

**University of Carthage
Ecole Polytechnique de Tunisie**



Data Analysis Project

Realized by :

MAHDI MCHEIHI

TAIB CHARFI

Supervised by :

AMOR MESSAOUD

Academic Year : 2023/2024

Rue Elkhawarezmi BP 743 La Marsa 2078

Tel : 71 774 699 Fax : 71 748 843

Site web : www.ept.rnu.tn

Contents

General introduction	1
1 Data Description	1
1.1 Features Overview	1
2 EDA:Exploratory data analysis	3
2.0.1 Correlation Matrix	3
2.0.1.1 Strong Linear Relationship	3
2.0.1.2 High correlations reveal strong relationships between features .	3
2.0.2 Chi Student Test	3
3 Data Visualization	6
3.1 Employment Status Client	6
3.2 Level of education client	7
3.3 Outlier Analysis	7
3.3.1 Level Of Education Clients	8
3.3.2 Employment Status Clients	8
3.3.3 Distribution Of Employment Status Client	10
3.3.4 Data Histogram	11
4 Feature engineering	12
4.1 Log Function	12
5 PCA	14
5.1 PCA visualisation	14
6 Tests	16
6.1 Fisher Test	16
6.1.1 Employment Status Clients And Level Of Education Clients	16
6.1.2 Referredby And Bank Account Type	16
6.2 Anova Test	17
7 Model Training and Evaluation	18
7.1 CatBoostClassifier:	18
7.2 XGBClassifier:	18

List of Figures

2.1	Correlation Matrix	4
2.2	chi student test:goodbadflag	5
3.1	employment client	6
3.2	Level of education client	7
3.3	Outlier Analysis:Level Of Education Clients	8
3.4	Outlier Analysis:Employment Status Clients	9
3.5	Distribution Of Employment Status Client	10
3.6	Data Histogram	11
4.1	Distribution after log function	12
4.2	relation between loanamont and totaldue	13
5.1	PCA	14
5.2	PCA1	14
5.3	PCA2	15
5.4	PCA1 and 2	15
6.1	PValue	16
6.2	PValue2	16
6.3	PValue and FStatistic	17
7.1	CatBoostClassifier	18
7.2	XGBClassifier	18

List of Tables

Chapter 1

Data Description

This dataset, sourced from a financial institution, encompasses a variety of variables that capture key client and financial details. Each record in the dataset represents a unique client profile, including aspects such as the type of bank account they hold, the amount of loan disbursed, and personal demographic information.

1.1 Features Overview

- **id**: unique identifier for each client
- **systemloanid**: uniquely identifies a specific loan transaction within the system.
- **loannumber**: refers to the sequence or count of loans associated with an individual client or a specific account
- **approveddate**: refers to the date on which a loan application was formally approved
- **creationdate**: refers to the date and possibly time when a loan account or a customer account was created in the institution's system.
- **Loanamount**: Refers to the principal amount of money borrowed by a client from a financial institution under the terms of a loan agreement.
- **Totaldue**: represents the total amount that a borrower is obligated to repay to a lender.
- **termdays**: Refers to the duration of a loan
- **referredby**: storing IDs or codes that link back to existing clients or affiliate partners
- **good bad flag** : used to categorize the creditworthiness or repayment behavior of borrowers
- **birthdate**: the date of birth of each client
- **bank account type**: the classification of a client's bank account based on its characteristics and functionalities
- **longitude gps**: longitudinal coordinate of a location
- **latitude gps**: latitudinal coordinate of a location
- **bank name clients**: the name of the bank where clients hold accounts.
- **bank branch clients**: refers to the specific branch location of the bank where clients have their accounts
- **employment status clients**: records the employment status of clients
- **level of education clients**: level of education clients
- **loanamount mean**: The mean of loan

- **totaldue mean:** The mean of the total due.
- **termdays mean:**the term day's mean

The dataset comprises both numerical and categorical features. Categorical features will be encoded numerically to enhance their interpretability by the model. These numerical features provide valuable information for statistical analysis and modeling, allowing us to examine relationships, trends, and patterns in the data.

Chapter 2

EDA:Exploratory data analysis

Exploratory data analysis (EDA) is used to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

2.0.1 Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table represents the correlation coefficient between two variables. The correlation coefficient measures the strength and direction of the linear relationship between two variables. There is a high correlation between the features ::

"loanamount" , "loannumber" with the value of 0.83

"loanamount" , "totaldue" with the value of 0.99

"totaldue" , "loannumber" with the value of 0.82

Those high correlation values can inform us about two things whether ::

2.0.1.1 Strong Linear Relationship

Features with high correlation coefficients move almost perfectly together. When one feature increases, the other tends to increase predictably, and vice versa. This suggests multicollinearity, which can affect model performance. Removing one of the features may simplify the model and improve its generalization. Feature Engineering Opportunities:

2.0.1.2 High correlations reveal strong relationships between features

This presents opportunities for feature engineering, where transformations or combinations of correlated features can capture complex relationships, enhancing predictive power. Further exploration is needed to determine the significance of these relationships.

2.0.2 Chi Student Test

this function provides a convenient way to perform and interpret the results of a chi-squared test for independence between categorical variables. In the analysis of the dataset, a Chi-squared test was employed to ascertain the relationship between various features and the target variable "goodbadflag". The results suggest that most features do not exhibit a statistically significant relationship with the target variable. However, four features have been identified as significantly associated with "goodbadflag":

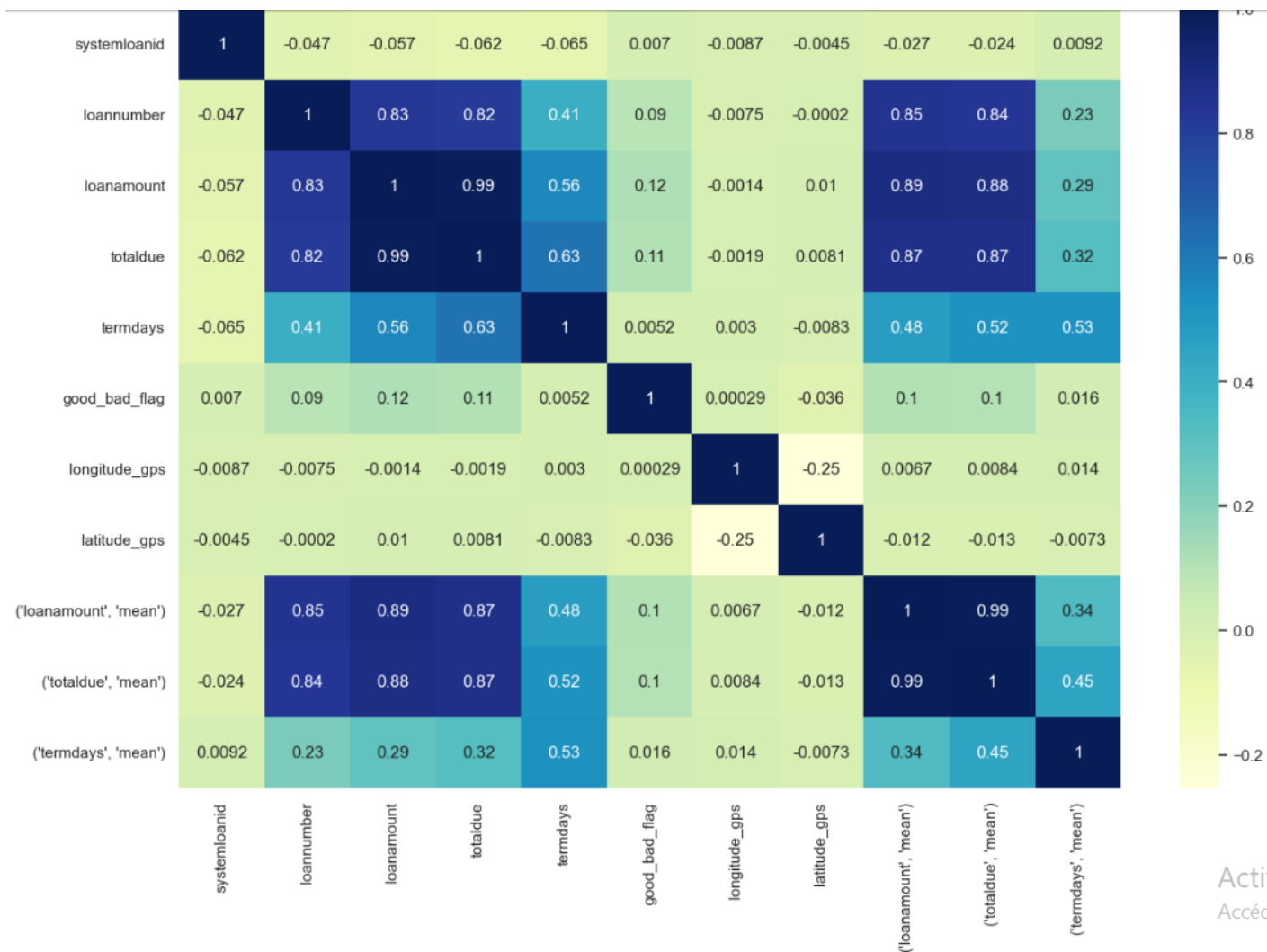


Figure 2.1 – Correlation Matrix

"Loannumber", "Loanamount", "Bankaccounttype", "Totaldue"

These findings indicate that while many personal and demographic details (like GPS data, employment status, and level of education) do not directly impact the creditworthiness as defined in this context, specific loan characteristics and financial behaviors (such as the amount and number of loans, and the type of bank account) are key indicators of loan performance.


```
customerid does not have a significant relationship with the good_bad_flag variable.
systemloanid does not have a significant relationship with the good_bad_flag variable.
loannumber has a significant relationship with the good_bad_flag variable.
approveddate does not have a significant relationship with the good_bad_flag variable.
creationdate does not have a significant relationship with the good_bad_flag variable.
loanamount has a significant relationship with the good_bad_flag variable.
totaldue has a significant relationship with the good_bad_flag variable.
termdays does not have a significant relationship with the good_bad_flag variable.
referredby does not have a significant relationship with the good_bad_flag variable.
good_bad_flag has a significant relationship with the good_bad_flag variable.
birthdate does not have a significant relationship with the good_bad_flag variable.
bank_account_type has a significant relationship with the good_bad_flag variable.
longitude_gps does not have a significant relationship with the good_bad_flag variable.
latitude_gps does not have a significant relationship with the good_bad_flag variable.
bank_name_clients does not have a significant relationship with the good_bad_flag variable.
bank_branch_clients does not have a significant relationship with the good_bad_flag variable.
employment_status_clients does not have a significant relationship with the good_bad_flag variable.
level_of_education_clients does not have a significant relationship with the good_bad_flag variable.
('loanamount', 'mean') does not have a significant relationship with the good_bad_flag variable.
('totaldue', 'mean') does not have a significant relationship with the good_bad_flag variable.
('termdays', 'mean') does not have a significant relationship with the good_bad_flag variable.
```

Figure 2.2 – chi student test:goodbadflag

Chapter 3

Data Visualization

Extensive data visualization is conducted to understand the distribution of features, identify potential outliers, and explore relationships between variables.

3.1 Employment Status Client

"Employment status of clients is a key factor to consider in loan behavior analysis. The following pie chart provides an overview of the various employment categories within our client database.

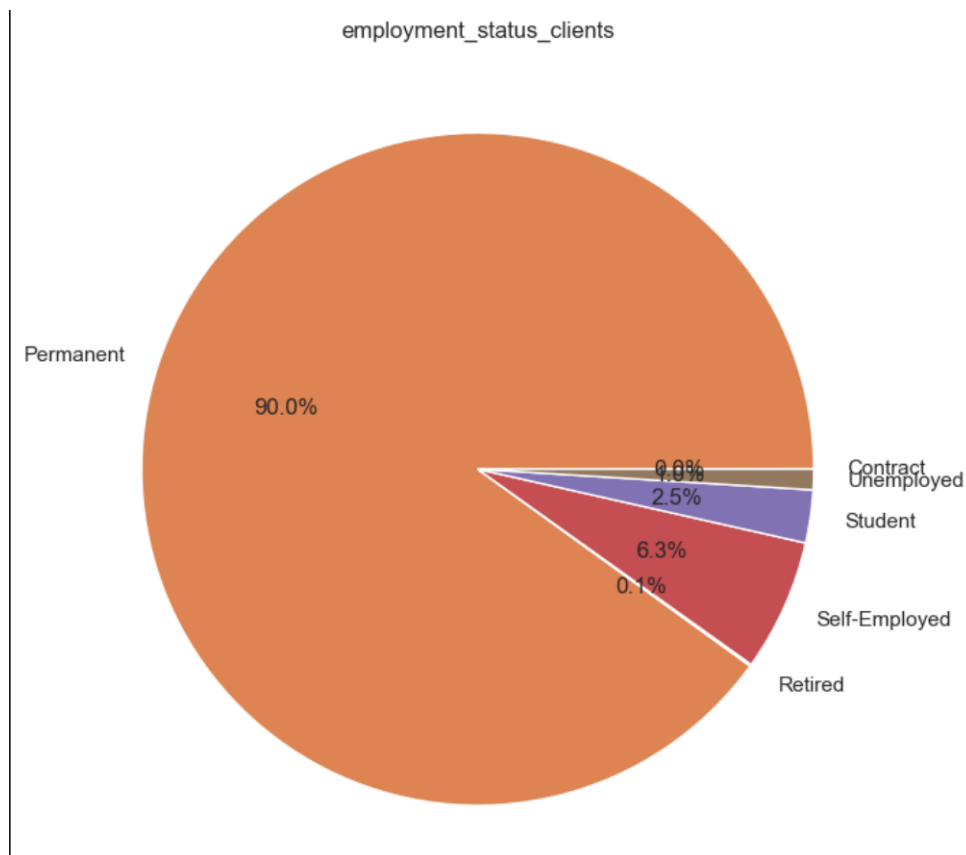


Figure 3.1 – employment client

Based on the pie chart, it's evident that the majority of employees, constituting 90 percent of the total, are engaged in permanent employment. Following this category, we observe self-employment accounting for 6.3 percent, and students comprising 2.5 percent. Finally, the minority groups, including retirees and the unemployed, represent just 1.0 percent of the distribution.

3.2 Level of education client

Now, we will export the education level of the clients. This feature provides essential information for understanding our data. As observed, the majority of the distribution com-

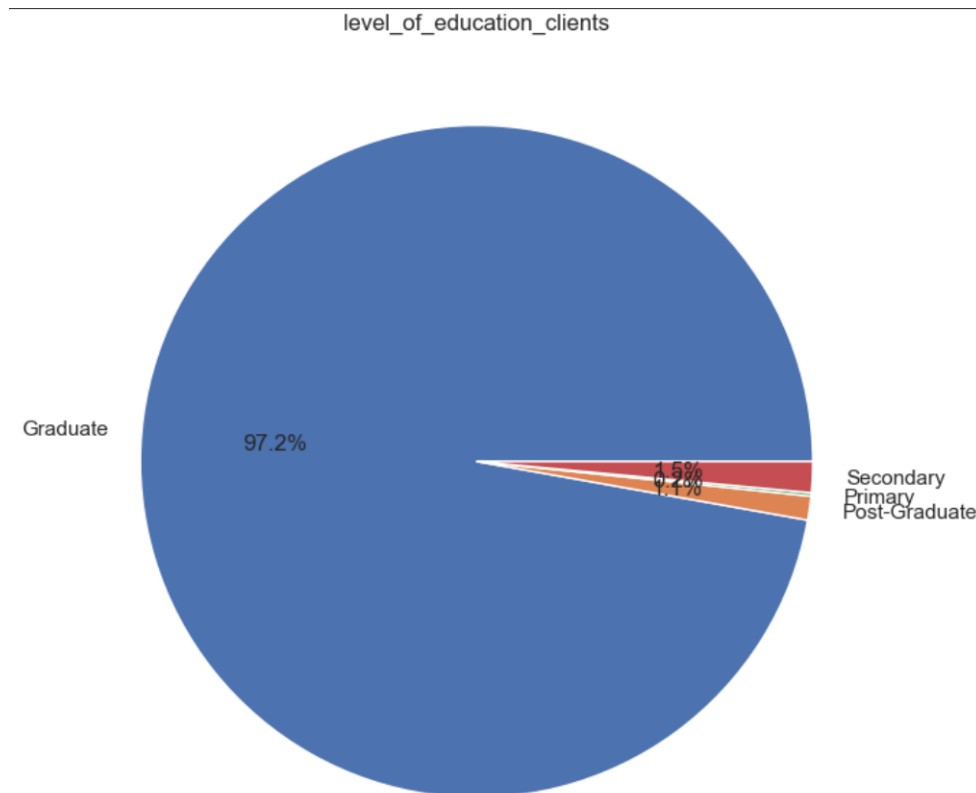


Figure 3.2 – Level of education client

prises graduate clients, accounting for 1.1 percent. Following this, we find 1.1 percent of post-graduates and 1.5 percent of individuals with a secondary education. Lastly, those with primary education represent a smaller portion, making up 0.2 percent of the total.

3.3 Outlier Analysis

Boxplots and Kernel Density Estimation (KDE) plots reveal outliers in various features, notably in loanamount, termdays, and level of education clients. These outliers may necessitate careful consideration during data preprocessing to prevent undue influence on model performance.

3.3.1 Level Of Education Clients

We opted to designate the loan amount as the y-axis label, term days as the x-axis label, and Level Of Education Clients as the hue for visualization.

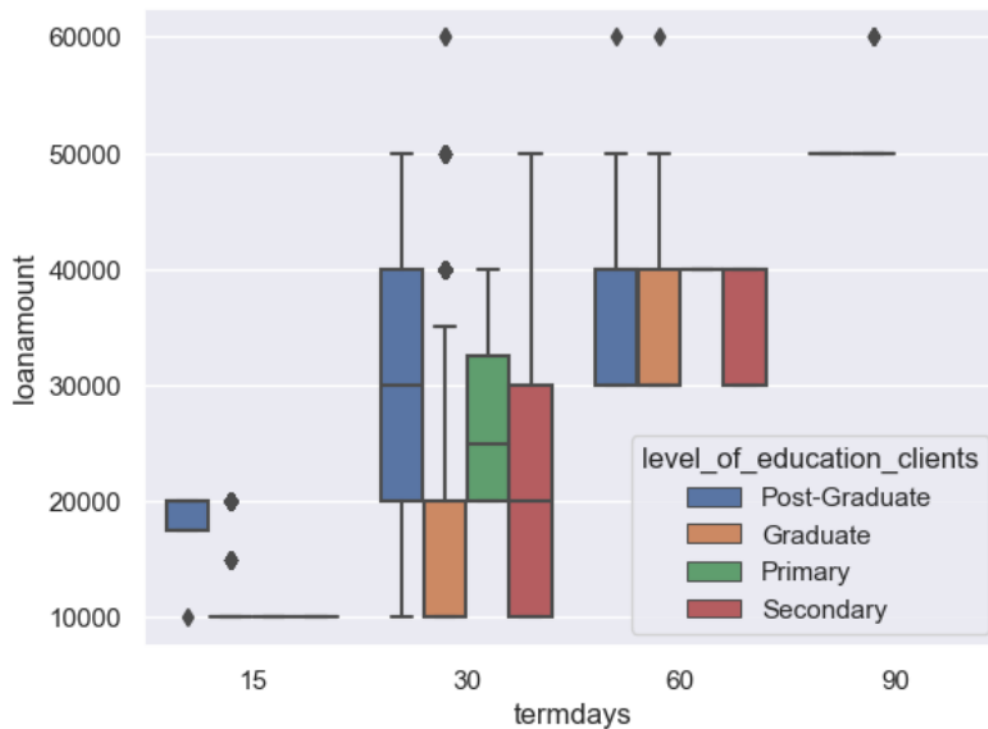


Figure 3.3 – Outlier Analysis:Level Of Education Clients

Upon analyzing the distribution of loan amounts based on term days, we discovered insightful patterns. Approximately half of the clients, whose loans must be repaid within 15 months, exhibit loan amounts ranging between 17,500 and 20,000 units, with the majority being post-graduates.

For loans with a term of 30 months, the distribution shifts slightly. About half of the post-graduates' loans fall within the range of 20,000 to 40,000 units, while graduated clients generally have loans below 20,000 units. Notably, there are outliers in this category.

In the case of loans with a 60-month term, a consistent trend emerges across all education levels. Roughly 50 percent of loans, regardless of educational background, fall between 30,000 and 40,000 units, with occasional outliers. However, loans for primary education show a deviation, clustering between 20,000 and 32,500 units.

3.3.2 Employment Status Clients

We opted to designate the loan amount as the y-axis label, term days as the x-axis label, and employment status of clients as the hue for visualization.

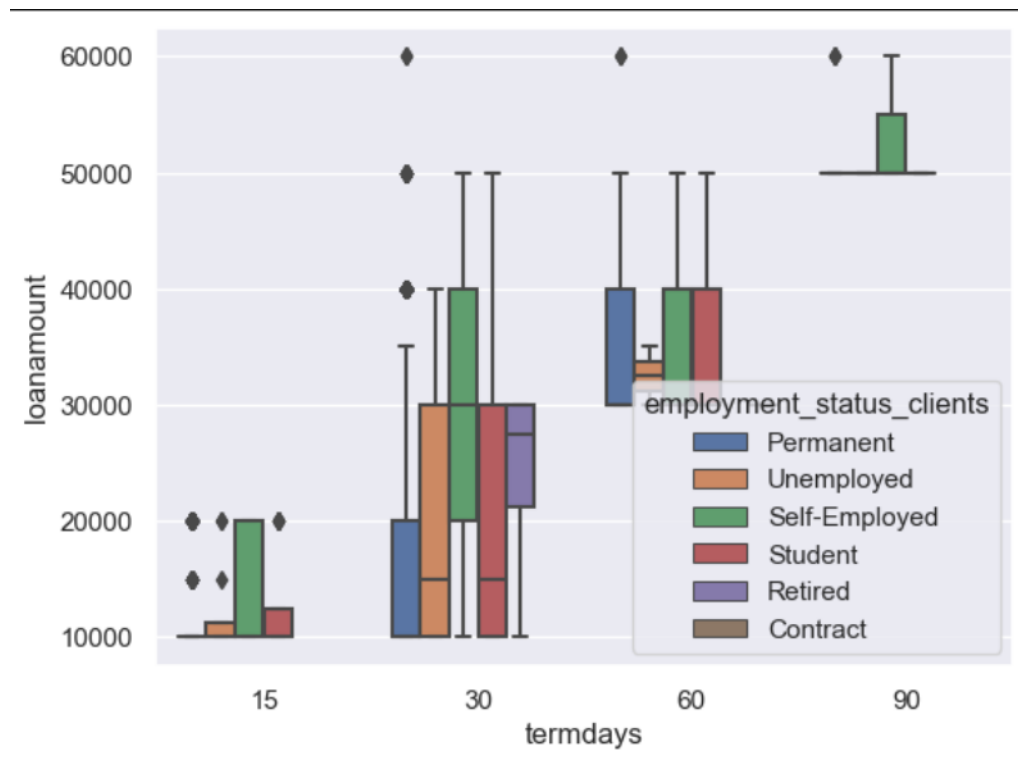


Figure 3.4 – Outlier Analysis: Employment Status Clients

Upon scrutinizing the loan amount distribution based on term days and clients' employment statuses, intriguing insights emerged. For loans with a 15-month repayment term, roughly half of the clients fall under the self-employed category, with loan amounts spanning from 1,000 to 20,000 units. Notably, students and unemployed individuals tend to have loans below 12,500 units, with occasional outliers

With a term of 30 months, a nuanced shift in distribution occurs. Approximately half of the loans held by permanent employees range from 10,000 to 20,000 units, while students and unemployed individuals see loan amounts reaching up to 30,000 units, maintaining a consistent median. Noteworthy is the range of loan amounts for self-employed individuals, which spans from 20,000 to 40,000 units, and retirees typically receive loans between 20,000 and 30,000 units.

Moving to a 60-month term, a uniform trend emerges across all employment statuses. Roughly 50 percent of loans, regardless of employment status, fall within the 30,000 to 40,000 unit range, punctuated by occasional outliers. However, loans designated for primary education exhibit a slight deviation, clustering between 31,000 and 32,500 units

For a 90-month term, only self-employed individuals are observed, receiving loans ranging from 50,000 to 55,000 units. This distinct pattern highlights the higher loan amounts sought by self-employed clients for extended-term commitments.

3.3.3 Distribution Of Employment Status Client

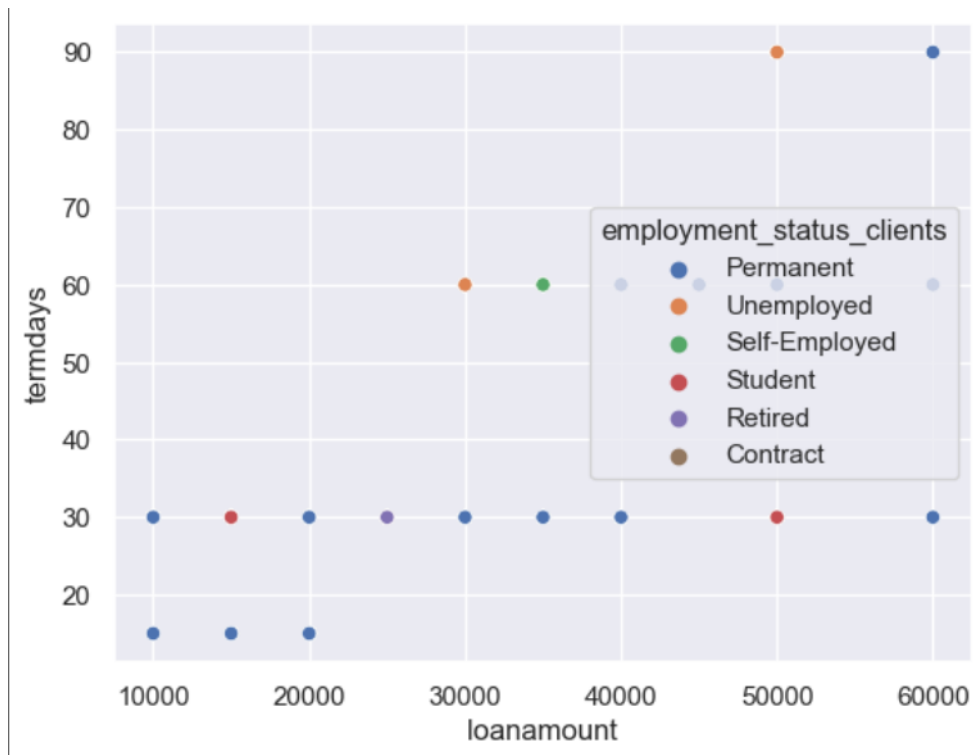


Figure 3.5 – Distribution Of Employment Status Client

3.3.4 Data Histogram

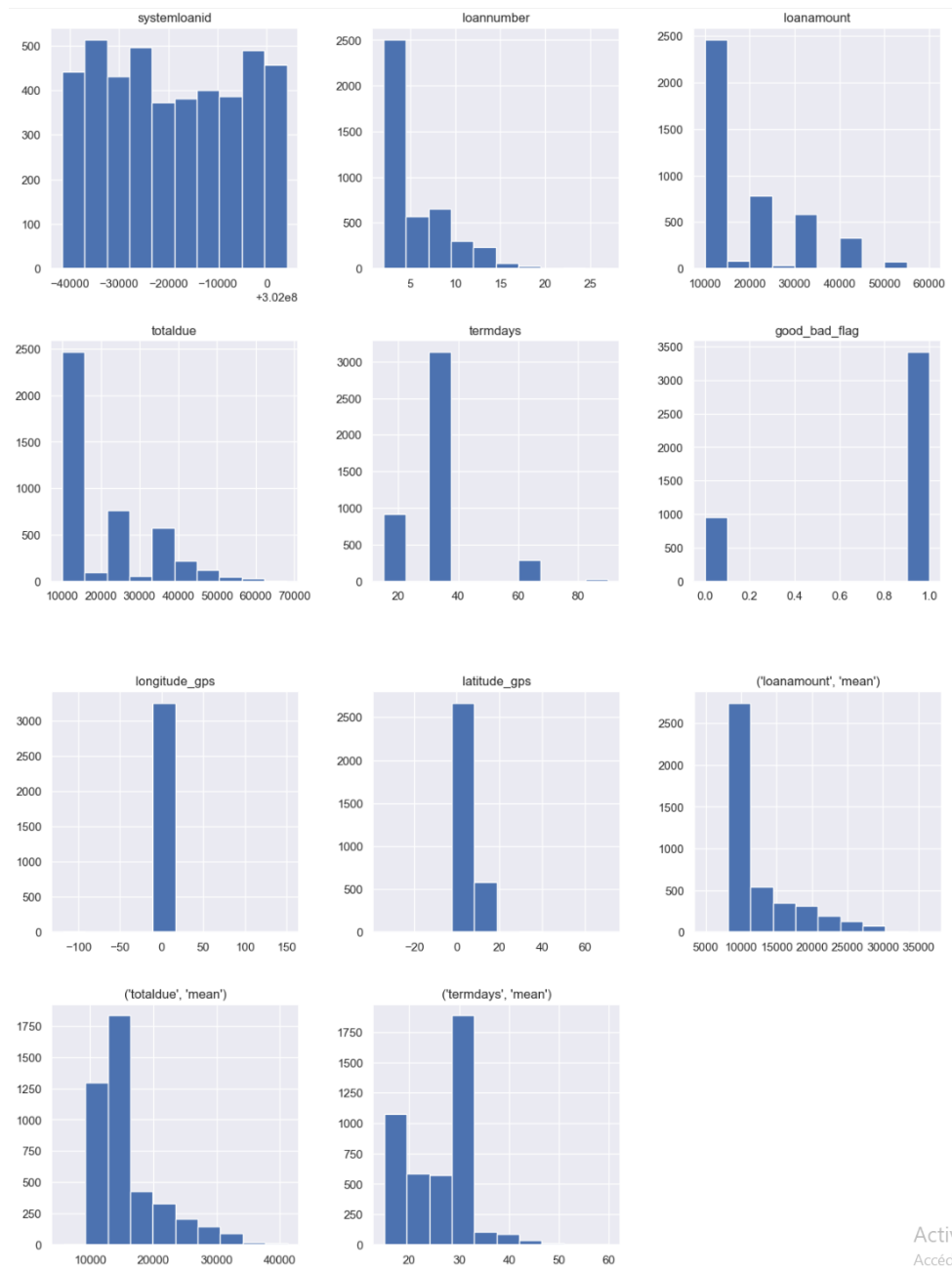


Figure 3.6 – Data Histogram

Chapter 4

Feature engineering

4.1 Log Function

Applying a logarithmic function as part of feature engineering in data analysis helps in managing skewed data, enhancing the effectiveness of various statistical and machine learning models, and improving the overall interpretability and robustness of the analysis.



Figure 4.1 – Distribution after log function

This histogram represents the distribution of logarithmic transformations of three different loan-related variables from a dataset: the average total due (totaldue), the average loan amount (loanamount), and the total number of loans (loannumber). By applying a logarithmic scale, the data is normalized.

Now we will show the new relation between "loanamount" and "totaldue"

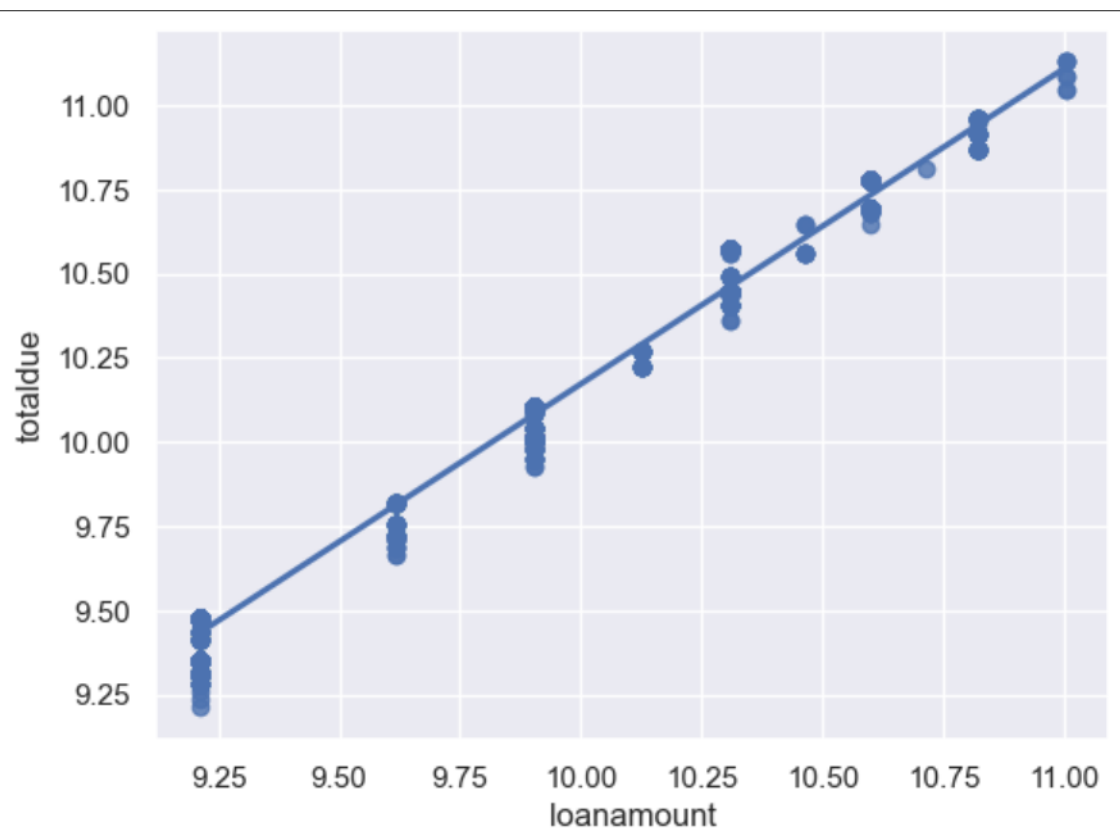


Figure 4.2 – relation between loanamont and totaldue

This scatter plot illustrates a strong linear relationship between the log-transformed loanamount and totaldue, indicating that increases in the loan amount are proportionally related to increases in the total due. The consistent spread of data points suggests a uniform variance, highlighting the effectiveness of the log transformation in stabilizing variance and reducing skewness. The presence of a direct correlation is beneficial for predictive modeling, as it allows for straightforward estimations of total due based on given loan amounts. This linear model could serve as a reliable tool for financial institutions to predict financial outcomes and optimize loan strategies.

Chapter 5

PCA

PCA stands for Principal Component Analysis. It's a technique used in data analysis and machine learning to simplify the complexity of high-dimensional data while retaining trends and patterns. Explained variance ratio is $[9.99991411e-01, 3.23588439e-06]$: indicates that the

```
Original data shape: (4368, 599)
Projected data shape: (4368, 2)
Components shape: (2, 599)
Explained variance ratio: [9.99991411e-01 3.23588439e-06]
```

Figure 5.1 – PCA

first principal component explains approximately 99.999 percent of the variance in the data, while the second principal component explains approximately 0.0003236percent of the variance.

To conclude , i can delete the second component and only use the first one .

5.1 PCA visualisation

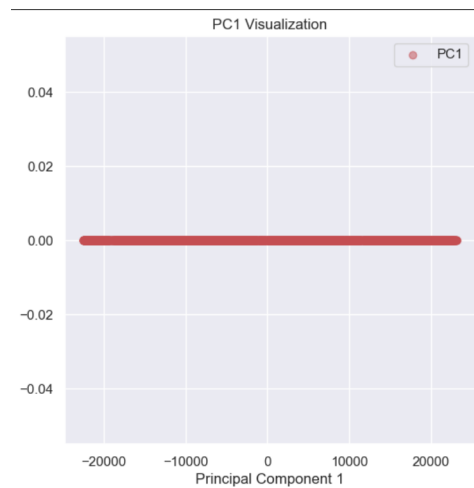


Figure 5.2 – PCA1

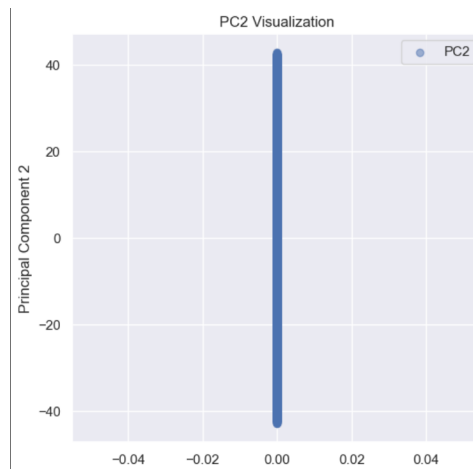


Figure 5.3 – PCA2

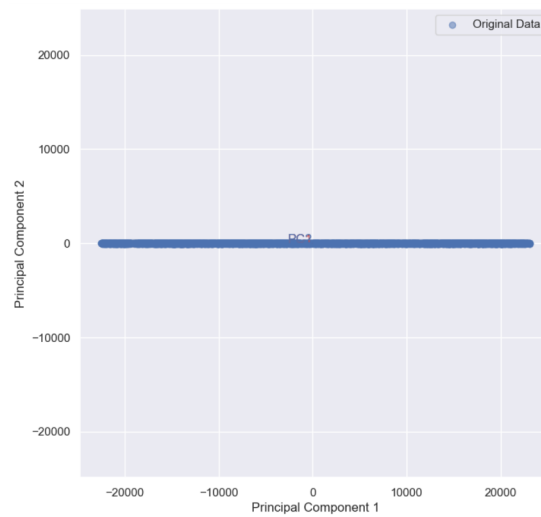


Figure 5.4 – PCA1 and 2

Chapter 6

Tests

6.1 Fisher Test

This statistical test is particularly valuable for small sample sizes or when the data in a contingency table is unevenly distributed, potentially leading to unreliable results from chi-squared tests due to small expected frequencies. The test examines the null hypothesis, which posits no association between the variables, against an alternative hypothesis suggesting a relationship exists. Through calculating the exact probability of observing the data as it is (or more extreme) under the assumption of independence, this test helps to clarify whether any observed correlation between the variables is statistically significant.

***hypothesis (H0)** that asserts there is no association between the variables, and an alternative.

***hypothesis (H1)** there is an association between the variables

6.1.1 Employment Status Clients And Level Of Education Clients

```
[57]: p_value = fisher_exact_custom(contingency_table)
      print("P-value:", p_value)

P-value: 0.002289377289377289
```

Figure 6.1 – PValue

P Value is less than 0.05, then we conclude that the null hypothesis is rejected and there is a significant association between the variables.

6.1.2 Referredby And Bank Account Type

```
[59]: p_value = fisher_exact_custom(contingency_table)
      print("P-value:", p_value)

P-value: 0.0013736263736263735
```

Figure 6.2 – PValue2

The same for the second case P value is less than 0.05, Then we conclude that the null hypothesis is rejected and there is a significant association between the variables.

6.2 Anova Test

In this segment of the analysis, we utilize the ANOVA (Analysis of Variance) test to determine whether there are statistically significant differences between the mean loan amounts across different types of bank accounts: Savings, Other, and Current.

hypothesis (H0) posits that there is no difference in mean loan amounts among these groups

hypothesis (H1) suggests that at least one group's mean is distinct. Pvalue is less than 0.05, then

```
F-statistic: 310.02233125523935  
P-value: 1.2673943953085132e-126
```

Figure 6.3 – PValue and FStatistic

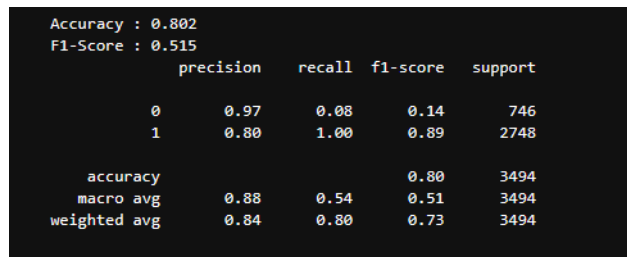
we conclude that the null hypothesis is rejected and there is a significant association between the variables.

Chapter 7

Model Training and Evaluation

7.1 CatBoostClassifier:

: For the CatBoost Classifier, we used the following parameters: random state=42,max depth=4,reg lambda=2. CatBoost automatically handles categorical features and offers robust performance with relatively fewer parameter adjustments. We see here the importance of the



```
Accuracy : 0.802
F1-Score : 0.515
      precision    recall  f1-score   support

      0       0.97       0.08       0.14       746
      1       0.80       1.00       0.89      2748

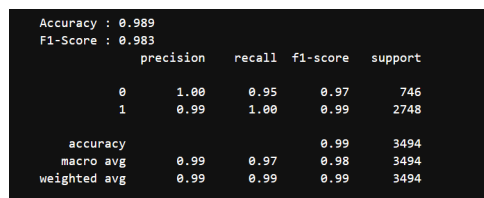
 accuracy
macro avg       0.88       0.54       0.51      3494
weighted avg       0.84       0.80       0.73      3494
```

Figure 7.1 – CatBoostClassifier

F1-score underscores a crucial aspect of model evaluation. While accuracy provides a straightforward measure of overall correctness, it can sometimes be misleading. A high accuracy might initially suggest that the model is performing well and on the right track. However, the F1-score, which takes into account both precision and recall, offers a more nuanced perspective. The F1-score, with its balanced consideration of precision and recall, offers a more reliable measure, revealing potential shortcomings that accuracy alone might overlook.

7.2 XGBClassifier:

Similar to LightGBM, we tuned the XGBoost Classifier parameters: n estimators = 1700, max depth = 7, learning rate = 0.09, colsample bytree = 0.8, reg lambda = 0.5. The additional reg lambda parameter was used to control regularization and further prevent overfitting.



```
Accuracy : 0.989
F1-Score : 0.983
      precision    recall  f1-score   support

      0       1.00       0.95       0.97       746
      1       0.99       1.00       0.99      2748

 accuracy
macro avg       0.99       0.97       0.98      3494
weighted avg       0.99       0.99       0.99      3494
```

Figure 7.2 – XGBClassifier

The increase in the F1 score with the new model is a positive sign of potential improvement. However, upon closer examination, particularly regarding precision for the 'good bad flag' class labeled as '0', which shows a precision value of '1.00', we suspect overfitting. This perfect precision suggests that the model may be fitting too closely to the training data, potentially hindering its ability to generalize to unseen data. Detecting overfitting is critical as it can lead to inflated performance metrics during training but poor performance on new data. To address this, we may need to revisit the model architecture, regularization techniques, or dataset balancing methods to promote better generalization, ensuring effectiveness in real-world applications.