Probing the potential loan risk indicators

Practicum Final Report

Author: Maria Regina Hartono

Creation Date: April 28, 2023

Table of Content

1. Objective

When working with large datasets, selecting relevant parameters can enhance the efficiency of our prediction model in the future, without the need to combine or test all possible parameters, which would be time-consuming and resource-intensive. In this report, two parameters (attributes) will be tested using financial data from the PKDD 99 competition to determine their potential usage as indicators for predicting the likelihood of good or bad loans.

Hypothesis 1:
The presence of a disponent (2nd client under the same account_id) is expected to have a positive influence on loan outcomes.

Hypothesis 2:
The monthly variation in the debtor's credit (income) is expected to influence loan outcomes.

2. General Approaches

The following general steps were taken to test the hypotheses:

1. The data was loaded into MariaDB.
2. The explanations provided on the PKDD 99 website were read.
3. Two rough hypotheses were formulated to test.
4. The relevant tables and attributes for joining were identified.
5. A schema was created for an overview of the relationships among the relevant entities.
6. The data from the relevant tables was cleaned and transformed.
7. A strategy was formulated and SQL code was used to filter the relevant results.
8. Snippets of the code were provided for the report.
9. The resulting table was exported to Excel/Pandas to create graphs for visualization.

Tools used: MariaDB, HeidiSQL, dbdiagram, codebeautify, Excel, pandas, and other online resources.

3. Schema

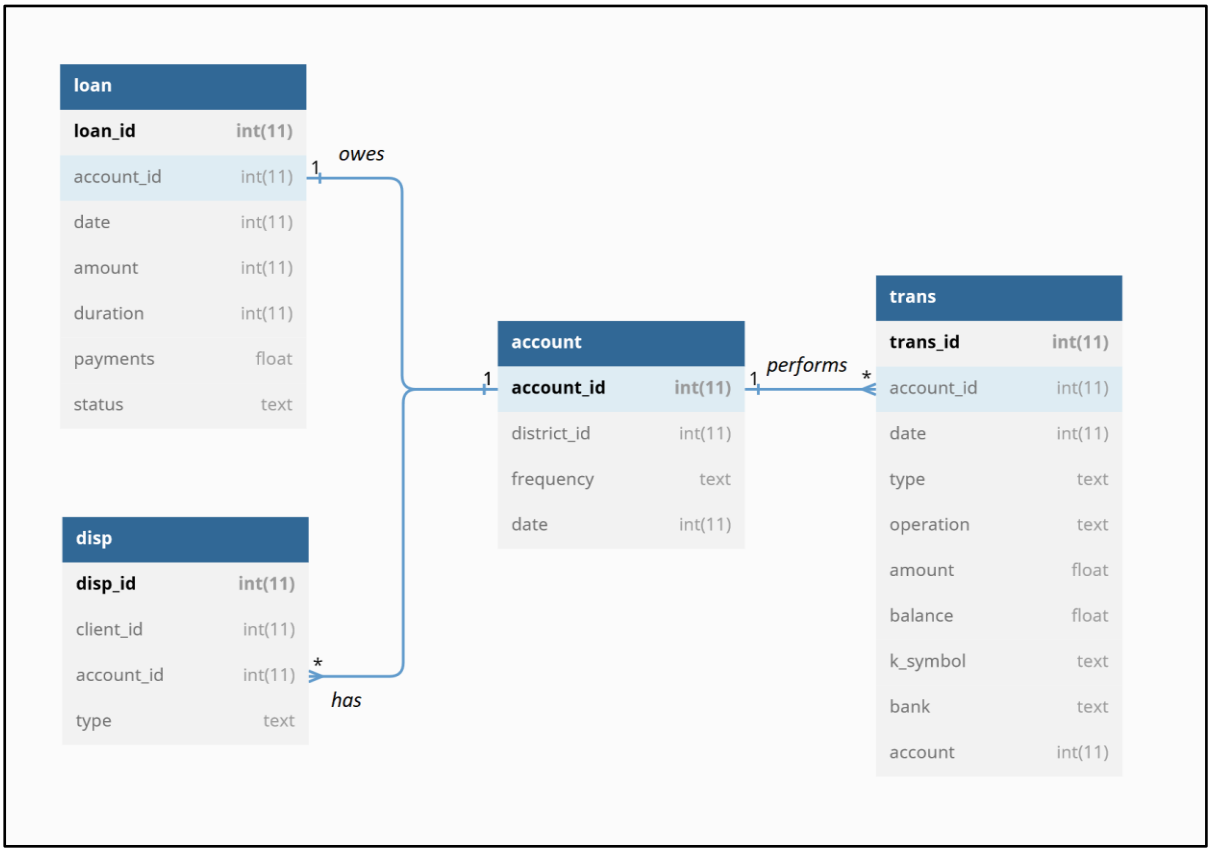

Figure 3.1 The ER Schema containing the selected tables relevant for this report. This Schema was created using https://dbdiagram.io/d and accessed on 29th April 2023.

4. Hypothesis 1:

The presence of a disponent (2nd client under the same account_id) would have positive influence on the loan outcomes

4.1. Background for Hypothesis

Reasoning:

- From a psychological standpoint, the inclusion of an additional person associated with the same ***account_id*** for loan repayment purposes, which typically necessitates a checking account, could potentially act as a form of reinforcement. This reinforcement, in turn, may create an additional level of pressure that motivates timely repayment of the loan.
- From a logistical standpoint, the probability of the loan being fully repaid increases with the number of individuals capable to access the account_id and contributing to its repayment.

4.2. Method

Tables used: loan, disp

```sql
SELECT
  ownership.status_disp,
  loan.status AS c_rate,
  COUNT(*) AS count,
  COUNT(*) * 100.0 / SUM(COUNT(*)) OVER (PARTITION BY ownership.status_disp) AS percentage
FROM
  loan
  JOIN (
    SELECT
      account_id,
      CASE
        WHEN COUNT(DISTINCT type) > 1 THEN 'with disponent'
        ELSE 'without disponent'
      END AS status_disp
    FROM
      disp
    GROUP BY
      account_id
  ) ownership
  ON loan.account_id = ownership.account_id
GROUP BY
  ownership.status_disp,
  loan.status
ORDER BY
  ownership.status_disp,
  FIELD(loan.status, 'A', 'B', 'C', 'D')
```

Figure 4.1 Screenshot of the SQL scripts. The script was formatted prior to application on HeidiSQL using https://codebeautify.org/sqlformatter for easier reading.

This SQL script analyzes loan information from the ***loan*** and ***disp*** tables to compare loan status for accounts with or without another authorized user (disponent). The subquery creates a table called ***ownership*** that groups the ***disp*** table by ***account_id*** and assigns each account to either "with disponent" or "without disponent". The main query then joins the ***loan*** and ***ownership*** tables to calculate the loan count and percentage for each combination of account ownership status and loan status. Results are grouped and ordered by account ownership status (***status_disp***) and loan status (***c_rate***).

Table 4.1 Screenshot of the resulted table

status_disp	c_rate	count	percentage
with disponent	A	55	37.93103
with disponent	C	90	62.06897
without disponent	A	148	27.56052
without disponent	B	31	5.77281
without disponent	C	313	58.28678
without disponent	D	45	8.37989

On **Table 4.1** under c_rate:

'A' stands for contract finished, no problems,
'B' stands for contract finished, loan not payed,
'C' stands for running contract, OK so far,
'D' stands for running contract, client in debt

'A' & 'C' are considered **good loans** whereas 'B' & 'D' are categorized as **bad loans**.

4.3. Results

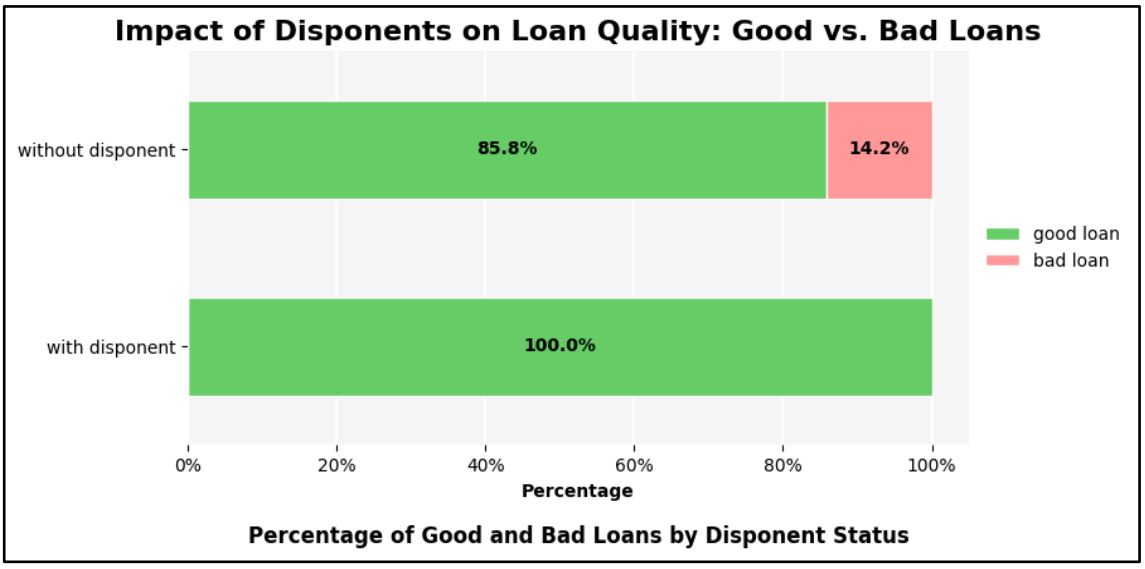

Figure 4.2 Impact of Disponents on Loan Quality: Good vs Bad Loans.

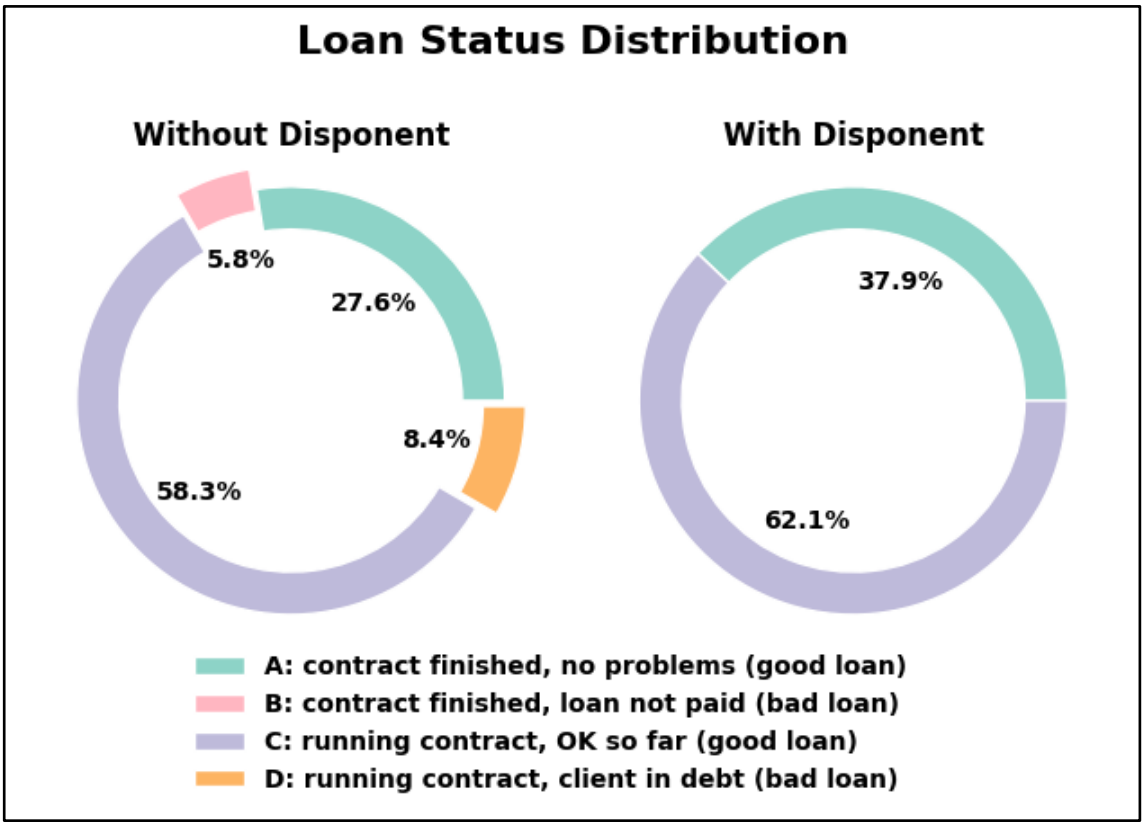

Figure 4.3 Breakdown of the loan rating category based on the presence of disponent.

Based on the information depicted in **Figure 4.2** and **Figure 4.3**, it is evident that:

- Accounts in which the disponent (2nd client) is present have a 0% bad loan among those taking loans.
- Conversely, 14% of accounts that do not have a disponent and are taking loans ended up with bad loans.

These observations support our initial hypothesis that having a disponent can act as a deterrent and reduce the likelihood of a loan becoming a bad loan.

5. Hypothesis 2:

The monthly credit (income) variation of the debtor would influence the outcome of a loan

5.1. Background for Hypothesis

Reasoning:

- A debtor with a large variation (standard deviation) in monthly credit may find it more difficult to stick to paying off their loan, as it would be harder to predict their monthly expenses.
- The actual monthly credit of borrowers may be affected by external economic conditions, such as recession or inflation. Therefore, instead of focusing on the average salary of users, it may be more useful to look at the variation of their monthly credit.

5.2. Method

Tables used: trans, loan

```sql
SELECT
    monthly_incomes.account_id,
    AVG(monthly_income) AS "average monthly credit",
    STD(monthly_income) AS "standard deviation monthly credit",
    STD(monthly_income)/AVG(monthly_income) * 100 AS percent_dev,
    loan.status
FROM (
    SELECT
        account_id,
        DATE_FORMAT(date, '%Y-%m') AS month,
        SUM(amount) AS monthly_income
    FROM trans
    WHERE type = 'credit'
    GROUP BY account_id, month
) AS monthly_incomes
JOIN loan ON monthly_incomes.account_id = loan.account_id
GROUP BY monthly_incomes.account_id
ORDER BY FIELD(loan.status, 'A', 'B','C','D'), monthly_incomes.account_id
```

Figure 5.1 Screenshot of the first SQL scripts. The script was formatted prior to application on HeidiSQL using https://codebeautify.org/sqlformatter for easier reading.

The script depicted in **Figure 5.1** retrieves and computes financial data for each individual account ID from two tables named ***trans*** and ***loan***. It calculates the average and standard deviation of monthly credit, as well as a percentage value called ***percent_dev*** (refer to *Eq. 1*). The script then joins the information with the ***loan*** table and groups results by ***account_ID***, ordered by loan ***status*** and ***account_ID***. Output includes the ***account_ID***, ***average monthly credit***, ***standard deviation of monthly credit***, ***percent_dev*** value, and loan ***status***.

$$\boldsymbol{monthly_{credit} = SUM\ (amount)_{type='credit'}} \qquad (Eq.\ 1a)$$

$$\boldsymbol{percent_{dev} = \frac{stdev(monthly_{credit})}{avg(monthly_{credit})} \times 100} \qquad (Eq.\ 1b)$$

Aggregation function was utilized for ***monthly credit*** calculation (GROUP BY ***account_ID, month***) to get the total monthly credit based on sum of ***amount*** under each transaction with **type 'credit'** in specific month for specific account. It was also being used in the ***percent dev*** calculation (GROUP BY ***account_ID***) to get the average and standard deviation of the monthly credit associated with each ***account_ID***.

Table 5.1 Screenshot of the resulted table based on script on Figure 5.1 (first set of rows)

account_id	average monthly credit	standard deviation monthly credit	percent_dev	status
2	12,516.161971078793	10,964.842683898676	87.60547130370504	A
25	41,660.40333429972	33,081.38713627092	79.40726562537729	A
67	27,194.15686341828	23,439.43900503322	86.19292417395765	A
97	10,504.506249189377	9,644.383961437647	91.81187323470719	A
132	32,652.209375023842	31,961.64187940786	97.88508187092478	A
173	13,562.48548396941	10,382.90102983496	76.55603423212762	A

The calculation of the percentage value called ***percent_dev*** might serve as a good indicator on how much the monthly credit values deviate from the average monthly credit.

```sql
SELECT
    t1.status,
    AVG(t1.percent_dev) AS "average percent deviation",
    STD(t1.percent_dev) AS "standard deviation of percent deviation"
FROM (
    SELECT
        monthly_incomes.account_id,
        AVG(monthly_income) AS "average monthly credit",
        STD(monthly_income) AS "standard deviation monthly credit",
        loan.status,
        STD(monthly_income)/AVG(monthly_income) * 100 AS percent_dev
    FROM (
        SELECT
            account_id,
            DATE_FORMAT(date, '%Y-%m') AS month,
            SUM(amount) AS monthly_income
        FROM trans
        WHERE type = 'credit'
        GROUP BY account_id, month
    ) AS monthly_incomes
    JOIN loan ON monthly_incomes.account_id = loan.account_id
    GROUP BY monthly_incomes.account_id, loan.status
) AS t1
GROUP BY t1.status;
```

Figure 5.2 Screenshot of the second SQL scripts. The script was formatted prior to application on HeidiSQL using https://codebeautify.org/sqlformatter for easier reading.

Figure 5.2 displays a script that performs calculations on the ***average*** and ***standard deviation*** of ***percent deviation values*** after grouping them according to loan status.

Table 5.2 Screenshot of the resulted table based on the script on Figure 5.2

status	average percent deviation	standard deviation of percent deviation
A	85.48674986794165	20.242554446907892
B	105.14779940545729	24.654077430150505
C	85.845729243875	22.724274669768718
D	102.18124522224925	29.719902262617932

```sql
SELECT
    t1.status AS loan_category,
    COUNT(CASE WHEN t1.percent_dev > 150 AND t1.percent_dev <= 200 THEN t1.account_id END) AS percent_dev_200,
    COUNT(CASE WHEN t1.percent_dev > 100 AND t1.percent_dev <= 150 THEN t1.account_id END) AS percent_dev_150,
    COUNT(CASE WHEN t1.percent_dev > 50 AND t1.percent_dev <= 100 THEN t1.account_id END) AS percent_dev_100,
    COUNT(CASE WHEN t1.percent_dev <= 50 THEN t1.account_id END) AS percent_dev_50,
    COUNT(t1.account_id) AS total_accounts_in_category
FROM (
    SELECT
        monthly_incomes.account_id,
        AVG(monthly_income) AS "average monthly credit",
        STD(monthly_income) AS "standard deviation monthly credit",
        loan.status,
        STD(monthly_income)/AVG(monthly_income) * 100 AS percent_dev
    FROM (
        SELECT
            account_id,
            DATE_FORMAT(date, '%Y-%m') AS month,
            SUM(amount) AS monthly_income
        FROM trans
        WHERE type = 'credit'
        GROUP BY account_id, month
    ) AS monthly_incomes
    JOIN loan ON monthly_incomes.account_id = loan.account_id
    GROUP BY monthly_incomes.account_id
) AS t1
GROUP BY t1.status;
```

Figure 5.3 Screenshot of the third SQL scripts. The script was formatted prior to application on HeidiSQL using https://codebeautify.org/sqlformatter for easier reading.

The 3rd SQL script calculates the number of accounts in each loan category that fall within specific ***percentage deviation*** ranges and also provides the total number of accounts in each category. This data might be valuable in anticipating whether a forthcoming loan will be categorized as "good" or "bad" based on the borrower's prior monthly credit records. "Good" loans are classified as those in categories A and C, whereas "bad" loans are those in categories B and D.

Table 5.3 Screenshot of the resulted table based on the script on Figure 5.3

loan_category	percent_dev_200	percent_dev_150	percent_dev_100	percent_dev_50	total_accounts_in_category
A	0	42	155	6	203
B	1	17	13	0	31
C	3	102	284	14	403
D	3	20	20	2	45

5.3. Results

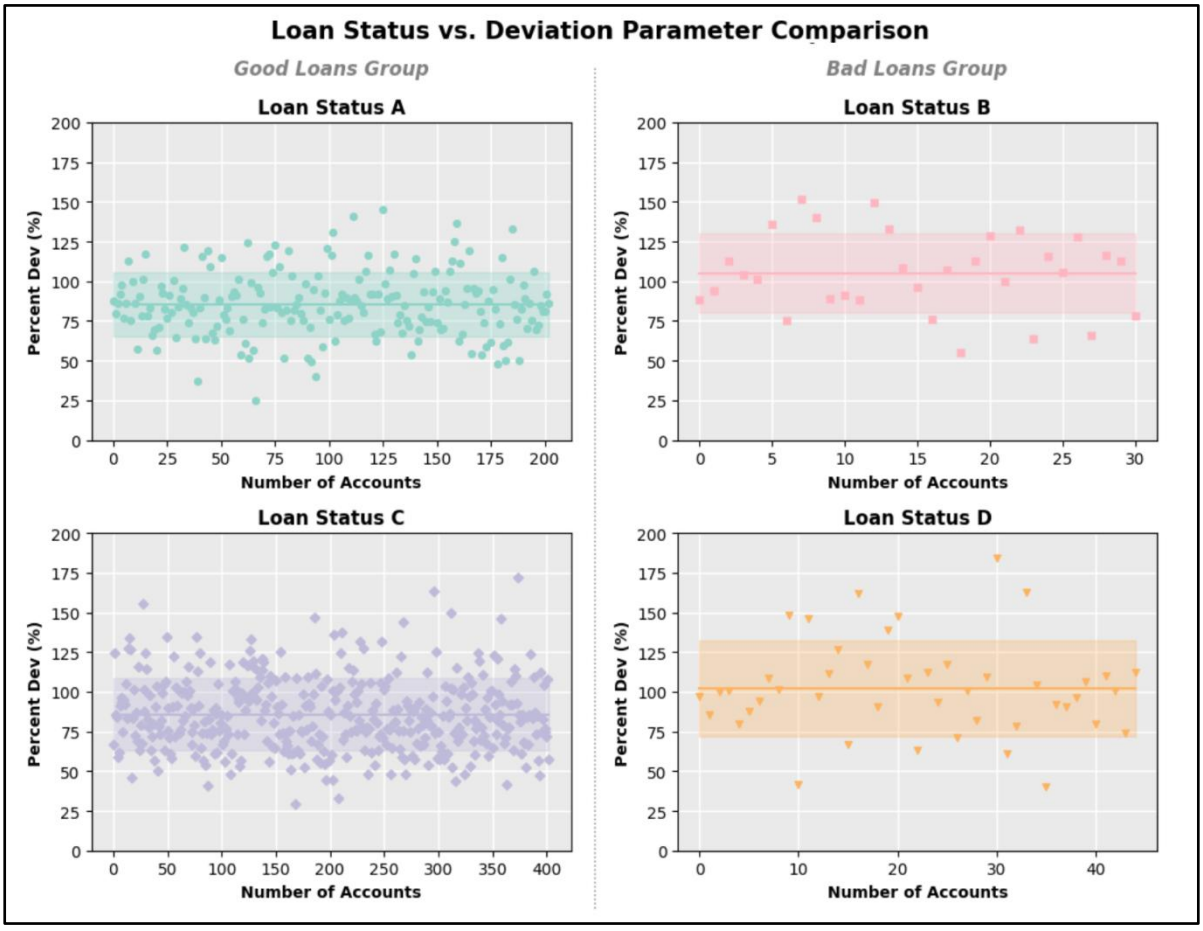

Figure 5.4 Loan Status vs. Deviation Parameter Comparison. This graph shows the distribution of individual data for percent_dev for both *Good Loans Group* and *Bad Loans Group*. The dotted line indicates the average value and standard deviation of percent_dev values in each loan status.

Figure 5.4 displays how the percentage deviation data for each loan category is spread out among individual accounts. **Figure 5.5** provides a clearer representation of the average percentage deviation differences between each loan group, while **Figure 5.6** presents the results of an Analysis of Variance (ANOVA) analysis conducted between the bad loans group and the good loans group. The analysis revealed a P-value of 6.37E-06, which is way less than 0.05, indicating a statistically significant difference between the percent_dev values of good loans and bad loans groups.

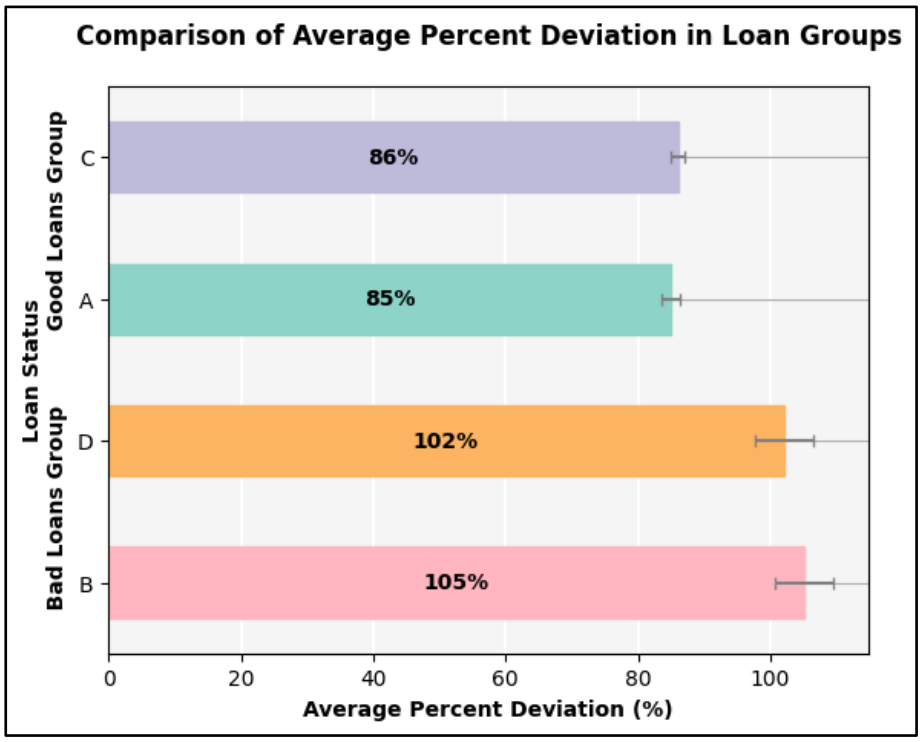

Figure 5.5 Comparison of Average Percent Deviation in Loan Groups. The error bar represents standard error which takes into account the number of samples for each loan status.

DATA SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
good loan	76	64.57523	0.849674	0.039466
bad loan	76	78.57738	1.033913	0.078335

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.28987	1	1.28987	21.89918	6.37E-06	3.904202
Within Groups	8.835055	150	0.0589			
Total	10.12493	151				

Figure 5.6 ANOVA: Single-Factor Analysis results for good loans and bad loans group. 76 sampling data were randomly selected from good loans group, whereas all the data from bad loans group were used for analysis.

Based on the visualization and ANOVA results, it appears that our original hypothesis is satisfied:

- The variation in the monthly income, here represented as percent_dev which takes into account the standard deviation of the monthly fund credited into the account divided by the average monthly credit, do appear to influence the outcome of the loan risk
- Higher percentage of deviation in the monthly income (higher ***percent_dev*)** means that the credit values for a particular account vary more widely from the average monthly credit, indicating that the account may have less stable financial behaviour and would increase the risk of bad loan and vice versa.

5.4. Predicting the loans risk based on percent_dev values

By analyzing the percentage deviation of accounts in each loan category, it might provide us with valuable information that can be used to calculate the likelihood of an account being classified as "good" or "bad" based on its deviation percentage. **Table 5.4** shows the number of accounts that fell under the specified ***percent_dev*** range and their loan category.

Table 5.4 Number of accounts under different range of percent_dev and their associated loan status. 50% was chosen as the lower limit as there were only few accounts with bad loan status below this value.

#Account that fall under the specified percent_dev range and their associated loan status					
loan_status	**150 < percent_dev ≤ 200**	**100 < percent_dev ≤ 150**	**50 < percent_dev ≤ 100**	**0 < percent_dev ≤ 50**	**total #account**
A	0	42	155	6	203
B	1	17	13	0	31
C	3	102	284	14	403
D	3	20	20	2	45
probability of bad loans	57%	20%	7%	9%	
risk factor for bad loans	10.63	2.05	0.60	0.80	

The additional analysis involving probability and risk factor for an account having certain percent_dev value to become bad loans is calculated based on the following equations:

$$probability\ of\ bad\ loans = \frac{number\ of\ accounts\ at\ specified\ percent_{dev} range\ \ in\ D\ +\ B\ category}{total\ number\ of\ accounts\ at\ the\ specified\ percent_{dev} range} \qquad (Eq.2)$$

$$risk\ factor\ of\ bad\ loans = \frac{odds\ of\ being\ a\ bad\ loan}{odds\ of\ being\ a\ good\ loan} \qquad (Eq.3)$$

$$odds\ of\ being\ a\ bad\ loans = \frac{number\ of\ accounts\ with\ specified\ percent\_dev\ in\ D\ +\ B\ category}{total\ number\ of\ accounts\ in\ D\ +\ B\ category} \qquad (Eq.3a)$$

$$odds\ of\ being\ a\ good\ loans = \frac{number\ of\ accounts\ with\ specified\ percent\_dev\ in\ A\ +\ C\ category}{total\ number\ of\ accounts\ in\ A\ +\ C\ category} \qquad (Eq.3b)$$

Table 5.4 indicates that an account with a percent_dev value of 120 is predicted to have roughly 20% chance of becoming a bad loan. Compared to being categorized as a good loan, an account with a percent_dev of 120 is about 2.05 times more likely to be classified as a bad loan.

The percent_dev (variation in monthly income/credit) might have the potential to serve as a relevant parameter for predicting the risk of bad loans. **Figure 5.7** demonstrates a rough prediction of the probability of bad loans based on percent_dev value using a 2nd order polynomial model.

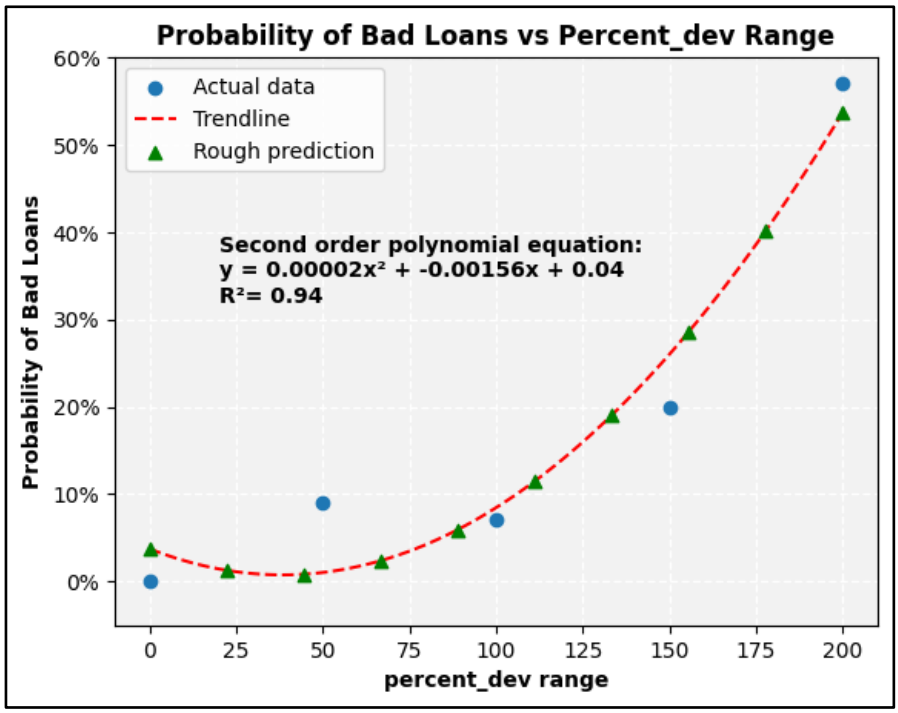

Figure 5.7 Rough prediction of the probability of bad loans with respect to percent_dev. The upper limits of the percent_dev range shown on Table 5.4 were used to construct the 2nd order polynomial model.

The fitting remains insufficient (R^2 = 0.94), particularly for lower spectrum of percent_dev values, and necessitates the inclusion of multi-parameters and adoption of more advanced models (such as machine learning) to establish a more accurate evaluation of loan risk.

6. Conclusion

In conclusion, the current analysis has revealed two key factors that can play a crucial role in determining the risk of bad loans - the presence of a disponent and monthly credit variation. Based on the findings, it is recommended that lenders take both of these factors into account when assessing loan applications. Specifically:

1. Having a disponent (2nd client) in the same account could potentially reduce the likelihood of a loan becoming a bad loan, as accounts with a disponent had a 0% bad loan rate among those taking loans, while accounts without a disponent had a 14% bad loan rate.
2. Monthly credit variation, as represented by percent_dev, has a statistically significant impact on the likelihood of a loan risk. Higher percent_dev indicates less stable financial behavior and increased risk of bad loans.
3. To make more informed decisions and reduce the risk of bad loans, lenders should also consider both the presence of a disponent and monthly credit variation as some of the good indicators when assessing loan applications.