PROJECT

# CREDIT DEFAULT PREDICTION

# CONTENT

# PROJECT OBJECTIVE

## PROBLEM DESCRIPTION -

- BANKS ALL AROUND THE WORLD RECEIVE COUNTLESS APPLICATIONS FOR CREDIT EVERY DAY. SOME OF THEM ARE GOOD AND WILL BE REPAID, BUT THERE IS A STILL HIGH RISK THAT ONE CREDITOR DEFAULTS HIS/HER LOANS.
- CREDIT DEFAULTS POSE A SIGNIFICANT RISK TO A COMPANY'S FINANCIAL STABILITY AND LONG-TERM SUSTAINABILITY. IT LEADS TO FINANCIAL LOSSES AS THE COMPANY MAY NOT RECOVER THE FULL AMOUNT OF THE LOAN, IMPACTING ITS PROFITABILITY AND CASH FLOW. THE RESOURCES INVESTED IN EVALUATING AND GRANTING LOANS TO DEFAULTING BORROWERS GO TO WASTE. MOREOVER, THE COMPANY'S REPUTATION AND CREDIBILITY CAN BE DAMAGED IF IT IS KNOWN FOR HAVING A HIGH DEFAULT RATE, LEADING TO A LOSS OF TRUST FROM CUSTOMERS AND INVESTORS.
- HENCE PREDICTING CREDIT DEFAULTS IS CRUCIAL FROM A BUSINESS PERSPECTIVE AS IT ENABLES BANKS TO EFFECTIVELY MANAGE RISK, REDUCE COSTS, ALLOCATE RESOURCES EFFICIENTLY, GAIN A COMPETITIVE ADVANTAGE, AND COMPLY WITH REGULATORY REQUIREMENTS.
- BY USING MACHINE LEARNING TECHNIQUES, BANKS CAN ASSESS THE CREDITWORTHINESS OF BORROWERS AND IDENTIFY THOSE WITH A HIGH PROBABILITY OF DEFAULT, ALLOWING THEM TO MAKE INFORMED LENDING DECISIONS, PROTECT THEIR LOAN PORTFOLIOS, AND MAINTAIN FINANCIAL STABILITY.

## PROBLEM OBJECTIVE

- THE OBJECTIVE OF THIS SOLUTION IS TO DEVELOP A MACHINE LEARNING MODEL THAT WILL ACCURATELY PREDICT IF A CREDITOR WILL DEFAULT ON THE LOAN WITH THE AIM TO MINIMIZE FALSE POSITIVES AND FALSE NEGATIVES WHILE MAXIMIZING THE RECALL RATE AND AREA UNDER THE CURVE (AUC) IN THE MODEL'S PERFORMANCE.
- FALSE POSITIVES OCCUR WHEN THE MODEL PREDICTS A BORROWER AS A DEFAULTER WHEN THEY ACTUALLY REPAY THE LOAN, POTENTIALLY LEADING TO LOST BUSINESS OPPORTUNITIES. FALSE NEGATIVES, ON THE OTHER HAND, HAPPEN WHEN THE MODEL PREDICTS A BORROWER AS NON-DEFAULTING, BUT THEY END UP DEFAULTING, RESULTING IN FINANCIAL LOSSES FOR THE BANK.
- MAXIMIZING THE RECALL RATE ENSURES THAT A HIGHER PROPORTION OF ACTUAL DEFAULTERS ARE CORRECTLY IDENTIFIED, REDUCING THE RISK OF GRANTING LOANS TO HIGH-RISK BORROWERS.
- PREDICTING DEFAULTS AS NON-DEFAULTS CAN SEVERELY HURT THE BANK AS IT MAY LEAD TO INCREASED DEFAULT RATES, FINANCIAL LOSSES, AND DAMAGE TO THE BANK'S REPUTATION AND CREDIBILITY.
- THEREFORE, ACCURATELY PREDICTING DEFAULTS IS CRUCIAL FOR RISK MANAGEMENT AND ENSURING THAT THE BANK CAN MAKE INFORMED DECISIONS TO MITIGATE THE POTENTIAL NEGATIVE IMPACT OF DEFAULTING BORROWERS.

## PLAN TO SOLVE

- TO SOLVE THE CREDIT DEFAULT PREDICTION PROBLEM USING MACHINE LEARNING, WE CAN FOLLOW THESE STEPS:
- DATA CLEANING: THE FIRST STEP IS TO CLEAN THE DATASET BY HANDLING MISSING VALUES, REMOVING DUPLICATES, AND ADDRESSING ANY INCONSISTENCIES OR ERRORS IN THE DATA. THIS ENSURES THAT THE DATA USED FOR TRAINING THE MODEL IS ACCURATE AND RELIABLE.
- DATA PREPROCESSING: THE DATASET NEEDS TO BE PREPROCESSED TO TRANSFORM AND PREPARE THE FEATURES FOR TRAINING. THIS INVOLVES ENCODING CATEGORICAL VARIABLES, SCALING NUMERICAL FEATURES, AND HANDLING OUTLIERS OR SKEWED DISTRIBUTIONS. PREPROCESSING TECHNIQUES LIKE FEATURE SCALING AND NORMALIZATION CAN ENHANCE THE PERFORMANCE OF MACHINE LEARNING MODELS.
- DATA SPLITTING: THE DATASET IS THEN SPLIT INTO TRAINING AND TESTING SETS. THE TRAINING SET IS USED TO TRAIN THE MODEL, WHILE THE TESTING SET IS USED TO EVALUATE ITS PERFORMANCE. IT IS IMPORTANT TO MAINTAIN AN APPROPRIATE BALANCE BETWEEN THE TRAINING AND TESTING DATA TO AVOID OVERFITTING OR UNDERFITTING OF THE MODEL.
- MODEL SELECTION: FOR CREDIT DEFAULT PREDICTION, RANDOM FOREST IS USED WHICH COMBINES MULTIPLE DECISION TREES AND LEVERAGES THEIR COLLECTIVE PREDICTIONS TO MAKE ACCURATE PREDICTIONS. IT HANDLES NON-LINEAR RELATIONSHIPS AND PERFORMS WELL EVEN WITH LARGE AND COMPLEX DATASETS.

- MODEL TRAINING: THE RANDOM FOREST MODEL IS TRAINED ON THE TRAINING DATASET. DURING TRAINING, THE MODEL LEARNS THE PATTERNS AND RELATIONSHIPS WITHIN THE DATA TO PREDICT CREDIT DEFAULTS ACCURATELY. THE ALGORITHM USES AN ENSEMBLE APPROACH, WHERE MULTIPLE DECISION TREES ARE BUILT AND COMBINED TO MAKE PREDICTIONS.
- Model Evaluation: Once the model is trained, it is evaluated using the testing dataset. Metrics such as accuracy, precision, recall, F1 score and AUC are calculated to assess the model's performance. A high recall and AUC indicate the model's ability to identify defaulting borrowers accurately while minimizing false positives and false negatives.
- Model Optimization: The model can be further optimized by fine-tuning hyperparameters, such as the number of trees in the Random Forest or the maximum depth of each tree. This process involves performing cross-validation or grid search to find the optimal combination of hyperparameters that yields the best performance

THE SCOPE OF THE MODEL IS TO ACCURATELY PREDICT WHETHER A BORROWER IS LIKELY TO DEFAULT ON THEIR LOAN. THE MODEL WILL TAKE INTO ACCOUNT VARIOUS FEATURES AND HISTORICAL DATA RELATED TO BORROWERS, SUCH AS LIMIT BALANCE, AGE AND REPAYMENT STATUS. THE MODEL'S PURPOSE IS TO ASSIST IN THE DECISION-MAKING PROCESS OF LOAN APPROVALS BY PROVIDING A PREDICTION OF DEFAULT RISK.

# DATA DESCRIPTION

DEFAULT OF CREDIT CARD CLIENTS PROVIDED BY [UCI Machine Learning]() IS A dataset CONTAINING 24 features, ranging from basic information like sex, Balance limit and repayment statements, of 30000 creditors. The features and their descriptions are listed below

```
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 24 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   LIMIT_BAL                   30000 non-null  int64
 1   SEX                         30000 non-null  int64
 2   EDUCATION                   30000 non-null  int64
 3   MARRIAGE                    30000 non-null  int64
 4   AGE                         30000 non-null  int64
 5   PAY_0                       30000 non-null  int64
 6   PAY_2                       30000 non-null  int64
 7   PAY_3                       30000 non-null  int64
 8   PAY_4                       30000 non-null  int64
 9   PAY_5                       30000 non-null  int64
 10  PAY_6                       30000 non-null  int64
 11  BILL_AMT1                   30000 non-null  int64
 12  BILL_AMT2                   30000 non-null  int64
 13  BILL_AMT3                   30000 non-null  int64
 14  BILL_AMT4                   30000 non-null  int64
 15  BILL_AMT5                   30000 non-null  int64
 16  BILL_AMT6                   30000 non-null  int64
 17  PAY_AMT1                    30000 non-null  int64
 18  PAY_AMT2                    30000 non-null  int64
 19  PAY_AMT3                    30000 non-null  int64
 20  PAY_AMT4                    30000 non-null  int64
 21  PAY_AMT5                    30000 non-null  int64
 22  PAY_AMT6                    30000 non-null  int64
 23  default payment next month  30000 non-null  int64
```

## THE DATASET LOOKS SOMEWHAT LIKE THIS

| | ID | LIMIT_BAL | SEX | EDUCATION | MARRIAGE | AGE | PAY_0 | PAY_2 | PAY_3 | PAY_4 | ... | PAY_AMT5 | PAY_AMT6 | default.payment.next.month |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 20000.0 | 2 | 2 | 1 | 24 | 2 | 2 | -1 | -1 | ... | 0.0 | 0.0 | 1 |
| **1** | 2 | 120000.0 | 2 | 2 | 2 | 26 | -1 | 2 | 0 | 0 | ... | 0.0 | 2000.0 | 1 |
| **2** | 3 | 90000.0 | 2 | 2 | 2 | 34 | 0 | 0 | 0 | 0 | ... | 1000.0 | 5000.0 | 0 |
| **3** | 4 | 50000.0 | 2 | 2 | 1 | 37 | 0 | 0 | 0 | 0 | ... | 1069.0 | 1000.0 | 0 |
| **4** | 5 | 50000.0 | 1 | 2 | 1 | 57 | -1 | 0 | -1 | 0 | ... | 689.0 | 679.0 | 0 |

5 rows × 25 columns

WHERE ['ID','LIMIT_BAL','AGE','BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6'] IE. 16 COLUMNS ARE CONTINUOUS AND ['SEX', 'EDUCATION', 'MARRIAGE','PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6','DEFAULT.PAYMENT.NEXT.MONTH'] IE. 9 COLUMNS ARE DISCRETE .

HENCE Y='DEFAULT.PAYMENT.NEXT.MONTH' IS THE TARGET VARIABLE AND REST X ARE THE PREDICTOR VARIABLES. 1 IN TARGET VARIABLE INDICATES A DEFAULT PAYMENT AND 0 INDICATES A NON-DEFAULTER.

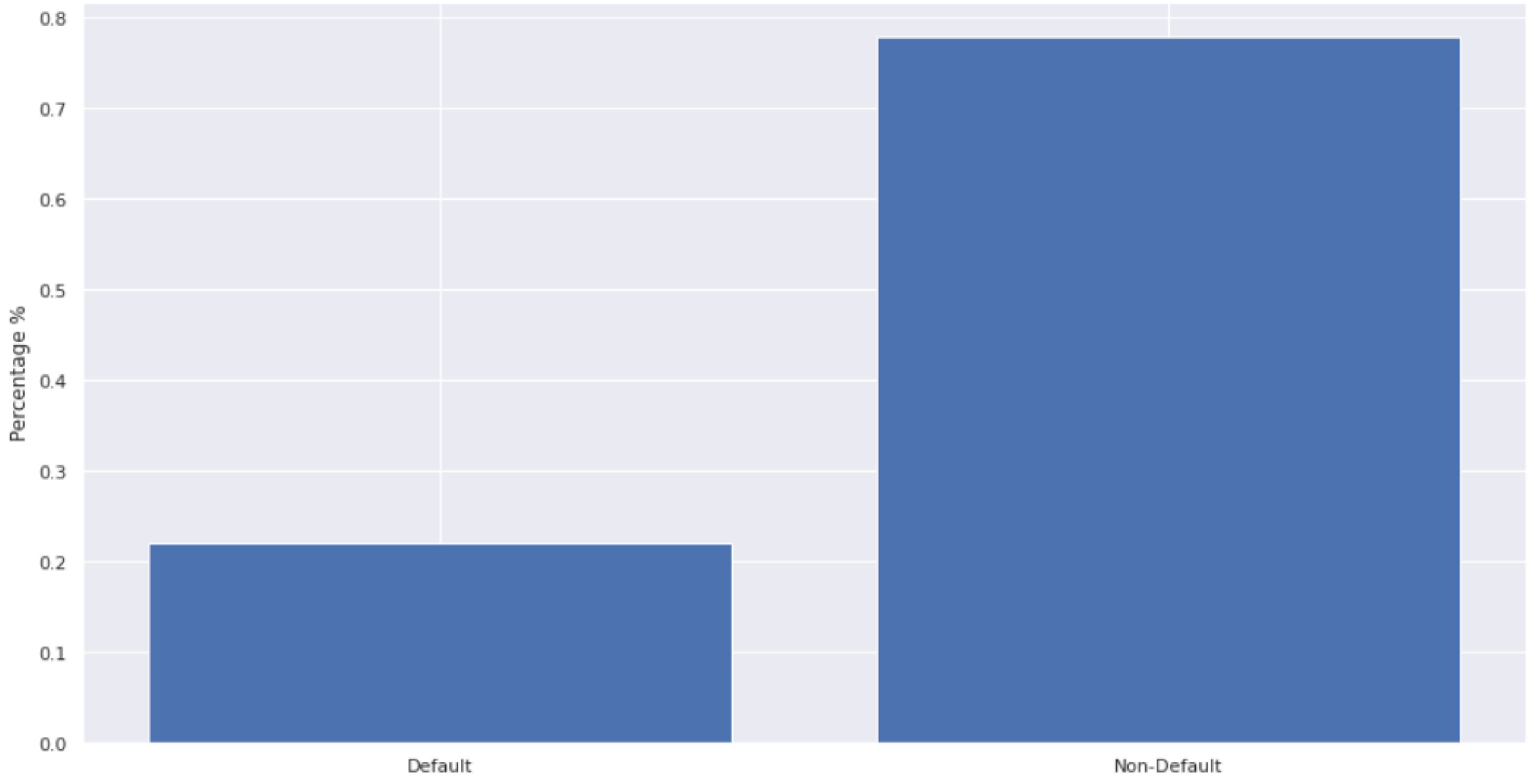
ALL THE COLUMNS ARE NUMERIC IN NATURE AND THERE ARE NO NULL VALUES. THEY HAVE SIMILAR SCALE AS WELL.

# EXPLORATORY DATA ANALYSIS

In [94]:
```
df.isna().sum()
#Data has no null values
```

Out[94]:
```
ID                            0
LIMIT_BAL                     0
SEX                           0
EDUCATION                     0
MARRIAGE                      0
AGE                           0
PAY_0                         0
PAY_2                         0
PAY_3                         0
PAY_4                         0
PAY_5                         0
PAY_6                         0
BILL_AMT1                     0
BILL_AMT2                     0
BILL_AMT3                     0
BILL_AMT4                     0
BILL_AMT5                     0
BILL_AMT6                     0
PAY_AMT1                      0
PAY_AMT2                      0
PAY_AMT3                      0
PAY_AMT4                      0
PAY_AMT5                      0
PAY_AMT6                      0
default.payment.next.month    0
dtype: int64
```

In [ ]:
```
num_data = len(credit_data["default payment next month"])
num_def = len(credit_data[credit_data["default payment next month"]== 1])
percent_def = len(credit_data[credit_data["default payment next month"]== 1])/len(credit_data["default payment next month"])
percent_non_def = 1- percent_def
label = ["Default", "Non-Default"]
percent = [percent_def, percent_non_def]
plt.bar(label, percent)
plt.ylabel('Percentage %')
plt.show()
```

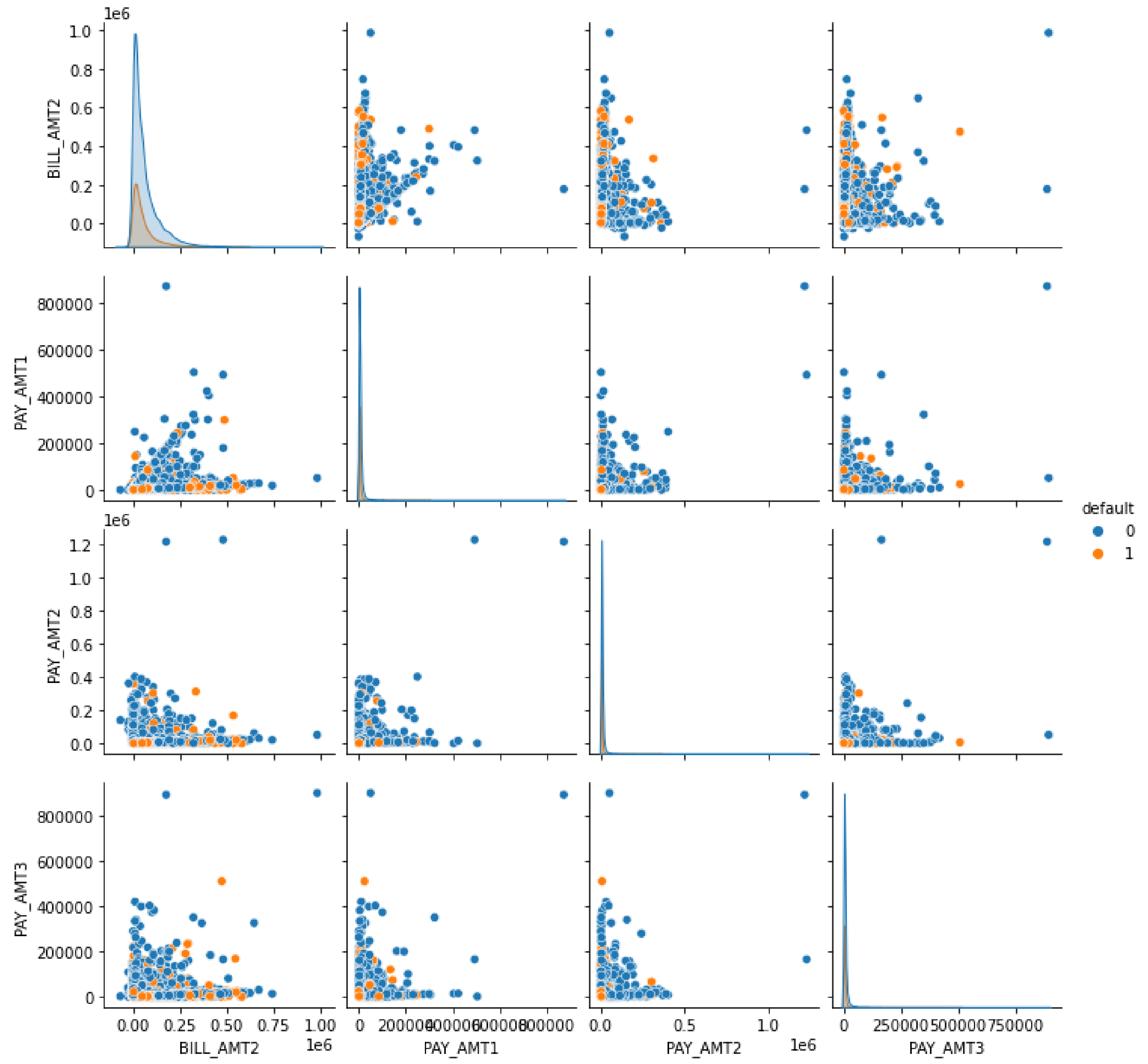


In [100]:
```python
pd.Series(model.feature_importances_,index=Xtrain.columns).sort_values(ascending=False)
```

Out[100]:
```
PAY_1        0.089666
BILL_AMT1    0.079182
BILL_AMT2    0.068533
BILL_AMT3    0.065613
PAY_AMT3     0.060447
PAY_AMT4     0.057392
PAY_AMT5     0.056970
BILL_AMT6    0.056692
BILL_AMT4    0.056523
BILL_AMT5    0.056461
PAY_AMT6     0.052927
PAY_AMT1     0.052702
PAY_AMT2     0.050305
AGE          0.045835
LIMIT_BAL    0.041577
PAY_2        0.026937
EDUCATION    0.014566
PAY_4        0.013459
PAY_3        0.012034
PAY_5        0.011575
MARRIAGE     0.010678
PAY_6        0.010614
SEX          0.009308
dtype: float64
```

In [31]:
```python
sns.pairplot(data=df,hue='default',vars=['BILL_AMT2','PAY_AMT1','PAY_AMT2','PAY_AMT3'])
```

Out[31]:
```
<seaborn.axisgrid.PairGrid at 0x7fdd201dad90>
```



# MODEL BUILDING

THE BASE MODEL LOOKS LIKE THIS –

```
model=ensemble.RandomForestClassifier(random_state=42,n_estimators=200,n_jobs=-1)

model.fit(Xtrain,ytrain)
predtrain=model.predict(Xtrain)
predtest=model.predict(Xtest)

def printscores(act,pred):
    print("Accuracy :",metrics.accuracy_score(act,pred))
    print("Recall :",metrics.recall_score(act,pred))
    print("Precision :",metrics.precision_score(act,pred))
    print("F1 :",metrics.f1_score(act,pred))
    print("AUC :",metrics.roc_auc_score(act,pred))

print("TRAINING METRICS :-")
printscores(ytrain,predtrain)
print("============")
print("TEST METRICS :-")
printscores(ytest,predtest)
```

AND YIELD OVERFIT RESULTS –

```
TRAINING METRICS :-
Accuracy : 0.9998838289962825
Recall : 0.9997676579925651
Precision : 1.0
F1 : 0.9998838154990124
AUC : 0.9998838289962826
============
TEST METRICS :-
Accuracy : 0.7583487940630798
Recall : 0.725417439703154
Precision : 0.7765640516385303
F1 : 0.7501199040767386
AUC : 0.7583487940630798
```

## SUCCESSIVE MODEL RESULT AFTER GRID SEARCH TO FIND THE OPTIMAL COMBINATION OF HYPERPARAMETERS THAT YIELDS THE FOLLOWING PERFORMANCE

```
TRAINING METRICS :-
Accuracy : 0.831454069363544
Recall : 0.433317843866171
Precision : 0.7228682170542635
F1 : 0.5418361417780361
AUC : 0.6918480398455013
============
TEST METRICS :-
Accuracy : 0.815558880102586
Recall : 0.3868274582560297
Precision : 0.6736672051696284
F1 : 0.4914555097230406
AUC : 0.6653659646181567
```

## SUCCESSIVE MODEL RESULT AFTER RECURSIVE FEATURE ELIMINATION AND DATASET BALANCING –

```
TRAINING METRICS :-
Accuracy : 0.7710269516728625
Recall : 0.6758828996282528
Precision : 0.8347202295552367
F1 : 0.7469508280908974
AUC : 0.7710269516728625
============
TEST METRICS :-
Accuracy : 0.7523191094619666
Recall : 0.6716141001855288
Precision : 0.8008849557522124
F1 : 0.7305751765893038
AUC : 0.7523191094619666
```

# FINAL MODEL

CONSIDERING MOST IMPORTANT COLUMNS BASED ON WEIGHT ['LIMIT_BAL', 'AGE', 'PAY_1', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5','PAY_6', 'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4','BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3','PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6'], THE FINAL MODEL RUNS AS FOLLOWS :-

```
model=ensemble.RandomForestClassifier(max_depth=40, min_samples_split=70, n_estimators=150,
                                      random_state=42,n_jobs=-1)
model.fit(Xtrain,ytrain)
predtrain=model.predict(Xtrain)
predtest=model.predict(Xtest)
```

AND YIELD THE FOLLOWING METRICS :-

```
TRAINING METRICS :-
Accuracy : 0.807784911717496
Recall : 0.7651150347779562
Precision : 0.8365019011406845
F1 : 0.7992175492524801
AUC : 0.8077849117174961
============
TEST METRICS :-
Accuracy : 0.7597323600973236
Recall : 0.7311435523114356
Precision : 0.775483870967742
F1 : 0.7526612398246714
AUC : 0.7597323600973237
```

# FUTURE SCOPE

THE FUTURE SCOPE OF THE CREDIT DEFAULT PREDICTION PROJECT HOLDS SEVERAL POSSIBILITIES FOR FURTHER IMPROVEMENT AND EXPANSION -

- INSTEAD OF RELYING SOLELY ON RANDOM FOREST, EXPERIMENTING WITH ENSEMBLE MODELS LIKE GRADIENT BOOSTING OR STACKING CAN POTENTIALLY IMPROVE THE PREDICTIVE ACCURACY OF THE CREDIT DEFAULT MODEL.
- NEURAL NETWORKS, PARTICULARLY DEEP LEARNING ARCHITECTURES SUCH AS MULTI-LAYER PERCEPTRON (MLP), CONVOLUTIONAL NEURAL NETWORKS (CNNS), OR RECURRENT NEURAL NETWORKS (RNNS), CAN BE EXPLORED AS ALTERNATIVE MODELS FOR CREDIT DEFAULT PREDICTION.
- EXPLORING ALTERNATIVE DATA SOURCES, SUCH AS SOCIAL MEDIA PROFILES, TRANSACTION HISTORY, OR ONLINE BEHAVIOR, CAN PROVIDE SUPPLEMENTARY INFORMATION FOR ASSESSING A BORROWER'S CREDIT RISK. INTEGRATING THESE DIVERSE DATA SOURCES AND APPLYING TECHNIQUES LIKE NATURAL LANGUAGE PROCESSING AND SENTIMENT ANALYSIS CAN ENRICH THE MODEL'S UNDERSTANDING AND PREDICTIVE CAPABILITIES.

# THANK YOU