Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

Secure Extreme Programming Model (SXP) on Data Mining Applications: SXP-DM Methodology

Mr.Sandepaga Vijay Hemanth[1,] Mr.K.Nagaraju[2]
Asst.Professor,Computer Technology Dept. Kavikulguru Institute of Tecchnology and Science,Ramtek. RTM Nagpur University,Maharashtra,India

Abstract - As the world becomes increasingly dynamic, the traditional static modeling may not be able to deal with it. One solution is to use agile modeling that is characterized with flexibility and adaptability. On the other hand, data mining applications require greater diversity of technology, business skills, and knowledge than the typical applications, which means it may benefit a lot from features of agile software development. In this paper, we will propose a framework named SXP-DM based on Secure Extreme Programming Model (SXP) that can easily adapt with discriminantdata mining applications. A case study in automotive manufacturing domain was explained and experimented to evaluate SXP-DM methodology.

I. INTRODUCTION

Data mining is the search for relationships and distinct patterns that exist in datasets but are “hidden" among the vast amount of data. A data mining task involves determining useful patterns from collected data or determining a model that fits best on the collected data.

Although the idea of applying data mining techniques on software engineering data has existed since mid 1990s, only lately the idea has especially attracted a large amount of interest within software engineering [1]. Data mining techniques are applied to analyze the problems raised during the life cycle of a software project development [3, 7], also to determine if two software components are related or not [16]. They were also used for software maintenance [2,9], software testing [15], software reliability analysis [8, 15], and software quality [6].

Many questions arise when trying to apply data mining techniques on software engineering field. What types of SE data are available to be mined?, which SE tasks can be held using data mining?, how are data mining techniques used in SE? are all important questions that a lot of researches were trying to find, have relevant responses.

On the other hand, few researchers were trying to find the SE processes modeling that fit better with data mining applications.

In this paper, we will focus on using agile modeling for discriminantdata mining applications, focusing on SXP (Secure Extreme Programming Model) modeling, which replaced the static Plan-Design-Build lifecycle, with Speculate-Collaborate-Learn lifecycle. The main characteristics of SXP lifecycle are the continuous learning, intense collaboration among developers, testers, and customers, and it can easily adapt with uncertain future [17].

We will start by viewing the characteristics of data mining applications, and the most widely methodology used for process modeling for data mining applications (CRISP-DM methodology), then we will present the characteristics of agile modeling, and suggest a new framework named SXP-DM for data mining processes using Secure Extreme Programming Model (SXP) method. The new framework was tested using a case study in the automobile manufacturing domain.

II. WHAT CHARACTERIZES DATA MINING Applications?

Data mining applications are characterized by the ability to deal with the explosion of business data

and accelerated market changes. These characteristics help providing powerful tools for decision makers. Such tools can be used by business users (not only PhDs, or statisticians) for analyzing huge amount of data for patterns and trends [19].

The most widely used methodology when applying data mining processes is named CRISP-DM. It was one of the first attempts towards standardizing data mining process modeling [18]. CRISP-DM has six main phases, starting by business understanding that can help in converting the knowledge about the project objectives and requirements into a data mining problem definition, followed by data understanding by performing different activities such as initial data collection, identifying data quality problems, and other preliminary activities that can help users be familiar with the data.

The next and important step is data preparation by performing different activities to convert the initial raw data into data that can be fed into modeling phase. This phase includes tasks such as data cleansing and data transformation. Modeling is the core phase which can use a number of algorithmic techniques (decision trees, rule learning, neural networks, linear/logistic regression, association learning, instance-based/nearest-neighbor learning, unsupervised learning, and probabilistic learning, etc.) available for each data mining approach, with features that must be weighed against data characteristics and additional business requirements.

The final two modules focus on evaluation of module results, and deployment of the models into production. Hence, users must decide on what and how they wish to disseminate/deploy results, and how they integrate data mining into their overall business strategy [18, 19].

III. APPLYING SXP METHOD ON DISCRIMINANT DATA MINING APPLICATIONS: SXP-DM METHODOLOGY

Software is intangible and more easily adapted than a physical product. Also, software processes depend on how a firm competes, and may be more adaptable than manufacturing processes bound by machinery, raw materials, and physical plants.

Technologies such as agile methods may make it less costly to customize and adapt development processes. Agile modeling has many process centric software management methods, such as: Secure Extreme Programming Model (SXP), Extreme Programming (XP), Lean Development, SCRUM, and Crystal Light methods.

Adaptive approaches are best fit when requirements are uncertain or volatile; this can happen due to business dynamism, and rapid evolving markets. It's difficult to practice traditional methodologies in such unstable evolving markets [11]. SXP modeling is one of such adaptive approaches. It replaces the static Plan-

Design-Build lifecycle, with the dynamic Secure Planning -Design-Coding-Test-Deliverance life cycle. Testing Phase recognizes individual unit tests are organized sinto"universal testing suite" integration and validation testing of the system can occur on a daily basis. This provides the SXP team with a continual indication of progress and also can raise SXP acceptance tests, also called customer tests ,are specified by the customer and focus on overall system features and functionality that are visible and reviewable by the customer, Acceptance tests are derived from user stories that have been implemented as part of deliverance.

The uncertain nature of complex problems such as discriminant data mining, and encourages exploration and experimentation. Discriminant data mining problems require a huge volume of information to be collected, analyzed, and applied; they also require advanced knowledge, and greater business skills than typical problems, which need "Planning" among different stakeholders, in order to improve their decision making ability. That decision making ability depends on "Secure

Testing" component in order to test knowledge raised by practices iteratively after each cycle, rather than waiting till the end of the project. Testing organizations can adapt more easily with SXP life cycle [17].

Hence the core of SXP is the premises were outcomes are naturally unpredictable, therefore, planning is a paradox. It is very difficult to successfully plan in a fast moving and unpredictable business environment, which is one of the main characteristics of discriminant data mining application.
This is one of the major points that we are using to create a data mining process framework based on SXP methodology (figure1). We call our new methodology SXP-DM as it combines the characteristics of SXP methodology, with the Discriminant data mining solution steps.
Secure Planning phase includes business and data understanding, and data preparations including ETL (Extract/Transform/Load) operations. This phase is the most important one as it takes considerable time and resources. This preparation phase will end by creating the enterprise data warehouse, and the required data marts and cubes.

Secure Design phase ensures the high communication in a diversity of experienced stakeholders in order to use the best modeling algorithm for Discriminant data mining process. Testing and evaluation of such algorithms occur in the "Testing" phase, the results will be discussed among the members of the project team, if the results are acceptable, a new release can be deployed in a form of discriminant scoring reports, otherwise a new Design phase will be used in order to chose better data mining algorithm.
The cyclic nature of the whole framework can respond to the business dynamic changes, a new data sources can be added to the preparation pase, and the cycle will move again

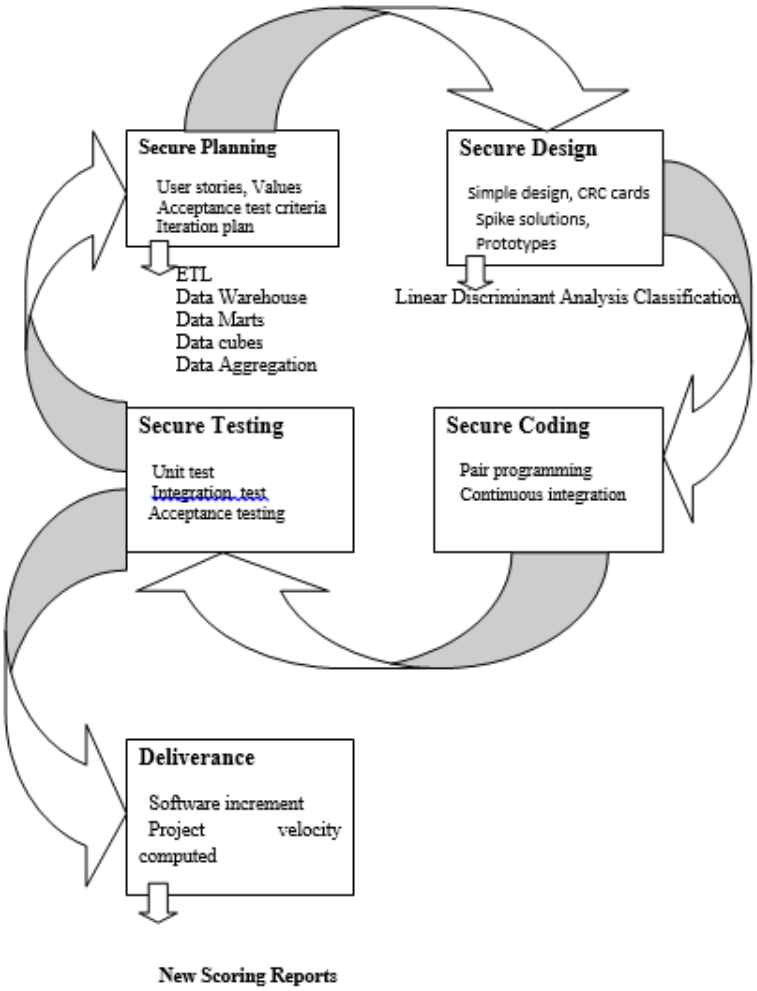


Figure1: SXP-DM a discriminant data mining process framework based on SXP methodology.

IV. A DATA MINING CASE STUDY IN

Census Services Concerning Housing Domain

The Boston Housing dataset contains the information collected by the U.S. Bureau Census concerning housing in the area Boston, Massachusets.The dataset include information on 506 census housing tracts in the Boston area the goal is to predict the median house price in new tracts based on information such as crime rate, pollution and number of rooms the dataset contains fourteen predictors and the response is the median house price (MEDV).

There are 14 attributes in each case of the dataset. They are:

1. CRIM: per capita crime rate by town.
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town.
4. CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
5. NOX: nitric oxides concentration (parts per 10 million).
6. RM: average number of rooms per dwelling.
7. AGE : proportion of owner-occupied units built prior to 1940
8. DIS : weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways.
10. TAX: full-value property-tax rate per $10,000.
11. PTRATIO: pupil-teacher ratio by town.
12. B: 1000(Bk - 0.63) ^2 where Bk is the proportion of African-Americans by town.
13. LSTAT: % lower status of the population.
14. MEDV: Median value of owner-occupied homes.

The main goal for such data mining solution was to get some initial positive results on prediction and to measure the prediction score of different data sources.

Using our proposed SXP-DM methodology, the enterprise data warehouse was created as a result of the Planning phase, and the ETL package was defined and developed.

The Design phase was one of the most important phases as it needed a lot of discussions and intensive collaborative team work. The method needed to model our prediction solution was not specified.with the consumer directly, yet a fundamental understanding of the market, the trends, the moods, and the changing consumer tastes and preferences are fundamental to competitive.

The information gathered in order to produce Census services on data mining solution.

A data warehouse is built to hold CRIM data, ZN data, INDUS data and CHAS data to MEDV analyzing and predicting median house prices, managing census ratio and planning for census predicative.data in support of the discriminant goals and objectives. Envision an analytic environment that will improve their ability to support planning and census management,in addition to enable them to meet the expectations of their decision-making process which is supported by appropriate data and trends. Regardless of functional boundaries and type of analysis needed, their requirements focus on improving access to detailed data, more consistent and more integrated information.

Having a data warehouse that combines online and offline behavioral data for decision-making purposes is a strategic tool which business users can leverage to improve census prediction demand forecasting, improve model/trim level mix planning, adjust body model/trim level mix with RAD data, and reduce days on lot.

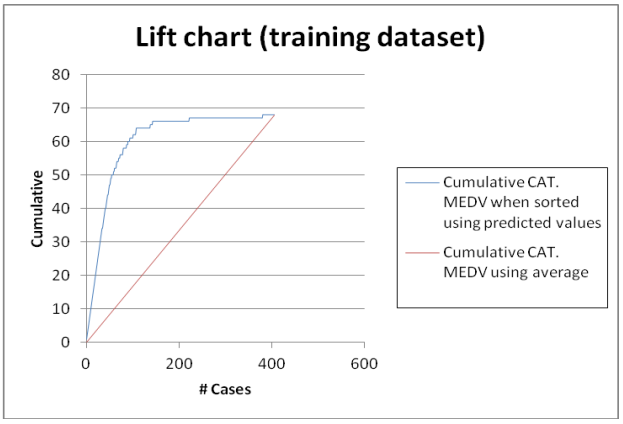
V. RESULT

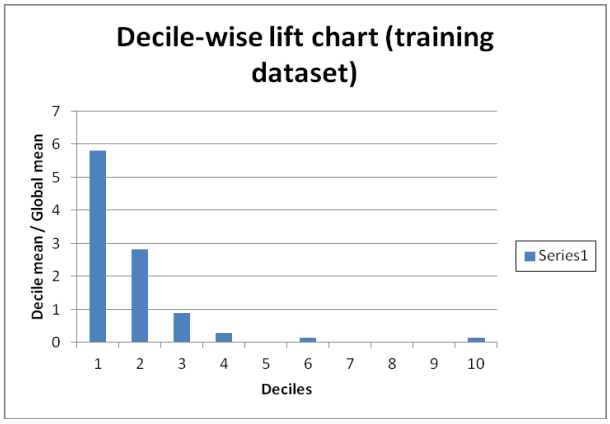
VI.

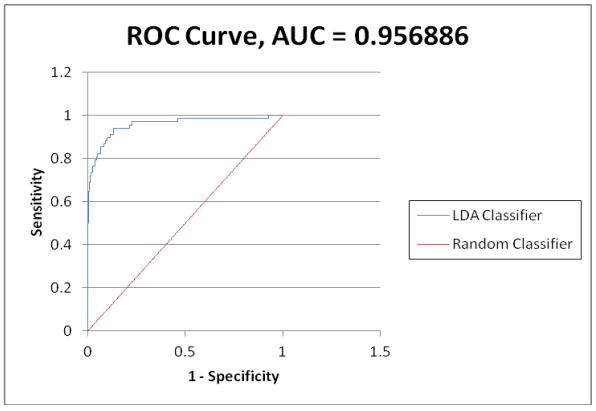
Linear Discriminant Analysis Classification - Lift Chart for Training Data

Output Navigator					
Inputs	Prior Class Probabilities	Class Funs	Canonical Variate Loadings	Train. Score - LDA Summary	Valid. Score - LDA Summary
Train. Canon. Scores	LDA Train. Lift Chart	LDA Train. Detail Rpt.	Valid. Canon. Scores	LDA valid. Lift Chart	LDA Valid. Detail Rpt.

Elapsed Times in Milliseconds			
Reading Data	Computation	Writing Data	Total
0	0	62	62







Decile	Mean	Std.Dev.	Min.	Max.
1	0.975	0.158114	0	1
2	0.475	0.505736	0	1
3	0.15	0.36162	0	1
4	0.05	0.220721	0	1
5	0	0	0	0
6	0.025	0.158114	0	1
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0.025	0.158114	0	1

VII. CONCLUSION

In this paper, we explained the use of data mining techniques in software engineering tasks such as programming, testing, maintenance, reliability, and quality. Due to the uncertain nature of discriminantdata mining application requirements, we proposed a new framework SXP-DM based on agile

REFERENCES

[1] Hassan A. E., Mockus A., Holt R. C., and Johnson P. M., "Guest editor's introduction: Special issue on mining software repositories". *IEEE Trans. Softw. Eng.*, 31(6):426– 428, 2005

[2] Riquelme J. C., Polo M., Aguilar_Ruiz J. S., Piattini M., Francisco J. and Francisco Ruiz F. T., "A omparison of Effort Estimation Methods for 4GL Programs: Experiences with Statistics and Data Mining", *International Journal of*

Software Engineering and Knowledge Engineering, Vol. 16, No. 1 (2006) 127-140.
[3] Nayak R., Qiu T., "A Data Mining Application: Analysis of Problems Occurring During A Software Project Development Process", *International Journal of Software Engineering and Knowledge Engineering*, Vol. 15, No. 4 (2005) 647-663.
[4] Slaughter S. A., Levine L., Ramesh B., Pries-Heje J., and Baskerville R., "Aligning Software Processes with Strategy", *MIS Quarterly* Vol. 30 o. 4, Pp. 891-918/December 2006.
[5] Giraud-Carrier C. And Povel O., "Characterizing Data Mining Software", *Intelligent Data Analysis* 7 (2003) 181–192, IOS Press.
[6] Khoshgoftaar T. M., Allen E. B., Jones W. D., Hudepohl J. P., "Data Mining for Predictors of Software Quality", *International Journal of Software Engineering and Knowledge Engineering*, Vol. 9, No. 5 (1999) 547-563.
[7] Alvarez-Mac´Ias J. L. and Mata-V´Azquez J., "Data Mining for the Management of Software Development Process", *International Journal of Software Engineering and Knowledge Engineering*, Vol. 14, No. 6 (2004) 665–695.
[8] Last M., Friedman M., and Kandel A., "Using Data Mining for Automated Software Testing", *International Journal of Software Engineering and Knowledge Engineering*, Vol. 14, No. 4 (2004) 369-393.
[9] Mattsson M. K., Chapin N., "Data Mining for Validation in Software Engineering: an Example", *International Journal of Software Engineering and Knowledge Engineering* Vol. 14, No. 4 (2004) 407-427.
[10] Yan X., Zhang C. and Zhang H., "Identifying Software Component Association with Genetic Algorithm", *International Journal of Software Engineering and Knowledge Engineering*, Vol. 14, No. 4 (2004) 441-447.
[11] C. Talbot, "Conference Review", CRISP-DM Special Interest Group 4th Workshop, March 18th 1999, Brussels, Belgium.
[12] Software Engineering,A practitioner's Approach by RogerS. Pressman, McGrawHill International Edition, 6th Edition
[13] Software Engineering by Sommerville, Pearson Education, 7th edition.
[14]Software Engineering by K.K. Aggarwal & Yogesh Singh, New Age International Publishers
[15]Software Engineering, An Engineering Approach by James F. Peters, Witold Pedrycz, John Wiley.
[16]Software Engineering principles and practice by Waman S Jawadekar, The McGraw-Hill Companies.
[17] Neil Maiden, Sarah Jones. "Agile Requirements: Can We Have Our Cake and Eat It Too?" IEEE Software May/June 2010, pp.87-88.
[18] David Budgen. *Software Design*. Pearson Addison Wesley; 2nd edition (May 15, 2003).1st Edition: 1994.
[19] Diomidis Spinellis. *Code Reading: The Open Source Perspective*. Addison-Wesley Pub Co; Book and CD-ROM edition (May 27, 2003).
[20] Stephen P. Berczuk (with Brad Appleton). *Software Configuration Management Patterns: Effective Teamwork, Practical Integration*. Addison-Wesley, 2003.

