**Big Data**

Project Proposal

**Team 7**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Sec** | **BN** | **ID** | **Email** |
| بموا عريان عياد | 1 | 17 | 9202391 | bemoi.tawadros00@eng-st.cu.edu.eg |
| مارك ياسر نبيل | 2 | 14 | 9203106 | mark.ibrahim00@eng-st.cu.edu.eg |
| بيتر عاطف فتحي | 1 | 18 | 9202395 | peter.zaki00@eng-st.cu.edu.eg |
| كريم محمود كمال | 2 | 12 | 9203076 | kairm.mohamed003@eng-st.cu.edu.eg |

# **Idea**

## **Business Problem**

Should a Loan be Approved or Denied?

## **Problem Statement**

Given the dataset from the U.S. Small Business Administration (SBA) comprising loan application information, the challenge is to develop a predictive model that effectively evaluates loan applications to determine whether they should be approved or denied. By leveraging historical data on both successful and defaulted loans, the goal is to create a robust decision-making tool that balances the promotion of small business growth with the need to minimize credit risk. This model should aid lending institutions in making informed decisions, ultimately contributing to the sustainability of small businesses and the broader economy.

# **Dataset**

Number of features : **27**

Dataset size :  **179.43 MB (899164 rows)**

Link : <https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied/data>

# 

[Big Data Project - Google Drive](https://drive.google.com/drive/folders/1zq0MBvshM81FszCgAoTLz2dwT-razDqr)

# **Planned Approach | Proposed Solution**

**1. Data Exploration and Preprocessing:**

* Perform exploratory data analysis (EDA) to understand the characteristics of the data, distribution of features, missing values, and outliers using a descriptive analysis method like clustering.
* Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features if necessary.

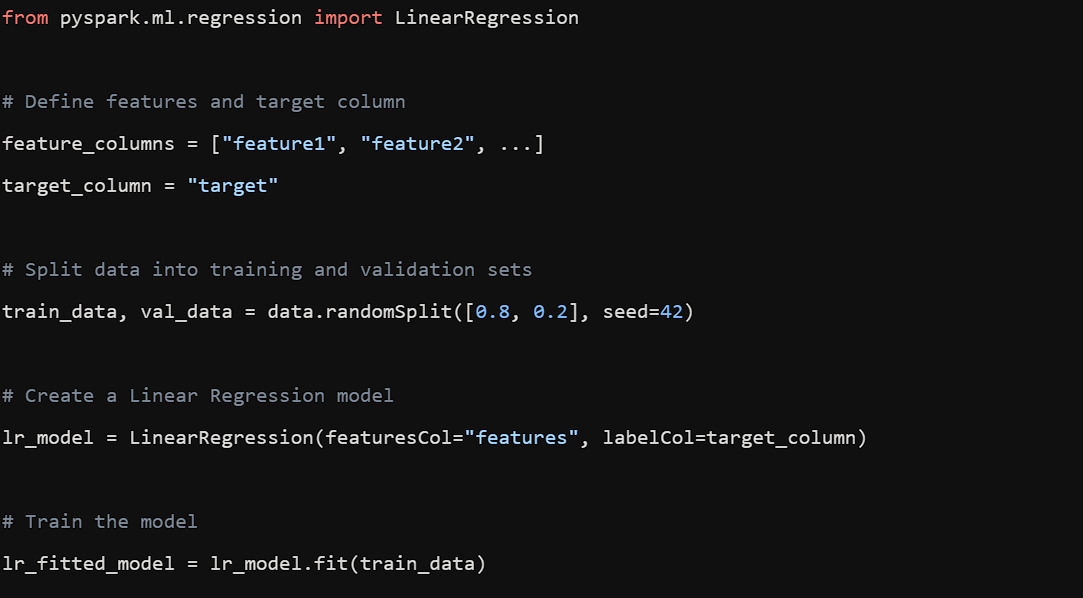
**2. Feature Engineering:**

* Extract relevant features from the dataset that could influence loan approval decisions, such as **loan amount**, **purpose of the loan**, **debt-to-income ratio**, etc.
* Create new features, if applicable, such as **loan-to-income ratio**, etc., which provide additional insights.

**3. Model Selection and Training:**

* Split the dataset into training and testing sets.
* Experiment with machine learning algorithms suitable for binary classification tasks, such as logistic regression, decision trees, and random forests.
* Train multiple models using the training dataset and evaluate their performance using appropriate metrics such as accuracy, precision, recall, F1-score, AUC-ROC.

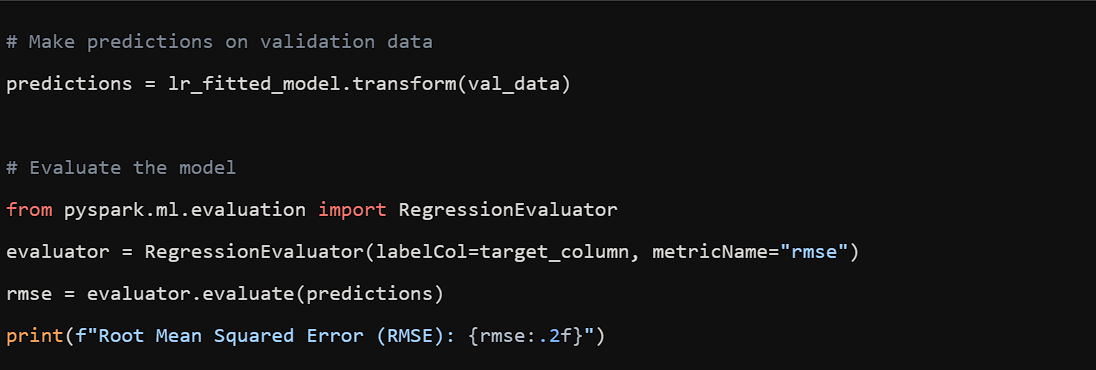
**4. Distributed training:**

* Use Spark or Hadoop for distributed training using map-reduce to reduce training time

**[PySpark distributed training example]**

**5. Model Evaluation and Tuning:**

* Perform model evaluation on the testing dataset to assess the generalization performance of the trained models.
* Fine-tune the hyperparameters of the selected model using techniques like grid search or random search to optimize their performance.
* Address any issues related to overfitting or underfitting by adjusting model complexity or regularization parameters.



**[PySpark distributed evaluation example]**

**6. Model Interpretation:**

* Interpret the trained models to gain insights into the factors influencing loan approval decisions. Techniques such as feature importance analysis, and partial dependence plots can help understand the model's decision-making process.