Customer Churn Prediction for a Software-as-a-Service Inventory Management Software Company : A Case Study in Thailand

Phongsatorn Amornvetchayakul
Department of Industrial Engineering
Chulalongkorn University
Bangkok, Thailand
e-mail: phongsatorn.amornvetchayakul@sasin.edu

Naragain Phumchusri
Department of Industrial Engineering
Chulalongkorn University
Bangkok, Thailand
e-mail: Naragain.P@chula.ac.th

***Abstract*—Software-as-a-Service is the fast growth and high market values as a new emerging online business. Customer churn is a critical measure for this business. Thus, this paper focuses on seeking a customer churn prediction model for a Software-as-a-Service inventory management software company in Thailand which is facing a high churn rate. This paper executes the prediction models with four machine learning algorithms: logistic regression, support vector machine, decision tree and random forest. The random forest model is capable to provide lowest error with 10-fold cross validation average scores of 91.6% recall and 92.6% F1-score. Moreover, feature importance scores can highlight useful insights of case-study that business metrics are significantly related to churn behavior. As a result, this paper is beneficial to the case-study company to help indicate real churn customer and enhance the effectiveness in executive decision and marketing campaign.**

Keywords-component; churn prediction; Machine learning; Software-as-a-Service (SaaS); Random forest

I. INTRODUCTION

Software-as-a-Service (SaaS) is a model where software is licensed over the internet on a subscription basis by a third-party provider. Due to its flexible and scalable nature, the use of cloud computing is exponentially rising, especially in this digital transformation landscape. According to Gartner [1], SaaS is the largest market segment and is expected to maintain this top position. The SaaS sector is forecasted to reach $150 billion in 2022 in terms of public cloud service revenue. SaaS's leading benefit is mainly revenue recurrence meaning that revenue opportunities switch from traditional one-sale transactions to subscription based. Hence, for SaaS businesses, higher customer retention or lower customer churn are crucial towards profitability.

Inventory management software is used for tracking inventory levels, purchase orders, sales, delivery, and much more. With SaaS model, inventory management software assist SMEs enhancing the business competitiveness in terms of scalability, versatility, and advanced security in affordable prices [2]. In Asia, the emerging economies brings a concern in productivity improvement by using the technology which is a part of the growing factor of this market while South East Asia including Thailand becomes a new target market of the inventory management software providers.

Regarding these growth factors, a case-study, SaaS inventory management software company in Thailand, is expected to capture the trends and scale up their market. However, acquiring a new customer is significantly more expensive than retaining an existing one [3]. Increasing customer retention rates by 5% can increases profits by 25% to 95% since repeat customers are more likely to retain relationship with a company in long term and potentially provide a massive marketing advertisement by referring to the company as a part of loyalty [4]. They further tend to spend more in each order after purchasing with the same company [5]. Consequently, customer churn rate, the percent of customers who have cancelled their subscription or use of a product or service, should be focused for SaaS businesses.

The case-study company is currently having over 50% churn rate. this issue leads to the objective of this paper that is to search for suitable customer churn prediction model for the case-study in Thailand since a proper prediction of customer churn as the ability to identify customers who are likely to churn is important to reduce churn rate. Moreover, it also enhances the efficiency and effectiveness of managerial decision for the case-study company to handle these risky customers, especially in marketing, by providing an insight or features influencing churn customer. In order to create a good-performance prediction model, this paper gathers a huge amount of information and collects the data from the case-study company. Then, statistical feature selections both Chi-squared and ANOVA are applied. Several machine learning algorithms, i.e., logistic regression, support vector machine, decision tree and random forest are explored. Eventually, the results are evaluated by Recall, F1-score, Accuracy, and Precision.

II. LITERATURE REVIEW

Over the past few decades, many papers have tried to research for a customer churn prediction in terms of industries, churn definitions and machine learning techniques. Regarding the type of business, customer churn prediction has been studied in various industries. Most of researches commonly focus on Telecommunication industry [6, 7, 8, 9]. As a major infrastructure, customer churn prediction in Telecommunication has been developed along with changing in technology, i.e., landline, mobile and wireless, respectively. Another famous industry in which studies concentrated is Banking [10, 11] in which churn rate is directly caused by various industry factors, e.g., customer

978-1-7281-6785-5/20/$31.00 ©2020 IEEE

portfolio's size or frequency of transactions. Retailing industry had also been investigated by many papers [12, 13, 14, 15, 16]. Those studies were exposed in the different sectors considering the customer churn prediction in the different angles. [12] focused on retailing in fast-moving consumer goods (FMCG). Recent papers tend to concentrate in online retailing according to market trends [13, 15, 16]. For SaaS industry, there are only few studies compared to the market size of this industry due to limited resources and data access authorization [17, 18, 19]. Most of studies in this SaaS industry were based on open data sources that could not catch up the changing in customer behavior in the present time.

Customer churn definition is defined differently for different industry. It is not only mentioned in the terms of customer terminating [9, 16, 19], switching to competitors [6, 8, 9, 13] or deactivating any activities [14, 15, 18] but it is also defined in the other aspects of interest, e.g., customer's portfolio size being lower than a specific threshold value [11], customer changing purchasing behavior [12], decrease in customer's lifetime value (CLV) [10].

In Thailand, the study related to customer churn prediction is concentrated on Telecommunication industry. [9] studied the customer churn analysis for a case study on the telecommunication industry in Thailand and found that C4.5 Decision Tree algorithm performed well with the most accurate result among other techniques; Logistic Regression and Neural Network. Additionally, in term of feature category, most of selected features are related to usage matric.

In early stage of emergence of SaaS business model, [17] applied churn analysis in telecommunications as a baseline with K-means and Decision Tree model to Software-as-a-Service (SaaS). Despite of the fact that this paper provided the benefit of features comparison between Software-as-a-Service (SaaS) and Telecommunication, there was still a gap in term of the number or variety of attributes used in analysis. Then [18] applied Logistic Regression, Random Forest and also ensemble learner of XGBoost to the model under the churn definition; customer suspended payment for at least 2 months. Recently, [19] studied the customer churn prediction in SaaS company with various machine learning algorithms, the result is nevertheless shown that Support Vector Machine is the most precise model in this case while precision is values as the most importance metric among others: accuracy, recall, and F1-score. The case study company of this paper was a digital marketing campaign management company offering marketing solution and its quality of service highly depended on other online platform policy.

The existing researches mostly concentrated on major industries, i.e., telecommunication, retailing and banking, while only few papers focused on SaaS industry. Thus, this paper studies on prediction of customer churn for a SaaS inventory management software company based on case-study in Thailand. Moreover, this paper defines churn according to the case-study company marketers' requirement as a customer who have been inactive consecutively for more than 14 days which is different from previous papers.

III. METHODOLOGY

A. Data Collection

This paper is encouraged by the case-study company's intention in improvement the customer churn issues which directly affect the company balance sheet. Due to experience and competency, the insights of business provided by the case-study company marketers are derived into 23 features variables of raw data in terms of customer usage behavior and business metric such as transactions. This paper extracted data from the case-study company database based on the data collecting from October 2015 to October 2019. The whole raw data consist of 1788 observations. It is noticeable that this paper defines churn as a customer who have not been active consecutively for more than 14 days according to the case-study company marketers' requirement.

B. Data Transformantion

1) Data cleaning

Regarding the raw data, missing data, meaning that there is no data value provided for the observed variable in a sample, are under consideration. This paper applied the listwise deletion technique to handle this problem [20]. This method is also applied in several customer churn prediction papers [9, 21]. In this case, Listwise deletion reduce the missing data from the 1788 observations of the whole raw data to 1718 observations. The data also includes both churn and non-churn customer records; 927 churn samples and 791 non-churn samples. This original dataset seems to be likely balanced distribution with 53.96% of churn samples and 46.04% of non-churn samples.

2) Data processing

The collected raw dataset is typically provided in the two different types of variables, i.e., quantitative variables and categorical variables. Due to the nature of categorical data provided by the case-study company, this paper uses one hot encoding method to convert the categorial data. Furthermore, standardization of dataset is applied in this case. Regarding the benefits in implementation, function scale in standardization neglects the shape of distribution by scaling to the feature's standard deviation after removing the mean of a certain feature.

3) Feature selection

This paper focuses on two types of univariate feature selection or statistical filter method for classification, i.e., Chi-squared score and ANOVA F-values. Both of them statistically measure each features' importance separately and rank these features variables in an order of importance. Then, the top ranks of highest-score features are selected to be used in any machine learning in the next step while the number of selected features is a parameter needed to be assigned. As a result, this paper executes all possible numbers of selected features.

C. Machine Learning

The appropriate techniques for this paper are deliberately selected regarding the functionality, reliability and performance of the algorithms performed successfully in

classification, especially from the papers related to SaaS business. Algorithms are functioned using Scikit-learn focusing on machine learning in Python [22].

1) Logistic regression

Logistic regression is one of the well-known classification methods in machine learning. The logistic regression model represents the relation between independent variables and dependent variable in probabilities while dependent variable has two possible outcomes, i.e., churn and non-churn. This method performed effectively in customer churn problem for SaaS business [18].

2) Support vector machine

Support vector machine is one of the classifiers that distinguish the class of the output with hyperplane. This technique can handle the linear and non-linear classification problems. Many studies were successful in solving classification problems. [13], focused on retailing industry, is one of those papers that getting a favorable outcome for Support vector machine as well as [7] who studied in telecommunication industry. Moreover, support vector machine had been found to be effective in the results of algorithm comparison for churn prediction in SaaS industry.

3) Decision tree

Decision tree is one of the most common classification methods and widely used in various fields of supervised machine learning problems. This technique can be easily with breakdown in smaller portions [14]. Decision tree was performed accurately for churn prediction problem studied by [6] while, in Thailand, [9] presented a good result for churn prediction in telecommunication industry. [17] also applied decision tree classification to predicting customer churn in SaaS industry.

4) Random forest

Random forest is generated by combining with a numerous decision trees in process of machine learning as one of ensemble learning classifier techniques. This technique is also stable in terms of diverse data and better reduction in risk of overfitting effect. [19], who studied in churn prediction in SaaS, found that this random forest technique provides a great performance, similar to results from support vector machine.

D. Cross Validation

This paper applies two general methods including with holdout method and k-fold cross validation. The dataset is partitioned in to 80% of training data and 20% of testing data under holdout method. This holdout testing set is used to evaluate the final model while the training set is proceeded k-fold cross validation. K-fold cross validation, k value is popularly assigned to be equaled to 10. [23] mentioned that there is not any specific rule for valuing k while a value of k is commonly equal to 10. It is noticeable that this value of k = 10 is valid to high variance data. As a result, in this paper, k is defined as equal to 10, meaning that the cross-validation dataset is randomly split into 10 portions, processing 10 iterations with different validation datasets. Overall, this paper tries to negate the effect of bias and overfitting by applying k-fold cross validation. All results of evaluation metrics from k-fold cross validation are used to compare the performance of prediction models. Then, the top performance as the final model using handout testing dataset which imitated the real data is evaluated to confirm the capability of the model performance on unknown dataset.

E. Evaluation Metrics

In order to evaluate and compare the performance or accuracy of prediction models, confusion matrix is a suitable tool to describe the outcome between predication and actuals in a binary classification. the confusion matrix can be used to calculate in 4 different dimensions that are Accuracy, Precision, Recall and F1-score.

Accuracy is the ratio of total number of correct predictions to all observations which generally indicate the performance of predication in terms of frequency. Its calculation is shown in (1):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision is the ratio of correctly predicted churns to all predicted churns. it is normally used in the problem focusing on false positive results or incorrectly churn prediction. Its calculation is shown in (2):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall is the ratio of correctly predicted churns to all churns meaning that the performance in prediction of how much real churns are correctly indicated. Its calculation is shown in (3):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1-score is the weighted average of precision and recall which also indicates the overall prediction accuracy. Its calculation is shown in (4):

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Regarding the case-study business objectives, the various measures of confusion matrix are considered in the order of importance. Due to the definition of churn, the paper aims to mitigate the cost of a new customer acquisition which is significantly high compared to the cost of a customer retentions caused by churn customers. It can be concluded that it is necessary to correctly identify a churned customer.

This paper considers Recall as an essential metric because, in customer churn detection, if a real churn customer is predicted as a non-churn customer, the consequence will severely affect the balance sheet of the company as the cost of false negative. Precision which values false positive as the most significant is considered as the last order of importance among other measures. F1-score and Accuracy are likely the same in term of overall performance indicator. The disadvantage of Accuracy is the dependence on data classification balance while F1-score can handle the issue of imbalanced data. Even though accuracy can be affected by imbalanced data, this case-study dataset,

combining with 53.96% of churn samples and 46.04% of non-churn samples, can be implied that the data are balanced. Therefore, Accuracy is still applicable. However, this research considers F1-score as another important metric which also predominates over Accuracy on account of hereinabove mentioned.

IV. Experimental Results

This research applies four machine learning algorithms on the data granted by the case-study company. After data cleaning, 1718 observations of 23 explanatory variables are separated into 2 portions: 80% is training dataset and 20% is testing dataset. Then, the training dataset is optimized by two types of univariate feature selection or statistical filter method for classification, i.e., Chi-squared score and ANOVA F-values. After that, machine learning models including logistic regression, support vector machines, decision tree, and random forest are fit with selected data. 10-fold cross validation are applied to evaluate the models.

Comparison of evaluation metrics for k-fold cross validation from various models is illustrated in Fig 1. Based on recall and F1-score, random forest algorithm performs better than other algorithms.

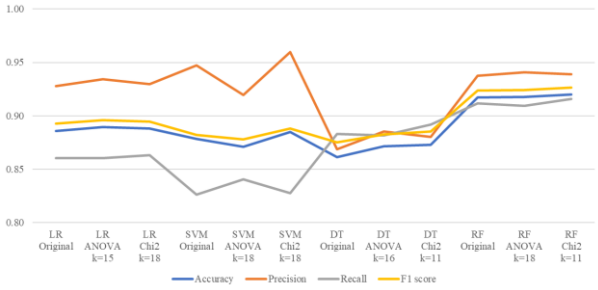

Figure 1. Results of Various Models with 10-fold Cross Validation

Table 1 also shows the result of each classifiers which is optimized models. Random forest seems to outperform other models with score of 91.6% recall and 92.6% F1-score. Even though support vector machine can provide a good result in precision with 96.0 percent, other measures are severely poor to perform.

TABLE I. Results of Classifiers

Models	Evaluation matric			
	Accuracy	*Precision*	*Recall*	*F1 score*
LR	0.889	0.934	0.860	0.896
SVM	0.885	0.960	0.827	0.888
DT	0.873	0.880	0.892	0.885
RF	0.920	0.939	0.916	0.926

As a result of classifiers comparison in previous step, the model that performed outstandingly apart from other algorithms is random forest model fitting the 11 highest scores features selected by chi-squared filter method. In order to confirm the performance of the final model, this model is investigated holdout cross validation with firstly separated testing dataset. The results are shown in Table 2. The final model is capable to perform similarly to previously k-fold cross validation dataset with 94.4% of recall and 94.6% of F1-score.

TABLE II. Results of the Final Model

RF with chi-squared k = 11	Evaluation matric			
	Accuracy	*Precision*	*Recall*	*F1 score*
k-fold	0.920	0.939	0.916	0.926
holdout	0.939	0.948	0.944	0.946

In addition to evaluate the classification model, the final model can also describe the related features. Table 3 shows the summary of the top 5 most importance features sorted by feature importance scores which derived from the final classification model. Overall, the important features impacting to the prediction model are mainly related to business features like transaction and usage frequency.

TABLE III. Top Five Important Features in The Final Model

Rank	Feature name	Description	Feature importance
1	currMonthTrans	Number of transactions in current month	0.273
2	prevMonthTrans	Number of transactions in previous month	0.207
3	amountSpend	Amount of customer spending	0.098
4	numAct	Number of actions per customer	0.058
5	UserAct	Average number of actions per user	0.055

V. Conclusion

In SaaS industry, the market is continuously expanding due to the global economic growth as well as inventory management software market. This case-study, SaaS inventory management software company in Thailand, is also expected to capture the trends and scale up their market. However, the company is facing a high customer churn rate problem. Developing the solution models which improves the customer churn rate becomes an important issue.

This paper found the most effective model is random forest model using the 11 highest scores features selected by chi-squared filter method. It shows the capability in prediction with 10-fold cross validation average scores of 91.6% recall and 92.6% F1-score. This final model is enough to satisfy the case-study objective indicating real churn customer correctly and perform better than others prediction models. Moreover, the contribution of feature importance scores provides beneficial insights to the case-study company. It highlights that business metrics such as transactions are the top feature importance. It can also imply if customer is online more frequent and books more amount of transactions via the platform, that customer is less likely to churn. As a result of the final model and feature importance, the case-study company will be able to indicate the right risky churn customers and handle them with effective marketing campaigns. Therefore, this paper can

enhance the efficiency and effectiveness of managerial decision for case-study company.

VI. Limitation and Future work

Due to limitation of case-study company's information management system, most of available features had been exposed in business-related feature usage area, e.g., transactions and number of various types of usages while these data still lack of quality metrics like service rating or customer satisfaction in many dimensions and also internal information like outage incidents, software errors. These all kind of data can enhance the prediction model to understand further dimensions of case-study related factors.

In addition to cross validation, although this paper applies both holdout method and k-fold cross validation to confirm the performance of prediction models, the final prediction model should also be tested with the updated actual data from the case-study company in order to satisfy practically the target of the company in practice.

Moreover, after this paper provides the customer churn prediction model satisfied the target in churn customer identification, it can be extended to forecast the number of customers, the future transaction and also the revenue of the company [24]. As a result of that extensions, the company's executives can practically apply this more concrete information like financial data in a managerial decision.

References

[1] "Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17% in 2020," Gartner Inc., 13 November 2019. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020. [Accessed 14 February 2020].

[2] MarketWatch, "Global Inventory Management Software Market to Hit USD 3 billion by 2024," MarketWatch, Inc., 26 November 2019. [Online]. Available: https://www.marketwatch.com/press-release/global-inventory-management-software-market-to-hit-usd-3-billion-by-2024-2019-11-26. [Accessed 14 February 2020].

[3] A. Gallo, "The Value of Keeping the Right Customers," harvard business school publishing corporation, 29 October 2014. [Online]. Available: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers. [Accessed 14 February 2020].

[4] F. Reichheld, "BAIN & COMPANY, INC.," September 2001. [Online]. Available: http://www2.bain.com/Images/BB_Prescription_cutting_costs.pdf. [Accessed 14 February 2020].

[5] Bain & Company, Inc., "the value of online customer loyalty and how you can capture it," 1 April 2000. [Online]. Available: https://www.bain.com/insights/the-value-of-online-customer-loyalty-and-how-you-can-capture-it/. [Accessed 14 February 2020].

[6] S. Hung and H. Wang, "Applying data mining to telecom churn management," in Pacific Asia Conference on Information Systems (PACIS), 2004.

[7] X. Guo-en and J. Wei-dong, "Model of customer churn prediction on support vector machine," Systems Engineering–Theory and Practice, vol. 28, no. 1, pp. 71-77, 2008.

[8] V. Umayaparvathi and K. Iyakutti, "Applications of data mining techniques in telecom churn prediction," International Journal of Computer Applications, vol. 42, no. 20, pp. 5-9, 2012.

[9] P. Wanchai, "Customer churn analysis : A case study on the telecommunication industry of Thailand," in 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), Cambridge, 2017.

[10] N. Glady, B. Baesens and Croux, "Modeling churn using customer lifetime value," European Journal of Operational Research, vol. 197, no. 1, pp. 402-411, 2009.

[11] O. Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: the case of private banking," Expert Systems with Applications, vol. 41, no. 17, pp. 7889-7903, 2014.

[12] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," European Journal of Operational, vol. 164, no. 1, pp. 252-268, 2005.

[13] X. Yu, S. Guo, J. Guo and X. Huang, "An extended support vector machine forecasting framework for customer churn in e-commerce," Expert Systems with Applications, vol. 38, no. 3, pp. 1425-1430, 2011.

[14] A. Tamaddoni Jahromi, S. Stakhovych and M. Ewing, "Managing B2B customer churn, retention and profitability," Industrial Marketing Management, vol. 43, no. 7, pp. 1258-1268, 2014.

[15] R. Vadakattu, B. Panda, S. Narayan and H. Godhia, "Enterprise subscription churn prediction," in 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, California, 2015.

[16] P. Berger and M. Kompan, "User Modeling for Churn Prediction in E-Commerce," IEEE Intelligent Systems, vol. 34, no. 2, pp. 44-52, 2019.

[17] B. Frank and J. Pittges, "Analyzing Customer Churn in the Software as a Service (SaaS) Industry," Radford university, Virginia, 2009.

[18] Y. Ge, S. He, J. Xiong and D. E. Brown, "Customer churn analysis for a software-as-a-service company," in 2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2017.

[19] A. Rautio, "Churn Prediction in SaaS using Machine Learning," Tampere University, Tampere, Finland, 2019.

[20] P. Allison, "Missing data, Quantitative applications in the social sciences," SAGE Publications, Inc., Thousand Oaks, California, 2002.

[21] S. Khodabandehlou and M. Zivari Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," Journal of Systems and Information Technology, vol. 19, no. 1/2, pp. 65-93, 2017.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher and M. Perrot, "Scikit-learn: Machine Learning in Python," The Journal of Machine Learning Research (JMLR), vol. 12, pp. 2825-2830, 2011.

[23] M. Kuhn and K. Johnson, "Over-Fitting and Model Tuning," in Applied Predictive Modeling, New York, Springer, 2013, pp. 61-92.

[24] A. Sukow and R. Grant, "Forecasting and the Role of Churn in Software-as-a-Service Business Models," iBusiness, vol. 05, no. 01, pp. 49-57, 2013.