Naive Bayes Based Approach on Telecommunication Churn Prediction Using Financial Data

Buddhika Livera [1], D D M Ranasinghe [2]
[1,2] Department of Electrical and Computer Engineering, The Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka.
email: [1] wbslivera@slt.com.lk, [2] ddran@ou.ac.lk

***Abstract*—In the telecommunication industry, identifying customers who tend to leave as early as possible has become a mandatory task for survival. The motive behind the early detection of churn is that with the oversaturated market, it costs more resources to gain a new market share than to recover a lost market share. Therefore, the most effective approach to sustain is the prevention of churn. In general, in the telecom industry revenue generation system is sepa-rated from the management systems of the company. However, through the revenue realization process, finance data received gives some insight into customer data. This study shows how those insights could be applied with a Naive Bayes based approach to predict churn with the accuracy of over 85%. As the suggested approach is based on financial data, it enables the integration with management systems such as ERP compared to others which are based on consumer usage**

Keywords— Telecom Churn, ERP, Naïve Bayes, Fintech

I. INTRODUCTION

At present, there are many telecommunication service providers, and one single service provider may not be able to maintain the market share as the most popular among the customers throughout the timeline. It is very often that people express dissatisfaction with some aspect of their service provider, it could be either quality issues, unwanted product activations, poor customer service, unfair pricing plans, or confusing billing schedules. Due to the business nature, the industry is more vulnerable to churn. Also, it is not only the frustration that causes churn, in an over penetrated market availability of options also contributes to churn.

The churn itself is complex and a challenging problem because it causes more ad-verse effects on the service providers in terms of effort and cost required to attract new customers than to regain the left customers [1]. Due to this very reason, all most all the leading companies are spending a considerable amount of resources to retain their customer base to achieve a consistent customer base.

There have been many efforts and research carried out to identify the potential of churn with various success rates. Though there are many prediction models with different approaches such as Yu Zhao, Bing Li with One-Class Support Vector Machine [2], and NNA Sjarif, NF Azmi with Multilayer Perceptron [3], the main source of data is based on the consumer usages and patterns generated by revenue generation (usually the billing system data) system. Because of the decoupling nature between management systems and billing systems in the business domain, relating prediction outcomes with other systems become complex or an impossible task. Instead of using consumer usage data, this paper proposes an approach based on financial information generated during a typical revenue realization process to predict the churn.

A Typical Telecommunication Business Model

In a modern telecommunication business setup, there is a billing system that drives core revenue generation with other facilitating systems such as Customer Relation-ship Management (CRM), Operation Support (OSS), and cashiering. However, to manage supportive business operations, such as finance, human resources, procurement, and inventory-related transactions, it is common to use Enterprise Resources Planning (ERP) systems or similar, in place by integrating with other systems [4]. One such typical setup is presented in figure 1, which is the setup of one of the leading telecommunication service providers in Sri Lanka.

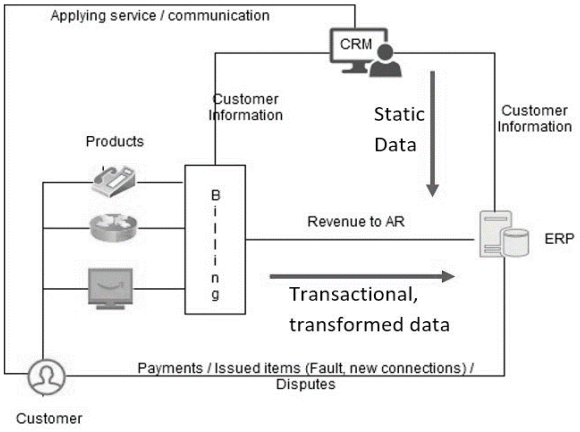


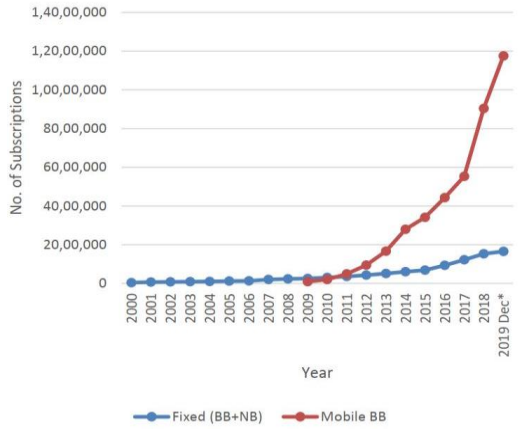
Fig. 13. Set up of a modern telecommunication service provider

Set up of a modern telecommunication service provider

As illustrated in the figure. 1, there are various activities related to different systems. Yet, some insights of all most all the activities can be obtained by analyzing the ERP management system.

A. Business Environment

This study was carried out in a rapidly changing business environment with many opportunities for growth due to the recent overwhelming extension of the ICT sector of the country, during the last decade. The statistics of the following figure 2, from the Telecommunications Regulatory Commission of the country indicates the growth of the telecommunication industry during the past two decades [5].



BB – Broadband / NB – Non-Broadband

Fig. 2. Fixed & Mobile Broadband Growth (2000 – 2019)

The approach and the model discussed in this study are based on historical data obtained from an ERP system of a leading telecommunication service provider in the south Asian region located in Sri Lanka. As per the company profile presented in its annual report of 2019, it has a market share of 1.4 million fixed voice, 0.5 million IPTV, and 1 million fixed broadband connections [6].

This paper contains four main sections describing the various aspect of the study. In the subsequent section, the related work is discussed to elaborate theoretical grounds, followed by the methodology explaining the data and suggested approaches to realize a prediction model. The paper is concluded with the results and discussion describing the findings and future work.

II. Related Work

Ever since the problem of telecommunication churn appears, there have been many efforts to identify churn on an individual and organizational scale. In the published paper by Dahiya, K., and Talwar K, reviews and summarizes the most common approaches that could achieve high accuracy rates on predicting churn [7]. In their study, they concluded that the decision tree-based techniques, Neural Network-based techniques, and regression techniques are generally applied in predicting customer churn in the telecommunication domain. It also highlights how decision tree-based techniques outperformed some of the existing data mining techniques such as regression in terms of accuracy while the neural networks exceed the performance of decision tree-based techniques due to the size of datasets used and the capability of different feature selection methods.

A group of researchers has also done a similar evaluation of various models on sub-scriber dissatisfaction and improving retention in the wireless telecommunications industry [8]. In their paper, they categorized features from a dataset of 46,000 primary business subscribers into five categories, network, billing, application for service, Market, and demographics. They noticed that the features under the market could be ignored when the prediction is made for shorter periods of intervals as for a shorter period market remain the same. As the outcome of the research, the paper claims a reduction in the churn of 40% because of early identification and treatment over a period of six weeks. Following table 1, summarize the categorization of features they identified and a point to highlight is that like many other efforts, the source of the features is the billing system.

Though there are many proposals to predict churn, based on the datasets, the different model tends to result in various success rates. For instance, in the research Simsek Gursoy, Tugba [9] has compared regression techniques with decision tree-based techniques and found that in logistic regression, the accuracy of the analysis churn prediction is 66%, wherein the case of decision trees the accuracy measured is 71.76%. However, in the paper published by Saini, Nisha, Monika, and Garg, Kanwal claim that with the decision tree-based approach, they could achieve 90% accuracy [10].

In the area of focused revenue and customer behavior in the telecommunication domain, few papers highlight the effective application of Bayesian networks. A re-search paper published by Geng Cui and M Leung Wong [11] proposes a Bayesian network model together with evolutionary programming for effective direct marking in the same domain.

Many organizations increasingly move towards ERP systems because it unites the entire functions of the company by maintaining many transactions under a single system. Due to this very nature of the ERP systems, there has been much interest in deriving intelligence from ERP systems. According to the findings by Rouhani, Saeed & Mehri, M [12] proposes that, ERP implementation promotes information-based decision making of an organization that enhances the organizations' business

Table 1. Factor Importance and Nature of data required for prediction

Factor	Importance	Nature of data
call quality	21%	network
pricing options	18%	market, billing
corporate capability	17%	market, customer service
customer service	17%	customer service
credibility / customer communications	10%	market, customer service
roaming / coverage	7%	network
handset	4%	application
billing	3%	billing

intelligence readiness.

Based on the above presented research work, it is possible to conclude that majority of efforts on predicting telecommunication churn depends on features extracted from consumer usage data. As a result of the nature of data that is mostly numerical, logistic regression-based techniques have become the preferred techniques over other techniques for predicting churn in the telecommunication industry. Finally, an interesting fact to be highlighted is that the prediction of the churn has been an individual activity that caused difficulties when integrating or extending prediction models with other advanced intelligence techniques.

III. METHODOLOGY

In this study, the methodology is developed in four main stages, data acquisition and transformation, detection of important attributes, model construction, and model validation. Also, to evaluate the performance of the suggested model in this paper critically, the transformed and filtered data was processed with an SVM model because in previous research SVM based approaches could provide reasonable results for predicting churn [2].

A. Data Acquisition

The acquisition of data was carried out from all geographical regions of the service provider for the period of 2019-20. Following the research domain, relevant information of the customers available in ERP (which was the organizations' business operation management system) was identified and reviewed with domain experts who manage the regional and billing operations. Those attributes with details are presented in table 2. The dataset extracted includes 30,000 subscribers covering all geographical operational area of the service provider. As the suggested method is a probabilistic model, the average of the ratios between disconnected customers and active customers for the last six months was maintained while acquiring data.

Another important fact in data acquisition is the timestamp of data extraction. The suggested approach tries to predict the probability of churning before three billing cycles (each cycle usually 30 days). As illustrated in figure 3, for each subscriber the extraction point was considered as three billings cycles before the date of termination of

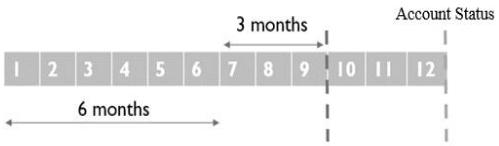


Fig. 14. Timeline of data extraction for terminated customer.

Table 2. Identified features for the prediction of churn.

Attribute	ERP Module	Description	Type
Billing Account	AR Account	Unique identification of the customer.	N/A
Connection Media	AR Account	Copper, LTE or FTTH, connection media. [COPPER, LTE, FTTH]	Categorical
Internet	AR Account	whether Subscribed for internet service or not [YES, NO]	Categorical
IPTV	AR Account	whether Subscribed for IPTV service or not [YES, NO]	Categorical
RTOM	AR Account	Geographical region of service providers' operation. (38 Regions)	Categorical
City	AR Account	City within each RTOM.	Categorical
Loyalty	AR Account	[TX date OR current date] - Subscribed date in years	Continuous
Faults	Inventory	Number of faults based on inventory issues for maintenance.	Continuous
Bill Totals	AR Invoice	Bill amounts of last 12 months	Continuous.
Payments	AR Receipts	Payment amounts of last 12 months	Continuous
Status	AR Account	[Label] Whether active (OK) or terminated customer (TX)	Target

service.

Timeline of data extraction for terminated customer.

B. Data Transformation

Even though all attributes have a potential to maintain a good relationship with the target value, bill totals and payments over a period of one year is not possible to apply directly due to following factors,

Over the time, tariff rates can be changed.

A customer has credit period to after bill generation to pay.

Usage can be high or low due to a temporary period such as Christmas or similar cultural events.

Due to the above factors, two features, Bill Totals, and payments were transformed by the following two equations. First equation (1) is the ratio of billing to payments for the last three months referred to as ratio3 while the second equation (2) is the ratio between bill to payments for the last three to six months referred to as ratio6.

$$Ratio3 = \sum_{i=1}^{3} \frac{B_i}{P_i} \quad (1)$$

$$Ratio6 = \sum_{i=4}^{9} \frac{B_i}{P_i} \quad (2)$$

C. Feature Importance

According to Jasmina Novakovic [13] in the paper titled "The Impact of Feature Selection on the Accuracy of Bayes Classifier" indicates that efficiency in all aspects of a Naïve Bayes model depends on the method of feature selection. Further, it concludes that depending on the features, the best feature selection method for the Naïve Bayes model may differ. Also, as elaborated in the related work section, there is evidence of previous research that some of the features are sensitive to market changes. Therefore, in this study minimum description method is employed before feeding the training data set into model building.

Minimum Description Length

Minimum description length (MDL) treats both models and data as codes. The key idea is that any data set can be appropriately encoded with the help of a model by uncovering underlying regularities in the data. Thus, the code length is directly related to the generalization capability of the model, where the model that provides the shortest description of the data should be chosen [14].

Following table 3, illustrates the feature importance calculated on a small sample via MDL, It indicates that according to the chosen sample attributes internet connectivity and the city has zero importance, and also it could correctly identify account number has no importance as it must be.

Table 34. Feature importance calculated via MDL algorithm

Name	Type	Rank	Importance
RATIO3	NUMBER	1	0.2971
RATIO6	NUMBER	2	0.2177
LOAYALTY	NUMBER	3	0.1731
FAULTS	NUMBER	4	0.0282
CONN_MEDIA	VARCHAR2	5	0.0099
RTOM	VARCHAR2	6	0.0063
IPTV	VARCHAR2	7	0.0005
BB	VARCHAR2	8	0.0000
BILLING_ACCOUNT	VARCHAR2	8	0.0000

Table 3 above illustrates the feature importance measured by the MDL algorithm on one instance of the data set. However, as discussed earlier in this section, the feature importance of each attribute varies based on the season. Therefore, it is possible for some attributes from the original data set to be excluded by the MDL algorithm, making refined data set with high feature importance to input for building the model.

Naïve Bayes

A Bayesian network represents the causal probabilistic relationship among a set of random variables, their conditional dependencies, and it provides a compact representation of a joint probability distribution. It consists of two major parts: a directed acyclic graph and a set of conditional probability distributions.

Mechanism of Naïve Bayes is based on joint probability distributions, for example let $\{x_1, x_2, x_3 \dots x_n\}$ be some events just like being voice only customer, already disconnected customer. In Bayesian Network, they can be represented as nodes. Now if a node has some dependency on another node then an arrow/arc is drawn from one node to another. It is interpreted as the child node's occurrence is influenced by the occurrence of its parent node. So, Bayesian Network represents a directed acyclic graph and now via conditional probability and chain rule shown in below equation (3), it is possible to get the full joint distribution i.e., the probability of the final event (churn customer) given all other dependent events.

$$P(x_1, x_2, x_3, \dots\dots, x_n) = P(x_1).P(x_2|x_1).P(x_3|x_2, x_1) \dots\dots\dots\dots P(x_n|x_1, x_{2,\dots\dots,} x_n)$$
$$= \pi(x_i|Parents(x_i)) \quad (3)$$

Zero Frequency problem

One of the disadvantages of the Naive-Bayes approach is that when there are no occurrences of a class label and a certain attribute value together in the training data set, then the frequency-based probability estimate will be zero. Hence, in a scenario where there are missing class labels, predictions can be completely wrong unless other features can overcome the zero influence.

The algorithm illustrated in figure 4 was employed to avoid the zero-frequency problem. The key idea behind the algorithm is to ensure that the categorical feature, RTOM which is a geographical attribute is fairly distributed in the training set while a numerical feature loyalty distribution is balanced in the training set.

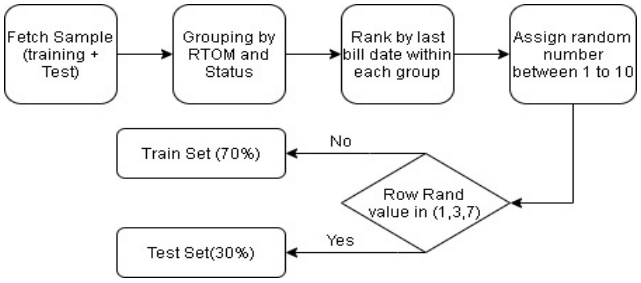


Fig.4. Flow Chart of Sampling Algorithm Used to Avoid Zero Frequency Problem

Flow Chart of Sampling Algorithm Used to Avoid Zero Frequency Problem

IV. Model Architecture

Following Fig 5 is the suggested architecture to predict customer churn in the telecommunication industry based on financial data. It consists of two major components; the first component is for automated sampling by filtering the importance of a feature at a given instance. Remaining part is the Naive Bayes model that calculates the probability of churning.

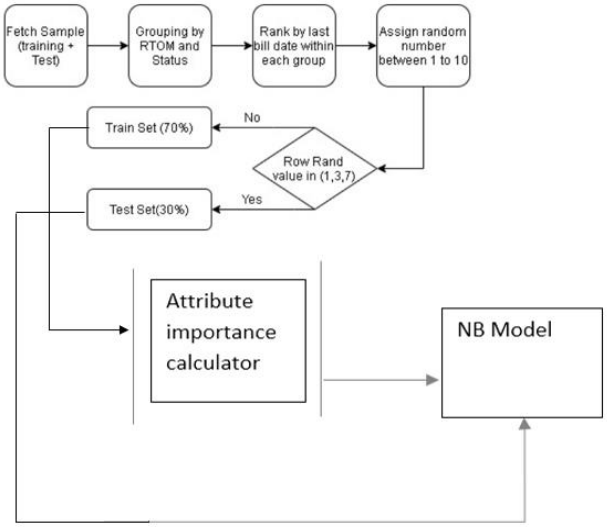


Fig. 5 Suggested Architecture of the complete prediction model

V. Results

The developed model was applied around 16,000 instances, and the performance was evaluated based on criteria related to binary classification problems which include, Rate of true positives (TP), Rate of false positives (FP), Precision, and Recall. Finally, the same set up was tested by switching the model to SVM and compare the performance between models.

A. Model Performance

Following confusion matrix summarizes the performance of the model. In this table TX refer to terminated customer, OK refers otherwise.

	OK	TX	Total
OK	6,419	1,368	7,787
TX	1,164	6,392	7,556
Total	7,583	7,760	15,343

Fig. 6 Confusion Matrix for the suggested approach

B. Confusion Matrix for the suggested approach

According to the confusion matrix illustrated in the figure 6, the performance of the suggested approach can be evaluated as follows.

Recall = TP / (TP + FN) = 84.5% (4)

Accuracy = (TP + TN) / (TP + TN + FP + FN) = 84% (5)

Precision = TP / (TP + FP) = 82.4% (6)

Specificity = TN / (TN + FP) = 82.3% (7)

Comparison with SVM

Once the NB model replaced with SVM model, efficiency was greatly reduced as the below presented confusion matrix in Fig 7.

	OK	TX	Total	Correct %
OK	5,913	3,676	9,589	61.6644
TX	4,733	4,947	9,680	51.1054
Total	10,646	8,623	19,269	
Correct %	55.5420	57.3698		

Fig. 7 Confusion Matrix with SVM model

Further, calculating classification model efficiency parameters as below highlights the drop of efficiency of prediction model due to introduction of SVM.

Recall = TP / (TP + FN) = 55.54% (8)

Accuracy = (TP + TN) / (TP + TN + FP + FN) = 57% (9)

Precision = TP / (TP + FP) = 61.2% (10)

Specificity = TN / (TN + FP) = 57.3% (11)

By comparing the results of NB model given in equations (4~7) with SVM model (8~11), the probabilistic approach of NB model has clear edge in the problem domain discussed in this paper.

C. Receiver operating characteristic curve

The receiver operating characteristic curve (ROC) summarizes the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds of each model. Following ROC clearly indicate that NB model also involve with lower trade of compared to SVM base module with respect to the problem discussed in this paper.

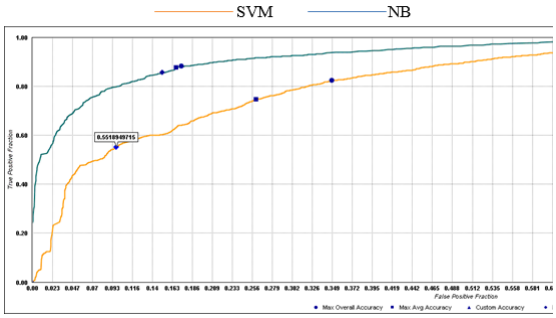


Fig. 6 ROC on Prediction [OK]

Conclusion and Future Work

In the telecommunication industry, early detection of the churn has become a crucial factor for survival. The recent development of the data science domain facilitates and provides various means to intercept the problem of the churn. However, there are no standard modules for the churn problem in any business domain as the nature of the problems and attributes vary between the environments where the business operates. It should be also noted that nowadays, many telecommunication service providers have identified that not only their revenue generation system is vital for survival, but other supportive systems are also important. Therefore, many companies tend to move towards implementing ERP systems to handle their operations. Considering all these concerns this research proposed an alternative method of predicting churn using financial information.

The study was carried out with a sample of 30, 000 subscribers, and their financial influence in the revenue generation process was analyzed to make predictions. Not like generic approaches based on consumer usage patterns that mostly govern by numerical features, this study was balanced with numerical and categorical features where there were four to numerical and six categorical features. It was the motive for selecting the probabilistic technique over logistic regression or SVM methods. The selection of the Naive Bayes method was justified at the end of the research as it yields around 85% accuracy in predicting churn.

This study also suggests that the predicting churn in telecommunication industry is not a static process, as many features related to churn is very sensitive to the change of market. Therefore, it is essential to reevaluate / fine tune models in fair frequency to keep up the accuracy of the prediction. As a solution, this paper suggests an application of minimum description length algorithm to evaluate the features of each sample before input to the model. The suggested approach is possible to apply for any dataset that has sensitivity to external variables. However, depending on the nature of the features, different algorithms may need to be employed to sustain model accuracy.

Upon completion of the research, the author concludes that similar intelligent models can be developed around the ERP system due to its integrated nature. However, it could be a challenging task to identify features to describe the problem as features could be shattered across the integrated modules. Though it could be a challenging task, building intelligence in decision making system (which is the ERP in this study) could be more effective for any industry, as it opens the way to measure decisions effectiveness and predict future impact. For example, predicting the churn in this study could be very useful for management to know the marketing waste, and compare marketing opportunity with finance status. The suggested approach is more likely building the intelligence in the same language which management system uses (such as finance data, inventory data compare to billing data like voice usage, data usage).

From the derived work so far in this research, there is a handful of future work that seems to be promising. For example, since the prediction has been done in the finance module of the ERP system, the same can be extended for revenue budgeting, forecasting, and evaluating the performance and return of marketing expenses.

REFERENCES

[1] Mbarek, R., Baeshen, Y.: Telecommunications Customer Churn and Loyalty Intention. Marketing and Management of Innovations. 110-117. 10.21272/mmi.2019.4-09 (2019).

[2] Zhao, Y., Li, B., Li, X., Liu, W., Ren, S.: Customer Churn Prediction Using Improved One-Class Support Vector Machine. 300-306. 10.1007/11527503_36 (2005).

[3] Sjarif, N.N.A., Mohd A., Nurulhuda., Sarkan, HM., Sam, S., Osman, M.: Predicting Churn: How Multilayer Perceptron Method Can Help with Customer Retention in Telecom Industry. IOP Conference Series: Materials Science and Engineering. 864. 012076. 10.1088/1757-899X/864/1/012076 (2020).

[4] Kadir, R. A., Yatin, S.: The benefits of implementing ERP system in telecommunications. Procedia-Social and Behavioral Sciences, 211, 1216-1222 (2015).

[5] Telecommunications Regulatory Commission of Sri Lanka, statistics 2019 quarter 4, http://www.trc.gov.lk/images/pdf/statis_q4_03032020.pdf, last accessed 2020/11/01.

[6] Sri Lanka Telecom PLC, Annual Report 2019. https://www.slt.lk/reports-html/annual/2019/index.html, last accessed 2020/11/01.

[7] Dahiya, K., Talwar, K.: Customer churn prediction in telecommunication industries using data mining techniques a review. Int. J. Adv. Res. Comput. Sci. Softw. Eng, 5(4), 417-433 (2015).

[8] Mozer, M., Wolniewicz, R., Grimes, D., Johnson, E., Kaushansky, H.: Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. Neural Networks, IEEE Transactions on. 11. 690 - 696. 10.1109/72.846740 (2000).

[9] Simsek G., Tugba.: Customer Churn Analysis in Telecommunication Sector. Istanbul University Journal of The School of Business Administration (2010).

[10] Saini, N., Monika., Garg, K.: Churn Prediction in Telecommunication Industry using Decision Tree. International Journal of Engineering Research and. V6. 10.17577/IJERTV6IS040379 (2017).

[11] Cui, G., Wong, M., Lui, H.: Machine Learning for Direct Marketing Response Models: Bayesian Networks with

Evolutionary Programming. Management Science. 52. 597-612. 10.1287/mnsc.1060.0514 (2006).

[12] Rouhani, S., Mehri, M.: Does ERP have benefits on the business intelligence readiness? An empirical study. 8. 81-105. 10.1504/IJISCM.2016.079559 (2016).

[13] Novakovic, J.: The Impact of Feature Selection on the Accuracy of Bayes Classifier (2010).

[14] Myung I.J.: Computational Approaches to Model Evaluation. International Encyclopedia of the Social & Behavioral Sciences (2001).