DSA5102 PROJECT
Loan Default Prediction

CHAI JIAYING (A0296685E)

Introduction

Financial loan services are widely used by banks, financial institutions, and government entities to manage lending risks. By using machine learning to **predict potential loan defaults,** companies can proactively identify high-risk individuals, enabling timely interventions to **minimize financial losses** and **improve repayment compliance**.

Goal: To predict loan default risk accurately using machine learning models trained on sample data.

DATASET

The dataset contains **255,347 rows** and **18 columns.**

Default	
0	225694
1	29653

The dataset is **unbalanced**, so we take a **subsample** with an equal ratio of positive and negative samples.

Default	
0	25000
1	25000

FEATURES

17 features: 9 numerical features, 7 categorical features (including 3 binary features)

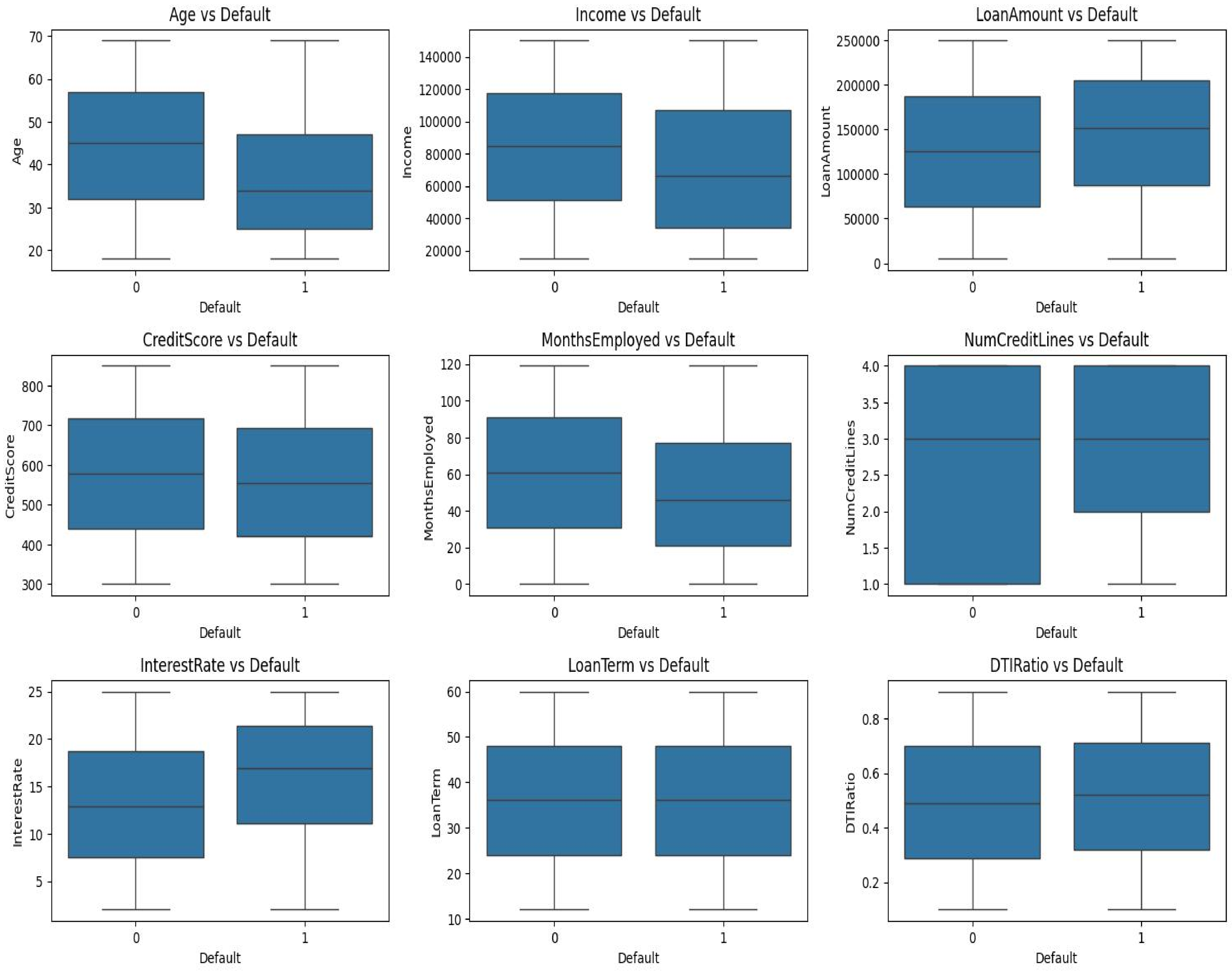
	Column.name	Description	
0	Loan ID	A unique identifier for each loan.	
1	Age	The age of the borrower.	
2	Income	The annual income of the borrower.	
3	LoanAmount	The amount of money being borrowed.	
4	CreditScore	The credit score of the borrower, indicating their creditworthiness.	
5	MonthsEmployed	The number of months the borrower has been employed.	numerical
6	NumCreditLines	The number of credit lines the borrower has open.	
7	InterestRate	The interest rate for the loan.	
8	LoanTerm	The term length of the loan in months.	
9	DTI Ratio	The Debt-to-Income ratio, indicating the borrowers debt compared to their income.	
10	Education	The highest level of education attained by the borrower.	
11	EmploymentType	The type of employment status of the borrower.	categorical
12	MaritalStatus	The marital status of the borrower.	
13	LoanPurpose	The purpose of the loan.	
14	HasMortgage	Whether the borrower has a mortgage.	
15	HasDependents	Whether the borrower has dependents.	binary
16	HasCoSigner	Whether the loan has a co-signer.	
17	Default	The binary target variable indicating whether the loan defaulted (1) or not (0).	

Visualize Data

numerical features

1	**Age**	The age of the borrower.
2	**Income**	The annual income of the borrower.
3	**LoanAmount**	The amount of money being borrowed.
4	**CreditScore**	The credit score of the borrower, indicating their creditworthiness.
5	**MonthsEmployed**	The number of months the borrower has been employed.
6	**NumCreditLines**	The number of credit lines the borrower has open.
7	**InterestRate**	The interest rate for the loan.
8	**LoanTerm**	The term length of the loan in months.
9	**DTI Ratio**	The Debt-to-Income ratio, indicating the borrowers debt compared to their income.

- Older people are seen less likely to default
- The dafaulters are seen with lower avg income
- The dafaulters are seen with larger Loan Amount and lesser credit score
- People having higher Interest rate are more likely to default

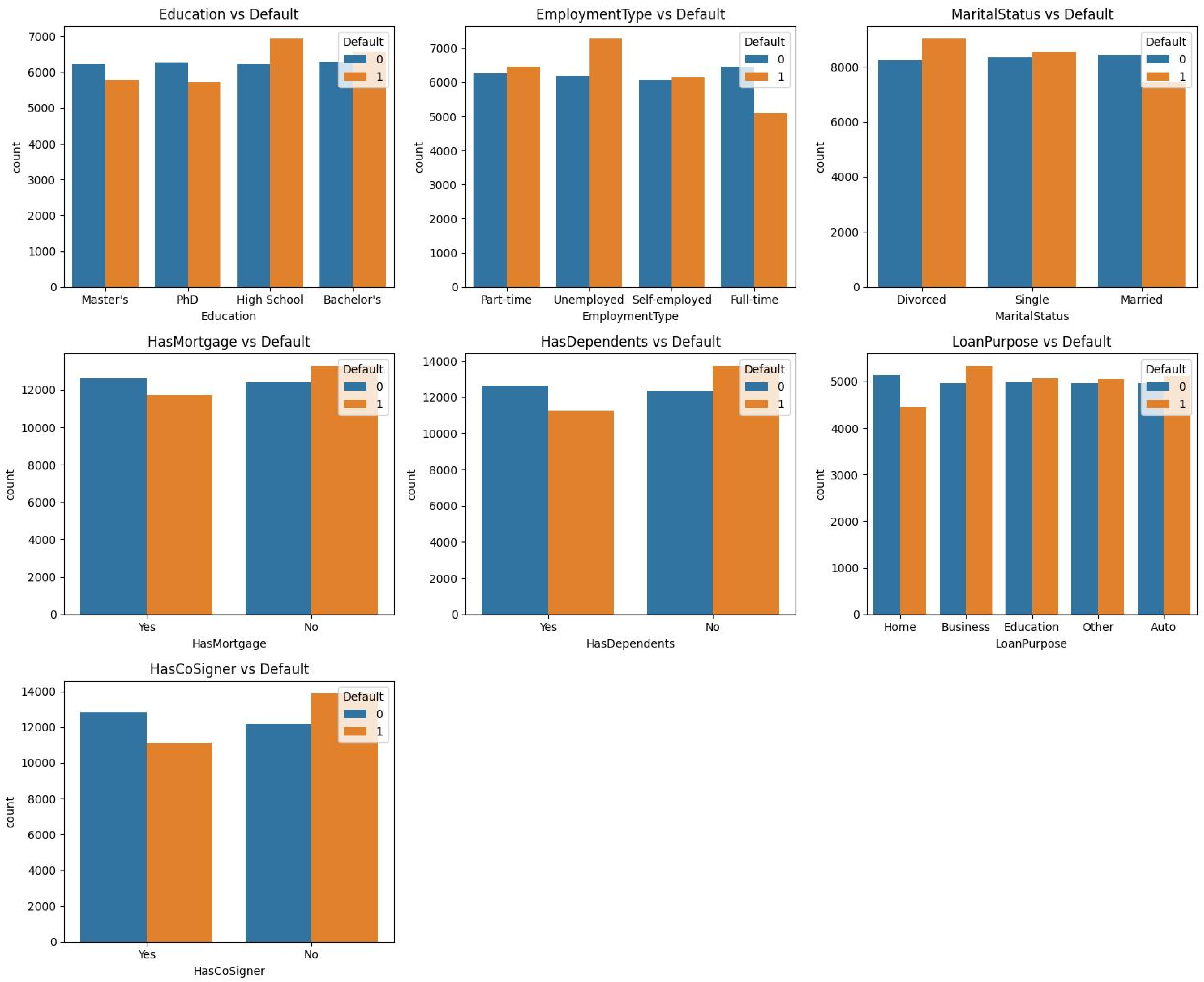


Visualize Data

categorical features

10 Education	The highest level of education attained by the borrower. **(PhD, Masters, Bachelor's, High School)**
11 EmploymentType	The type of employment status of the borrower. **(Full-time, Part-time, Self-employed, Unemployed)**
12 MaritalStatus	The marital status of the borrower. **(Single, Married, Divorced)**
13 LoanPurpose	The purpose of the loan. **(Home, Auto. Education. Business, Other)**
14 HasMortgage	Whether the borrower has a mortgage.**(Yes or No)**
15 HasDependents	Whether the borrower has dependents.**(Yes or No)**
16 HasCoSigner	Whether the loan has a co-signer.**(Yes or No)**

- Higher education people are seen less likely to default
- Unemployed people are seen more likely to default
- People without mortgage or dependent or co-signer are more likely to default



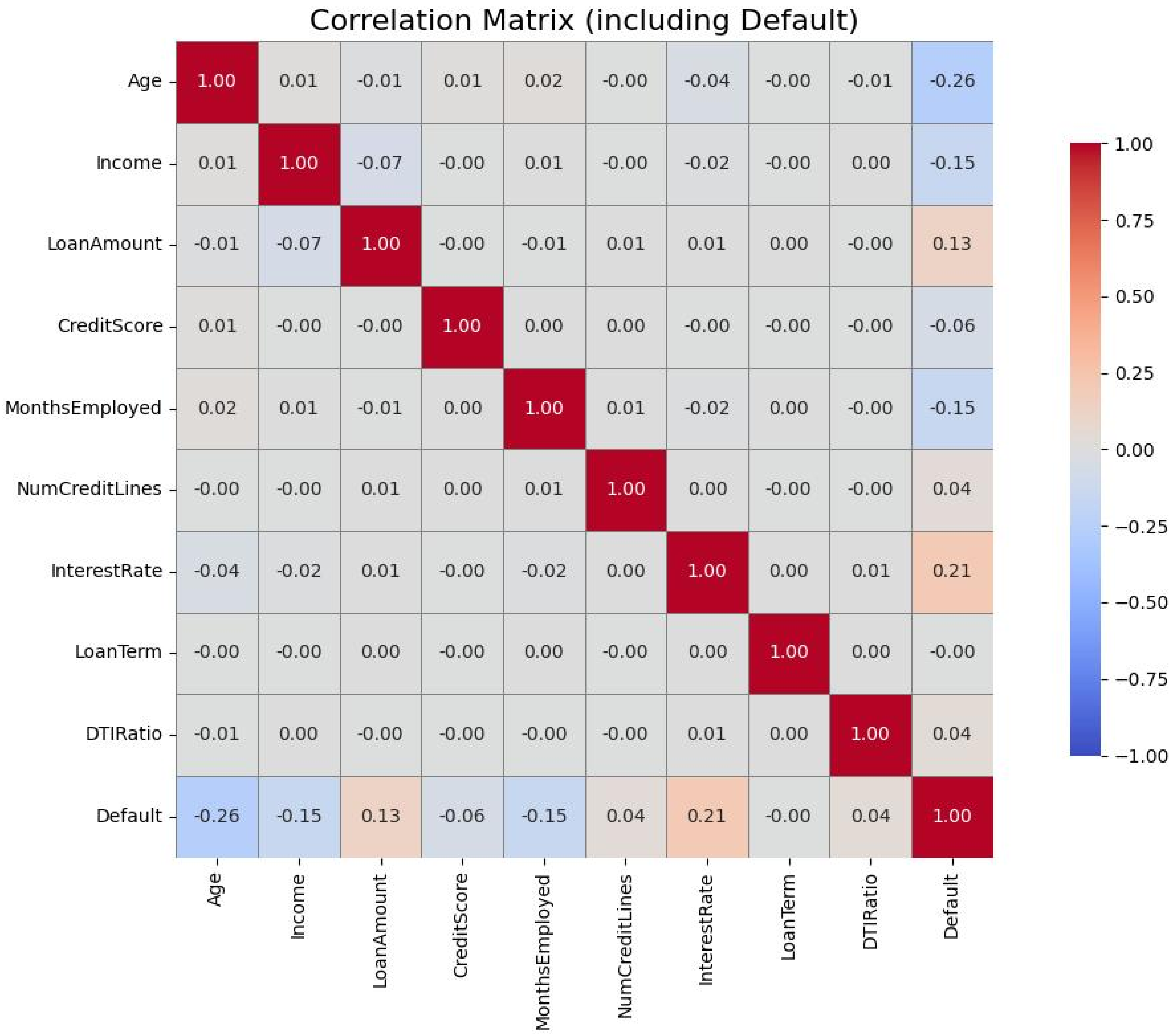
Visualize Data

Correlation Matrix

No significant correlation found

Importance:

Age	-0.26
Interest Rate	0.21
Income	-0.15
MonthsEmployed	-0.15
LoanAmount	0.13



Preprocessing Data

Feature Engineering

$$\text{LoanToIncomeRatio} = \frac{\text{LoanAmount}}{\text{Income}}$$

$$\text{CreditUtilizationRate} = \frac{\text{LoanAmount}}{\text{CreditScore}}$$

Lable Encoder

- convert categorical and binary data into numerical values
- Apply LabelEncoder to binary columns
  'HasMortgage', 'HasDependents', 'HasCoSigner'
- Apply one-hot encoding to categorical columns
  'Education', 'EmploymentType', 'MaritalStatus', 'LoanPurpose'

Standard Scaler

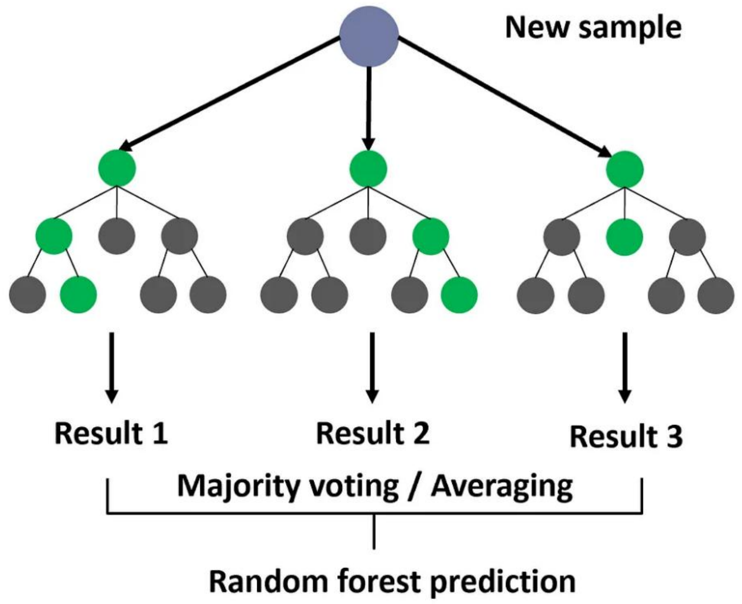
The order of magnitude difference between the different features is very large, so consider normalizing the data.

Splitting the dataset

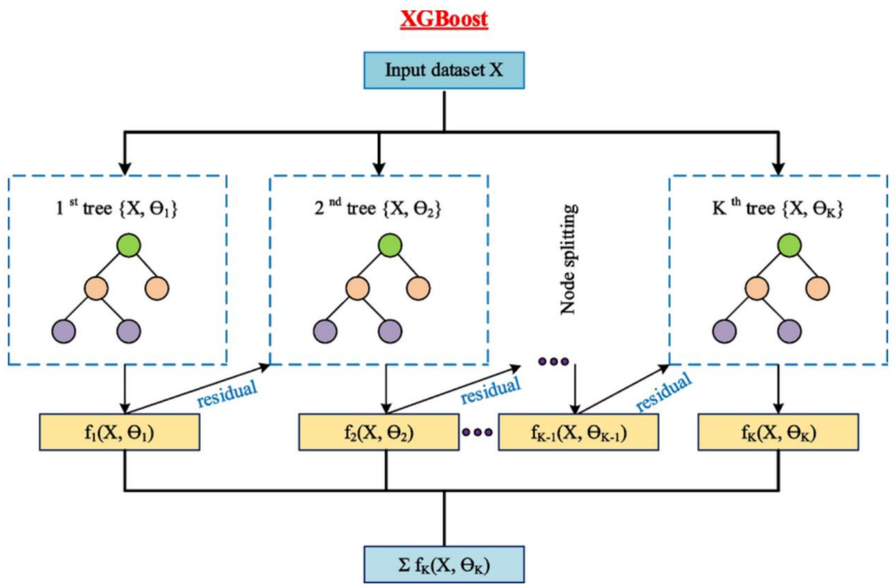
- Split the dataset into features and target variable
- Train-test split (80% train, 20% test)

Model Selection

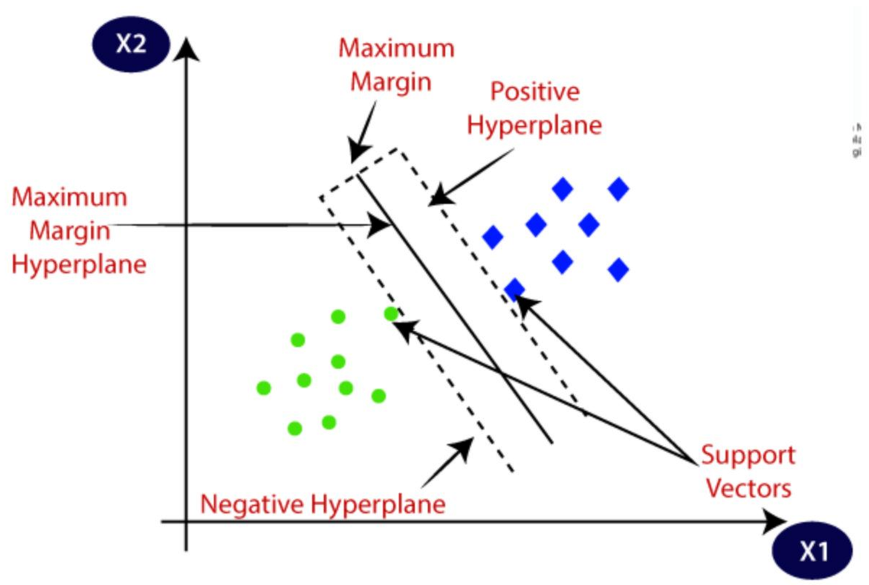
Supervised Learning



Random Forest

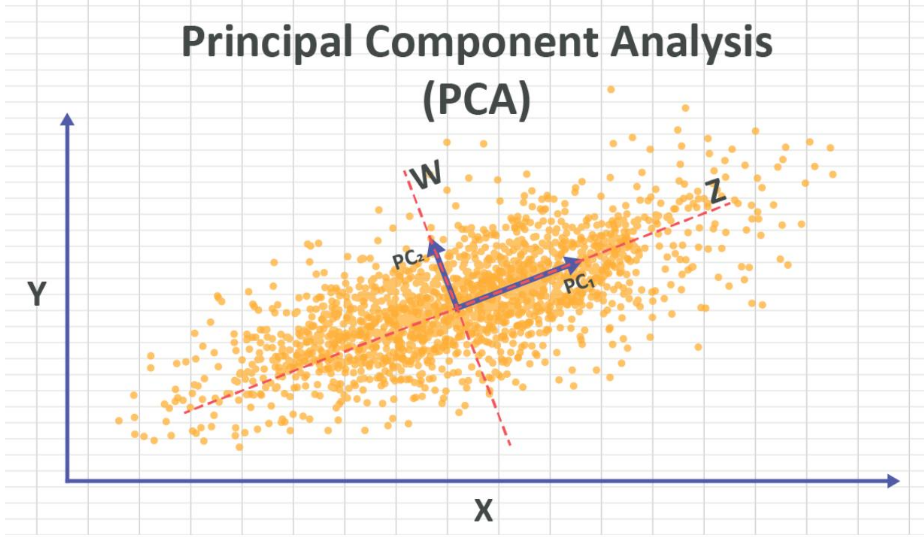


XGBoost

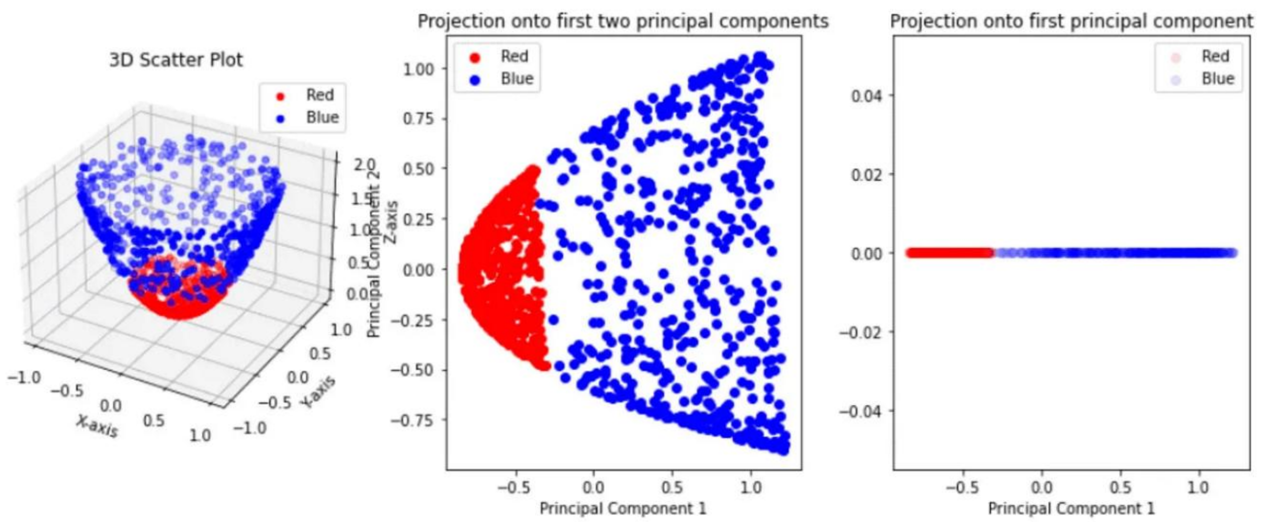


SVM (Support Vector Machine)

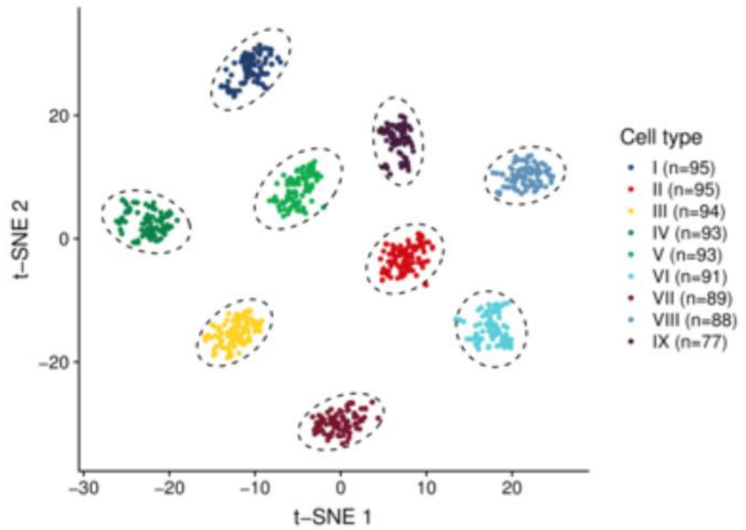
Unsupervised Learning



PCA



Kernel PCA



t-SNE

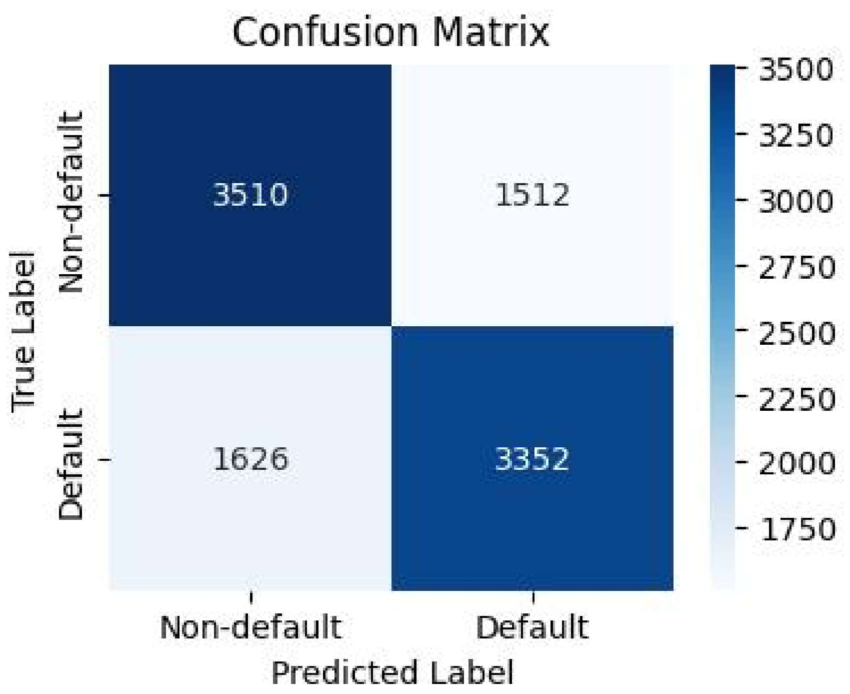
Supervised Learning - Random Forest

Model

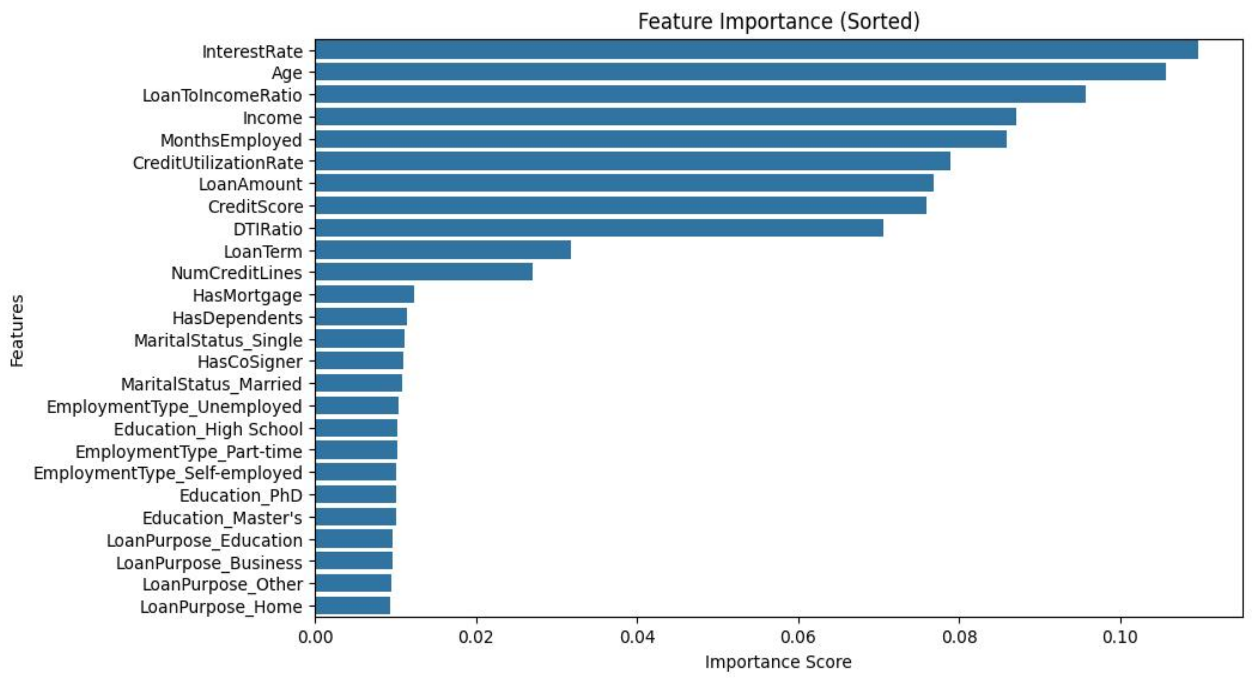
rf_model = RandomForestClassifier (n_estimators=1000, random_state=1, max_features="sqrt", max_depth=34, criterion="entropy")

KFold: n_splits=5

Confusion Matrix



Feature Importance



Classification Report

```
Random Forest Classifier with K-Fold Cross-Validation:
Test Accuracy: 0.6862
Classification Report:
              precision    recall  f1-score   support

         0.0       0.68      0.70      0.69      5022
         1.0       0.69      0.67      0.68      4978

    accuracy                           0.69     10000
   macro avg       0.69      0.69      0.69     10000
weighted avg       0.69      0.69      0.69     10000
```

Evaluation

Model	Accuracy	Recall	AUC	PR AUC
Random Forest	0.686	0.70/0.67	0.748	0.745

· The model performs fairly balanced in capturing both default and non-default samples.

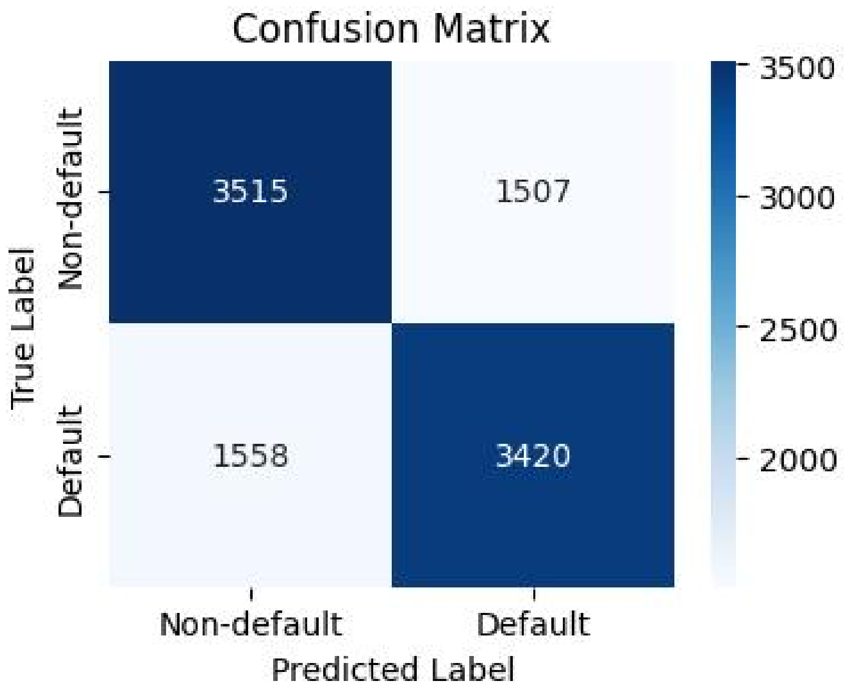
· TOP 3 features: Interest rate, Age, Loan-to-Income ratio

Supervised Learning - XGBoost

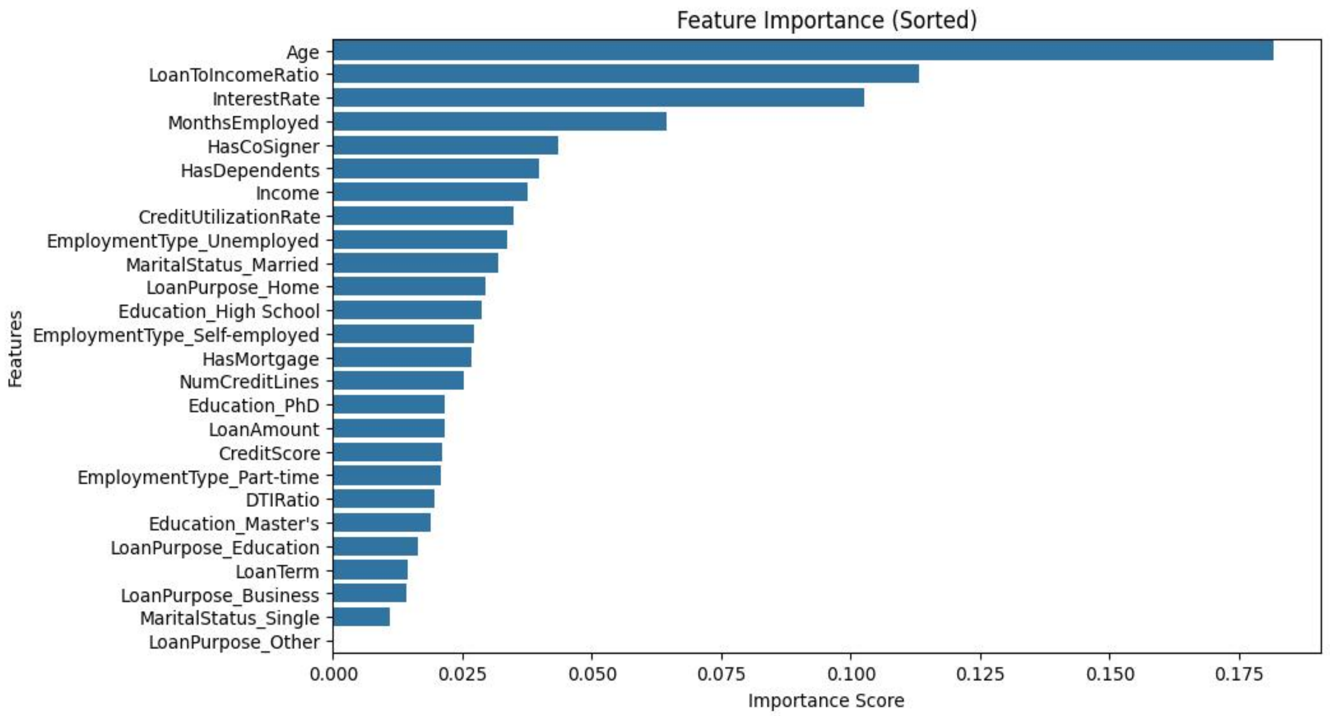
Model

xgb_model = XGBClassifier(max_depth=4, min_child_weight=6, gamma=0.1, subsample=0.8, colsample_bytree=0.8, learning_rate=0.1,n_estimators=100, use_label_encoder=False, eval_metric='logloss', random_state=42)
KFold: n_splits=5

Confusion Matrix



Feature Importance



Classification Report

```
XGBoost Classifier with K-Fold Cross-Validation:
Test Accuracy: 0.6935
Classification Report:
              precision    recall  f1-score   support

         0.0       0.69      0.70      0.70      5022
         1.0       0.69      0.69      0.69      4978

    accuracy                           0.69     10000
   macro avg       0.69      0.69      0.69     10000
weighted avg       0.69      0.69      0.69     10000
```

Evaluation

Model	Accuracy	Recall	AUC	PR AUC
Random Forest	0.686	0.70/0.67	0.748	0.745
XGBoost	0.694	0.70/0.69	0.758	0.755

·The model performs fairly balanced in capturing both default and non-default samples.

· TOP 3 features: Age, Loan-to-Income ratio,Interest rate.

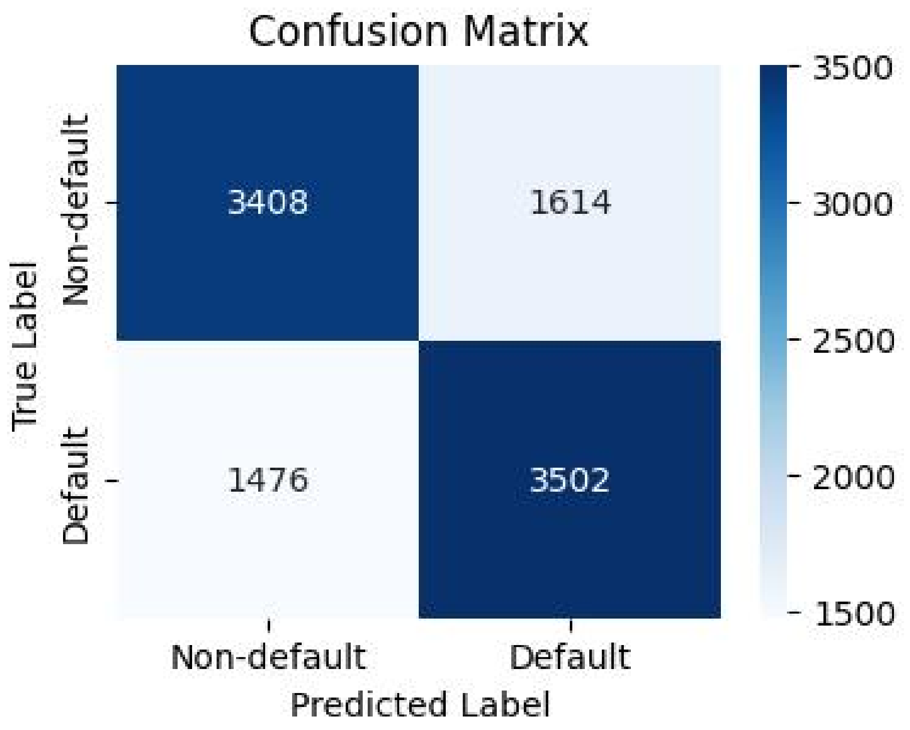
Supervised Learning - SVM

Model

svm_model = SVC(kernel='poly',C=0.03,gamma='scale', probability=True)

KFold: n_splits=5

Confusion Matrix



Classification Report

```
SVM Classifier with K-Fold Cross-Validation:
Test Accuracy: 0.691
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.68      0.69      5022
         1.0       0.68      0.70      0.69      4978

    accuracy                           0.69     10000
   macro avg       0.69      0.69      0.69     10000
weighted avg       0.69      0.69      0.69     10000
```

Evaluation

Model	Accuracy	Recall	AUC	PR AUC
Random Forest	0.686	0.70/0.67	0.748	0.745
XGBoost	0.694	0.70/0.69	0.758	0.755
SVM	0.691	0.68/0.70	0.754	0.745

SVM has lower False Negative (FN) than previous model, indicates that the model has fewer missed judgments for positive class samples.

Unsupervised Learning - PCA

Model

pca = PCA(n_components=5)

kmeans = KMeans(n_clusters=2, random_state=42)

Classification Report

Unbalanced Data

```
Clustering Accuracy with best alignment: 0.575644123486863
              precision    recall  f1-score   support

           0       0.91      0.58      0.71    225694
           1       0.15      0.58      0.24     29653

    accuracy                           0.58    255347
   macro avg       0.53      0.58      0.47    255347
weighted avg       0.82      0.58      0.65    255347
```

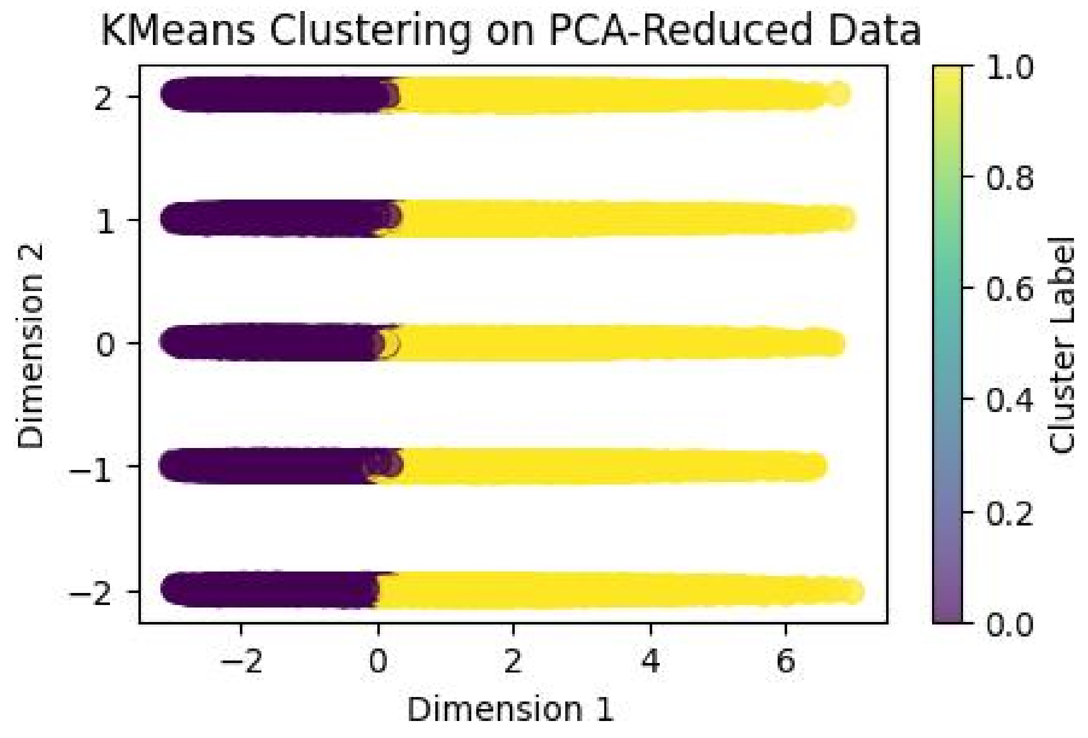
Balanced Data

```
Clustering Accuracy with best alignment: 0.58382
              precision    recall  f1-score   support

           0       0.57      0.69      0.62     25000
           1       0.61      0.48      0.53     25000

    accuracy                           0.58     50000
   macro avg       0.59      0.58      0.58     50000
weighted avg       0.59      0.58      0.58     50000
```

Visualization



Evaluation

Model	Imbalanced data		Balanced data	
	Accuracy	F1-Score	Accuracy	F1-Score
PCA	0.575	0.71/0.24	0.583	0.62/0.53

- **Model use unbalanced data does not perform well on default=1 class.**
- **After balanced the dataset, the model performs better than before, but still worse than supervised learning method.**

Unsupervised Learning - Kernel PCA

Model

```
kpca =
KernelPCA(kernel="rbf",
gamma=0.1, n_components=3)

kmeans =
KMeans(n_clusters=2,
random_state=42)
```

Classification Report

Unbalanced Data

Clustering Accuracy with best alignment: 0.5714

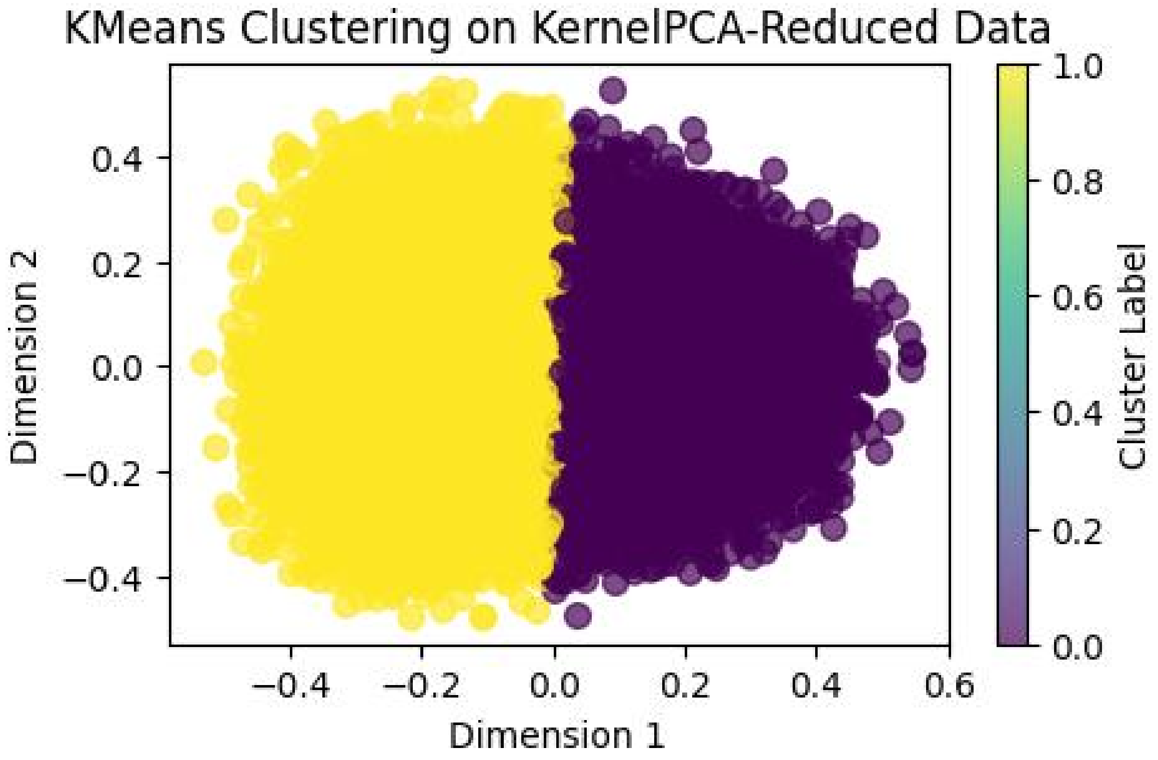
	precision	recall	f1-score	support
0.0	0.58	0.56	0.57	5022
1.0	0.57	0.59	0.58	4978
accuracy			0.57	10000
macro avg	0.57	0.57	0.57	10000
weighted avg	0.57	0.57	0.57	10000

Balanced Data

Clustering Accuracy with best alignment: (0.4237,)

	precision	recall	f1-score	support
0	0.42	0.43	0.43	5000
1	0.42	0.42	0.42	5000
accuracy			0.42	10000
macro avg	0.42	0.42	0.42	10000
weighted avg	0.42	0.42	0.42	10000

Visualization



Evaluation

Model	Imbalanced data		Balanced data	
	Accuracy	F1-Score	Accuracy	F1-Score
PCA	0.575	0.71/0.24	0.583	0.62/0.53
Kernel PCA	0.571	0.57/0.58	0.424	0.43/0.42

- This model appears more stable with imbalanced data and less sensitive to class imbalance.
- And can better capture nonlinear factors in the data.

Unsupervised Learning - t-SNE

Model

tsne = TSNE(n_components=2, random_state=42)

kmeans = KMeans(n_clusters=2, random_state=42)

Classification Report

Unbalanced Data

```
Clustering Accuracy with best alignment: 0.5323657611015599
              precision    recall  f1-score   support

           0       0.91      0.52      0.66    225694
           1       0.14      0.61      0.23     29653

    accuracy                           0.53    255347
   macro avg       0.53      0.56      0.45    255347
weighted avg       0.82      0.53      0.61    255347
```

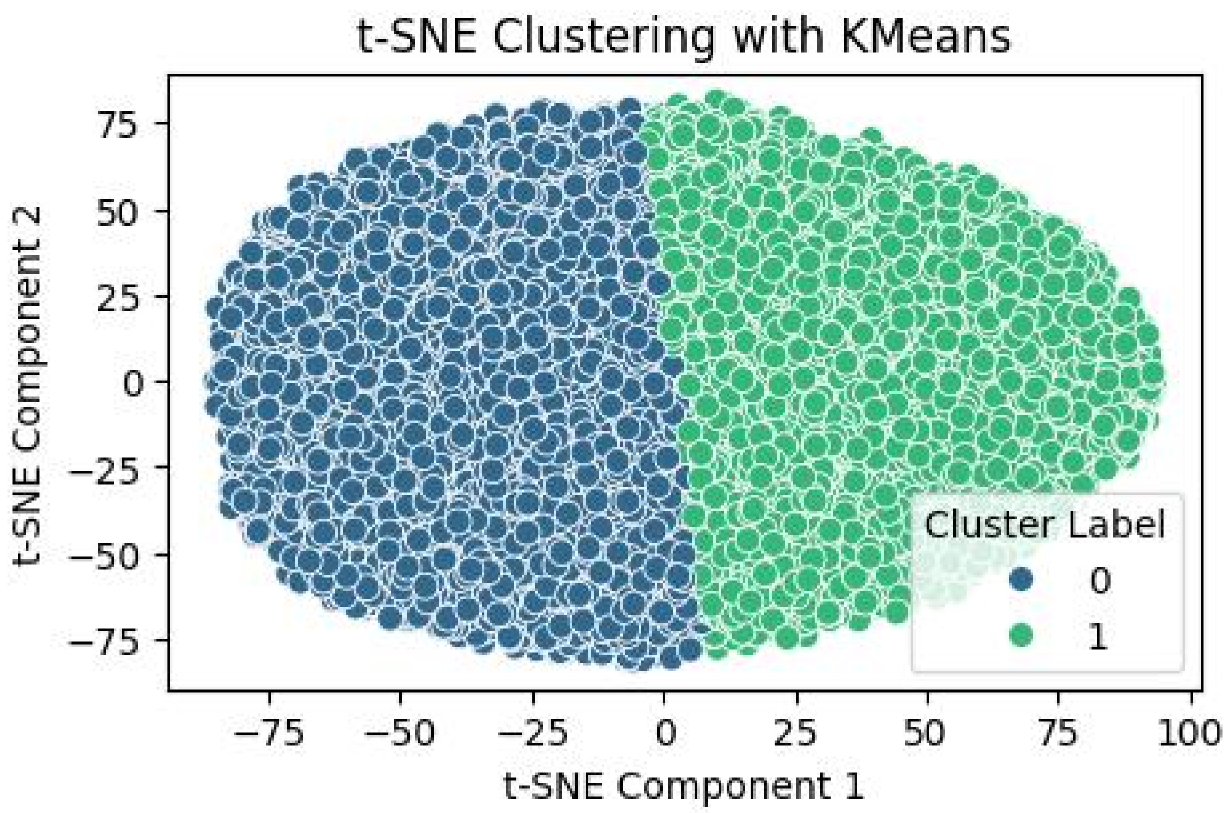
Balanced Data

```
Clustering Accuracy with best alignment: 0.58214
              precision    recall  f1-score   support

           0       0.58      0.61      0.59     25000
           1       0.59      0.56      0.57     25000

    accuracy                           0.58     50000
   macro avg       0.58      0.58      0.58     50000
weighted avg       0.58      0.58      0.58     50000
```

Visualization



Evaluation

Model	Imbalanced data		Balanced data	
	Accuracy	F1-Score	Accuracy	F1-Score
PCA	0.575	0.71/0.24	0.583	0.62/0.53
Kernel PCA	0.571	0.57/0.58	0.424	0.43/0.42
t-SNE	0.532	0.66/0.23	0.582	0.59/0.57

t-SNE, like PCA, is more influenced by data imbalance but shows improvements with balanced data.

CONCLUSION

Model comparison

Model	Accuracy	Recall	AUC	PR AUC
Random Forest	0.686	0.70/0.67	0.748	0.745
XGBoost	0.694	0.70/0.69	0.758	0.755
SVM	0.691	0.68/0.70	0.754	0.745

Model	Imbalanced data		Balanced data	
	Accuracy	F1-Score	Accuracy	F1-Score
PCA	0.575	0.71/0.24	**0.583**	**0.62/0.53**
Kernel PCA	0.571	0.57/0.58	0.424	0.43/0.42
t-SNE	0.532	0.66/0.23	**0.582**	**0.59/0.57**

- For imbalanced data, Kernel PCA is the best choice as it performs well without a significant drop in F1-scores between classes, showing robustness to imbalance.
- For balanced data, PCA and t-SNE both perform comparably well, with t-SNE slightly edging out on F1-scores.

In summary, supervised learning models perform significantly better than unsupervised models due to their ability to learn directly from labeled data, and among the supervised models, XGBoost achieves the best overall performance.

EVALUATION

Top 3 Feature Impact on Loan Default Prediction:

- **Age:** Younger borrowers may have less financial stability and credit history, leading to a higher risk of default.
- **Loan-to-Income Ratio:** Higher values indicate a larger debt burden relative to income, increasing default likelihood due to financial strain.
- **Interest Rate:** Higher interest rates increase monthly payments, making it harder for borrowers to keep up with repayments, which can raise the risk of default.

Economic Value of Loan Default Prediction:

- Enables lenders to assess borrower risk effectively, helps reduce financial losses, improve capital allocation.
- Identifying high-risk loans can support lenders in setting appropriate interest rates or collateral requirements, ultimately contributing to a healthier credit market .

THANK YOU!