Linear Discriminant

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns

#input balance and income (Float)
#output default (Categorical data)
```

In [2]:
```
tbl=pd.read_excel("Default1.xlsx")
tbl
```

Out[2]:

	Unnamed: 0	default	student	balance	income
0	1	No	Yes	729.526495	44361.625074
1	2	No	Yes	817.180407	12106.134700
2	3	No	No	1073.549164	31767.138947
3	4	No	No	529.250605	35704.493935
4	5	No	No	785.655883	38463.495879
...	...	...	...	...	...
9995	9996	No	No	711.555020	52992.378914
9996	9997	No	No	757.962918	19660.721768
9997	9998	No	No	845.411989	58636.156984
9998	9999	No	No	1569.009053	36669.112365
9999	10000	No	Yes	200.922183	16862.952321

10000 rows × 5 columns

In [3]:
```
tbl.head()
```

Out[3]:

	Unnamed: 0	default	student	balance	income
0	1	No	Yes	729.526495	44361.625074
1	2	No	Yes	817.180407	12106.134700
2	3	No	No	1073.549164	31767.138947
3	4	No	No	529.250605	35704.493935
4	5	No	No	785.655883	38463.495879

In [4]:
```
tbl.shape
```

Out[4]:
```
(10000, 5)
```

In [5]:
```
tbl.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  10000 non-null  int64  
 1   default     10000 non-null  object 
 2   student     10000 non-null  object 
 3   balance     10000 non-null  float64
 4   income      10000 non-null  float64
dtypes: float64(2), int64(1), object(2)
memory usage: 390.8+ KB
```

In [6]:
```
tbl.describe()
```

Out[6]:

	Unnamed: 0	balance	income
count	10000.00000	10000.000000	10000.000000
mean	5000.50000	835.374886	33516.981876
std	2886.89568	483.714985	13336.639563
min	1.00000	0.000000	771.967729
25%	2500.75000	481.731105	21340.462903
50%	5000.50000	823.636973	34552.644802
75%	7500.25000	1166.308386	43807.729272
max	10000.00000	2654.322576	73554.233495

In [7]:
```
tbl.isnull().sum
```

Out[7]:
```
<bound method NDFrame._add_numeric_operations.<locals>.sum of       Unname
d: 0  default  student  balance  income
0          False    False    False    False   False
1          False    False    False    False   False
2          False    False    False    False   False
3          False    False    False    False   False
4          False    False    False    False   False
...          ...      ...      ...      ...     ...
9995       False    False    False    False   False
9996       False    False    False    False   False
9997       False    False    False    False   False
9998       False    False    False    False   False
9999       False    False    False    False   False

[10000 rows x 5 columns]>
```

Statistical Analysis

In [8]:
```
#unique => default value 2 that means Yes or No

tbl.describe(include='all')
```

Out[8]:

	Unnamed: 0	default	student	balance	income
count	10000.00000	10000	10000	10000.000000	10000.000000
unique	NaN	2	2	NaN	NaN
top	NaN	No	No	NaN	NaN
freq	NaN	9667	7055	NaN	NaN
mean	5000.50000	NaN	NaN	835.374886	33516.981876
std	2886.89568	NaN	NaN	483.714985	13336.639563
min	1.00000	NaN	NaN	0.000000	771.967729
25%	2500.75000	NaN	NaN	481.731105	21340.462903
50%	5000.50000	NaN	NaN	823.636973	34552.644802
75%	7500.25000	NaN	NaN	1166.308386	43807.729272
max	10000.00000	NaN	NaN	2654.322576	73554.233495

Analysis of Zero Values in Predictors

In [9]:
```
#499 rows of the balance variable contain is 0 value

#499 default

(tbl.balance==0).sum(axis=0)
```

Out[9]:
```
499
```

Categorical Variable Analysis

In [10]:
```
tbl.student.value_counts()
```

Out[10]:
```
student
No     7055
Yes    2945
Name: count, dtype: int64
```

Response Variable Analysis

In [11]:
```
tbl.default.value_counts()
```

Out[11]:
```
default
No     9667
Yes     333
Name: count, dtype: int64
```

Encode Categorical Variables

In [12]:
```
tbl['default2']=tbl.default.factorize()[0]

tbl['student2']=tbl.student.factorize()[0]

tbl.head(3)
```

Out[12]:

	Unnamed: 0	default	student	balance	income	default2	student2
0	1	No	Yes	729.526495	44361.625074	0	0
1	2	No	Yes	817.180407	12106.134700	0	0
2	3	No	No	1073.549164	31767.138947	0	1

Graphical Representation

In [24]:
```
tbl_dfno=tbl[tbl.default2==0].sample(frac=0.15)

tbl_dfyes=tbl[tbl.default2==1]

tbl_df=tbl_dfno._append(tbl_dfyes)
```

In [25]:
```
tbl_df
```

Out[25]:

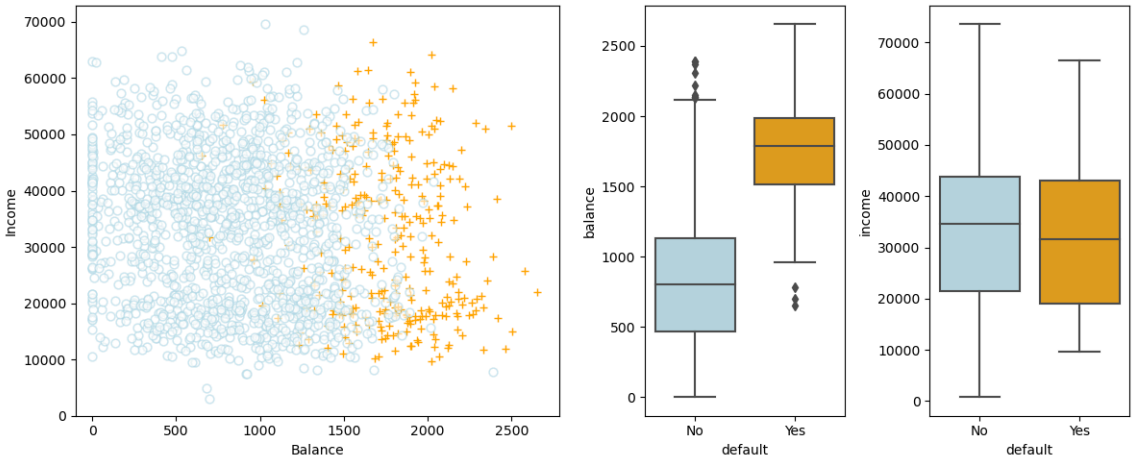
	Unnamed: 0	default	student	balance	income	default2	student2
4157	4158	No	Yes	1084.044084	13680.113299	0	0
7645	7646	No	Yes	207.925970	19078.891905	0	0
767	768	No	Yes	1029.681549	15977.321139	0	0
2249	2250	No	No	988.667818	28486.339752	0	1
7149	7150	No	No	561.989745	40446.008780	0	1
...	...	...	...	...	...	...	...
9912	9913	Yes	No	2148.898454	44309.917173	1	1
9921	9922	Yes	Yes	1627.898323	17546.997016	1	0
9949	9950	Yes	No	1750.253150	51578.940163	1	1
9951	9952	Yes	No	1515.606239	48688.512086	1	1
9978	9979	Yes	No	2202.462395	47287.257108	1	1

1783 rows × 7 columns

In [26]:
```
fig=plt.figure(figsize=(12,5))
gs=mpl.gridspec.GridSpec(1,4)
ax1=plt.subplot(gs[0,:2])
ax2=plt.subplot(gs[0,2:3])
ax3=plt.subplot(gs[0,3:4])
ax1.scatter(tbl_df[tbl_df.default=='Yes'].balance,
            tbl_df[tbl_df.default=='Yes'].income,s=40,c='orange',marker='+
ax1.scatter(tbl_df[tbl_df.default=='No'].balance,
            tbl_df[tbl_df.default=='No'].income,s=40,marker='o',linewidths
            edgecolors='lightblue',facecolors='white',alpha=.6)


#--------------------------------------------------------------------

ax1.set_ylim(ymin=0)
ax1.set_ylabel('Income')
ax1.set_xlim(xmin=-100)
ax1.set_xlabel('Balance')
c_palette={'No':'lightblue','Yes':'Orange'}
sns.boxplot(x='default',y='balance',data=tbl,orient='v',ax=ax2,palette=c_pa
sns.boxplot(x='default',y='income',data=tbl,orient='v',ax=ax3,palette=c_pal
gs.tight_layout(plt.gcf())
```



Linear Discriminant Analysis

50% Threshold

In [30]:
```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
```

In [31]:
```
X=tbl[['balance','income','student2']]

y=tbl.default2

lda=LinearDiscriminantAnalysis(solver='svd')

y_pred=lda.fit(X,y).predict(X)

tbl_df=pd.DataFrame({'True default status':y,'Predicted default status':y_p

tbl_df.replace(to_replace={0:'No',1:'Yes'},inplace=True)

tbl_df.groupby(['Predicted default status','True default status']).size().u
```

Out[31]:

True default status	No	Yes
Predicted default status		
No	9645	254
Yes	22	79

20% Threshold

In [32]:
```
decision_prob=0.2

y_prob=lda.fit(X,y).predict_proba(X)

tbl_df=pd.DataFrame({'True default status':y,'Predicted default status':y_p

tbl_df.replace(to_replace={0:'No',1:'Yes','True':'Yes','False':'No'},inplac

tbl_df.groupby(['Predicted default status','True default status']).size().u
```

Out[32]:

True default status	No	Yes
Predicted default status		
False	9435	140
True	232	193

In [ ]:
```

```