Bank Loan Status Prediction Using ML Techniques

Jaya lakshmi Gunji,Naveen Yadav Gongati,Priyanka Akula,Somna Sattoor

I. Motivation

The Bank Loan Status Prediction System is motivated to change the current loan disbursement of the banking sector by applying advanced ensemble learning, with its sophisticated machine learning algorithms for predicting borrower performance, to strengthen the selection of risk and resources for better risk performance, which will minimize defaults and bolster portfolio performance. This intelligent and scalable solution will not only automate and optimize the process of loan approval but also adjust with changeable regulations and financial indicators in time to provide financial institutions with real-time and accurate up-to-date risk management tools. The outcome will be a more efficient, transparent, and effective lending process, enabling the setting of new industry benchmarks by both the borrower and lender. Loan approval processes are altered with the amalgamation of the different model predictions for better accuracy and a decrease in bias. These will offer real-time risk management and change industry benchmarks.

II. Project Background

The banking sector is evolving rapidly due to advancements in technology and data analytics, making predictive modeling essential for optimizing financial services, especially loan allocation. Traditional methods are being replaced by machine learning techniques that promise enhanced accuracy and efficiency. Our project focuses on developing an advanced ensemble learning model to improve loan allocation by combining predictions from multiple models, thus reducing bias and increasing accuracy.

Our model analyzes borrower profiles and loan attributes such as credit history, income, employment status, and loan terms to predict loan outcomes. This capability is crucial for minimizing defaults, optimizing portfolio performance, and ensuring resources are allocated correctly. The system is scalable and adaptive, responding to changing borrower behavior and market conditions, creating an intelligent loan allocation mechanism that sets new benchmarks in risk management.

III. Literature Survey

The objective of the project ”Bank Loan Status Prediction using Machine Learning” is to use sophisticated machine learning algorithms and data analysis to modernize the loan approval procedure. The study will make use of ensemble techniques, Random Forest, LightGBM, CatBoost, XGBoost, and other algorithms with a dataset that was acquired from Kaggle. The dataset comprises eighteen variables, one of which is the crucial loan status and the performance measures including recall, precision, ROC curve, F1 score, accuracy, and recall will be used to assess how well the prediction models work. The methodology and strategy of the project will be based on important research publications in the fields of credit risk assessment and loan status prediction. In order to identify credit defaulters and improve the loan lending process, for example, ”Prediction of Loan Status in Commercial Banks using Machine Learning Classifiers” emphasizes the significance of accurate predictive modeling. Similar to this, ”Prediction of Loan Pricing on the basis of Area Location using K-Nearest Neighbour and Support Vector Machine Learning Algorithms” emphasizes the importance of precise loan pricing forecasts and favors the KNN method due to its higher accuracy levels. Furthermore, ”Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector” shows how useful ensemble models are for predicting loan eligibility and lowering default rates. In particular, decision trees using AdaBoost work exceptionally well in this regard.The project’s methodology is shaped by these influential studies as a whole, which highlight the critical role that precise predictive modeling plays in enhancing decision-making procedures and lowering risks in the banking industry

IV. Methodology

Data collection, feature extraction, and various pre-processing techniques in data make the process more refined and properly structured. Data preparation is done here to make sure that data is suitable for the stage of modeling, where building and training of a machine learning model will be done.

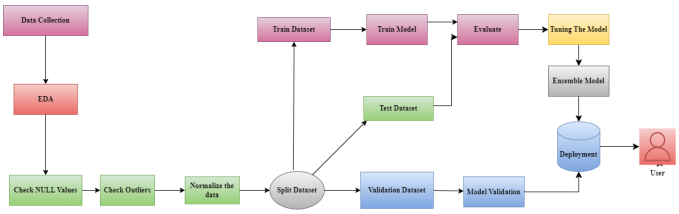

Fig. 1. Project Methodology.

A. Data Collection

The dataset used for classifying loan status is from kaggle, containing about 100514 rows and 19 columns. Further, basic EDA is done in order to understand the data, which leads to the finding of imbalanced data with missing and null values, duplicate values, and outliers.

B. Data Pre-Processing

In this project, data pre-processing has been very careful in handling possible duplicates and missing values not to lose any quality and integrity the data has. The following steps outline the pre-processing activities executed.

Removing Duplicate Rows: This was a data set with 100,514 rows in total, with 19 columns. After dropping all identified columns which were not needed, the dataset retained 16 columns but with the same amount of rows as before. Duplicate entries within the dataset were looked into and 10,728 duplicate rows were deleted, thus resulting in a final dataset refinement of 89,786 rows and 16 columns. This is such an important step because it ensures each data entry is distinct and not duplicates of each other, which could introduce biases.

Handling Missing and Null Values: The dataset was further checked for missing or null values, which could have a bad effect on machine learning models. Rows that had any values missing were well identified and systematically dropped. With such a rigorous process, 22,296 rows had to be dropped in total. Hence, the dataset was further cleaned to be in the shape of 67,490 rows by 16 columns. This is necessary to retain the integrity of the dataset and to guarantee that the model is trained with complete and exact data, enhancing the reliability and robustness of the model's predictions.

Outlier Detection: The dataset was examined for

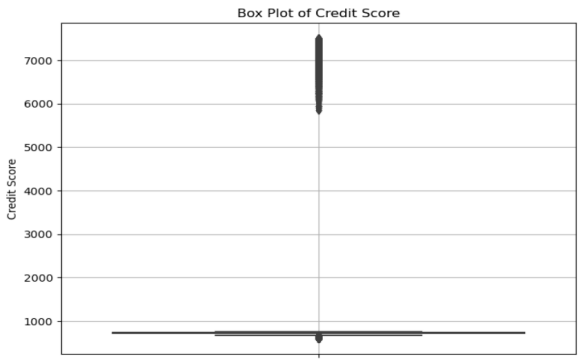

Fig. 2. Outlier Detection.

outliers in the 'Credit Score' column to ensure data consistency and accuracy. Outliers can significantly skew the results of machine learning models, hence their identification and removal is crucial. The dataset was analyzed to identify outliers using the Inter-Quartile Range (IQR) method. Outliers were defined as values that lie significantly outside the range of typical values in the 'Credit Score' column. A total of 6,883 outlier rows were identified and removed from the 'Credit Score' column. The removal of these outliers resulted in a refined dataset with a new shape of 60,607 rows and 16 columns.

One - Hot Encoding: The categorical data from preprocessing has been one-hot encoded in order to convert the categorical variables into a format that makes it pertinent to the machine learning models. It turned out that the categorical columns were 'Loan Status', 'Term', 'Years in current job', 'Home Ownership', and 'Purpose'. We did not encode the target column 'Loan Status'; it stayed in its original representation for model training and validation. The other categorical columns were one-hot encoded, to reflect each value of a category, through a new binary column that tells the presence of this category, with a 1 or 0. The 'dummies()' function was utilized to perform this encoding, resulting in an expanded dataset with additional columns repre-

senting the binary values for each category. This transformation ensured that all categorical features were represented numerically, making the dataset fully compatible with various machine learning models and enhancing the overall model training and prediction processes.

Balancing the Training Data with SMOTE: In the data preprocessing phase, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance the training data after splitting the dataset into training and testing sets. Initially, the dataset was divided into training and testing sets, with the training set containing 48,485 rows and 44 columns, and the testing set containing 12,122 rows and 44 columns. The class distribution before applying SMOTE revealed a significant imbalance, with 39,767 instances of the majority class (Fully Paid) and only 8,718 instances of the minority class (Charged Off).

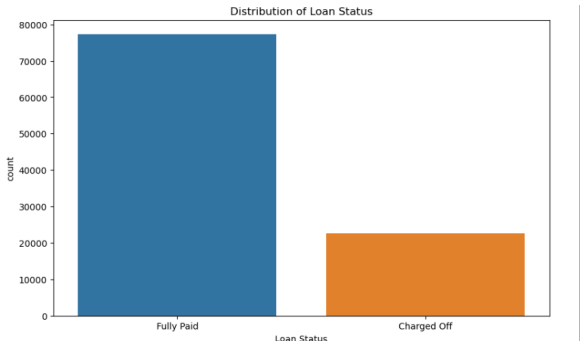

Fig. 3. Distribution of Loan Status.

To address this imbalance, SMOTE was applied to the training data. SMOTE works by generating synthetic samples for the minority class to create a more balanced dataset. After applying SMOTE, the training data was balanced with an equal number of instances for both classes, resulting in 39,767 instances each for the Fully Paid and Charged Off classes. The balanced training dataset now contained 79,534 rows and 44 columns.

Subsequently, the balanced training dataset was further divided into training and validation sets. This additional split ensures that a portion of the data is reserved for validating the model during training, helping to fine-tune hyperparameters and assess the model's performance before testing on unseen data.This balancing process is crucial as it prevents the model from being biased towards the majority class, thereby improving the model's ability to accurately predict both classes. By using SMOTE, the training dataset becomes more representative of the actual distribution, leading to better performance and generalization of the machine learning model.

Feature Scaling: This was applied to standardize the numerical features of the dataset, ensuring equal contribution to the model and enhancing algorithm performance. Initially, features like 'Current Loan Amount', 'Credit Score', 'Annual Income', and 'Monthly Debt' had different ranges, potentially biasing model training. Using the StandardScaler from the sklearn library, each feature was transformed to have a mean of 0 and a standard deviation of 1. Post-scaling, numerical features were centered around 0, with a standard deviation of 1, ensuring consistent feature ranges. This standardization process prevents biases from original feature scales, leading to more accurate and reliable model predictions.

C. Data Modeling and Model Details

In the data modeling phase, various machine learning algorithms were employed to build predictive models. The models trained include logistic regression, k-nearest neighbors (KNN), random forest and XGBoost. Each model underwent hyperparameter tuning to identify the best-performing parameters. Logistic regression, KNN, random forest and XGBoost models were individually evaluated for their performance. After assessing these individual models, an ensemble approach was implemented, combining logistic regression, KNN, random forest and XGBoost, resulting in improved overall metrics and performance.

D. Training Data

The training data used in this project was carefully preprocessed and balanced to ensure robust model training. Initially, the dataset was split into training and testing sets. Due to a significant class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE generated synthetic samples for the minority class, resulting in a balanced training dataset with an equal number of instances for both classes. This balanced training dataset was then further split into training and validation sets to fine-tune hyperparameters and assess model performance during training. This step ensured that the models were trained on data that accurately represented

both classes, thereby improving their performance and generalization capabilities.

V. Experiments and Results

A. Base Model

In order to provide a benchmark for future advancements, the experiments' first phase was devoted to assessing baseline models' performance. Logistic regression, Random Forest, K-Nearest Neighbors (KNN), and XGBoost were the four models that were put to the test. These models were selected because they are widely used and take different approaches to categorization tasks. Because of its ease of use and interpretability, the linear model known as logistic regression offers a solid basis. KNN is a non-parametric technique that was chosen for its straightforward and natural way of classifying data according to how close it is to training samples. Because of its versatility and capacity to manage feature interactions and non-linearity, Random Forest, an ensemble of decision trees, was chosen. Because of its versatility and efficiency in managing many kinds of data and structures, the gradient boosting technique XGBoost was added. Standard measures including ROC-AUC, accuracy, precision, recall, F1 score, and log loss were used to evaluate each model. With ROC-AUC of 0.8736, accuracy of 0.7914, and F1 score of 0.7949, the Logistic Regression model outperformed the other baselines. On the other hand, with a ROC-AUC of 0.6048 and an accuracy of 0.5852, the KNN model performed the worst. The ROC-AUCs of 0.7334 and 0.7084, respectively, for the Random Forest and XGBoost models indicated modest performance as shown in Figure 4.

Base Model	ROC-AUC	Accuracy	Precision	Recall	F1 Score	Log Loss
Logistic Regression	0.8736	0.7914	0.7841	0.8059	0.7949	0.4479
KNN	0.6048	0.5852	0.5850	0.5942	0.5896	1.4468
Random Forest	0.7334	0.6680	0.7016	0.5882	0.6399	0.6105
XGBoost	0.7084	0.6485	0.6698	0.5899	0.6273	0.8646

Fig. 4. Based Models Results

B. Hyperparameter Tuning

Hyperparameter tuning was done on valid data to maximize each model's performance after the baseline evaluation. In order to improve the models' predictive power, this approach entailed modifying model-specific parameters. The regularization value (C) for logistic regression was tuned to strike a compromise between fitting the training set and preventing overfitting. By choosing the ideal number of neighbors (N-neighbors), which establishes how many neighboring points affect the classification decision, the KNN model's performance was enhanced. Critical parameters for the Random Forest model were adjusted, including the amount of features evaluated for splitting at each node (max-features), the maximum depth of the trees (max-depth), and the total number of trees in the forest (n-estimators). Grid search was used for hyperparameter adjustment in order to enhance the baseline models' performance. All of the models had notable improvements as a result of the tuning process. The ROC-AUC of Logistic Regression rose to 0.9302 when the regularization parameter (C=1) was ideal as shown in Figure 5. The KNN model's ROC-AUC increased to 0.9222 when the optimal number of neighbors was set at 15. With max=depth=None, max-features='log2', and n-estimators=200 as its parameter settings, the Random Forest model produced a ROC-AUC of 0.9471. Significant improvement was also demonstrated by XGBoost, which achieved a ROC-AUC of 0.9381 with learning=rate=0.1, max-depth=7, and n-estimators=200. Significant improvements in other parameters, including accuracy, precision, recall, and F1 score, were also produced by these changes.

Base Model	ROC-AUC	Accuracy	Precision	Recall	F1 Score	Log Loss	Best Parameters
Logistic Regressio n	0.9302	0.8801	0.8085	0.9971	0.8930	0.2699	C:1
KNN	0.9222	0.8687	0.8081	0.9682	0.8809	0.3896	N_neighbors=15
Random Forest	0.9471	0.8919	0.8341	0.9793	0.9009	0.2815	Max_depth=None, max_features='log2', n_estimators= 200
XGBoost	0.9381	0.8877	0.8216	0.9912	0.8985	0.2642	Learning_rate=0.1, Max_depth= 7,n_estimators=200

Fig. 5. Hyperparameter tuning results

C. Ensembling

The use of an Voting Classifier ensemble technique allowed for even better prediction performance as shown in Figure 7. To achieve this, individual ensemble models for KNN, Random Forest, XGBoost, and Logistic Regression were combined with their tuned variants. An ensemble strategy makes sense since it maximizes the advantages of individual models while reducing their drawbacks. An even wider range of patterns

and relationships within the data can be captured by the ensemble through the combination of predictions from several models. Generalizability and robustness are frequently enhanced by this method. This study's ensemble method ensured that the various approaches (boosting, non-parametric, tree-based, and linear) were successfully integrated by utilizing a voting process in which each base model contributed to the final prediction. With ROC-AUC of 0.9374, accuracy of 0.8827, and precision of 0.8305, the ensemble performed well. Significant improvements were made on the recall and F1 score, which ended up at 0.9625 and 0.8917, respectively. In addition, the ensemble model's log loss decreased to 0.2934 from the individual models' 0.2934, demonstrating improved calibration and prediction reliability. The effectiveness of mixing numerous models to provide reliable and excellent predicted outcomes is demonstrated by this ensemble approach.

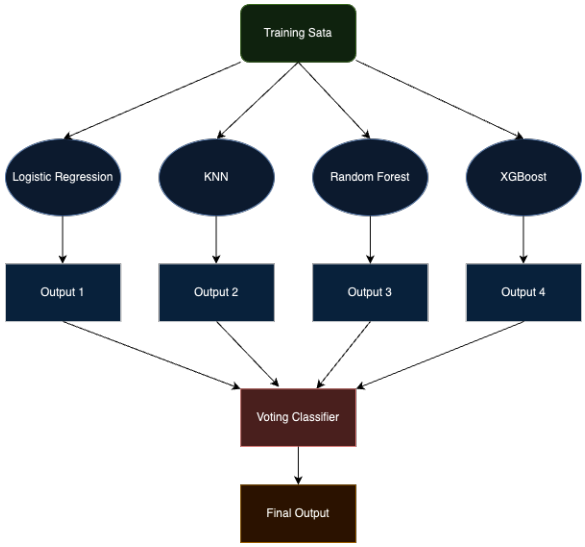

Fig. 6. Ensemble Model Work flow

Model type	Models	Metrics
Ensemble Technique	Logistic Regression, KNN, Random Forest, XGBoost	ROC-AUC : 0.9374, Accuracy : 0.8827, Precision : 0.8305, Recall : 0.9625, F1 Score : 0.8917, Log Loss : 0.2934

Fig. 7. Report of Ensemble Models

D. Application

Gradio was used to design an intuitive application once the ensemble model for forecasting loan status was successfully developed. This program provides a convenient user interface for those looking to determine their eligibility for a loan.Figure 8 shows a variety of relevant factors, such as the current loan amount, loan length, credit score, annual income, years in current employment, house ownership status, loan purpose, monthly debt, and credit history, are captured using input components specifically designed for the interface. Users can quickly receive an estimate of their loan status by entering these facts into the application, enabling them to make well-informed decisions about their financial futures. The Gradio-based system captures the results of data analysis, model creation, and user interface design. Gradio's simplicity and versatility provide accessibility for a wide variety of users, regardless of technical skill level. Whether people are asking for a new loan, refinancing an existing one, or just looking into their possibilities, this application is a useful tool for determining their chances of loan acceptance. This program, with its user-friendly interface and reliable ensemble model predictions, represents a big step forward in democratizing access to financial knowledge and helping consumers to confidently navigate the complex environment of borrowing.

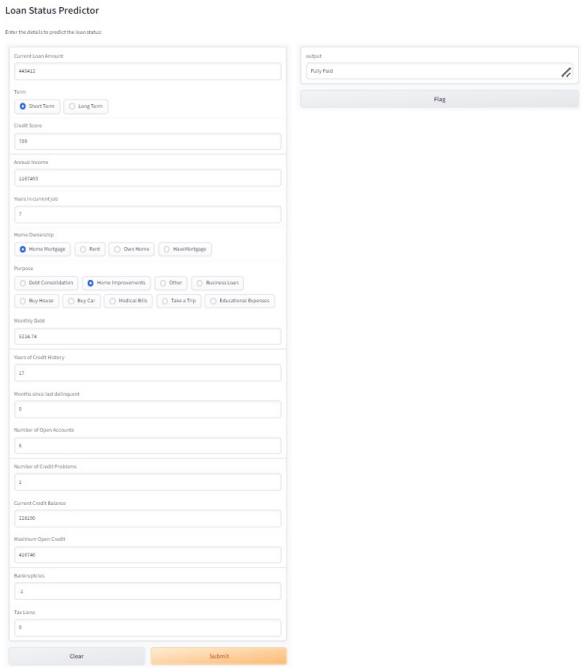

Fig. 8. Gradio Application

VI. Discussion

In this Project, out of all the machine learning classifiers, the highest accuracy for Random Forest and XGBoost stands at 89.19% and 88.77%, respectively, with their corresponding F1 values of 90.09% and 89.85%, which enable robust model prediction at an acceptable recall-precision trade-off. It has a greater performance compared to the K-Nearest Neighbors and Logistic Regression methods for the status of the loan.

A breakthrough advancement in the loan approval process, the combination of sophisticated machine learning algorithms and ensemble techniques in our system offers a more precise and effective means of forecasting results. Thus, by using such sophisticated models, time and resources are saved since the loan assessment and decision-making process is really made more effective and human error is decreased. This implies that loan approval accuracy and dependability have increased, indicating that the system is capable of satisfying the intricate needs of contemporary banking operations. The mentioned flexibility is crucial to guarantee that all banking operations keep the same level of quality and service. Therefore, as we employ effective ensemble methodologies, our system would perform well in a variety of financial situations and shifting market conditions. Furthermore, the general methods and guiding ideas in our study make it possible to go beyond procedures associated to loan acceptance. In summary, this study establishes a new industry benchmark and creates opportunities for future technological developments and applications by combining efficiency, dependability, and adaptability to significantly push the current loan status prediction systems to the new performance levels.

VII. Future Scope

Future scope of this project is huge and promising. As confidence and dependency on data-driven financial decisions grow, adaptive algorithms that update and retrain new models with new data will keep the system in sync with market patterns and regulatory standards. All of these will improve real-time loan evaluation accuracy. Domain understanding adds much and creates new functionalities. Advanced clustering can classify customers by financial behavior and preferences for personalized and relevant financing options. Predictive analytics will improve risk assessment and loan approval intelligence and efficiency. The solution can be expanded to include all financial products and integrated with other banking systems to support management and decision-making, future-proofing financial services from the many challenges and opportunities the sector will face.

References

[1] R. Manglani and A. Bokhare, "Logistic Regression Model for Loan Prediction: A Machine Learning Approach," 2021 Emerging Trends in Industry 4.0 (ETI 4.0), Raigarh, India, 2021, pp.1-6, doi: 10.1109/ETI4.051663.2021.9619201.

[2] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu and P. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.

[3] L. Zheng, Q. Han and Z. Junhu, "A Combination Method of Resampling and Random Forest for Imbalanced Data Classification," 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), Suzhou, China, 2022, pp. 1-5, doi: 10.1109/CTISC54888.2022.9849803.

[4] A. Mueankoo, M. Eso and S. Musikasuwan, "Performance of a Loan Repayment Status Model Using Machine Learning," 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Prachuap Khiri Khan, Thailand, 2022, pp. 1-4, doi: 10.1109/ECTI-CON54298.2022.9795570.

[5] M. Meenaakumari, P. Jayasuriya, N. Dhanraj, S. Sharma, G. Manoharan and M. Tiwari, "Loan Eligibility Prediction using Machine Learning based on Personal Information," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 1383-1387, doi: 10.1109/IC3I56241.2022.10073318.

[6] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.