Building Customer Churn Prediction Models in Fitness Industry with Machine Learning Methods

Min Shan

Min Shan
VT 2017
Examensarbete, 15 hp
Supervisor: Juan Carlos Nieves
Examiner: Lars-Erik Janlert
Kandidatprogrammet i datavetenskap, 180 hp

Abstract

With the rapid growth of digital systems, churn management has become a major focus within customer relationship management in many industries. Ample research has been conducted for churn prediction in different industries with various machine learning methods. This thesis aims to combine feature selection and supervised machine learning methods for defining models of churn prediction and apply them on fitness industry. Forward selection is chosen as feature selection methods. Support Vector Machine, Boosted Decision Tree and Artificial Neural Network are used and compared as learning algorithms. The experiment shows the model trained by Boosted Decision Tree delivers the best result in this project. Moreover, the discussion about the findings in the project are presented.

Acknowledgements

I would first like to thank my supervisor Juan Carlos Nieves of computer science department at Umeå University. His guidance helped me in all the time of research and writing of this thesis. I would also like to thank Peter Johansson and Oskar Lindberg from Xlent Norr for their knowledge and helps at conducting the experiment.

Contents

1 Introduction

1.1 Problem Statement

Nowadays, customer relationship management (CRM) has become more and more important for companies to run successful business; because CRM targets the development of profitable, long-term relationships with key customers and stakeholders [28]. The better the relationship, the easier it is to conduct business and generate revenue. Thus, a lot of companies do realize that mining their existing database and associated information technologies provide enhanced opportunities to understand customers and maintain a good customer relationship [6]. Therefore, developing technology to improve CRM makes good business sense.

Customer churn management, as a part of CRM, has received increasing attention over the past time. Customer churn prediction aims at detecting customers with a high propensity to cut ties with a service or a company [38]. An accurate prediction allows a company to take actions to the targeting customers who are most likely to churn, which can improve the efficient use of the limited resources and result in significant impact on businesses [37]. Churn management strategies consist of two steps:

1. Ranking customers based on the estimated likelihood that they will churn.
2. Offering incentives to a core group of customers at the top of the churn ranking.

This thesis will focus on the former. Machine learning models will be build to predict the churn possibility of customers.

1.2 Related Works

Ample research has been conducted to predicting customer churn in different industries, including telecom industry [25][18], credit card providers [1], banking [42], wireless industry [44], etc. different data mining technologies have been used for building models, such as support vector machines [6], AdaBoost [25], neural networks [32], and ensemble of hybrid methods [36][17]. Among other, the results of previous works have showed great potentials of applying machine learning methods to customer churn prediction problems. For instance, the authors of [25] showed that AdaBoost algorithm successfully provides an opportunity to define a high risk customer group in telecom industry. In [18], decision trees and neural network methods were used for modeling. The result shows that data mining techniques can effectively assist telecom service providers to improve the Accuracy of churn prediction. In [6], Support Vector Machines were used in order to construct a churn model with a higher predictive performance in a newspaper subscription context. The result shows SVM

outperforms a logistic regression only when the appropriate parameter selection technique (feature selection) is applied.

Based on [25][19], the standard procedure of churn prediction can be summarized into three main steps. Firstly, doing feature selection to select relevant attribute to prediction. Secondly, building a machine learning model. The attributes selected in previous step are used as input of the model. Thirdly, evaluating models.

1.3 Research Question

As mentioned above, customer churn prediction is popular in a number of industries. To the best of our knowledge, there is no research about modeling customer churn prediction in fitness industry. There are similarities between fitness industry and other industries concerning customer relations. However, there may exist some unique factors which can affect customer relations in fitness industry. Among all customers, the group of customers who subscribe the service monthly and pay automatically from their bank accounts (monthly-paying customers) are especially important since they contribute to a big proportion of companies' income. However, this group has high churn rate according to a Swedish fitness gym. Against this background, this thesis aims to explore several questions regarding customer churn prediction in fitness industry based on statistic and machine learning methods. The purpose of this thesis is finding the feature selection methods and modeling methods which can contribute to customer churn predictions in fitness industry.

The research question of this thesis is :

Concerning Boosted Decision Trees, Artificial Neural Networks and Supported Vector Machines as supervised machine learning algorithms in this thesis project, which combination of feature selection method and supervised machine learning algorithm contributes to a high-quality and effective classifier of customer churn prediction for monthly-paying customers in fitness industry?

1.4 Methodology

The research of this thesis will be conducted in four steps: literature study, data collection, feature selection and modeling. The workflow process of the research of this thesis is described by Figure 1.

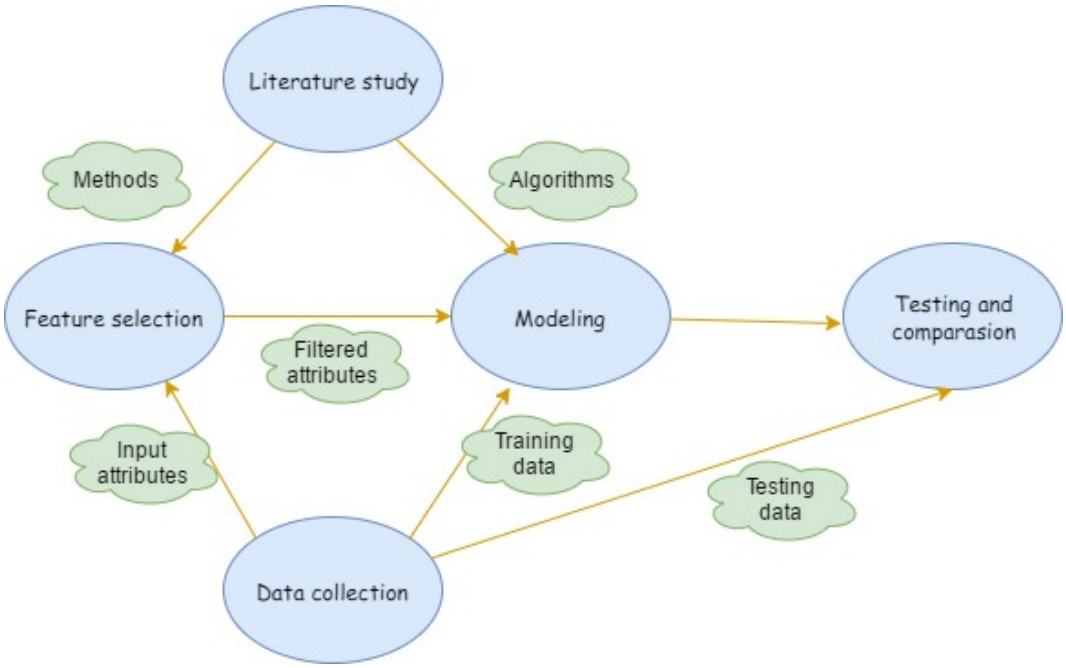


Figure 1: Process of experiment

1.4.1 Literature Study

There are plenty of feature selection methods and modeling methods which have been used in similar application domains (see Section 1.2). The purpose of literature study is to find the candidate methods which fit the properties of the data from different perspectives. For example, for the large amount of data, the methods which require less computing space and time are more suitable.

1.4.2 Data Collection and Cleansing

A data set from a fitness company which includes all the customers' data will be provided during the execution of this thesis. Since imbalanced data have been seen as a barrier to the performance of some standard classifiers [22][20], therefore, creating balanced dataset is important. 3000 customers' data will be selected randomly as the samples of which half will be customers who have churned. Since the churn prediction model may be biased due to a particular observed point of time, the studies will focus on the annual churn records of the year 2016.

As the authors of [3] claim: the effect on churn is greatly decreasing after six months despite statistically significant trends. Therefore five continuous months under 2016 are selected randomly in this research. The output of the model is whether the customer churns voluntarily in the latter two-month period. The input is first three-month relevant information, which is the attributes left after feature selection. The started month of observed period for each sample is randomly selected. For example, if started month is February, the observed period for input is from February to April, and output period is May and June. Meanwhile, for another sample, the started month of observed period can be August. An example of data selection for one sample is given in Figure 2.

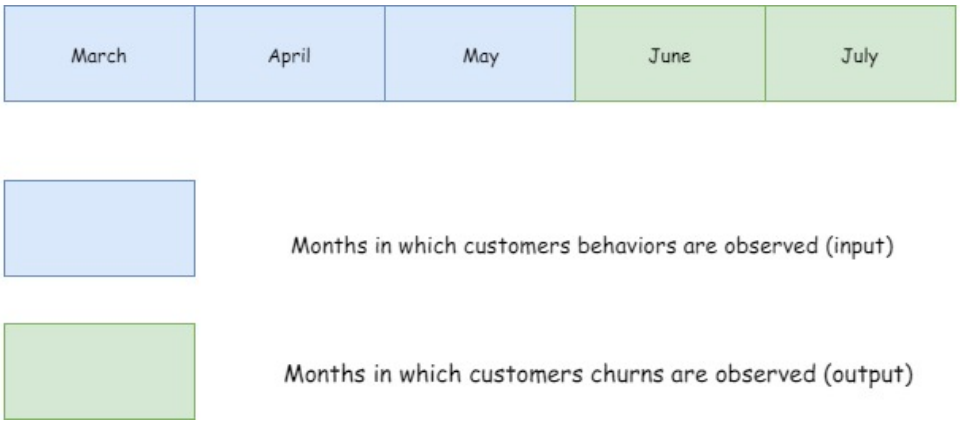


Figure 2: An example of timeline of prediction model

3000 samples will be divided to training samples and testing samples, which possess 70% and 30% of total samples respectively since the amount of dataset is adequate.

1.4.3 Feature Selection

One method will be selected based on literature study. Feature selection is conducted with the help of Microsoft Azure machine learning studio, which is an analysis tool that can be used to build, test, and deploy predictive analytics solutions. The results will be input attributes to the models.

1.4.4 Modeling and Testing

2 - 3 machine learning algorithms will be selected based on literature study. The training data which are filtered after data collection and feature selection are used to train the models which are built with different methods. Microsoft Azure machine learning studio is used to build the models. The testing samples will be tested on the models. At the end, the conclusions will be drawn based on the evaluation of models.

1.5 Structure of the Thesis

The remainder of this thesis is organized as follows. In Chapter 2, literature study related to feature selection, supervised machine learning algorithms and evaluation methods used in this thesis will be described. Chapter 3 describes the experimental results, and Chapter 4 includes discussion about findings of experiment and suggestions for future research. Finally, conclusions and future work are outlined in Chapter 5.

2 Literature Study

This chapter will present the result of literature study. Firstly, several feature selection methods will be presented. Secondly, several machine learning models will be described. At the end, the evaluation methods will be presented.

2.1 Feature Selection

Feature selection in machine learning refers to identifying a representative set of features from which to construct a classifier for a particular task [13]. The goal of feature selection is to choose a subset X_s of the complete set of input features $X = \{x_1, x_2, x_3, ..., x_M\}$ (M is the dimension of X) so that the subset X_s can predict the output Y with Accuracy comparable to the performance of the complete input set X, and with great reduction of the computational cost.

There are four reasons that feature selection is conducted [41]:

1. Make learning models easier to interpret.
2. Shorter training times.
3. Avoid the curse of dimensionality.
4. Reduce overfitting of models. In overfitting, a model describes random error or noise instead of the underlying relationship.

Two feature selection methods are considered in this thesis:

2.1.1 Forward Selection

In forward selection, variables are incorporated into larger and larger subsets step by step [12]. The procedure begins by evaluating all feature subsets which consist of only one of the input attributes: $\{X_1\}$, $\{X_2\}$, ...,$\{X_M\}$. Then select the best individual feature X and doing the evaluation by including one other feature from the remaining $M-1$ input attributes to find the best input subset with two features. Afterwards, the input subsets with more features are evaluated progressively [7]. The subset includes the variables which gives the most statistically significant improvement of the fit. The prespecified criterion can be F-tests, t-tests, adjusted R-square, Akaike information criterion, Bayesian information criterion, etc [23].

2.1.2 Backward Elimination

Backward elimination is another feature selection method. Unlike forward selection, Backward elimination starts by building the full model using the whole dataset. It eliminates one variable at a time based on the least deterioration in model fit [12]. Like the forward selection, the goal of backward elimination is find the subset which matches statistical criterion. However, some claim that backward elimination is computational expensive [16][12]. Due to this reason, forward selection will be used in this thesis project.

2.2 Supervised Machine Learning

Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances [21]. The goal of supervised learning is to build a learning model of the distribution of class labels in terms of predictor features. The resulting classifier (model) is afterwards used to assign class labels to the testing data whose values of the predictor features are known, but the value of the class label is unknown.

2.2.1 Algorithm Selection

A number of supervised learning algorithms have been examined, like decision tree, Naive Bayes, support vector machine (SVM), artificial neural network (ANN), random forest and Boosted Decision Tree. Three of them are chosen to be applied in this thesis: Boosted Decision Tree, SVM and ANN. There are several reasons. Naive Bayes requires zero training time and little storage space during both the training and classification stages. However, it has poor Accuracy in general compared with other supervised learning algorithms [21]. Furthermore, naive Bayes models are poor at predicting calibrated probabilities because of the unrealistic independence assumption [5], which makes it an unsuitable algorithm for the thesis project.

Decision tree is a simple and logic-based supervised learning algorithm. Decision trees do not have better performance when dealing with multi-dimensions and continuous features as SVMs and ANNs do. As [43] described: an individual decision tree is a weak learner which does not perform well when it works individually. Therefore, decision tree is excluded based on its working weakness. Two other algorithms which are related to decision tree are random forest method and Boosted Decision Tree. They are so-called ensemble learning methods which will generate many classifiers and aggregate their results [24]. Random Forest is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction [34]. Boosted Decision Tree uses boosting algorithm to design a procedure that combines many decision trees to achieve a final powerful classifier [43]. The authors of [5] suggest Boosted Decision Trees have overall better performance than other algorithms including random forest. Hence, Boosted Decision Tree is chosen as one of the modeling algorithms.

As mentioned above, SVMs and neural networks tend to perform much better when dealing with continuous features and they have high Accuracy in general compared with other supervised learning algorithms [21]. Therefore, they are also chosen as modeling algorithms in this thesis.

2.2.2 Boosted Decision Tree

The boosting algorithm is one of the most powerful learning techniques introduced in the past decade [31][10]. Motivation for the boosting algorithm is to design a procedure that combines many "weak" classifiers to achieve a final powerful classifier [43] [31] [10] [30]. Decision trees are known as weak learners since they are not stable. Small fluctuations in the data can make huge differences. This feature makes decision trees perfect "working partners" to boosting algorithm because boosting is used to reduce the error of any weak learning algorithm.

The given number of decision trees are constructed on weighted events. For the first round, all events are assigned equal weights. For the second round, the weights are increased for the events misclassified by the first decision tree and decreased for the events correctly classified by the first decision tree. Therefore, the second decision tree focuses on the samples misclassified by the first tree. In other words, the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. This procedure is repeated until all trees are created. In the group, each member has its own specialty arising from its special training. It leads to that the group of diverse members can do a better job than any single member. Boosting algorithm aims at reducing test error in the training set [27]. The pseudo code of Boosted Decision Tree algorithm is described by Algorithm 1:

Algorithm 1: Boosted decision tree algorithm

Function $I(X)$
  $I(X)$ = 1 if X is true, 0 otherwise
Function *Boosted Decision Tree*
  Data: Training sample $\mathbb{S} = \{ S_1, S_2, ..., S_N \}$, a group of classifiers $\{C_1(x), C_2(x)....C_T(x)\}$
  Initialize the observation weights $w_i = 1/N$, $i = 1, 2, ..., N$;
  //Iterate T times
  for *t = 1 to T* **do**
    Train classifier $C_t(x)$ on $\mathbb{S}_t$;
    Compute weighted error of newest tree;
    $err_t = \frac{\sum_{i=1}^{N} w_i I(y_i \neq C_t(x_i))}{\sum_{i=1}^{N} w_i}$
    Compute $\alpha_t = log[(1 - err_t)/err_t]$
    //Update weights
    for *i = 1,...,N* **do**
      $w_i \leftarrow w_i \cdot exp[\alpha_t \cdot I(y_i \neq C_t(x_i))]$;
    end
    Normalize the weights;
  end
  $C(x) = sign[\sum_{t=1}^{T} \alpha_t C_t(x)]$

Boosted Decision Tree is one of the algorithms that do not need pre-feature selection procedure since it is doing the selection while training. Under the training, a feature summary is created internally and features with weight 0 are not used by any tree splits [4].

2.2.3 Support Vector Machine (SVM)

SVMs have strong theoretical foundations and excellent empirical successes [35]. SVM method is one of the most recommended algorithms in classification problem [21]. In this algorithm, each data item is plotted as a point in n-dimensional space where n is the number of features, with the value of each feature being the value of a particular coordinate. The goal is finding the hyper-plane that differentiate the classes very well [33]. It means SVMs maximize the margin around the separating hyperplane (Figure 3).

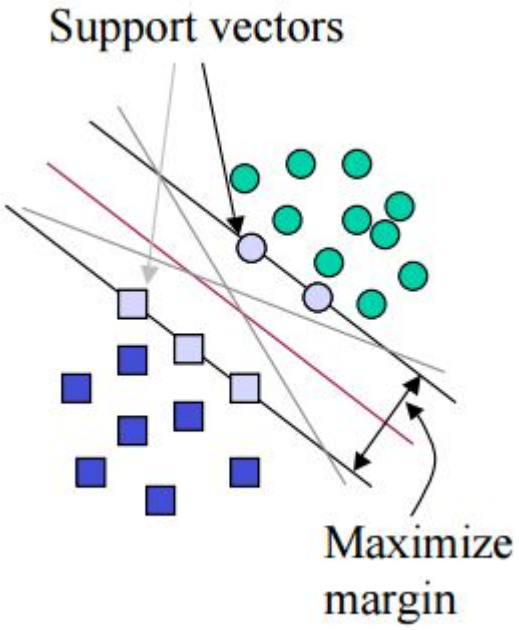


Figure 3: Hyper-plane of support vector machine

If the training data is linearly separable, then a pair (w,b) exists such that

$$w^T \cdot x_i + b \geq 1, \text{ for all } x_i \in \text{Positive instances}$$
$$w^T \cdot x_i + b \leq -1, \text{ for all } x_i \in \text{Negative instances}$$

Where w is a weight vector and b is bias.
The width of margin will be $2/||w||$. Hence, the maximization can be converted to minimizing the $||w||$. Therefore, the minimization can be set up as a convex quadratic programming problem:

$$min\ f : 1/2||w||^2,$$
$$\text{subject to } y_i(w^T x_i + b) \geq 1.$$

Once the optimum separating hyperplane is found, data points that lie on its margin are

known as support vector points and the solution is represented as a linear combination of only these points. Other data points are ignored.

Most real-world problems involve non-separable data for which no hyperplane exists that successfully separates the positive from negative instances in the training set. One solution to the inseparability problem is to map the data onto a higher dimensional space using kernel tricks [39] and define a separating hyperplane there. With an appropriately chosen kernel functions of sufficient dimensionality, any consistent training set can be made separable [21].

SVMs have high-variance which may lead to overfitting problem [21]. Therefore, feature selection procedure is needed before modeling.

2.2.4 Artificial Neural Network

Artificial neural networks are mathematical models inspired by the organization and functioning of biological neurons. One of the advantages of ANNs over statistic methods is that ANNs can be mathematically shown to be universal function approximators [15]. This means that artificial neural networks can automatically approximate whatever functional form best characterizes the data. There are widely usages of ANNs, like modeling real neural networks, pattern recognition, forecasting and data compression [11].

In ANNs, weights assigned with each arrow represent information flow. The weights are initialized with random values. Each training set is then presented for the perceptron in turn. An activation function Act is introduced for calculating the expected output a_i from each related input $x_0, x_1 ... x_j$:

$$a_i = Act(\sum x_j w_{j,i})$$

For every input set, the output from the perceptron is compared to the desired output y_i. If the output is not as desired, the weights will be adjusted on the currently active inputs towards the desired result by doing [29]:

$$\begin{aligned} e_i &= y_i - a_i \\ \bigtriangledown w_{ji} &= \alpha e_i x_j \\ w_{ji} &\leftarrow w_{ji} + \bigtriangledown w_{ji} \end{aligned}$$

In which α is the learning rate that decides the speed of learning.

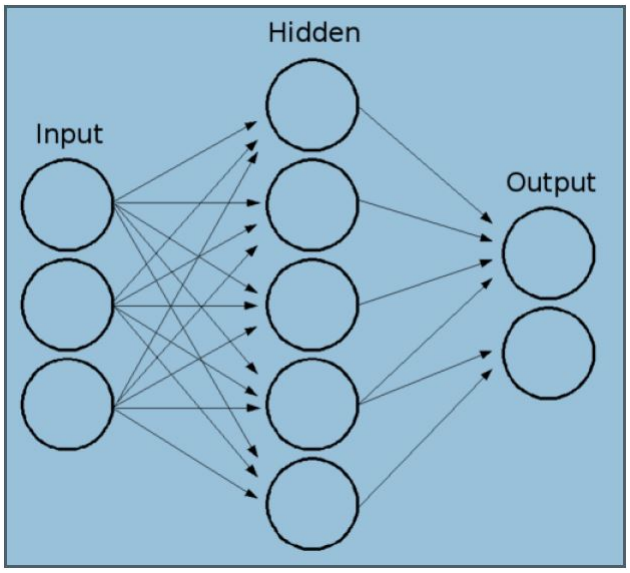


Figure 4: Multi-layer perceptron

The ANNs which have one or more hidden layers called multi-layer perceptron (MLP) (Figure 4). The MLP is divided into three layers: the input layer, the hidden layer and the output layer, where each layer in this order gives the input to the next. The extra layers gives the structure needed to recognize non-linearly separable classes [26].

Similar to SVMs, ANNs have high-variance as well [21]. Therefore, feature selection procedure will be applied before modeling.

2.3 Evaluation Methods

Since more and more researchers has been realized that simple classification Accuracy is often a poor metric for measuring performance [9] [8], two advanced methods are chosen in this thesis to evaluate the quality and effectiveness of machine learning models.

2.3.1 ROC Curve and AUC

A receiver operating characteristics (ROC) chart is a technique for visualizing, organizing and selecting classifiers based on their performance. Recent years an increase in the use of ROC graphs in the machine learning domain has been seen [8]. The ROC chart shows false positive rate (1-specificity) on X-axis, the probability of target is 1 when its true value is 0, against true positive rate (sensitivity) on Y-axis, the probability of target is 1 when its true value is 1. Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases. The diagonal line in Figure 5 is for a random model. Each binary classifier (for a given test set of examples) is represented by a point on the graph. By varying the threshold of the probabilistic classifier, a set of binary classifiers are obtained, represented with a set of points on the graph. ROC curve is independent of the threshold value of classifiers and is therefore suitable for comparing classifiers when the threshold may vary [40].

Area under ROC curve (AUC) is often used as a measure of quality of the classification models [40]. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1 (Figure 5). An area under the ROC curve of 0.8, for example, means that a randomly selected case from the group with the target equals 1 has a score larger than that for a randomly chosen case from the group with the target equals 0 in 80% of the time.

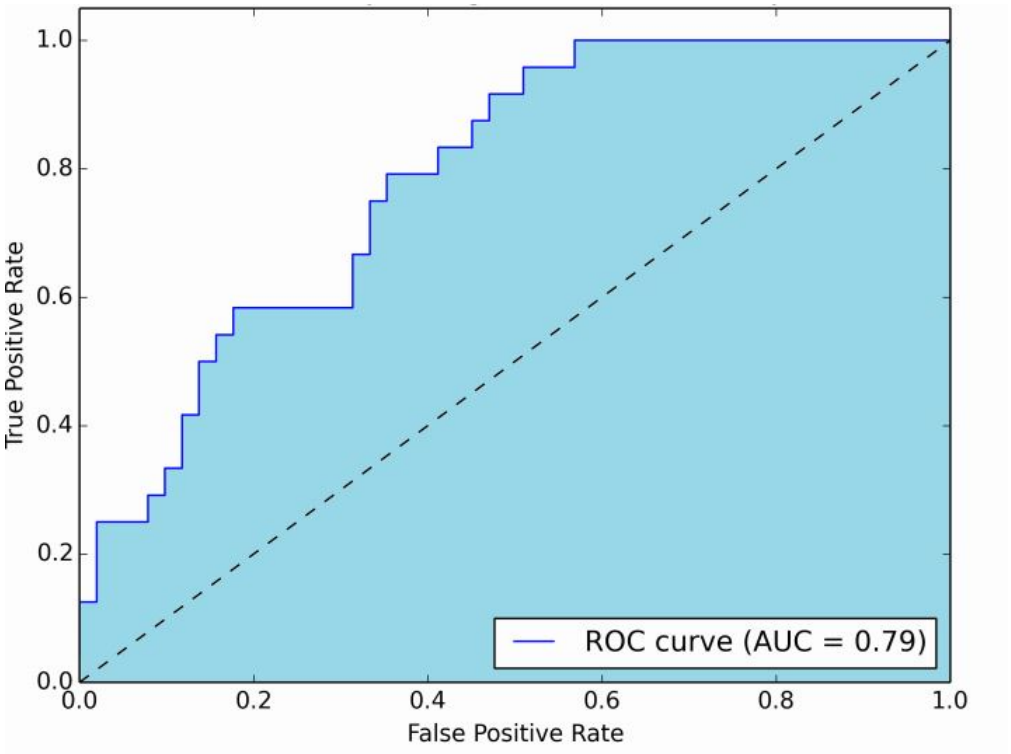


Figure 5: ROC and AUC

2.3.2 Lift Charts

Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model [14]. For example, suppose a population has an average positive rate of 10%, but a certain model has identified a segment with a positive rate of 30%. Then that segment would have a lift of 3.0 (30%/10%). In a cumulative lift chart (gains chart), the input of a lift curve is sorted by the scores from the training model from high to low. The y-axis shows the true positive numbers. The x-axis shows the percentage of total data in the sorted order. The baseline (red line in Figure 6) represents the expected number of positives we would predict if we did not have a model but simply selected cases at random. It provides a benchmark against which we can see performance of the model. The green line represents the perfect lift curve. Lift curves in reality usually lies between the green line and red line.

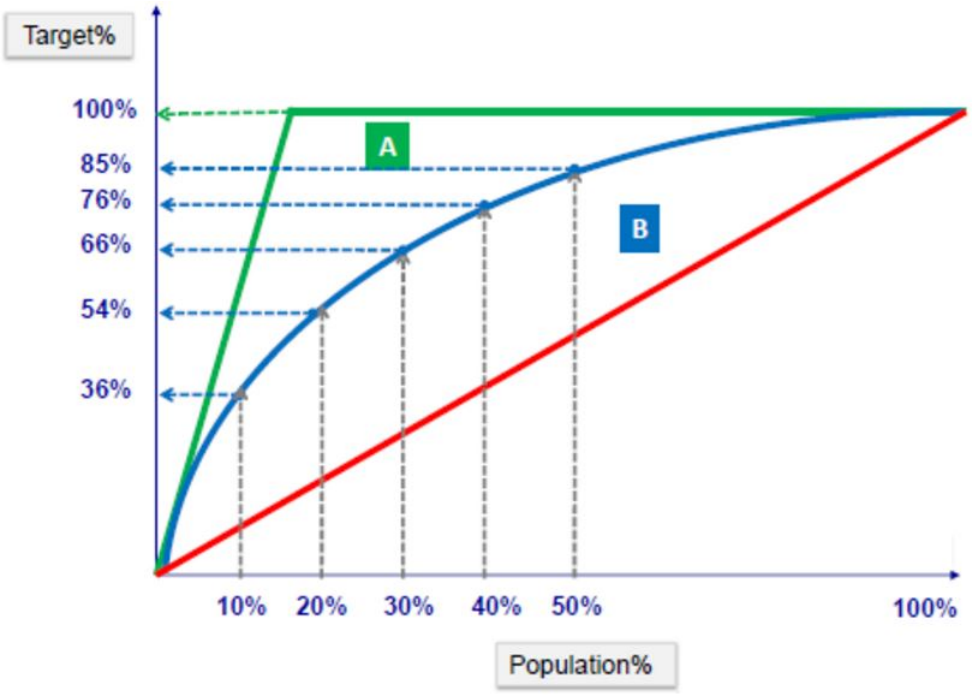


Figure 6: cumulative lift chart

2.4 Summary of Literature Study

Based on the studies in this chapter, the learning process and methods shows in Figure 7. Forward selection will be applied before training models with SVM and ANN, while Boosted Decision Tree will be used without feature selection. At the end, ROC chart and lift chart will be used to evaluate the models.

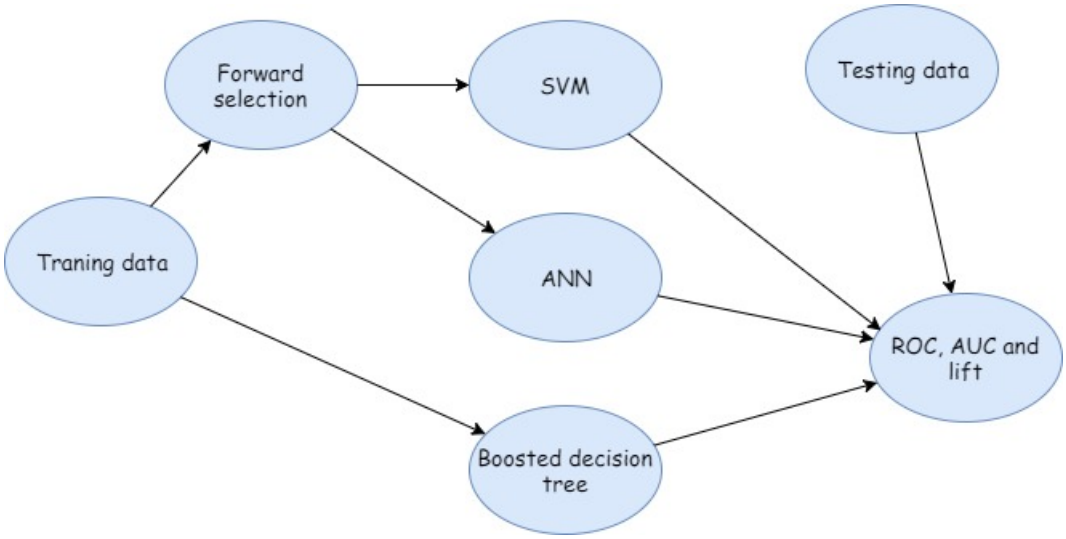


Figure 7: Learning process and methods

3 Methods and Experimental Result

The main experimental findings are presented in this chapter. A brief description about data collection and cleansing will be presented in the first section. Learning evaluation will be described in the second section.

3.1 Data Collection and Cleansing

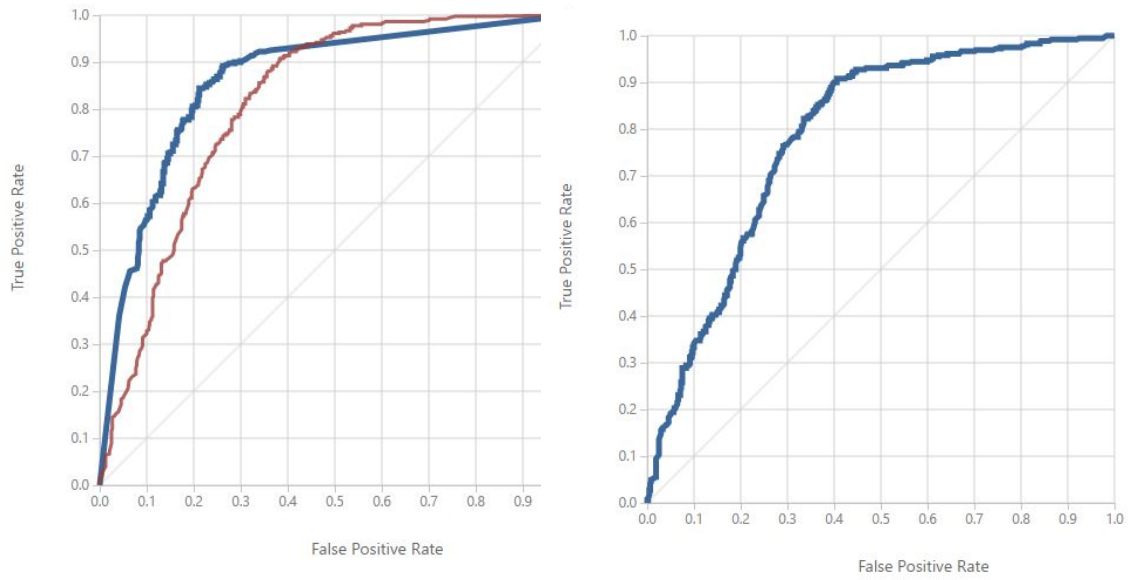
As it is described in Section 1.4, 3000 balanced samples are collected. Each sample has a table which contains all training records under year 2016, from which arbitrary three-month records will be used as input for non-churn samples and the record contains the latest three months before churning will be selected for churned samples. A database is built to store and process the raw data to an organized dataset which models can recognize. In the final dataset, each row represents a customer and columns present the attributes related to customers. Most of attributes generates as training frequency by certain period of certain form.

After aggregation, 198 columns are generated. The data points with empty value by any attribute were removed from the dataset. At the end, 2803 data points are valid among which 1219 are customers who are churned.

3.2 Learning Evaluation

Section 2.4 shows the learning process based on literature study. Under feature selection, different numbers of relevant attributes were tested for training the models based on ANN and SVM algorithm in order to avoid the overfitting problem. Even though it was mentioned that the Boosted Decision Tree model does not need feature selection, feature selection was applied to test if this assumption holds in this project. Different parameter choices for algorithms were tested on each models. The results below show the optimal combination of parameters of each algorithm.

The ROC curves at Figure 8 show that ROC curve of Boosted Decision Tree is closest to the left-top corner of charts. The AUC values of the models which are generate by these three algorithms are presented at Table 1. The values indicate the model trained by Boosted Decision Tree scores best of all three. And the feature selection procedure affect little learning process for all three algorithms if the number of features is greater and equals to 20.



(a) ROC curves of Boosted Decision Tree (blue curve) and ANN (red curve)

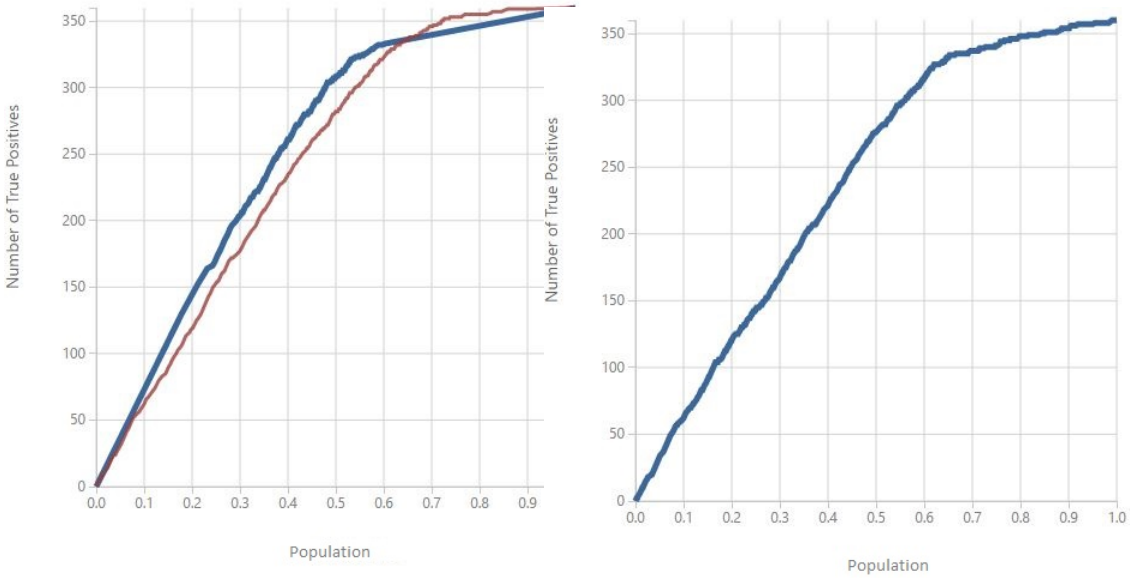
(b) ROC curve of SVM

Figure 8: ROC curves

Table 1 AUC Values of Models

number of features	Boosted Decision Tree	ANN	SVM
10	0.817	0.813	0.797
20	0.876	0.814	0.800
30	0.866	0.816	0.795
40	0.868	0.814	0.799
50	0.864	0.818	0.798
All	0.867	0.820	0.786

Figure 9 presents lift charts of three models. Among top-10%-percentile, the model of Boosted Decision Tree catches about 70 churners while models of ANN and SVM catches around 60 respectively. Among top-50%-percentile, the model of Boosted Decision Tree catches more than 300 churners while models of ANN and SVM catches around 280 and 275 respectively. This indicates that the Boosted Decision Tree model has the highest lift value; thus, it is the most effective classifier among all three in general.



(a) Lift charts of Boosted Decision Tree (blue curve) and ANN (red curve)

(b) Lift chart of SVM

Figure 9: Lift charts

4 Discussion

It is described in Section 1.2, feature selection is one of the steps of churn prediction. However, as the evaluation presented in Chapter 3 indicates, the feature selection procedure has little impact on the learning process for all three algorithms in this project. This can possibly due to two reasons. Firstly, the dataset has high quality; thus, there is no significant overfitting problem in the dataset. Secondly, the feature selection method used in this project is unsuitable for all three learning algorithms. Due to the time limitation, more feature selection methods could not be tested in this project. Therefore, one suggestion for the future research is to compare different feature selection methods for churn prediction models.

The result also indicates the Boosted Decision Tree algorithm creates the most effective classifier with high quality based on AUC values and Lift charts in this thesis project. Unlike the result presented in [6] that SVMs work well as classifiers, SVM is actually the worst algorithm compared with the other two. I assume there is big difference of datasets from [6] and this thesis.

On the other hand, there is no guideline which can claim certain feature selection methods or machine learning algorithms are absolutely best. The answer is always ”It depends!”. Even the most experienced data scientists cannot tell which algorithm will perform best before trying them [2]. It makes sense in a way that each dataset has unique features and factors which makes it hard to reproduce others learning process. Moreover, parameter settings of most learning algorithms which affect the learning speed and learning Accuracy are different and it is difficult to know the optimal combination without testing many times. Therefore, one suggestion regarding method choice is choosing several suitable methods based on previous experiments and surveys and try them.

In this thesis, ROC, AUC and Lift were used for evaluation as Accuracy Rate does not give much information as evaluation merit. Which model is relatively better can easily be seen under comparison. However, there is no research which has been done to indicate how good a model is based on its AUC and lift value. Thus, it is hard to identify whether a model is ”good enough”. On the other hand, it is important to know how good a model is which can to some extent imply the value of the model from business' perspective. Therefore, one suggestion for future work is to develop and explore a more complete system for evaluation of models' value.

The models built in this project used one-time learning: only certain period data is used and models are trained one time. The prediction ability of this kind of learning model may decrease as time goes by. Hence, lifelong machine learning models can be a big help to solve this problem. As in lifelong machine learning, models can automatically take in newgenerated data and automatically modify themselves in order to keep the timeliness of the models.

Churn prediction models are essential for CRM since it may help the companies to save the customers who want to leave. In this thesis, machine learning has showed its capability

of using and creating customer-specific data. In fact, this capability can lead to machine learning being involve in entire customers' life circle. For example, when a customer joins a company, the company can recommend the most suitable products according to specific profile through learning process. During the membership period, actions can be take to enhance user experience based on data like shopping record. Hereafter, churn prediction can be applied to prevent the customer from terminating his membership. In a word, machine learning holds significant potential regarding developing customer relations which may be a big help for companies concerning CRM and decision making.

5 Conclusions

In this thesis, forward feature selection was used for data preprocessing in Chapter 3. Boosted Decision Tree, ANN and SVM were applied to build three machine learning models of customer churn prediction for monthly-paying customers in fitness industry. ROC charts, AUC value and lift charts were used for evaluation. According to the experimental result, the model based on Boosted Decision Tree has the best quality and effectiveness of all three models. During the experiment, several problems like algorithm selections and model evaluations were addressed and suggestions were made. More completed evaluation system for model value evaluation can be considered in the future research. Furthermore, machine learning has capacity to be engaged more in business management especially CRM. Hopefully, more advanced and complex models can be developed to assist more industries and companies.

References

[1] Dudyala Anil Kumar and Vadlamani Ravi. Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1):4–28, 2008.

[2] Microsoft Azure. How to choose algorithms for microsoft azure machine learning, 2015. https://docs.microsoft.com/sv-se/azure/machine-learning/machine-learning-algorithm-choice//, accessed May 2017.

[3] Microsoft Azure. Analyzing customer churn by using azure machine learning, 2016. https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-azure-ml-customer-churn-scenario/, accessed March 2017.

[4] Microsoft Azure. Feature selection modules, 2017. https://msdn.microsoft.com/en-us/library/azure/dn905912.aspx/, accessed April 2017.

[5] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[6] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1):313–327, 2008.

[7] Kan Deng and Andrew W Moore. *On Greediness of Feature Selection Algorithms*.

[8] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[9] Tom Fawcett and Foster Provost. Adaptive fraud detection. *Data mining and knowledge discovery*, 1(3):291–316, 1997.

[10] Yoav Freund, Robert Schapire, and N Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[11] Neha Gupta. Artificial neural network. *Network and Complex Systems*, 3(1):24–28, 2013.

[12] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[13] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[14] Howard Hamilton. Cumulative gains and lift charts, 2012. http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html//, accessed April 2017.

[15] Tim Hill, Leorey Marquez, Marcus O'Connor, and William Remus. Artificial neural network models for forecasting and decision making. *International journal of forecasting*, 10(1):5–15, 1994.

[16] Xia Hong and SA Billings. Givens rotation based fast backward elimination algorithm for rbf neural network pruning. *IEE Proceedings-Control Theory and Applications*, 144(5):381–384, 1997.

[17] Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, Hossam Faris, et al. Hybrid data mining models for predicting customer churn. *International Journal of Communications, Network and System Sciences*, 8(05):91, 2015.

[18] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006.

[19] Adnan Idris, Muhammad Rizwan, and Asifullah Khan. Churn prediction in telecom using random forest and pso based data balancing in combination with various feature selection strategies. *Computers & Electrical Engineering*, 38(6):1808–1819, 2012.

[20] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[21] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[22] Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.

[23] Machine learning algorithms. Stepwise regression, 2017. http://trymachinelearning.com/machine-learning-algorithms/regression/stepwise-regression/, accessed March 2017.

[24] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[25] Ning Lu, Hua Lin, Jie Lu, and Guangquan Zhang. A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, 10(2):1659–1665, 2014.

[26] Fiona Nielsen. Neural networks – algorithms and applications. *Niels Brock Business College*, 2001.

[27] Yinsheng Qu, Bao-Ling Adam, Yutaka Yasui, Michael D Ward, Lisa H Cazares, Paul F Schellhammer, Ziding Feng, O John Semmes, and George L Wright. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical chemistry*, 48(10):1835–1843, 2002.

[28] Werner J Reinartz and Vita Kumar. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of marketing*, 67(1):77–99, 2003.

[29] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25:27, 1995.

[30] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[31] Robert E Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[32] Anuj Sharma, Dr Panigrahi, and Prabin Kumar. A neural network based approach for predicting customer churn in cellular network services. *arXiv preprint arXiv:1309.3945*, 2013.

[33] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[34] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[35] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[36] Chih-Fong Tsai and Yu-Hsin Lu. Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10):12547–12553, 2009.

[37] Dirk Van den Poel and Bart Lariviere. Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, 157(1):196–217, 2004.

[38] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3):2354–2364, 2011.

[39] Analytics Vidhya. Understanding support vector machine algorithm from examples, 2015. `https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code//`, accessed April 2017.

[40] Miha Vuk and Tomaz Curk. Roc curve, lift chart and calibration plot. *Metodoloski zvezki*, 3(1):89, 2006.

[41] Wikipedia. Feature selection, 2017. `http://en.wikipedia.org/wiki/Feature_selection`, accessed March 2017.

[42] Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.

[43] Hai-Jun Yang, Byron P Roe, and Ji Zhu. Studies of boosted decision trees for miniboone particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 555(1):370–385, 2005.

[44] Yu Zhao, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren. Customer churn prediction using improved one-class support vector machine. In *International Conference on Advanced Data Mining and Applications*, pages 300–306. Springer, 2005.