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/

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.