Emmanuel Pedernal

1. Generate, discuss and interpret the results on the tests of Equality of Class Means and Equality of Covariance Matrices.

Tests of Equality of Class Means

	F	df1	df2	p
Age	0.298	1	4998	0.585
Experience	0.275	1	4998	0.600
Income	1688.005	1	4998	< .001
Family	18.893	1	4998	< .001
CCAvg	777.413	1	4998	< .001
Education	95.206	1	4998	< .001
Mortgage	102.994	1	4998	< .001
Securities Account	2.410	1	4998	0.121
CD Account	555.829	1	4998	< .001
Online	0.197	1	4998	0.657
CreditCard	0.039	1	4998	0.843

Note. The null hypothesis specifies equal class means.

The following features; Income, Family, CCAvg, Education, Mortgage, and CD Account are all strong features to use when we classify if person will get a loan or not while Age, Experience, CreditCard are not strong indicator with this dataset.

Class Means in Training Data

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard
0	45.390	20.147	66.411	2.363	1.747	1.833	52.036	0.101	0.037	0.599	0.299
1	44.767	19.534	143.829	2.598	3.888	2.236	97.249	0.132	0.290	0.619	0.288

For Income, the average income from class 1 (144 income) has a huge difference from class 0 (66.2) who didn't get the loan. As is in real life where banks are more inclined to give loan to higher income than not.

A bit higher family attainment is approved of loan because a higher number of family members mean increase in capacity compared to lower number families.

People with higher average CC spending are more prone to be approved than lower average spending just like in real life because they have higher credit rating.

Education in general could not mean that a higher educational attainment equates to higher income but the individual has a capacity for to earn which are more prone to have lone approved.

High mortgage is approved more, the bank has assurance that they have the collateral which reduces risk of losing money from not paying the loan.

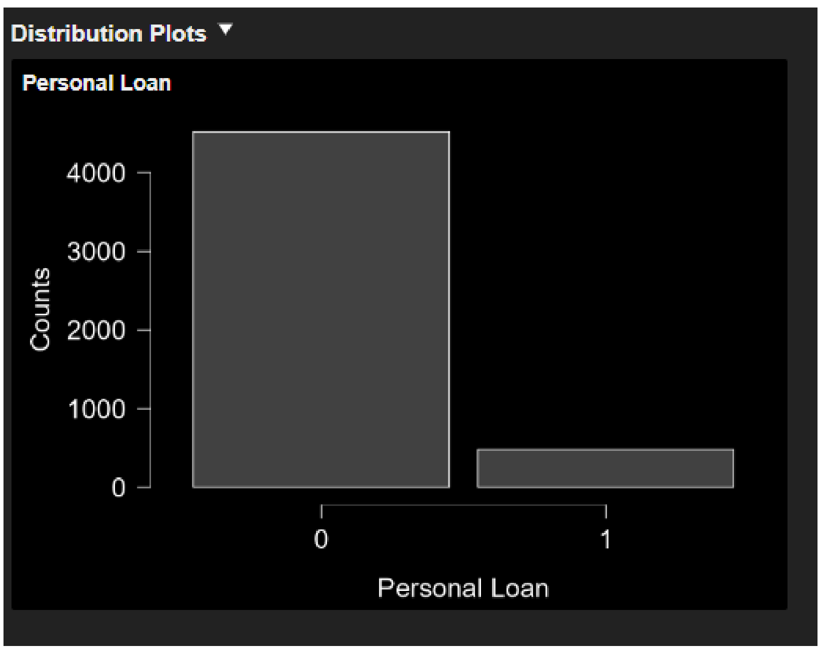
Higher CD account individuals who have account(s) are approved more because they already have financial profile from financial institutions this could mean possible existing credit ratings, bank accounts, mortgages, or collateral to avert risk.

Tests of Equality of Covariance Matrices

	χ^2	df	p
Box's M	2128.204	66	< .001

Note. The null hypothesis specifies equal covariance matrices.

Having a p val less than 0.05 we reject null hypothesis where the covariance matrices are equal. The dataset used are not equal between the not approved and approved loan. Our model leans more on a certain group than (people who are NOT approved).



2. Based on the previous result generated, what can you say about the relationships among the variables (Pooled within-Class Matrices Correlations)?

Pooled Within-Class Matrices Correlations

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard
Age	1.000										
Experience	1.000	1.000									
Income	−0.254	−0.238	1.000								
Family	−0.181	−0.187	−0.488	1.000							
CCAvg	−0.269	−0.258	0.856	−0.407	1.000						
Education	0.020	−0.003	−0.629	0.148	−0.529	1.000					
Mortgage	−0.192	−0.186	0.234	−0.166	0.096	−0.213	1.000				
Securities Account	−0.179	−0.177	−0.128	−0.039	−0.110	−0.096	−0.146	1.000			
CD Account	−0.238	−0.232	−0.114	−0.125	−0.128	−0.177	−0.111	0.457	1.000		
Online	−0.138	−0.135	−0.082	−0.057	−0.121	−0.098	−0.140	−0.069	0.178	1.000	
CreditCard	−0.155	−0.151	−0.114	−0.053	−0.133	−0.088	−0.142	−0.099	0.390	−0.090	1.000

Focusing on important pairings we have;

Income and CCavg – is the only pair that has strong positive correlation – usually we either drop this BUT because income and CCavg is important (answer no.1) we retain this instead.

Income/ CCavg – Like in real life Individuals who are high earners tend to have more CC balance

Income/Education – Higher income does not mean higher education, this could mean either old people who managed to attain wealth vs newly employed OR business owner's vs employees.

CCAvg/Education – Suggests lower CC spending the higher education attainment could mean individual have high financial knowledge than those who don't or low salary from employees tend to not use CC while people who other means who aren't college graduates (or above) use CC more.

Other pairings have small correlation, variables are weak enough to influence the groupings.

3. Generate and explain the class proportions.

Class Proportions

	Data Set	Training Set	Test Set
0	0.904	0.906	0.894
1	0.096	0.093	0.106

Our dataset is split in 80/20 and leans more on individuals that are not approved than approved. Class representation between the data set to train/test are consistent.

Class Proportions	Data Set	Training Set	Test Set
0	4,520	3,624	896
1	480	376	104

4. Generate and explain briefly and precisely the confusion matrix.

Confusion Matrix

		Predicted	
		0	1
Observed	0	879	15
	1	44	62

Class	Observed
0	The model correctly predicted 879 or correct 98% of the time for individuals who did not qualify for loan while approving erroneously 2% of the time
1	The model correctly predicted 62 (58.5%) out of 106 in approving an individual loan while incorrectly denying 44 (41.5%) of individuals applying for loan

5. Generate the model measurement. Explain the following values: Accuracy, Precision, and Recall.

Model Performance Metrics

	0	1	Average / Total
Support	894	106	1000
Accuracy	0.941	0.941	0.941
Precision (Positive Predictive Value)	0.952	0.805	0.937
Recall (True Positive Rate)	0.983	0.585	0.941
False Positive Rate	0.415	0.017	0.216
False Discovery Rate	0.048	0.195	0.121
F1 Score	0.968	0.678	0.937
Matthews Correlation Coefficient			NaN
Area Under Curve (AUC)	0.951	0.951	0.951
Negative Predictive Value	0.805	0.952	0.879
True Negative Rate	0.585	0.983	0.784
False Negative Rate	0.017	0.415	0.216
False Omission Rate	0.195	0.048	0.121
Threat Score	8.534	0.838	4.686
Statistical Parity	0.923	0.077	1.000

Note. All metrics are calculated for every class against all other classes.

Metric	Observation
Accuracy – (% of correct prediction)	(TP + TN)/ Total = 62 + 879 / 1000 = 94.1% Of (2) classes the model gets the correct prediction 94.1% of the time
Precision – Out of all approved predictions, how many are truly approved?	TP/ TP + FP = 62 / 62 + 15 = 80.5%

Sa lahat ng tama ilang yung actual na tama Increase if we want to avoid mistakenly handing out loans to individuals who shouldn't get it	If the model predicts that someone is eligible to get a loan it is correct 80.5% of the time.
Recall – Out of all people who truly deserve approval, how many did the model actually catch? Sa lahat ng class 1 talaga ilan ang tama Important if we don't want to miss individuals who are qualified to get the loan	TP/ TP + FN = 62 / 62 + 44 = 58.5% Of the actual 106 who are eligible for the model only got 58.5% correctly, this could be attributed to our data set leaning to the 0 class (not approved)

6. Based on the result of linear discriminant function, which among the variable has the highest influence or contribution in the approval of loan application? Explain.

Linear Discriminant Coefficients

	LD1
(Constant)	−0.009
Age	−0.416
Experience	0.468
Income	0.970
Family	0.265
CCAvg	0.140
Education	0.472
Mortgage	0.036
Securities Account	−0.138
CD Account	0.536
Online	−0.082
CreditCard	−0.151

DA_score=−0.009+(−0.416·Age)+(0.468·Experience)+(0.970·Income)+(0.265·Family)+(0.140·CCAvg)+(0.472·Education)+(0.036·Mortgage)+(−0.138·Securities)+(0.536·CD Account)+(−0.082·Online)+(−0.151·CreditCard)

Income has the highest coeff with 0.97, according to our model the feature "income" has the greatest weight when we are predicting if we should grant a loan or not. An increase in income means an increase of 0. 97 points to be granted a loan. Like in real life people who have high income are more likely to be granted loan.

7. If you are to talk to the bank officer, give possible recommendations on how they will decide on whether to approve a loan application or not.

As a consultant I would recommend to focus with the following features in order; Income, CD account and education/experience level. Since high income less risk of none payment, as well as having multiple CD accounts means the individual knows that banking system and have previous records as for education/experience, in general higher education/experience does not really mean higher pay, but should still be taken into consideration if an individual is borderline for both Income and CD account.

Age (Highest negative) of the applicant should be taken cautiously since there's a risk with higher age but should be considered ethics wise.

Model wise, with an accuracy of 94.1% the model performs well but we also missed 41.5% of deserving applicants (Recall 58.5%) we could optimize our model more by either increasing the weight or reducing the weight of features or increase the number of data and balance our dataset. Lastly, we need to update our model regularly (2 -3 times a year) since we are observing economic activities of an individual few shifts in its features could result to greater loss or gain to an individual.