

Credit Exploratory Data Analysis: Case Study – Assignment

The objective of this case study is to identify patterns that indicate whether a customer may have difficulty repaying their loan installments. These insights can help the company make informed decisions, such as denying loans, adjusting loan amounts, or offering higher interest rates to higher-risk applicants. This approach ensures that customers who are capable of repaying the loan are not unfairly rejected.

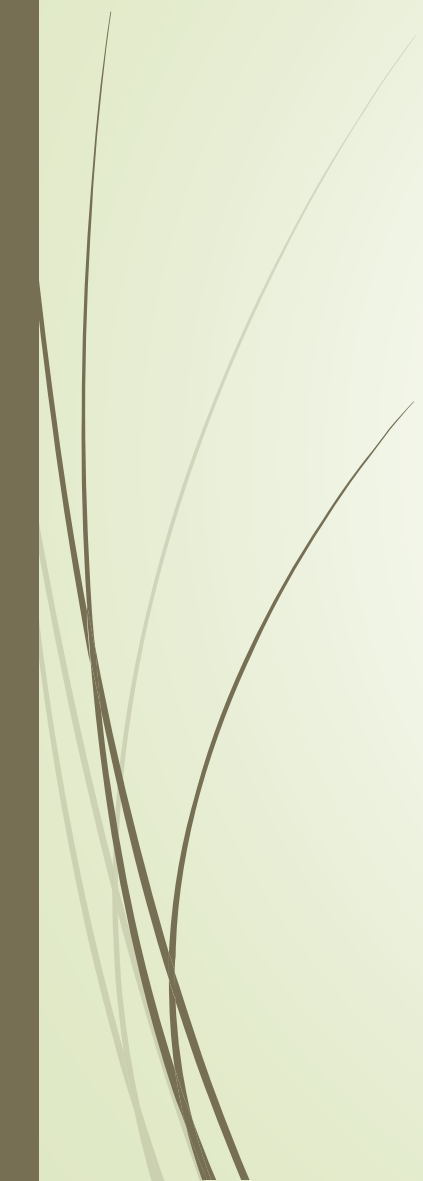
There are two key risks involved:

- **If the applicant is likely to repay the loan**, rejecting the loan results in a loss of potential business for the company.
- **If the applicant is likely to default**, approving the loan can lead to financial losses for the company.

Using **Exploratory Data Analysis (EDA)**, this case study aims to identify such applicants and mitigate these risks.



Problem Statement



A consumer finance company that provides loans to customers wants to identify patterns among clients who struggle to repay their installments. By analyzing these patterns, the company can make better lending decisions, such as rejecting high-risk applications, reducing loan amounts, or offering higher interest rates to risky borrowers. This will help improve their overall portfolio and manage financial risk more effectively



Datasets Provided for Analysis:

The analysis involves two main datasets:

1. **Application Data**

2. **Previous Application Data**

- Both datasets are provided in **CSV (Comma Separated Values)** format for easy handling.
- Additionally, there is a file with **Column Descriptions**, which helps in understanding the role and contribution of each column in the analysis.
- **Prerequisites:**
- To perform the analysis, the following tools and libraries are required:
 - **Programming Language:** Python
 - **Platform:** Jupyter Notebook
 - **Libraries:** Pandas, Numpy, Matplotlib, Seaborn, itertools, and handling warnings
- These tools will support a smooth and effective data analysis process.

Data Understanding

1. Application_data.csv

- Number of Columns: 122
- Number of Rows: 307511
- Data Types: Integer, Float, Strings
- Descriptive view of Data file: There were anomalies like negative numbers, Null values, Days and Years were not in proper Format
 - Float64: 65
 - Int64: 41
 - Object: 16 1.

2. Previous_Application_data.csv

- Number of Columns: 37 2.
- Number of Rows: 1670214
- Data Types: Integer, Float, Strings
- Descriptive view of Data file: There were anomalies like negative numbers, Null values, Days and Years were not in proper
 - Format Float64: 15
 - Int64: 06
 - Object: 16

Overview

- **Analysis Design**

- The analysis will be conducted in three main steps:

- 1. Exploratory Data Analysis (EDA):**

1. Understand and clean the dataset.
2. Identify key columns that are relevant for the analysis.

- 2. Univariate, Bivariate, and Multivariate Analysis:**

1. Analyze the target variable (categorical) and numerical variables in isolation and in combination.

- 3. Identify High-Risk Variables:**

1. Determine which variables are most useful for predicting high-risk customers.

- **Key Variables for Analysis**

- **TARGET:** Customers who have missed at least one payment. This will be the primary variable for analysis, with other variables compared against it.
- **NAME_CONTRACT_STATUS:** The status of previous loan applications. This will serve as a crucial variable in understanding loan history and risk.

- **Assumption:**

- Since we don't have information on the number of missed payments, we will treat all customers with at least one missed payment as having similar risk levels.

Exploring Data Sets

Analysis : Focusing on preparing the dataset for predictive modeling.

Understanding the Data Structure:

- Examining data dimensions and column types.
- Identifying missing values and outliers.

Data Cleaning & Preparation:

- Removing columns with excessive missing values.
- Developing a strategy to handle outliers.
- Correcting incorrect data types.

Data Transformation:

- Creating new columns through binning.
- Removing irrelevant columns.
- Structuring the dataset for further analysis.



Data Cleaning & Manipulation for Application Data:

How we did that?

- Rectify the null values.
- Filtering unwanted data columns.
- Filling the missing values.
- Sorting the data.
- Fixing the datatype

To Remove unwanted or irrelevant columns,

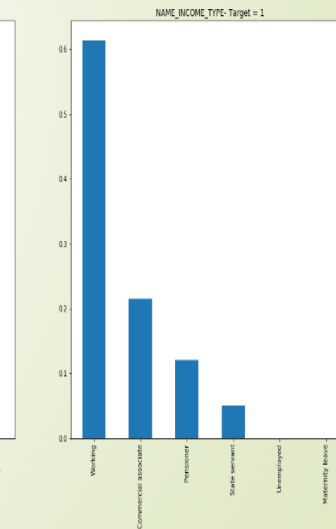
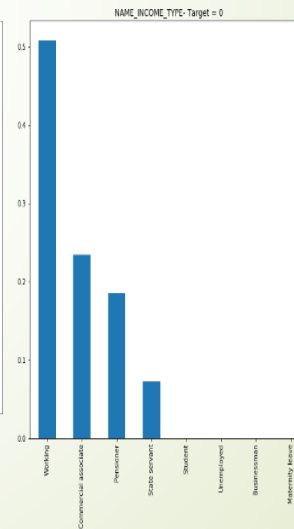
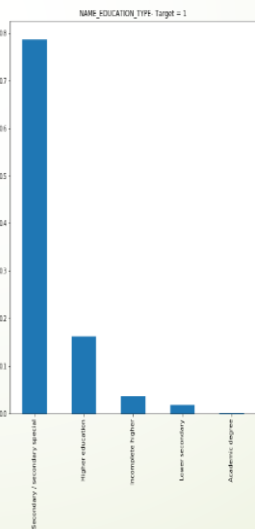
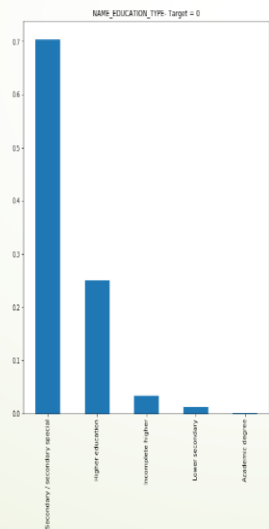
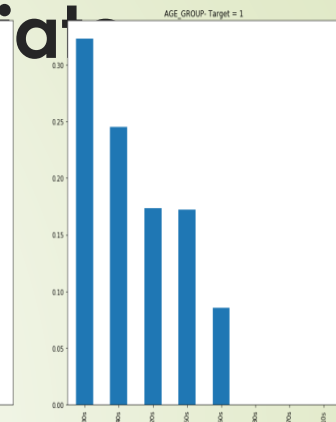
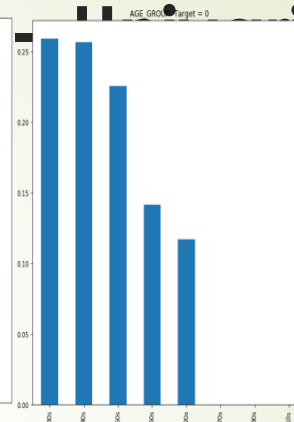
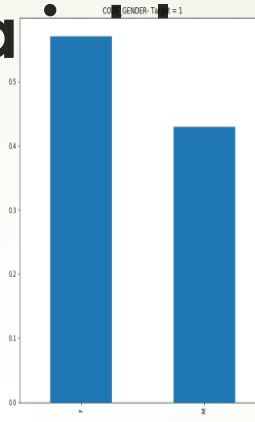
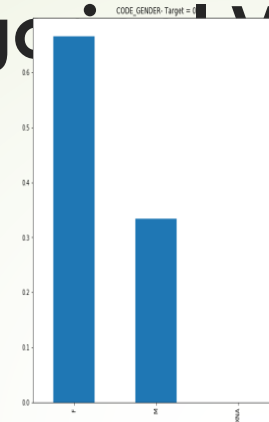
- First, I have calculated null values “nulls(application_data)”
- Then calculated the values in term of percentage.
- Found that there were above 41 columns which consists more than 50% Null values.
- By comparing the columns contribution with given csv file (Columns_Description.csv), I have removed the irrelevant columns.
- Similarly, After removing the 50% data there were 10 Columns which are more than 15% Null Values.

Exploring Categorical Variables Analysis

we examine how defaulters (**Target = 1**) and non-defaulters (**Target = 0**) are distributed across different categorical variables. This helps us identify trends and characteristics associated with loan repayment behavior.

Key Insights:

- **Gender (Fig 1):** A larger share of defaulters are male.
 - **Age Group (Fig 2):** Most defaulters fall within their **30s**, indicating a higher risk in this age group.
 - **Education Level (Fig 3):** A significant number of defaulters have a **secondary education**, suggesting education level may play a role in financial stability.
 - **Income Type (Fig 4):** The majority of defaulters come from the **working-class segment**, which could indicate a correlation between job type and repayment risk.
- Understanding these patterns can help in making better lending decisions and minimizing financial risk.



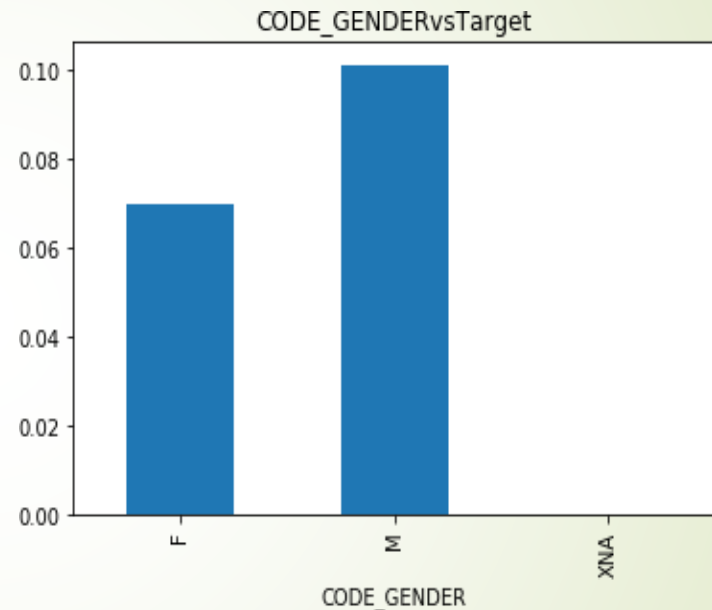
Exploring Categorical Variables –

Bivariate Analysis

In this step, we examine the relationship between loan default and gender to identify potential risk patterns.

Key Insight:

- **Gender & Default Probability:**
 - **Males** have a higher likelihood of defaulting (~10%) compared to **females** (~7%).
 - This suggests that gender may be a contributing factor in assessing credit risk.
 - Understanding such trends can help in refining risk assessment strategies and making more informed lending decisions.

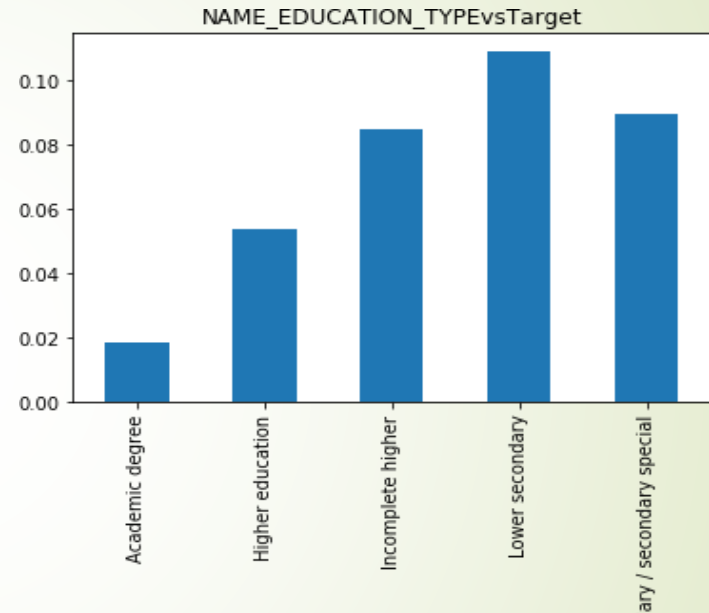


Exploring Categorical Variables – Bivariate Analysis

In this step, we analyze the relationship between **loan default and education type** to identify potential risk factors.

Key Insight:

- **Education Level & Default Probability:**
 - Borrowers with a **lower secondary education** have the **highest default rate (~11%)**, compared to other education levels.
 - This suggests that individuals with lower educational qualifications may face greater financial instability, impacting their ability to repay loans.
 - These insights can help in refining risk assessment models and tailoring lending strategies accordingly.

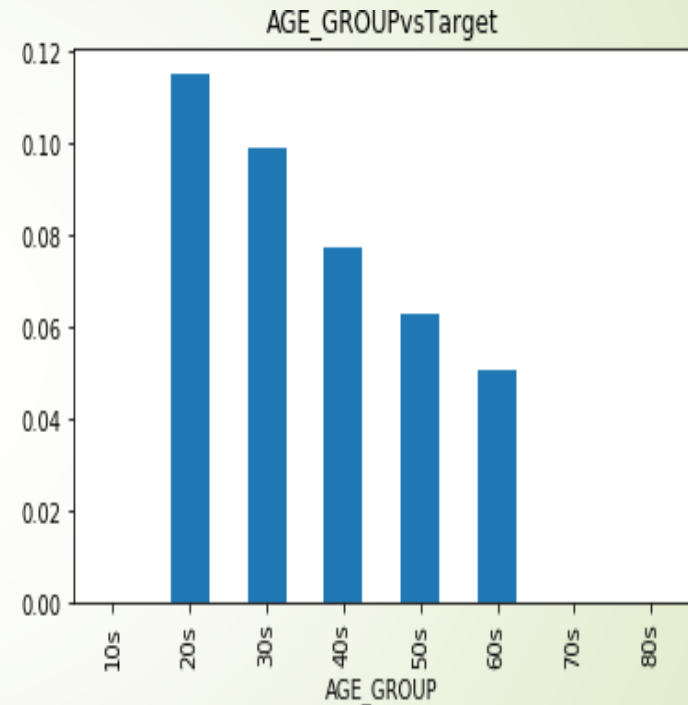


Exploring Categorical Variables – Bivariate Analysis

In this step, we investigate the relationship between **loan default** and **age group** to better understand risk patterns.

Key Insight:

- **Age Group & Default Probability:**
 - Customers in their **20s and 30s** have a **higher risk of defaulting (>10%)** compared to other age groups.
 - This indicates that younger customers may face greater financial challenges, making them more likely to default on loans.
 - These findings can help tailor risk assessment strategies for different age demographics, ensuring more accurate lending decisions.



Exploring Categorical Variables – Bivariate Analysis

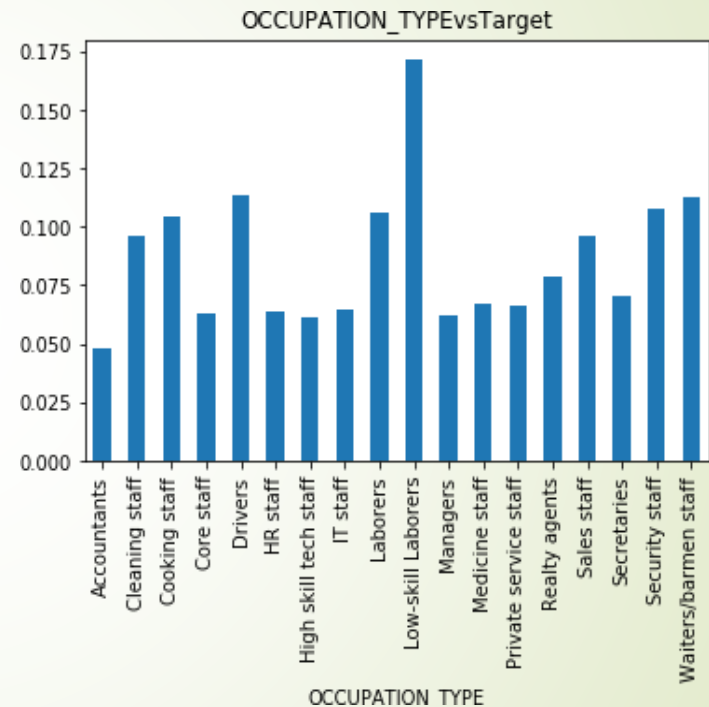
In this step, we examine the relationship between **loan default** and **occupation type** to identify any relevant risk factors.

Key Insight:

•Occupation Type & Default Probability:

- Customers working as **low-skill laborers** have a **higher risk of defaulting (~17%)** compared to other occupation types.
- This suggests that individuals in lower-skilled jobs may face more financial instability, which increases their likelihood of defaulting on loans.

These insights highlight the importance of occupation type in evaluating loan risk, helping lenders make more informed decisions.

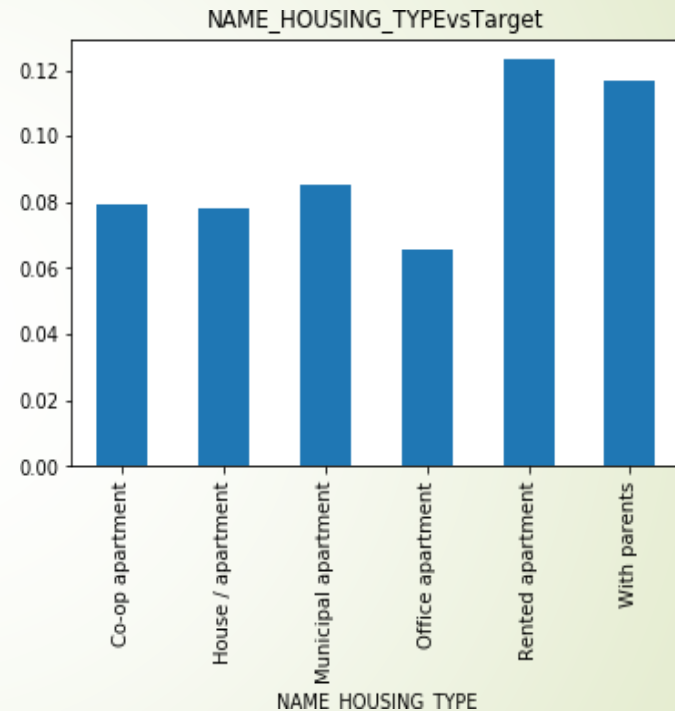


Exploring Categorical Variables – Bivariate Analysis

In this step, we analyze the relationship between **loan default** and **housing type** to understand how living arrangements may influence the likelihood of default.

Key Insight:

- **Housing Type & Default Probability:**
 - Customers who live **with parents** or in a **rented apartment** have a **higher risk of defaulting (>10%)**.
 - This suggests that housing stability may be linked to financial stability, with these living arrangements possibly indicating greater financial strain or less secure financial situations.
 - These findings can be used to better assess financial risk based on customers' living situations.

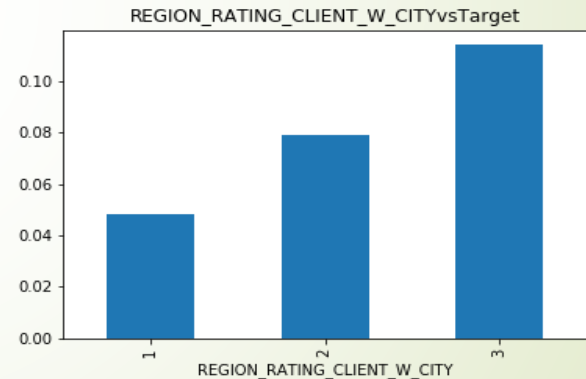
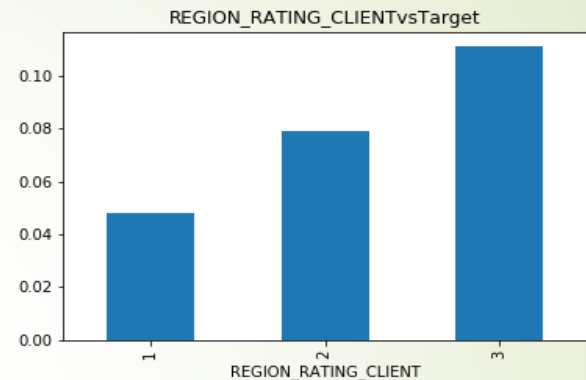


Exploring Categorical Variables – Bivariate Analysis

In this step, we explore the relationship between **loan default** and **region rating** to understand how the geographical area might impact loan repayment behavior.

Key Insight:

- **Region Rating & Default Probability:**
 - Customers residing in **Region Rating 3** have a **higher risk of defaulting (~10%)** compared to other regions.
 - This suggests that customers in regions with lower ratings may face more financial challenges, which increases their likelihood of defaulting on loans.
- These insights emphasize the importance of considering regional factors in risk assessment for better lending decisions.



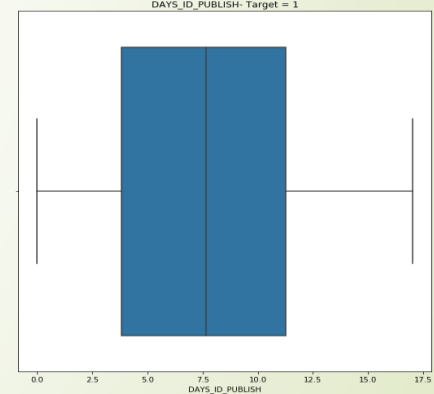
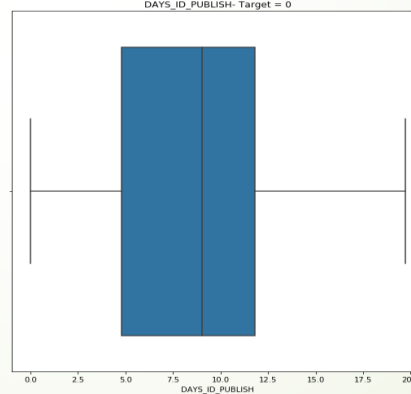
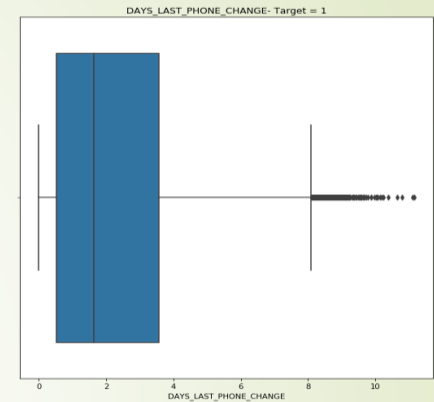
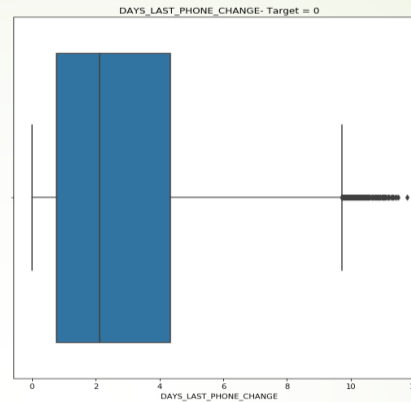
Exploring Numerical Variables – Univariate Analysis

In this step, we examine the distribution of **defaulters (Target = 1)** and **non-defaulters (Target = 0)** across various numerical variables. This helps to uncover patterns that may indicate default risk.

Key Insights:

- **DAYS_LAST_PHONE_CHANGE (Fig 1):**
 - The **median** and **75th percentile** values for defaulters are lower than those of non-defaulters.
 - This suggests that defaulters are more likely to change their phone numbers before applying for a loan, possibly indicating instability or attempts to avoid contact.
- **DAYS_ID_PUBLISH (Fig 2):**
 - Defaulters tend to change their IDs more frequently than non-defaulters.
 - This could indicate potential issues with identity consistency or a higher level of financial instability among defaulters.

These patterns provide valuable insights into behaviors that could signal higher default risk.

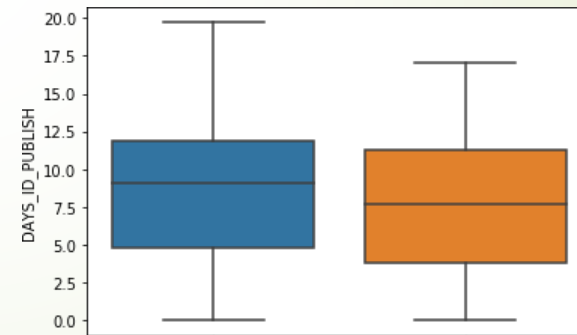
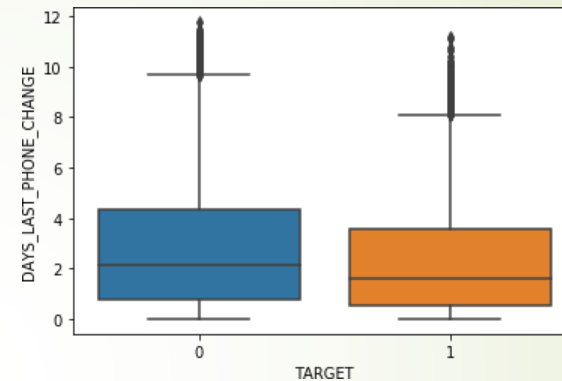


Exploring Numerical Variables – Bivariate Analysis

In this step, we analyze the relationship between **loan default (Target = 1)** and **non-default (Target = 0)** categories, and how they are related to various numerical variables.

Key Insights:

- **DAYS_LAST_PHONE_CHANGE (Fig 1):**
 - Defaulter customers tend to change their phone numbers **closer to the submission of the loan application**.
 - This could indicate instability or an attempt to hide their identity, which might signal higher risk.
- **DAYS_ID_PUBLISH (Fig 2):**
 - Defaulter customers also tend to change their IDs **closer to the submission of the application**.
 - This behavior may be a sign of financial instability or potential fraudulent activity, increasing the likelihood of default.
- These patterns suggest that changes in personal information closer to the application date might be a red flag in assessing loan repayment risk.



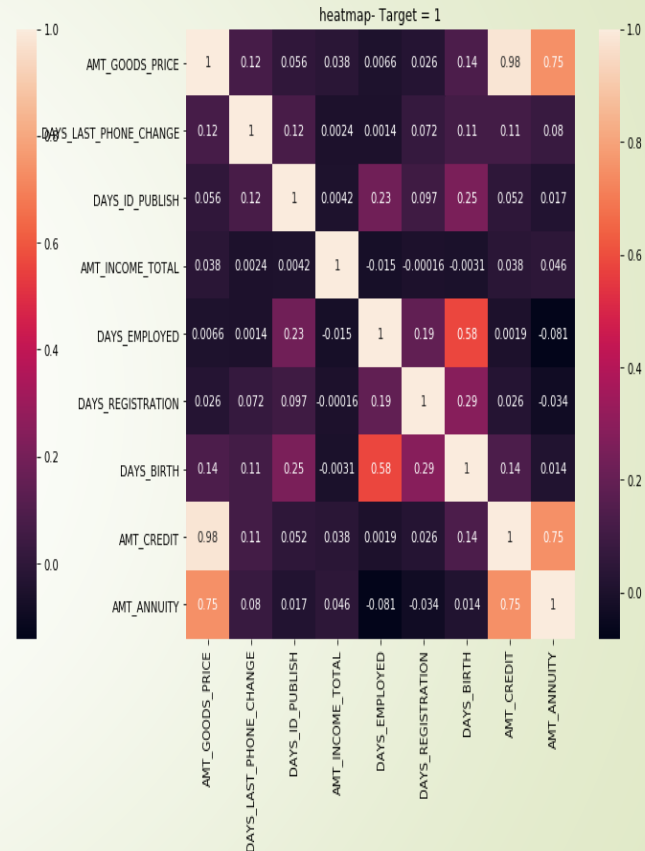
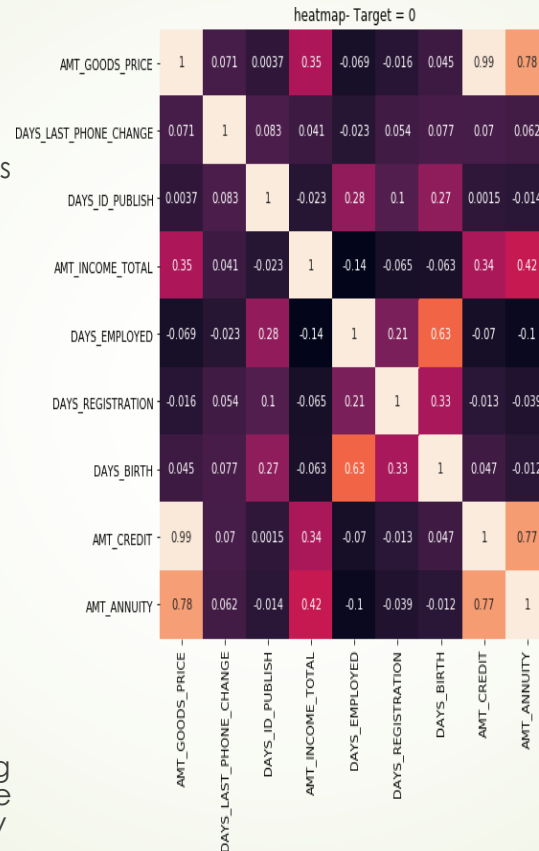
Exploring Numerical Variables – Correlation Analysis

In this step, we analyze the correlation between variables for **Target = 0** (non-defaulters) and **Target = 1** (defaulters) to identify the top 10 most correlated variable pairs in each group. This helps us understand if the relationships between variables differ based on the loan default status.

Key Insights:

- **Correlation for Target = 0 (Non-Defaulters) (Fig 1):**
 - This shows the top correlated variables for non-defaulters.
- **Correlation for Target = 1 (Defaulters) (Fig 2):**
 - This shows the top correlated variables for defaulters.

Although the correlation values for both datasets appear to be similar, presenting the values in a **tabular format** will provide a clearer comparison, helping to identify any differences in the relationships between variables based on the default status.



Exploring Numerical Variables – Correlation Analysis

In this step, we examine the correlation between variables for **Target = 0** (non-defaulters) and **Target = 1** (defaulters) to check if the top 10 correlated variable pairs are consistent across both groups.

Key Insights:

- **Top 10 Correlated Variables for Target = 0 (Non-Defaulters) (Fig 1):**
 - Displays the most correlated variables for non-defaulters.
- **Top 10 Correlated Variables for Target = 1 (Defaulters) (Fig 2):**
 - Displays the most correlated variables for defaulters.

Observation:

- **8 out of the top 10 correlated variables are common** across both the non-defaulters and defaulters datasets, indicating that some relationships between variables remain consistent regardless of default status.
- These findings provide a strong basis for further analysis, showing that the same key variables influence both groups.

	VAR1	VAR2	Correlation	Correlation_abs
414	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
337	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96	0.96
277	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89	0.89
440	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
129	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
207	DAYS_EMPLOYED	DAYS_BIRTH	0.58	0.58
415	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34	0.34
389	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33	0.33

	VAR1	VAR2	Correlation	Correlation_abs
414	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
337	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96	0.96
277	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89	0.89
440	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
129	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
207	DAYS_EMPLOYED	DAYS_BIRTH	0.58	0.58
415	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34	0.34
389	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33	0.33

Overview

Based on the analysis conducted, the following insights can aid in predicting customers who are likely to default:

S.No.	Variable	Variable Type
1	CODE_GENDER	Categorical
2.	NAME_EDUCATION_TYPE	Categorical
3.	AGE_GROUP	Categorical
4.	NAME_HOUSING_TYPE	Categorical
5.	NAME_INCOME_TYPE	Categorical
6.	OCCUPATION_TYPE	Categorical
7.	REGION_RATING_CLIENT	Categorical
8.	REGION_RATING_CLIENT_W_CITY	Categorical
9.	DAYS_LAST_PHONE_CHANGE	Numerical
10.	DAYS_ID_PUBLISH	Numerical



Thank you

Submitted By: Kaomudie Mukhopadhyay
Gmail id:Kaomudiemukhopadhyay702@gmail.com