Accion Technical Assessment

Using Machine Learning to Predict Default_20_6

Flag if Next Loan Renewal has Maximum DPD ≥ 20 within First 6 Months (1 = Yes, 0 = No)

- Yash Kanakia

Problem Statement

Objective: Predict whether a loan renewal will default (DPD ≥ 20) within the first 6 months.

Goal: Help reduce the risk of lending to potentially defaulting borrowers by identifying key indicators early.

Dataset Overview

Number of records: 18029

Key variables:

- **Customer Stability**: Age, marital status, vintage.
- **Loan Characteristics**: Loan cycle, loan amount, interest rate, tenure.
- **Past Performance**: Max DPDs, arrears (principal and interest), recovery patterns.
- **Delinquency Measures**: Amount and count of DPDs across different ranges

Imbalance in the target variable: There is an imbalance of class here. 90% values are 0 i.e. Non-Defaulters and 10% are 1 that i.e. Defaulters.

Exploratory Data Analysis (EDA)

Demographics Overview:

- **Marital Status**: Married: 94%, Widow: 6%, Unmarried: <1%
- **Education: Elementary**: 44%, High School: 26%, Middle School: 25%, No School: 3%, Higher Ed: 2%
- **Province Groups**: A1: 45%, A2: 17%, B2: 14%, Others: 24%
- **Income Source**: Business: 99.7%, Others: 0.3%

Key Business Fields:

- **Trade**: 63%, **Agriculture**: 16%, **Home Industry & Livestock**: 16%

Statistical Highlights:

- **Significant Factors**: Marital Status, Education, Business Field, Province Group ($p < 0.05$).
- **Non-Significant**: Source of Income ($p > 0.05$).

Distribution Observations:

- **High Variance**: Installment Amount, Business Length, Other Income.
- **Skewed Features & Outliers**: Need treatment before modeling.

Multivariate Insights:

- **Correlations**: Loan, Installment & Principal Amounts.
- **Defaults**: Higher among widows, no education, and customers from B2.

Data Cleaning and Feature Selection

- Dropping Duplicates:
  - **Total Rows** (After Removing Duplicates) : 9775
- Feature Selection:
  - **Categorical**: Marital, Education, Income Source, etc.
  - **Numerical**: Loan Cycle, Amount, Delinquency, etc.
- Feature Elimination:
  - Default Columns (Except Default_20_6)
  - ID, Date, Note -> **Nominal**
  - Group, Weight -> **Redundant**
  - Columns with binary values (**redundant**)
  - Columns where that are **Multicollinear**/Too Many Values (Increase **Dimensionality**)/ Random Values (length of business) / No Information (Diff_Disb_LOD, data_jangkawaktu)

Model Building and Tuning

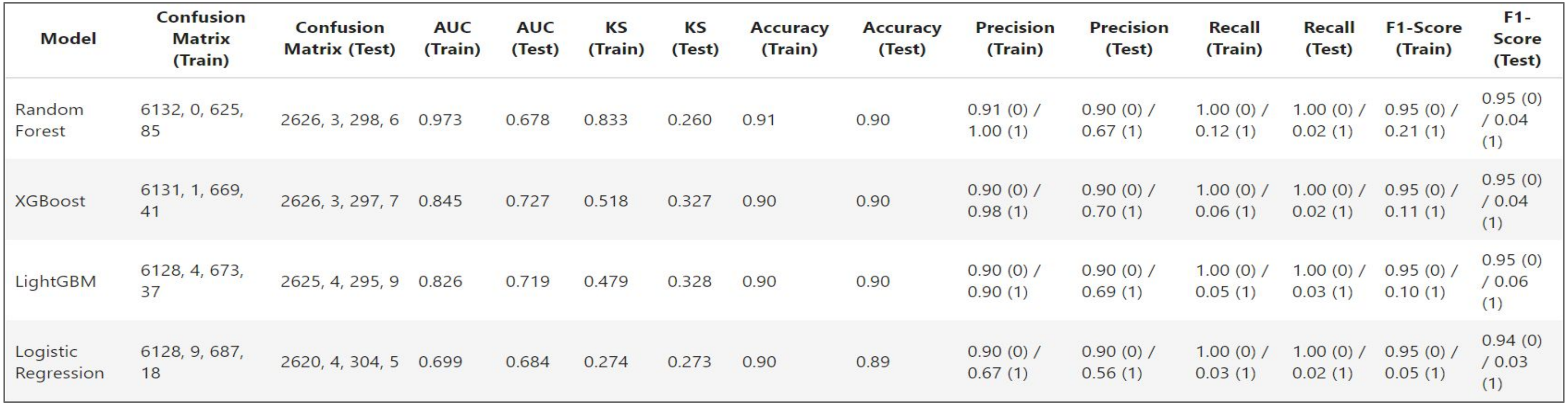
Models Tested:

- Random Forest - Base, Tuned, and Most Important Feature
- XGBoost - Base, Tuned, and Most Important Feature
- LightGBM - Base, Tuned, and Most Important Feature
- Logistic Regression

Parameter Tuning:

- Used techniques like **grid search**, **cross-validation**, **stratification**, and **scale_pos_weight** tuning in XGBoost for class imbalance.
- Adjusted Threshold

Model Performance Comparison



Model	Confusion Matrix (Train)	Confusion Matrix (Test)	AUC (Train)	AUC (Test)	KS (Train)	KS (Test)	Accuracy (Train)	Accuracy (Test)	Precision (Train)	Precision (Test)	Recall (Train)	Recall (Test)	F1-Score (Train)	F1-Score (Test)
Random Forest	6132, 0, 625, 85	2626, 3, 298, 6	0.973	0.678	0.833	0.260	0.91	0.90	0.91 (0) / 1.00 (1)	0.90 (0) / 0.67 (1)	1.00 (0) / 0.12 (1)	1.00 (0) / 0.02 (1)	0.95 (0) / 0.21 (1)	0.95 (0) / 0.04 (1)
XGBoost	6131, 1, 669, 41	2626, 3, 297, 7	0.845	0.727	0.518	0.327	0.90	0.90	0.90 (0) / 0.98 (1)	0.90 (0) / 0.70 (1)	1.00 (0) / 0.06 (1)	1.00 (0) / 0.02 (1)	0.95 (0) / 0.11 (1)	0.95 (0) / 0.04 (1)
LightGBM	6128, 4, 673, 37	2625, 4, 295, 9	0.826	0.719	0.479	0.328	0.90	0.90	0.90 (0) / 0.90 (1)	0.90 (0) / 0.69 (1)	1.00 (0) / 0.05 (1)	1.00 (0) / 0.03 (1)	0.95 (0) / 0.10 (1)	0.95 (0) / 0.06 (1)
Logistic Regression	6128, 9, 687, 18	2620, 4, 304, 5	0.699	0.684	0.274	0.273	0.90	0.89	0.90 (0) / 0.67 (1)	0.90 (0) / 0.56 (1)	1.00 (0) / 0.03 (1)	1.00 (0) / 0.02 (1)	0.95 (0) / 0.05 (1)	0.94 (0) / 0.03 (1)

Overall Performance:

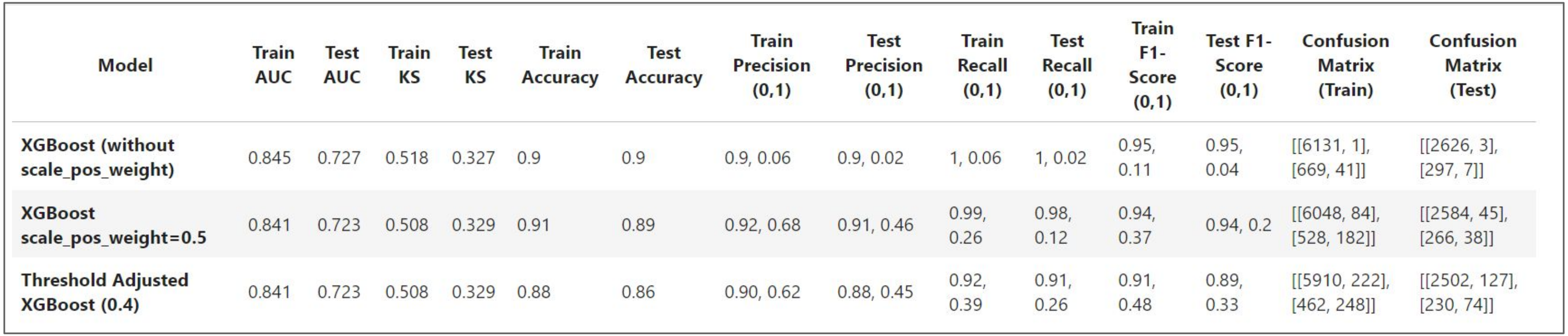
XGBoost is the most balanced model overall, having the highest Test AUC and Test KS, while maintaining a high Test Accuracy and Precision. **LightGBM** is also a strong candidate, showing competitive performance with the highest Test KS and a good balance in other metrics.

Class Imbalance:

All models struggled with the positive class (1) prediction, as indicated by the confusion matrices showing many true negatives and few true positives. This suggests a possible class imbalance issue in the dataset.

Further Model Improvement

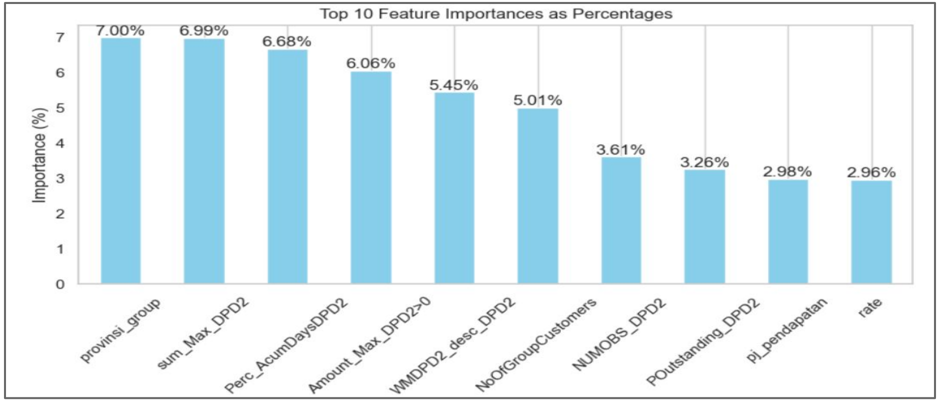
1. XGBoost is the best performing comparatively in terms of AUC, KS and Precision-Recall.
2. Although, the recall is low and needs further improvement in terms of balancing the target variable class - using `scale_pos_weight`
3. Comparing the performance XGBoost Model with and without the `scale_pos_weight` parameter and adjusted threshold.
4. Summarize
   - **Recall Improvement**: The tuned model with adjusted threshold significantly improves recall for defaults in both train and test sets.
   - **Precision Trade-off**: Improved precision for defaults comes at the cost of decreased recall for non-defaults to some extent

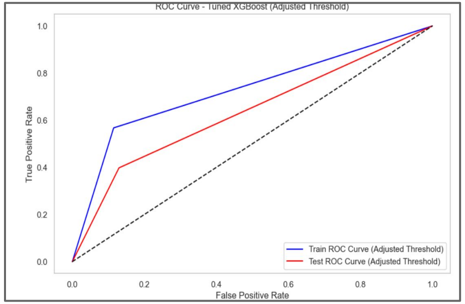


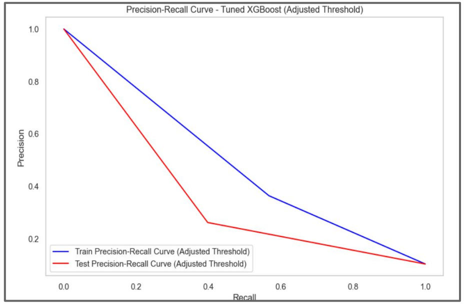
Model	Train AUC	Test AUC	Train KS	Test KS	Train Accuracy	Test Accuracy	Train Precision (0,1)	Test Precision (0,1)	Train Recall (0,1)	Test Recall (0,1)	Train F1-Score (0,1)	Test F1-Score (0,1)	Confusion Matrix (Train)	Confusion Matrix (Test)
XGBoost (without scale_pos_weight)	0.845	0.727	0.518	0.327	0.9	0.9	0.9, 0.06	0.9, 0.02	1, 0.06	1, 0.02	0.95, 0.11	0.95, 0.04	[[6131, 1], [669, 41]]	[[2626, 3], [297, 7]]
XGBoost scale_pos_weight=0.5	0.841	0.723	0.508	0.329	0.91	0.89	0.92, 0.68	0.91, 0.46	0.99, 0.26	0.98, 0.12	0.94, 0.37	0.94, 0.2	[[6048, 84], [528, 182]]	[[2584, 45], [266, 38]]
Threshold Adjusted XGBoost (0.4)	0.841	0.723	0.508	0.329	0.88	0.86	0.90, 0.62	0.88, 0.45	0.92, 0.39	0.91, 0.26	0.91, 0.48	0.89, 0.33	[[5910, 222], [462, 248]]	[[2502, 127], [230, 74]]

Final Model Summary - XGBoost (threshold 0.4)

The top 10 features for identifying potential loan defaulters include geographic location (**provinsi_group**), cumulative maximum days past due (**sum_Max_DPD2**), and the percentage of accumulated days past due (**Perc_AcumDaysDPD2**), which all indicate the likelihood of default. Additional indicators such as the presence of any days past due (**Amount_Max_DPD2>0**) and outstanding principal amounts (**POutstanding_DPD2**) highlight customers with troubling repayment histories. Customer income (**pj_pendapatan**) and interest rates (**rate**) further provide insight into financial strain that may increase default risk. Focusing on these features can help businesses implement targeted risk management strategies to mitigate potential losses.







Insights and Recommendations

- **Focus on Risk Identification**: Leverage the tuned model to pinpoint high-risk customers for proactive risk management measures.
- **Threshold Adjustmen**t: Adjust the classification threshold to align with business goals, potentially capturing more defaults for cautious lending.
- **Feature Importance Analysis**: Utilize key features like geographic location and days past due to enhance risk management strategies for identifying potential defaulters.
- **Targeted Communication**: Develop tailored communication plans for high-risk customers, offering personalized financial advice and support.
- **Ongoing Monitoring and Updates**: Continuously assess the model's performance and retrain as needed to adapt to changing economic conditions and borrower behavior.
- **Test New Features**: Explore additional predictive features, such as behavioral data and recent spending patterns, to improve model accuracy.
- **Evaluate Costs of False Positives**: Analyze the implications of false positives to ensure that the costs of investigations do not outweigh potential losses from missed defaults.