MDS | CAPSTONE PROJECT

# LOAN DEFAULT PREDICTION

## Milestone 2

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 04
- Business Problem Overview 05
- EDA Summary 06
- Data Preprocessing Summary 07
- Model Building 08
- Model Comparison & Selection 16
- Key Insights and Recommendations 19

## Milestone 2

Loan Default Prediction

# EXECUTIVE SUMMARY

We will be developing machine learning models to predict loan defaults, improving loan approval accuracy through data preprocessing, analysis, and model building, focusing on performance and interpretability.

**Milestone 1:**

- **Problem Statement:** Automate loan approval processes to reduce human error and bias by predicting loan defaults using machine learning.
- **Solution Approach:** Conducted exploratory data analysis (EDA), treated missing values and outliers, encoded categorical data, and scaled numerical features.
- **Data Report:** Analyzed the dataset's structure, missing values, and key features, ensuring data suitability for predictive modeling.
- **Exploratory Data Analysis (EDA):** Identified key patterns, outliers, and relationships between features and target variables.
- **Data Preprocessing:** Handled missing data, outliers, and applied one-hot encoding and scaling for model readiness.

**Milestone 2:**

- **Model Building:** Developed linear and ensemble models to predict loan defaults, focusing on interpretability and accuracy.
- **Performance Evaluation:** Assessed models using metrics like accuracy, precision, recall, and AUC-ROC to gauge effectiveness.
- **Model Selection:** Chose the best model based on balanced performance and interpretability, aligned with project objectives.
- **Feature Importance Analysis:** Revealed significant predictors guiding the bank's focus on critical factors during loan approvals.
- **Next Steps:** Continuous refinement of the chosen model and exploring additional algorithms to improve predictive accuracy and adaptability.

# PROBLEM OVERVIEW

**Problem Overview:**

The current process for home equity loan approval is time-consuming and prone to human error and biases, as it relies on manual review of applicants' creditworthiness. To streamline decision-making, this project aims to develop a machine learning model to accurately predict loan defaults, ensuring fairness by adhering to the **Equal Credit Opportunity Act** (ECOA) guidelines. The ECOA mandates that credit-scoring models be statistically sound, non-discriminatory, and empirically derived, ensuring equal access to credit for all applicants. This will reduce reliance on human judgment and eliminate potential biases in the loan approval process.

**Objectives:**

- Automate the home equity loan approval process by developing a machine learning model to predict loan defaults.
- Provide interpretable models that justify rejections or approvals of loan applications.

**Need of the Present Study:**

- Manual loan approvals are subject to human bias and inefficiency. By adopting a data-driven approach, banks can reduce errors, process applications faster, and improve the accuracy of predictions.
- Adherence to **ECOA** guidelines.

**Business/Social Opportunity:**

- Improved decision-making in loan approvals increases profitability by reducing non-performing assets (NPAs) and defaults.
- Socially, it ensures a fairer, unbiased loan approval system, complying with guidelines such as the Equal Credit Opportunity Act, benefiting both customers and the financial institution.

# EDA – KEY INSIGHTS

The analysis highlighted significant patterns in both numerical and categorical features, offering a clearer understanding of loan default risk.

**Numerical Features:**

- Higher loan amounts, mortgage dues, and debt-to-income ratios are linked to higher default risk.
- Shorter job tenure and a higher number of derogatory reports or delinquencies increase default likelihood.
- Most features have mild skewness, with a few notable outliers, especially in loan and debt-related variables.

**Categorical Features:**

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

**Feature Correlation:**

- There is a strong correlation between property value and mortgage dues, suggesting a relationship between higher-value properties and larger mortgages.
- No major multicollinearity issues were found, though MORTDUE and VALUE exhibit a moderate correlation.

**Conclusion:**

Key predictors for loan default risk are larger loan amounts, higher debt-to-income ratios, and poor credit history, while categorical variables like loan purpose and job type also provide valuable insights into default risk. The data is largely clean and ready for modelling, with minimal concerns around multicollinearity.

# DATA PRE-PROCESSING - SUMMARY

**Missing Value Imputation:**

- We found a few missing values, so we filled them in before building our models. For numbers, we used KNN, which looks at similar data points. For categories, we used the most common value to fill in gaps.

**Outlier Treatment:**

- We addressed outliers to prevent extreme values from skewing results. Using the IQR method, we capped values outside of a defined range. Some columns, like DEROG, DELINQ, and NINQ, weren't treated due to their unique characteristics.

**Encoding Categorical Columns:**

- We transformed categorical data into binary format using One-hot Encoding, which is effective because the categories don't have a clear order.

**Train-Validation-Test Split:**

- We divided our dataset into training, validation, and testing sets to ensure accurate model performance. We kept 80% for training and 20% for testing, with a portion of the training set reserved for validation.

**Multicollinearity Check:**

- We checked for multicollinearity in linear models like Logistic Regression and LDA. Although some features had higher values, they were still acceptable.

**Scaling:**

- Scaling is crucial for models sensitive to feature size, like Logistic Regression and LDA. We applied StandardScaler to ensure consistent feature scaling across our training and test sets. This approach helps us prepare the data effectively, ensuring our models perform reliably and meet project goals.

# MODEL BUILDING

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

**Linear Models**

| | |
|---|---|
| Logistic Regression Models | 10 |
| Linear Discriminant Analysis Models | 11 |

**Ensemble Techniques**

| | |
|---|---|
| Random Forest Models | 12 |
| ADA Boost Models | 13 |
| Gradient Boosting Models | 14 |
| XGBoost Models | 15 |

Milestone 2

Loan Default Prediction

# LOGISTIC REGRESSION MODELS

We implemented several Logistic Regression models to improve decision-making processes:

1. **Original Data:** Built on scaled dataset, model performed poorly, especially identifying defaulters.
2. **Oversampled Data:** Showed better results by identifying defaulters but fell short in identifying non-defaulters. Model was also overfitting, creating a reasonable doubt on its accuracy on future data.
3. **Optimal Threshold:** Same as Oversampled Data.
4. **Hyperparameter Tuning:** No improvements over Oversampled Data.

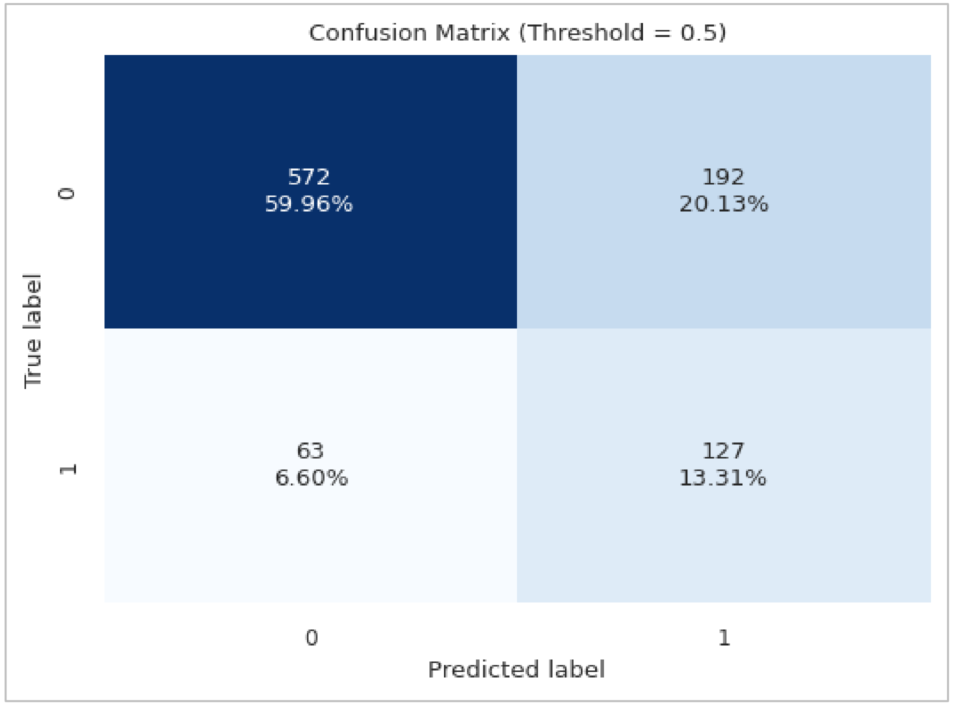
**Best Performing Model:**

The **Logistic Regression model on oversampled data** achieved around 70% recall during training and 67% in validation, with low validation precision (40%), indicating potential misclassification of non-defaulters.

**Next Steps:** We will proceed with building the Linear Discriminant Analysis (LDA) model to further enhance our predictive capabilities in loan default prediction.

Best Model: Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 73% | 70% | 74% | 72% |
| Validation | 73% | 67% | 40% | 50% |



# LINEAR DISCRIMINANT ANALYSIS (LDA) MODELS

We also implemented several LDA models, to improve our ability to predict loan defaults without biases:

1. **Original Data:** Built on scaled dataset, model exhibited moderate accuracy, however similar to previous models struggled to identify defaulters.
2. **Oversampled Data:** Demonstrated improved default prediction but misclassified significant numbers of non-defaulters. This model gave better performance as compared to other LDA models.
3. **Optimal Threshold:** Same performance as of oversampled model after changing threshold value to 0.1.
4. **Hyperparameter Tuning:** Tuned model showed no improvements over Oversampled Data.

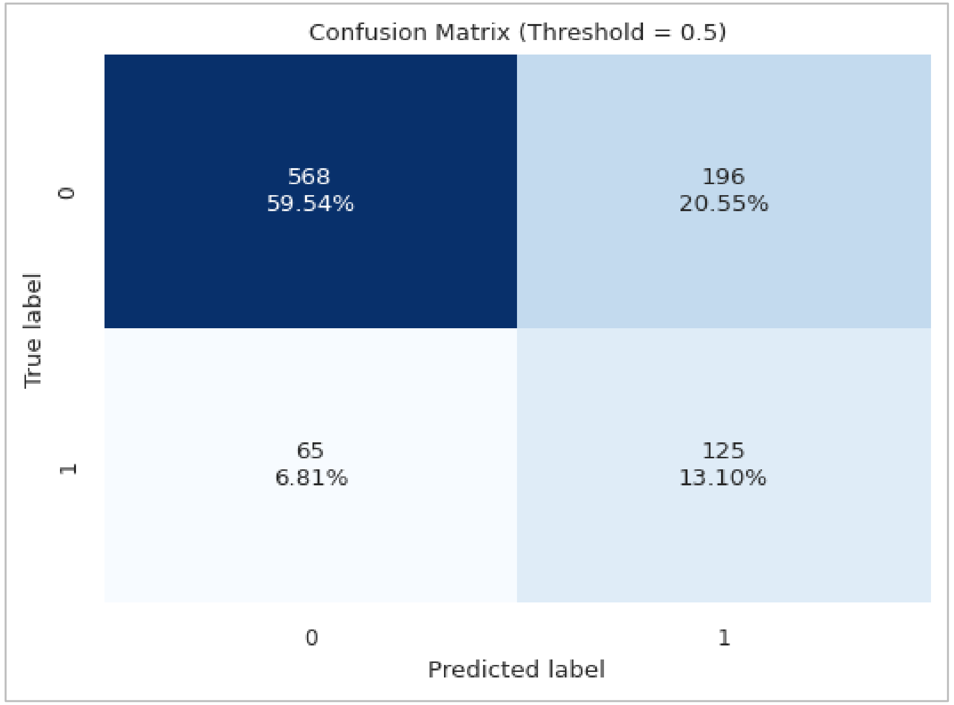
**Best Performing Model:**

The **LDA model on oversampled data** model shows strong Recall in identifying defaulters (69% training, 66% validation) but struggles with precision, resulting in a low F1 score of 49%. Further refinement is needed.

**Next Steps:** We will proceed to build Random Forest models to further refine our predictive capabilities in loan default prediction.

## Best Model - Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 72% | 69% | 73% | 71% |
| Validation | 73% | 66% | 39% | 49% |



# RANDOM FOREST MODELS

We implemented following three Random Forest models to enhance loan default predictions through diverse training approaches:

1. **Original Data:** Achieved high training accuracy, but recall (59.47%) indicates challenges in identifying defaulters.
2. **Oversampled Data:** Improved recall to 67.90%, enhancing defaulter detection, yet precision concerns remain. This model has performed best and effectively mitigates the impact of class imbalance, enhancing its predictive reliability.
3. **Hyperparameter Tuning:** Maintained high training accuracy, but validation recall (67.37%) suggests potential overfitting and generalization issues.

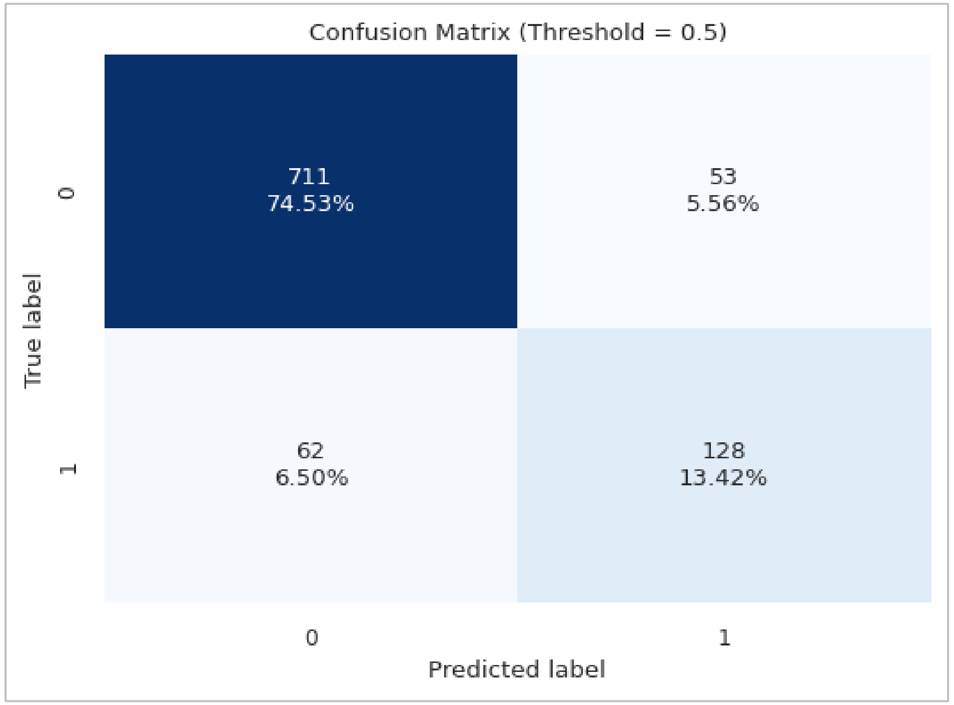
**Best Performing Model:**

The **Random Forest on Tuned Oversampled data** model demonstrates acceptable recall of 67% and precision of 71% on validation data, indicating effective detection of potential defaulters. However, overfitting on training data raises concerns about its reliability for future predictions. Further tuning or alternative modeling approaches needed.

**Next Steps:** We will proceed to build Random Forest models to further refine our predictive capabilities in loan default prediction.

## Best Model - Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 98% | 98% | 99% | 98% |
| Validation | 88% | 67% | 71% | 69% |



# ADA BOOST MODELS

Following three ADA Boost models were developed to enhance loan default predictions:

1. **Original Data:** The model performed decently with a training accuracy of 85.95%, but its recall of 44.02% indicates limited ability to identify defaulters.
2. **Oversampled Data:** Although training accuracy dropped to 83.92%, recall improved to 83%, demonstrating better identification of potential defaulters, albeit with lower precision.
3. **Hyperparameter Tuning:** This model showed significant improvements, achieving a training accuracy of 91.71% and a recall of 88.31%, indicating stronger performance in detecting defaulters.

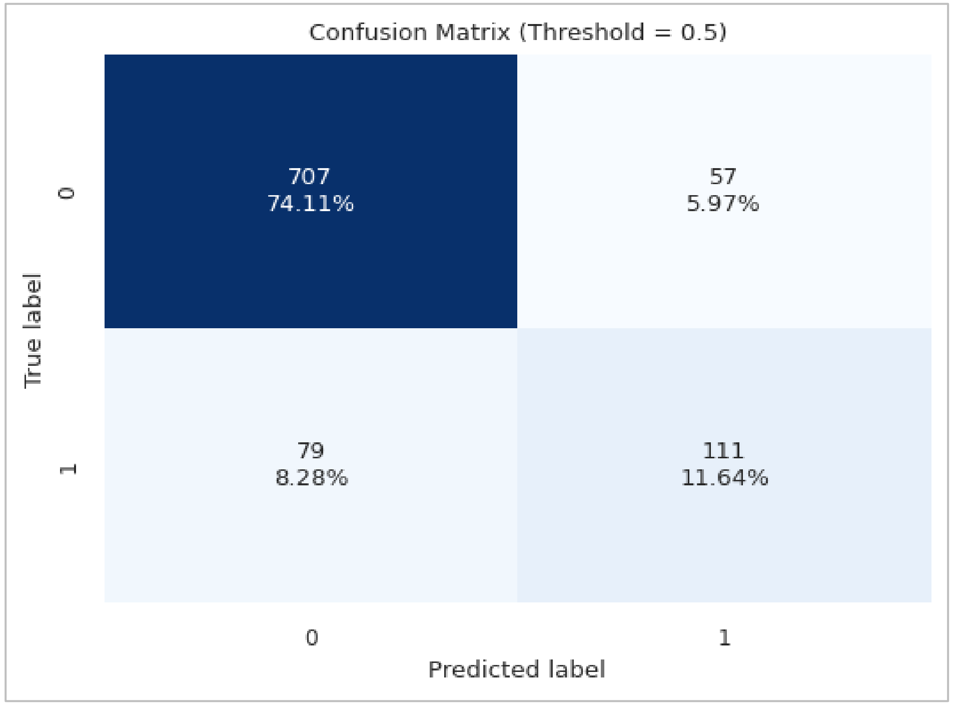
**Best Performing Model:**

The **ADA Boost model Tuned on Oversampled data** stands out, balancing high recall and precision, with a validation recall of 58.42%. However, the model also is overfitting and may not be able to predict defaulters with required efficiency.

**Next Steps:** We will proceed to build the Gradient Boost model to further enhance predictive accuracy.

## Best Model - Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 92% | 89% | 95% | 91% |
| Validation | 86% | 58% | 66% | 62% |



# GRADIENT BOOSTING MODELS

The Gradient Boosting model effectively captures complex relationships in the data but exhibits varying performance across different datasets:

1. **Original Data:** Training accuracy is at 90% with a recall of 54%, indicating a struggle in identifying defaulters. Same is true for its performance on Validation data.
2. **Oversampled Data:** Achieves high training recall of 89%, but validation shows lower performance with 60% recall.
3. **Hyperparameter Tuning:** Perfect training performance at 100%, yet validation recall is only 63%, indicating overfitting.

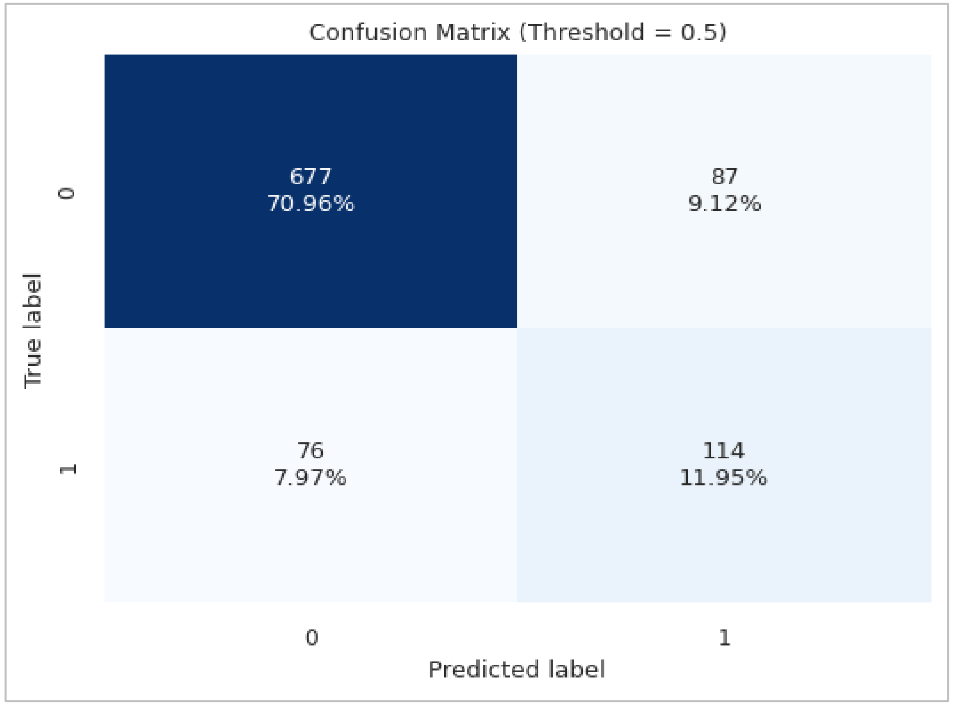
**Best Performing Model:**

The **Gradient Boosting model on Oversampled data** demonstrates strong performance, achieving a training recall of 88% and a validation recall of 60%. This indicates its effectiveness in identifying defaulters, although the drop in validation recall suggests room for improvement in generalization to unseen data.

**Next Steps:** We will proceed to build the XG Boost model to further enhance predictive accuracy.

Best Model - Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 89% | 88% | 90% | 89% |
| Validation | 83% | 60% | 57% | 74% |



# XGBOOST MODELS

The XGBoost model showcases remarkable training performance but presents significant challenges in validation, indicating overfitting concerns:

1. **Original Data:** Achieved perfect training scores, with 100% across all metrics; however, validation recall dropped to 61%, suggesting difficulty in generalizing to unseen data.
2. **Oversampled Data:** Maintained a perfect training performance while validation recall improved to 67%, indicating better identification of defaulters but still reflecting potential overfitting.
3. **Hyperparameter Tuning:** The model shows troubling signs of severe overfitting, with a significant drop to 50% accuracy in training and only 20% in validation, indicating a lack of robustness.

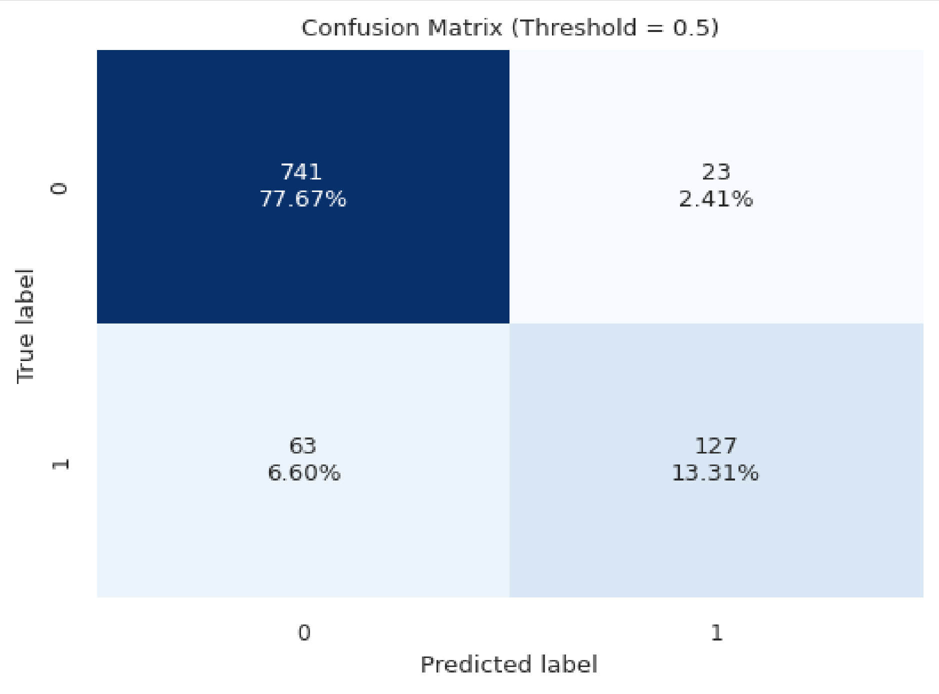
**Best Performing Model:**

**XGBoost - Resampled** achieved a good balance with a validation recall of 67%, which indicates better performance in identifying defaulters compared to the other models. It also maintained a high level of accuracy at 91% and an F1 score of 75%.

**Next Steps:** Compare all models and select an appropriate, best performing algorithm.

Best Model - Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 100% | 100% | 100% | 100% |
| Validation | 91% | 67% | 85% | 75% |



# 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.

# MODEL SELECTION

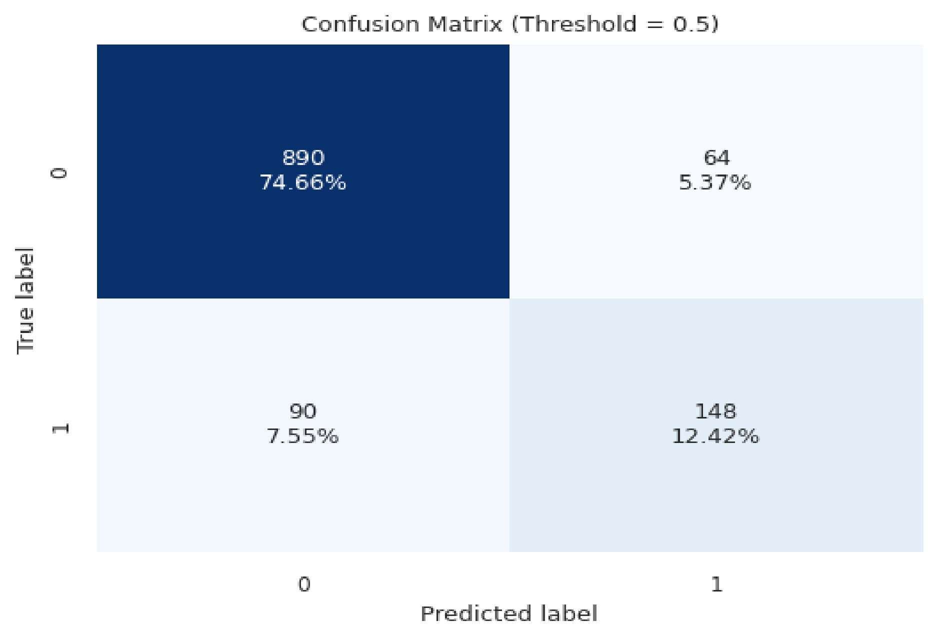
Based on the performance metrics of all 20 models developed, the **Random Forest Tuned on Oversampled Data** stands out as the most effective model for balancing recall and generalization. Here are some key reasons for this selection:

- **Recall:** This model achieved a **recall** of approximately **67.37%** on the validation set, effectively identifying a significant number of actual defaulters. This aligns with the project's objective to minimize missed defaulters.
- **F1 Score:** The **F1 score** of **69%** indicates a better balance between precision and recall compared to other models, making it a reliable choice for loan default predictions.
- **Generalization:** While **Random Forest** models showed some **overfitting** tendencies during training, the tuned version demonstrated improved validation performance, suggesting it generalizes better than many other models tested.

In conclusion, the Tuned Random Forest Model provides a favorable balance between recall and precision, ensuring effective identification of defaulters while maintaining acceptable performance on unseen data, making it the recommended choice for further application in loan default prediction scenarios.

Best Model - Stats for Data Science Team

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| Training | 98% | 98% | 99% | 98% |
| Validation | 88% | 67% | 71% | 69% |
| Test | 87% | 62% | 70% | 66% |

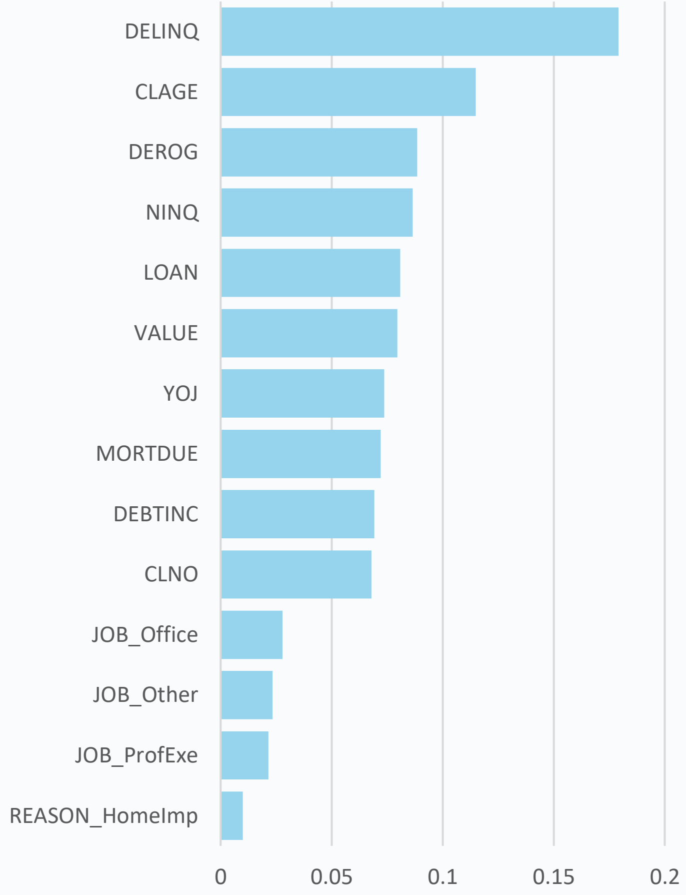


# FEATURE IMPORTANCE

Following features influence loan default outcome:

- **Top Features:** The most influential features are DELINQ and CLAGE, indicating that the number of delinquencies and the age of credit accounts are crucial for predicting loan defaults.
- **Moderate Importance:** Features like NINQ, VALUE, and DEROG also contribute significantly, suggesting that credit inquiries, property value, and derogatory marks play a role in creditworthiness.
- **Lower Importance:** Features related to job type (e.g., JOB_Office, JOB_ProfExe) and reason for the loan (e.g., REASON_HomeImp) have lesser importance, highlighting that employment type may not significantly influence default risk compared to credit history.
- **Insights for Model Improvement:** Given the importance of DELINQ and CLAGE, further analysis on credit history may enhance model performance. Addressing features with low importance could streamline the model without losing predictive power.

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

- **Model Performance:** The Random Forest model tuned on oversampled data achieved high training performance but demonstrated overfitting, indicating a need for further adjustments to improve generalization.
- **Recall Focus:** Recall metrics were prioritized throughout model development, aligning with the objective to identify potential defaulters effectively.
- **Data Imbalance:** Oversampling was effective in improving recall for minority class (defaulters) but led to precision trade-offs, highlighting the challenges of handling imbalanced datasets.
- **Feature Importance:** EDA revealed key features impacting loan default predictions, including loan amount and credit history, which should be monitored during the approval process.
- **Confusion Matrix Analysis:** Misclassifications of defaulters and non-defaulters indicated the necessity for refined thresholds and better model training strategies.
- **Model Interpretability:** Emphasized the importance of interpretability in models, especially for justifying decisions under the Equal Credit Opportunity Act.
- **ROC_AUC Insights:** The ROC_AUC score indicated fair discrimination ability; however, improvements are needed to enhance model robustness.
- **Validation vs. Training Discrepancy:** The significant performance gap between training and validation metrics signals potential overfitting and necessitates further model validation.
- **Data Pre-processing Impact:** Effective missing value imputation and outlier treatment positively influenced model accuracy, underscoring the importance of comprehensive data pre-processing.
- **Comparison of Models:** Different models, including Logistic Regression and LDA, were evaluated, with Random Forest providing the most balanced results despite overfitting concerns.

# BUSINESS RECOMMENDATIONS

- **Refine Model Selection:** Continuously evaluate and refine model parameters to improve performance metrics, focusing on enhancing recall while maintaining acceptable precision.
- **Implement Rigorous Testing:** Conduct thorough validation on unseen data to ensure model robustness before deployment in real-world scenarios.
- **Monitor Loan Applications:** Utilize model insights to develop monitoring tools for loan applications, ensuring that key features influencing default risk are regularly assessed.
- **Important Features:** It's evident from the model built that number of delinquent lines, age of oldest credit line, number of recent inquires and debt-to-income ratio appear to have significant impact on default prediction, this information is critical in default prediction.
- **Regular Model Updates:** Establish a schedule for re-evaluating and updating models as more data becomes available to maintain accuracy and effectiveness.
- **Cross-Department Collaboration:** Foster collaboration between data science teams and credit departments to align model outputs with operational realities and improve decision-making processes.
- **Training and Awareness:** Train staff on the implications of model predictions and the importance of objective criteria in loan approval to mitigate biases.
- **Consider Alternative Models:** Explore additional algorithms and ensemble methods that may enhance prediction accuracy, particularly in areas where current models underperform.
- **Feedback Loop:** Create a feedback mechanism to gather insights from the model's performance in practice, facilitating continuous learning and adaptation to evolving financial landscapes.

# Milestone 2

## Loan Default Prediction

# THANK YOU

TEAM 4 – SYNTEGRITY - DAISY | DESMOND | SODEEQ | SHAISHAV

MDS | CAPSTONE PROJECT