INTERNSHIP ON DATA MINING COMBINATION WITH DATA SCIENCE AT Z SOFT SOLUTION LIMITED

BY

Md. Mansur Islam Bhuiyan

ID: 153-15-6560

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Fahad Faisal

Senior Lecturer

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER, 2018

APPROVAL

This Internship titled **"Internship On Data Mining Combination With Data Science At Z Soft Solution Limited**", submitted by **Md. Mansur Islam Bhuiyan** ID: 153-15-6560 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents.

BOARD OF EXAMINERS

`

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

`

Narayan Ranjan Chakraborty — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

`

Md. Tarek Habib — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

`

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this internship report is prepared under the supervision of **Fahad Faisal Senior Lecturer, Department of CSE,** Daffodil International University. I also declare that neither this internship report nor any part of this internship report has been submitted elsewhere for award of any Degree or Diploma. I also declare that, I collect information from Z Soft Solution Limited, Data Industry Service Provider Based Company and Internet.

Supervised by:

`

_____

Fahad Faisal
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

Md. Mansur Islam Bhuiyan
ID:153-15-6560
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish my profound indebtedness to **Fahad Faisal**, Senior Lecturer Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of supervisor in the field of "Internship On Data Mining Combination With Data Science At Z Soft Solution Limited" to carry out this internship. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this internship.

I would like to express our heartiest gratitude to **Dr. Syed Akther Hossain, Professor and Head, Department of CSE**, for his kind help to finish my internship and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Facing the problem of variation and chaotic behavior of customers, the lack of sufficient information is a challenge to many business organizations. Human analysts lacking an understanding of the hidden patterns in business data, thus, can miss corporate business opportunities. In order to embrace all business opportunities, enhance the competitiveness, discovery of hidden knowledge, unexpected patterns and useful rules from large databases have provided a feasible solution for several decades. While there is a wide range of financial analysis products existing in the financial market, how to customize the investment portfolio for the customer is still a challenge to many financial institutions. This paper aims at developing an intelligent Financial Data Mining Model (FDMM) for extracting customer behavior in the financial industry, so as to increase the availability of decision support data and hence increase customer satisfaction. The proposed financial model first clusters the customers into several sectors, and then finds the correlation among these sectors. It is noted that better customer segmentation can increase the ability to identify targeted customers, therefore extracting useful rules for specific clusters can provide an insight into customers' buying behavior and marketing implications. To validate the feasibility of the proposed model, a simple dataset is collected from a financial company in Hong Kong. The simulation experiments show that the proposed method not only can improve the workflow of a financial company, but also deepen understanding of investment behavior. Thus, a corporation is able to customize the most suitable products and services for customers on the basis of the rules extracted.

TABLE OF CONTENTS

CONTENTS **PAGE**

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

Daffodil International University is a good opportunity, which was the internship system. We got it during the last I would like to say that in this internship system I have a lot of practical experience. Due to the subprime mortgage crisis of 2015 and the global financial crisis of 2016, many investors suffered from financial products losses. The stock markets fell and large financial institutions collapsed. Such a series of financial events not only caused panic in the financial markets but also encouraged investors to take their money out of risky mortgage bonds and unstable equities etc. Investors now have to a more cautious attitude towards investments. Dealing with the financial events, unpleasant investor experience has become common and these personal experiences are demonstrated in risk and attitude to risk. This situation creates factors impacting investors in respect to returns on expectations. Despite the economy recovering gradually, investors have had to reassess the high risk of investment. At the same time, customer groups and discovering interesting relations among the variables for each data item.

1.2 Motivation

I am currently pursuing my Bachelors in Computer Science Engineering at **Daffodil International University**, I understand that it is important to have practical knowledge that complements the knowledge of the textbook and helps the student to gain a broader perspective on topics. Data mining is the field where large amount of data is collected and through process to some useful data information. What motivates it, the need of area motivates. Everything wants the precise information which is possible through it. Some of the fields where data mining is used are: Financial Service, Banking Area, Marketing, In Industries (To know the reviews of people and likes of people), Tech Companies, Retail Companies, Manufacturing Companies, Healthcare and Telecommunication industry.

1.3 Internship Objectives

The ultimate goal of my internship program is to prepare me for a competitive job market. So this is very effective in developing know-how. I would like to collect some extra features to give me a qualified person.

The internship in computer science is designed to provide work experience when students are still in school, coordinate work experience with academic education, and help students move from classroom to work. I am result oriented, Self-motivating and independent desire to work.

1.4 Introduction to the Company

Z Soft Solution Ltd is working as a data science and machine learning industry since 2016. They have team with highly skilled on several development areas like Web and MobileApps, Disaster Recovery with Micro Data System in Big Data or even for service serving their clients with best effort physically and technically. Z Soft Solution started in February 2015 officially but started its journey from 2014 to redefine data science and machine learning with Artificial Intelligence. Guided by the combined experiences of their team, Z Soft Solution has grown into a moderate version in the industry. The teams of serial developers either in business or in technology are passionate about democratizing the power of data science and machine learning for the mass market.

1.5 Report Layout

In the chapter one I have described objective of internship, Motivation of internship and Introduction to the company.

In the chapter two I have described the methodology of my internship. And this chapter gives the information about where the internship has been attached to undertake this program. Also included about how did perform the internship works, about the company, what are the data science service offered in Z Soft Solution Ltd and what are the roles of in jobs market of Data Mining.

In the chapter three I have described about daily task and activities, Events and Activities and Challenges.

In the chapter four I have described is Competencies Earned, Smart Plan, Reflections.

In the chapter five I have described is Conclusion and Future Scope. I discuss Future Scopes of Data Mining and write conclusion.

CHAPTER 2

Organization

2.1 Introduction

Z Soft Solution Ltd is introducing to be one of the multinational company in UK, Lithuania and Bangladesh. There are providing data science, machine learning, artificial intelligence services and solution since July 2015.It has its use of tools Weka, Phython, R, Scala, Julia, Rapid Miner, Orange, Knime, connecting to Data Sources and also Configuring data with excel sheet. The company worked with many multinational projects and international organizations with high appreciation from all concerned. It using the latest technologies and upgrading the services wherever it is required.

The centers Corporate data science and machine learning Solution department is capable of providing domain financial and telecommunication industry solutions with a group of highly efficient technical experts. Z Soft Solution has a very strong professional engineering and management team certified and associated with SAS Academy, IBM, Coursera, Harvard Extension School, DataCamp, and Dataquest and actively involved with world leading data science and machine learning associations including Udemy, Big Data University, TensorFlow, EdX.

2.2 Product and Market Situation

Z Soft Solution Ltd. prides itself as one of the valuable international data science, machine learning, artificial intelligence Company in Bangladesh and UK. They are the most experienced company in the ICT field where they are basic business ethics is Long Term Relationship with their customers. As they look at the growth over the decade our inception, they are extremely proud of what they have achieved, and even more excited about our outlook for an equally promising future. Z Soft Solution Ltd also provides different IT Services. These are given below.

IT Services and Supports

- Data Science
- Machine Learning
- Artificial Intelligent
- Cyber Security and Disaster recovery
- Web Design
- Mobile Apps Development
- Plugin Development
- Data mining with specific domain
- Cyber Security Services collaboration with Fire Eye and Siltech Network Solution
- Website Development.
- Telemarketing and Lead Generation.
- Virtual office Management
- Email Marketing and Automation
- Digital Marketing
- Animation and Video

2.3 Target Group:

The company's customer base includes all consumers and all small and medium-sized companies, including start-ups. The company intends to focus on building customer relationships works better when it’s driven by data as these are the perfect targets for our new high accuracy data offerings and have the company's largest growth potential. Web Solutions believes that these market segments have special pricing and service needs and create a dedicated, more reliable customer.

2.4 Data Analysis

Data Analysis is a useful technique for understanding your Strengths and Weaknesses, and for identifying both the Opportunities open to you.

Strengths:

- Communication between customer separated by distance (at official and within online customer)
- Working from official data mining
- Setting up an online data gathering system
- Gathering information (valuable asset in business)

Weaknesses:

- New entrants underestimate levels of expertise needed to survive the market
- Large sums of money required to set up businesses
- Competition for small resellers
- Numerous pricing tariffs and service potions

Opportunities:

- Good data predication with more industry
- Number of data in the UK increasing
- Some Research Directions in Data Mining
- Revision of regulatory tools

2.5 Organizational Structure

Organizational Structure of Z Soft Solution Ltd. Shown below in fig 2.5:

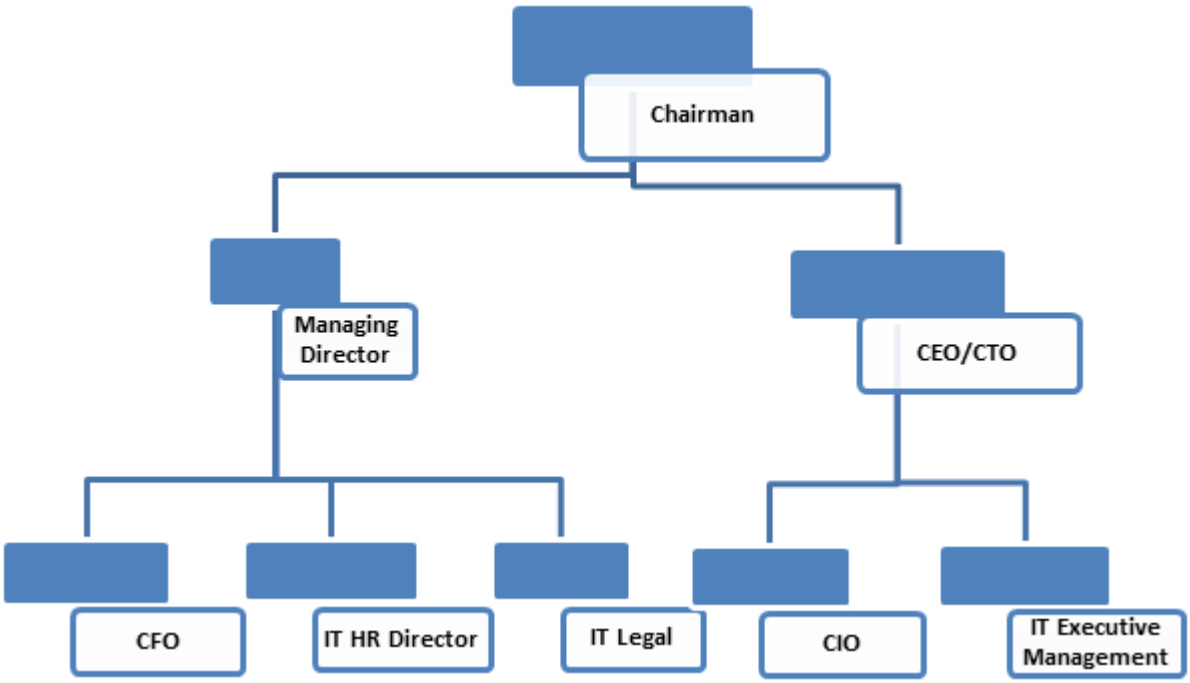


Figure 2.5: Organizational structure of Z Soft Solution Limited

CHAPTER 3

Tasks, Projects and Activities

3.1 Daily Task and Activities

Month - 1: In the first month of internship on Z Soft solution limited I have learned and performed the following tasks:

- Learning & understanding About Data Components.
- Learning & understanding Basics of Data Mining.
- Understanding Data Mining Tools.
- Learning & understanding R Programming.
- IDE Installation.
- Startup (After R is downloaded and installed)

Month - 2: In the second month of internship on Z Soft solution limited I have learned and performed the following tasks:

- Basic Table of Useful R commands.
- Entering Commands.
- The Workspace.
- Graphic User Interface.
- Operator in R, Data Types, Function, Importing Data, Plotting Data.
- Descriptive Statistics
- SQL.

Month – 3: In the third month of internship on Z Soft Solution limited I have learned and performed the following tasks:

- File and Directory Permissions.
- Learn Descriptive Statistics
- Python
- Learn Inferential statistics
- Predictive Model (Learning ANOVA, Linear and Logistic Regression on SAS and use of Salesforce Administration tools.)

Month – 4: The last month of internship on Z Soft Solution Limited I have learned and performed the following tasks:

- Installing the Python and SciPy platform
- Loading the dataset.
- Summarizing the dataset.
- Visualizing the dataset.
- Evaluating some algorithms.
- Making some predictions.

3.2 Events and Activities

- Monitor and Maintain Data Systems and online dataset.
- Financial Prediction Dataset
- Retail Company Sales Data Set.
- Marketing domain Data Set.
- Black Friday dataset
- HealthCare Data Set
- Text Mining Data Set.
- Travel Industry Data Set.
- Twitter Mining Data

3.3 Architecture of Financial Data Mining Model (FDMM)

According to Mark K.Y Mak et al; A Financial Data Mining Model, Sage Journal; Jan 01 2011; In order to gain a better understanding on investors' behavior and achieve higher customer satisfaction, a Financial Data Mining Model (FDMM) is proposed. It is crucial for financial planners to understand what and when the clients' need so as to devise the most appropriate asset allocation strategies. The framework is designed to select the related data from different available databases. Such data is converted into qualitative and quantitative data and then is used for building a centralized data warehouse. In order to find out which set of investment products that clients might be interested in, one year of historical data is required for data analysis. The generic architecture of FDMM is illustrated in Figure 1. The FDMM consists of three modules, namely, the Data Selection and Preprocessing Module (DSPM), the Clustering Module (CM) and

the Rules Discovery Module (RDM). The DSPM is used to select the relevant data and prepare the proper format for the mining process. The CM is to identify the most influencing factors that affect customer behavior and then such data is segmented into different groups. Clustering the database is aimed at identifying the target group of customers so as to discover useful rules for these segmented groups. Following the CM, the RDM can discover useful rules for each specific group. George S; The 5 Clustering Algorithms Data Scientists Need to Know; Feb 5 2018; https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

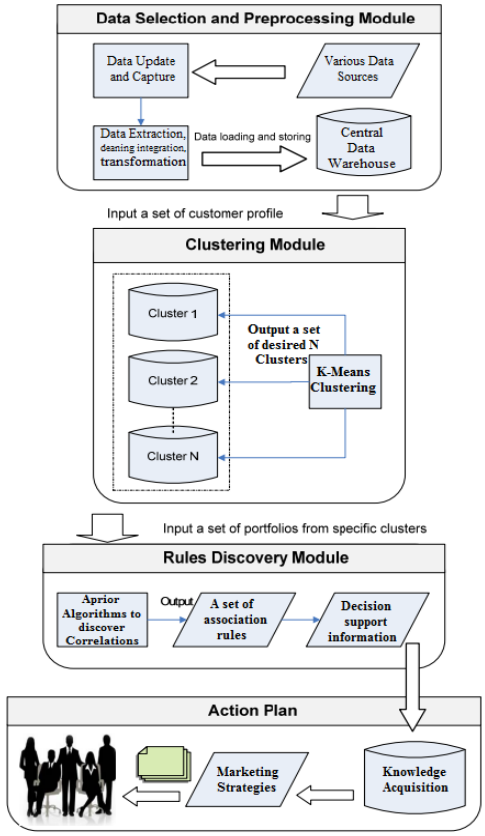


Figure 1. Architecture of the FDMM

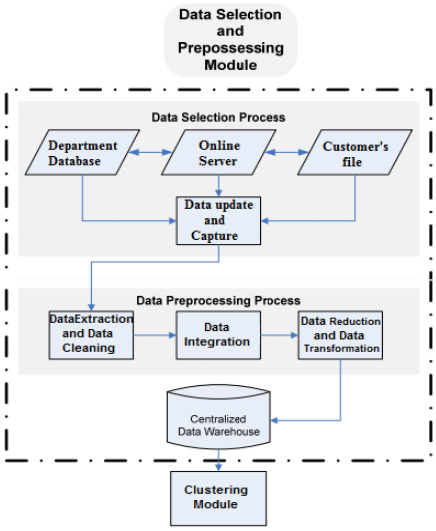


Figure 2. Workflow of DSPM

Figure 3.1: Architecture of FDMM [9]

3.4 Data Selection and Preprocessing Module (DSPM)

The Data Selection and Preprocessing Module (DSPM) aims at building a central data warehouse for supporting data mining tasks and quality information sharing. DSPM is connected to various data sources. The departmental database is a major component of DSPM, which holds valuable data, such as the customers' profiles, asset allocation and past investment records. As shown in Figure 2, there are two processes involved in DSPM, namely data selection process and data preprocessing.

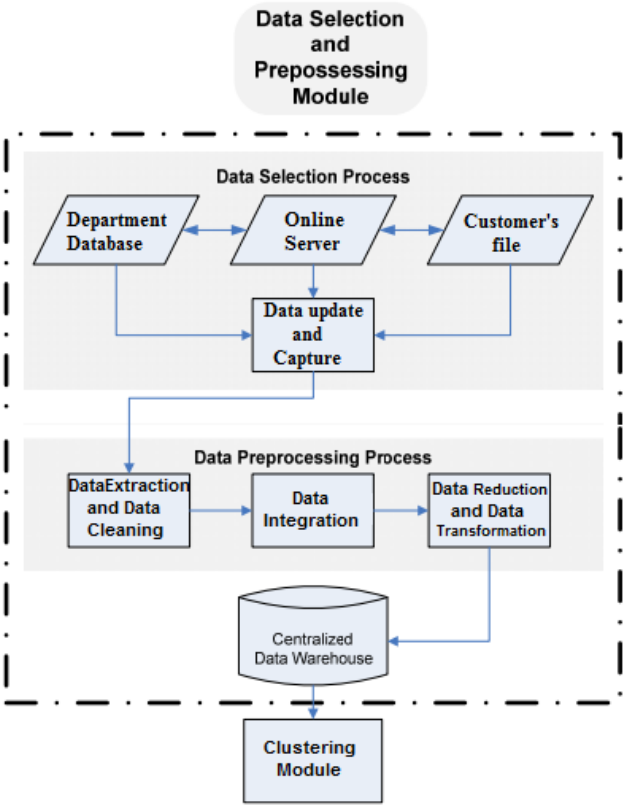


Figure 3.2: Workflow of DSPM [9]

3.4.1 Data Selection Process

The data selection process is used for selecting the data from heterogeneous data sources, including department databases, systems, customer credit files and the company online server. First of all, relevant data is updated in different data sources according to the daily operation. Then, the DSP determines where the data can be collected. For instance, the movement of the

stock market can be collected from the Product and Research Department. On the other hand, market information such as government policy, market trends, and competitors' analysis can be collected from the marketing department. All the updated data is captured and collected. After the relevant data is gathered from various data sources, the next phase is data preprocessing. The data preprocessing plays a significant role in the entire data mining process as it can ensure the quality of the data.

3.4.2 Data Preprocessing

The data preprocessing is an essential part of data mining. The purpose of data preprocessing is to clean selected data for better quality. This is because some selected data may have different formats. This stage enables all necessary information be extracted from the data selection process to have an appropriate format for further mining process. All the data will go through data cleaning to increase the accuracy of the mining result. In general, data cleaning includes filtering, aggregating and filling in missing values. The outliers may be caused by human errors or technical errors. For example, the age of a customer should be "21" but it is recorded as "12". This is likely a human error. Data cleaning can reduce the presence of harmful data such as noisy data, inconsistent data, and missing data that affect the results. All the data will then go through the transformation process for converting the data into appropriate forms for mining. The data transformation can enhance the capability of reading different data. The data transformation can involve different types of normalization. Each attribute has its own initial range. Take income and age as an example.

3.5 Clustering Module (CM)

After the data preprocessing, the data is well prepared for mining. The purpose of CM is to shorten the processing time for RDM. It not only improves the efficiency of the performance but also makes the rules easier to find. The k-means algorithm is applied to partition such data into different groups. If there is any change of data from the data sources, such data will be preprocessed through the DPM, and the CM can then divide them into the most appropriate groups. When clusters are available for processing in the RDM, the RDM can simply generate rules for a particular cluster. Therefore, the RDM does not used to determine which target data

that will generate rules. When there are huge amounts of data, the CM divides the target clusters at first, that enables the procedure of RDM to be more accurate and proceed more easily.

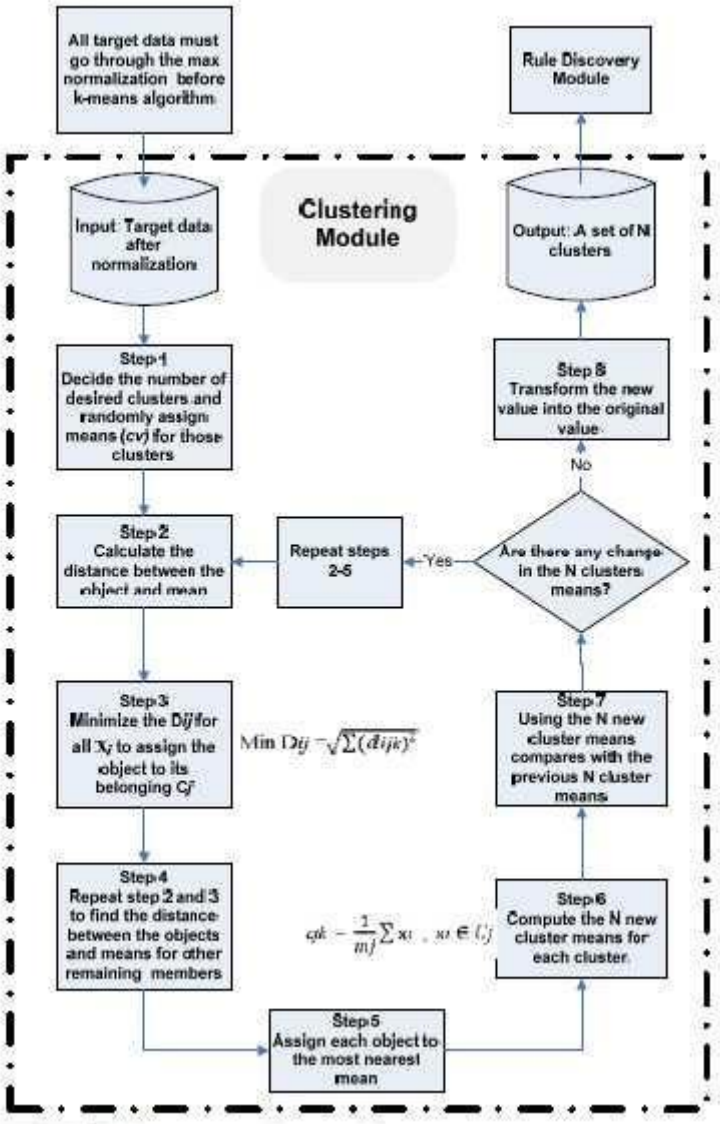


Figure 3.3: The Workflow of CM [9]

3.6 Rules Discovery Module (RDM)

In this stage, the RDM aims at discovering the relationships in a specific cluster. The RDM can directly extract an input data set from the CM to generate useful rules. In this module, the Apriori algorithm (Agrawal et al., 1993) is applied to find the frequent patterns, correlations and associations. Such rules can indicate which groups or sets of items customers are likely to

purchase in a given set of clusters. After the generation of the rules, the rules will allow management to make an evaluation. Then, the sales and marketing department can use such rules for decision making in regard to a specific cluster.

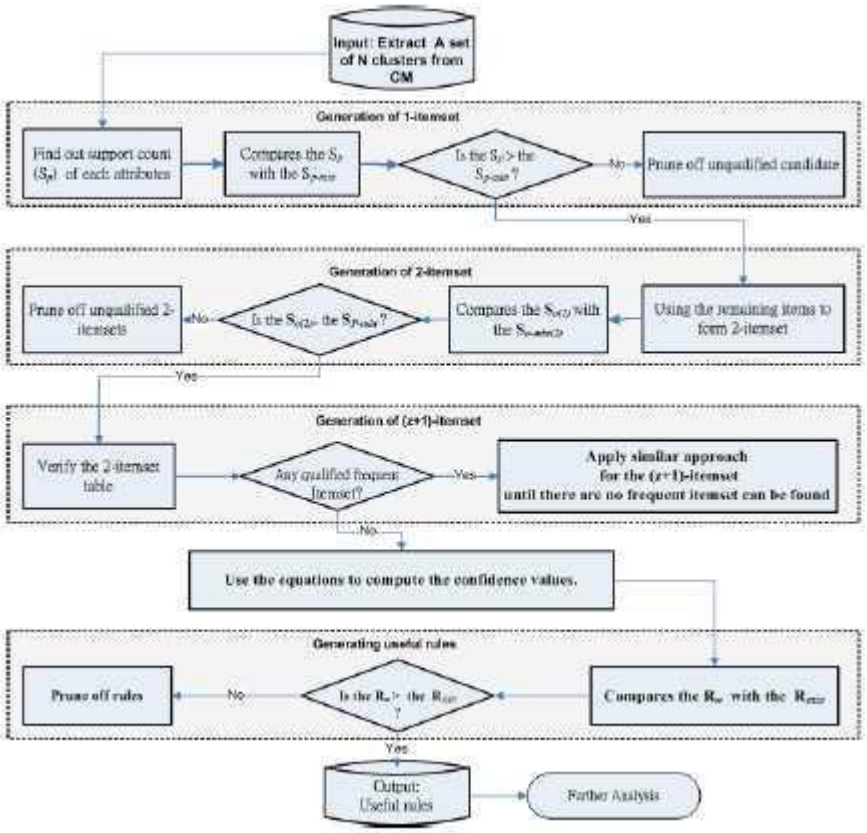


Figure 3.4: Workflow of RDM [9]

It’s working like below

```
from rdm.db import DBVendor, DBConnection, DBContext, AlephConverter
from rdm.wrappers import Aleph

# Provide connection information
connection = DBConnection(
    'ilp',              # User
    'ilp123',           # Password
    'workflow.ijs.si', # Host
    'ilp',              # Database
)

# Define learning context
context = DBContext(connection, target_table='trains', target_att='direction')

# Convert the data and induce features using Aleph
conv = AlephConverter(context, target_att_val='east')
aleph = Aleph()
theory, features = aleph.induce('induce_features', conv.positive_examples(),
                                conv.negative_examples(),
                                conv.background_knowledge())
print theory
```

Figure: 3.4.1 PRDM [10]

(Anze V et al, Matic P et al, Aug 5 2016); Python-RDM Documentation; https://media.readthedocs.org/pdf/rdm/latest/rdm.pdf

3.6.1 Apriori Algorithm

Generally, the Apriori algorithm consists of two phases: mining of frequent item sets and generation of association rules. shows table the notion used in the Apriori algorithm and the steps taken are as follows.

Step 1a: Transform a set of transactions from a cluster. This set of transactions is denoted as T, where Ta represents the ath transaction. The number of transactions is indicated as b, where Ta = {Ta|a = 1, 2…b}. Each transaction consists of

different attributes. Here J denotes the attributes, where p represents as pth attribute. The number of attributes is indicated as q, where Jp= { Jp|p = 1, 2…q}. J is a subset of T. Hence, T = { Jp|p = 1, 2…q}. Note that if an itemset is frequent, any of its subsets is frequent as well. The first step is to use T to form a table in order to find the frequency of occurrence (support count) of different attributes in the transactions. Here the support count of each attribute is denoted as Sp, where Sp = ∑ Sap', p = p'. Sap' is used to indicate the absence (0) or presence (1) of an attribute in a transaction. Here, the predetermined threshold support counts all attributes, and potential item

sets are set to be Sp-min. Step 1b: Sp is compared with Sp-min. A candidate is retained only when it is equal to or greater than the predefined threshold support count (Sp-min). If Sp is smaller than Sp-min, the corresponding candidates will be removed.
Step 2a: Merge the remaining candidates to form an item set with two items. The combination of these two item sets are called L, where Lo is represented as the oath 2- item set. Create a 2-itemset table and find the support count of these 2-itemsets. The support count of 2-itemset is S(2), where So(2) indicates as the support count of the oath 2-itemset combination. Scan for support counts of these 2- item sets (S(2)) in the 2-itemset table.

Step 2b: Compare the support count of these 2-itemsets (S(2)) with the predetermined threshold support count (Sp- min) of the candidates and prune off the unqualified 2- item set candidates. The 2-itemset table only contains combinations which have the support count (So(2)) equal to or greater than the threshold support(Sp-min).

Step 3a: Verify the 2-itemset table as to whether there are any qualified combinations.

Step 3b: If there are still qualified combinations, the algorithm is continued. Apply a similar approach to step 2 to form a table for the (z + 1)-itemset, where z is set to be 2 initially, that is (2+1) as 3 itemset, (3+1) as 4-itemset etc. Then za = zp + 1 for further combinations. za is the number of candidates of the next itemset table and zp is the number of present candidates. Here, the support count of the gth (z + 1)-itemset combination is set to be Sg(z+1). The algorithm is said to be terminated when no frequent itemsets can be found.
In addition to the confidence values, the lift ratio of the rules is used to measure how well the rules generated are for predicting the results as compared to a single itemset. The equations of the lift ratio are shown in Table 4. This is not the end of the algorithm as the incremental data continuously record and enter the algorithm of RDM. The existing data and the newly formed data are loaded into the algorithm for rules mining. After the treatment of the algorithm, some new rules may be found, as the incremental data may have some preferences, customers' requirements etc. These implications point out that some rules need to be evaluated.

Symbol	Description
T	A set of transactions
T_a	The a^{th} transaction
b	The number of the transactions
A	The attributes of the data set.
Ap	The p^{th} attribute ($p = p'$)
q	The number of the attributes
S_p	The support count of each attributes
$S_{p\text{-}min}$	The minimum support count of all attributes and potential frequent itemsets
L_o	The o^{th} 2-itemset
$S_{o(2)}$	The support count of o^{th} 2-itemset
$S_{g(z+1)}$	The support count of the g^{th} (z+1)-itemset combination (A_{ep}, A_{ep}…)
$S_{p\text{-}condition}$	The support count of the 1-itemset condition of the rules for 2-itemset
$S_{ep\text{-}condition}$	The support count of the 1-itemset condition of the rule for (z+1)-itemset
$S_{o(2)\text{-}condition}$	The support count of the 2-itemset condition of the rule for (z+1)-itemset
$S_{g(z)\text{-}condition}$	The support count of the z-itemset condition of the rule for (z+1)-itemset
$S_{p\text{-}result}$	The support count of the 1-itemset result of the rules for 2-itemset
$S_{ep\text{-}result}$	The support count of the 1-itemset result of the rule for (z+1)-itemset
$S_{o(2)\text{-}result}$	The support count of the 2-itemset result of the rule for (z+1)-itemset
$S_{g(z)\text{-}result}$	The support count of the z-itemset result of the rule for (z+1)-itemset
R_w	The confidence value of the w^{th} association rule
R_{min}	The threshold confidence value of all association rules.
I	The lift ratio of rules of all association rules.
I_{mp}	The lift ratio of rules of the mp^{th} association rule.

Figure 3.5: Apriori algorithm notation table [10]

3.6.2 Rules Evaluation

As mentioned above, all the extracted rules must go through the rules evaluation process to assess the feasibility. The rules evaluation needs to acquire and accumulate the knowledge. Rules extracted can be used for reference. Users can adjust the rules on the basis of customers'

requirements, latest market trends, professional industry knowledge and experience. After modification of the rules, data are formed and loaded continuously. Hence, the entire proposed model is said to be worked continuously, without termination. This allows knowledge to have continuous improvement and results in high customer satisfaction.

3.6.3 Rules Classification

The purpose of the RDM aims at discovering the useful relations in investment behavior in the financial industry. Such useful rules can be used for understanding investment behavior. For example, the investment behavior of younger customers who are willing to take more risky financial products product as compared to the older investors. The Product and Research Department manager can use such useful rules to devise the coverage of the portfolio to be included and what types of customers should be promoted. Another type of rule is the trivial rules. It seems to be known by common sense and the influence of such kind of rules are minimal. For instance, the higher risk of a financial product has a higher return. Thus the rules extracted can just show the evidence but not a new idea in an improvement. The last one is the inexplicable rule. Some results are not easily interpreted by the rules.

Confidence Value (R_w) for 2-itemset candidates.	
(with 1-itemset condition)	Confidence(R_w) = $S_{o(2)}$ / $S_{p\text{-}condition}$
Confidence Value (R_w) for (z + 1)-itemset candidates.	
(with 1-itemset condition)	Confidence(R_w) = $S_{g(z+1)}$ / $S_{p\text{-}condition}$
(with 2-itemset condition)	Confidence(R_w) = $S_{g(z+1)}$ / $S_{o(2)\text{-}condition}$
(with z-itemset condition)	Confidence(R_w) = $S_{g(z+1)}$ / $S_{g(z)\text{-}condition}$

Generating Association Rules

Lift Ratio of the Rules (I_{mp}) for 2-itemset candidates.	
(with 1-itemset result)	Lift ratio (I_{mp}) = $S_{o(2)}$ / $S_{p\text{-}result}$
Lift Ratio of the Rules (I_{mp}) for (z + 1)-itemset candidates.	
(with 1-itemset result)	Lift ratio (I_{mp}) = $S_{g(z+1)}$ / $S_{up\text{-}result}$
(with 2-itemset result)	Lift ratio (I_{mp}) = $S_{g(z+1)}$ / $S_{o(2)\text{-}result}$
(with z-itemset result)	Lift ratio (I_{mp})) = $S_{g(z+1)}$ / $S_{g(z)\text{-}result}$

Equations of Lift Ratio

Figure 3.6: Generating Association Rules & Equation of Left Ratio [10]

3.7 Case Study

3.7.1 Company Background

To validate the feasibility of FDMM, a case study is conducted in a financial company. Convoy Financial Services Limited was founded in 1993. It is wholly owned by Convoy Financial Services Holdings Limited (CFS), and is an independent insurance and MPF schemes brokerage broker firm in Hong Kong. Convoy provides a wide range of financial products including investment linked assurance schemes, insurance products and Mandatory Provident Fund (MPF) schemes. To provide the best suited financial products and services, Convoy insists on communication with its clients and business partners. In order to build an effective communication channel between clients and product providers, it has set up close relationships with over 18 product providers. Convoy also offers a variety of independent financial services customized to the needs of its clients, which includes providing clients' financial needs analysis tailored financial solutions, regularly reviewing and managing clients' plans. With the vision to become the best of Hong Kong Best Company for Financial Planning Excellence (IFA), Convoy promises to improve its services by achieving its mission- Respect, Care, Lead and Contribute to their clients. Thus, Convoy acts in the clients' best interest to meet clients' financial objectives.

3.7.2 Challenges Faced by the Company

To maintain competiveness in a challenging environment, Convoy Limited needs to concentrate on the quality of products and services in order to increase the customer satisfaction. In the existing workflow of the company, a customers' enquiry can be collected from different sources including the online customer zone, walk-in customers and the sales department. Each department only stores its desired data in its own database. Thus,
there is a lack of information sharing among the company. The customer services representative will then handle the basic customer enquiry. After that, the customer representative will transfer the enquiry information to the financial consultant. The financial consultant will prepare the relevant information on the basis of customer enquiry. Then, the financial consultant will meet with the customer to understand the actual needs. In general, the prepared information is not always useful. Since the customers' needs might be different to their original enquiry. Therefore, the financial consultant usually needs to communicate deeply with the customers to gain more

understanding of the needs. When the customer is satisfied with the investment plan, the financial consultant will create a customer profile, and such a profile record will be stored in the company's database. The financial consultant will then send the information of the customers' options to the product providers. If the customer is not satisfied, the financial consultant will follow up the case.

3.8 FDMM Implementation

In order to implement FDMM in Convoy Limited, a prototype based on the system architecture in Figure 1 was developed in XLMinerTM (Shmueli et al., 2007). XLMinerTM is a simple and user friendly data mining add- in for Excel. Based on the case information of Convoy Limited, it is found that the major problems of the company are poor information sharing and data management. The company does not apply any data mining tools to manage their business data. Therefore, it is difficult to identify the patterns and relationships in the data set. Faced with overwhelming amounts of business data, it needs a discovery-driven data analysis technology to improve data management. In order to solve these problems, the company needs to consider following objectives: (a) to improve the information sharing in the organization; (b) to enhance the data management; and (c)
to increase customer satisfaction. The company can follow these objectives and then find out the best solutions for improving the existing workflow of the company.

Period	Total number of customers (65) (Cluster 1)	Total number of customers (52) (Cluster 2)
1-3 years	2	8
3-5 years	4	14
5-10 years	7	21
10-20 years	13	5
20 or above	39	4

Investment Horizon

Code	Name	Description
S1	Dynamic Evergreen (Medium risk)	Seek to achieve returns over an 18-months term
S2	Dynamic Growth (Medium to high risk)	Seek to achieve competitive long term capital growth
S4	Global Opportunity (High risk)	Seek to achieve returns in the medium to long term.

Descriptions of Investment Linked Assurance Scheme

Figure 3.7: Investment Horizon & Descriptions of Investment Linked Assurance Scheme [10]

	A	B	C	D	E
1	Client ID	Age	Income	Investment Experience	Portfolio
2	001	26	30,000	4	Life Insurance, MPF, China Growth Fund
3	002	30	63,000	0	Fixed Deposit, Currency Linked Deposit, Technology Stock
4	003	40	45,000	3	Currency Linked Deposit, Global Equity Fund, Life Insurance
5	004	24	23,000	0	Technology Stock, Korea Equity Fund, MPF
6	005	33	32,000	8	Emerginign Market Equity Fund, Currency Linked Deposit
7	006	52	66,000	3	Fixed Deposit, Material Stock, MPF, Life Insurance
8	007	38	38,000	5	Energy Stock, China Growth Equity Fund, Medical Insurance, MPF
9	008	25	28,000	1	Stocks, Currency-linked Deposits, Mutual Funds, MPF
10	009	46	55,000	6	Fixed Deposit, Stocks, Insurance, MPF
11	010	29	26,000	1	Stocks, Insurance, Currency-linked Deposit, MPF
12	011	30	29,000	9	Stocks, Mutual Funds, Insurance, MPF
13	012	44	60,000	7	Mutal Funds, Insurance, Currency-linked deposit
14	013	57	49,000	13	Fixed Deposit, Cash Reserve, Insurance
15	014	48	56,000	7	Fixed Deposit, Currency linked Deposit, Insurance, Malaysia Equity Fund, MPF
16	015	21	14,000	0	Cash Reserve, Insurance, Pacific Technology Equity Fund, MPF
17	016	29	36,000	10	Fixed Deposited, Investment Linked Insurance, Energy Stock
18	017	50	49,000	7	Cash Reserve, Russia Equity Fund, MPF, Life Insurance
19	018	35	29,000	5	Material Stock, Emerging Europe Fund, Global Equity Fund
20	019	36	60,000	8	Investment Linked Insurance, Technology Stock
21	020	36	98,000	13	China Growth Fund, Korea Equity Fund, Technology Stock

Raw Data from Convoy Limited

Figure 3.8: Raw Data Convoy Limited [10]

3.8.1 K-means Clustering Phase

Phase The objectives of FDMM are to identify the influencing factors on investment and gain further understanding of investor behavior. Thus, basic customer information, including age, experience and the income, is used for analysis. The reasons for selecting these three variables are that these three variables mainly influenced investment behavior. Many studies reveal that experience is the most significant factor influencing investing behavior. Also, the asset allocations of each portfolio depend on the income of the clients. In order to prove that higher income can drive more diversification of the choices of the portfolio, the income variable is selected. In addition to income, age is considered as a variable. Since the categories of the portfolio mainly cover the life insurance and MPF, age is an influencing factor that affects the asset allocation. These two products might be varied due to differences of age. Figure 5 shows the data set, in which each row represents a customer profile. In K-means clustering phase, customers are segmented into appropriate groups based on their characteristics. Considering the numerical measurement of the K-means algorithm, only the interval variables (age, income and investment experience) are selected (i.e. the portfolio is ignored). However, it would be used for generating the useful rules in a later part. By defining the cluster as 2 and using the "K-Means

Clustering" function in XLMinerTM, interpreting the results of the two clusters are highlighted as follows:

Cluster 1: Table 5 shows the investment horizon of both clusters. Over a half the customers in this group are more likely to seek a long term investment horizon (20 years or above). To achieve a long term capital growth, this group tends to select riskier portfolios. It implies that such customers can tolerance a high risk. Table 6 shows the description of investment linked assurance schemes

offered by the company. It can be seen that there is a significant proportion of customers prefer choosing the S2 investment and almost a half the customers selected S4. These two types of investment style are riskier compared with S1. Thus, this group can be considered more aggressive than

cluster 2. The results provide an insight into the marketing implication of this group. The company can identify such a group as potentially valued customers. Cluster 2: This cluster indicates that the higher age of customers means higher investment experience. The past investment experience might provide such customers with more risk awareness. Thus, this group tends to diversify their portfolios. The main bulk of the portfolios consists of more than 75% of bond funds such as global bond, US bonds, and they are expected to seek steady and slow returns. Therefore, the mainly selected type of portfolio is S1. In addition to bond funds, this group mainly selects life insurance and medical insurance. Thus, the company can offer a conservative portfolio to this group of customers.

Types of Attributes Extracted from Cluster 1's Portfolio (Input)		
Name of Attributes	**Risks level**	**No of Customer Selected**
Global Equity Fund	Medium	2
Korea Equity Fund	Medium	24
Russia Equity Fund	High	2
Asia Pacific Equity Fund (Exclude Japan)	High	1
Emerging Europe Fund	High	3
Global Emerging Markets Equity Fund	High	32
Material Stock	Medium	28
Saving Insurance	Low	3
Malaysia Equity Fund	High	9
Global Technology Fund	High	12
Taiwan Equity Fund	High	25
Emerging Markets Bond Fund	Low	11
Global Bond Fund	Low	5
Fixed Income	Low	5
Currency Linked Deposit	Low	3
Life Insurance	N/A	6
China Growth Fund	Low	24
Medical Insurance	N/A	2
Investment Linked Insurance	Medium	2
MPF	N/A	49
Technology Stock	High	53
Energy Stock	High	37

Investment Products Selected from Cluster 1

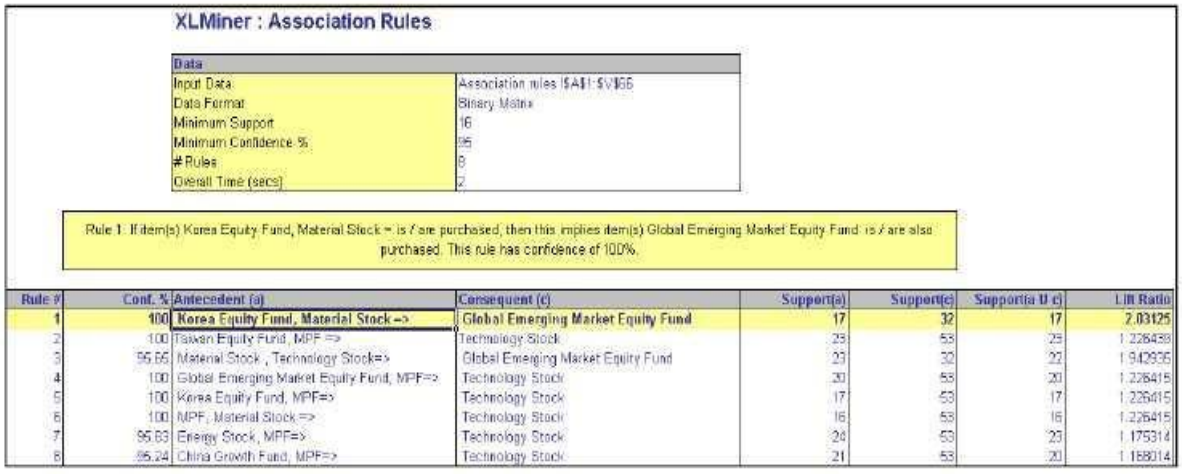
Figure 3.9: Investment Products Selected from Cluster 1 [10]

3.8.2 Association Rules

This module aims at applying association rules to discover the hidden patterns in the clusters segmented. For illustrative purposes, cluster 1 is selected to show how to discover unexpected rules in the XLMinerTM. This is because cluster 1 indicates the most similar variable behavior compared with cluster 2. If useful rules are extracted from this cluster, it can help the company to customize the best suited portfolios for such customers, as well as achieving high customer satisfaction. the list of investment products selected in the cluster 1. There are 7 types of investment products in cluster 1, including equity funds, bond funds, stocks, fixed deposits, currency linked deposits, MPF and insurance. Each product indicates the level of risk. Over half the products are rated as higher risk, with about a fifth of products rated as low risk. This means that this group tends to mainly select higher risk products for their portfolios. It is justified as the

implications found in the previous part show that this group is aggressive in taking more risk in order to achieve returns in medium to long term capital growth.

Similar to the clustering analysis, the "Association Rules" function of XLMinerTM is applied to interpret the results. Figure 6 shows that there are eight useful rules generated. As shown in the figure, five rules (rule 1, 2, 4, 5, and 6) have a confidence level of 100%. It implies that such rules are highly reliable in indicating the success rate of investment decision making, based on these rules. All rules' lift ratios are greater than 1. It provides an insight into the prediction so as to increase the probability of the "THEN" (result) and the "IF (condition) parts. It also indicates that all items in the generated rules are positively correlated with other items. Take rule 1 as an example, the Korea equity fund and material stock have a positive correlation with the global emerging markets equity fund. The support of the items also provides marketing implications. There are 53 observations of technology stocks in the cluster 2, which is 81.54% of total observations. All the rules contain technology stock. This means that this group will buy technology stocks under particular conditions. A strong relationship between MPF and technology stocks is explained by rule 2, 4, 5, 6, 7, and 8. (IF the clients select MPF, THEN they will select technology stocks as well). This gives information to the company so as to facilitate decision making based on these rules.



XLMiner : Association Rules

Data	
Input Data	Association rules!$A$1:$V$65
Data Format	Binary Matrix
Minimum Support	16
Minimum Confidence %	95
# Rules	8
Overall Time (secs)	2

Rule 1: If item(s) Korea Equity Fund, Material Stock = is / are purchased, then this implies item(s) Global Emerging Market Equity Fund: is / are also purchased. This rule has confidence of 100%.

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	Korea Equity Fund, Material Stock =>	Global Emerging Market Equity Fund	17	32	17	2.03125
2	100	Taiwan Equity Fund, MPF =>	Technology Stock	23	53	23	1.226439
3	95.65	Material Stock , Technology Stock=>	Global Emerging Market Equity Fund	23	32	22	1.942935
4	100	Global Emerging Market Equity Fund, MPF=>	Technology Stock	20	53	20	1.226415
5	100	Korea Equity Fund, MPF=>	Technology Stock	17	53	17	1.226415
6	100	MPF, Material Stock =>	Technology Stock	16	53	16	1.226415
7	95.83	Energy Stock, MPF=>	Technology Stock	24	53	23	1.175314
8	95.24	China Growth Fund, MPF=>	Technology Stock	21	53	20	1.168014

Figure 3.10: Summary Report of Association Rules [10]

CHAPTER 4
Competencies and Smart Plan

4.1 Competencies Earned

Skills Knowledge as a result or outcome is a statement of what the student is expected to know, understands, or may lead to as a result of the learning process. Installation and clustering data and combination machine learning with R language. The key role of advanced learning & understanding R programing, entering commands, descriptive statistics outcomes, loading the dataset, and internships. The Student Association office funds many student Intern Learning Outcomes: Gathering and organizing information in practice Project Description: Take part in the unstructured large amount of data to structured data. Learning Objectives Run and data combination with machine learning; Install, configure, and providing good data.

4.2 Smart Plan

Every company should have an intelligent plan to succeed. Basically, some of the ordinary things combine an intelligent plan. How to make a large amount of data deployed? Each company must have a specific plan that helps them rise above. Find out your problems and solve them. Be able to take any kind of risk.

4.3 Reflections

Z Soft Solution Ltd. started operations in 2015. In recent years, they have offered their work and service offering on the basis of customer recommendations and have taken into account the time requirements. They have been working with many international projects and an small portion of works for national organizations and have received a reputation. They use the latest technology and update the services they need. This is the Corporate Data Science Solution that provides advanced Machine Learning and Artificial Intelligent with a highly technical team of experts.

CHAPTER 5

Conclusion and Future Career

5.1 Discussion and Conclusion

After the financial crisis of 2014 and the global crisis of 2015, investors are becoming more cautious towards investments, especially in high risk financial products. These financial issues make it more difficult to devise a portfolio. In fact, customer orientation is becoming a trend in today's business. Many companies want to understand customers' needs, requirements and preferences in order to achieve high customer satisfaction. Customer satisfaction relies on superior products and services that the company provides. To customize the products and services, a company needs to gain more understanding of customer behavior. However, many companies lack a decision support system. In addition to providing superior products or services, many companies are facing the challenge of handling a huge growing amount of data in daily transactions. Attempting to address these challenges, the aim of the paper is to develop an intelligent Financial Data Mining Model (FDMM) that can help financial companies to tackle the problems. The fundamentals of FDMM are that firstly, all the relevant quality data are collected and preprocessed through DSPM, so each department can share information across the organization. Secondly, the CM is developed to partition the customers into specific groups. The segmented groups provide marketing implications to the sales managers so as to develop customer values for highly profitable customers. Thirdly, the RDM is developed to generate useful rules for the target clusters. The unique characteristic of RDM is that it will keep going as newly formed data are continually loaded into RDM for rules mining. The new data may be of great interest for the company. The useful rules can be integrated with industry knowledge, experience, customer needs, etc. Finally, the introduced data mining software - XLMinerTM - not only validates the proposed method but also provides a road map for implementing FDMM in real world practice. With the fundamentals of data mining methods and software implementation, a company can discover the patterns on how to build a customized portfolio based on the rules extracted and can learn more about investors' behavior in financial markets. This can effectively support the company for a long term sustainable success development.

5.2 Scope for Further Career

There are career opportunities in different areas of Data Scientist. Professional opportunities in Data Mining not only mention the Data Science and Machine Learning platform, but cover many different areas, such as Structured the large amount of data, Decision Making, and Artificial Intelligent Nowadays, many companies have moved to Data Science and Machine Learning. A company such as Google, IBM, SpaceX, Facebook, Microsoft, SAP, Cognizant Technology, Tata Consulting has moved many companies to a data science and machine learning solution. Data Mining, Data Science and Machine Learning are the most sexist job of the 21 century.

References and Biography:

[1] Get idea about Z Soft Solution, Available at <<https://zsoftsolution.co.uk/about-us.php>>, last accessed on 11September 2018,10.30pm.

[2] Mark K.Y Mak; A Financial Data Mining Model, Sage Journal; Jan 01 2011; Available at <<https://journals.sagepub.com/doi/full/10.5772/50937>> last accessed on 12 September 2018, 08.30pm.

[3] George S; The 5 Clustering Algorithms Data Scientists Need to Know; Feb 5 2018; <<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>> last accessed on 13 October 2018,09.30pm.

[4] (Anze V et al, Matic P et al, Aug 5 2016); Python-RDM Documentation; <<https://media.readthedocs.org/pdf/rdm/latest/rdm.pdf>> last accessed on 13 October 2018,10.30pm.

[5] About Recommended Partitioning Scheme, Available at: Agrawal, R., Imielinski, T., Swami, A.N., 1993, Data mining: A performance perspective, IEEE Transactions on Knowledge and Data Engineering, vol. 5, no. 6, pp. 914–925.

[6] Get Concept about Data Mining, Available at:

Berry, M.J.A., Linoff, G.S., 2004, Data mining techniques for marketing, sales, and customer relationship management, Ind: Wiley, Indianapolis.

[7] Get Concept Available at:

Liang, D., Christos, T., 2003, Experiences of using a quantitative approach for mining association rules, in: Lecture Notes Computer Science, vol. 2690, pp. 693- 700.

[8] Get Concept Available at:

Ngai, E.W.T., Xiu, L., Chau, D.C.K., 2009, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, Vol. 36, no. 2, pp. 2592-2602

[9] Get Figure Available at: Vazirgiannis, M., Halkidi, M., Gunopulos, D., 2003, Uncertainty handling and quality assessment in data mining, Hong Kong: Springer, London.

[10] Get Figure Available at: Kwan, I.S.Y., Fong, J., Wong, H.K., 2005, An e-customer behavior model with online analytical mining for internet marketing planning, Decision Support Systems, vol. 41, no. 1, pp. 189-204.

Appendices

Appendix A: Internship Reflection

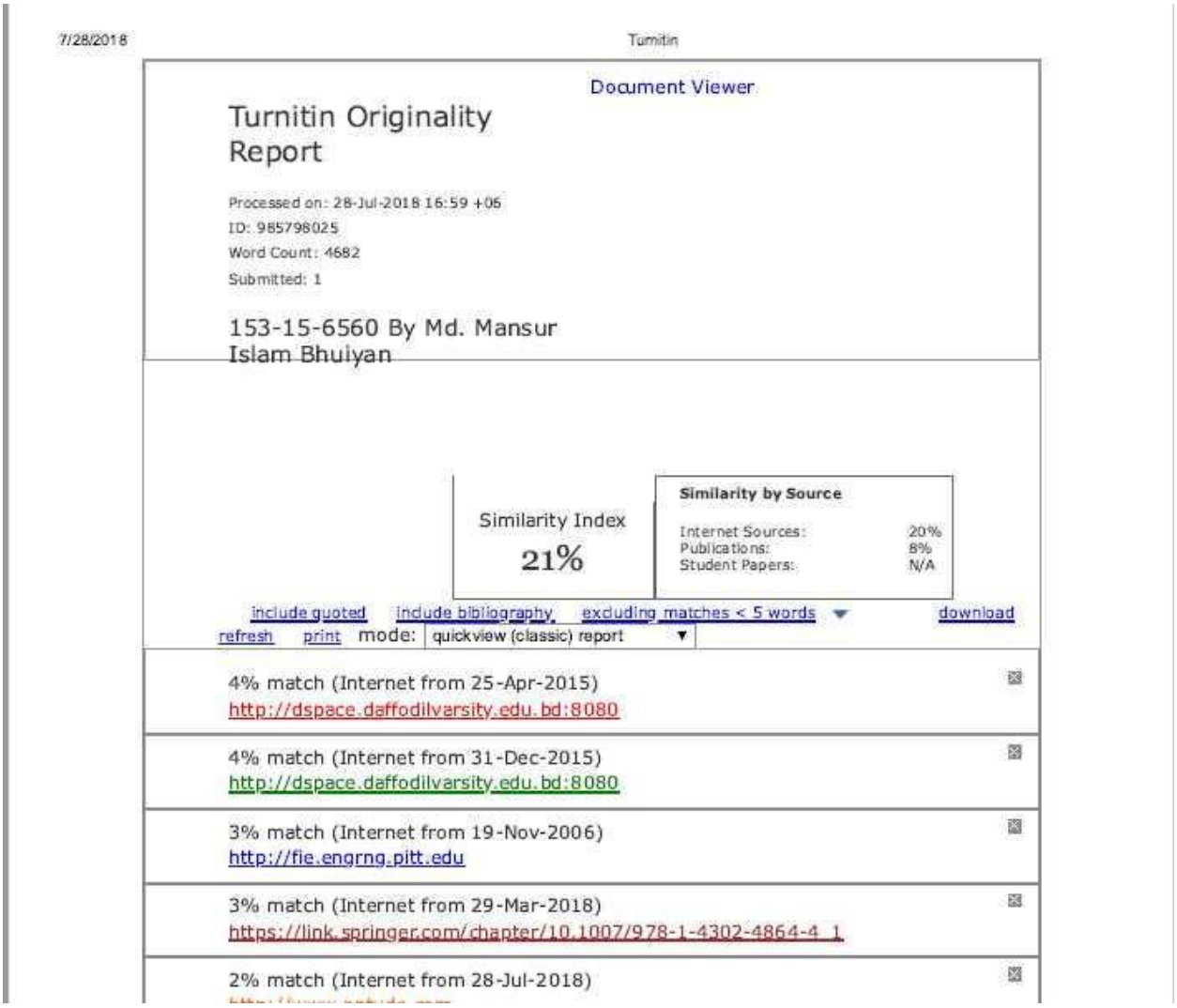
The main objective of internship in a professional environment, practical solutions to Global problems involve applying the information in the classroom. Appropriate skills and relationships in the professional environment, learn to master new knowledge, skills, and ability to decide on refining and developing plans. Add to the network with other professional managers and professional relationships. To do the exercises I use correctly business etiquette. Mission / vision for establishing a humanitarian organization, how to contacting colleagues, how power is shared, how it is structured, how decision share, how to understand the professional organization's culture and how much responsibility and feedback to the organization. Assessment after internships and internships personally conferences offer the opportunity to make professional opinions. Internship experience preparing to live in a global society, leadership and service, my gift to practice it.

Appendix B: Company Detail

Head Office

Name	Z Soft Solution Limited
Address	123 Northmoor Road Greater Manchester, M12 5RS, United Kingdom
Telephone	44 (0)161 818 7559
Fax	880-2-8116103
E-mail	info@zsoftsolution.co.uk
Website	www.zsoftsolution.co.uk
Type of Organization	Data Industry
Employees	56

Plagiarism Report



7/28/2018 Turnitin

Document Viewer

Turnitin Originality Report

Processed on: 28-Jul-2018 16:59 +06
ID: 985798025
Word Count: 4682
Submitted: 1

153-15-6560 By Md. Mansur Islam Bhuiyan

Similarity Index
21%

Similarity by Source

Internet Sources:	20%
Publications:	8%
Student Papers:	N/A

include quoted include bibliography excluding matches < 5 words download
refresh print mode: quickview (classic) report

4% match (Internet from 25-Apr-2015)
http://dspace.daffodilvarsity.edu.bd:8080

4% match (Internet from 31-Dec-2015)
http://dspace.daffodilvarsity.edu.bd:8080

3% match (Internet from 19-Nov-2006)
http://fie.engrng.pitt.edu

3% match (Internet from 29-Mar-2018)
https://link.springer.com/chapter/10.1007/978-1-4302-4864-4_1

2% match (Internet from 28-Jul-2018)