A Data Mining Approach to Credit Default Risk Prediction

Abstract— This study develops a robust statistical framework for predicting credit default risk using approximately 32,000 applications. Following missing data imputation and outlier treatment, the research utilized EDA and dimensionality reduction (PCA/t-SNE) to optimize the feature space. Modeling involved comparing baseline Logistic Regression against regularized variants (Ridge, Lasso, Elastic Net), with SMOTE addressing class imbalance. The Ridge Logistic Regression model emerged as the optimal solution, achieving a ROC-AUC of ~98% and high coefficient stability. Analysis identified lower credit grades, higher interest rates, and larger loan amounts as primary risk drivers, while income and home ownership were key mitigating factors . These findings offer an interpretable decision support tool for early credit risk detection in financial institutions.

Keywords— Credit Risk Assessment, Logistic Regression, Default Prediction, Imbalanced Data, SMOTE, Regularization, Principal Component Analysis, Financial Data Mining.

I. INTRODUCTION

Credit default risk is a serious problem confronting banking institutions and peer-to-peer (P2P) lending platforms. When a borrower defaults on the loan, the lender faces substantial financial losses. In the context of this work (i.e., we assume it to be collected using any online lending service (e.g., Lending Club)), the primary goal is to estimate whether borrowers will default. The dataset, titled LoanDataset.csv, contains all attributes of each customer and loan application. The attributes used are demographic information (e.g., age, income), financial characteristics (homeownership status, employment length, loan purpose, credit grade, loan amount, interest rate, loan term, prior default history, and credit history length), and a dependent variable Current_loan_status. This dependent variable distinguishes between DEFAULT and NO DEFAULT loan status. The research will develop a powerful statistical model with this target variable.

In this study, the data mining process has been implemented end-to-end. After this introduction, we give the problem definition and general dataset information. In the Methods section, we describe data cleaning work, exploratory and confirmatory analyses, feature engineering, and modeling. The Results section shows the comparative performance of different models in the form of tables, indicates interpretation and coefficient analysis of the best model. Lastly, the Conclusion summarizes the findings and provides recommendations for further studies.

II. DATA CLEANING & PREPROCESSING

As the first step of the study, the customer ID variable was removed from the dataset because it would not be used in the later stages of the project and did not provide an analytical contribution. Subsequently, the data types of the variables were examined, and incorrect definitions were corrected. Specifically, it was determined that the loan amount variable was perceived as *object* due to the "£" symbol in its values; the variable was converted to *float* type by cleaning the relevant symbol. Similarly, the customer income variable was also converted from *object* type to *float* type.

The presence of duplicate observations in the dataset was checked, and 6 duplicate observations identified were removed from the dataset. For categorical variables, spelling consistency was checked, possible spelling differences and inconsistencies were examined, and necessary adjustments were made. Additionally, observations evaluated as logically inconsistent in the basic context were corrected or removed from the dataset.

Considering realistic constraints regarding credit usage, 11 observations containing individuals outside the 18–80 age range were removed from the dataset. In addition, in the check performed under the assumption that individuals' credit

history (on a yearly basis) cannot date back to before 18 years of age,

$$customer\ age\ -\ credit\ history\ length\ <\ 18$$

775 observations (approximately 2.6% of the dataset) satisfying the condition were removed from the dataset. The assumption that the employment duration considered in customers' credit assessment must start after legal adulthood was also checked, and it was observed that all observations satisfied this condition.

In order to prevent **data leakage** in subsequent modeling stages, the dataset was split into **train** and **test** subsets using the **hold-out method** at this point. During the split, by using the *stratify* parameter, it was ensured that the class distribution of the **response variable** was preserved in the train and test sets in accordance with the percentages in the original dataset.

Due to the nature of financial data exhibiting skewed distributions, a commonly preferred approach in the literature was followed, and an outlier analysis was performed for the income and loan amount variables using the 1%–99% percentile range. During this process, potential outliers were flagged; however, as it was evaluated that high income and high loan amounts could be probable and meaningful within the context of credit data and do not constitute erroneous observations on their own, the observations in question were kept in the dataset.

In order to examine the structure of missing values in the dataset, primarily correlation heatmap and missingness matrix visualizations were created.

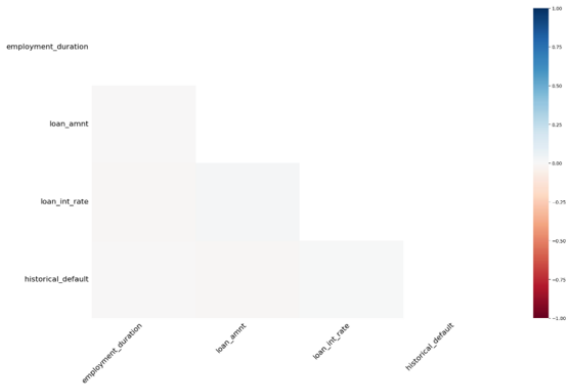

Figure 1 – Missing Correlation Heatmap

Correlation heatmap results showed that there was no significant correlation between the variables with missing values.

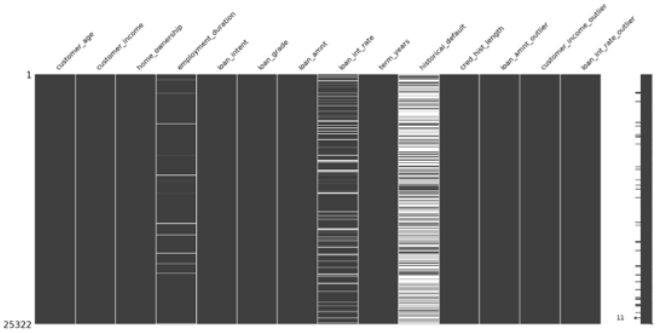

Figure 2 – Missingness Matrix

In addition, when the **missingness matrix** was examined, it was observed that the missing values did not depend on a specific observational pattern and appeared to be randomly distributed within the dataset.

In order to support these visual examinations and to decide more formally whether the missing values are **MCAR (Missing Completely At Random)** or **MAR (Missing At Random)**, statistical tests were applied for each variable showing missingness. In this context, a missingness indicator (missing, 1 / not missing, 0) was created for the variables where missingness was observed; based on this indicator, the distributions of other variables were compared. The **Welch two-sample t-test** was used for numeric variables, and the **Chi-square test** was used for categorical variables.

The basic logic used in the **MCAR–MAR** distinction is as follows: If the missingness in a variable is completely at random (MCAR), the distribution of other variables in observations where this variable is missing should not statistically differ from observations where it is not missing. For example, in the case that the missing values for the loan_int_rate variable are MCAR, no significant difference is expected between the income average of individuals with missing interest rates and the income average of individuals without missing interest rates. This situation is supported by obtaining **p > 0.05** as a result of a **two-sample t-test** and is interpreted as the missingness being completely at random.

In contrast, if individuals with missing interest rates are observed to be systematically different in terms of another variable (for example, if the loan_grade values of individuals with missing loan_int_rate are statistically significantly lower than those without missing values and **p < 0.05** is obtained), in this case, the missingness is not random. The fact that missingness depends on another observed variable causes us to classify this situation as MAR (Missing At Random). As a result of the statistical tests performed, it was determined that the missingness in the **loan_int_rate** variable has an **MCAR (Missing Completely At Random)** structure; conversely, it was determined that the missingness in the **employment_duration** variable occurred under the **MAR (Missing At Random)** mechanism. This distinction has been the primary determinant in defining the **imputation** strategy to be applied in handling missing values.

Due to the presence of approximately 20,000 missing observations in the **historical_default** variable and the categorical nature of this variable, it was evaluated that direct deletion of missing values would lead to significant loss of information. Furthermore, with the assumption that the missingness could point to unobservable past credit behaviors, it was deemed appropriate to treat the missing values in this variable as a separate category, and the missing observations were filled under the "**Unknown**" class.

For the **loan_int_rate** variable, which was determined to have an **MCAR** structure, simple imputation methods were considered sufficient. Taking into account the skewed distribution of financial data and sensitivity to outliers, this variable was filled with the **median** value using **Simple Imputer**. This approach was preferred because it is less affected by extreme values compared to mean imputation.

For the **employment_duration** variable, which has an **MAR** structure, a more advanced imputation method was applied because the missingness is related to other observed variables. In this context, the **K-Nearest Neighbors (KNN) imputation** approach, which considers similarity between observations and is distance-based, was adopted. During the imputation process, **employment_duration** values were estimated based on similar observations by considering 5 neighbors (k = 5) for each missing observation.

In this way, by selecting imputation methods suitable for the missing value mechanism, both the loss of information in the dataset was minimized and the risk of **bias** that could occur in later modeling stages was reduced.

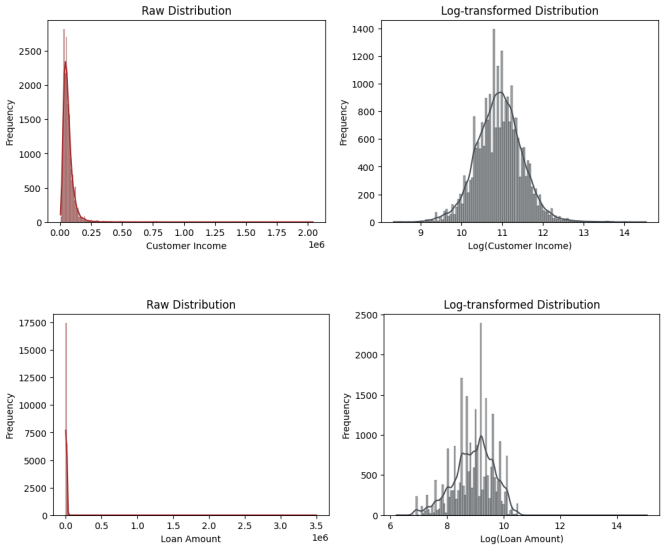

Figure 3 – Density Plots

Due to the nature of financial data exhibiting **right-skewed** distributions, in order to reduce the extreme skewness in the **income** and **loan_amount** variables, a widely preferred approach in the literature, **logarithmic transformation (log transformation)**, was applied. With this transformation, it was aimed to make the distributions of the variables more symmetric, reduce the impact of outliers, and minimize assumption violations in subsequent modeling stages.

III. Data Description & Exploratory Data Analysis (EDA)

A. Data Description

Variable	Description
customer_id	Unique identifier for each customer
customer_age	Age of the customer
customer_income	Annual income of the customer
home_ownership	Home ownership status
employment_duration	Duration of employment in months
loan_intent	Purpose of the loan
loan_grade	Grade assigned to the loan
loan_amnt	Loan amount requested
loan_int_rate	Interest rate of the loan
loan_int_rate	Interest rate of the loan
term_years	Loan term in years
historical_default	Indicates if the customer has a history of default
cred_hist_length	Length of the customer's credit history in years
Current_loan_status	Current status of the loan

Table 1 – Data Description

B. Exploratory Data Analysis

Within the scope of exploratory data analysis, primarily the **univariate distribution** of the variables was examined. For numerical variables, the central tendency and dispersion characteristics of the variables were evaluated by calculating basic descriptive statistics such as **mean, median, standard deviation, and minimum and maximum values.**

Variables	Count	Mean	Std	min	25%	50%	75%	Max
Customer Age	25321	27.83	6.22	20	23	26	30	80
Customer Income	25321	66,223.27	53,502.42	4200	39,000	55,200	79,992	2,039,784
Employment Duration	25321	4.78	4.02	0	2	4	7	41
Loan Amount	25321	9,799.32	23,666.02	500	5,000	8,000	12,375	3,500,000
Loan Interest Rate	25321	11.01	3.08	5.42	8.5	11	13.11	23.22
Term Years	25321	4.78	2.45	1	3	4	7	10
Credit History Length	25321	5.78	4.07	2	3	4	8	30

Table 2 – Descriptive Statistics

When Table 1 is examined, it is observed that the dataset has a **generally young customer profile**. The median value of customer ages is 26, and observations are primarily concentrated in the 20–30 age range. Distinct differences between the mean and median in the **Customer income** and **loan amount** variables, along with high maximum values, show that these variables have a **right-skewed** distribution. This situation is consistent with the income and loan amount heterogeneity commonly observed in financial data.

The fact that the **Employment duration** and **credit history length** variables have relatively low median values indicates that a significant portion of the individuals in the dataset have a limited work and credit history. The **Loan interest rate** variable, on the other hand, is distributed over a narrower range, indicating that interest rates have a relatively homogeneous structure.

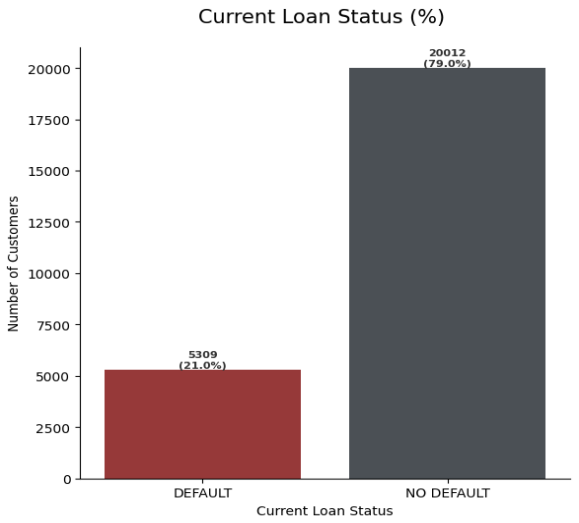

Figure 4 – Bar Plot of Response

When the figure is examined, it is observed that the response variable **Current Loan Status** has an **imbalanced** distribution. Approximately **79%** of the observations belong to the **"NO DEFAULT"** class, and **21%** belong to the **"DEFAULT"** class. This indicates that sampling methods such as SMOTE can be utilized during the modeling stage.

Within the scope of exploratory data analysis, following the **univariate examination**, a transition was made to bivariate examination in order to evaluate the relationships between variables, and first, **correlation analysis** was performed.

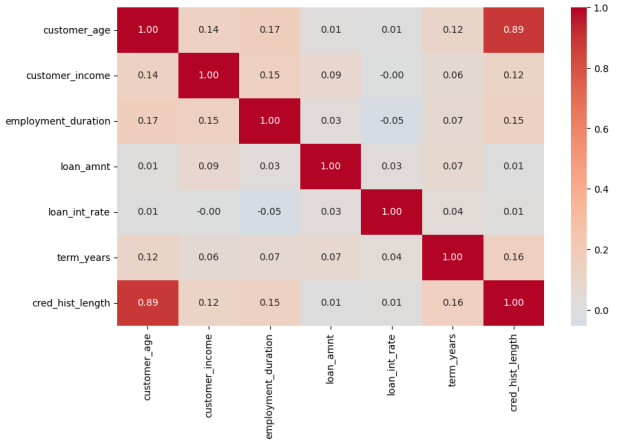

Figure 5 – Correlation Heatmap

When the correlation matrix is examined, it is observed that there are **weak linear relationships** between the vast majority of variables. The most notable relationship is the **high positive correlation (r ≈ 0.89)** observed between **customer_age** and **credit_history_length**; this situation is consistent with the natural lengthening of the credit history duration as age increases. Other than this, no correlation strong enough to negatively affect the modeling process was observed; this indicates that there is **no serious multicollinearity problem** in the dataset.

Figure 6 – Hexbin Density of Income vs Loan

The density distribution in the graph shows that there is a positive and logically consistent relationship between **loan amount** and **customer income**; the dark-colored **'sweet spot'** region in the center of the data proves that the bank's main portfolio is clustered in the upper-middle income group and medium-scale loan amounts. In this visualization, where variance is stabilized thanks to **logarithmic transformation**, the thinning out of hexagons at the extremities indicates that **outliers** do not disturb the main trend, while the alignment of density along a linear line expresses that **financial risk appetite** increases in a balanced manner in parallel with income.

1. Is there a relationship between **loan grade** and **default status**?

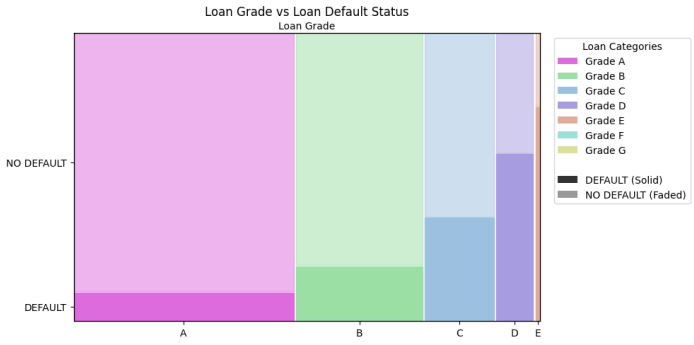

Figure 7 – Mosaic Plot of Loan Grade

The graph shows that there is a distinct relationship between **loan grade** and **loan default status**. While the **NO DEFAULT** rate is higher in individuals with low-risk credit grades (A–B), it is observed that the **DEFAULT** rate increases as the credit grade decreases (C–D–E). This finding shows that the **loan grade** variable is a strong and distinctive variable in explaining loan default.

2. Does **customer income** vary according to **loan default status**?

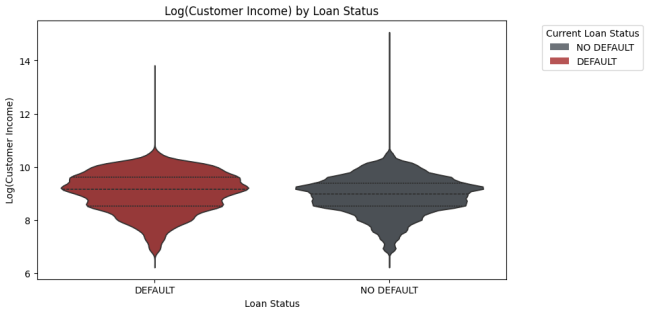

Figure 8 – Violin Plot of Income by Response

The graph shows that the income distributions of the **DEFAULT** and **NO DEFAULT** groups overlap significantly, but the median income in the **NO DEFAULT** group is relatively higher. This situation indicates that income level may have a limited but potentially significant impact on default risk, and an appropriate **two-sample test** is required to statistically verify this difference.

3. Does the **loan amount** vary according to **loan default status**?

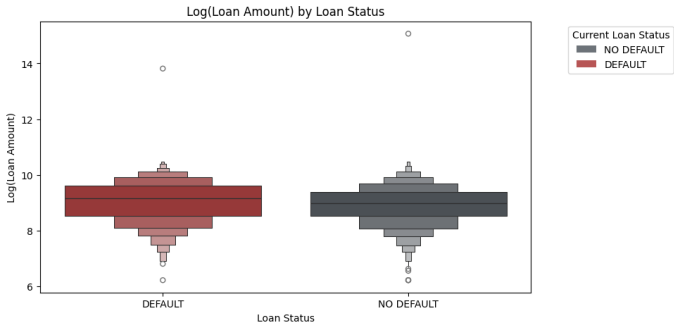

Figure 9 – Box Plot of Loan by Response

When the graph is examined, it is observed that the loan amount distributions of the **DEFAULT** and **NO DEFAULT** groups **overlap** to a large extent, but the median of the loan amounts of **defaulted customers** is **relatively higher**. Due to the persistence of outliers and skewness in the distributions, it will be appropriate to prefer formal tests to statistically verify this difference.

4. Is there a statistically significant relationship between **home ownership status** and **loan default status** ?

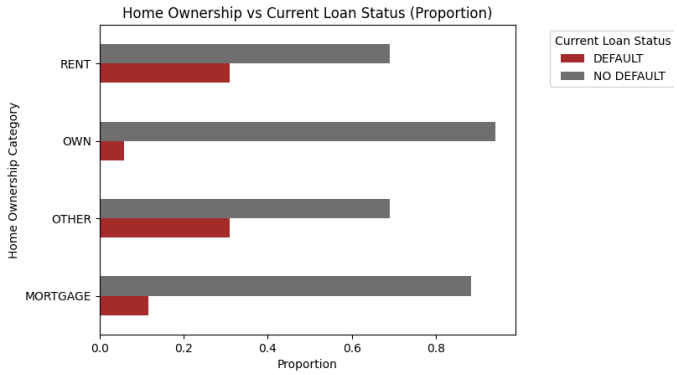

Figure 10 – Bar Plot of Home Ownership

When the graph is examined, it is observed that there are distinct differences between **home ownership status** and **loan default rates**. While the **NO DEFAULT** rate is high in the **OWN** and **MORTGAGE** categories, it is noteworthy that the **DEFAULT** rates are relatively higher in the **RENT** and **OTHER** categories. This finding indicates that home ownership could be a factor that reduces the risk of loan default.

IV. METHODOLOGY

As a result of the data cleaning, exploratory data analyses performed in previous stages, along with feature engineering and dimensionality reduction steps, a suitable and consistent dataset for modeling has been obtained. Since the dependent variable focused on in the project is categorical (binary classification: default vs. non-payment), **logistic regression** was preferred as the primary modeling technique. Logistic regression is a widely used statistical model to explain the probability of an event (in our case, the probability of credit default). Before building the model, the assumptions of logistic regression were checked.

Before proceeding to the modeling stage, the fundamental assumptions of the logistic regression model were comprehensively evaluated in terms of **multicollinearity**, **logit linearity**, and **influential observations**.

Multicollinearity

VIF values calculated for numerical and encoded categorical variables remained within **acceptable limits for all variables**. The highest VIF values were observed in the loan_grade_ord (VIF ≈ 3.31) and loan_int_rate (VIF ≈ 2.92) variables, and these levels do not point to a serious multicollinearity problem. The fact that the VIF values of all other variables are **below 2** indicates that there is no linear dependency at a level that would disrupt the coefficient estimates in the model.

Linearity in the Logit

The assumption of logit linearity was tested for continuous variables using the **Box–Tidwell approach**. In this context, the contributions of the created $X \times \log(X)$ terms to the model were examined, and it was observed that the assumption of linear logit weakened in some variables. This finding supports why **log transformations**, **feature engineering**, and **dimensionality reduction (PCA)** approaches applied in earlier stages were necessary during the modeling process.

Influence Diagnostics

Leverage (hat values) and **Cook's Distance** metrics were used to analyze influential observations. It was observed that the highest **Cook's D** and **leverage** values examined remained at low levels; thus, no **high-influence observations** were found that disproportionately directed the model. For this reason, no intervention

such as observation deletion or additional weighting was needed.

During the modeling stage, different logistic regression-based approaches were systematically compared using the feature set obtained in previous steps. All models were evaluated under **5-fold Stratified Cross-Validation (shuffle = True)** to account for class imbalance.

Scaling and Pre-processing

Prior to modeling, numerical variables were scaled using **StandardScaler**. Since the dataset contains dummy encoding and a sparse structure, **mean subtraction was not performed** (**with_mean = False**) during scaling; only **variance scaling (dividing by standard deviation)** was applied. This approach was preferred to ensure compatibility with one-hot/dummy encoded variables.

Modeling Approaches

The following modeling scenarios were established and compared within the scope of the study:

- **Baseline Logistic Regression:** The fundamental logistic regression model built using all features was taken as the reference model for comparisons.
- **Cost-Sensitive Logistic Regression:** To address class imbalance, a cost-sensitive logistic regression model was established using **class_weight = balanced**.
- **SMOTE + Logistic Regression:** As a sampling-based approach, **SMOTE** was applied to the training data, followed by the training of a logistic regression model. The SMOTE process was performed within the cross-validation loop.
- **PCA + Logistic Regression:** To evaluate the effect of dimensionality reduction, PCA was used in conjunction with logistic regression. The **PCA component count** was selected from pre-determined candidate values via **GridSearchCV**, using **ROC-AUC** as the performance metric.
- **Regularized Logistic Regression Models:** To control overfitting and increase model generalizability, the following regularized models were established:
    - **Ridge (L2) Logistic Regression**
    - **Lasso (L1) Logistic Regression**
    - **Elastic Net Logistic Regression**
- In these models, regularization parameters (**C** and **l1_ratio**) were optimized using **GridSearchCV**.
- **Embedded Feature Selection (L1-Selected):** Variables with non-zero coefficients were selected using the best Lasso model; a **plain logistic regression** model was then re-established using these selected features.

Evaluation Metrics

Due to the **imbalanced** structure of the target variable, model performance was not evaluated solely based on accuracy. Instead, the following metrics were reported:

- **Sensitivity (Recall – DEFAULT class)**
- **Specificity**
- **F1-score**
- **Cohen's Kappa**
- **ROC-AUC**

These metrics were used together to more robustly evaluate both the models' success in capturing the minority class and their overall discriminative power.

V. Results & Findings

A. Confirmatory Data Analysis

At this stage, relationships observed and patterns emerging during **exploratory data analysis (EDA)** were formally evaluated through **statistical hypothesis tests**. The main purpose of **CDA** is to reveal whether the findings obtained during the **EDA** process are coincidental or statistically significant and to quantitatively verify the relationships between variables. Accordingly, appropriate tests were applied depending on the type and distribution characteristics of the variables; the obtained results were interpreted within the framework of the relevant hypotheses.

1. Is there a statistically significant relationship between **loan grade** and **loan default status**?

- *H_0:* : **Loan grade** and **loan default status** are independent.
- *H_1*: **Loan grade** and **loan default status** are not independent.

Chi-square test results show that there is a statistically significant and moderate-to-high level relationship between credit grade (**loan_grade**) and loan default status (**Current_loan_status**) ($\chi^2_4 = 3533.45$, $p < 0.001$; *Cramer's V = 0.37*). This finding supports that credit grade is a strong and distinctive variable in explaining default risk.

2. Does **customer income** vary in a statistically significant way according to **loan default status**?

- **H_0**: There is no statistically significant difference between the **income** distributions of defaulted and non-defaulted customers.
- **H_1**: There is a statistically significant difference between the **income** distributions of defaulted and non-defaulted customers.

Prior to the relevant comparison, parametric test assumptions were formally evaluated. The **normality** assumption was tested using the **Shapiro–Wilk test**, and the **homogeneity of variance** assumption was tested using the **Levene test**. In both tests, the **H_0** hypotheses were **rejected**; therefore, it was determined that the income distributions are not normally distributed and that the homogeneity of variance between groups is not satisfied. For these reasons, instead of the independent samples t-test, the **Mann–Whitney U test**, which is less sensitive to assumptions, was preferred. The test results show that there is a statistically significant difference between the income distributions of defaulted and non-defaulted customers (*Mann–Whitney U = 33,542,796.5, $p < 0.05$).* This finding supports that income level is associated with loan default status and can be a significant variable in explaining default risk.

3. Does the **loan amount** vary in a statistically significant way according to **loan default status**?

- **H_0:**: There is no statistically significant difference between the **loan amount** distributions of defaulted and non-defaulted customers.
- **H_1**: There is a statistically significant difference between the **loan amount** distributions of defaulted and non-defaulted customers.

Prior to the relevant comparison, parametric test assumptions were formally tested; it was observed that the **normality** and **homogeneity of variance** assumptions were not satisfied, and the H_0 hypotheses were rejected in both cases. For this reason, instead of parametric tests, the **Mann–Whitney U test** was preferred. Test results show that there is a **statistically significant difference** between the **loan amounts** of defaulted and non-defaulted customers (*Mann–Whitney U = 58,840,985.5, $p < 0.05$*).

4. Is there a statistically significant relationship between home ownership status and loan default status?

- *H_0*: **Home ownership** and **loan default status** are independent.
- *H_1*: **Home ownership** and **loan default status** are not independent.

Chi-square test results show that there is a statistically significant relationship between home ownership status and loan default status ($\chi^2_3 = 1613.50$, $p < 0.001$). The calculated **Cramer's V = 0.25** indicates that this relationship has a moderate effect. This finding supports that home ownership could be a significant factor in explaining loan default risk.

B. Feature Engineering & Dimension Reduction

In this section, **encoding processes** have been applied for categorical variables; using **filtering methods**, variables that contain low information for modeling or that need to be removed due to high correlation have been identified. Additionally, new features (**feature engineering**) have been derived from existing variables, and **dimensionality reduction techniques** such as **PCA** and **t-SNE** have been applied to reduce the data structure to a more compact representation.

Firstly, home_ownership, loan_intent, and historical_default variables have been converted to numerical form using the **dummy encoding** method. To prevent multicollinearity and the use of redundant information in the model, the **N-1 level approach** has been adopted; therefore, **one-hot encoding** was not preferred. Since the loan_grade variable has a structurally **ordinal** quality ($A > B > C > D > E$), **ordinal encoding** was applied for this variable.

In this study, a **filter-based feature selection methodology** was adopted. Within the scope of this approach, the relationships of the variables with the target variable were evaluated using **model-independent measures**. For numerical variables, **Spearman correlation analysis**, which is robust to outliers and can capture monotonic relationships, was applied. Results revealed that the **loan_int_rate** variable has the **strongest positive relationship** with the probability of default, while the **customer_income** variable shows a **negative and statistically significant** relationship with default. While the **employment_duration** and **loan_amount** variables exhibited weak-to-moderate relationships, it was observed that the **customer_age** and **credit_history_length** variables provided quite **weak signals** on their own. Considering that p-values were low due to the large sample size, the contribution of variables to the model was evaluated based on **effect size** rather than **statistical significance**.

Relationships with the target variable for categorical variables were analyzed using the **Cramer's V** measure, again in line with the filter-based approach. As a result of this analysis, it was determined that the **historical_default** variable has a **very strong relationship** with the target variable ($Cramer's\ V \approx 0.75$), the **home_ownership** variable has a **moderate level**, and the **loan_intent** variable exhibits a **weak level** relationship. These findings show that categorical variables provide information with different weights to the modeling process.

As part of the **filter-based selection** process, **inter-feature correlations** among numerical variables were also examined. In this analysis, a high correlation was observed between **customer_age** and **credit_history_length**, and this situation was evaluated as a risk of **information redundancy**. However, in accordance with the objective of the **filter-based** approach, instead of directly eliminating the variables, a new feature was derived to preserve the information content. In this context, the variable

$$credit\_hist\_ratio = \frac{credit\_history\_length}{customer\_age}$$

representing how much of an individual's adult life has passed with a credit history was created; subsequently, the **customer_age** and **credit_history_length** variables were removed from the dataset. With this approach, within the framework of **filter-based feature selection**, both information loss was minimized and the risk of **multicollinearity** was reduced.

Before proceeding to the **dimensionality reduction** step, **z-score standardization** was applied so that variables measured on different scales contribute equally to the analysis. This standardization process was performed **only for the PCA application, and no changes were made to the original dataset**. The scaled data were used only in the dimensionality reduction analyses.

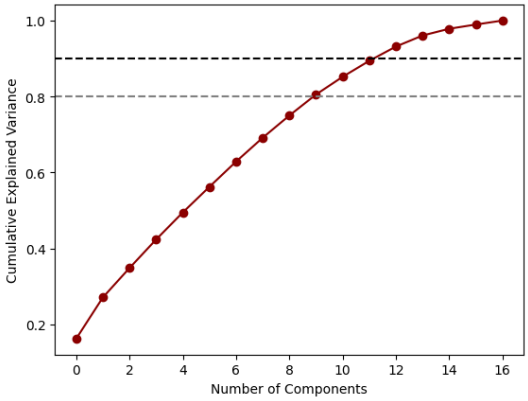

Figure 11 – Scree Plot of PCA

When the **cumulative explained variance graph** is examined, it is observed that the **80%** and **90% explained variance thresholds** are exceeded with approximately **10** and **12 components**, respectively. Within the scope of this study, an **85% explained variance threshold** was preferred to maintain the effect of dimensionality reduction while minimizing information loss. In line with this threshold, it was determined that at least **11 principal components (k = 11)** should be used in

the dataset containing a total of **17 variables**, and a **cumulative explained variance** level of **85.19%** was reached with these components.

Furthermore, it was evaluated that the **interpretability** was insufficient because visualizations made only on the **PC1–PC2 plane** represented a limited portion of the total variance in the dataset. Therefore, a two-component PCA graph was not included in the report. When the **variance ratios explained by each of the determined k = 11 principal components** are examined, it is seen that the first component explains approximately **16.2%** of the total variance, and the second component explains **10.9%**. The contributions of the following components decrease gradually, ranging between **4.7%–7.7%**. This distribution shows that the variance is not concentrated in one or two components; rather, the information is spread across multiple components. This situation supports the idea that interpretations made only via **PC1–PC2** would be insufficient to represent the data structure and that a multi-component PCA representation is more appropriate.

Before proceeding to the **t-SNE analysis**, **standard scaling (z-score standardization)** was applied so that the different scales of the variables would not affect the distance-based structure. This scaling process was performed **only for t-SNE visualization**, and no permanent changes were made to the original dataset. t-SNE was used to visualize **local patterns** in the high-dimensional data structure.

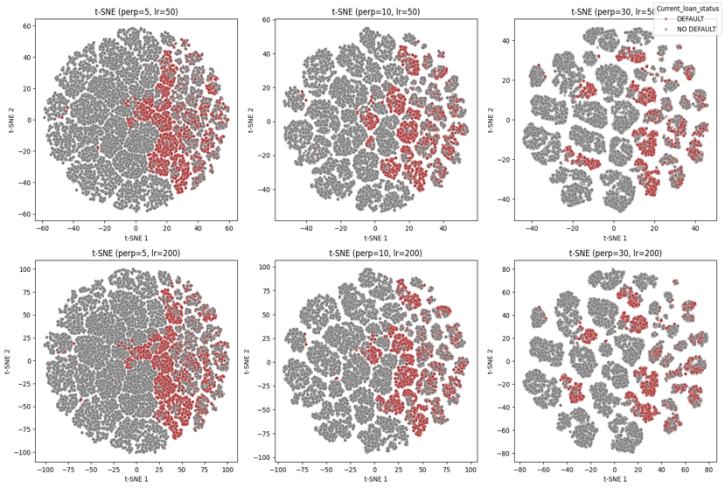

Figure 12 – t-SNE Plots

When t-SNE visualizations obtained under different **perplexity (5, 10, 30)** and **learning rate (50, 200)** values are examined, it is observed that as **perplexity** increases, the data structure separates into a larger number of small and compact clusters. Especially in the visualizations using **perp = 30**, it is noteworthy that observations form **numerous local islands (island-like clusters)**. This situation indicates that the observations in the dataset have a **highly heterogeneous** structure in multi-dimensional space.

At lower perplexity values (**perp = 5**), while observations form more continuous and dense regions, it is observed that some local details are lost. At the **perp = 10** level, a more **interpretable representation** that balances both the local neighborhood structure and the global distribution was obtained. When comparing learning rates, it is seen that the **lr = 50** value provides more stable and organized clustering, whereas the visual dispersion increases at the **lr = 200** value.

In line with these evaluations, the combination of **perp = 10** and **lr = 50** stands out as the setting that reflects the data structure in the most balanced way. However, in all parameter combinations, it is observed that the **DEFAULT** and **NO DEFAULT** classes significantly overlap, showing limited concentrations only in some local areas. This situation supports the finding that the default status cannot be clearly separated in a low-dimensional representation and that the problem has a **multivariate and complex structure**. Therefore, t-SNE results were used not for classification purposes, but as an **exploratory visualization** tool.

C. Model Interpretation

In this study**, Cross-Validation (CV)** was used for model selection, hyperparameter tuning, and generalizability assessment. The tuning parameters for Ridge, Lasso, Elastic Net, and PCA-based models (e.g., C, l1_ratio, PCA component count) were determined using 5-fold Stratified Cross-Validation. The results show that Baseline Logistic Regression, Ridge, Lasso, Elastic Net, and L1-selected models exhibited nearly the same level of high performance on both cross-validation and the test set.

Model	CV_AUC	CV_F1	CV_Sensitivity	CV_Specificity	CV_Kappa	Test_AUC	Test_F1	Test_Sensitivity	Test_Specificity	Test_Kappa
Ridge (best C=10)	0.98	0.845	0.842	0.96	0.804	0.981	0.845	0.849	0.958	0.804
Baseline Logistic	0.98	0.845	0.842	0.96	0.804	0.981	0.846	0.849	0.958	0.805
L1 Selected Plain Logistic (17 feats)	0.98	0.845	0.842	0.96	0.804	0.981	0.846	0.849	0.958	0.805
Lasso (best C=10)	0.98	0.846	0.842	0.96	0.805	0.981	0.845	0.847	0.958	0.803
Elastic Net (C=10)	0.98	0.845	0.842	0.96	0.804	0.981	0.845	0.847	0.958	0.803
Balanced Logistic	0.98	0.809	0.959	0.891	0.749	0.981	0.806	0.961	0.888	0.744
SMOTE+Logistic	0.98	0.815	0.952	0.898	0.757	0.981	0.809	0.953	0.893	0.748
PCA (11)+Logistic	0.889	0.526	0.467	0.918	0.42	0.884	0.511	0.457	0.911	0.4

Table 3 – Model Metric Results

In all of these models, metrics were at levels of **ROC-AUC ≈ 0.98, F1-score ≈ 0.845, Sensitivity ≈ 0.85,** and **Specificity ≈ 0.96**, indicating that defaulting and non-defaulting customers were successfully distinguished. The high consistency between test set metrics and cross-validation results demonstrates that the models do not suffer from **overfitting** and are reliable in terms of generalizability. **Cohen's Kappa** values around **0.80** show that the predictions are significantly superior to random classification.

In the **class_weight = balanced** and **SMOTE + Logistic** approaches applied to address class imbalance, a **significant increase in Sensitivity values** (≈ 0.95–0.96) for the default class was observed. However, this increase occurred alongside a **decrease in Specificity and Kappa values**. This situation indicates that the overall model balance weakened due to an increase in the false positive rate. Therefore, while these approaches may be preferred in scenarios where missing a default is very costly, they were not found optimal for general risk scoring purposes.

The **PCA(11) + Logistic** model exhibited clearly lower performance compared to all other models. Specifically, the drop in **F1-score (≈ 0.51)** and **Sensitivity (≈ 0.46)** values indicates that the dimensionality reduction process performed with PCA caused the loss of some discriminative information related to the target variable in this problem. This result demonstrates that PCA's variance-preservation-oriented structure does not always maximize classification performance.

Model	Test	LR_Stat	df	P-Value	Decision ($\alpha = 0.05$)
Ridge (best C=10)	LRT	18684.673	17	0.00	Reject H_0
Baseline Logistic	LRT	18682.42	17	0.00	Reject H_0
L1 Selected Plain Logistic (17 feats)	LRT	18682.42	17	0.00	Reject H_0
Lasso (best C=10)	LRT	18685.88	17	0.00	Reject H_0
Elastic Net (C=10)	LRT	18684.874	17	0.00	Reject H_0
Balanced Logistic	LRT	16578.938	17	0.00	Reject H_0
SMOTE+Logistic	LRT	NaN	NaN	NaN	Skipped (SMOTE)
PCA (11) +Logistic	LRT	8785.275	11	0.00	Reject H_0

Table 4 – Model Significance Results

The **Likelihood Ratio Test (LRT)** results applied to all models show that **all logistic regression models, including the PCA model (excluding SMOTE), are statistically significant** ($p < 0.001$). This finding confirms that all established models offer significant explanatory power compared to the null model.

In light of these findings, the **Ridge Logistic Regression** model was selected as the final model due to:

- High and stable performance metrics,
- Statistically significant whole-model result (LRT),
- Coefficient stability provided by regularization.

Interpretation of Coefficients (Ridge Model)

Feature	Coef	Feature	Coef
loan_amnt	0.489	employment_duration	-0.123
loan_int_rate	0.067	home_ownership_OWN	-0.471
term_years	0.081	loan_intent_EDUCATION	-0.518
home_ownership_OTHER	0.027	loan_intent_HOMEIMP	-0.206
home_ownership_RENT	0.342	loan_intent_MEDICAL	-0.247
loan_grade_ord	1.039	loan_intent_PERSONAL	-0.463
customer_income	-0.912	loan_intent_VENTURE	-0.555
historical_default_Unknown	-21.343	historical_default_Y	-1.377
Credit_hist_ratio	-0.889		

Table 5 – Ridge Model Coefficents

When the coefficients of the selected Ridge model are examined, loan_grade_ord is the **leading variable increasing default risk the most.** This result shows that as the credit grade decreases (numerical value increases), the probability of default increases significantly. The positive coefficients of loan_amnt and home_ownership_RENT reveal that high loan amounts and being a tenant increase the risk of default. The variables loan_int_rate and term_years also provide more limited but positive contributions.

On the other hand, among the variables **decreasing** default risk, home_ownership_OWN has the strongest negative effect. This finding indicates that home ownership is associated with financial stability. Furthermore, the negative coefficients of the loan_intent_PERSONAL, loan_intent_MEDICAL, and loan_intent_HOMEIMPROVEMENT variables show that these loan purposes carry a lower default risk compared to others. The negative coefficient of the employment_duration variable supports the idea that long-term employment reduces the risk of

default. Overall, the directions and magnitudes of the coefficients are highly consistent with financial intuition and credit risk literature.

VI. Conclusion

In this study, a comprehensive statistical modeling process was conducted to predict individual credit default risk. As a result of data cleaning, exploratory and confirmatory analyses, feature engineering, and dimensionality reduction steps, a consistent and reliable dataset was created for modeling. Logistic regression was chosen as the fundamental method for the binary classification problem; different approaches such as regularization (Ridge, Lasso, Elastic Net), strategies for handling class imbalance (class-weighted and SMOTE), PCA-based dimensionality reduction, and embedded feature selection were systematically compared.

The models were evaluated using multiple performance metrics that account for the imbalanced class structure. The results obtained show that logistic regression has **high discriminative power** for this problem. Although regularized and non-regularized models exhibited similar performances, **Ridge Logistic Regression** was preferred as the final model due to its coefficient stability, statistical significance, and strong generalizability.

In conclusion, this study provides an interpretable, statistically sound, and applicable modeling framework for forecasting credit default risk. The developed approach is directly applicable in areas such as credit risk management, early warning systems, and customer segmentation. In future studies, model performance could be further enhanced by using data structures that include a time dimension or alternative non-linear methods.

References

[1] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[2] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models*, 5th ed. New York, NY, USA: McGraw-Hill/Irwin, 2005.

[3] Prakash Raushan, "Loan dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/prakashraushan/loan-dataset

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.