Phase 1:

The data for HCDR Project entails 10 csv file ranging from 37 KB to 690 MB Dataset size with following features:

* + (688 meg compressed) with millions of rows of data
  + 2.71 Gig of data uncompressed

Following is the file sizes which we plan to use in our project:

**name [ rows cols] MegaBytes**

application\_train [ 307,511, 122] 158MB

application\_test [ 48,744, 121] 25MB

bureau [ 1,716,428, 17] 162MB

bureau\_balance [ 27,299,925, 3] 358MB

credit\_card\_balance [ 3,840,312, 23] 405MB

installments\_payments [ 13,605,401, 8] 690MB

previous\_application [ 1,670,214, 37] 386MB

POS\_CASH\_balance [ 10,001,358, 8] 375MB

Each file serves its own purpose, and the details of these files are mentioned below:

1. **HomeCredit\_columns\_description.csv:**
   * Contains descriptions for the columns in the various data files.
2. **POS\_CASH\_balance.csv:**
   * Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
   * One row for each month of history of every previous credit in Home Credit related to loans in the sample.
3. **application\_test.csv:**
   * Test dataset for loan applications.
   * Static data for various loan applications, with each row representing a loan.
4. **application\_train.csv:**
   * Train dataset for loan applications.
   * Similar to the test dataset, it contains static data for various loan applications, including the target variable.
5. **bureau.csv:**
   * Information about a client's previous credits from other financial institutions, reported to the Credit Bureau.
   * For clients with a loan in the sample, the number of rows equals the number of credits the client had in the Credit Bureau before the application date.
6. **bureau\_balance.csv:**
   * Monthly balances of previous credits in the Credit Bureau.
   * Each row corresponds to each month of history for every previous credit reported to the Credit Bureau.
7. **credit\_card\_balance.csv:**
   * Monthly balance snapshots of previous credit cards that the applicant had with Home Credit.
   * One row for each month of history for every previous credit in Home Credit related to loans in the sample.
8. **installments\_payments.csv:**
   * Repayment history for the previously disbursed credits in Home Credit related to the loans in the sample.
   * One row for every payment made and an additional row for each missed payment.
9. **previous\_application.csv:**
   * All previous applications for Home Credit loans of clients who have loans in the sample.
   * One row for each previous application related to loans in the data sample.
10. **sample\_submission.csv:**
    * A sample submission file for predictions for the project.

**Gantt Chart:**

**A screenshot of a calendar

Description automatically generated**

**Phase 2:**

**Project Description:**

This project targets to serve the financial services to underserved customers who are unbanked population by providing a positive and safe borrowing experience. While this project will be using various statistical and machine learning methods to make these predictions, It will also serve to contribute this expertise in developing more effective models and strategies for credit risk assessment. The ultimate objective is to enhance financial inclusion and provide fair and positive loan experiences for a broader segment of the population.

**The project is divided into following major Project activities:**

**1. Project Setup and Exploration**

* Set up the development environment.
* Understand the dataset structure and features.
* Explore data statistics, missing values, and distributions.
* Visualize relationships between variables.

**2. Data Preprocessing**

* Handle missing values.
* Encode categorical variables.
* Deal with outliers.
* Feature scaling or normalization if needed.
* Create new relevant features if possible.

**3. Baseline Model**

* Build a baseline model using the provided starter notebook.
* Submit the baseline model to Kaggle for initial evaluation.
* Understand the performance metrics provided by Kaggle.

**4. In-Depth Data Analysis**

* Explore correlations between features and the target variable.
* Identify patterns that might be indicative of creditworthiness.
* Consider feature engineering based on insights gained.

**5. Model Improvement**

* Experiment with different machine learning algorithms (e.g., Random Forest, Gradient Boosting, Logistic Regression).
* Fine-tune hyperparameters using cross-validation.
* Implement feature selection techniques.
* Utilize advanced methods (ensemble models, stacking) for improved performance.

**6. Model Evaluation and Interpretability**

* Evaluate models using appropriate metrics (accuracy, precision, recall, F1-score).
* Interpret model results and understand feature importance.
* Create visualizations to explain model predictions.

**8. Documentation and Presentation**

* Create a well-documented notebook explaining your process, choices, and results.
* Prepare a presentation summarizing your approach, challenges faced, and outcomes.