A study of Performance of loan default prediction model Using Machine Learning Techniques

Prakash S. Chougule [1], Tejaswi S.Kurane[2] , Mrs. Varsha C.Shinde[3] ,Rahual. H. Waliv[4], Ramesh D.Shinde[5] , Shruti N. Tiware Patil[6]

[1] Associate Professor, Rajarshi Chhatrapati Shahu College, Kolhapur(MS), India

[2,]Assistant Professor, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India

[3] Associate Professor, Vivekanand College, Kolhapur(MS), India

[4] Associate Professor, Kisan Veer Mahavidyalaya, Wai Dist.Satara (MS), India

[5]Assistant Professor, Jaysingpur College, Jaysingpur (MS), India

[6] Research Student, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India

Abstract: Loan lending has been playing significance role in financial world through out the year. It is quite profitable for the both lenders and borrower's. In the banking sector a loans have become a key component that steers the economy and directly impact of the growth of a nation economy. The loan default predication is to p redict rather the borrower will delay the repayment or not. This is an important problem of banking and fi nance companies. Now a days there are numerous risk identified with bank sector regarding giving loan to the client and for the individual who get the loan. In our study our main aim is that to build up the loan defaulter predication model based on machine learning technique. Three machine learning models like, Rando m Forest, Logistic Regression, Light GBM to predict whether a customer may get loan or not.In this study we compute correlation Heatmap and VIF factor it shows that there is no correlation between the variables also we used random forest to identify top ten features which are highly affected on default prediction.Using SMOTE and SMOTE ENN methods we constructed above three models then our study shows that all of these three models based on Smote ENN technique perform excellent with high accuracy to accurately predict loan defaulter as compared to model based on SMOTE .This paper also shows that machine learning models may be a better option than traditional techniques for organizations trying to forecast the failure of loans

Key words: Loan default, Logistic Regression, Random Forest, Light GBM, Smote, Smote- ENN

Introduction: In world All people needs a loan there are many loan trading banks institutes etc people take help of these for there financial problems or personal issues economical Competition in world makes a person or individual to have a loan. To keep their transactions fluid and earn revenue to sustain themselves through economic periods small to large scale banking organisations rely on borrowing activities.Due to the interest earned from loans and therefore, very important in the banking sector financial risk may arise for the banks. Based on advance non repayment by institutes amount big, and the each year people paying from lending institutes bad loans to borrowing financial institutes at the upfront so in such a way huge loss suffered and the financial distress impact on the financial sectors all over the world.Loans predictive model building this can be very helpful for Financial Institute to come up with the challenges of Lending history reducing the chances of large losses due to Loan defaulter and trusted money defaulters not repaying. So lending is a win-win for both the lender and the borrower but also it exposes both the lender and the borrower to significant risk which can basically be boiled down to the inability of the borrower to pay back the loan in time. This is a mutual decision as to the lenders and the borrowers and is known as 'Credit Risk'. In conventional lending model banking officers primarily uses 5C Principal (collection of character, capital, capacity, collateral and terms) to access the capacity of customer must be availed before the loan would granted. This judgement however, is subjective to each person reading, and there are a lot of factors influencing how a consumer may experience the transaction. So considering above problem this paper aims to build and machine learning model to lend the loan to a non defaulter customer which will help to identify whether approved the loan to an particular individual or not.

Literature Review: Practitioners in the banking and financial sectors have increasingly turned to machine learning (ML) and deep learning (DL) models to automate and enhance the process of credit risk assessment and loan prediction. These advanced algorithms are able to process vast amounts of data far more efficiently and accurately than traditional methods, providing more reliable insights into a candidate's creditworthiness. Machine learning models such as decision trees, random forests, support vector machines (SVM), and k-nearest neighbours (KNN) have been commonly used for this purpose. More recently, deep learning algorithms like neural networks, particularly those based on recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, are gaining attention due to their ability to capture complex patterns in large datasets, especially when there is temporal or sequential data involved (e.g., historical financial behaviour over time).The Random Forest Algorithm was adopted by Lin Zhu et al. in paper [4] and Nazeeh Ghatasheh in paper [5] in order to build a model for predicting the likelihood of loan default. According to paper [4], the accuracy of random forest(98%) was better than other algorithms such as logistic regression(73%), decision trees(95%), and support vector machines(75%). Random Forest Algorithm is a good methodology for credit risk prediction from the findings of the paper [5] Paper [5] mentioned that competitive classification accuracy and simplicity are the main advantages of the algorithms.A wide range of popular methods including logistic regression, k-nearest neighbours, random forest, neural networks, support vector machines, stochastic gradient boosting, Naive Bayes, etc. have been reported in [6] and it is concluded about that it is close to impossible to determine the overall best method. In their paper [7] Nikhil Madane and Siddharth Nanda carried out a review on credit scoring of mortgage loans based on which they found the following results: In most cases,

- Credit applications that fail to meet certain requirements will not be accepted because the probability that the application is going to pay back is too low.
- Applicants from lower-income areas are more likely approved and more likely to repay their loans on time.

Pidikiti Supriya et al. Decision Trees as a [8] machine learning tool to implement their model. Load the datasets Data Analysis They initiated their analysis from data cleaning pre-processing, and missing value imputation, at last exploratory data analysis, and finally model building and evaluation. On a public test set, they achieved a best accuracy of 81%.They tested using the C4. The result of maximum precision was 78.08% when the data partition was 90:10 and the largest recall value was 96.4% when the data partition was 80:20 based on Decision Trees using C4.5 algorithm in [9]. Hence, partition of 80:20 was decided to perform best in terms of highest accuracy and high recall value. In paper [10], the authors performed exploratory data analysis. The primary objective of the paper was to categorize and analyze the nature of borrowers. Seven unique graphs were plotted and visualized and the authors concluded the majority of loan applicants preferred short-term loans using these graphs.Syed Zamil Hasan Shoumo et al. [11], suggested that the Support Vector Machines outperform many of the models including the logistic regression, random forest etc., applied in the paper, for comparative performance uitvoeren analysis.The authors in paper [12] selected 4 different models:M1: Logistic Regression model ,M2: Random Forest model ,M3: Gradient Boosting model and D1-D4: Multilayer Neural Network models (deep learning)and using these models they demonstrated that data quality check is crucial, that is, analysis of data and cleaning before modelling to exclude redundant variables. According to the paper the major aspects of deciding whether to give individual a loan or not are the selection of features and the algorithm. Aboobyda Jafar Hamid et. al. proposed a model for classifying loan risk by using Data Mining in paper [13]. It accomplished this using three algorithms: J48 ,Bayes Net and Naive Bayes.They concluded J48 was the best algorithm for the purpose as it had a high accuracy (78.3784%) and a low mean absolute error (0.3448). Aditi Kacheria et al. So [14] applied the Naive Bayesian algorithm for their model. And to enhance the classification performance, they implemented the k-NN and binning algorithms to enhance the data quality. Missing values in dataset were addressed using K NN and binning algorithm was used to handle the anomalies.According to a study conducted in the Czech Republic and Slovakia by [15] Martin Vojtek and Evzen Kocenda, most local banks are using the logit method-based models. Other approaches such as CART or neural networks are mainly considered as either support tools in the variable selection step or in the model quality evaluation step. The authors also found that k-NN isn't ever or very rarely used.A comprehensive study comparing the performance of the XGBoost algorithm with the Yu Li in paper [2] logistic regression performance Model discrimination and model with the stability of the XGBoost model being significantly greater than that of the logistic regression model In this Paper by utilizing exploratory data analysis (EDA) methods such as bar plots, Variance Inflation Factor, and correlation matrices, we can discover which relationships between variables are significant. A random forest classifier is applied to performing feature selection to gauge the more significant features. Before checking model performance I have checked whether the data is balanced data or not. Data found to be imbalance then by using Synthetic Minority Over-sampling Technique (SMOTE) combined data has been converted to balanced data set. And while fitting the model Synthetic Minority Over-sampling Technique plus Edited Nearest Neighbours (SMOTEENN) this techniques helps to generates synthetic data points for the minority class to balance the dataset and to Cleans the dataset by removing mislabelled or noisy samples, mainly from the majority class. SMOTE and SMOTEENN are implemented on all three models logistic regression, random forest, and Light GBM.

Theory: Nearly every sector in the world is advancing towards complete automation. Various concepts and methods are being developed every day to achieve this goal and many fields have been under study for many years. One of the most upcoming fields that have grabbed the attention and excitement of scientists, researchers, and technologists is Artificial Intelligence (AI).

Random Forest : Random Forest belongs to the supervised learning algorithm. Like decision trees, they are also used for classification and regression. A predictor ensemble is built with several decision trees that expand in randomly selected data subspaces [12].

Logistic Regression: It is a statistical analysis method to predict a binary outcome, such as Yes or No, based on prior observations or a learning set of data points. A logistic regression model predicts a dependent variable by analysing the relationship of the dependent variable to one or more independent variables. Logistic regression is a predictive model and as such is an important tool in machine learning. It allows algorithms to classify incoming data based on historical data. These binary outcomes allow a straightforward decision between the two alternatives. A better fit to the data is the logistic function (also called the sigmoid function or inverse logit function).The term “logistic” refers to the fact that the model is working with logarithms. Its derivation does NOT come from terms for “logic” or “logical.”The objective of logistic regression is to find the sigmoid curve that best fits the sample data. This process will consist of finding the best values for the intercept and the coefficients that yield the closest fit to the data points.

Light GBM (Light Gradient-Boosting):

Light GBM, short for Light Gradient-Boosting Machine, is a free and open source distributed gradient boosting framework for machine learning, originally developed by Microsoft. It is based on decision tree algorithms and used for ranking, classification and other machine learning tasks.The Light GBM framework supports different algorithms including GBT, Light GBM has many of XG Boost's advantages, including sparse optimization, parallel training, multiple loss functions, regularization, bagging, and early stopping. A major difference between the two lies in the construction of trees. Exclusive feature bundling (EFB) is a near-lossless method to reduce the number of effective features. In a sparse feature space many features are nearly exclusive, implying they rarely take nonzero values simultaneously. One-hot encoded features are a perfect example of exclusive features. EFB bundles these features, reducing dimensionality to improve efficiency while maintaining a high level of accuracy. The bundle of exclusive features into a single feature is called an exclusive feature bundle.

Methodology:

a) Data Collection and Preprocessing:

This dataset originates from Kaggle, a famous machine learning community. It offers personal information from borrowers across 18 different aspects, providing a multidimensional depiction of their living conditions. Datasets is more closely aligned with the interactive information that can occur in real-world lending scenarios. Hence, it possesses a higher degree of authority.

The dataset has features like: age, income, whether the individual has taken a personal loan, and a few other financial and demographic characteristics.

b) Data Pre-Processing:

Checking data distribution before modelling is essential to understand the characteristics of the dataset. It was observed that the data is imbalanced, meaning there are significantly more non-default cases compared to default cases. To address this issue**,** SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest neighbours) is applied. SMOTE generates synthetic samples for the minority class, while ENN removes noisy and borderline samples from the majority class, ensuring a more balanced and cleaner dataset for training the machine learning models.

c) **Exploratory Data Analysis (EDA):**Bar plots are used to analyse connections between different categorical variables.Correlation Matrix: We also build a correlation matrix to identify the linear relationships between numerical variables.

d) **Model Implementation and Evaluation**

Partial Dependence plots that help you interpret how a feature (x) affects the target (y) when the machine learning model is trained.A logistic regression with cross-validation, random forest and Light GBM classifiers is done as a baseline model with SMOTE.The logistic regression, random forest and Light GBM classifiers are trained using SMOTEENN and their performance is evaluated based on accuracy and other metrics.

e) **Real-Time Prediction Model**

The serial best model are Light GBM with SMOTEENN, Random Forest with SMOTEENN and logistic regression with SMOTEENN are deployed for the prediction of loan default in real-time and comparison is done between this three models.The system makes predictions in real-time when new data is entered, improving decision making for financial institutions.

Statistical Analysis:

a) **Exploratory Data Analysis (EDA):** We Start with creating subplot for numerical features with more appropriate binning. Here bins are created to present numerical features more clearly.

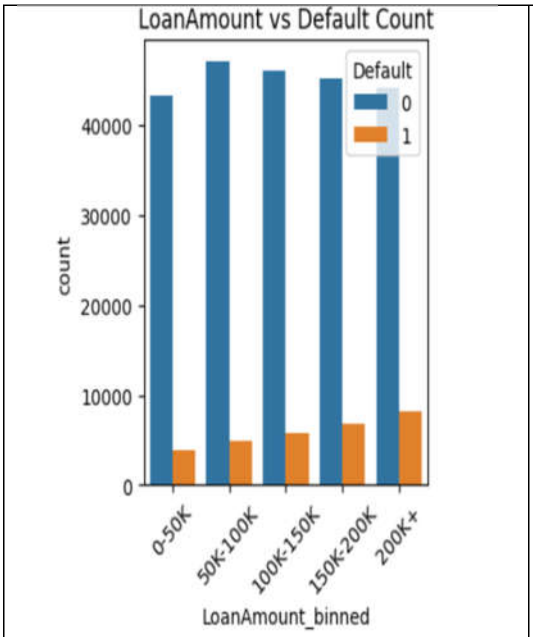


Fig. a

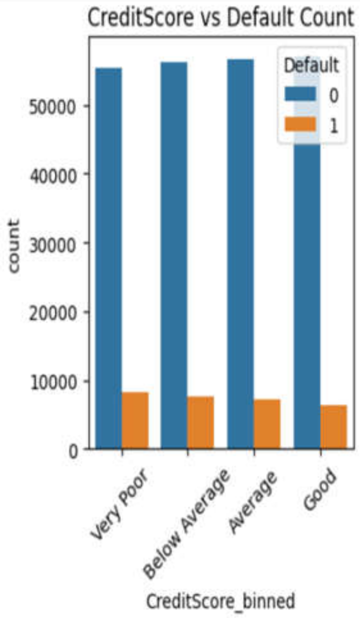


Fig. b

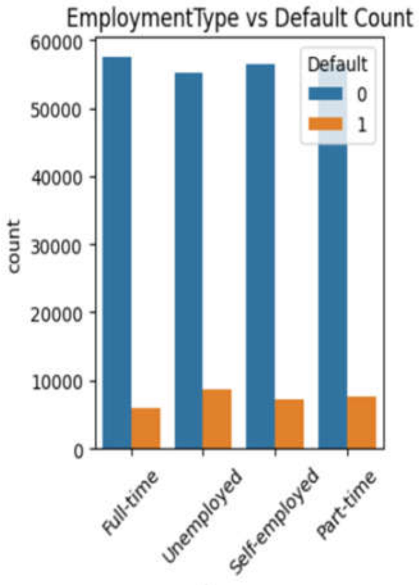


Fig. c

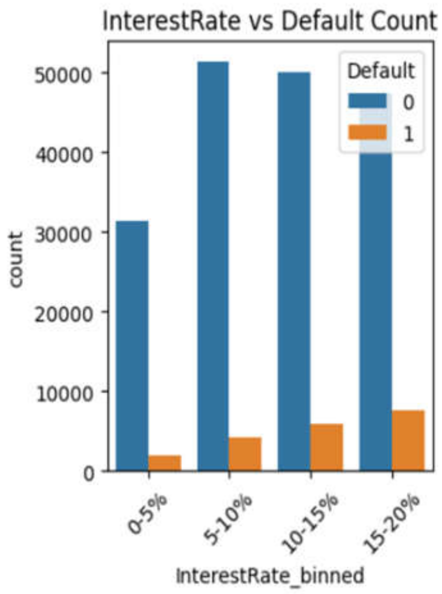
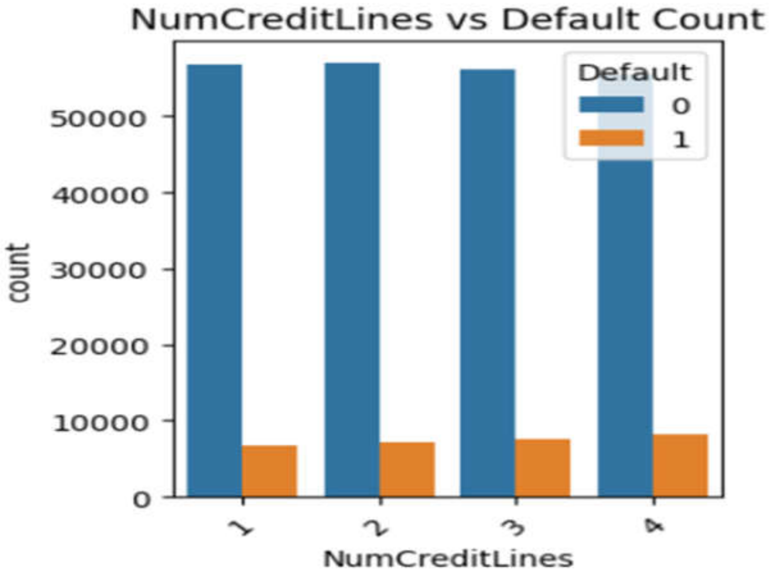
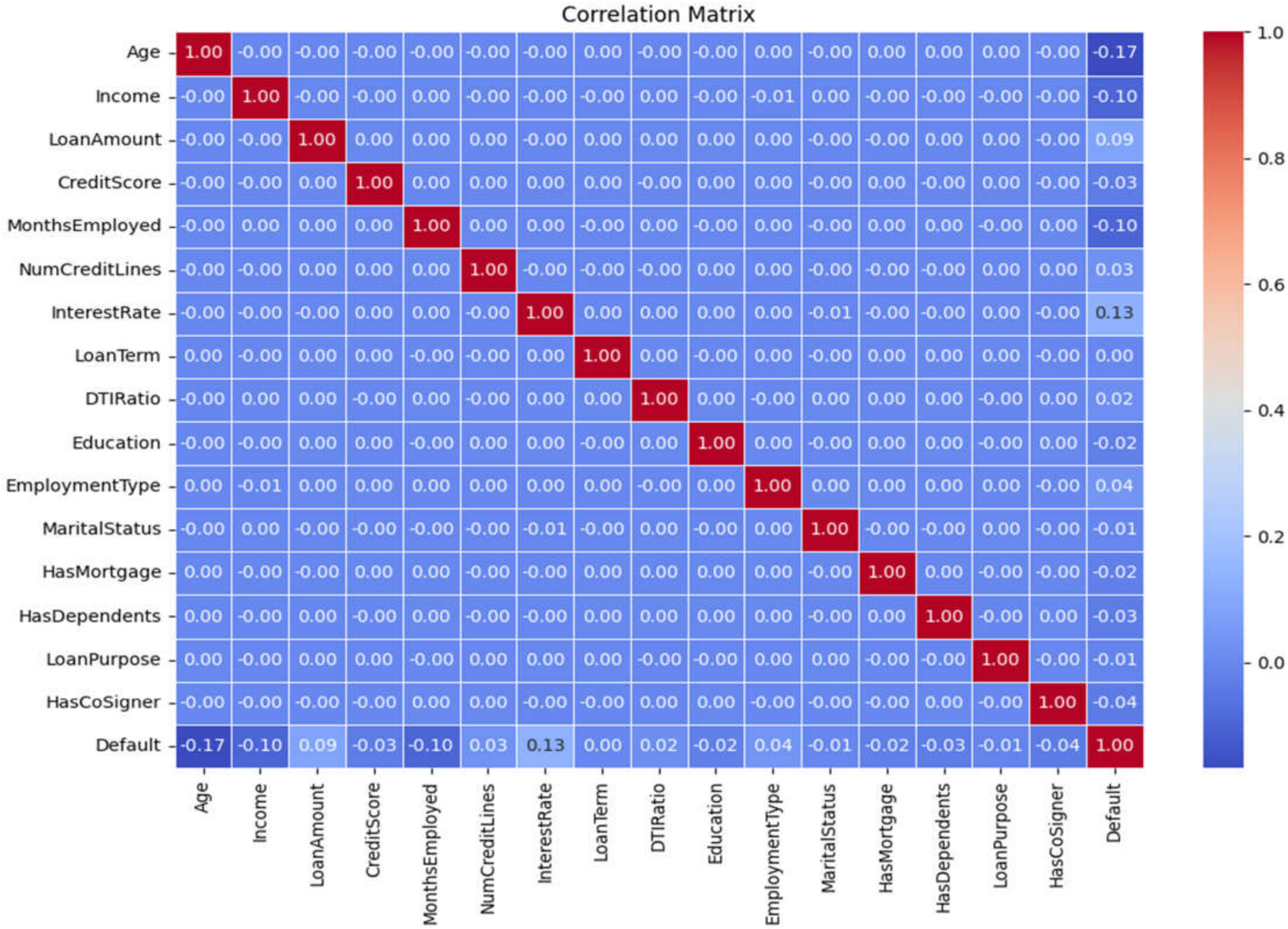
	

Fig. d	Fig. e

b) **Correlation Heat Map**: To visually represent the strength and direction of relationships between multiple variables in a dataset



	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio	Education	EmploymentType	MaritalStatus	HasMortgage	HasDependents	LoanPurpose	HasCoSigner	Default
Age	1.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.17
Income	-0.00	1.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.01	0.00	-0.00	-0.00	-0.00	-0.00	-0.10
LoanAmount	-0.00	-0.00	1.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.09
CreditScore	-0.00	-0.00	0.00	1.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.03
MonthsEmployed	-0.00	0.00	0.00	0.00	1.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.10
NumCreditLines	-0.00	-0.00	0.00	0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.03
InterestRate	-0.00	-0.00	-0.00	0.00	0.00	-0.00	1.00	0.00	0.00	0.00	0.00	-0.01	-0.00	-0.00	0.00	-0.00	0.13
LoanTerm	0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	1.00	0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00
DTIRatio	-0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	1.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	0.02
Education	-0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	1.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.02
EmploymentType	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	1.00	0.00	0.00	0.00	0.00	-0.00	0.04
MaritalStatus	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.00	0.00	-0.00	0.00	1.00	-0.00	-0.00	0.00	-0.00	-0.01
HasMortgage	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	1.00	0.00	-0.00	-0.00	-0.02
HasDependents	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	1.00	-0.00	0.00	-0.03
LoanPurpose	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	1.00	-0.00	-0.01
HasCoSigner	-0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	1.00	-0.04
Default	-0.17	-0.10	0.09	-0.03	-0.10	0.03	0.13	0.00	0.02	-0.02	0.04	-0.01	-0.02	-0.03	-0.01	-0.04	1.00

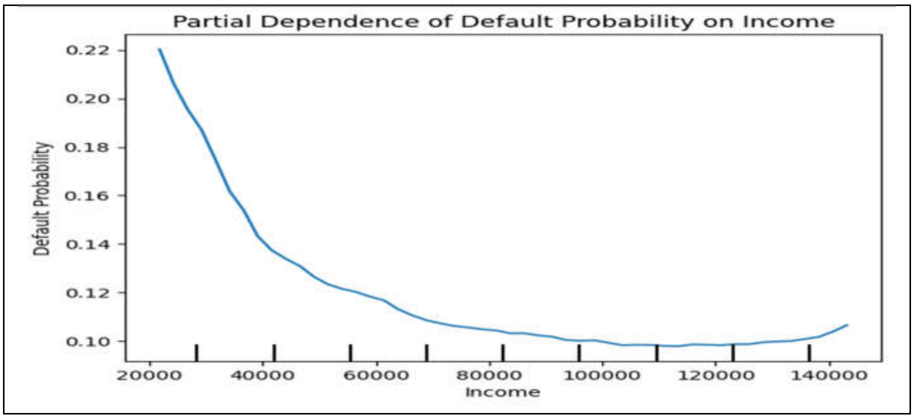
Since there is no strong correlation of default with any other variable we will use feature selection using random forest. Correlation matrix does not show high absolute correlations among the variables, it means pairwise linear relationships between variables are not significant.

c) To check the same, we will Check Variance Inflation Factor

All variables have VIF values close to 1, which suggests that none of the variables is highly correlated with others.

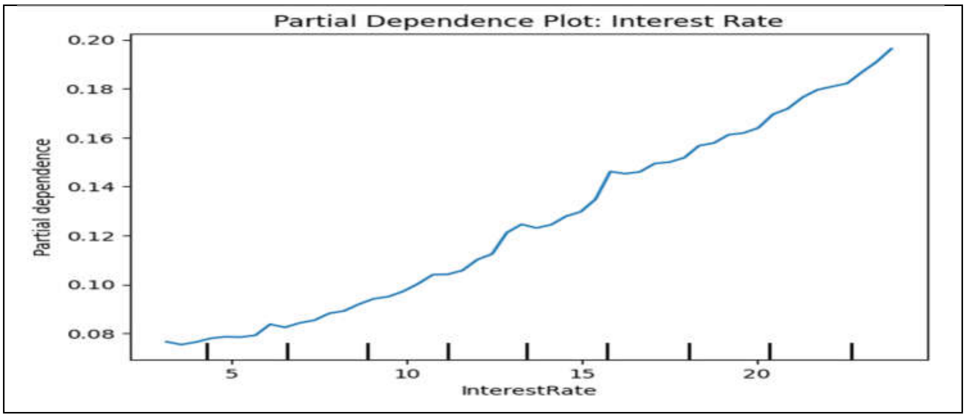
Features	const	Age	Income	Loan Amount	Credit Score	Months Employed	Num Credit Lines	Interest Rate	Loan Term	DT Ration
VIF	51.59	1	1	1	1	1	1	1	1	1

d) **Partial Dependence Plot: To** Analysing the Effect of Income on Default ProbabilityA PDP allows us to visualize the relationship between a feature and the predicted outcome, while keeping all other features constant. This can help us interpret the effect of a particular feature on the model's predictions. In this analysis, we'll explore how the Income feature impacts the likelihood of Default in a loan dataset, using a Random Forest Classifier model. By visualizing the partial dependence of Income, we can draw conclusions about whether higher income is associated with a higher or lower probability of loan default.



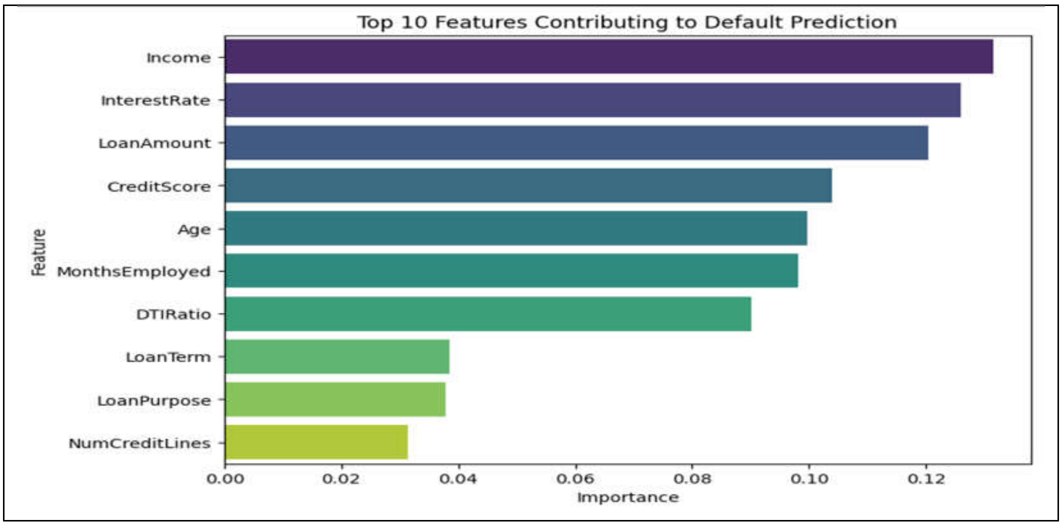
The y-axis (Partial Dependence) indicates the predicted probability of default, independent of the effect of other features. This means how the model predicts the probability of default when income changes, keeping other features constant.The graph shows a decreasing trend of partial dependence as Income increases. This suggests that, according to the model, higher-income individuals are associated with a lower probability of default, whereas lower-income individuals have a higher probability of default. Essentially, income plays a protective role against default risk, where low-income individuals are more likely to default compared to their high-income counterparts.

e) Partial Dependence on Interest Rate:



As visible in graph, as interest rate increases, probability of default for customer increases

f) **Feature Selection using Random Forest:**A random forest classifier is employed to rank the features based on their importance.The most significant features are selected for model training to enhance performance and reduce complexity.



The top 10 features listed contribute to the prediction of default according to their importance scores derived from the Random Forest model. Income and Interest Rate are the most influential features, indicating that financial capacity and loan affordability play a critical role in predicting defaults.

Model Fitting: SMOTE (Synthetic Minority Over-sampling Technique):

Logistic Regression with cross validation:
Cross-validated AUC-ROC scores: [0.73224896 0.71295249 0.71176041 0.71536291 0.74964025]
Mean AUC-ROC: 0.724393003831781

	precision	**recall**	**f1-score**	**support**
0	0.94	0.68	0.79	67681
1	0.21	0.65	0.32	8924
accuracy			0.68	76605
macro avg	0.58	0.67	0.56	76605
weighted avg	0.85	0.68	0.74	76605

AUC-ROC on test set: 0.7289756635180312

Interpretation:
Recall for class 1 (Default) is low (0.65). This indicates the model misses many of the positive cases.

Precision for class 1 is also low (0.21), meaning many of the predicted defaults are actually false positives.

Random Forest Classifier: Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.90	0.84	0.87	67681
1	0.21	0.31	0.25	8924
accuracy			0.78	76605
macro avg	0.56	0.58	0.56	76605
weighted avg	0.82	0.78	0.8	76605

AUC-ROC on test set: 0.6742163025426495

Interpretation:

The recall for class 1 (Default) is low (0.31). The model misses a large number of actual positive cases (defaults).

The precision for class 1 is also low (0.21), indicating many false positives.

Light GBM model:

Light GBM report:

	precision	recall	f1-score	support
0	0.9	0.95	0.92	67681
1				8924
accuracy			0.86	76605
macro avg	0.61	0.56	0.58	76605
weighted avg	0.83	0.86	0.84	76605

Light GBM Accuracy: 0.8623196919261145

Interpretation:

Class 0 (Non-default):

Precision (90%): Out of all predicted non-default cases, 90% were correctly classified.

Recall (95%): The model successfully identified 95% of actual non-default cases.

F1-Score (92%): A good balance between false positives and false negatives, indicating high confidence in non-default predictions.

Class 1 (Default):

Precision (33%): Out of all predicted default cases, only 33% were actual defaulters.

Recall (17%): The model captured only 17% of actual defaulters, missing a significant portion.

F1-Score (23%): Indicates low reliability in identifying defaulters, with a high trade-off between false positives and false negatives.

SMOTEENN (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors):

Logistic Regression : Classification Report:

	precision	recall	f1-score	support
0	0.69	0.57	0.62	21803
1	0.72	0.82	0.77	30377
accuracy			0.71	52180
macro avg	0.71	0.69	0.69	52180
weighted avg	0.71	0.71	0.71	52180

AUC-ROC: 0.77205694415503

Interpretation:

Class 0(Non-default):

Precision (69%): Out of all predicted non-default cases, 69% were correct.

Recall (57%): The model correctly identified 57% of actual non-defaulters.

F1-Score (62%): A moderate balance between precision and recall, suggesting some misclassification of non-defaulters as defaulters.

Class 1 (Default):

Precision (72%): Out of all predicted default cases, 72% were actual defaulters.

Recall (82%): The model successfully identified 82% of actual defaulters, demonstrating strong sensitivity to high-risk borrowers.

F1-Score (77%): A good balance between false positives and false negatives, showing reliability in identifying defaulters.

Random Forest : Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.90	0.78	0.84	21803
1	0.86	0.93	0.89	30377
accuracy			0.87	52180
macro avg	0.88	0.86	0.86	52180
weighted avg	0.87	0.87	0.87	52180

AUC-ROC: 0.9440075434887428

Interpretation:

Class 0 (Non-default):

Precision (90%): Out of all predicted non-default cases, 90% were correct.

Recall (78%): The model identified 78% of the actual non-default cases correctly.

F1-Score (84%): This is a harmonic mean of precision and recall, indicating good balance between false positives and false negatives.

Class 1 (Default):

Precision (86%): Out of all predicted default cases, 86% were correct.

Recall (93%): The model captured 93% of the actual default cases.

F1-Score (89%): Indicates high reliability in identifying default cases with minimal trade-off between false positives and false negatives.

Light GBM : Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.85	0.91	0.88	21803
1	0.93	0.88	0.91	30377
accuracy			0.89	52180
macro avg	0.89	0.90	0.89	52180
weighted avg	0.90	0.89	0.89	52180

AUC-ROC: 0.9570254992976994

Interpretation Class 0 (Non-default):

Precision (85%): Out of all predicted non-default cases, 85% were correctly identified.

Recall (91%): The model successfully identified 91% of the actual non-default cases.

F1-Score (88%): A balanced metric showing good performance in terms of both precision and recall.

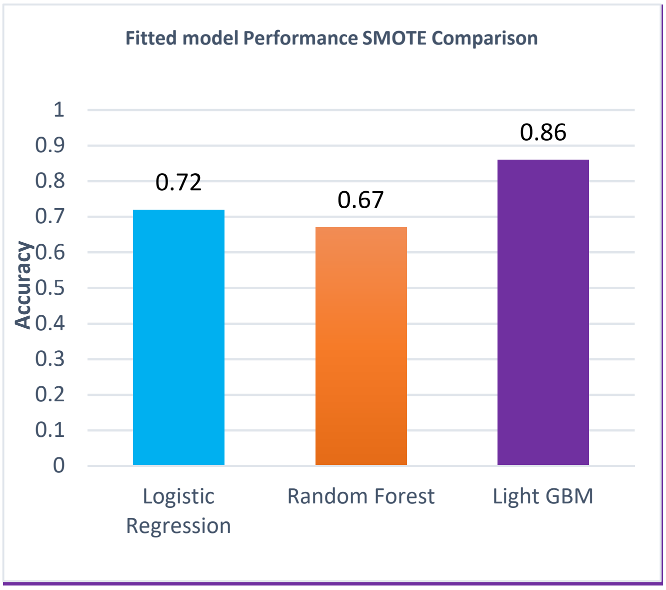
Class 1 (Default):

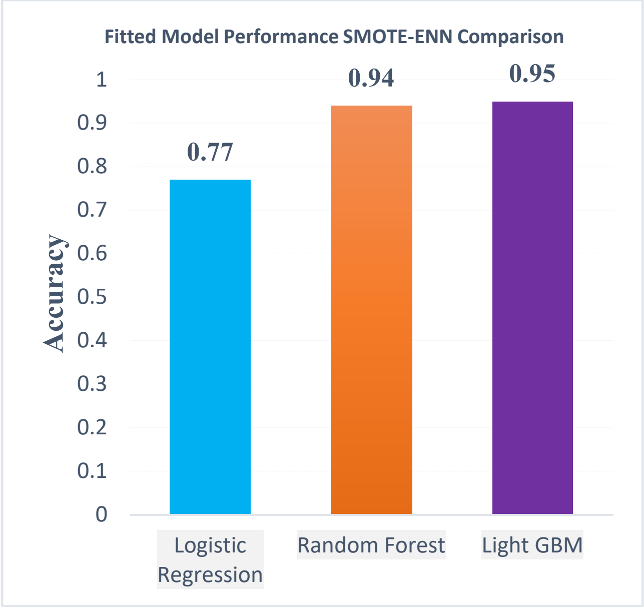
Precision (93%): Out of all predicted default cases, 93% were correct.

Recall (88%): The model captured 88% of the actual default cases.

F1-Score (91%): Demonstrates strong reliability in identifying default cases with minimal false negatives or positives.

Comparison Plot::





Real Time Prediction Model:

1. For light GBM

- **Train model**

```
from lightgbm import LGBMClassifier

# Train the model on your dataset (assuming X and y are already defined)
lgbm = LGBMClassifier(class_weight='balanced', random_state=42)
lgbm.fit(X_train, y_train)

# Save the trained model
joblib.dump(lgbm, 'Loan_Default_Prediction.pkl')
```

- **Input**

```
new_customer = pd.DataFrame({
    'Age': [35],
    'Income': [50000],
    'LoanAmount': [15000],
    'CreditScore': [700],
    'MonthsEmployed': [24],
    'InterestRate': [5],
    'LoanTerm': [36],
    'DTIRatio': [0.4],
})
# Predict using the trained model
probability = lgbm.predict_proba(new_customer)[:, 1]

# Classify as Defaulter (1) or Non-Defaulter (0)
threshold = 0.5
prediction = (probability >= threshold).astype(int)

# Print result
```

```
if prediction[0] == 1:
    print("The customer is likely a DEFAULTER.")
else:
    print("The customer is NOT a defaulter.")
print(f"Default Probability: {probability[0]:.2f}")
```

- **Output**

```
The customer is likely a DEFAULTER.
Default Probability: 0.92
```

2. For Random Forest

- **Train model**

```
from sklearn.ensemble import RandomForestClassifier

# Train the model on your dataset (assuming X and y are already defined)
rf = RandomForestClassifier(n_estimators=100, class_weight='balanced', random_state=42)
rf.fit(X_train, y_train)

# Save the trained model
joblib.dump(rf, 'Loan_Default_Prediction.pkl')
```

- **Input**

```
new_customer = pd.DataFrame({
    'Age': [35],
    'Income': [50000],
    'LoanAmount': [15000],
    'CreditScore': [700],
    'MonthsEmployed': [24],
    'InterestRate': [5],
    'LoanTerm': [36],
    'DTIRatio': [0.4],
})
# Predict using the trained model
probability = rf.predict_proba(new_customer)[:, 1]

# Classify as Defaulter (1) or Non-Defaulter (0)
threshold = 0.5
prediction = (probability >= threshold).astype(int)

# Print result
if prediction[0] == 1:
    print("The customer is likely a DEFAULTER.")
else:
    print("The customer is NOT a defaulter.")

print(f"Default Probability: {probability[0]:.2f}")
```

- **Output**

```
The customer is likely a DEFAULTER.
Default Probability: 0.90
```

3. For Logistic Regression

- **Train model**

```
from sklearn.linear_model import LogisticRegression

# Train the model on your dataset (assuming X and y are already defined)
```

```
log_reg = LogisticRegression(class_weight='balanced', random_state=42)
log_reg.fit(X_train, y_train)
# Save the trained model
joblib.dump(log_reg, 'Loan_Default_Prediction.pkl')
```

- **Input**

```
new_customer = pd.DataFrame({
   'Age': [35],
   'Income': [50000],
   'LoanAmount': [15000],
   'CreditScore': [700],
   'MonthsEmployed': [24],
   'InterestRate': [5],
   'LoanTerm': [36],
   'DTIRatio': [0.4],
})

# Predict using the trained model
probability = log_reg.predict_proba(new_customer)[:, 1]
# Classify as Defaulter (1) or Non-Defaulter (0)
threshold = 0.5
prediction = (probability >= threshold).astype(int)
# Print result
if prediction[0] == 1:
   print("The customer is likely a DEFAULTER.")
else:
   print("The customer is NOT a defaulter.")
print(f"Default Probability: {probability[0]:.2f}")
```

- **Output**

```
The customer is NOT a defaulter.
Default Probability: 0.28
```

Conclusion:

From graphical representation we conclude that if loan amount increases, the cases for default increases and if credit score improves, count for default decreases also employment type, unemployed people have highest number of default cases as well as if interest rate increases, count for default also increases and as count of no. of credit lines increases, default count increases. we observed that from correlation Heatmap and VIF factor it shows that there is no correlation between the variables also we to identify top ten features which are highly affected on default prediction using random forest . Using SMOTE and SMOTE ENN methods we constructed three models Random Forest, Logistic Regression, Light GBM from this our study shows that all of these three models based on Smote ENN technique perform excellent with high accuracy to accurately predict loan defaulter as compared to model based on SMOTE.

References

[1] Aslam U, Aziz H I T, Sohail A and Batcha N K 2019 An empirical study on loan default prediction models Journal of Computational and Theoretical Nanoscience 16 pp 3483–8

[2] Li Y 2019 Credit risk prediction based on machine learning methods The 14th Int. Conf. on Computer Science & Education (ICCSE) pp 1011–3

[3] Ahmed M S I and Rajaleximi P R 2019 An empirical study on credit scoring and credit scorecard for financial institutions Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET) 8 275–9

[4] Zhu L, Qiu D, Ergu D, Ying C and Liu K 2019 A study on predicting loan default based on the random forest algorithm The 7th Int. Conf. on Information Technol. and Quantitative Management (ITQM) 162 pp 503–13

[5] Ghatasheh N 2014 Business analytics using random forest trees for credit risk prediction: a comparison study Int. Journal of Advanced Science and Technol. 72 pp 19–30 [6] Breeden J L 2020 A survey of machine learning in credit risk

[7] Madane N and Nanda S 2019 Loan prediction analysis using decision tree Journal of The Gujarat Research Society 21 pp214–21

[8] Supriya P, Pavani M, Saisushma N, Kumari N V and Vikas K 2019 Loan prediction by using machine learning models Int. Journal of Engineering and Techniques 5 pp144–8

[9] Amin R K, Indwiarti and Sibaroni Y 2015 Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (case study: bank pasar of yogyakarta special region) The 3rd Int. Conf. on Information and Communication Technol. (ICoICT) pp 75–80

[10] Jency X F, Sumathi V P and Sri J S 2018 An exploratory data analysis for loan prediction based on nature of the clients Int. Journal of Recent Technol. and Engineering (IJRTE) 7 pp 176–9

[11] Shoumo S Z H, Dhruba M I M, Hossain S, Ghani N H, Arif H and Islam S 2019 Application of machine learning in credit risk assessment: a prelude to smart banking TENCON 2019 – 2019 IEEE Region 10 Conf. (TENCON) pp 2023–8

[12] Addo P M, Guegan D and Hassani B 2018 Credit risk analysis using machine and deep learning models Risks 6 p 38

[13] Hamid A J and Ahmed T M 2016 Developing prediction model of loan risk in banks using data mining Machine Learning and Applications: An Int. Journal (MLAIJ) 3 pp 1–9 [14] Kacheria A, Shivakumar N, Sawkar S and Gupta A 2016 Loan sanctioning prediction system Int. Journal of Soft Computing and Engineering (IJSCE) 6 pp 50–3

[15] Vojtek M and Kocenda E 2006 Credit scoring methods Finance a uver - Czech Journal of Economics and Finance 56 pp 152–167

[16] Russel S and Norvig P 1995 Artificial intelligence - a modern approach

[17] Alshouiliy K, Alghamdi A and Agrawal D P 2020 AzureML based analysis and prediction loan borrowers creditworthy The 3rd Int. Conf. on Information and Computer Technologies (ICICT) 1 pp 302–6

[18] Li M, Mickel A and Taylor S 2018, "Should this loan be approved or denied?": a large dataset with class assignment guidelines Journal of Statistics Education 26 pp 55–66

[19] Vaidya A 2017 Predictive and probabilistic approach using logistic regression: application to prediction of loan approval The 8th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT) 1 pp 1–6

[20] Murphy K P 2012 Machine learning: a probabilistic approach