- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

DATA ANALYTICS MASTER PROGRAM
PRACTICUM PROJECT

BY MANDY NGUYEN

Case Studies: Bankruptcy
Charles Book Club
Home Equity

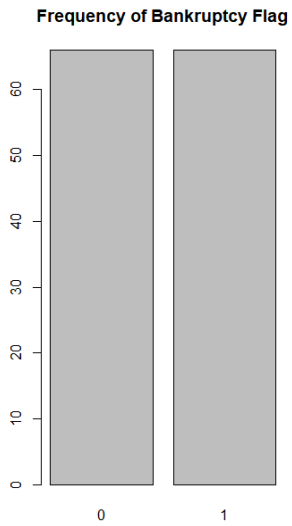
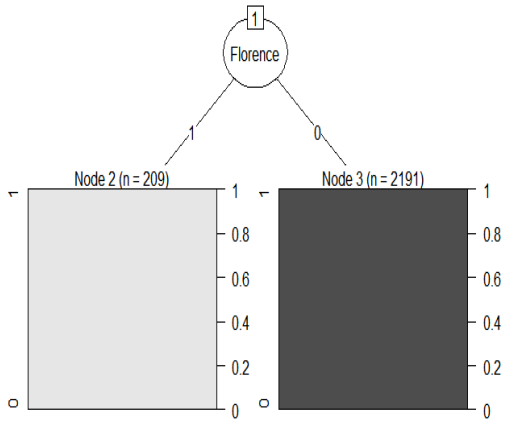
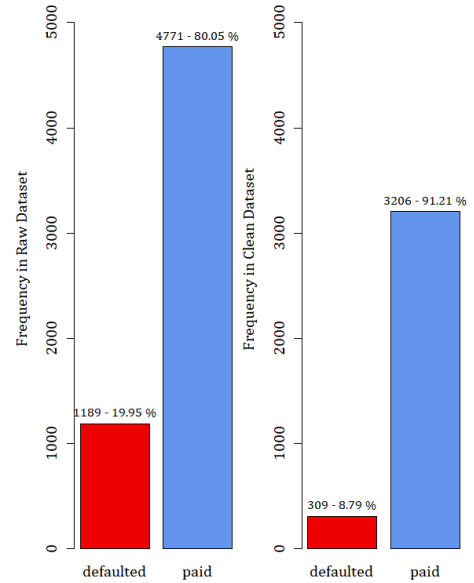
December 08, 2021

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

DATA

- **DATA**

	BANKRUPTCY	BOOK CLUB	HOME EQUITY
GOAL	Find best models to classify target variables.		
OBSERVATIONS	132	4,000	5,960
VARIABLES	27	19	13
PREDICTORS	24	16	12
TARGET	“D” Class: 1 = bankrupt Class : 0 = healthy	“Florence” Class: 1 = Yes to Florence Class: 0 = No to Florence	“BAD” Class: 1 = Defaulted (loan) Class: 0 = Paid (loan)
TARGET’s CLASS DISTRIBUTION CHART			

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

PREPARATION

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

BANKRUPTCY	BOOK CLUB	HOME EQUITY
ATTRIBUTE ANALYSIS	ATTRIBUTE ANALYSIS	ATTRIBUTE ANALYSIS
RANDOMIZE DATA SET	COMPUTE RFM SCORE	REMOVE MISSING VALUES
SELECT VARIABLES SET	SELECT VARIABLES SET	COMBINE CLASSES
NORMALIZE DATA	PARITION	NORMALIZE DATA
PARTITION	BALANCE	PARTITION
		BALANCE TRAINING

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

REMOVE MISSING VALUES

COMBINE CLASSES

Frequency of Values	Categories												
	0	1	2	3	4	5	6	7	8	9	10	11	13
	DEROG												
Before	3188	212	80	23	4	1	2	2	1	1	1		
After	2302												
Grouped	3188	212	115										
	DELINGQ												
Before	2964	333	126	50	21	6	7	6	1		1		
After	2302												
Grouped	2964	333	218										
	CLNO												
Before	1763	853	470	241	81	31	22	15	12	7	17	2	1
After	1127	556	304	143	51	14							
Grouped	1763	853	470	429									

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

VARIABLES SELECTION

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

• BANKRUPTCY	• BOOK CLUB	• HOME EQUITY
LOGISTIC REGRESSION CORRELATION	RFM GIVEN ATTRIBUTES LOGISTIC REGRESSION	LOGISTIC REGRESSION CORRELATION INFORMATION VALUES (SCORECARD BY WOE)

GLM

Original Dataset

Improved Original Dataset (step = backward = both)

Retained Ratios: R3, R5, R6, R9, R10, R16, R17, R18, R22, R23, R24

Original Training Dataset

Improved Original Training Dataset (step = backward = both)

Retained Ratios: R2, R3, R5, R6, R9, R10, R12, R14, R15, R16, R19, R21

Normalized Dataset

Improved Normalized Dataset (step = backward = both)

Retained Ratios: R3, R5, R6, R9, R10, R16, R17, R18, R22, R23, R24

Normalized Training Dataset

Improved Normalized Training Dataset (step = backward = forward = both)

Retained Ratios: R5, R6, R7, R10, R11, R12, R14, R16, R17, R18, R22, R24

Reduced Dataset

Improved Reduced Dataset (step = backward = both)

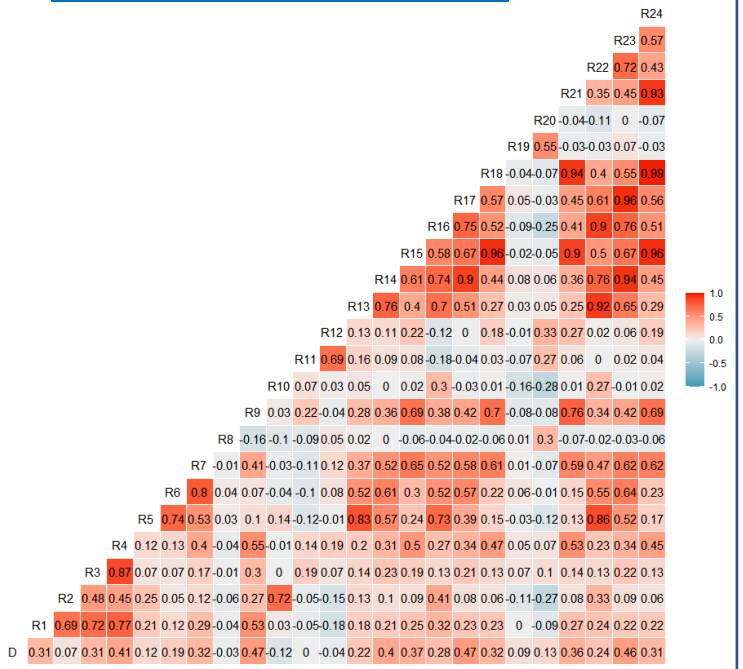
Retained Ratios: R9, R10, R17

Reduced Training Dataset

Improved Original Training Dataset (step = backward = both)

Retained Ratios: R9, R10, R17

CORRELATION



"RFM" SCORE Concatenate R__code, F_code, M_code to creates RFM classes

R_CODE

Recency ("Rcode" attribute):

- 0–2 months (Rcode = 1)
- 3–6 months (Rcode = 2)
- 7–12 months (Rcode = 3)
- 13 months and up (Rcode = 4)

F_CODE

Frequency ("Fcode" attribute):

- 1 book (Fcode = l)
- 2 books (Fcode = 2)
- 3 books and up (Fcode = 3)

M_CODE

Monetary ("Mcode" attribute):

- $0–$25 (Mcode = 1)
- $26–$50 (Mcode = 2)
- $51–$100 (Mcode = 3)
- $101–$200 (Mcode = 4)
- $201 and up (Mcode = 5)

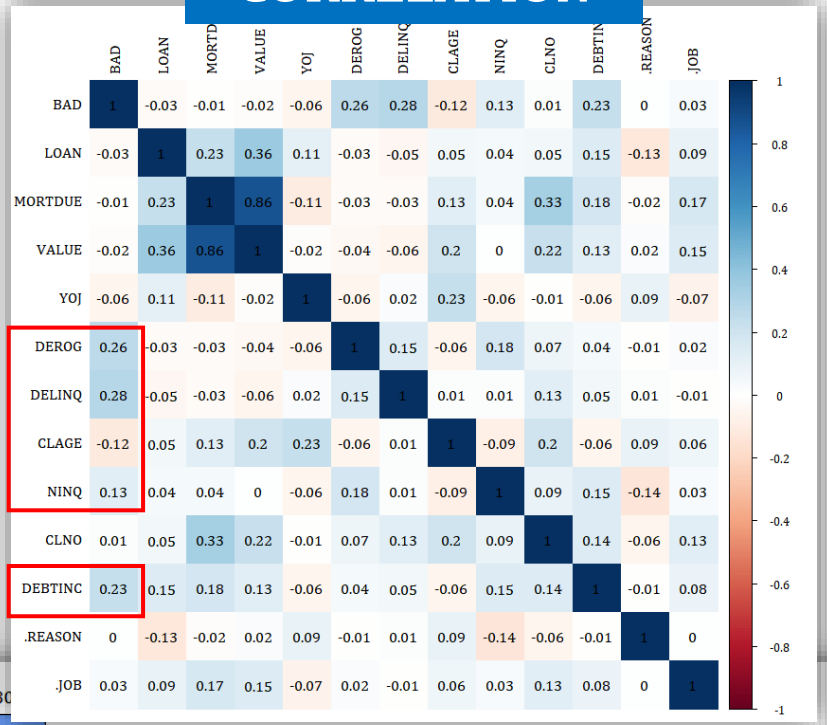
Seq.	ID	M	R	F	Mcode	Rcode	Fcode	Yes_Florence	No_Florence	RFM_score
1	25	297	14	2	5	4	2	0	1	425

RFM = "4_5_2"

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

CORRELATION



GLM

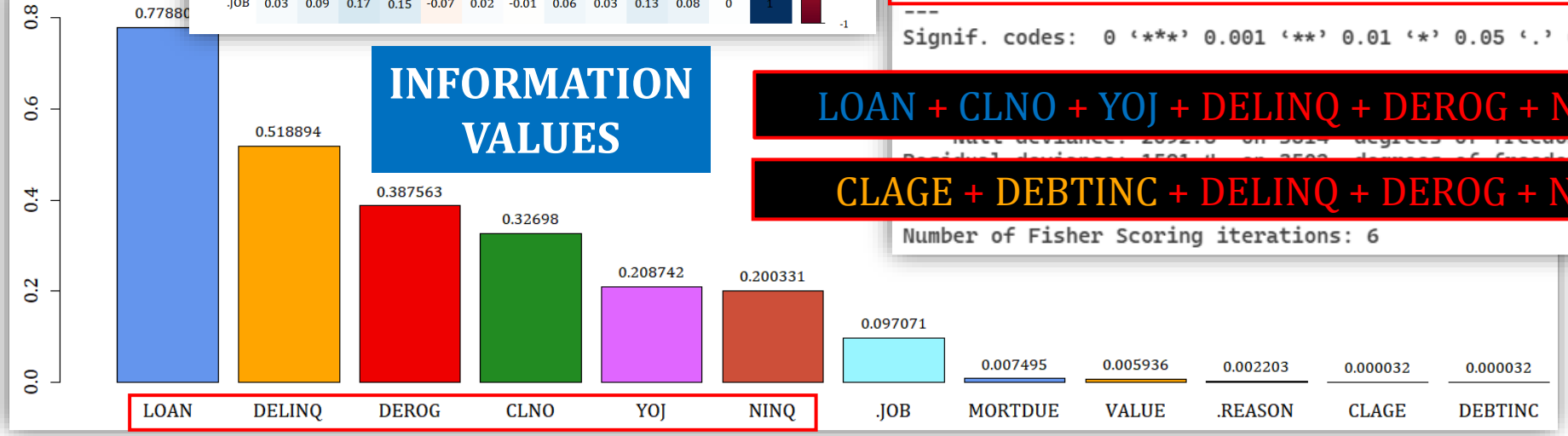
```
Call:
glm(formula = STATUS ~ LOAN + JOB + DEROG + DELINQ + CLAGE +
    NINQ + CLNO + DEBTINC, family = "binomial", data = X)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9354   0.1913   0.2816   0.3999   1.6910

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.100e+00  4.661e-01  10.943  < 2e-16 ***
LOAN         1.480e-05  6.992e-06   2.117 0.034282 *
JOBOffice    6.049e-01  2.636e-01   2.295 0.021725 *
JOBOther     8.124e-02  2.065e-01   0.393 0.694035
JOBProfExe  -3.175e-02  2.346e-01  -0.135 0.892357
JOBSales    -1.321e+00  4.462e-01  -2.961 0.003063 **
JOBSelf     -7.832e-01  3.953e-01  -1.981 0.047558 *
DEROG       -7.427e-01  1.012e-01  -7.336 2.20e-13 ***
DELINQ      -7.511e-01  6.907e-02 -10.874  < 2e-16 ***
CLAGE        5.540e-03  1.036e-03   5.346 9.01e-08 ***
NINQ        -1.269e-01  3.739e-02  -3.395 0.000687 ***
CLNO         1.980e-02  7.643e-03   2.590 0.009592 **
DEBTINC     -1.044e-01  1.031e-02 -10.119  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations: 6
```

INFORMATION VALUES



LOAN + CLNO + YOJ + DELINQ + DEROG + NINQ

CLAGE + DEBTINC + DELINQ + DEROG + NINQ

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

PARTITION & CLASSIFICATION

6 SUBSETS

- **Original Dataset:** Original Randomized dataset with all predictors
- **Original Training Dataset:** 60% Original Dataset
- **Normalized Dataset:** Original Randomized dataset with all predictors normalized by Min-Max method
- **Normalized Training Dataset:** 60% Normalized Dataset
- **Reduced Dataset:** Original Randomized dataset with 4 predictors R9, R10, R17, R20 (selected from correlation plot)
- **Reduced Training Dataset:** 60% Reduced Dataset

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

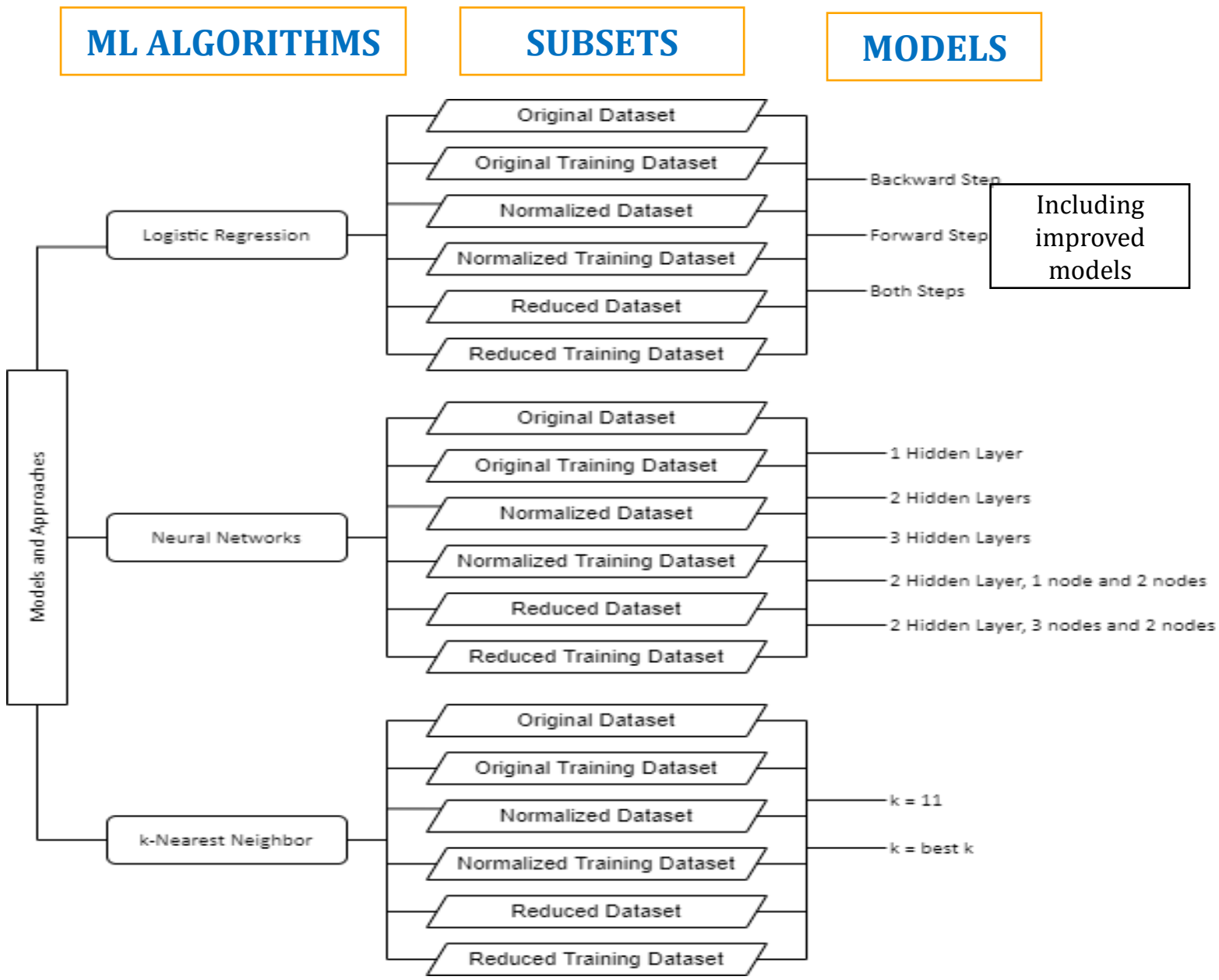
- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

69 Models

6 Subsets

60 % Training

40 % Validation



- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY
- CLASSIFICATION

Models	Accuracy	95% CI	No Info. Rate	P-Value	Kappa	Sensitivity	Specificity
LOGISTIC REGRESSION MODELS							
Improved Original Dataset (step = backward = both)	0.8679	(0.7466, 0.9452)	0.5283	1.709e-07	0.7333	0.9286	0.8000
Retained Ratios: R3, R5, R6, R9, R10, R16, R17, R18, R22, R23, R24							
Normalized Dataset	0.9434	(0.8434, 0.9882)	0.6604	1.006e-06	0.8721	0.9714	0.8889
Reduced Dataset							
Reduced Dataset	0.9245	(0.8179, 0.9791)	0.6604	6.766e-06	0.8271	0.9714	0.8333
Improved Reduced Dataset (step = backward = both)	0.9245	(0.8179, 0.9791)	0.6604	6.766e-06	0.8271	0.9714	0.8333
Retained Ratios: R9, R10, R17							
NEURAL NETWORKS							
Improved Original Dataset (hidden: layer = 1; node = 3)	0.9623	(0.8702, 0.9954)	0.5283	2.356e-12	0.9243	0.9643	0.9600
Normalized Dataset (hidden: layer = 1; node = 1)	0.9623	(0.8702, 0.9954)	0.6604	1.104e-07	0.9159	0.9714	0.9444
Improved Normalized Dataset (hidden: layer = 1; node = 3)	0.9811	(0.8993, 0.9995)	0.6604	7.945e-09	0.9585	0.9714	1.0000
Reduced Dataset (hidden: layer = 1; node = 1)	0.9623	(0.8702, 0.9954)	0.6604	1.104e-07	0.9159	0.9714	0.9444
K-NEAREST NEIGHBOR							
Improved Original Dataset (k = 1)	1	(0.9328, 1)	0.5283	2.055e-15	1	1.0000	1.0000
Improved Original Dataset (k = 3)	0.8113	(0.6803, 0.9056)	0.5283	1.709e-05	0.6198	0.8571	0.7600
Improved Normalized Dataset (k = 1)	1	(0.9328, 1)	0.6604	2.812e-10	1	1.0000	1.0000
Improved Normalized Dataset (k = 2)	0.8868	(0.7697, 0.9573)	0.6604	0.0001552	0.7665	0.8286	1.0000
Improved Reduced Dataset (k = 1)	1	(0.9328, 1)	0.6604	2.812e-10	1	1.0000	1.0000
Improved Reduced Dataset (k = 3)	0.9057	(0.7934, 0.9687)	0.6604	3.58e-05	0.7979	0.8857	0.9444

SIGNIFICANT CLASS: "1" CUTOFF (C): 0.5	RFM Score	K-Nearest Neighbors	LOGISTIC REGRESSION		
	C >mean	K14	C0.5	C0.5	C0.5
	ALL	SET 1	ALL	SET 2	SET 3
ACCURACY	0.81	0.92	0.91	0.91	0.91
SENSITIVITY	0.00	0.00	0.04	0.02	0.00
SPECIFICITY	0.87	1.00	0.99	0.99	1.00

SET 1: PREDICTORS: "FirstPurch", "Related.Purchase"; TARGETS: "Yes_Florence", "No_Florence"
SET 2: "ArtBks", "ItalArt", "ItalAtlas", "ItalCook"
SET 3: "Rcode", "Fcode", "Mcode"

BALANCED TRAINING

LOGISTIC REGRESSION

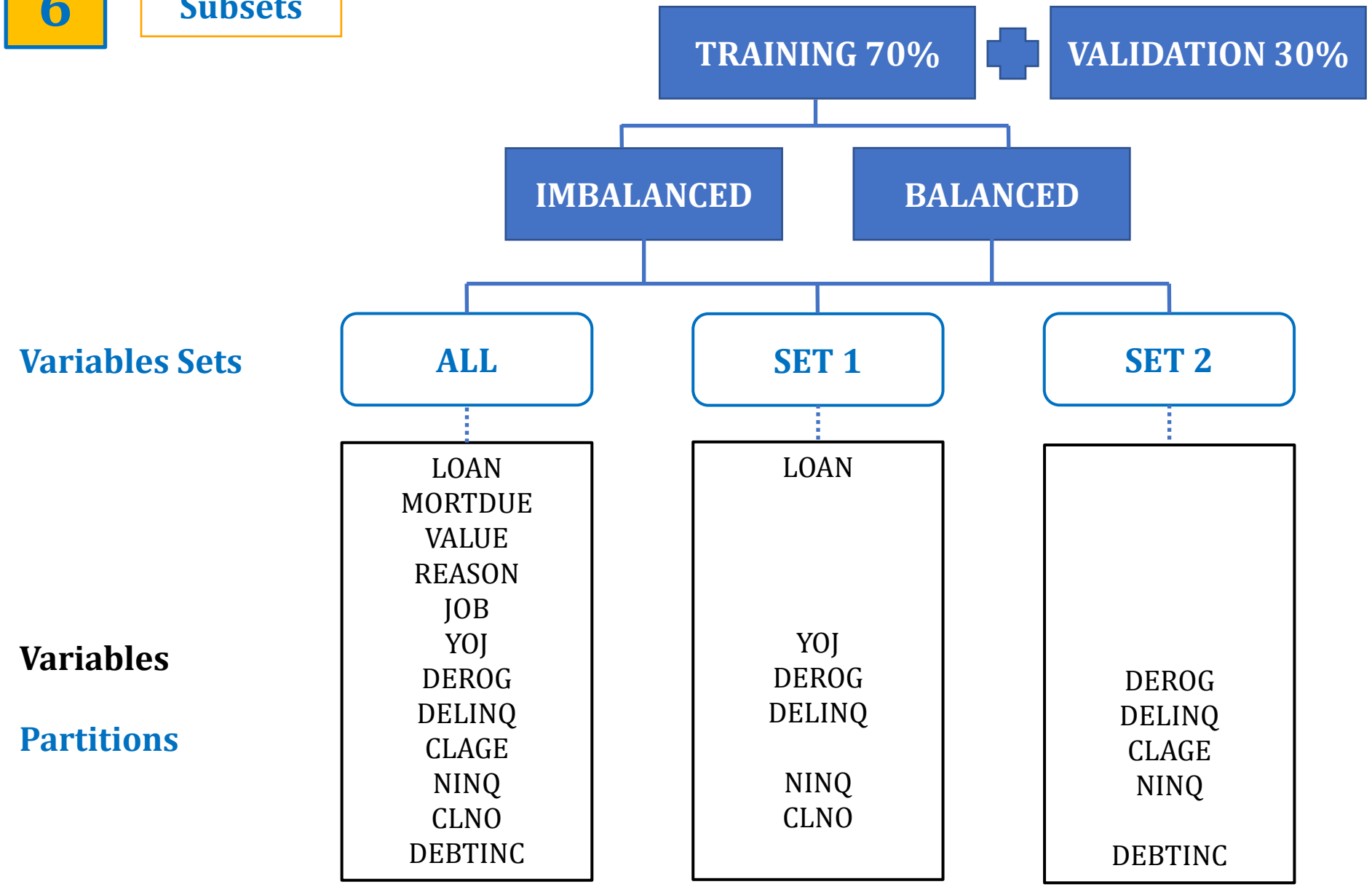
SIGNIFICANT CLASS: "1"

CUTOFF (C): 0.5

	C0.5 P ALL	C0.5 Q SET 2	C0.5 R SET 3
ACCURACY	0.61	0.72	0.46
SENSITIVITY	0.51	0.30	0.63
SPECIFICITY	0.62	0.76	0.45

SET 1: PREDICTORS: "FirstPurch", "Related.Purchase"; TARGETS: "Yes_Florence", "No_Florence"
SET 2: "ArtBks", "ItalArt", "ItalAtlas", "ItalCook"
SET 3: "Rcode", "Fcode", "Mcode"

6 Subsets



- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

18 MODELS

		LOGISTIC REGRESSION	K-NEAREST NEIGHBORS	NEURAL NETWORS
IMBALANCED TRAINING SET	ALL VARIABLES	MODEL A	MODEL G	MODEL M
	VARIABLES SET 1	MODEL B	MODEL H	MODEL N
	VARIABLES SET 2	MODEL C	MODEL I	MODEL O
BALANCED TRAINING SET	ALL VARIABLES	MODEL D	MODEL J	MODEL P
	VARIABLES SET 1	MODEL E	MODEL K	MODEL Q
	VARIABLES SET 2	MODEL F	MODEL L	MODEL R

LOGISTIC REGRESSION

SIGNIFICANT CLASS: "**PAID**"
CUTOFF: **0.5**

	IMBALANCED TRAINING SET			BALANCED TRAINING SET		
	A ALL	B SET 1	C SET 2	E ALL	F SET 1	G SET 2
ACCURACY	**1.00**	**0.92**	**0.93**	**0.11**	**0.12**	**0.11**
SENSITIVITY	**1.00**	**0.99**	**0.99**	**0.07**	**0.07**	**0.07**
SPECIFICITY	**1.00**	**0.09**	**0.27**	**0.55**	**0.63**	**0.54**

BEST MODEL: B: **IMBALANCED TRAING SET**
SET 1: LOAN, JOY, CLAGE, DEBTINC, NINQ, CLNO

K-Nearest Neighbors

SIGNIFICANT CLASS: "**PAID**"
K: **various**

	IMBALANCED TRAINING SET			BALANCED TRAINING SET		
	K5	K5	K5	K15	K7	K15
	G	H	I	J	K	L
	ALL	SET 1	SET 2	ALL	SET 1	SET 2
ACCURACY	**0.92**	**0.91**	**0.92**	**0.73**	**0.83**	**0.84**
SENSITIVITY	**0.99**	**0.99**	**0.99**	**0.78**	**0.87**	**0.91**
SPECIFICITY	**0.08**	**0.21**	**0.10**	**0.35**	**0.41**	**0.40**

BEST MODEL: I: **IMBALANCED TRAING SET**
K5
SET 2: DEROG, DELINQ, CLAGE, NINQ, DEBTINC

NEURAL NETWORKS

SIGNIFICANT CLASS: "**PAID**"
LAYER (L) CUTOFF: **0.5**

	IMBALANCED TRAINING SET			BALANCED TRAINING SET		
	L(1)	L(1,1)	K5	K15	K7	K15
	M	N	O	P	Q	R
	ALL	SET 1	SET 2	ALL	SET 1	SET 2
ACCURACY	0.92	0.92	0.62	0.08	0.08	0.15
SENSITIVITY	0.99	1.00	0.66	0.00	0.00	0.11
SPECIFICITY	0.00	0.00	0.2	1.00	1.00	0.61

BEST MODEL: N: **IMBALANCED TRAING SET**
(1,1): 2 LAYERS, 1 NODE EACH LAYER
SET 1: LOAN, JOY, CLAGE, DEBTINC, NINQ, CLNO

SIGNIFICANT CLASS: "**PAID**" CUTOFF: **0.5**	**LOGISTIC REGRESSION** IMBALANCED B SET 1	**K-Nearest Neighbors** IMBALANCED K5 I SET 2	**NEURAL NETWORKS** IMBALANCED L(1,1) N SET 1
ACCURACY	**0.92**	**0.92**	**0.92**
SENSITIVITY	**0.99**	**0.99**	**1.00**
SPECIFICITY	**0.09**	**0.10**	**0.00**

MODEL B: IMBALANCED TRAING SET - SET 1: LOAN, JOY, CLAGE, DEBTINC, NINQ, CLNO
MODEL I: IMBALANCED TRAING SET - K5 - SET 2: DEROG, DELINQ, CLAGE, NINQ, DEBTINC
MODEL N: IMBALANCED TRAING SET - (1,1): 2 LAYERS, 1 NODE EACH LAYER - SET 1: LOAN, JOY, CLAGE, DEBTINC, NINQ, CLNO

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

RESULTS

• RESULTS	• BANKRUPTCY	• BOOK CLUB	• HOME EQUITY
• ALGORITHM	**NEURAL NETWORKS**	**LOGISTICS REGRESSION**	**K-Nearest Neighbors**
• TRAINING SET	FULL ORIGINAL RANDOMIZED	BALANCED TRAINING	IMBALANCED TRAINING
• TARGET	"D" Class: 1 = bankrupt Class : 0 = healthy	"Florence" Class: 1 = Yes to Florence Class: 0 = No to Florence	"BAD" Class: 1 = Defaulted (loan) Class: 0 = Paid (loan)
• PREDICTORS	**ALL**	**SET 3** "Rcode", "Fcode", "Mcode"	**SET 2:** DEROG, DELINQ, CLAGE, NINQ, DEBTINC
• CUTOFF	**0.5**	**0.5**	**0.5**
• OTHERS	**LAYER 1; NODE 3**		**K5**

- BANKRUPTCY
- BOOK CLUB
- HOME EQUITY

- DATA
- PREPARATION
- VARIABLES SELECTION
- PARTITION
- CLASSIFICATION
- RESULTS
- SUMMARY
- CONCLUSION

CONCLUSION & KEY LEARNING

- **DATA**

	BANKRUPTCY	BOOK CLUB	HOME EQUITY
GOAL	Find best models to classify target variables.		
OBSERVATIONS	132	4,000	5,960
VARIABLES	27	19	13
PREDICTORS	24	16	12
TARGET	“D” Class: 1 = bankrupt Class : 0 = healthy	“Florence” Class: 1 = Yes to Florence Class: 0 = No to Florence	“BAD” Class: 1 = Defaulted (loan) Class: 0 = Paid (loan)
TARGET’s CLASS DISTRIBUTION CHART