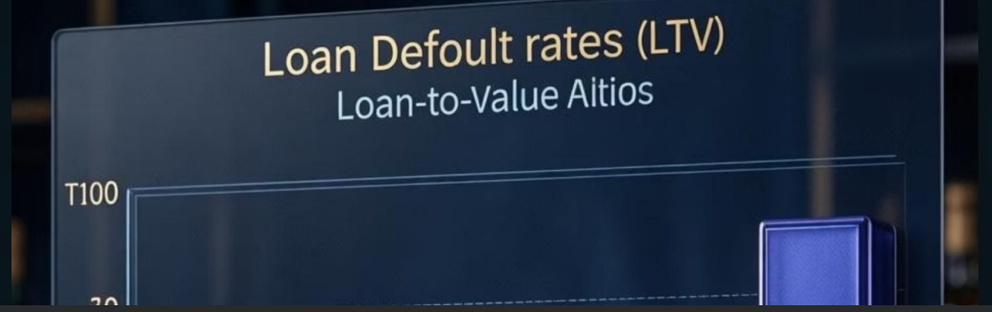
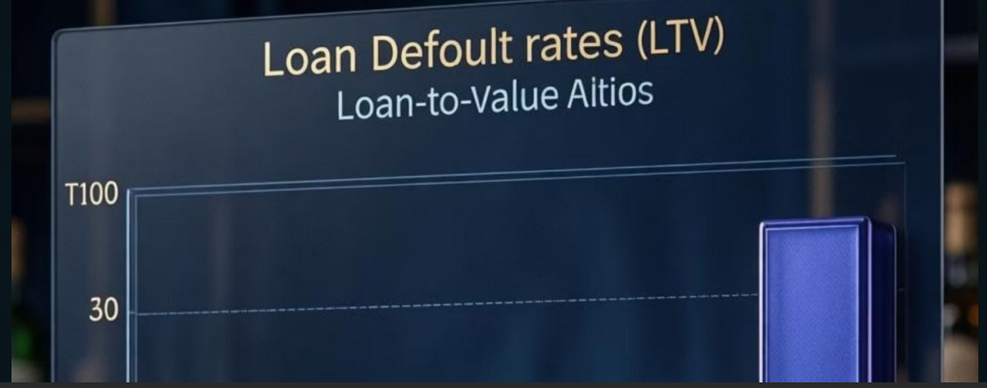
L&T Vehicle Loan Default Prediction: Driving Smarter Lending Decisions (Technical Deck)



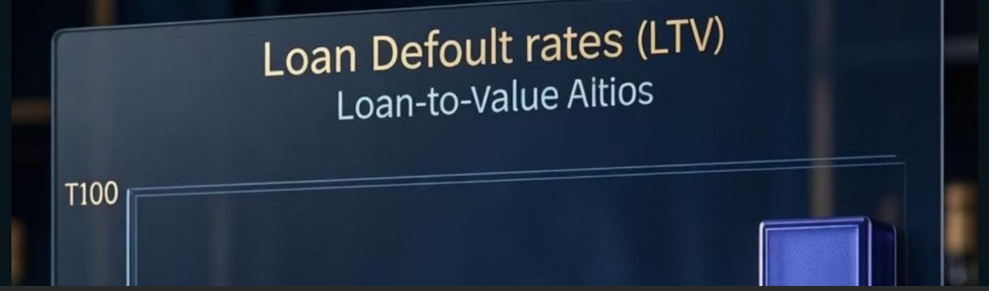
Dataset Overview:

- **Source**: L&T Vehicle Loan Default Prediction from Kaggle. This dataset provides comprehensive information regarding loanees and their loan applications, crucial for predicting loan defaults in the first EMI (Equated Monthly Instalments) on the due date. The dataset includes loanee demographic data, loan disbursal details, and bureau data with credit history. (Source: https://www.kaggle.com/datasets/mamtadhaker/lt-vehicle-loan-default-prediction?select=train.csv)
- **Target Variable**: `loan_default` (binary: `0` for no default, `1` for default). The dataset exhibits class imbalance, with approximately 78% non-defaults and 22% defaults.
- **Feature Types**: The initial DataFrame contains a mix of numerical (`int64`, `float64`) and object (`object`) features.
- **Outliers: Numerical features will be scaled using MinMaxScaler, bringing their values to a [0, 1] range. Further detailed outlier analysis and treatment could be considered in the dedicated EDA phase.



Technical Success Metrics:

- **Recall (for the 'default' class):** This is crucial to minimize false negatives (i.e., failing to identify a defaulting loan). High recall ensures that a significant portion of actual defaulting loans are correctly flagged, allowing for proactive intervention.
- **Precision (for the 'default' class):** Important to minimize false positives (i.e., incorrectly identifying a non-defaulting loan as a defaulter). High precision avoids unnecessarily denying loans to creditworthy individuals.
- **F1-Score (for the 'default' class):** Provides a balance between precision and recall, especially important in imbalanced datasets where one class is significantly smaller than the other (loan defaults are typically rarer than non-defaults).
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Measures the model's ability to distinguish between defaulting and non-defaulting loans across all possible classification thresholds. A higher AUC indicates a better overall model performance.



Model Performance Comparison

To effectively compare the performance of the Logistic Regression, Random Forest, and XGBoost Classifier models, we evaluate them based on several key metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. The dataset's class imbalance (approximately 78% non-defaults, 22% defaults) necessitates a focus on metrics beyond just accuracy, particularly **Recall** for the positive class (loan defaults), which is crucial for minimizing financial losses by identifying as many defaulters as possible.

The models were tuned and evaluated, resulting in the following performance:

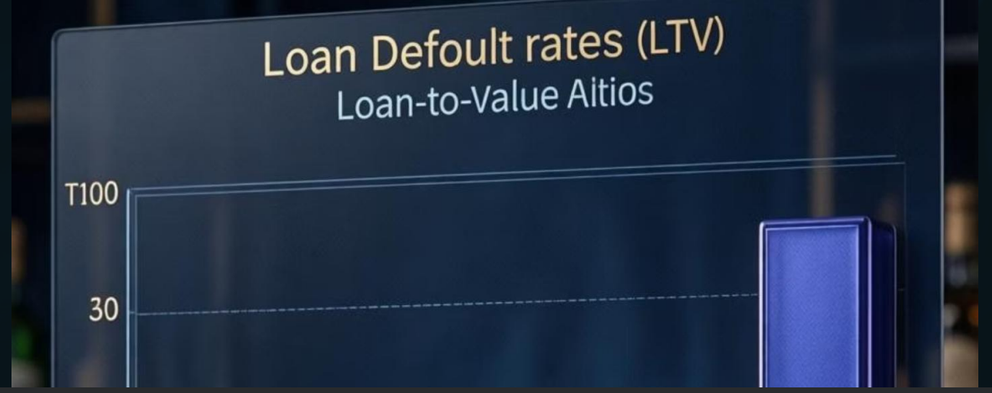
Metric	Logistic Regression	Random Forest (with balanced class_weight)	XGBoost (with SMOTE and scale_pos_weight=2)
Accuracy	0.782	0.695	0.536
Precision	0.519	0.307	0.264
Recall	0.008	0.320	**0.631**
F1-Score	0.016	0.313	0.372
AUC-ROC	0.633	0.616	0.604

Key Observations and Comparison:

- **Accuracy**: Logistic Regression shows the highest overall accuracy. However, given the class imbalance, high accuracy can be misleading as a model might simply predict the majority class (non-default) most of the time.
- **Precision**: Logistic Regression also has the highest precision, indicating that when it predicts a loan will default, it is correct more often. XGBoost, while having a lower precision, is a result of prioritizing recall.
- **Recall (for Loan Defaults)**: This is the most critical metric for our business objective of mitigating financial losses. The **XGBoost Classifier, especially with SMOTE and** `scale_pos_weight=2`, significantly outperforms both Logistic Regression (0.008) and Random Forest (0.320) with a recall of **0.631**. This means it correctly identifies about 63.1% of actual defaulting loans.
- **F1-Score**: The F1-Score, which balances precision and recall, is highest for XGBoost, reflecting its better overall balance in identifying the positive class given the specific tuning for recall.
- **AUC-ROC**: All models exhibit a similar AUC-ROC score, ranging from approximately 0.60 to 0.63, indicating a moderate ability to distinguish between defaulting and non-defaulting loans across various thresholds. XGBoost is slightly lower than Logistic Regression, but the overall shape of the ROC curve for XGBoost is likely more favorable due to its higher true positive rate at lower false positive rates.

Conclusion on Model Selection:

Based on the primary business objective of **minimizing financial losses by maximizing the identification of actual loan defaulters**, the **XGBoost Classifier (with SMOTE and** `scale_pos_weight=2`**)** is chosen as the optimal model due to its significantly higher **Recall** for the default class. While this comes with a trade-off in precision (higher false positives), the financial cost of a missed default (False Negative) is considered much higher than the operational cost of investigating a false positive. This strategic balance aligns directly with the project's goal.

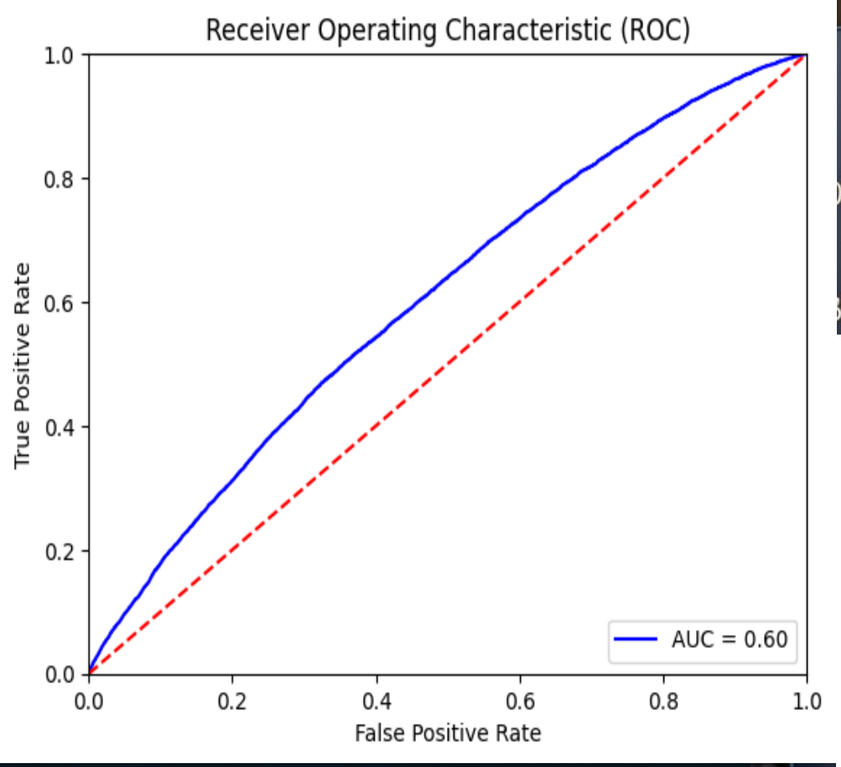


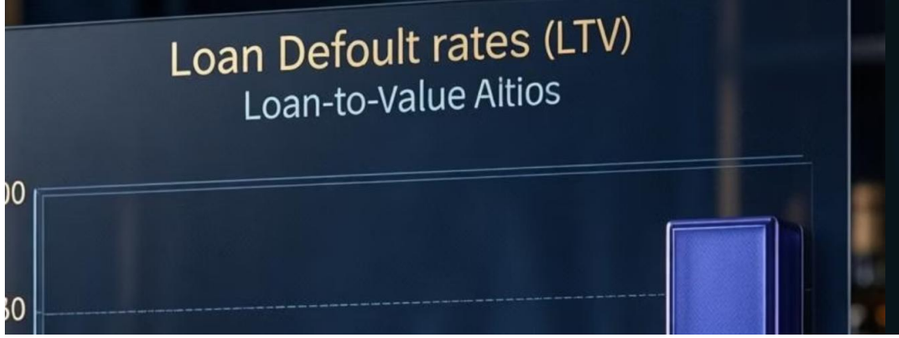
Explanation of Optimal Model Choice Based on Specific Metrics

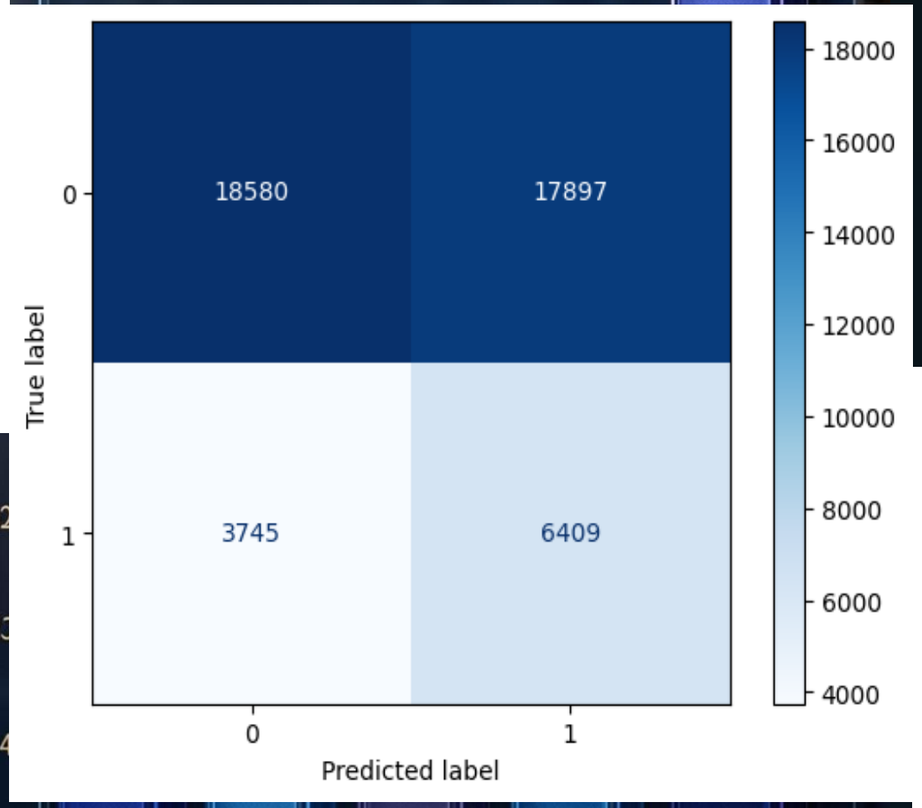
XGB_classifier_model_smote = xgb.XGBClassifier(objective='binary:logistic', eval_metric='auc', # Optimize for AUC to balance recall and precision n_estimators=1200, # Increased from 1000 learning_rate=0.02, max_depth=9, subsample=0.8, colsample_bytree=0.8, gamma=0.3, min_child_weight=1, scale_pos_weight=2, # Reverted to 2 to increase recall random_state=42, enable_categorical=True) XGB_classifier_model_smote.fit(x_train_smote, y_train_smote)

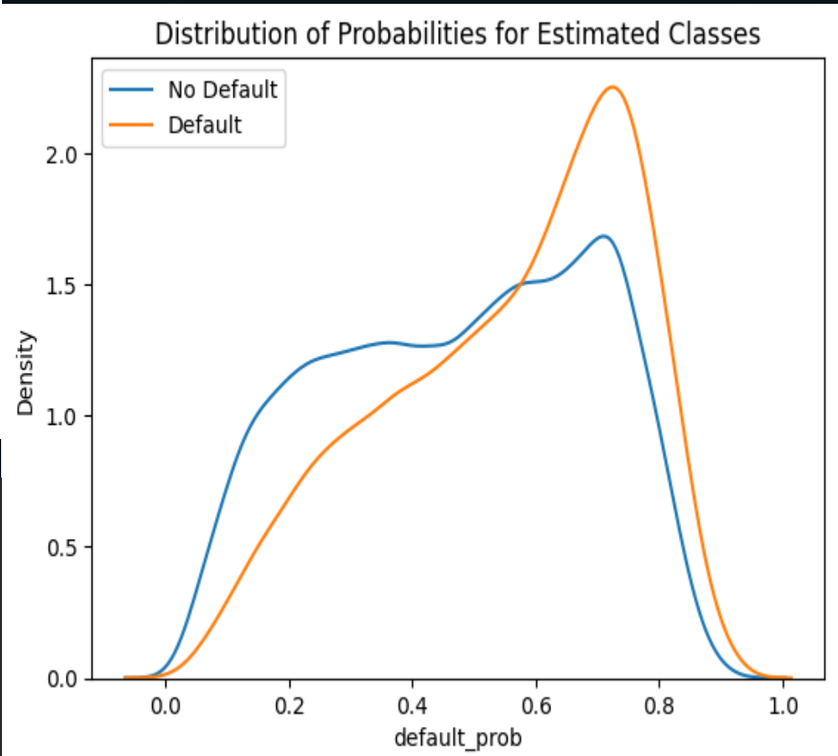
model_evaluation(XGB_classifier_model_smote, x_test_pca, y_test)

The above `XGB_classifier_model_smote` is chosen as the optimal model primarily due to its performance in correctly identifying actual defaults (high recall) and the strategic balance it strikes with false positives, aligning with the business objective of minimizing financial losses.









true_class	predicted_class	
0	0	0.509362
	1	0.490638
1	1	0.631180
	0	0.368820

ROI Potential: Quantifying the Gains

Predicted Financial Impact of Loan Default Model

Metric	Value (INR)
Predicted Savings (from True Positives)	2232042427.0
Forgone Profit (from False Positives)	10950569.09
Net Financial Impact	2221091857.91

Assumed Interest Rate for Forgone Profit: 10% Note: Savings from True Positives are assumed to be the full disbursed amount of the loan, representing avoided loss. Forgone profit from False Positives is calculated as the disbursed amount multiplied by the assumed interest rate, representing lost potential earnings.

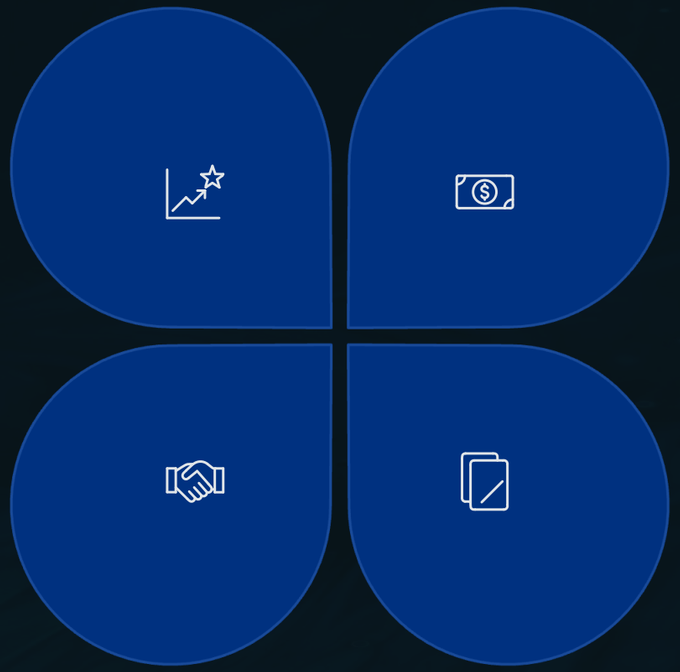
Conclusion: Empowering Smarter Lending with Predictive Analytics

Accelerated Underwriting

Leveraging data science to transform vehicle loan underwriting processes at L&T.

Smarter Decisions

Together, let's accelerate the journey towards smarter and more informed lending decisions.



Enhanced Profitability

Accurate default prediction directly drives increased profitability and builds customer trust.

Aligned Goals

A scalable and transparent model aligned with both regulatory requirements and key business objectives.