# Loan Default Prediction

COSC 522 Final Project (Group 3)

Andrew Penny | Cody Lee Viscardis | Peter Mansfield | Soe Thet Ko

THE UNIVERSITY OF TENNESSEE KNOXVILLE

# Overview

Problem Statement

Inaccurate risk assessment on loan applications

- ➢ cause financial losses for lending institutions
- ➢ hinder financial inclusion for qualified borrowers

Traditional loan assessment methods rely heavily on credit scores and financial history.

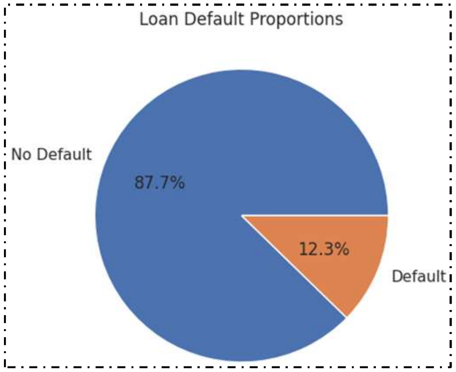
Proposed Solution

To develop a machine learning model that leverages customer demographics and predicts potential loan defaults with accuracy over 90%

Potential Benefits

- ➢ loan providers can make informed decisions and minimize risks
- ➢ qualified applicants get access to loan

# Methodology




Dataset

"Loan Prediction Based on Customer Behavior" from Kaggle

- historical data of over 250,000 borrowers
- 11 features
- target: Risk_Flag indicates loan defaulted or not

(https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior)

| | Id | Income | Age | Experience | Married/ Single | House_Ownership | Car_Ownership | Profession | CITY | STATE | CURRENT_JOB_YRS | CURRENT_HOUSE_YRS | Risk_Flag |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1303834 | 23 | 3 | single | rented | no | Mechanical_engineer | Rewa | Madhya_Pradesh | 3 | 13 | 0 |
| 1 | 2 | 7574516 | 40 | 10 | single | rented | no | Software_Developer | Parbhani | Maharashtra | 9 | 13 | 0 |
| 2 | 3 | 3991815 | 66 | 4 | married | rented | no | Technical_writer | Alappuzha | Kerala | 4 | 10 | 0 |
| 3 | 4 | 6256451 | 41 | 2 | single | rented | yes | Software_Developer | Bhubaneswar | Odisha | 2 | 12 | 1 |
| 4 | 5 | 5768871 | 47 | 11 | single | rented | no | Civil_servant | Tiruchirappalli[10] | Tamil_Nadu | 3 | 14 | 1 |

**Dataset Snippet**

# Methodology - cont.

Algorithms

The following four architectures are chosen to train a binary classifier with supervised learning on the dataset.

- Gradient Boosting
  - can handle complex data
  - high accuracy by combining predictions of weaks learners sequentially

- Neural Network
  - can detect complex nonlinear relationships
  - quite resistant to label-noise

- Logistic Regression
  - easy to implement and efficient to train
  - high interpretability

- Random Forest
  - features importance interpretability
  - high accuracy by averaging multiple decision trees constructed in parallel

Evaluation

The following metrics and plots are used to visualize and evaluate the model's performance on the test set.

➢ Accuracy, Precision, Recall, F1 score, Confusion Matrix, ROC Curve, ROC AUC

# Results - Gradient Boosting Classifier

Initial LogReg for analysis - model predicting all negatives (majority class)

Accuracy: 0.8759325396825397
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
Confusion Matrix:
[[44147     0]
 [ 6253     0]]

Undersampled dataset and retrained logreg model- still predicting all negatives

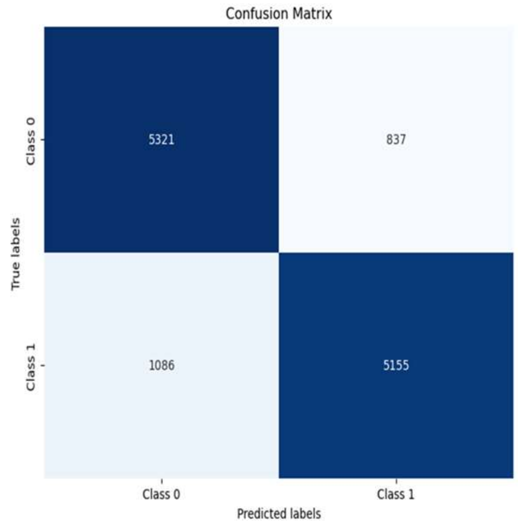
Accuracy: 0.49665295588353897
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
Confusion Matrix:
[[6158    0]
 [6241    0]]

Undersampled to balance and used standard scaler - trained GRADIENT BOOSTING CLASSIFIER

Accuracy: 0.6291636422292121
Precision: 0.6316717422663889
Recall: 0.6314693158147733
F1 Score: 0.6315705128205128
Confusion Matrix:
[[3860 2298]
 [2300 3941]]

Implemented GridSearch on GBC model- best model results:

Best Hyperparameters: {'learning_rate': 0.2, 'max_depth': 7, 'min_samples_split': 4, 'n_estimators': 150}
Evaluation Metrics with Best Model:
Accuracy: 0.8449068473263973
Precision: 0.8603137516688919
Recall: 0.8259894247716713
F1 Score: 0.8428022561922669
Confusion Matrix:
[[5321  837]
 [1086 5155]]



# Results - Neural Network

**Model Fitting:**

Early Stopped At 25 Epochs

```
loss: 0.3652
accuracy: 0.8774
val_loss: 0.3656
val_accuracy: 0.8750
```

**Model Evaluation:**

```
loss: 0.3647
accuracy: 0.8754
```

**Optimizer**: Adam

**Learning rate**: 0.005

**Loss Function**: Binary_crossentropy

**Metrics**: Accuracy

**Train/Test/Validation Split**: 80/10/10

**Normalization**: Z-Score

**Categorical Encoding**: Label

**Network Architecture:**

```
 Layer (type)                Output Shape              Param #
=================================================================
 dense_5 (Dense)             (None, 16)                192

 dropout_4 (Dropout)         (None, 16)                0

 dense_6 (Dense)             (None, 16)                272

 dropout_5 (Dropout)         (None, 16)                0

 dense_7 (Dense)             (None, 16)                272

 dropout_6 (Dropout)         (None, 16)                0

 dense_8 (Dense)             (None, 16)                272

 dropout_7 (Dropout)         (None, 16)                0

 dense_9 (Dense)             (None, 1)                 17

=================================================================
Total params: 1025 (4.00 KB)
Trainable params: 1025 (4.00 KB)
Non-trainable params: 0 (0.00 Byte)
```

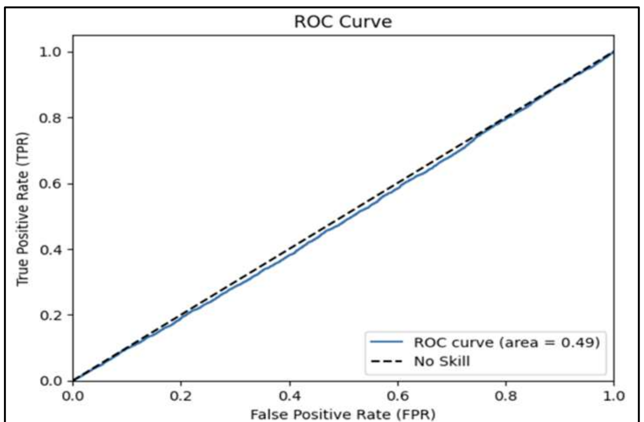
# Results - Logistic Regression

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 1.00 | 0.93 | 24407 |
| 1 | 0.00 | 0.00 | 0.00 | 3593 |
| accuracy | | | 0.87 | 28000 |
| macro avg | 0.44 | 0.50 | 0.47 | 28000 |
| weighted avg | 0.76 | 0.87 | 0.81 | 28000 |

```
[[24407     0]
 [ 3593     0]]
```

## Nonviable Solution

- Multiple feature removal
- 4-8x Oversampling minority class
- 90-15% Undersampling majority class




THE UNIVERSITY OF TENNESSEE KNOXVILLE

# Results - Random Forest Classifier

Model 1 (use all features,
no undersampling/oversampling)

```
Classification Report
              precision    recall  f1-score

           0       0.94      0.95      0.94
           1       0.60      0.54      0.57

    accuracy                           0.90
   macro avg       0.77      0.74      0.75
weighted avg       0.89      0.90      0.90

Confusion Matrix
[[41937  2210]
 [ 2905  3348]]

ROC AUC score: 0.7426814851162876
```

Model 2 (use all features,
undersample the majority class)

```
Classification Report
              precision    recall  f1-score

           0       0.83      0.89      0.86
           1       0.88      0.81      0.84

    accuracy                           0.85
   macro avg       0.85      0.85      0.85
weighted avg       0.85      0.85      0.85

Confusion Matrix
[[5548  705]
 [1174 5079]]

ROC AUC score: 0.8497521189828883
```

Model 3 (use all features,
oversample the minority class)

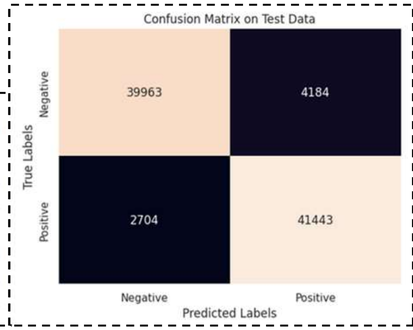
```
Classification Report
              precision    recall  f1-score

           0       0.93      0.90      0.92
           1       0.90      0.93      0.92

    accuracy                           0.92
   macro avg       0.92      0.92      0.92
weighted avg       0.92      0.92      0.92

Confusion Matrix
[[39731  4416]
 [ 2898 41249]]

ROC AUC score: 0.9171631141413912
```

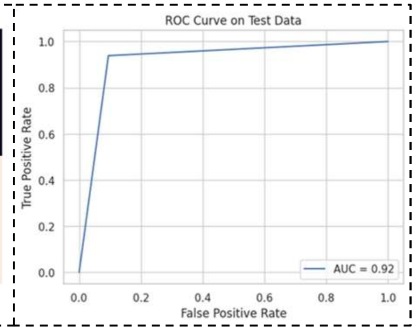
Model 4 (drop two features due to correlation,
oversampling still applied )

```
Classification Report
              precision    recall  f1-score

           0       0.93      0.91      0.92
           1       0.91      0.93      0.92

    accuracy                           0.92
   macro avg       0.92      0.92      0.92
weighted avg       0.92      0.92      0.92

Confusion Matrix
[[39957  4190]
 [ 3013 41134]]

ROC AUC score: 0.9184202777085645
```

Final Model (drop three less important features based on
feature importance analysis, oversampling still applied )

```
Classification Report
              precision    recall  f1-score

           0       0.94      0.91      0.92
           1       0.91      0.94      0.92

    accuracy                           0.92
   macro avg       0.92      0.92      0.92
weighted avg       0.92      0.92      0.92
```




Numerical Features: Standard Scalar Normalization
Categorical Features: Binary and One-Hot Encoding
High Cardinal Categorical Features: Target Encoding

Train/Test Split: 80/20
Undersampling: RandomUnderSampler
Oversampling: SMOTE

‘Experience’ & ‘Current Job Yrs’ correlation: 0.64
‘City’ & ‘State’ correlation: 0.34
‘Car Ownership’, ‘Marital Status’, ‘House Ownership’ < 0.025 Importance

THE UNIVERSITY OF TENNESSEE KNOXVILLE

# Conclusion

Random Forest:

- Only Model to achieve ≥90% benchmark
- Overall, effective method for actuarial analysis

Lessons Learned:

- Rebalancing is important for skewed dataset
- Can get the same or even higher performance by dropping correlated features and less important features

Solution Improvements:

- Larger dataset with additional features
- Analysis of model on various other loan sub-categories (vehicles, revolving lines, etc.) and accompanying credit risk to derive model generalizability