Loan Approval Prediction

Dharam Singh Vislavath
DharamSinghVislavath@my.unt.edu

Prathyusha Gangisetty
PrathyushaGangisetty@my.unt.edu

Divya Sai Sri Murugudu
Divyasaisrimurugudu@my.unt.edu

Kusuma Kumari Dama
kusumakumaridama@my.unt.edu

Manoj Kumar Unnam
ManojKumarUnnam@my.unt.edu

Siri Ranabothu
Siriranabothu@my.unt.edu

Abstract

Banks use loan approval prediction to determine applicants' eligibility. For this project, we employed stacked ensemble models with a meta-learner, Graph Neural Networks (GNN), and machine learning models like XGBoost. For improved analysis, XGBoost was also implemented using the Orange Tool, and feature engineering was improved with K-Means clustering. The models were created and trained on a training dataset before being evaluated on a different dataset that lacked target labels in order to forecast the likelihood and status of loan acceptance. The model's performance was examined by visualizing the variations in prediction probabilities between the training and testing datasets. Reliable and effective predictions were ensured by assessing the models' efficacy using metrics such as accuracy and ROC AUC.

1. Introduction

For financial organizations to assess applicants' creditworthiness, accurate loan approval predictions are crucial. Strong tools for analyzing application data and producing accurate predictions are provided by machine learning (ML). Based on demographic and financial characteristics, this study use machine learning approaches to forecast loan approval results. Graph Neural Networks (GNN), XGBoost, and ensemble approaches are used to increase prediction accuracy and offer lucid insights.

1.1. Objective

This project's objective is to create and evaluate machine learning models that forecast loan approval status. Additionally, it seeks to display outcomes and analyze prediction probabilities in order to facilitate well-informed financial services decision-making.

1.2. Dataset

A training dataset with labeled loan approval data and a testing dataset without target labels serve as the foundation for our analysis. The datasets offer information on the credit-related, financial, and demographic characteristics of applicants. To get the data ready for machine learning model training and assessment, preprocessing and feature engineering techniques were used. By taking these actions, consistency was guaranteed, missing data were addressed, and feature representation was improved for more accurate predictions.

1.3. Methodology

1. **Data Pre-processing:** The mean for numerical features and the mode for categorical data were imputated to fill in missing values. Label encoding was used for categorical data, and feature scaling was used to normalize numerical features. To enhance model performance, further features were designed, including total income and loan-to-income ratio.
2. **Model Development:**
    - **XGBoost Classifier:** Models with hyperparameter tweaking were created utilizing SMOTE, grid search, and randomized search to address class imbalance.
    - **Graph Neural Networks (GNN):** Utilizing relational data between loan applicants, a GNN model was developed.
    - **Stacked Ensemble Model:** The Stacked Ensemble Model employed a meta-learner to increase accuracy after combining predictions from three base learners.
    - **Feature Augmentation with K-Means:** By adding cluster labels to the dataset, the feature space was improved.

- **Orange Tool:** For visual analysis, XGBoost was implemented within the Orange Tool.

3. **Evaluation of the Model:** To ensure the accuracy and robustness of predictions, the performance of each model was assessed using metrics such as ROC AUC and accuracy. The training and testing datasets' prediction probabilities were compared using visualization approaches.

2. Related Work

1. **Loan Approval Prediction Based on Machine Learning Approach [2]**: An approach based on machine learning is suggested in this study to predict loan approvals. To find crucial elements affecting loan approval decisions, the authors employ decision tree algorithms. The suggested model's analysis of applicant demographic and financial data yields a high degree of accuracy, proving its usefulness in helping banks automate loan approval procedures.

2. **Prediction of Loan Approval Using Machine Learning [3]**:This work builds predictive models for loan approvals using machine learning approaches like logistic regression, support vector machines (SVM), and decision trees. The study highlights how various applicant characteristics impact loan eligibility and evaluates the effectiveness of alternative algorithms, emphasizing the advantages of automation in terms of boosting efficiency and decision-making accuracy.

3. **Bank Loan Approval Prediction Using Artificial Neural Networks [8]**: The use of artificial neural networks (ANN) to forecast bank loan approvals is examined in this work. To train the ANN model, the authors take into account a number of application characteristics, including income, credit score, and work history. In contrast to conventional machine learning models, the study shows that neural networks can handle intricate patterns in data and achieve high predicted accuracy. These studies offer a framework for using machine learning algorithms to automate loan approval procedures, and they served as motivation for our project's investigation of cutting-edge methods including ensemble learning, XGBoost, and GNN for improved predictions.

3. Proposed Method

The suggested approach uses a systematic workflow that combines data mining tools, feature engineering, and machine learning algorithms to forecast loan approval status. The following steps are part of the methodology:

1. **Data Cleaning and Pre-processing:**For the purpose to handle missing values, encode category variables, and normalize numerical features, the project starts by cleaning the loan datasets. A training dataset that contained the target variable and a test dataset that did not were both used. To guarantee comparability with the training dataset, the test dataset underwent preprocessing.

2. **Feature Engineering:** In order to improve predicted accuracy, new features were developed, such as Total Income (which combines the incomes of the applicant and co-applicants), Loan Income Ratio (which is the ratio of the loan amount to total income), and Income Per Dependent (which is the income per dependent family member). These derived attributes gave more detailed information about the applicants' financial profiles.

3. **Model Development:**

   - **Variants of XGBoost :** To investigate performance enhancements, standard XGBoost, XGBoost with SMOTE for class imbalance, and hyperparameter-tuned XGBoost models were used.
   - **Graph Neural Network (GNN):** Using graph-based representations, a GNN model was created to integrate relational insights across candidates, such as feature similarities.
   - **Stacked Ensemble Learning:** Using Logistic Regression as a meta-learner, this ensemble learning technique integrated predictions from the XGBoost, LightGBM, and CatBoost models to ensure higher predictive accuracy.
   - **Feature Augmentation with K-Means and XGBoost:** Patterns in applicant data were found using K-Means clustering, and the cluster labels that were produced were used as extra inputs to improve the predictions made by the XGBoost model.

4. **Model Evaluation:**Accuracy, precision, recall, F1-score, and ROC AUC metrics were used to evaluate the models. Visual comparisons between training and test predictions offered additional insights into model reliability, and validation findings were meticulously examined to guarantee the models' predictability.

5. **Integration with Orange Tool:** The Orange data mining tool was utilized for feature selection, interactive data exploration, and initial model assessment. Faster experimentation with conventional algorithms and feature correlations was made possible by its visual and drag-and-drop interface.

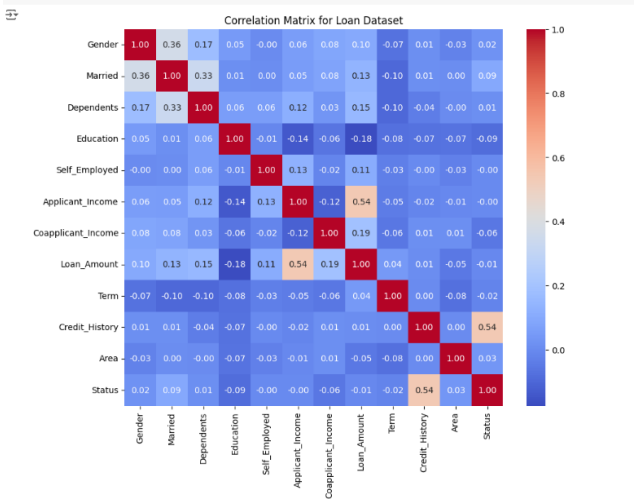

Figure 1. Correlation Matrix for Loan Dataset Features

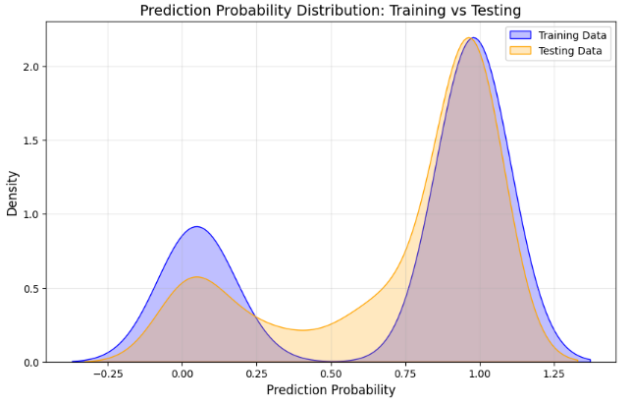

Figure 2. Prediction Probability Distribution: Validation vs Test Data (XGBoost Model)

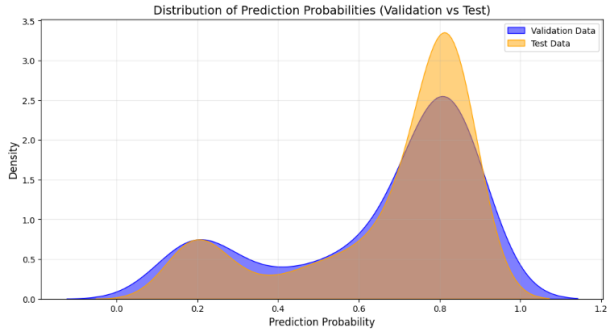

Figure 3. Prediction Probability Distribution: Validation vs Test Data (Stacked Ensemble Model)

6. **Visualization and Analysis:** To compare training and testing results, comprehensive visualizations were employed, including bar graphs, KDE plots, and prediction probability distributions. In addition to validating predictions, these visual aids provided insightful information on model behavior.

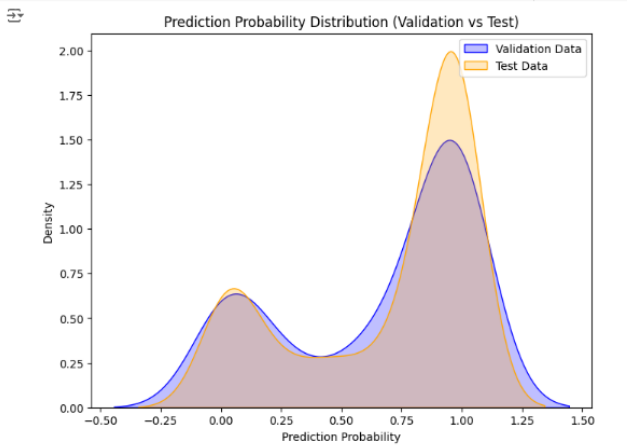

Figure 4. Prediction Probability Distribution: Validation vs Test Data (Feature Augmentation with K-Means + XGBoost)

4. Experiments

4.1. Dataset

This study's dataset offers a wide range of features for forecasting loan approval status. Python was used for feature engineering and preprocessing on the dataset. To improve the dataset and increase forecast accuracy, three more columns were added that capture important financial variables. Following processing, this dataset was used to build and assess models in Python and Orange. An summary of the dataset is provided below:

- **Gender**: The applicant's gender is indicated. This tool offers information about demographic patterns in loan approvals.
- **Married**: Indicates whether the applicant is married. Financial stability can be influenced by marital status, which may have an effect on judgments about loan acceptance.
- **Dependents**: shows the applicant's total number of dependents. This function assists in identifying any debts that might affect the applicant's capacity to pay back repayment.
- **Education**:Indicates the applicant's level of education, classifying them as either graduate or not. Income potential and repayment capacity can be correlated with educational attainment.
- **Self Employed**: Indicates if the candidate works for themselves. Income stability and creditworthiness can be strongly impacted by one's employment type.

- **Applicant Income**: This is the primary applicant's income. When evaluating the borrower's capacity to repay the loan, this is a crucial factor.
- **Coapplicant Income**: gives the relevant co-applicant's income, if any. The ability to repay is improved by total income.
- **Loan Amount**: The entire loan amount that the applicant has asked for. This accurately depicts the amount of money that is being evaluated for approval.
- **Term**: Shows how many months are needed to pay back the loan. The length of the payback period affects the monthly installment amounts and financial strain.
- **Credit History**: Displays the applicant's credit history as a numerical value (e.g., 1.0 if the history is positive). The credit history of an applicant is an important measure of their capacity to repay debt.
- **Area**: Indicates whether the land is semi-rural, urban, or rural. Because market circumstances and property values vary by location, loan approval may be affected.
- **Status**: The target variable, which shows if the loan was granted (Y) or denied (N).

  This dataset, which captures important financial and credit characteristics of applicants, was chosen for its applicability in determining loan eligibility. In order to ensure interoperability with machine learning models, it underwent meticulous preprocessing to address missing values and encode categorical variables.To improve the dataset and model predictions, three more features were extracted:

- **Total Income**: The total income of the co-applicant and the applicant. This characteristic aids in determining the applicants' aggregate financial capacity:
- **Loan Income Ratio**: A measure of the financial load in relation to income, calculated as the loan amount divided by total income:
- **Income Per Dependent**: This measure of the applicant's income divided by the number of dependents aids in determining their financial capacity:

4.2. Software

The software used in this project includes:

1. Python: Python was the main programming language used for analysis, model building, and data preprocessing.
2. Pandas: Used for data management and manipulation, enabling effective loan dataset structure and cleansing.
3. NumPy: Used for mathematical calculations and array operations, among other numerical computations.
4. Matplotlib and Seaborn: Applied to the development of visuals for the analysis of model performance and prediction probabilities.
5. Scikit-learn: Used to put machine learning strategies like data partitioning, scaling, and model performance evaluation into practice.
6. PyTorch Geometric: A dedicated library for building and refining the model of a Graph Neural Network (GNN).
7. XGBoost and LightGBM: Crucial for developing and optimizing ensemble models based on trees in order to improve prediction accuracy.
8. Orange Tool: A data mining tool for feature selection, visualization, and interactive exploratory data analysis.
9. Streamlit: A user-friendly online application was developed using Streamlit, enabling the model to be accessed for predictions of loan acceptance in real time.
10. Google Colab: facilitated the training of complex models by providing a collaborative development environment and computing resources.

A thorough and effective workflow for the loan approval prediction project was ensured by the combined facilitation of data processing, model construction, evaluation, visualization and accessibility provided by these software tools.

4.3. Hardware

The following hardware resources were used for this project:

- Google Colab: The primary tool used to construct this project was Google Colab, which offered a virtual machine environment with a GPU. For computationally demanding models such as the Graph Neural Network (GNN) and ensemble learning techniques, the GPU greatly shortened training times. Moreover, the cloud-based infrastructure of Google Colab guaranteed smooth cooperation and resource-demanding work execution.

  Setting up the environment, installing required libraries, and controlling dependencies by hand would be essential if the project were to be run on a personal computer. Additionally, without a GPU-accelerated setup, training times for models like XGBoost and GNN may be prolonged.

4.4. Experiment: Evaluating Model Variants for Loan Approval Prediction

4.4.1 Objective

Evaluating machine learning models' performance on a feature-engineered dataset was the aim of the trials. The processed dataset, which was produced in Python and included further features, was used on both the Orange and Python platforms. Advanced model implementations like XGBoost, GNN, and ensemble learning were done with Python, but Orange allowed for quick creation and validation of baseline models, guaranteeing a thorough assessment of the dataset's predictive capacity.

4.4.2 Feature Selection

The first set of data comprised demographic and financial characteristics of the applicant, including credit history, income, and education. To design other features like Total Income, Loan Income Ratio, and Income Per Dependent, feature engineering was used. Cluster labels produced using K-Means were included as an extra feature to further evaluate the effects of enhanced data. The goal was to increase forecast accuracy by uncovering hidden patterns among candidates.

4.4.3 Evaluation Metrics

The models' performance was evaluated using the following evaluation metrics:

- Accuracy: To calculate the percentage of loans that are approved and denied with the proper classification.
- Precision:To calculate the percentage of loans that are approved and denied with the proper classification.
- Recall:To evaluate how well the model can recognize real loans that have been approved.
- F1-Score:A statistic that combines recall and precision.
- ROC-AUC:The model's ability to differentiate between loans that are authorized and those that are denied is measured by the ROC-AUC.

4.4.4 Experiment Considerations

Avoiding Data Leakage: When training the models, care was taken to make sure no future knowledge was used. The computation of features such as Total Income and Loan Income Ratio was limited to data that would be accessible at the time of projection.

Feature Usefulness: Each attribute was evaluated for importance to make sure it improved the model. Confusion and overfitting were prevented by eliminating features that were too similar or didn't offer fresh insights.

Balancing Classes: TheMethods such as SMOTE were employed to balance the data because the dataset contained more authorized loans than denied ones, making it easier for the model to forecast both approvals and denials.

Model Simplicity: TheEven though sophisticated models like GNN and ensemble learning were employed, an attempt was made to maintain the models' simplicity and usefulness for everyday applications.

4.4.5 Experiment Results

Model	Accuracy	ROC AUC
Graph Neural Network (GNN)	85.37%	
Stacked Ensemble of Tree-Based Models	84.55%	85.17%
Feature Augmentation with K-Means + XGBoost	84.55%	85.88%

Figure 5. Final Model Comparison: Accuracy and ROC AUC Metrics

Loan Approval Prediction Model Metrics

	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
1	XGBoost Classifier	76.42	77.41	90.0	83.23	75.4
2	XGBoost with GridSearchCV and StratifiedKFold	76.42	80.0	85.0	82.42	75.26
3	XGBoost with RandomizedSearch CV and StratifiedKFold	75.61	78.41	86.25	82.14	77.41
4	XGBoost with SMOTE	76.42	78.65	87.5	82.84	74.8
5	Orange Data Mining Tool (Gradient Boosting)	82.4	84.3	82.4	80.4	83.5

Figure 6. Performance Metrics of Loan Approval Prediction Models

5. Results and Analysis

- XGBoost, Graph Neural Networks (GNN), and the stacked ensemble technique were among the models that were successfully used to forecast loan approval status, according to the experiment results.
- The models' performance was greatly improved by feature engineering, which included developing features like Total Income and Loan Income Ratio that captured important financial patterns.
- Through the use of GridSearchCV and RandomizedSearchCV, hyperparameter tuning improved XGBoost's predictive accuracy and recall.

- To assess conventional models like Gradient Boosting and visualize performance indicators, the Orange tool offered an intuitive user interface, confirming their efficacy in this field.

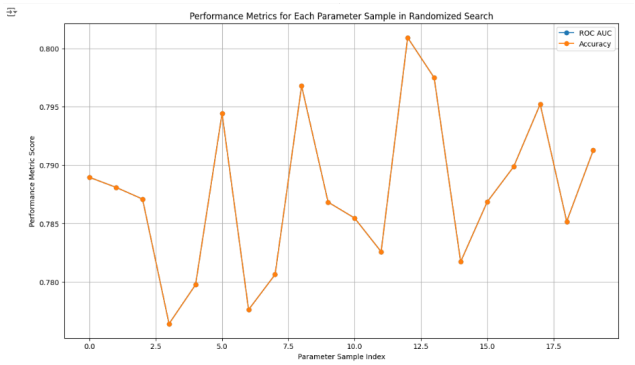

Figure 7. Performance Metrics for Parameter Tuning in Randomized Search

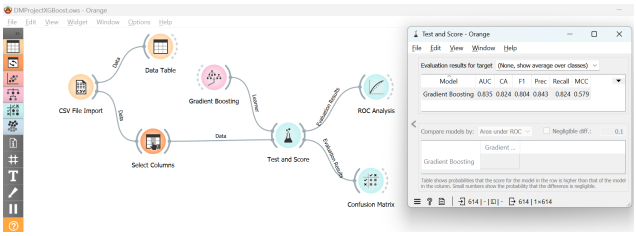

Figure 8. Orange Workflow for Gradient Boosting Evaluation

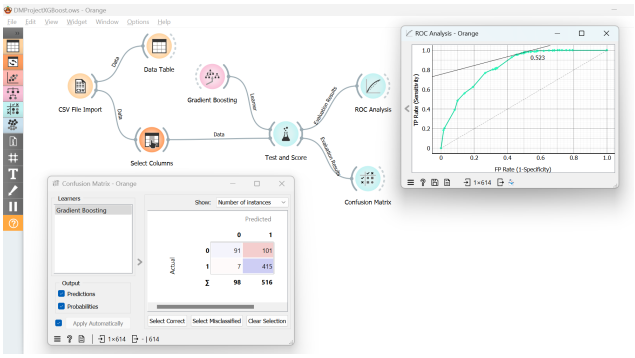

Figure 9. Evaluation Metrics for Gradient Boosting in Orange

5.1. Final Results

The below figure shows the performance of the key models based on accuracy and ROC AUC

5.2. Discussion

Graph-based associations between applicants, including similarity in credit history or financial patterns, were used by the Graph Neural Network to achieve the maximum accuracy, making it the most successful model. Through the combination of XGBoost, LightGBM, and CatBoost's advantages, the stacked ensemble model provided a well-rounded strategy that consistently produced results in accuracy and AUC measures. The efficiency of XGBoost was increased overall when hidden patterns in the data were revealed by feature augmentation using K-Means clustering. Using SMOTE in conjunction with XGBoost also addressed class imbalance, enhancing recollection and more effectively recognizing candidates from minority classes. The quality and scope of the dataset limited the models, even though hyperparameter modification greatly improved XGBoost's performance.The Orange tool made quick experiments possible, showcasing the potential of gradient boosting as evidenced by its remarkable AUC score in comparison to baseline models.

Model	Accuracy	ROC AUC
XGBoost Classifier	76.42%	75.40%
XGBoost with GridSearchCV	76.42%	75.26%
XGBoost with RandomizedSearchCV	75.61%	77.41%
XGBoost with SMOTE	76.42%	74.80%
Graph Neural Network (GNN)	85.37%	N/A
Stacked Ensemble of Tree-Based Models	84.55%	85.17%
Feature Augmentation with K-Means + XGBoost	84.55%	85.88%

Figure 10. Performance Metrics of Loan Approval Prediction Models

6. Project Web Interface

In order to improve accessibility and show how our loan approval prediction algorithms are used in real-world scenarios, we created an interactive online interface that is hosted on Ngrok. The portal lets users explore the insights produced by the algorithm and engage with our predictive model XGBoost in real-time..

6.1. Features of Web Interface

The user interface is intended to be simple and easy to use. Key applicant data, including income, loan amount, credit history, and other specifics, can be entered by users to get immediate estimates regarding the likelihood and status of loan acceptance. Visual feedback is another feature of the application that helps users better comprehend model outputs and probabilities.

6.2. Technical Implementation

With Streamlit powering the web interface, rapid deployment and a flawless user experience are guaranteed. We made the application publicly available at a secure URL using Ngrok. The trained machine learning model and the backend are directly connected, enabling effective data interchange and dynamic predictions. This configuration closes the gap between sophisticated analytics and user-

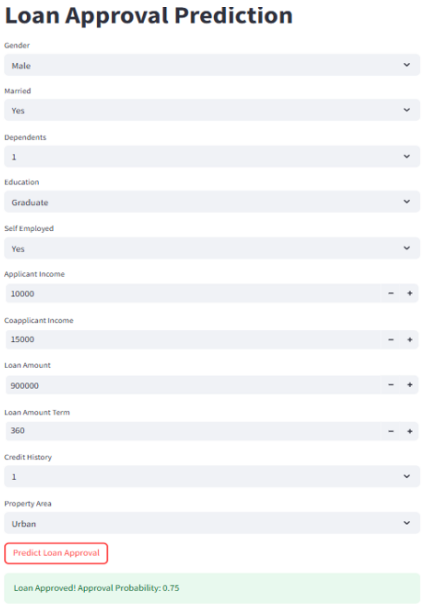

Figure 11. Interactive Loan Approval Prediction Interface

focused apps by making the loan approval prediction system accessible and useful.

7. Conclusions

The Graph Neural Network, which successfully takes use of graph-based associations like credit history similarities, had the greatest accuracy of 85.37 in this project's evaluation of sophisticated machine learning models for loan acceptance prediction. While XGBoost's performance was improved by feature augmentation using K-Means clustering, which resulted in an accuracy of 84.55 and an AUC of 85.88, the stacked ensemble model produced an accuracy of 84.55. Hyperparameter-tuned XGBoost models showed considerable performance gains, and SMOTE successfully addressed class imbalance, increasing recall. Quick experimentation was made easier by the Orange tool, and among baseline models, Gradient Boosting had the greatest AUC of 83.5. Future improvements might include behavioral and economic data from outside sources, allowing for even more reliable and precise loan approval prediction algorithms.

7.1. Future Directions

- **Integration of External Economic Data:**To increase the model's predictive abilities and flexibility in a range of situations, more external data should be incorporated, such as area economic indicators, credit bureau data, and market movements.

- **Development of Real-Time Predictive Systems:**developing models that dynamically process fresh applicant data to enable financial institutions to make decisions more quickly and predict loan approvals instantly.

- **Federated Learning for Privacy-Preserving Training:**By using federated learning, several financial institutions can work together to train models without exchanging sensitive data, improving security and privacy.

For financial organizations, precise loan approval forecasts can decrease default risks, cut operating expenses, and expedite decision-making. They also encourage financial inclusion by guaranteeing equitable evaluations and giving applicants faster, more individualized loan approvals.

8. Limitations

There are limits to this project. Because the models were created and evaluated using a particular loan dataset, their applicability to datasets from other financial institutions or geographical areas with different practices may be limited. The precision of forecasts may be impacted by economic shifts like changes in interest rates or policy changes. Periodically retraining the models is also necessary to keep them current and adjust to changes in application behavior or credit patterns. Furthermore, because the models rely on past loan data, they might not be able to accurately forecast results for unusual or unknown situations.

9. Further Research

To improve the predictability of loan approval models, future studies should concentrate on integrating other data sources, such as macroeconomic indicators or comprehensive credit histories. Decision-making efficiency may be further increased by using real-time adaptive systems that alter forecasts in response to new applicant data. It could be beneficial to investigate the incorporation of sophisticated graph-based models to better represent the relationship dependencies between lenders and applicants. Furthermore, evaluating the models on a variety of datasets from various geographical locations or financial institutions will offer insightful information about their scalability and resilience. Another important area of attention would be the application of interpretability strategies to increase the transparency and explainability of predictions for stakeholders.

10. Applications

The financial industry's decision-making procedures might be greatly improved by using accurate loan approval prediction models. By automating loan approvals, banks and other financial institutions can cut down on processing time and manual labor. Lenders can investigate data trends, pinpoint important elements impacting loan approvals, and enhance feature selection by incorporating data

mining technologies. By reducing defaults and guaranteeing objective decision-making, these insights facilitate more efficient credit risk assessment. Predictions can also be utilized to tailor loan offers according to the profiles of applicants, which would increase client satisfaction. More broadly, these developments support financial inclusion by improving credit availability for marginalized groups and promoting economic growth.

11. Contributions

1. **Dharam Singh Vislavath:** Responsible for gathering, cleaning, and preprocessing the dataset using Python (pandas) and Orange.

2. **Kusuma Kumari Dama:** Managed integration with the Orange tool for data exploration, feature selection, and visual implementation of XGBoost.

3. **Divya Sai Sri Murugudu:** Designed and implemented the Graph Neural Network model and model optimization to handle relational dependencies in the data

4. **Prathyusha Gangisetty:** Worked on feature engineering to improve model performance.Developed the XGBoost models with various enhancements.

5. **Siri Ranabothu:** Focused on model performance evaluation and comparison of XGBoost, GNN, and ensemble methods.

6. **Manoj Kumar Unnam:** Consolidated project documentation, including methodology, results, and references.

References

[1] D. Cheng et al. Critical firms prediction for stemming contagion risk in networked-loans through graph-based deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8350–8358, 2023.

[2] A. Kumar, I. Garg, and S. Kaur. Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering*, 18(3):18–21, 2016.

[3] R. Kumar et al. Prediction of loan approval using machine learning. *International Journal of Advanced Science and Technology*, 28:455–460, 2019.

[4] Q. Liu et al. Rmt-net: Reject-aware multi-task network for modeling missing-not-at-random data in financial credit scoring. *arXiv preprint*, 2022.

[5] T. Ndayisenga et al. Bank loan approval prediction using machine learning techniques: Application of data science in financial industry. In *Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management*, Orlando, Florida, USA, June 2022.

[6] M. Tejaswini et al. An approach for prediction of loan approval using machine learning algorithm. In *IEEE Xplore*, 2020.

[7] J. D. Turiel and T. Aste. P2p loan acceptance and default prediction with artificial intelligence. *arXiv preprint*, 2019.

[8] V. Viswanatha, A. Ramachandra, K. Vishwas, and G. Adithya. Prediction of loan approval in banks using machine learning approach. *International Journal of Engineering and Management Research*, 13(4), August 2023.

[9] H. Wang and L. Cheng. Catboost model with synthetic features in application to loan risk assessment of small businesses. *arXiv preprint*, 2021.