Live Presentation

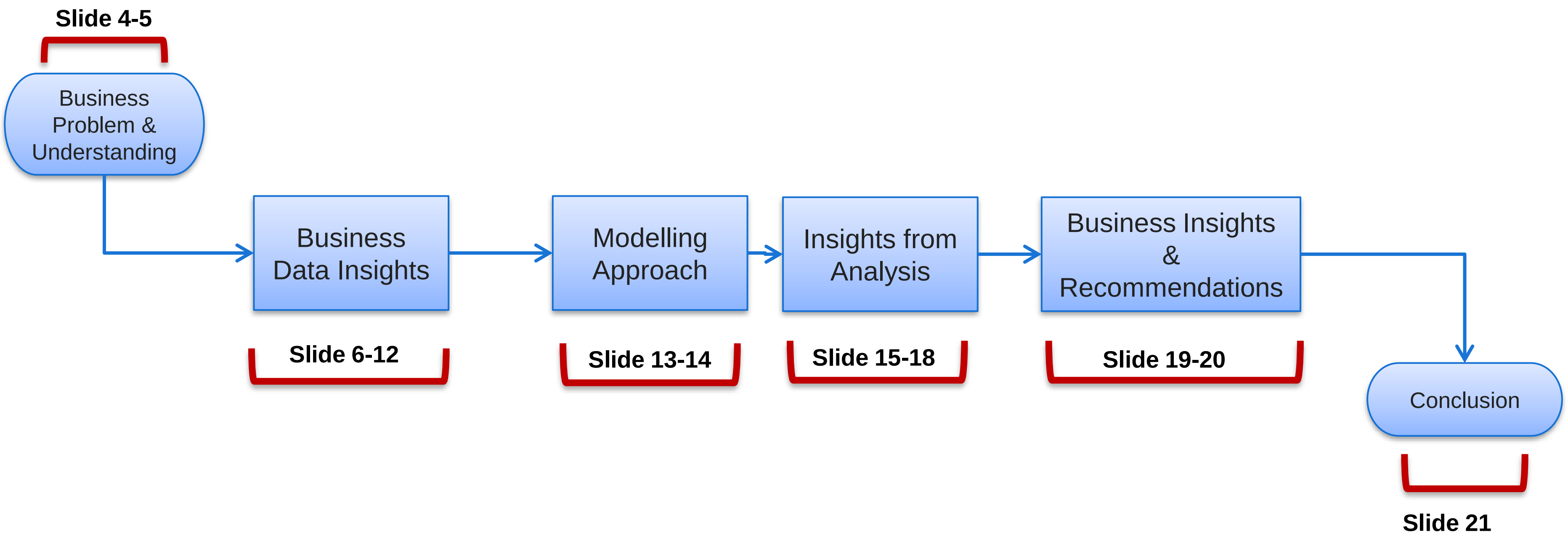
Loan Default Prediction - MDS

Date: 30th Mar 2025

Group 4 :
Hanan Ahmad Arar
Amneh Mohamad Ibrahim Ghanem
Mohamed Hassan Sharaf
Mohammad AHM Alrashed

Contents / Agenda

Project Delivery Methodology



Business Problem Understanding

Introduction

- The project focuses on improving **loan approval processes** in banks by leveraging machine learning to enhance risk assessment and reduce non-performing loans (NPLs).

- Why? Traditional manual and rule-based methods are **inefficient** and **biased**.

- The goal is to develop an **accurate**, **interpretable**, and **fair credit scoring model** that ensures **regulatory compliance** and **optimizes loan approvals**.

Business Problem Understanding

Executive Summary

Objective

- **Business Need**: Predict loan defaults to reduce financial risk and optimize lending decisions.
- **Key Question**: Can machine learning improve credit risk assessment for more efficient lending?

Key Insights from Business Data Analysis:

Loan default is significantly influenced by:

- Debt-to-Income Ratio (DEBTINC)
- Delinquency History (DELINQ)
- Credit Age (CLAGE)
- Derogatory Marks (DEROG)

Categorical factors (Job, Reason) have lower impact but were encoded for completeness.
Missing values and outliers were properly handled.

Goal: Leverage AI to enhance risk-based lending strategies & minimize defaults.

Business Data Insights

Data Sources

- Loan applications,
- credit bureau reports,
- employment history

Total Entries:

- 5,960 rows, with 13 columns (1 target variable + 12 features).

Key Variables:

- **Target Variable**: BAD (Loan Default - Yes/No)
- **Key Predictors**: DEBTINC, DELINQ, CLAGE, DEROG (More might be discovered during the study)

Business Data Insights

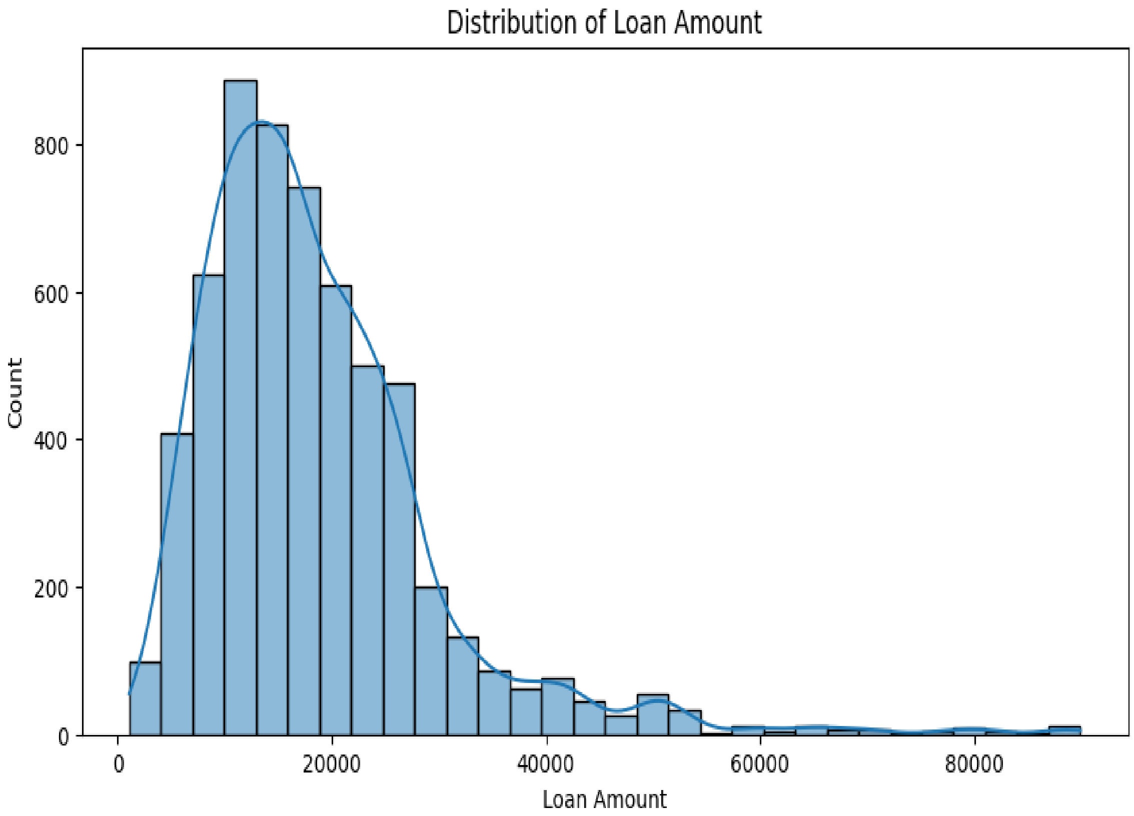
Key Findings from Data Analysis:

- Debt-to-Income Ratio (DEBTINC) is a strong predictor of loan default.
- Past credit behavior (DEROG, DELINQ) directly impacts default rates.
- Job Type (JOB) and Loan Purpose (REASON) have a smaller but notable influence.

Key Visuals for Univariate & Bivariate Analysis:

- Histogram: Distribution of Loan Amounts, Distribution of Derogatory Marks
- Scatterplot: Loan amount versus Mortgage Due
- Heatmap of Feature Correlations

Business Data Insights



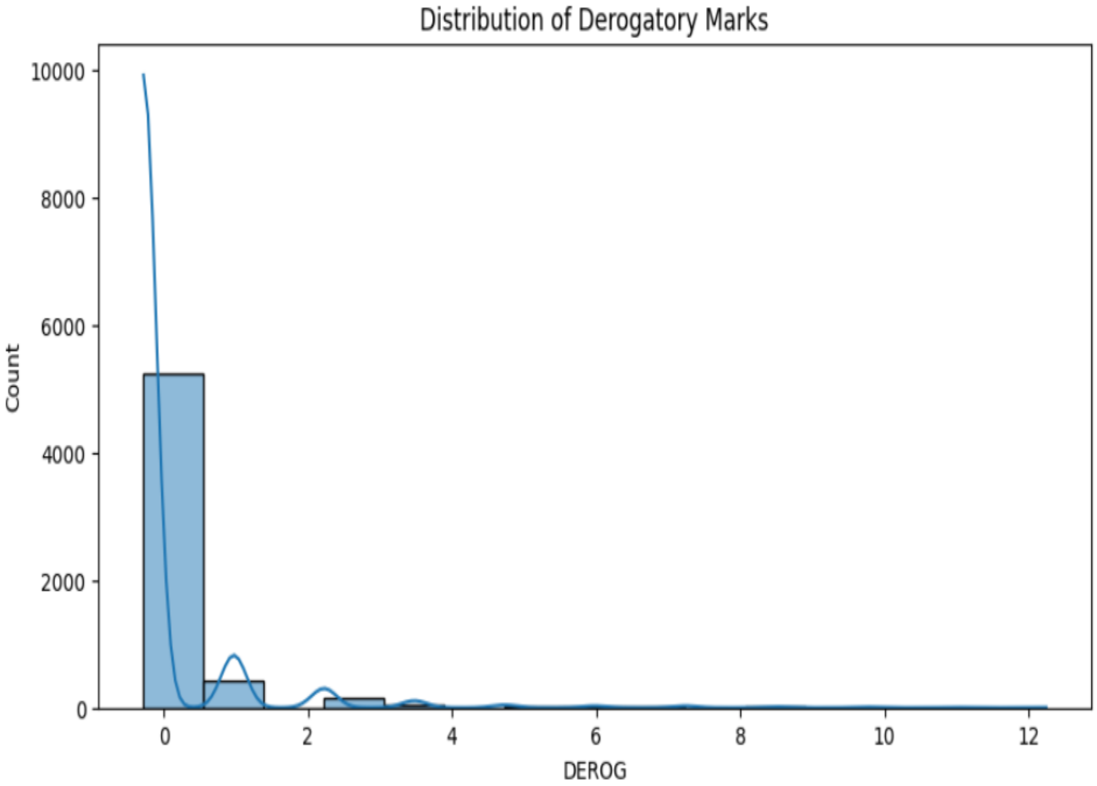
Univariate Analysis

- The **peak** occurs around $10,000–$20,000, indicating that **most approved loans** fall within this range.

Business Insight:

- CMO: Create premium loan offers for high-value borrowers ($40,000+), ensuring adequate risk assessment.
- COO: Borrowers with larger loan amounts may be at higher risk of default, as larger financial obligations increase the chance of repayment issues.

Data Pre-processing & EDA



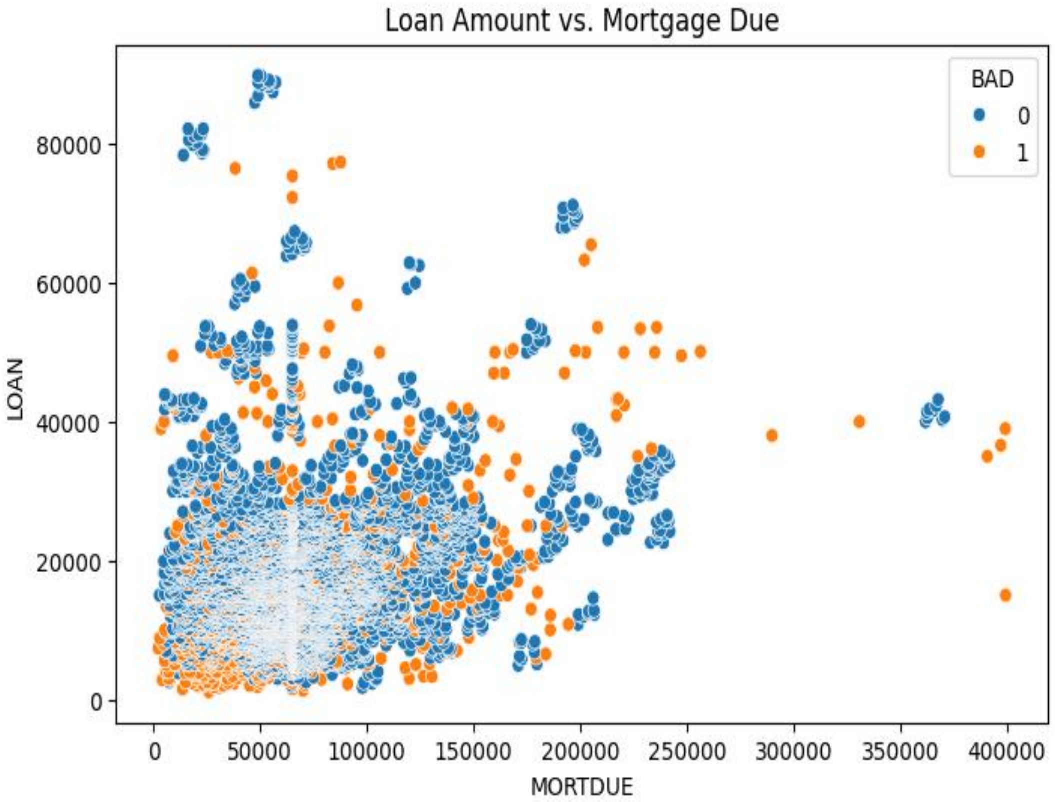
Univariate Analysis

- The distribution of derogatory marks (DEROG) is **highly skewed** to the right.
- **Most borrowers** have **zero derogatory marks**, indicating a clean credit history.

Business Insight:

- CMO: Most borrowers have good credit histories, meaning banks can offer competitive interest rates to attract low-risk customers.
- COO: Develop stricter loan approval policies for customers with multiple derogatory marks.

Business Data Insights



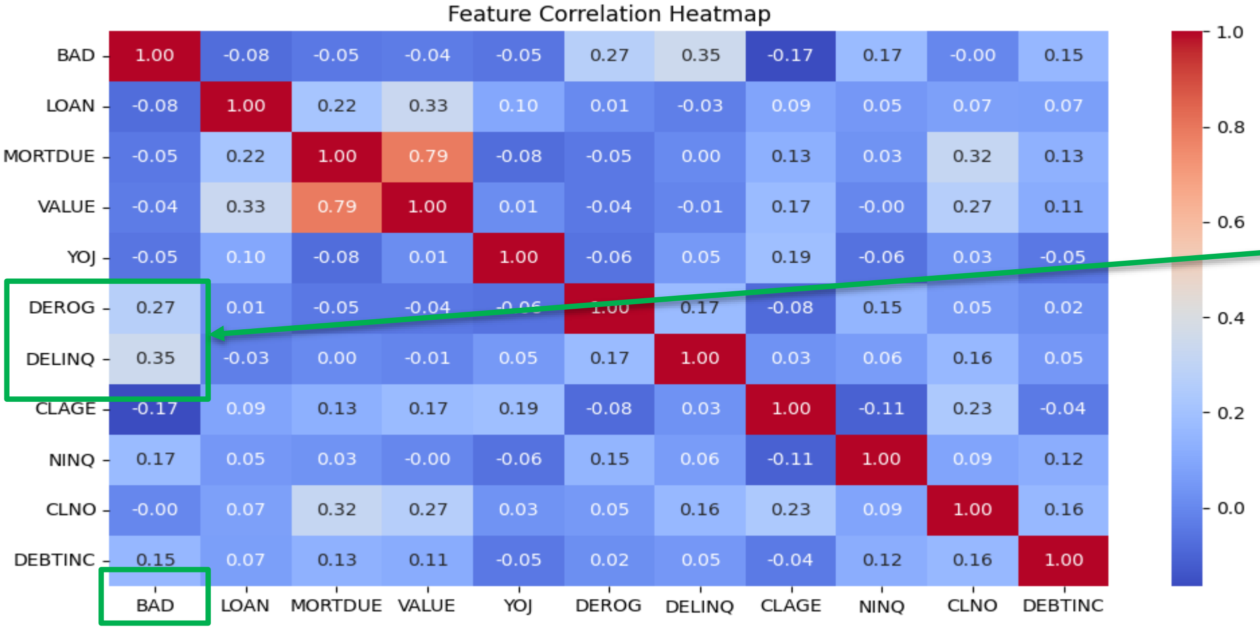
Bivariate Analysis

- **Most loans** are **concentrated in the lower ranges** of Loan Amount (< $30,000) and Mortgage Due (< $100,000).
- A **positive correlation** exists between LOAN and MORTDUE, meaning that higher mortgage dues generally correspond to higher loan amounts.

Business Insight:

- CMO: Prioritize applicants with low MORTDUE for standard loan approvals, as they are lower-risk borrowers.
- COO: Introduce additional credit checks for borrowers applying for high LOAN amounts (> $40,000) and high mortgage obligations (> $150,000).

Business Data Insights



Multivariate Analysis – Feature Correlation Heatmap

- Loan Default (BAD) has the **strongest correlation** with DELINQ (0.35) and DEROG (0.27), confirming that past delinquencies and derogatory marks are the **biggest indicators** of loan default risk.

Business Data Insights

Statistical Tests & Feature Selection:

ANOVA Test Results

“We used ANOVA to see if the **average delinquency and debt-to-income ratios** differ significantly between people who defaulted and those who didn’t. High F-values and near-zero p-values told us these features are **critical** and should go into the model.”

- DELINQ ($F = 812.95$, $p < 10^{-16}$) → Strong predictor of loan default.
- DEBTINC ($F = 145.78$, $p < 10^{-33}$) → Debt-to-Income ratio plays a key role in default behavior.

Conclusion: These features should be prioritized in predictive models.

Chi-Square Test Results

Chi-Square helps us understand if features like **employment type or loan reason** are **related** to defaulting. We found that people’s jobs strongly **correlate** with default risk.”

- JOB ($\chi^2 = 73.82$, $p < 10^{-14}$) → Employment type impacts loan default probability.
- REASON ($\chi^2 = 8.19$, $p = 0.0042$) → Loan purpose has a minor but notable influence.

Conclusion: Useful for risk segmentation strategies in lending policies.

Modelling Approach

Machine Learning Approach

- Baseline Model: Logistic Regression (to establish performance benchmark).
- Advanced Models: Random Forest, Gradient Boosting (to capture complex relationships).

Feature Selection Strategy

- Used ANOVA for numerical features.
- Used Chi-Square for categorical features.

Data Preprocessing

- **Standardized** numerical variables to improve model performance.
- **Encoded** categorical variables using Label Encoding.
- Split dataset into Train (80%) and Test (20%) for model evaluation.

Model Optimization

- Applied **SMOTE** (Synthetic Minority Over-sampling Technique) to handle class imbalance and improve recall.
- **Hyperparameter tuning** performed on Random Forest & Gradient Boosting using RandomizedSearchCV to improve accuracy and reduce overfitting.

Modelling Approach – After SMOTE

Final Model Performance After SMOTE & Hyperparameter Tuning

Model	Train Accuracy	Train Precision	Train Recall	Train F1-Score	Train AUC-ROC	Test Accuracy	Test Precision	Test Recall	Test F1-Score	Test AUC-ROC
Logistic Regression	72%	75%	66%	71%	72%	76%	42%	61%	50%	70%
Random Forest	100%	100%	100%	100%	100%	90%	70%	90%	79%	90%
XGBoost	95%	90%	99%	95%	95%	88%	66%	88%	75%	88%
SVM	73%	76%	67%	71%	73%	76%	43%	62%	51%	71%
Decision Tree	100%	100%	100%	100%	100%	85%	62%	69%	65%	79%

Key Takeaways

Random Forest

- Best recall (90%) → Strong at **catching defaulters** (**Recall Value**)
- Good generalization (90% AUC-ROC) → Balances overfitting with performance
- Moderate precision (70%) → Can still produce **false positives**.
- Best for: High-risk environments where missing defaulters is costly.

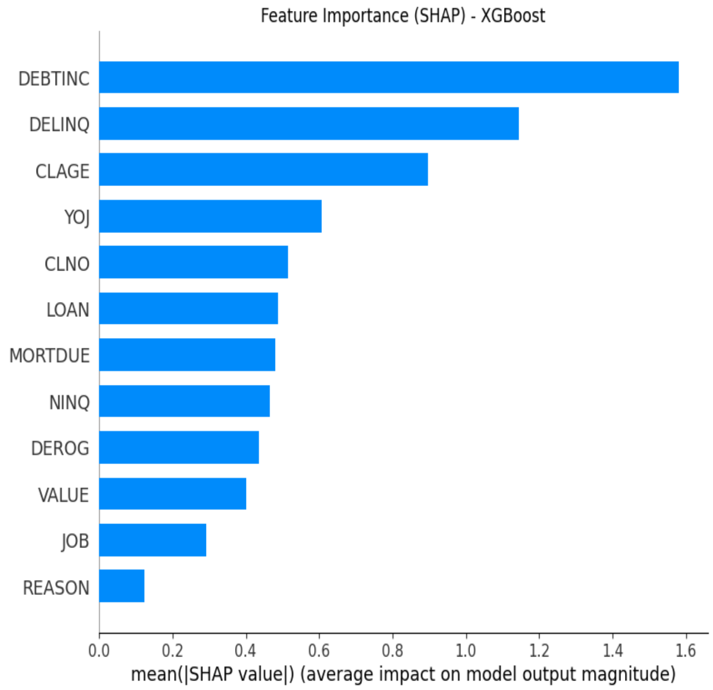
XGBoost

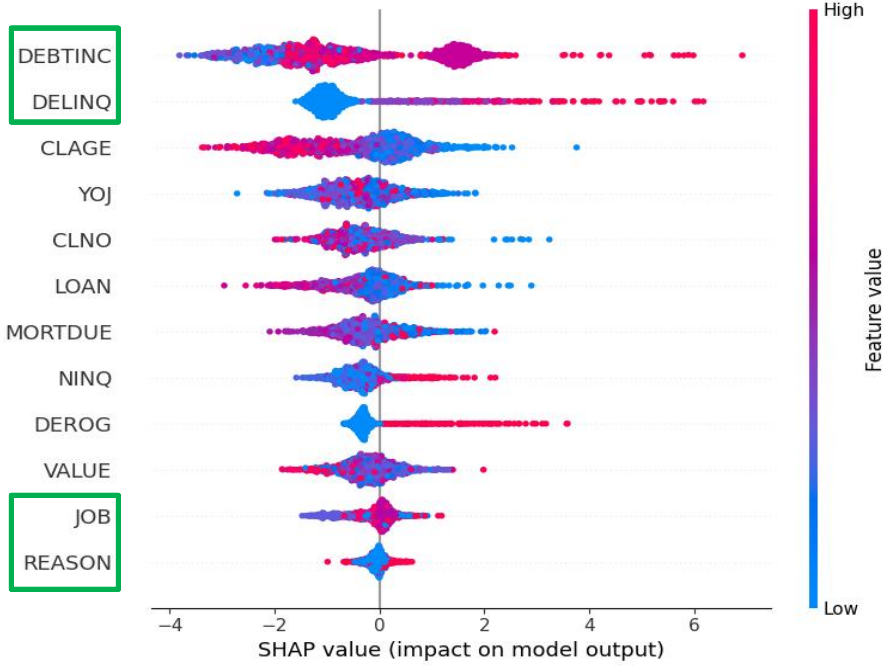
- Good recall (88.0%) → Decent at capturing defaulters.
- Balanced performance (88% AUC-ROC) → **Trades off** precision & recall well.
- Best for: When reducing false alarms is critical (e.g., avoiding unnecessary loan rejections)

Insights from Analysis

SHAP Feature Importance

The SHAP Feature Importance plot shows which features have the highest impact on predicting loan defaults.





Insights from Analysis

Statistical Tests vs. Machine Learning Insights

Statistical Test	Findings	Machine Learning Outcome
ANOVA	DELINQ, DEROG, DEBTINC, CLAGE are significant	✅ All included in ML model
Chi-Square	JOB, REASON impact default probability	✅ Encoded but had low SHAP importance
PCA	First 3 PCs explain ~40% variance	❌ PCA not used; models naturally select features

SHAP Feature Importance

Top 3 most important factors influencing default:

- Debt-to-Income Ratio (DEBTINC)
- Delinquency History (DELINQ)
- Credit Age (CLAGE)

Derogatory Marks (DEROG) & Loan Amount (LOAN) also contribute but with **lower impact**.

Insights from Analysis

Final Business Recommendations

For CMO (Marketing & Customer Acquisition):

- Target low-risk borrowers (low DEBTINC, long CLAGE) for premium loans.
- Offer better rates to financially **stable** customers (high YOJ, low DELINQ).

For COO (Risk & Operations):

- Apply **stricter** underwriting for high-risk borrowers (DEBTINC > 50, DELINQ > 2).
- Monitor high MORTDUE applicants before approving large loans.
- **Implement** early intervention policies to prevent defaults.

For Future Data Science Efforts (AI & Modeling):

- Use DEBTINC, DELINQ, and CLAGE as key predictors for loan default.
- Deploy XGBoost for best accuracy, with Random Forest for explainability.
- Enhance model explainability and fairness with SHAP & AI-driven credit scoring.

Insights from Analysis

- **Debt-to-Income Ratio** (DEBTINC) is the strongest predictor of loan default: Higher DEBTINC significantly increases default risk, Borrowers with (DEBTINC > 50) should be **flagged for stricter evaluation**.

- **Delinquency** (DELINQ) and Derogatory Marks (DEROG) indicate **high-risk borrowers**: More than 2 past delinquencies (DELINQ > 2) significantly **increase** default probability, Borrowers with major derogatory reports (DEROG > 1) are at higher risk.

- **Loan Amount** (LOAN) **alone does not strongly** predict default, but large loans (> $40,000) **combined** with high DEBTINC **increases** risk.

- **Credit Age** (CLAGE) acts as a protective factor: Borrowers with (CLAGE > 300) months have lower default rates, Long credit history should be **rewarded** with better loan terms.

- **Employment Stability** (YOJ) **impacts** default risk: Borrowers with (YOJ < 3) years are more likely to default.

- **Mortgage Obligations** (MORTDUE) **correlate** with loan amounts, but high mortgage dues > $200,000 increase default risk.

- **XGBoost** **outperforms** **Random Forest** in predictive power and captures stronger feature interactions, making it the best choice for default prediction, yet we've decided to obtain the benefits from both (**Future Ensemble application**).

Business Insights & Recommendations

Monetary Projections of Implementing Models

This slide shows a **business-level comparison** of machine learning models used to predict loan defaults, and it **quantifies the monetary** impact of each model's performance based on false negatives (FN) and false positives (FP)

Key Assumptions:

- False Negative (FN) Loss = 100% of loan value is lost if a defaulter is wrongly accepted (i.e., full loan is **written off**).
- False Positive (FP) Loss = 10% of loan value is **missed profit** when a good customer is wrongly rejected.
- Test Set Size = 1,192 loan applicants, with ~20% actual defaults (i.e., ~238 defaulters and ~954 non-defaulters).
- Average Loan Value = $18,608, calculated from the actual dataset.
- **Est. Financial Revenue** = Only earned from correctly approved loans (True Negatives), adjusted by removing FP and FN
- **Total projected revenue** for this specific dataset 1192 customer = (Actual FN-TEL) + EFR
  where, Actual FN = (original Wrongly Accepted – model specific Total Estimated Loss) +

Model	FN Loss ($) Wrongly Accepted	FP Loss ($) Wrongly Rejected	Total Estimated Loss FN + FP ($)	Actual FN Loss from Data ($) Write offs	Est. Financial Revenue ($) (TN−FP−FN) × (10%×Avg Loan)	Total Projected Gain in Revenue (Actual FN-TEL) + EFR
Random Forest	$446,591	$169,333	$615,924	$20,120,400	**$1,391,876**	**$19,504,476**
XGBoost	$539,631	$199,105	$738,736	$20,120,400	**$1,323,027**	**$19,381,664**
Logistic Regression	$1,730,541	$372,159	$2,102,701	$20,120,400	$857,827	$18,017,699
SVM	$1,693,325	$360,995	$2,054,320	$20,120,400	$883,879	$18,066,080
Decision Tree	$1,376,990	$186,080	$1,563,069	$20,120,400	$1,265,342	$18,557,331

Business Insights & Recommendations

Strategic Takeaways

AI-driven risk assessment will:

- **Reduce bad loan approvals**

→ Lower default rates
→ Improved Capital Allocation
→ **Operational Efficiency** Gains

- **Enhance risk-based lending**

→ Holistic Customer View
→ Personalized credit scoring
→ **Competitive Differentiation**

- **Improve customer segmentation**

→ **Target** high-risk clients proactively
→ Risk-Tiered Marketing

- **Implementation Roadmap**

→ Define approval criteria
→ Integrate AI-driven decision-making into workflows

Conclusion

- Our **AI-driven risk assessment** model successfully **enhances** monetary benefits, loan approval accuracy, reducing bad loan approvals and lowering default rates.

- Through data analysis and machine learning, we **improve** risk-based lending with personalized credit scoring and better customer segmentation.

- The **implementation roadmap** ensures **seamless** integration, optimizing decision-making for sustainable, fair, and efficient lending.

APPENDIX

Thank You, Questions?

Data Background and Contents

Data Dictionary

The Home Equity dataset (HMEQ) contains baseline and loan performance information for recent home equity loans. The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. There are 12 input variables registered for each applicant.

- **BAD**:1=Client defaulted on loan, 0 = loan repaid
- **LOAN**: Amount of loan approved
- **MORTDUE**: Amount due on the existing mortgage
- **VALUE**: Current value of the property
- **REASON**: Reason for the loan request (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- **JOB**: Thetype of job that loan applicant has such as manager, self, etc.
- **YOJ**: Years at present job kewlfunky@hotmail.com
- **DEROG**: Number of major derogatory reports (which indicates serious delinquency or late payments). GXO6TL4JN8
- **DELINQ**: Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due)
- **CLAGE**: Age of the oldest credit line in months
- **NINQ**: Number of recent credit inquiries
- **CLNO** :Number of existing credit lines
- **DEBTINC**: Debt-to-income ratio (all monthly debt payments divided by gross monthly income. This number is one of the ways lenders measure a borrower’s ability to manage the monthly payments to repay the money they plan to borrow)

Happy Learning !