IMPLEMENTATION OF CRISP-DM MODEL IN ORDER TO DEFINE THE SALES PIPE LINES OF PT X

Dadan Umar Daihani[1], Dina Feblian[2]

[1]Lecturer of Master Program in Industrial Engineering, Faculty of Industrial Technology, University of Trisakti, Jakarta, Indonesia
[2]Alumny Master Program in Industrial Engineering, Faculty of Industrial Technology, University of Trisakti, Jakarta, Indonesia
dadan@trisakti.ac.id, dina.feblian@gmail.com

ABSTRACT
PT X is a joint venture subsidiary of a bank in Indonesia, which is a company, engaged in general insurance business. PT X is required to increase its share of the parent company by the shareholders in developing its business potential. In developing the potential of the business, companies need enough information to be analyzed further. Sources of information that can be used is the business data of the company, both of PT X as a subsidiary and its holding company. By utilizing business data in large numbers, the company will find a variety of information needed. For the business data processing required a technique of data mining. In a study conducted, the implementation of the model refers to the six stages of Cross-Industry Standard Process for Data Mining or commonly known by the acronym CRISP-DM. In doing data mining, the six phases of the CRISP-DM is business understanding, understanding the data, the data preparation, modeling, evaluation, and deployment. At this stage of deployment, the report will be made by the customer segmentation by region. The expected result is the information required in the form of a new customer pipeline.
Key words: Data Mining, CRISP-DM, Customer Segmentation, Customer Pipeline

1. INTRODUCTION

1.1. Background

Data is a valuable asset. In accordance with the rapid development of Information technology especially the methode of collection and storing of data, the database built is often too large. As a result, the data that has been collected is dificult to interpete because there are too many, too long, too unattractive and take a long time to process them. Finally, a decision that should be made based on the data, often returning only made based on the intuition of decision makers. This prompted the creation of a branch of data mining.

Data mining is a process of data search in automatic manner, sourced from a database on a large scale to obtain useful information [12]. The development of data mining can not be separated from the development of information technology that is able to accumulate large amounts of data. According Kaena, data mining has now progressed in various business sectors which require support of data in very large quantities.

In today's global era, the business world is very dynamic and the competition is very tight. Therefore, businesses are required to have the right strategy, so that the business can survive and develop as expected. In order to keep their exixtency, the enterprise should increase their production capacity, diversify its business, improve efficiency and expand its marketing network. To formulate and choose the appropriate strategy, the company must mastering the bussiness environment, the economic development and their internal strength. Accordingly they require a different kind of data and information relating to its business environment, political environment, global economic developments and conditions, and internal capabilities. One way to meet the above requirement should be built a data base that will support the analysis of the situation and the state of progress of the company and its environment. This condition is also faced by PT X which is a joint venture company of a bank in Indonesia which was

established in Jakarta in 2011. The company is engaged in general insurance. In this business line, of course, not only the PT X engaged in these efforts, there are still quite a lot of other companies engaged in similar. This raises a tight business competition between insurance companies. Competition is happening in the world of insurance business is forcing the perpetrators to be more creative in formulating a strategy to ensure business continuity insurance that they run as well as increase the company's revenue. As a subsidiary company, the shareholders of the holding company has always demanded PTX to always increase market share and develop their business potential.

In order to do this, the company needs enough information to be analyzed further. Sources of information that can be used is the company business data, both from within the PT X itself or sourced from its parent company. Business data in large quantities is one of the valuable assets owned by the company, but if the data is not managed properly, the data and the information is simply not worth the setupuk data. But if the company is good at managing, selecting, extracting, processing, it will get very useful information to map the peruhaan, of competition and the development of the business environment in general. To take advantage of the data are numerous, varied and, the company needs a management system, as well as the withdrawal of pengkobinasian data so obtained information according to need. One technique pengelohanan data is now quite powerful and well developed is a data mining technique.

1.2. Research Objectives

Based on the issue, as noted above, this study focused to carry out three things:
1. Develop a model formulation of potential market share for PT X by using the techniques of data mining.
2. Finding the required information in the form of a sales pipeline of the modeling done in the business data are sourced from both fruit companies namely PT X and its parent company.

3. Knowing the size of the market share in the Commercial segment that has been achieved by PT X.

2. LITERATURE

2.1. Sales Pipeline

According Cefkin (2007), sales pipeline management used in commercial enterprises as a means to direct the focus and attention of the personnel as well as to anticipate the performance in the future. Accounting systems in the sales pipeline is used to monitor the potential revenues, investment decisions, and to establish a future projection. Therefore, the sales pipeline can be used as a guide by the shareholders and market analysts [2].

Sales Pipeline according Kimla (2013) was made mandatory and are used for sales as well as forecasting and analysis. Periodically evaluated for effectiveness and improvement. Flexible Pipeline is the most successful method used by some companies. Consistent feedback from salespeople needed for the accuracy of the pipeline and as a reflection of the real conditions that exist. Pipeline adjusted each time it needs to reflect the required changes in the sales process [5].

Model pipeline operations are generally described by Yan et al., (2015), ie as identifying new sales leads, then the seller enter sales opportunities prospects in pipeline management system. Prospects that have been identified are then evaluated and assigned several potential sales opportunities. A sales opportunity consists of a set of one or more products or services that sellers are trying to convert into a client's purchase of real [14].

Based on the explanation above, it can be concluded that the sales pipeline describes an approach to sales, is based on the underlying principle of the sales process that made the sale. Describing the move sales from potential customers or prospects to qualify the prospect of a guidance and further validate that leads to sales opportunities. Sales pipeline is used so that sales can be controlled and work in a well-organized sales

2.2. Market Segmentation

According to Lamb, Hair, and McDaniel (2010), the market segment is a subset of people or organizations that share one or more characteristics that cause have similar product needs [7]. From a marketing perspective, the market segment can be described as somewhere between the two extremes. Market Segmentation or market segmentation according to Kotler and Keller (2012), is a market which has been divided consists of a group of customers who have a set of needs and wants something similar [6]. The process of dividing the market into segments that are meaningful, relatively the same, or a group that can be identified is called market segmentation. The purpose of market segmentation is to allow marketers to adjust the marketing mix to meet the needs of a specific segment [7].

Segmentation can be performed based on key variables such as the following, namely: [6]

1. The geographic segmentation, segmentation divides the market into geographic units such different countries, autonomous regions, municipalities, climate, or residential areas.
2. Demographic segmentation, segmentation divides the market into different groups based on variables such as age, family size, family life cycle, gender, income, occupation, religion, race, nationality generations, and social classes.
3. Psychographic segmentation, market segments is done by grouping the consumer or buyer becomes part of the market according to the variables pattern or lifestyle and personality.
4. Segmentation behavior, in this market segment is classified into groups that are distinguished based on knowledge, attitude, use of, or response to a product.

In this study, to determine the segmentation and knowing the size of the market share of PT X against the parent company that is focused on one segment, namely the Commercial segment by using business data from both sides of the company. In the business data processing using the application of data mining.

2.3. Data Mining

According to Gartner Group, data mining is a process of identifying a set of data in large quantities and stored in a storage area, through techniques such as pattern recognition mathematical and statistical techniques [8]. Data mining is a term used to find hidden knowledge in database. Data mining is a semi-automatic process that uses statistical techniques, mathematics, artificial intelligence and machine learning to extract and identify potential knowledge and useful information that is helpful is stored in large databases [13].

Data mining by David Hand, Heikki Mannila and Padhraic Smyth from MIT is an analysis of the data is usually large to find a clear link and summarize the results of the analysis of the data that has not been known previously by way of the latest to be understood and can be useful for owners of such data [8].

In addition to the above definition, data mining is the analysis of data to build relationships and identify patterns, focusing on identifying relationships in data. Data mining is an information extraction activity whose goal is to find the hidden facts contained in the database. Using a combination of machine learning, statistical analysis, modeling techniques and database technology. Data mining find patterns and relationships in data and infer rules that allow a prediction for the future [9].

Data mining is a process of data search is automatically sourced from a database on a large scale to obtain useful information [12]. Data mining is the analysis of the survey data sets to find unsuspected relationships and summarize data in a way that is different, that is understandable and useful to the data owner. Data mining is a field of science that brings together some of the techniques from machine learning, pattern recognition, statistical, database and visualization for handling problems in the retrieval of information from a large database [8].

Some of the drivers that support the remarkable progress in the field of data mining, among others: [8]
1. The rapid growth in the data set.
2. Storage of data in the data warehouse so that the entire company has good access to the existing database.
3. An increase in data access through web and intranet navigation.
4. The pressure of business competition to increase market share in the economic globalization.
5. The development of software technology for data mining.
6. Rapid development in the advancement of computing and storage capacity development.

Based on the definition that has been delivered, the important thing related to data mining, namely:
1. Data mining is an automated process of data that already exists.
2. The data will be processed in the form of data on a large scale.
3. The purpose of data mining is to get the relationship or pattern that can be either an information which may provide a useful indication.

Search a relationship or pattern in the data mining could be a relationship between two or more in one dimension. For example in the dimensions of the product, it can be seen linkage purchase of a product from other products. In addition, the relationship can also be seen between the two attributes and two or more objects or more [10].

Meanwhile, the discovery of a pattern is another output of data mining. For example, a banking company that will increase loan facilities to customers. The company will look for a pattern from the list of existing customers to find customers who have potential and potential customers who do not have, so the company can also perform analysis of customers that can be raised about the loan facilities [10].

Some explanation of the initial definition of data mining include a focus on process automation. Berry and Linoff, (2004) in the book Data Mining Technique for Marketing, Sales, and Customer Support defines data mining as a process of exploration and

analysis of automatic and semi-automatic to the large amounts of data with the goal of finding meaningful patterns or rules. Three years later in the book Mastering Data Mining, they give a redefinition of the notion of data mining and give a statement that "if we deplore is the phrase automatically or semi-automatically, because we felt it focuses excessive automated techniques and less on exploration and analysis ". This gives the false notion that data mining is a product that can be purchased than science that must be mastered [8].

The statement confirms that the data mining automation does not replace human intervention. Humans must actively participate in every phase of the data mining process. Greatness capability data mining algorithms contained in the analysis software are now allowing the use of error that can be fatal. Users may apply improper analysis of our data collection by using a different approach. Therefore, it takes an understanding of the statistical and mathematical models of the structure of the underlying performance of the software [8].

Data Mining is not a science entirely new [1] [4] and the theory was already much discussed in the literature. Theories include that, Naive-Bayes and Nearest Neighbour, Decision Trees, Association Rules, K-Means Clustering, and Text Mining [1]. While presenting the latest developments in the developed algorithms, among others, Artificial Neural Network (ANN), Genetic Algorithm, Fuzzy C-Means, Support Vector Machine (SVM), and others [8]. Since the development of the 1990s, data mining formerly controlled by the statisticians, apparently in the 2000s developed by computer science experts who are trying to improve the performance of statistical theories exist. According to Daryl Pregibon mentioned that data mining is a blend of Statistics, Artificial Intelligence and Database [3].

2.4. Cross-Industry Standard Process for Data Mining
In this study, the basic model will be referenced on the CRISP-DM . Cross-Industry Standard Process or CRISP-DM was developed in 1996 by analysts from several industries such as Daimler Chrysler,

SPSS and NCR [8] [11]. CRISP-DM is the standardization of the data mining process as a general problem-solving strategies of business or unit of study [8].

In CRISP-DM data mining project has a life cycle that is divided into six phases [8] [11]. Relations between fhase and activity cycles referred to can be seen in Figure 2.1
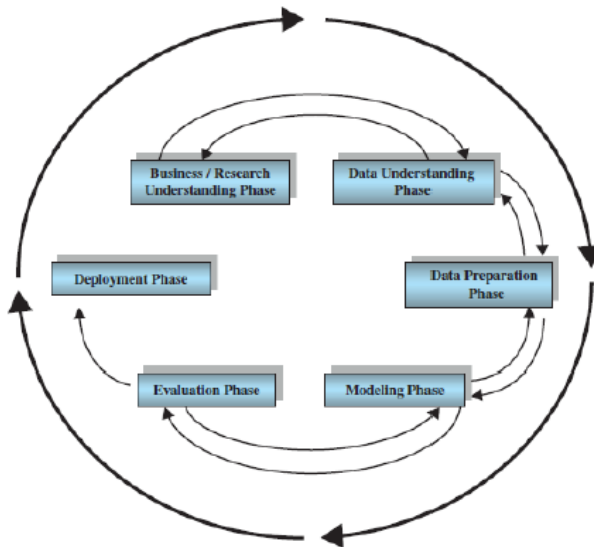

Figure 2.1 Lifecycle process CRISP-DM
(Larose, 2014)

Overall sequence of this phase is adaptive and affect each other. Phase one is strongly influenced by the results of previous fases. The relationship between the phases illustrated by arrows. For example, the modeling phase will depend on the results of the data preparation and evaluation of data. Likewise, these phases affects the other phases. If one phase is changed, then the previous phase should also be adjusted to developments / outcome of the previous phases. Therefore, this process is called a cycle interconnected. Based on the behavior and characteristics of the model, the process may go back to the data preparation phase for further improvement of the data or move forward to the evaluation phase [8].

From Figure 2.1 it can be seen that the life cycle of the CRISP-DM consists of six phases, namely: [8]
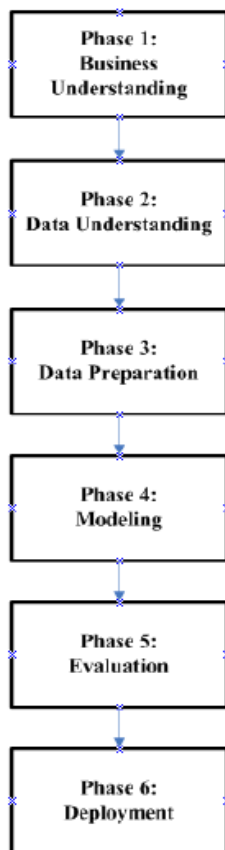
1. Business Understanding Business Understanding Phase or Phase
   a. Determination of project goals and requirements in detail within the scope of the business or research unit as a whole.
   b. Translating the objectives and constraints into a formula and the definition of data mining problems.
   c. Setting up the initial strategy to achieve the goal.
2. Data Understanding Data Understanding Phase or Phase
   a. Collecting data.
   b. Using analysis of data inquiry to identify further regarding the data and an initial search of knowledge that is therein.
   c. Evaluating the quality of data.
   d. If desired, select a small number of data that may contain a pattern of problems.
3. Data Preparation Phase or Phase Data Processing
   a. Preparing the data available from the beginning, because it is a collection of data that will be used for the entire next phase. This phase is hard work that needs to be carried out intensively.
   b. Selecting cases and variables that you want to analyze and which according to the analysis to be performed.
   c. Make changes to multiple variables if needed.
   d. Setting up the initial data that is ready for device modeling.
4. Modeling Phase or Phase Modelling
   a. Selecting and applying appropriate modeling techniques.
   b. Calibration model rules to optimize the results.
   c. To note that some of the techniques likely to be used on the same data mining problem.
   d. If necessary, the process can return to the processing phase of data to make the data into a form that is in accordance with the specification requirements of specific data mining techniques.
5. Evaluation Phase or Phase Evaluation
   a. Evaluating one or more models used in the modeling phase to get the quality and effectiveness before being disseminated to use.

b. Establish whether there are models that meet the objectives in the initial phase.

c. Determine whether there are important issues of business or research that is not handled properly.

d. Take decisions relating to the use of data mining results.

6. Deployment Phase or Phase Deployment

a. Using the resulting model. The formation of the model does not signify that terselesaikannnya project.

b. A simple example of deployment that is making the report.

c. Examples of complex deployment process is the application of data mining in parallel to other departments.

3. RESEARCH METHOD

Referring to the special reference model CRISP-DM, this study is divided in six (6) steps, as illustrated in Figure 3.1.

In detail, the steps can be described as follows:

1. Business Understanding, is an understanding of the substance of data mining activities to be undertaken, needs from a business perspective. Its activities include setting goals or business goals, understand the business situation, translating business goals into data mining purposes.

2. Data Understanding, is collecting data, studying the data to understand the data that will be used in research, identify problems related to the data.

3. Data Preparation, at this stage the structure of the database will be prepared to facilitate the process of mining.

4. Modeling Phase, is a decisive phase of data mining techniques are used, determining data mining tools, data mining algorithms, determining the optimal parameter values.

5. Evaluation Phase, is the phase of the interpretation of the data mining results shown in the modeling process contained in the previous phase. In-depth evaluation with the aim of adjusting the model obtained to match your goals to be achieved in the first phase.

6. Deployment Phase, or the deployment phase is the phase of preparation of the report or presentation of knowledge gained from the evaluation of the data mining process.

Furthermore each phase is described as follows:
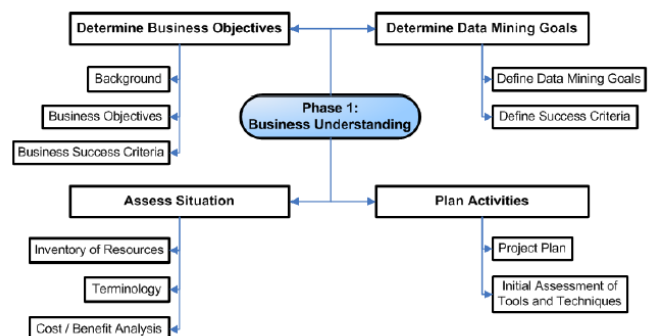
3.1 *Phase 1*

Figure 3.2 The First Phase Reference Model CRISP-DM

In this phase there are four processes, namely:

1. Determine business objectives undertaken to understand the objectives to be achieved by PT X from a business perspective to the parent company as a whole.
2. Asses situation were made to include the facts that existed at PT X about the resources, constraints, assumptions and other factors that must be considered in determining the purpose of data analysis and plans to do the research.
3. Determine data mining goals that state the purpose of data mining is technically against the research undertaken.
4. Plan activities that explain the plan aimed to achieve the goal of data mining and thus can achieve business objectives, and then determine the tools that will be used.
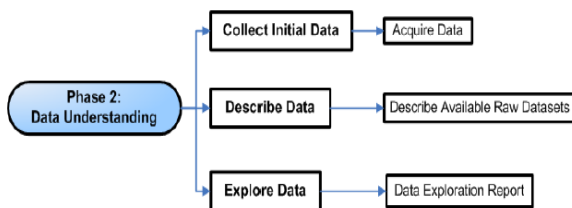
3.2 Phase 2

Figure 3.3 The Second Phase Reference Model CRISP-DM

In this phase there are three processes, namely:

1. Collect initial data that obtaining data or access to open data to be used in research. This initial collection includes data loading if necessary for understanding the data.
2. Describe the data that is examining the data obtained and reported the results.
3. Explore the data that is exploration of the content of the data obtained, then gives a description of the contents of the data.
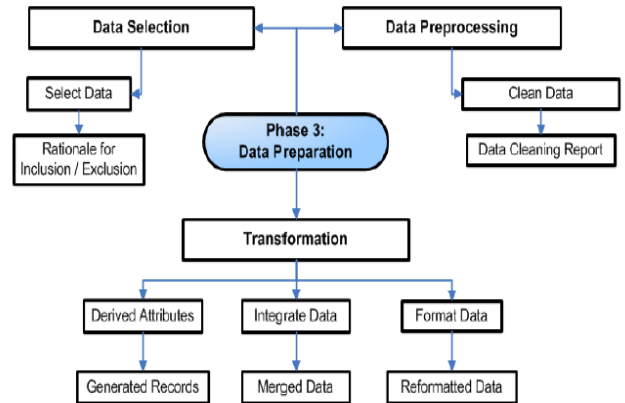
3.3 Phase 3

Figure 3.4 The Third Phase Reference Model CRISP-DM

In this phase there are three processes, namely:

1. Data selection is choosing the data that will be used in the data mining process. In the process of the election is done also attributes that are tailored to the data mining process.
2. Data preprocessing is to ensure the quality of data that have been at the stage of data selection, at this stage the problem that must be addressed is if there are noisy data and missing values. The process of data cleansing or cleansing carried out to find data anomalies that may be still present in the data.
3. Transformation is the grouping of the attributes or fields that have been elected into the new database for data mining materials.
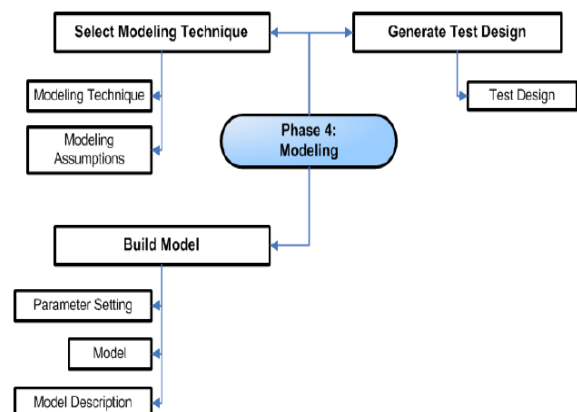
3.4 Phase 4

Figure 3.5 Phase four CRISP-DM reference model

In this phase there are three processes, namely:
1. Select modeling technique is the first step in modeling, using modeling techniques that have been assigned to the business understanding phase.
2. Build a model that is run in accordance with the procedures of modeling tools.
3. Generate test design is to test the quality and validity of the model using a dataset that has been prepared.
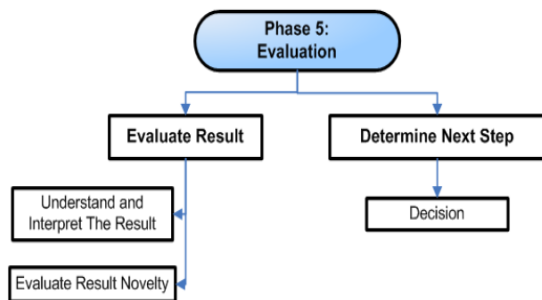
3.5 Phase 5

Figure 3.6 The fifth phase of CRISP-DM reference model in Research

In this phase there are two processes, namely:
1. Evaluate the result that summarizes the results of the assessment in terms of business success criteria, including the final statement on whether the research meets the business objectives.
2. Determine the next step is to give a decision whether the modeling techniques used can be used as a standard in determining the purpose of research.
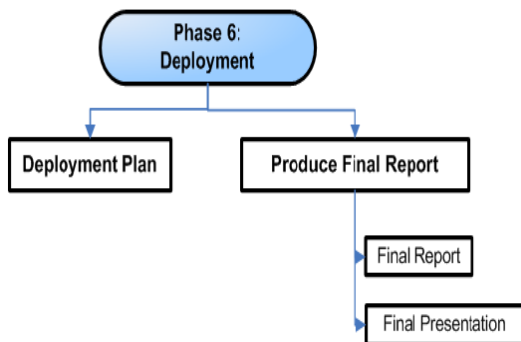
3.6 Phase 6

Figure 3.7 Phase Sixth CReEISP-DM reference model in Research

In this phase there are two processes, namely:
1. Deployment plan that describes an overview of the plan to the preparation of reports to be made.
2. Produce final report which provides a visualization of the statements that have been made based on a deployment plan

4. RESULT AND DISCUSSION

The data used in the modeling process amounted to 34 files, data files consist of 17 primary and 17 secondary data files. Both pieces of data files are then merged into one and sorted according to the order of the file name and also adjusted according to the contents and file naming. From the results of fuzzy lookup process is done in modeling stage, then there are 17 files that have been saved. The following table displays a summary that describes how many number of customers that are included in the pipeline of new customers who have been grouped by month and be segmented by region division of labor of the Commercial segment at PT X. The result of this grouping are presented in Table 4.1 below.

| NO	WILAYAH	JUMLAH CUSTOMER												TOTAL
		JAN	FEB	MAR	APR	MEI	JUN	JUL	AGT	SEPT	OKT	NOV	DES	
1	MEDAN	17	26	16	23	19	15	10	17	18	9	15	12	197
2	PALEMBANG	19	11	13	4	7	18	4	1	7	7	2	1	94
3	JAKARTA 1	11	17	22	14	27	20	16	11	10	10	6	2	166
4	JAKARTA 2	10	9	6	12	17	14	9	16	9	19	9	8	138
5	JAKARTA 3	7	10	13	11	6	0	6	6	3	2	5	0	85
6	BANDUNG	7	14	16	17	11	14	3	7	8	6	8	5	116
7	BANJARMASIN	26	16	23	7	13	15	8	2	11	6	7	3	137
8	SEMARANG	10	14	14	12	17	6	8	2	2	11	3	5	104
9	SURABAYA	10	23	34	21	18	24	20	20	11	9	15	16	221
10	DENPASAR	1	4	8	10	5	3	12	5	5	5	5	4	67
11	MAKASSAR	5	9	15	15	12	12	7	7	9	5	6	8	110
	GRAND TOTAL	123	153	180	146	152	149	103	94	93	89	81	72	1435

Table 4.1 Number of New Customer Pipeline

Based on Table 4.1 it can be seen that the number of pipeline of new customers in total for Medan area amounted to 197 customers, Palembang amounted to 94 customers, Jakarta 1 totaled 166 customer, Jakarta 2 amounted to 138 customers, Jakarta 3 amounted to 85 customers, Bandung numbered 116 customer, Banjarmasin region numbered 137 customer, region amounted to 104 customer Semarang, Surabaya region amounted to 221 customers, amounted to 67 customer

Denpasar and Makassar region amounted to 110 customers. So that the number of new customer pipeline as a whole amounted to 1,435 customers.

Customer categorization process outlined selanjunya seen from the status which is already a customer or are not yet a customer. Results pengelohan data is presented in the following graph (Figure 4.1)
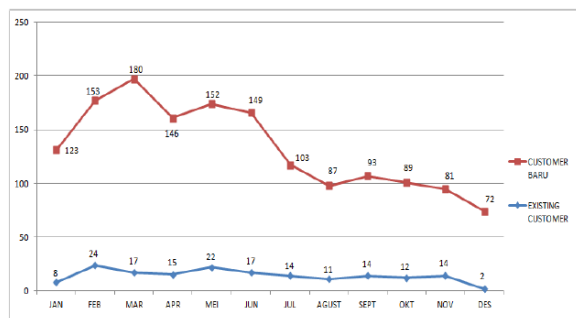
Graph Figure 4.1 Comparison of New and Existing Customers

Based on the chart that has been shown, it can be seen that in January there were eight the number of customers who have become customers of PT X and as many as 123 customers who have not become a customer of PT X of a total of 131 customers. In February, there were 24 the number of customers who have become customers of PT X and as many as 153 customers who have not become a customer of PT X of a total of 177 customers. In March there were 17 the number of customers who have become customers of PT X and as many as 180 customers who have not become a customer of PT X of a total of 197 customers. In April, there were 15 the number of customers who have become customers of PT X and as many as 146 customers who have not become a customer of PT X of a total of 161 customers. In May there were 22 the number of customers who have become customers of PT X and as many as 152 customers who have not become a customer of PT X of a total of 174 customers. In June there were 17 the number of customers who have become customers of PT X and as many as 149 customers who have not become a customer of PT X of a total of 166 customers. In July there were 14 the number

of customers who have become customers of PT X and as many as 103 customers who have not become a customer of PT X of a total of 117 customers. In August there were 11 the number of customers who have become customers of PT X and as many as 87 customers who have not become a customer of PT X from a total of 98 customers. In September there were 14 the number of customers who have become customers of PT X and a total of 93 customers who have not become a customer of PT X of a total of 107 customers. In October there were 12 the number of customers who have become customers of PT X and as many as 89 customers who have not become a customer of PT X of a total of 101 customers. In November there were 14 the number of customers who have become customers of PT X and as many as 81 customers who have not become a customer of PT X from a total of 95 customers. In December there were two the number of customers who have become customers of PT X and as many as 72 customers who have not become a customer of PT X from a total of 74 customers.

By knowing this categorization, then the marketing can menkonsentrasikan to menggalam the prospectus only by prospective customers, and maintain a good relationship with a company that has become a customer.

5. CONCLUSION

As a conclusions, we have tree element importants as follows:
1. Model CRISP-DM is very helpful in developing data mining models to solve a special case, since CRSP-DM is very systematic and provides detailed guidance on every step.
2. Using the approach of CRSP-DM, the preparation of the sales pipe line at the company X can be formulated properly and easelly.
3. Data mining can create an environment to support business intelligence, which is efficient in getting the required information, provide a special tool for the analytical

capabilities are much better, reducing the strain on the operating system to create reports and analytical activities, as well as provide the ability for users to the scope of the business processes.

6. REFERENCES

[1] Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction To Data Mining.* New York: Pearson.

[2] Lamb, C. W., Hair, J. F., & McDaniel, C. (2010). *Marketing.* Boston, MA: South- Western College.

[3] Kotler, P., & Keller, K. (2012). *Marketing Management.* Upper Saddle River, New Jersey: Prentice Hall.

[4] Larose, D. T. (2014). *Discovering Knowledge in Data : An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc.

[5] Turban, E., Aronson, J. E., & Liang, T.-P. (2005). *Decision Support Systems and Intelligent Systems.* New Jersey: Prentice-Hall, Inc.

[6] Pal, N. R., & Jain, L. C. (2005). *Advanced Techniques in Knowledge Discovery and Data Mining.* New York: Springer.

[7] Ponniah, P. (2011). *Data Warehousing Fundamentals for IT Professionals.* New Jersey: John Wiley & Sons, Inc.

[8] Bramer, M. (2007). *Principles of Data Mining.* London: Springer-Verlag.

[9] Kantardzic, M. (2011). *Data Mining : Concepts, Models, Methods, and Algorithms.* Hoboken, New Jersey: John Wiley & Sons, Inc.

[10] Gorunescu, F. (2011). *Data Mining - Concepts, Models, and Techniques.* Berlin: Springer-Verlag Berlin Heidelberg.

[11] Shafique, U., & Qaiser, H. (2014). International Journal of Innovation and Scientific Research. *A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) , 12* (1), 217-222.

[12] Cefkin, M. (2007). Ethnographic Praxis in Industry Conference. *Numbers may speak louder than words, but is anyone listening? The rhythmscape and sales pipeline management* , 187–199.

[13] Kimla, N. (2013). *Pipeline Management: Success with CRM.* Los Angeles: Pipelinersales Inc.

[14] Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., et al. (2015). Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. *On Machine Learning towards Predictive Sales Pipeline Analytics* , 1945-1951.