UNIVERSITEIT GENT

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Reconciling Performance and Interpretability in Customer Churn Prediction using Ensemble Learning based on Generalized Additive Models

Koen W. De Bock[*]

Dirk Van den Poel[†]

August 2012

2012/805

[*] IESEG School of Management – Université Catholique de Lille (LEM, UMR CNRS 8179), Department of Marketing,IESEG School of Management
[†] Ghent University: http://www.crm.UGent.be

D/2012/7012/38

Abstract

To build a successful customer churn prediction model, a classification algorithm should be chosen that fulfills two requirements: strong classification performance and a high level of model interpretability. In recent literature, ensemble classifiers have demonstrated superior performance in a multitude of applications and data mining contests. However, due to an increased complexity they result in models that are often difficult to interpret. In this study, GAMensPlus, an ensemble classifier based upon generalized additive models (GAMs), in which both performance and interpretability are reconciled, is presented and evaluated in a context of churn prediction modeling. The recently proposed GAMens, based upon Bagging, the Random Subspace Method and semi-parametric GAMs as constituent classifiers, is extended to include two instruments for model interpretability: generalized feature importance scores, and bootstrap confidence bands for smoothing splines. In an experimental comparison on data sets of six real-life churn prediction projects, the competitive performance of the proposed algorithm over a set of well-known benchmark algorithms is demonstrated in terms of four evaluation metrics. Further, the ability of the technique to deliver valuable insight into the drivers of customer churn is illustrated in a case study on data from a European bank. Firstly, it is shown how the generalized feature importance scores allow the analyst to identify the importances of churn predictors in function of the criterion that is used to measure the quality of the model predictions. Secondly, the ability of GAMensPlus to identify nonlinear relationships between predictors and churn probabilities is demonstrated.

Keywords: Database marketing, customer churn prediction, ensemble classification, generalized additive models (GAMs), GAMens, model interpretability

1. Introduction

Many companies are currently operating in an environment of intensified competition, shortening product life cycles and decreasing customer brand loyalty (Cooil, Keiningham, Aksoy, & Hsu, 2007). In an effort to tighten the relationship that exists with a customer, many companies increasingly turn to the concepts of Customer Relationship Management (CRM) (Reinartz & Kumar, 2002; Winer, 2001) and, more specifically, database marketing (Blattberg, Kim, & Neslin, 2008). While both concepts aim at enhancing the relationship between a company and its customers, database marketing formally emphasizes the importance of customer data, such as demographical and psycho-graphical information, purchase history and survey responses, to allow for more effectively targeted marketing actions (Blattberg, et al., 2008).

An important discipline within database marketing is customer retention management, or the prevention of customer churn, defined as the propensity of customers to end the relationship with the company, and to switch to the competition. Several authors report the close link between customer retention and firm profitability (Gupta, Lehmann, & Stuart, 2004; Larivière & Van den Poel, 2005). Moreover, it is generally accepted that prolonging relationships with existing customers generates a higher return on investment than attracting new customers (Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000; Rust & Zahorik, 1993). A well-documented approach to improve customer retention is the practice of customer churn prediction, in which a classification model is built to identify those customers that are most likely to demonstrate churning behavior (Xie, Li, Ngai, & Ying, 2009).

Technically, customer churn prediction involves binary classification, which intends to generalize the relationship between churning behavior on the one hand, and information describing the customer on the other hand in a model that can be used for prediction purposes (Xie, et al., 2009). Consider the following notation. T is a training data set with information on *n* customers; $T = \{(x_1, y_1),...,(x_n, y_n)\}$. Each customer vector *x* is an element of X , being a set of *D* predictive features, $X = \{X_1,...,X_D\}$ and Y denotes a binary churn behavior outcome. Training samples (x_i, y_i) are a combination of an input vector x_i and a class membership y_i with $y_i \in \{0,1\}$, where class *1* identifies churning behavior. A standard classification model is a function that maps a given instance *x* to one of both classes in Y . However, in this study, confidences in class memberships (i.e., churn probabilities) are considered rather than exact classifications. This enables companies to produce a ranking of customers based on their proneness to churn, and to focus retention strategies on a certain proportion of 'riskiest' customers. Hence,

the desired output of a classifier F is the conditional class membership probability $P(Y=1|X)$. A classifier F then becomes a function $F: X \rightarrow P(Y|X)$ that maps an instance $x \in X \subset R^D$ to a confidence estimation that x belongs to class *1*.

Strong classification performance is generally perceived as a vital element of a customer churn prediction model. While Neslin et al. (Neslin, Gupta, Kamakura, Lu, & Mason, 2006) indicate that several steps within the modeling process determine the success of a churn prediction project, they emphasize that the estimation technique has a considerable impact upon the return of investment of retention actions. Consequently, a large body of literature is devoted to the evaluation of different modeling techniques for the prediction of customer churn. Techniques that have been suggested in literature include statistical techniques (e.g., logistic regression (Smith, Willis, & Brooks, 2000), generalized additive models (GAMs) (Coussement, Benoit, & Van den Poel, 2010), survival analysis (Van den Poel & Larivière, 2004)) and classifiers originating from data mining literature (e.g., neural networks (Mozer, et al., 2000), support vector machines (Coussement & Van den Poel, 2008a) and decision trees (Smith, et al., 2000)).

In recent literature on churn prediction, special interest has been devoted to ensemble classification (Lemmens & Croux, 2006). An ensemble classifier, or multiple classifier system (MCS), combines *M* classifiers into one aggregated model $E = \{F_1, F_2, F_3, ..., F_M\}$ and produces predictions as combinations of the outputs of its ensemble members using a certain fusion rule. Ensemble classifiers have been shown to demonstrate superior performance over uncombined models in several domains, such as image classification (Giacinto & Roli, 2001), cancer classification (Dettling, 2004), gene selection (Diaz-Uriate & de Andres, 2006), face recognition (X. Y. Tan, Chen, Zhou, & Zhang, 2005) and credit scoring (Paleologo, Elisseeff, & Antonini, 2010). It is generally accepted that ensemble classifiers are effective only if the constituent ensemble members exhibit strong classification performance and if they are diverse, i.e. if there is a level of disagreement on some proportion of the predictions to be made (Giacinto & Roli, 2001). Moreover, a tradeoff exists between both elements, indicated as the accuracy-diversity dilemma (Chandra, Chen, & Yao, 2006).

Several ensemble classifier algorithms have been proposed, each aiming at an injection of diversity between the ensemble members while maintaining member and overall classification performance. A possible classification of ensemble methods involves member classifier algorithm choice, and member training organization. In this study, non-hybrid ensemble classifiers (i.e., all ensemble members belong to the same algorithm family) with parallel member training are considered. One of the earliest ensemble methods within this category, proposed by Breiman (Breiman, 1996), is Bagging, an acronym for Bootstrap aggregation. In Bagging, each member classifier

$F_l; l = 1,...,M$ in the ensemble E is trained on a bootstrap sample of the training data, i.e., a random sample taken with replacement and with a size that is equal to that of the training data set. Aggregate predictions are obtained by means of majority voting, where the final classification is equal to the most frequently predicted class among the ensemble members (Kuncheva, 2004). Bagging especially enhances performance if its base classifiers are unstable (i.e., small variations in training data result in a significantly different classifier), decreasing variance. Two well-known related methods are the Random Subspace Method and Random Forests. In the Random Subspace Method (RSM; (Ho, 1998)), also known as Attribute Bagging (Bryll, Gutierrez-Osuna, & Quek, 2003), *R* features are randomly sampled (without replacement) instead of instances for the training of ensemble members. Successfully applied ensemble classifiers in customer churn prediction include Bagging (Lemmens & Croux, 2006), Stochastic Gradient Boosting (Burez & Van den Poel, 2009), Random Forests (Larivière & Van den Poel, 2005), Rotation Forests (De Bock & Van den Poel, 2011) and AdaCost (Glady, Baesens, & Croux, 2009).

In several domains, such as medical diagnostics (K. C. Tan, Yu, Heng, & Lee, 2003) or credit scoring (Martens, Baesens, Van Gestel, & Vanthienen, 2007; Setiono, Baesens, & Mues, 2009), model comprehensibility is extremely important. Also in database marketing literature is model interpretability advocated by several authors as an additional requirement for successful churn prediction models (Qi, et al., 2009; Shaw, Subramaniam, Tan, & Welge, 2001). Interpretable, intuitive models enable marketing decision makers to gain insight into customer behavior and identify factors with an impact upon customer loyalty and churning behavior (Masand, Datta, Mani, & Li, 1999). Techniques that have been suggested to deliver interpretable models include logistic regression (Buckinx & Van den Poel, 2005; Kim & Yoon, 2004), decision trees (Kim & Yoon, 2004) and, more recently, generalized additive models (Coussement, et al., 2010). Unfortunately, as Neslin et al. (Neslin, et al., 2006) suggest, explanation and prediction are two distinct functions of churn prediction models that can hardly be reconciled. Moreover, classification performance within this category of models has been found to be inferior to more strongly performing techniques, such as ensemble classifiers.

In this paper, an ensemble classifier is presented that reconciles interpretability with strong classification performance. Based on a recently proposed ensemble classifier (De Bock, Coussement, & Van den Poel, 2010a) based on Bagging and RSM, and implementing generalized additive models (GAMs) as base classifiers, GAMensPlus is presented. This technique extends GAM-based ensemble classifiers with two instruments that allow model interpretation: (i) generalized feature importance scores, and (ii) bootstrap smoothing spline confidence intervals.

The remainder of this paper is structured as follows. Section 2. reviews generalized additive models and the GAMens classifier. Section 3. then presents GAMensPlus. Section 4. is devoted to an experimental comparison of classification performance of GAMensPlus to a selection of benchmark techniques. This section contains subsections on the experimental setup, evaluation criteria, and experimental results. In Section 5., interpretability of GAMensPlus is assessed in a specific churn prediction context. Finally, conclusions are made and directions for future research are suggested.

2. Related literature

2.1. Generalized additive models

The ensemble classifier proposed in this study is based upon generalized additive models (GAMs) (De Bock, et al., 2010a). Generalized additive models have been successfully applied in several domains as a flexible technique for nonparametric regression (Berg, 2007; Coussement & Van den Poel, 2008b; Lado, Cadarso-Suarez, Roca-Pardinas, & Tahoces, 2006). GAMs extend the framework of generalized linear models (GLMs; (McCullagh & Nelder, 1989)) which comprises a broad range of parametric regression models, characterized by (i) a response variable belonging to any distribution within the exponential family (the random component), (ii) a fixed function that represents any functional relationship between the combined linear effect of the predictors and the expected value of the outcome (the link function) and (iii) to the assumption of a combined linear effect of the explanatory features (the systematic component) (Lado, et al., 2006). In the context of customer churn prediction involving binary classification, a GLM would take the form of a logistic regression, in which the response variable *Y* is described by a binomial distribution, and the logistic link function is applied:

$$\text{logit}(P(Y=1|X)) \equiv \log\left\{\frac{P(Y=1|X)}{1-P(Y=1|X)}\right\} = \sum_{k=1}^{p} \beta_k X_k \qquad \textbf{(1)}$$

In generalized additive models, proposed by Hastie and Tibshirani (Hastie & Tibshirani, 1986), the influence of an explanatory feature is no longer subject to any linear or other parametric specification, but instead fit using an arbitrary nonparametric function. GAMs replace the linear combination $\sum_{k=1}^{p} \beta_k X_k$ by the additive form $\sum_{k=1}^{p} f_k(X_k)$, where each partial function f_k is a unspecified smooth function. In order to accommodate a binary response variable and the inclusion of categorical variables, the GAM specification that is considered in this study is a logistic, semi-parametric additive model:

$$\text{logit}(P(Y=1|X)) \equiv \log\left\{\frac{P(Y=1|X)}{1-P(Y=1|X)}\right\} = \sum_{j=1}^{p_c} s_j(X_j) + \sum_{k=1}^{p_b} \beta_k X_k \quad \textbf{(2)}$$

where features X_j, $j=1,...,p_c$ are continuous variables, $X_k = 1,..., p_b$ are dummy-coded components of categorical variables and the smooth functions $s_1(X_1), s_2(X_2),...,s_{p_c}(X_{p_c})$ are smoothing splines that estimate the nonparametric trend for the dependence of the logit on $X_1, X_2,...X_{p_c}$. We kindly refer the reader to (Hastie & Tibshirani, 1990) for more details on GAMs and smoothing splines.

2.2. GAM-based ensemble classification

In (De Bock, et al., 2010a), ensemble classification based on generalized additive models is presented. GAMens is an algorithm based on Bagging, the Random Subspace Method (RSM), and adopts GAMs as constituent classifiers. The technique is based upon a logistic, semi-parametric additive model specification as in (2). Ensemble predictions are obtained using mean combination, takes the average of the posterior class membership probabilities output by the individual ensemble members:

$$E(x) = \frac{1}{M}\sum_{l=1}^{M} F_l(x) = \frac{1}{M}\sum_{l=1}^{M} P_l(Y=1|x) \quad \textbf{(3)}$$

A limited number of parameters is required to be specified for GAMens. First, the *M* parameter designates the number of desired GAM base classifiers to be included in the ensemble classifier. Second, the desired number of variables to be selected as random feature subspaces is required (*R* parameter). Finally, specification of the number of degrees of freedom to be used in the smoothing spline estimation is required (*DF* parameter).

3. GAMensPlus

Based on GAMens, *GAMensPlus* is presented as a modeling technique for customer churn prediction that combines strong classification performance with model interpretability. The pseudo code of GAMensPlus is presented in Figure 1. and Figure 2. GAMensPlus combines the training and prediction phases of GAMens (Figure 1.) with an explanation phase (Figure 2.), in which two heuristics are introduced to allow model interpretation and enable marketing decision makers to better understand the influence and relative importance of descriptive features: generalized feature importance scores, and bootstrap confidence intervals for smoothing splines. Both concepts are explained in this section.

[INSERT FIGURE 1. HERE]

[INSERT FIGURE 2. HERE]

3.1. Generalized feature importance scores

As a first interpretability heuristic in GAMensPlus, generalized feature importance scores are introduced. Generalized feature importance scores are based upon the concept of variable importance measures, as introduced by Breiman as a by-product of Random Forests (Breiman, 2001). While different types of variable importance measures have been proposed, here permutation accuracy importances are considered, which are reported in (Strobl, Boulesteix, Zeileis, & Hothorn, 2007) as most advanced and reliable importance measure available in Random Forests.

Permutation accuracy importance scores are calculated using out-of-bag data. As every member tree F_j, $j=1,...,M$ within a Random Forest is trained using a bootstrap sample, approximately one-third of the training instances are not selected to build that respective tree. These instances are called the out-of-bag (oob) instances for tree F_j. and can be used to reliably estimate variable importances. Permutation accuracy importance scores for feature X_d are then obtained by calculating, for every member tree F_j, the average difference in accuracy for tree F_j. before and after permuting the values of variable X_d in the out-of bag data, and averaging the result over all trees F_j., $j=1,...,M$.

In customer churn prediction, the performance of a classifier is evaluated differently according to the specifics of the business setting and marketing objectives of retention-increasing efforts. Depending on the situation, different performance metrics are relevant for the evaluation of a churn prediction model, such as accuracy, AUC or lift. Consequently, the relative importance of predictive features should be measured differently according to the evaluation criterion that is being optimized. Hence, generalized feature importance scores $FI_{PC}(X_d)$ are introduced that measure the importance of feature X_d the average decrease in performance evaluation criterion *PC*.

In (Strobl, et al., 2007), variable importance measures in Random Forests are found to be biased in situations that involve data with different scales of measurement and the number of categories of categorical variables. Two responsible factors are identified: biased variable selection at node splits in CART decision trees, and effects induced by bootstrap sampling with replacement. The generalized feature importance scores in GAMensPlus are not affected by these deficiencies for two reasons. Firstly, GAMensPlus implements GAMs as base classifiers that

do not involve (biased) feature selection. Secondly, experiments in (Strobl, et al., 2007) demonstrate that the inclusion of categorical features only introduces bias if they include more than two categories, while the GAM specification for GAMensPlus only allows continuous and binary features.

3.2. Bootstrap smoothing spline confidence intervals

As second instrument for model interpretation, bootstrap confidence intervals for smoothing splines are introduced in GAMensPlus. The Bagging component of GAMensPlus, introducing the use of bootstrap samples of the data as training data for ensemble members, simultaneously allows for the construction of bootstrap confidence intervals that summarize the nonparametric trends captured within the ensemble member GAMs. These allow model users not only to identify the relationship that exists between a predictive feature and the probability to churn, but also to evaluate the precision of the identified relationship in particular regions within the range of values of a feature.

Several bootstrap confidence intervals for smoothing splines have been defined (Wang & Wahba, 1995). In this study, bootstrap percentile intervals (Efron, 1982) are considered. However, more advanced approaches, as suggested in (Wang & Wahba, 1995), could also be implemented. To identify the 95% bootstrap confidence interval of the smoothing splines for feature X_d, at each value $x_{d,i}$ of X_d, the empirical distribution of $s_d^*(x_{d,i})$, the random variable of bootstrapped smoothing splines has to be identified. A bootstrap confidence interval is then constructed by points at the 2.5^{th} and 97.5^{th} percentile of this empirical distribution. Repeating this process for each unique value of X_d then results in the 95% confidence band of the nonparametric regression line (Efron & Tibshirani, 1993). A similar approach is applied to obtain bootstrap confidence intervals for regression coefficients of the dummy features.

4. Experimental comparison

This section is devoted to a comparison of classification performance of GAMensPlus versus a selection of benchmark algorithms in the context of customer churn prediction modeling. Subsequent subsections discuss evaluation criteria, experimental setup and results.

4.1. Evaluation criteria

To evaluate the classification performance of GAMensPlus and the selection benchmark classifiers, four evaluation criteria are considered: (i) accuracy and (ii) AUC, which are both often used to assess and compare

generic classification quality, and (iii) top-decile lift and (iv) lift index reflect which are particularly suited to evaluate models for customer targeting.

Accuracy, or the percentage of correctly classified instances, is the first evaluation criterion. While well-accepted to evaluate classifier models, and intuitive, accuracy is considered an inappropriate metric for churn modeling, for a number of reasons: (i) it does not take into account predicted class membership probabilities but instead assumes a threshold to obtain classifications from probabilities, and (ii) it is unreliable in a situation of class imbalance (Lemmens & Croux, 2006).

The second evaluation criterion is Area Under the Receiver Operating Characteristics curve (AUC or AUROC) which is often used in churn prediction literature (Coussement & Van den Poel, 2008b; Lemmens & Croux, 2006). Several authors like Provost et al. (2000) or Langley (2000) advocate AUC as an objective performance criterion, well-suited for the comparison of classifier performance. Unlike accuracy, it evaluates the ability of a classifier to distinguish between the two classes based on the predicted class membership probabilities, and is therefore suitable for imbalanced classification problems such as customer churn prediction.

Lift focuses on the segment of customers with the highest risk to the company, i.e. customers with the highest probability to churn. Two alternative variations of lift are considered: top-decile lift and lift index. Suppose that a company is interested in the top 10% of most likely churners, based on predicted churn probabilities. The top decile lift then equals the ratio of the proportion of churners in the top decile of ordered posterior churn probabilities, $\pi_{10\%}$, to the churn rate in the total customer population, π (Lemmens & Croux, 2006):

$$Top-decile\ lift = \frac{\pi_{10\%}}{\pi} \tag{4}$$

A fourth and final evaluation metric is lift index (Crone, Lessmann, & Stahlbock, 2006; Ling & Li, 1998). Suppose *S* is a ranked list of customers based on their churn probability. Lift index is then calculated as

$$Lift\ index = \frac{\left(1.0 \cdot S_1 + 0.9 \cdot S_2 + 0.8 \cdot S_3 + ... + 0.1 \cdot S_{10}\right)}{\sum_{i=1}^{10} S_i} \tag{5}$$

where S_i is the number of churning customers in the *i*th decile of *S*. The lift index takes a value between 0.5 and 1, where a value of 0.5 indicates random identification of customers as churners, and a value of *1* with $S_1 = \sum_{i=1}^{10} S_i$ if the churn rate is smaller than 10%.

4.2. Data

Experiments are conducted on data sets from six real-life churn prediction projects originating from large European companies. Table 1. summarizes characteristics of the data sets. For reasons of confidentiality, company names are not disclosed.

[INSERT TABLE 1. HERE]

These data sets have a number of common features. First, they all (with the exception of the first data set) exhibit rather large dimensionalities, both in terms of number of instances and the number of descriptive features. Second, they are characterized by considerable class imbalance, most notably the data sets originating from a bank, a European telecommunications operator and a mail-order garments company. Predictive features among these data sets capture information on customer demographics, historical transactional data and financial information. In all of these data sets, churn is defined as the absence of at least one product purchase or renewal within a certain time period. The data sets Bank and Telecom correspond to a contractual setting where partial churn is measured, i.e. defection in at least one product category. The DIY supplies, Supermarket chain I and II, and mail-order garments data sets correspond to non-contractual settings where total churn is measured in a certain time period.

To deal with class imbalance, which is known to distort classifier performance for classification algorithms that are not particularly designed to deal with this problem, undersampling is applied, as suggested by Weiss (2004) and applied to customer churn prediction by Burez and Van den Poel (2009). Undersampling involves randomly removing instances from the majority class from the training data until both classes are balanced.

4.3. Experimental setup

To evaluate the predictive performance of GAMensPlus, an experimental comparison is made with a selection of benchmark algorithms. Classification performance of GAMensPlus is compared to five benchmark algorithms: three ensemble classifiers (Bagging, Random Subspace Method (RSM) and Random Forests) and two uncombined classifiers (logistic regression and generalized additive models). As outlined earlier, classification performance is evaluated in terms of four performance metrics: accuracy, AUC, top-decile lift and lift index.

GAMensPlus is implemented in R (R Development Core Team, 2009) based upon the GAMens package (De Bock, Coussement, & Van den Poel, 2010b). Bagging and Random Forest results are obtained using the adabag (Alfaro, Gámez, & García, 2006) and randomForest (Liaw & Wiener, 2002) packages in R. All remaining

algorithms are programmed in R. Ensemble sizes of GAMensPlus, Bagging, RSM and Random Forest are set to 100 constituent members per ensemble. All ensemble classifiers except GAMensPlus are combinations of unpruned CART base classifiers.

Algorithm parameters settings are based on default or recommended values. First, the size of random feature subsets in Random Forests are set equal to the square root of the number of features in the respective data set, as suggested in (2001). This setting is also used for RSM, and for GAMensPlus, as recommended in (De Bock, et al., 2010a). Second, the *DF* parameter in GAMensPlus, denoting the number of degrees of freedom for smoothing spline estimation is fixed to four, as suggested in (De Bock, et al., 2010a) based on examples provided by Hastie and Tibshirani (Hastie & Tibshirani, 1990). Experimental results are all based following five times twofold cross-validation (5 x 2cv). This involves five replications of a twofold cross-validation. In each replication, instances in the data set are randomly assigned to two parts of equal size. Moreover, stratified random sampling is applied in order to maintain the original class distributions. One part is once used as training data for a classifier and the performance is calculated for the other part, acting as a test set. This process is then repeated, switching the roles of the two data set parts. The same splits are used for all classifier algorithms. Note that the undersampling of the training data sets is applied after the division of the data for the cross-validation. In order to test for significant differences among classifiers' performances, the obtained results are analyzed as suggested by Demšar (2006), using a nonparametric Friedman test followed by Holm's procedure (Holm, 1979) to make post-hoc pairwise comparisons between GAMensPlus and the benchmark algorithms.

4.4. Experimental results

This section presents the results of the experimental comparison of GAMensPlus to a selection of benchmark algorithms, for data sets from six real-life customer churn prediction projects. Tables 2., 3., 4. and 5. report result averages and standard deviations of results in terms of accuracy, AUC, top-decile lift and lift index respectively based on runs from a five times twofold cross-validation (5x2cv). The best and second best results per data set are indicated in bold and italic fonts, respectively.

[INSERT TABLE 2. HERE]

[INSERT TABLE 3. HERE]

[INSERT TABLE 4. HERE]

[INSERT TABLE 5. HERE]

[INSERT TABLE 6. HERE]

Table 6. provides a summary of the results of the Friedman tests for the four evaluation metrics and Holm's post-hoc tests for comparisons between GAMensPlus and Bagging, RSM, and Random Forest as ensemble classifier benchmarks and GAM and logistic regression as uncombined, interpretable classifier benchmark algorithms. Entries in Table 6 denote average rankings of the respective algorithms over all 6 considered datasets. For every performance metric, the lowest average rank, received by the best performing algorithm, is indicated in bold. Ranks that differ significantly at significance levels of 90% and 95% percent are indicated by one or two asterisk symbols, respectively.

Overall, these average rankings and post-hoc tests confirm the highly competitive performance of GAMensPlus. In detail, the following observations can be derived from these results.

First, it is clear that building an ensemble of GAMs is an effective strategy to increase the predictive performance of an uncombined generalized additive model. GAMensPlus significantly outperforms GAM for all four performance evaluation metrics.

Second, a comparison of GAMensPlus to other, well-established, ensemble classifiers reveals that GAMensPlus is superior for the most relevant performance metrics. GAMensPlus consistently outperforms Bagging and RSM. The strongest competition is delivered by Random Forests. This confirms findings in previous experiments that revealed the strong predictive performance of Random Forests in the context of customer churn prediction (Coussement & Van den Poel, 2008b; Larivière & Van den Poel, 2005). GAMensPlus significantly outperforms Random Forests in terms of AUC, top-decile lift and lift index at the 90% significance level.

Third, the results indicate the good performance of logistic regression, which obtains the second highest average rankings for AUC and the two lift measures. Logistic regression consistently outperforms the three benchmark ensemble algorithms; bagging, RSM and Random Forests. These observations confirms previous research findings (Burez & Van den Poel, 2009), stating that logistic regression, despite its simplicity, performs competitively when compared to more advanced techniques, in the discipline of customer churn prediction modeling.

5. Case Study: churn prediction in a European financial services company

In this section, the interpretability component of GAMensPlus is demonstrated. To this end, a case study is presented related to customer churn prediction in financial services. This analysis is based upon a churn prediction project for a major European bank that was elaborated early 2005. The objective of the project involved the creation of a customer churn model to predict partial churn over a 12-month period, i.e. the closing of a checking account by a customer, regardless of possession of other accounts or loans at the bank. To this end, the company provided data that includes customer information at the end of March 2005. This data was used to measure churn behavior over a one-year period, from March 1st, 2004 to March 31st, 2005. All customer information that was available at March 1st, 2004 was then considered to create a set of predictive features. See Table 7. for a summarization of these features. Note that this data set also featured earlier in the experimental comparison in Section 4. All features are calculated using the reference date of March 1st, 2004, while the monetary features are expressed in euro.

[INSERT TABLE 7. HERE]

Two interpretability instruments are demonstrated: generalized feature importance scores, and bootstrap smoothing spline confidence intervals. Keeping all algorithm settings is in Section 4., a GAMensPlus model is trained and the results of the interpretation phase are presented. First, feature importance scores are considered. Table 8. provides the ten most importance features according to a ranking of AUC-based feature importance scores, while Table 9. considers feature importance scores based on top-decile lift.

[INSERT TABLE 8. HERE]

[INSERT TABLE 9. HERE]

Table 8. provides insight into the most important predictive features for churn behavior assuming that the churn prediction model is evaluated in AUC. The list indicates the importance of features related to account services usage, such as the balance of the checking account, or the total number of credit transactions. Most of the features in the list are RFM variables (Cullinan, 1977), or features related to recency (R), frequency (F) and monetary value (M) of product purchase or service usage. These features have been identified as strong predictors in customer churn prediction modeling in several studies (Kim & Yoon, 2004; Lemmens & Croux, 2006). Table 9. shows the

top ten of features in terms of feature importance scores based on top-decile lift. A comparison between Tables 8. and 9. reveals that many features appear in both rankings. However, the order is different, and new features emerge, such as the number of overdraft days, i.e., the number of days with a negative account balance, and the average overdraft amount. This indicates that different features are identified as most important by the feature importance scores when the performance of the customer churn prediction model is measured differently.

Figure 3. provides average trends and 95% confidence bands based on bootstrap confidence intervals for a selection of features. The figures also include histograms that provide an indication of the data density at a particular region of a feature. Dark-colored bars represent frequencies for non-churners, while light-colored bars represent frequencies for churning customers.

[INSERT FIGURE 3. HERE]

Overall, the plot reveals the ability of GAMensPlus to summarize the nonparametric smoothing spline trends of its member GAMs using the confidence bands based on bootstrap confidence intervals. The average trend represents the overall relationship of a feature to the probability to churn, while the confidence bands indicate the reliability of the trend at different regions of the feature range. Consider the trend of the length of relationship of the customer, i.e. the number of years that a person has been customer. The trend reveals an overall negative relationship between the length of relationship and the probability to churn. However, in the range of 0 to 5 years and 15 to 25 years, the relationship is positive on average. At these intervals, broader confidence intervals are observed indicating that there is more variation among the estimated trends, and that the estimated trend is less reliable at these intervals. The relationship between recency, or the number of months since the last change in balance amount, is quasi-linearly positive. As expected, the confidence bands become wider as the feature takes values in regions with less density. This is also observed for the trend of age, and the percentage of loan repaid. The trend of age is quasi-linearly negative: older customers are more loyal. The trend of the percentage of loan repaid follows a U-shape. Finally, a dummy feature is considered, indicating whether the customer is retired. This type of feature is fit in a linear fashion, as specified in the GAM specification for GAMensPlus. Expectedly, the trend of this feature confirms the trend of age: the probability to churn decreases if the customer is retired.

6. Conclusion, limitations and future work

Customer churn prediction is an important discipline within database marketing, aiming to identify those customers that have the highest probability to churn, i.e., to cease the relationship with the company and move their business elsewhere. In this study, GAMensPlus, a classification algorithm that combines strong classification performance with a high degree of model interpretability, is presented for customer churn prediction. The former factor allows marketing decision makers to effectively identify churners, while the latter enables them to extract valuable insights from the model and increase understanding, and thus, acceptance, of the model throughout their organization.

GAMensPlus extends GAMens, an ensemble classifier for binary classification using generalized additive models (GAMs) as constituent classifiers and combining two classical ensemble strategies, Bagging and the Random Subspace Method (RSM). While earlier work demonstrated the strong performance of GAMens on a selection of data sets from varying domains, it is in this study considered and evaluated for customer churn prediction. The extension involves the addition of an interpretation phase that incorporates two instruments for model interpretability: (i) generalized feature importance scores, and (ii) bootstrap confidence intervals for smoothing splines.

Generalized feature importance scores are inspired by the variable importance measures in Random Forests. However, they are not necessarily calculated as contributions of features to accuracy, but instead require the analyst to specify the actual performance metric that is used to measure the performance of the churn prediction model. Hence, features are properly evaluated according to their contribution to the measured quality of the churn prediction model.

The Bagging component in GAMensPlus allows the construction of bootstrap smoothing spline confidence intervals that summarize the nonparametric trends captured within the ensemble member GAMs. When calculated over the entire range of a feature, confidence bands and average trends can be constructed that allow interpretation of the overall relationship between the feature and the churn probability, and the reliability of the estimated trend at different regions of the feature range.

The evaluation of GAMensPlus involves two parts: an evaluation of classification performance including a comparison to other algorithms, and a demonstration of the interpretability mechanisms in a case study. First, in an experimental comparison of classification performance over data sets from six real-life churn prediction projects, GAMensPlus is compared to three ensemble classifiers (Bagging, RSM and Random Forests) and two individual

classifiers (logistic regression, and GAM). Moreover, performance is compared using four evaluation criteria: accuracy, AUC, top-decile lift and lift index. The results indicate that GAMensPlus obtains strong classification performance that is performing at least as good as the benchmark algorithms. On average, GAMensPlus is ranked second in terms of accuracy, and first considering AUC, top-decile lift and lift index results.

Further, the interpretability instruments of GAMensPlus have been demonstrated in a case study, involving churn prediction at a European financial services company.

Certain limitations and directions for future research to this study can be identified. Firstly, the study does not take into account more advanced bootstrap confidence intervals for smoothing splines. Future research could for example consider alternative techniques, as proposed and compared in [46]. Secondly, the current study does not compare the interpretability techniques to other techniques, such as logistic regression, decision trees, or rule extraction techniques for ensemble methods.

Acknowledgements

The authors thank all former and current Ph.D. researchers at the modeling cluster of the Department of Marketing who contributed the real-life business data sets they gathered and processed during their Ph.D.'s, the developers of R, the randomForest and adabag packages. Further, the authors acknowledge Ghent University for funding the Ph.D. project of Koen W. De Bock and the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy). Finally, we express our gratitude to the reviewers for their useful remarks.

References

Alfaro, E., Gámez, M., & García, N. (2006). adabag: Applies Adaboost.M1 and Bagging. In (pp. R Package version 1.1).

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry, 23*(2), 129-143.

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

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

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

Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition, 36*(6), 1291-1302.

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

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626-4636.

Chandra, A., Chen, H., & Yao, X. (2006). Trade-off between diversity and accuracy in ensemble generation. In Y. Jin (Ed.), *Multi-objective Machine Learning*. New York: Springer-Verlag.

Cooil, B., Keiningham, T. L., Aksoy, L., & Hsu, M. (2007). A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. *Journal of Marketing, 71*(1), 67-83.

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. (2008a). 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. (2008b). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems, 44*(4), 870-882.

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.

Cullinan, G. J. (1977). *Picking them by their batting averages' recency-frequency-monetary method of controlling circulation*. New York: Manual Release 2103, Direct Mail / Marketing Association.

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

De Bock, K. W., Coussement, K., & Van den Poel, D. (2010b). GAMens: Applies GAMens, GAMrsm and GAMbag ensemble classifiers. In (pp. R Package version 1.11).

De Bock, K. W., & Van den Poel, D. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications, 38*(10), 12293-12301.

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

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics, 20*(18), 3583-3593.

Diaz-Uriate, R., & de Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*(3.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Giacinto, G., & Roli, F. (2001). An approach to the automatic design of multiple classifier systems. *Pattern Recognition Letters, 22*(1), 25-33.

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research, 197*(1), 402-411.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of Marketing Research, 41*(1), 7-18.

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.

Holm, S. (1979). A simple sequantially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Kim, H. S., & Yoon, C. H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*(9-10), 751-765.

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

Lado, M. J., Cadarso-Suarez, C., Roca-Pardinas, J., & Tahoces, P. G. (2006). Using generalized additive models for construction of nonlinear classifiers in computer-aided diagnosis systems. *IEEE Transactions on Information Technology in Biomedicine, 10*(2), 246-253.

Langley, P. (2000). Crafting papers on Machine Learning. In P. Langley (Ed.), *17th International Conference on Machine Learning (ICML-2000)* (pp. 1207 - 1216): Stanford University.

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), 472-484.

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

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22.

Ling, C. X., & Li, C. (1998). Data mining for Direct Marketing: Problems and Solutions. In *Fourth International Conference on Knowledge Discover and Data Mining (KDD-98)* (pp. 73-79).

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466-1476.

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

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. London, UK: Chapman & Hall.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks, 11*(3), 690-696.

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.

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*(2), 490-499.

Provost, F., Fawcett, T., & Kohavi, R. (2000). The Case against Accuracy Estimation for Comparing Induction Algorithms. In J. Shavlik (Ed.), *15th International Conference on Machine Learning (ICML 1998)* (pp. 445-453). Madison, Wisconsin.: Morgan Kaufman.

Qi, J. Y., Zhang, L., Liu, Y. P., Li, L., Zhou, Y. P., Shen, Y., Liang, L., & Li, H. Z. (2009). ADTreesLogit model for customer churn prediction. *Annals of Operations Research, 168*(1), 247-265.

R Development Core Team. (2009). R: A Language and Environment for Statistical Computing. In *R Development Core Team* (pp. Vienna, Austria). Vienna, Austria.

Reinartz, W., & Kumar, V. (2002). The mismanagement of customer loyalty. *Harvard Business Review, 80*(7), 86-94.

Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. *Journal of Retailing, 69*(2), 193-215.

Setiono, R., Baesens, B., & Mues, C. (2009). A note on knowledge discovery using neural networks and its application to credit card screening. *European Journal of Operational Research, 192*(1), 326-332.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems, 31*(1), 127-137.

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.

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*(25), doi:10.1186/1471-2105-1188-1125.

Tan, K. C., Yu, Q., Heng, C. M., & Lee, T. H. (2003). Evolutionary computing for knowledge discovery in medical diagnosis. *Artificial Intelligence in Medicine, 27*(2), 129-154.

Tan, X. Y., Chen, S. C., Zhou, Z. H., & Zhang, F. Y. (2005). Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble. *IEEE Transactions on Neural Networks, 16*(4), 875-886.

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.

Wang, Y. D., & Wahba, G. (1995). Bootstrap confidence-intervals for smoothing splines and their comparison to bayesian confidence-intervals. *Journal of Statistical Computation and Simulation, 51*(2-4), 263-279.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations, 6*(1), 315-354.

Winer, R. S. (2001). A framework for customer relationship management. *California Management Review, 43*(4), 89-108.

Xie, Y. Y., Li, X., Ngai, E. W. T., & Ying, W. Y. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445-5449.

Data set	Instances	Number of features	Minority class percentage
Supermarket chain II	32,371	46	25.15
DIY supplies	3,827	15	28.14
Bank	20,456	137	5.99
Supermarket chain II	8,453	36	47.38
Telecom	35,550	529	2.76
Mail-order garments	43,305	244	1.76

Table 1.: Data set description

Dataset	Algorithm											
	GAMensPlus		*Bagging*		*Random Forest*		*RSM*		*Logistic regression*		*GAM*	
Supermarket chain I	*0.7449*	(0.0096)	0.7199	(0.0173)	0.7375	(0.0112)	0.7397	(0.0327)	0.7404	(0.0194)	**0.7476**	(0.0055)
DIY supplies	0.6657	(0.0070)	**0.6760**	(0.0360)	0.6497	(0.0085)	*0.6735*	(0.0415)	0.6324	(0.0268)	0.6647	(0.0088)
Bank	0.7291	(0.0153)	*0.7507*	(0.0272)	0.7422	(0.0145)	**0.7993**	(0.0186)	0.7118	(0.0493)	0.6490	(0.0826)
Supermarket chain II	*0.6890*	(0.0084)	0.6888	(0.0221)	0.6708	(0.0098)	**0.6904**	(0.0176)	0.6772	(0.0212)	0.6566	(0.0423)
Telecom	*0.6409*	(0.0321)	0.6113	(0.0154)	0.6168	(0.0154)	**0.6907**	(0.0116)	0.6244	(0.0260)	0.5975	(0.0160)
Mail-order garments	**0.8182**	(0.0092)	0.7527	(0.0429)	0.7662	(0.0122)	*0.7960*	(0.0398)	0.7805	(0.0075)	0.7646	(0.0067)

Table 2.: Experimental results: accuracy (averages and standard deviations)

Dataset	Algorithm											
	GAMensPlus		*Bagging*		*Random Forest*		*RSM*		*Logistic regression*		*GAM*	
Supermarket chain I	*0.8135*	(0.0101)	0.7724	(0.0284)	0.8055	(0.0109)	0.7883	(0.0099)	0.8089	(0.0127)	**0.8174**	*(0.0124)*
DIY supplies	**0.7580**	(0.0108)	0.7483	(0.0228)	0.7174	(0.0066)	0.7348	(0.0173)	*0.7567*	(0.0118)	0.7532	(0.0090)
Bank	0.7819	(0.0149)	*0.7823*	(0.0137)	**0.8093**	(0.0138)	0.7807	(0.0140)	0.7525	(0.0396)	0.6929	(0.0657)
Supermarket chain II	**0.7486**	(0.0181)	*0.7437*	(0.0294)	0.7270	(0.0069)	0.7418	(0.0146)	0.7350	(0.0254)	0.7046	(0.0266)
Telecom	**0.6349**	(0.0144)	0.6168	(0.0210)	0.6273	(0.0166)	0.6130	(0.0213)	0.6307	(0.0122)	0.6238	(0.0104)
Mail-order garments	**0.8474**	(0.0043)	0.8152	(0.0095)	0.8386	(0.0062)	0.8345	(0.0068)	0.8311	(0.0073)	0.7986	(0.0309)

Table 3.: Experimental results: AUC (averages and standard deviations)

Dataset	Algorithm											
	GAMensPlus		*Bagging*		*Random Forest*		*RSM*		*Logistic regression*		*GAM*	
Supermarket chain I	**2.7313**	(0.092)	2.1200	(0.3372)	2.6824	(0.1287)	2.2918	(0.1717)	2.7262	(0.4268)	2.7812	(0.4570)
DIY supplies	**2.2278**	(0.0835)	1.9896	(0.1526)	1.9319	(0.1036)	1.9207	(0.1601)	2.2148	(0.1380)	2.1813	(0.0994)
Bank	*4.0209*	(0.2104)	3.6890	(0.2984)	**4.2024**	(0.1935)	3.8917	(0.2068)	3.2442	(0.6238)	2.2876	(0.8665)
Supermarket chain II	*1.8261*	(0.1341)	**1.8310**	(0.3412)	1.7476	(0.154)	1.7329	(0.2442)	1.8261	(0.1344)	1.7181	(0.5412)
Telecom	**2.2110**	(0.1829)	1.9984	(0.1828)	2.1497	(0.2329)	2.0945	(0.1727)	*2.2090*	(0.0967)	2.1742	(0.1240)
Mail-order garments	**5.1380**	(0.1816)	4.5796	(0.3359)	5.0070	(0.2003)	4.8521	(0.3303)	4.9703	(0.1318)	4.5611	(0.6048)

Table 4.: Experimental results: top-decile lift (averages and standard deviations)

Dataset	Algorithm											
	GAMensPlus		*Bagging*		*Random Forest*		*RSM*		*Logistic regression*		*GAM*	
Supermarket chain I	*0.7823*	(0.0097)	0.7496	(0.0243)	0.7710	(0.0098)	0.7557	(0.0209)	0.7789	(0.0294	**0.7851**	(0.0532)
DIY supplies	**0.7338**	(0.0081)	0.7296	(0.0165)	0.7009	(0.0050)	0.7078	(0.0170)	*0.7329*	(0.0084)	0.7299	(0.0067)
Bank	*0.8097*	(0.0139)	0.8073	(0.0155)	**0.8264**	(0.0139)	0.7985	(0.0158)	0.7840	(0.0363)	0.7186	(0.0745)
Supermarket chain II	*0.6777*	(0.0101)	0.6687	(0.0195)	0.6621	(0.1657)	**0.6781**	(0.0199)	0.6716	(0.0097)	0.6549	(0.0090)
Telecom	**0.6778**	(0.0145)	0.6528	(0.0267)	0.6623	(0.0149)	0.6452	(0.0243)	*0.6742*	(0.0124)	0.6680	(0.0102)
Mail-order garments	**0.8827**	(0.0044)	0.8433	(0.0106)	*0.8700*	(0.0076)	0.8624	(0.0096)	0.8672	(0.0079)	0.8404	(0.0253)

Table 5.: Experimental results: lift index (averages and standard deviations)

Algorithm	Average rank			
	Accuracy	*AUC*	*Top-decile lift*	*Lift index*
GAMensPlus	2.33	**1.50**	**1.58**	**1.50**
Bagging	3.83	4.00**	4.33**	4.50**
Random Forest	4.33**	3.5*	3.33*	3.67*
RSM	**1.83**	4.33**	4.67**	4.17**
Logistic regression	4.00*	3.33	2.92	3.0*
GAM	4.67**	4.33**	4.17**	4.17*

* p<0.10; ** p<0.05

Table 6.: Average algorithm rankings and post-hoc test results (Holm's procedure)

Category	*Feature*
Demographical customer information	Gender
	Age
	Language
	Marital status
	Occupation
General customer information	Number of accounts, per account type
	Number of debet payment cards
	Number of credit cards
	Number of debit cards
	Overall length of relationship (LOR)
	Number of bank accounts
	Number of checking accounts
Features related to bank account usage	Recency based on change in balance amount
	Recency based on last transaction
	Length of relationship (LOR)
	Current account balance (in €)
	Total number of credit transactions
	Total number of debit transactions
	Total credit movement
	Total debet movement
	Number of days with credit interest owed
	Number of days with debit interest due
	Number of overdraft days (balance < 0 €)
	Average overdraft amount
Features related to other products	Number of loans
	Percentage of loan repaid
	Total loan amount
	Remaining loan amount
	Number of mortgage loans
	Percentage of mortgage loan paid
	Total morgage loan amount
	Remaining mortgage loan amount

Table 7.: Selection of features in the *Bank* dataset

Rank	*Feature*	*Feature importance score*
1	Checking account balance	0.1276
2	Average amount of credit transactions	0.1177
3	Number of debit transactions	0.1106
4	Recency based on last transaction	0.0963
5	Recency based on change in balance amount	0.0910
6	Total accounts balance	0.0853
7	Total number of credit transactions	0.0721
8	Number of bank accounts	0.0674
9	Number of checking accounts	0.0611
10	Total credit movement	0.0599

Table 8.: Ten most important features with feature importance scores based on AUC

Rank	*Feature*	*Feature importance score*
1	Recency based on change in balance amount	0.3488
2	Number of overdraft days (balance < 0 €)	0.3258
3	Number of days that overdraft interest is due	0.2754
4	Recency based on change in balance amount	0.2613
5	Number of bank accounts	0.2530
6	Number of checking accounts	0.2241
7	Average overdraft amount	0.2188
8	Average amount of debit transactions	0.2146
9	Number of debit transactions	0.2086
10	Checking account balance	0.1815

Table 9.: Ten most important features with feature importance scores based on top-decile lift

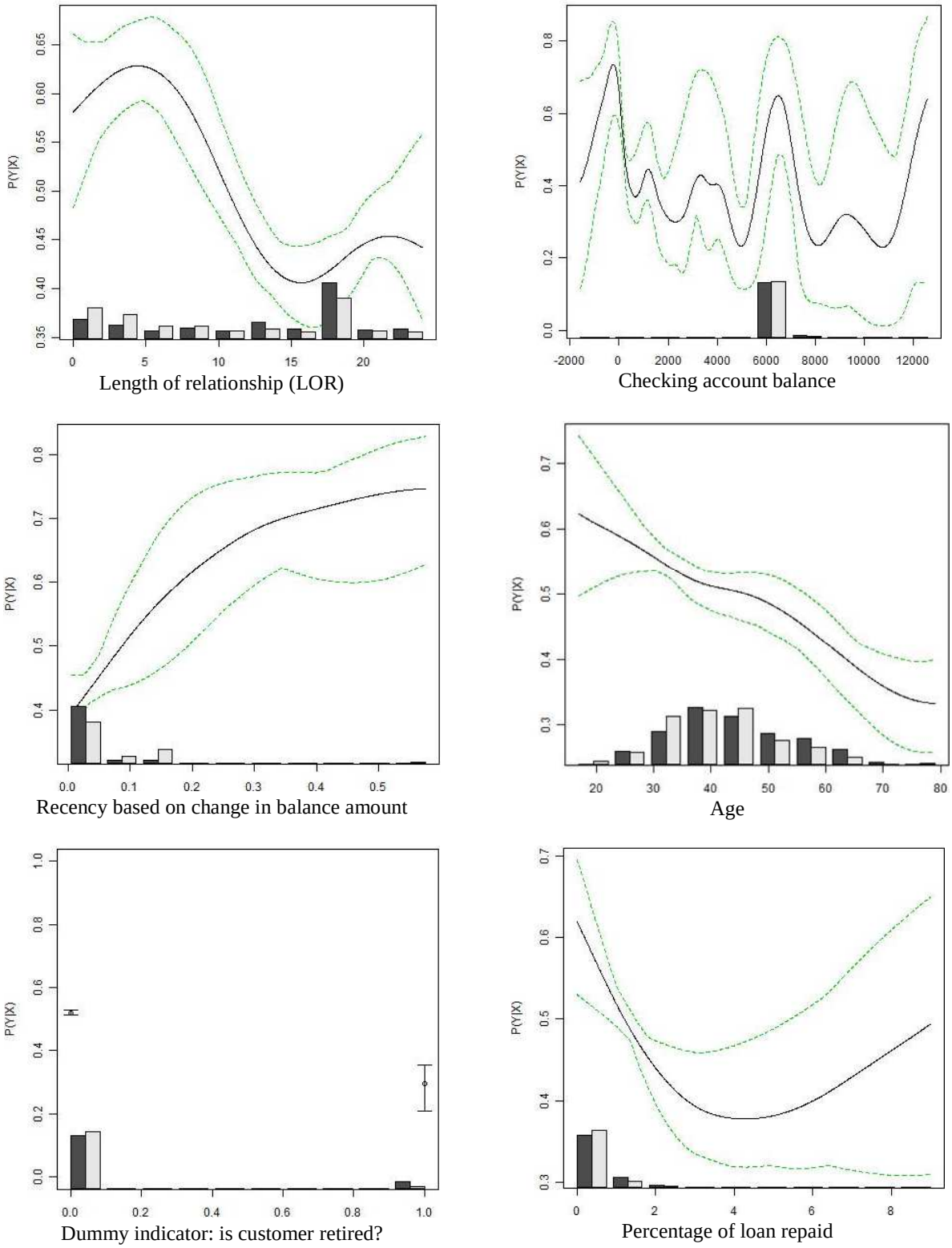

Figure 3.: Bootstrap confidence intervals and average trends for a selection of predictive features