Home Loan Defaults

Prediction Methods - Neural Networks

Contents / Agenda

- Business Problem Overview & Solution Approach
- Data Dictionary
- Exploratory Data Analysis (EDA) Results
- Data Preprocessing
- Model Performance Summary
- Best Model Selection
- Conclusions & Recommendations
- Appendix / Screenshots

Data Dictionary

The data contains different attributes of customers' loan and credit details. The detailed data dictionary is given below:

- BAD: Whether the client defaulted on the loan (0 = No, 1 = Yes)
- 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)
- JOB: Type of job the applicant has (e.g., Manager, Self-employed)
- YOJ: Years at present job

Data Dictionary

- DEROG: Number of major derogatory reports (serious delinquency/late payments)
- DELINQ: Number of delinquent credit lines (past due payments)
- 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)

Business Problem Overview

Consumer Home Loans represent a substantial percentage of retail bank profits. Loan defaults (NPA – bad loans) can put those profits at risk. It is in a bank's best interest to thoroughly vet prospective loan applicants to minimize loan default occurrences.

Loan Underwriting is a multi-phase process involving:

- **Credit Analysis/Assessment** – credit scores, history, existing debt, **DTI** ratio
- **Employment/Income Verification** employment status/history, income confirmation
- **Asset Assessment** – verification of down payment, investments, properties, liquid assets
- **Collateral Evaluation** – property appraisal, loan-to-value (**LTV**) ratio, title search
- **Risk Assessment** – borrower profile, red flags, risk rating

Business Problem Overview

Verification Process the existing process is a manual one with a lot of hands-on human interactions.

The current process is characterized by:

- Human-dependent
- Labor/Resource-intensive
- Time-consuming
- Manual / manually-driven
- Paper-heavy / document-heavy
- Prone to human error / inconsistency
- Subjective (underwriter discretion varies)
- Difficult to scale

Business Problem Overview

Limitations of the current approach include:

- Effort-intensive
- Clerical errors
- Data/document errors
- Prone to errors in judgment
- Prone to incorrect approvals
- Susceptible to human error
- Vulnerable to bias
- Documentation/reasoning consistency
- Bottlenecks, long turnaround times – reduces potential for scalability

Project Goals and Objectives

Goal:

A bank's consumer credit department seeks to streamline the decision-making process for home equity lines of credit. In compliance with the Equal Credit Opportunity Act (ECOA), the department requires an empirically derived and statistically sound credit scoring model.

The model will be developed using data collected through the existing loan underwriting process from recent applicants who were granted credit. The model needs to be interpretable to provide reasonings for loan acceptances/rejections.

Objective:

Build a classification model using an Artificial Neural Network (ANN) to predict clients who are likely to default on loans. The model should provide recommendations on the key features to consider during loan approval.

Solution Approach

1. Exploratory Data Analysis (EDA)

- Perform general **Exploratory Data Analysis** (**EDA**)
- Perform supplementary **EDA** on charts rendered through Rapid Miner.

2. Perform Follow up EDA after Feature Engineering

Assuming that **Feature Engineering** was provided for the project:

- Repeat EDA and supplementary EDA analysis of preprocessed (normalized) data

3. Basic Neural Network (NN) Analysis

- Analyze initial Model performance results

4. Neural Network with Parameter Tuning

Hand Tune selected NN parameters and rerun the model (iteratively)

- Comment on parameter tuned NN Model performance

Solution Approach

5. Neural Network Tuned Using Grid Search

GridSearchCV Evaluates all combinations of hyperparameters to find the best-performing configuration

- Analyze hyperparameter tuned **NN Model** performance.

6. Model Performance Summary

- Compare NN, parameter and hyperparameter tuned Model results

7. Conclusions and Recommendations

- Final observations, insights and recommendations

Expected Outcomes

- **Model Development**:
  - Build a highly performant classification model using ANN to predict loan defaults.
  - Make the model interpretable so approvals/rejections can be easily justifiable
  - Select optimal model parameters through hyperparameter tuning (**GridSearchCV**)
- **Risk Identification**:
  - Identify the key features that are the strongest predictors of loan default
  - Establish acceptable risk thresholds
- **Process Improvement**:
  - Minimize human subjectivity and bias in loan decisions
  - Streamline decision making time by adding automation to the process
  - Establish standardized assessments for all applicants
- **Increase Business Value**:
  - Identify high-risk applicants to minimize loan defaults
  - Reduce labour/manpower required in the current approval process
  - Increase scalability (process more loans) while keeping costs low

Initial Exploratory Data Analysis (EDA)

Loan Default Prediction Description Statistics

Feature	count	unique	top	freq	mean	std	min	25%	50%	75%	max
BAD	5960.00	nan	nan	nan	0.20	0.40	0.00	0.00	0.00	0.00	1.00
LOAN	5960.00	nan	nan	nan	18607.97	11207.48	1100.00	11100.00	16300.00	23300.00	89900.00
MORTDUE	5442.00	nan	nan	nan	73760.82	44457.61	2063.00	46276.00	65019.00	91488.00	399550.00
VALUE	5848.00	nan	nan	nan	101776.05	57385.78	8000.00	66075.50	89235.50	119824.25	855909.00
REASON	5708	2.00	DebtCon	3928	nan	nan	nan	nan	nan	nan	nan
JOB	5681	6.00	Other	2388	nan	nan	nan	nan	nan	nan	nan
YOJ	5445.00	nan	nan	nan	8.92	7.57	0.00	3.00	7.00	13.00	41.00
DEROG	5252.00	nan	nan	nan	0.25	0.85	0.00	0.00	0.00	0.00	10.00
DELINQ	5380.00	nan	nan	nan	0.45	1.13	0.00	0.00	0.00	0.00	15.00
CLAGE	5652.00	nan	nan	nan	179.77	85.81	0.00	115.12	173.47	231.56	1168.23
NINQ	5450.00	nan	nan	nan	1.19	1.73	0.00	0.00	1.00	2.00	17.00
CLNO	5738.00	nan	nan	nan	21.30	10.14	0.00	15.00	20.00	26.00	71.00
DEBTINC	4693.00	nan	nan	nan	33.78	8.60	0.52	29.14	34.82	39.00	203.31

Summary statistics for features including count, mean, standard deviation, and distribution percentiles.

Missing values — There are a lot of missing values (**DEBTINC** has 4,693 vs 5,960 total records); imputation needed

High variance — **VALUE** and **MORTDUE** have large standard deviations (**STD**) relative to means

Potential outliers — **CLAGE** max of 1,168 months (~97 years), **DEBTINC** max of 203% seem extreme

Skewed distributions — Several features show large gaps between 75% and max (**LOAN, VALUE, MORTDUE**)

Categorical features — **REASON** and **JOB** will need encoding

DEBTINC most incomplete — Only 4,693 records; key feature for loan risk so imputation strategy matters

Initial Exploratory Data Analysis (EDA)

Basic Type Information

Feature	Non-Null Count	Null Count	Dtype
BAD	5960.00	0.00	int64
LOAN	5960.00	0.00	int64
MORTDUE	5442.00	518.00	float64
VALUE	5848.00	112.00	float64
REASON	5708.00	252.00	object
JOB	5681.00	279.00	object
YOJ	5445.00	515.00	float64
DEROG	5252.00	708.00	float64
DELINQ	5380.00	580.00	float64
CLAGE	5652.00	308.00	float64
NINQ	5450.00	510.00	float64
CLNO	5738.00	222.00	float64
DEBTINC	4693.00	1267.00	float64

Suggested Type Conversion Map

Feature	Current Type	Converted Type
BAD	int64	int8
LOAN	int64	
MORTDUE	float64	
VALUE	float64	
REASON	object	
JOB	object	
YOJ	float64	int8
DEROG	float64	int8
DELINQ	float64	int8
CLAGE	float64	
NINQ	float64	int8
CLNO	float64	int8
DEBTINC	float64	

Analyze the raw data types and identify any that can be converted for better cpu performance.

Initial Exploratory Data Analysis (EDA)

Dataset Shape

Rows	Columns
5960	13

Based on the initial observations of the number of missing feature values, it's important to note the shape of the imported data set. The dataset has **5960** rows of data. It's important to keep that fact in mind when examining missing values and exploring imputation techniques.

The number of missing values calls into question the quality of the dataset and given the required number of imputations, how truly accurate the model performance, predictions and results will be.

It's important to truly understand the dataset, especially if it comes from an unknown source.

Initial Exploratory Data Analysis (EDA)

Missing Numeric Data

Feature	NaN Count	NaN %
MORTDUE	518.00	8.69
VALUE	112.00	1.88
YOJ	515.00	8.64
DEROG	708.00	11.88
DELINQ	580.00	9.73
CLAGE	308.00	5.17
NINQ	510.00	8.56
CLNO	222.00	3.72
DEBTINC	1267.00	21.26

Missing Categorical (String) Data

Feature	Missing Count	Missing %
REASON	252.00	4.23
JOB	279.00	4.68

A break down of the **Numeric** and **Categorical** feature missing data counts.

Initial Exploratory Data Analysis (EDA)

Summary of Missing Feature Row Counts

Missing Features	Row Count
0	3364
1	1589
2	449
3	219
4	64
5	83
6	66
7	25
8	39
9	49
10	11
11	2

Summary table of data completeness by showing how many rows have missing values. A lot of rows have 1 or more missing feature values. Given a data set of **5960** rows, only **3364** rows are feature complete. This once again calls into question the quality of the dataset. An imputation strategy is required to fill the missing values.

Initial Exploratory Data Analysis (EDA)

Loan Default Prediction Random Data Sample

BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
0.00	12700.00	79448.00	91004.00	DebtCon	Mgr	20.00	0.00	0.00	227.51	10.00	23.00	34.46
0.00	32300.00	235343.00	45270.00	DebtCon	Mgr	3.00	0.00	1.00	210.92	3.00	49.00	41.26
1.00	15000.00	106239.00	135942.00	DebtCon	ProfExe	19.50	0.00	1.00	9.10	2.00	24.00	nan
0.00	13000.00	72931.00	96803.00	DebtCon	Other	6.00	0.00	0.00	208.67	2.00	18.00	41.05
0.00	14700.00	10098.00	51653.00	DebtCon	ProfExe	3.00	0.00	0.00	302.65	0.00	16.00	18.42
0.00	14000.00	51721.00	61193.00	DebtCon	Other	0.00	0.00	0.00	187.00	2.00	12.00	25.81
0.00	5000.00	70470.00	77908.00	DebtCon	nan	5.00	nan	nan	nan	nan	nan	nan
0.00	26700.00	129559.00	170042.00	DebtCon	Mgr	1.00	0.00	0.00	232.65	1.00	24.00	34.12
0.00	7700.00	20887.00	26958.00	DebtCon	Other	0.00	0.00	0.00	17.46	10.00	6.00	30.23
0.00	20900.00	111464.00	144487.00	DebtCon	Office	11.00	0.00	1.00	206.68	0.00	25.00	40.07
0.00	13900.00	63831.00	81378.00	DebtCon	nan	9.00	nan	nan	nan	nan	nan	24.37
0.00	18000.00	105000.00	172500.00	DebtCon	Office	6.00	0.00	0.00	219.17	4.00	24.00	nan
0.00	15700.00	83830.00	122719.00	nan	nan	nan	nan	nan	nan	nan	nan	35.17
0.00	6800.00	136951.00	160306.00	HomeImp	ProfExe	11.00	0.00	0.00	219.74	1.00	18.00	20.89
0.00	14500.00	78503.00	90558.00	DebtCon	Mgr	19.00	0.00	0.00	250.45	10.00	24.00	36.88
0.00	32100.00	120725.00	168783.00	DebtCon	ProfExe	15.00	0.00	0.00	121.56	1.00	17.00	27.90
0.00	12000.00	84191.00	100654.00	HomeImp	ProfExe	0.00	0.00	0.00	79.68	1.00	20.00	33.77
0.00	18000.00	7051.00	66200.00	DebtCon	Office	nan	0.00	0.00	250.57	0.00	22.00	nan
0.00	47100.00	48062.00	107824.00	HomeImp	Self	9.00	0.00	0.00	203.69	2.00	35.00	41.53
0.00	17500.00	23288.00	94904.00	DebtCon	ProfExe	8.00	0.00	0.00	145.12	0.00	20.00	29.68

Table containing a random sampling of 20 rows from the imported raw dataset.

(EDA) – Complete Statistical Analysis

The original project specification slide deck contains statistical analysis for a few select features (**BAD**, **CLAGE**, **CLNO**). They provide a overview of some basic statistics:

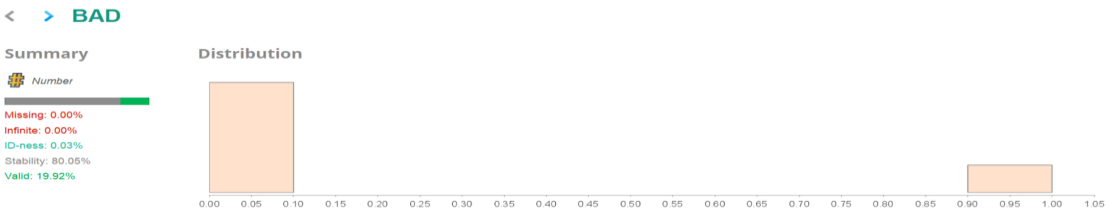
- **BAD**: Whether the client defaulted on the loan (0 = No, 1 = Yes)
- **CLAGE**: Age of the oldest credit line in months
- **CLNO**: Number of existing credit lines

While a high-level overview is a good place to start, it's important (as data scientists) to gain a thorough understanding of the complete dataset and its tendencies.
It is not in the best interest of model success to do anything else but.

Refer to the slides titled "Complete Statistical Analysis" for a more detailed examination of the data. Those sections provide a thorough exploratory data analysis of all features, distributions relationships, and key characteristics to ensure a comprehensive understanding of the data before the commencement of model development.

Statistics after Retrieve (Initial EDA):

- At this stage, the statistics were used to understand the raw data before any modifications. This helped in identifying missing values, data types, and the overall distribution of features. It represents the initial univariate exploratory data analysis (EDA).



Statistics

Name	Value
Minimum	0
Maximum	1
Average	0.199
Standard Deviation	0.400

BAD – represents whether a loan was defaulted or not.

Observations:

- Feature Label we are training the model to predict. No normalization or modifications to be done.
- **20%** of loans were defaulted / **80%** of loans were not
- Class imbalance exists (for non-defaulting results)
- Class weighting will have to be calculated and applied during model training

Statistics after Retrieve (Initial EDA):

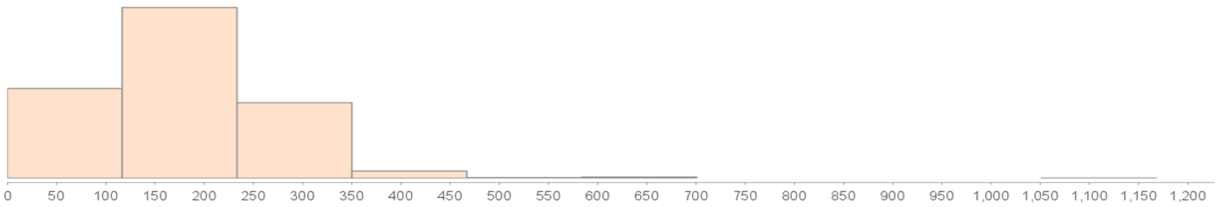
CLAGE

Summary

Number

Missing: 5.17%
Infinite: 0.00%
ID-ness: 0.49%
Stability: 0.12%
Valid: 94.22%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	1168.234
Average	179.766
Standard Deviation	85.810

CLAGE – age of oldest credit line (months).

Issues:

- Right-skewed with outliers – apply normalization.
- Data errors – (Min 0 years?) (Max 97 years?) – Cap/Clip

Statistics after Retrieve (Initial EDA):

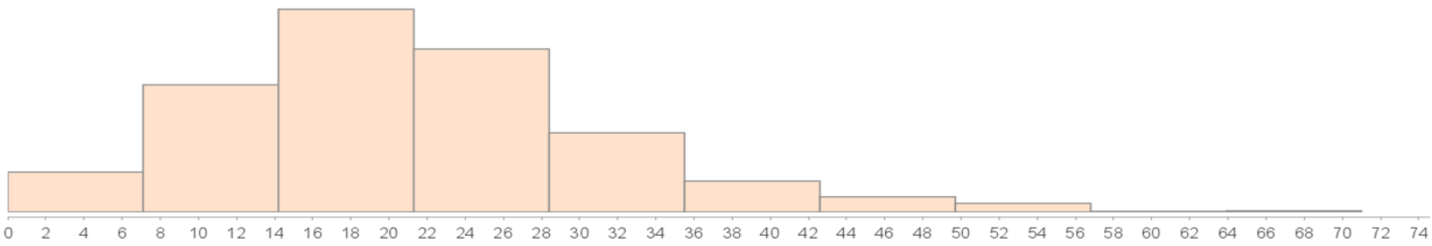
CLNO

Summary

Number

Missing: 3.72%
Infinite: 0.00%
ID-ness: 1.04%
Stability: 5.51%
Valid: 89.73%

Distribution



Statistics

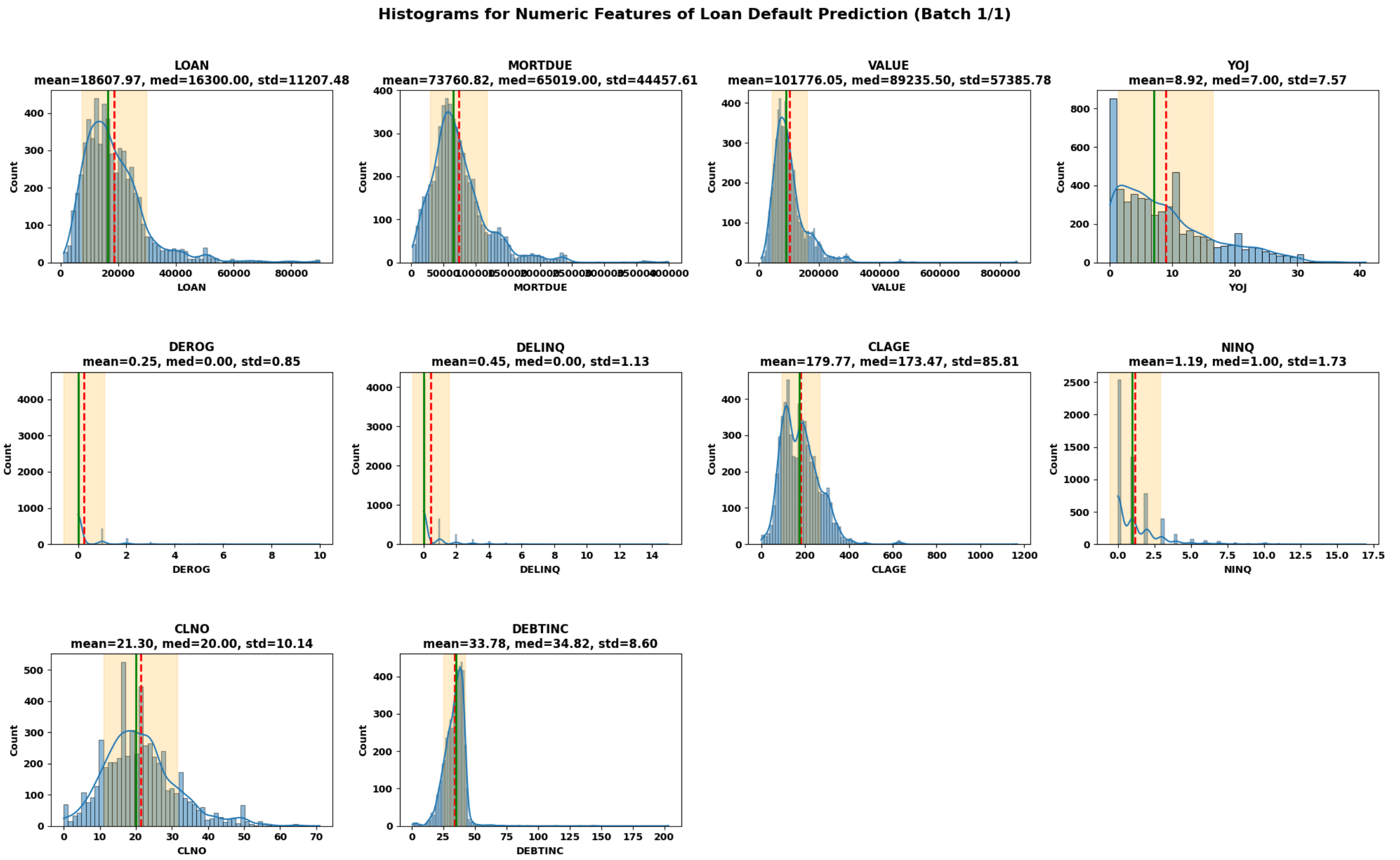
Name	Value
Minimum	0
Maximum	71
Average	21.296
Standard Deviation	10.139

CLNO – number of existing credit lines

Issues:

- Near-normal distribution with right-skewed with outliers – apply normalization.
- Data errors – (Min 0 credit lines?) (Max 71 credit lines?) – Cap/Clip

(EDA) – Complete Statistical Analysis



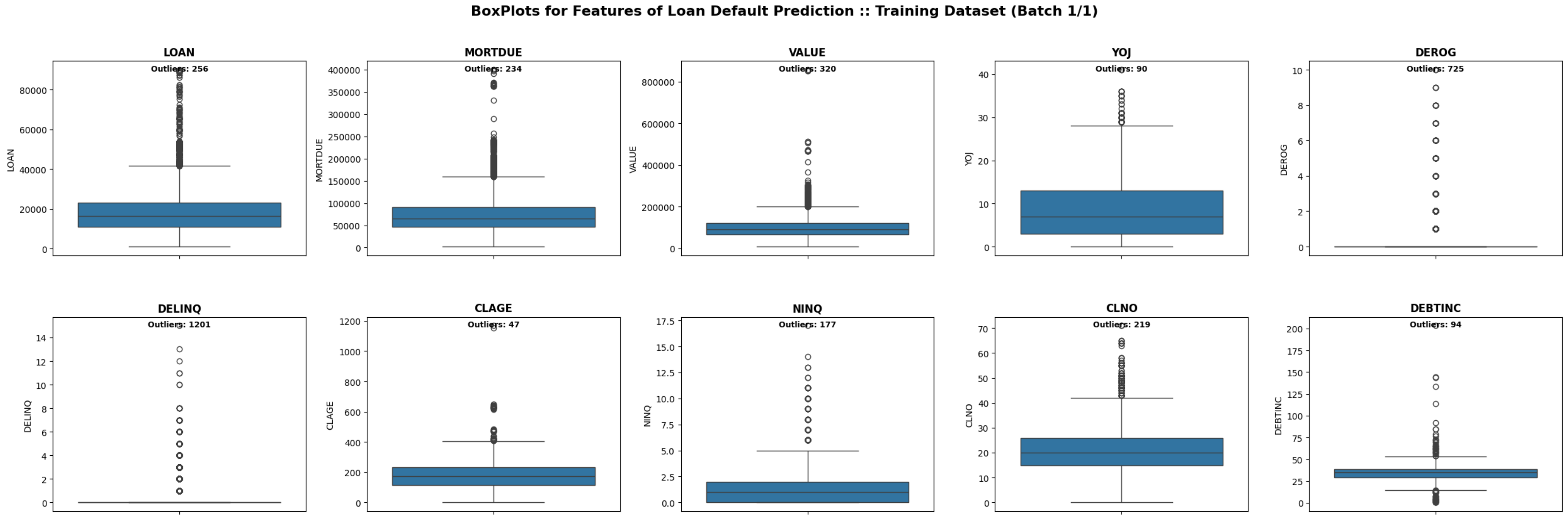
(EDA) – Complete Statistical Analysis

Numeric Features Statistics

Feature	Min	Max	Mean	Median	Mode	Std	Range	Quantile	Skew	Kurtosis
LOAN	1100.00	89900.00	18607.97	16300.00	[15000]	11207.48	88800.00	[11100.0, 16300.0, 23300.0,...	2.02	6.93
MORTDUE	2063.00	399550.00	73760.82	65019.00	[42000.0]	44457.61	397487.00	[46276.0, 65019.0, 91488.0,...	1.81	6.48
VALUE	8000.00	855909.00	101776.05	89235.50	[60000.0]	57385.78	847909.00	[66075.5, 89235.5, 119824.2...	3.05	24.36
YOJ	0.00	41.00	8.92	7.00	[0]	7.57	41.00	[3, 7, 13, 21, 24]	0.99	0.37
DEROG	0.00	10.00	0.25	0.00	[0]	0.85	10.00	[0, 0, 0, 1, 2]	5.32	36.87
DELINQ	0.00	15.00	0.45	0.00	[0]	1.13	15.00	[0, 0, 0, 2, 3]	4.02	23.57
CLAGE	0.00	1168.23	179.77	173.47	[102.5, 206.96666667]	85.81	1168.23	[115.11670223, 173.46666667...	1.34	7.60
NINQ	0.00	17.00	1.19	1.00	[0]	1.73	17.00	[0, 1, 2, 3, 4]	2.62	9.79
CLNO	0.00	71.00	21.30	20.00	[16]	10.14	71.00	[15, 20, 26, 34, 40]	0.78	1.16
DEBTINC	0.52	203.31	33.78	34.82	[0.5244992154, 0.7202950067...	8.60	202.79	[29.140031372, 34.818261819...	2.85	50.50

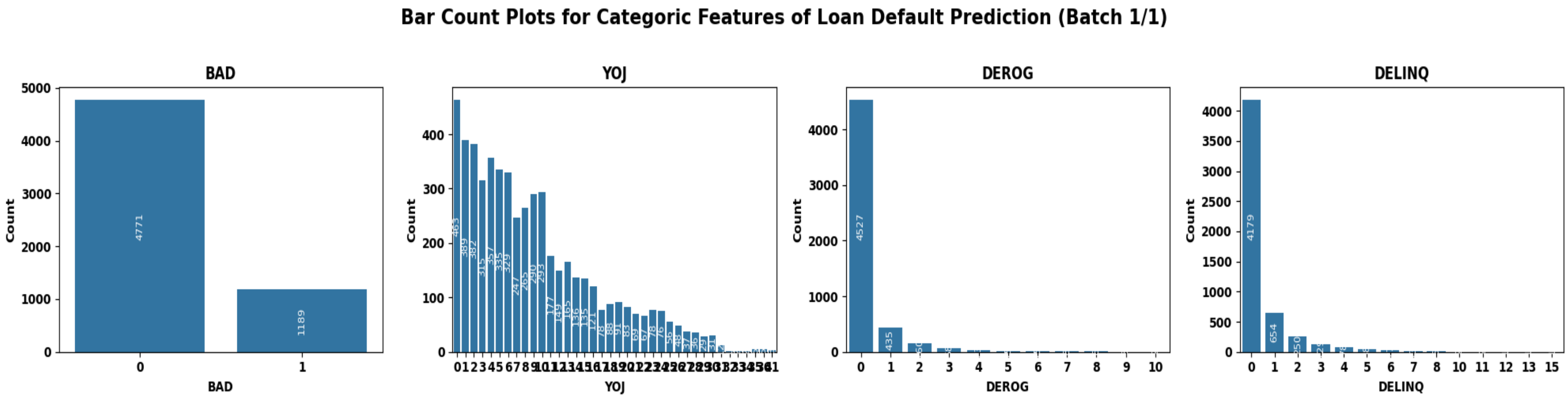
- **LOAN** – Right-skewed, clustered around $15k, large outliers
- **MORTDUE** – Right-skewed, clustered around $70k
- **VALUE** – Highly Right-skewed
- **YOJ** – Most normal distribution, mode=0
- **DROG** – Zero inflated (most values=0)
- **DELINQ** – Zero inflated (most values=0)
- **CLAGE** – Relatively symmetric, extreme outliers
- **NINQ** – Highly Right-skewed, most have 0-2 inquires
- **CLNO** – Close to normal distribution, but still has right skew
- **DEBTINC** – High kurtosis, extreme outliers

(EDA) – Complete Statistical Analysis



- **LOAN** – Wide **Inter Quartile Range (IQR)**, outliers > $40k
- **MORTDUE** – Tight **IQR**, high outliers > $45k
- **VALUE** – Tight **IQR**, high outliers > 200k
- **YOJ** – Wideset **IQR**, still high outliers
- **DROG** – Box collapsed at 0
- **DELINQ** – Box collapsed at 0
- **CLAGE** – Tight **IQR**, fewer outliers
- **NINQ** – Tight **IQR**
- **CLNO** – Tight **IQR,** high outliers
- **DEBTINC** – Tight **IQR**, +/- outliers exist

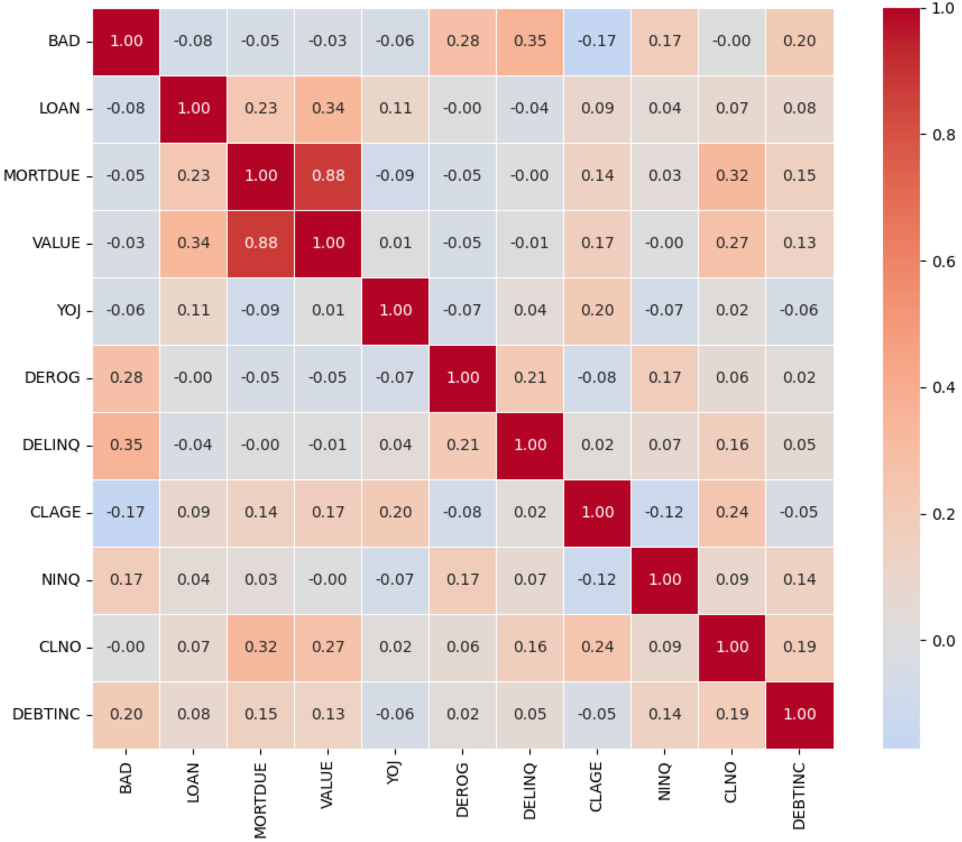
(EDA) – Complete Statistical Analysis



- **BAD** – predicted feature label, class imbalance, 80% of loans don't default
- **YOJ** – Right-skewed, declines after 4 years
- **DEROG** – Most values are 0 (**zero-inflated**)
- **DELINQ** – Zero-inflated, Right-skewed

(EDA) – Correlation Matrix

Correlation Matrix



(EDA) – Correlation Matrix

Feature Correlation

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
BAD	1.00	-0.08	-0.05	-0.03	-0.06	0.28	0.35	-0.17	0.17	-0.00	0.20
LOAN	-0.08	1.00	0.23	0.34	0.11	-0.00	-0.04	0.09	0.04	0.07	0.08
MORTDUE	-0.05	0.23	1.00	0.88	-0.09	-0.05	-0.00	0.14	0.03	0.32	0.15
VALUE	-0.03	0.34	0.88	1.00	0.01	-0.05	-0.01	0.17	-0.00	0.27	0.13
YOJ	-0.06	0.11	-0.09	0.01	1.00	-0.07	0.04	0.20	-0.07	0.02	-0.06
DEROG	0.28	-0.00	-0.05	-0.05	-0.07	1.00	0.21	-0.08	0.17	0.06	0.02
DELINQ	0.35	-0.04	-0.00	-0.01	0.04	0.21	1.00	0.02	0.07	0.16	0.05
CLAGE	-0.17	0.09	0.14	0.17	0.20	-0.08	0.02	1.00	-0.12	0.24	-0.05
NINQ	0.17	0.04	0.03	-0.00	-0.07	0.17	0.07	-0.12	1.00	0.09	0.14
CLNO	-0.00	0.07	0.32	0.27	0.02	0.06	0.16	0.24	0.09	1.00	0.19
DEBTINC	0.20	0.08	0.15	0.13	-0.06	0.02	0.05	-0.05	0.14	0.19	1.00

- Strongest Predictors of loan default (BAD)
  - **DELINQ** (0.35)
  - **DEROG** (0.28)
  - **DEBTINC** (0.20)
- Multicollinearity:
  - **MORTDUE** and **VALUE**

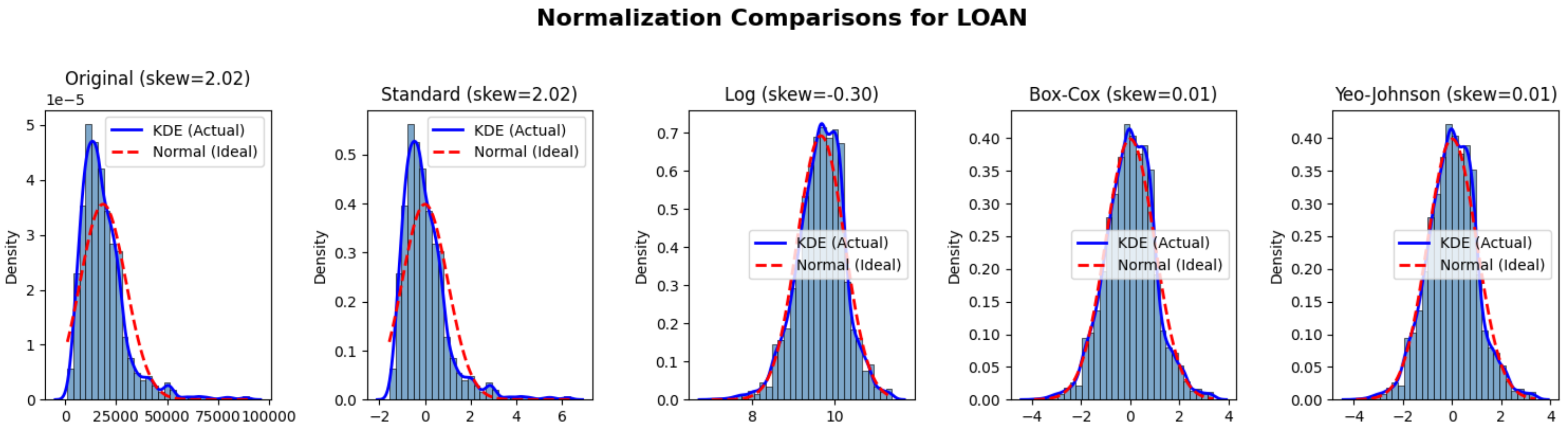
Complete Statistical Analysis: Normalization Comparisons

Preliminary examinations revealed the existence of heavily right-skewed features with extreme outliers. The effective course of action is to apply normalization techniques to produce more normalized distributions, reduce the skewness and effect of outliers. Capping/clipping should also be considered as corrective actions for outliers.

The following slides compare 4 normalization functions and compare their results against the original distributions:

- **Standard Scaling**
- **Log (log1p)**
- **Yeo-Johnson**
- **Box-Cox**

(EDA) – Complete Statistical Analysis

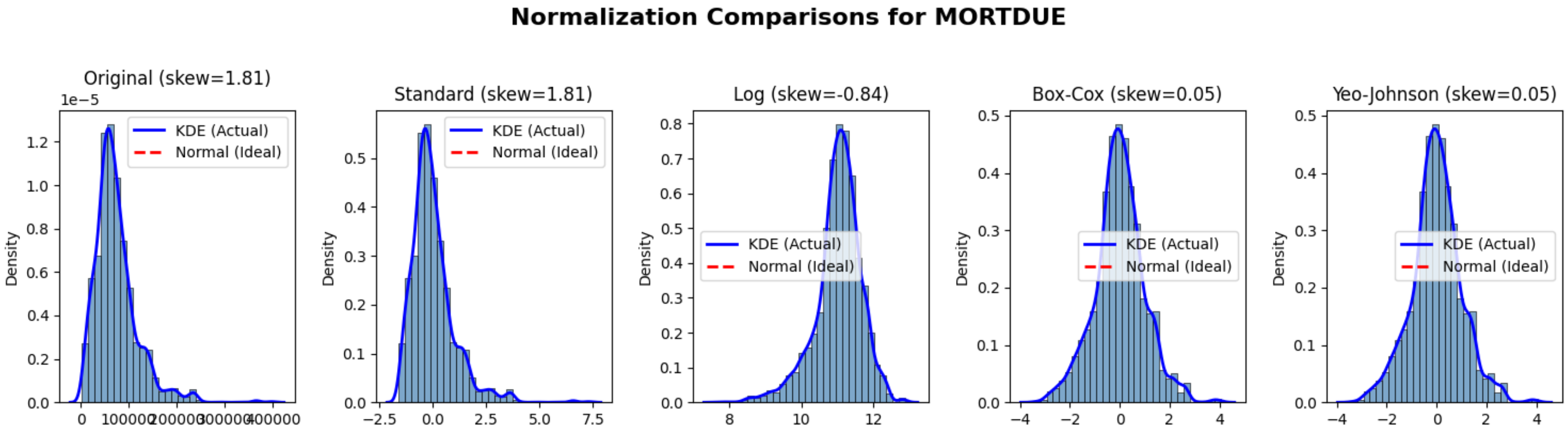


Normalization Comparisons for LOAN

Method	Skew
Original	2.02
StandardScaler	2.02
Log	-0.30
Box-Cox	0.01
Yeo-Johnson	0.01

- **Yeo-Johnson/Box-Cox** – both score best, near-normal distribution with **skew** of 0.01
- **Yeo-Johnson** – For consistency, use **Yeo-Johnson**

(EDA) – Complete Statistical Analysis

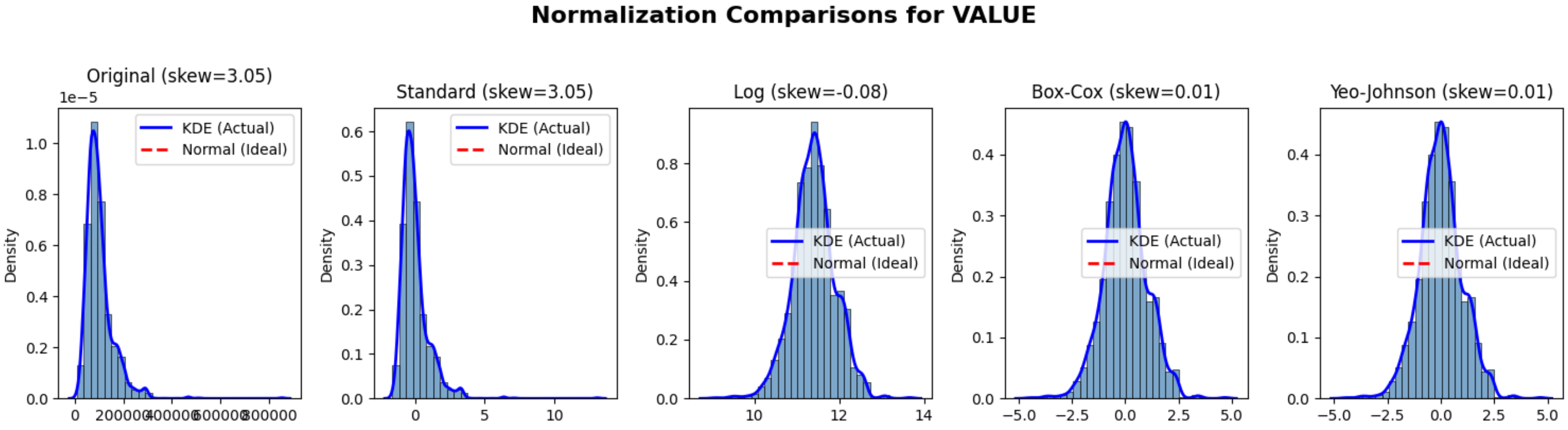


Normalization Comparisons for MORTDUE

Method	Skew
Original	1.81
StandardScaler	1.81
Log	-0.84
Box-Cox	0.05
Yeo-Johnson	0.05

- **Yeo-Johnson/Box-Cox** – both score best, near-normal distribution with **skew** of 0.01
- **Yeo-Johnson** – For consistency, use **Yeo-Johnson**

(EDA) – Complete Statistical Analysis

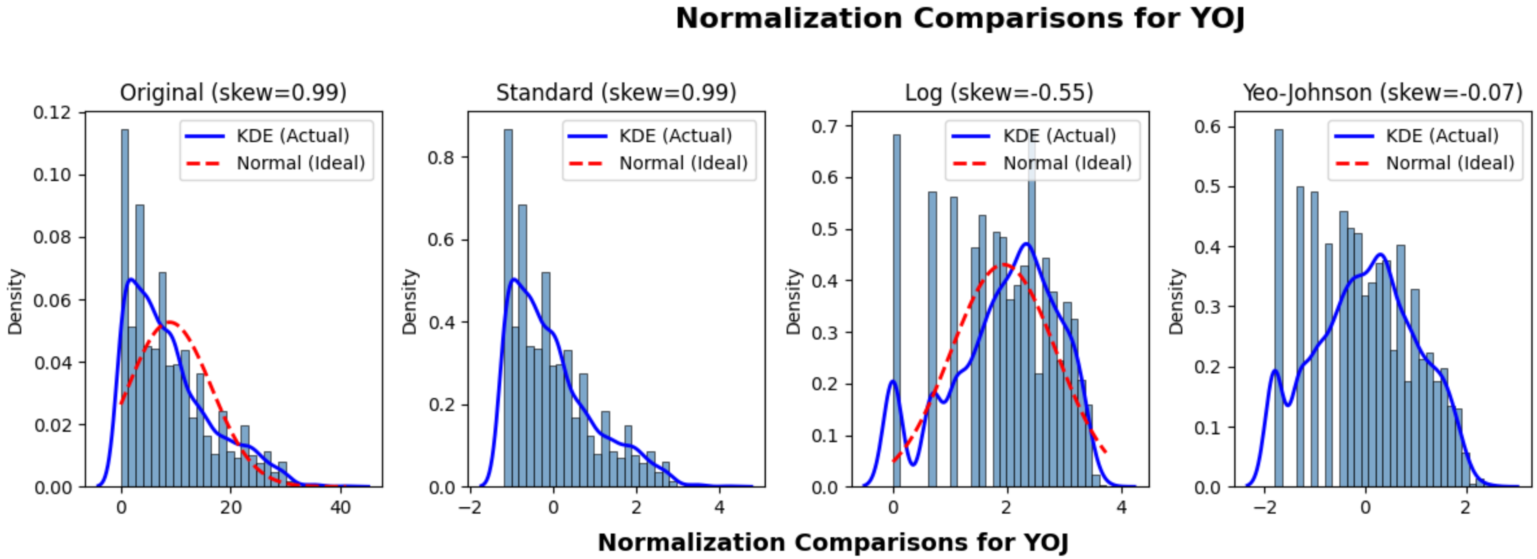


Normalization Comparisons for VALUE

Method	Skew
Original	3.05
StandardScaler	3.05
Log	-0.08
Box-Cox	0.01
Yeo-Johnson	0.01

- **Yeo-Johnson/Box-Cox** – both score best, near-normal distribution with **skew** of **0.01**
- **Yeo-Johnson** – For consistency, use **Yeo-Johnson**

(EDA) – Complete Statistical Analysis

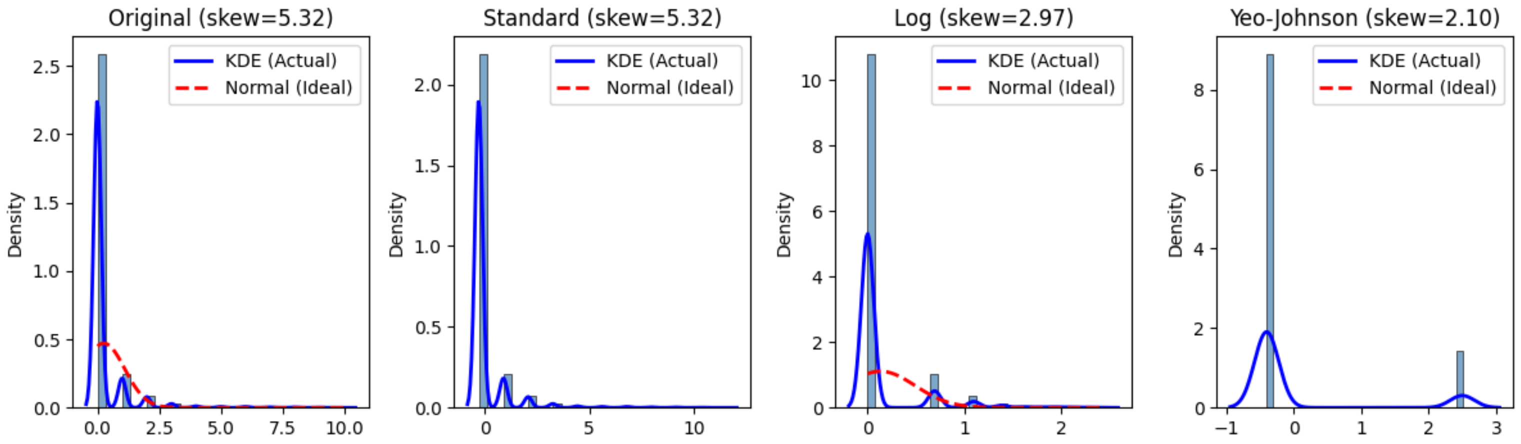


Method	Skew
Original	0.99
StandardScaler	0.99
Log	-0.55
Yeo-Johnson	-0.07

- **Yeo-Johnson** – best score best, near-normal distribution with **skew** of **-0.07**

(EDA) – Complete Statistical Analysis

Normalization Comparisons for DEROG



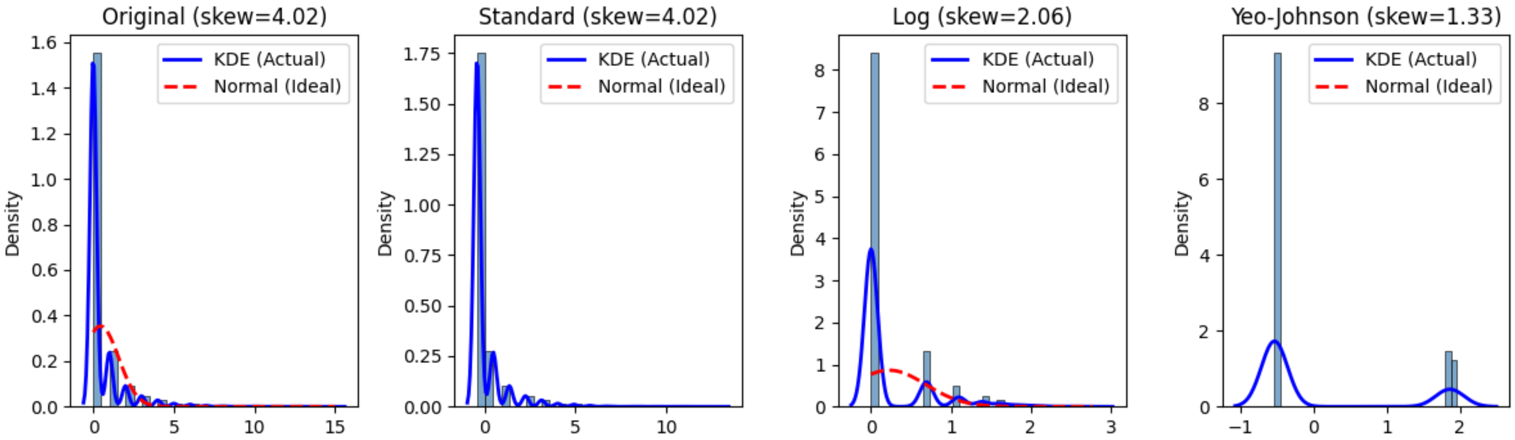
Normalization Comparisons for DEROG

Method	Skew
Original	5.32
StandardScaler	5.32
Log	2.97
Yeo-Johnson	2.10

- **Yeo-Johnson** – both score best with **skew** of **2.10**

(EDA) – Complete Statistical Analysis

Normalization Comparisons for DELINQ

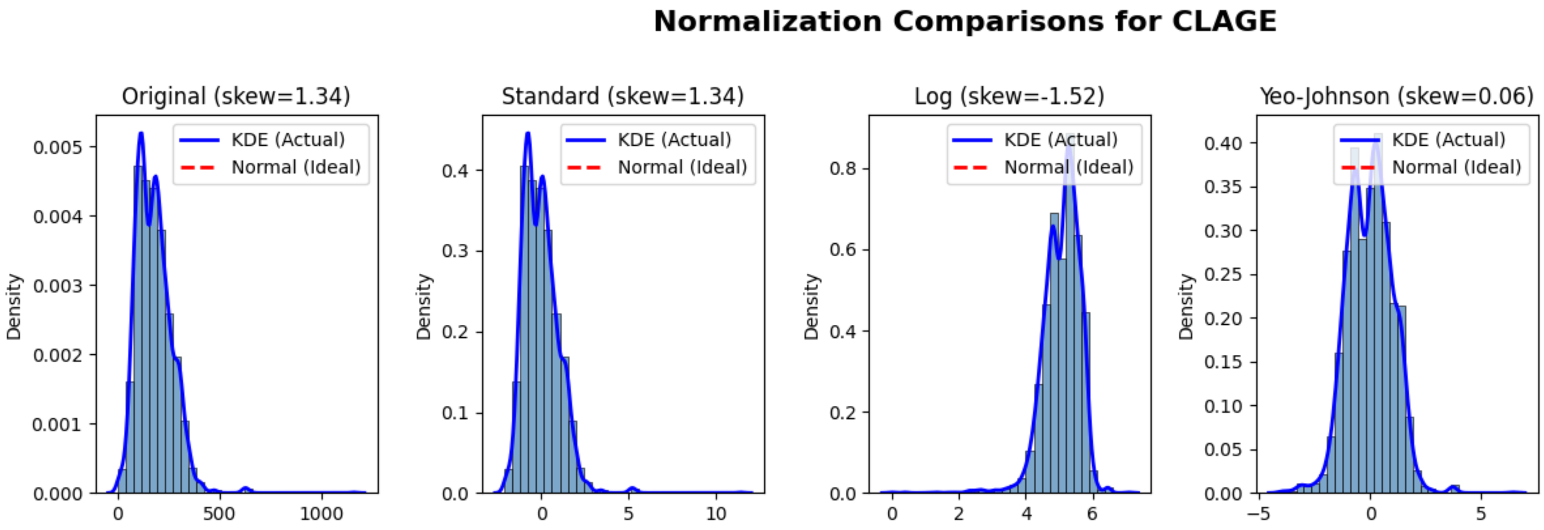


Normalization Comparisons for DELINQ

Method	Skew
Original	4.02
StandardScaler	4.02
Log	2.06
Yeo-Johnson	1.33

- **Yeo-Johnson** – both score best with **skew** of **1.33**

(EDA) – Complete Statistical Analysis

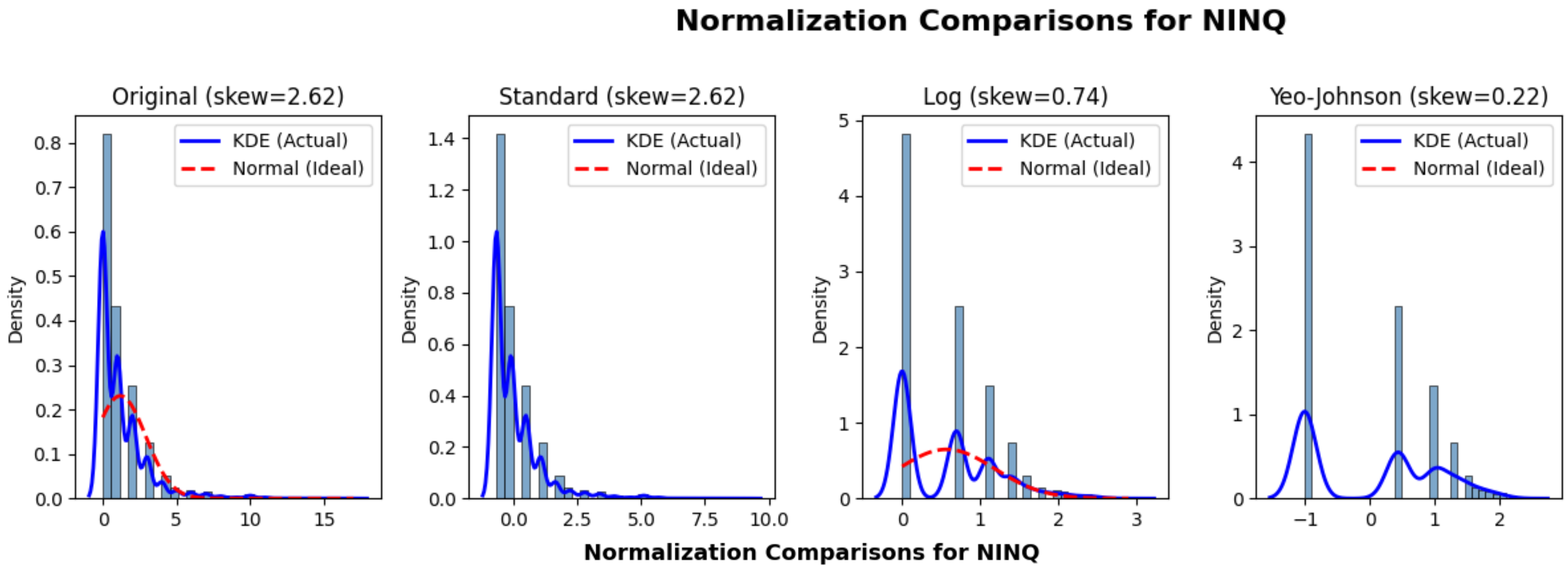


Normalization Comparisons for CLAGE

Method	Skew
Original	1.34
StandardScaler	1.34
Log	-1.52
Yeo-Johnson	0.06

- **Yeo-Johnson** – best score, near-normal distribution with **skew** of **0.06**

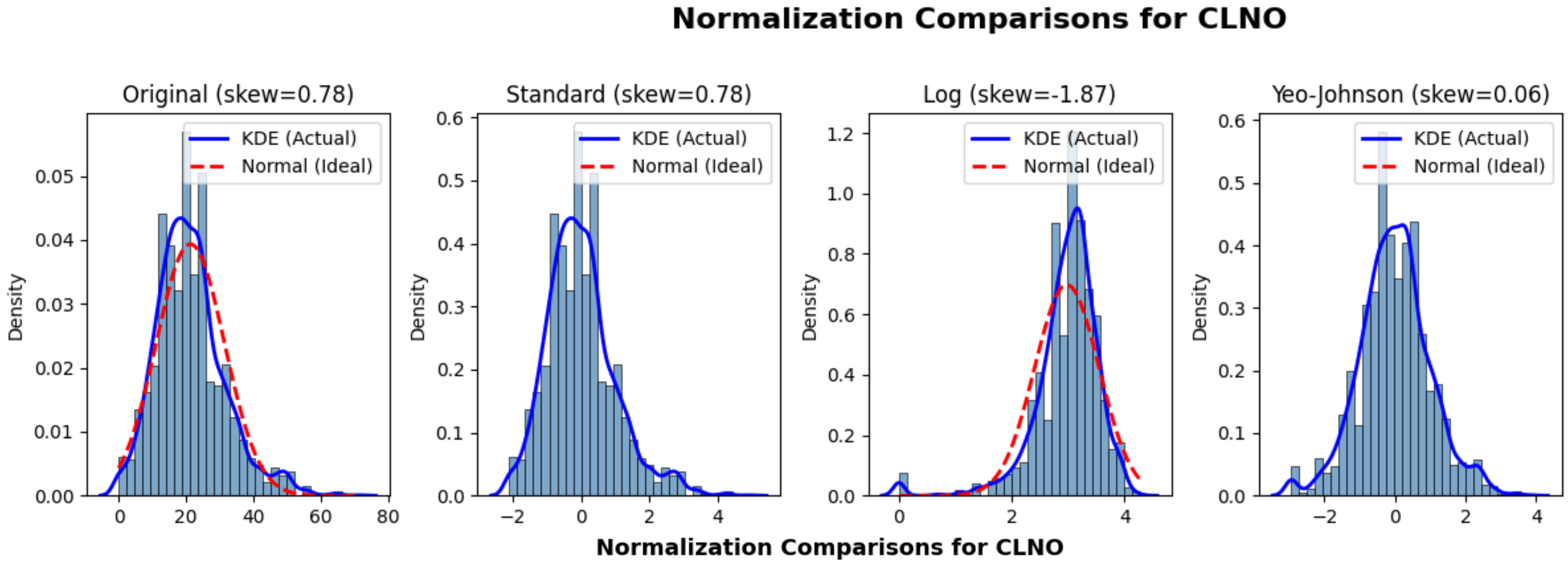
(EDA) – Complete Statistical Analysis



Method	Skew
Original	2.62
StandardScaler	2.62
Log	0.74
Yeo-Johnson	0.22

- **Yeo-Johnson**– best score with **skew** of **0.22**

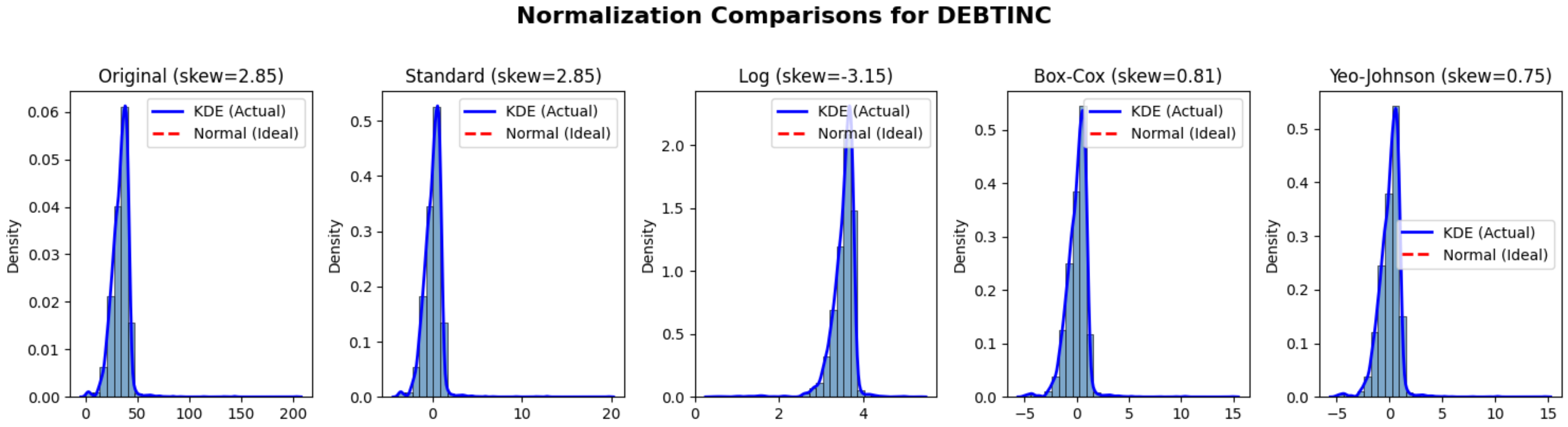
(EDA) – Complete Statistical Analysis



Method	Skew
Original	0.78
StandardScaler	0.78
Log	-1.87
Yeo-Johnson	0.06

- **Yeo-Johnson** – best score, near-normal distribution with **skew** of **0.06**

(EDA) – Complete Statistical Analysis



Normalization Comparisons for DEBTINC

Method	Skew
Original	2.85
StandardScaler	2.85
Log	-3.15
Box-Cox	0.81
Yeo-Johnson	0.75

- **Yeo-Johnson** – best score with **skew** of **0.75**

Statistics after Preprocessing:

- After performing preprocessing steps such as missing value replacement, nominal-to-numerical conversion, and normalization, the statistics were again used to verify the changes and improvements. This helped confirm that data cleaning and transformations were applied correctly and that the dataset was now ready for modeling.
- Based on the information available in this slide deck, it is impossible to verify what normalization techniques were used in feature engineering.
- Analysis here is best guess. For more detailed insights, review to the subsections entitled: **Statistical Analysis – Normalized Features**

Feature Engineering – Derived Features:

Raw features don't always capture the full picture (in this case of borrower risks). Derived features combine existing data to create metrics that are more directly tied to the outcome being predicted.

Benefits:

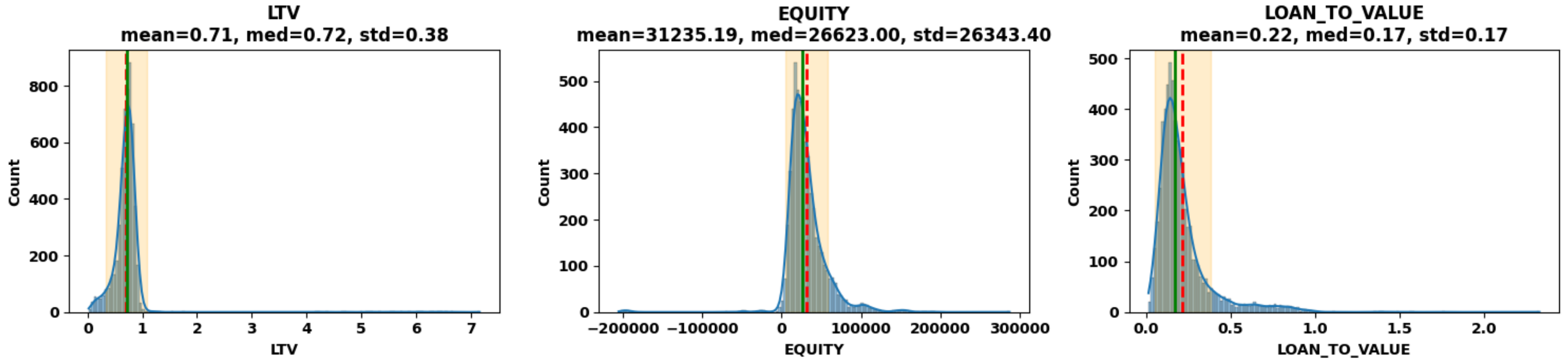
- Captures relationships amongst features
- Domain Knowledge
- Stronger predictive signals
- Better Interpretability
- Expose Hidden Patterns
- Improves Model Performance

Suggested Derived Features:

- **LTV** (Loan to Value Ratio) – MORTDUE/VALUE
- **EQUITY** – Value – MORTDUE
- **LOAN_TO_VALUE** – LOAN/VALUE

Feature Engineering – Derived Features:

Histograms for Numeric Features of Loan Default Prediction :: Derived Features (Batch 1/1)



Numeric Features Statistics

Feature	Min	Max	Mean	Median	Mode	Std	Range	Quantile	Skew	Kurtosis
LTV	0.02	7.16	0.71	0.72	[0.8]	0.38	7.14	[0.6230004794490694, 0.7188...	10.61	145.62
EQUITY	-205445.00	287300.00	31235.19	26623.00	[18000.0]	26343.40	492745.00	[17355.0, 26623.0, 39757.0,...	-0.56	22.10
LOAN_TO_VALUE	0.02	2.33	0.22	0.17	[0.15]	0.17	2.31	[0.12232428987971776, 0.170...	2.87	13.46

Derived Features Summary

- Generated Derived Features display right skew and outliers. They should be normalized before including.
- Additional derived features should be explored to find hidden patterns and relationships not visible in original data.

Statistics after Preprocessing:

CLAGE

Summary

Number

Missing: 0.00%

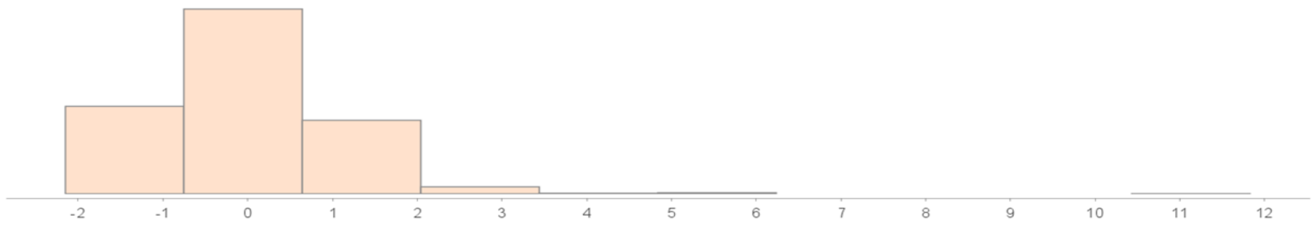
Infinite: 0.00%

ID-ness: 0.02%

Stability: 5.17%

Valid: 94.82%

Distribution



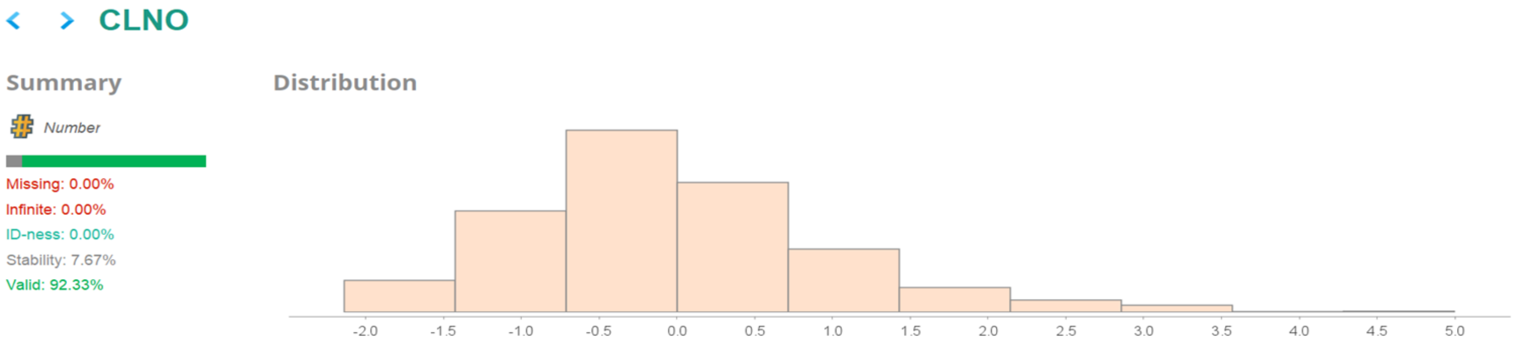
Statistics

Name	Value
Minimum	-2.151
Maximum	11.829
Average	0.000
Standard Deviation	1.000

Observations:

- Missing values imputed.
- Distribution more normalized
- Outliers compressed

Statistics after Preprocessing:



Statistics

Name	Value
Minimum	-2.140
Maximum	4.997
Average	0.000
Standard Deviation	1.000

Observations:

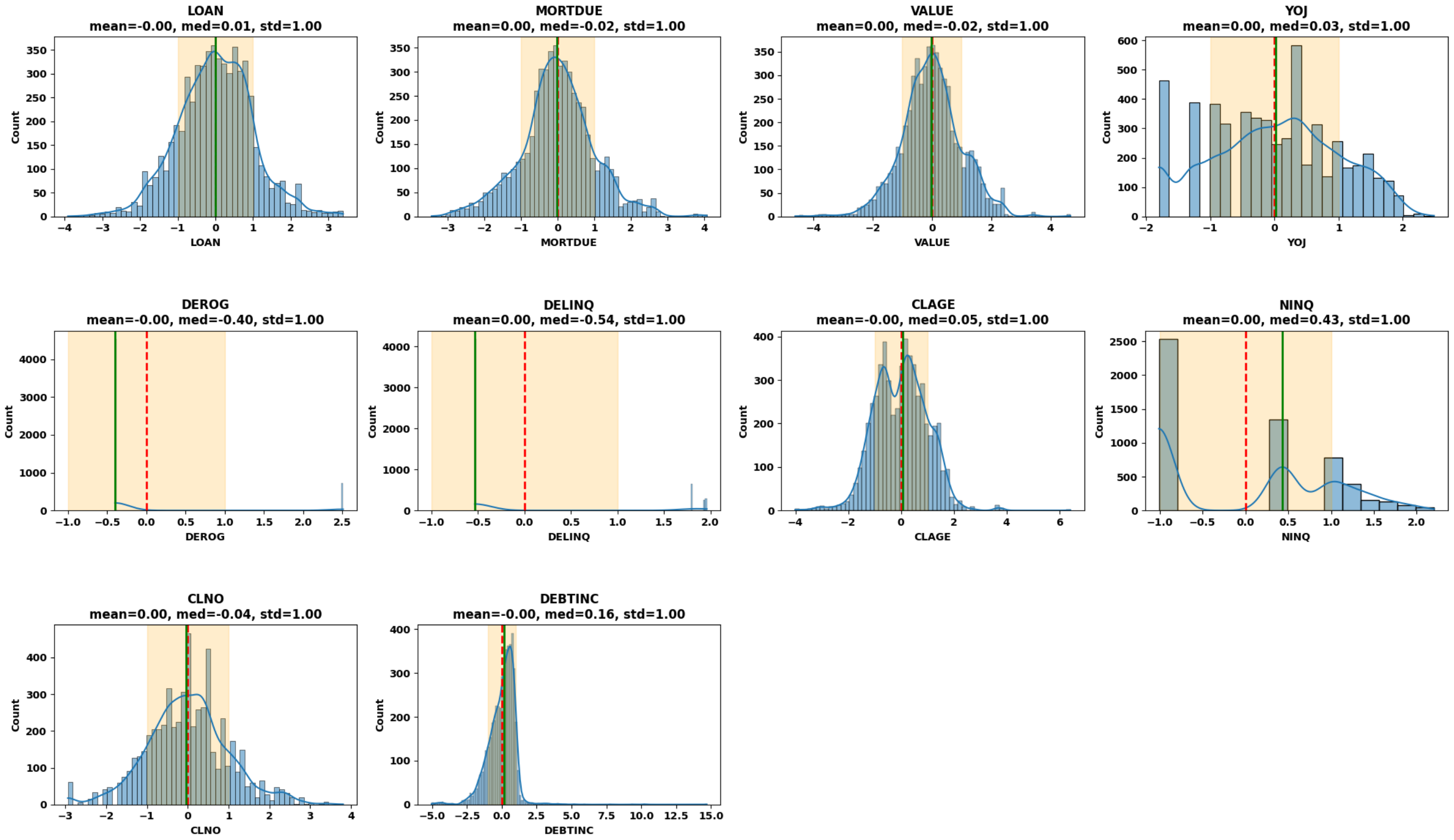
- Missing values imputed.
- Distribution more normalized
- Some outliers still exist

Statistics after Preprocessing:

- After performing preprocessing steps such as missing value replacement, nominal-to-numerical conversion, and normalization, the statistics were again used to verify the changes and improvements. This helped confirm that data cleaning and transformations were applied correctly and that the dataset was now ready for modeling.
- Based on the information available in this slide deck, it is impossible to verify what normalization techniques were used in feature engineering.
- Analysis here is best guess. For more detailed insights, review to the subsections entitled: **Statistical Analysis – Normalized Features**

(EDA) – Statistical Analysis – Normalized Features

Histograms for Numeric Features of Loan Default Prediction :: Normalized (Batch 1/1)



(EDA) – Statistical Analysis – Normalized Features

Numeric Features Statistics

Feature	Min	Max	Mean	Median	Mode	Std	Range	Quantile	Skew	Kurtosis
LOAN	-3.92	3.39	-0.00	0.01	[-0.1381151630478908]	1.00	7.31	[-0.6482956726395991, 0.006...	0.01	0.52
MORTDUE	-3.43	4.07	0.00	-0.02	[-0.6874825596089248]	1.00	7.50	[-0.5465742287336288, -0.01...	0.05	0.77
VALUE	-4.61	4.67	0.00	-0.02	[-0.8022544228748186]	1.00	9.28	[-0.6127659215276012, -0.01...	0.01	0.90
YOJ	-1.80	2.49	0.00	0.03	[-1.8007719010441925]	1.00	4.29	[-0.7063284847342738, 0.025...	-0.07	-0.78
DEROG	-0.40	2.51	-0.00	-0.40	[-0.40018597729696453]	1.00	2.91	[-0.40018597729696453, -0.4...	2.10	2.41
DELINQ	-0.54	1.96	0.00	-0.54	[-0.5357755820929123]	1.00	2.50	[-0.5357755820929123, -0.53...	1.33	-0.21
CLAGE	-4.01	6.43	-0.00	0.05	[-0.9115351930218034, 0.434...	1.00	10.43	[-0.7187314871508779, 0.053...	0.06	0.88
NINQ	-1.01	2.20	0.00	0.43	[-1.0093593188755363]	1.00	3.21	[-1.0093593188755363, 0.430...	0.22	-1.50
CLNO	-2.94	3.81	0.00	-0.04	[-0.4506302371339534]	1.00	6.75	[-0.5602989859696038, -0.03...	0.06	0.62
DEBTINC	-5.04	14.70	-0.00	0.16	[-5.040196536763477, -4.986...	1.00	19.74	[-0.5216928731921547, 0.156...	0.75	19.53

- **LOAN** – Perfect normal distribution, minimal skew
- **MORTDUE** – Perfect normal distribution, minimal skew
- **VALUE** – Close to normal distribution, reduced skew/kurtosis
- **YOJ** – Already normal but now reduced skew/kurtosis
- **DROG** – Zero inflated, binarize?
- **DELINQ** – Zero inflated, binarize?
- **CLAGE** – Normal, but bimodal? Reduced skew/kurtosis
- **NINQ** – Better. Reduced skew/kurtosis.
- **CLNO** – Already normal but now reduced skew/kurtosis
- **DEBTINC** – Normal-ish. Kurtosis reduced, but still high

Scenario 1: Basic Neural Network

Metric	Accuracy	Kappa	Weighted Recall	Weighted Precision
Basic Neural Network	84.96	0.507	74.20	76.71

- **Definitions**
  - **Accuracy** – percentage of model predictions that are correct
  - **Kappa** – **(0<->1.0)** how much model is better than random guessing
  - **Weighted Recall**
    - **Recall** – percentage of defaults actually caught by model
    - **Weighted Recall** – average recall across both classes, weighted by class size
  - **Weighted Precision**
    - **Precision** – of all predicted defaults, what percentage were actually correct?
    - **Weighted Precision** – average precision across both classes, weighted by class size.
- **Why Weight?**
  - To account for class imbalance, weighting gives more importance to majority (non-default) so scores are reflection of population proportions

Scenario 1: Basic Neural Network

- Analysis
  - Accuracy
    - ≈**85%** is roughly **5%** better than **80%** baseline (if model predicted non-default for every row)
    - Model is learning
    - Not overfitting
  - Kappa
    - Moderate agreement, better than random guessing (≈**0.5**)
    - Room for improvement through tuning
    - Confirms weighting offsets class imbalances
  - Weighted Recall
    - Model catches ≈**74%** of actual cases
    - Lower than **Accuracy**, missing some defaults
  - Weighted Precision
    - Model is ≈**77%** correct when it makes a prediction (weighted)
    - Average precision across both classes, weighted by class size.
    - Room for improvement via parameter tuning (or additional derived features/polynomial features)

Scenario 2: Neural Network With Parameter Tuning

Metric	Accuracy	Kappa	Weighted Recall	Weighted Precision
Neural Network with Parameter Tuning	87.92	0.583	76.37	83.17

Scenario 2: Neural Network With Parameter Tuning

- Analysis
  - Accuracy
    - ≈**88%** is roughly **8%** better than **80%** baseline (if model predicted non-default for every row)
    - Model is learning better (≈**3%** improvement over **Basic NN**)
    - Tuning improved performance
    - Not overfitting
  - Kappa
    - Better agreement, approaching good range (>**0.6**)
    - ≈**15%** improvement over **Basic NN**
  - Weighted Recall
    - Model catches ≈**76%** of actual cases – more cases overall
    - ≈**2%** improvement over **Basic NN**
    - Tuning responsible for slight improvement
  - Weighted Precision
    - Model is ≈**83%** correct when it makes a prediction (weighted)
    - Significant improvement over **Basic NN**
    - Tuning most benefited this metric

Scenario 3: Neural Network Using Grid Search

Metric	Accuracy	Kappa	Weighted Recall	Weighted Precision
Grid Search Optimized Model	88.120	0.546	72.24	89.25

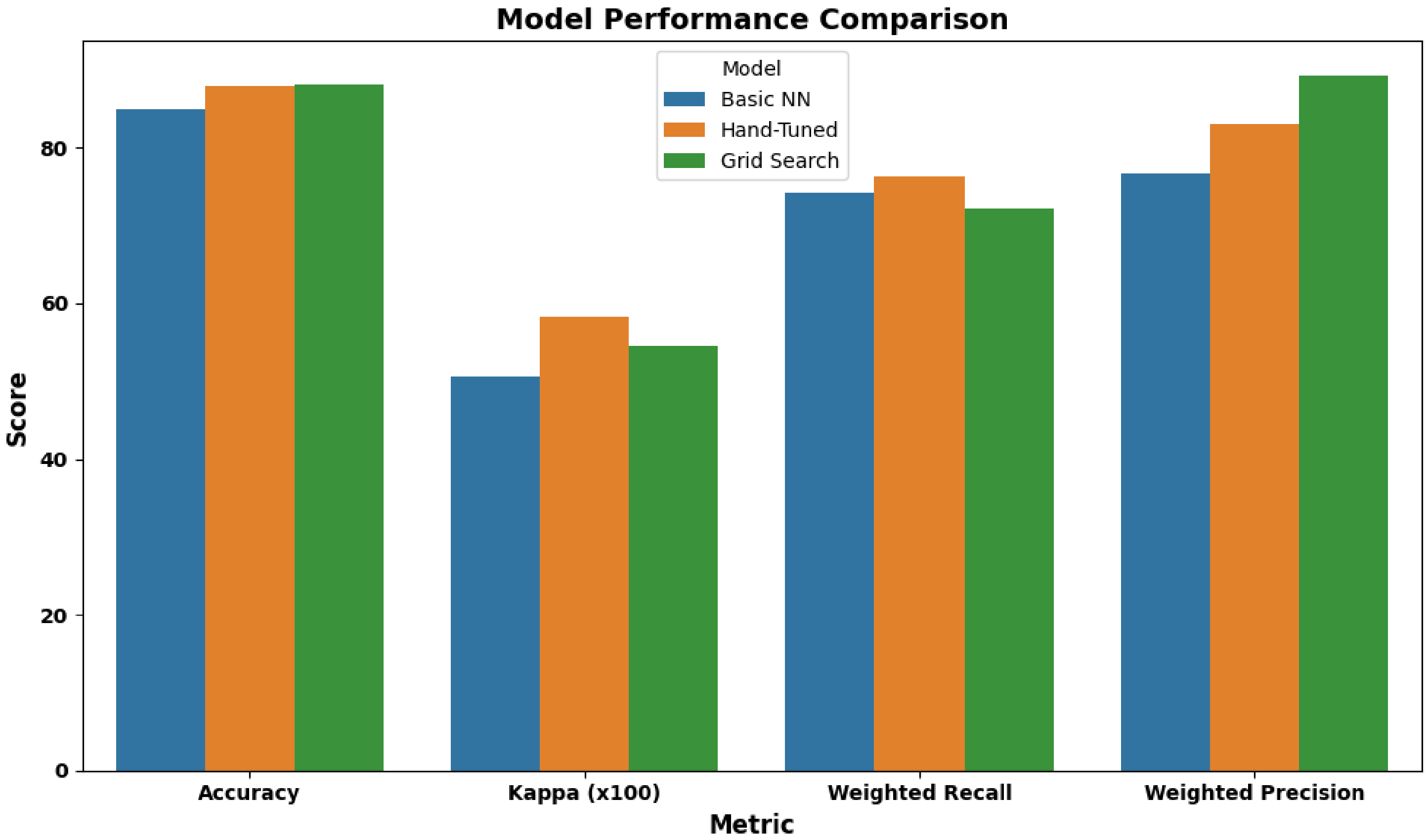
Scenario 3: Neural Network Using Grid Search

- Analysis
  - Accuracy
    - ≈**88%** is roughly **8%** better than **80%** baseline
    - Model is learning slightly better over tuned **Basic NN**
    - Best overall accuracy
    - Not overfitting
  - Kappa
    - Decreased from hand tuned parameters
    - Less balanced (weighted) prediction
  - Weighted Recall
    - Lowest overall
    - Catching fewer cases overall
    - Trade-off for higher precision
  - Weighted Precision
    - Model is ≈**89%** correct when it makes a prediction (weighted)
    - Significant improvement over tuned **Basic NN**
    - **GridSearchCV** most benefited this metric

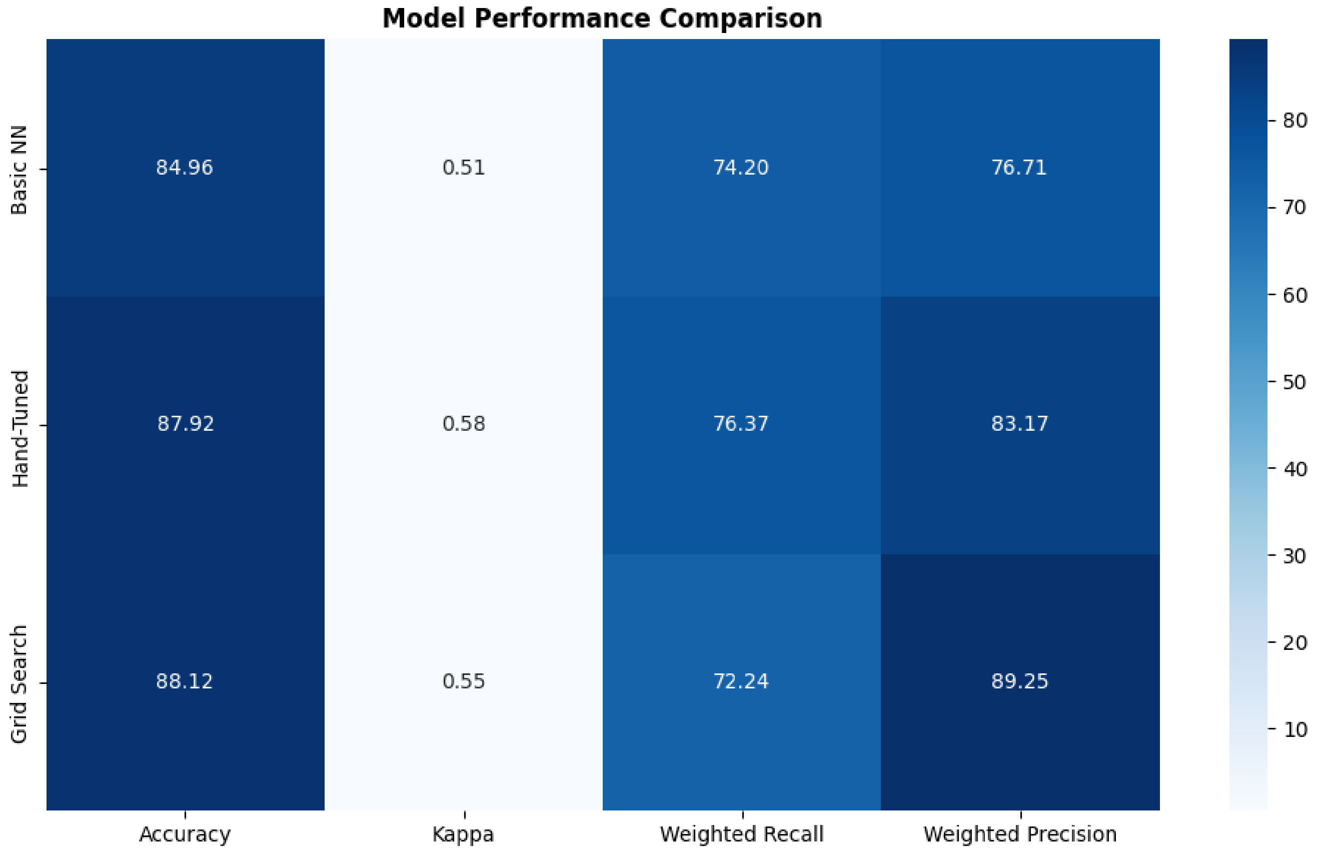
Model Performance Summary

Metric	Accuracy	Kappa	Weighted Recall	Weighted Precision
Without parameter tuning	84.96	0.507	74.20	76.71
With parameter tuning	87.92	0.583	76.37	83.17
Grid Search Optimized Model	88.120	0.546	72.24	89.25

Model Performance Summary



Model Performance Summary



Model Performance Summary

Comparison Analysis

Hand-Tuned (parameter tuned) Model is the best for predicting loan defaults (even though **GridSearchCV** has higher weighted precision). **GridSearchCV** sacrificed recall for higher precision, which is wrong for the business requirements.

Weighted Recall is the most important determination metric because it measures the actual number of loan defaults. From a business perspective, the cost of a loan defaulting far exceeds that of (potentially good) applicant rejections. Higher recall == fewer bad loans (fewer financial losses).

Hand-Tuned also has the highest **Kappa** scores as it has the most balanced prediction across models.

Other Notes:

All models exceed the 80% baseline, proving they were actually learning rather than just predicting.

None of the models exhibited overfitting behaviours.

General Business Advice

Use the **Hand-Tuned Model** to red flag higher risk applications for a manual review. Do not solely rely upon the model for making blanket automated decisions.

Conclusions and Recommendations

- Please mention actionable insights & recommendations

Great Learning

Happy Learning !