Tab 1

`📌 High-Level Project Goal:  
 Identify and pre-select low-risk NTC individuals using ML modeling on credit bureau and alternative data for targeted credit product invitations.

🧠 Project Framework for ML-Driven NTC Credit Targeting

1. Project Planning & Definition  
   * Objective: Pre-select NTC Segment with PD < 25% and dollar NCL% < 10%.
   * **Outcome of machine learning development:** Binary prediction attribute to indicate trade default or not. Dollar prediction on how much dollar of that charge-off trade.### chat gpt tim xem no hieu cai gi (<https://www.kaggle.com/datasets/laotse/credit-risk-dataset/code>)   
     Description .cac buoc build hieu qua

\*(overfitted ) bi nghieng ve 1 thang

\*step 2 EDA -> classify chat gpt on classification and regression model

\*simple EDA

* + **Outcome of project**: utilize cross-tab analysis to look at PD% and NCL% of segments: Asset X Score, Tenure X Asset, DTI X Score, and Total Debt X Score. -> our risk threshold is PD% < 25% and NCL% < 10%. IF a segment has a higher loss than this threshold we are not sending out pre-selected mail.
  + Tools: ML development: Python
  + Tools: Cross-tab analysis: Tableau .

1. Data Collection & Understanding: **Hoai will create and supply the data ( thank you hoai)**
   * Data Sources:  
     + Credit bureau data: demographics, tradelines, delinquency flags, charge-off indicators.
     + Alternative data: assets, employment, utility payments, digital footprints (if any).
   * Key Variables: income, DTI, assets, inquiries, age, number of trades, score, charge-off indicator, total dollar loss.
2. Data Preparation & Binning  
   * Clean data: Handle missing values, standardize formats.
   * Binning strategy (convert numerical variables to categorical bands): Down here are example
     + Income → <20k, 20k–50k, 50k+
     + Assets → <500, 500–2,000, 2,000+
     + DTI → <20%, 20–35%, 35%+
     + Score → <600, 600–680, 680+
     + Age → <25, 25–30, 30+
     + Employment status, zip risk level, etc.
3. Outcome Variable Definition  
   * Binary Target: Prediction of PD  
     + PD\_forecast = 1 / 0
   * Continuous Target: Prediction of loss dolar regard to particular PD  
     + Expected loss = dollar amount
4. Exploratory Data Analysis (EDA)  
   * Analyze relationships:  
     + Correlation heatmaps
     + Bivariate analysis: bad rate and dollar loss by bands
     + Identify stable variables across time
   * Visualize with Tableau or seaborn (Python)
5. Model Building (Please refer to Tab 2)  
   * Model Choice:  
     + Baseline: Logistic Regression
     + Advanced: Random Forest, XGBoost, LightGBM
   * Split data:  
     + Train (70%) / Validation (15%) / Test (15%)
   * Metrics for evaluation:  
     + AUC-ROC, KS-statistic, Precision-Recall
     + Gini for segmentation power
6. Model Interpretation & Variable Importance  
   * Use SHAP or permutation importance for explainability.
   * Identify which features best predict PD / dollar charge-off.
7. Risk Segmentation & Business Rule Design  
     
   * Group into segments using:  
     + Asset X Score - (
     + Tenure X Asset (month the first credit) ( 0-6 , 6-12)-36+
     + DTI(0-10%,10%-20) X Score (debt-to-income)
     + Total Debt X Score (nhu asset) debt cao -> tuy theo cai distribution
   * Apply Unit PD% < 25% and NCL% Dollar loss < 10% thresholds ( deo approve)
8. Cross-tab & Strategy Analysis  
   * Example: For PD% and NCL%. If both X and Y are below thresholds, GREEN, otherwise RED.
   * For example, possible outcome, based on the table below, our cut-off is **$500+ Asset AND 681+ Score**

| Asset/Score | 250-660 | 661-680 | 681-700 | 701-800 | 801-900 | Grand Total |
| --- | --- | --- | --- | --- | --- | --- |
| 0-$500 | Total account: 100  Total CO: 25.  Total loan: $100K.  Total NCL:  $11K.  X=25/100=25%  y=11K/100k=11%  X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |
| $500 - $1K | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |
| $1K - $3K | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |
| $3K - $5K | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |
| $5K - $10K | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |
| $10K+ | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |
| Grand Total | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% | X% | Y% |



1. Final Output & Business Recommendation

* Define rule-based pre-screening logic based on score and NCL bands.
* Recommend criteria for emailing “pre-qualified” customers.
* Document guardrails to revisit quarterly (e.g., macroeconomic shocks).

1. Optional – Simulate Impact: Leave this step later, refocus later.

* Estimate approval rate, expected loss, and profit at each segment.
* Model ROI from inviting vs. excluding a group.

Dung doc nua anh hoai oi nghe dth di !

Tab 2

Train Data before mapping (table 1):

| ID | Month with since 1st credit | Score | Income | Asset | Total Debt | Credit Line | 30+ other attributes | Actual Charge-Off Tr  ade | Actual Charge-Off Dollar |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 5 | 696 | $15000 | $2000 | $0 | $1000 | … | 0 | 0 |
| 2 | 10 | 730 | $5000 | $300 | $20000 | $750 | … | 0 | 0 |
| … | … | 655 | $30000 | $90 | $6900 | $1500 | … | 1 | $690 |
| 1000 | 17 | 715 | $7890 | $0 | $2500 | $500 | … | 0 | 0 |

Train Data After mapping (table 2):

| ID | All Raw attributes | Score Bands | Month since 1st credit band | Income Band | Asset Band | Total Debt band | Debt-to-income | Debt-to-income band | Credit line band |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | … | 691-700 | 0-6M | $10K-$20K | $1K-$3k | No Debt | 0 | 0 | $500-$1K |
| 2 | … | 721-730 | 7-12M | <= $5K | <$500 | $20K-$30K | 400% | 200%+ | $500-$1K |
| … | … | 651-660 | … | $20K-$30K | <$500 | $5K-$10K | 23% | 15%-30% | $1K-$2K |
| 1000 | … | 721-730 | 12-24M | $5K-$10K | No Asset | $2.5K-$$5K | 32% | 30%-50% | $500-$1K |

**Train data** with machine learning outcomes (table 3, basically random 70% of table 2): learn pattern and correlation between attributes toward actual charge-off trade and actual charge-off dollar

| ID | All Raw attributes | Score Bands | Month since 1st credit band | Income Band | Asset Band | Total Debt band | Debt-to-income | Debt-to-income band | Credit line band |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | … | 691-700 | 0-6M | $10K-$20K | $1K-$3k | No Debt | 0 | 0 | $500-$1K |
| 2 | … | 721-730 | 7-12M | <= $5K | <$500 | $20K-$30K | 400% | 200%+ | $500-$1K |
| … | … | 651-660 | … | $20K-$30K | <$500 | $5K-$10K | 23% | 15%-30% | $1K-$2K |
| 700 | … | … | … | … | … | … | … | … | ... |

**Validation data** with machine learning outcomes (table 4, random 15% of table 2): Have machine learning to output 2 prediction variable like below.

| ID | All other variables from Table 2 | Actual Charge-Off Trade | Actual Charge-Off Dollar | Prediction of CO Trades | Prediction of CO Dollar |
| --- | --- | --- | --- | --- | --- |
| 701 | .. | 0 | 0 | 1 | $200 |
| 702 | .. | 0 | 0 | 0 | 0 |
| … | .. | 0 | 0 | 0 | 0 |
| 850 | .. | 1 | $ 470 | 0 | 0 |

**Test data** with machine learning outcomes (table 5): Have built models to predict PD and PD Dollar here. And compare with actual PD and actual PD Dollar

| ID | All other variables from Table 2 | Actual Charge-Off Trade | Actual Charge-Off Dollar | Prediction of CO Trades | Prediction of CO Dollar |
| --- | --- | --- | --- | --- | --- |
| 851 | .. | 0 | 0 | 0 | 0 |
| 852 | .. | 0 | 0 | 1 | $500 |
| … | .. | 1 | $690 | 1 | $700 |
| 1000 | .. | 0 | 0 | 0 | 0 |

Tab 3

* 28 borrower attributes columns(features)
* 2 label columns:  
  + Binary indicator (charge-off: yes/no)
  + Dollar amount of charge-off (can be 0)

This means you're working with a supervised learning problem where your labels are fully observed in all three data splits.

Here is the updated breakdown of your framework, reflecting this corrected structure:

🔁 Updated Model Development Framework with Pre-Labeled Inputs

STEP 1: DATA PREPARATION & SPLIT

* Dataset: 10,000 rows (each row = one borrower)
* Columns:  
  + 28 borrower attributes (income, assets, DTI, etc.)
  + charge\_off\_flag: 1 if ever charged-off, else 0
  + charge\_off\_amount: actual $ loss, 0 if none

Split data into:

* Train set (e.g., 60%)
* Validation set (e.g., 20%)
* Test set (e.g., 20%)

All sets already contain the actual labels (charge\_off\_flag, charge\_off\_amount).

STEP 2: MODEL TRAINING (Train Set)

* Input: 28 attributes (X\_train), charge\_off\_flag & charge\_off\_amount (Y\_train)
* Train:  
  1. A binary classification model to predict PD% (probability of default)
  2. A regression model to predict expected dollar loss (EAD × LGD = NCL)
* Output: Trained models

Note: You do not use predictions here; this is the learning phase using true labels.

STEP 3: MODEL VALIDATION (Validation Set)

* Input: 28 attributes (X\_valid)
* Run prediction using trained models:  
  + prediction\_pd = model.predict\_proba(X\_valid)
  + prediction\_ncl = model.predict\_loss(X\_valid)
* Append predictions:  
  + Actuals: charge\_off\_flag, charge\_off\_amount
  + Predictions: prediction\_pd, prediction\_ncl
* Analyze:  
  + Group data into segments using attributes (e.g., income band + score)
  + Cross-tab segments by:  
    - Actual PD% = sum(actual charge\_off\_flag) / total accounts in segment
    - Actual NCL% = sum(actual charge\_off\_amount) / total loaned $
    - Predicted PD% = average(prediction\_pd) in segment
    - Predicted NCL% = average(prediction\_ncl) in segment
* Use results to fine-tune model or segment strategy (e.g., apply decision threshold)

STEP 4: MODEL TESTING (Test Set)

* Repeat same process as validation:  
  + Run predictions
  + Compare actual vs predicted
  + Assess final model performance using:  
    - Classification metrics: ROC AUC, accuracy, recall, precision
    - Regression metrics: RMSE, MAE, R²
    - Business metrics: segment-level PD% and NCL% for target bands

STEP 5: SEGMENT STRATEGY & ACTION

* Use crosstab analysis to identify low-risk segments (e.g., asset > $500 and score > 680)
* Apply business guardrails:  
  + Only invite segments where:  
    - PD% < 25%
    - NCL% < 10%
* Send pre-qualified email offers only to these segments