Decision Support Systems 82 (2016) 26–34

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

The added value of Facebook friends data in event attendance prediction

Matthias Bogaert [a], Michel Ballings [b,*], Dirk Van den Poel [a]

[a] Ghent University, Department of Marketing, Tweekerkenstraat 2, 9000 Ghent, Belgium
[b] The University of Tennessee, Department of Business Analytics and Statistics, 916 Volunteer Blvd., 249 Stokely Management Center, 37996 Knoxville, TN, USA

ARTICLE INFO

Article history:
Received 23 April 2015
Received in revised form 20 November 2015
Accepted 21 November 2015
Available online 28 November 2015

Keywords:
Facebook
Network data
Events
Predictive models
Social media

ABSTRACT

This paper seeks to assess the added value of a Facebook user's friends data in event attendance prediction over and above user data. For this purpose we gathered data of users that have liked an anonymous European soccer team on Facebook. In addition we obtained data from all their friends. In order to assess the added value of friends data we have built two models for five different algorithms (Logistic Regression, Random Forest, Adaboost, Neural Networks and Naive Bayes). The baseline model contained only user data and the augmented model contained both user and friends data. We employed five times two-fold cross-validation and the Wilcoxon signed rank test to validate our findings. The results suggest that the inclusion of friends data in our predictive model increases the area under the receiver operating characteristic curve (AUC). Out of five algorithms, the increase is significant for three algorithms, marginally significant for one algorithm, and not significant for one algorithm. The increase in AUC ranged from 0.21%-points to 0.82%-points. The analyses show that a top predictor is the number of friends that are attending the focal event. To the best of our knowledge this is the first study that evaluates the added value of friends network data over and above user data in event attendance prediction on Facebook. These findings clearly indicate that including network data in event prediction models is a viable strategy for improving model performance.

1. Introduction

Facebook is a large-scale social media platform with 1.55 billion monthly active users and 894 million daily active users [32] and has grown to the point of becoming an important channel for social contact [30,64] and product promotion [15,11]. Among other things, it enables businesses to schedule meetings and gatherings using a functionality called Facebook Events [33]. With Facebook Events promoters can manage event participants and notify participants' friends [33]. The downside of this functionality's popularity is that many companies are using it and hence there are a lot of co-occurring events [5]. In order to make a user's Facebook experience more enjoyable and to avoid information overload, Facebook predicts whether or not the user will attend the event. It logically follows then, that a very important task is to try and make those predictions as accurate as possible.

While there is a considerable body of research on event modeling in other fields and networks [23,51,67], little research has been done on Facebook Events specifically, despite the platform's aforementioned size and success. A very common and important research question in event predictions pertains to the importance of specific sets of predictors. If a set of predictors does not improve predictive performance it should be removed from the model so as to prevent from slowing down the modeling process. In the case of Facebook data, a meaningful question is whether friends data should be included in the model. If a typical user has 300 friends, and we have 1000 users in our sample, including friends data would imply analyzing an additional 300,000 users. If these data do not improve the predictive model significantly, adding them would imply an unnecessary lag in the modeling process.

This paper seeks to fill this gap in literature by studying the added value of friends data over and above user data in event prediction on Facebook. We focus on predicting whether a soccer fan will attend a given event or not. For this purpose we developed a Facebook application to extract a user's data along with a user's friends data. In total 5010 users and 1,102,573 friends authorized our application to collect their relevant data. To investigate the added value of friends data we build and compare two models. The first one includes only user data and the second one includes both user data and friends data. The difference in performance between both models yields the added value of friends data. If the performance increase is significant, friends data should be incorporated in future models. If not, it should be excluded for the sake of parsimony and execution speed. Furthermore, we benchmark these two models for five state-of-the-art classification algorithms namely Logistic Regression, Random Forest, Adaboost, Neural Networks and Naive Bayes.

In the remainder of this article we first provide an overview of extant literature. Second, we provide details on the methodology. Third, we elaborate on our findings and their implications. Finally, we discuss limitations and avenues for future research.

* Corresponding author.
E-mail addresses: Matthias.Bogaert@UGent.be (M. Bogaert), Michel.Ballings@utk.edu (M. Ballings), Dirk.VandenPoel@UGent.be (D. Van den Poel).

http://dx.doi.org/10.1016/j.dss.2015.11.003
0167-9236/© 2015 Elsevier B.V. All rights reserved.

Table 1
Overview of events literature.

Study	Case	Facebook data	User data	Network data	Added value network
Mynatt and Tullio [68]	Company meetings		X		
Horvitz et al. [47]	Company meetings		X	X	
Lovett et al. [63]	Company meetings			X	
Tullio and Mynatt [80]	Company meetings			X	
Daly and Geyer [23]	Company meetings		X	X	
Pessemier et al. [72]	Cultural activities	X	X		
Coppens et al. [20]	Cultural activities	X	X		
Lee [58]	Cultural activities		X	X	
Kayaalp et al. [51]	Concerts		X	X	
Minkov et al. [67]	Academic events		X		
Klamma et al. [52]	Academic events			X	
Zhang et al. [87]	Facebook events and Academic events	X	X	X	
Our study	Facebook	X	X	X	X

2. Literature overview

The addition of social network information has proven to achieve good performance in several applications (other than event prediction). On Facebook, examples can be found in the field of activities [86], users [19], movies [74] and interests [42]. On Twitter, network information has proven to be useful in predicting user behavior [71] and tweet popularity [46,79]. On other social network sites, including social relationship data has improved results in peer recommendations [61,85]. Despite the importance of network data in social media prediction, literature on event attendance prediction remains scarce as discussed in the next paragraph.

Literature on event prediction can be classified according to the data that is used in the model. In this typology there are three classes: predictive models that are enriched with (1) user data (e.g., [67]), (2) network data (e.g., [80]), or (3) both user and network data (e.g., [47]). User data are defined as specific profile characteristics that represent the preferences of the user. Examples are the interests of the user [20], demographics [72] and past event-history [87]. Network data are defined as data that contain information about the user's social network. Examples are the number of peers that are attending the event [63], and event preferences of their friends [52].

Table 1 provides a literature review on event prediction literature with a focus on data sources and platforms. It is clear that, to the best of our knowledge, our study is the only one that evaluates the added value of network data over and above user data on Facebook. Even more so, Table 1 indicates that the added value of network data has not been evaluated on other platforms either. The study of Zhang et al. [87] is of special interest as it focuses on user and network data from Facebook, just as our study.

In their research, three large groups of event predictors and corresponding approaches are proposed. First, in a similarity-based approach (SBA) they use event profile data (e.g., topic and location) and user profile data (e.g., interests and activity history) to compute similarities. Second, in an approach that they call the relationship-based approach (RBA), they include network data such as whether or not friends will attend the event. Third, in their history-based approach (HBA) they add users' historic event attendances. The authors subsequently propose a hybrid approach (SRH), which is a combination of the three other approaches and data sources. Their research concludes that indeed the combination of all three data sources (SRH) yields the most precise and accurate results, followed by RBA, SBA and HBA.

Just as in the other studies in Table 1, Zhang et al. [87] do not assess the added value of network data over and above user data. They only investigate the difference in precision between the hybrid approach and the other methods. They have not made pairwise comparisons between the three different data sources by solely comparing the combined sources with the individual sources. Their results suggest that the SRH approach significantly outperforms the three other approaches. For the three other models, their study only states that they perform better than a random model, thereby neglecting to investigate whether the models are significantly different from one another. With this approach, they are also unable to detect whether the increase in performance is due to network data or not. Regarding these results, it is clear that their study does not incorporate a comprehensive assessment of the added value of friends data. Furthermore, their research doesn't disclose which variables should be included or not in order to make predictive models as efficient as possible. Such assessment is necessary because including friends data implies a certain computational cost. From that perspective, one could argue that including friends data is only reasonable if the results improve significantly.

To fill this gap in literature, this study focuses on one such pairwise comparison: it will assess the extra value of friends data over and above user profile data. By doing so, we can precisely isolate the impact of our network variables. To make the comparison we build two models, a first one — the baseline model — containing user predictors and a second one — the augmented model — with network predictors in addition to the user predictors[1]. Examples of user variables are the number of groups, posts, events and photos. Network variables are operationalized as the number and percentage of friends that are attending a certain event. Furthermore, we assess several algorithms to determine if the increase in prediction performance is consistent.

We have three hypotheses about why network variables might improve event recommendations. First, the theory of homophily [3,65,82], also called endogenous group formation [44], states that like-minded people group together and often share the same tastes and opinions [41,78,84]. Second, and closely related to homophily, is the idea of social influence [35] and selection [65]. The former states that persons tend to follow the decisions of their peers [21]. The latter states that people mostly select friends who are similar [34]. Third, network variables capture the concept of trust. Trust-based theories state that friends' actions will be more easily followed and hence be more accurate if they are sourced from a trustworthy connection or friend. This is especially important in the case of events because trust and acceptance are critical factors for actual event attendance [48,59,70]. In addition, Facebook friends are often real-life friends [30] and can therefore be deemed trustworthy ties.

Various studies confirm the result that adding social relationships increases the performance of predictive models in Facebook applications relating to romantic partnership [6] and link prediction [50]. Chang and Sun [18] also found evidence that network variables play an important role in location check-ins. Using Facebook data, they conclude that previous check-in behavior of the user and the check-ins of friends are the most relevant predictors of check-in behavior. Thus, if a friend is attending a Facebook Event, a user may be more inclined to attend as well. It is clear that from the theories of homophily, social influence and selection that the probability of adopting a given behavior

[1] In the remainder of this paper, we will always refer to the model with only user data as the baseline model and to the model with user and friends data as the augmented model.

Table 2
Overview of predictors.

Variable category	Variable
Demographic and identification variables	Age
	IND(gender)
	IND(email)
	IND(website)
Geographical variables	IND(hometown)
	IND(location)
Professional/ educational variables	COUNT(languages)
	COUNT(work)
	COUNT(educations)
	IND(education type)
Social variables	COUNT(family)
	IND(sexual orientation)
	IND(relationship status)
	COUNT(OF 23 family relationship types) (e.g., aunt)
	COUNT(friend connections)
	COUNT(groups)
Personal variables	COUNT(favorite teams)
	COUNT(sports)
	COUNT(television)
	COUNT(music)
	COUNT(movies)
	COUNT(books)
	COUNT(activities)
	COUNT(inspirational people)
	COUNT(interests)
	COUNT(OF 10 television categories) (e.g., show)
	COUNT(activity category)
	IND(OF 14 interests) (e.g., design)
	IND(OF 23 sports) (e.g., fitness)
	IND(bio)
	IND(quotes)
	IND(political)
	IND(religion)
General Facebook account variables	Length Facebook membership
	Recency last update = REC(profile update created)
	MEAN(album privacy)
	Profile completeness = SUM(IND(37 profile variables))
	IND(username)
	Time ratio = SDIET(all actions)/MIET(all actions)
Likes	COUNT(OF 188 like categories) (e.g., musician/band)
	COUNT(likes)
	REC/MIET/SDIET(like created)
	COUNT(posts likes)
Statuses	COUNT(statuses)
	REC/MIET/SDIET(status updated)
Photos	COUNT(photos)
	REC/MIET/SDIET(photo created)
Videos	COUNT(videos)
	REC/MIET/SDIET(video created)
Albums	COUNT(albums)
	REC/MIET/SDIET(album created)
Events	COUNT(events)
	MIET/SDIET(event created)
	IND(event time == start day)
	IND(event time == end day)
	IND(event time == month)
	IND(event time == season)
	IND(event time == year)
	IND(event time == weekend)
	IND(event location)
	LENGTH(event time)
Links	COUNT(links)
	REC/MIET/SDIET(link created)
Check-ins	COUNT(check-ins)
	REC/MIET/SDIET(check-in created)
	IND(check in app)
Notes	COUNT(notes)
	REC/MIET/SDIET(note created)
Games	COUNT(games)
	REC/MIET/SDIET(game created)
Tags	REC/MIET/SDIET(photo user tags)
	COUNT(video user tags)
	COUNT(photo user tags)
	COUNT(check-in user tags)
	REC/MIET/SDIET(video user tags)
Comments made	REC/MIET/SDIET(photos/albums/statuses/links/check-ins comments)
	COUNT(photos/albums/statuses/links/check-ins comments)
Comments received	REC/MIET/SDIET(photos/albums/statuses/links/check-ins comments received)
	COUNT(photos/albums/statuses/links/check-ins comments received)

rises when others in one's network have already adopted that behavior [2,4,21].

To summarize, we found strong indications in extant literature that the augmentation of user data with network data can improve the predictive power of our model. To the best of our knowledge this is the first study to look into this issue for the social network site Facebook. In the next section of this paper we will elaborate on our methodology.

3. Methodology

3.1. Data

In order to extract data from Facebook, we made a Facebook application for a European soccer team. To stimulate usage of our application we offered a prize (i.e., a signed shirt of a famous soccer player) to the participants and asked three questions to determine the winner. The application was advertised several times on the Facebook fan page of the soccer team. In addition, the application was added to the main page tabs for added visibility. Application users were presented with an authorization box in which they had to give their permission before the data were gathered from their profile. The data were collected between May 7, 2014 and June 9, 2014. In total we collected 5315 event observations (2368 unique events) from 978 users. We also gathered data of 194,639 friends, which are used for the creation of network predictors. The response variable in our models is binary, with the value 1 if users indicated that they were attending and 0 otherwise. Of all our event observations attendance is 78.2%.

3.2. Predictors

The user-related variables are summarized in Table 2. The ‘Like’ variables in our study only relate to likes generated by users. ‘Likes’ are also only available for a page, band, app, or leisure activity. In the photo and video variables the affix ‘created’ points out that the photo or video was uploaded, or created and immediately uploaded with the Facebook app. Tags in photos refer to tags of the user himself/herself. The variable ‘username’ captures if a user has upgraded his/her username to an alphabetic identifier from the standard numeric identifier. Due to regulations on Facebook, we could only gather the twenty-five last albums, photos, videos, links, status updates, notes and check-ins. In order to alleviate this restraint, we calculated the frequency by time as to no users in our database reached this restriction. For the last seven days, we computed the frequency of status updates, photo and link uploads, for the last four months album uploads and check-ins were computed, and for the last year notes and video uploads were computed.

In Table 2 IND: indicator, COUNT: frequency, REC: recency, MIET: mean inter-event time, SDIET: standard deviation inter-event time, and LENGTH: length of the time interval. MIET is the mean time that passes between two subsequent events (e.g., album uploads). SDIET is defined as the standard deviation of the time between two subsequent events.

Within our user variables, we are particularly interested in event-related user variables. The majority of the user-event variables are calculated as time indicator variables (see Table 2 section Events). These variables resolve to 1 if the event took place at a certain time and 0 otherwise. Applying this logic we computed dummies for the day of the week (for both start day and end day of the event), the weekend, the month, and the season. Other event variables such as the duration and location were also added. We denote that we didn't include dummies for the type of event, since our database mainly contains soccer events. Other popular events were related to parties and festivals. In total we calculated 540 user variables for our first model.

In order to create our second model, we augmented the first model with friends-related variables. Next to our users we also gathered data from their friends (194,639). We computed five variables that are important for the event that we are predicting, namely the total and relative number of friends that are going to the focal event and the average number of total, soccer, and team events the user's friends attended.

3.3. Classification algorithms

In this section, we elaborate on the choice of our classification algorithms. In total, we use five single classifier and ensemble techniques: Naive Bayes (NB), Logistic Regression (LR), Neural Networks (NN), Random Forest (RF), and Adaboost (AB). Naive Bayes is the least complex algorithm because it only estimates the joint probability $p(x,y)$. In contrast Logistic Regression estimates the conditional probability $p(y|x)$ and this can result in better performance [69]. Neural networks are similar to logistic regression if the logistic activation function is employed but add additional complexity by incorporating a hidden layer. This increases flexibility and this can result in better performance. Random Forest adds additional complexity by using an ensemble of trees. Trees are inherently nonlinear and incorporate interactions. Using many trees and combining them often improves performance. Finally adaptive boosting (Adaboost) adds complexity by incorporating a weighting mechanism that focuses on incorrectly classified instances in the previous iteration. We will evaluate the added value of network variables for all these algorithms. This will allow us to draw conclusions across a range of complexity levels. In the following paragraphs we will provide more details about the different algorithms.

3.3.1. Naive Bayes

We use the original Naive Bayes algorithm as a method for probabilistic classification. This method applies Bayes' Theorem to classify new observations and naively assumes conditional class independences [55]. Despite the fact that the conditional independence assumption is rarely satisfied, it achieves reasonable performance and low computation times [55]. Several authors have tried to overcome the problem of conditional dependency by introducing randomness such as random feature selection and bagging [56,73]. The function *naiveBayes* was used from the R-package *e1071* [66]. Gaussian distributions were assumed for the predictors.

3.3.2. Logistic Regression

We use regularized Logistic Regression with the lasso approach to cope with overfitting. The lasso (least absolute shrinkage and selection operator) sets a bound on the sum of the absolute values of the coefficients forcing the coefficients to shrink towards zero [49, p. 219]. In this regard, the value of the shrinkage parameter λ determines the amount of shrinkage. The higher the value of λ the smaller the coefficients will be. We use cross-validation to determine the optimal shrinkage parameter. The statistical R-package *glmnet* by Friedman et al. [37] is used to create our model. We set the parameter α to 1 to obtain the lasso approach and we set the *nlambda* parameter to 100 (default) to compute the sequence of λ.

3.3.3. Neural Networks

We use the feed-forward artificial neural network optimized by BFGS with one hidden layer. This approach is considered much more reliable, efficient and convenient than back propagation and has proven to be sufficient in a variety of cases [27]. Before implementing the neural network, we rescale the numerical variables to $[-1, 1]$ [12]. The binary variables are disregarded and coded as {0, 1}. Scaling is necessary to avoid local optima and numerical problems and to ensure efficient training. The statistical R-package *nnet* is used to build the neural network [75]. The network weights are randomized at the start of the iterative procedure [76, p. 154]. This implies that the results change for subsequent neural networks, which mimics the development of the human brain [81]. We follow the recommendations of Ripley [76, p. 149] and set the *entropy* parameter to the maximum likelihood method. The *rang* parameter which manages the range of initial random weights was set to 0.5 (default). The parameters *abstol* and *rel* were also left at their default $1.0e^{-4}$ and $1.0e^{-8}$. Weight decay was used to avoid overfitting [27] and the maximum number of weights (*MaxNWts*) and maximum number of iterations (*maxit*) were set at a very large number (5000) in order to avoid early stopping. Finally a grid search was performed in order to determine the weight decay and the number of nodes in the hidden layer [27]. In accordance to Ripley [76, p. 163, p. 170] we sequenced over all combinations of $decay = \{0.001, 0.01, 0.1\}$ and $size = [1,\ldots, 20]$ to determine the optimal combination.

3.3.4. Random Forest

Random Forest combines bagging with random feature selection to build an ensemble of trees [16]. Each tree is grown on an independent bootstrap sample and at each node of each tree a randomly selected subset of features is evaluated [16]. To grow the ensemble all the trees are aggregated by means of majority voting [16]. As a result, Random Forest copes with the instability and the suboptimal performance of decision trees [29]. Two parameters have to be provided: the number of trees and the number of predictors randomly selected at each node of each tree [57,28]. We follow the recommendation of Breiman [16] to use a large number of trees (500) and the square root of the total number of predictors as the number of predictors to be evaluated at each node. We use the statistical R-package *randomForest* provided by Liaw and Wiener [62].

3.3.5. Adaboost

The original Adaboosting algorithm [36] sequentially reweights the training data [45, pp. 337–340]. In each iteration the observations that were misclassified in the previous iteration are given more weight, whereas the correctly classified observations are given lower weight. Hence, instances that are hard to classify are given more importance in each iteration. The final model is a linear combination of all the previous models [45, pp. 337–340]. We use stochastic boosting, one of the most recent boosting variants which introduces randomness as an integral part of the procedure [39]. Randomness is induced by making bootstrap samples in which the propensity of an observation being selected is proportional to the current weight [39]. There are two important parameters: the number of iterations and the number of terminal nodes in the base classifier. In accordance with Friedman [39] we determine the number of terminal nodes by setting the maximum depth of the trees to 3 and we set the number of iterations to 500. To fit our model we use the statistical R-package *ada* [22].

3.4. Performance evaluation

We use the area under the receiver operating characteristic curve (AUC or AUROC) to evaluate the performance of our classification models. AUC is argued to be an objective performance measure for classification problems by several authors [54]. The receiver operating characteristic curve (ROC) is a graphical representation of the sensitivity against one minus specificity for all possible cut-off values [43]. AUC is a more adequate measure of classifier performance than PCC (percentage correctly classified) [7] whenever the cut-off value that will be used at model deployment is unknown, because AUC evaluates the entire range of cut-off values [8]. AUC is defined as follows:

$$\mathrm{AUC} = \int_0^1 \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \mathrm{d}\frac{\mathrm{FP}}{(\mathrm{FP}+\mathrm{TN})} = \int_0^1 \frac{\mathrm{TP}}{\mathrm{P}} \mathrm{d}\frac{\mathrm{FP}}{\mathrm{N}} \qquad (1)$$

with TP: True Positives, FN: False Negatives, FP: False Positives, TN: True Negatives, P: Positives (event), and N: Negatives (non-event).

AUC is restricted between the values of 0.5 and 1, where the former denotes that the model does not perform better than random and the latter indicates a perfect prediction [43]. If the AUC is below 0.5 in the test set, this is a strong indication of overfitting.

3.5. Cross-validation

We use five times two-fold cross-validation (5 × 2 cv) to make sure our results are not overly optimistic or pessimistic [25,1]. 5 × 2 cv starts by randomly dividing the sample in two parts where each part is used once as a training sample and once as a test sample. This process is repeated five times and results in 10 AUCs per model [25]. We take the median of the results to obtain the overall AUC of our models. As a measure of dispersion, the interquartile range (IQR) is used.

In order to test whether two models are significantly different from each other we follow Demšar's [24] suggestion to use the Wilcoxon signed rank test [83]. The Wilcoxon signed-rank test [83] is a non-parametric test that ranks the differences in performance of two models while ignoring the signs. Ranks are assigned from low to high absolute differences, and equal performances get the average rank. The ranks of both the positive and negative differences are summed and the minimum of those two is compared to a table of critical values. To be significant the smallest sum of ranks should be smaller than the critical value.

This test is considered safer than a parametric *t*-test because the assumptions of normality and homogeneity of variance [24] do not need to be met. However, when the assumptions of a *t*-test can be satisfied, the Wilcoxon signed rank test has less power than a paired *t*-test. When the sample size equals 10 verifying normality and homogeneity is problematic and thus the Wilcoxon signed rank test is preferred [24].

3.6. Variable importance evaluation

Because we are using a lot of predictors in our sample, it is important to know which variables have great predictive power [77]. One way to do so is by calculating the variable importance. In tree-based methods such as Random Forest we can evaluate the importance of our predictors by using the total decrease in node impurities from splitting on the variable, averaged over all trees. The Gini index is used as a measure of node impurity [17]. The importances are then averaged over the 10 folds by taking the median of the 5 × 2 cv variable importances. We used the *importance* function in the *randomForest* package [62].

3.7. Partial dependence plots

Partial dependence plots allow one to graphically depict the relationship between an independent and a dependent variable, after eliminating the average effect of the other independent variables [38,40]. This is analogous to multiple linear regression of y on all x_j, where the coefficient x_1 accounts for the effect of x_1 on y with the other variables kept constant. Partial dependence plots are mostly used on decision tree-based methods and allow one to gain insight in how classification variables relate to the most important predictors [40,45, pp. 369–370]. In order to create a partial dependence plot we follow the method described by Berk [14, p. 222].

For each value v in the range of a predictor x we create a novel data set where x only takes on that value. All the other variables are left untouched. Next, for each novel data set, we score all the instances using a Random Forest model that is built on the original data. Subsequently the mean of half the logit of the predictions is calculated yielding one single value for all instances called p. The final step in creating the partial dependence plot is plotting all the values v of x against their corresponding p. All partial dependence plots are five times twofold cross-validated using the *interpretR* R-package [10].

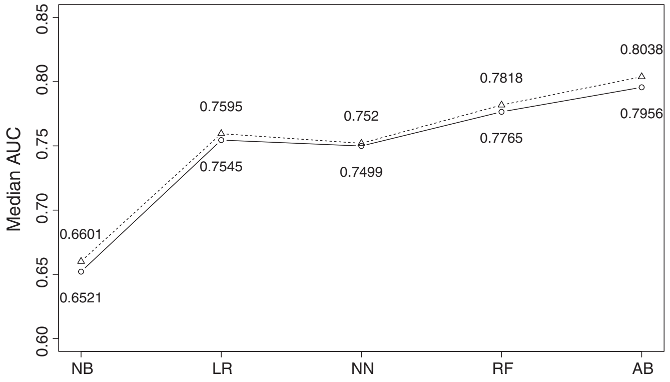


Fig. 1. Cross-validated AUC. The solid line represents the baseline model, the dashed line the augmented model. NB = Naive Bayes, LR = Logistic Regression. NN = Neural Networks. RF = Random Forest. AB = Adaboost.

4. Discussion of results

4.1. Model performance

The cross-validated results are summarized in Fig. 1 and Table 3. The main research question of this study was to assess if friends (i.e., network) data add value over and above user data in event prediction. We find that the inclusion of network variables results in an improvement of the AUC for all our classifiers. For the baseline model the AUC ranges from 65.21% to 79.56%, for the augmented model from 66.01% to 80.38%. The increase in AUC ranges from 0.21%-points to 0.82%-points. Fig. 1 also reveals that Adaboost (AB) is the top performing algorithm, followed by Random Forest (RF), Logistic Regression (LR), Neural Networks (NN) and Naive Bayes (NB). However, for computational reasons one might prefer RF since it allows parallel execution whereas AB is sequential in nature.

The Wilcoxon tests (Table 3) indicate that the results are significantly different for three out of five classifiers. The results show a significant difference on the 1% significance level for RF, AB and NB and on the 10% significance level for LR. We found no significant difference for our NN classifier. Adding friends data results in a slight increase in interquartile range (IQR; Table 4). Nevertheless the IQRs are low for all classifiers, indicating that all classifiers have stable results. The IQR also confirms that Adaboost is the top performer because it has the smallest IQR. These findings confirm our hypothesis that Facebook friends data can significantly improve the predictions in event attendance prediction systems. It has to be noted though that for some classifiers results are not significant.

4.2. Predictors

In order to uncover what the main drivers of predictive performance are we first look at a scree plot of the predictors (Fig. 2). In the scree plot, the 200 top predictors of the model with friends data are plotted against the median 5 × 2 cv mean decrease in Gini in a descending order. It is clear from this plot that predictors with rank higher than twelve only add little to our predictions. Hence, we focus on the top twelve predictors in the rest of this discussion.

Table 3
Summary of cross-validated median AUC.

	NB	LR	NN	RF	AB
Base model	0.6521	0.7545	0.7499	0.7765	0.7956
Augmented model	0.6601	0.7595	0.7520	0.7818	0.8038
Wilcoxon test	V = 0	V = 10	V = 21	V = 0	V = 0
	p < 0.01	p < 0.10	p < 0.6	p < 0.01	p < 0.01

Table 4
Summary of cross-validated median IQR.

	NB	LR	NN	RF	AB
Base model	0.0039	0.0046	0.0086	0.0039	0.0035
Augmented model	0.0072	0.0160	0.0170	0.0057	0.0047

Table 5
Median cross-validated variable importance.

No.	Variable name	Median decrease Gini
1	IND(event time == end day Mon)	20.107
2	IND(event time == start day Sun)	19.460
3	IND(event time == start month May)	18.023
4	COUNT(events)	16.447
5	IND(event time == weekend)	13.379
6	PERCENTAGE(friends event attending)	12.600
7	COUNT(friends event attending)	12.010
8	IND(event time == end day Sun)	10.814
9	IND(event time == start season Spring)	9.909
10	IND(event time == start day Sat)	9.878
11	IND (event time == start season Summer)	9.025
12	IND(event time == start month June)	8.146

Table 5 contains the importance of the top twelve predictors and Fig. 3 the partial dependence plots of selected variables. In Table 5 we observe that most of the top predictors are related to the timing of the event and the friends variables. The most important predictor of event attendance is whether the event ends on a Monday. In Fig. 3a we clearly observe a positive relationship between that predictor and event attendance. A plausible explanation can be found in the specific nature of our data. Major soccer events are mostly held on a Sunday. Hence event promoters on Facebook mostly set the ending of the event one day later (Monday). We also ran a plot (not shown) of whether the event starts on a Sunday and found the same positive relationship. Conversely, plots related to whether the event starts in the weekend depict a negative relationship with the probability of attending (not shown). This reinforces our explanation that important soccer games take place on Sunday (and their end time is always set to Monday on Facebook), minor soccer games are mostly played on other days in the weekend and receive less public attention. We denote that events with their end time on Monday, were not denoted as weekend events, this explains the negative relationship with the response variable. In Fig. 3b, we note a positive relationship between whether the event starts in the month May and event attendance. The month May is also traditionally the play-off season in European soccer leagues. The same logic can be applied to explain the positive relationship between the Spring season and our response variable, since the month May lies in Spring season (not shown).

The results in Table 3 and Fig. 1 already clearly indicated that friends data improve model performance. These results are substantiated in that network predictors (the total and relative number of friends that indicated their attendance) are among the top ten predictors (sixth and seventh variable in Table 5). Looking at the partial dependence plots in Fig. 3d and e, we first observe a positive and afterwards a negative effect, when more friends (more than 12 or 1.8%) are attending. The main reason for this relationship can be found in the News Feed Algorithm (NFA). Each time a friend interacts with something on Facebook, such as replying to an event invitation, a user gets notified in his or her News Feed. However, in order to avoid information overload Facebook limits the number of notifications for the same event. If a lot of friends are going, the NFA will stop propagating the message through the News Feed due to anti-spam regulations [31]. This implies that the probability of attending will first rise with every (close) additional friend that indicates attendance, and then decrease to normal once a given number of friends has been reached. Generally, these findings are partially different from the findings of Aral et al. [2] and Backstrom [4] who state that the adoption probability rises when friends already adopted. This partial difference is undoubtedly due to the many changes the NFA has undergone since these studies have been published. For example, Facebook recently increased their anti-spam regulations by hiding promotional posts in the user's News Feed [26]. In addition, Facebook users now have more control over what they see in their News feed [26].

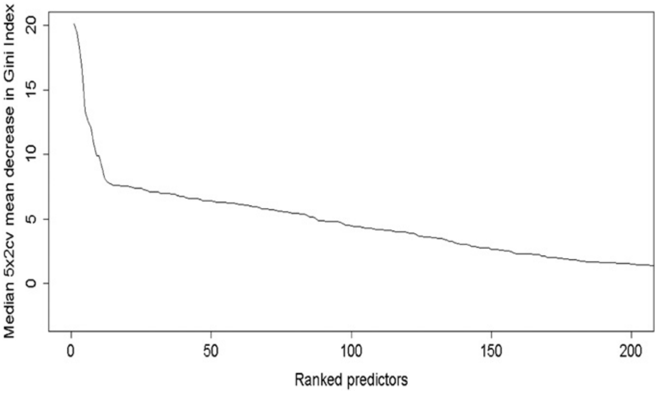


Fig. 2. Scree plot of the 200 most important predictors.

The total number of events the user attended (Fig. 3c) is constant in the beginning and afterwards negatively related to our dependent variable. This implies that people will have an equal propensity of attending until they attend too many events. This is a plausible relationship. People don't have an unlimited amount of time to attend events. The more events the user attends, the less time he or she has to attend other events.

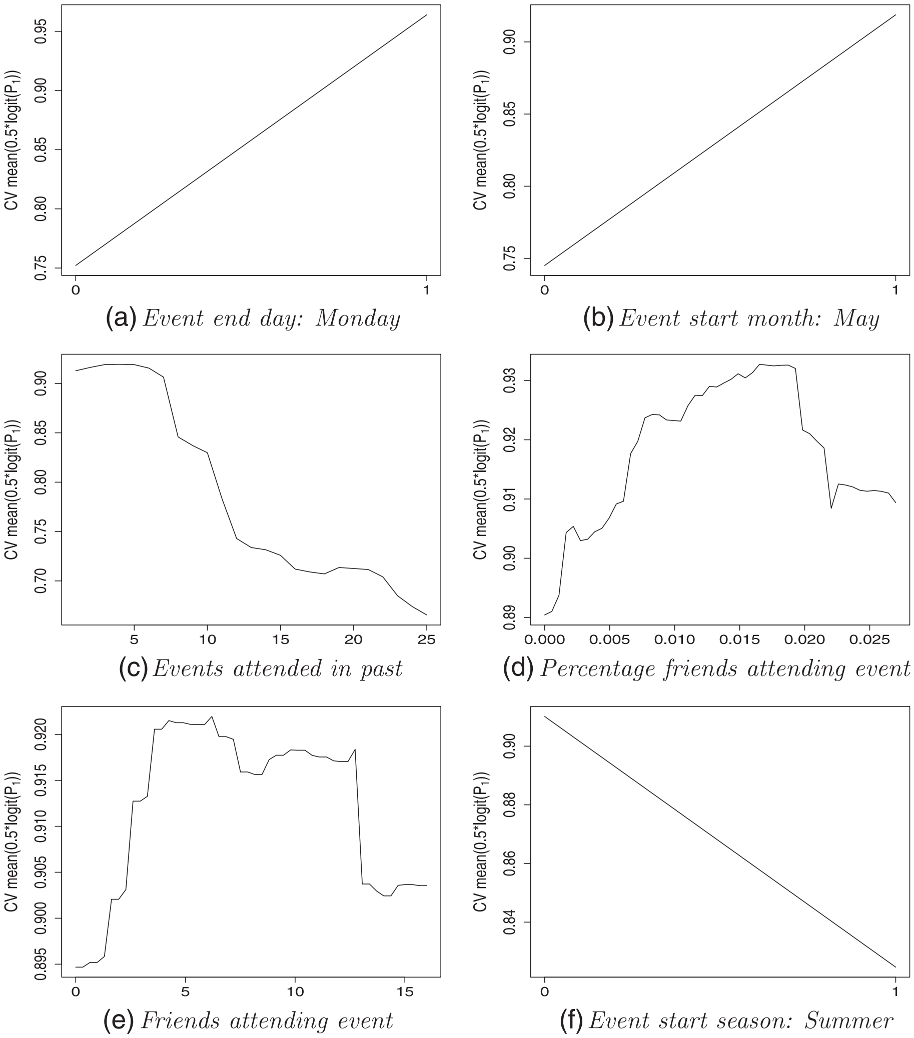
Finally, the predictors related to whether the event takes place in the Summer are negatively related with our dependent variable (see Fig. 3f). Again, we refer to the specific nature of our data. In the Summer, the soccer season has ended and hence there are no important soccer events taking place. A diagnostic plot of whether the event is held in June and our response variable supports our hypothesis (not shown).

5. Conclusion and practical implications

In this study we set out to (1) evaluate the added value of a Facebook user's friends data over and above user data in event predictions and (2) gain more insight in the top predictors of event attendance.

The results suggest that augmenting the data with network variables increases the AUC between 0.22%-points and 0.82%-points. This is in line with the conclusion of Benoit and Van den Poel [13] where the AUC also significantly rose with the inclusion of network effects. The top performing algorithm is Adaboost, closely followed by Random Forest. This is similar to findings of Ballings and Van den Poel [9], where Adaboost and Random Forest came out as the best-in-class classifiers in a social media application. The top predictors are mainly related to the event time such as the start day and end day of the event. Network variables were also top predictors of event attendance. More specifically the absolute and relative number of friends that are attending the event are very important. We also provided a list of the top twelve predictors in Table 5 and partial dependence plots in Fig. 3.

Our findings provide important insights for (1) Facebook Inc., (2) event promoters alike that want to increase the number of attendees, and (3) companies that want to build event prediction apps on Facebook. Facebook Inc. could incorporate our findings to adapt the News Feed Algorithm (NFA) for events. Recently, Facebook has fine-tuned the NFA algorithm to give more control to the user as to what he or she wishes to see and not to see in his or her News Feed [26]. Most of these updates are related to Facebook Pages and spam. Events however, are not specifically mentioned. A useful update could be to ask users to which extent they want to be informed about events,



(a) *Event end day: Monday*

(b) *Event start month: May*

(c) *Events attended in past*

(d) *Percentage friends attending event*

(e) *Friends attending event*

(f) *Event start season: Summer*

Fig. 3. Partial dependence plots.

thereby giving them more control. Users seem to be positively influenced by a certain group of attendees. For example, users could indicate a threshold for the number of friends attending an event that controls when events appear and disappear in their News Feed.

Also event schedulers and promoters on Facebook can utilize our findings. Event organizers would benefit from more explicitly providing information about the attendees of an event. They can, for example, send invitations to friends of the attendees and include in the notification the number of friends that will attend.

Companies that want to create a Facebook app for event scheduling and promoting can also benefit from our results. We have proven that the inclusion of friends data significantly improves the accuracy of the prediction system. For example, when building an app that recommends events to a certain user, one could calculate which events are attended by his or her friends to generate more accurate predictions.

6. Limitations and future research

First, this study is limited because of selection effects. We extracted our data with a custom-built Facebook app via the Facebook Page of a European soccer team by offering a chance to win a prize. It might be the case that some users were not interested in this prize and hence were not willing to share their data. Another way of collecting data from Facebook is web-crawling as proposed by Lampe et al. [53] and Lewis et al. [60]. Nevertheless, web-crawling also suffers from the limitation that data cannot be extracted from private Facebook profiles. Generally, the collected data from web-crawling and a Facebook application largely overlap. Our approach is less intrusive since we ask permission from the user and provide a 'rules and regulations' section in the app with our contact information. We also ensured the user that we anonymize all information and do not extract private messages. Finally, we also provided a disclaimer explaining the purpose of our academic research. Therefore, we believe that our approach is superior to web-crawling. Since we only limit our data to a subsample, our results do suffer from generalizability issues. However, regardless of this limitation our study is the first to investigate the added value of friends data in event attendance prediction. Hence, we consider this study a valuable contribution to literature. An avenue for future research can be to obtain a broader sample and more representative results.

A second limitation is that some of our predictors are limited in the number of values. Facebook only allows to extract the 25 most recent entries for specific variables. To mitigate this problem we computed the frequency of a specific time period as to no variable reaches this limit. The frequency of status updates, photo uploads and link uploads was calculated for the last 7 days, album uploads and check-ins for the last 4 months and video uploads and notes for the last year.

A third limitation is that we only include a limited number of friends variables in our analyses, mostly the ones that are related to the focal event. Following Zhang et al. [87], a possible avenue for future research could be to add more friends variables. We could investigate which type of predictors yields the biggest increase in model performance. This would help practitioners understand which elements in event attendance prediction systems make them as accurate and efficient as possible.

References

[1] E. Alpaydin, Combined 5 × 2 cv F test for comparing supervised classification learning algorithms, Neural Computation 11 (1998) 1885–1892.

[2] S. Aral, L. Muchnik, A. Sundararajan, Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks, Proceedings of the National Academy of Sciences 106 (2009) 21544–21549.

[3] S. Aral, D. Walker, Tie strength, embeddedness, and social influence: a large-scale networked experiment, Management Science 60 (2014) 1352–1370.

[4] L. Backstrom, Group formation in large social networks: membership, growth, and evolution, KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press 2006, pp. 44–54.

[5] L. Backstrom, News Feed FYI: A Window Into News Feed, https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed. 2013.

[6] L. Backstrom, J. Kleinberg, Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on Facebook, Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing, ACM, New York, NY, USA 2014, pp. 831–841.

[7] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen, G. Dedene, Bayesian neural network learning for repeat purchase modelling in direct marketing, European Journal of Operational Research 138 (2002) 191–211.

[8] M. Ballings, D. Van den Poel, Customer event history for churn prediction: how long is long enough? Expert Systems with Applications 39 (2012) 13517–13522.

[9] M. Ballings, D. Van den Poel, CRM in social media: predicting increases in Facebook usage frequency, European Journal of Operational Research 244 (2015) 248–260.

[10] M. Ballings, D. Van den Poel, R-package interpretR: Binary Classifier Interpretation Functions, 2015.

[11] M. Ballings, D. Van den Poel, M. Bogaert, Social media optimization: identifying an optimal strategy for increasing network size on Facebook, Omega 59 (2016) 15–25.

[12] M. Ballings, D. Van den Poel, N. Hespeels, R. Gryp, Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications 42 (2015) 7046–7056.

[13] D.F. Benoit, D. Van den Poel, Improving customer retention in financial services using kinship network information, Expert Systems with Applications 39 (2012) 11435–11442.

[14] R.A. Berk, Statistical Learning From a Regression Perspective, Springer, 2008.

[15] S. Bhagwat, A. Goutam, Development of social networking sites and their role in business with special reference to Facebook, Journal of Business and Management 6 (2013) 15–28.

[16] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32.

[17] L. Breiman, Manual on Setting Up, Using, and Understanding Random Forests V3.1, Statistics Department University of California Berkeley, CA, USA, 2002.

[18] J. Chang, E. Sun, Location 3: how users share and respond to location-based data on social networking sites, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media 2011, pp. 74–80.

[19] J. Chen, W. Geyer, C. Dugan, M. Muller, I. Guy, Make new friends, but keep the old: recommending people on social networking sites, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA 2009, pp. 201–210.

[20] S. Coppens, E. Mannens, T.D. Pessemier, K. Geebelen, H. Dacquin, D.V. Deursen, R.V.d Walle, Unifying and targeting cultural activities via events modelling and profiling, Multimedia Tools and Applications 57 (2011) 199–236.

[21] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, S. Suri, Feedback effects between similarity and social influence in online communities, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM 2008, pp. 160–168.

[22] M. Culp, K. Johnson, a.G. Michailidis, ada: An R Package for Stochastic Boosting, 2012.

[23] E.M. Daly, W. Geyer, Effective event discovery: using location and social information for scoping event recommendations, Proceedings of the Fifth ACM Conference on Recommender Systems, ACM, New York, NY, USA 2011, pp. 277–280.

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

[25] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (1998) 1895–1923.

[26] S. Dredge, Facebook squeezes 'overly promotional page posts' out of news feeds, http://www.theguardian.com/technology/2014/nov/17/facebook-page-posts-news-feeds. 2014.

[27] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics 35 (2002) 352–359.

[28] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.

[29] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97 (2002) 77–87.

[30] N.B. Ellison, C. Steinfield, C. Lampe, The benefits of Facebook "friends:" social capital and college students' use of online social network sites, Journal of Computer-Mediated Communication 12 (2007) 1143–1168.

[31] Facebook, News Feed FYI | Facebook Newsroom, https://newsroom.fb.com/news/category/news-feed-fyi/. 2014.

[32] Facebook, Company Info | Facebook Newsroom, https://newsroom.fb.com/company-info/. 2015.

[33] Facebook, Products | Facebook Newsroom, http://newsroom.fb.com/products/. 2015.

[34] X. Fang, P.J.H. Hu, Z.L. Li, W. Tsai, Predicting adoption probabilities in social networks, Information Systems Research 24 (2013) 128–145.

[35] L. Festinger, A theory of social comparison processes, Human relations 7 (1954) 117–140.

[36] Y. Freund, R.E. Schapire, et al., Experiments with a new boosting algorithm, ICML 1996, pp. 148–156.

[37] J. Friedman, T. Hastie, N. Simon, R. Tibshirani, glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, 2015.

[38] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of Statistics 1189–1232 (2001).

[39] J.H. Friedman, Stochastic gradient boosting, Computational Statistics and Data Analysis 38 (2002) 367–378.

[40] J.H. Friedman, J.J. Meulman, Multiple additive regression trees with application in epidemiology, Statistics in Medicine 22 (2003) 1365–1381.

[41] G. Groh, Recommendations in taste related domains: collaborative filtering vs. social filtering, Proc ACM Group'07 2007, pp. 127–136.

[42] X. Han, L. Wang, N. Crespi, S. Park, Á. Cuevas, Alike people, alike interests? Inferring interest similarity in online social networks, Decision Support Systems 69 (2015) 92–106.

[43] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[44] W.R. Hartmann, P. Manchanda, H. Nair, M. Bothner, P. Dodds, D. Godes, K. Hosanagar, C. Tucker, Modeling social interactions: identification, empirical methods and policy implications, Marketing Letters 19 (2008) 287–304.

[45] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, Springer, 2009.

[46] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in Twitter, Proceedings of the 20th International Conference Companion on World Wide Web, ACM, New York, NY, USA 2011, pp. 57–58.

[47] E. Horvitz, P. Koch, C.M. Kadie, A. Jacobs, Coordinate: probabilistic forecasting of presence and availability, Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA 2002, pp. 224–233.

[48] H. Itoga, G.T. Lin, et al., Using Facebook for event promotion-implementing change, African Journal of Business Management 7 (2013) 2788–2793.

[49] G. James, D. Witten, T. Hastie, R. Tibshirani, 2013. An Introduction to Statistical Learning: With Applications in R. 1st ed., 2013. Corr. 4th printing 2014 edition ed., Springer, New York.

[50] I. Kahanda, J. Neville, Using Transactional Information to Predict Link Strength in Online Social Networks, ICWSM 2009, pp. 74–81.

[51] M. Kayaalp, T. Ozyer, S.T. Ozyer, A Collaborative and Content Based Event Recommendation System Integrated With Data Collection Scrapers and Services at a Social Networking Site, IEEE, New York, 2009.

[52] R. Klamma, P.M. Cuong, Y. Cao, You never walk alone: recommending academic events based on social network analysis, in: J. Zhou (Ed.), Complex Sciences, Springer, Berlin Heidelberg 2009, pp. 657–670 (number 4 in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering).

[53] C.A. Lampe, N. Ellison, C. Steinfield, A familiar face(book): profile elements as signals in an online social network, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA 2007, pp. 435–444.

[54] P. Langley, Crafting papers on machine learning, Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000, pp. 1207–1212.

[55] P. Langley, Iba, K. Thompson, An analysis of Bayesian classifiers, Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI Press, San Jose, California 1992, pp. 223–228.

[56] P. Langley, S. Sage, Induction of selective Bayesian classifiers, Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA 1994, pp. 399–406.

[57] B. Larivière, D. Van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems with Applications 29 (2005) 472–484.

[58] D.H. Lee, PITTCULT: Trust-based Cultural Event Recommender, Assoc Computing Machinery, New York, 2008.

[59] W. Lee, L. Xiong, C. Hu, The effect of Facebook users' arousal and valence on intention to go to the festival: applying an extension of the technology acceptance model, International Journal of Hospitality Management 31 (2012) 819–827.

[60] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, N. Christakis, Tastes, ties, and time: a new social network dataset using Facebook.com, Social Networks 30 (2008) 330–342.

[61] H.Y. Liao, K.Y. Chen, D.R. Liu, Virtual friend recommendations in virtual worlds, Decision Support Systems 69 (2015) 59–69.

[62] A. Liaw, M. Wiener, Classification and regression by randomForest, R news 2 (2002) 18–22.

[63] T. Lovett, E. O'Neill, J. Irwin, D. Pollington, The calendar as a sensor: analysis and improvement using data fusion with social networks and location, Proceedings of the 12th ACM International Conference on Ubiquitous Computing, ACM, New York, NY, USA 2010, pp. 3–12.

[64] W.G. Mangold, D.J. Faulds, Social media: the new hybrid element of the promotion mix, Business Horizons 52 (2009) 357–365.

[65] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annual Review of Sociology 415–444 (2001).

[66] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2015.

[67] E. Minkov, B. Charrow, J. Ledlie, S. Teller, T. Jaakkola, Collaborative future event recommendation, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM Press 2010, p. 819.

[68] E. Mynatt, J. Tullio, Inferring calendar event attendance, Proceedings of the 6th International Conference on Intelligent User Interfaces, ACM, New York, NY, USA 2001, pp. 121–128.
[69] A.Y. Ng, M.I. Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, MIT Press 2002, pp. 841–848.
[70] C.M. Paris, W. Lee, P. Seery, The Role of Social Media in Promoting Special Events: Acceptance of Facebook 'Events', Springer-Verlag Wien, Vienna, 2010.
[71] M. Pennacchiotti, A.M. Popescu, A Machine Learning Approach to Twitter User Classification, 11ICWSM, 2011 281–288.
[72] T.D. Pessemier, S. Coppens, K. Geebelen, C. Vleugels, S. Bannier, E. Mannens, K. Vanhecke, L. Martens, Collaborative recommendations with content-based filters for cultural activities via a scalable event distribution platform, Multimedia Tools and Applications 58 (2012) 167–213.
[73] A. Prinzie, D. Van den Poel, Random multiclass classification: generalizing random forests to random MNL and random NB, in: R. Wagner, N. Revell, G. Pernul (Eds.), Database and Expert Systems Applications, Springer, Berlin Heidelberg 2007, pp. 349–358 (number 4653 in. Lecture Notes in Computer Science).
[74] J.A. Recio-García, L. Quijano, B. Díaz-Agudo, Including social factors in an argumentative model for Group Decision Support Systems, Decision Support Systems 56 (2013) 48–55.
[75] B. Ripley, W. Venables, nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models, 2015.
[76] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
[77] M. Sandri, P. Zuccolotto, Variable selection using random forests, in: P.S. Zani, P.A. Cerioli, P.M. Riani, P.M. Vichi (Eds.), Data Analysis, Classification and the Forward Search, Springer, Berlin Heidelberg 2006, pp. 263–270 (Studies in Classification, Data Analysis, and Knowledge Organization).
[78] P. Singla, M. Richardson, Yes, there is a correlation: — from social networks to personal behavior on the web, Proceedings of the 17th International Conference on World Wide Web, ACM, New York, NY, USA 2008, pp. 655–664.
[79] B. Suh, L. Hong, P. Pirolli, E.H. Chi, Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network, 2010 IEEE Second International Conference on Social Computing (SocialCom) 2010, pp. 177–184.
[80] J. Tullio, E.D. Mynatt, Use and implications of a shared, forecasting calendar, in: C. Baranauskas, P. Palanque, J. Abascal, S.D.J. Barbosa (Eds.), Human-Computer Interaction — INTERACT 2007, Springer, Berlin Heidelberg 2007, pp. 269–282 (number 4662 in. Lecture Notes in Computer Science).
[81] K. Venkatesh, V. Ravi, A. Prinzie, D. Van den Poel, Cash demand forecasting in ATMs by clustering and neural networks, European Journal of Operational Research 232 (2014) 383–392.
[82] B. Wellman, The network is personal: introduction to a special issue of social networks, Social Networks 29 (2007) 349–356.
[83] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 80–83 (1945).
[84] Y. Xie, Z. Chen, K. Zhang, C. Jin, Y. Cheng, A. Agrawal, A. Choudhary, Elver: Recommending Facebook Pages in Cold Start Situation Without Content Features, IEEE, New York, 2013.
[85] Y. Xu, X. Guo, J. Hao, J. Ma, R.Y.K. Lau, W. Xu, Combining social network and semantic concept analysis for personalized academic researcher recommendation, Decision Support Systems 54 (2012) 564–573.
[86] A. Zanda, S. Eibe, E. Menasalvas, SOMAR: A SOcial Mobile Activity Recommender, Expert Systems with Applications 39 (2012) 8423–8429.
[87] Y. Zhang, H. Wu, V. Sorathia, V.K. Prasanna, Event recommendation in social networks with linked data enablement, Proceedings of 15th International Conference on, Enterprise Information Systems 2013, pp. 371–379.

Matthias Bogaert is a PhD student at Ghent University. He has received his B.S. in Business Engineering and M.S. in Business Engineering: Marketing Engineering/Data Analytics from Ghent University. His research interests are social media analytics, predictive analytics and big data.

Michel Ballings (PhD) is Assistant Professor of Business Analytics at The University of Tennessee (Knoxville). He teaches Data Mining and Customer Analytics. His research interests are in social media analytics, customer analytics, and machine learning. He has co-authored several peer-reviewed publications in journals such as European Journal of Operational Research, Omega, and Expert Systems with Applications.

Dirk Van den Poel (PhD) is a Professor of Business Analytics/Big Data at Ghent University, Belgium. He teaches courses such as Statistical Computing, Big Data, Analytical Customer Relationship Management, Advanced Predictive Analytics, Predictive and Prescriptive Analytics. He co-founded the advanced Master of Science in Marketing Analysis, the first (predictive) analytics master program in the world as well as the Master of Science in Statistical Data Analysis and the Master of Science in Business Engineering/Data Analytics. His major research interests are in the field of analytical CRM (Customer Relationship Management): customer acquisition, churn, upsell/cross-sell, and win-back modeling. His methodological interests include ensemble classification methods and big data analytics. He has co-authored 70+ international peer-reviewed (ISI-indexed) publications in journals such as Journal of Applied Econometrics, Applied Geography, European Journal of Operational Research, and Decision Support Systems.