**DATA SCIENCE – PRACTICAL INTERVIEW DURATION 3 DAYS**

**Description of variables in Train/Test Data**

0. \*\*ID\*\*: A unique identifier for each entry in the dataset.

1. \*\*customer\_id\*\*: Unique identifier for each customer in the dataset.

2. \*\*country\_id\*\*: Identifier or code representing the country where the customer resides or where the loan was issued.

3. \*\*tbl\_loan\_id\*\*: Unique identifier for each loan associated with the customer.

4. \*\*Total\_Amount\*\*: The total loan amount initially disbursed to the customer.

5. \*\*Total\_Amount\_to\_Repay\*\*: The total amount the customer is expected to repay, including principal, interest, and fees.

6. \*\*loan\_type\*\*: The category or type of loan.

7. \*\*disbursement\_date\*\*: The date when the loan amount was disbursed to the customer.

8. \*\*duration\*\*: The length of the loan term, typically expressed in days

9. \*\*lender\_id\*\*: Unique identifier for the lender or institution that issued the loan.

10. \*\*New\_versus\_Repeat\*\*: Indicates whether the loan is the customer's first loan ("New") or if the customer has taken loans before ("Repeat").

11. \*\*Amount\_Funded\_By\_Lender\*\*: The portion of the loan funded directly by the lender.

12. \*\*Lender\_portion\_Funded\*\*: Percentage of the total loan amount funded by the lender.

13. \*\*due\_date\*\*: The date by which the loan repayment is due.

14. \*\*Lender\_portion\_to\_be\_repaid\*\*: The portion of the outstanding loan that needs to be repaid to the lender.

15. \*\*target\*\*: This variable takes the value 0 or 1. 1 means the customer defaulted on the loan, whereas 0 means, the customer paid the loan.

**Description of variables in Economic Indicators**

1. Country – Name of the country
2. Period – Years running from 2001 to 2023
3. Indicator – Selected economic indicators

* Inflation, consumer prices (annual %)
* Official exchange rate (LCU per US$, period average)
* Real interest rate (%)
* Average precipitation in depth (mm per year)
* Deposit interest rate (%)
* Lending interest rate (%)
* Interest rate spread (lending rate minus deposit rate, %)
* Fossil fuel energy consumption (% of total)
* Unemployment rate.

**Task 1: Data Preparation - In python**

1. Combine the train and test datasets, ensure you add a column to show if the record is from test or train.
2. For the economic indicators transform the dataset from wide to long ie to have three columns the Country, Indicator and the Year.
3. Extract the following specific temporal features and name them as follows.

**Month:**

* Name: disbursement\_month, due\_month Extraction: Extract the month (1-12) from the respective date fields.

**Day:**

* Name: disbursement\_day, due\_day Extraction: Extract the day of the month (1-31) from the respective date fields.

**Year:**

* Name: disbursement\_year, due\_year Extraction: Extract the year (e.g., 2024) from the respective date fields.

**Weekday:**

* Name: disbursement\_weekday, due\_weekday Extraction: Extract the weekday as a number (0-6, where 0 is Sunday, 1 is Monday, etc.).

**Quarter:**

* Name: disbursement\_quarter, due\_quarter Extraction: Extract the quarter (1-4) from the respective date fields.

1. Create two temporal features to capture cyclicality i.e. seasonality and continuity on the month variable for both disbursement and due dates.
2. Create a variable for repayment ratio, log transform total amount, percentage change rate in the economic indicators over the recent 3 years, repayment\_ratio × duration and amount/ duration ratio.
3. Filled missing values using country- and indicator-specific mean.
4. Include the average **Inflation and Interest Rates** in thecombine the train and test datasets in (1) above ie
   1. Average\_Inflation\_3Yrs – Average inflation for 3 years
   2. Average\_Inflation\_5Yrs – Average inflation for 5 years
   3. Average\_Interest\_Rates\_3Yrs – Average interest rates for 3 years
   4. Average\_ Interest\_Rates \_5Yrs - Average interest rates for 5 years
5. Document the steps followed inside the python file

**Task 2: Power BI Dashboard Design for Loan Portfolio Analysis**

Create a well-structured Power BI dashboard [**Use Power BI Desktop Report Server Version (May 2024)**] based on this dataset to provide insights into loan performance, default trends, and lender exposure. Below are suggested visuals and their respective chart types:

1. **Loan Portfolio Overview (Summary Cards & KPI Indicators)**

* Total Loans Disbursed (Sum of Total\_Amount) → Card Visualization
* Total Amount to be Repaid (Sum of Total\_Amount\_to\_Repay) → Card Visualization
* Average Loan Duration (Average of duration) → Card Visualization
* Default Rate (%) (SUM(target) / COUNT(tbl\_loan\_id) \* 100) → Card Visualization

1. **Loan Disbursement Trends**

* Loan Disbursements Over Time (disbursement\_date vs. Total\_Amount) → Line Chart
* Loan Repayments Over Time (due\_date vs. Total\_Amount\_to\_Repay) → Line Chart
* New vs. Repeat Loan Trends (disbursement\_date vs. count of customer\_id, segmented by New\_versus\_Repeat) → Stacked Column Chart

### **Loan Performance by Customer & Country**

* **Top 10 Customers by Loan Amount (customer\_id vs. Sum of Total\_Amount) → Bar Chart**
* **Loan Distribution by Country (country\_id vs. Sum of Total\_Amount) → Map or TreeMap**
* **Loan Defaults by Country (country\_id vs. count of target where target = 1) → Clustered Bar Chart**
* **Loan Type Distribution (loan\_type vs. Count of tbl\_loan\_id) → Pie Chart or Donut Chart**

1. **Lender Performance & Exposure**

* Total Amount Funded by Lender (**lender\_id vs. Sum of Amount\_Funded\_By\_Lender**) → Bar Chart
* Lender Portfolio Exposure (**lender\_id vs. Sum of Lender\_portion\_Funded**) → 100% Stacked Bar Chart
* Lender Repayment Performance (**lender\_id vs. Sum of Lender\_portion\_to\_be\_repaid**) → Table or Bar Chart

1. **Default Analysis**

* Default Rate by Loan Type (**loan\_type vs. Target**) → Clustered Bar Chart
* Default Rate by Loan Duration (**duration vs. Target**) → Scatter Plot
* New vs. Repeat Loan Defaults (**New\_versus\_Repeat vs. Target**) → Column Chart
* Total Defaults Over Time (**due\_date vs. Sum of target**) → Line Chart with Trendline

1. **Loan Status Drill-Down Table**

**Table with key fields:**

* customer\_id, tbl\_loan\_id, loan\_type, Total\_Amount, Total\_Amount\_to\_Repay, due\_date, target
* **Conditional Formatting:** Highlight rows where target = 1 (defaulted loans)

1. **Interactive Filters (Slicers)**

* country\_id (Country Selection)
* loan\_type (Filter by Loan Category)
* lender\_id (Filter by Lender)
* disbursement\_date (Date Range Selector)
* New\_versus\_Repeat (Filter by New vs. Repeat Borrowers)

**Task 3: Development of a Probability of Default model in Python**

Aggregate customer level variables and develop a PD model. Apply one-hot encoding to the feature loan type. Check and remove variables that are highly correlated with the target variable use a threshold of 70%.

For the model **do not split** the data to train and test instead use **StratifiedKFold cross-validation.**

**Create two baseline models and access the following accuracy parameters AUC, Precision, Accuracy, Recall and GINI**

* 1. Logistic Regression Model
  2. Random Forest Model
  3. K-Nearest Neighbors
  4. Decision Tree
  5. Naïve Bayes
  6. Support Vector Machine

For the model provide a data frame showing the results for the models above as per the measurement metrics **AUC, Precision, Accuracy, Recall and GINI**

Document the steps followed inside the python file

**Expected files required for the practical interview assessment:**

1. A Jupyter notebook file with the workings for Task 1 and Task 3
2. A Power BI file ensure you import the data to power BI for visualization use DAX to carry out any intermediate computations for Task 2.

**END**