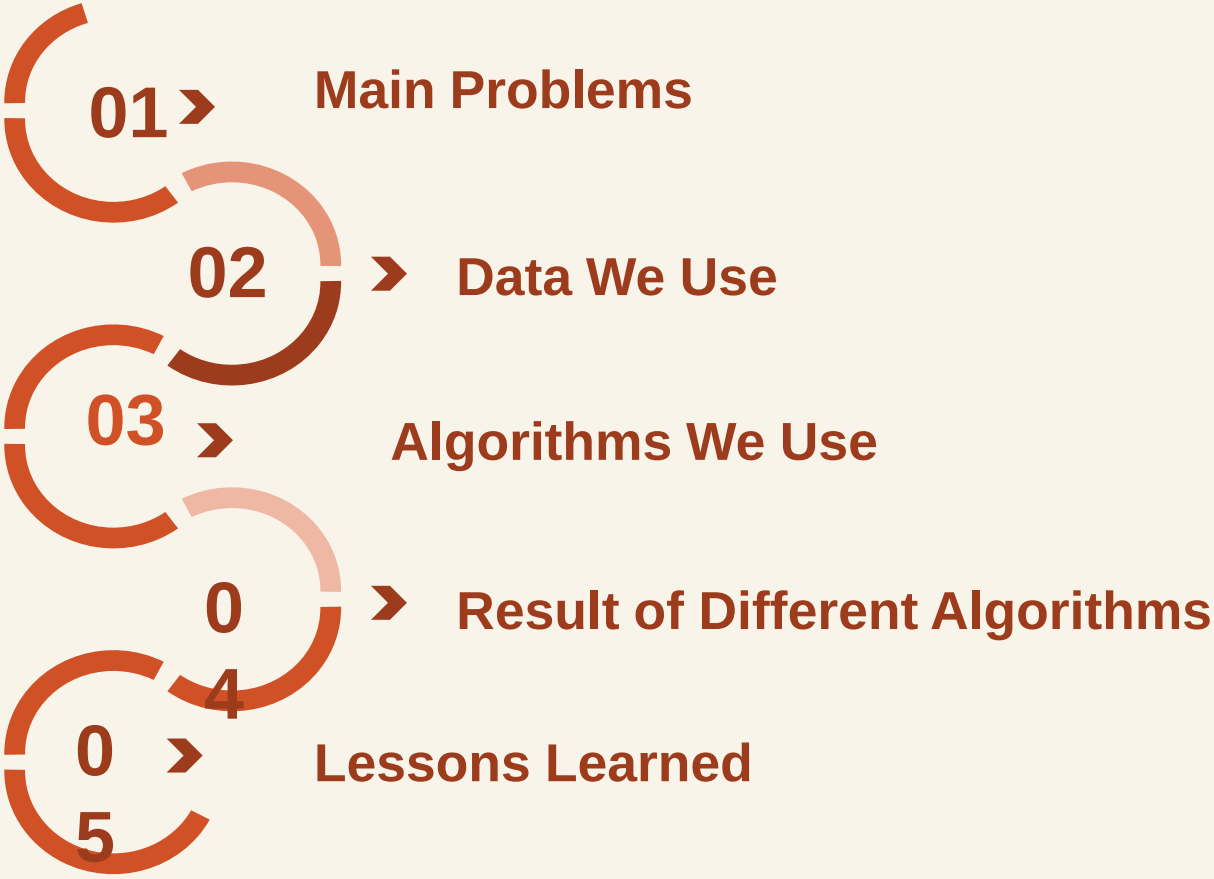
Data Mining for CFS

Team 7 Epitome

Amey Jain
Bingyu Zhang
Jhalak Goyal
Jiali Chen
Qianwen Li

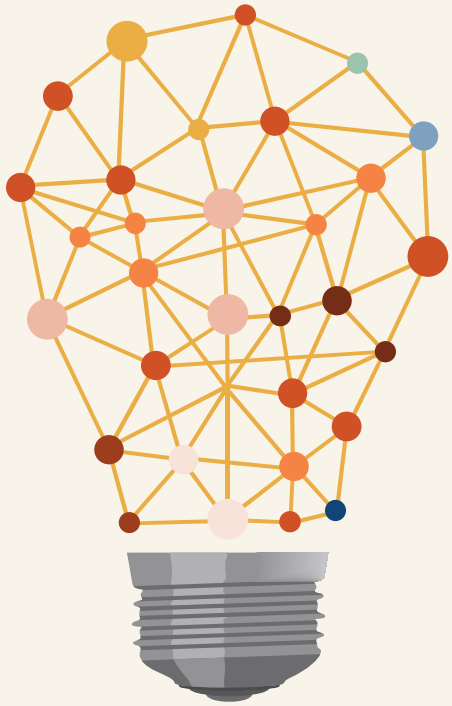
Agenda


01
Main Problems
02
Data We Use
03
Algorithms We Use
04
Result of Different Algorithms
05
Lessons Learned

3 Main Problems

01

Product Selection



02

Product Introduction

Data We Use

---- Product Selection

Existing Customer Data

- 186 records
  6 attributes
  1 class label

New Customer Data

- 21 records
- Missing class

Class 1: Education Loan, Overdraft Checking Account

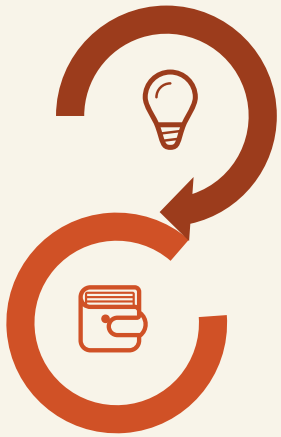
Type	Life-style	Vacation	eCredit	Salary	Property Value	Class
Student	Spending>>Saving	6	40	13.62	3.2804	C1

Data We Use

---- Product Introduction

Existing Product Data

- 160 records
  8 attributes
  2 class representations



New Product Data

- 40 records
- Missing label and missing score

Label: 1 ---> success
0 ---> failure
Score: sales of first year

Service _Type	Customer	Monthly_Fee	Budget	Size	Promotion	Interest_ Rate	Period	Label	Score
Fund	Student	0.64	0.95	Small	Full	0	10	1	26.72

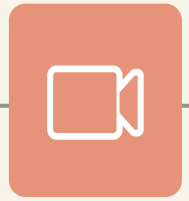
Two Algorithms Used

K Nearest Neighbors

KNN

Implement KNN algorithm to find the nearest records with the one we need to classify

Decision Tree



Weka

Using the C4.5 algorithm built in weka

C4.5 implemented

Implement the decision tree using C4.5 and reduced error pruning

Two Algorithms Used

--- KNN

KNN

Implement KNN algorithm to find the nearest instances with the one we need to classify

- Preprocess data

  Normalize numeric value

- Calculate similarity between instances

  Apply similarity matrix for non-numeric attribute

- Using weighted voting to determine the class

- Adjust the weights of different attributes

Type	Life-style	Vacation	eCredit	Salary	Property Value	Class
Student	Spending<Saving	0.079	0.107	0.220	0.183	C1

Two Algorithms Used

--- Decision Tree

- Generate a tree-like graph
- C4.5 which can handle continuous value
- Post-prune the tree to overcome over-fitting

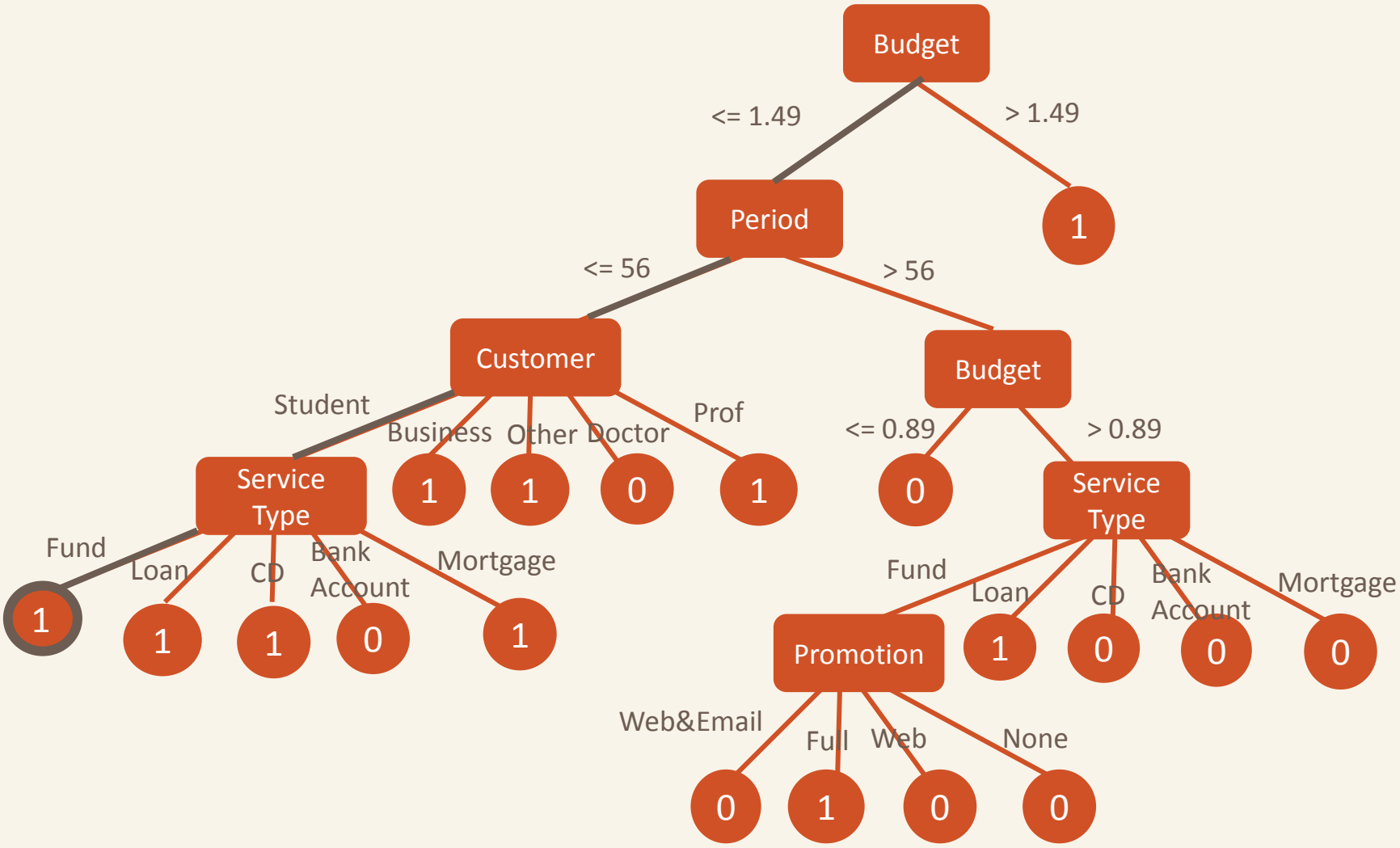
Weka

Using the J48 algorithm built in weka

C4.5 Implemented

Implement the decision tree using C4.5 and reduced error pruning

Two Algorithms Used --- Decision Tree



Result of Product Selection

KNN

Decision Tree in Weka

C4.5 Implemented

The Weight of Attributes:
Most Important: eCredit
Least Important: Life Style

Cross Validation Accuracy			
1	94.4%	6	88.9%
2	100%	7	94.4%
3	100%	8	83.3%
4	100%	9	77.8%
5	88.9%	10	88.9%

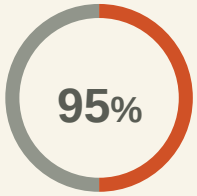
=== Confusion Matrix ===

```
 a  b  c  d  e   <-- classified as
23  2  1  9  1 |  a = C1
 1 25  0  0  0 |  b = C2
 2  0 39  0  0 |  c = C3
10  0  0 36  1 |  d = C4
 0  0  0  2 34 |  e = C5
```

Cross Validation Accuracy			
1	89.5%	6	89.5%
2	73.4%	7	73.7%
3	78.9%	8	78.9%
4	94.7%	9	89.5%
5	89.5%	10	86.7%

Result of Product Introduction

--- Binary Label



KNN

Decision Tree In Weka

C4.5 Implemented

The Weight of Attributes:
Most Important: Budget

Cross Validation Accuracy			
1	93.75%	6	100%
2	93.75%	7	100%
3	87.5%	8	100%
4	100%	9	81.25%
5	100%	10	93.75%

=== Confusion Matrix ===

```
 a  b   <-- classified as
63  1 |  a = 0
 6 90 |  b = 1
```

Cross Validation Accuracy			
1	87.5%	6	100%
2	93.75%	7	81.25%
3	93.75%	8	100%
4	93.75%	9	93.75%
5	100%	10	81.25%

Result of Product Introduction

--- Real Label

Output of KNN:

Service_Type	Customer	Monthly_ Fee	Budget	Size	Promotion	Interest_R ate	Period	Label	Score
Fund	Student	0.75	0.93	Small	Web&Email	1	5	1	21.542
Fund	Business	1.1	0.93	Small	Web&Email	1	65	1	23.588
Loan	Other	2.17	3.07	Small	Full	1	89	1	27.6459
Mortgage	Business	1.2	1.17	Small	Web	4	10	1	28.1259
CD	Business	1.2	1.09	Small	Web	0	26	0	20.666
Bank_Account	Professional	2.02	0.94	Large	None	3	15	1	21.798
Bank_Account	Doctor	4.11	1.07	Large	Web	1	20	0	21.616
Bank_Account	Student	4.08	0.98	Large	None	0	15	0	19.266
Loan	Business	14.17	4.83	Medium	Web	3	84	1	32.166
Loan	Professional	11.12	5.19	Large	Web	4	103	1	32.078
Mortgage	Professional	10.68	6.01	Large	Web	2	85	1	31.6579
Mortgage	Doctor	12.99	5.21	Medium	Web	3	87	1	33.63
Mortgage	Business	13.65	3.71	Large	None	1	87	1	31.3659
CD	Business	5.63	7.15	Medium	Web	1	88	1	33.2940

Lessons Learned

01 Data Requirements

- ***Training data where the attributes is correlated with the class we need to classify.***
- ***Test data with the same attributes as the training data.***

02 Highlights of KNN

- ***Normalization of the data***
- ***Weight of the attributes is very important when using KNN***

  84.4% -----> 92%

03 Highlights of Decision Tree

- ***Using C4.5 to handle continuous value***
- ***Post-pruning to avoid overfit***

Lessons Learned

04 Choice of Classifiers

Decision Tree

STRENGTH

- Easy to interpret
- Performs well when dataset is large

KNN

STRENGTH

- Easy to implement
- Can predict continuous and discrete values
- Robust to noisy data

Decision Tree

WEAKNESS

- Only can predict discrete values
- Over-fitting problem

KNN

WEAKNESS

- Low performance when dataset is large
- Similarity matrix is required for categorical value

Thank You For Listening!

Team 7