German Credit Risk Analysis and Classification

Overview

This report provides an extensive exploratory data analysis (EDA) and classification model development conducted on the German Credit dataset, sourced from the UCI Machine Learning Repository. The aim of the project was to thoroughly analyze dataset features, visualize insights, and construct predictive models capable of classifying applicants into distinct categories of credit risk (Good Risk vs. Bad Risk).

Dataset Description

The German Credit dataset contains 1,000 loan application records. Each record is characterized by various numeric, categorical, ordinal, and encoded features related to the applicant's financial status, personal information, loan purpose, and more. The main classification target is the applicant's credit risk:

- **Good Credit Risk (Class 1)** – approximately 70% of the dataset
- **Bad Credit Risk (Class 2)** – approximately 30% of the dataset

For more details, refer to the dataset source: UCI German Credit Data.

Data Preparation and Preprocessing

The dataset was initially loaded from a CSV file $data/raw/german_credit.csv$ containing a mix of numerical and categorical variables, many of which were **custom-coded** using abbreviations or numerical codes. Below is a preview of the raw dataset before any transformation:

	Status_of_existing_checking_account	Duration_in_month	Credit_history	Purpose	Credit_amount	Savings_account_bonds	Present_employment_since	Installment_rate_in_percentage_of_disposable_income	Personal_status_and_sex	Other_debtors_guarantors	Present_residence_since	Property	Age_in_years	Other_installment_plans	Housing	Number_of_existing_credits_at_this_bank	Job	Number_of_people_being_liable_to_provide_maintenance_for	Telephone	foreign_worker	Credit_risk
0	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	1
1	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	2
2	A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	1
3	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1
4	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	2

To make the data human-readable and ready for analysis, a SQL script was executed to **decode the custom features** using mappings from the original dataset documentation. Additionally, columns were renamed using a JSON-based mapping to improve interpretability.

The resulting **processed dataset** contains **98 columns**:

- **21 original features**
- **77 derived features** from one-hot encoding and engineered variables

This transformed dataset was saved to: $data/processed/credit_features.csv$

Cleaning and Encoding

- **Missing categorical values** were imputed with a default `"Unknown"` label to preserve completeness without data loss.
- The column `Personal Status & Sex` was **binary encoded** to extract a simple `Sex` feature.
- The target label `Risk Label` was mapped to a new column `Target`, where:

- 1 represents **Good Credit Risk**
- 0 represents **Bad Credit Risk**

This encoding made the target suitable for binary classification tasks.

Exploratory Data Analysis (EDA)

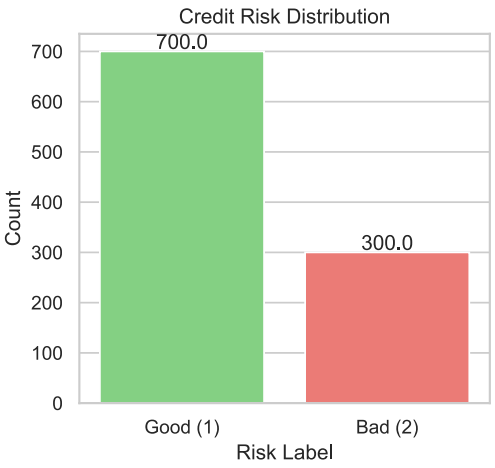
Data Overview

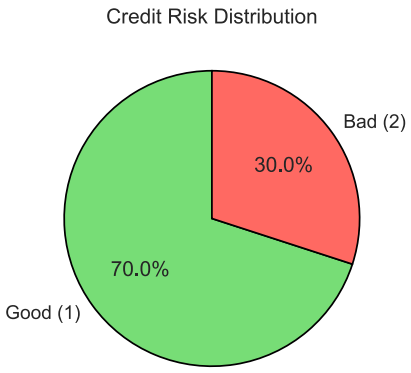
To gain a quick overview of the dataset, we computed descriptive statistics for the numerical features. Since the dataset has already been one-hot encoded, this summary focuses only on continuous variables. A few redundant or uninformative columns were excluded to streamline the analysis.

	count	mean	std	min	25%	50%	75%	max
Loan Duration (Months)	1000.00	20.90	12.06	4.00	12.00	18.00	24.00	72.00
Credit Amount (DM)	1000.00	3271.26	2822.74	250.00	1365.50	2319.50	3972.25	18424.00
Installment Rate (%)	1000.00	2.97	1.12	1.00	2.00	3.00	4.00	4.00
Years at Residence	1000.00	2.85	1.10	1.00	2.00	3.00	4.00	4.00
Age (Years)	1000.00	35.55	11.38	19.00	27.00	33.00	42.00	75.00
# Existing Credits	1000.00	1.41	0.58	1.00	1.00	1.00	2.00	4.00
Dependents	1000.00	1.16	0.36	1.00	1.00	1.00	1.00	2.00
Risk Label	1000.00	1.30	0.46	1.00	1.00	1.00	2.00	2.00
Credit per Month	1000.00	167.69	153.49	24.06	89.60	130.33	206.18	2482.67
Age < 25	1000.00	0.15	0.36	0.00	0.00	0.00	0.00	1.00
Age 25–40	1000.00	0.58	0.49	0.00	0.00	1.00	1.00	1.00
Age > 40	1000.00	0.27	0.45	0.00	0.00	0.00	1.00	1.00
Checking <0 DM	1000.00	0.27	0.45	0.00	0.00	0.00	1.00	1.00
Checking 0–200 DM	1000.00	0.27	0.44	0.00	0.00	0.00	1.00	1.00
Checking ≥200 DM	1000.00	0.06	0.24	0.00	0.00	0.00	0.00	1.00
Checking: No Checking Account	1000.00	0.39	0.49	0.00	0.00	0.00	1.00	1.00
No Credits	1000.00	0.04	0.20	0.00	0.00	0.00	0.00	1.00
All Paid Duly	1000.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00
Existing Paid Duly	1000.00	0.53	0.50	0.00	0.00	1.00	1.00	1.00
Past Delay	1000.00	0.09	0.28	0.00	0.00	0.00	0.00	1.00
Critical History	1000.00	0.29	0.46	0.00	0.00	0.00	1.00	1.00
Purpose: New Car	1000.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00
Purpose: Used Car	1000.00	0.10	0.30	0.00	0.00	0.00	0.00	1.00
Purpose: Furniture	1000.00	0.18	0.39	0.00	0.00	0.00	0.00	1.00
Purpose: Radio/TV	1000.00	0.28	0.45	0.00	0.00	0.00	1.00	1.00

Class Balance Analysis

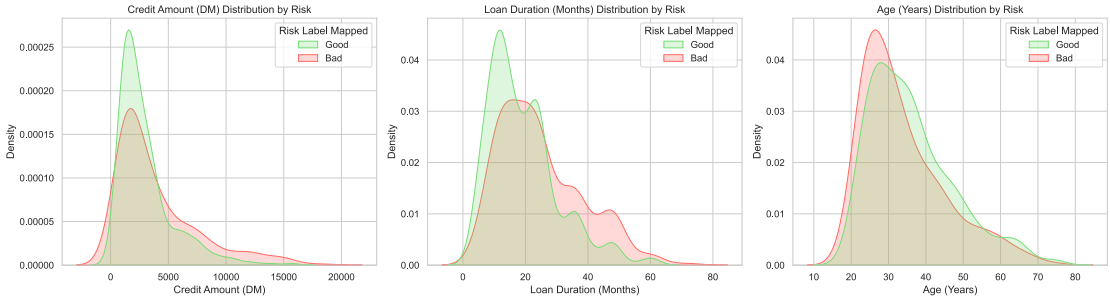
A significant imbalance was observed, with good credit risks constituting 70% of cases. This imbalance can bias predictive models toward the majority class. The imbalance was visualized using both count and pie charts:



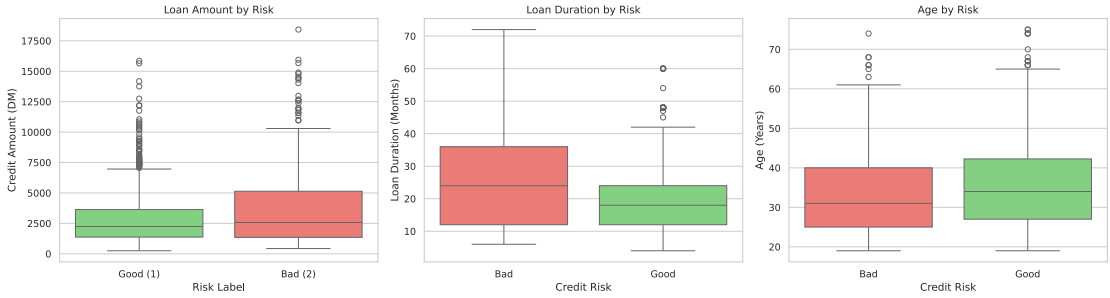


Numerical Distribution by Risk

Descriptive statistics for numerical features such as Credit Amount, Loan Duration, and Age revealed skewed distributions, especially notable in the Credit Amount and Loan Duration features. Density plots helped visualize differences between Good and Bad risk classes:



Boxplots further clarified that Bad Risk loans were typically higher in amount and longer in duration:

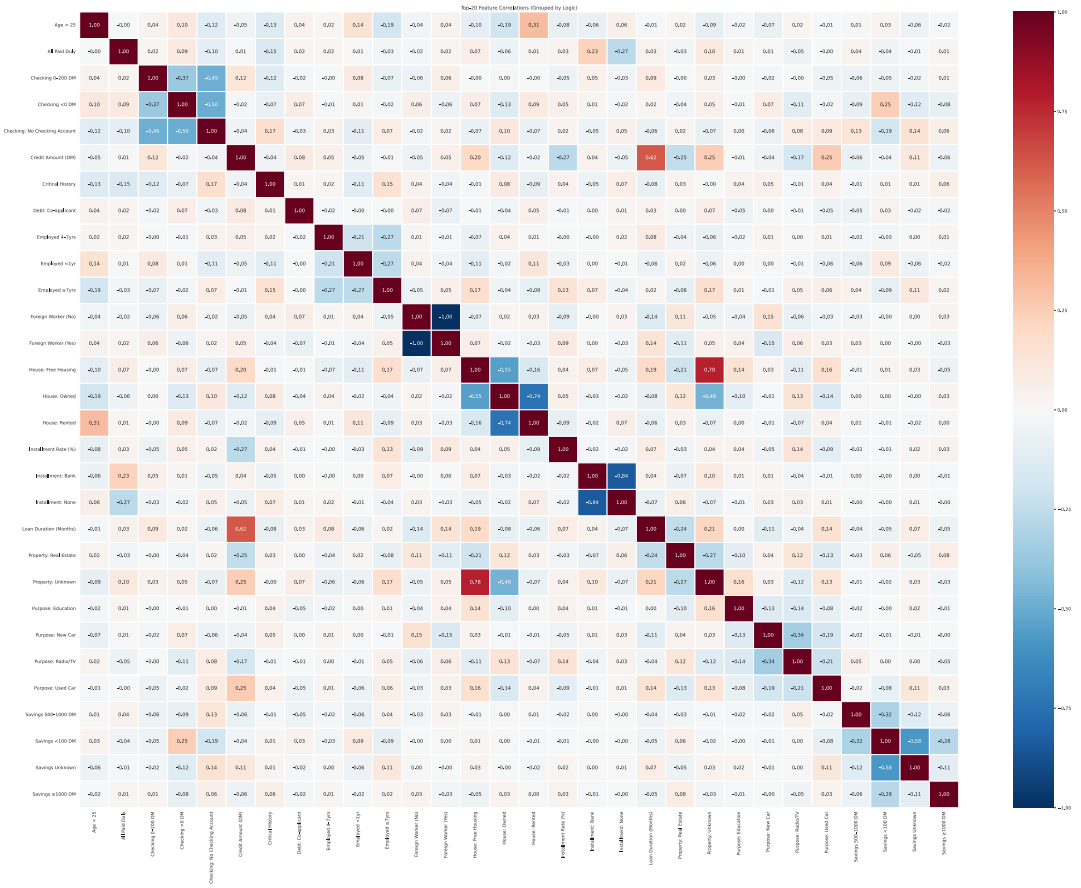


Categorical Feature Analysis

The analysis of categorical features such as Loan Purpose, Housing Type, and Personal Status & Sex highlighted key trends. For instance, people with rented housing or those seeking loans for used cars and retraining exhibited higher proportions of bad credit risk:

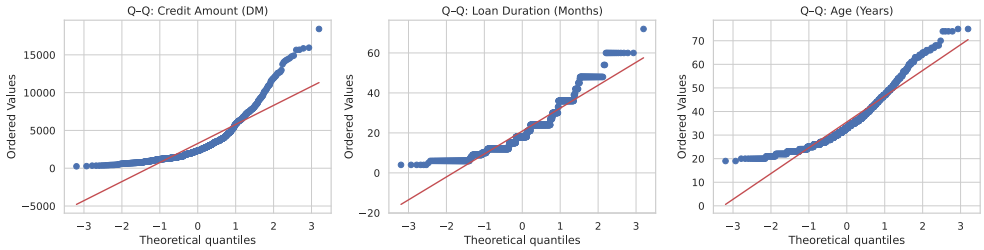
Correlation Analysis

A correlation heatmap showed moderate internal correlations within grouped features (one-hot encoded categories) but relatively weak correlations across different feature groups, indicating minimal multicollinearity.



Normality Checks (Q-Q Plots)

Quantile-Quantile plots indicated significant deviations from normality in numerical features, particularly evident in the distribution tails of Credit Amount:



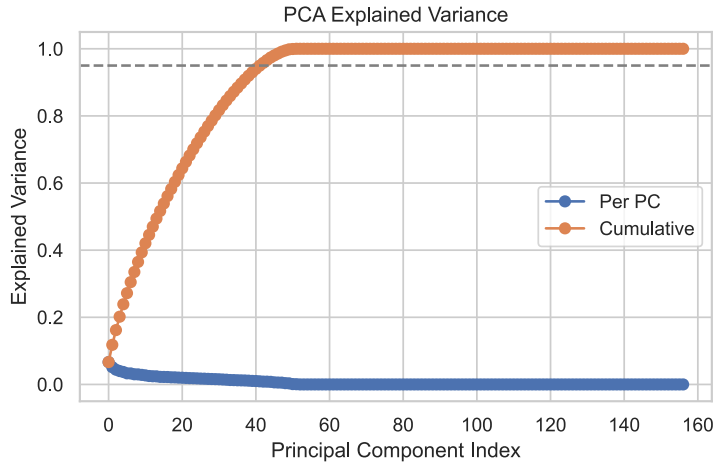
Model Development

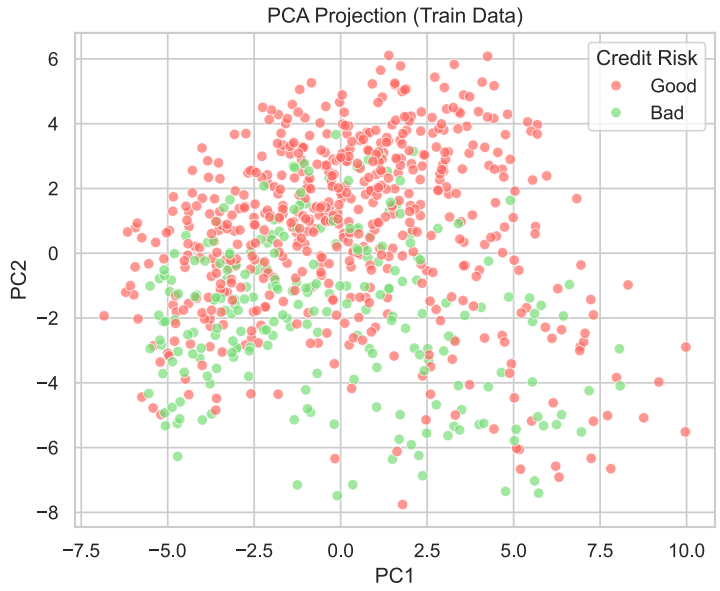
Train-Test Split and Feature Scaling

The dataset was split into training (80%) and test sets (20%) using stratified sampling to preserve class distribution. Feature scaling was performed using both standardization and min-max scaling techniques to ensure effective model training.

Dimensionality Reduction (PCA)

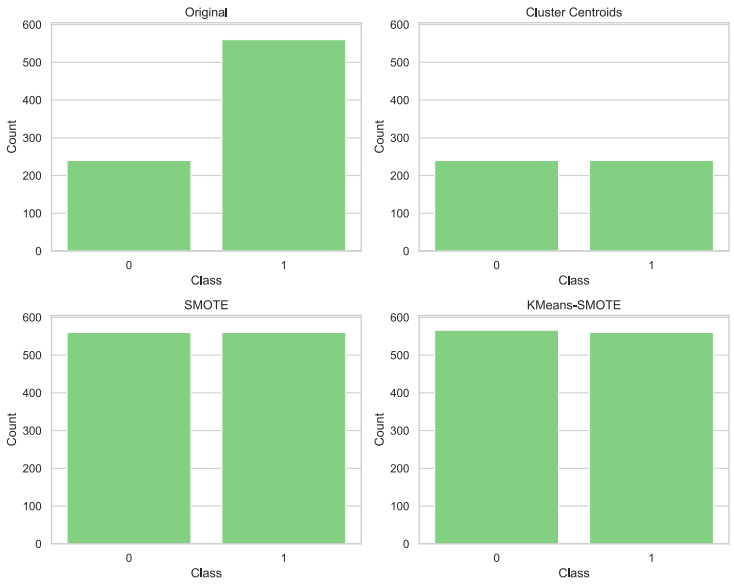
Principal Component Analysis (PCA) was applied to reduce feature dimensionality, capturing 95% of variance in fewer components. PCA effectively visualized class separations in a 2-dimensional space:





Addressing Class Imbalance

To handle class imbalance, various resampling techniques such as Cluster Centroids (undersampling), SMOTE, and KMeans-SMOTE (oversampling) were employed, as shown in the comparative visualization below:

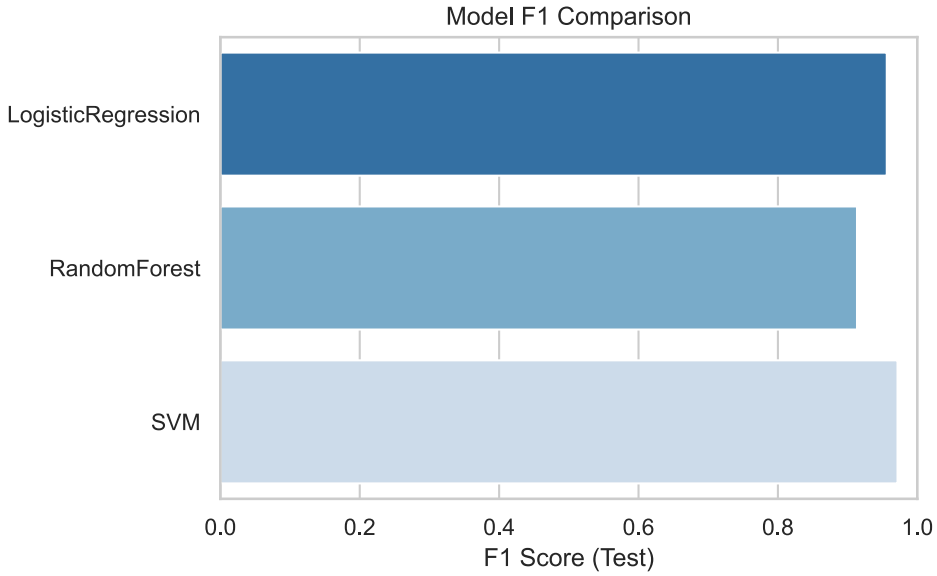


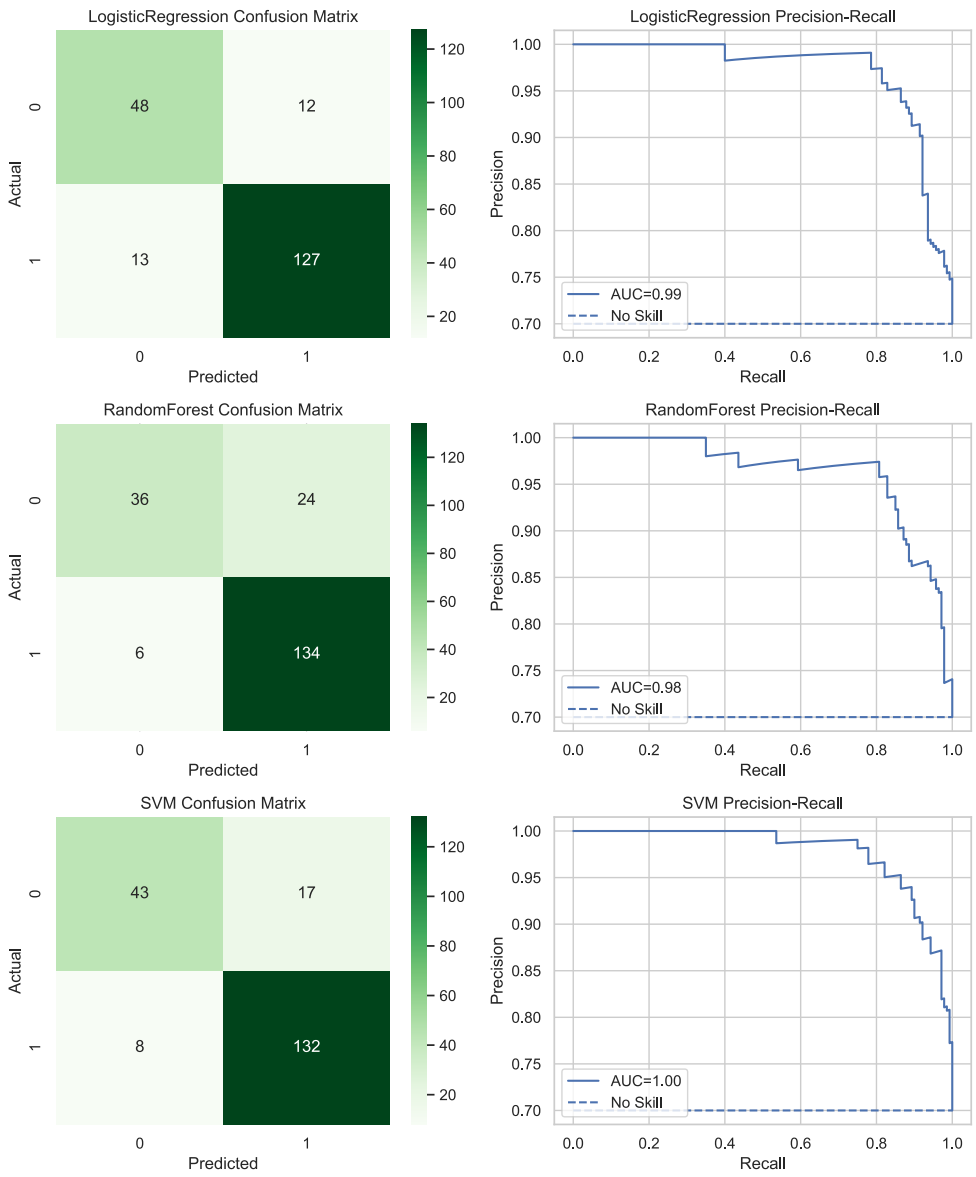
Classification Models and Evaluation

Three classification models—Logistic Regression, Random Forest, and Support Vector Machine (SVM)—were trained and evaluated using 5-fold cross-validation optimized for the F1 score.

Model Performance

Random Forest emerged as the best-performing model, exhibiting the highest F1 score, precision, recall, and area under the Precision-Recall curve (PR-AUC). Performance across models is summarized:





Conclusions

Random Forest demonstrated superior predictive capability, leveraging key features such as Checking Account Status, Loan Duration, and Credit History. Addressing class imbalance significantly enhanced model effectiveness, highlighting the critical role of proper data preprocessing.

Recommendations and Future Work

Future improvements could involve additional feature engineering, advanced hyperparameter tuning, and exploring ensemble methods like Gradient Boosting Machines (GBM). Additionally, addressing outliers and applying transformations to skewed data might further improve model performance and stability.

Project Structure

```
.
  data
     raw
        german_credit.csv
     processed
         credit_features.csv
        model_summary.csv
        rf_feature_importances.csv
  sql
     feature_engineering.sql
  plots
     credit_eda.svg
     numerical_distribution.svg
     box_plots.svg
     categorical_distribution_Loan_Purpose.svg
     categorical_distribution_Housing_Type.svg
     categorical_distribution_Personal_Status_And_Sex.svg
     corrlation_heatmaps.svg
     qq_plots.svg
     pca_explained_variance.svg
     pca_explained_variance_2d.svg
     kmeans-smote.svg
     f1-score.svg
     precision-recall.svg
  credit_eda.ipynb
  README.md
  requirements.txt
```

This detailed report combines robust statistical analysis and visual insights to provide comprehensive understanding and actionable recommendations for credit risk assessment.