Journal of Business Research 66 (2013) 1629–1636

Contents lists available at SciVerse ScienceDirect

Journal of Business Research

Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning

Kristof Coussement [*], Koen W. De Bock

IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 8179), Expertise Center for Database Marketing (ECDM), Department of Marketing, 3 Rue de la Digue, F-59000, Lille, France

ARTICLE INFO

Article history:
Received 1 January 2012
Received in revised form 1 September 2012
Accepted 1 October 2012
Available online 10 January 2013

Keywords:
Customer relationship management
Online gambling
Customer churn prediction
Ensemble algorithms
GAMens
Random forests

ABSTRACT

The online gambling industry is one of the most revenue generating branches of the entertainment business, resulting in fierce competition and saturated markets. Therefore it is essential to efficiently retain gamblers. Churn prediction is a promising new alternative in customer relationship management (CRM) to analyze customer retention. It is the process of identifying gamblers with a high probability to leave the company based on their past behavior. This study investigates whether churn prediction is a valuable option in the CRM palette of the online gambling companies. Using real-life data of poker players at bwin, single algorithms, CART decision trees and generalized additive models are benchmarked to their ensemble counterparts, random forests and GAMens. The results show that churn prediction is a valuable strategy to identify and profile those customers at risk. Furthermore, the performance of the ensembles is more robust and better than the single models.

1. Introduction

Online gambling is an upcoming trend in today's society due to the digitalization and availability of new technologies, and its popularity is reflected by a tremendous amount of academic research studies (e.g. Smith, Levere, & Kurtzman, 2009). For instance, the online poker market has grown everywhere in the world, accounting for an estimated $3.6 billion in total revenues (Fiedler & Wilcke, 2011). Popular online gambling activities involve sports betting, virtual poker, and traditional casino games.

Previous research mainly focuses on the psychological facets of gambling behavior (Bleichrodt & Schmidt, 2002; Lam, 2006, 2007; McDaniel & Zuckerman, 2003; Mowen, Fang, & Scott, 2009), while other research papers investigate the determinants of individual differences amongst gamblers (e.g. Cronce and Corbin (2010)). Thus far, business and marketing literature approaches gambling from a purely exploratory insight perspective. In detail, former research gives insight into the individual betting behavior (Andrade & Iyer, 2009; Grant & Xie, 2007; Seybert & Bloomfield, 2009; Smith et al., 2009), the legitimation and development process of gambling and the impact on residents' perceptions (Humphreys, 2010; Roehl, 1999), the forecasting possibility of the outcomes of sporting events (Stekler, Sendor, & Verlander, 2010), the gambling agents' learning processes (Dana & Knetter, 1994), etc.

Although Oh and Hsu (2001) report that past gambling behavior impacts future gambling activities and that later on, several researchers tested the impact of past (aggregated) gambling behavior on future gambling behavior (Jolley, Mizerski, & Olaru, 2006; Lam, 2006; Lam & Mizerski, 2009; Mizerski, Miller, Mizerski, & Lam, 2004), no research study has explored the possibility of using real, individual past gamblers' behavior to predict future behavior to improve the customer relationship management (CRM) of the gambling company. However, gambling companies are searching new approaches to retain customers using their CRM strategy, because the rapid growth of the Internet does not facilitate the choice of the gambling website amongst the many options available (Jolley et al., 2006). Consequently, the CRM literature is expanding (Cooper, Gwin, & Wakefield, 2008; Ko, Kim, Kim, & Woo, 2008) and the fight to maintain the current players' base is a very important challenge within the gambling industry nowadays (Lal, Martone Carrolo, & Harrah's entertainment, 2001).

Since the concept of problem gambling is associated with the retention of gamblers, major gambling companies are actively approaching their players with targeted marketing actions to continue playing (Jolley et al., 2006; National Gambling Impact Study Commission, 1999; Productivity Commission, 1999). Indeed, in accordance with retailers and financial service providers, a lot of gambling companies realize that collecting customer data and storing the information into a customer database is an essential element in their retention strategy (Athanassopoulos, 2000).

Customer churn prediction is one of the key activities of a proactive retention strategy (Keaveney & Parthasarathy, 2001). Customer churn prediction is the process of assigning a probability of future churning

[*] Corresponding author.
E-mail addresses: k.coussement@ieseg.fr (K. Coussement), k.debock@ieseg.fr (K.W. De Bock).

0148-2963/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbusres.2012.12.008

behavior to each gambler in the database by building a prediction model based on past information. Intuitively, customers' past behavior is in line with their future gambling behavior. A good churn prediction model is a model that is able to assign high churn probabilities to real churners, and low churn probabilities otherwise. Practically, the marketing analyst uses the churn probabilities of the prediction model to rank the customers from most likely to leave the company to least likely to leave the company. The top X% of customers, containing on average more churners than the random model, are used for the company's targeted retention campaign.

Neslin, Gupta, Kamakura, Lu, and Mason (2006) indicate that numerous steps in the customer churn prediction process have an impact on its success. However, they strongly suggest focusing on the prediction technique due to its huge impact on the return on investment of subsequent marketing actions. As a result, academic literature on the optimization of churn prediction algorithms has exploded in the last few years. Prediction algorithms that are used in customer churn prediction literature include *single algorithms* like decision trees (Lemmens & Croux, 2006), generalized additive models (Coussement, Benoit, & Van den Poel, 2010), logistic regression (Smith, Willis, & Brooks, 2000) or support vector machines (Coussement & Van den Poel, 2008), etc., and *ensemble learners* that combine the predictions of multiple single algorithms (e.g. random forests (Coussement & Van den Poel, 2009)).

This paper introduces the concept of customer churn prediction in the gambling literature, whereas its value as an additional instrument in the CRM toolbox is investigated. Section 2 describes the predictive modeling methodology. Moreover, Section 3 digs into the churn prediction algorithms. Section 4 describes the evaluation metrics used, while Section 5 gives more insight into the real-life gambling dataset. Furthermore, the results are described in Section 6. Finally, the managerial implications for customer churn prediction and the use of ensemble methods are discussed in Section 7, while Section 8 concludes this research paper.

2. Predictive modeling methodology

The purpose of predictive modeling is to predict what is likely to happen in the future based upon what happened in the past (Blattberg, Kim, & Neslin, 2008). In the context of CRM, predictive modeling uses historical transactions and characteristics of a customer to predict future customer behavior.

In predictive modeling, two distinct phases are identified, i.e. a training phase and a prediction or test phase. In the training phase, a model that links future customer behavior to historical customer information is created using a training dataset. This dataset consists of input variables and a target variable for a range of customers. The input variables describe the customers' profiles and their past behavior in a certain time period, and usually include demographic characteristics and historical transactions. The latter category includes the popular RFM variables (McCarty & Hastak, 2007), i.e. the recency of customer purchases (R), the purchase frequency (F) and the monetary value of the historical transactions (M). Several studies show that RFM variables are amongst the strongest performing variables in explaining future customer behavior (e.g. Bose & Chen, 2009). The target variable reflects the behavior of interest to be predicted in a subsequent period of time. In the case of customer churn prediction, the target variable is represented by a binary variable representing whether or not a customer churned during the observation period. The purpose of training a predictive model is trying to capture the relationship between the input variables and the target variable.

In the prediction or test phase, the trained model is deployed on data not used during model training, i.e. the test set. Indeed, a test set of customers that are not included in the training process is used to assess the quality of the predictions output by the model. Measuring the performance on unseen data better reflects the true prediction performance, because measuring the performance on the training dataset often results in finding idiosyncratic patterns in the training dataset that do not hold up in a real-life test setting. As such, the original dataset is randomly split into a training set and a test set. In order to disclose the true capabilities of a prediction algorithm, it is good practice to repeat the training phase multiple times in order to reduce the risk of erroneously identifying a particular model as a good model (Malthouse, 2001). This objective is pursued in the cross-validation methodology as suggested by Witten and Frank (2000) and already implemented in a variety of predictive modeling settings in marketing (e.g. Cui, Wong, & Lui, 2006). In this study, results are based upon a five times twofold cross-validation procedure (5×2-fold cv). This involves five replications of a twofold cross-validation. In each replication, customers of the original dataset are randomly assigned to one of two equally-sized cross-validation parts. One part is once used as training data to build the prediction model, while the performance is calculated for the other part acting as a test dataset. This process is then repeated, switching the roles of the two cross-validation parts. Furthermore before putting a prediction model into practice, it is important to compare its performance to alternative types of prediction models in order to find the best algorithm (Neslin et al., 2006). As such, Section 3 describes the different prediction algorithms used in this research study.

3. Churn prediction algorithms

3.1. Single algorithms

3.1.1. CART decision tree

Decision trees are popular techniques in predictive modeling due to their simplicity and transparency (Duda, Hart, & Stork, 2001). For instance, Murthy (1998) found more than 300 academic references that use decision trees in a variety of settings. A decision tree is composed of a set of rules that divide a dataset that is heterogeneous in terms of the target variable into smaller, more homogeneous sets. The technique is best explained by referring to its prediction phase. To determine the churn probability of a customer in the test dataset, the customer enters the decision tree at the start or root node at the top of the decision tree. This node represents a test that attributes the customer to one of the lower-level or child nodes. The test is a logical question formulated in terms of the input variable, chosen in such a way that it discriminates maximally between churners and non-churners. The result of the test determines which child node is chosen for the customer. Tests at subsequent child nodes redirect the customer through the decision tree until a terminal or leaf node is reached. The collection of unique paths between the root node and the leaf nodes makes up the rules used to generate predictions. All customers from the test set reaching the same leaf node receive the same prediction probability, calculated as the percentage of churners from all training customers in that leaf node.

This study employs the CART algorithm, acronym for Classification and Regression Trees, because it is a popular decision-tree algorithm using the *Gini coefficient* as splitting criterion (Breiman, Friedman, Olsen, & Stone, 1984). It implements binary recursive partitioning; it uses *two-way splits* that split a node into exactly two child nodes, while its growing process continues recursively until splits are no longer possible because all customers in a node are identical in terms of the target variable or the input variables. Finally, the decision-tree pruning that aims at reducing the decision-tree complexity is necessary, because decision trees are susceptible to overfitting. In CART, pruning is done via the *maximum performance objective criterion* (Breiman et al., 1984; Steinberg & Colla, 1998).

3.1.2. Generalized additive model

The most popular technique among predictive modeling techniques that have been applied to the prediction of customer churn is incontestably logistic regression (Buckinx & Van den Poel, 2005). Its popularity is mainly attributed to its ability to combine the simplicity of a linear model with favorable performance, and its widespread availability in many statistical and data-mining software environments like SAS, IBM SPSS Modeler, R, WEKA, etc. However in several studies (e.g. Coussement et al., 2010), it has been shown how more advanced prediction techniques like generalized additive models (GAM) significantly outperform logistic regression. A particular limitation of logistic regression is that it assumes that the functional form of the relationship between the (logit-transformed) target variable and the input variables is known and linear.

A noteworthy category of nonparametric techniques that overcomes these drawbacks are GAM (Hastie & Tibshirani, 1986). GAM determine that the functional form of the prediction model depend on the training data itself resulting in a complex non-linear relationship between the input variables and the target variable.

Statistically speaking, GAM are more flexible than a logistic regression as the influence of an input variable on the target variable is no longer subject to any linear parametric specification (Hastie & Tibshirani, 1986, 1990). It is fit using an arbitrary nonparametric function, called a smooth function. Therefore, GAM replace the linear combination in logistic regression $\sum_{k=1}^{p} \beta_k X_k$ by the additive form $\sum_{k=1}^{p} f_k(X_k)$, where each partial function f_k is an unspecified smooth function. In this study, GAM are configured as logistic, semi-parametric additive models that accommodate for the inclusion of dummy-coded input variables (such as gender). The GAM model is represented as follows via

$$\operatorname{logit}(P(Y=1|X)) \equiv \log\left\{\frac{P(Y=1|X)}{1-P(Y=1|X)}\right\} = \sum_{j=1}^{p_c} s_j\left(X_j\right) + \sum_{k=1}^{p_d} \beta_k X_k \qquad (1)$$

where input variables $X_j, j=1,\ldots,p_c$ are continuous variables, $X_k=1,\ldots,p_d$ are dummy-coded variables and the smooth functions $s_1(X_1)$, $s_2(X_2),\ldots,s_{p_c}\left(X_{p_c}\right)$ are smoothing splines that estimate the nonparametric trend for the dependence of the logit on $X_1, X_2, \ldots X_{p_c}$. More information on GAM and smoothing splines is found in Hastie and Tibshirani (1990).

3.2. Ensemble learners

In recent years, the practice of combining predictions from single algorithms has become a popular topic in theoretical and applied research (van Wezel & Potharst, 2007). The predictions of ensemble learners are taken as combinations of the individual ensemble member probabilities (Kuncheva & Rodríguez, 2007). The main factor defining the popularity of ensemble algorithms is the strong prediction performance that is observed within multiple comparative studies in various domains and applications (e.g. Bauer & Kohavi, 1999; Dietterich, 2000). An ensemble of individual prediction models is likely to generate better and more robust predictions than a single algorithm when accuracy and diversity are present amongst the ensemble members.

Many well-known ensemble learners are proposed that are classified along three dimensions (De Bock, Coussement, & Van den Poel, 2010): (i) the manipulation of the original training dataset; (ii) the choice of the single algorithm for the ensemble members; and (iii) the combination rule used to aggregate the ensemble member predictions. The first dimension is the strategy in which different training datasets are created for the ensemble members. Instead of using the original training dataset for each ensemble member, some transformation is applied to create different versions of the original training set for each ensemble member. Two notorious strategies that are applied to manipulate the training data are Bagging (Breiman, 1996) and the Random Subspace Method (RSM) (Ho, 1998). Bagging, acronym for Bootstrap Aggregating, prescribes the use of bootstrap samples of the original customers taken with replacement and of equal size as the original training dataset, as new training data for each ensemble member. RSM, the second strategy, randomly picks a predetermined number of variables from the original training dataset in an iterative way to create new datasets for training. Both strategies, sampling and variable selection are used to increase diversity in ensemble learners.

A second dimension in the classification of the ensemble learners is the choice of the single algorithm for use in the ensemble members. This study uses single algorithms, CART and GAM, as the members for the ensemble algorithms. The combination of Bagging, RSM and the use of CART decision trees in the construction of an ensemble learner is known as the random forests (RF) ensemble algorithm (Breiman, 2001). RF is known to perform very well in many domains including predictive modeling settings in CRM (Coussement & Van den Poel, 2008). When GAM are considered as ensemble members in ensemble algorithms in combination with Bagging and RSM, the ensemble learner is called GAMens (De Bock et al., 2010). Their experiments prove that GAMens is very competitive with the popular RF algorithm. In sum, the strong performance of ensemble algorithms RF and GAMens in previous research is the motivation to include both ensemble techniques in this study.

Finally, a combination rule to aggregate the individual ensemble member predictions is chosen. This rule defines how the predictions of ensemble members are combined into an aggregated prediction for a particular customer in the test dataset. If probability estimates are desired, the average of the ensemble member predictions is calculated to obtain an ensemble prediction for a customer in the test dataset (Kuncheva, 2004).

4. Evaluation metrics

To evaluate the quality of the predictions of the churn prediction models, two evaluation metrics are considered, i.e. the top-decile lift (TDL) and the lift index (LI). Both are popular choices in academic literature on customer churn prediction because of their relevance for business (Crone, Lessmann, & Stahlbock, 2006; Lemmens & Croux, 2006). Indeed, these evaluation metrics measure the ability of a prediction model to produce a good ranking of the customers based upon their predicted churn probabilities. In detail, a good prediction model is able to assign higher churn probabilities to real churners.

TDL focuses on the customers indicated by the prediction algorithm as being most risky to leave the company (Lemmens & Croux, 2006). Practically, the customers are first sorted from predicted most likely to churn to predicted least likely to churn based on the churn probabilities obtained by the prediction model. Afterwards, the proportion of churners in the top ten percent most likely to churn is compared with the proportion of churners in the total dataset. This increase in churn density is called the TDL. For example, a TDL of two means that the density of churners in the top 10% is twice the density of churners in the total dataset. The higher the TDL, the better the prediction algorithm is. A TDL higher than one identifies a model that outperforms a random selection of customers and that has heterogeneity with respect to the churn probabilities. This top decile of customers is very interesting to target via a customer retention campaign, because it contains a higher proportion of churners.

LI is the second evaluation metric considered in this study (Crone et al., 2006). LI generalizes the TDL measure, because it does not require any specification of an arbitrary percentage to calculate the increase in churn density. Assuming that L is a list of customers that are sorted based on their predicted churn probabilities. The formula for LI is

$$LI = \frac{(1.0 \cdot L_1 + 0.9 \cdot L_2 + 0.8 \cdot L_3 + \ldots + 0.1 \cdot L_{10})}{\sum_{i=1}^{10} L_i} \qquad (2)$$

with L_i the number of actual churning customers in the ith decile of L. The LI for a random model takes a value of 0.5. That means that any value higher than 0.5 indicates a better than random prediction.

5. Dataset

This study uses a dataset on actual gambling behavior delivered by bwin Interactive Entertainment, an online gambling operator and recently used in other gambling related studies (e.g. Braverman and Shaffer (2010) or LaBrie and Shaffer (2011)). The dataset is made freely available through the website of the Cambridge Health Alliance. The data includes two years of recorded Internet betting activity by a cohort of gamblers who opened for the first time and account with bwin during February 2005. The sample includes 3729 gamblers who played at least four times on three different dates in the input variables creation period. Fig. 1 shows the time-line used within this research study.

The time period in which the input variables or churn predictors are created is given in black, while the time span in which one observes whether a gambler does (not) come back to the website to gamble is represented in white. The in-sample dataset contains 2486 players, while an out-of-period (OOP) test dataset is considered containing 1243 players. The in-sample dataset is used for the 5×2-fold cv. In detail, ten prediction models are built during the training phase of the 5×2-fold cv, which are then validated on the in-sample test data of the 5×2-fold cv and the OOP test dataset. The OOP test dataset is added to the experiment, because a validation on unseen test data from another time frame is a more difficult exercise. A time window of 17 months is used to collect and construct the churn predictors for both the in-sample dataset and the OOP test dataset.

Furthermore, a four month period is considered to measure the churn behavior for each and every gambler in the datasets. In other words, a gambler is considered a churner when not having played during this four month period and vice versa. The churn incidence or the number of gamblers that stopped betting in the four month period equals to 30.64% for the in-sample dataset and 23.74% for the OOP test set. 60 churn predictors or gamblers' characteristics of the dataset are constructed and they are divided into two broad categories: behavioral information and demographic information. Appendix 1 gives an overview of all gamblers' information available to predict customers' churn.

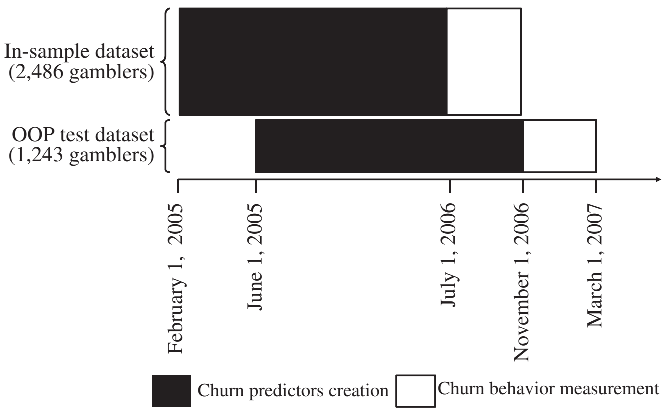

Fig. 1. Graphical display of the time window.

6. Results

6.1. Classification performance

The results in Tables 1 and 2 follow from using the following experimental parameters. Per fold in the 5×2-fold cv, the RF and GAMens ensemble algorithms aggregate predictions from 1000 ensemble members. In other words, 1000 single prediction models are built whereby each is trained on a bootstrap sample from the original training dataset (Bagging), while Breiman's (2001) suggestion is followed to randomly select the square root of the total number of variables for each ensemble member model (RSM). More specifically, a random selection of the square root of 60 or eight variables for each ensemble member is considered.

Table 1 provides a performance overview of the churn prediction algorithms for the 5×2-fold cv. The results on the in-sample test set (TEST) and the OOP test set are reported. In detail, the performances of the individual cross-validation runs are given, i.e. Run 1 till Run 10, the averaged performance over the ten runs (Average) and the standard error of the performance measure over the ten runs (Standard Error) are considered for TDL and LI.

A cell in Table 2 gives the average performance of an algorithm over the 5×2-fold cv. The significance between two algorithms is obtained by applying the nonparametric Wilcoxon-signed rank test which ranks the differences in performance over the ten cross-validation runs, ignoring the signs, and comparing the ranks for the positive and the negative differences (Demšar, 2006; Wilcoxon, 1945). The Bonferroni correction is applied to address the problem of multiple comparisons, six in this setting, and to take into account the family-wise error rate (Abdi, 2007). In any row, performance measures that do share a common subscript are not significantly different at the Bonferroni-corrected p value of 0.05/6.

Table 1 reveals that churn prediction in this gambling setting is highly beneficial and worthwhile because the predictive performance measures substantially exceed the thresholds of the respective random models. In detail, the TDL measures exceed one, while the LI of the models overstep the threshold of 0.5. It is clear from Table 1 that churn prediction is a valuable approach in the global CRM strategy of a gambling company.

Table 1 provides insight into the stability of the algorithms over the cross-validation runs by means of the standard error. In particular, Table 1 shows that the single algorithms, CART and GAM, are more sensitive to small changes in the training set than their ensemble counterparts, RF and GAMens. The standard errors for TDL and LI are systematically higher for CART and GAM than for RF and GAMens in both test cases, i.e. TEST and OOP.

Table 2 reveals the beneficial effect of ensemble algorithms over the corresponding single models. It is clear from Table 2 that GAMens has a significantly better performance than GAM, while RF performs substantially better than a single CART for TDL and LI in both test situations. The added value of CART over GAM or vice versa is not confirmed, while Table 2 confirms that the ensemble models, RF and GAMens, are very competitive, and thus their performances are not statistically different in this gambling setting.

In general, one can conclude that churn prediction is a very valuable tool to the traditional CRM palette in the gambling industry. The following section explores the churn predictors' importance of the strongest algorithms, i.e. random forests and GAMens.

6.2. Churn indicators importance

Next to the ability to generate accurate predictions, churn prediction should deliver insight into the drivers of customer churn. Hence, comprehensibility is an important characteristic of customer churn prediction models (Masand, Datta, Mani, & Li, 1999; Verbeke, Martens, Mues, & Baesens, 2011). However, while generally recognized for their

Table 1
Performance overview of the churn prediction algorithms for the 5×2-fold CV.

			5×2-fold CV Runs									
Evaluation metric	Set	Algorithm	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average	Standard Error
TLD	Test	CART	2.1333	2.2101	2.2123	2.3679	2.0279	2.3679	2.6864	2.1837	2.1860	2.0785	2.2454	0.1794
		RF	2.8181	2.7363	2.8444	2.8152	2.7390	2.7100	2.7917	2.7926	2.8971	2.8678	2.8245	0.0581
		GAM	2.3440	2.4732	2.4757	2.5784	2.4493	2.5784	2.5547	1.8943	2.6600	2.5784	2.4586	0.2062
		GAMens	2.8707	2.7363	2.7363	2.7390	2.9204	2.7390	2.7917	2.7626	2.7917	2.7626	2.7824	0.0624
	OOP	CART	2.7184	3.0922	2.9903	3.1941	2.4126	2.9563	3.3980	2.9563	3.0582	2.8543	2.9631	0.2537
		RF	3.3641	3.5000	3.6359	3.3641	3.3980	3.5679	3.4660	3.3641	3.5340	3.5000	3.4694	0.0905
		GAM	2.7184	2.8543	3.0242	2.8543	2.8543	2.9563	2.9223	1.8349	2.7864	2.7184	2.7524	0.3195
		GAMens	3.3641	3.3301	3.4660	3.3301	3.4660	3.3980	3.2961	3.5000	3.4660	3.3980	3.4014	0.0670
LI	TEST	CART	0.7024	0.7213	0.6938	0.7179	0.7132	0.7159	0.7233	0.7171	0.6894	0.7089	0.7103	0.0110
		RF	0.7593	0.7533	0.7558	0.7504	0.7577	0.7629	0.7585	0.7614	0.7613	0.7538	0.7574	0.0039
		GAM	0.7071	0.7131	0.7376	0.7268	0.7053	0.7394	0.7303	0.6701	0.7200	0.7289	0.7178	0.0195
		GAMens	0.7526	0.7549	0.7597	0.7470	0.7500	0.7530	0.7524	0.7512	0.7603	0.7488	0.7530	0.0041
	OOP	CART	0.7486	0.7227	0.7101	0.7247	0.7554	0.6973	0.7453	0.7057	0.7215	0.7292	0.7260	0.0181
		RF	0.7787	0.7747	0.7799	0.7739	0.7847	0.7746	0.7750	0.7768	0.7824	0.7746	0.7775	0.0036
		GAM	0.7132	0.6881	0.7614	0.7136	0.7403	0.7315	0.7214	0.5823	0.7281	0.6908	0.7071	0.0464
		GAMens	0.7797	0.7739	0.7814	0.7698	0.7844	0.7719	0.7742	0.7810	0.7803	0.7749	0.7772	0.0046

strong classification performance, ensemble methods are sometimes reproached their black box nature, i.e., their reduced ability to deliver insights due to an increased model complexity. This study shows that this need not be the case by reporting variable importance scores, a by-product of the random forest algorithm that allows the analyst to measure the relative importance of each variable in the data set (Breiman, 2001).

This study concentrates on permutation-based variable importance scores, generally accepted as the most sophisticated and desirable variation (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). The variable importance score of a particular variable is obtained by, for every member in the ensemble, calculating the difference in predictive accuracy on the out-of-bag data, and the predictive accuracy after randomly permuting the variable's values within the same data set. The out-of-bag data is simply the set of customers that were not selected in the bootstrap sample used as training data for that respective ensemble member. The global importance score is then the average difference in prediction performance, calculated over all the members in the ensemble (Breiman, 2001).

In Table 3, both for random forest and GAMens models, the 20 most important customer churn drivers are identified based on permutation-based variable importance scores. To offer additional insight into how churners and non-churners differ in terms of these characteristics, average values are provided for both the churners and non-churners, and these are compared using significance tests.

The following observations emerge from Table 3. First, overall, random forests and GAMens report the same variables as most important predictors for customer churn. Only 3 out of 20 variables for the random forest model do not emerge in the GAMens ranking, and vice versa. This high level of agreement is confirmed when considering the Pearson correlation coefficient between the importance scores for both techniques which mounts up to 90.86%. Second, the rankings clearly identify so-called RFM variables (recency, frequency and monetary value) as the major variables to predict customer churn. With some exceptions, overall, recency-related variables appear at the top of both rankings and seem to outperform variables related to frequency and, finally, variables expressing monetary value of past transactions. Differences between random forests and GAMens most clearly relate to variables related to the length of relationship which emerge at the top for random forest and appear to be relatively less important in the GAMens model. Third, a comparison of average feature values between churners and non-churners reveals significant differences for most variables and offers a simple method to get insight into the relationship between a variable and churn behavior.

Table 2
Summary table of the average performance measures over the 5x2-fold cv.

Evaluation metric	Set	Algorithm			
		CART	RF	GAM	GAMens
TDL	TEST	2.2454_a	2.8245_b	2.4586_a	2.782_b
TDL	OOP	2.9631_a	3.4694_b	2.7524_a	3.4014_b
LI	TEST	0.7109_a	0.7574_b	0.7178_a	0.7530_b
LI	OOP	0.7260_a	0.7775_b	0.7071_a	0.7772_b

Note: in any row, performance measures that do share a common subscript are not significantly different at $p<0.05/6$.

7. Managerial implications

The results prove that employing a churn prediction model to identify those customers at risk of stopping gambling is more beneficial than randomly selecting customers out of the database for targeting purposes. The search towards more advanced algorithms by means of ensemble algorithms to optimize the churn prediction performance is a very meticulous task that could have a huge impact on the bottom-line profit of the company (Van den Poel & Larivière, 2004). Based on the formula of Neslin et al. (2006), the gain in profitability between two classification techniques for a change in TDL is defined as follows

$$GAIN = N\alpha\{[\gamma CLV + \ \delta(1-\gamma)]\beta_0\}\Delta_{TDL} \tag{3}$$

with N the total number of customers, α the percentage of the customer base contacted, β_0 the base-line churn rate, δ the cost of the customer incentive to the firm, γ the fraction of targeted would-be churners who decide to remain because of the incentive (i.e., the success rate of the incentive), CLV the customer lifetime value (i.e., the value to the firm if the customer is retained) and Δ_{TDL} the difference in top decile lift between two classifiers. Based on the real-life bwin Interactive Entertainment dataset, and taking into account the different between CART versus RF and GAM versus GAMens in terms of TDL on the OOP, Table 4 summarizes the results with different γ and CLV, a cost of 50

Table 3
20 most important variables for random forests and GAMens based on variable importance scores and average variable value comparison between churners and non-churners.

	Ramdom forest				GAMens			
Rank	Variable	Variable important score	Average feature value non-churners	Average feature value churners	Variable	Variable important score	Average feature value non-churners	Average feature value churners
1	Recency	0.0422	94.57	249.42**	Recency	0.0201	94.57	259.42**
2	Recency_lastloss	0.0327	109.34	271.47**	Recency_lastloss	0.0149	109.34	271.47**
3	IPT_std	0.0162	39.87	36.23**	Frequency_sessions_lastmonth	0.0115	2.42	0.2**
4	rFrequency_sessions	0.0161	0.1	0.04*	rFrequency_sessions	0.0083	0.1	0.04**
5	IPT_max	0.0145	129.56	102.11**	**Frequency_sessions_lastweek**	0.0067	0.55	0.02**
6	Frequency_sessions	0.0138	34.63	15.2**	Frequency_sessions	0.0058	34.63	15.2**
7	Frequency_bets_lastmonth	0.0132	284.26	10.59**	Frequency_bets_lastmonth	0.0057	284.86	10.59
8	Recency_lastwin	0.0130	123.32	259.21**	Frequency_sessions_lost	0.0053	22.55	9.85**
9	**IPT_mean**	0.0126	23.64	21.72	Recency_lastwin	0.0048	123.32	259.21**
10	Frequency_sessions_lastmonth	0.0118	2.42	0.2**	**Frequency_bets_lastweek**	0.0042	56.77	0.91**
11	Frequency_sessions_lost	0.0113	22.55	9.85**	rMonetary_stakes	0.0037	107.6	38.23**
12	rMonetary_stakes	0.101	107.6	38.23**	rMonetary_winnings	0.0026	104.5	37.02**
13	Monetary_stakes	0.0098	41887.13	16592.26**	rMonetary_losses	0.0032	6.25	2.24**
14	rMonetary_winnings	0.0095	104.5	37.02	IPT_max	0.0030	129.56	102.11**
15	Monetary_winnings	0.0092	40694.26	16046.49	**Frequency_sessions_won**	0.0030	9.86	4.23**
16	rMonetary_winnings	0.0095	107.5	37.02**	Monetary_stakes	0.0026	41887.13	16592.96**
17	**Monetary_winnings**	0.0092	40694.26	16046.49**	Monetary_winnings	0.0024	40694.26	16046.49**
18	Monetary_losses	0.0090	6.25	2.24**	IPT_std	0.0024	39.87	36.23
19	**Frequency_bets**	0.0071	4926.35	2062.985**	rFrequency_bets	0.0021	13.5	4.85**
20	Lor_firstplay	0.0055	395.14	437.8**	Lor_firstplay	0.0019	395.14	437.8**

Note 1: Statiscally significant differences between churnes and non-churnes are indicated ($p<0.1$(*) & $p<0.05$(**)) based on paired T-tests for the comparison of means.
Note 2: Variables in bold appear uniquely for the respective technique.

Euros (δ), β_0 equal to 23.74% the base-line churn rate in the OOP and by setting $N\alpha = 1$ to put the gain on the customer level.

Based on the results of Table 4 and knowing that gambling databases contain a large number of customers; this study encourages marketing analysts to focus on ensemble algorithms for customer churn prediction in order to optimize the identification of those customers at risk, and accordingly the bottom-line contribution per customer.

8. Conclusions

Nowadays gambling companies are searching for new alternative ways of building and retaining customer relationships. Churn prediction modeling is one of these alternatives and it ranks customers from most to least likely to churn based on the allocation of a churn probability to each customer in the customer database. Based on a real-life online gambling dataset obtained from bwin Interactive Entertainment, this study confirms that customer churn prediction is a valuable approach in the CRM portfolio of a gambling company.

The study shows the beneficial impact of ensemble algorithms over single prediction models in this customer churn prediction setting. This study highlights the advantages of the ensembles RF and GAMens, that is, robust and better prediction performance, than the single algorithms, CART and GAM. Although the performance of a prediction algorithm is the most important criterion to decide which algorithm to apply in a real-life setting, it is seldom considered as the only decision criterion. Three other criteria are placed in the balance when putting prediction algorithms into practice: the required expertise of the analyst, the running time of the algorithm and the availability in standard software packages. Firstly, running the ensemble learners RF and GAMens requires additional and specialized knowledge from the analyst in order to fully understand the functioning of the ensemble algorithms.

Additional training is necessary and costly. Secondly, the ensemble learners have, on average, a larger running time than the single models which is an additional element to consider when employing ensemble learners. Finally, ensemble learners are freely available in open source software environments like R or WEKA. However, an additional effort is required to use the ensemble toolbox and to integrate it with the existing predictive modeling software package used by the company.

Several avenues for further research are given. First, the results of this research should be extended to other types of gambling environments offering virtual poker tables. Our results are obtained by using data from bwin, in nature a sports betting operator. The dataset most probably contains players whose primary interest is sports gambling, while they probably choose poker as an additional game. Building a churn prediction model on data that includes customers who first shifted towards online gambling/sports betting in general would be interesting. Finally, churn prediction is only one side of the coin. Additional efforts could be spent to combine the event of churning with the profitability of the customers in order to truly identify high value churners.

Table 4
Comparison of profitability gains per customer (in Euro) by using ensembles over single algorithms.

Algorithm	TDL	P_{diff}	CLV		γ	
				0.1	0.3	0.5
CART	2.9631		300	9.0147	15.0245	21.0342
RF	3.494	0.5063	600	12.6205	25.8421	39.0636
			1200	19.8323	74.4773	75.1223
GAM	2.7524		300	11.5554	19.2591	26.9627
GAMens	3.4014	0.6490	600	16.1776	33.1256	50.0736
			1200	25.4220	60.8587	96.2954

Acknowledgments

The authors thank "bwin" and the Cambridge Health Alliance for freely distributing the data, the editors for handling this paper through the reviewing process, and the reviewers for their insightful comments.

Appendix 1. Overview of gamblers' characteristics

Variable type	Variable subtype	Variable name	Description
Behavioral	Recency	Recency	Number of days since last betting acti
		Recency_lastwin	Number of days since last netto win
		Recency_lastloss	Number of days since last netto loss
	Interpurchase time	IPT_mean	Average number of days between online betting sessions
		IPT_std	Standard deviation of the number of days between betting sessions
		IPT_min	Minimum number of days between betting sessions
		IPT_max	Maximum number of days between betting sessions
		IPT_CV	Coefficient of variation of interpurchase time (ratio of IPT_std to IPT_mean)
	Length of Relationship	Lor_registration	Number of days since user registration
		Lor_firstpay	Number of days since first money deposit
		Lor_firstplay	Number of days since first betting
	Frequency	Frequency_bets	Number of bets Frequency_sessions Number of betting sessions
		Freq_bets_lastweek	Number of bets during the last week
		Freq_bets_lastmonth	Number of bets during the last month
		Freq_sessions_lastweek	Number of betting sessions during the last week
		Freq_sessions_lastmonth	Number of betting sessions during the last month
		Freq_sessions_won	Number of betting sessions with a positive netto betting result
		Freq_sessions_lost	Number of betting sessions with a negative netto betting result
		rFrequency_bets	Number of bets relative to the length of relationship (Lor_firstplay)
		rFrequency_sessions	Number of betting sessions relative to the length of relationship (Lor_firstplay)
	Monetary	Montary_stakes	Total monetary amount of stakes
		Monetary_wins	Total monetary amount of netto wins
		Monetary_losses	Total monetary amount of netto losses
		rMonetary_stakes	Total monetary amount of stakes relative to the length of relatinoship (Lor_firstplay)
		rMonetary_wins	Total monetary amount of netto wins relative to the length of relatinoship (Lor_firstplay)
		rMonetary_losses	Total monetary amount of netto losses relative to the length of relatinoship (Lor_firstplay)
	Promotion	d_promofunds	Did the customer receive promotion playing funds?
Demographic	Language	d_[language]	Dummy indicating customer language; language = {Catalan Croatian Czech Danish English French German Greek Hungarian Italian Norwegian Polish Portuguese Russian Slovak Slovenian Spanish Swedish Turkish}
	Region	d_[region]	Dummy indicating region of origin = {Australia_and_New_Zealand Central_Asia Eastern_Asia Eastern_Europe Northern_America Northern_Europe South_America Southern_Africa Southern_Asia Southern_Europe Western_Asia Western_Europe}
	Gender	d_gender	Dummy indicating gender (0 = female, 1 = male)

Note: a session is defined as a day on which one or more bets are made.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage.

Andrade, E. B., & Iyer, G. (2009). Planned versus actual betting in sequential gambles. *Journal of Marketing Research*, *46*(3), 372–383.

Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research*, *47*(3), 191–207.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, *36*(1–2), 105–139.

Blattberg, R. C., Kim, B. -D., & Neslin, S. A. (2008). *Database marketing: Analyzing and managing customers*. New York (NY): Springer.

Bleichrodt, H., & Schmidt, U. (2002). A context-dependent model of the gambling effect. *Management Science*, *48*(6), 802–812.

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16.

Braverman, J., & Shaffer, H. J. (2010). How do gamblers start gambling: Identifying behavioral markers for high-risk Internet gambling. *European Journal of Public Health*, *22*(2), 273–278.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olsen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. : Chapman & Hall/CRC.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviorally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252–268.

Cooper, M. J., Gwin, C. F., & Wakefield, K. L. (2008). Cross-functional interface and disruption in CRM projects: Is marketing from Venus and information systems from Mars? *Journal of Business Research*, *61*(4), 292–299.

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, *37*(3), 2132–2143.

Coussement, K., & Van den Poel, D. (2008). 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.

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, *36*(3), 6127–6134.

Cronce, J. M., & Corbin, W. R. (2010). Effects of alcohol and initial gambling outcomes on within-session gambling behavior. *Experimental and Clinical Psychopharmacology*, *18*(2), 145–157.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781–800.

Cui, G., Wong, M. L., & Lui, H. -K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612.

Dana, J. D., & Knetter, M. M. (1994). Learning and efficiency in a gambling market. *Management Science*, *40*(10), 1317–1328.

De Bock, K. W., Coussement, K., & Van den Poel, D. (2010). Ensemble classification based on generalized additive models. *Computational Statistics & Data Analysis*, *54*(6), 1535–1546.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157.

Duda, R. O., Hart, E., & Stork, D. G. (2001). *Pattern classificaiton*. New York, NY: John Wiley & Sons.

Fiedler, I., & Wilcke, A. -C. (2011). *Der Markt für Onlinepoker: Spielerherkunft und Spielerverhalten*. Norderstedt: BoD.

Grant, S. J., & Xie, Y. (2007). Hedging your bets and assessing the outcome. *Journal of Marketing Research*, *44*(3), 516–524.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–318.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman and Hall.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.

Humphreys, A. (2010). Megamarketing: The creation of markets as a social process. *Journal of Marketing*, *74*(2), 1–19.

Jolley, B., Mizerski, R., & Olaru, D. (2006). How habit and satisfaction affects player retention for online gambling. *Journal of Business Research*, *59*(6), 770–777.

Keaveney, S. M., & Parthasarathy, M. (2001). Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors. *Journal of the Academy of Marketing Science*, *29*(4), 374–390.

Ko, E., Kim, S. H., Kim, M., & Woo, J. Y. (2008). Organizational characteristics and the CRM adoption process. *Journal of Business Research*, *61*(1), 65–74.

Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms.* Hoboken, New Jersey: John Wiley & Sons.

Kuncheva, L. I., & Rodríguez, J. J. (2007). Classifier ensembles with a random linear oracle. *IEEE Transactions on Knowledge and Data Engineering*, *19*(4), 500–508.

LaBrie, R. A., & Shaffer, H. J. (2011). Identifying behavioral markers of disordered Internet sports gambling. *Addiction Research & Theory*, *19*(1), 56–65.

Lal, R., Martone Carrolo, P., & Harrah's entertainment (2001). *Inc. Harvard Business School Publishing Case Series*. (pp. 1–16).

Lam, D. (2006). The influence of religiosity on gambling participation. *Journal of Gambling Studies*, *22*(3), 305–320.

Lam, D. (2007). An exploratory study of gambling motivations and their impact on the purchase frequencies of various gambling products. *Psychology and Marketing*, *24*(9), 815–827.

Lam, D., & Mizerski, R. (2009). An investigation into gambling purchases using the NBD and NBD-Dirichlet models. *Marketing Letters*, *20*(3), 263–276.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*(2), 276–286.

Malthouse, E. C. (2001). Assessing the performance of direct marketing scoring models. *Journal of Interactive Marketing*, *15*(1), 49–62.

Masand, B., Datta, P., Mani, D., & Li, B. (1999). CHAMP: A prototype for automated cellular churn prediction. *Data Mining and Knowledge Discovery*, *3*(2), 219–225.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656–662.

McDaniel, S. R., & Zuckerman, M. (2003). The relationship of impulsive sensation seeking and gender to interest and participation in gambling activities. *Personality and Individual Differences*, *35*(6), 1385–1400.

Mizerski, R. W., Miller, R., Mizerski, K., & Lam, D. (2004). The stochastic nature of purchasing a state's lottery products. *Australasian Marketing Journal*, *12*(3), 56–69.

Mowen, J. C., Fang, X., & Scott, K. (2009). A hierarchical model approach for identifying the trait antecedents of general gambling propensity and of four gambling-related genres. *Journal of Business Research*, *62*(12), 1262–1268.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, *2*, 345–389.

National Gambling Impact Study Commission (1999). *Final report* (Washington, DC), http://govinfo.library.unt.edu/ngisc/index.html.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. X., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211.

Oh, H., & Hsu, C. H. C. (2001). Volitional degrees of gambling behaviours. *Annals of Tourism Research*, *28*(3), 618–637.

Productivity Commission (1999). *Australia's gambling industries: Inquiry reports*, *vol. 1–3*. (pp. 10)Melbourne, Australia: Media and Publications.

Roehl, W. S. (1999). Quality of life issues in a casino destination. *Journal of Business Research*, *44*(3), 223–229.

Seybert, N., & Bloomfield, R. (2009). Contagion of wishful thinking in markets. *Management Science*, *55*(5), 738–751.

Smith, G., Levere, M., & Kurtzman, R. (2009). Poker player behavior after big wins and big losses. *Management Science*, *55*(9), 1547–1555.

Smith, K. A., Willis, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the Operational Research Society*, *51*(5), 532–541.

Steinberg, D., & Colla, P. (1998). *CART (Classification and Regression Trees)*. San Diego, CA: Salford Systems.

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*(3), 606–621.

Strobl, C., Boulesteix, A. -L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 8–25.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, *157*(1), 196–217.

van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, *181*(1), 436–452.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.

Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations.* San Francisco, CA: Morgan Kaufmann.