MDS | CAPSTONE PROJECT

LOAN DEFAULT PREDICTION

Milestone 1

TEAM 4

CAPSTONE – MILESTONE 1

AGENDA

Executive Summary	03
Business Problem Overview	05
Solution Approach	07
Data Report	09
Exploratory Data Analysis	12
Data Preprocessing	24
Model Building Approach	26

Milestone 1

Loan Default Prediction

EXECUTIVE SUMMARY

This project aims to automate home equity loan approvals by building machine learning models to predict loan defaults. Through data analysis, preprocessing, and model building, we streamline decision-making and improve loan approval accuracy, balancing model performance with interpretability and fairness.

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.

Project Summary:

- **Problem Statement:** Automate loan approval processes to reduce human error and bias by predicting loan defaults using machine learning.
- **Solution Approach:** Conducted 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, understanding how the data was collected and ensuring its suitability for predictive modelling.
- **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.

Next Steps:

- **Model Building:** Develop both linear and ensemble models to predict loan defaults, focusing on interpretability and accuracy.
- **Performance Evaluation:** Assess models using metrics like accuracy, precision, recall, and AUC-ROC to gauge effectiveness.
- **Model Selection:** Choose the best model based on balanced performance and interpretability, aligned with project objectives.

BUSINESS PROBLEM OVERVIEW

The current manual home equity loan approval process is slow and prone to human error. Banks aim to automate it using machine learning to predict loan defaults, reducing biases and improving accuracy. This project focuses on developing a model to streamline decision-making and ensure fair, efficient approvals.

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.

SOLUTION APPROACH

This project aims to automate home equity loan approvals by building machine learning models to predict loan defaults, ensuring efficient and fair decisions. The approach involves data analysis, preprocessing, and model development to streamline the decision-making process and improve accuracy.

SOLUTION APPROACH

The solution approach includes a thorough analysis of transaction data through Exploratory Data Analysis (EDA) for initial insights, followed by data transformation to optimize model building and predictions. It concludes with an introduction to the model-building process, outlining its purpose and benefits.

EDA (Exploratory Data Analysis):

- Analyzed data to understand distributions, relationships, and outliers.
- Identified missing values and trends across key variables.

Data Preprocessing:

- Treated missing values and outliers.
- Applied one-hot encoding for categorical variables.
- Scaled numerical features for linear models.

Model Building:

- Built linear models (Logistic Regression, LDA) for interpretability.
- Developed ensemble methods (Random Forest, Boosting) for better accuracy and robustness.
- Tuned hyperparameters to optimize model performance.

DATA REPORT

The Data Report section summarizes the dataset used for analysis, focusing on its structure, key variables, and the treatment of missing values. This ensures a strong foundation for predicting loan defaults and conducting further analysis by maintaining data integrity and relevance.

DATA COLLECTION

The dataset used for this project contains key financial and demographic information about applicants, likely gathered through the bank's loan application and underwriting process. It includes variables such as loan amount, mortgage dues, income, and job type, as well as past credit history and delinquency details.

Data Collection Methodology:

- The data appears to be sourced from the bank's internal systems during the loan application and approval process, capturing both the applicant's financial profile and loan performance over time.

Possible Data Collection Methods:

- Data may have been collected through a combination of applicant-provided information (income, job type) and external sources like credit bureaus (delinquency and credit inquiries) to assess creditworthiness.
- Loan performance data (whether the applicant defaulted) could have been tracked over the course of the loan's lifecycle within the bank’s internal systems.

This collection methodology ensures the data is representative of the factors influencing loan default, supporting accurate predictive modelling.

DATA OVERVIEW

Field Name	Data Type	Description
BAD	int64	1 = Client defaulted, 0 = loan repaid
LOAN	int64	Amount of loan approved
MORTDUE	float64	Amount due on the existing mortgage
VALUE	float64	Current value of the property
REASON	object	Reason for the loan request (home improvement or Debt consolidation)
JOB	object	Type of job loan applicant has
YOJ	float64	Years at present job
DEROG	float64	Number of major derogatory reports
DELINQ	float64	Number of delinquent credit lines
CLAGE	float64	Age of the oldest credit line in months
NINQ	float64	Number of recent credit inquiries
CLNO	float64	Number of existing credit lines
DEBTINC	float64	Debt-to-income ratio

Dataset Information:

Key Insights:

- **Data Shape:** The dataset contains **5,960 rows** and **13 columns**, including numerical and categorical variables.
- **Data Types**: The dataset contains 11 numerical and 2 categorical columns, such as LOAN (numerical) and REASON (categorical).
- **Target Variable**: The target variable, BAD, is binary, indicating whether the applicant defaulted (1) or repaid (0) the loan.
- **Key Numerical Stats:** Loan amounts range from $1,100 to $89,900, with a mean of $18,607.
- **Key Categorical Stats:** Job type has six categories, with 'Other' being the most frequent.
- **Missing Values:** Missing data found in columns like MORTDUE, YOJ, and DEBTINC, with up to 21.26% missing in DEBTINC.
- **Outliers:** Outliers identified in columns like LOAN and CLAGE, treated based on interquartile range.
- **Other Information:** Numerical variables include financial indicators, and categorical data reflects applicant employment and loan reasons.

RECORDS	COLUMNS	DATA TYPE
5,960	13	float64(9), int64(2), object(2)

EXPLORATORY DATA ANALYSIS (EDA)

EDA (Exploratory Data Analysis) will reveal trends and key features of the dataset through univariate, bivariate and multivariate analysis, providing a deeper understanding of the underlying patterns and relationships within the data.

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.

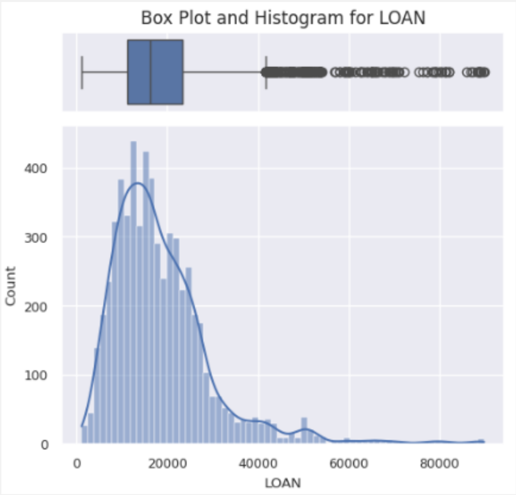
Conclusion:

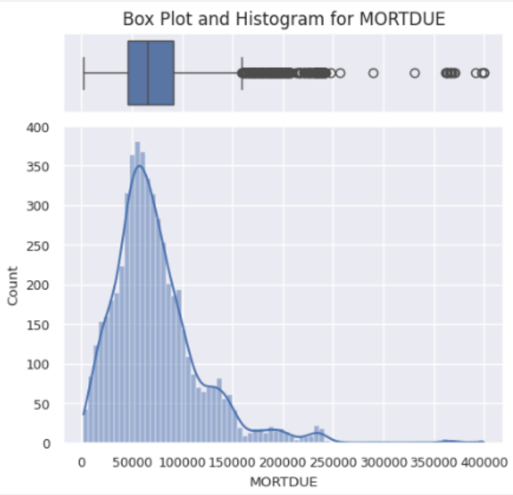
Key predictors for loan default risk are larger loan amounts, higher debt-to-income ratios, and poor credit history, while categorical variables like loan purpose and job type also provide valuable insights into default risk. The data is largely clean and ready for modelling, with minimal concerns around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

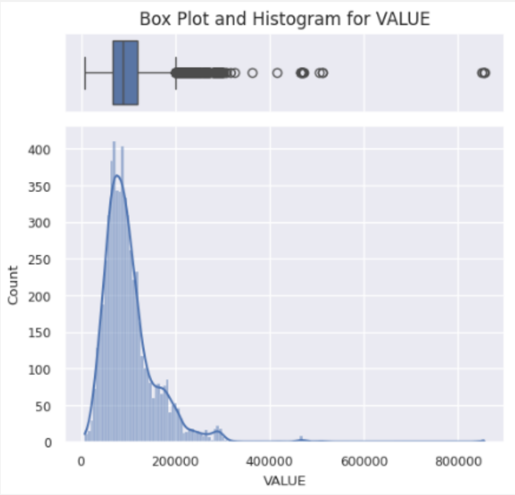
UNIVARIATE ANALYSIS (EDA)

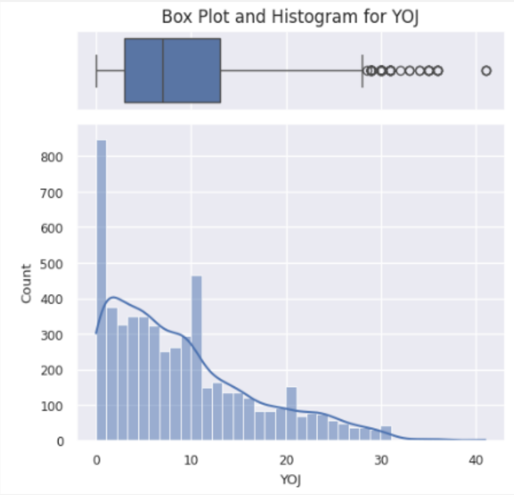
Univariate analysis examines individual variables to understand their distribution, central tendency, and spread. It helps identify key patterns, outliers, and missing values, providing initial insights into the data that guide further preprocessing and model-building steps.

UNIVARIATE ANALYSIS 1 OF 3





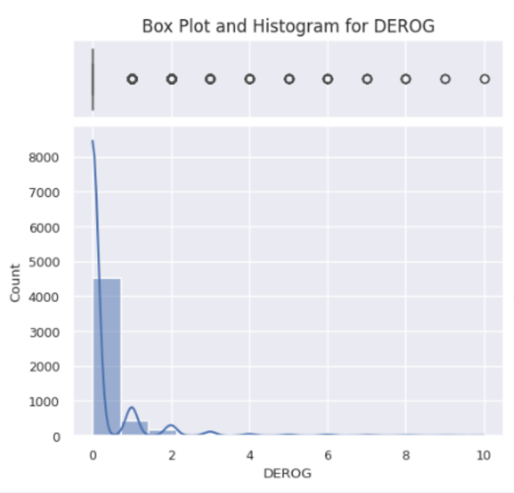


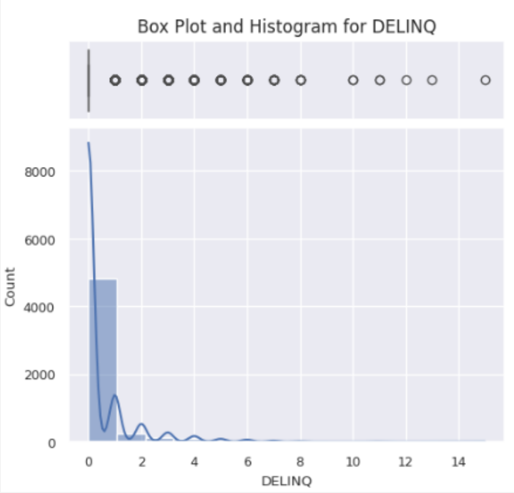


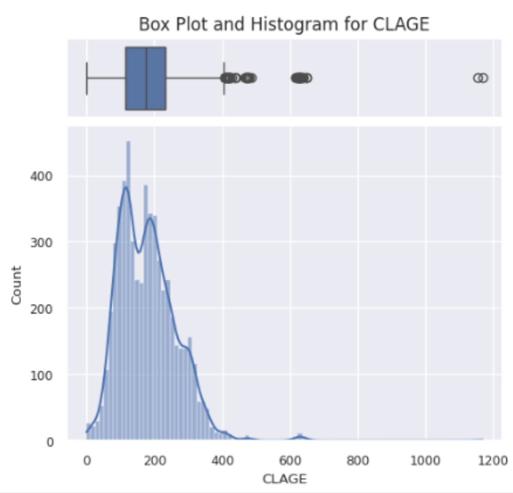
Loan Approved (LOAN)	Mortgage Due (MORTDUE)	Property Value (VALUE)	Years at Present Job (YOJ)
Key Insights: • The average loan amount is around $18,607, with a range from $1,100 to $89,900. • The distribution is skewed towards lower values, with a few high loan amounts acting as outliers.	**Key Insights:** • Mortgage dues range widely, with a mean of $73,760 and a maximum of $399,550. • The data has several high outliers, but the majority of mortgage dues are concentrated below $100,000.	**Key Insights:** • Property values have a mean of $101,776, with a maximum of $855,909. • The distribution shows a positive skew, with most properties valued under $200,000.	**Key Insights:** • The median number of years at the current job is 7, with values ranging from 0 to 41. • A significant portion of applicants have fewer than 5 years at their job.

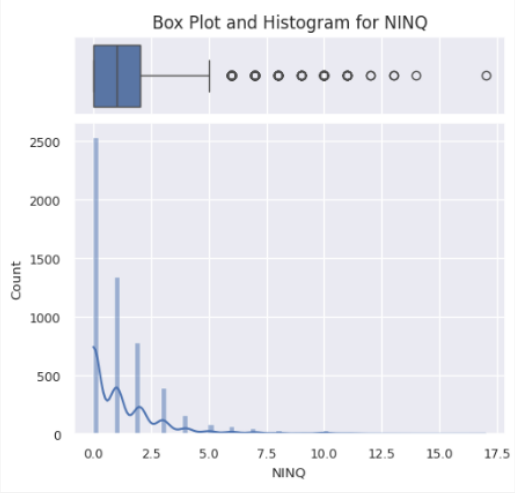
CAPSTONE – MILESTONE 1

UNIVARIATE ANALYSIS 2 OF 3









Derogatory Reports (DEROG)

Key Insights:

- Most applicants have no derogatory reports, with a mean of 0.25.
- There are a few cases with multiple derogatory reports, with outliers reaching up to 10.

Delinquent Cr. Lines (DELINQ)

Key Insights:

- The majority of applicants have no delinquencies, with a mean of 0.45.
- A few applicants have a high number of delinquencies, indicating financial stress.

Age of Oldest Cr. Line (CLAGE)

Key Insights:

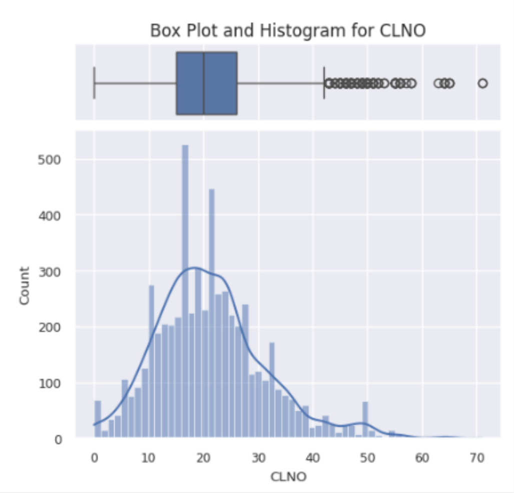
- The average age of the oldest credit line is 179 months (~15 years), with some as old as 97 years.
- The distribution shows a few outliers with very old credit lines..

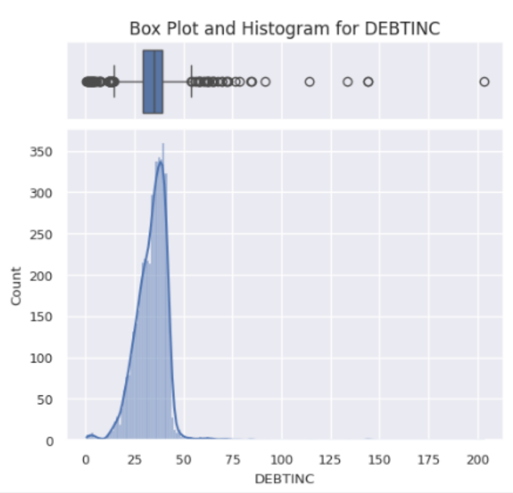
Recent Cr. Inquiries (NINQ)

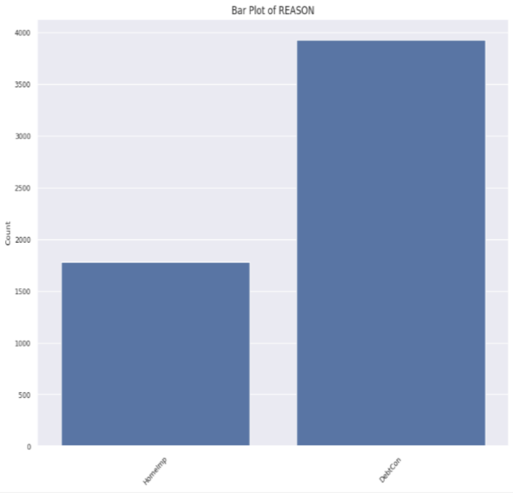
Key Insights:

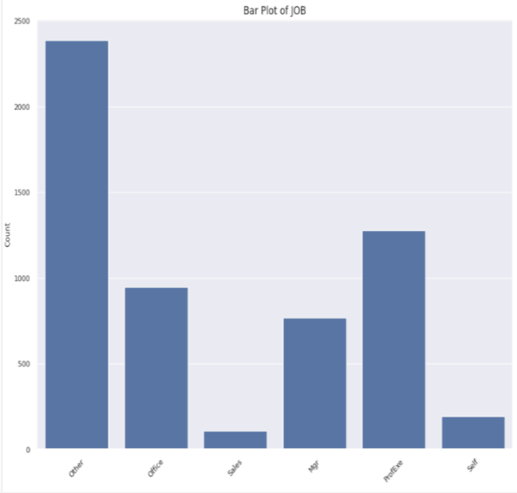
- Most applicants have fewer than 2 recent credit inquiries, with a maximum of 17.
- The data shows a skew towards lower numbers, with high inquiries considered risky.

UNIVARIATE ANALYSIS 3 OF 3


Box Plot and Histogram for CLNO


Box Plot and Histogram for DEBTINC


Bar Plot of REASON


Bar Plot of JOB

Existing Cr. Lines (CLNO)

Key Insights:

- The number of credit lines ranges widely, with an average of 21.
- The majority of applicants have between 15 to 26 credit lines, with a few outliers.

Debt/Income Ratio (DEBTINC)

Key Insights:

- Debt-to-income ratio averages around 33.7%, with a few outliers exceeding 200%.
- Most applicants have a debt-to-income ratio below 40%, with outliers signalling potential financial stress.

Reason for Loan Req (REASON)

Key Insights:

- The majority of applicants request loans for **Debt Consolidation** (~69%), with the rest for **Home Improvement**.
- Debt consolidation is a significant driver of loan applications, indicating applicants are likely looking to manage existing debt.

Type of Job (JOB)

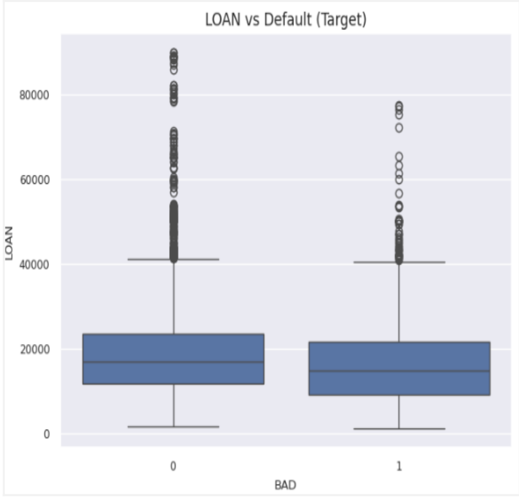
Key Insights:

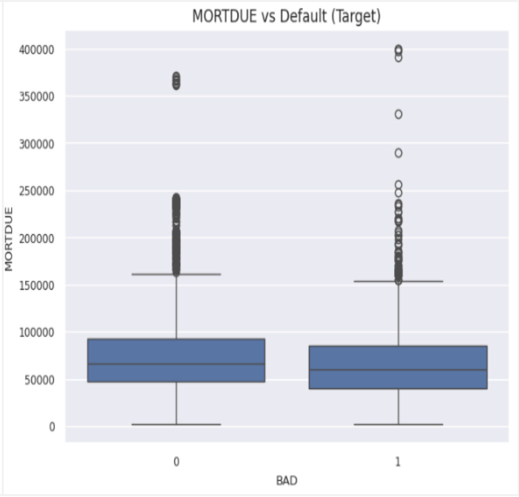
- The '**Other**' category is the most common job type, followed by **Professional/Executive** roles.
- The distribution suggests a diverse applicant base, with significant representation from office workers and managers.

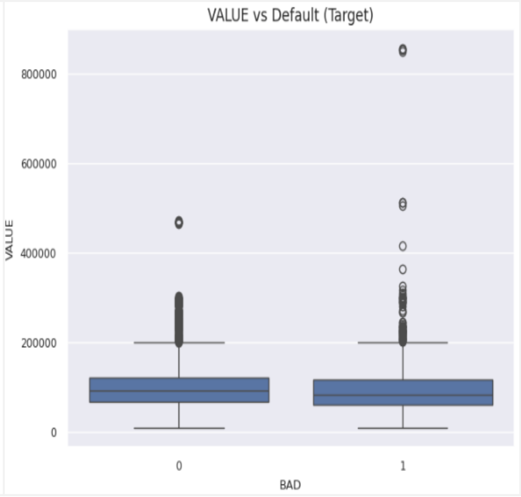
BIVARIATE ANALYSIS (EDA)

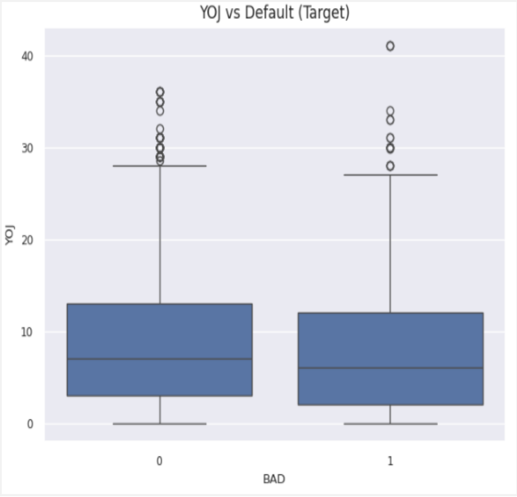
Bivariate analysis assesses the relationship between each feature and the target variable (BAD). It helps identify patterns, correlations, and feature importance by comparing how different features impact loan default.

BIVARIATE ANALYSIS 1 OF 3



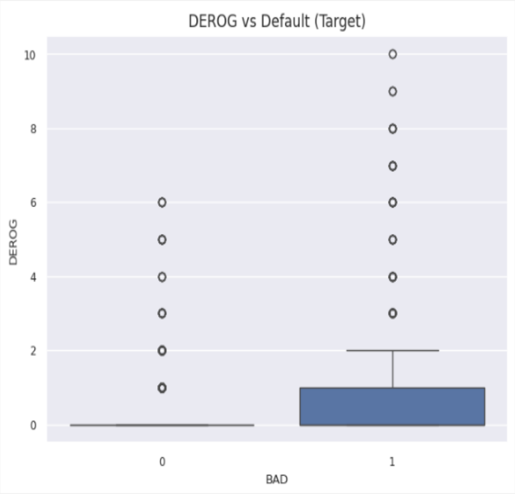


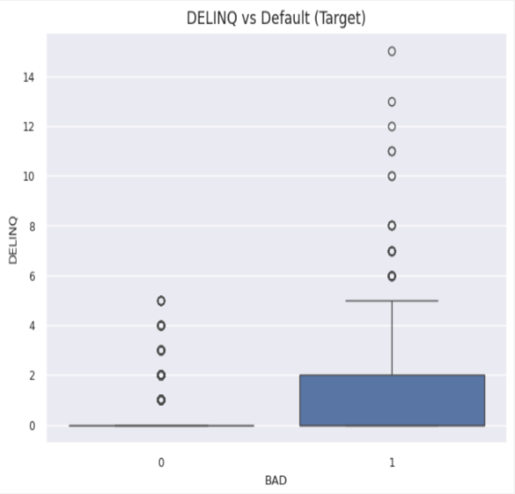


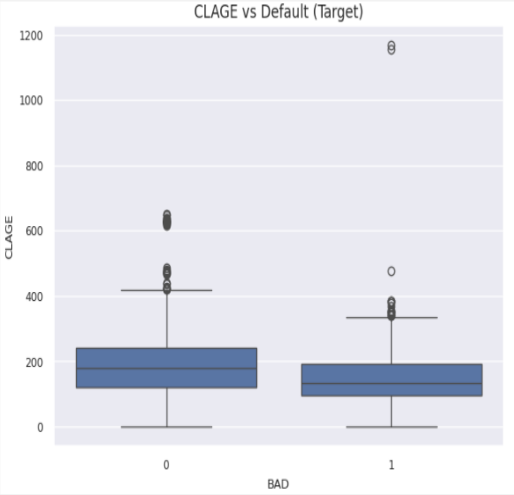


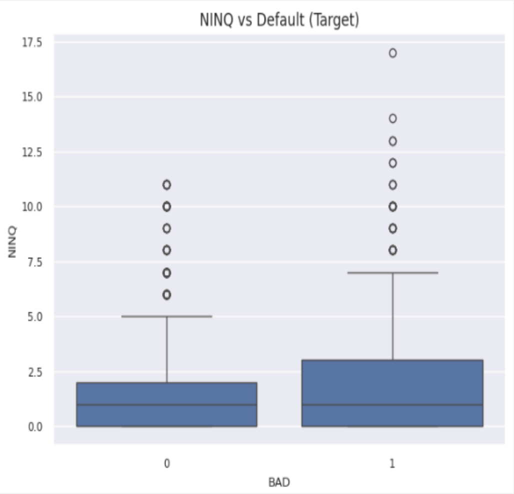
LOAN vs Default	MORTDUE vs Default	VALUE vs Default	YOJ vs Default
Key Insights: • Higher loan amounts tend to have a slightly higher probability of default. • Defaults are more frequent in the mid-to-high loan ranges compared to lower amounts.	**Key Insights:** • Higher mortgage dues show a moderate increase in default risk. • Applicants with low mortgage dues tend to default less frequently.	**Key Insights:** • Properties with lower values are associated with higher default rates. • Higher property values tend to correlate with lower default rates, showing applicants with higher equity are less likely to default.	**Key Insights:** • Applicants with fewer years on the job (YOJ < 5) show a higher likelihood of default. • Longer job tenure correlates with a lower risk of default, indicating stability in income.

BIVARIATE ANALYSIS 2 OF 3









DEROG vs Default

Key Insights:

- Applicants with more derogatory reports have a significantly higher chance of defaulting.
- Zero derogatory reports strongly correlate with loan repayment.

DELINQ vs Default

Key Insights:

- Higher numbers of delinquencies show a strong correlation with default.
- Applicants with no delinquency history are much less likely to default.

CLAGE vs Default

Key Insights:

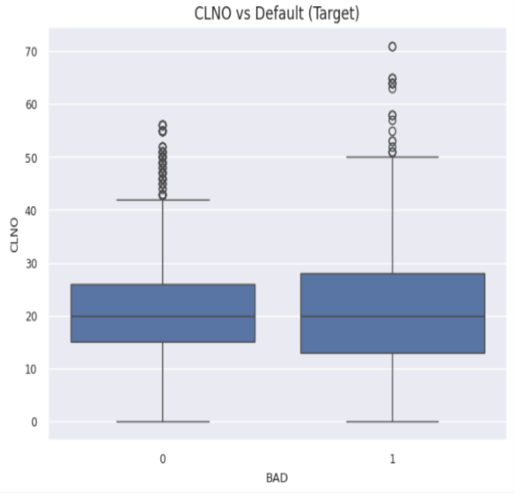
- Older credit lines generally correlate with a lower default risk.
- Newer credit lines show a higher chance of default, likely due to a shorter credit history.

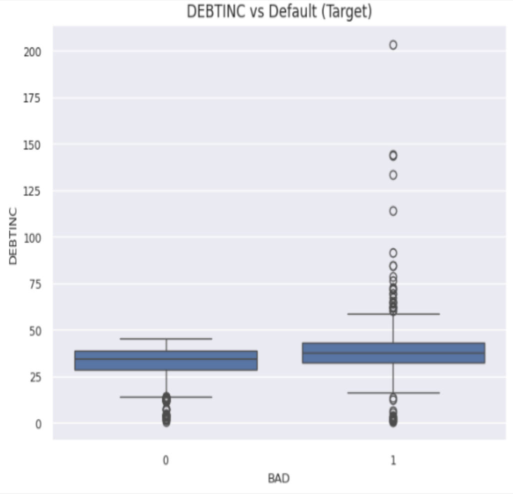
NINQ vs Default

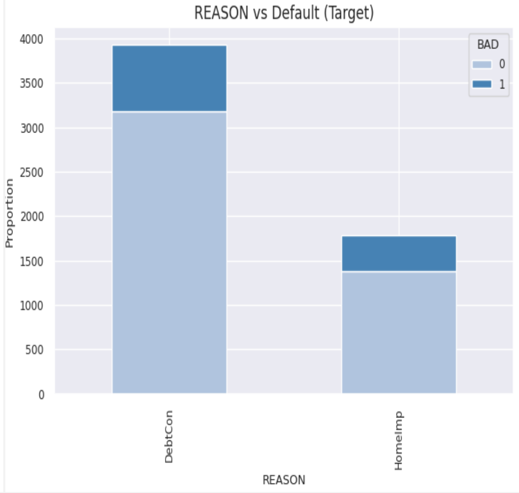
Key Insights:

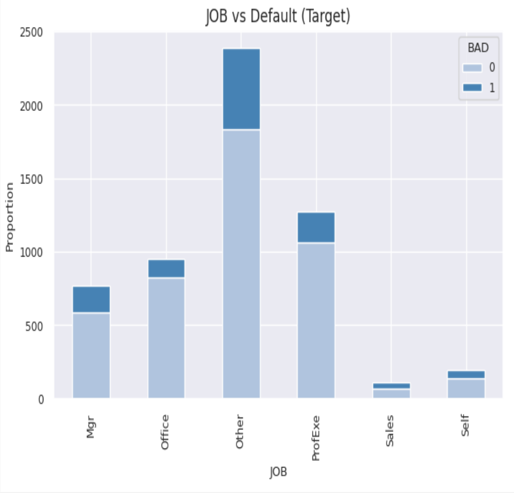
- A higher number of recent credit inquiries correlates with a higher probability of default.
- Applicants with fewer or no recent inquiries have lower default rates.

BIVARIATE ANALYSIS 3 OF 3









CLNO vs Default

Key Insights:

- Having more credit lines is associated with a moderate decrease in default probability.
- Applicants with fewer credit lines may have limited access to credit, leading to higher default rates.

DEBTINC vs Default

Key Insights:

- A higher debt-to-income ratio strongly correlates with default risk.
- Lower debt-to-income ratios indicate better financial stability and a lower chance of default.

REASON vs Default

Key Insights:

- Applicants applying for Debt Consolidation have a higher default rate than those applying for Home Improvement loans.
- Home Improvement applicants generally display lower default risks, suggesting they are financially more secure.

JOB vs Default

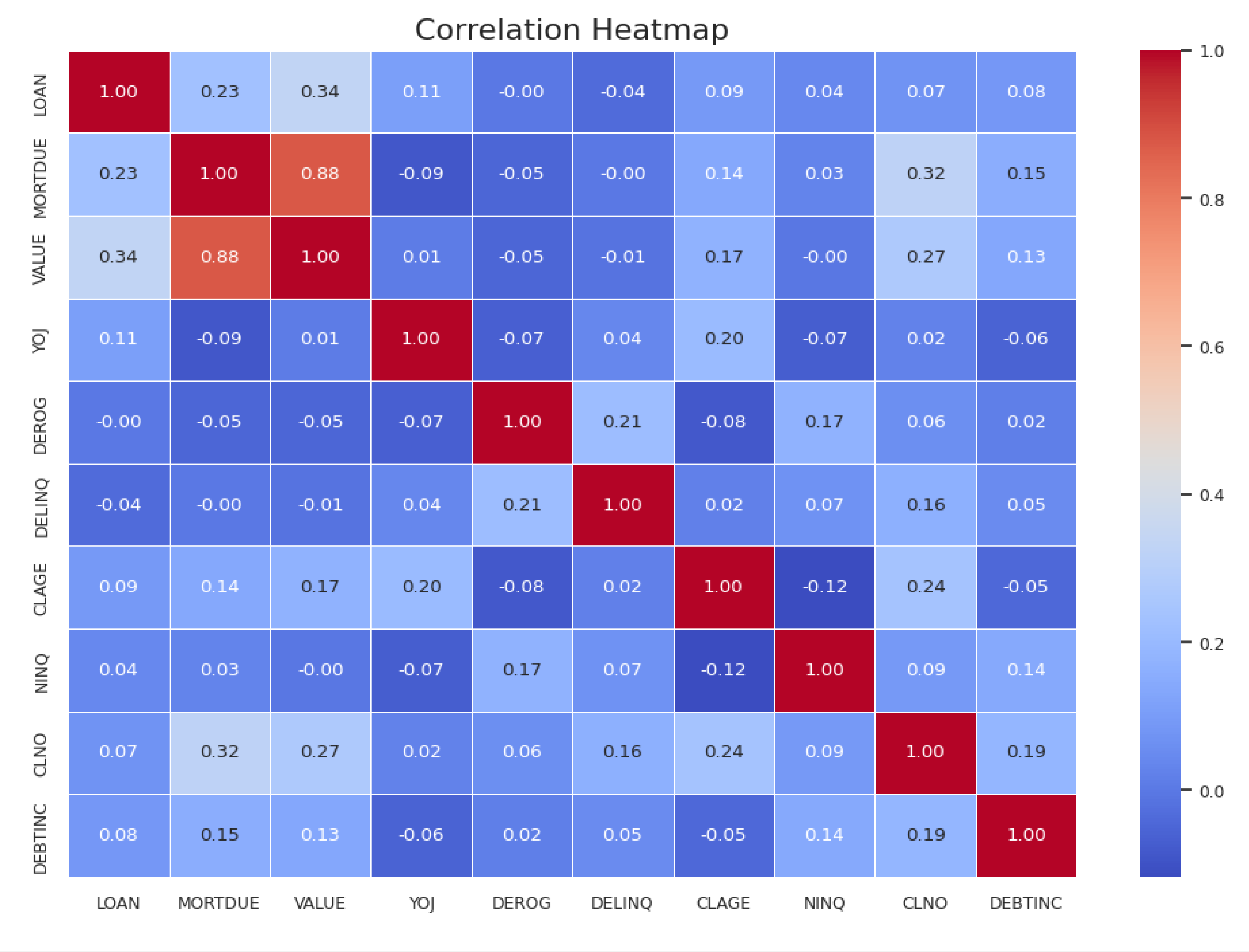
Key Insights:

- 'Other' job category shows a higher default rate compared to professional and managerial roles.
- Professional/Executive roles and Managers tend to have lower default risks, indicating more financial stability.

MULTIVARIATE ANALYSIS (EDA)

In multivariate analysis, a correlation map was used to identify relationships between features and detect multicollinearity. This ensures that no highly correlated features distort model performance, aiding in better feature selection for building robust models.

BIVARIATE ANALYSIS – NUMERICAL TO NUMERICAL



Correlation Heatmap

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
LOAN	1.00	0.23	0.34	0.11	-0.00	-0.04	0.09	0.04	0.07	0.08
MORTDUE	0.23	1.00	0.88	-0.09	-0.05	-0.00	0.14	0.03	0.32	0.15
VALUE	0.34	0.88	1.00	0.01	-0.05	-0.01	0.17	-0.00	0.27	0.13
YOJ	0.11	-0.09	0.01	1.00	-0.07	0.04	0.20	-0.07	0.02	-0.06
DEROG	-0.00	-0.05	-0.05	-0.07	1.00	0.21	-0.08	0.17	0.06	0.02
DELINQ	-0.04	-0.00	-0.01	0.04	0.21	1.00	0.02	0.07	0.16	0.05
CLAGE	0.09	0.14	0.17	0.20	-0.08	0.02	1.00	-0.12	0.24	-0.05
NINQ	0.04	0.03	-0.00	-0.07	0.17	0.07	-0.12	1.00	0.09	0.14
CLNO	0.07	0.32	0.27	0.02	0.06	0.16	0.24	0.09	1.00	0.19
DEBTINC	0.08	0.15	0.13	-0.06	0.02	0.05	-0.05	0.14	0.19	1.00

Correlation Analysis Insights

Strong Positive Correlation:

- MORTDUE and VALUE: Higher property values are strongly associated with higher mortgage dues, indicating that applicants with more valuable properties tend to carry larger mortgages.

Moderate Positive Correlation:

- LOAN with MORTDUE and VALUE: Larger loan amounts are moderately linked to higher mortgage dues and property values, suggesting that applicants borrowing larger sums often have more expensive properties.

Low Positive Correlation:

- CLNO and CLAGE: More credit lines are slightly correlated with older credit histories, implying that applicants with more credit lines have more established credit histories.

Negative Correlation:

- DEBTINC and CLAGE: Older credit lines are inversely related to debt-to-income ratio, indicating that applicants with longer credit histories tend to have lower debt burdens, reflecting better financial stability.

DATA PREPROCESSING

In the data preprocessing step, we treated missing values, handled outliers, applied one-hot encoding to categorical variables, and scaled numerical features for linear models. This ensured data integrity, consistency, and readiness for model building, allowing for accurate and unbiased predictions in the next phase.

DATA PREPROCESSING

1. Missing Value Imputation:

Missing values exists but relatively at lower percentage, so we will impute them before building models. We will use **KNN** for numerical columns and **SimpleImputer** with the most frequent value for categorical columns. Reasons for choosing imputation methods as follows:

- **KNN Imputation (Numerical Data):** Accounts for relationships between variables like loan amounts and debt-to-income ratios, leveraging patterns in similar rows.
- **Most Frequent Value (Categorical Data):** Replaces missing values with the most common category, which is reasonable for features like REASON and JOB.

2. Outlier Treatment:

Outlier treatment is essential to prevent extreme values from distorting model performance, especially in financial data. The following code uses IQR-based outlier treatment to cap outliers within acceptable limits.

- No Treatment for DEROG, DELINQ, and NINQ due to their discrete nature.
- **Identifying Outliers:** The IQR method calculates the range between the 25th and 75th percentiles, defining outliers as values outside this range.
- **Capping Bounds:** The lower and upper bounds are set at 1.5 times the IQR below Q1 and above Q3, respectively and Outliers were replaced with Upper and Lower bounds.

3. Encoding Categorical Columns:

We will encode the categorical columns and replace the output with 0s and 1s in preparation for model building.

- **One-hot Encoding:** It will help us convert categorical values into binary representations.
- No Clear Ordinal Values: We chose to use One-Hot encoding as features do not demonstrate clear ordinal values.

4. Train-Test Split:

In the Train-Test Split step, we separated the dataset into training and testing sets to evaluate model performance.

- **Feature/Target Separation:** We segregated predictors (features) and the target variable, ensuring the model is trained on relevant data.
- **Train/Test Split: The** split ensures the model is tested on unseen data for accurate performance assessment.

5. Multicollinearity Check:

For Linear models like Logistic Regression and LDA, it is crucial to check for multicollinearity among the features and address it if necessary.

- **No Immediate Concern:** While VALUE and MORTDUE have relatively higher VIF values, they are still below the common threshold of 5, meaning they do not present a severe multicollinearity problem.

6. Dataset Scaling:

Scaling is essential for models that are sensitive to the magnitude of features, such as Logistic Regression and LDA. These models can be influenced by features with larger values, leading to biased model performance.

- **Our Approach:** After encoding and splitting the dataset, we apply StandardScaler to the numerical features in both the training and test sets, ensuring consistency in feature scaling across the datasets.

MODEL BUILDING APPROACH

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

In this phase, we will focus on building both linear and ensemble models to predict loan defaults, utilizing the pre-processed data. Here's the planned approach:

Linear Models on Scaled Dataset:

- **Logistic Regression:** We will build this model after removing insignificant features based on statistical tests (like p-values) to improve model efficiency.
- **Linear Discriminant Analysis:** LDA will be built after feature selection to improve interpretability and ensure robust classification.

Ensemble Methods:

- **Random Forest:** We will first build a basic Random Forest model, followed by hyperparameter tuning to optimize performance using GridSearchCV method.
- **ADA Boost:** We will build an Adaptive Boosting (ADA Boost) model and fine-tune its parameters to improve model performance.
- **Gradient Boosting:** This model will be trained and then optimized through hyperparameter tuning to enhance its ability to handle complex decision boundaries.
- **XGBoost:** We will start with the default XGBoost model and tune its hyperparameters to improve its predictive power and speed.

Benefits of this approach:

- **Linear Models:** These provide easy interpretability and are useful for understanding the importance of features, especially after removing insignificant variables.
- **Ensemble Methods:** These models are more flexible, can handle feature interactions better, and are less prone to overfitting due to techniques like bagging and boosting. Hyperparameter tuning will further enhance their performance.

Performance Metric Check and Model Selection:

- Models will be evaluated using performance metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **AUC-ROC** to identify the best-performing model.
- We will compare results, focusing on predictive accuracy and generalization.
- The final model will be chosen based on its balance of performance, interpretability, and alignment with the project's objectives.

Great Learning
POWER AHEAD

Milestone 1

Loan Default Prediction

THANK YOU

TEAM 4: DAISY | DESMOND | SODEEQ | SHAISHAV
MDS | CAPSTONE PROJECT