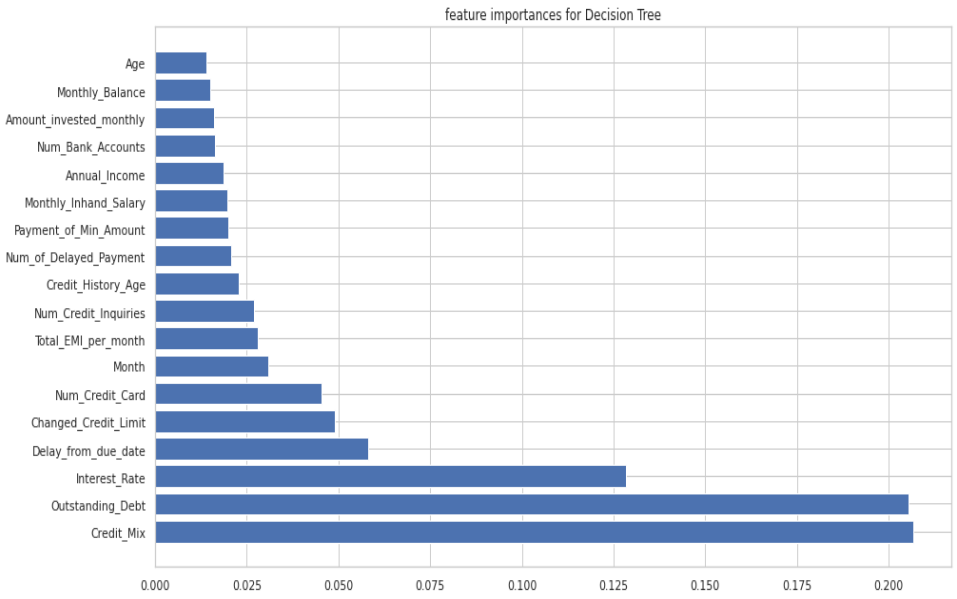
Payment and EMI Details:

1. Payment_of_Min_Amount, Total_EMI_per_month**:** Adherence to minimum payments and total monthly EMI.
2. Amount_invested_monthly, Payment_Behaviour**:** Monthly investments and payment behavior.

9. Monthly Balance and Credit Score:

1. Monthly_Balance, Credit_Score: Monthly financial balance and credit score.

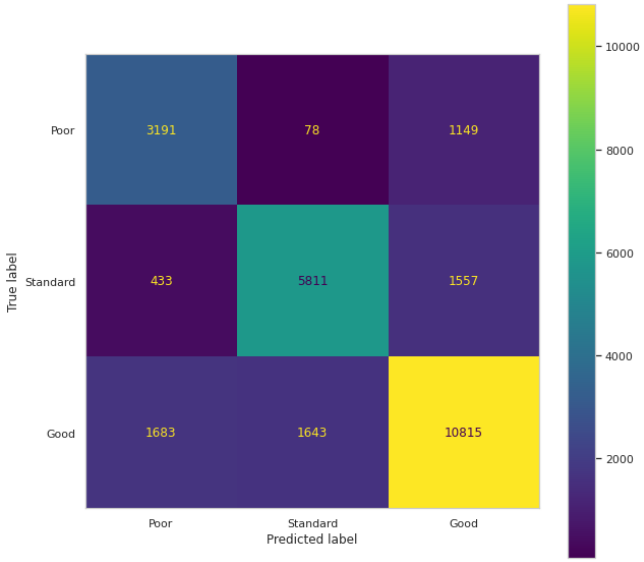
Experimental results –



Important features determined by Random Forest

	Train score	Test score
Random forest	0.793952	0.751783
Voting	0.769141	0.737822
Decision Tree	0.763905	0.725683
KNN	0.803170	0.723217
XGBOOST	0.720348	0.717640
Logistic Regression	0.653296	0.652466
adaboost	0.637184	0.634560

Final Results



Algorithms

Conclusion –

One's creditworthiness is mostly determined by their financial stability and credit management practices. Better credit ratings are frequently linked to higher annual income and a sizable monthly take-home pay, which indicates a person's capacity for sound money management. Higher credit scores are generally associated with a lower credit usage ratio, which reflects responsible credit use. Longer credit histories with prompt payments are also preferred by lenders because they represent a stable and trustworthy financial background. Maintaining timely payments and preventing late payments has a beneficial effect on credit scores and creates goodwill with creditors.

Additionally, lenders see it favorably when borrowers maintain a moderate number of credit accounts in good standing, which adds to a positive creditworthiness signal. On the other hand, a high volume of late payments or repeated credit inquiries may be taken adversely and result in a credit score reduction. Furthermore, high debt to income ratios or large monthly payments may indicate financial pressure and have a negative effect on credit scores. Regarding credit score prediction models, Random Forest and KNN provide great performance in evaluating credit risk and demonstrate effective learning from training data with high training scores (0.793952 and 0.803170, respectively). All things considered, these variables work together to mold a person's credit profile and affect their ability to access advantageous financial options.

Reference –

1. https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=credit%20scoring%20model
2. https://www.kaggle.com/code/yogidsba/personal-loan-logistic-regression-decision-tree/notebook.

3. Google Scholar- Google Scholar is a specialized search engine that indexes scholarly literature, including peer-reviewed articles, theses, books, conference papers, and patents. This approach allows you to access credible and authoritative sources to deepen your understanding of these topics and stay informed about the latest advancements in credit scoring methodologies.
4. JSTOR - JSTOR is a widely recognized and reputable digital library that offers a diverse range of academic resources, including journals, books, and primary source materials. For those delving into the subject of credit score models, JSTOR provides a rich collection of scholarly articles and research papers that cover various aspects of credit scoring.