Bank Loan Case Study

Final Project-2

Project Description:

A data analyst at a finance company specializes in lending various types of loans to urban customers. The company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval. Being a data Analyst at a finance company our job is to use Exploratory Data Analysis (EDA) to analyse patterns in the data and ensure that capable applicants are not rejected.

When a customer applies for a loan, the company faces two risks:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

The dataset provided contains information about loan applications. It includes two types of scenarios:

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y instalments of the loan.
2. All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

Approach:

Exploratory Data Analysis (EDA) is crucial for understanding the data, uncovering patterns, and identifying anomalies or relationships that can inform further analysis. Here is a step-by-step approach for performing EDA on a bank loan case study:

1. Data Collection and Preparation:

First off, all the dataset provided was downloaded in excel format. The dataset "application data" provides details about the current loan applications like the type of contract, annuity amount, credit amount etc.

The Dataset details are:
- Number of Data-Points: 49,999
- Number of Features: 122

2. Understand the Data

- **Objective:** Grasp the business context and the purpose of the analysis.
- **Columns:** AMT_INCOME_TOTAL, AMT_CREDIT, NAME_EDUCATION_TYPE, etc.
- **Goal:** Determine the characteristics of applicants who default on loans versus those who do not.

3. Data Cleaning

- **Missing Values:** Identify missing values using
- **Duplicates:** Check for and handle duplicate rows .
- **Outliers:** Detect and manage outliers using boxplots and statistical methods.
- **Data Types:** Ensure each column has the correct data type.

4. Univariate Analysis

- **Descriptive Statistics:** Get a summary of numerical columns.
- **Distribution Plots:** Plot histograms, box plots for numerical features.
- **Categorical Plots:** Plot bar charts and count plots for categorical features like NAME_EDUCATION_TYPE.

5. Bivariate Analysis

- **Correlation:** Use `Corr()` and heatmaps to find relationships between numerical features.
- **Scatter Plots:** Explore relationships between pairs of numerical features.
- **Box Plots:** Compare distributions of numerical features across different categories.
- **Crosstabs and Chi-Square Tests:** Analyse relationships between categorical features.

6. Multivariate Analysis

- **Group by Analysis:** Group data by categorical features and analyze aggregated statistics.
- **Heatmaps:** Visualize correlations and interactions between multiple features.

7. Feature Engineering

- **Create New Features:** Based on domain knowledge, create new features that may be useful.
- **Transformation:** Apply log transformations or scaling to handle skewed distributions.
- We created new columns which showed duration in years.

Tech Stack Used:

1. **Microsoft Excel 2021** — A spreadsheet editor software used mainly by professionals to enter data in table format, perform computations, plot graphs etc. Here Microsoft Excel is used to filter data and plot graphs to get insights about the movies.
2. **Microsoft Word:** A word processing application for preparing report.

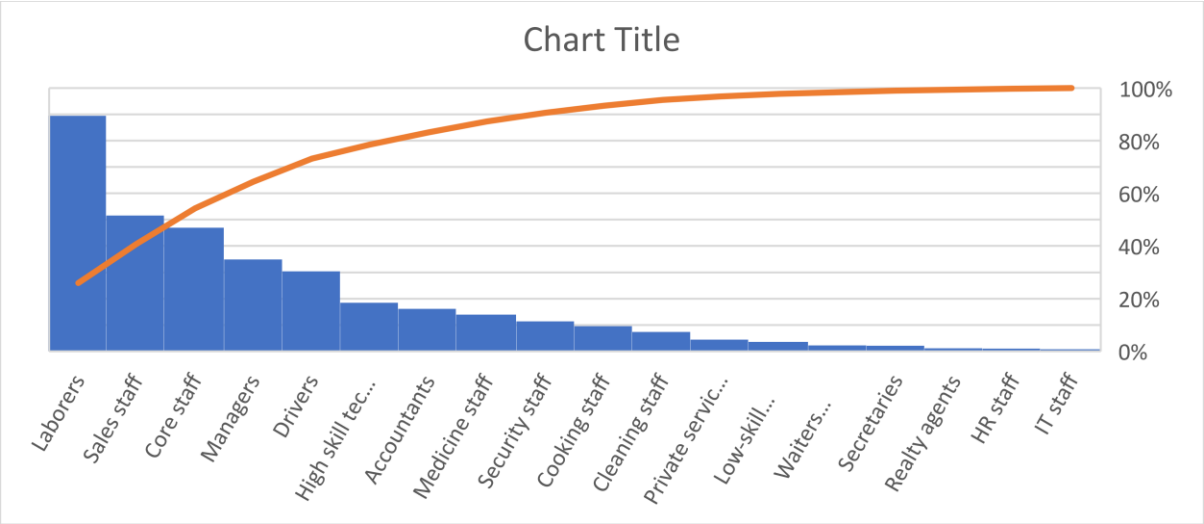
DATA ANALYTICS TASKS:

A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.
Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
INSIGHTS:

- Dropped all the columns where number of null values is greater than 40% and also all the unimportant columns.

No. of columns with null values >40%	49	Dropped
No. of columns with null values >40%	73	Remained

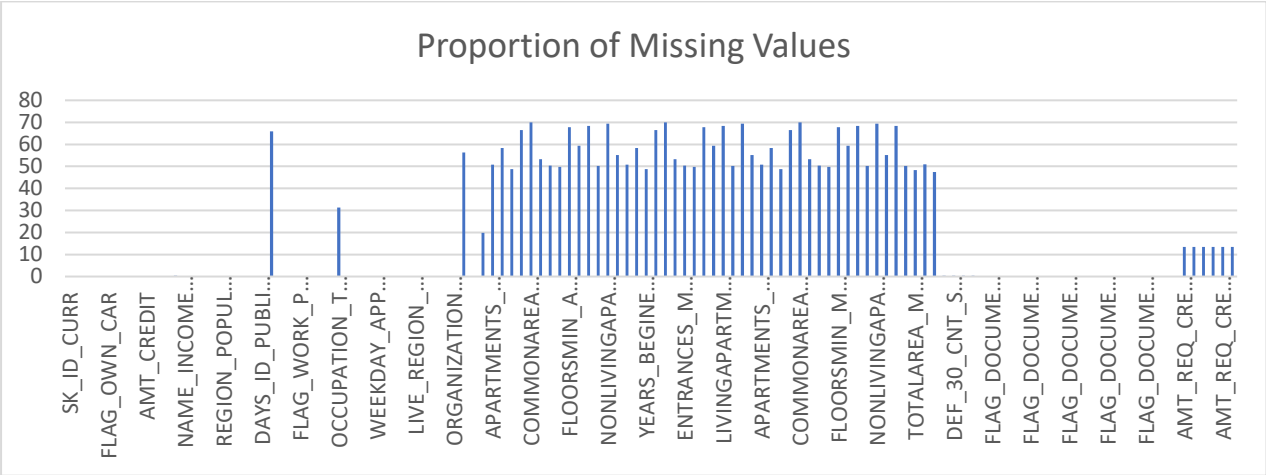
- For null values in **OCCUPATION_TYPE** column, we replaced them with values which have the maximum count of the value of **NAME_INCOME_TYPE** and **NAME_EDUCATION_TYPE** of the corresponding null value. For the null values in **OCCUPATION_TYPE** column which are still present after above step, we replaced them with **Laborers** which is the most common Occupation type**. (Refer Excel sheet Task A.1 and Task A.1.1)**. Took help of pivot to calculate the occurrences of each occupation according to income type and education type **(Sheet 1).**



- For null values in **NAME_TYPE_SUITE** column, we replaced them with values which have the maximum count of the value of **NAME_INCOME_TYPE** and **NAME_FAMILY_STATUS** of the corresponding null value. For all the column values of **NAME_INCOME_TYPE** and **NAME_FAMILY_STATUS,** the most common column value of **NAME_TYPE_SUITE** is **Unaccompanied**. So replaced all null values of **NAME_TYPE_SUITE** with **Unaccompanied. (Refer Excel Task A.1.2)**

- For null values in **AMT_GOODS_PRICE** column, we replaced them with the corresponding row value of **AMT_CREDIT** column. **(Refer Excel Task A.1.3)**
- For null values in **AMT_ANNUITY** column, we replaced them with the median value of **AMT_ANNUITY** for all rows with corresponding value of **AMT_CREDIT** and **AMT_INCOME_TOTAL. (Refer Excel Task A.1.3)**
- For null values in **CNT_FAM_MEMBERS** column, we replaced them with the median value of **CNT_FAM_MEMBERS.**
- We found some error values in **GENDER** column. Replaced them with **F** (Female) which is the most common gender.
- We found some error values in **ORGANIZATION_TYPE** column. Replaced them with **No Work** as these people were pensioners and unemployed.

RESULT:



B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
INSIGHTS:

- For Outliers in **CNT_CHILDREN** and **CNT_FAM_MEMBERS** column, we checked for number of Parents for all row values. Found no issues.
- For Outliers in **DAYS_BIRTH** column, we checked the maximum and minimum age. No issue found.
- For Outliers in **DAYS_EMPLOYED** column, we plotted a box plot, and found a positive value as an outlier. We checked for distribution of **AGE** for all rows which have positive value of **DAYS_EMPLOYED**. We found that most of the **AGE** values were above 50. So we replaced the positive value of **DAYS_EMPLOYED** with median value of **DAYS_EMPLOYED** for rows which have **AGE** greater than equal to 50.
- For Outliers in **DAYS_REGISTRATION** column, we plotted a box plot, and considered value less than **-18000** as outliers. Calculated Age in years, Registration in Years and Difference between Age and Registration for all those rows. Found no issues.
- For Outliers in **DAYS_ID_PUBLISH** column, we checked the maximum and minimum age and difference between age and **DAYS_ID_PUBLISH** in years. No issue found.
- For Outliers in **AMT_INCOME_TOTAL** column, we plotted a box plot, and considered value greater than 5000000 as outlier. We compared various row values of the particular row with median values of the columns and found large differences. So replaced the above outlier value with median value.

**box plots or scatter plots to visualize the distribution of numerical variables for all the observations made above are provided in excel sheet Task B OUTLIERS.

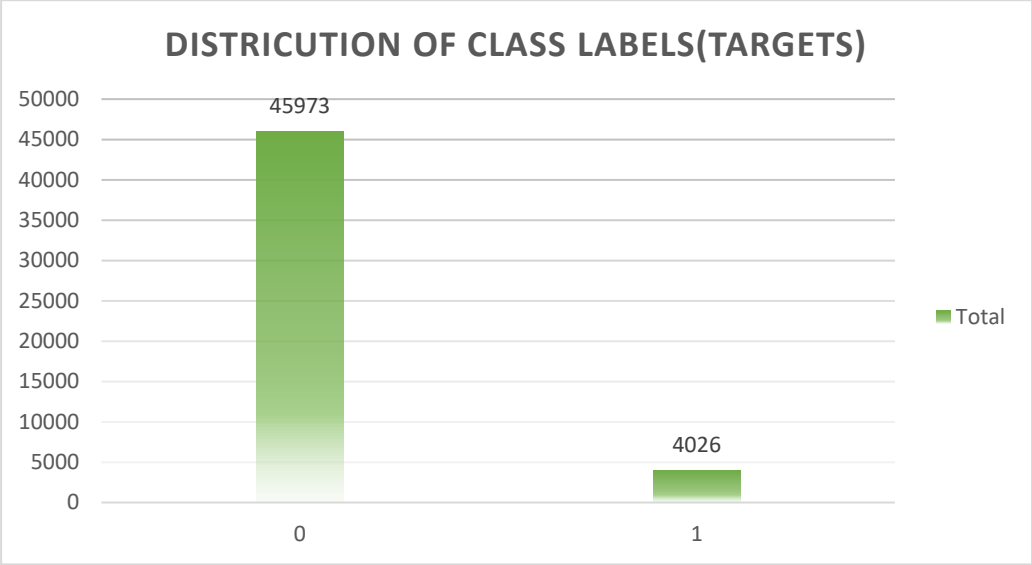
C. Analyse Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

INSIGHTS:

- The Dataset is **highly imbalanced**, skewed more towards **Class Label 0** (No Payment Difficulty). From above Bar Chart, we can see that around **92%** of Applicants didn't had any difficulty in paying loan instalments and around **8%** (**Class Label 1**) of Applicants had difficulty in paying loan instalments.
- This data imbalance may give wrong predictions during modelling. So it needs to be taken care of by upscaling Data of **Class Label 1** or downscaling Data of **Class Label 0**.

	Count of 0's and 1's	Ratio	Contribution
0	45973	11.41903	91.95
1	4026		8.05



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

INSIGHTS:

Univariate Analysis:

- Most of the loan applications are for **Cash Loans** and very less are for **Revolving Loans**. This is True in reality as well.
- Most of the loan applicants are **Females** and less are **Males**.

- Most of the applicants were not accompanied by anyone else followed by applicants who were accompanies by family members.
- Most of the applicants had **Working** Income Type followed by **Commercial Associate** Income Type.
- Most of the applicants had education up to **Secondary** followed by **Higher education.**
- Most of the applicants were normally **Married** followed by **Single or Not Married.**
- Most of the applicants had **own House** followed by applicant who were living **with parents**.
- Most of the applicants were **Laborers** followed by **Waiters/barmen staff**.
- Most of the applicants worked in **Business Entity Type 3** followed by applicants who **didn't work** at the time. Most of these applicants were **Pensioners** and few were **Unemployed**.
- Most of the applicants **didn't own car.**
- Most applicants' income was less than **Unit 400000** but a lot of their Credit Amount is more than **Unit 400000** showing that they have applied for loans of amount greater than their Income.
- Most applicants' Annuity amount is less than **Unit 50000** which is close to **10%** of Income of most applicants.
- The distribution of **AMT_GOODS_PRICE** closely follows the distribution of **AMT_CREDIT**.
- We can observe that **AGE** column somewhat follows a normal distribution and most applicants are between age **27** and **65** i.e. most were in the working age group.
- Also, most applicants had less than **8** years of work experience.

**Graph/Charts for all the above observations is provided in excel sheet Task D UNIVARIATE ANALYSIS.

Segmented Univariate Analysis:

- For columns **AMT_INCOME_TOTAL**, we can observe that for both **0** and **1** of **TARGET** most of the applicants has income between 125000-150000

**Graph/Charts for the above observations is provided in excel sheet Task D SEGMENTED UNIVARIATE.

Bivariate Analysis:

- We can observe that most clients applied for Loans during **Weekdays** and between **9 A.M** and **4 P.M**. But there are also very few clients who applied for Loans late at night as well.
- For columns with two unique values, we can observe that for both **0** and **1** of **TARGET:**
  — Most Contract types are **Cash loans.**
  — Most applicants are **Female.**
- For categorical columns, we can observe that for both **0** and **1** of **TARGET:**
  — Most applicants came **unaccompanied** followed by **family**.
  — Most applicants income type was **Working** followed by **Commercial associate**.
  — Most applicants education was up to **Secondary** followed by **Higher education**.
  — Most applicants were **Married** followed by **Single**.
  — Most applicants were living in their own **House/Apartment** followed by applicants living **With Parents.**
- For categorical columns, we can observe that:
  — Most applicants were **Laborers** followed by **Core Staff** for **TARGET** 0. Most applicants were **Laborers** followed by **Sales staff** for **TARGET 1**.
  — Most applicants were working in **Business Entity Type 3** followed by applicants who had **No Work** (Pensioners and Unemployed) for **TARGET 0** whereas most applicants were

working in **Business Entity Type 3** followed by applicants who were **Self-employed** for **TARGET 1**.

**Pivot Tables/Heatmaps for all the above observations is provided in excel sheet Task D BIVARIATE ANALYSIS.

E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

INSIGHTS: For this task we filtered the variables based on the target (0 or 1).We have considered 10 variables here for our observation

Considering **0.5 (absolute value)** as threshold for high Correlation, we can observe that:

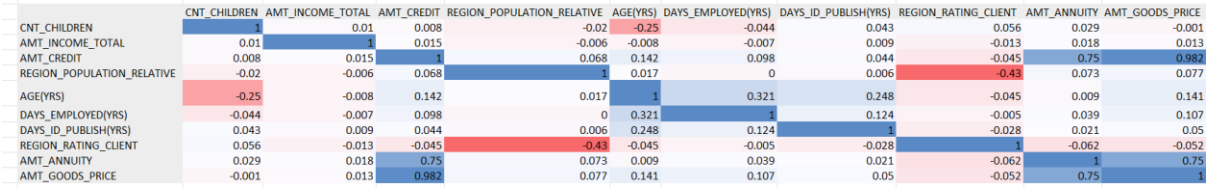
- **AMT_CREDIT** and **AMT_GOODS_PRICE** are **highly and positively correlated** as the Credit amount request is for the Goods whose price is in **AMT_GOODS_PRICE** column.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	AGE(YRS)	DAYS_EMPLOYED(YRS)	DAYS_ID_PUBLISH(YRS)	REGION_RATING_CLIENT	AMT_ANNUITY	AMT_GOODS_PRICE
CNT_CHILDREN	1	0.036319722	0.005705458	-0.024912809	-0.335689	-0.067895528	0.032534853	0.021288992	0.026404227	0.001367977
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.181941261	-0.073642	0.030601987	-0.032065618	-0.205031899	0.451114602	0.384454763
AMT_CREDIT	0.005705458	0.377965752	1	0.095539444	0.051244	0.084860082	0.007964812	-0.102556478	0.770758443	0.986810608
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539444	1	0.030385	-0.005581805	0.002304099	-0.539333113	0.117260594	0.09905037
AGE(YRS)	-0.335688904	-0.073641965	0.051244108	0.030385201	1	0.278495911	0.27025402	-0.009115905	-0.009660692	0.049369802
DAYS_EMPLOYED(YRS)	-0.067895528	0.030601987	0.084860082	-0.005581805	0.278496	1	0.079995141	0.015124574	0.05426097	0.085865736
DAYS_ID_PUBLISH(YRS)	0.032534853	-0.032065618	0.007964812	0.002304099	0.270254	0.079995141	1	0.007512774	-0.009660942	0.00950579
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.102556478	-0.539333113	-0.009116	0.015124574	0.007512774	1	-0.129913111	-0.104474609
AMT_ANNUITY	0.026404227	0.451114602	0.770758443	0.117260594	-0.009661	0.05426097	-0.009660942	-0.129913111	1	0.775831652
AMT_GOODS_PRICE	0.001367977	0.384454763	0.986810608	0.09905037	0.04937	0.085865736	0.00950579	-0.104474609	0.775831652	1

CORRELATION OF VARIABLES FOR APPLICATIONS WHO DO NOT HAVE PAYMENT DIFFICULTY (Target 0)

Considering **0.5 (absolute value)** as threshold for high Correlation, we can observe that:

- **AMT_CREDIT** and **AMT_GOODS_PRICE** are **highly and positively correlated** as the Credit amount request is for the Goods whose price is in **AMT_GOODS_PRICE** column.



	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	AGE(YRS)	DAYS_EMPLOYED(YRS)	DAYS_ID_PUBLISH(YRS)	REGION_RATING_CLIENT	AMT_ANNUITY	AMT_GOODS_PRICE
CNT_CHILDREN	1	0.01	0.008	-0.02	-0.25	-0.044	0.043	0.056	0.029	-0.001
AMT_INCOME_TOTAL	0.01	1	0.015	-0.006	-0.008	-0.007	0.009	-0.013	0.018	0.013
AMT_CREDIT	0.008	0.015	1	0.068	0.142	0.098	0.044	-0.045	0.75	0.982
REGION_POPULATION_RELATIVE	-0.02	-0.006	0.068	1	0.017	0	0.006	-0.43	0.073	0.077
AGE(YRS)	-0.25	-0.008	0.142	0.017	1	0.321	0.248	-0.045	0.009	0.141
DAYS_EMPLOYED(YRS)	-0.044	-0.007	0.098	0	0.321	1	0.124	-0.005	0.039	0.107
DAYS_ID_PUBLISH(YRS)	0.043	0.009	0.044	0.006	0.248	0.124	1	-0.028	0.021	0.05
REGION_RATING_CLIENT	0.056	-0.013	-0.045	-0.43	-0.045	-0.005	-0.028	1	-0.062	-0.052
AMT_ANNUITY	0.029	0.018	0.75	0.073	0.009	0.039	0.021	-0.062	1	0.75
AMT_GOODS_PRICE	-0.001	0.013	0.982	0.077	0.141	0.107	0.05	-0.052	0.75	1

CORRELATION OF VARIABLES FOR APPLICATIONS WHO HAD PAYMENT DIFFICULTY (Target 1)

- **Conclusion:**

  Through this project, I was able to understand the importance of **Data Analytics** in **Bank Loan Analysis** as it provides valuable insights which helps in making **Data-Driven Decisions**.
  In this project I was able to get insights like which features are important to predict loan defaulters, correlation between various features like income amount, loan amount, personal belongings details etc. which can be communicated to relevant stakeholders as per the requirements.

- **EXCEL FILE LINK:**
  https://docs.google.com/spreadsheets/d/1O6H9xwdpTKcTAENhR11n3-nznu-QrWBC/edit?usp=drive_link&ouid=104742351045324653369&rtpof=true&sd=true