BIA 652

FINAL PROJECT PROPOSAL

NAME	CWID
DISHANT NAIK	10454341
VIREN GHORI	10456691

INTRODUCTION

The advancement in technology has brought in convenience and made our life simpler. Today we are just one swipe away to make any kind of payment. Now a days we do not need to carry bunch of notes and coins whenever we go out for shopping. All you need is a thin plastic card to make a payment. After the invention of payment methods like Apple pay and Samsung pay you don't even need to carry a card as well.

A credit card is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's promise to the card issuer to pay them for the amounts plus the other agreed charges.

While making it easy for us to do shopping while sitting on our couch, it has on the other hand made it easy for people to buy unnecessary items they don't need or just buy things which they cannot repay. And this raises concerns of credit card holders defaulting to the credit card landers. So instead of dealing it after credit card holder default it is better to do some analysis and get alerted when credit card holder shows sign of defaulting in advance.

Project statement:

Defaulting credit card is one of the major concerns in finance world. Banks and other credit card landers always need to be alerted to detect such person and take appropriate actions against them to avoid defaulting. And that is what motivated us to build our project around it. In this project we will try to classify a credit card user as **"Credible clients"** or **"Not credible clients".** To do this first we will closely analyze data in hand and will look for any strong or strange relationship between dependent and independent variables. Then we will use **"Logistic Regression"** to classify the dependent variable.

Data set description:

Data set source: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

About data set (Used for the project) source:

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

About Data set:

This data set aimed at the case of customer default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Description of response variable Y:

"default.payment.next.month" is the response variable for our analysis. Its datatype is "int64" and it only contains 0s and 1s. This variable indicates 1 when client is not credible and 0 when client is credible.

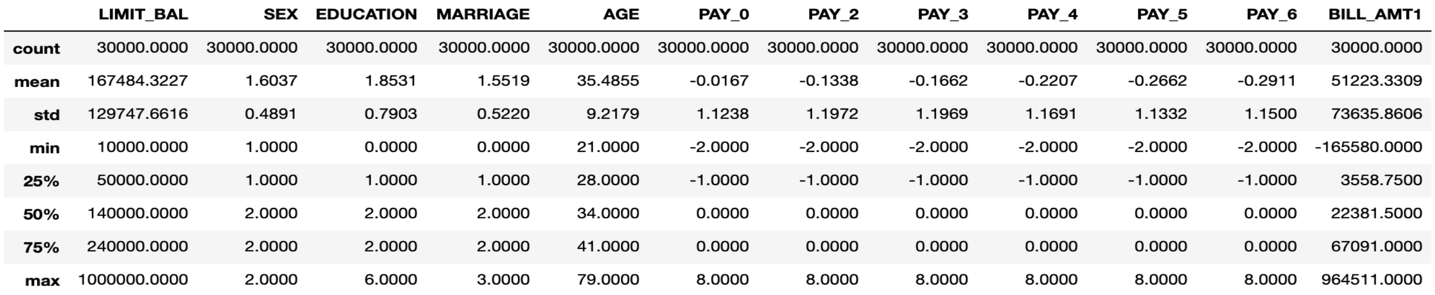
Description of predictor variables X:

There are in total 23 unique predictors. Description of each is as below.

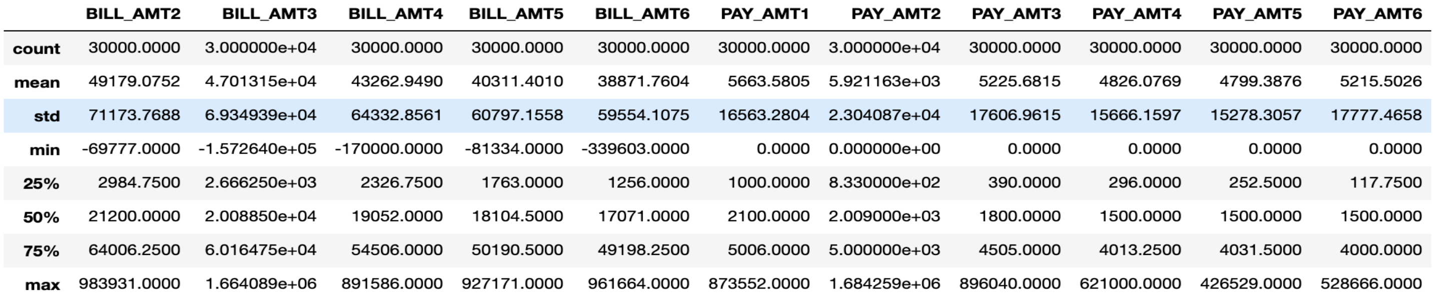
1. LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit [Data type: float64]
2. SEX: Gender (1=male, 2=female) [Data type: int64]
3. EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) [Data type: int64]
4. MARRIAGE: Marital status (1=married, 2=single, 3=others) [Data type: int64]
5. AGE: Age in years [Data type: int64]
6. PAY_0: Repayment status in September 2005 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 8 = payment delay for eight months, 9 = payment delay for nine months and above) [Data type: int64]
7. PAY_2: Repayment status in August 2005 (scale same as above) [Data type: int64]
8. PAY_3: Repayment status in July 2005 (scale same as above) [Data type: int64]
9. PAY_4: Repayment status in June 2005 (scale same as above) [Data type: int64]
10. PAY_5: Repayment status in May 2005 (scale same as above) [Data type: int64]
11. PAY_6: Repayment status in April 2005 (scale same as above) [Data type: int64]
12. BILL_AMT1: Amount of bill statement in September 2005 (NT dollar) [Data type: float64]

13. BILL_AMT2: Amount of bill statement in August 2005 (NT dollar) [Data type: float64]
14. BILL_AMT3: Amount of bill statement in July 2005 (NT dollar) [Data type: float64]
15. BILL_AMT4: Amount of bill statement in June 2005 (NT dollar) [Data type: float64]
16. BILL_AMT5: Amount of bill statement in May 2005 (NT dollar) [Data type: float64]
17. BILL_AMT6: Amount of bill statement in April 2005 (NT dollar) [Data type: float64]
18. PAY_AMT1: Amount of previous payment in September 2005 (NT dollar) [Data type: float64]
19. PAY_AMT2: Amount of previous payment in August 2005 (NT dollar) [Data type: float64]
20. PAY_AMT3: Amount of previous payment in July 2005 (NT dollar) [Data type: float64]
21. PAY_AMT4: Amount of previous payment in June 2005 (NT dollar) [Data type: float64]
22. PAY_AMT5: Amount of previous payment in May 2005 (NT dollar) [Data type: float64]
23. PAY_AMT6: Amount of previous payment in April 2005 (NT dollar) [Data type: float64]

Data set summary statistics:

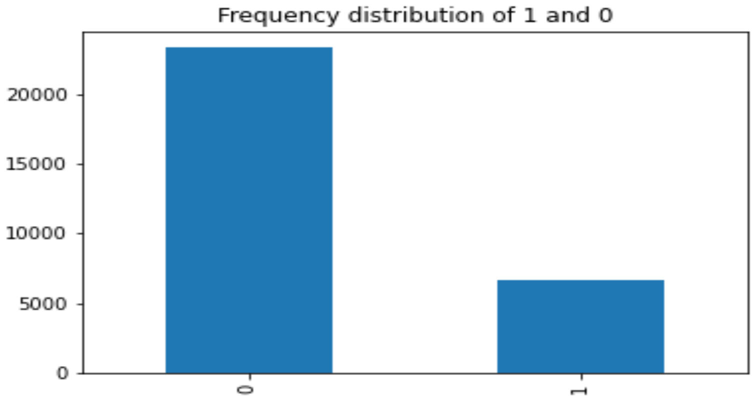


	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1
count	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000	30000.0000
mean	167484.3227	1.6037	1.8531	1.5519	35.4855	-0.0167	-0.1338	-0.1662	-0.2207	-0.2662	-0.2911	51223.3309
std	129747.6616	0.4891	0.7903	0.5220	9.2179	1.1238	1.1972	1.1969	1.1691	1.1332	1.1500	73635.8606
min	10000.0000	1.0000	0.0000	0.0000	21.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-165580.0000
25%	50000.0000	1.0000	1.0000	1.0000	28.0000	-1.0000	-1.0000	-1.0000	-1.0000	-1.0000	-1.0000	3558.7500
50%	140000.0000	2.0000	2.0000	2.0000	34.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	22381.5000
75%	240000.0000	2.0000	2.0000	2.0000	41.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	67091.0000
max	1000000.0000	2.0000	6.0000	3.0000	79.0000	8.0000	8.0000	8.0000	8.0000	8.0000	8.0000	964511.0000

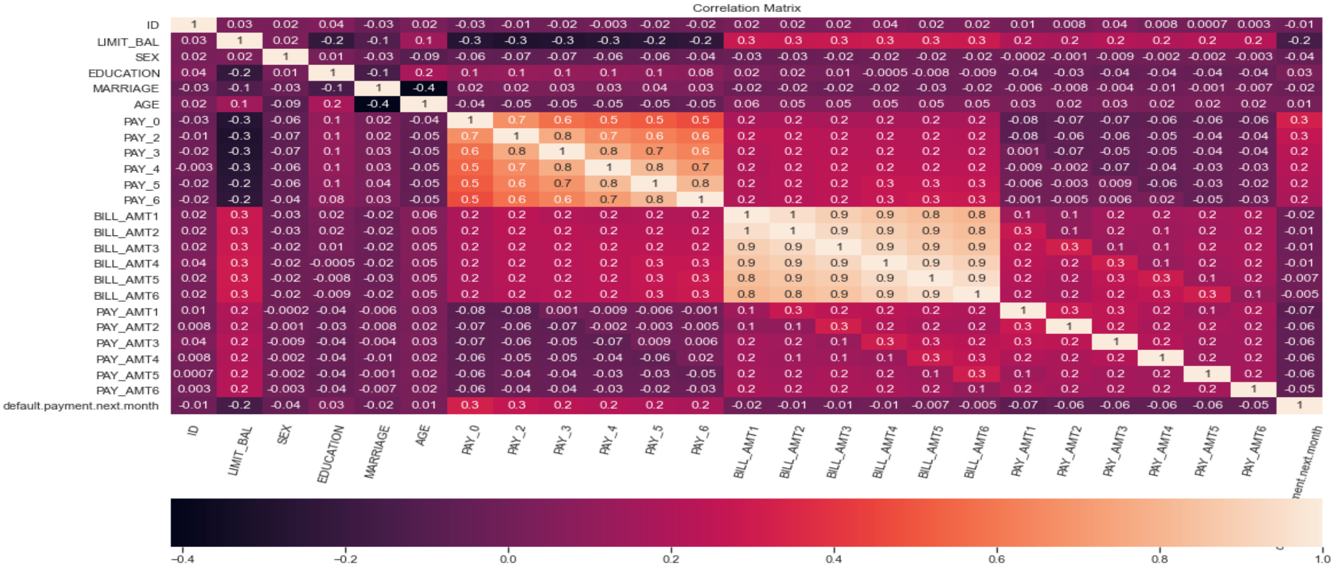


	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
count	30000.0000	3.000000e+04	30000.0000	30000.0000	30000.0000	30000.0000	3.000000e+04	30000.0000	30000.0000	30000.0000	30000.0000
mean	49179.0752	4.701315e+04	43262.9490	40311.4010	38871.7604	5663.5805	5.921163e+03	5225.6815	4826.0769	4799.3876	5215.5026
std	71173.7688	6.934939e+04	64332.8561	60797.1558	59554.1075	16563.2804	2.304087e+04	17606.9615	15666.1597	15278.3057	17777.4658
min	-69777.0000	-1.572640e+05	-170000.0000	-81334.0000	-339603.0000	0.0000	0.000000e+00	0.0000	0.0000	0.0000	0.0000
25%	2984.7500	2.666250e+03	2326.7500	1763.0000	1256.0000	1000.0000	8.330000e+02	390.0000	296.0000	252.5000	117.7500
50%	21200.0000	2.008850e+04	19052.0000	18104.5000	17071.0000	2100.0000	2.009000e+03	1800.0000	1500.0000	1500.0000	1500.0000
75%	64006.2500	6.016475e+04	54506.0000	50190.5000	49198.2500	5006.0000	5.000000e+03	4505.0000	4013.2500	4031.5000	4000.0000
max	983931.0000	1.664089e+06	891586.0000	927171.0000	961664.0000	873552.0000	1.684259e+06	896040.0000	621000.0000	426529.0000	528666.0000

Distribution of "Credible clients" and "Not credible clients" in Dataset:



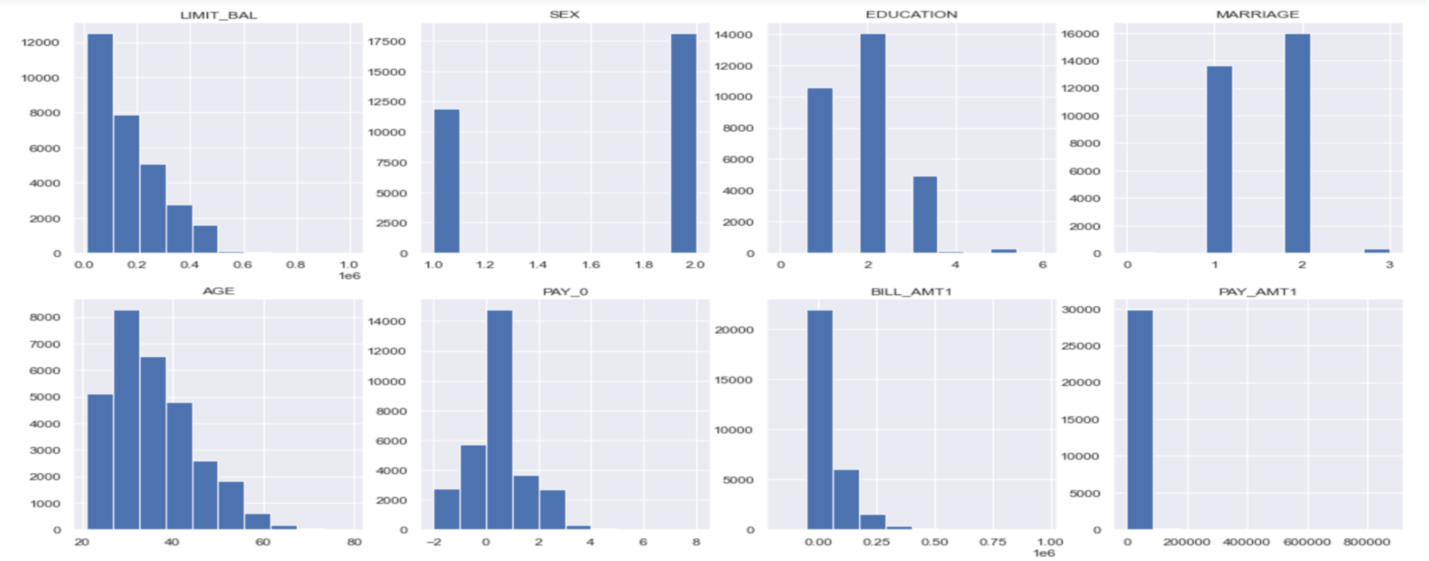
Correlation matrix of dataset:



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	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	ment.next.month
ID	1	0.03	0.02	0.04	-0.03	0.02	-0.03	-0.01	-0.02	-0.003	-0.02	-0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.01	0.008	0.04	0.008	0.0007	0.003	-0.01
LIMIT_BAL	0.03	1	0.02	-0.2	-0.1	0.1	-0.3	-0.3	-0.3	-0.3	-0.2	-0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2	-0.2
SEX	0.02	0.02	1	0.01	-0.03	-0.09	-0.06	-0.07	-0.07	-0.06	-0.06	-0.04	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.0002	-0.001	-0.009	-0.002	-0.002	-0.003	-0.04
EDUCATION	0.04	-0.2	0.01	1	-0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.08	0.02	0.02	0.01	-0.0005	-0.008	-0.009	-0.04	-0.03	-0.04	-0.04	-0.04	-0.04	0.03
MARRIAGE	-0.03	-0.1	-0.03	-0.1	1	-0.4	0.02	0.02	0.03	0.03	0.04	0.03	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.006	-0.008	-0.004	-0.01	-0.001	-0.007	-0.02
AGE	0.02	0.1	-0.09	0.2	-0.4	1	-0.04	-0.05	-0.05	-0.05	-0.05	-0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.03	0.02	0.03	0.02	0.02	0.02	0.01
PAY_0	-0.03	-0.3	-0.06	0.1	0.02	-0.04	1	0.7	0.6	0.5	0.5	0.5	0.2	0.2	0.2	0.2	0.2	0.2	-0.08	-0.07	-0.07	-0.06	-0.06	-0.06	0.3
PAY_2	-0.01	-0.3	-0.07	0.1	0.02	-0.05	0.7	1	0.8	0.7	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.2	-0.08	-0.06	-0.06	-0.05	-0.04	-0.04	0.3
PAY_3	-0.02	-0.3	-0.07	0.1	0.03	-0.05	0.6	0.8	1	0.8	0.7	0.6	0.2	0.2	0.2	0.2	0.2	0.2	0.001	-0.07	-0.05	-0.05	-0.04	-0.04	0.2
PAY_4	-0.003	-0.3	-0.06	0.1	0.03	-0.05	0.5	0.7	0.8	1	0.8	0.7	0.2	0.2	0.2	0.2	0.2	0.2	-0.009	-0.002	-0.07	-0.04	-0.03	-0.03	0.2
PAY_5	-0.02	-0.2	-0.06	0.1	0.04	-0.05	0.5	0.6	0.7	0.8	1	0.8	0.2	0.2	0.2	0.3	0.3	0.3	-0.006	-0.003	0.009	-0.06	-0.03	-0.02	0.2
PAY_6	-0.02	-0.2	-0.04	0.08	0.03	-0.05	0.5	0.6	0.6	0.7	0.8	1	0.2	0.2	0.2	0.3	0.3	0.3	-0.001	-0.005	0.006	0.02	-0.05	-0.03	0.2
BILL_AMT1	0.02	0.3	-0.03	0.02	-0.02	0.06	0.2	0.2	0.2	0.2	0.2	0.2	1	1	0.9	0.9	0.8	0.8	0.1	0.1	0.2	0.2	0.2	0.2	-0.02
BILL_AMT2	0.02	0.3	-0.03	0.02	-0.02	0.05	0.2	0.2	0.2	0.2	0.2	0.2	1	1	0.9	0.9	0.9	0.8	0.3	0.1	0.2	0.1	0.2	0.2	-0.01
BILL_AMT3	0.02	0.3	-0.02	0.01	-0.02	0.05	0.2	0.2	0.2	0.2	0.2	0.2	0.9	0.9	1	0.9	0.9	0.9	0.2	0.3	0.1	0.1	0.2	0.2	-0.01
BILL_AMT4	0.04	0.3	-0.02	-0.0005	-0.02	0.05	0.2	0.2	0.2	0.2	0.3	0.3	0.9	0.9	0.9	1	0.9	0.9	0.2	0.2	0.3	0.1	0.2	0.2	-0.01
BILL_AMT5	0.02	0.3	-0.02	-0.008	-0.03	0.05	0.2	0.2	0.2	0.2	0.3	0.3	0.8	0.9	0.9	0.9	1	0.9	0.2	0.2	0.3	0.3	0.1	0.2	-0.007
BILL_AMT6	0.02	0.3	-0.02	-0.009	-0.02	0.05	0.2	0.2	0.2	0.2	0.3	0.3	0.8	0.8	0.9	0.9	0.9	1	0.2	0.2	0.2	0.3	0.3	0.1	-0.005
PAY_AMT1	0.01	0.2	-0.0002	-0.04	-0.006	0.03	-0.08	-0.08	0.001	-0.009	-0.006	-0.001	0.1	0.3	0.2	0.2	0.2	0.2	1	0.3	0.3	0.2	0.1	0.2	-0.07
PAY_AMT2	0.008	0.2	-0.001	-0.03	-0.008	0.02	-0.07	-0.06	-0.07	-0.002	-0.003	-0.005	0.1	0.1	0.3	0.2	0.2	0.2	0.3	1	0.2	0.2	0.2	0.2	-0.06
PAY_AMT3	0.04	0.2	-0.009	-0.04	-0.004	0.03	-0.07	-0.06	-0.05	-0.07	0.009	0.006	0.2	0.2	0.1	0.3	0.3	0.2	0.3	0.2	1	0.2	0.2	0.2	-0.06
PAY_AMT4	0.008	0.2	-0.002	-0.04	-0.01	0.02	-0.06	-0.05	-0.05	-0.04	-0.06	0.02	0.2	0.1	0.1	0.1	0.3	0.3	0.2	0.2	0.2	1	0.2	0.2	-0.06
PAY_AMT5	0.0007	0.2	-0.002	-0.04	-0.001	0.02	-0.06	-0.04	-0.04	-0.03	-0.03	-0.05	0.2	0.2	0.2	0.2	0.1	0.3	0.1	0.2	0.2	0.2	1	0.2	-0.06
PAY_AMT6	0.003	0.2	-0.003	-0.04	-0.007	0.02	-0.06	-0.04	-0.04	-0.03	-0.02	-0.03	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	1	-0.05
default.payment.next.month	-0.01	-0.2	-0.04	0.03	-0.02	0.01	0.3	0.3	0.2	0.2	0.2	0.2	-0.02	-0.01	-0.01	-0.01	-0.007	-0.005	-0.07	-0.06	-0.06	-0.06	-0.06	-0.05	1

Histograms of independent variables:

Note: Not all independent variables are included in this plot.



Regression method:

Logistic regression is used when dependent variable is binary variable in nature or in simpler words when outcome of dependent variable is either success or failure. As dependent variable of this dataset is binary, we will be using Logistic regression. Logistic regression for 2 class classification can be represented by following formula.

$$p(C_1|\emptyset) = y(\emptyset) = \sigma(W^T\emptyset) \quad \text{[Note: } \sigma(.)\ is\ a\ Sigmond\ function\text{]}$$

$$p(C_2|\emptyset) = 1 - p(C_1|\emptyset)$$

$$Classify\ C_1\ if\ \ p(C_1|\emptyset)\ >\ p(C_2|\emptyset)\ Otherwice\ C_2$$