



Retail Credit

Estimating probability of default with XGBoost

Motivation

- Liquidity
- IFRS 9 & Basel II compliance
- Modern



Data Pre-Processing I

1- Create mapping column for target variable (Multi-class classification)

Class 0 (Non-default)	Class 1 (At risk)	Class 2 (Default)
"Current" "Fully Paid"	"Late (31-120 days)" "Late (16-30 days)" "In Grace Period"	"Charged Off" "Default"

2- Drop columns with excessive missing values

Threshold = 30%	Threshold = 10%	Threshold = 5%
49 columns dropped	50 columns dropped	52 columns dropped

3- Drop insignificant columns (e.g. loan_url, loan_id)

Data Pre-Processing II

4- Apply transformations

7.1%

→

0.071

5- Convert data to correct types (e.g. numbers, dates)

6- Replace missing values

7- Convert categorical data to numerical

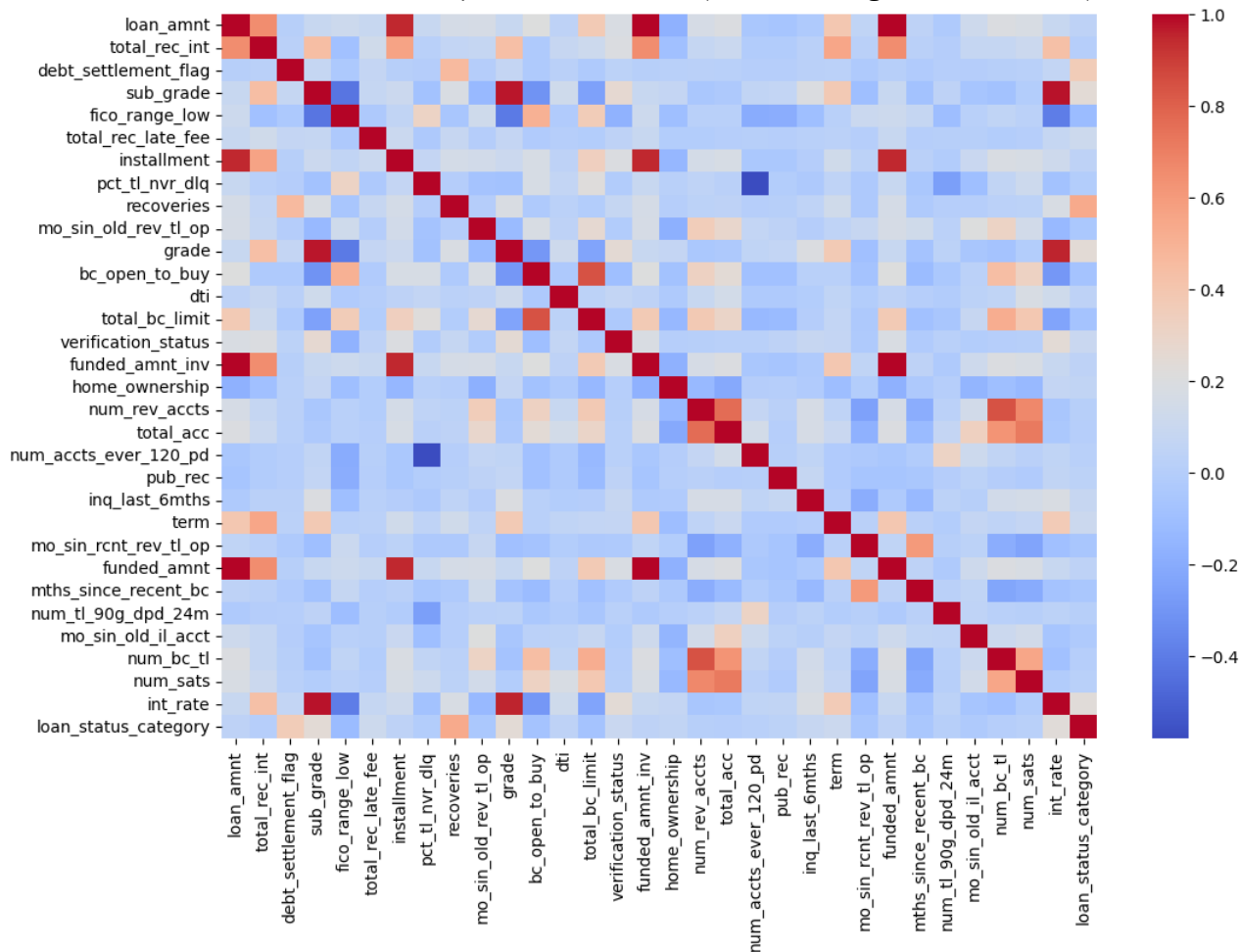
sub_grade (categorical)	→	sub_grade (numerical)
A1	→	1
A2	→	2

Feature Selection

- 1- Selected most relevant features using LASSO
- 2- Optimal LASSO regularisation parameter α obtained using k-fold cross-validation
- 3- TOP 30 feature selection – avoid overfitting & computational trade-off
- 4- Remove highly correlated features
- 5- Economic intuition
 - Lasso selected features
 - Keep the Interest Rate predictor (Basel II compliance)

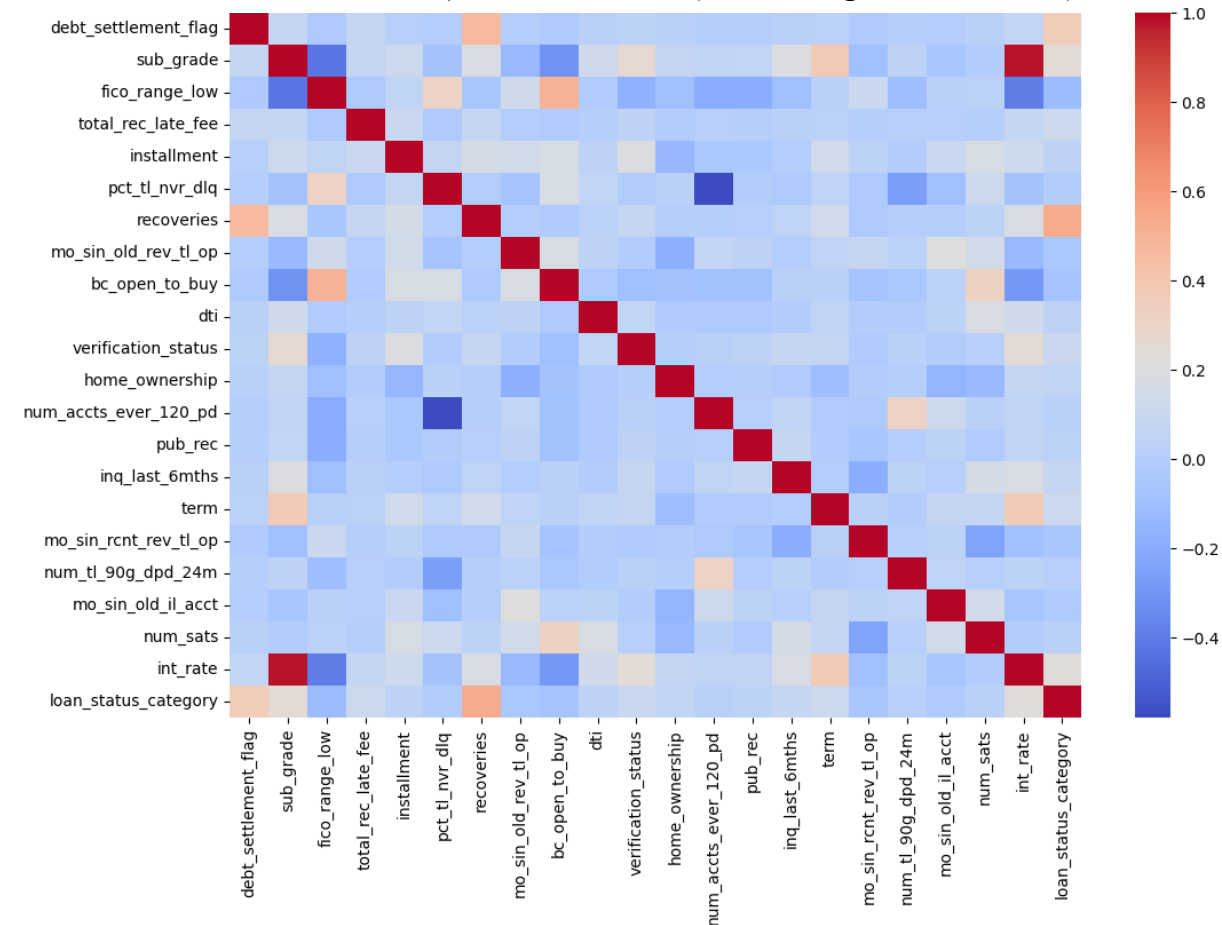
Feature Selection – Correlation Heatmap

Correlation Heatmap of Selected Features (before removing correlated features)



30 features selected with LASSO
(+ loan int. rates)

Correlation Heatmap of Selected Features (after removing correlated features)



20 features selected after correlation analysis
(+ loan int. rates)

Data Balancing

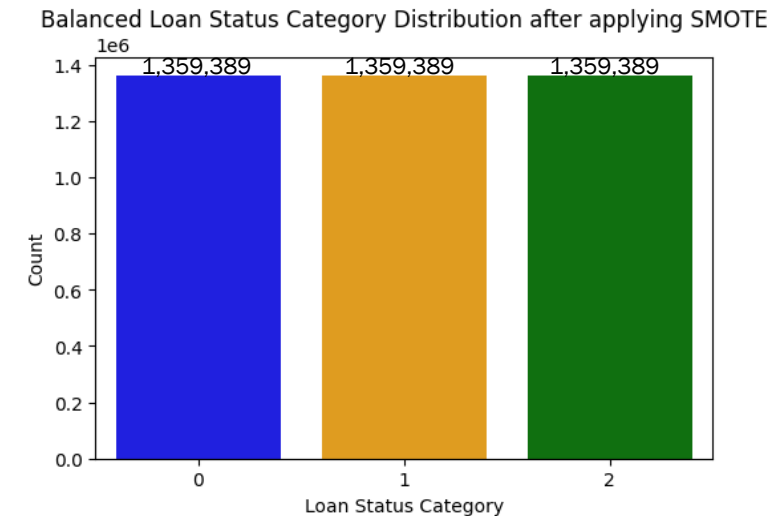
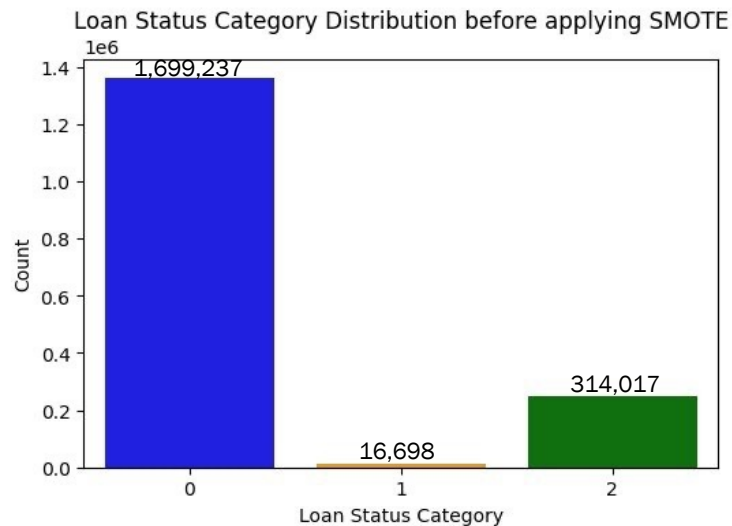
1- Split dataset into 80% training and 20% testing

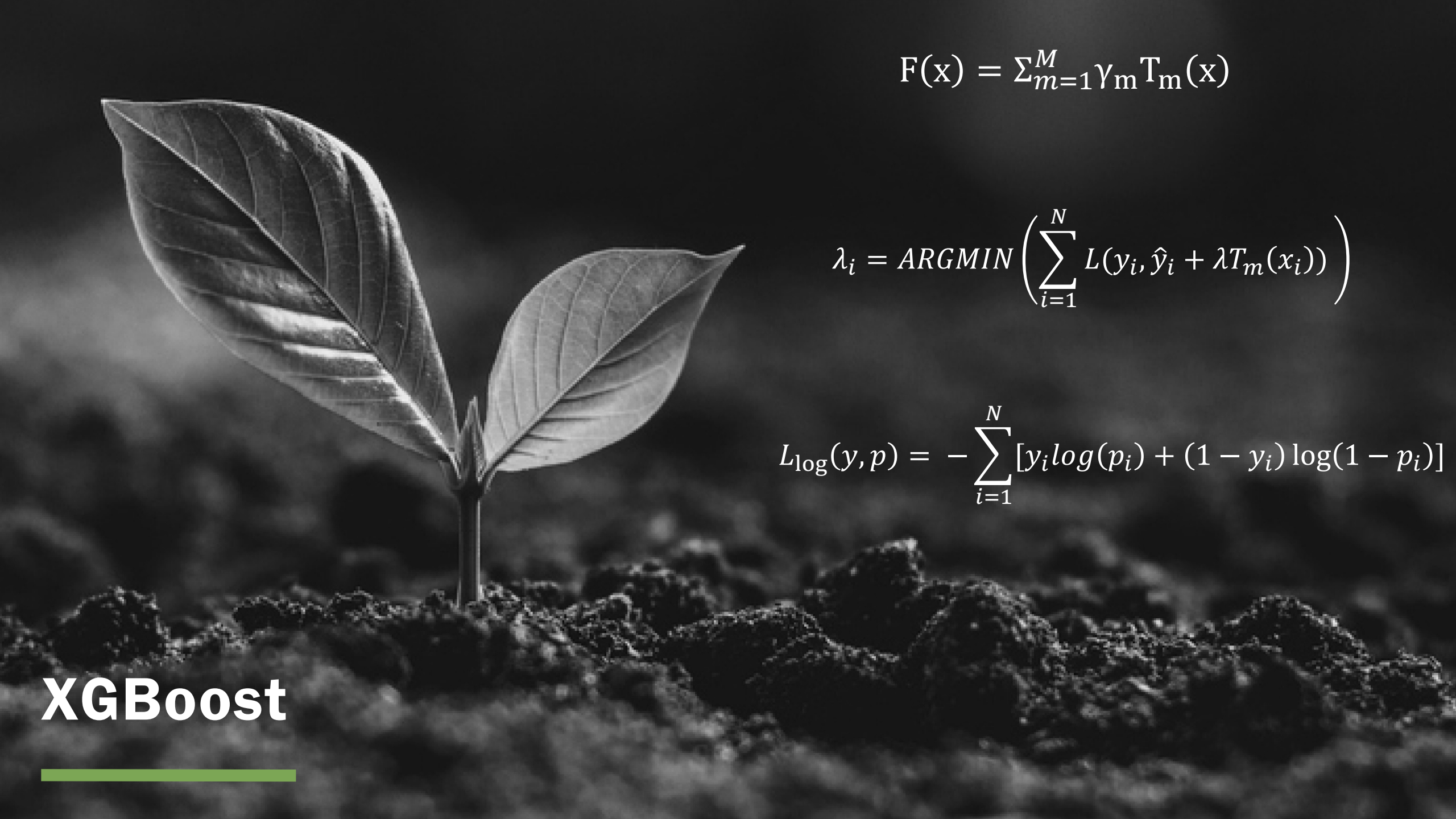
Ensure train and test sets maintain same class distribution as original dataset

2- Apply SMOTE to **training data**

Why SMOTE? - Dataset highly imbalanced

Why not apply to testing set too? - Prevent introducing artificial data, prevent data leakage when testing



A black and white photograph of a young plant with two leaves growing out of dark, textured soil. The plant is positioned on the left side of the frame, with its stem and leaves clearly visible against the dark background of the soil.
$$F(x) = \sum_{m=1}^M \gamma_m T_m(x)$$

$$\lambda_i = \text{ARGMIN} \left(\sum_{i=1}^N L(y_i, \hat{y}_i + \lambda T_m(x_i)) \right)$$

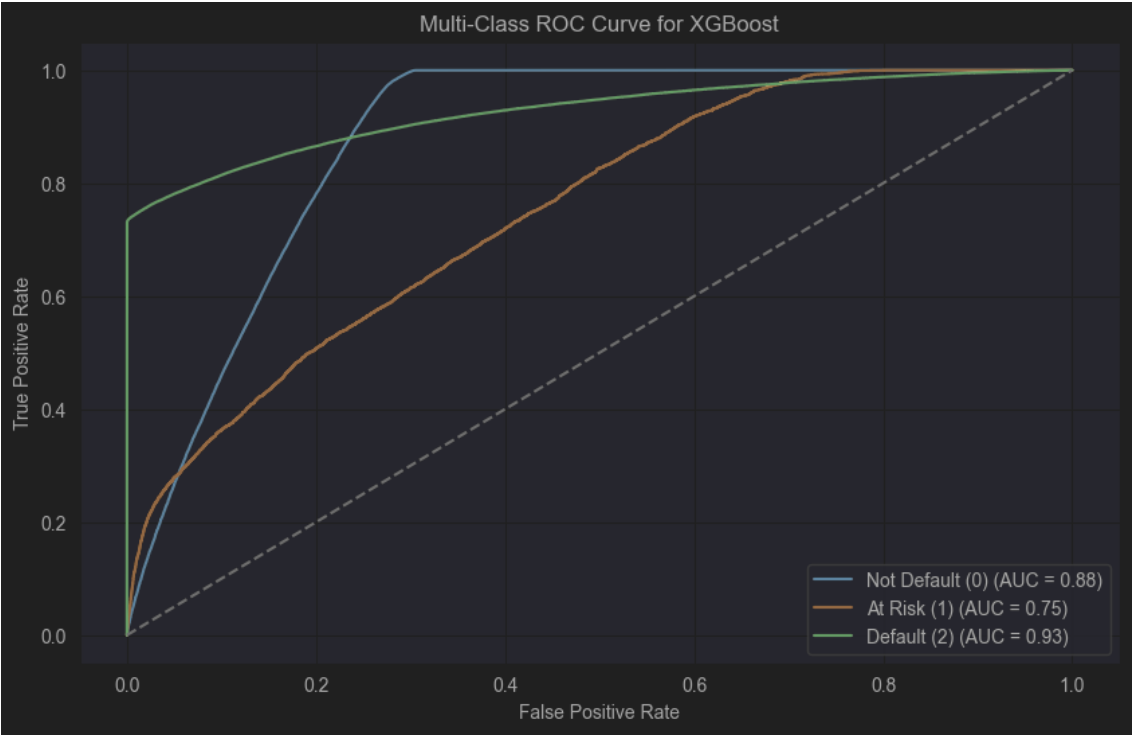
$$L_{\log}(y, p) = - \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

XGBoost

Model Performance

Class	Precision	Recall	F1
No Default	0.95	0.98	0.96
At Risk	0.09	0.17	0.11
Default	0.99	0.74	0.84

		Predicted		
		No Default	At Risk	Default
Actual	No Default	333,916	5413	519
	At Risk	2741	564	35
	Default	16,004	578	46,221





Next steps

- Rebalancing
- Macroeconomic enhancement