The Hang Seng University of Hong Kong
2022-2023 Semester 2

MSIM4311 :
Business Intelligence and Data Mining

Group Project Report

Class: L03
Group Members:

LAU Hei Yee	s197671
CHAN Ka Lam	s198167
CHOI Katarina Kai Ru	s198179
HO Yan Wa	s198184
LAM Yuet Ying	s198193

Table of Contents

Question 1 – Data Analytics

Data set link:
https://www.kaggle.com/datasets/mirzahasnine/loan-data-set?resource=download&select=loan_train.csv
Google share drive link: https://drive.google.com/drive/folders/0ALBDfRxfF3MxUk9PVA
Google share folder link:
https://drive.google.com/drive/folders/1nWux0TKiZqz_hGwNvTUqb4VrvLHuT3VY?usp=sharing

Introduction

In recent years, the culture of lending is growing and the industry is becoming larger. As we all know, loans can be a lifesaver and more and more people are borrowing loans for things such as buying a new home, going for a world tour, or completing higher education from the best colleges, which make their dreams come true. However, different categories of people will show the characteristics of people borrowing loans. Therefore, finding out what element will influence the loan amount borrowed is our objective in this project.

In our dataset, it contains 615 records and 11 variables. We have also found out the mean, maximum, minimum, and standard deviation of the loan amount in the dataset, which is shown in the following table:

Variable(s)	Detail
Gender	Female / Male
Married	Yes/No
Dependents	Number of Family member
Education	Graduate / Not graduate
Self_employed	Yes/ No
Applicant_Income	Applicant income
Coapplicant_Income	Coapplicant
Loan_Amount	Borrowed amount
Term	Borrowed period
Credit_histroy	1 / 0
Area	Living area
Status	Repay or not

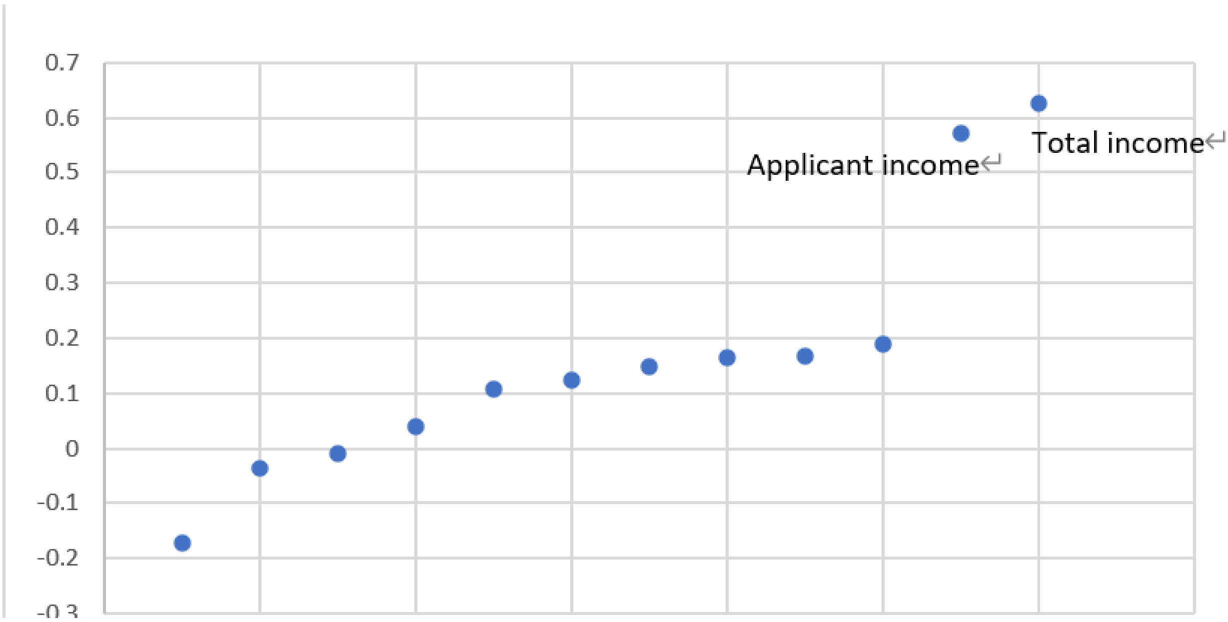
	Amount
Mean	1414042.35
Maximum	70000000
Minimum	900000
Standard Deviation	8815682.47

Data cleaning

The data cleaning and transformation is the process to convert raw data source into another format.It is needed before we do the further analyses for the data set. There are total 4 steps to execute in this part. Firstly, we will reomove the reocords if the loan amount is equal to 0 since it is useless for our analysis. Secondly, we will convert all the blank value into missing value. Thirdly, convert all of the character variables to numeric variables. It included the data from gender, married, dependents, education, self employee, area and status total 7 variables because it would be used in arithmetic calculations. Last but not least, we will add two new variables. One is add the applicant income and coapplicant income and result it as total income. Second is debt to income ratio by dividing the loan amount by total income. These two variables will help us in further analyses.

Correlation

Before go for the regression part, we are going to find out which variables are linerly related to our objective which is loan amount. From the result, we can see that applicant income and total income have higher correlation coefficient than other with 0.57 and 0.62 respentivey. It means these two variables have a strong relationship with loan amount.



Pearson 相關係數
觀測值數目

	GenderN	MarriedN	DependentsN	EducationN	Self_EmployedN	Applicant_Income	Coapplicant_Income	Term	Credit_History	StatusN	TotalIncome	Debt_ratio
Loan_Amount	0.10698 580	0.14919 591	0.16387 580	-0.17113 593	0.12388 562	0.57091 593	0.18853 593	0.03946 579	-0.00838 544	-0.03726 593	0.62459 593	0.16667 593

Regression

From the perspective of the bank, the company would like to know the customer repayment ability and the borrow amount in order to reduce the risk. So, 2 regression models are constructed for predicting the loan amount and the status of prediction. Before constructing the model, feature selection is managed to make sure that the models only have related variables. In this case, a stepwise selection method is adopted. The p-value limit is changed from smaller than 0.15 to smaller than 0.05 to make sure that all variables have a strong relationship with the dependent variable.

In terms of forecasting loan amounts, multiple regression models are adopted. The following images show the result of stepwise selection and regression.

Stepwise Selection: Step 7

Variable GenderN Removed: R-Square = 0.5444 and C(p) = 2.3012

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.690108E16	3.380216E15	113.50	<.0001
Error	475	1.414589E16	2.978082E13		
Corrected Total	480	3.104697E16			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-4478184	1012383	5.827081E14	19.57	<.0001
TotalIncome	9.64849	0.46454	1.284711E16	431.39	<.0001
Debt_ratio	474018	32089	6.498563E15	218.21	<.0001
MarriedN	1780861	523762	3.442932E14	11.56	0.0007
EducationN	-1640895	628264	2.031487E14	6.82	0.0093
Self_EmployedN	1565530	734952	1.35127E14	4.54	0.0337

Regression for loan amount

The REG Procedure
Model: MODEL1
Dependent Variable: Loan_Amount

Number of Observations Read	593
Number of Observations Used	560
Number of Observations with Missing Values	33

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.400642E16	4.801284E15	161.87	<.0001
Error	554	1.643226E16	2.966112E13		
Corrected Total	559	4.043868E16			

Root MSE	5446203	R-Square	0.5936
Dependent Mean	14596964	Adj R-Sq	0.5900
Coeff Var	37.31052		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5702975	951287	-6.00	<.0001
TotalIncome	1	10.80571	0.42166	25.63	<.0001
Debt_ratio	1	498456	30146	16.53	<.0001
MarriedN	1	1591706	486694	3.27	0.0011
EducationN	1	-1418740	567260	-2.50	0.0127
Self_EmployedN	1	1470611	667948	2.20	0.0281

After 7 steps in the stepwise selection, 5 features are chosen. They are Totallncome, Debt_ratio, MarriedN, EducationN and Self_EmployedN. By running the SAS code, an ANOVA table and parameter table is generated. From the ANOVA table, the p-value is smaller than 0.0001. This implies that we have enough evidence to reject the null hypothesis which is there is no relationship between features and loan amount. Also, the r square is almost 0.6 which means the strength of the prediction equation is strong. Based on these evidence, it is believed that this model is a predictive model.

Following is the regression equation of the loan amount:

$$\begin{aligned} Loan\ amount = {} & -5702975 + 10.80571\,(Total\ Income) + 498456\,(Debt\_ratio) \\ & + 1591706(MarriedN) - 1418740\,(EducationN) \\ & + 1470611\,(Self\_EmployedN) + Residual \end{aligned}$$

In terms of repayment status of customers, logistic regression is used as status is a binary variable. The following images show the result of stepwise selection and regression.

Step 4. Effect MarriedN is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	595.787	461.839
SC	599.962	474.367
-2 Log L	593.787	455.839

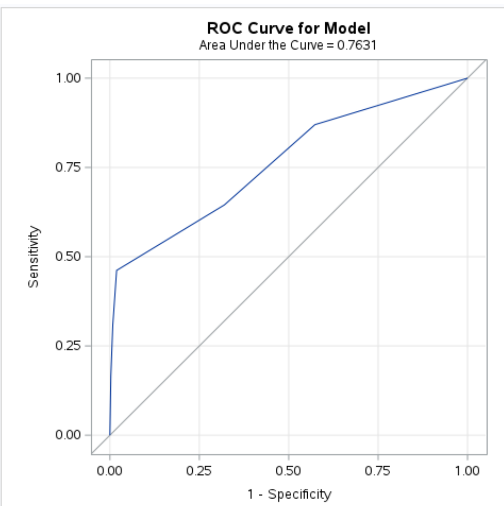
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	137.9472	2	<.0001
Score	141.0658	2	<.0001
Wald	76.4325	2	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
19.9782	10	0.0295

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	2.8368	0.4270	44.1411	<.0001	17.061
Credit_History	1	-3.8659	0.4166	86.1244	<.0001	0.021
AreaN	1	-0.2743	0.0910	9.0849	0.0026	0.760

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	67.8	Somers' D	0.526
Percent Discordant	15.2	Gamma	0.634
Percent Tied	17.1	Tau-a	0.226
Pairs	63375	c	0.763

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Credit_History	1.0000	0.021	0.009	0.047
AreaN	1.0000	0.760	0.636	0.909



Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
6.4832	3	0.0903

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.120	169	0	375	0	31.1	100.0	0.0	31.1	.
0.140	147	160	215	22	56.4	87.0	42.7	40.6	87.9
0.160	147	160	215	22	56.4	87.0	42.7	40.6	87.9
0.180	147	160	215	22	56.4	87.0	42.7	40.6	87.9
0.200	147	160	215	22	56.4	87.0	42.7	40.6	87.9
0.220	109	255	120	60	66.9	64.5	68.0	47.6	81.0
0.240	109	255	120	60	66.9	64.5	68.0	47.6	81.0
0.260	78	255	120	91	61.2	46.2	68.0	39.4	73.7
0.280	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.300	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.320	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.340	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.360	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.380	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.400	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.420	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.440	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.460	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.480	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.500	78	368	7	91	82.0	46.2	98.1	91.8	80.2
0.520	78	368	7	91	82.0	46.2	98.1	91.8	80.2

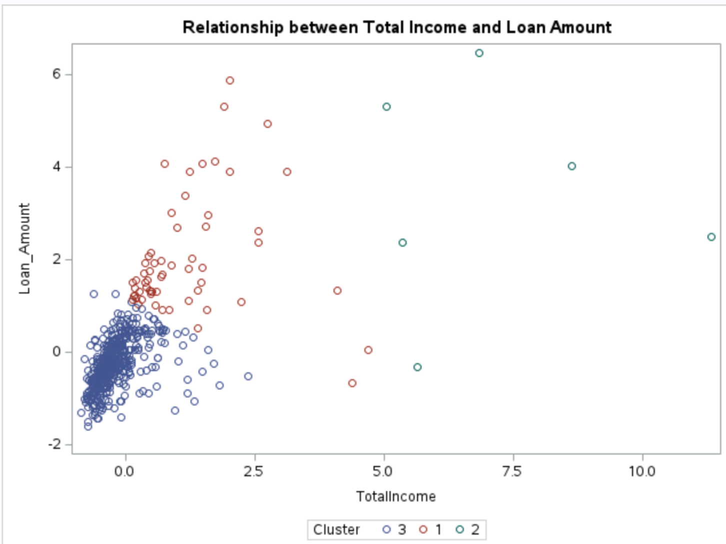
After 4 steps in the selection, Credit_History and AreaN are chosen for the model. Both parameters have p-value smaller than 0.05 which means that they are good predictors. The data in the fit test is 0.0903 and greater than 0.05 which reflect that the model is a good fitting model. Moreover, take 0.50 as the cut-off point in the classification table, there is 46.2% of sensitivity and 98.1% of specificity. 82% of the correct cases are classified using this cut-off point. In addition, the ROC score is 0.7631 which is close to 1 and shows the predictive strength of the model is strong. According to these information, it is believed that this model is a predictive model. Following is the regression equation of the status:

$$Status == 2.6073 - 3.5809\,(Credit\_History) - 0.2704\,(AreaN) + Residual$$

Clustering and Recommendations

In the following part, we aim to group different sets of objects into classes of similar characteristics.

For the first analysis, we hope to find out the relationship between loan amount and total income. From the scatter plot, we can see there are 3 clusters, the red dots cluster 1, blue dots cluster 2 and green dots cluster 3.



Now, have a deeper analysis by looking at the means table. From this table, we can identify those specific characteristics of a particular cluster.

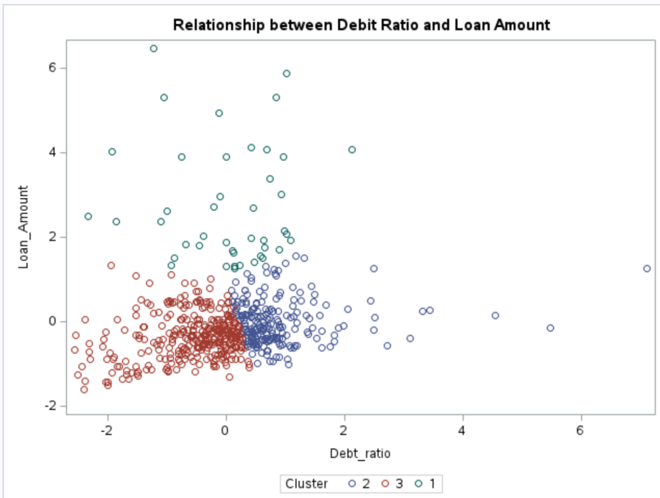
Cluster Means		
Cluster	**TotalIncome**	**Loan_Amount**
1	1.162198552	2.001296139
2	7.147265403	3.394117943
3	-0.211085809	-0.262199204

Therefore, we can understand that, cluster 3, the blue dots are representing the low income group, while loaning the least amount of money. Their mean of total income and loan amount are -0.21& -0.26. Revealed that their mean are below average.

Then, cluster 1, the relatively average income group (1.162) with average loan amount (2.001). Although this cluster has a positive mean, compared with cluster 2 , the high income group (7.147) with the most loan amount (3.394), their mean is much less. As a result, this analysis can show that cluster 2 are the high income group while loaning the most money.

For the second analysis, we aim to find out the relationship between loan amount and debt ratio. Make a quick review, debt ratio means loan amount divided by the total income, we aim to find out the ability for each group to do repayment.

From the scatter plot, we can see 3 clearly identical groups.



Move on to the means table.

Cluster Means		
Cluster	**Debt_ratio**	**Loan_Amount**
1	0.033854716	2.700106470
2	0.873549240	-0.001967049
3	-0.553952660	-0.351298725

All of the debt ratios are low. Cluster 1 = 0.033, Cluster 2 = 0.874 and Cluster 3 are the lowest, a negative result of -0.554. But compare them together, cluster 2 is comparatively high. So, we can draw a conclusion that the high income group (cluster 2) owns the most ability to repay.

With these 2 results, we can draw a conclusion that cluster 2, the high income group's demand for loaning money, is the most outstanding group. Therefore, in the following part, we will make recommendations.

If banks launch a loan package, we encourage them to loan to the high income group.

Firstly, they have needs. From the result of the first clustering analysis, they borrow the most amount of money. Reveals there are needs for this group of people. If Bank loans more to them, the bank can earn more by receiving interest.

Secondly, this group is safe to loan. We are so sure that the biggest concern for banks to loan is the ability for repayment. The high-income group is the safest group to loan. As they have good credit history which we find out from before analysis. They must be able to repay on time.

<u>Conclusion</u>

In conclusion, we found that people who are married and self-employed loan more. New family members may increase the living cost. It causes them to need loans. Those who are self-employed may need funds to start their businesses and operations, which may also lead to a large number of loans. It seems that people may take out loans for a better quality of life with family and career.

Before taking out a loan, people should calm down. There are a lot of loan advertisements to avoid being affected by these. Making loan decisions should be according to your demands. Only take out a loan if you can pay it back in order to avoid unnecessary financial pressure.

Question 2 – Data anonymisation

Description

About our data anonymisation policy, we have anonymised the customer name, telephone number, district, region and address, which are sensitive personal data that may cause harm or embarrassment to the individual. We have used different anonymization methods in this case.

For the customer number and telephone number, we have used Masking that hides personal identifiers to ensure that the data cannot refer back to a certain person. For the customer name, we hide the whole name without the first character so that customers can realize themselves but strangers cannot know the exact customer name. For the telephone number, we masked the last four digits to keep the confidentiality of the customer.

	full_name	new_full_name
1	SALLY HUI	Sxxxx Hxx
2	STEVEN WANG	Sxxxxx Wxxx
3	SUSAN WANG	Sxxxx Wxxx
4	JOHN LI	Jxxx Lx
5	THERESA CHAU	Txxxxxx Cxxx
6	SARA LEE	Sxxx Lxx
7	DEREK CHOI	Dxxxx Cxxx
8	STEVEN AU	Sxxxxx Ax
9	MARY NG	Mxxx Nx
10	JOANNA TANG	Jxxxxx Txxx
11	RICHARD WONG	Rxxxxxx Wxxx
12	TED LOK	Txx Lxx
13	RITZ TANG	Rxxx Txxx

	telephone	new_telephone
1	98421838	9842xxxx
2	91731955	9173xxxx
3	91481676	9148xxxx
4	98342063	9834xxxx
5	98451668	9845xxxx
6	98301276	9830xxxx
7	93721974	9372xxxx
8	98282015	9828xxxx
9	92701337	9270xxxx
10	98182107	9818xxxx
11	98471786	9847xxxx
12	95661989	9566xxxx
13	91251500	9125xxxx

After that, as the district and region may contain sensitive information about which the individual is staying, we have replaced the district in which the individual is living , with a real district area in Hong Kong. For example, the original district and region are wai chai and Hong Kong Island, it will be replaced by eastern and Hong Kong Island. It refers to the incoming file HK_districts.

	new_district	district
1	eastern	wan chai
2	kwai tsing	sai kung
3	tuen mun	yuen long
4	wan chai	tuen mun
5	southern	islands
6	tai po	north
7	sham shui po	sai kung
8	north	tai po
9	sai kung	tuen mun
10	yau tsim mong	sham shui po
11	central and western	wan chai
12	north	north
13	tuen mun	eastern

	new_region	region
1	hong kong island	hong kong island
2	new territories	new territories
3	new territories	new territories
4	hong kong island	new territories
5	hong kong island	new territories
6	new territories	new territories
7	kowloon	new territories
8	new territories	new territories
9	new territories	new territories
10	kowloon	kowloon
11	hong kong island	hong kong island
12	new territories	new territories
13	new territories	hong kong island

What's more, the address was anonymised by the fake flat. We set some fake flat and floor numbers with numeric numbers and letters. Those fake numbers will be randomly distributed. The data can show fake floor numbers and fake flats respectively. The new address will consist of fake data to protect users from hidden addresses.

	address	new_flat_no	new_floor_no	new_address
1	flat 6, 55/F	flat A	23/F	flat A, 23/F
2	flat B, 60/F	flat G	21/F	flat G, 21/F
3	flat 4, 35/F	flat 2	65/F	flat 2, 65/F
4	flat 10, 22/F	flat A	34/F	flat A, 34/F
5	flat 7, 19/F	flat D	67/F	flat D, 67/F
6	flat K, 45/F	flat 8	43/F	flat 8, 43/F
7	flat 3, 31/F	flat 1	54/F	flat 1, 54/F
8	flat 3, 21/F	flat G	56/F	flat G, 56/F
9	flat 12, 31/F	flat 7	21/F	flat 7, 21/F
10	flat 12, 1/F	flat D	56/F	flat D, 56/F
11	flat 9, 56/F	flat 4	54/F	flat 4, 54/F
12	flat F, 24/F	flat G	43/F	flat G, 43/F
13	flat 11, 23/F	flat 2	56/F	flat 2, 56/F