A Framework for the Prediction of User Potential in Marketing
Silvio MARCOVIC\textsuperscript{a} Tatsuya ASUKAI\textsuperscript{b} Bernd HOLLERIT\textsuperscript{b} Kenji TANAKA\textsuperscript{b} Kazuhiro AOYAMA\textsuperscript{b}

\textsuperscript{a}Graz University of Technology, Institut für Softwaretechnologie, Inffeldgasse 16b/II, 8010 Graz Austria
\textsuperscript{b}Department of Systems Innovation, School of Engineering, The University of Tokyo, Tokyo, 113-8654 Japan

Abstract
Managing marketing funds in order to maximize profits while minimizing costs is an inherent objective of any business. A key to optimizing this relation is to select a subset of customers to target marketing expenses. An approach is presented, which helps identifying such a subset at an early stage of their life cycle. We suggest a measure to rate the users’ value relative to others, and present an approach how to predict the user rating. This work compares and evaluates two algorithms (naive Bayes and random forest) which predict the user rating. Two groups were used, the first representing the top rated users, which are the part of the subset on whom the marketing funds should be focused, and a second one containing the users which may be neglected. The dataset is skewed, due to the amount of users of interest being low. It is shown how the algorithms can deal with such an imbalanced set. The algorithms are evaluated with a precision-recall curve. The framework is compared to other prediction models and metrics such as recency, frequency, monetary model and customer lifetime value. Since the framework is flexible, it encourages including and expanding other established metrics.

Key words: Machine learning, Class imbalance problem, Customer relationship marketing, User rating

1 Introduction

Customer relationship marketing is the process of creating, maintaining, and enhancing strong, value-laden relationships with customers and other stakeholders [1]. Companies have changed their focus of attention from transaction based selling platforms to relational based approaches. The focus now lies in customer relationship marketing and loyalty. More emphasis is put on involving customers in the long-term. Examples of these are loyalty bonuses, value offers, deals, and social media services as a way to create an interactive relationship. By developing customer loyalty, it is believed that steady sales over the lifetime of a customer can be achieved [2].

If a customer’s revenue over time exceeds the cost of acquisition and servicing the customer, they become profitable. An important task in the field is to analyse which customers could be profitable to invest in, and which are less likely. Marketing campaigns are only as successful as their targeted customers. Ideally, the targeted subset of customers should only contain potential and/or valuable customers [3].

In this paper an approach is proposed to determine such a set of potential customers. The target is to identify which customers are likely to be loyal to the business. In order to identify such users, a metric to evaluate the worth of a customer is presented, which is denoted as user rating (UR) in the following. Through a data-mining method, this value is predicted at an early-stage of the customers’ lifetime.

To show the credibility of our framework, a case study was conducted. A database of a Japanese e-commerce company was analysed, where the URs were calculated and predicted. The presented approach identifies around 80 percent of the top 1 percent ranked users based on the UR, while maintaining an error rate between 5 and 7 percent, depending on the algorithm used. Since there is a vast amount of possibilities of defining the UR and evaluating it, the case study should be seen as one possibility on how to apply our proposed process.
The rest of this paper is organized as follows: Section 2 explains the proposed framework. Section 3 shows the application of our framework on actual data. The discussion of the main differences between related research and the proposed method is found in Section 4. Finally, the research is concluded in Section 5.

2 A framework for potential user classification

Figure 1 shows a flow chart on how the provided data are processed to receive the final result, a subset of users. The numbers in Figure 1 correspond to the subsection numbers of the current section. They will not be referenced explicitly further.

2.1 Preparation

The whole process is split into two main parts: the preparation and the application part. The former covers the goal formulation, which users are of interest and which are not. How to decide on features and evaluate them is covered in this part.

2.1.1 Identify users of interest

Firstly, it is essential to understand how an average user behaves. This can be done by clustering users and plotting their purchase behaviour over time, to acquire an understanding of which segments currently exist in the service.

The next step is to identify on which users to focus the marketing expenses. Objective criteria can be used such as users that have a high customer lifetime value [4], or focusing on a certain segment, such as commercial users. Also, identifying which market strategies are in place is an important factor. Customer attraction techniques such as offering unprofitable products for the business, but cheap for the customer, have to be taken into account.

2.1.2 Decide on features

With this information, features for the users can be derived, which have the potential to separate users with potential from average ones.

User rating: A ranking formula based on multiple attributes is used to evaluate the customers’ value. The formula is multiplicative, since we try to focus on customers that excel in every feature. In an additive formula a user can receive a very low score in one feature and still perform good overall. Well-balanced solutions seem to be closer to how a human would pick between several options presented [5].

As a general form, UR can be expressed as in Equation 1.

\[ UR = \prod_{i=1}^{n} (f_i^{w_i}) \]  

Using \( F = \{f_1, f_2, ..., f_n\} \) as a set of features where each feature uses a range normalized score \( RNC = (x - x_{min}) (x_{max} - x_{min}) \) and therefore \( \forall f \in F : f \in \mathbb{R}, 0 \leq f \leq 1 \) is true. \( W = \{w_1, w_2, ..., w_n\} \) is a set of weights respectively, where \( \forall w \in W : w \in \mathbb{R}, 0 < w < 1 \) and \( \sum_{i=1}^{n} w_i = 1 \) ensuring that \( UR \in \mathbb{R}, 0 \leq UR \leq 1 \). n is the number of features.
number of features used.

2.1.3 Adjust weights by objective criteria

Further, the importance of each criterion used in the UR is evaluated. This can be done by any objective or subjective criterion. An objective criterion can be formulated as shown in Equation (2). The optimal weight set \( W_{opt} \) is chosen through finding the maximum of a decision function, taking the weights as an input.

\[
W_{opt} = \arg\max_{w_1, \ldots, w_n \in \mathbb{R}} f(w_1, \ldots, w_n) \quad (2)
\]

Minimizing the cost of a decision function is possible as well. General discussion on how to find an optimum will not be part of this research. The interested reader may refer to [6].

2.2 Application

The application covers the part where the previously defined UR is predicted. By deciding on independent variables, using appropriate algorithms for classification and adjusting precision recall values, the final subset of users is acquired.

2.2.1 Decide on set of independent variables

Independent variables are collected for a defined period of time. In practical applications, this could be done multiple times after certain intervals. To choose an appropriate time frame is one of the key aspects to determine the utility of the model. If the chosen time frame is too short, the performance of the algorithms used decreases significantly. If the time frame is too long, the benefit of predicting is lower; the users are already closer to their real value.

The type of independent variables can be divided into two parts. The first type of data contains customers’ geographic and demographic data. The second type of data is based on the customers’ interactive behaviour with the business. They include transactions, feedback from customers, and web records [7].

2.2.2 Apply classification algorithms

Applications that try to predict rare, but important samples, suffer from the class imbalance problem [8]. Since the users of interest are relatively few, the classification algorithms used will suffer from this problem. This section will first explain the class imbalance problems and then introduce the two algorithms used in this research: naive Bayes classifier and random forest.

For an imbalanced dataset, the class having greater numbers of instances is called the majority class while the one having relatively fewer number of instances is called a minority class. Literature may also refer to the classes as major and minor class. Most classifiers are biased towards the majority class and show poor classification rates on the minority class [9, 8].

There are four groups of methods proposed for solutions of the class imbalance problem. The first one is sampling methods. They construct balanced training data sets. Undersampling may cause loss of useful information, while oversampling may cause over-fitting. There are additional techniques that generate synthetic samples such as SMOTE (Synthetic Minority Over-sampling Technique). The second group contains cost sensitive methods. They assign different costs of misclassification to the minority and majority class. Cost sensitive learning may lead to over-fitting. Ensemble-based methods use multiple classifiers to improve generalization and prediction accuracy. Recognition-based methods learn on the minority class samples [8]. However, since decision trees and naive Bayes cannot be built by one class learning, recognition-based methods are not further discussed in this research.

Naive Bayes classifier: As a first model a naive Bayes classifier was chosen, as it works fast and often competes with more sophisticated algorithms [10]. For categorical variables a multinomial naive Bayes was used; for the continuous variables a Gaussian naive Bayes model was used. A final Gaussian naive Bayes model was used to predict the final result using the probabilities of the multinomial and Gaussian model as independent variables. To address the class imbalance problem, weights were used for each sample depending on the class.

Random forest: The second model is a random forest [11], which uses oversampling on the minority class. In literature, there are different approaches on how to use the mtry value [12, 13]. In the model used, mtry was tuned based on the lowest out-of-bag error. The number of trees was set to 500. The model predicts the UR of each user. Users above a chosen threshold will be part of the top group, the rest will be part of the bottom group.

2.2.3 Evaluation of model results

The most common evaluation metric for traditional applications is accuracy. Since accuracy is not suitable to evaluate imbalanced data sets, new metrics have been proposed. Precision and recall are often used to measure the classification quality of binary classifiers. Precision
measures the fraction of predicted samples from the relevant class. Recall measures the fraction of samples from a relevant class that are predicted correctly. False positive and true positive rates are often used as well. False positive rate measures the percentage of misclassified negative samples. True positive rate measures the percentage of correctly classified samples, which is the same as recall [14, 15].

Receiver operating characteristics (ROC) graphs are useful for unbalanced class distributions. ROC graphs are two-dimensional graphs that plot the true positive rate on the y-axis and the false positive rate on the x-axis. A precision-recall graph is a graph that has the precision plotted on the y-axis and the recall on the x-axis. To compare classifiers with each other a single metric from the graphs is used, the area under the curve (AUC) [16]. The differences between ROC and precision-recall curves lies in how they treat true negatives. If true negatives are not important for the task, which is arguably the case for our problem statement, precision-recall is typically more useful.

In order to calculate AUC, interpolation between points is performed. In ROC space this can be done through linear interpolation and is therefore straightforward. In precision-recall space, interpolation should not be done linearly. As recall varies, precision does not necessarily change linearly. If linear interpolation is used, this can yield to an overly-optimistic estimation. It is proposed to create new points by applying the formula in (3) for all integer points between $A$ and $B$, where $A$ and $B$ are points that are far apart in the precision-recall space [17].

$$TP_A + x \over TP + FN \cdot TP_A + x \over TP_A + x \cdot FP_A + \over FP_A - FP_B \cdot x$$ (3)

For further information on how to evaluate AUC values for imbalanced data sets, refer to [18]. A weighted AUC value is suggested, since the misclassification of rare events can be costly and more expensive than a false alarm.

Upon looking at the evaluation metrics, especially the AUC values, one can choose the appropriate algorithm for the problem. A threshold for the binary classification should be chosen to retrieve the final result, a subset of users.

Classifying a high percentage of potential users may be of interest, therefore, a point with high recall value should be chosen for the minority class, while keeping the overall misclassification error in mind.

2.2.4 Investigate misclassifications

Optionally, if the results of the algorithm show poor performance, the incorrectly classified samples could be analysed. There may be some additional variables that have the power to distinguish true positives from false positives or respectively true negatives from false negatives. The new independent variables should be included into the model and the algorithms should be applied again. It is also possible to set a higher time frame for data collection as discussed previously. This process can be done until the results show significance and are satisfying.

3 Case Study

The research was conducted on transaction data of an e-commerce company operating in Japan. Table 1 displays the characteristics of the data set. We used a subset of users to focus on long-life customers. Long-life customers are defined as customers that were starting to use the service in the first 6 months, they make up about 17 percent of the whole user base. Further in this research, user and customer refers to the defined subset.

| Duration       | 18 months       |
|----------------|-----------------|
| Timespan       | 2012/10 - 2014/04 |
| Number of Categories | 14             |
| Number of Transactions | 2.7 M of 10 M  |
| Number of Users       | 169 K of 1.1 M  |

Table 1: Data set characteristics

Each user’s rating was calculated. The top 1 percent URs form the first class of users, while the bottom 99 percent form the second class.

One half of the data set was used for training and testing, and the other half was used for the evaluation of the classifiers. Both partitions have the same amount of users from each class. The train and test set was further randomly split into their sets respectively using different sizes depending on the algorithm used.

3.1 Model use

This subsection shows which precise methods were used for the case study with regard to the proposed model in Section 2.

3.1.1 Identify users of interest

For the case study we tried to focus on users who use the service steadily and do not have preferences for certain articles. The dataset suggested that users who exhibit this
behaviour tend to stay at the service for a longer amount of time compared to the average ones.

3.1.2 Decide on features

The UR chosen in this research uses the following three features.

- $f_1$: revenue
- $f_2$: number of orders
- $f_3$: number of distinct categories of orders

A category is counted if its revenue is above a margin based on the users’ highest category revenue. The weights and the margin were optimized to rate highly those users whose purchases value’ grow steadily over time. In addition, users, who utilize the service for business, were rated low.

3.1.3 Adjust weights by objective criteria

We decided to adjust the weights in a way that the users of interest tend to grow faster over time in comparison to an average user. Therefore the AverageAnnualGrowthRate as defined in Equation (4) was used, to show the average growth rate of a subset of users [19].

$$r_{AAGR} = \frac{1}{T - 1} \sum_{t=2}^{T} \frac{Y_t - Y_{t-1}}{Y_{t-1}}$$

The AverageAnnualGrowthRate was calculated for the average users, as well as for the top 1 percent users. The decision function is the difference between those two growth rates. In order to calculate the growth rate, all users have to be rated and grouped beforehand. This has to be done for multiple weight sets. Since the computation is expensive, the functions true optimal weight set was not calculated. A parameter sweep for the weights was performed, with the restriction that the sum of weights have to be 1 and each individual weight has to be bigger than 0.

For the used data set a high weight on the feature Number of distinct categories showed good results. Also the standard deviation of the annual growths for the top users was significantly lower for the top users in this area. This is shown in Table 2. The weight set used in the research was the one that had the highest growth rate, therefore $W_{opt} = \{0.15, 0.15, 0.7\}$.

Figure 2 and Figure 3 show the amount of money in JPY, spent over a period of 360 days. The colours represent the individual categories; the order of categories, from bottom to top, corresponds to the order of decreasing revenues in the whole population. The x-axis represents the time after a user creates an account. The y-axis represents the amount of money spent on average per user in JPY. Since the transaction data consist of 18 months and only long-life customers (user that exist for at least 12 months) were taken into account, the graph only shows a duration of 12 months. Otherwise there would be a declining part in the graph over the last 6 months. For example, only those few users that exist from the beginning of the dataset, and hence have 18 months of purchase histories, could possibly be in the 18th time bin, which is significantly less than the users presented in an earlier bin. The first few days of the purchase behaviour of each user have been excluded. Since most users purchase on the day they create an account, there would be a peak in the beginning of the graph. With our optimal weight set, we obtain a user separation, where the top users’ purchase history have a significantly higher growth rate than average users’, while also spending and ordering more than the average users.

3.1.4 Decide on a set of independent variables

The results presented here use a collection time of three months. The prediction results for one month were not convincing due to the fact that purchase frequency for most users was too low for meaningful observations during that time frame. Table 3 shows all independent variables that were used for the classification task. The check-boxes indicate if the variables were actually used in the final prediction or not. Since there is a total of 14 categories, the total amount of independent variables used was 37. Some of them were inspired upon investigating [20].

The data set used contained demographic data; however, the accumulated data seemed inconsistent and sometimes plain wrong. The data set contained a fair amount of users with an age of over 200 years and without specifying their gender. Since the customers were free to put any data into their accounts, the data is prone to bias through self-reporting [21]. Beel et. al [22] show the importance of using demographic data for recommendations. Even though we

| weight set | growth rate diff. | STD top1% |
|------------|-------------------|-----------|
| \{0.1,0.1,0.8\} | 0.0374 | 0.059 |
| \{0.15,0.15,0.7\} | 0.0385 | 0.06 |
| \{0.1,0.2,0.7\} | 0.0377 | 0.0581 |
| \{0.33,0.33,0.33\} | 0.0232 | 0.1952 |
| \{0.8,0.1,0.1\} | 0.0159 | 0.1766 |

Table 2: Growth rate calculation for different feature sets

Vol. 10 No.1 (2018)
understand the importance of demographic variables, we refrained from using them since we could not confirm the validity of the data.

The independent variables listed below were found to be the strongest predictors and are ordered by importance. The ranking of the variables was performed with a widely used score of the random forest framework. The importance of a variable is given by an increasing mean of the error of a tree in the forest, when the observed values of this variable are randomly permuted in the out-of-bag samples [23]. Another strong indicator has been the revenue and number of orders in the water, beverages, alcohol category. An average user tends to purchase more water than the top user. The e-commerce company of this study uses cheap products in this category to attract customers. However, those products itself are not profitable for the company.

1. $C$: number of distinct categories of orders
2. $R_R$: the ratio between the revenue from the top category and the revenue from all categories
3. $R_O$: the ratio between the number of products ordered from the top category and the number of products ordered from all categories

### 3.1.5 Evaluation of model results

The algorithms were evaluated with the precision-recall curve, and their AUC values were calculated. The interpolation of the points was done as proposed previously. Figure 4 shows the performance of the minority classes. The random forest outperforms the naive Bayes classifier throughout all recall values, however, only slightly. It is important to note that the AUC curves presented in the
figure do not contain the overall performance of the algorithm since their values lie at about 99% and fail to show differences between the algorithms.

The threshold for class prediction was adjusted so that recall will be 80 percent for the minority class. As Table 4 shows, 80 percent of the top users were identified, while maintaining a big number of irrelevant users. The weights for the naive Bayes model were adjusted to achieve a high recall value for the top rated users. With using a weight of 0.8 and 0.2 for the two classes respectively, the percentage of mislabelling was 7 percent.

The classification threshold for the random forest was chosen in the same manner, so that the recall value for the minor class is 80 percent. With this approach, a lower percentage of mislabelling was achieved, resulting in 5.2 percent as shown in Table 5.

4 Discussion and comparison to related research

One of the key contributions of this research is the UR discussed in Section 2 and how to predict it with machine learning algorithms. In this section, we discuss and compare our framework, especially the presented metric UR, to other already existing methods and point out the novelty and uniqueness of our approach.

4.1 RFM score

The recency, frequency and monetary (RFM) score has common aspects to UR. The RFM model has received interest from academia and the industry. In the field of direct marketing, they have been recognized as important in prediction of the future responses by customers to potential
### Table 3: Usage of independent variables per algorithm

| Description                  | Bayes | Forest |
|------------------------------|-------|--------|
| R               | ✔     |        |
| O               | ✔     | ✔      |
| C               | ✔     | ✔      |
| r_i             | ☐     | ☐      |
| o_i             | ☐     | ☐      |
| RF              | ☐     | ☐      |
| OF              | ☐     | ☐      |
| R_{RF}          | ☐     | ☐      |
| R_{OF}          | ☐     | ☐      |
| TC              | ☐     | ☐      |
| L               | ☐     | ☐      |

### Table 4: Naive Bayes results

|        | Precision | Recall | F1 Score |
|--------|-----------|--------|----------|
| Top 1  | 0.11      | 0.80   | 0.19     |
| Bottom 99 | 1.00   | 0.93   | 0.96     |
| Average| 0.99      | 0.93   | 0.96     |

### Table 5: Random forest results

|        | Precision | Recall | F1 Score |
|--------|-----------|--------|----------|
| Top 1  | 0.14      | 0.80   | 0.24     |
| Bottom 99 | 1.00   | 0.95   | 0.97     |
| Average| 0.99      | 0.95   | 0.97     |

Figure 4: Precision-recall curve for minority classes of both algorithms

The process of RFM is as follows: First, the data are sorted by each dimension and then the customers are divided in segments, usually five equal segments are used. Recency is commonly defined by the number of periods since the last purchase. The more recent the last purchase, the higher the score. Frequency stands for the number of purchases in a certain period. For monetary, the total amount of money spent during a period is counted. Based on their scores, customers can be grouped into segments [25].

One approach for a scoring method is hard coding. For each of the three variables, a weighted score for each person is applied. Then, a formula is used to sum the individual parts to receive an RFM score. Assigning the weights is usually a function of judgement of database marketers. Therefore, the method is often referred to as judgement based RFM [26][25].

While RFM uses its three variables to segment users, UR can be seen as a more general form. The variables used in the UR are not restricted to a certain amount of variables and type. UR, as presented in this research, uses multiplicative scoring, however additive scoring is also possible. If an additive scoring function is performed and the features used are recency, frequency and monetary, then, the UR would be almost equal to the RFM score. One remaining difference is the representation of the result, UR has a value between 0 and 1 where 1 is the theoretical best user of the data set and 0 the theoretical worst. One of the main advantages of the RFM model is that customers are not only segmented by the score but also by the cells of an $R \ast F \ast M$ matrix.

### 4.2 Customer lifetime value

The lifetime value of a customer (CLV) is the net of revenues from a single customer minus the costs of acquisition, selling and servicing the customer while taking into account the time value of money. It is a well recognized concept in research and business. CLV models provide a way to understand which customers to focus on [27].

It is common in direct marketing to rank customers on a particular metric. CLV is one of many customer selection strategies and Venkatesan and Kumar propose CLV as a metric to evaluate other customer selection strategies [28]. The comparison was done based on the profits in future periods of the selected customers.

Since CLV models try to predict the net income, the performance can be easily measured. However, UR tries to include other factors into the metric. Therefore, measuring performance based on net income is meaningful only if the purpose of the segmentation is maximizing the net income.
4.3 Customer potential value

Potential value is the profit or value by a customer, if the customer behaves ideally, meaning, the customers purchases all products or services he currently purchases in a market at full price at the business. Verhoef and Donkers propose a model to segment the customers based on the current value and the potential value into four categories [29].

The objective criterion in our case study, which takes into account the growth rates of the customers, focuses on users that have comparable low current value and a high value in the future. While the potential value, as defined by Verhoef and Donkers, takes into account external factors, our presented framework focuses on internal factors only [29].

4.4 Summary of novelty and uniqueness

RFM uses its three variables (recency, frequency, and monetary) to segment users, the UR in this study can be seen as a more general approach, as UR is not restricted to a certain amount of variables and type. UR is differentiated further from RFM and similar approaches by the use of multiplicative instead of additive scoring, and focusing on internal factors only. Other models, such as CLV, try to predict net income only. UR, however, includes other factors into the metric. UR takes into account the growth rates of customers and therefore focuses on users that have comparable low current value, but high value in the future.

5 Conclusion and future research

The goal of this research is to identify customers, who are likely to remain loyal to a business. In this paper, we demonstrated how a subset of “good” users can be selected in order to aim marketing funds at them. The performance of both algorithms show that the selected subset identifies most top 1 percent users, but contain a huge number of low ranked users as well. A correct classification of the minority class was attributed more value than the correct classification of the majority class, because of their purchasing behaviour.

UR is defined in a way that does not aim to maximize the net income, therefore there is no objective explanation why a certain set of features for UR should be better compared to another set. If maximizing net income is the main goal, other metrics are more appropriate to use. UR should be seen as a descriptive model with the ability to pursue other goals than profit maximization. Since UR does not have fixed variables, it arguably encourages businesses to expand predefined models, such as RFM, by other variables.

The customers of interest are usually a small fraction of the whole customer base. Therefore, further research on the class imbalance problem, especially in how to use metrics to evaluate the algorithms, could be conducted in the future.

Since misclassified users have similar purchase behaviour at the beginning as good ones, it can be argued that giving attention to those may help preventing future drop outs. However, this assumption may be hard to prove and should be further investigated. In addition, the framework could be extended so that it estimates the likelihood that a user will stop using the service. This could enable more advanced marketing strategies, using not only linear segmentation, but multi-dimensional segmentation.

ACKNOWLEDGMENTS

The authors would like to thank Marc-Andre Chavy-Macdonald, who reviewed this paper and contributed through continuous discussions. We would further like to express gratitude to Alexandra Ospanova for her proof-reading, which has profoundly improved the language of this paper.

References

[1] P. Kotler and G. Armstrong, Principles of marketing. Pearson Education, 2010.

[2] B. Nguyen and D. S. Mutum, “A review of customer relationship management: successes, advances, pitfalls and futures,” Business Process Management Journal, vol. 18, no. 3, pp. 400–419, 2012.

[3] K. W. Wong, S. Zhou, Q. Yang, and J. M. S. Yeung, “Mining customer value: From association rules to direct marketing,” Data Mining and Knowledge Discovery, vol. 11, no. 1, pp. 57–79, 2005.

[4] P. D. Berger and N. I. Nasr, “Customer lifetime value: Marketing models and applications,” Journal of interactive marketing, vol. 12, no. 1, pp. 17–30, 1998.

[5] C. Tofallis, “Add or multiply? a tutorial on ranking and choosing with multiple criteria,” INFORMS Transactions on Education, vol. 14, no. 3, pp. 109–119, 2014.
[6] J. Snyman, *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms*. Springer Science & Business Media, 2005, vol. 97.

[7] I. Bose and X. Chen, “Quantitative models for direct marketing: A review from systems perspective,” *European Journal of Operational Research*, vol. 195, no. 1, pp. 1–16, 2009.

[8] S. M. A. Elrahman and A. Abraham, “A review of class imbalance problem,” *Journal of Network and Innovative Computing*, vol. 1, pp. 332–340, 2013.

[9] R. Longadge and S. Dongre, “Class imbalance problem in data mining review,” *arXiv preprint arXiv:1305.1707*, 2013.

[10] I. Rish, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.

[11] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[12] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, “Random forests for classification in ecology,” *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[13] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, “Conditional variable importance for random forests,” *BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[14] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, “On the class imbalance problem,” in *Natural Computation, 2008. ICNC’08. Fourth International Conference on*, vol. 4. IEEE, 2008, pp. 192–201.

[15] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

[16] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[17] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[18] C. G. Weng and J. Poon, “A new evaluation measure for imbalanced datasets,” in *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*. Australian Computer Society, Inc., 2008, pp. 27–32.

[19] P. Mawson et al., “Measuring economic growth in new zealand,” New Zealand Treasury, Tech. Rep., 2002.

[20] B. Baesens, G. Verstraeten, D. Van den Poel, M. Egmont-Petersen, P. Van Kenhove, and J. Vanthienen, “Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers,” *European Journal of Operational Research*, vol. 156, no. 2, pp. 508–523, 2004.

[21] P. C. Kendall and J. S. Comer, “Research methods in clinical psychology,” *The Oxford Handbook of Clinical Psychology: Updated Edition*, p. 52, 2014.

[22] J. Beel, S. Langer, A. Nürnberg, and M. Genzmehr, “The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems,” in *Research and Advanced Technology for Digital Libraries*. Springer, 2013, pp. 396–400.

[23] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[24] D. L. Olson and B. Chae, “Direct marketing decision support through predictive customer response modeling,” *Decision Support Systems*, vol. 54, no. 1, pp. 443 – 451, 2012.

[25] J.-T. Wei, S.-Y. Lin, and H.-H. Wu, “A review of the application of rfm model,” *African Journal of Business Management*, vol. 4, no. 19, p. 4199, 2010.

[26] J.-T. Wei, S.-Y. Lin, and H.-H. Wu, “A review of the application of rfm model,” *African Journal of Business Management*, vol. 4, no. 19, p. 4199, 2010.

[27] J. A. McCarty and M. Hastak, “Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression,” *Journal of Business Research*, vol. 60, no. 6, pp. 656 – 662, 2007.

[28] R. Venkatesan and V. Kumar, “A customer lifetime value framework for customer selection and resource allocation strategy,” *Journal of Marketing*, vol. 68, no. 4, pp. 106–125, 2004.
[29] P. C. Verhoef and B. Donkers, “Predicting customer potential value an application in the insurance industry,” Decision support systems, vol. 32, no. 2, pp. 189–199, 2001.
