MDS | CAPSTONE PROJECT

LOAN DEFAULT PREDICTION

Final Presentation

TEAM 4 - SYNTEGRITY

THE TEAM – SYNTEGRITY (TEAM 4)

Shaishav Merchant
Singapore

Key Contributions
- Solution Approach
- Exploratory Data Analysis
- Methodology

Desmond Muzuva
Zimbabwe

Key Contributions
- Model Building
- Hyperparameter Tuning
- Performance Metrics

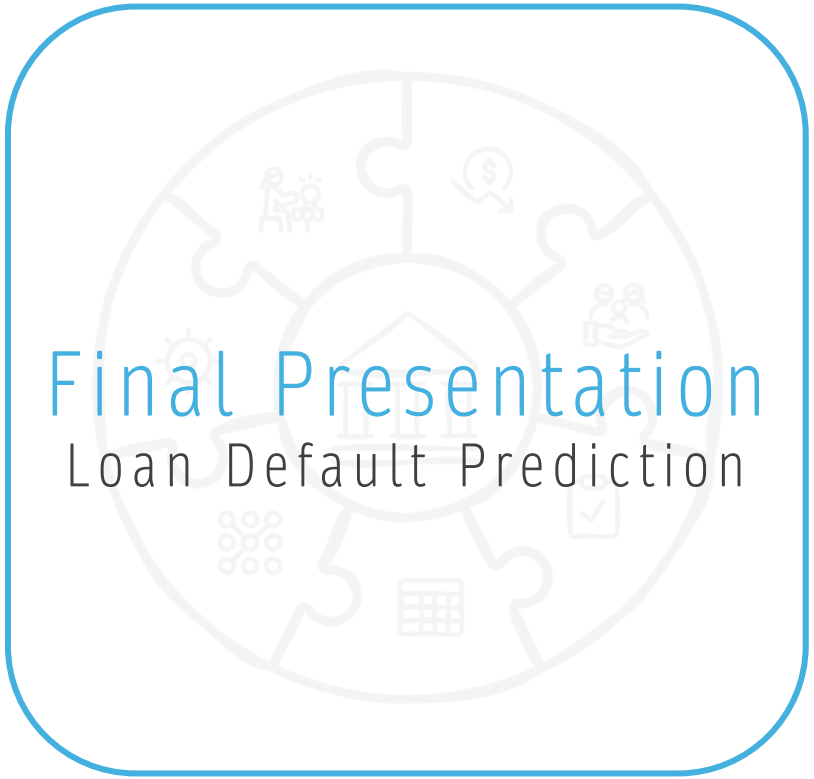
Monsuru Sodeeq
Nigeria

Key Contributions
- Model Comparison
- Final Model Selection
- Feature Importance

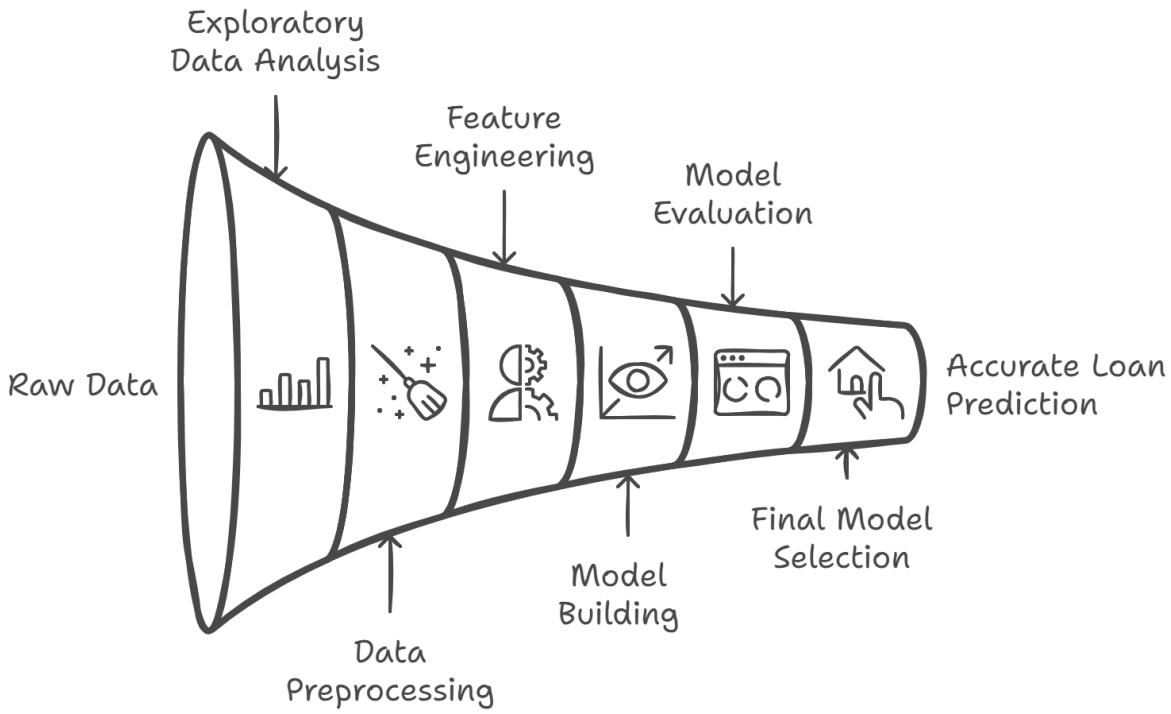
Vu Thi Ai Duyen (Daisy)
Singapore

Key Contributions
- Key Insights
- Business Recommendations

AGENDA



EXECUTIVE SUMMARY – PROBLEM AND APPROACH



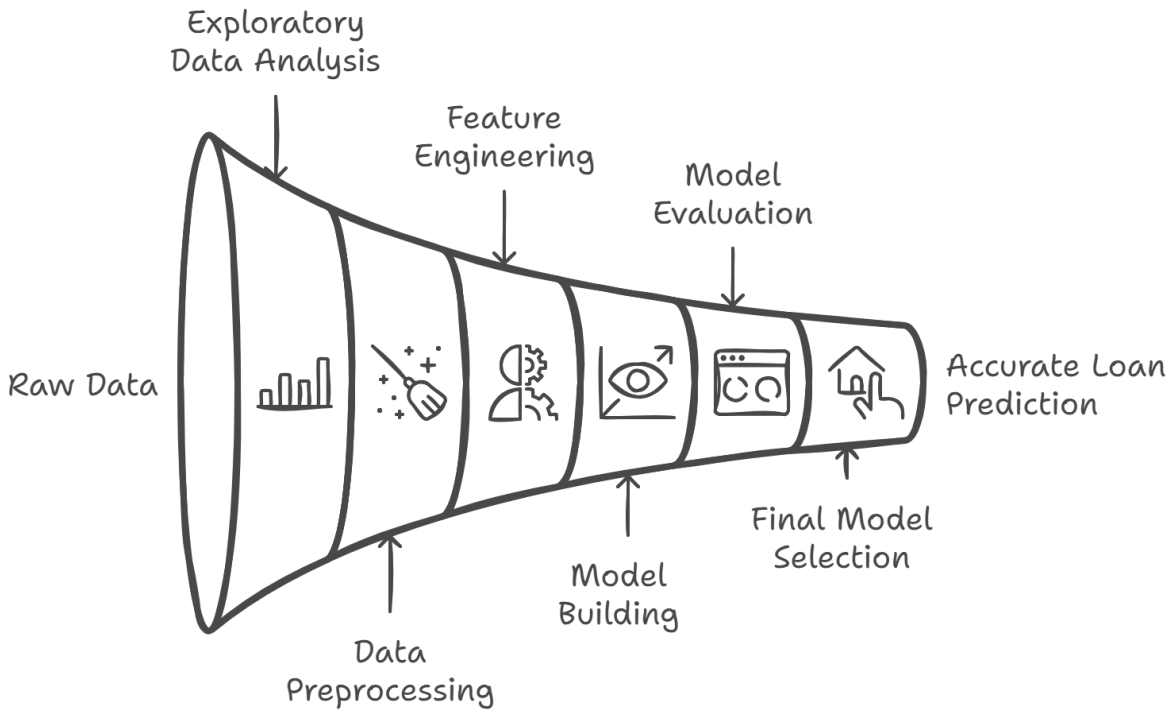
Problem and Solution Overview:

- Automated loan approval using machine learning to reduce human error and bias.
- Exploratory Data Analysis, Data Processing ensured data readiness.

Model Development and Evaluation:

- Built linear and ensemble models, focusing on accuracy and interpretability.
- Evaluated performance with metrics like recall, balanced precision as well value preposition.
- **Random Forest model** trained on Original Data has been proposed as it offered best chance to identify defaulters, minimizing loss.

EXECUTIVE SUMMARY – KEY TAKEAWAYS

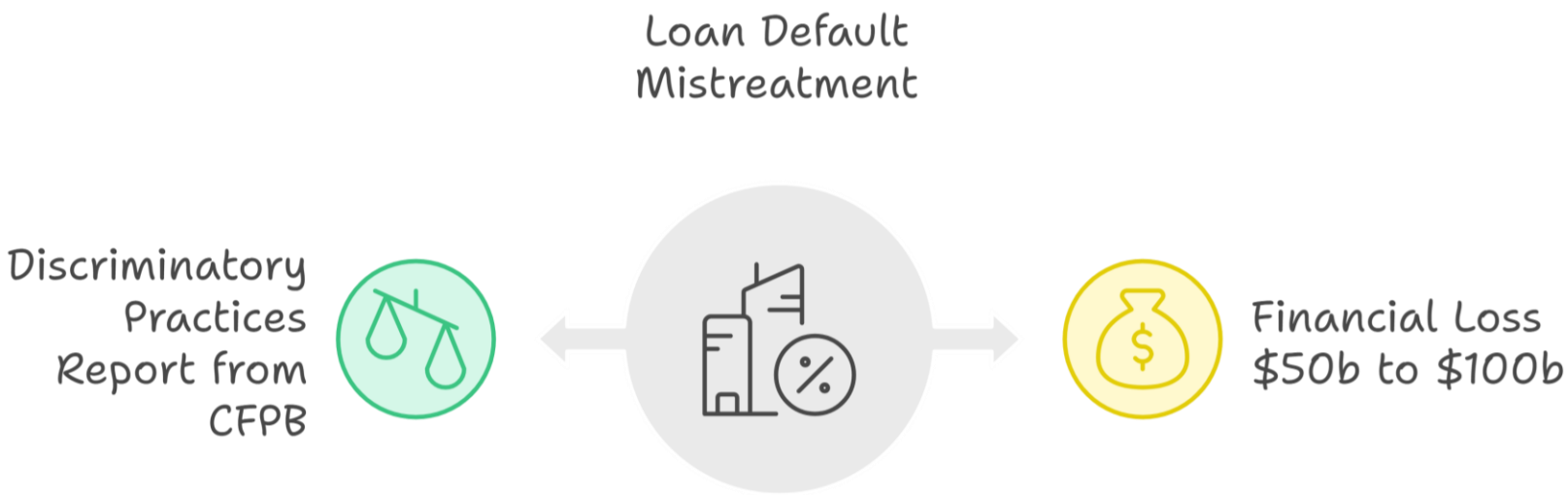


Key Insights and Next Steps:

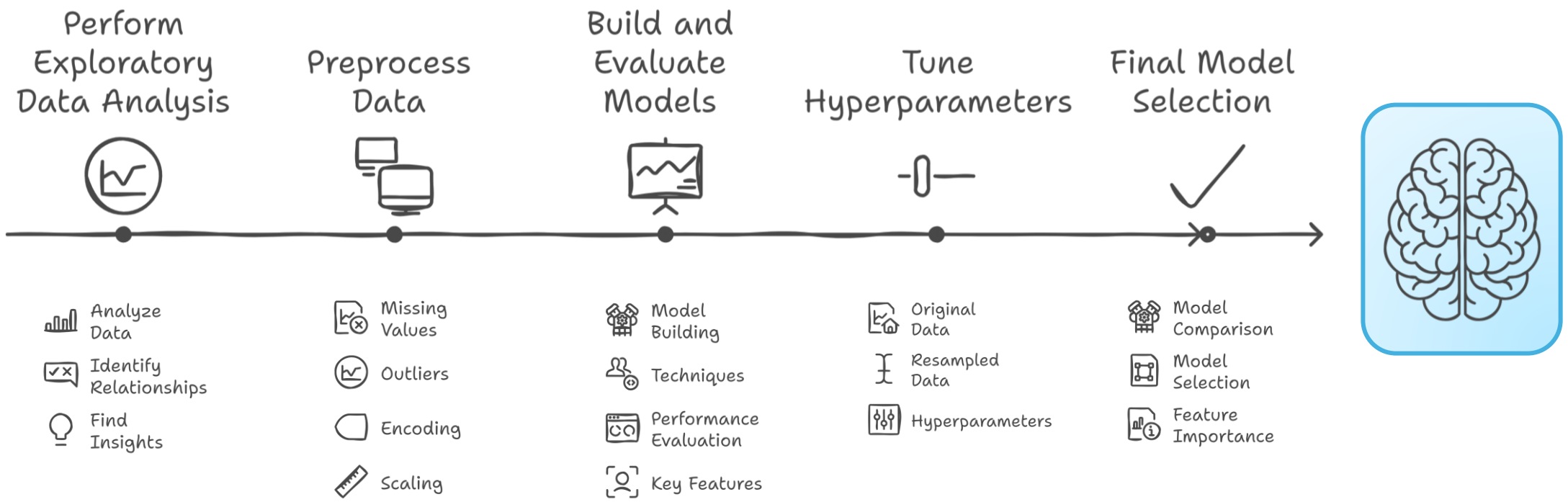
- **Deploy the model** within the bank's underwriting system, followed by regular monitoring.
- **Update the model** periodically with fresh data to maintain its effectiveness and profitability.
- Refine the **loan approval strategy** based on model insights to further reduce defaults and increase profits.

PROBLEM OVERVIEW

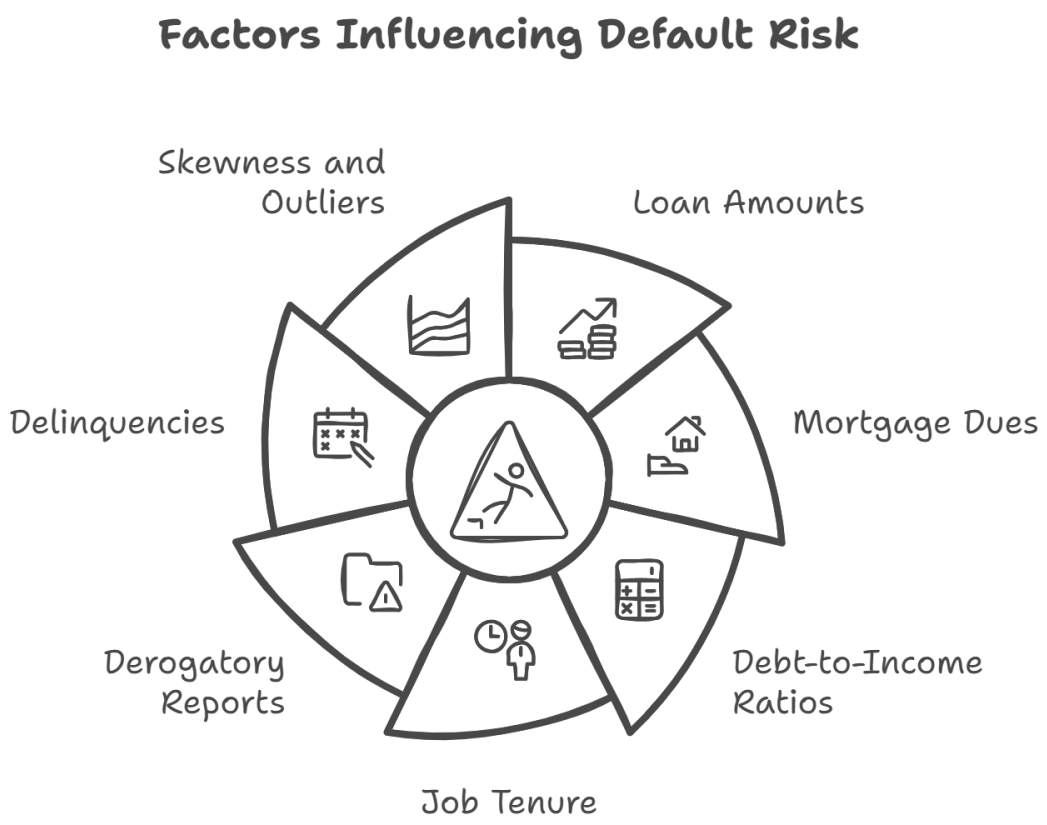
- **Financial Loss:** Banks are facing increasing challenges in accurately identifying loan defaulters, with an estimated $50-$100 billion lost during economic downturns.
- **Bias Concerns:** The Consumer Finance Protection Bureau (CFPB) also highlights existing biases in loan approvals.
- **Leveraging Technology:** This project aims to automate the loan approval process using machine learning to reduce human bias, ensure compliance with the Equal Credit Opportunity Act (ECOA), and improve fairness, transparency, and decision-making in lending.



SOLUTION APPROACH



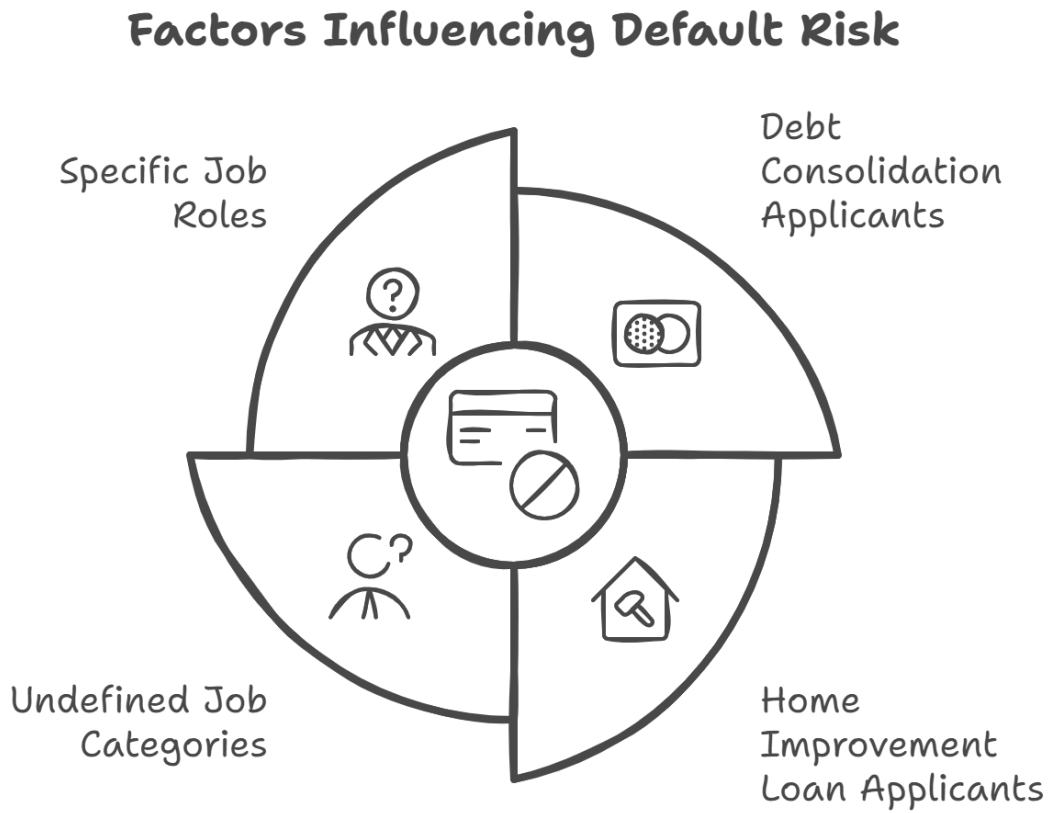
EDA – KEY INSIGHTS – NUMERICAL FEATURES



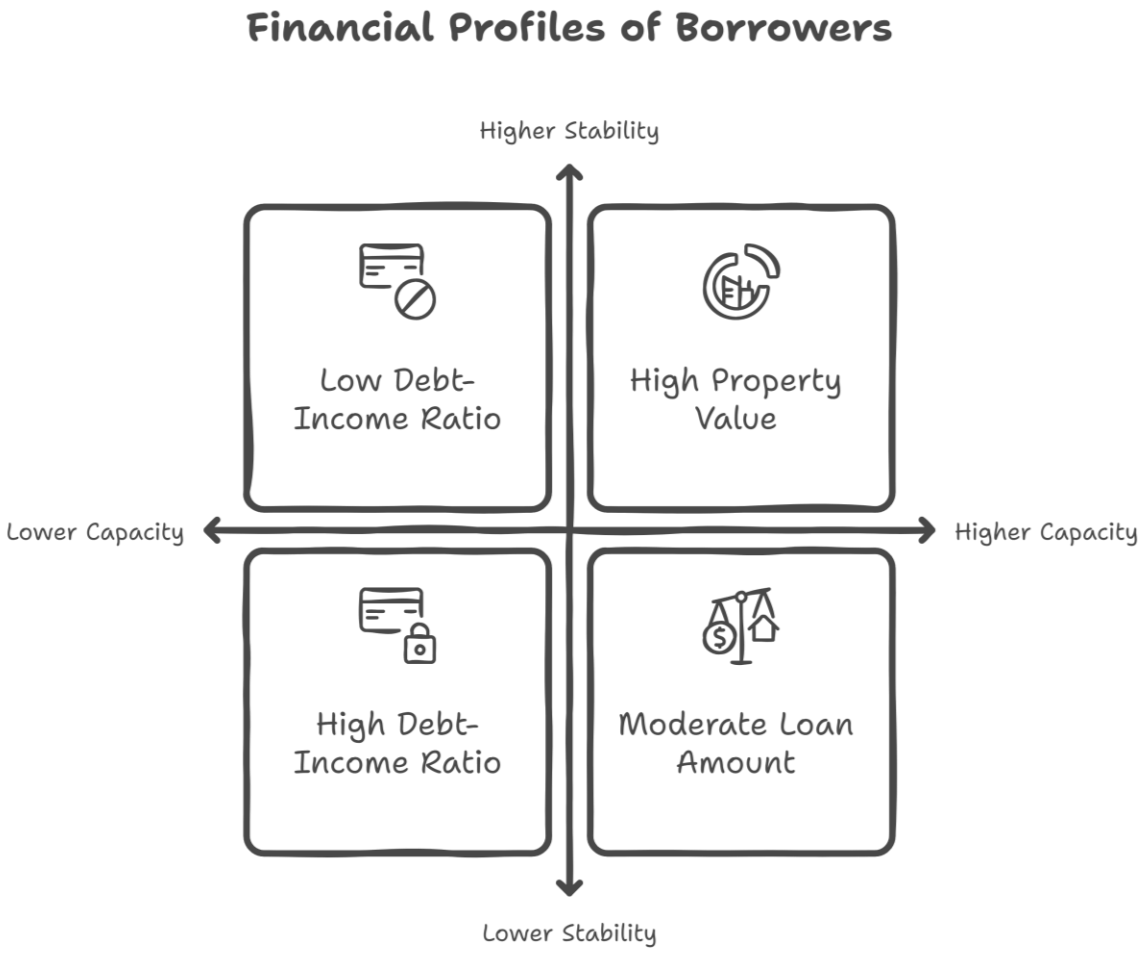
- **Loan and Debt Influence:** Higher loan amounts, mortgage dues, and debt-to-income ratios are linked to higher default risk.
- **Employment and Credit History:** Shorter job tenure and a higher number of derogatory reports or delinquencies increase default likelihood.
- **Data Distribution and Outliers:** Most features have mild skewness, with a few notable outliers, especially in loan and debt-related variables.

EDA – KEY INSIGHTS – CATEGORICAL FEATURES

- **REASON:** Applicants applying for debt consolidation are more likely to default than those applying for home improvement loans.
- **JOB:** Undefined job categories (classified as "**Other**") show a higher default rate compared to specific roles like Manager or Office jobs.



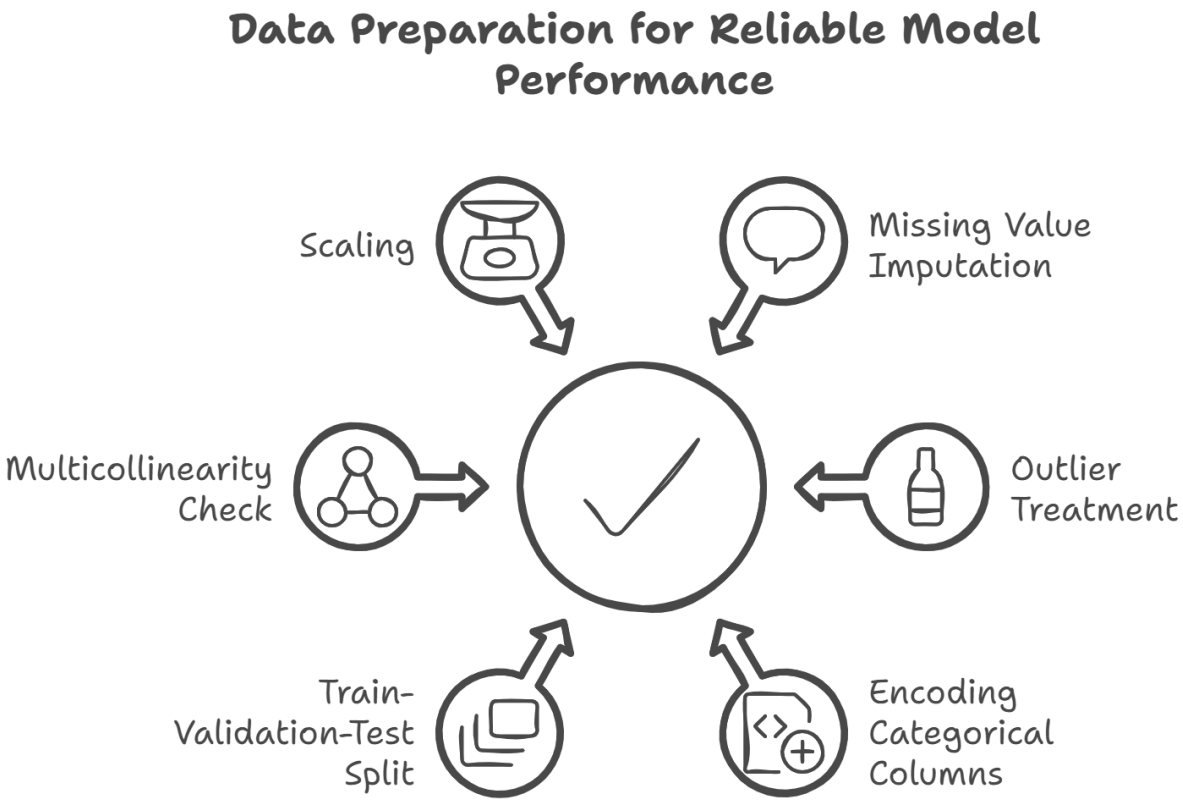
EDA – KEY INSIGHTS - CORRELATION



- **Property Value and Mortgage:** Higher property values (VALUE) are strongly correlated with larger mortgage dues (MORTDUE), reflecting applicants' financial capacity.
- **Loan Amount and Property:** Larger loan amounts are moderately linked to higher property values and mortgage dues, indicating more valuable properties among borrowers.
- **Debt and Credit History:** Longer credit histories (CLAGE) are associated with lower debt-to-income ratios (DEBTINC), suggesting better financial stability.

DATA PRE-PROCESSING & FEATURE ENGINEERING

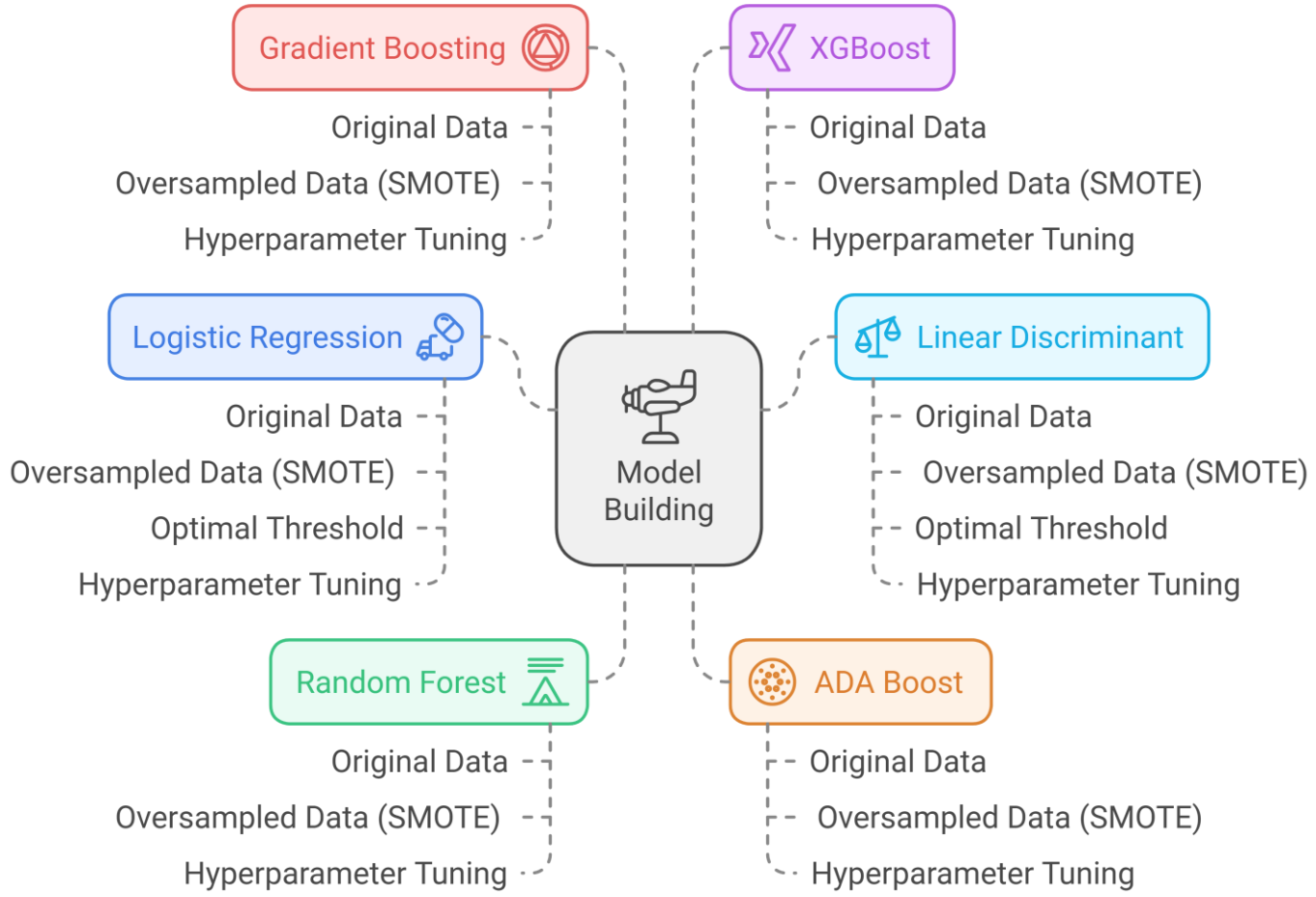
- **Missing Values:** numerical values were imputed using KNN for numerical and most common value for categorical.
- **Outliers:** were capped to prevent skewed results, columns with discrete values were exempted.
- **One-hot Encoding:** was applied to convert categorical data into binary format.
- **StandardScaler:** was used to scale numerical data, ensuring consistency for models sensitive to feature size like Logistic Regression and LDA.



MODEL BUILDING & EVALUATION

In the model building approach, we develop linear models (Logistic Regression, LDA) for interpretability and ensemble methods (Random Forest, Gradient Boost, XGBoost) for accuracy. Hyperparameter tuning is applied to optimize performance and ensure robust predictions.

MODEL BUILDING APPROACH & METHODOLOGY

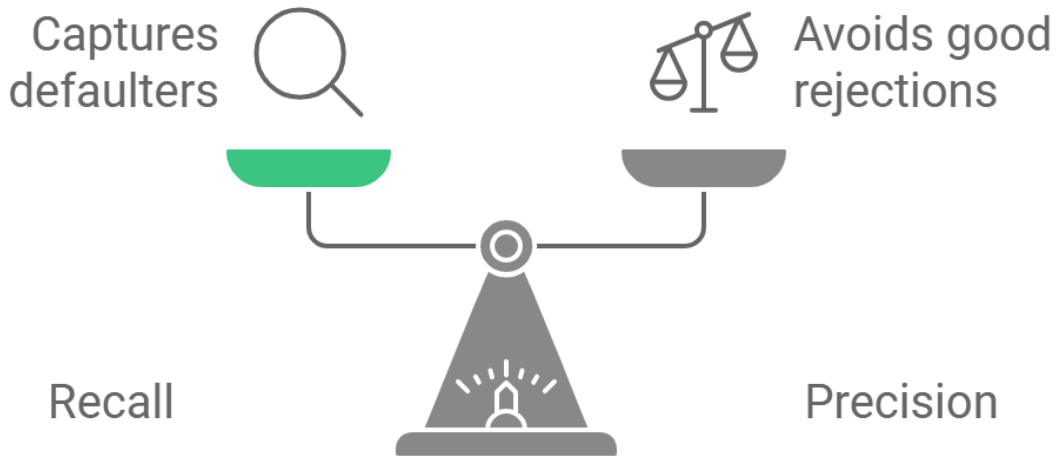


Machine Learning: We explored linear (Logistic Reg., LDA) and ensemble models (Random Forest, Gradient Boosting, XGBoost), testing them on original and oversampled datasets to handle class imbalance.

Tuning: Hyperparameter tuning was applied to improve model performance and ensure robustness, focusing on recall without overfitting.

The best model, **Random Forest built on Original data**, demonstrated high recall (default prediction) on both validation and test datasets, meeting the project objective of accurately identifying defaulters.

MODEL EVALUATION – RECALL & PRECISION



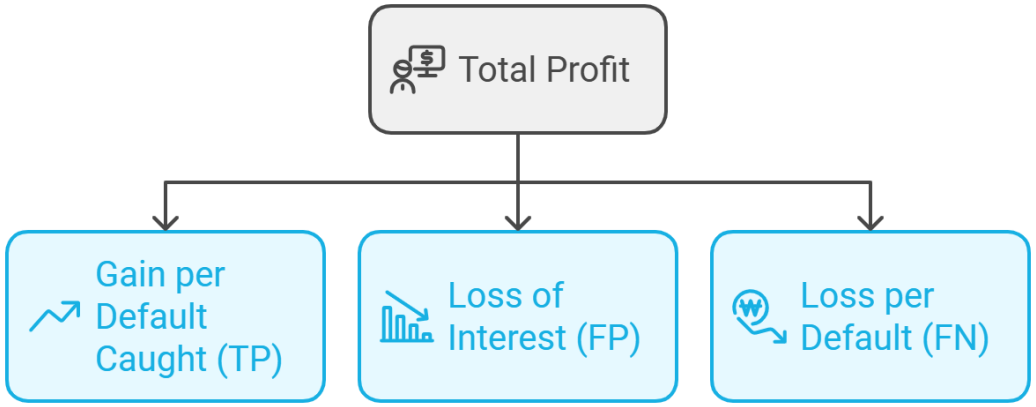
Why Recall Matters?

- Recall refers to the model's ability to capture high-risk borrowers (those likely to default). It's crucial for banks because missing a defaulter means losing the entire loan principal.
- The Random Forest model has a Recall of 81% on unseen data, meaning it identifies 8 out of 10 defaulters, ensuring fewer bad loans.

Balanced Precision:

- The model ensures we don't over-reject good borrowers. With a Precision rate of 35%, the model finds a middle ground by avoiding the rejection of good borrowers while still capturing defaulters.

MODEL EVALUATION – PROFIT CALCULATION



Profit = **TP** (Gain per default caught) *less* **FP** (Loss of interest) *less* **FN** (Loss per default)

Profitability:

- The model delivers the highest test-set profit (**$779K**), which represents the savings from avoiding bad loans and maximizing approved loans from reliable customers.
- Identifying more defaulters, while minimizing misclassification of non-defaulters, ensure protection against loss due to default and lost opportunity to earn interest.

Key Terms:

- Gain (True Positive): $19K.
- Loss of Interest (False Positive): 4% pa earning x 10 years.
- Loss per default (False Negative): $19K + 1% Cost of Loan servicing.
- Refer to **Appendix** for detailed working on Profit calculation.

MODEL COMPARISON & SELECTION

Model comparison and selection involved evaluating algorithms based on recall and F1 score to identify the best-performing model for predicting loan defaults. The Random Forest model tuned on oversampled data emerged as the top choice, demonstrating a strong balance between performance metrics and generalization capabilities.

TOP 5 BEST PERFORMING MODELS

The top 5 models based on **high Recall** (to capture defaulters effectively), **balanced Precision** (to minimize false positives), **Profit**, and **minimal overfitting** between training and validation sets.

1. **Random Forest**: Chosen for its **high Recall (85%)**, **low overfitting**, and substantial **Profit ($859,040)**. It balances performance and generalization effectively.
2. **Random Forest Tuned**: Like the base Random Forest model, this tuned version delivers consistent **Recall (85%)** with **low overfitting** and slightly improved **Profit ($5,142,778)**, making it a strong candidate.
3. **XGBoost Tuned**: Offers **high Recall (78%)** and reasonable **Profit ($537,428)**. This model balances profitability and defaulter identification effectively.
4. **Logistic Regression Scaled**: Provides **consistent performance** with **low overfitting (5% Recall difference)** and **Recall (76%)**. Though its **Profit ($206,718)** is lower, it remains stable and reliable.
5. **Logistic Regression Tuned**: Similar to the scaled version, this model shows **low overfitting** and **Recall (76%)** with stable **Profit ($202,679)**, making it another reliable choice for deployment.

Next Steps: Test Random Forest model on the Test dataset to validate performance, focusing on Recall, Precision, and Profit before proposing it for production use.

Top 5 Models Comparison

	Accuracy	Recall	Precision	F1	Profit
1. Random Forest model built on Original Data					
Training	**70%**	**89%**	**39%**	**54%**	**$5,035K**
Validation	**66%**	**85%**	**35%**	**50%**	**$859K**
2. Random Forest model on Original Data with Hyperparameter Tuning					
Training	70%	89%	39%	54%	$5,143K
Validation	66%	85%	35%	50%	$859K
3. XGBoost Model on Original Data with Hyperparameter Tuning					
Training	74%	91%	43%	58%	$6,134K
Validation	68%	78%	36%	49%	$537K
4. Logistic Regression Model on Scaled Dataset					
Training	65%	80%	34%	48%	$2,162K
Validation	63%	76%	32%	45%	$207K
5. Logistic Regression Model (Scaled) using Hyperparameter Tuning					
Training	65%	81%	34%	48%	$2,230K
Validation	63%	76%	32%	45%	$203K

FINAL MODEL SELECTION

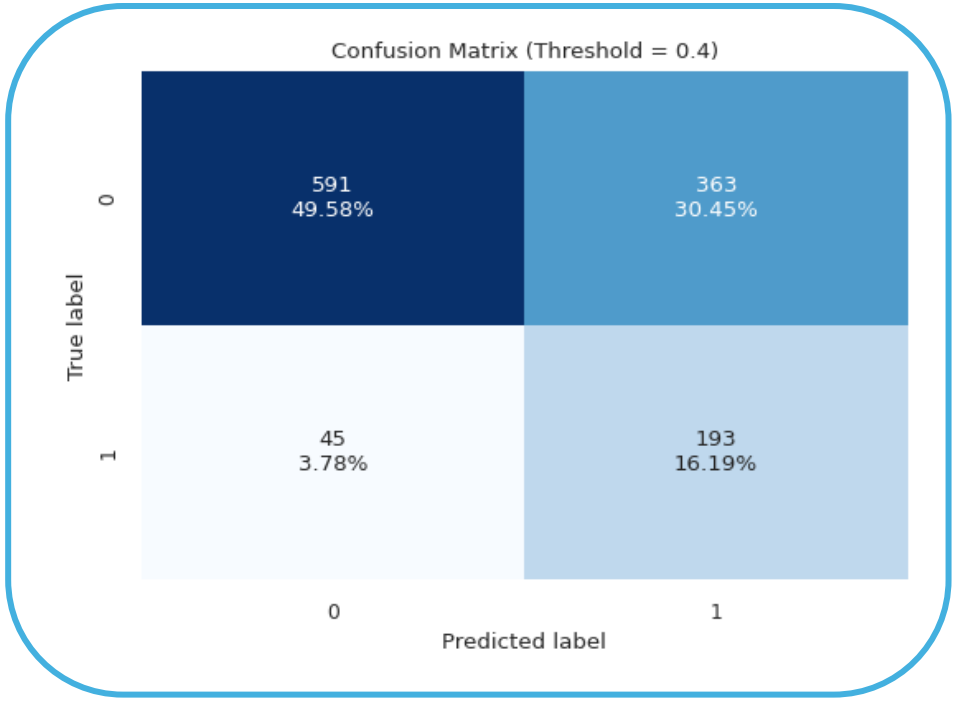
Random Forest Model Built on Original Data as Final Model

- **Best Recall on Test Set**: The **RF model** achieves the highest **Recall (81.09%)** on the test dataset, making it the most effective at capturing defaulters compared to the other models. This aligns with the objective of maximizing defaulter identification.
- **Balanced Precision and F1 Score**: While **Precision (34.71%)** is moderate, the **F1 score (48.61%)** strikes a balance between Recall and Precision, ensuring the model is practical for real-world application without excessive false positives.
- **High Profit on Test Set**: The Random Forest model generates the highest **Profit ($779,451.23)** on the test set, outperforming all other models in terms of financial return, which is a crucial criterion for the bank's decision-making.
- **Low Overfitting**: The model shows minimal overfitting between the training, validation, and test sets, with consistent Recall and Precision across the datasets. This confirms that the model generalizes well to unseen data.
- **Reliable Performance**: Compared to the tuned and other models, the **Random Forest model built on original data** maintains robust performance across all key metrics (Recall, Precision, Profit), making it the most suitable for deployment in production.

Conclusion: Given its high Recall, balanced Precision, strong financial performance, and consistent generalization across datasets, the Random Forest model built on original data is the ideal candidate for final deployment.

Random Forest Model on Original Data

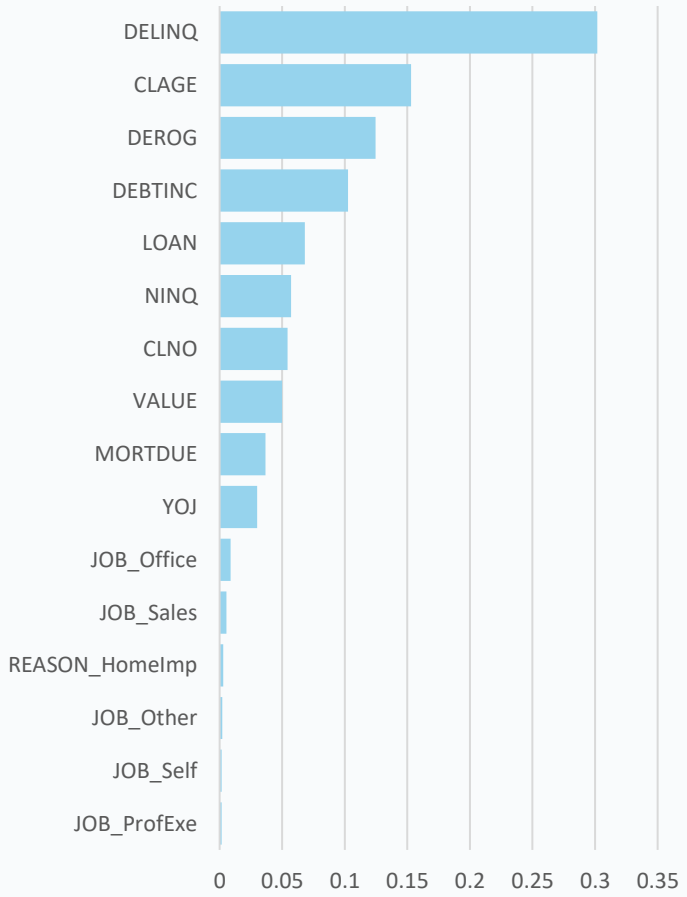
	Accuracy	Recall	Precision	F1	Profit
Training	70%	89%	39%	54%	$5,035K
Validation	66%	85%	35%	50%	$859K
Test	66%	81%	35%	49%	$779K



FEATURE IMPORTANCE

Following features influence loan default outcome:

- **DELINQ (Delinquent Credit Lines)** is the most important feature, contributing **30.17%** to the model's decision-making process. This suggests that past delinquency is a critical indicator of loan default risk.
- **CLAGE (Age of Credit Line)** and **DEROG (Derogatory Reports)** are also highly influential, with **15.29%** and **12.45%** importance, respectively. These events are strong predictors of a customer's ability to manage their loans.
- **DEBTINC (Debt-to-Income Ratio)** holds **10.26%** importance, highlighting the relevance of a borrower's debt burden relative to their income when predicting loan default.
- **LOAN Amount** contributes **6.80%** to the model, indicating that the size of the loan also plays a role in assessing the risk of default, though less so than credit history-related features.
- Features related to **job roles** (e.g., **JOB_Office**, **JOB_Sales**) and **loan reason** (e.g., **REASON_HomeImp**) have relatively low importance, suggesting that employment type and loan purpose are less critical in predicting loan defaults compared to credit history and financial metrics.



Conclusion: This feature importance analysis provides valuable insights for refining credit scoring models and making informed decisions in the loan approval process.

KEY INSIGHTS AND RECOMMENDATIONS

Key insights indicate that the Random Forest model tuned on oversampled data effectively predicts loan defaults, balancing recall and precision. Continuous refinement and exploring additional algorithms are recommended for improved performance and adaptability.

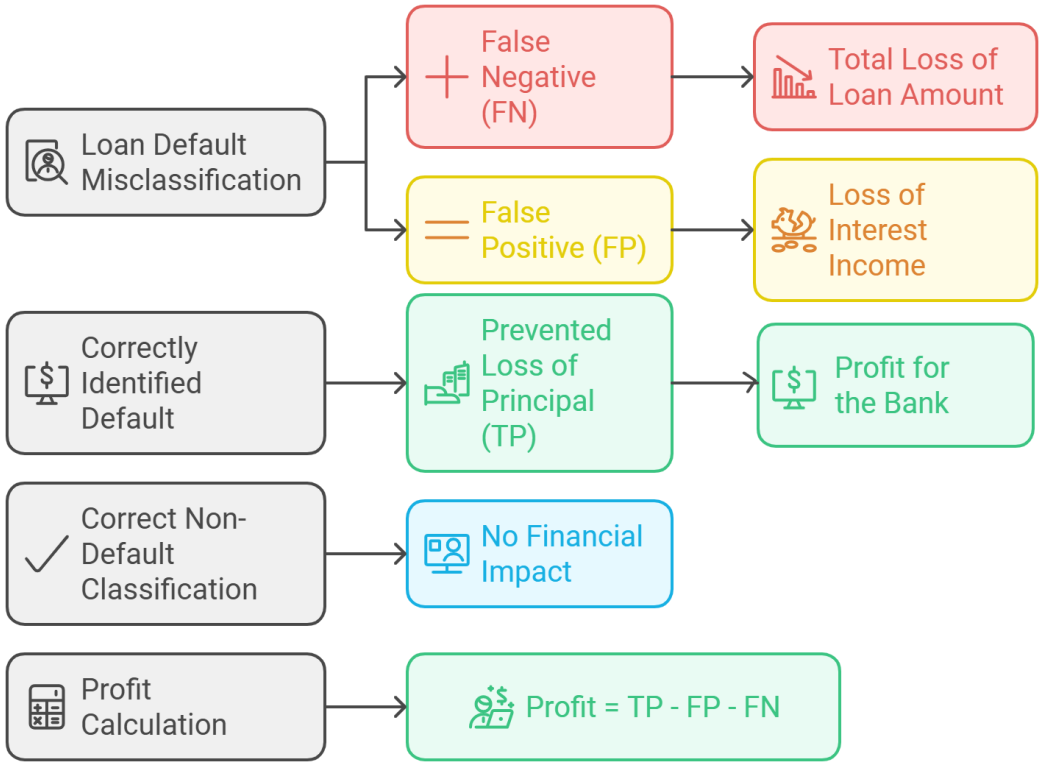
KEY INSIGHTS

- **Objective Alignment:** The machine learning model effectively predicts loan defaults, automating the bank's decision-making process and reducing bias in loan approvals.
- **Top Features:** Key features like DELINQ, delinquent credit lines, (30.17% importance), CLAGE, age of credit lines, (15.29%), and DEROG, number of derogatory reports, (12.45%) are the most influential in predicting loan defaults, emphasizing the borrower's credit history.
- **Model Performance:** Random Forest built on original data showed the best overall performance, with high Recall (81%) and the highest profit ($779K), indicating its effectiveness in identifying defaulters while ensuring profitability.
- **Profit Calculation:** The profit model aligns well with the bank's objectives by calculating savings from correctly predicted defaulters (True Positives) and potential lost revenue due to misclassified non-defaulters (False Positives).
- **EDA Findings:** Higher loan amounts, mortgage dues, and debt-to-income ratios are linked to higher default risk, highlighting key risk factors.
- **Balanced Precision:** The model maintains moderate Precision (34.71%) with a balanced F1 score, indicating the model captures defaulters well while managing a reasonable level of false positives.
- **Interpretability:** Features like loan amount, debt-to-income ratio, and derogatory reports provide clear insights into default risks, ensuring the model remains interpretable.
- **Low Overfitting:** The selected model shows consistent performance across training, validation, and test sets, confirming good generalization to unseen data.
- **Loan Default Impact:** The impact of loan defaults on profitability is significant, making the prediction of defaults critical for financial stability.
- **Business Objective Alignment:** The selected model supports the bank's goals of reducing non-performing assets (NPA) while complying with regulatory standards like ECOA.

BUSINESS RECOMMENDATIONS

- **Deploy Random Forest Model:** Deploy the Random Forest model built on original data for production. It strikes the best balance between high Recall, profitability, and low overfitting, ensuring effective loan default prediction.
- **Focus on Top Features:** Use key features like delinquent credit lines, age of credit lines, and derogatory reports to refine loan approval criteria and mitigate default risk.
- **Monitor Model Performance:** Regularly track model performance in production, particularly Recall and Precision, to ensure consistent default prediction and minimize loan losses.
- **Profit Calculation Review:** Continuously review and refine the profit calculation to ensure it reflects current interest rates, loan servicing costs, and default recovery strategies.
- **Improve Loan Risk Assessment:** Integrate the model into the bank's loan underwriting process, automating decisions and reducing reliance on manual, error-prone evaluations.
- **Data Refresh:** Regularly update the model with new data to ensure accuracy and adaptability to changing market conditions and borrower profiles.
- **Compliance and Interpretability:** Ensure the model remains interpretable and transparent, in line with regulatory guidelines like the Equal Credit Opportunity Act (ECOA).
- **Test on New Data:** Further validate the model on additional datasets to ensure robust generalization before full deployment.
- **Enhance Client Engagement:** Use insights from the model to offer better financial advice to clients, helping them improve creditworthiness and avoid default.
- **Refine Risk Mitigation Strategy:** Incorporate the model's results into broader risk mitigation strategies, enabling the bank to set more informed loan approval policies.

APPENDIX: PROFIT CALCULATION



- **Loan Default (False Negatives - FN):** If a defaulter is misclassified as a non-defaulter, the bank loses the entire loan amount.
  Example: *A $100,000 loan default leads to a $100,000 loss.*
- **Missed Opportunity (False Positives - FP):** When a non-defaulter is wrongly classified as a defaulter, the bank loses interest income, not the principal.
  Example: *A $100,000 loan at 5% interest over 10 years* ***less*** *1% loan servicing cost, nett 4%, leads to $62,889 in lost interest.*
- **Correct Defaults (True Positives - TP):** Correctly identifying defaulters saves the bank from losing the loan principal.
  Example: *Each True Positive saves $100,000.*
- **Correct Non-Defaults (True Negatives - TN):** Correct classification of non-defaulters has no direct financial impact.
- **Profit Calculation Formula:**
  Profit = TP (Gain per default caught) - FP (Loss of interest) - FN (Loss per default)

Key Assumptions:

- Loan default results in loss of total amount, can be fine-tuned.
- Average loan tenor is 10 years (source: MoneyGeek.com)
- Interest rate assumed at 5% (Source: Finance Smarter).
- Bank incurs 1% annual loan servicing cost (Source: Finance Smarter).

Final Presentation

Loan Default Prediction

THANK YOU

TEAM 4 – SYNTEGRITY

MDS | CAPSTONE PROJECT