PREDICTING LOAN DEFAULTERS FOR BAJAJ FINSERV LTD CAPSTONE PROJECT

Presentation by Tendai Jonhasi

July 30th 2024

Executive Summary

In the financial industry, accurately assessing the creditworthiness of borrowers is paramount for lenders to mitigate risks and maintain a healthy lending portfolio. Identifying potential defaulters—individuals at a higher risk of failing to repay their loans—helps lenders make informed decisions, minimize financial losses, and ensure financial stability. This project aims to develop a predictive model that can classify borrowers as defaulters or non-defaulters based on various financial and demographic factors. Utilizing a dataset provided by Bajaj Finserv, we embarked on a comprehensive data analysis and model-building process to achieve this objective. The dataset includes information on loan types, loan amounts, interest rates, loan terms, employment types, income levels, credit scores, gender, marital status, and education levels of borrowers.

Objectives

1. Data Preparation and Analysis
2. Model Development and Comparison
3. Feature Importance and Recommendations

Key Findings

Significant Predictors

Credit Score, employment type, interest rate, loan amount, and income level emerged as significant predictors of loan default. Borrowers with unstable employment, high-interest rates, and large loan amounts were more likely to default.

Model Performance

The Random Forest model, especially when applied with SMOTE, demonstrated improved recall for defaulters, highlighting its effectiveness in identifying high-risk borrowers.

Data Imbalance

Addressing the class imbalance in the dataset using SMOTE significantly enhanced the model's ability to predict defaulters, though it still showed challenges in achieving high precision and recall simultaneously.

Recommendations

1. **Incorporate Employment Type in Risk Assessment**: Lenders should integrate employment stability into their risk models, considering borrowers with full-time jobs as lower risk.
2. **Focus on Financial Indicators**: Close monitoring of interest rates and loan amounts is essential, as these are critical indicators of default risk.
3. **Enhanced Data Collection and Model Review**: Collect comprehensive borrower information and regularly update predictive models to capture current trends and behaviors.
4. **Financial Literacy Programs**: Implement educational programs for borrowers to improve financial management and reduce default rates.

Predicting Loan Defaulters for Bajaj Finserv Ltd

01.

Data Preparation and Exploration

The objective is to transform raw data into a suitable format for modeling and to understand the underlying patterns. From handling missing values to some basic exploratory data analysis to understand the data.

02.

Machine Learning Model Development and Evaluation

The objective is to develop initial predictive models and evaluate their performance without addressing class imbalance and compare with the outcomes when we improve model performance by addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

03.

Feature Importance and Recommendations

A concise overview of the best-performing model and the resulting actionable recommendations based on the focus areas and suggestions on how to further enhance predictive power of model.

Data Preparation and Exploration

Handling Missing Values

Imputed missing values using the mean for numerical features and the mode for categorical features.

```
loan_type            0
loan_amount          0
interest_rate        0
loan_term            0
employment_type      0
income_level         0
credit_score         0
gender               0
marital_status       0
education_level      0
application_date     0
approval_date        0
disbursement_date    0
due_date             0
default_status       0
```

Encoding Categorical Variables

Used Label Encoding for categorical features, ensuring that no encoded value is 0 and maintaining a mapping for clarity.

```
Mapping for loan_type: {1: 'Car Loan', 2: 'Education Loan', 3: 'Home Loan', 4: 'Personal Loan'}
Mapping for employment_type: {1: 'Full-time', 2: 'Part-time', 3: 'Self-employed'}
Mapping for gender: {1: 'Female', 2: 'Male'}
Mapping for marital_status: {1: 'Divorced', 2: 'Married', 3: 'Single'}
Mapping for education_level: {1: 'Bachelor', 2: 'High School', 3: 'Master', 4: 'PhD'}
Mapping for income_level: {'Low': 1, 'Medium': 2, 'High': 3}
```

Scaling Numerical Features:

Applied Standard Scaler to standardize numerical features so that it has a mean of 0 and a standard deviation of 1.

Exploratory Data Analysis

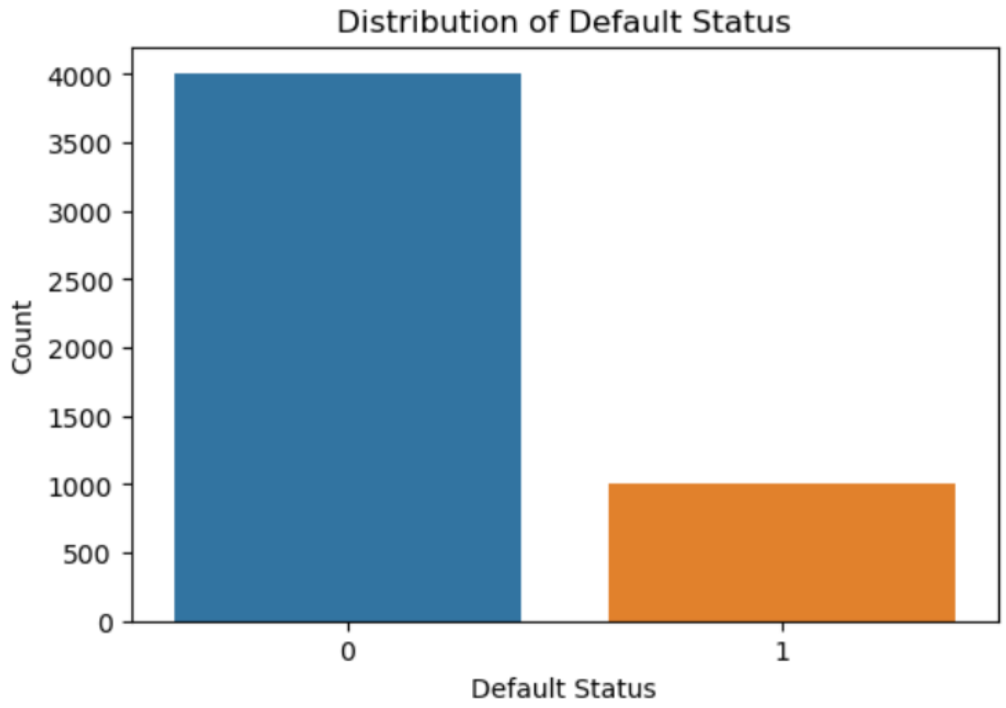
Checked for data imbalance in the target variable. Created visualizations like scatter plots and pie charts to analyze the distribution and relationships in the data.

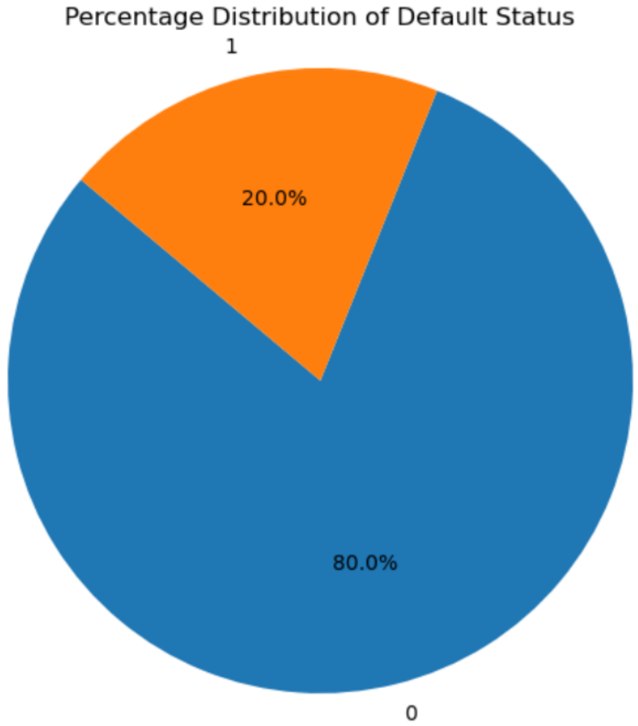
- Check for Imbalance in target variable
- Check for feature distributions
- Check Linear Relationship between target variable and features

Process Flow of the Explanatory Analysis

Imbalance in Target Variable

The variable/attribute "Default Status" represents whether or not the customer has paid or defaulted on loan. This variable is the target or dependent variable that will be predicted in the model. The following graphs provides a visual representation of the distribution of data for this variable.



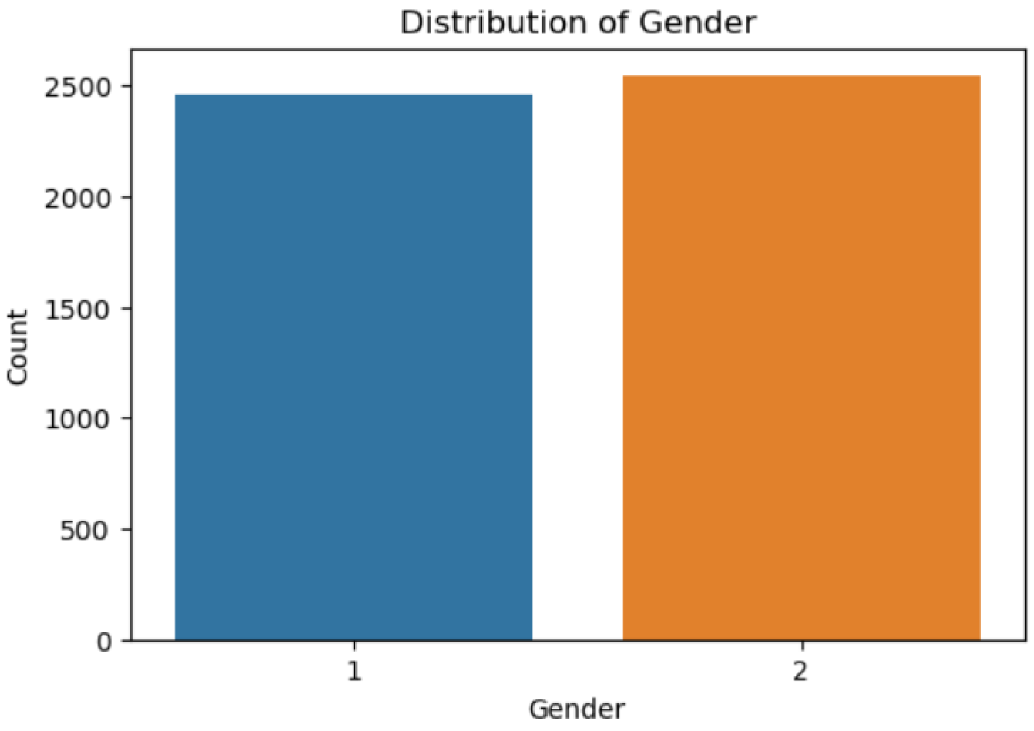


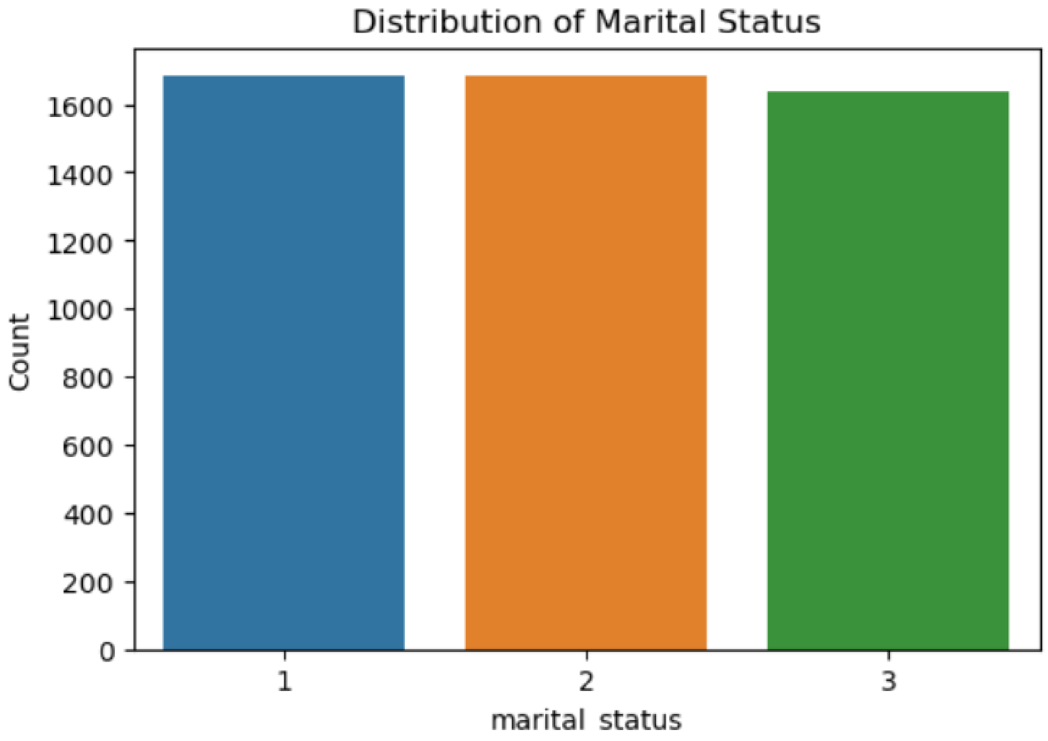
Based on the chart, it is evident that out of the total 5,000 records, 4,001 (80%) are classified as 0 and 999 (20%) are classified as 1. This indicates that the majority of customers has paid the loan and did not default. This distribution is known as imbalanced, which could cause modeling issues and low performance levels for the imbalanced value. In the upcoming phases, this issue will be tackled.

Process Flow of the Explanatory Analysis

Feature Distribution

The variables gender and marital status are control variables, that make sure our dataset is representative of the demography of the customers for the bank.



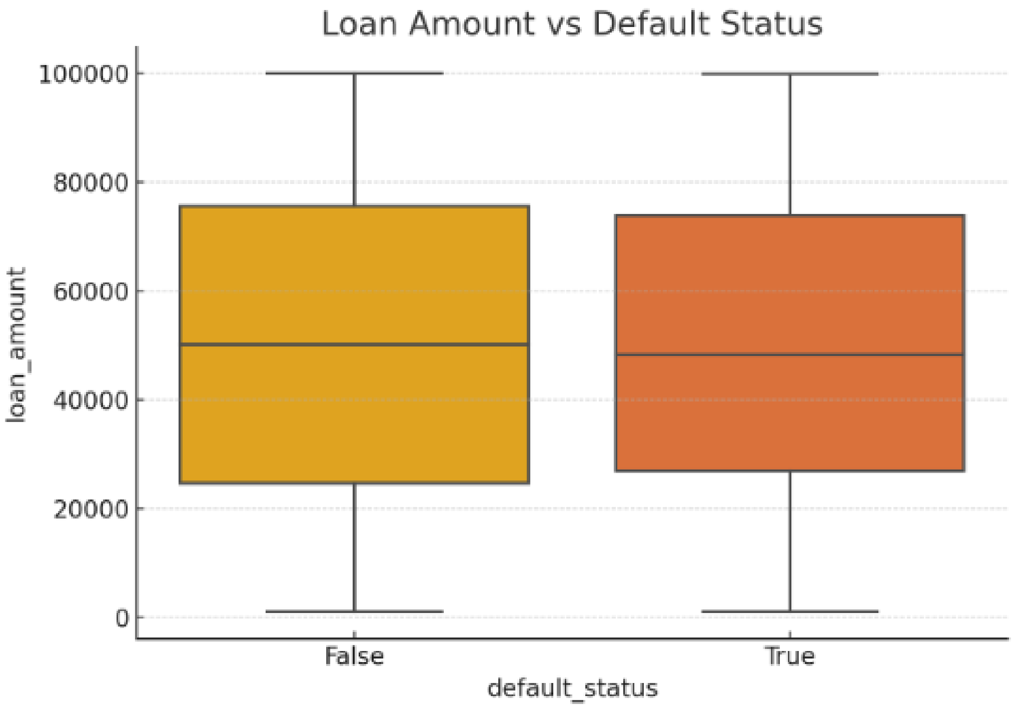


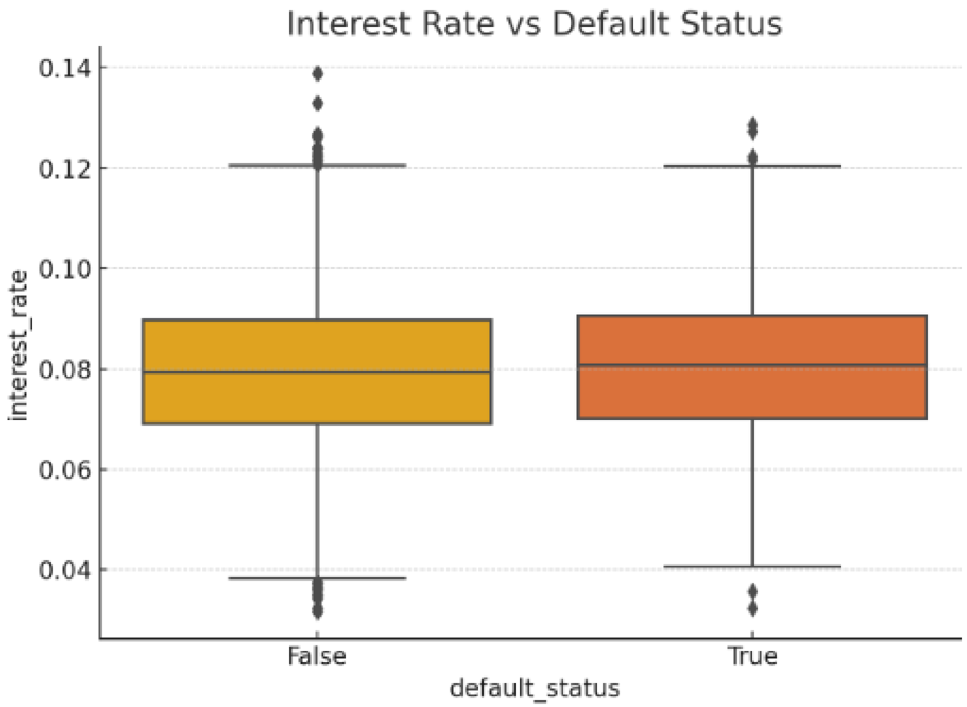
Here we see that, the distribution of the gender and martial status is evenly distributed across the dataset. Eliminating any skewness in the data towards a particular gender, or martial status. Normalising the dataset has reduced the likelihood of confounded bias.

Process Flow of the Explanatory Analysis

Relationship between Features and the Target Variable

The boxplots reveal the following insights regarding the relationship between numerical features and the default status:
Loan Amount vs Default Status: There is a slight difference in loan amounts between defaulters and non-defaulters, with defaulters tending to have higher loan amounts.
Interest Rate vs Default Status: Defaulters tend to have higher interest rates compared to non-defaulters.

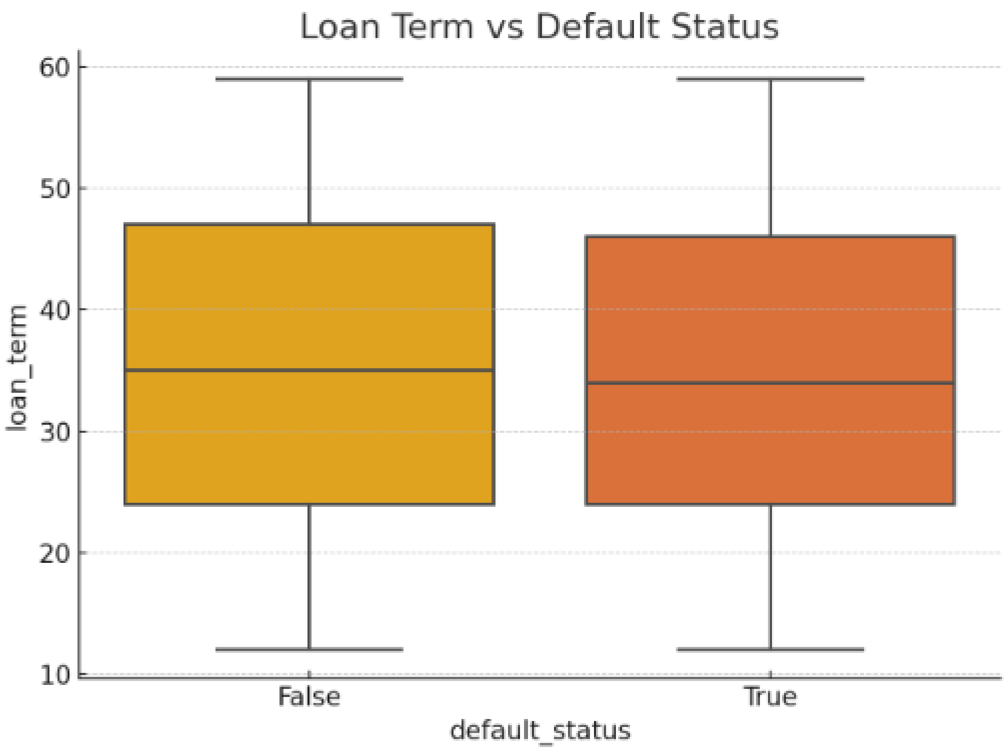


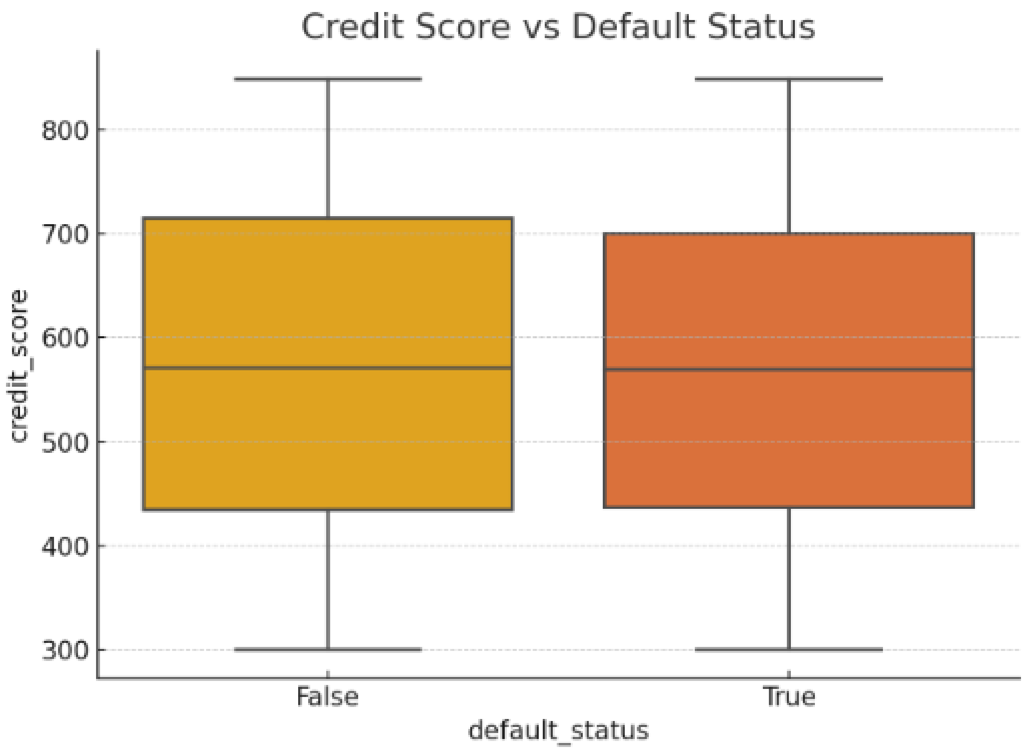


Process Flow of the Explanatory Analysis

Relationship between Features and the Target Variable

The boxplots reveal the following insights regarding the relationship between numerical features and the default status:
Loan Term vs Default Status: There is no significant difference in loan term between defaulters and non-defaulters**.**
Credit Score vs Default Status: Defaulters tend to have lower credit scores compared to non-defaulters.





Predicting Loan Defaulters for Bajaj Finserv Ltd

01.

Data Preparation and Exploration

The objective is to transform raw data into a suitable format for modeling and to understand the underlying patterns. From handling missing values to some basic exploratory data analysis to understand the data.

02.

Machine Learning Model Development and Evaluation

The objective is to develop initial predictive models and evaluate their performance without addressing class imbalance and compare with the outcomes when we improve model performance by addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

03.

Feature Importance and Recommendations

A concise overview of the best-performing model and the resulting actionable recommendations based on the focus areas and suggestions on how to further enhance predictive power of model.

Model Training and Evaluation

To build a model that predicts loan defaults, it needs to be trained using algorithms. In this study, various algorithms are used and analyzed to determine which one is the best performing one. The algorithms used in this research are:

1

Logistic Regression

- This algorithm is a statistical method that predicts the probability of each class, and the algorithm displays the probability of that prediction (e.g., an 80% chance that the class is 1).
- "The Logistic Regression models were recognized as the most appropriate models in deciding to grant credit to individuals and regarded as the industry standard in credit scoring model development" (Victor & Raheem, 2021).

2

Decision Tree

- Decision Tree works by separating into different nodes, where each node represents a variable. The variable that affects the prediction the most is placed at the beginning and then continuing with the other variables below which are of less important (hierarchy).
- The **Random Forest** algorithm is one of the most flexible, powerful and widely-used algorithms built as an ensemble of Decision Trees.

3

Ensemble Algorithms

- "Ensemble techniques are the methods that use multiple learning algorithms or models to produce one optimal predictive model. The model produced has better performance than the base learners taken alone" (Khandelwal, 2021). For this project, we only focus on boosting.

3

Adaptive Boost

- AdaBoost is one of the first boosting algorithms to have been introduced.
- It makes use of weighted errors to build a strong classifier from a series of weak classifiers.

Gradient Boosting

- Gradient boosting works by building simpler (weak) prediction models sequentially where each model tries to predict the error left over by the previous model.

4

Support Vector Machine (SVM)

- "It works on the principle of fitting a boundary to a region of points that are all alike (that is, belong to one class)" (Kotu & Deshpande, 2019).
- So, the model is trained on some historical data and based on similarities it is divided into regions, thus when the new records comes it goes to the region where the features are similar.

5

Neural Network

- This algorithm is more advanced than the classical algorithms.
- "A neural network consists of a set of neurons that are connected together. A neuron takes a set of numeric values as input and maps them to a single output value. At its core, a neuron is simply a multi-input linear-regression function." (Kelleher & Tierney, 2018).

To determine the effectiveness and performance of the predictive model, it must be evaluated using the following metrics; ACCURACY, PRECISION, RECALL and F1-SCORE

Synthetic Minority Over-sampling Technique

To address class imbalance in the training data, we used SMOTE (Synthetic Minority Over-sampling Technique).

The output Counter {0: 3201, 1: 3201} indicates the number of instances for each class in the default_status variable after applying the Synthetic Minority Over-sampling Technique (SMOTE).

Class Balance

0: 3201 means there are now 3201 instances of the Non-Defaulter class (default status 0). *1: 3201 means there are now 3201 instances of the Defaulter class (default status 1).

Before applying SMOTE, the dataset was imbalanced, with one class having significantly fewer instances than the other. SMOTE generates synthetic samples for the minority class to balance the dataset. After applying SMOTE, both classes have the same number of instances, achieving a balanced dataset.

```python
from imblearn.over_sampling import SMOTE

# Define features and target variable
X = loan_data.drop(columns=['application_date', 'approval_date', 'disbursement_date', 'due_date', 'default_status'])
y = loan_data['default_status']

# Encode categorical variables
X = pd.get_dummies(X, drop_first=True)

# Split the dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Apply SMOTE to balance the training data
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)

# Add a constant to the independent variables for the intercept term
X_resampled_const = sm.add_constant(X_resampled)
```

[No Title]

SMOTE first selects a minority class instance, a, at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors, b, at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.

Logistic Regression

Results without SMOTE

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7974	0.264	-6.800	0.000	-2.315	-1.279
loan_type	-0.0416	0.038	-1.106	0.269	-0.115	0.032
loan_amount	-0.0239	0.043	-0.561	0.575	-0.107	0.060
interest_rate	0.1155	0.042	2.740	0.006	0.033	0.198
loan_term	-0.0271	0.043	-0.631	0.528	-0.111	0.057
employment_type	0.1070	0.052	2.051	0.040	0.005	0.209
income_level	-0.0067	0.052	-0.129	0.898	-0.109	0.095
credit_score	-0.0349	0.043	-0.820	0.412	-0.118	0.049
gender	0.1577	0.085	1.849	0.064	-0.009	0.325
marital_status	0.0784	0.052	1.505	0.132	-0.024	0.181
education_level	-0.0372	0.038	-0.983	0.326	-0.111	0.037

Classification Report:

	precision	recall	f1-score	support
0	0.80	1.00	0.89	1197
1	0.00	0.00	0.00	303
accuracy			0.80	1500
macro avg	0.40	0.50	0.44	1500
weighted avg	0.64	0.80	0.71	1500

Accuracy Score:
0.798

Interpretation

The Pseudo R-squared value is very low (0.005989), suggesting that the model explains only a small portion of the variability in the data.

Significant Features:
Interest Rate and Employment Type are the only statistically significant predictors of default status. Gender is statistically at a higher p-value > 0.1. Income Level, Loan Amount, Loan Type, Loan Term, Credit Score, Gender, Marital Status, and Education Level are not statistically significant predictors (p-values > 0.05).

Results with SMOTE

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.0806	0.168	12.410	0.000	1.752	2.409
loan_type	-0.1813	0.024	-7.517	0.000	-0.229	-0.134
loan_amount	-0.0110	0.028	-0.391	0.696	-0.066	0.044
interest_rate	0.1249	0.028	4.414	0.000	0.069	0.180
loan_term	0.0065	0.028	0.228	0.819	-0.049	0.062
employment_type	-0.1123	0.034	-3.285	0.001	-0.179	-0.045
income_level	-0.2001	0.034	-5.881	0.000	-0.267	-0.133
credit_score	-0.0265	0.028	-0.945	0.345	-0.081	0.028
gender	-0.2249	0.055	-4.112	0.000	-0.332	-0.118
marital_status	-0.1276	0.034	-3.726	0.000	-0.195	-0.060
education_level	-0.1938	0.024	-7.916	0.000	-0.242	-0.146

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.58	0.67	1197
1	0.21	0.43	0.28	303
accuracy			0.55	1500
macro avg	0.50	0.50	0.47	1500
weighted avg	0.68	0.55	0.59	1500

Accuracy Score:
0.5486666666666666

Interpretation

The Pseudo R-squared value is still very low (0.02698), suggesting that the model explains only about 2.7% of the variability in the default status.

Significant Features:
Features like loan_type, interest_rate, employment_type, income_level, gender, marital_status, and education_level are highly significant predictors of default status. While SMOTE helps address the class imbalance, *the model still has difficulty predicting defaulters accurately, as seen in the high number of false positives and false negatives.*

Random Forest Classifier

Without SMOTE

Classification Report:

	precision	recall	f1-score	support
0	0.80	1.00	0.89	1197
1	0.33	0.00	0.01	303
accuracy			0.80	1500
macro avg	0.57	0.50	0.45	1500
weighted avg	0.70	0.80	0.71	1500

Accuracy Score:
0.7973333333333333

Feature Importances:

	importance
interest_rate	0.203087
loan_amount	0.194650
credit_score	0.192945
loan_term	0.154600
education_level	0.054554
loan_type	0.050846
marital_status	0.044175
income_level	0.043346
employment_type	0.038379
gender	0.023417

With SMOTE

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.82	0.80	1197
1	0.18	0.16	0.17	303
accuracy			0.68	1500
macro avg	0.49	0.49	0.49	1500
weighted avg	0.67	0.68	0.68	1500

Accuracy Score:
0.684

Feature Importances:

	importance
credit_score	0.190819
loan_amount	0.187924
interest_rate	0.179316
loan_term	0.177911
loan_type	0.055447
education_level	0.052782
employment_type	0.044967
marital_status	0.044195
income_level	0.041435
gender	0.025204

Random Forest Classifier is the **best-performing model with SMOTE** as identified based on the highest accuracy of 68.4%, 79% precision, and 80% F1 scores.

Model Performance

Without SMOTE, the accuracy is higher (79.7%) compared to with SMOTE (68.4%). This is because the model is biased towards the majority class (non-defaulters) in the imbalanced dataset. The recall for class 1 (defaulters) is 0.00 without SMOTE, meaning the model fails to identify any defaulters. With SMOTE, the recall for class 1 improves to 0.16, indicating the model is better at identifying defaulters when the dataset is balanced. Without SMOTE, the model heavily favors the majority class (non-defaulters), as seen in the very high recall for class 0 and almost zero for class 1. With SMOTE, the model shows a more balanced performance, with improved recall for class 1, but at the cost of a decrease in overall accuracy.

The important features remain consistent in both cases, with interest_rate, loan_amount, credit_score and loan_term being the top four. However, their relative importances shift slightly with SMOTE applied.

Gradient Boosting and AdaBoost with SMOTE

Gradient Boosting

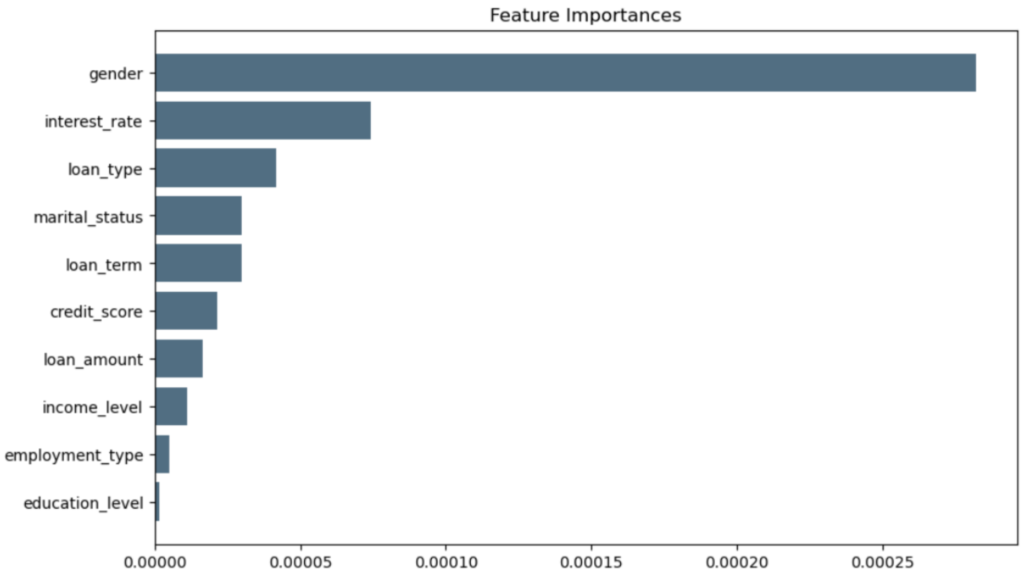
```
Gradient Boosting with SMOTE
Confusion Matrix:
[[768 429]
 [188 115]]

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.64      0.71      1197
           1       0.21      0.38      0.27       303

    accuracy                           0.59      1500
   macro avg       0.51      0.51      0.49      1500
weighted avg       0.68      0.59      0.62      1500

Accuracy Score:
0.5886666666666667
```

AdaBoost

```
AdaBoost with SMOTE
Confusion Matrix:
[[694 503]
 [169 134]]

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.58      0.67      1197
           1       0.21      0.44      0.29       303

    accuracy                           0.55      1500
   macro avg       0.51      0.51      0.48      1500
weighted avg       0.68      0.55      0.60      1500

Accuracy Score:
0.552
```

Support Vector Machine

Results without SMOTE

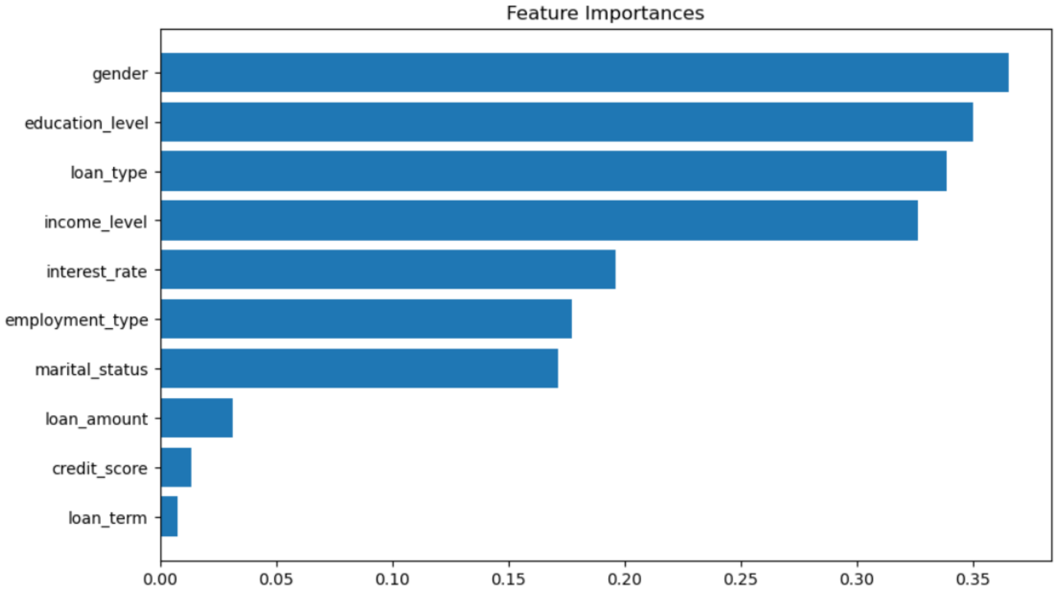



```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89      1197
           1       1.00      0.00      0.00       303

    accuracy                           0.80      1500
   macro avg       0.90      0.50      0.44      1500
weighted avg       0.84      0.80      0.71      1500
```

```
Accuracy Score:
0.798
```

The model without SMOTE has a higher accuracy score, indicating it performs better overall. However, accuracy alone can be misleading, especially if the data is imbalanced. The top three important features are interest_rate, loan_type and loan_term.

Results with SMOTE




```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.56      0.66      1197
           1       0.20      0.44      0.28       303

    accuracy                           0.53      1500
   macro avg       0.50      0.50      0.47      1500
weighted avg       0.68      0.53      0.58      1500
```

```
Accuracy Score:
0.5333333333333333
```

The model with SMOTE has a better chance predicting the non-defaulters more accurately than without SMOTE despite a lower accuracy score. The top three important features are now education_level, loan_type and income_level.

Neural Network

Without SMOTE

Classification Report without SMOTE:

	precision	recall	f1-score	support
0	0.80	0.99	0.88	1197
1	0.18	0.01	0.02	303
accuracy			0.79	1500
macro avg	0.49	0.50	0.45	1500
weighted avg	0.67	0.79	0.71	1500

Accuracy Score without SMOTE:
0.7906666666666666

```
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7988 - loss: 0.4835
Epoch 12/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8031 - loss: 0.4680
Epoch 13/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.8052 - loss: 0.4694
Epoch 14/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7966 - loss: 0.4779
Epoch 15/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7922 - loss: 0.4763
Epoch 16/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7995 - loss: 0.4653
Epoch 17/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8118 - loss: 0.4435
Epoch 18/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8030 - loss: 0.4643
Epoch 19/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.7905 - loss: 0.4716
Epoch 20/20
110/110 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - accuracy: 0.7906 - loss: 0.4783
47/47 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
```

With SMOTE

```
176/176 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.6776 - loss: 0.5911
Epoch 12/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - accuracy: 0.7123 - loss: 0.5744
Epoch 13/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - accuracy: 0.7121 - loss: 0.5706
Epoch 14/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 930us/step - accuracy: 0.7217 - loss: 0.5579
Epoch 15/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 943us/step - accuracy: 0.7293 - loss: 0.5519
Epoch 16/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - accuracy: 0.7335 - loss: 0.5428
Epoch 17/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 964us/step - accuracy: 0.7331 - loss: 0.5425
Epoch 18/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 936us/step - accuracy: 0.7427 - loss: 0.5363
Epoch 19/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.7487 - loss: 0.5297
Epoch 20/20
176/176 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.7412 - loss: 0.5292
47/47 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
```

Classification Report with SMOTE:

	precision	recall	f1-score	support
0	0.79	0.63	0.70	1197
1	0.19	0.35	0.25	303
accuracy			0.57	1500
macro avg	0.49	0.49	0.48	1500
weighted avg	0.67	0.57	0.61	1500

Accuracy Score with SMOTE:
0.5746666666666667

Predicting Loan Defaulters for Bajaj Finserv Ltd

01.

Data Preparation and Exploration

The objective is to transform raw data into a suitable format for modeling and to understand the underlying patterns. From handling missing values to some basic exploratory data analysis to understand the data.

02.

Machine Learning Model Development and Evaluation

The objective is to develop initial predictive models and evaluate their performance without addressing class imbalance and compare with the outcomes when we improve model performance by addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

03.

Feature Importance and Recommendations

A concise overview of the best-performing model and the resulting actionable recommendations based on the focus areas and suggestions on how to further enhance predictive power of model.

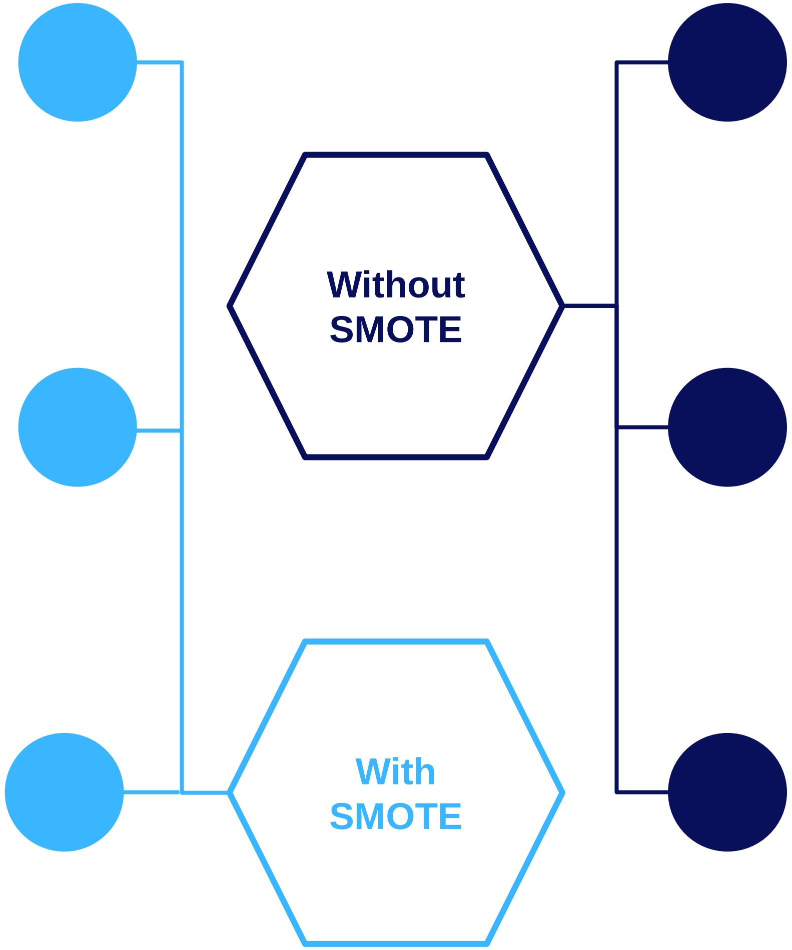
Literature on Predicting Loan Defaulters

Random Forest Classifier is our **best-performing model with SMOTE** while the Logistic Regression model dominates over the others without SMOTE.

Zhu, et al., (2019): Different ML models were applied, such as Decision Tree, SVM, Logistic Regression, and Random Forest , the SMOTE method was used together with the ***Random Forest*** to address the issue of data imbalance, achieving the highest accuracy of 98%

Platur Gashi, (2023): This paper presents several models for predicting loan defaults using a variety of ML algorithms with SMOTE. The results demonstrate that ensemble algorithms outperform individual ones in predicting loan defaults. The top-performing algorithms were ***Boosted Decision Trees (Boosting)*** and ***Random Forest (Bagging).***

Abedin, et al., (2023): In small business credit risk assessment, the default and nondefault classes are highly imbalanced. Using Weighted-SMOTE ensemble to overcome the imbalance, this study proposes that the ***Random Forest Classifier*** provides a good trade-off between the performance on default class and that of nondefault class



Tariq, et al., (2019): Using the SEMMA (Sample, Explore, Modify, Model and Assess) methodology, the best performance was chosen ***Logistic Regression***, since it was the best in predicting "True Negatives" (how accurate is the model for predicting loans that default)

Owusu, et al., (2023): This study addresses loan default in online peer-to-peer lending activities. Due to the imbalanced nature of the dataset, the adaptive synthetic (ADASYN) oversampling algorithm is used to balance the data. ***Deep neural network (DNN)*** is used for prediction with accuracy of 94.1% is realized

Victor & Raheem, (2021): The "Genetic Algorithm" used here is based on Darwin's theory of evolution, which after many calculations selects the variables that have the most impact on the target variable. Logistic Regression, Random Forest, and SVM algorithms were used for predictive modeling, the most successful algorithm was ***Random Forest***, achieving 82% F1 score

Important Features in predicting Default Status

We determine the important features in predicting default status after building and evaluating the best performing model with SMOTE (Random Forest) and the one without SMOTE (Logistic Regression). While, controlling for other factors, an increase in the interest rate or the loan amount, or a decrease in the credit score or; self-employed and part-time employees are associated with an increased probability of default.

The insights derived from this analysis highlight the importance of incorporating various financial and demographic factors into predictive models, as well as the benefits of using advanced machine learning techniques and data balancing methods like SMOTE. By integrating these recommendations, lenders can enhance their ability to accurately assess borrower risk, leading to better lending decisions and reduced default rates.

Credit Score

Credit score is a crucial predictor for defaulting on a loan due to its strong relationship with a borrower's creditworthiness and financial behavior. Credit scores are based on a borrower's credit history, which includes data on past loans, repayment behavior, defaults, bankruptcies, and other credit-related activities. Lenders use credit scores as a primary measure of creditworthiness. A lower score can indicate higher risk of defaulting on their loan.

Loan Amount

Larger loan amounts often carry higher risk. Borrowers who take out larger loans may have higher monthly payments, which can strain their finances. If the loan amount is disproportionately large compared to the borrower's income, it increases the likelihood of default. Borrowers with higher loan amounts relative to their income may struggle to meet repayment obligations.

Interest Rate

Lenders often assign higher interest rates to borrowers they deem to be higher risk, or have a higher likelihood of default based on their credit history, income stability, etc. High interest rates can also be reflection of economic downturns, inflation, or tighter monetary policy which can increase the cost of borrowing and living expenses, making it harder for borrowers to meet their loan obligations. Borrowers with higher monthly obligations may struggle more to keep up with payments

Employment Type

Employment type can be a significant predictor for defaulting on a loan due to several factors related to job stability, income consistency, and overall financial security. Individuals in less stable jobs may have fewer financial cushions or resources and lack job security, making it difficult for individuals to guarantee steady income over time, leading to higher default rates.

THANK YOU.

Tendai Jonhasi
tendaimjonhasi@gmail.com
LinkedIn profile
GitHub