**Problem statement:**

Build a loan model to predict bad loans. Use the dataset provided in the file for the same

**Data source:**

In sqlite3 database. There are multiple tables in the sqlite3 object you have to read the database and import the tables

List of tables in the database:

|  |
| --- |
| 1. loan\_information |
| 1. enterprise\_information |
| 1. scores\_information |
| 1. disbursement\_information |

**Steps**

1. Build a single dataset for multiple sources
2. Clean the dataset
3. Perform EDA
4. Build a model to predict the label

**Technology and tools allowed:**

Use PySpark and Koalas wherever possible. Use Spark ML library. You can use google colab if required.

**Expected output:**

Output should be submitted in jupyter / colab notebooks/ any other formats that can be easily reviewed.