Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method

S Arifin¹, F Samopa¹,²

¹ Magister Management of Technology Department, InstitutTeknologi SepuluhNopember, Surabaya, Jawa Timur 60264, Indonesia.
² Research Center Building, InstitutTeknologiSepuluhNopember, Teknik Kimia Street, Surabaya, Jawa Timur 60117, Indonesia.

Abstract. This research was intended to know factors that influenced the churn rate significantly in telecommunication company through research of historical billing and profiling data of customers. This study consisted of seven variables which are four billing historical variables and three profiling variables. The data was taken from a telecommunication company in Indonesia by taking active customer with minimum 6 months and historical data form January until March 2018. The data were tested using Support Vector Machine method by taking the result of classification performance either on performance of total variable or performance in each variable. The significance threshold was defined 5%. The results show that there were three variables that influenced the churn rate significantly namely voice usage, data usage and reload with performance percentage less than 5% of total performance. Those three variables were historical billing data.

1. Introduction

In the telecommunications industry, customer is the most valuable asset for the company to sustain the business process. Reducing number of customer is certainly unexpected thing for the companies. Thus, companies need to analyze the customers profile to facilitate the conduct of business segmentation in support of a proper decision making. The right decision is expected to generate a value that can improve the performance and profit of the company. In analysis customer data, the data history is absolutely necessary. It can describe the customer’s value in the future whether the customer is potentially churn or non-churn. In this case, Support Vector Machine method can be used for classification [1].

Several methods of classification lately have been used for main purpose especially in the data processing. The popular method that used by researchers was data mining based; either using classification or clustering such as Decision Tree and K-Mean. However, most papers only discuss churn prediction without analyzing historical accounts of customer billing. In this study, in order to predict churn and also analyzed historical billing that included the use of voice in minutes, short message service in the event, and the internet data in kb. During the last 3 months in 2017, Support Vector Machine method was applied by taking voice, short message service (sms) and the data service as the main variable in the test. Based on the test results, it could be used as a reference in analyzing the factors that affect customer churn [2].
Table 1. Churn percentage in one of Telecommunication company in Indonesia 2017

| Region | Q1-Q2 | Q2-Q3 | Q3-Q4 | Q4-Q1 |
|--------|-------|-------|-------|-------|
| Central | 15%  | 7%    | 16%   | 17%   |
| East   | 39%  | -19%  | 1%    | 23%   |
| Capital | 2%   | 7%    | 9%    | 6%    |
| North  | 19%  | 11%   | 32%   | 30%   |
| West   | 3%   | 5%    | 22%   | 36%   |
| National | 15% | 0%    | 13%   | 20%   |

Table 1 illustrated that one of the internal sources of telecommunication companies from 1st quarter, 2nd quarter, 3rd quarter and 4th quarter explained significantly customers churn numbers which each region also quite large numbers. If it was viewed more detail as a whole data during 2017, there was only a region that experienced decreasing in churn rate, namely the East region in the second quarter. Nationally, the churn rate increase in each quarter grew by 20%. This was a corporate homework that requires evaluation and analysis of factors that can potentially be affected to churn rate.

Using the prediction of customer churn and customer behavior analysis, it were expected to help the marketing and CRM division in segmenting customers who could simplify the product campaign to minimize the existence of customer churn and increase the profit of the company [3]. The analysis conducted to produce variables which have great potential to make the customer into churn. Based on these variables, it can be used as a benchmark in determining a particular focus product in accordance with the profile of the customer.

Based on the background above, the problem in this study was unknown factors that affect the churn rate significantly. The existing services in telecommunication companies have not reflected their effectiveness in influencing the high churn rate. So, it takes effort in analyzing the billing behavior of consumers in determining the potential of churn rate [4]. From this case, the formulation of the problem in this study is "analysis of churn rate significantly factors in telecommunication industry using Support Vector Model ".

The purpose of this study was to determine what factors have a significant effect on the occurrence of churn rate using customer billing services and customer profiling within focus on the correlation between independent variables with dependent variable. In addition, this study aimed to classify and predict future churn rates based on patterns in the data history.

2. Support Vector Machine Principle and Algorithm

Support Vector Machine (SVM) is defined as a set of related learning methods that analyze data and recognize patterns, which are used for classification and regression analysis [5]. SVM takes a set of input data and predicts for each given input, derived from two classes that are classified by finding the best hyperplane value. Hyperplane is separator between the two classes that can be found by measuring the hyperplane's margins and searching for the maximum point[6].

For more detail about SVM, here given an example using dummy data test in the table 2.

Table 2. Data set

| Voices | Data Usage | y |
|--------|------------|---|
| 2.947814 | 6.626878 | 1 |
| 2.530388 | 7.785050 | 1 |
| 3.566991 | 5.651046 | 1 |
| 3.156983 | 5.467077 | 1 |
| 2.582346 | 4.457777 | -1 |
Table 2 describe that the dummy data consists of two attributes namely Voices and Data usage and there are two classes 1 as non-churn and -1 as churn. The next step is to determine the constraint for both attributes. Constraint is used for quadratic programming to looking for the value of w1, w2 and b value by using the formula:

\[ y_i (w^T x_i + b) \]  

(1)

The constraint results can be described in the table 3 below:

**Table 3. Finding the constraint.**

| W1. Voices  | W2. Data Usage | Constraint |
|-------------|----------------|------------|
| 8.560183    | 7.960593       | 1.785812   |
| 7.348017    | 9.351858       | 1.96491    |
| 10.35822    | 6.788367       | 2.41162    |
| 9.167592    | 6.567373       | 1          |
| 7.498897    | 5.354943       | 1.881125   |
| 6.260322    | 7.474642       | 1          |
| 9.50571     | 4.229255       | 1          |

Table 3 mention how the constraint obtained by multiple w1 with voice variable and w2 with data usage variable as in the formula the plus b. After the constraint have been obtained, then find the new value of w1, w2 and b value by using quadratic programming that subject to constraint value. The quadratic programming can be done by using the below formula:

\[
\min_{w,b} \frac{1}{2} w^T W
\]

subject to

\[ y_i (w^T x_i + b) \geq 1 \text{ for } \forall i \]

After compute the quadratic programming formula, the value of w1, w2 and b can be found in the table 4 below.

**Table 4. new w1, w2 and b value**

| w1          | w2          | b            |
|-------------|-------------|--------------|
| 2.903909    | 1.201258    | -14.735      |

Table 4, explain the result of quadratic programming value to looking for the new value of w1, w2 and b. Those value are used to determine the margin score. The margin score of each data are used to find the classification classes. Margin score indicates the not-normalised distance between each data point to the decision hyperplane. Margin scoring data can be found by the formula below:

\[ \Delta^i = W^T X_i + b \]  

(3)
Where b is the value of the new b after quadratic programming. The margin score value can be found in the table 5 below.

**Table 5. Margin score**

| W1. Voice | W2. Data Usage | Margin Score |
|-----------|----------------|--------------|
| 8.560183  | 7.960593       | 1.7858       |
| 7.348017  | 9.351858       | 1.9649       |
| 10.35822  | 6.788367       | 2.4116       |
| 9.167592  | 6.567373       | 1            |
| 7.498897  | 5.354943       | -1.8811      |
| 6.260322  | 7.474642       | -1           |
| 9.50571   | 4.229255       | -1           |

Table 5 explain how margin score can be found. The margin score are used to determine classification classes by simply take the sign of margin score in each data using function below.

\[
sign(x) = \begin{cases} 
  +1 & \text{if } x \geq 0 \\
  -1 & \text{if } x < 0 
\end{cases}
\]  

(4)

Sign function will return 1 if the input argument is positive or zero and return -1 if the input argument is negative. Finally classification result as shown in table 6 below.

**Table 6. Classification result.**

| Score     | Classification |
|-----------|----------------|
| 1.785812  | 1              |
| 1.96491   | 1              |
| 2.41162   | 1              |
| -1        | 1              |
| -1.88113  | -1             |
| -1        | -1             |
| -1        | -1             |

In the table 6 show that classification data are same with original data set. There are four data for 1 class and three data for -1 class. It can be concluded that SVM 100% classified correctly.

### 3. Research Methodology

In this research, the data were tested amounted to 819 MSISDN with 7 variables. The data taken from a telecommunication company in Indonesia with MSISDN active at least 6 months and there were usage for 3 months from January until March 2018. In this test, it was conducted and supervised a learning technique, that in each variable there are churn and non churn data that become data reference learning.

**Table 7. Data set and the index.**

| Variables | Index |
|-----------|-------|
| POC       | Consist of 22 POC Area |
| Handset   | 2.5G, 2.75G, 2G, 3.5G, 3.75G, 3G, 4G |
Voice in Minutes 0>1000, 1001-2000, 2001-3000, 3001-4000, >4000 in Minutes
SMS in Event 0>1000, 1001-2000, 2001-3000, 3001-4000, >4000 in Event
Data in Kb <=20GB, >20GB - <=40GB, >40GB - <=60GB, >60GB
Package Name More than 12 package
Reload 19 Reload categories from 5000 to 200000

Table 7 described that there were seventh variables, namely POC (point of control), Handset in devices, Voice usage in Minutes, Short Message Services in event, data usage in kb, packet name that attached in the MSISDN and reload which each variable has an index. The index means the categorical data in every single rows. However, the data obtained was not in single data but in the form of categorical. After variables data was defined, and then it was continued to the research flow. The research flow explains how data is processed and produce a result for getting knowledge analysis as shown in Figure 1 and Figure 2.

![Figure 1. Data Processing.](image1)

![Figure 2. SVM Model.](image2)

In Figure 1, it was explained that from each attribute in the tested data, consist of training data and testing data. Both data would be tested with SVM modelling in Figure 2 for getting performance that would be used for knowledge analysis.

In figure 2, the threshold was determined at 5% [7] of the total performance, if the attribute performance tested was smaller than the threshold, then the attribute had a significant effect, otherwise if greater than 5%, then the attribute did not significantly influence.

4. Finding and Discussion
In this section, it would be explained how to analysis the data and find the churn rate significantly variables based on hypothesis that had been decided. Firstly, the zero hypothesis was determined as shown in figure 3.
In Figure 3, there were seven hypotheses as zero hypothesis[8] with POC as hypothesis 1, Handset as hypothesis 2, Voice in Minutes as hypothesis 3, SMS in Event as hypothesis 4, Data in Kb as hypothesis 5, Packet Name as hypothesis 6 and Reload as hypothesis 7 with respectively H0 as the comparative hypothesis. The initial hypothesis was all variables which not significantly affected to churn rate.

To prove the hypothesis, then it was conducted a testing influence of each variable on the overall result of the variable. It would determine the significant value based on the predetermined threshold that was 0.05 or 5% [9].

**Table 7. Research Result.**

| Hypothesis | Performance Classification | Distance | Sig (5%) | ROC Area | Mean absolute error |
|------------|---------------------------|----------|----------|----------|---------------------|
| H1         | 60.20%                    | 22.80%   | FALSE    | 0.40     | 0.53                |
| H2         | 58.90%                    | 24.00%   | FALSE    | 0.41     | 0.51                |
| H3         | 79.30%                    | 3.70%    | TRUE     | 0.78     | 0.20                |
| H4         | 74.00%                    | 8.90%    | FALSE    | 0.71     | 0.26                |
| H5         | 82.50%                    | 0.40%    | TRUE     | 0.82     | 0.17                |
| H6         | 64.60%                    | 18.30%   | FALSE    | 0.61     | 0.35                |
| H7         | 80.60%                    | 2.30%    | TRUE     | 0.75     | 0.19                |

Based on table 7, it could be concluded that there were three attributes that significantly influence the churn rate of total 7 attributes which have been tested. The usage data in kb (H5), voice in minutes (H3) and reload (H7) had a performance smaller value than 5% [10] of the total performance of the overall variables. Meanwhile, the variable was approaching the significant in SMS (H4) that was equal to 9%. Based on these results, the telecommunications companies XYZ should maintain Data and Voice services to minimise the churn rate.
After the test, it was obtained a new research model that included three variables as the result namely of voice in minutes, data in Kb and reload. Those variables showed significant effect to the churn rate.

5. Conclusion
Based on the result, it could be concluded that variable voice in minutes, data in Kb and reload significantly influenced the occurrence of churn rate. The telecommunications company especially in Indonesia needed to maintain three main services, namely voice, data and reload service.

6. Reference
[1] Guo-en XIA 2008 Model of Customer Churn Prediction on Support Vector Machine SETP28(1) pp. 71-77.
[2] Shaaban E 2012 A Proposed Churn Prediction Model MUST University 24
[3] Brandusoiu I 2013 Churn Prediction In The Telecommunications Sector Using Support Vector Machines Fascicle of Management and Technological Engineering 13
[4] Yudhistira A 2017 (online) Analisa Customer Churn pada perusahaan Internet Service Provider XYZ menggunakan Backpropagation Neural Network (http://repository.its.ac.id/2564/)
[5] Srivastava D 2009 Data Classification using Support Vector Machine Journal of Theoretical and Applied Information Technology JATIT (2) pp. 1-7.
[6] Yin Y, Han D, and Cai Z 2011 Explore Data Classification Algorithm Based on SVM and PSO for Education Decision Journal of Convergence Information Technology 6(10) 122–128
[7] Herawati M 2016 Prediksi Customer Churn Menggunakan Algoritma Fuzzy Iterative Dichotomiser 3 J. Math. and Its Appl. 13(1)
[8] Tanudjaja H 2007 Analisis Hubungan Dan Pengaruh Variabel Makroekonomi Terhadap Kredit Bermasalah Perbankan Indonesia (Jakarta: Universitas Indonesia)
[9] Ardiansya, IH, Lubis D 2017 Pengaruh Variabel Makroekonomi terhadap Pertumbuhan Sukuk Korporasi di Indonesia IAEI 5(1)
[10] Fraticasari S 2018 Optimasi Pemodelan Regresi Linier Berganda Pada Prediksi Jumlah Kecelakaan Sepeda Motor Dengan Algoritme Genetika Journal Pengembangan Teknologi Informasi dan Ilmu Komputer 2(5)
[11] Teknomo K 2012 Support Vector Machine Tutorial (Columbia: Revoledu.com publishing)

Acknowledgments
This research has been acknowledged by advisors of Magister Management of Technology Department in Institut Teknologi Sepuluh Nopember (ITS), Surabaya and Revoledu.com by Teknomo Kardi [11].