In this project, you are a data analyst at a financial company specializing in lending loans. Your company faces challenges with some customers defaulting on their loans. The aim is to use Exploratory Data Analysis (EDA) to identify patterns in the data to make informed loan approval decisions.

Understanding the Files

a. previous_application.csv: Contains information about previous loan applications.

b. application_data.csv: Provides details about the current loan applications.

c. columns_description.csv: Describes the columns present in the other datasets, explaining what each column represents.

Data Description:

previous_application.csv

Identification

SK_ID_CURR: ID of loan in our sample

Contract Information

NAME_CONTRACT_TYPE: Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application

AMT_ANNUITY: Annuity of previous application

AMT_APPLICATION: For how much credit did client ask on the previous application

AMT_CREDIT: Final credit amount on the previous application. This differs from AMT_APPLICATION in a way that the AMT_APPLICATION is the amount for which the client initially applied for, but during our approval process he could have received different amount - AMT_CREDIT

AMT_GOODS_PRICE: Goods price of good that client asked for (if applicable) on the previous application

AMT_DOWN_PAYMENT: Down payment on the previous application

Application Details

WEEKDAY_APPR_PROCESS_START: On which day of the week did the client apply for previous application

HOUR_APPR_PROCESS_START: Approximately at what day hour did the client apply for the previous application

FLAG_LAST_APPL_PER_CONTRACT: Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract

NFLAG_LAST_APPL_IN_DAY: Flag if the application was the last application per day of the client. Sometimes clients apply for more applications a day. Rarely it could also be error in our system that one application is in the database twice

Interest Rates

RATE_DOWN_PAYMENT: Down payment rate normalized on previous credit

RATE_INTEREST_PRIMARY: Interest rate normalized on previous credit

RATE_INTEREST_PRIVILEGED: Interest rate normalized on previous credit

Purpose and Status

NAME_CASH_LOAN_PURPOSE: Purpose of the cash loan

NAME_CONTRACT_STATUS: Contract status (approved, cancelled, ...) of previous application

DAYS_DECISION: Relative to current application when was the decision about previous application made

NAME_PAYMENT_TYPE: Payment method that client chose to pay for the previous application

CODE_REJECT_REASON: Why was the previous application rejected

Goods and Client Details

NAME_TYPE_SUITE: Who accompanied client when applying for the previous application

NAME_CLIENT_TYPE: Was the client old or new client when applying for the previous application

NAME_GOODS_CATEGORY: What kind of goods did the client apply for in the previous application

Portfolio and Product

NAME_PORTFOLIO: Was the previous application for CASH, POS, CAR, …

NAME_PRODUCT_TYPE: Was the previous application x-sell o walk-in

CHANNEL_TYPE: Through which channel we acquired the client on the previous application

NAME_SELLER_INDUSTRY: The industry of the seller

CNT_PAYMENT: Term of previous credit at application of the previous application

NAME_YIELD_GROUP: Grouped interest rate into small medium and high of the previous application

PRODUCT_COMBINATION: Detailed product combination of the previous application

Timeline

DAYS_FIRST_DRAWING: Relative to application date of current application when was the first disbursement of the previous application

DAYS_FIRST_DUE: Relative to application date of current application when was the first due supposed to be of the previous application

DAYS_LAST_DUE_1ST_VERSION: Relative to application date of current application when was the first due of the previous application

DAYS_LAST_DUE: Relative to application date of current application when was the last due date of the previous application

DAYS_TERMINATION: Relative to application date of current application when was the expected termination of the previous application

Insurance

NFLAG_INSURED_ON_APPROVAL: Did the client requested insurance during the previous application

application_data.csv

Identification

SK_ID_CURR: ID of loan in our sample

Loan Outcome

TARGET: Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

Client Information

CODE_GENDER: Gender of the client

FLAG_OWN_CAR: Flag if the client owns a car

FLAG_OWN_REALTY: Flag if client owns a house or flat

CNT_CHILDREN: Number of children the client has

AMT_INCOME_TOTAL: Income of the client

Contract Information

NAME_CONTRACT_TYPE: Identification if loan is cash or revolving

AMT_CREDIT: Credit amount of the loan

AMT_ANNUITY: Loan annuity

AMT_GOODS_PRICE: For consumer loans it is the price of the goods for which the loan is given

Property Details

NAME_TYPE_SUITE: Who was accompanying client when he was applying for the loan

NAME_INCOME_TYPE: Clients income type (businessman, working, maternity leave,…)

NAME_EDUCATION_TYPE: Level of highest education the client achieved

NAME_FAMILY_STATUS: Family status of the client

NAME_HOUSING_TYPE: What is the housing situation of the client (renting, living with parents, ...)

Client's Occupation

OCCUPATION_TYPE: What kind of occupation does the client have

ORGANIZATION_TYPE: Type of organization where client works

Application Details

WEEKDAY_APPR_PROCESS_START: On which day of the week did the client apply for the loan

HOUR_APPR_PROCESS_START: Approximately at what hour did the client apply for the loan

Client's Region

REGION_POPULATION_RELATIVE: Normalized population of region where client lives (higher number means the client lives in more populated region)

REGION_RATING_CLIENT: Our rating of the region where client lives (1,2,3)

REGION_RATING_CLIENT_W_CITY: Our rating of the region where client lives with taking city into account (1,2,3)

Client's Age

DAYS_BIRTH: Client's age in days at the time of application

Client's Employment

DAYS_EMPLOYED: How many days before the application the person started current employment

External Sources

EXT_SOURCE_1: Normalized score from external data source

EXT_SOURCE_2: Normalized score from external data source

EXT_SOURCE_3: Normalized score from external data source

Hints for using the Files:

To ensure that you make the most of the datasets provided, it's crucial to know which file to reference for each task. Here's a breakdown:

A. Handle Missing Data

- **Primary File**: **application_data.csv**
    - This file will likely contain the majority of the data required for your analysis. Start by identifying and handling missing values in this dataset.
- **Reference File**: **columns_description.csv**
    - If you're unsure about the significance of a column with missing values, refer to this file for a detailed description.

B. Identify Outliers

- **Primary File**: **application_data.csv**
    - Focus on the numerical columns in this dataset to detect outliers. This file contains most of the applicant and loan details that will be crucial for this task.

C. Analyze Data Imbalance

- **Primary File**: **application_data.csv**
    - The target variable indicating payment difficulties will be in this file. Use it to analyze any data imbalance in loan repayment scenarios.

D. Various Analyses

- **Primary File**: **application_data.csv**
    - This file provides a comprehensive view of both the applicant and the loan, making it ideal for univariate, segmented univariate, and bivariate analyses.
- **Secondary File**: **previous_application.csv**
    - For a deeper dive or more nuanced insights, especially when considering previous loan applications, this file can be invaluable.

E. Identify Correlations

- **Primary File**: **application_data.csv**
    - Given that correlations will largely focus on the relationship between various attributes and the likelihood of default, this file will be the primary source of data.

When working through the tasks, always ensure that you understand the context and meaning of each column.

Tips and Tricks

- Always back up your original data before making any changes.
- While handling missing data, consider the business context. Sometimes, deleting the record might be more appropriate than imputation.
- When identifying outliers, always cross-check with business knowledge. Some outliers might be genuine data points.
- While analyzing correlations, remember that correlation does not imply causation. Use the insights to generate hypotheses and not conclusions.

Points to Remember:

1. Understanding of the Financial Sector:

- **Credit Systems**: Familiarity with how credit systems operate, including credit scores, credit histories, and their significance in loan approval processes.
- **Loan Types**: Knowledge of different types of loans such as cash loans, revolving loans, and their characteristics.
- **Risk Management**: An understanding of how financial institutions manage risks associated with lending.

2. Loan Repayment and Default:

- **Repayment Schedules**: Awareness of how repayment schedules work, including terms like annuities, due dates, and down payments.
- **Default**: Understand what constitutes a loan default, its implications for both the borrower and the lender, and the factors that typically lead to a default.
- **Late Payments**: The significance of late payments, how they impact credit scores, and their relevance in predicting potential defaults.

3. Customer Segmentation in Banking:

- **Demographics**: Recognizing the importance of demographic factors like age, gender, income levels, and their influence on loan eligibility and repayment capability.
- **Employment and Income**: Understanding how a person's employment type, organization type, and income levels can impact their creditworthiness.

4. External Data Sources:

- **Credit Bureaus**: Familiarity with how credit bureaus operate and the kind of data they provide which can be valuable in loan decision processes.
- **Significance of External Ratings**: Understanding ratings from external sources and their relevance in predicting loan defaults.

5. Application Process in Banking:

- **Documentation**: Awareness of the typical documents required during a loan application process.
- **Approval Workflow**: Understanding the steps involved in loan approvals, including verification checks, credit checks, and final approval or rejection.

6. Business Implications:

- **False Positives and Negatives**: Recognizing the business implications of falsely approving or rejecting a loan.
- **Customer Relationships**: Understanding the importance of maintaining good relationships with customers, even when rejecting a loan application.
- **Operational Efficiency**: The significance of streamlining the loan approval process for operational efficiency without compromising on risk assessments.

7. Market Trends and Economic Indicators:

- **Economic Indicators**: Familiarity with macroeconomic indicators that can influence credit markets, such as interest rates, inflation rates, and unemployment rates.
- **Market Trends**: Keeping an eye on trends in the credit market, such as changing default rates, which can provide context to the data analysis.

8. Regulatory and Compliance Aspects:

- **Financial Regulations**: Awareness of local and international financial regulations that govern loan approvals and risk assessments.
- **Data Privacy**: Understanding the importance of data privacy regulations when dealing with personal financial data.

Important Hypothesis:

1. Demographic Factors:

- **Gender & Payment Difficulty**: Male clients might exhibit different payment behaviors than female clients.
- **Income & Payment Difficulty**: Clients with lower total incomes might face more challenges in making timely payments.
- **Age & Payment Difficulty**: Younger clients might face more payment difficulties than older clients.

2. Loan Characteristics:

- **Loan Amount & Payment Difficulty**: Higher loan amounts might be associated with increased payment difficulties.
- **Loan Type & Payment Difficulty**: Certain types of loans, like cash loans, might be associated with more payment difficulties than revolving loans.
- **Loan Term & Payment Difficulty**: Short-term loans might see more initial payment difficulties than long-term loans.

3. Historical Data & Behavior:

- **Previous Applications & Payment Difficulty**: Clients with multiple previous loan applications might have different payment behaviors than those with fewer applications.
- **Previous Loan Approval & Payment Difficulty**: Clients who have had their previous loan applications approved might have different payment behaviors than those who were denied.

4. External Factors & Creditworthiness:

- **External Ratings & Payment Difficulty**: Lower scores from external sources might correlate with increased payment difficulties.

5. Employment & Occupation:

- **Employment Duration & Payment Difficulty**: Clients with shorter employment durations might face more payment difficulties.
- **Occupation Type & Payment Difficulty**: Some occupation types might be associated with increased payment difficulties.

6. Application Details:

- **Application Timing & Payment Difficulty**: Applications made during certain times, like weekends or late hours, might be associated with increased payment difficulties.

7. Property & Reality:

- **Property Ownership & Payment Difficulty**: Clients who own real estate or a car might face different payment challenges than those who don't.

8. Family Status:

- **Family Size & Payment Difficulty**: Clients with larger families (more children or dependents) might face more payment difficulties.

9. Regional Factors:

- **Client's Region & Payment Difficulty**: Clients from certain regions or with specific regional ratings might exhibit different payment behaviors.

10. Previous Application Details (from previous_application.csv):

- **Previous Loan Purpose & Current Payment Difficulty**: The purpose of previous loans might influence the likelihood of payment difficulties for the current loan. For instance, if a previous loan was for urgent medical needs, the client might face more financial strain.

11. Contract & Product Details (from previous_application.csv):

- **Previous Loan Contract Type & Current Payment Difficulty**: The type of contract for previous loans (like cash loans, consumer loans) might influence current payment behaviors.
- **Interest Rates & Payment Difficulty**: Clients with higher interest rates on their previous loans might face more payment difficulties.

12. Client Behavior (from previous_application.csv):

- **Previous Loan Cancellations & Current Payment Difficulty**: Clients who have previously canceled loan applications might have different payment behaviors.

These hypotheses, rooted in the project's context, can guide the exploratory data analysis process. They can be validated or refuted using the datasets provided, which will offer insights into the factors influencing payment difficulties.