Import Libraries & Load Dataset

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:

```
df = pd.read_csv('CreditCardDefault.csv')
```

In [3]:

```
df.shape
```

Out[3]:

```
(30000, 25)
```

In [4]:

```
df.head()
```

Out[4]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BI
0	1	20000	2	2	1	24	2	2	-1	-1	...	
1	2	120000	2	2	2	26	-1	2	0	0	...	
2	3	90000	2	2	2	34	0	0	0	0	...	
3	4	50000	2	2	1	37	0	0	0	0	...	
4	5	50000	1	2	1	57	-1	0	-1	0	...	

5 rows × 25 columns

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  30000 non-null  int64
 1   LIMIT_BAL           30000 non-null  int64
 2   SEX                 30000 non-null  int64
 3   EDUCATION           30000 non-null  int64
 4   MARRIAGE            30000 non-null  int64
 5   AGE                 30000 non-null  int64
 6   PAY_0               30000 non-null  int64
 7   PAY_2               30000 non-null  int64
 8   PAY_3               30000 non-null  int64
 9   PAY_4               30000 non-null  int64
 10  PAY_5               30000 non-null  int64
 11  PAY_6               30000 non-null  int64
 12  BILL_AMT1           30000 non-null  int64
 13  BILL_AMT2           30000 non-null  int64
 14  BILL_AMT3           30000 non-null  int64
 15  BILL_AMT4           30000 non-null  int64
 16  BILL_AMT5           30000 non-null  int64
 17  BILL_AMT6           30000 non-null  int64
 18  PAY_AMT1            30000 non-null  int64
 19  PAY_AMT2            30000 non-null  int64
 20  PAY_AMT3            30000 non-null  int64
 21  PAY_AMT4            30000 non-null  int64
 22  PAY_AMT5            30000 non-null  int64
 23  PAY_AMT6            30000 non-null  int64
 24  next_month_payment  30000 non-null  int64
dtypes: int64(25)
memory usage: 5.7 MB
```

In [6]:

```
df.describe()
```

Out[6]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
mean	15000.500000	167484.322667	1.603733	1.853133	1.551867	35.485500
std	8660.398374	129747.661567	0.489129	0.790349	0.521970	9.217904
min	1.000000	10000.000000	1.000000	0.000000	0.000000	21.000000
25%	7500.750000	50000.000000	1.000000	1.000000	1.000000	28.000000
50%	15000.500000	140000.000000	2.000000	2.000000	2.000000	34.000000
75%	22500.250000	240000.000000	2.000000	2.000000	2.000000	41.000000
max	30000.000000	1000000.000000	2.000000	6.000000	3.000000	79.000000

8 rows × 25 columns

In [7]:

```
df.columns
```

Out[7]:

```
Index(['ID', 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0',
       'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT
2',
       'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1',
       'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6',
       'next_month_payment'],
      dtype='object')
```

Feature Engineering

In [8]:

```
X = df.drop(labels=['next_month_payment', 'ID'], axis=1)
```

In [9]:

```
y = pd.DataFrame(df.iloc[:,-1])
```

In [10]:

```
from sklearn.model_selection import train_test_split
```

In [11]:

```
from sklearn.preprocessing import StandardScaler
```

In [12]:

```
scaler = StandardScaler()
```

In [13]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, str
```

In [14]:

```
X_train = scaler.fit_transform(X_train)
```

In [15]:

```
X_test = scaler.transform(X_test)
```

Learning Algorithm

In [16]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
```

In [17]:

```
model1 = LogisticRegression()
model2 = SVC()
model3 = DecisionTreeClassifier()
model4 = GaussianNB()
```

In [18]:

```
model1.fit(X_train, y_train)
model2.fit(X_train, y_train)
model3.fit(X_train, y_train)
model4.fit(X_train, y_train)
```

```
C:\Users\15516\anaconda3\lib\site-packages\sklearn\utils\validation.py:760:
DataConversionWarning: A column-vector y was passed when a 1d array was exp
ected. Please change the shape of y to (n_samples, ), for example using rav
el().
  y = column_or_1d(y, warn=True)
C:\Users\15516\anaconda3\lib\site-packages\sklearn\utils\validation.py:760:
DataConversionWarning: A column-vector y was passed when a 1d array was exp
ected. Please change the shape of y to (n_samples, ), for example using rav
el().
  y = column_or_1d(y, warn=True)
C:\Users\15516\anaconda3\lib\site-packages\sklearn\naive_bayes.py:206: Data
ConversionWarning: A column-vector y was passed when a 1d array was expecte
d. Please change the shape of y to (n_samples, ), for example using ravel
().
  y = column_or_1d(y, warn=True)
```

Out[18]:

```
GaussianNB(priors=None, var_smoothing=1e-09)
```

In [19]:

```
y_predm1 = model1.predict(X_test)
y_predm2 = model2.predict(X_test)
y_predm3 = model3.predict(X_test)
y_predm4 = model4.predict(X_test)
```

In [20]:

```
from sklearn.metrics import accuracy_score
```

In [21]:

```
accuracy1 = accuracy_score(y_test, y_predm1)
accuracy2 = accuracy_score(y_test, y_predm2)
accuracy3 = accuracy_score(y_test, y_predm3)
accuracy4 = accuracy_score(y_test, y_predm4)
```

In [22]:

```
print(accuracy1, accuracy2, accuracy3, accuracy4)
```

```
0.8101111111111111 0.8197777777777778 0.7298888888888889 0.6842222222222222
```

Hard Voting Classifier

In [23]:

```
from sklearn.ensemble import VotingClassifier
```

In [24]:

```
estimators = [('log', model1),('svc',model2),('tree',model3),('naive',model4)
hard_classifier = VotingClassifier(estimators, voting='hard')
```

In [25]:

```
hard_classifier.fit(X_train, y_train)
```

```
C:\Users\15516\anaconda3\lib\site-packages\sklearn\preprocessing\_label.py:
235: DataConversionWarning: A column-vector y was passed when a 1d array wa
s expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)
C:\Users\15516\anaconda3\lib\site-packages\sklearn\preprocessing\_label.py:
268: DataConversionWarning: A column-vector y was passed when a 1d array wa
s expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)
```

Out[25]:

```
VotingClassifier(estimators=[('log',
                              LogisticRegression(C=1.0, class_weight=None,
                                                 dual=False, fit_intercept=
True,
                                                 intercept_scaling=1,
                                                 l1_ratio=None, max_iter=10
0,
                                                 multi_class='auto',
                                                 n_jobs=None, penalty='l2',
                                                 random_state=None,
                                                 solver='lbfgs', tol=0.000
1,
                                                 verbose=0, warm_start=Fals
e)),
                             ('svc',
                              SVC(C=1.0, break_ties=False, cache_size=200,
                                  class_weight=None, coef0=0.0,...
                                                     criterion='gini',
                                                     max_depth=None,
                                                     max_features=None,
                                                     max_leaf_nodes=None,
                                                     min_impurity_decrease=
0.0,
                                                     min_impurity_split=Non
e,
                                                     min_samples_leaf=1,
                                                     min_samples_split=2,
                                                     min_weight_fraction_le
af=0.0,
                                                     presort='deprecated',
                                                     random_state=None,
                                                     splitter='best')),
                             ('naive',
                              GaussianNB(priors=None, var_smoothing=1e-0
9))],
                 flatten_transform=True, n_jobs=None, voting='hard',
                 weights=None)
```

In [26]:

```
y_pred_hard = hard_classifier.predict(X_test)
```

In [27]:

```
accuracy_score(y_test, y_pred_hard)
```

Out[27]:

```
0.8177777777777778
```

Soft Voting Classifier

In [28]:

```
from sklearn.ensemble import VotingClassifier
```

In [29]:

```
model2_soft = SVC(probability=True)
estimators = [('log', model1),('svc',model2_soft),('tree',model3),('naive',mo
hard_classifier = VotingClassifier(estimators, voting='soft')
```

In [30]:

```
hard_classifier.fit(X_train, y_train)
```

```
C:\Users\15516\anaconda3\lib\site-packages\sklearn\preprocessing\_label.py:
235: DataConversionWarning: A column-vector y was passed when a 1d array wa
s expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)
C:\Users\15516\anaconda3\lib\site-packages\sklearn\preprocessing\_label.py:
268: DataConversionWarning: A column-vector y was passed when a 1d array wa
s expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)
```

Out[30]:

```
VotingClassifier(estimators=[('log',
                              LogisticRegression(C=1.0, class_weight=None,
                                                 dual=False, fit_intercept=
True,
                                                 intercept_scaling=1,
                                                 l1_ratio=None, max_iter=10
0,
                                                 multi_class='auto',
                                                 n_jobs=None, penalty='l2',
                                                 random_state=None,
                                                 solver='lbfgs', tol=0.000
1,
                                                 verbose=0, warm_start=Fals
e)),
                             ('svc',
                              SVC(C=1.0, break_ties=False, cache_size=200,
                                  class_weight=None, coef0=0.0,...
                                                     criterion='gini',
                                                     max_depth=None,
                                                     max_features=None,
                                                     max_leaf_nodes=None,
                                                     min_impurity_decrease=
0.0,
                                                     min_impurity_split=Non
e,
                                                     min_samples_leaf=1,
                                                     min_samples_split=2,
                                                     min_weight_fraction_le
af=0.0,
                                                     presort='deprecated',
                                                     random_state=None,
                                                     splitter='best')),
                             ('naive',
                              GaussianNB(priors=None, var_smoothing=1e-0
9))],
                 flatten_transform=True, n_jobs=None, voting='soft',
                 weights=None)
```

```
In [31]:  y_pred_hard = hard_classifier.predict(X_test)
```

```
In [32]:  accuracy_score(y_test, y_pred_hard)
```

```
Out[32]: 0.8058888888888889
```

```
In [ ]:
```