Data Mining Module Code: B9DA103 (CA1)

Submitted by:
Prasad Tambe(10515513)

Table of Contents

A) Critique Analysis on CRISP-DM model

1.1Introduction:

There has been an increase in demand for how the data is been structured using different data modeling techniques. CRISP-DM is being the most favored methodology for data mining.

Data Mining is also called “Knowledge Discovery of Data”, where it is the process to determine and analyze some advanced, critical information/Patterns from a data set (which can be small or large) and the conclusion can result in the storing data for futurity use and can be developed for a new and good use.

The following critique is being written with reference to this article. Colin Shearer in 2000 was the author who published this article, where the article recites about how non-proprietary, augmented defined for developing and resulting in improved and superior-conclusion for the data mining called “CRISP-DM known as Cross-Industry Standard Process for Data Mining. [Shearer,2000].

The main goal behind developing the tool, industry, operational neutral tool which was led by four companies as a data mining group: DaimlerChrysler, Integral Solutions Ltd., NCR, and OHRA in 1996[Shearer,2000].

So according to the information and fine-points which are mutually shared on this article, I can say that there are many advantages as it is in favor from the past 20 years and one of the preferred ones for Data Scientists Projects, Data Mining and Prediction Study, but there are certain inconveniences as well as benefits which makes it a bit neutral.

1.2 Summary:

Being précised, by reference to the article in detailed of the CRISP-DM Model which Subsists of 6 Phases which include “Business Understanding”, “Data Understanding”, “Data Preparation”,” Modelling”,” Evaluation”,” Deployment” with the tasks and the outputs gained which is resulting with it. It generally summarizes the brief stop about the evolution and the author where characterizes the type of Data Mining Techniques, which helps to sort out many business problems and resolves it.

1.3 Critique Evaluation:

The CRISP-DM constructs 6 phases of the data mining process: "business understanding," "data understanding," "data preparation," "modeling," "deployment,“evaluation."[Shearer,2000].

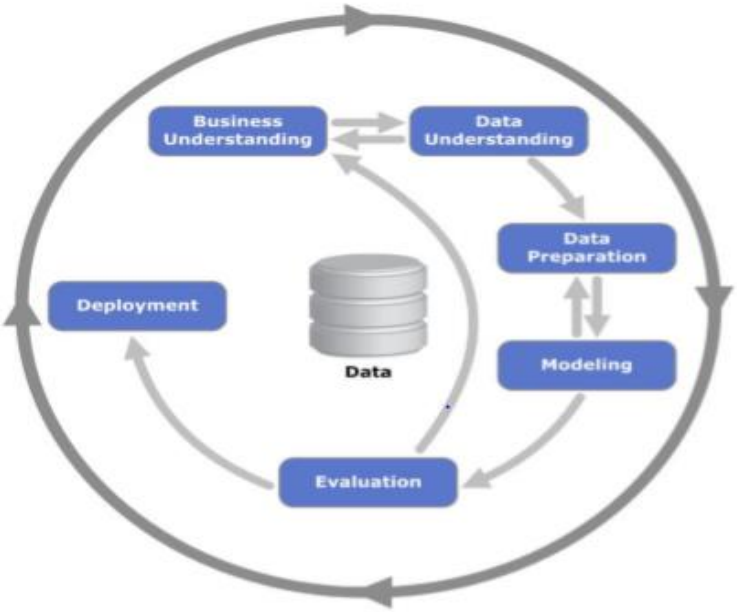
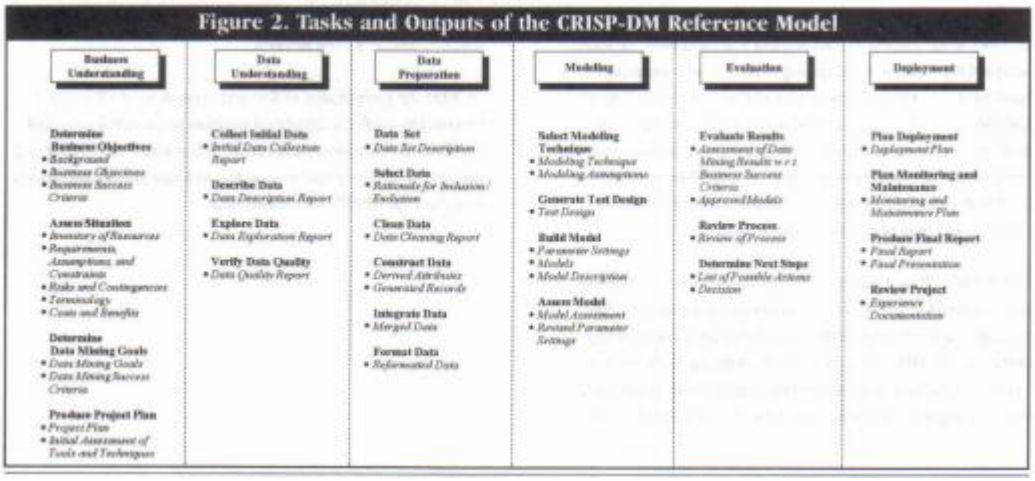


Fig 1. Phases of CRISP-DM Reference Model



1.3.1 First Phase:

1.3.1.1 Business Understanding: -

To begin any project at the start we need to understand the business requirements.
While understanding the business requirements it is important to clear all the critical factors which may affect the business project and also the background objective. Business understanding is broken into various aspects.

- Here, we need to interpret what are the intentions of the client.
- The factors play an important role as to how it may affect the business.
- It also depends on the success criteria of how the business objectives are regulated.
- There shouldn’t be a case where the results generated are inaccurate regardless of the business objective and solutions.
- We need to find out the inventory resources from personnel to software which is applicable to support the data mining project and also, we need to list the assumptions and constraints which are required.
- And also, we need to list all the project issues and possible solutions to the issues.
- Hence, we need to create a summary of business and data mining tools that constructs cost benefits reasoning for the project.
- Data mining success criteria are assumed for achieving a level of accuracy for the project.

- If the business goal cannot be completely rendered into a data mining tool.it may be a rational way to reconsider the issue of the business project.
- A project plan is needed to characterize the calculation plan for gaining the data mining goals which includes an assessment of potential risks and outlining specific steps for a proposed timeline.
- Assessments of the tools and techniques which are used are important in producing a proper project plan. [Shearer, C., 2000]

1.3.2 Phase Two:

1.3.2.1 Data Understanding:

- The analyst should be familiar with the data and discover the initial visions and data quality problems.
- Integrating and storing data is necessary while collecting the initial data.
- After collecting, the initial data the next step provides a basic knowledge of the data on which proper steps will be built.
- Exploring the data addresses data mining problems that can be resolved by visualizing, querying, and reporting.
- It is necessary to verify the data quality as the data isn't incomplete or there are no missing values as such and review the attributes. [Shearer, C., 2000]

1.3.3 Phase Three:

1.3.3.1 Data Preparation:

- Data Preparation: This step consists of 5 steps which create the final data set to be fed into the modeling tool.
- Select data: This step concludes which attributes are important.
- Clean Data: At this stage, the analyst must either select a clean subset of data or preferable techniques for identifying the missing data.
- Construct data: A derived data is needed for producing generalized records.
- Integrate data: Here, we are combining data from various tables and records to create new values.
- Format Data: This step concludes making data suitable for a specific modelling tool. [Shearer, C., 2000]

1.3.4 Phase Four:

1.3.4.1 Modelling:

- Various modeling methods are chosen and applied, calibrating their parameters to optimum values. There's few modelling steps that include.
- Select Modelling Technique: We choose one or more modeling techniques that reassumed which should be used for modelling.
- Generate Test Design: The test model's quality and validity are separated by train and test sets.
- Build Model: after testing, the analyst runs the modeling tool on the data set to generate more records.
- Assess Model: The model is ranked with the desired text design and data mining success criteria. [Shearer, C., 2000]

1.3.5 Phase Five:

1.3.5.1 Evaluation:

- In this stage, the model is evaluated and reviewed using two steps namely evaluate results and review process.
- The project leader should decide whether to use the model or move back to previous stages.
- A data analyst should be ambitious and determined the next steps using the success criteria. [Shearer, C., 2000]

1.3.6 Phase Six:

1.3.6.1 Deployment:

- The whole information gained must be taught and presented to the customer so that he /she can smoothly run the process.
- There are four main steps in the phase:

✓ Planning deployment:
✓ Planning monitoring and maintenance.
✓ Producing a final report
✓ Reviewing the project [Shearer, C., 2000]

1.4 Conclusion:

I hereby conclude that when you google ‘CRISP-DM’ would beneficial for both academic and industry usage because it breaks down every phase into the root level which would in turn help to meet your business objective. Hence, I have concluded that data mining data science now and twenty years back is much more goal focused and applies to the process. CRISP-DM holds a pullback in the data mining process but it needs to be monitored more as it lacks maintenance.

For example, Tourism recommender: A potential pathway is to create a recommendation system of position and operation as soon as the requirement is met in the discovery phase of the target. A potential roadmap for the implementation of a location and activity recommendation framework (Section 4.1) may suggest that, once the goal is established as a first step(goal exploration), the organization will want to use the position and activity history of the users as relevant data (data value exploration) from the data obtained from the services and networks based on the location of third parties.[

Figure4.1 shows the process for the recommendation process.

Then, the data preparation activity begins to process using the tensor which creates user-activity which can we equip and train the recommendations .when one of the preferred models are selected and evaluated, the company would recommend the most preferred location from the user (product exploration), through a reporting or developing a mobile app/web.[Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J.H., Kull, M., Lachiche, N., Quintana, M.J.R. and Flach, P.A., 2019.]

Another example includes the CRISP-DM approach on evaluation which is also known as a classifier.

Indoor positioning access is established using fingerprinting sensors and signal models. The desired position is regulated by data mining techniques and heuristic for Indoor positioning methods for fingerprinting. The evaluation was

based on the CRISP method which is a traditional way of determining a classifier's accuracy. Every misclassification was regarded as the same as in the CRISP process. Statistical indicators such as duration, recall and accuracy may be calculated using the confusion matrix.in these 5 practically well-known classifiers are compared: Decision Tree, k–NN, Rule Induction, Naive Bayes, and Artificial Neural Network (ANN) classifiers.

TABLE I
CONFUSION MATRIX

Predicted	Actual			Precision
	C_1	C_2	C_3	
C_1	.	.	.	%
C_2	.	.	.	%
C_3	.	.	.	%
Recall	%	%	%	

Figure1.1 Confusion Matrix

TABLE II
SUMMARY OF TESTED CLASSIFIERS

Name	Accuracy in %	Miss	Close	Far
ANN	96.77	5	4	1
3NN W	92.26	12	8	4
9NN W	92.26	12	8	4
Naive Bayes 1 kernel	91.61	13	10	3
9NN	90.97	14	8	6
5NN W	90.32	15	10	5
11NN W	90.32	15	12	3
13NN W	90.32	15	11	4
13NN	90.32	15	11	4
1NN W	89.68	16	12	4
5NN	89.03	17	11	6
Naive Bayes	87.10	20	15	5
Decision Tree	84.52	24	11	13
ID3	83.87	25	15	10
Rule Induction	80.65	30	18	12

Figure 1.2 Concludes the Summary of Tested Classifiers

TABLE III
CONFUSION MATRIX OF 1ST CASE SELECTED ZONES WITH 9NN CLASSIFIER

Actual	Predicted						Total Result
	1st Floor West Corr.	Lab 102	Lab 103	Lab 104	Other Close	Far	
1st Floor West Corr.	**4**				1	1	6
Lab 102	1	**0**	1	1			3
Lab 103			**10**	1			11
Lab 104				**6**			6
Total Result	5	0	11	8	1	1	26

Figure 1.3 Confusion Matrix with 9NN Classifier.

TABLE IV
CONFUSION MATRIX OF 1ST CASE SELECTED ZONES WITH NAIVE BAYES CLASSIFIER

Actual	Predicted						Total Result
	1st Floor West Corr.	Lab 102	Lab 103	Lab 104	Other Close	Far	
1st Floor West Corr.	**2**				3	1	6
Lab 102		**3**					3
Lab 103			**11**				11
Lab 104			3	**3**			6
Total Result	2	3	14	3	3	1	26

Figure 1.4 Confusion Matrix with Naïve Bayes

TABLE V
CONFUSION MATRIX OF 2ND CASE SELECTED ZONES WITH 9NN CLASSIFIER

Actual	Predicted					Total Result
	2nd Floor East Corr.	2nd Floor North Corr.	2nd Floor West Corr.	Lecture Hall 205	Far	
2nd Floor East Corr.	**8**				1	9
2nd Floor North Corr.		**2**				2
2nd Floor West Coor.			**9**			9
Lecture Hall 205				**6**		6
Total Result	8	2	9	6	1	26

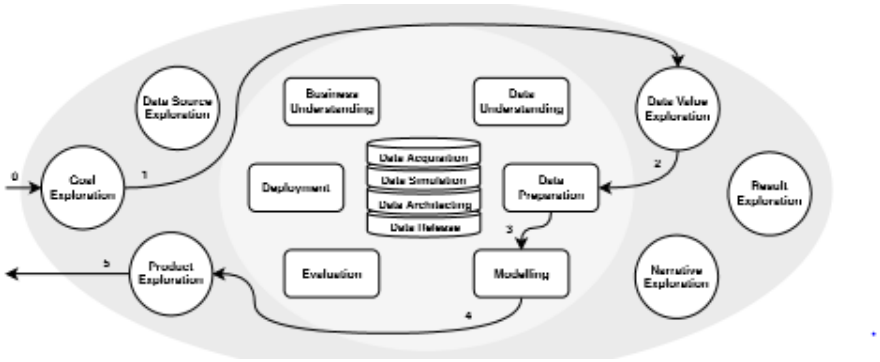
Figure 1.5 Confusion Matrix with 9NN Classifier.

TABLE VI
CONFUSION MATRIX OF 2ND CASE SELECTED ZONES WITH NAIVE BAYES CLASSIFIER

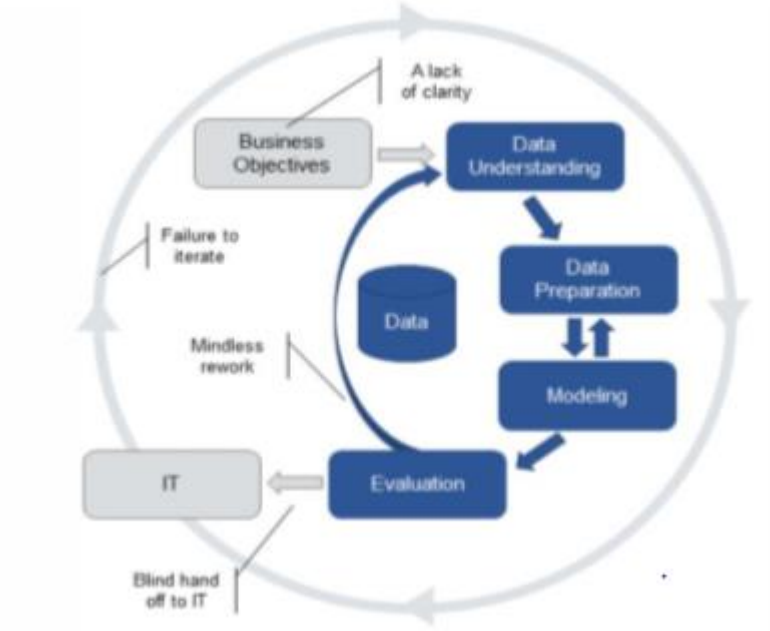
Actual	Predicted					Total Result
	2nd Floor East Corr.	2nd Floor North Corr.	2nd Floor West Corr.	Lecture Hall 205	Far	
2nd Floor East Corr.	**7**			1	1	9
2nd Floor North Corr.		**2**				2
2nd Floor West Coor.			**9**			9
Lecture Hall 205		3		**3**		6
Total Result	7	5	9	4	1	26

Figure 1.6 Confusion Matrix with Naïve Bayes Classifier.

The 9–NN and the Naive Bayes classifiers are more effectively compared using the confusion matrix given in the figures1.1,1.2,1.3,1.4,1.5,1.6 as they are inaccuracy. In the same amount of cases, there are two classifiers that are clearly misclassified. Concluding the following example CRISP-DM ignores the account of topology. Since CRISP-DM does not take into consideration it is not efficient that CRISP-DM should be followed for the indoor room positioning method. [Tamas, J. and Toth, Z., 2018, January.]



As CRISP-DM is used from a quite lengthy time, there is a certain problem such as the monitoring and the maintenance which is not being updated as the site is no longer used or monitored. As there are few problems which are related to CRISP-DM which includes:



- A lack of clarity
- Mindless rework
- Failure to iterate.

- No maintenance.
- Blind handoff to IT.

which was acknowledged by James Taylor, in a blog that says people are taking alternatives that bents the Cycle of CRISP-DM.

Hence, I hereby conclude that CRISP-DM is a great framework for analytics, but it should be upgraded so that it can be used more effectively in prediction analysis and there should more new technologies which can analyze the models efficiently and big data can be used in a more advanced way.

B) Critical Analysis of Data Mining in E-Commerce

2.1 Introduction:

The development of computer and Internet Technology, one of the man issues in the applications of Internet Technology is how to pick up useful information that benefits us from these complex data sets. E-commerce has changed most of the business functions in competitive enterprises. In the growing phases of technologies which have made the consolidated process between the client and dealers, dealers and traders, traders and industry, and industries and their large suppliers. In simpler terms, online -Transactions are empowered by e-commerce and e-business. It is not easy to achieve real-time data on a large scale. As the data are associated with many aspects of business transactions that are easily accessible, it is only the duty of the data mining process to function these data sets.

Data mining tools generate new information for decision-makers from very larger profiles of a database. There are many mechanisms in this origination that builds abstracts, summarizes, evaluates, aggregations of data. Having a huge amount of data makes some problems for detections of hidden relationships among various attributes of data and between several snapshots of data over a period of time. Data Mining has been pursued as a research topic by at least three communities: The artificial Intelligence researchers, the statisticians, and the database engineers. Although much work has been up to date, there is a need

to study more in-depth where there is a variety of problems related to e-commerce. The purpose of this critical analysis is to present data mining techniques and express the applications of data mining in e-business and how it is benefited in the business fields. [Rastegari, H. and Sap, M.N.M., 2008.]

2. 2 Key Highlights:

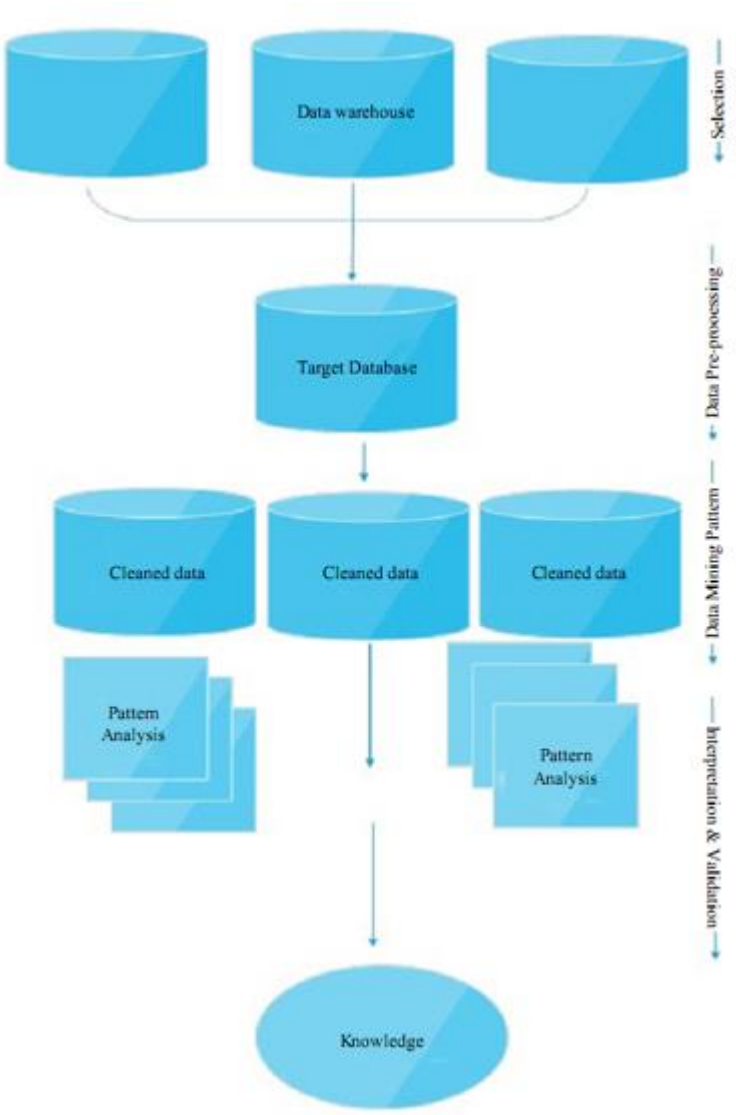
- Application of data mining in e-commerce sites in the field to gain all possible areas of e-commerce where data mining can be appropriate for an increase the company.

    - As we all acknowledge, customers usually leave behind such detail's businesses will store in their files while going shopping online.

    • Facts as these are unstructured or standardized data that can be processed to provide the organization with a competitive advantage.

2.3 Data Mining Process

- Data pre-and before-data mining methods are the first and simplest approach for data mining, by extracting inappropriate data which is not relevant to appropriate analysis.
- Data mining phase:
  The process therefore improves the efficiency of the entire process of data mining and also increases the exactness of the data and decreases fairly the time needed for actual mining.[Ismail, M., Ibrahim, M.M., Sanusi, Z.M. and Nat, M., 2015.]



In general data mining process iterates from the following five basic steps, which are:

- Data collection: This step is all about determining the type of data to be gathered, the criteria for it and the method used by the project.
- Data selection process typically is based on the following five basic steps: At the end the right input and output attributes are chosen to represent the function.
- • Data transformation: This step is all about arranging the data on demand by eliminating noise, transforming one type of data, normalizing the data if appropriate, as well as specifying the technique for managing missing data.

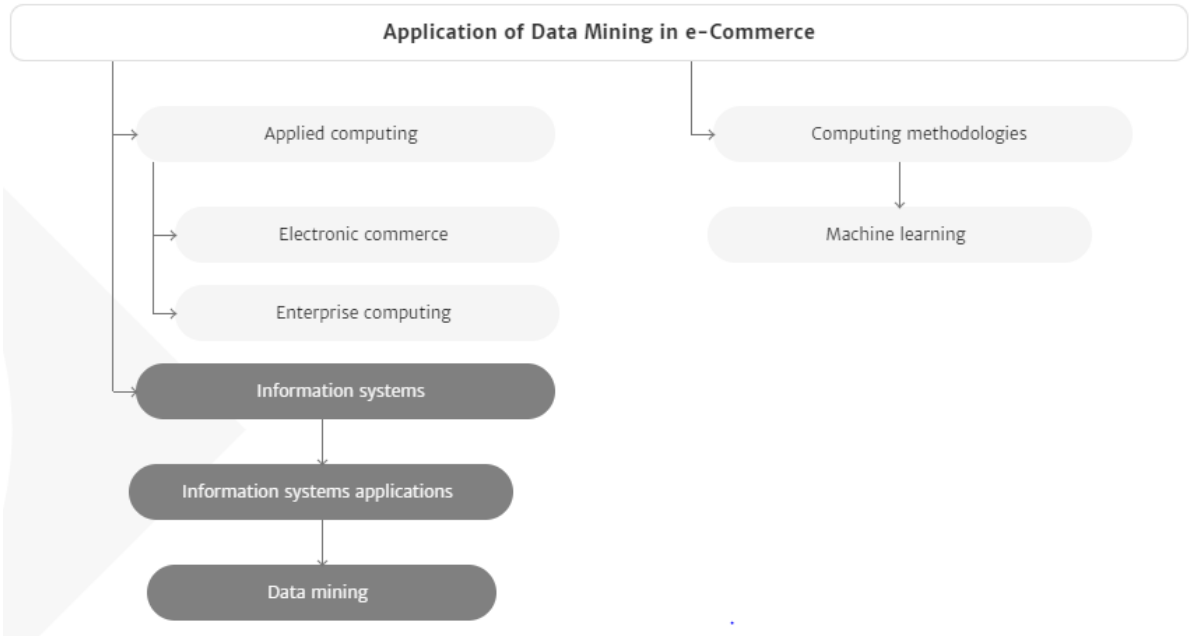
- Information extraction per se: the extracted information has been collected using one of the methods used to automate data mining by correctly executing the procedural procedures. • Analysis and testing of the results: Robustness is tested by the data mining algorithm check for a better understanding of the data and the synthesized information together with their validity. The extracted knowledge can also be analyzed by matching it with the previous technology area expertise and the Use of the information discovered:
- The goal is to introduce to decision makers the results for discovered information, so that a newly discovered model can be implemented in contrast with previously identified knowledge. [Ismail, M., Ibrahim, M.M., Sanusi, Z.M. and Nat, M., 2015.]

2.4 Benefits of Data Mining in E-Commerce

The below are fields where e-commerce data mining can be used for company benefits:

1) Consumer Profiling
   • This is also known as a consumer-oriented e-commerce approach. It helps businesses to use business intelligence to prepare their business activities and processes, and to develop new insights on goods or services for profitable e-commerce through the mining of customer information.
   • Buy recognition of customers: the tour results will help to reduce selling expenses for businesses. Companies can use surfing information from

consumers to decide whether it is.



- Classifying the customers of purchasing:

- Shop or just shop or buy something new with or with. This helps businesses prepare their services and develop it.

- Customer personalization is the process of supplying people with content and services based on their needs and behaviours.

- Personalization data mining research focuses mainly on applications suggestion and related topics such as shared filters

3) Basket Analysis :

• Single basket shopper has a history and a market basket analysis (MBA) that lets retailers better understand their consumers.
Basket review - There are various ways of getting the most out of the cup.

4) cross-selling and up-selling campaigns:

• Display products bought together; it may inspire consumers to buy a printer and obtain luxury paper or ink cartridges of the highest quality.

• Shoppers profile:

- Purchaser profile: In evaluating the shoppers ' basket with the assistance of data mining overnight, you gain insight into their age, income level,

purchasing habits, desires to purchase, exploit this and give the customer experience.

4) Demand predictions

• The time a buyer spends purchasing an item is a component of demand which helps to determine whether a consumer will again purchase it.

• Specify a method for the intended detection of selling goods by using this form of analyzes.

6) Market Segmentation

- Customer segmentation:
  Data mining is one of the main ways of consumer segmentation. It can be broken down into specific and relevant segments such as income, age, gender, consumer profession, from the lots of data collected, and that can be used when either the companies run email marketing campaigns.

The segmenting of a retail company's inventory would boost sales rates, because the company will target its marketing on a close-fit and highly sought-after market.

2.5 Challenges of Data Mining in E-Commerce

1) Spider Identification

- The web pages are the main data base for e-commerce businesses. Therefore, the awareness by the 17 search machines of how rapid things arise, how they occur and when improvements in the search engines become essential for e-commerce firms.
- Spiders are programs sent to find new knowledge by the search engine. They can also be called bots or crawlers.

2) Data Transformations

Data Transformations

Data engineering is a threat to data mining software. Data transformations. Today, only two different sources can provide the data needed for transformation, of which one is an active and operating program that can be implemented in the data warehouse, and secondly, certain operations that include allocating new tables, binning and aggregating data.

3) Scalability of Data Mining Algorithms

Data Mining algorithms with the use of yahoo, which has more than 1,2 billion pages views in a single day when large amounts of information are available, scalability is a major factor. The data mining algorithm can manage or process it, particularly because of the nonlinal size, or the vast amount of data collected from the web site at a reasonable time.

4) Make Data Mining Models Comprehensible to Business Users

- The goal is to plan and identify alternate models and a practical way for business users to understand which model regression can be built and how can they be introduced.
- How, for instance, do we implement neighboring models? How do we present the results of assembly rule algorithms with tens of thousands of rules without frustrating users?
- The demographic component of change of tourists through marriage, rises in wages or incomes, rapid growth of their offspring, needs which shape the pillars on which improvements have been modelled.

5) Help Slowly Changing Dimensions:
   • As a result, the features of goods can change, the architecture, the labeling of products or services and also the price improvement or deterioration may be possible in new choices.

6) Market users can access data transformation and model development.
   • The willingness of business users to provide consistent answers to questions needs facets of data transition but also technological expertise into the methods used in analytics

E.g. the delivery of the knowledge by consultants or by a software association for internet computational computing cubes and recommended transitions for mining).

2.6 CONCLUSION:

- Customer and purchase details obtained, which is the major assets of e-commerce firms, needs to be used for the benefit of businesses deliberately.
- Everyone in e-commerce may overcome these problems by using and implementing the right technologies, although the difficulty and grossness of the described challenges varies.
- To help companies overcome the problem of recognition of spider search engines, for example, by creating an e-commerce platform so that search engines can read and download the latest version.
- Data collected on consumers and their purchases, the main resource of companies involved with e-commerce.
- Needs to be actively used for corporations ' benefit. Data mining is an important part of these businesses in delivering user-oriented services in order to boost customer satisfaction.
- It is obvious that in this globally competitive world the use of data mining software is a must for e-commerce firms.

2.7 References:

[1] Rastegari, H. and Sap, M.N.M., 2008. Data mining and e-commerce: methods, applications, and challenges. Jurnal Teknologi Maklumat, 20(2), pp.116-128.

[2] Ismail, M., Ibrahim, M.M., Sanusi, Z.M. and Nat, M., 2015. Data mining in electronic commerce: benefits and challenges. International Journal of Communications, Network and System Sciences, 8(12), p.501.

[3] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J.H., Kull, M., Lachiche, N., Quintana, M.J.R. and Flach, P.A., 2019. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering.

[] Tamas, J. and Toth, Z., 2018, January. Limitation of CRISP accuracy for evaluation of room-level indoor positioning methods. In 2018 IEEE International Conference on Future IoT Technologies (Future IoT) (pp. 1-6). IEEE.

[6] Shearer, C., 2000. A Survey of Sequential Pattern Mining. The CRISP-DM Model: The New Blueprint for Data Mining, 5, pp.13-22.