**Phase 1: Problem Understanding and Dataset Exploration**

**Assumptions:**

* Banks want to minimize loan defaults and assess the creditworthiness of borrowers.
* Machine learning models can help capture patterns and predict defaults more effectively than traditional scoring methods.
* Home Credit Group's dataset contains information relevant to loan defaults (e.g., client demographics, loan history).

**Overall Objective:**

Understand the dataset and the business problem, identify key features, and validate the potential of machine learning for predicting loan defaults.

**Specific Objectives:**

1. **Dataset Loading & Inspection**:
   * Load the Home Credit dataset and explore its structure.
   * Identify missing values, data types, and overall data quality.
   * Understand the distribution of key variables (e.g., target variable like loan default).
2. **Define the Business Problem**:
   * Define what constitutes "default" and the target variable.
   * Set initial goals for the model (e.g., predict whether a loan will default based on borrower characteristics).
3. **Data Cleaning & Preprocessing**:
   * Handle missing values, outliers, and data imbalances (e.g., many more non-defaulters than defaulters).
   * Feature engineering: Create new features from existing ones, like debt-to-income ratio, loan duration, or client risk profiles.

**Phase 2: Baseline Model Development**

**Assumptions:**

* A baseline model (simple model) can provide a starting point for comparison.
* Basic machine learning algorithms (e.g., logistic regression) can establish the model's predictive power.

**Overall Objective:**

Develop a baseline machine learning model to predict loan defaults, providing a starting point for more advanced models.

**Specific Objectives:**

1. **Feature Selection**:
   * Identify and select features that are likely to impact loan defaults (e.g., income, loan history).
   * Address multicollinearity or correlations between features.
2. **Split Data**:
   * Split data into training and testing sets (e.g., 80/20 or using cross-validation) to ensure proper model evaluation.
3. **Baseline Model**:
   * Train a simple model like logistic regression or decision tree.
   * Evaluate the model's accuracy, precision, recall, and AUC-ROC score to understand how well it captures defaults.

**Phase 3: Advanced Modeling & Experimentation**

**Assumptions:**

* More sophisticated models (e.g., random forest, gradient boosting) will improve performance.
* Feature importance and interactions can uncover hidden insights into loan defaults.

**Overall Objective:**

Build and evaluate more advanced models, tuning parameters and using a variety of approaches to maximize prediction accuracy.

**Specific Objectives:**

1. **Model Selection & Tuning**:
   * Experiment with more complex models (e.g., Random Forest, Gradient Boosting, XGBoost).
   * Use hyperparameter tuning (e.g., GridSearchCV) to improve model performance.
2. **Model Comparison**:
   * Compare models based on key metrics (e.g., accuracy, AUC-ROC, precision, recall).
   * Use feature importance to identify which factors contribute most to the prediction of loan defaults.
3. **Handling Class Imbalance**:
   * Address class imbalance through techniques like SMOTE (Synthetic Minority Oversampling) or adjusting the decision threshold.

**Phase 4: Model Validation & Interpretation**

**Assumptions:**

* Banks will require explainable models to justify lending decisions.
* Model validation is critical to ensure the model generalizes well on unseen data.

**Overall Objective:**

Validate the chosen model(s), ensuring accuracy, fairness, and interpretability, and prepare to present results to potential clients.

**Specific Objectives:**

1. **Cross-Validation & Testing**:
   * Use k-fold cross-validation to ensure the model generalizes well.
   * Test the model on unseen data and evaluate performance metrics.
2. **Model Explainability**:
   * Use techniques like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) to explain the model's decisions.
   * Create clear visualizations of feature importance and model predictions.
3. **Prepare for Client Presentation**:
   * Develop a presentation of key findings, explaining how the model can help banks reduce loan defaults.
   * Emphasize the accuracy, interpretability, and impact of the solution.

**Phase 5: Feedback & Iteration**

**Assumptions:**

* Banks will have specific needs and concerns, such as regulatory requirements or risk tolerances.
* The POC might need customization based on feedback from potential clients.

**Overall Objective:**

Iterate on the POC based on client feedback, making necessary adjustments to the models, features, or outputs.

**Specific Objectives:**

1. **Client Feedback**:
   * Gather feedback on the model's utility, performance, and potential improvements.
2. **Model Iteration**:
   * Refine the model, adjust features, or adapt the solution based on specific client needs (e.g., focusing on particular risk factors).
3. **Finalize POC**:
   * Ensure the POC is robust, accurate, and client-ready, with clear documentation and visualizations.

'SK\_ID\_CURR',

 'TARGET',

 'NAME\_CONTRACT\_TYPE',

 'CODE\_GENDER',

 'FLAG\_OWN\_CAR',

 'FLAG\_OWN\_REALTY',

 'CNT\_CHILDREN',

 'AMT\_INCOME\_TOTAL',

 'AMT\_CREDIT',

 'AMT\_ANNUITY',

 'AMT\_GOODS\_PRICE',

 'NAME\_TYPE\_SUITE',

 'NAME\_INCOME\_TYPE',

 'NAME\_EDUCATION\_TYPE',

 'NAME\_FAMILY\_STATUS',

 'NAME\_HOUSING\_TYPE',

 'REGION\_POPULATION\_RELATIVE',

 'DAYS\_BIRTH',

 'DAYS\_EMPLOYED',

 'DAYS\_REGISTRATION',

 'DAYS\_ID\_PUBLISH',

 'OWN\_CAR\_AGE',

 'FLAG\_MOBIL',

 'FLAG\_EMP\_PHONE',

 'FLAG\_WORK\_PHONE',

 'FLAG\_CONT\_MOBILE',

 'FLAG\_PHONE',

 'FLAG\_EMAIL',

 'OCCUPATION\_TYPE',

 'CNT\_FAM\_MEMBERS',

 'REGION\_RATING\_CLIENT',

 'REGION\_RATING\_CLIENT\_W\_CITY',

 'WEEKDAY\_APPR\_PROCESS\_START',

 'HOUR\_APPR\_PROCESS\_START',

 'REG\_REGION\_NOT\_LIVE\_REGION',

 'REG\_REGION\_NOT\_WORK\_REGION',

 'LIVE\_REGION\_NOT\_WORK\_REGION',

 'REG\_CITY\_NOT\_LIVE\_CITY',

 'REG\_CITY\_NOT\_WORK\_CITY',

 'LIVE\_CITY\_NOT\_WORK\_CITY',

 'ORGANIZATION\_TYPE',

 'EXT\_SOURCE\_1',

 'EXT\_SOURCE\_2',

 'EXT\_SOURCE\_3',

 'APARTMENTS\_AVG',

 'BASEMENTAREA\_AVG',

 'YEARS\_BEGINEXPLUATATION\_AVG',

 'YEARS\_BUILD\_AVG',

 'COMMONAREA\_AVG',

 'ELEVATORS\_AVG',

 'ENTRANCES\_AVG',

 'FLOORSMAX\_AVG',

 'FLOORSMIN\_AVG',

 'LANDAREA\_AVG',

 'LIVINGAPARTMENTS\_AVG',

 'LIVINGAREA\_AVG',

 'NONLIVINGAPARTMENTS\_AVG',

 'NONLIVINGAREA\_AVG',

 'APARTMENTS\_MODE',

 'BASEMENTAREA\_MODE',

 'YEARS\_BEGINEXPLUATATION\_MODE',

 'YEARS\_BUILD\_MODE',

 'COMMONAREA\_MODE',

 'ELEVATORS\_MODE',

 'ENTRANCES\_MODE',

 'FLOORSMAX\_MODE',

 'FLOORSMIN\_MODE',

 'LANDAREA\_MODE',

 'LIVINGAPARTMENTS\_MODE',

 'LIVINGAREA\_MODE',

 'NONLIVINGAPARTMENTS\_MODE',

 'NONLIVINGAREA\_MODE',

 'APARTMENTS\_MEDI',

 'BASEMENTAREA\_MEDI',

 'YEARS\_BEGINEXPLUATATION\_MEDI',

 'YEARS\_BUILD\_MEDI',

 'COMMONAREA\_MEDI',

 'ELEVATORS\_MEDI',

 'ENTRANCES\_MEDI',

 'FLOORSMAX\_MEDI',

 'FLOORSMIN\_MEDI',

 'LANDAREA\_MEDI',

 'LIVINGAPARTMENTS\_MEDI',

 'LIVINGAREA\_MEDI',

 'NONLIVINGAPARTMENTS\_MEDI',

 'NONLIVINGAREA\_MEDI',

 'FONDKAPREMONT\_MODE',

 'HOUSETYPE\_MODE',

 'TOTALAREA\_MODE',

 'WALLSMATERIAL\_MODE',

 'EMERGENCYSTATE\_MODE',

 'OBS\_30\_CNT\_SOCIAL\_CIRCLE',

 'DEF\_30\_CNT\_SOCIAL\_CIRCLE',

 'OBS\_60\_CNT\_SOCIAL\_CIRCLE',

 'DEF\_60\_CNT\_SOCIAL\_CIRCLE',

 'DAYS\_LAST\_PHONE\_CHANGE',

 'FLAG\_DOCUMENT\_2',

 'FLAG\_DOCUMENT\_3',

 'FLAG\_DOCUMENT\_4',

 'FLAG\_DOCUMENT\_5',

 'FLAG\_DOCUMENT\_6',

 'FLAG\_DOCUMENT\_7',

 'FLAG\_DOCUMENT\_8',

 'FLAG\_DOCUMENT\_9',

 'FLAG\_DOCUMENT\_10',

 'FLAG\_DOCUMENT\_11',

 'FLAG\_DOCUMENT\_12',

 'FLAG\_DOCUMENT\_13',

 'FLAG\_DOCUMENT\_14',

 'FLAG\_DOCUMENT\_15',

 'FLAG\_DOCUMENT\_16',

 'FLAG\_DOCUMENT\_17',

 'FLAG\_DOCUMENT\_18',

 'FLAG\_DOCUMENT\_19',

 'FLAG\_DOCUMENT\_20',

 'FLAG\_DOCUMENT\_21',

 'AMT\_REQ\_CREDIT\_BUREAU\_HOUR',

 'AMT\_REQ\_CREDIT\_BUREAU\_DAY',

 'AMT\_REQ\_CREDIT\_BUREAU\_WEEK',

 'AMT\_REQ\_CREDIT\_BUREAU\_MON',

 'AMT\_REQ\_CREDIT\_BUREAU\_QRT',

 'AMT\_REQ\_CREDIT\_BUREAU\_YEAR'

'SK\_ID\_CURR',  
 'SK\_ID\_BUREAU',  
 'CREDIT\_ACTIVE',  
 'CREDIT\_CURRENCY',  
 'DAYS\_CREDIT',  
 'CREDIT\_DAY\_OVERDUE',  
 'DAYS\_CREDIT\_ENDDATE',  
 'DAYS\_ENDDATE\_FACT',  
 'AMT\_CREDIT\_MAX\_OVERDUE',  
 'CNT\_CREDIT\_PROLONG',  
 'AMT\_CREDIT\_SUM',  
 'AMT\_CREDIT\_SUM\_DEBT',  
 'AMT\_CREDIT\_SUM\_LIMIT',  
 'AMT\_CREDIT\_SUM\_OVERDUE',  
 'CREDIT\_TYPE',  
 'DAYS\_CREDIT\_UPDATE',  
 'AMT\_ANNUITY'

'SK\_ID\_BUREAU',

'MONTHS\_BALANCE',

'STATUS'

'SK\_ID\_PREV',

 'SK\_ID\_CURR',

 'NAME\_CONTRACT\_TYPE',

 'AMT\_ANNUITY',

 'AMT\_APPLICATION',

 'AMT\_CREDIT',

 'AMT\_DOWN\_PAYMENT',

 'AMT\_GOODS\_PRICE',

 'WEEKDAY\_APPR\_PROCESS\_START',

 'HOUR\_APPR\_PROCESS\_START',

 'FLAG\_LAST\_APPL\_PER\_CONTRACT',

 'NFLAG\_LAST\_APPL\_IN\_DAY',

 'RATE\_DOWN\_PAYMENT',

 'RATE\_INTEREST\_PRIMARY',

 'RATE\_INTEREST\_PRIVILEGED',

 'NAME\_CASH\_LOAN\_PURPOSE',

 'NAME\_CONTRACT\_STATUS',

 'DAYS\_DECISION',

 'NAME\_PAYMENT\_TYPE',

 'CODE\_REJECT\_REASON',

 'NAME\_TYPE\_SUITE',

 'NAME\_CLIENT\_TYPE',

 'NAME\_GOODS\_CATEGORY',

 'NAME\_PORTFOLIO',

 'NAME\_PRODUCT\_TYPE',

 'CHANNEL\_TYPE',

 'SELLERPLACE\_AREA',

 'NAME\_SELLER\_INDUSTRY',

 'CNT\_PAYMENT',

 'NAME\_YIELD\_GROUP',

 'PRODUCT\_COMBINATION',

 'DAYS\_FIRST\_DRAWING',

 'DAYS\_FIRST\_DUE',

 'DAYS\_LAST\_DUE\_1ST\_VERSION',

 'DAYS\_LAST\_DUE',

 'DAYS\_TERMINATION',

 'NFLAG\_INSURED\_ON\_APPROVAL'

'SK\_ID\_PREV',  
 'SK\_ID\_CURR',  
 'MONTHS\_BALANCE',  
 'CNT\_INSTALMENT',  
 'CNT\_INSTALMENT\_FUTURE',  
 'NAME\_CONTRACT\_STATUS',  
 'SK\_DPD',  
 'SK\_DPD\_DEF'

'SK\_ID\_PREV',  
 'SK\_ID\_CURR',  
 'NUM\_INSTALMENT\_VERSION',  
 'NUM\_INSTALMENT\_NUMBER',  
 'DAYS\_INSTALMENT',  
 'DAYS\_ENTRY\_PAYMENT',  
 'AMT\_INSTALMENT',  
 'AMT\_PAYMENT'

'SK\_ID\_PREV',  
 'SK\_ID\_CURR',  
 'MONTHS\_BALANCE',  
 'AMT\_BALANCE',  
 'AMT\_CREDIT\_LIMIT\_ACTUAL',  
 'AMT\_DRAWINGS\_ATM\_CURRENT',  
 'AMT\_DRAWINGS\_CURRENT',  
 'AMT\_DRAWINGS\_OTHER\_CURRENT',  
 'AMT\_DRAWINGS\_POS\_CURRENT',  
 'AMT\_INST\_MIN\_REGULARITY',  
 'AMT\_PAYMENT\_CURRENT',  
 'AMT\_PAYMENT\_TOTAL\_CURRENT',  
 'AMT\_RECEIVABLE\_PRINCIPAL',  
 'AMT\_RECIVABLE',  
 'AMT\_TOTAL\_RECEIVABLE',  
 'CNT\_DRAWINGS\_ATM\_CURRENT',  
 'CNT\_DRAWINGS\_CURRENT',  
 'CNT\_DRAWINGS\_OTHER\_CURRENT',  
 'CNT\_DRAWINGS\_POS\_CURRENT',  
 'CNT\_INSTALMENT\_MATURE\_CUM',  
 'NAME\_CONTRACT\_STATUS',  
 'SK\_DPD',  
 'SK\_DPD\_DEF'