

Assignment 2 Data Mining

by Sameeran Raveenthiran

Submission date: 10-May-2024 12:47PM (UTC+0100)

Submission ID: 233110332

File name: 33405_Sameeran_Raveenthiran_Assignment_2_Data_Mining_565911_602359442.docx (803.12K)

Word count: 3318

Character count: 20162



ASSIGNMENT 2: DATA MINING



Sameeran Raveenthiran
2114454 UWTSD

Contents

Introduction:.....	2
Data Mining:.....	2
Data Set:	2
• Description of Attributes:.....	3
• Number of Instances:	3
• Identification of Missing Values and Other Relevant Characteristics:	3
• Summary of Data Familiarization, Exploration, and Visualization Process:	3
Pre-processing and Post-processing:.....	4
• Pre-processing	Error! Bookmark not defined.
• Post-processing	Error! Bookmark not defined.
Analysis Methods:	4
EDA Findings:	4
Insights:	17
Business Recommendations:.....	18
Conclusion:	19
References:.....	19

Introduction:

Data Mining:

Data mining has become integral in the modern business sector, notably for extracting valuable insights from extensive datasets. It involves using algorithms to detect hidden patterns and relationships that guide client decision-making, such as segmenting populations or evaluating probabilities to enhance efficiency and mitigate risks. The South German Credit Data Set from the UCI Machine Learning Repository provides a rich source of data, revealing factors that determine a borrower's reliability or default likelihood. Through detailed analysis of demographics and financial behaviors, this information is vital for financial institutions to refine risk assessments and optimize lending strategies, thereby reducing defaults, and enhancing loan management. Our analysis aims to uncover key drivers behind loan repayment outcomes, aiding financial institutions in making more informed decisions.

Data Set:

The South German Credit Data Set, hosted by the UCI Machine Learning Repository and originating from a German bank, is pivotal in finance and machine learning for studying credit risk assessment. Rich in data on financial conditions, employment, and credit histories, it supports the development of predictive models and decision-making tools for evaluating loan default probabilities. Its accessibility has fostered numerous algorithms and systems, enhancing both academic and industry practices in credit scoring and risk analysis. This data set significantly advances our understanding of credit risk modelling and financial decision-making, marking a crucial resource in evaluating borrower creditworthiness [7]:

1. 'Status' indicates the checking account status, categorized as 'no checking,' '< 0 DM,' '0 to 200 DM,' or '> 200 DM.'
2. 'Duration' signifies the duration of the credit in months.
3. 'Credit history' denotes the individual's credit history, with options such as 'delay in paying off in the past' or 'existing credits paid back duly till now.'
4. 'Purpose' reflects the purpose of the credit, ranging from 'car (new)' to 'business.'
5. 'Amount' represents the credit amount requested.
6. 'Savings' indicates the status of savings accounts or bonds, categorized similarly to the checking account.
7. 'Employment duration' denotes the duration of current employment, ranging from 'unemployed' to '>= 7 yrs.'
8. 'Installment rate' signifies the percentage of disposable income allocated to loan installments.
9. 'Personal status and sex' reflect the personal status and gender of the individual.
10. 'Other debtors' indicates the presence of other debtors or guarantors, categorized as 'none,' 'co-applicant,' or 'guarantor.'
11. 'Present residence since' denotes the duration of current residence in years.

12. 'Property' signifies the type of property owned, ranging from 'unknown / no property' to 'real estate.'
13. 'Age' represents the age of the individual in years.
14. 'Other installment plans' reflects the presence of other installment plans, categorized as 'bank,' 'stores,' or 'none.'
15. 'Housing' denotes the type of housing, such as 'own,' 'rent,' or 'for free.'
16. 'Number of credits' indicates the number of existing credits at the current bank, categorized as '1,' '2-3,' '4-5,' or '>= 6.'
17. 'Job' represents the type of job held by the individual, ranging from 'unemployed/unskilled - non-resident' to 'manager/self-empl./highly qualif. employee.'
18. 'People liable' signifies the number of people dependent on the individual, categorized as '3 or more' or '0 to 2.'
19. 'Telephone' reflects the presence of a telephone, categorized as 'no' or 'yes (under customer name).'
20. 'Foreign worker' denotes whether the individual is a foreign worker, categorized as 'yes' or 'no.'
21. 'Credit risk' (target column) represents the credit risk, categorized as 'bad' or 'good.'

- **Description of Attributes:**

The dataset comprises various attributes providing insights into individuals' financial status, employment details, personal demographics, and loan-related characteristics.

- **Number of Instances:**

The dataset contains a total of 1000 instances, each representing a unique credit application.

- **Identification of Missing Values and Other Relevant Characteristics:**

Throughout the data exploration process, efforts will be made to identify and manage missing values, ensuring data integrity. Additionally, relevant characteristics such as data distribution and potential correlations between attributes will be scrutinized to provide a holistic view of the dataset.

- **Summary of Data Familiarization, Exploration, and Visualization Process:**

The data will undergo thorough examination and visualization using descriptive statistics and various visualization techniques. These include histograms, box plots, and correlation matrices, among others. This exploration aims to reveal underlying patterns

and relationships within the data, laying the groundwork for subsequent analysis and interpretation.

Pre-processing and post-processing:

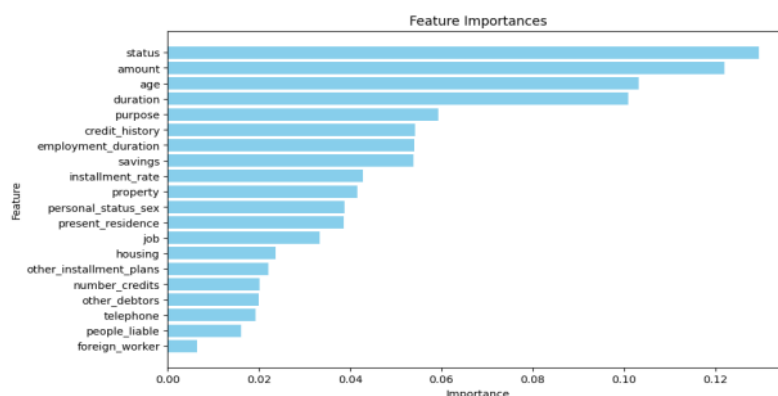
Pre-processing and post-processing are crucial steps in refining data quality and extracting actionable insights from the South German Credit Data Set.

Pre-processing: Upon examining the dataset, no missing values or duplicates were found, simplifying the initial steps. The first action involved translating column names from Dutch to English to enhance understanding. Furthermore, all columns were initially encoded as integers due to their categorical nature. We reverted these columns to their original categorical values for clarity and better interpretability.

Post-processing: Using the original dataset, which was free of missing values and duplicates, eliminated the need for re-converting categorical columns into numerical formats, streamlining the analysis. We also balanced the dataset to ensure fair representation of both classes in the target variable. To discern the factors impacting credit risk, we employed the Random Forest algorithm, a powerful tool for understanding the importance of each feature in predicting reliability as a borrower. This analysis provided crucial insights into what drives creditworthiness assessments.

Analysis Methods:

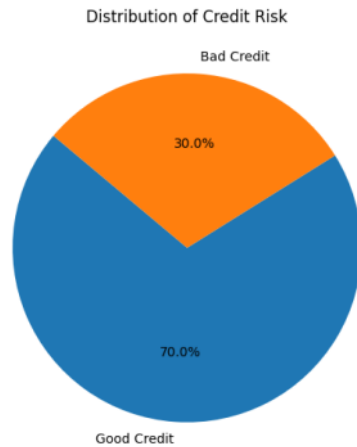
To thoroughly understand our dataset, we relied extensively on Exploratory Data Analysis (EDA), exploring each attribute through histograms, box plots, and scatter plots. This crucial step allowed us to appreciate the dataset's nuances, identify trends, and examine relationships between variables. Following EDA, we utilized the Random Forest algorithm on a balanced dataset to delve deeper into the data's complexities and determine key predictive features for credit risk.



This method highlighted the significant roles of 'status,' 'amount,' 'age,' and 'duration' in predicting credit risk, offering deep insights, and guiding our feature prioritization for future modeling and decision-making.

EDA Findings:

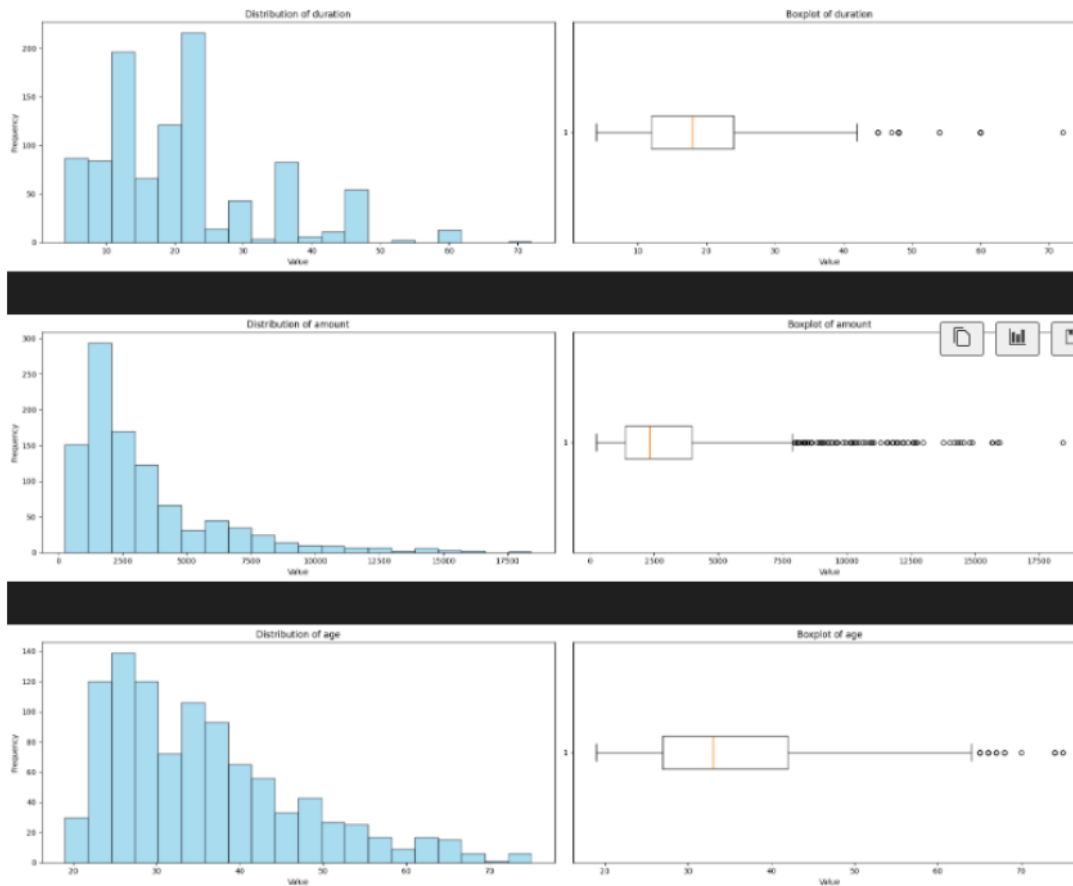
In our data exploration, we delved into the dataset to find out what factors might affect whether someone repays a loan. We looked at things like age, credit amount, and employment duration to see how they relate to credit risk. Here's what we discovered.



In the first phase we examined the dataset for any abnormalities and found that the dataset is quite clean with no missing values, infinity values or duplicates although the dataset is highly imbalanced where the number of bad credit records is only 30%.

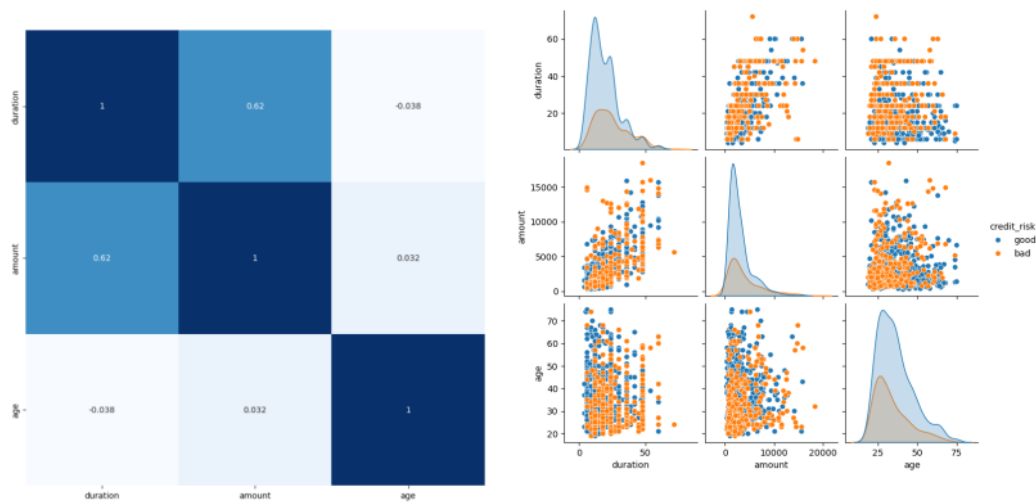
Acknowledging the importance of conducting thorough exploratory data analysis (EDA), we have decided to temporarily set aside any preprocessing steps. Our focus remains on understanding the relationships between various columns and credit risk to avoid potential errors in the data.

In our initial analysis, we examined the columns related to duration, age, and amount:



- **Duration:** The duration of credit spans from 4 to 72 months, with an average of about 20.9 months. We noticed a right-skewed distribution, with few outliers.
- **Amount:** Credit amounts requested range from 250 to 18,424 DM, with an average of 3,271.25 DM. The data shows significant right-skewness and numerous outliers.
- **Age:** Applicant ages range from 19 to 75 years, with an average of 35.54 years. Although some outliers were detected, the distribution is only slightly right-skewed.

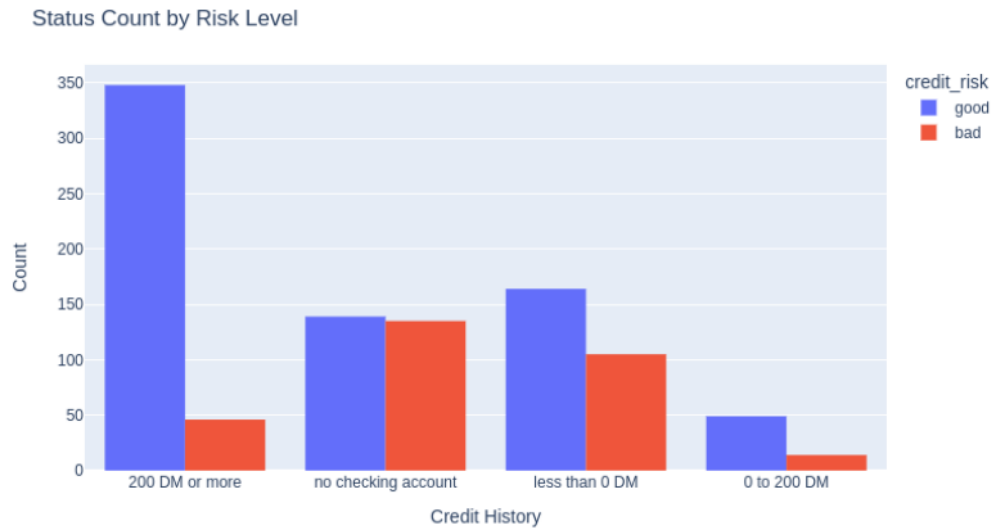
Investigating the relationships between these variables:



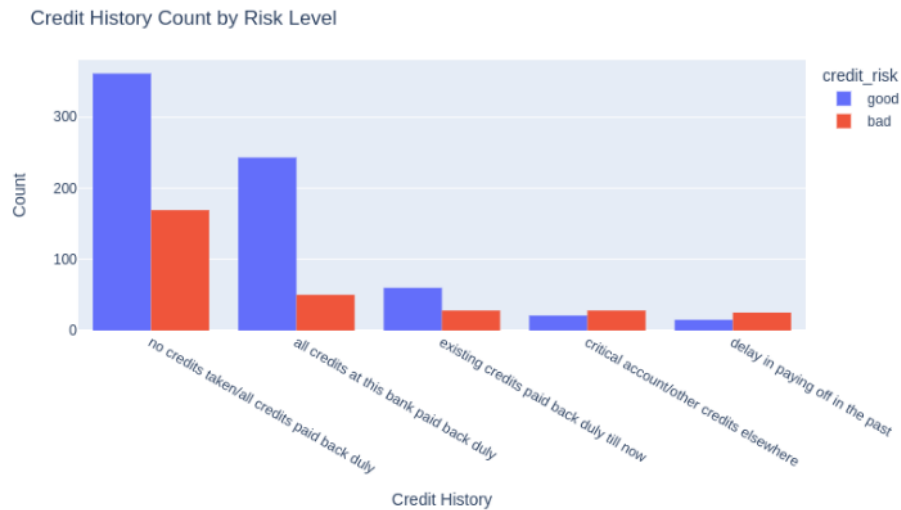
- Duration vs. Amount: We observed a moderate positive correlation of approximately 0.625 between credit duration and requested amount, indicating that longer durations are associated with higher amounts.
- Duration vs. Age: A negligible negative correlation of around -0.038 suggests little to no relationship between credit duration and applicant age.
- Amount vs. Age: A very weak positive correlation of about 0.032 hints that older applicants may request slightly higher credit amounts on average, but the relationship is minimal.

Additionally, all plots exhibited a positive skew. Notably, most loan applicants fell within the age range of 20 to 30 years, preferring lower credit amounts, typically repaid within 20 months. Interestingly, the bank received more loan applicants aged 20 to 30 years, often requesting amounts between 250 and 2500 DM.

After this we look at the categorical columns and their relationship with credit_risk.

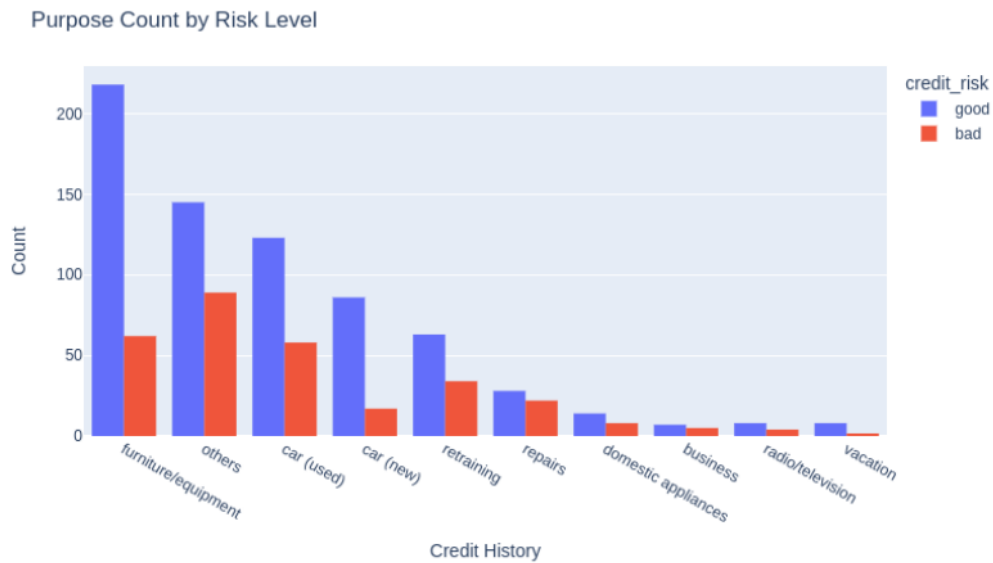


Approximately 80% of bad credit cases are linked to individuals without checking accounts or those with negative balances. Accounts with balances over 200 DM typically show the highest good-to-bad credit ratios. Conversely, accounts with balances between 0 and 200 DM also demonstrate a relatively favorable credit risk distribution.

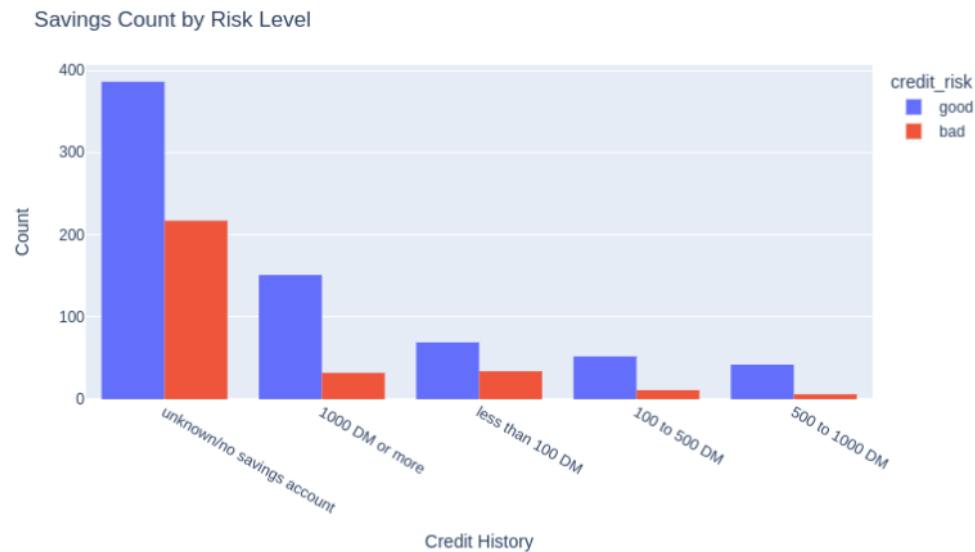


About 56.3% of bad credit cases involve individuals who have consistently repaid previous credits on time or have not taken any credits from the current bank. Notably, those with a history of delayed payments represent 71.4% of bad credit cases, underscoring the impact of payment history on credit assessments.

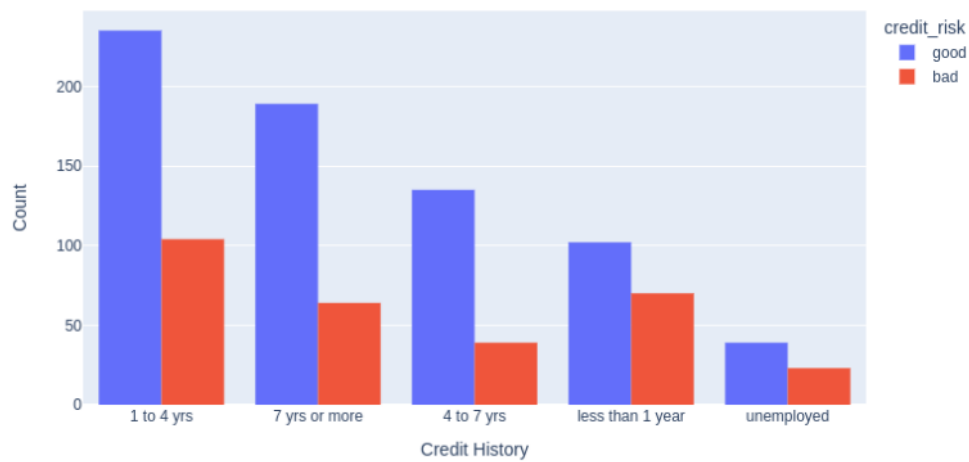
Among the individuals who have no history of taking credits or have paid back all credits duly, a significant portion, accounting for approximately 53.0%, took the most amount of loans.



It's evident that loans designated for purposes such as furniture/equipment exhibit a favorable ratio of good credit risk compared to bad credit risk. On the other hand, loans for purposes like used cars, new cars, and others tend to have a higher proportion of bad credit risk.

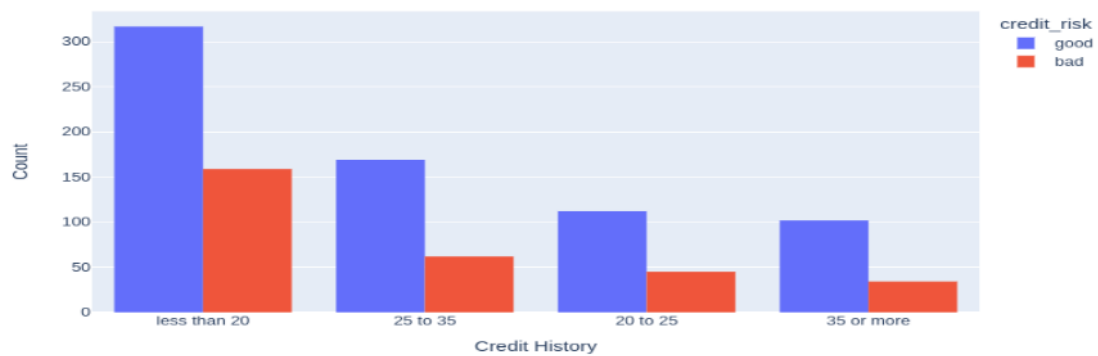


Employment Duration Count by Risk Level



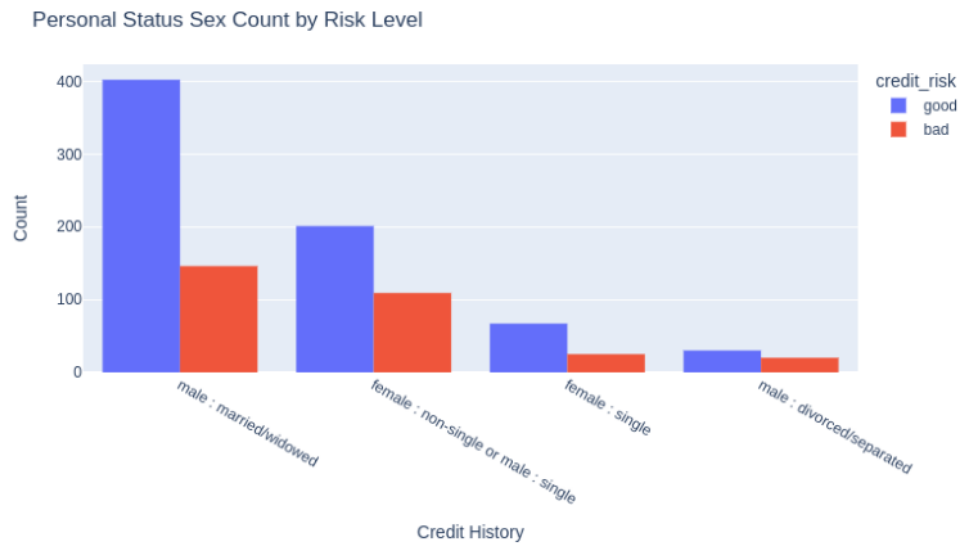
In our initial review of the dataset, we found it clean with no missing or infinite values and no duplicates, though it is highly imbalanced with only 30% bad credit records.

Installment Rate Count by Risk Level

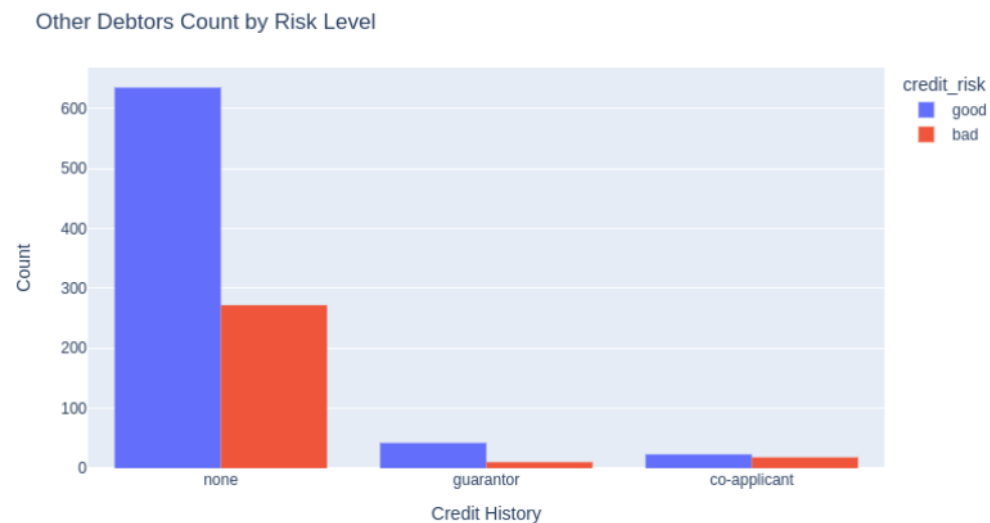


Analysis reveals that 47.6% of individuals prefer installment rates under 20%. Lower installment rates are often linked to higher bad credit risks, suggesting that limited net income may hinder

repayment abilities and increase credit risk.



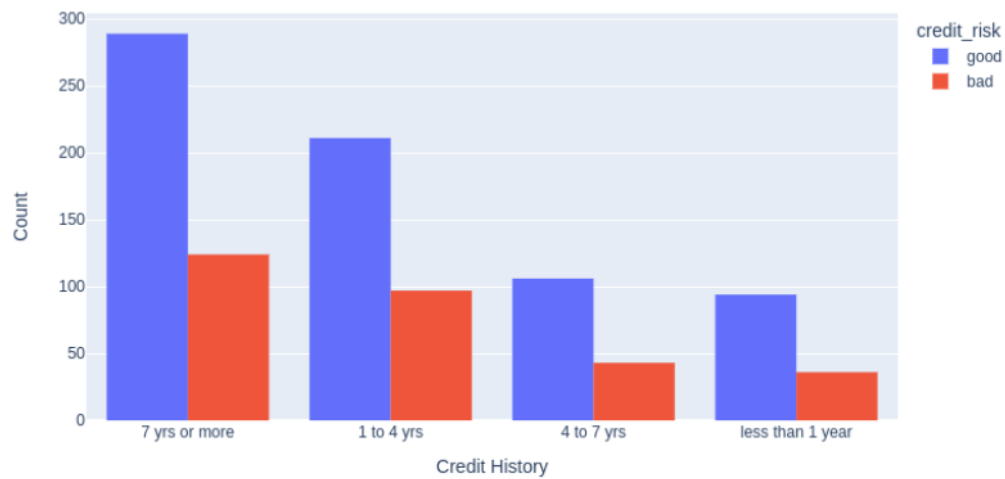
Data analysis shows 54.8% of the dataset comprises married or widowed men, with a significant 31.0% being non-single women or single men. Notably, married or widowed individuals display higher creditworthiness compared to divorced or separated ones, suggesting marital status impacts credit ratings.



Data reveals 90.7% of individuals have no other debtors. Only 5.2% have a guarantor, and 4.1% a co-applicant. Interestingly, co-applicants have nearly equal good and bad credit risks,

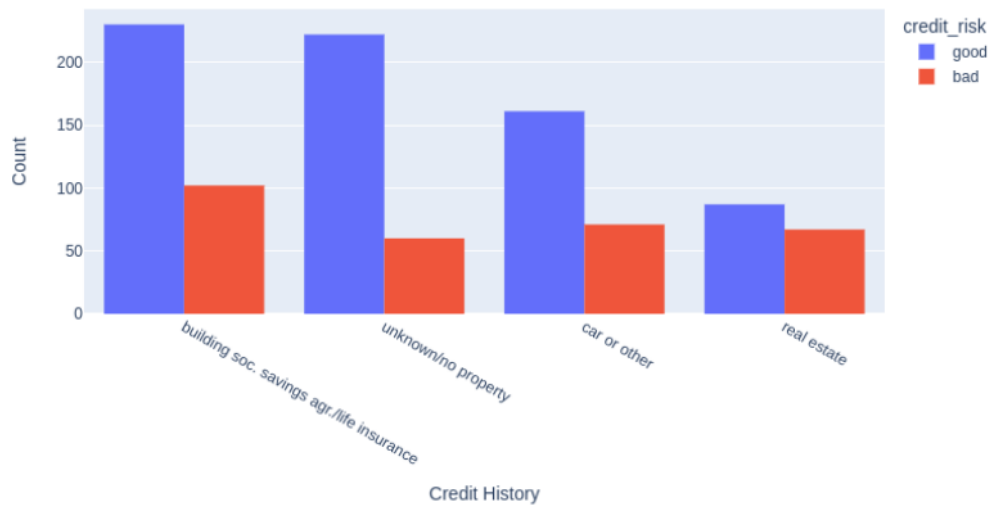
while those without any debtors show the highest number of bad credit cases.

Present Residence Count by Risk Level



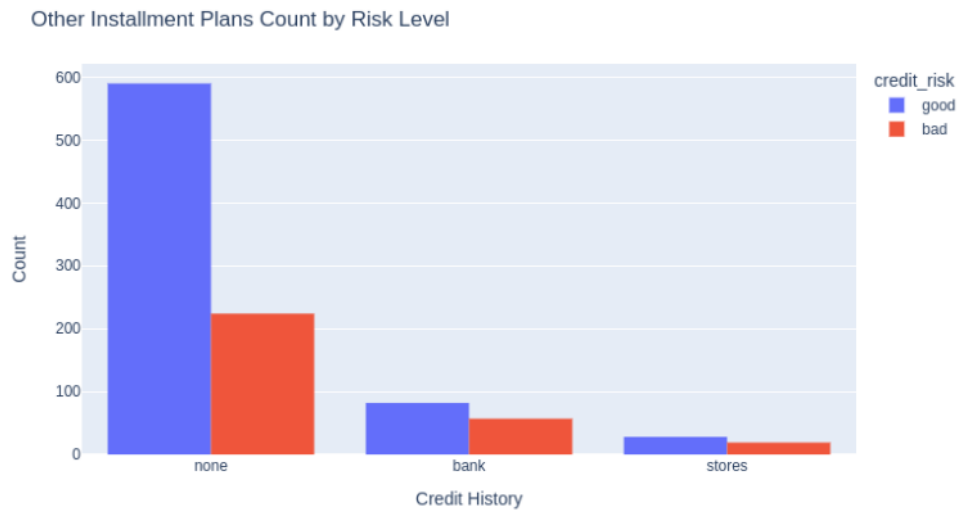
The data shows 41.3% of individuals have lived in their current home for over 7 years, and 30.8% for 1 to 4 years. Most bad credit cases, 73.6%, are among residents of 1 to 4 years or more than 7 years. Notably, the highest number of entries is in these groups, with those residing less than a year having a higher ratio of bad to good credit cases.

Property Count by Risk Level

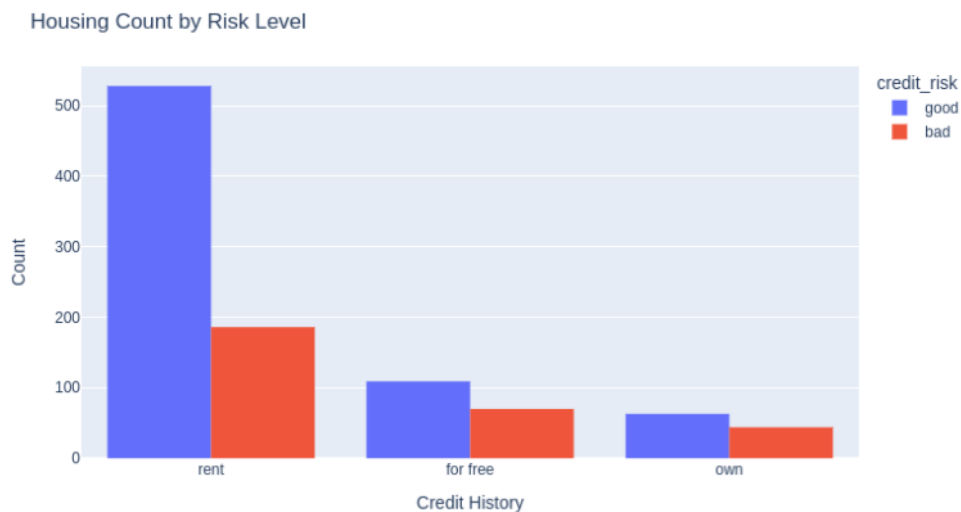


Out of the individuals in the dataset, 33.2% have building society savings or life insurance, 28.2% have unknown or no property, 23.2% own a car or other property, and 15.4% own real estate. Interestingly, among those classified as owning real estate, there were 67 bad credit risk

cases and 87 good credit risk cases. Conversely, individuals categorized as having unknown or no property exhibited the best ratio of good to bad credit, favoring good credit.



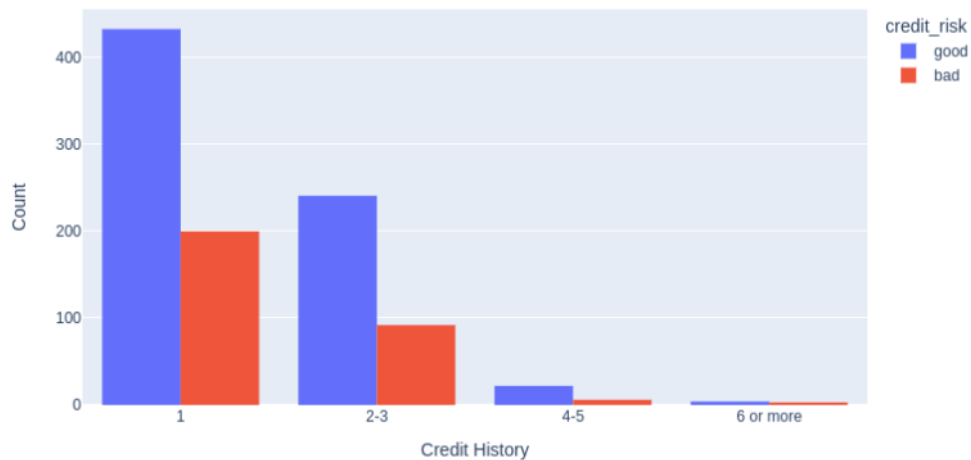
Data analysis shows that 81.4% of individuals have only bank-related installment plans, with 75% of bad credit risks in this group, unexpectedly. For those with store installment plans, bad credit instances (19) outnumber good (28). Similarly, bank installment plans show worse (57) than good (82) cases, indicating that non-bank plans correlate with higher credit risks.



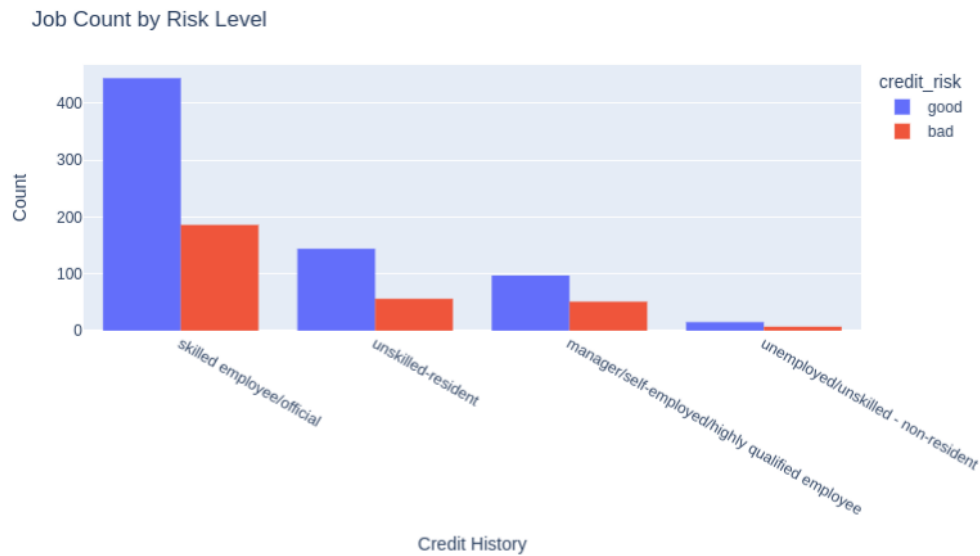
Analysis of the dataset reveals that 71.4% of individuals rent their homes, while 17.9% live rent-free, and 10.7% own their homes. Despite their credit status, most choose renting. Intriguingly,

renters display a higher incidence of both good and bad credit records compared to other groups. Renters generally exhibit higher bad credit risks, suggesting renting may correlate with increased credit risk. Conversely, homeowners show the lowest bad-to-good credit ratio, indicating lower risk associated with homeownership.

Number Credits Count by Risk Level



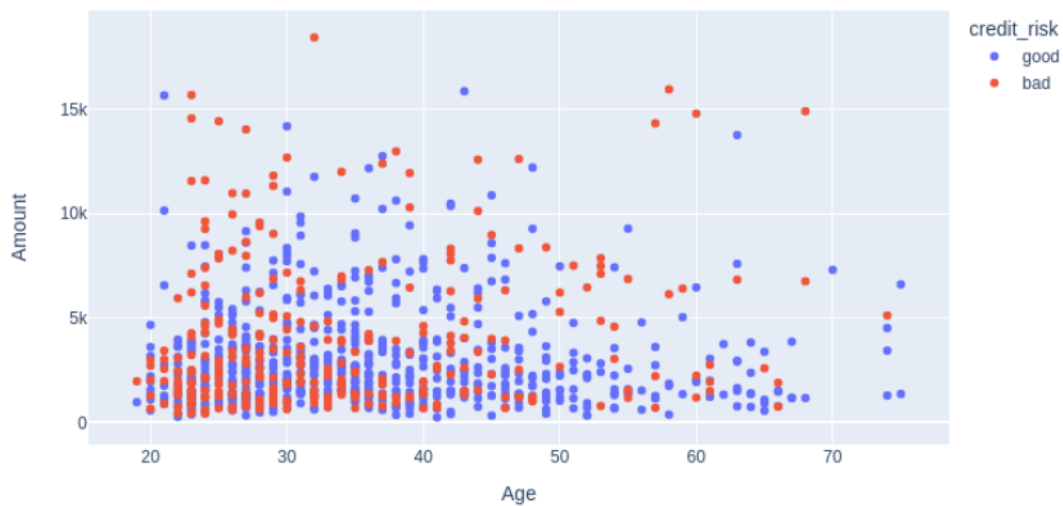
The dataset reveals that 63.3% of individuals have just one credit, while 33.3% possess 2 to 3 credits. Fewer have 4 to 5 credits (2.8%), and only 0.6% have six or more. Interestingly, as the number of credits increases, the likelihood of being categorized as a bad credit risk decreases. Specifically, 31.6% of those with one credit are considered risky, but this figure drops to 27.6% for those with 2 to 3 credits, 21.4% for those with 4 to 5 credits, and paradoxically rises to 33.3% for those with six or more credits. This suggests that managing multiple credits responsibly might enhance creditworthiness, highlighting the impact of credit history on financial reputation.



The dataset analysis shows diverse employment statuses: 63.0% are skilled employees or officials, 20.0% are unskilled residents, 14.8% are managers, self-employed, or highly qualified employees, and a mere 2.2% are unemployed or unskilled non-residents. Despite variances in bad credit risks among these job categories, a significant number of individuals within each are considered good credit risks. Specifically, 70.5% of skilled workers, 65.5% of higher-level professionals, 68.2% of unemployed or unskilled non-residents, and 72.0% of unskilled residents are categorized as good credit risks. These findings indicate that while job type can affect credit risk, other factors are also crucial in assessing creditworthiness.

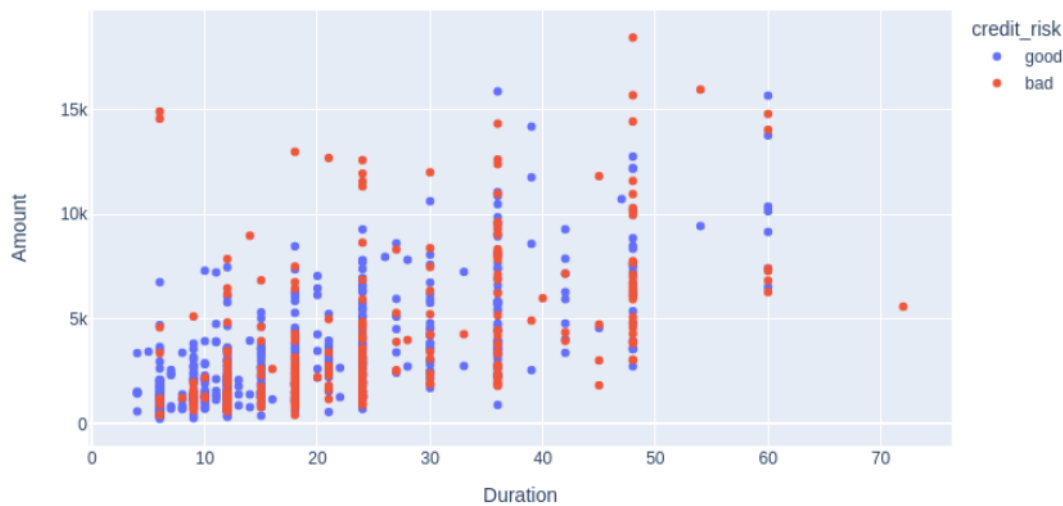
Additionally, having a telephone is associated with a lower proportion of bad credit risks (28.0%) compared to those without (31.3%). Renters face higher credit risks compared to homeowners or those living rent-free. Interestingly, individuals managing multiple credit lines exhibit lower bad credit risks, suggesting effective credit handling. Foreign workers also show a significantly lower bad credit risk (10.8%) compared to non-foreign workers (30.7%). Overall, although job categories and housing statuses show variations in bad credit risks, most individuals are deemed good credit risks, reflecting a generally positive credit profile across the dataset.

Scatter Plot of Age vs Amount



Most of the individuals identified as bad credit risks are aged between 20 and 60 years, indicating that credit risk is prevalent across a broad age range. Additionally, bad credit risks are commonly associated with loan amounts ranging from 0 to 12500 DM, suggesting that these individuals may struggle with managing higher loan amounts or may have lower financial stability.

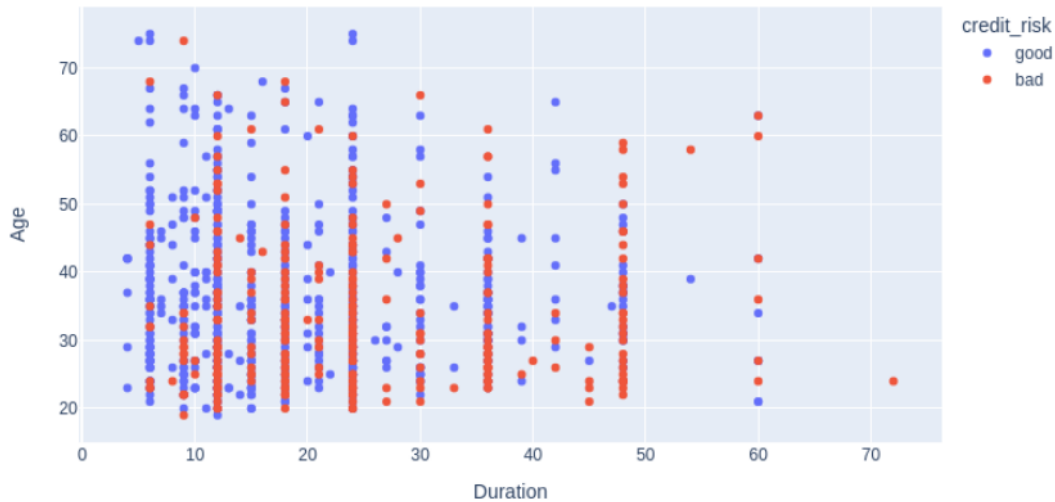
Scatter Plot of Amount vs Duration



Most bad credit risks are associated with credit durations ranging from 12 to 60 months, suggesting that these individuals may struggle with long-term loan commitments. Moreover,

there is a positive relationship between the duration of bad credit risks and the loan amount, indicating that individuals with longer credit durations tend to borrow larger amounts. This insight underscores the importance of considering both duration and loan amount when assessing credit risk.

Scatter Plot of Age vs Duration



Bad credit risks often coincide with loan durations spanning from 12 to 60 months, indicating potential challenges with long-term financial commitments. Additionally, the age distribution of individuals classified as bad credit risks predominantly falls within the 20 to 60-year range, highlighting the significance of age as a factor influencing creditworthiness.

Insights:

Based on the findings from the exploratory data analysis (EDA), several insights can be gleaned:

1. **Credit Duration and Amount Relationship:** There exists a moderate positive correlation between credit duration and loan amounts, with longer durations generally correlating with higher loan amounts. This pattern suggests that individuals often require larger loans for longer-term financial commitments.
2. **Demographic Factors:** Factors like age, marital status, and employment duration significantly impact credit risk. Younger individuals, those with shorter employment histories, and those who are divorced or separated show a higher propensity for bad credit. In contrast, married or widowed individuals and those with longer employment durations tend to have more favorable credit profiles.

3. **Credit History and Behavior:** A history of delayed payments or limited credit interactions with the current bank typically leads to poorer credit ratings. Additionally, the loan's purpose influences credit risk, with furniture/equipment loans often associated with lower risk than vehicle loans.
4. **Financial Stability Indicators:** Higher balances in savings accounts and lower instalment rates relative to one's net income serve as indicators of reduced credit risk. Individuals with healthier savings or smaller instalment burdens generally show better credit standings.
5. **Housing Status and Job Type:** Renters display a higher risk of bad credit compared to homeowners or those living rent-free. Employment type also varies in risk, with managers, self-employed individuals, and highly skilled workers showing lower credit risks compared to others.
6. **Number of Credits and Credit Risk:** Individuals holding multiple credit lines generally have a lower risk of bad credit, suggesting effective credit management across numerous obligations.
7. **Foreign Worker Status:** Foreign workers tend to have a significantly lower rate of bad credit compared to local workers, pointing to possible differences in financial management or stability between these groups.
8. **Loan Duration and Amount:** Higher instances of bad credit are often linked with loan durations of 12 to 60 months and amounts up to 12,500 DM, indicating difficulties in managing long-term and substantial financial commitments.

Overall, these insights highlight the multifaceted nature of credit risk assessment, encompassing demographic, financial, and behavioral factors. By considering these insights, businesses can refine their loan approval processes, identify risk factors early, and make more informed lending decisions to mitigate risks effectively.

Business Recommendations:

Based on the analysis findings, businesses can make several recommendations to enhance their loan approval processes and mitigate risks effectively:

1. Refine risk assessment models by integrating insights from our dataset analysis. Incorporate variables like employment duration, housing status, credit history, and job category to improve accuracy in predicting risks.
2. Implement segmentation strategies based on demographic and financial data from the analysis, categorizing applicants into different risk profiles and customizing loan terms accordingly.

3. Develop targeted marketing campaigns to attract customer segments likely to have good credit, focusing on individuals with stable employment, substantial savings, and positive credit histories.
4. Implement risk mitigation measures for higher-risk segments by setting stricter eligibility criteria, requiring additional collateral or guarantors, and offering lower credit limits to manage potential losses.
5. Establish a continuous monitoring and evaluation system to assess the performance of loan portfolios over time, regularly updating risk models based on market dynamics and customer behavior.
6. Launch customer education initiatives to improve financial literacy, particularly targeting higher-risk segments to encourage responsible borrowing and enhance repayment rates.
7. Strengthen collaboration with credit reporting agencies to access comprehensive credit histories, leveraging external data for deeper insights into customer creditworthiness and supporting informed lending decisions.

By implementing these recommendations, businesses can optimize their loan approval processes, minimize credit risks, and foster a healthier lending environment conducive to sustainable growth and profitability.

Conclusion:

In conclusion, our dataset analysis unveils critical insights into credit risk factors among loan applicants.

Firstly, essential variables such as credit duration, loan amount, and applicant age show distinct distributions and interconnections. Notably, longer credit periods often correlate with higher loan amounts. Older applicants typically seek larger loans, while the majority, aged between 20 and 30, prefer smaller amounts with an average repayment term of 20 months.

Secondly, categorical data like checking account status, credit history, loan purpose, and employment duration reveal significant links to credit risk. A considerable portion of negative credit outcomes are associated with no or overdrawn checking accounts. Additionally, individuals with histories of payment delays or minimal credit engagements are more susceptible to credit issues.

Moreover, factors such as housing status, job category, and the presence of co-debtors significantly impact credit evaluations. Renters show higher credit risk ratios compared to homeowners. In contrast, managers, self-employed, or highly skilled employees often demonstrate superior creditworthiness.

Demographically, variations in credit risk emerge across gender, marital status, and foreign worker status. Married or widowed applicants tend to have better credit profiles, unlike their divorced counterparts. Foreign workers also experience lower bad credit rates than locals.

Overall, these insights underscore the importance of a holistic approach in credit risk assessment, integrating demographic, financial, and behavioral data for optimal lending decisions and risk mitigation.

References:

- [1] <https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>
- [2] <https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/>
- [3] <https://medium.com/swlh/german-credit-risk-classification-modeling-and-metrics-19182b87f060>
- [4] <https://machinelearningmastery.com/calculate-feature-importance-with-python/>
- [5] <https://medium.com/swlh/german-credit-risk-classification-modeling-and-metrics-19182b87f060>
- [6] <https://towardsdatascience.com/german-credit-data-part-1-exploratory-data-analysis-8f5f266b2426>
- [7] <https://www.kaggle.com/datasets/uciml/german-credit>

Assignment 2 Data Mining

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

31 /100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17
