Faculty of Engineering
Ain shams University
Computer and Systems Department

APRIORI ALGORITHM

TERM PROJECT

CSE412 - Selected Topics in Computer Engineering

Team members:

Mohand Gamal Fawzy Helmy Hammad	1501516
Mohned Mohamed Abd El-Hamied Ahmed	1501519
Mostafa Mohamed Ali Radwan Badr	1501466
Mostafa Mohamed Ezz El-Din Asaad	1501464
Menna tullah Hussein Aly Mustafa	1501490
Mona Mahmoud Abd El-Hafez Abd El-Fatah	16X0129

Abstract:

This report describes the implementation of Apriori algorithm to find all the association rules in the given dataset. The dataset itself represents the customer data for an insurance company; it has 12 attributes with 5822 records. This implementation aims to find all the possible association rules for user-defined values of support and confidence, in addition to computing the lift and leverage for each rule. The output of the algorithm is generated into a text file containing all valid association rules along with their support, confidence, lift and leverage. The algorithm was implemented from scratch using R language and RStudio. This implementation was done by a group of 6 senior students at the faculty of engineering, ain shams university, department of computer and systems engineering as a term project for a big data course. This report is considered as one of the deliverables for the project alongside the source code, a demo representation of how the code implements the algorithm and presentation slides.

Table of Contents:

1. Introduction

1.1. Purpose

The purpose of this report is to describe the term project of the Apriori algorithm implementation, clarify the project aims, objectives and phases.

1.2. List of definitions

- **Apriori Algorithm:** It is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

- **Support:** It is an indication of how frequently the items appear in the data. It is measured by the proportion of transactions in which an itemset appears.

- **Confidence:** It explains how likely Y is purchased when X is purchased. It defines association between two items. This is measured by the proportion of transactions with item X, in which item Y also appears.

- **Lift:** This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. A lift value greater than 1 means that item Y is likely to be bought if item X is bought, while a value less than 1 means that item Y is unlikely to be bought if item X is bought.

- **Leverage:** It computes the difference between the observed frequency of X and Y appearing together and the frequency that would be expected if X and Y were independent. A leverage value of 0 indicates independence.

- **Association rules:** if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases.

- **Cross tabular data format:** It is a method to quantitatively analyze the relationship between multiple variables. Cross tabulation groups variables to understand the correlation between different variables. It also shows how correlations change from one variable grouping to another. It is usually used in statistical analysis to find patterns, trends, and probabilities within raw data.

- **Actual time:** is the total time taken for the actual work completed.

- **Planned value:** the approved duration to get the task completed.

- **Earned value:** the work completed to date. It shows you the value that the project has produced if it were terminated today.

- **SPI:** Schedule performance index. It shows how you are progressing compared to the planned project schedule. It is a measure of schedule efficiency, expressed as the ratio of earned value to planned value.

- **SV:** Schedule variance. It helps you complete on time. You are ahead of schedule if the Schedule Variance is positive, behind schedule if the Schedule Variance is negative, on schedule if the Schedule Variance is zero.

1.3. Overview

The report starts by stating the project beneficiaries, then describing the project aims and objectives. The report then gives a detailed description for the project, its architecture and development environment, also providing some test cases and their result. Finally, it provides the project conclusion and uses the earned value method to assess the progress in the project.

2. Beneficiaries

The Apriori algorithm is used for data mining and generation for association rules which can be very helpful to various beneficiaries like:

2.1. Insurance Companies

- Predict which customers are potentially interested in an insurance policy.
- Describe the actual or potential customers; and possibly explain why these customers buy a policy.

2.2. Retail Shops

- Determine which retail items are purchased together
- Describe the potential customers for certain retail items

2.3. Web Services

- Determine a set of links clicked on by one user in a single session
- Filtering the web advertisements that appears to the user according to his /her interests

3. Project Aims and Objectives

3.1. Project Aims

This project aims to calculate the Apriori algorithm and generate all valid association rules to predict which customers are potentially interested in an insurance policy and to describe the actual or potential customers; and possibly explain why these customers buy a policy.

3.2. Project Objectives

- Calculate Apriori algorithm to generate valid association rules based on user-defined minimum support and minimum confidence
- Calculate and output the support for each valid association rule
- Calculate and output the confidence for each valid association rule
- Calculate and output the lift for each valid association rule
- Calculate and output the leverage for each valid association rule

4. Detailed Project Description

The project was developed by a team of 6 whose names and ids are mentioned on the cover page.

The project was developed by using R language and RStudio with the help of some basic libraries. It is important to mention that NO ready libraries to calculate either the support, confidence, lift, leverage, or association rules were used, as requested in the project requirements.

The project algorithm is executed by using 6 functions (which will be discussed later in the system architecture):

- CalculateSupport
- CalculateRules
- CheckInput
- GetDataReady
- PrintRules
- Shifter

The program firstly prompts the user to enter the minimum support and keeps prompting him/her until they enter a value within the range which is greater than 0 and less than 100 by using a function to validate the input. Then it prompts him/her to enter the minimum confidence and keeps prompting until they enter a value within the range by using the same function mentioned above.

The program then calls a function to extract the specified attributes on which the association rules will be generated from the while data found in the “ticdata2000” file. These 12 attributes for our team are from index 11 till index 22. These attributes are then given their names according to the “TicDataDescr”.

The data and the minimum support are then passed to a function to calculate all the levels of support for the data and compare them to the user-defined minimum support, it then only returns the item sets having support equal or more than the user-defined minimum support.

Theses item sets are then passed along with the minimum confidence to a function to calculate all the possible rules for every and each item set and compare their confidence with the user-defined minimum confidence. Then it calculates the lift and leverage for the valid rules. Finally, it returns the valid rules with their support, confidence, lift, and the index of the attribute in the right-hand side of the rule (antecedent)

Eventually, these rules are passed to a function to print them in a proper and understandable form in a text file named “output.txt” and then open that file. If the user entered values of minimum support and minimum confidence that resulted in no valid association rules being generated, the text file notifies the user by writing in the text file “There was no valid association rules in the data to satisfy the conditions given”

5. Project Phases

5.1. Data Extraction and Inputs Validation:

Duration: 3 April – 9 April

Working on extracting the attributes from the index 11 till 22 and adding their names to them according to the data description file, also working on validating the user inputs. In addition to learning how to code using R language and getting more familiar with its syntax

5.2. Support Calculation:

Duration: 10 April – 17 April

Working on calculating the support for all available item sets in the data and comparing them to the user-defined minimum support and pruning the item sets that have support less than the minimum support

5.3. Confidence, Lift, and Leverage Calculations:

Duration: 18 April – 27 April

Working on calculating the confidence for all different rules that can be generated from each item set that has support more than or equal the minimum support, then compare them to the user-defined minimum confidence and pruning the rules that have confidence less than the minimum confidence. Then for each item set with valid rules calculate both the lift and the leverage.

5.4. Outputting the Rules:

Duration: 28 April – 4 May

Working on outputting all the previously calculated valid rules in an understandable format and saving them in a text file, also printing the support, confidence, lift and leverage for each rule.

5.5. Report, Presentation Slides and Demo:

Duration: 5 May – 10 May

Working on the project report, making sure that the code is well commented, recording a demo video for the code execution and output, documenting test cases and preparing presentation slides to finalize all the deliverables for the project delivery.

6. System Architecture

This Apriori algorithm is implemented using R language by implementing 2 main functions:

- **CalculateSupport <- function(data, minSupport):** which takes the data and the user-defined minimum support as parameters and returns a data frame containing all item sets of data that have support that exceeds or equal to the user-defined minimum support.

- **CalculateRules(support, minConfidence):** which takes the support data frame which was calculated by the CalculateSupport function and the user-defined minimum confidence as parameters and returns a data frame containing all possible rules and their confidence, support, lift and leverage and index to the right hand side column of the rule (antecedent).

In addition to 4 helping functions to assist in the execution of the main functions and to output the results:

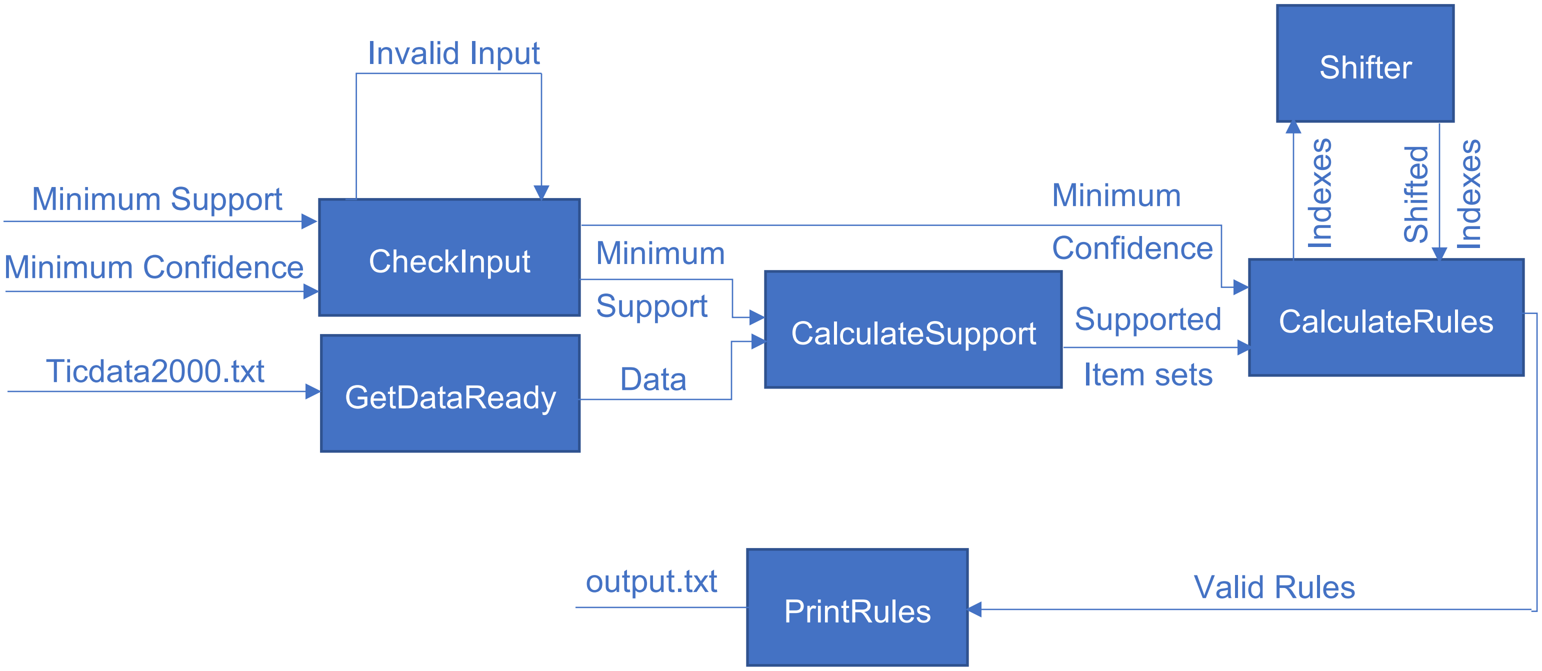
- **CheckInput <- function(input):** which takes a variable as a parameter to check whether it is a valid value or not and returns TRUE if there is an error and FALSE is the value is valid. It is used to check whether the user inputted minimum support and minimum confidence within the valid range or not (greater than 0% and less than 100%). If the input is invalid it notifies the user of the error and prompt him/her to enter a new valid input.

- **GetDataReady <- function():** which takes no parameters and return a data frame of the portion of the data that we should get the association rules from. It reads the whole data from the path of “ticdata2000” and extract the 12 attributes (columns) from them starting from a specific index (which is in our team’s case is 11). Then it gives each column a name based on what was written in “TicDataDescr”.

- **Shifter <- function (x, n = 1):** which takes a list and the amount of shifting as parameters and returns that same list but shifted cyclically by the shifting amount. If the shifting amount is not given it shifts by 1 as default. It is used by the CalculateRules function to get the different possible rules of the same item set to calculate the confidence for each.

- **PrintRules <- function(rules):** which takes the final valid association rules outputted by the CalculateRules function as a parameter and return nothing. It prints all valid association rules in a text file with name "output.txt". Each rule is generated and outputted in that file in an understandable format alongside its support, confidence, lift and leverage in a single line, followed by the next rule till all valid rules are outputted into the text file successfully and the file is opened. In case that the were no association rules for the data that satisfy the minimum support and the minimum confidence defined by the user, the output text file should notify the user by having the sentence "There was no valid association rules in the data to satisfy the conditions given"

The following diagram describes the whole system architecture



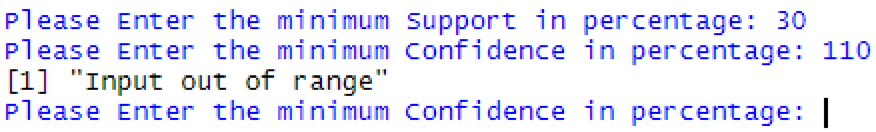
7. Development Environment

The Apriori algorithm was implemented using R, which is a free software environment for statistical computing and graphics.

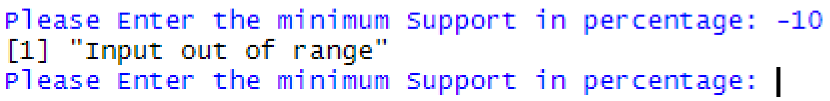
The project was completely developed and debugged by using RStudio integrated development environment for R language.

8. Testing Cases and Results

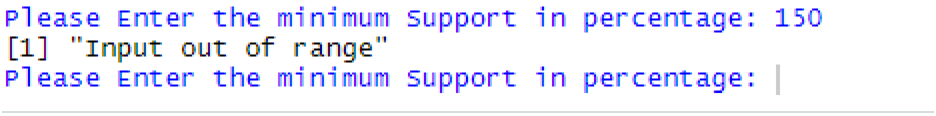
8.1. Handling user's input

- 

```
Please Enter the minimum Support in percentage: 30
Please Enter the minimum Confidence in percentage: 110
[1] "Input out of range"
Please Enter the minimum Confidence in percentage: 
```

- 

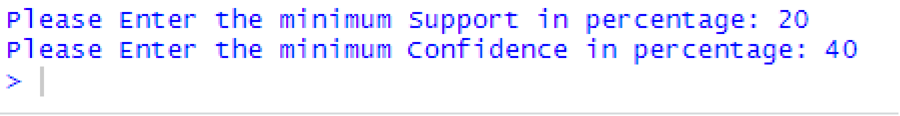
```
Please Enter the minimum Support in percentage: -10
[1] "Input out of range"
Please Enter the minimum Support in percentage: 
```

- 

```
Please Enter the minimum Support in percentage: 150
[1] "Input out of range"
Please Enter the minimum Support in percentage: 
```

-

8.2. Handling normal cases

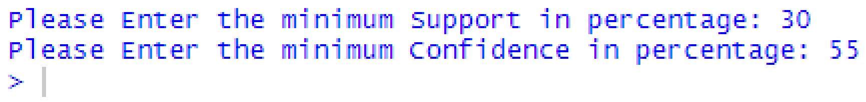
- 

```
Please Enter the minimum Support in percentage: 20
Please Enter the minimum Confidence in percentage: 40
> |
```

  - In output.txt file:

- MRELSA(0) ---> MBERZELF(0) Support:33.66541 Confidence:80.06536 Lift:1.117575 Leverage:0.03541784
- MBERZELF(0) ---> MRELSA(0) Support:33.66541 Confidence:46.99113 Lift:1.117575 Leverage:0.03541784
- MRELSA(1) ---> MBERZELF(0) Support:20.81759 Confidence:59.70443 Lift:0.8333714 Leverage:-0.04162377
- MRELSA(0) ---> MBERBOER(0) Support:30.7798 Confidence:73.20261 Lift:1.020559 Leverage:0.006200671
- MBERBOER(0) ---> MRELSA(0) Support:30.7798 Confidence:42.91188 Lift:1.020559 Leverage:0.006200671
- MRELSA(1) ---> MBERBOER(0) Support:22.17451 Confidence:63.59606 Lift:0.886629 Leverage:-0.02835399
- MRELOV(2) ---> MBERZELF(0) Support:20.45689 Confidence:67.8246 Lift:0.946715 Leverage:-0.01151397
- MRELOV(2) ---> MBERBOER(0) Support:21.3157 Confidence:70.67198 Lift:0.9852784 Leverage:-0.003184889
- MFALLEEN(0) ---> MBERZELF(0) Support:23.10203 Confidence:76.55094 Lift:1.06852 Leverage:0.01481436
- MFALLEEN(0) ---> MBERBOER(0) Support:25.36929 Confidence:84.06375 Lift:1.171981 Leverage:0.03722781
- MOPLHOOG(0) ---> MBERHOOG(0) Support:20.11336 Confidence:54.54122 Lift:2.083589 Leverage:0.1046013
- MBERHOOG(0) ---> MOPLHOOG(0) Support:20.11336 Confidence:76.83727 Lift:2.083589 Leverage:0.1046013
- MOPLHOOG(0) ---> MBERZELF(0) Support:32.44589 Confidence:87.98323 Lift:1.228095 Leverage:0.06026198

- MBERZELF(0) ---> MOPLHOOG(0) Support:32.44589 Confidence:45.2889 Lift:1.228095 Leverage:0.06026198
- MOPLHOOG(0) ---> MBERBOER(0) Support:26.55445 Confidence:72.00745 Lift:1.003897 Leverage:0.00103081
- MBERHOOG(0) ---> MBERZELF(0) Support:23.13638 Confidence:88.38583 Lift:1.233714 Leverage:0.04382947
- MBERHOOG(0) ---> MBERBOER(0) Support:20.31948 Confidence:77.62467 Lift:1.08221 Leverage:0.01543565
- MBERZELF(0) ---> MBERBOER(0) Support:54.75782 Confidence:76.43251 Lift:1.065589 Leverage:0.03370459
- MBERBOER(0) ---> MBERZELF(0) Support:54.75782 Confidence:76.341 Lift:1.065589 Leverage:0.03370459
- MRELSA(0) & MBERZELF(0) ---> MBERBOER(0) Support:24.64789 Confidence:73.21429 Lift:1.140734 Leverage:0.03040837
- MBERZELF(0) & MBERBOER(0) ---> MRELSA(0) Support:24.64789 Confidence:45.01255 Lift:1.140734 Leverage:0.03040837
- MRELSA(0) & MBERBOER(0) ---> MBERZELF(0) Support:24.64789 Confidence:80.07812 Lift:1.140734 Leverage:0.03040837
- MOPLHOOG(0) & MBERZELF(0) ---> MBERBOER(0) Support:23.85778 Confidence:73.53097 Lift:1.258966 Leverage:0.04907479
- MBERZELF(0) & MBERBOER(0) ---> MOPLHOOG(0) Support:23.85778 Confidence:43.56964 Lift:1.258966 Leverage:0.04907479
- MOPLHOOG(0) & MBERBOER(0) ---> MBERZELF(0) Support:23.85778 Confidence:89.84476 Lift:1.258966 Leverage:0.04907479

- 

```
Please Enter the minimum Support in percentage: 30
Please Enter the minimum Confidence in percentage: 55
> |
```

  - In output.txt file:

- MRELSA(0) ---> MBERZELF(0) Support:33.66541 Confidence:80.06536 Lift:1.117575 Leverage:0.03541784

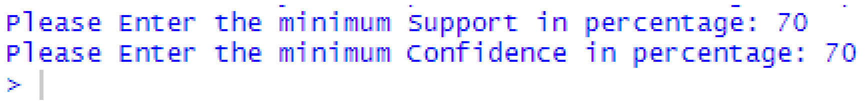
- MRELSA(0) ---> MBERBOER(0) Support:30.7798 Confidence:73.20261
  Lift:1.020559 Leverage:0.006200671

- MOPLHOOG(0) ---> MBERZELF(0) Support:32.44589 Confidence:87.98323
  Lift:1.228095 Leverage:0.06026198

- MBERZELF(0) ---> MBERBOER(0) Support:54.75782 Confidence:76.43251
  Lift:1.065589 Leverage:0.03370459

- MBERBOER(0) ---> MBERZELF(0) Support:54.75782 Confidence:76.341
  Lift:1.065589 Leverage:0.03370459

8.3. Handling special cases

- 

```
Please Enter the minimum Support in percentage: 70
Please Enter the minimum Confidence in percentage: 70
> 
```

  - In output.txt file:

There were no valid association rules in the data to satisfy the conditions given

9. Conclusion

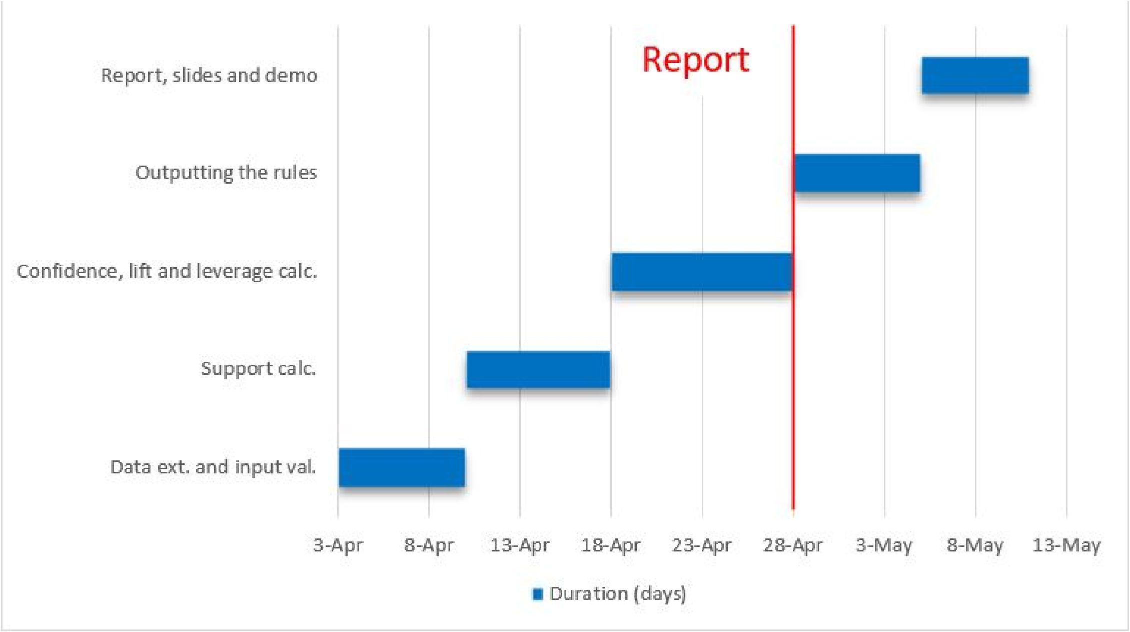
We can conclude from the project that data mining, even though it takes a lot of processing power, is considered an essential part for the success of many organizations and services. As for our case in this project, the association rules we generate help in predicting which customers are potentially interested in an insurance policy, in describing the actual or potential customers and in explaining why these customers buy a policy. All of which are critical information for any successful insurance company. Additionally, each association rule has its own support and confidence, and the more their values approach 100%, the more the rule is thought to be a fact. So, the user of the system can define his/her own minimum support and minimum confidence based on his/her preferences and act according to the association rules generated.

10. Earned Value Method

10.1. Activities Table

Activity	Predecessor	Duration(days)
Data extraction and input validation	-	7
Support calculation	Data extraction and input validation	8
Confidence, lift and leverage calculation	Support calculation	10
Outputting the rules	Confidence, lift and leverage calculation	7
Report, presentation slides and demo	Data extraction and input validation. Support calculation. Confidence, lift and leverage. calculation Outputting the rules.	6

10.2. Activities Gantt Chart



Fig(1): Gantt chart for the Earned value method for the progress of the project

10.3. Earned Value Analysis

Report at week 5

Activity	Actual % complete	Time taken (days)
Data ext. and validation	100%	8
Support calculation	100%	9
Confidence, lift and leverage calc.	80%	8
Outputting the rules	0%	0
Report, slides, and demo	0%	0

Activity	Actual time	Earned value	Planned value	SPI	SV
Data ext. and validation	8	7	7		
Support calculation	9	8	8		
Confidence, lift and leverage calc.	8	8	10		
Outputting the rules	0	0	0		
Report, slides, and demo	0	0	0		
Total	25	23	25	0.92	-2

SPI=Earned value/ Planned value

SV=Earned value – Planned value