Governors State University
OPUS Open Portal to University Scholarship

All Capstone Projects | Student Capstone Projects

Spring 2016

A Metric for Measuring Customer Turnover Prediction Models

Divyasri Kambhampati
Governors State University

Madhurika Reddy Kommidi
Governors State University

Prathyusha Metla
Governors State University

Follow this and additional works at: http://opus.govst.edu/capstones

 Part of the Computer Sciences Commons

Recommended Citation

Kambhampati, Divyasri; Kommidi, Madhurika Reddy; and Metla, Prathyusha, "A Metric for Measuring Customer Turnover Prediction Models" (2016). *All Capstone Projects*. 210.
http://opus.govst.edu/capstones/210

For more information about the academic degree, extended learning, and certificate programs of Governors State University, go to http://www.govst.edu/Academics/Degree_Programs_and_Certifications/

Visit the Governors State Computer Science Department
This Project Summary is brought to you for free and open access by the Student Capstone Projects at OPUS Open Portal to University Scholarship. It has been accepted for inclusion in All Capstone Projects by an authorized administrator of OPUS Open Portal to University Scholarship. For more information, please contact opus@govst.edu.

Table of Contents

1.Project description

1.1 Abstract

The interest for data mining techniques has increased tremendously during the past decades, and numerous classification techniques have been applied in a wide range of business applications. Hence, the need for adequate performance measures has become more important than ever. In this application, a cost-benefit analysis framework is formalized in order to define performance measures which are aligned with the main objectives of the end users, i.e., profit maximization.

A new performance measure is defined, the expected maximum profit criterion. This general framework is then applied to the customer churn problem with its particular cost-benefit structure. The advantage of this approach is that it assists companies with selecting the classifier which maximizes the profit. Moreover, it aids with the practical implementation in the sense that it provides guidance about the fraction of the customer base to be included in the retention campaign

1.2 Competitive Information

The interest for data mining techniques has increased tremendously during the past decades, and numerous classification techniques have been applied in a wide range of business applications. Hence, the need for adequate performance measures has become more important than ever. The advantage of this approach is that it assists companies with selecting the classifier which maximizes the profit. Moreover, it aids with the practical implementation in the sense that it provides guidance about the fraction of the customer base to be included in the retention campaign.

1.3 Relationship to Other Applications/Projects

In this application, a cost-benefit analysis framework is formalized in order to define performance measures which are aligned with the main objectives of the end users, i.e., profit maximization. A new performance measure is defined, the expected maximum profit

criterion. This general framework is then applied to the customer churn problem with its particular cost-benefit structure.

1.4 Assumptions and Dependencies

In the existing system we mainly concentrate on how to increase the sales and how to attract the customers.

When investigating and comparing these data mining techniques for customer churn prediction, it is imperative to have an adequate performance measure.

Here we concentrate on the new customers.

1.5 Future Enhancements

In the proposed system we mainly concentrate on the dropouts rather than attracting the new customers.

The H measure is a new approach to performance measurement, which overcomes this problem and focuses on misclassification costs In this paper, not only the misclassification costs, but also the benefits originating from a correct classification are explicitly taken into account

1.6 Definitions and Acronyms

Customer Turnover:

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.

H Measure:

The H measure can be motivated in terms of a prior on the severity of the two types of misclassification costs (false alarms versus missed positive cases). Under this interpretation, the prior weights implicitly imposed on the relative misclassification cost by the popular area

under the curve (see left most pivot) are classifier-dependent (rightmost plot), whereas in the case of the H-measure they can be controlled explicitly using a beta prior (centre plot).

Misclassification Costs:

In some contexts, certain kinds of errors are more costly than others. For example, it may be more costly to classify a high-risk credit applicant as low risk (one kind of error) than it is to classify a low-risk applicant as high risk (a different kind of error). Misclassification costs allow you to specify the relative importance of different kinds of prediction errors. Misclassification costs are basically weights applied to specific outcomes. These weights are factored into the model and may actually change the prediction (as a way of protecting against costly mistakes).

Customer Churn Prediction Models:

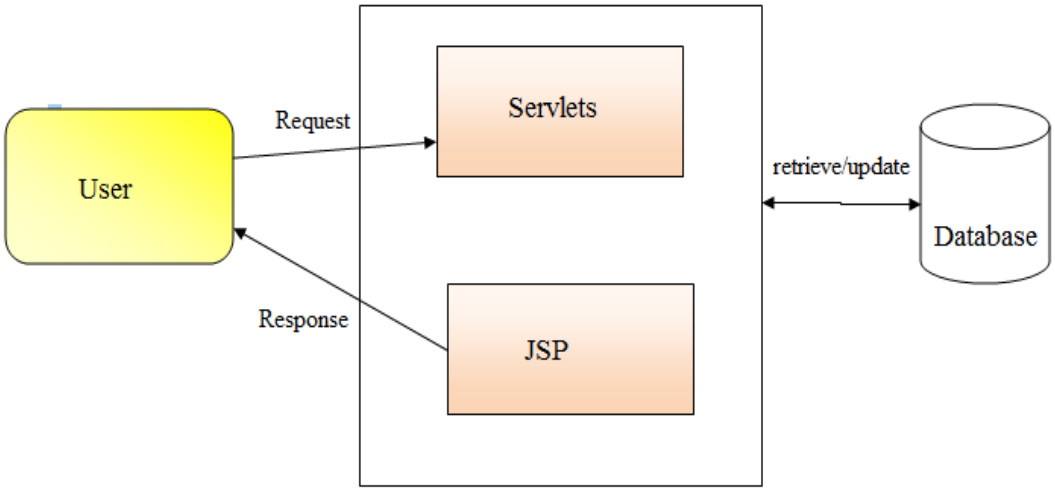
The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every online business. Besides the direct loss of revenue that results from a customer abandoning the business, the costs of initially acquiring that customer may not have already been covered by the customer's spending to date. (In other words, acquiring that customer may have actually been a losing investment.) Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.

2. Technical Description

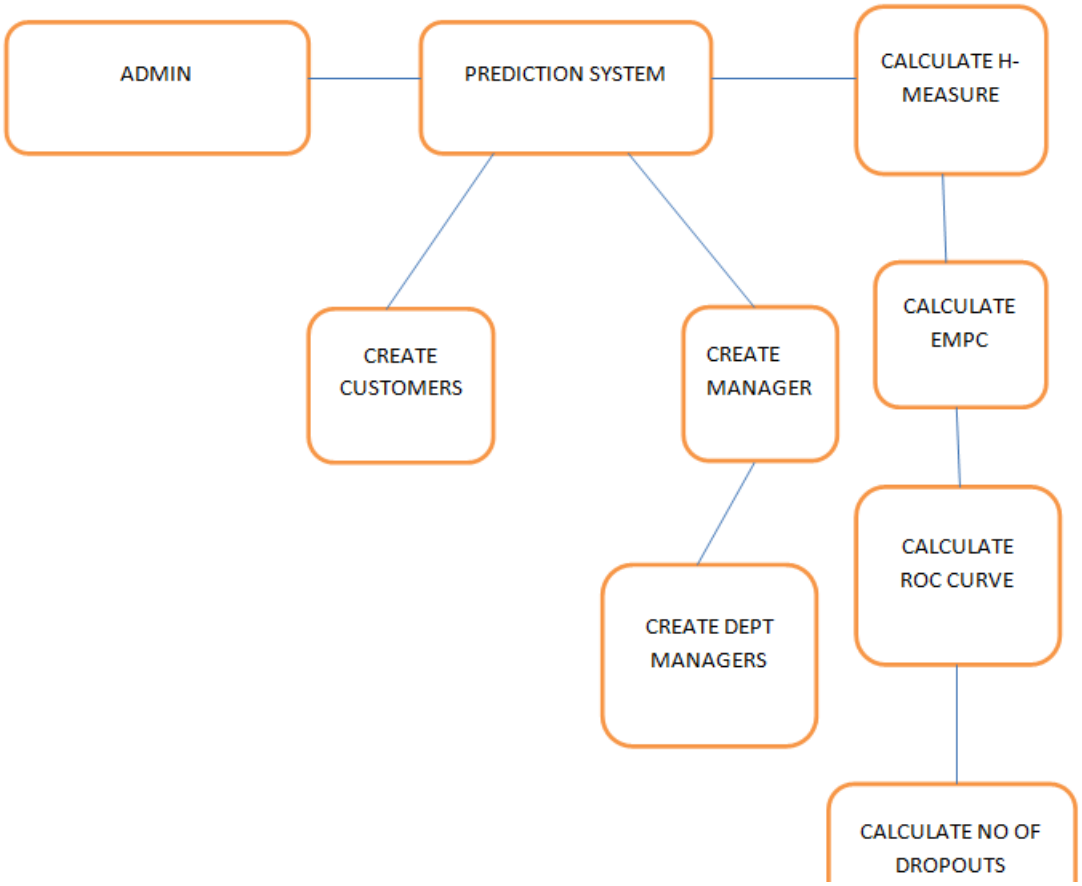
2.1 Project/Application Architecture:

The software architecture of a program or computing system is a depiction of the system that aids in the understanding of how the system will behave.

Software architecture serves as the blueprint for both the system and the project developing it, defining the work assignments that must be carried out by design and implementation teams. The architecture is the primary carrier of system qualities such as performance, modifiability, and security, none of which can be achieved without a unifying architectural vision. Architecture is an artefact for early analysis to make sure that a design approach will yield an acceptable system. By building effective architecture, you can identify design risks and mitigate them early in the development process**.**



2.2 Project/Application Information Flows:



2.3 Interaction with Other Applications:

Modern computer systems are based on design principles that have been used in industries all around the world for many decades. A technical architecture is the design and documentation of a software application. This provides a blueprint schematic for developers to use when they are building or modifying a computer system.

It is important to define the computer architecture before building a software application. The technical architecture typically defines the communication networks, security, hardware, and software that are used by the application. This ensures that all new systems are compatible with the existing computer devices and equipment used in the company.

Companies use enterprise technical architecture to design their computer systems. This plan outlines the hardware, network communication, and software tools that the company uses during daily operations. Having a defined technical architecture ensures that new products and software tools can easily connect to the existing systems of a company.

Many software protocols are used in computer systems. The technical architecture defines what communication protocols are permitted in a network or system. This ensures that communication protocols are flexible and compatible with modern devices and other network equipment used by a company.

2.4 Capabilities:

In this section we will discuss about the analysis of the project. Systems are created to solve problems. One can think of the systems approach as an organized way of dealing with a problem. In this dynamic world, the subject System Analysis and Design (SAD), mainly deals with the software development activities.

2.4.1 Feasibility Study:

All the projects are feasible if they have unlimited resources and infinite time. But the development of software is plagued by the scarcity of resources and difficult delivery rates. It is necessary and prudent to evaluate the feasibility of a project at the earliest possible time. The 3 considerations involved in the feasibility analysis are:

2.4.2 Economic Feasibility:

This procedure is to determine the benefits and savings that are expected from an application, system and compare it with cost. The benefits should outweigh cost. Then the decision made to design and implement the system. Otherwise further justifications and alterations are made in the proposed system that has to be done. This is an on-going effort that improves in accuracy of each phase of the system lifecycle. For the current project we cannot expect any economic feasibility because this project uses open source environments.

2.4.3 Technical Feasibility:

Technical feasibility concentrates on the existing mobile systems (hardware, software etc.) and to what extent it can support the proposed edition. If budget is the serious

constraint, then the project is judged not feasible. The technical feasibilities are important role in the current project is important because here we are dealing with an android operating system.

2.6 Risk Assessment and Management:

People are inherently resistant to change and mobiles have been known to facilitate the change that the people expect. Since our application deals with all the technical people around the world who use an android mobile, operational feasibility, risk assessment and management can be considered negligible.

3.Project Requirements

3.1 Identification of Requirements:

Churn prediction modeling techniques attempt to understand the precise customer behaviors and attributes which signal the risk and timing of customer churn. The accuracy of the technique used is obviously critical to the success of any proactive retention efforts. After all, if the marketer is unaware of a customer about to churn, no action will be taken for that customer. Additionally, special retention-focused offers or incentives may be inadvertently provided to happy, active customers, resulting in reduced revenues for no good reason.

Unfortunately, most of the churn prediction modeling methods rely on quantifying risk based on static data and metrics, i.e., information about the customer as he or she exists right now. The most common churn prediction models are based on older statistical and data-mining methods, such as logistic regression and other binary modeling techniques. These approaches offer some value and can identify a certain percentage of at-risk customers, but they are relatively inaccurate and end up leaving money on the table.

<customerattriction_708-5 user_capability_”0914786”>

3.1.1 Non Functional Requirements:

Performance requirements:

Requirements about resources required, response time, transaction rates, throughput, benchmark specifications or anything else having to do with performance.

3.2 Operations, Administrations, Maintenance, and Provisioning:

Operating constraints:

List any run-time constraints. This could include system resources, people, needed software, ...

Platform constraints:

Discuss the target platform. Be as specific or general as the user requires. If the user doesn't care, there are still platform constraints.

Accuracy and Precision:

Requirements about the accuracy and precision of the data.(Do you know the difference?) Beware of 100% requirements; they often cost too much.

Modifiability:

Requirements about the effort required to make changes in the software. Often, the measurement is personnel effort (person- months).

Portability:

The effort required to move the software to a different target platform. The measurement is most commonly person-months or % of modules that need changing.

Reliability:

Requirements about how often the software fails. The measurement is often expressed in MTBF (mean time between failures). The definition of a failure must be clear. Also, don't confuse reliability with availability which is quite a different kind of requirement. Be sure to specify the consequences of software failure, how to protect from failure, a strategy for error detection, and a strategy for correction.

Security:

One or more requirements about protection of your system and its data. The measurement can be expressed in a variety of ways (effort, skill level, time...) to break into the system. Do not discuss solutions (e.g. passwords) in a requirements document.

Usability:

Requirements about how difficult it will be to learn and operate the system. The requirements are often expressed in learning time or similar metrics.

Legal:

There may be legal issues involving privacy of information, intellectual property rights, export of restricted technologies, etc.

3.3 Security and Fraud Prevention:

Quality Constraints:

Correctness:

Each requirement must accurately describe the functionality to be delivered. The reference for correctness is the source of the requirement, such as an actual customer or a higher-level system requirements specification. A software requirement that conflicts with a corresponding system requirement is not correct (of course, the system specification could itself be incorrect).

Only user representatives can determine the correctness of user requirements, which is why it is essential to include them, or their close surrogates, in inspections of the requirements. Requirements inspections that do not involve users can lead to developers saying, "That doesn't make sense. This is probably what they meant. This is also known as guessing."

Feasible:

It must be possible to implement each requirement within the known capabilities and limitations of the system and its environment. To avoid infeasible requirements, have a developer work with the requirements analysts or marketing personnel throughout the elicitation process. This developer can provide a reality check on what can and cannot be done technically, and what can be done only at excessive cost or with other tradeoffs.

Necessary:

Each requirement should document something the customers really need or something that is required for conformance to an external requirement, an external interface, or a standard. Another way to think of "necessary" is that each requirement originated from a source you recognize as having the authority to specify requirements. Trace each requirement back to its origin, such as a use case, system requirement, regulation, or some other voice-of-the-customer input. If you cannot identify the origin, perhaps the requirement is an example of "gold plating" and is not really necessary.

Complete:

No requirements or necessary information should be missing. Completeness is also a desired characteristic of an individual requirement. It is hard to spot missing requirements because they aren't there. Organize the requirements hierarchically in the SRS to help reviewers understand the structure of the functionality described, so it will be easier for them to tell if something is missing.

Consistent:

Consistent requirements do not conflict with other software requirements or with a higher level(system or business) requirements. Disagreements among requirements must be resolved before development can proceed. You may not know which (if any) is correct until you do some research. Be careful when modifying the requirements, as inconsistencies can slip in undetected if you review only the specific change and not any related requirements.

Modifiable:

You must be able to revise the SRS when necessary and maintain a history of changes made to each requirement. This means that each requirement be uniquely labeled and expressed separately from other requirements so you can refer to it unambiguously. You can make an SRS more modifiable by organizing it so that related requirements are grouped together, and by creating a table of contents, index, and cross-reference listing.

Traceable:

You should be able to link each software requirement to its source, which could be a higher-level system requirement, a use case, or a voice-of-the-customer statement. Also link each software requirement to the design elements, source code, and test cases that are constructed to implement and verify the requirement. Traceable requirements are uniquely labeled and are written in a structured, fine-grained way, as opposed to large, narrative paragraphs or bullet lists

3.4 Release and Transition Plan:

3.4.1 Deployment Details:

When you deploy a multidimensional data mining solution, this solution creates your data mining objects within the same database as the source cube.

When you process the mining structure or mining model, you must process the source cube as well. For this reason, deploying a solution that uses OLAP mining models can take longer than relational data mining solutions.

Typically data mining objects also use the same data sources and data source views that are used for the cube. However, you can add data sources and data source views that are targeted specifically to data mining. For example, typically a cube would not contain data about prospective clients, or external data not used in the multidimensional objects.

4. Project Design Description

We are following Waterfall Model for the project implementation as water fall model is one of the most widely used Software Development Process. It is also called as "Linear Sequential model" or the "classic life cycle" or iterative model. It is widely used in the commercial development projects. It is called so because here, we move to next phase (step) after getting input from previous phase, like in a waterfall, water flows down to from the upper steps.

Requirement Gathering and analysis:

All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification doc.

System Design:

The requirement specifications from first phase are studied in this phase and system design is prepared. System Design helps in specifying hardware and system requirements and also helps in defining overall system architecture.

Implementation:

With inputs from system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing.

Integration and Testing:

All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.

Deployment of system:

Once the functional and non-functional testing is done, the product is deployed in the customer environment or released into the market.

Maintenance:

There are some issues which come up in the client environment. To fix those issues patches are released. Also to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

We opted for waterfall model because it is very simple, easy to understand and easy to use. It is easy to manage due to the rigidity of the model since in this model each phase has specific deliverables and a review process. Phases are processed and completed one at a time. Works well for smaller projects where requirements are very well understood. We generally use waterfall model when requirements are very well known, clear and fixed when product definition is stable when technology is understood when there are no ambiguous requirements and when ample resources with required expertise are available freely.

5.Project Internal/External Interface Impacts and Specifications

Project internal/external internal impacts and specifications is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. It implies a systematic and rigorous approach to design—an approach demanded by the scale and complexity of many systems problems.

Today, ideas from design methods and systems design may be more relevant to designers than ever before—as more and more designers collaborate on designing software and complex information spaces. Frameworks suggested by systems design are especially useful in modelling interaction and conversation. They are also useful in modelling the design process itself.

A systems approach to design asks:

- For this situation, what is the system?
- What is the environment?
- What goal does the system have in relation to its environment?
- What is the feedback loop by which the system corrects its actions?
- How does the system measure whether it has achieved its goal?
- Who defines the system, environment, goal, etc.—and monitors it?
- What resources does the system have for maintaining the relationship it desires?
- Are its resources sufficient to meet its purpose?

A systems approach to design is entirely compatible with a user-centred approach. Indeed, the core of both approaches is understanding user goals. A systems approach looks at users in relation to a context and in terms of their interaction with devices, with each other, and with themselves.

Project Internal Interface Impacts:

Admin:

- Admin logs into the system and creates the packages and makes it available for the company to buy.
- Admin can see the chart analysis to check the customers interacting with the company frequently and concentrate more on them.
- Chart is generated depending on the feedback taken from the customers i.e., based on their interaction with the company.

Company Head:

- Company registers by buying the packages needed for them. After registering a mail containing ID and Password is sent to the registered mail address.
- Company head logs in and registers both product managers and employees he needed.
- Company can also check the chart analysis to know about their customers.

Product Manager:

- Product manager is created by the Company Head.
- Product manager logs in and will add the products, view the products.

Employee:

- Employee is created by the Company Head.
- Employee logs in and will answer to the queries asked by the user.

User:

- User Registers and then logs in and then asks the queries if needed. Also he has the option to select the way in which he needs the answer.
- Depending on the option selected he gets the answer to the query.

Specifications:

Specifications is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The Specifications stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. Computers connected to the net are from many different manufacturers, running on different operating systems and they differ in architecture, computing power and capacity. By considering this point SUN Microsystems Corporation felt the need for a new programming language suitable for this heterogeneous environment and Java was the solution. This breaks barriers between different computers, chips and operating systems.

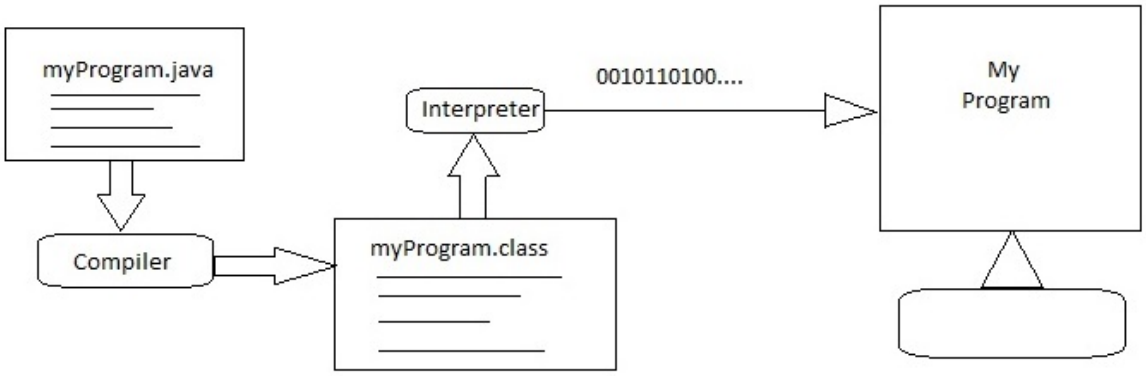
Hibernate ORM (Hibernate in short) is an object-relational mapping library for the Java language, providing a framework for mapping an object-oriented domain model to a traditional relational database. Hibernate solves object-relational impedance mismatch problems by replacing direct persistence-related database accesses with high-level object handling functions.

Hibernate's primary feature is mapping from Java classes to database tables (and from Java data types to SQL data types). Hibernate also provides data query and retrieval facilities. It generates SQL calls and relieves the developer from manual result set handling and object conversion. Applications using Hibernate are portable to supported SQL databases with little performance overhead.

Hibernate is a high-performance Object/Relational persistence and query service which is licensed under the open source GNU Lesser General Public License (LGPL) and is free to download. Hibernate not only takes care of the mapping from Java classes to database tables (and from Java data types to SQL data types), but also provides data query and retrieval facilities.

6.Project Design Units Impacts

Computers connected to the net are from many different manufacturers, running on different operating systems and they differ in architecture, computing power and capacity. By considering this point SUN Microsystems Corporation felt the need for a new programming language suitable for this heterogeneous environment and Java was the solution. This breaks barriers between different computers, chips and operating systems.



"At a time when companies in many industries offer similar products and use comparable technology, many of the previous bases for competition are no longer viable. In a global environment, physical location is frequently not a source of advantage, and protectionist regulation is increasingly rare. Proprietary technologies can often be rapidly copied and attempts to achieve breakthrough innovation in products or services often fail. What's left as a basis for competition is execution and smart decision making. An organisational commitment to and developed capability of Analytics is enabling market-leading companies to succeed in the rapidly evolving arena of global competition." (Davenport and Harris, 2007) The information and data mining software systems facilitate the analytical reasoning process, providing humans with means to deal with the enormous amount of data and information generated in various areas of human endeavour. Since its inception in the late 80s, data and information mining technologies have reached the level of embedded technology, coming as part of modern data management and analysis suites.

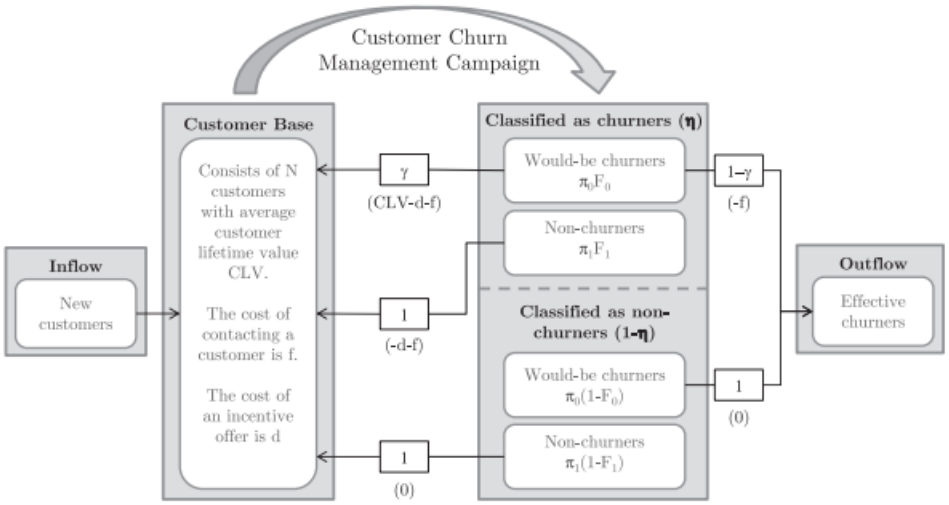
However, technology is only one of the necessary conditions for achieving competitive advantage. The issue of Analytics not being fully utilised in organisations due to lack of a clearly defined analytics process has been recognised for some time. In the late '90's – early 2000's a number of methodologies was developed to address this issue. Among them CRISP-DM (Chapman, et al., 2000) is perhaps the best known and broadly used iterative data mining methodology. However such methodologies are focussed primarily on the technical aspects of the data mining process with little attention to the business aspects of the overall Analytics process (Pyle, 2004). For instance, "Business Understanding" is part of CRISPDM, however, little is provided about how that actually can be done. There is an embedded assumption that the business analysts will somehow communicate with the data miners and in this communication the data mining models will be related to business key performance indicators (KPIs). However, industry leaders have pointed out the existence of a communication gap between data mining experts and business domain experts (Fayyad, 2004). This gap, together with some related issues, has been explored in Van Rooyen's (2004) critical evaluation of the project management utility of CRISP-DM and Data Mining Projects Methodology (DMPM) of the SAS Institute, in a business decision-support environment. Recently there have been attempts to improve existing Analytics methodologies and allow them to become more effective and reliable in providing useful insights in business contexts. A notable contribution to the field is Van Rooyen's (2005) Strategic Analytics Methodology (SAM). In the last few years there has been an improvement in the use of Analytics in business settings. However, it is important to stress that the potential value of Analytics has not been fully realised or utilised in business settings as yet. This paper presents an Analytics process approach that aims to maximise the value-add of Analytics projects. The process draws on the extensive knowledge and experience gained from industry, consulting, research and education activities in Analytics. The paper focuses specifically on the key process stages that have not been identified or given sufficient attention in previous Analytics methodologies. Further the paper is organised as follows. Section 2 considers the reasons why analytics projects may fail. Section 3 discusses solutions to the problems discussed in Section 2. Section 4 illustrates and reflects on the practical application of the stage model of the Analytics process.

6.1. Functional Area/Design UnitA:

Until now, the performance measures MP and EMP have been discussed in general terms, without any assumptions on the distribution of the cost and benefit parameters other than that their values are assumed to be positive. However, when the cost-benefit framework is to be applied, an interpretation has to be given to the parameters, which requires domain specific knowledge. Section 4.1, will discuss the deterministic approach for performance measurement of customer churn prediction. The probabilistic approach will be treated in detail. In this section, a case study will be presented to illustrate the use of the EMPC and H measure for measuring classification performance in a customer churn prediction context. A selection of 21 techniques has been tested on 10 different customer churn data sets. The data sets have been pre-processed to prepare them for the analysis. Table 2 summarizes the main characteristics of the data sets. The classification techniques used in the benchmarking study are listed in Table 3 and include many commonly applied data mining methods. A more extensive discussion regarding the data sets, the pre-processing and the employed techniques can be found in the original benchmarking study, carried out by Verb eke et al.

Each data set has been split into a training set and a test set, where the former was used to train the classifiers, and the latter to perform an out-of-sample test. Thus, each instance in the test set is assigned a score s, on which classification is based by setting a cut off value. Through the empirical score distributions, the AUC, H measure, MPC measure, and EMPC measure are then calculated. Per data set, each measure leads to a ranking of the 21 techniques. Since the ranking is the most important output of the classification performance measurement process, the rankings from the different measures will be compared to one another. As the 10 data sets involved in this study are independent, the 10 resulting rankings of techniques are considered independent observations.

Until now, the performance measures MP and EMP have been discussed in general terms, without any assumptions on the distribution of the cost and benefit parameters other than that their values are assumed to be positive. However, when the cost-benefit framework is to be applied, an interpretation has to be given to the parameters, which requires domain specific knowledge. We will discuss the deterministic approach for performance measurement of customer churn prediction in the below figure.



6.1.1 Functional Overview:

The EMPC measure incorporates uncertainty about the acceptance rate , but for the parameters CLV, , and , fixed values are assumed. In this paragraph, the sensitivity of the EMPC measure to small variations in these fixed parameters will be assessed. Therefore, the first derivative of (18) with respect to CLV is calculated Note that the optimal cut off, depends on the value of CLV. In what follows, the partial derivative within the integration is elaborated. Using yields which, by using can be applied, which leads to the last term between brackets being equal to zero, and thus: In other words, this means that when CLV is changed, holding and constant, EMPC changes proportionally. Note that holding and constant in fact means that the cost of the retention offer as a percentage of CLV, and the cost of contacting a customer as a percentage of CLV remain constant. Analogously, for variations in the following equation can be derived with empc, the expected fraction of the customer base which accepts the retention offer. In fact, this implies that the larger the optimal targeted fraction (or the fraction accepting the offer), the more sensitive the expected maximum profit for variations in (or). Also note that an increase in CLV, while holding d and f constant instead of and , corresponds to a parallel decrease in and . The sensitivity of the rankings will be analyzed in a real-life case study. A further indication of the agreement in ranking is

given in which shows the rank of each technique, averaged over the 10 data sets. The full line represents the ranking according to the EMPC measure, whereas the other points represent other performance metrics. Hence, a point plotted far from the full line shows strong disagreement with the EMPC-based ranking. One can immediately see that the disagreement of AUC with EMPC is larger than any other performance measure. The H measure with optimized parameters follows the EMPC ranking much closer, again indicating that it is a reasonable approximation. A second point of view is the profitability of a retention campaign. As explained before, EMPC ranks classifiers according to the expected profit based on assumptions about the classification costs and benefits. AUC, as shown by Hand [2], is also a measure of profit, but with invalid assumptions for the distribution of costs and benefits. Therefore, choosing a classifier based on AUC may lead to suboptimal model selection from a profit point of view, which is illustrated in Table 4. For each data set, the optimal technique is selected based on EMPC and on AUC, and the expected profit for that particular choice of classifier is given. As indicated in the last column, selection based on AUC leads in some cases to suboptimal model selection, with losses up to C ¼ 0:137 per customer, a substantial amount for telecom operators with millions of customers. Moreover, for selection based on EMPC, it is possible to calculate the fraction of the vast customer base which needs to be targeted to realize the maximal profit, which is also displayed in Table 4. This is one of the major advantages of the EMPC (and also the MPC) measure, as it gives guidance to practitioners about how many customers to include in a retention campaign. When selecting a model with AUC or the H measure, there is no such guidance, and deviating from the optimal fraction may again lead to suboptimal profits.

6.1.2 Impacts:

The implications of the findings presented in this paper are straightforward but essential. Companies rely more than ever on data mining techniques to support their decision making processes. When evaluating a classification technique which is to be used in a business context, it is imperative to base any evaluation criterion on the goal of the end user. Since companies strive for profit maximization, a performance measure evidently should take this into account. The very commonly used area under the ROC curve does have its merits and an interesting interpretation in the sense that the AUC of a classification method is the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. However, as has shown, AUC makes incorrect implicit

assumptions about the misclassification costs, and the use of this performance metric in a business environment leads to suboptimal profits. This paper outlines a theoretical framework which incorporates all gains and losses related to the employment of a data mining technique, and defines a probabilistic performance measure, the expected maximum profit. As each corporate environment has its own specificities, the framework is defined on a general level. To be applicable to a certain business problem, the particularities of its cost and benefit structure need to be incorporated. This process is worked out in detail for the problem of customer churn and an EMP measure for customer churn, EMPC, is derived. Also the link between EMPC and the H measure was investigated and it appears that the latter with appropriately chosen distribution parameters is a good approximation of the former. The performance measure for customer churn is validated in an extensive case study. The results clearly indicate that the use of AUC as a performance metric leads to suboptimal profits. The case study also points to one of the major advantages of the EMPC measure. It does not only select the classifier which maximizes the profit, but it also provides the practitioner with an estimate of the fraction of the customer base which needs to be targeted in the retention campaign. This optimal fraction varies from case to case, and deviating from this fraction again leads to suboptimal profits. Note that the H measure, although it is able to select the most profitable classifier, does not provide guidance on the optimal fraction of the customer base to be included in the retention campaign. Finally, a sensitivity analysis was carried out, to analyse how vulnerable the EMPC measure is to incorrect estimation of the fixed parameters CLV, , and . The outcome shows that the EMPC measure and its resulting ranking is relatively robust with regard to changes in these parameter values. An obvious though not straightforward direction for further research entails the extension of the framework to multiclass problems. Furthermore, besides the application of the cost-benefit analysis framework to the customer churn problem, there remain many business problems where a profit driven performance measure would add value. Among others, data mining techniques are employed for financial credit scoring , direct marketing response models , fraud detection [28], and viral marketing in social networks . With the explosive growth of data, and the increasing popularity of social network sites, companies will rely more than ever on data mining techniques to support their decision making process. In such cases, the cost-benefit analysis framework presented in this paper provides directions on how to develop performance measures tailored to specific business problems.

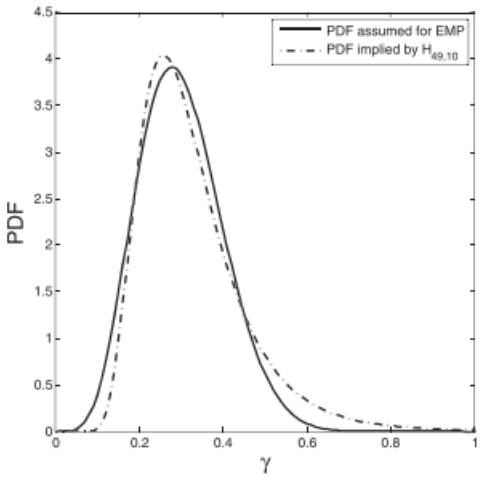
6.1.3 Requirements:

AS a result of the steep growth in computational power, the interest for data mining techniques has increased tremendously the past decades. A myriad of classification techniques has been developed and is being used in a wide range of business applications. As more and more methods are elaborated, the need for adequate performance measures has become more important than ever before. There has been a lot of attention for the receiver operating characteristic (ROC) curve, which is a graphical representation of the classification performance for varying thresholds [1]. However, rather than visually comparing curves, practitioners prefer to capture the performance of a classification method in a single number. A very popular and straightforward concept is the area under the ROC curve (AUC), which is closely related to the Gini coefficient and the KolmogorovSmirnov statistic. The problem with these measures is that they implicitly make unrealistic assumptions about misclassification costs. The H measure is a new approach to performance measurement, which overcomes this problem and focuses on misclassification costs In this paper, not only the misclassification costs, but also the benefits originating from a correct classification are explicitly taken into account. The main rationale is that the most important goal for practitioners is profit maximization. A cost-benefit analysis framework will be worked out, in which two types of performance measures will be defined. The first metric is the maximum profit (MP), whereas the second metric is the expected maximum profit (EMP). The difference between both measures is that MP is a deterministic approach, which assumes that all parameters related to the costs and benefits are accurately known. The EMP measure on the other hand defines a probability distribution for the cost and benefit parameters, and is a probabilistic approach. Analogously to the H measure, also EMP is related to the ROC curve of the classifier, as will be discussed in Section 3. Due to the multitude of business situations in which classification methods are employed, it is difficult, if not impossible, to define one single profit driven performance measure. Section 4 will focus on the prediction of customer churn, which has become a very important business problem. During the last decade, the number of mobile phone users has increased drastically, with expectations of 5.6 billion mobile phone subscribers in 2011,1 which is around 80 per cent of the world population. Hence, telecommunication markets are getting saturated, particularly in developed countries, and the emphasis is shifting from attraction of new customers to retention of the existing customer base. In this context, customer churn prediction models play a crucial role and they are increasingly being researched (see, e.g., the extensive literature overview given by

Verbeke et al.) and new approaches such as the use of social network data are explored (see among others). When investigating and comparing these data mining techniques for customer churn prediction, it is imperative to have an adequate performance measure. Therefore, with the guidance of the cost-benefit framework, a novel performance measure which consistently incorporates the losses and gains will be developed. The main advantage of this metric is that it not only unambiguously identifies the classifier which maximizes the profit for a given customer retention campaign, but also determines the fraction of the customer base which should be targeted to maximize the profit. This is a crucial help for practitioners, since deviating from the optimal fraction leads to suboptimal profits. Section 5 explores the link with the H measure and shows that the H measure can be used as an approximation for the EMP measure for customer churn. Finally, the developed performance measures will be tested in an extensive case study, of which the findings are reported.

6.2. Functional Area/Design UnitB:

To employ the MPC and EMPC measures, it is necessary to obtain values for the parameters in the profit function. The values for these parameters, such as CLV, are industry specific and may even vary from company to company. This paper focuses on customer churn prediction in the telecom industry and estimates for the parameters are based on values reported in the literature and information from telecom operators. When a specific company uses MPC or EMPC for selecting the optimal model, they can plug in their own estimates for the parameters. In this study, the values of the parameters CLV , d, and f are taken as C ¼ 200; C ¼ 10, and C ¼ 1, respectively. The parameter , representing the response rate, is much more difficult to estimate. For the MPC criterion, a single point estimate would be required, which corresponds to one degree of freedom. For the EMPC criterion, however, there are two degrees of freedom. This enables the practitioner to define an expected value and a standard deviation, where the latter accounts for the uncertainty in the practitioner's estimation. There is only one restriction, namely that need to be strictly greater than one in order to obtain a uni modal beta distribution. Neslin et al. assumed values ranging from 10 to 50 percent, therefore this paper proposes the expected value and standard deviation of the acceptance rate to be equal to 30 percent and 10 percent respectively. This leads to the parameters 0 being 6 and 14, respectively, which corresponds to the probability density function plotted in Fig. 2 (solid line). When MPC is calculated, a point estimate for y is required, in which case the expected value, equal to 30 percent, is taken.



6.2.1 Functional Overview:

The H measure with optimized parameter values, however, shows very high correlation and lowest variability, indicating that both rankings agree to a large extent and that this is a reasonable approximation to the EMPC measure. Finally, the correlation between MPC and EMPC is again lower, which can be attributed to the fact that the underlying cost and benefit assumptions are different, i.e., it is a deterministic versus a probabilistic approach. This also suggests that when the operator has accurate estimates of the response rate , it is preferable to use MPC as criterion. When there is more uncertainty involved with the estimation of a probabilistic approach, and thus EMPC, is recommended. Furthermore, the box plot shows an outlier for the correlation between EMPC and AUC, H2;2, and H49;10. These outliers correspond to the data set D2, where the expected maximum profit is zero for all techniques. As a result, EMPC and MPC consider all techniques to be unprofitable, and they all receive the same rank. AUC, H2;2, and H49;10 on the other hand will rank techniques differently. Therefore, the correlation is low for this data set. A further indication of the agreement in ranking is given in Fig. 4, which shows the rank of each technique, averaged over the 10 data sets. The full line represents the ranking according to the EMPC measure, whereas the other points represent other performance metrics. Hence, a point plotted far from the full line shows strong disagreement with the EMPC-based ranking. One can immediately see that the disagreement of AUC with EMPC is larger than any other performance measure. The H measure with optimized parameters follows the EMPC ranking much closer, again indicating that it is a reasonable approximation. A second point of view is the profitability of a retention campaign. As explained before, EMPC ranks classifiers according to the expected profit based on

assumptions about the classification costs and benefits. AUC, as shown by Hand [2], is also a measure of profit, but with invalid assumptions for the distribution of costs and benefits. Therefore, choosing a classifier based on AUC may lead to suboptimal model selection from a profit point of view, which is illustrated in Table 4. For each data set, the optimal technique is selected based on EMPC and on AUC, and the expected profit for that particular choice of classifier is given. As indicated in the last column, selection based on AUC leads in some cases to suboptimal model selection, with losses up to C ¼ 0:137 per customer, a substantial amount for telecom operators with millions of customers. Moreover, for selection based on EMPC, it is possible to calculate the fraction of the vast customer base which needs to be

targeted to realize the maximal profit.

6.2.2. Impacts:

The impact of variations in CLV, , and on the estimated expected profit and analytically derives first order approximations for this sensitivity. This yields some straightforward rules of thumb for the sensitivity, such as, e.g., the higher the targeted fraction, the more sensitive the profit to changes in . However, the question arises how the ranking between classification algorithms is affected. Therefore, the results of the case study are used to analyse this impact. First, the techniques have been ranked with the proposed values for CLV, , and . Then, each parameter has been multiplied with a constant (while holding the others equal), and the techniques have been ranked again with this new parameter value. The correlations between the ranking in the base scenario and the ranking in the new scenario have been plotted for varying values of the multiplier, ranging from 1/2 to 2 shows these results for all three fixed parameters. The median and the first and third quartile (over the 10 data sets) have been plotted. Note that the plot for CLV assumes that CLV is changed while d and f are held constant (not and). It can be seen that variations in both CLV and have a similar impact, with median correlations decreasing until 0.8. The impact of is virtually nonexistent. Furthermore, when the ranking changes, also the best performing technique (with the highest expected profit) may change. Therefore,displays the impact of suboptimal classifier selection due to incorrect parameter values on the profitability. The percentage loss is plotted on the right axis. Again, the impact is most significant for CLV and , whereas variations in do not impact the profitability. But even though there is an impact, it is relatively limited to losses of maximal 20 percent, and this for substantial variations in the parameters (doubling or halving the CLV or). These results, both in terms of correlation

between rankings and percentage loss due to incorrect classifier selection, indicate that the EMPC measure and corresponding rankings are robust to changes in the fixed parameters. The impact only becomes noticeable for multipliers smaller than 0.75 or larger than 1.5.

6.2.3 Requirements:

schematically represents the dynamical process of customer churn and retention within a customer base. New customers flow into the customer base by subscribing to a service of an operator, and existing customers flow out of the customer base by churning. When setting up a churn management campaign, the fraction of the customer base with the highest propensity to churn is contacted at a cost f per person and is offered an incentive with cost d. In this fraction, there are true would-be churners and false would be churners. In the latter group everyone accepts the incentive and does not churn, as they never had the intention. From the former group, a fraction accepts the offer and thus results in gained customer lifetime value (CLV), whereas the fraction ð1 Þ effectively churns. In the fraction ð1 Þ, which is not targeted, all would-be churners effectively churn, and all non churners remain with the company. The benefits per customer related to each flow are shown between brackets in Fig. 1. These are incremental benefits, as compared to not undertaking the customer churn management campaign. This process was described by Neslin et al. , who established the following expression for the total profit of a retention campaign.

with the fraction of the customer base that is targeted, CLV the customer lifetime value, d the cost of the incentive, f the cost of contacting the customer, and A the fixed administrative costs. The lift coefficient, , is the percentage of churners within the targeted fraction of customers, divided by the base churn rate, 0. Lastly, is the fraction of the would-be churners accepting the offer, or alternatively it is interpreted as the probability of a targeted churner accepting the offer and thus not churning. It is assumed that CLV , A, f, and d are positive, and that CLV > d. Note that depends on the choice for the threshold t, and thus the company has influence on the size of the targeted fraction. Equation (can be expressed in terms of the score distributions. Moreover, if the average rather than the total profit is considered, and the fixed cost A, irrelevant for classifier selection, is discarded, it is possible to obtain a functional form equivalent to the expression for P in . To work out the conversion.

A Metric for Measuring Customer Turnover Prediction Models

As pointed out by Verbeke et al. , the MPC criterion is preferred over the commonly used top-decile lift. Setting the targeted fraction of customers to, e.g., 10 percent is a purely arbitrary choice and most likely leads to suboptimal profits and model selection. Since the ultimate goal of a company setting up a customer retention campaign is to minimize the costs associated with customer churn, it is logical to evaluate and select a customer churn prediction model by using the maximum obtainable profit as the performance measure. Moreover, it has another advantage, which is very appealing to practitioners, in the sense that it is possible to determine the optimal MPC fraction. This quantity represents how many customers should be targeted for profit maximization.

The Functional Requirements include:

- Admin logs into the system and creates the packages and makes it available for the company to buy.
- Company registers by buying the packages needed for them. After registering a mail containing ID and Password is sent to the registered mail address.
- Company head logs in and registers both product managers and employees he needed.
- Product manager logs in and will add the products, view the products.
- Employee logs in and will answer to the queries asked by the user.
- User Registers and then logs in and then asks the queries if needed. Also he has the option to select the way in which he needs the answer. Depending on the option selected he gets the answer to the query.
- Now based on the queries asked and answered by the employees chart analysis is done based on the products of the company.
- This chart analysis is done separately for each company which can be seen both by the admin and the company.
- Admin can see the chart analysis to check the customers interacting with the company frequently and concentrate more on them.
- Company can also check the chart analysis to know about their customers.

7.Open Issues

The receiver operating characteristic curve is a concept which has been extensively used in the machine learning community. This Section discusses the ROC curve in the context of cost-benefit analysis, along with a derived performance measure, the Area under the ROC-curve, also known as AUC . However, showed recently that AUC is a flawed measure and proposed an alternative, the H-measure, which will be treated in The ROC-Curve and the Area under the ROC-Curve A ROC curve is a graphical representation of the classification performance with varying threshold t. It is a plot of the sensitivity versus one minus the specificity, i.e., F0ðtÞ as a function of F1ðtÞ. The interested reader is referred to for an extensive discussion on ROC-curves. Because ROC-curves are not convenient to compare classifiers, especially when their ROC-curves intersect, the area under the ROC-curve is often used to quantify the performance. A larger AUC would indicate superior performance. The area under the ROC curve, which is closely related to the Gini coefficient and the Kolmogorov-Smirnov statistic, has the following statistical interpretation: the AUC of a classification method is the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance . Also in the context of the cost-benefit analysis framework, the ROC-curve has its merit. Since each point on the ROC-curve corresponds to a threshold t, the optimal cut off T, is located somewhere on the curve.

This theorem is a direct result from the definition of the ROC curve and the first order condition as defined by. When the ROC-curve is not convex, and thus the slope is not monotonically decreasing, there may be multiple points on the ROC curve satisfying the first order condition. More specifically, points in the concave region are not optimal for any , in which case the convex hull needs to be defined. Theorem 2. The convex hull of a (nonconvex) ROC-curve defines a set of points where each point corresponds to the optimal cutoff for a certain value of the cost benefit ratio 2 ½0; þ1Þ. This theorem is based on the second order condition for profit maximization and is discussed in detail by Fawcett [1]. The interesting implication of this fact is that an integration over a range of values is equivalent to an integration over the corresponding part of the ROC curve. In fact, every value for 2 ½0; þ1Þ, has a corresponding isoperformance line with a slope equal to 1=0. The optimal cut off for a given value is then located where the isoperformance line is tangent to the ROC curve. Thus, by varying from zero to infinity, each point on the ROC curve is traversed. This interpretation of an integration over the ROC curve has another very important consequence,

as was revealed by Hand . He showed that the popular AUC is equivalent to an expected maximum profit measure. However, the probability density which is implicitly assumed when calculating the AUC depends on the empirical score distribution of the classifier itself. Therefore, AUC is seriously flawed as a performance measure, and Hand proposed an alternative, the H measure.

To calculate a concrete value for the expected minimum loss, assumptions have to be made regarding the probability density of b and c. A first assumption is the independence of b and c, namely that wðb; cÞ ¼ uðcÞvðbÞ holds true, where wðb; cÞ is the joint probability density function of b and c, whereas uðcÞ and vðbÞ are the marginal probability density functions of c and b, respectively. The expected minimum loss (L) is then equal to L.

8.Acknowledgements

It is our pleasure to express our whole hearted thanks to Mr.Nelson Chen for providing the necessary guidelines and facilities to undergo this project.

We reveal our sincere gratitude to the Advisor Dr.Soon-ok Park for giving us the opportunity to do this project.

We also very much thankful to the members of the Computer Science Department and also the members of our team for their extreme support and guidance in making this project successful.

We express our deep sense of gratitude to Mr.Nelson Chen for his valuable guidance and supervision at every stage.

9.References

[1] T. Fawcett, “An Introduction to ROC Analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006. [2]

[2] D. Hand, “Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve,” Machine Learning, vol. 77, no. 1, pp. 103-123, 2009.

[3] W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques,” Expert Systems with Applications, vol. 38, pp. 2354-2364, 2011.

[4] A.A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, “Analyzing the Structure and Evolution of Massive Telecom Graphs,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 5, pp. 703-718, May 2008.

[5] S. Ali and K. Smith, “On Learning Algorithm Selection for Classification,” Applied Soft Computing, vol. 6, no. 2, pp. 119-138, 2006.

[6] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, “Assessing the Accuracy of Prediction Algorithms for Classification: An Overview,” Bioinformatics, vol. 16, no. 5, pp. 412-424, 2000.

[7] N. Chawla, “Data Mining for Imbalanced Datasets: An Overview,” Data Mining and Knowledge Discovery Handbook, pp. 875-886, Springer, 2010.

[8] P. Domingos, “Metacost: A General Method for Making Classifiers Cost-Sensitive,” Proc. Fifth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 155-164, 1999.

[9] F. Provost and T. Fawcett, “Robust Classification for Imprecise Environments,” Machine Learning, vol. 42, no. 3, pp. 203-231, 2001.

[10] A. Bernstein, F. Provost, and S. Hill, “Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 4, pp. 503-518, Apr. 2005.

[11] Z. Zhou and X. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 63-77, Jan. 2006.

[12] C. Elkan, "The Foundations of Cost-Sensitive Learning," Proc. Int'l Joint Conf. Artificial Intelligence, vol. 17, no. 1, pp. 973-978, 2001.

[13] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," European J. Operational Research, vol. 218, no. 1, pp. 211-229, 2012.

[14] A. Bradley, "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[15] R. Prati, G. Batista, and M. Monard, "A Survey on Graphical Methods for Classification Predictive Performance Evaluation," IEEE Trans. Knowledge and Data Eng., vol. 23, no 11, pp. 1601-1618, Nov. 2011.

[16] S. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. Mason, "Detection Defection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," J. Marketing Research, vol. 43, no. 2, pp. 204-211, 2006.

[17] J. Burez and D. Van den Poel, "CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services," Expert Systems with Applications, vol. 32, pp. 277-288, 2007.

[18] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky, "Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry," IEEE Trans. Neural Networks, vol. 11, no. 3, pp. 690-696, May 2000.

[19] T. Verbraken, W. Verbeke, D. Martens, and B. Baesens, "Profit Optimizing Customer Churn Prediction with Bayesian Network Classifiers," accepted for publication in Intelligent Data Analysis, 2013.

[20] J. Hur and J. Kim, "A Hybrid Classification Method Using Error Pattern Modeling," Expert Systems with Applications, vol. 34, no. 1, pp. 231-241, 2008.

[21] A. Lemmens and C. Croux, "Bagging and Boosting Classification Trees to Predict Churn," J. Marketing Research, vol. 43, no. 2, pp. 276-286, 2006.

[22] E. Lima, C. Mues, and B. Baesens, "Domain Knowledge Integration in Data Mining Using Decision Tables: Case Studies in Churn Prediction," J. Operational Research Soc., vol. 60, no. 8, pp. 1096- 1106, 2009.

[23] A. Van der Vaart, Asymptotic Statistics. Cambridge Univ Press, 2000.

[24] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Pearson Education, 2006.

[25] I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, Inc., 2000.

Sites Referred

- http://java.sun.com
- http://www.sourcefordgde.com
- http://www.networkcomputing.com/
- http://www.roseindia.com/
- http://www.java2s.com/

10.Appendices

The following are included in this appendix:

1. Architecture and design flow of the whole project

2. How project iterates through a series of steps in carrying out the desired output generation.

3. Sample of features

4. The graphs that depicts the functioning of the project as expected.

5. Finding the Areas where there is scope for maximizing the profit by reducing the misclassification costs.