Predicting Self-declared Movie Watching Behavior using Facebook Data and Information-Fusion Sensitivity Analysis

Abstract

The main purpose of this paper is to evaluate the feasibility of predicting whether yes or no a Facebook user has self-reported to have watched a given movie genre. Therefore, we apply a data analytical framework that (1) builds and evaluates several predictive models explaining self-declared movie watching behavior, and (2) provides insight into the importance of the predictors and their relationship with self-reported movie watching behavior. For the first outcome, we benchmark several algorithms (logistic regression, random forest, adaptive boosting, rotation forest and naive Bayes) and evaluate their performance using the AUC. For the second outcome, we evaluate variable importance and build partial dependence plots using information-fusion sensitivity analysis for different movie genres. To gather the data we developed a custom native Facebook app. We resampled our dataset to make it representative of the general Facebook population with respect to age and gender. The results indicate that adaptive boosting outperforms all other algorithms. Time- and frequency-based variables related to media (movies, videos, and music) consumption constitute the list of top variables. To the best of our knowledge this study is the first to fit predictive models of self-reported movie watching behavior and provide insights into the relationships that govern these models. Our models can be used as a decision tool for movie producers to target potential movie-watchers and market their movies more efficiently.

Keywords: Facebook, movies, predictive models, social media, machine learning, information-fusion

1. Introduction

Movie marketing budgets have skyrocketed over the past years with expenditures reaching the 100 million dollar mark (Forbes, 2014). These advertising dollars are split across mass advertising via traditional media (e.g., television and radio) and targeted advertising on social media (Oh et al., 2016). While the latter approach is considered more cost effective because of the improved

targeting, there is still room for improvement. The underlying reason is that targeting options on social media are descriptive instead of predictive. We call them descriptive, because the options describe users in terms of their socio- demographics, location, and current preferences (Facebook, 2016). A marketer could contract a market research company to figure out the profile of consumers who might be interested in a specific movie genre. The marketer can then target those profiles on Facebook. In contrast, a predictive approach would consist in making available an option to Facebook advertisers to target social media users who are most likely to watch a given movie genre in the future. Facebook can implement a predictive approach, by letting advertisers choose their movie genre (or product) from a list, and subsequently fitting and deploying predictive models by comparing Facebook users that have and have not watched that given movie genre (or bought a given product). This predictive approach could be very effective but the question remains whether it is feasible to build such a system. Hence, an avenue for research is to develop such a targeting option, by estimating a model to predict whether yes or no a given user will watch a movie from a given movie genre, and compare its effectiveness with the descriptive targeting options that are readily available.

Despite the fact that movies and social media are extensively studied in the field of box office revenues (e.g., Rui et al., 2013) and recommender systems (e.g., Shapira et al., 2012), no study has evaluated the feasibility of developing such a model. Such a model would enable movie producers to identify and target potential customers. The development of such a model requires a data analytical methodology capable of providing answers to the questions: 'Is it feasible to develop such a targeted marketing approach, and if yes which algorithm performs best?','Which predictors are most important?', and 'What is the relationship between predictors and response?'.

In order to fill this gap in literature, we assess the feasibility of identifying social media users who will watch a given movie genre. In order to do so we predict self-declared movie watching behavior of Facebook users using all their available data. We note that we do not target the whole population of movie watchers (i.e., people who have watched a movie but did not declare it on Facebook). Instead, because of data availability, we are targeting a subset of the target population (i.e., people who have watched a movie and have listed it on their Facebook profile). We choose Facebook as our social media channel of interest since it has the richest data and targeting options, providing the strongest possible benchmark for our proposed model. To investigate the capacity of

identifying self-declared movie watchers on Facebook, we implement a data analytical methodology. The objectives of this data analytical framework are twofold. The first is related to the predictive performance of our framework. We build five prediction models (i.e., logistic regression, naive Bayes, random forest, adaboost and rotation forest) and investigate which models perform best. The second is related to the descriptive capacity of our framework. For that purpose we use information-fusion sensitivity analysis to determine which predictors are most important and assess their relationship with self-reported movie watching behavior.

The remainder of this paper is organized as follows. First, we elaborate on existing literature concerning social media and movies. Second, we discuss our methodology. Third, we provide an overview of the results. Fourth, we summarize our conclusions and their practical implications. Finally, we discuss the limitations and suggest avenues for future research.

2. Prior research

Based on extensive literature research concerning predictive modeling in social media and movies, we found that existing literature can be categorized into three types: i) movie sales prediction, (ii) recommender systems and (iii) predicting and explaining individual movie watching behavior. First, studies related to movie sales seize the predictive power of social media to forecast box-office revenues (Asur and Huberman, 2010). For example, Rui et al. (2013) found that people with more followers on Twitter and tweets expressing the intention to watch a movie have a high influence on movie sales. Second, recommender systems guide users in a large space of possible options to help them find movies of interest and produce individualized movie recommendations (Golbeck, 2006; Gupta et al., 2008). For example, Shapira et al. (2012) developed a recommender system that incorporated both profile characteristics and posting data from Facebook users and showed that enriching scarce rating data with these Facebook variables significantly improved recommendation results. Finally, movie watching behavior studies try to identify users that are most likely to watch a certain movie. Whereas recommender systems suggest movies to a given user, studies focusing on movie watching behavior recommend users to an advertiser of a given movie. In short, recommender systems assign movies to users, and customer acquisition systems assign users to movies. We think of recommender systems to be important tools for services such as Netflix, whereas the research that we propose is important for individual producers to promote

new releases.

To highlight our contribution to literature Table 1 provides an overview of all the studies concerning social media prediction and movies. From Table 1, it is clear that no study has conducted research on movie watching behavior with social media data. This is an important gap in literature because, in the case of movie producers, this application would enable the pursuit of a targeted marketing approach. Producers could identify users who are most likely to watch a certain movie genre and send an invitation to watch their movie of that genre in theaters to increase attendance.

In order to fill this gap in literature, we build a model that assists movie producers in predicting self-declared movie watching behavior using all the available Facebook data. Based on the characteristics of similar users, our model predicts which users have a high probability of watching a certain movie genre again. We note the difference between movie watching behavior prediction and movie recommender systems. On the one hand, movie recommender systems use ratings and past movie history to come up with relevant movies for a certain user (Gupta et al., 2008). On the other hand, our model matches users who watched a certain movie genre with other users who have not yet watched a certain movie type (or did not declare it) based on the similarity of behavioral characteristics on Facebook (e.g., the number of movie-related likes and the number of movies watched in the past). It then ranks the users who have not watched this type of movie, from high to low probability of watching. In that sense our model matches relevant users to a given movie genre, whereas recommender systems match relevant movies genres to a given user. This approach opens a lot of interesting opportunities for targeted marketing strategies on Facebook (Benedek et al., 2014). Nowadays advertisers on Facebook can decide to target consumers based on socio-demographics (e.g., age, gender and education), location (e.g., state or city), interests and behaviors (e.g., football). For example, in the case of movies, advertisers can decide, based on market research, to target males in the state of New York who are interested in a specific movie genre. A problem, however, is that these targeting options are descriptive and general. Our model offers a solution to movie producers that is more custom (i.e., specific) to their product. For example, advertisers can decide to target the users who have the highest probability of watching the specific genre of their movie.

To build such a model, we propose a data analytical system to predict self-declared movie watching behavior. The first objective of this system is to assess the capacity of our system

Table 1: Overview of social media prediction literature concerning movies

Study	Movie sales	Recommender system	Movie watching behavior
Basuroy et al. (2003)	X		
Golbeck (2006)		X	
Liu (2006)	X		
Mishne (2006)	X		
Dellarocas et al. (2007)	X		
Duan et al. (2008)	X		
Liu et al. (2007)	X		
Gupta et al. (2008)		X	
Asur and Huberman (2010)	X		
Goel et al. (2010)	X		
Liu et al. (2010)	X		
Moon et al. (2010)	X		
Said et al. (2011)		X	
Borsato and Polato (2012)		X	
David et al. (2012)		X	
Reddy et al. (2012)	X		
Shapira et al. (2012)		X	
Venkatesan and Mai (2012)		X	
Apala et al. (2013)	X		
El Assady et al. (2013)	X		
Jain (2013)	X		
Mestyán et al. (2013)	X		
Rui et al. (2013)	X		
Arias et al. (2014)	X		
Du et al. (2014)	X		
Hennig-Thurau et al. (2014)	X		
Liu et al. (2014)	X		
Pham et al. (2014)		X	
Dipak Damodar Gaikar et al. (2015)	X		
Kim et al. (2015)	X		
Ding et al. (2016)	X		
Lee et al. (2016)	X		
Oh et al. (2016)	X		
Our study			**X**

to accurately identify users with a high propensity of watching a certain movie genre based on Facebook data. In order to do so we evaluate the predictive performance of five different classifiers: logistic regression (Cox, 1958), naive Bayes (Langley et al., 1992), random forest (Breiman, 2001), adaboost (Friedman, 2002) and rotation forest (Rodriguez et al., 2006). By determining which algorithm is best on this type of problem, our findings will allow future research to focus on one algorithm instead of many. Previous work on predictive modeling in social media has shown that Facebook data yields accurate predictions in the field of events (Bogaert et al., 2016a), romantic ties (Bogaert et al., 2016b), movie recommendations (Shapira et al., 2012) and box office predictions (Oh et al., 2016). Overall, Facebook data have been shown to improve predictive performance.

The second objective of this system is to determine which variables are driving the predictive performance and to uncover their relationship with self-declared movie watching behavior. In line with previous literature, we believe that Facebook contains a number of variables that can be indicative of movie watching behavior. More specifically, there are four main data types on Facebook, which could be important in explaining movie watching behavior: (i) profile data, (ii) behavioral data, (iii) interests data and (iv) network data. First, Facebook profile data (e.g., age, gender and relationship status) have been shown to be a viable alternative to traditional ratings for movie recommendations (Gupta et al., 2008). Therefore we believe that general profile data can also have value in explaining movie watching behavior. For example, young adults may be more willing to watch a certain movie type. Second, behavioral Facebook data concerning movies can result in accurate movie suggestions when rating data are scarce (David et al., 2012). For example, David et al. (2012) show that the use of data related to TV shows a user watched or liked on Facebook and traditional rating data (e.g., on Netflix) results in similar movie recommendations. Hence, in the case of movie watching behavior, it could be that the number of previous movies someone has indicated to watch is an important predictor. Third, Shapira et al. (2012) showed that including a user's interests on Facebook significantly increases the accuracy of movie recommender systems. Gupta et al. (2008) confirm these findings and find that a given user's favorite books and music correlate with his/her movie preferences. Also in the field of box office predictions researchers found that likes on Facebook are related to box office sales (Ding et al., 2016; Oh et al., 2016). Therefore, we believe that variables that express a user's interests in media consumption (e.g., liking of TV shows or actors) can also impact performance of movie watching behavior predictions.

Finally, network data (e.g., whether users are a fan of the same movie or commented on the same movie) further improve the accuracy of movie recommendations (Said et al., 2011). Also, in other applications than movie predictions, Facebook Friends data are found to be important predictors of user behavior (e.g., Bogaert et al., 2016a).

To summarize, we found strong indications in extant literature that predicting self-reported movie watching behavior (i.e., identifying users with a high probability of watching a certain movie genre) is a viable application domain next to the well established applications of movie sales predictions and movie recommender systems. We also found strong indications that Facebook data could yield accurate results in predicting whether or not a user will watch a certain movie type. Finally, we expect that several variable types are important for predictive performance. In the next section, we will discuss our methodology.

3. Methodology

For our data analytical methodology we rely on the widespread CRISP-DM framework (Chapman et al., 2000). CRISP-DM stands for 'Cross-Industry Standard Process for Data Mining' and provides a methodological framework for conducting data analytics projects. CRISP-DM comprises six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. The stages are sequential in nature, however, in some cases one has to go back to previous stages. The first three stages (business understanding, data understanding, and data preparation) require the most time. The final output of these stages is the creation of a cleansed dataset, also referred to as the basetable. Once the basetable is created the following stages of the CRISP-DM framework (modeling, evaluation and deployment) can start.

Figure 1 provides an overview of our analysis beginning at the modeling step of the CRISP-DM framework. For each of our 10 movie genres, we build different classification models, evaluate performance and apply information-fusion analysis. First, we assess the predictive performance of our framework. Therefore, we build five classification models using a train/test split and evaluate the performance of 5 classification models using the AUC. The AUC is an aggregate measure of classifier performance and can be defined as the probability that a randomly chosen positive instance is ranked higher than a negative instance (Hand and Anagnostopoulos, 2013). Second, we evaluate the descriptive capacity of our system. Therefore, the predictions of all five classification

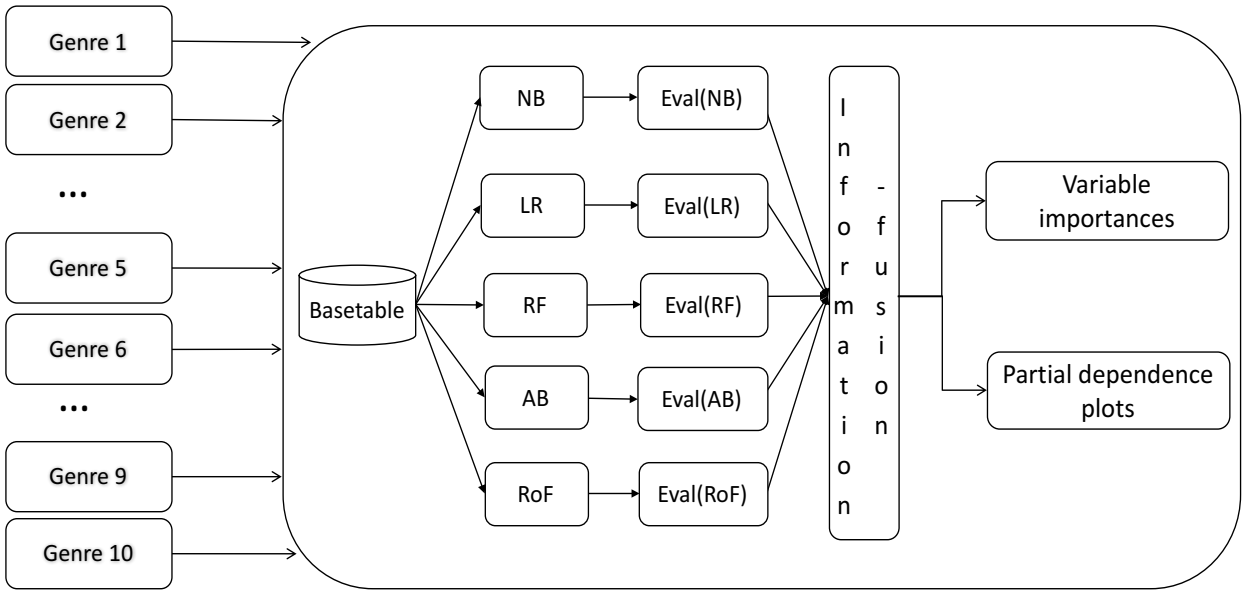


Figure 1: Analysis Overview

models are combined using information-fusion (see Section 3.4). Information-fusion aggregates the predictions of all classification models by taking the weighted average of their AUCs. Once the fusion model is built, we conduct a sensitivity analysis by calculating the variable importances as the mean decrease in AUC and building partial dependence plots for the top predictors. The variable importance ranks the predictors according to their influence on predictive performance. Partial dependence plots uncover the relationship between a certain predictor and response, while keeping all other predictors constant. As stated before, this process is repeated for all 10 genres. Hence, in total 50 models are built (5 classification models for 10 movie genres). The reported model performance results are the median values across our 10 genres. The reported sensitivity analysis results are the rescaled median values across our 10 genres. We also report the sensitivity analysis of three different genres, since the impact of certain variables can be different across different types of movies.

3.1. Data

In order to extract our Facebook data, we developed a custom native Facebook application. The Facebook application had both a front-end and a back-end. The back-end, on its part, included the creation of a database to store the collected data. The front-end included the functionalities to the users. The application was developed for a European soccer team, and was advertised several times on the Facebook page of the European soccer team. To increase awareness and interest, an incentive (i.e., a signed jersey) was offered to Facebook users to run the application. When users clicked the app they were presented with an authorization box that allowed the users to donate their data in exchange for entering the drawing of a prize. The authorization box also included a rules and regulation section, containing our contact information. In addition, we also ensured the users that all extracted data would be anonymous and that no private messages would be gathered. Afterwards, the users had to fill out several questions concerning the soccer team and the number of participants of the application to determine the winner of the jersey. Figure 2 clarifies which data were mined from a user's profile with our application (red boxes). The data were collected between May 7, 2014 and May 26, 2014. In total, we gathered user profile data of 5010 unique Facebook users. Furthermore our data contain 6738 unique movies from 3818 different users. Since we are interested in predicting declared movie watching behavior, we restrict our sample to the 3818 unique users.

Several selection effects could occur when mining the data. A first selection bias happens when the application is advertised via the Facebook page of the soccer team. Hence, people following the soccer team will be more enticed to click on the application. This implies that our sample can contain more soccer fans than the average Facebook sample. A second selection effect happens when the promotion is displayed through the News Feed via the News Feed Algorithms (NFA) of Facebook. The NFA determines who sees a certain advertisement, and when that happens, based upon the interactions and the interests of the users. We tried to mitigate this problem by targeting the advertisement to a representative sample of Facebook users. A final selection effect occurs because of the fact that if a user sees the application, only a small selection will be willing to proceed to the application and fill out the survey. The reason could be that the user is not willing to provide their data to the application or that the offered prize is not satisfactory. In order to cope with these selection effects, we tested whether our sample is representative for the

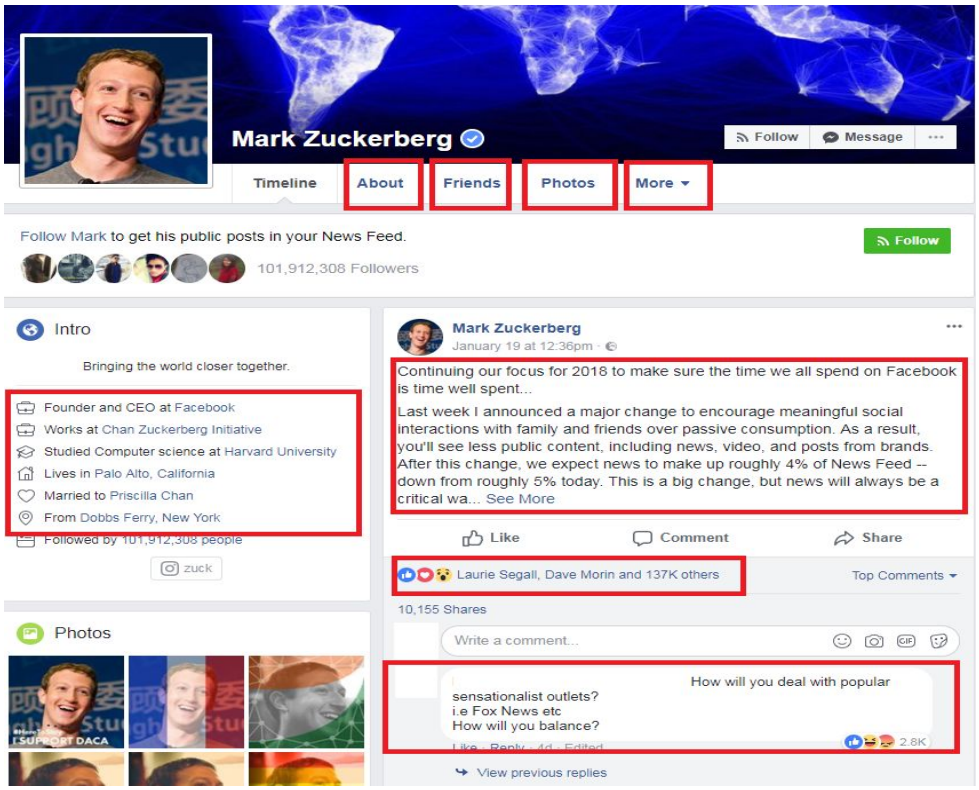


Figure 2: Example of the collected data from a user profile. The data in the red boxes are extracted through the API

general Facebook population. This task is non-trivial since Facebook only publishes a handful of demographic statistics of its user base on an aggregate level. Specifically, via Facebook's Advertising targeting system it is possible to extract the relative gender distributions across different age groups for the whole Facebook population (Facebook, 2018). Figure 3 displays the gender characteristics per age group for the general Facebook population (left pane) and our sample (center pane). We performed a χ^2 test which indicates that our sample is not representative on the age distribution for males ($\chi^2(4) = 106.61, p < 0.001$) and females ($\chi^2(5) = 163.60, p < 0.001$). Our sample clearly under-represents females in all age groups and the over-representation of males is especially large in the 13-24 age group. To alleviate this problem, we resample our data set to be representative of the general Facebook distribution. To do so, we determine the number of observations that were necessary in every age group per gender with a sample of 5010 unique users. For example, 32% of the Facebook population is male in the age group of 13-24, this corresponds with 538 ($\lfloor 5010 * 0.32 \rfloor$) users in our data set. Age groups that were over-represented in our data were undersampled. This means that we randomly selected a number of users from that age groups such that the desired distribution was achieved. Under-represented age groups were oversampled. Hence, certain users were randomly selected and replicated. After resampling, we performed a χ^2 test which indicated that there was no significant difference between the male and female age groups ($\chi^2(4) = 0.0003, p = 1$ and $\chi^2(4) = 0.0009, p = 1$).

Since it was not feasible to create a separate model for all of our 6738 unique movies, we decided to focus on the 10 most popular movie genres. We believe that predicting movie genre a more general proxy to predict movie watching behavior and hence more relevant for practitioners. Table 2 provides an overview of the distribution of the 10 most watched movie genres. The fraction of users who have watched the movie genre equal the total number of users who watched the movie genre divided by the number of unique users in our database (3818).

Table 2 also serves as an indication of the distribution of our dependent variable. Hence, we build 10 models for each algorithm where the response variable is binary and takes the value 1 if a user has declared on his/her Facebook profile whether he/she watched a particular movie genre and 0 otherwise. It is clear from Table 2 that our response variable suffers from a high class imbalance problem, namely a severe under-representation of the users who watched a movie. In order to cope with this class imbalance problem, we used two common data resampling techniques:

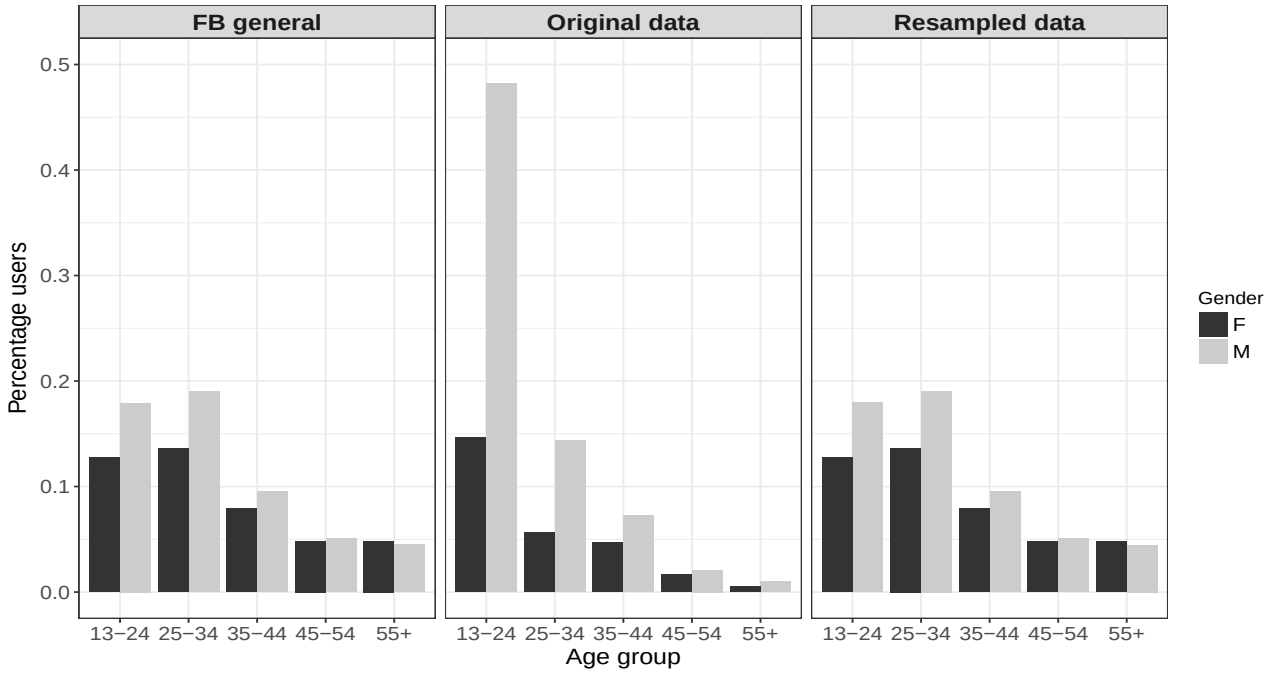


Figure 3: Sample

random oversampling (ROS) and synthetic minority over sampling technique (SMOTE) (Dag et al., 2016). In ROS we randomly duplicate the cases of the minority class (i.e., users who watched the movie) to obtain a balanced data set with an equal distribution between both classes (Estabrooks et al., 2004). In SMOTE we randomly remove cases from the majority class and create synthetic instances of the minority class until a balanced class distribution is reached (Chawla et al., 2002). The synthetic cases of the minority class are created as follows. First, take the difference between the minority class instances and their nearest neighbor. Next, multiply this difference by a random number between $[0, 1]$ and add this to the original minority class instances (Chawla et al., 2002).

3.2. Variables

In total we included 486 user-related variables in our models. 'Like' variables in this study refer to likes generated by users. These 'likes' are only available for Facebook pages, bands, applications or leisure activities. As a result of regulations on Facebook, only the last 25 items, such as photos, videos, check-ins and notes, could be extracted from the web page. In order to mitigate this limitation, the frequency by time was calculated, since none of the users in the data set reached this restriction. We computed the frequency of status updates, photo uploads and links created for

Table 2: Top ten most popular movie genres

Movie name	Fraction of users watched
Drama	0.1972
Adventure	0.1886
Action	0.1504
Fantasy	0.0846
Thriller	0.0684
Romance	0.0656
SciFi	0.0457
Horror	0.0410
Documentary	0.0308
Music	0.0289

the last 7 days, album uploads and check-ins for the last 4 months and notes and video uploads for the last year.

Table 3 gives an overview of the different categories of the variables included in our study and illustrates each category with two examples. The term 'posts' refers to the different posting objects that exist on Facebook, namely statuses, photos, albums, videos, check-ins and links. 'Comments' refers to both comments made and received by the user. Tags, likes and comments are included for all different post types. In Table 3 MIET stands for mean-inter event time (i.e., the average time between posts) and SDIET means standard deviation inter event time (i.e., the variation on the time between posts).

Since our variable of interest is whether or not a user watched a particular movie genre, we also included several variables related to videos, movies and television. For these variables we did not only compute user variables, but also added the network variables. Table 4 shows the data type (i.e., whether they are based on a single user's info or whether they are built on friends' data) and the created variables. We note that 'videos' includes both movies and videos from all kinds of categories, such as advertisements and home-made videos. The categories variable refers to the different types of videos that exist such as movies, TV shows, ads, home-made movies, news and movie and movie characters. The included user variables are related to recency and frequency

Table 3: Overview of predictors

Variable category	Example
Demographic and identification variables	Age
	Gender
Geographical variables	Hometown
	Location
Professional and educational variables	Languages
	Location
Social variables	Number of friends
	Relationship status
Personal variables (interests)	Favorite sports (e.g., soccer)
	Favorite music (e.g., jazz)
General Facebook account variables	Profile completeness
	Length of relationship
Events	Number of events attended
	Recency/MIET/SDIET of created events
Games	Number of games
	Recency/MIET/SDIET of created games
Notes	Number of notes
	Recency/MIET/SDIET of created notes
Posting variables	Number of posts
	Recency/MIET/SDIET of created posts
Likes	Number of post likes
	Recency/MIET/SDIET of created post likes
Tags	Number of post tags
	Recency/MIET/SDIET of created post tags
Comments made/received	Number of post comments
	Recency/MIET/SDIET of created post comments

Table 4: Overview of movie-related predictors

Data type	Variable
User data	Number of videos/movies/TV shows watched
	Number of categories watched
	Recency/MIET/SDIET of videos/movies/TV shows
Network data	Number of videos/movies/TV shows watched by friends
	Number of categories watched by friends
	Number of friends who watched the focal genre

variables in customer relationship mananagement (Ballings and Van den Poel, 2012). People with a high frequency and a shorter recency have a higher probability to exhibit repeat behavior (in our case movie watching behavior) (Van den Poel, 2003). The network variables are included because they are among the top predictors in social media prediction (Bogaert et al., 2016a).

3.3. Classification algorithms

In this section we discuss the different algorithms to model movie watching behavior. We opted to benchmark several single classifiers and ensemble techniques with a proven track record of strong performance in social media applications and with different ranges of complexity (Ballings and Van den Poel, 2015a; Bogaert et al., 2016b). The single classifiers are logistic regression (LR) and naive Bayes (NB). Naive Bayes is very inexpensive in terms of computational overhead and complexity, but makes many assumptions (Prinzie and Van den Poel, 2007). For example, naive Bayes assumes variables to be conditionally independent to estimate the joint probability $p(x, y)$. However, the assumption of conditional class independence is often violated. Logistic regression estimates the conditional probability $p(y|x)$ and is one step up in terms of complexity in that all coefficients are estimated jointly (Eren Demir, 2014). We implemented the following tree-based ensemble techniques: random forest (RF), adaboost (AB) and rotation forest (RoF). Tree-based methods can be specified as additive models and they add complexity by allowing for nonlinear relationships. In addition, the selected tree-based ensembles all tackle the increased complexity in different ways. Random forest adds complexity by combining a large number of bootstrapped decision trees. Moreover, random forest also selects a random number of variables at each node

split (Biau, 2012). Adaboost adds complexity by introducing an iterative procedure that favors misclassified instances (Friedman, 2002). Finally, rotation forest induces complexity by combining bootstrapped trees with principal components analysis (Kuncheva and Rodriguez, 2007).

3.3.1. Logistic regression

Logistic regression fits a function for the prediction of the probability of the occurrence of an outcome as (James et al., 2013, pp. 132):

$$p = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}} = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i)}} \tag{1}$$

In Eq. 1 p represents the probability of the interested outcome, α the intercept term and $\beta_1, \ldots, \beta_p$ the coefficients of the independent variables $X_1, \ldots, X_i$.

In this study, we use regularized logistic regression and we apply the lasso approach in order to avoid overfitting. The lasso (least absolute shrinkage and selection operator) minimizes the residual sum of squares for which the sum of the absolute value of the coefficients is less than a constant (Tibshirani, 1996). In other words, it imposes a bound on the sum of the absolute values of the coefficients and therefore forces coefficients to shrink towards zero (Guisan et al., 2002). To fit our models, we used the statistical *R*-package *glmnet* provided by Friedman et al. (2015). The parameter α is set to 1 to obtain the lasso method and the sequence of λ is computed by setting the parameter nlambda to 100 (default).

3.3.2. Naive Bayes

Naive Bayes is a simple induction algorithm that simplifies learning and is the most straightforward and most widely tested technique for probabilistic classification (Langley et al., 1992). Naive Bayes applies Bayes' theorem for classification of observations and assumes that features are class independent (Rish, 2001):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

In Eq. 2 $P(A)$ and $P(B)$ are the probabilities of two events A and B (both independent from each other), $P(A|B)$ represents the probability of event A given event B and $P(B|A)$ the probability of event B given event A.

Although, in general, conditional independence is a poor assumption and rarely true in real-world applications, the naive Bayes algorithm performs remarkably well on many learning tasks and manages to compete with much more sophisticated classifiers (Langley et al., 1992; Rish, 2001). We implemented our models using the *R*-package *e*1071 by Meyer et al. (2015).

3.3.3. Random forest

Significant improvements in classification and regression accuracy can be obtained by using a classification algorithm called random forest (Breiman, 2001). In a standard decision tree approach, each node is split using the best split among all variables (Breiman, 1996). Random forest changes the way in which classification or regression trees are grown: each tree is constructed using an independent bootstrap sample of the data and at each node of each tree a randomly selected subset of features is evaluated (Biau, 2012). Thus, random forest enhances bootstrap aggregating with an additional layer of randomness (Breiman, 2001).

Research has shown that random forest is one of the best techniques available and performs very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks (Biau, 2012; Coussement and Van den Poel, 2008).

Furthermore, it is very user-friendly in the sense that only two parameters have to be provided: the number of trees and the number of predictor variables in the random subset at each node of the tree (Ballings et al., 2016). We follow the recommendations of Breiman (2001) and use a large number of trees (500). To implement the algorithm, we used the statistical *R*-package *randomForest* provided by Liaw and Wiener (2002).

3.3.4. Adaboost

Boosting is a general approach for improving the accuracy of machine-learning algorithms and entails combining the outputs of many weak classifiers to produce a powerful predictor (Freund et al., 1996). The training data are sequentially reweighted and in each iteration the performance of the classifiers is evaluated, thereby giving a higher weight to misclassified observations, while correctly classified observations' weights are decreased (Friedman, 2002). Consequently, observations that are hard to classify receive an increasing influence. Finally, all the classifiers are added to the ensemble to build the final predictor (Freund et al., 1999). We use stochastic boosting, one of the most recent boosting variants, to improve on the original algorithms, through incorporating

randomization into the procedure (Friedman, 2002). Two important parameters are the number of iterations and the number of terminal nodes in the base classifiers. Following recommendations by Friedman (2002), the number of terminal nodes is determined by setting the maximum depth of the trees to 3, while the number of iterations is set to 500. We used the statistical *R*-package *ada* by Culp et al. (2012) to fit our models.

3.3.5. Rotation forest

Rotation forest is a powerful method for generating tree-based ensembles. A typical random forest requires a large number of trees in order to achieve good performance. By contrast, rotation forest can achieve similar performance with a smaller number of trees (Kuncheva and Rodriguez, 2007). This approach consists of taking a subset of features and a bootstrap sample of the data and carrying out a principal component analysis (PCA), where a small rotation of the axes of the feature space may lead to a very different tree (Rodriguez et al., 2006). Results from research by Rodriguez et al. (2006) show that when rotation forest was compared to bagging, adaboost and random forest on 33 different data sets, rotation forest outperformed the other three algorithms. Another paper by De Bock and Van den Poel (2011) also showed the superior performance of rotation forest in the field of customer churn. We use the statistical *R*-package *rotationForest* to implement the algorithm (Ballings and Van den Poel, 2015b).

3.3.6. Performance measures

As a measure of classifier performance we use the area under the receiver operating characteristic curve (AUC). The AUC alleviates the problems of the accuracy by aggregating the results over all possible thresholds and giving an equal weight to both positive and negative cases (Guo et al., 2015). Hence, the AUC is perfectly suited for situations where the data is unbalanced. The AUC is defined as the probability that a randomly chosen positive example is ranked higher than a randomly picked negative example (Hand and Anagnostopoulos, 2013). The AUC can also be seen as a graphical representation between the true positive rate (i.e., sensitivity) and the false positive rate (one minus specificity) (Hernandez-Orallo et al., 2012). Sensitivity, specificity and the AUC are defined as (He and Garcia, 2009; Hanley and McNeil, 1982):

$$Sensitivity = \frac{TP}{TP + FN}, \tag{3}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4}$$

$$AUC = \int_0^1 \frac{TP}{(TP + FN)} d\frac{FP}{(FP + TN)} = \int_0^1 \frac{TP}{P} d\frac{FP}{N}, \tag{5}$$

with P: Positives, and N: Negatives.

The AUC can take values ranging from 0.5 to 1. A value of 0.5 indicates that the predictions are not better than random, whereas a value of 1 indicates that the predictions are perfect (Eren Demir, 2014). An important characteristic of the AUC is that it is threshold independent. This means that AUC includes the entire range of possible thresholds (Hernandez-Orallo et al., 2012).

3.3.7. Cross-validation and statistical tests

To make sure our results are not optimistic or pessimistic, we build 10 different predictive models (one for each genre in our dataset) to perform the cross-validation. We believe that our results will gain in generalizability by fitting a single model for each of 10 genres as opposed to, for example, estimating 10 models predicting only one genre. Each predictive model of these 10 predictive models is built using the holdout set approach (James et al., 2013, pp.176-178). If no cross-validation of the parameters was necessary a 50/50 split was used for training and test, else a 25/25/50 split was applied for training, validation and test. In total we have 10 predictive models, one for each movie. We then take the median of the AUC scores of our 10 models to obtain the overall AUC. The interquartile range (IQR) is used as a measure of dispersion.

To test for significant differences between the algorithms, we follow the suggestions of Demšar (2006) and use Friedman's rank sum test together with the Bonferroni-Dunn post hoc test for comparison of the different classifiers. The classifiers are ranked, whereby the best performing classifier receives rank 1. If no ties are observed, the worst performing classifier receives a rank equal to the total number of classifiers. If ties do occur, they are handled by averaging the ranks. The average AUC ranks preserve the order of the folds while the median of the AUCs does not. Therefore, averaging ranks allows a stricter comparison than calculating the median (Ballings and Van den Poel, 2015a).

Two classifiers will perform significantly differently if the difference between their average ranks surpasses the critical difference. The critical difference (CD) can be written as (Demšar, 2006):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{6}$$

where q_α is the critical value for a certain p-value and number of classifiers, k is the number of classifiers and N the number of folds. In our study the critical difference for a p-value of 0.05, 5 classifiers (LR, NB, RF, RoF, AB), 10 folds (10 movie models) and a critical value of 2.498 equals 1.7661.

3.4. Information-fusion

3.4.1. Information-fusion model

Information-fusion intelligently combines the information extracted from several data mining algorithms (Oztekin et al., 2013). Therefore, information-fusion yields more useful and accurate information than single data mining models (Sevim et al., 2014). The reasoning behind information fusion is that there is no single best method that works for every problem. Hence it is better to integrate the results of several prediction models instead of using a single prediction model (Dag et al., 2016). If y represents the response variable and X a set of predictors with $X = \{x_1, x_2, ..., x_p\}$, then a classifier k can be formulated as:

$$\hat{y}_{individual_k} = f_k(x_1, x_2, ..., x_p) = f_k(X). \tag{7}$$

Let Ψ represent the information fusion operator of the individual classification models $f_k(X)$. If we then assume that we have 5 prediction models, then the information-fusion model $\hat{y}_{fusion}$ can be represented as:

$$\hat{y}_{fusion} = \Psi\left(\hat{y}_{individual_1}, \hat{y}_{individual_2}, ..., \hat{y}_{individual_5}\right) = \Psi\left(f_1(X), f_2(X), ..., f_5(X)\right). \tag{8}$$

If Ψ then represents a linear function of the classifiers f_k with β_k as the individual weighting coefficient of each classier f_k, then Eq. 8 can be rewritten as:

$$\hat{y}_{fusion} = \sum_{k=1}^{5} \beta_k f_k(X) = \beta_1 f_1(X) + \beta_2 f_2(X) + ... + \beta_5 f_5(X). \tag{9}$$

In Eq. 9 we assume that the weights β_k are normalized such that $\sum_{k=1}^{5} \beta_k = 1$ holds. The weights β_k are the rescaled median performances, across our 10 genres, of the individual classification

models $f_k(X)$: $\beta_k = AUC_k / \sum_{k=1}^{5} AUC_k$ (Oztekin et al., 2016). Hence, the higher the accuracy of a certain prediction model, the higher the weight in the fusion function $\hat{y}_{fusion}$ (Oztekin, 2016).

3.4.2. Information-fusion sensitivity analysis

After determining the information-fusion model (Eq. 9), another important question in data analytics is to examine which variables are the drivers of predictive performance (Oztekin et al., 2013). In that sense variable importance measures are seen as a form of sensitivity analysis where the independent variables are ranked according to their importance in prediction. In variable importance measures the effect of a certain variable on performance is examined by permuting that variable's values. The difference between the model's performance before and after permuting the variable is then taken as a measure of variable importance (Sandri and Zuccolotto, 2006). This variable importance measure can also be seen as a sensitivity measure since it shows us how sensitive the model is to permutation on the focal variable (Sevim et al., 2014). The higher the variable importance measure, the more sensitive the model is to changes in the predictor and the higher its impact on performance. Since we are working in a highly unbalanced setting, we decide to use the decrease in AUC as our measure of variable importance (Janitza et al., 2013). The decrease in AUC uses the AUC to determine the change in predictive performance and hence is more robust to changes in the underlying distribution of the data. The importance measures are averaged across our 10 genres by taking the median.

We apply the same logic of our information fusion model in order to come up with an information-fusion sensitivity score (S_{fusion}). We can then rewrite Eq. 9 in terms of the information-fusion sensitivity of a certain variable j with 5 classification models:

$$S_{fusion_j} = \sum_{k=1}^{5} \beta_k V_{kj} = \beta_1 V_{1j} + \beta_2 V_{2j} + ... + \beta_5 V_{5j}. \tag{10}$$

In Eq. 10 V_{kj} represents the median decrease in AUC of predictor k in prediction model j. The values of β are similar to these in Eq. 9, namely the rescaled median AUCs, across our 10 genres, of the five different classifiers. Hence, the sensitivity score S_{fusion_j} of variable j calculates median decrease in AUC when permuting variable j rescaled across all algorithms and averaged across all movie genres.

After determining the information-fusion model and evaluating the variable importance, the

final question is how the top predictors are related to the response variable. In linear regression, this is represented by the coefficients of the regression model, where the coefficient x_1 encapsulates the effect of x_1 on y while leaving all other variables constant. In data mining, this relationship can be graphically displayed by partial dependence plots (Meire et al., 2016). Partial dependence plots depict the relationship between predictor and response after controlling for the average effect of all other predictors (Friedman and Meulman, 2003). In order to create partial dependence plots, we follow the suggestions of Berk (2008, pp.222).

First we build our fusion model based on Eq. 9. By taking our fusion model as a basis for our partial dependence, we make sure that we account for the total effect of a predictor over all our models. Next, for every distinct value v of a predictor x, a new data sample is created that only takes on that one value v while controlling for the average effect of all other predictors. Next, we predict the output for every new data set using our fusion model. This is followed by taking the mean of half the logit of the predictions, resulting in one single value p for all instances. Finally, we plot all values v against their corresponding p (Berk, 2008).

4. Results

4.1. Model evaluation

Table 5 provides an overview of the cross-validated model performance values and interquartile ranges (IQR) for both data sampling techniques. Performance is calculated as the median AUC and accuracy for all five algorithms across our 10 movie genres. The results indicate that we can predict movie watching behavior with high predictive accuracy: the AUC ranges from 65.24% to 82.68% for ROS, for SMOTE the AUC from 64.57% to 82.55%. Table 5 also shows that adaboost (AB) is the top performer, followed by random forest (RF), rotation forest (RoF), logistic regression (LR), and naive Bayes (NB). Moreover, our models are very stable with IQRs ranging from 4.08% to 6.93% for ROS and from 1.71% to 5.66% for SMOTE . These results are in line with previous research regarding social media analytics, where adaboost also was the top performer in the field of usage increase (Ballings and Van den Poel, 2015a), events (Bogaert et al., 2016a) and romantic ties (Bogaert et al., 2016b). Finally, we also summarize the (relative) number of wins for each data sampling technique in the last row of Table 5 (Demšar, 2006). We note that ROS outperforms SMOTE in 60% (3 out of 5 times) of the cases for AUC. Hence, we use the ROS models for

Table 5: Cross-validated performance and IQR for ROS and SMOTE

Models		Performance	IQR
ROS	LR	0.7691	0.0693
	NB	0.6524	0.0408
	RF	0.8206	0.0530
	AB	0.8268	0.0629
	RoF	0.8179	0.0562
SMOTE	LR	0.7700	0.0500
	NB	0.6457	0.0171
	RF	0.8203	0.0443
	AB	0.8255	0.0566
	RoF	0.8194	0.0563
Wins (ROS/SMOTE)		3/2 (0.60/0.40)	

computing the variable importances and the partial plots in the sensitivity analysis.

In order to find out whether the differences between the algorithms are significant, we compare the average ranks using the Friedman test with the Bonferonni-Dunn post-hoc test in Table 6. From the Friedman test, we learn that we can reject the null hypothesis of no significant differences for all performance measures and sampling techniques (see Table 6). With the Bonferonni post-hoc test we investigate which classification models are significantly different from the top performer. Hence we are able to distinguish two groups: models that are equal to the top performer in statistical terms (i.e., the difference between the average ranks is smaller than 1.7761) and models that perform significantly worse (i.e., the difference between the average ranks is greater than 1.7761). In Table 6, adaboost is the top performer and random forest and rotation forest perform equally well in statistical terms (in bold). Logistic regression and naive Bayes both performed statistically worse than adaboost.

Wolpert's no free lunch theory states that when comparing two learning algorithms A and B, there are just as many situations where A is superior to B and vice versa. Hence, the superiority of a learning algorithm is dependent upon the assumptions of the algorithms and the characteristics of the data (Wolpert, 1996). The reasons why random forest, rotation forest, and adaboost are

Table 6: Average ranks based on AUC and accuracy

		LR	RF	AB	RoF	NB	Friedman χ^2 (4)
ROS	AUC	3.80	**2.40**	**1.50**	**2.30**	5.0	30.96, p<0.001
SMOTE	AUC	4.00	**2.20**	**1.30**	**2.50**	5.0	35.12, p<0.001

superior in this case are manifold. First, the methods are non-parametric methods that do not require the normality assumption to be met (Ballings and Van den Poel, 2015a). As in many real-life data sets the analyses suggests that the data is not normally distributed and non-linear. This is exemplified by the low performance of logistic regression. The superior performance of all tree-based methods (i.e., adaboost, random forest and rotation forest) provides support for this assumption. Another reason is that adaboost, rotation forest, and random forest are ensemble methods. Ensemble methods lower the total test set error by solving the representational, statistical and computational problem of single classifiers (Dietterich, 1996). Single classifiers (e.g., decision trees) often tend to be unstable and and have a high variance (Croux et al., 2007). Random forest reduces the variance of single decision trees by combining bootstrap aggregation with random subspaces (Breiman, 2001). Rotation forest lowers the variance by decorrelating the trees by applying PCA and bagging (De Bock and Van den Poel, 2011) Adaboost does not only decrease the variance but also lowers the bias (Bauer and Kohavi, 1999). Finally, when confronted with a large number of variables single classifiers tend to overfit (Babyak, 2004). Ensemble methods such random forest and rotation forest do not overfit (Breiman, 2001). These reasons explain why adaboost, random forest and rotation forest are the top performers.

4.2. Information-fusion sensitivity analysis

To determine which variables are important, we made a plot that depicts the top 100 predictors against their sensitivity score in decreasing order (Figure 4). This means that the variable with the highest sensitivity score receives rank 1, the second highest sensitivity score rank 2 and so on. Since decision makers are not only interested in the average effect on movie watching behavior, we conduct a sensitivity analysis for an action movie and a documentary[1]. In Figure 4 the black solid line represents the average (i.e., the cross-validated effect across our 10 genres), the red line plots

[1]Partial plots for other movie genres are available upon request

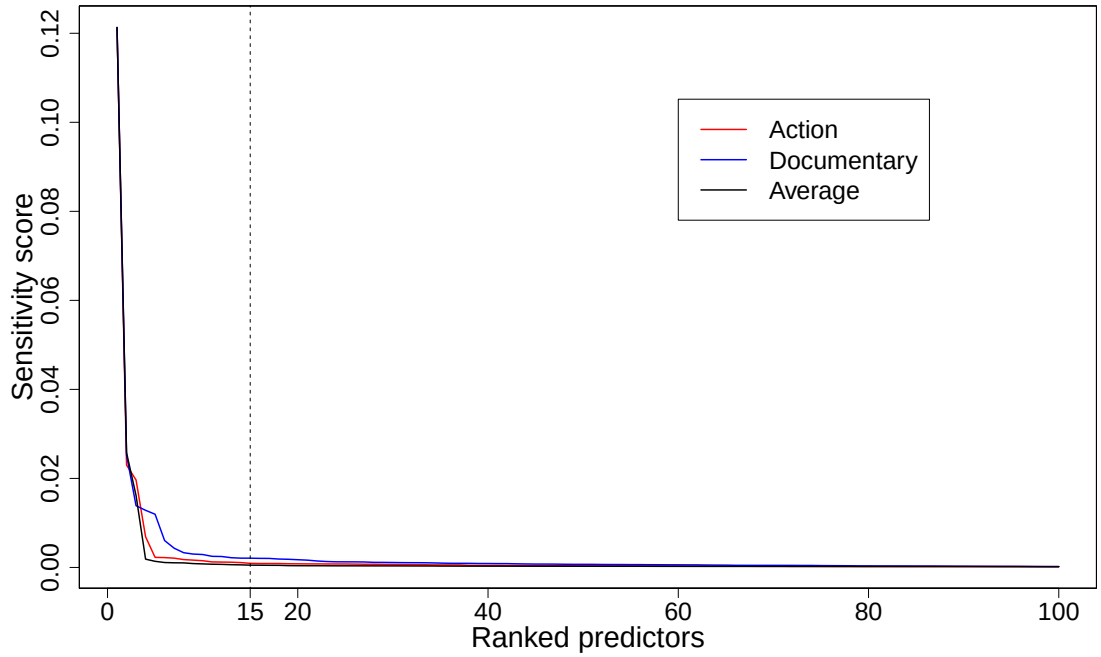


Figure 4: Scree plot of the predictors

the effect for an action movie and the blue line for documentary. We calculated the sensitivity scores based on Eq. 9. The β values are calculated as the rescaled AUCs of ROS from Table 5. The variable importances are calculated as the median decrease in AUC for each predictor for each algorithm. The final sensitivity scores are then computed by inserting the rescaled model performances and the variable importances for each algorithm in Eq. 9. For the average effect the β values and the variable importances are cross-validated across our 10 genres and averaged by taking the median. Figure 4 informs us that predictors with a rank higher than 15 (recall that a lower rank is better) do not add much to the predictive performance (black dashed line).

In Table 7 we summarize the top 15 predictors for the average, the action movie and the documentary. Recall that a higher sensitivity score means a higher variable importance, and hence means more impact on predictive performance. For example, the number of movies watched has a sensitivity score of 0.1153 for the average. This means that when permuting on the number of movies watched the AUC decreases on average with 11.53%-points across all 5 algorithms and 10 genres. For the action movie, the sensitivity score is 0.1213 which means that the AUC drops with

12.12%-points on average across all 5 algorithms. We also include the type of variable where B stands for behavioral data, I for interests, N for network data and P for profile data. We note the difference between behavioral data and interest data. Behavioral data, on the one hand, describes a given user's specific movie-related behavior (e.g., number of movies a user has watched) or general Facebook behavior (e.g., the liking behavior of that person). Interest data, on the other hand, capture the more general interests of the users, such as music or movie/TV categories. In Table 7 MIET stands for mean inter-event time and SDIET stands for standard deviation inter-event time. Movie-related behavioral variables are amongst the top predictors, together with interests and profile variables. For the average and documentary interest variables are more important than profile variables, whereas for the action movies profile variables are more important. Network variables were not found to be important only in the case of an action movie. Overall, we can state that the effect of network variables in the case of movies is less substantive than in other cases (Bogaert et al., 2017). In general, Table 7 informs us that Facebook variables that describe a user's specific movie watching behavior are amongst the top predictors. In addition we see that not only movies but also other video content (e.g., commercials) play an important role. Similarly to Gupta et al. (2008), we also found that a person's music and book interests and even interests in general are related to her movie watching behavior. For the average, we see that music and book interest variables were among the top predictors. For a documentary, television-related interests were important as well. This supports the theory that movie watchers are not only interested in movies, but show interest in a large variety of media-related topics. Finally, time-related variables (e.g., recency and mean inter-event time) play an important role in determining whether or not a user will watch a movie. Previous research on romantic tie prediction has shown that time-based predictors play an important role in social media predictions (Bogaert et al., 2016b). A final observation is that age has an influence on all movie genres, however for some genres more (e.g., documentary) more than others.

Table 7: Overview of the median sensitivity scores

Rank	Average			Action			Documentary		
	Variable	Sensitivity score	Type	Variable	Sensitivity score	Type	Variable	Sensitivity score	Type
1	**Number (movies)**	0.1153	B	**Number (movies)**	0.1213	B	**Number (videos)**	0.1010	B
2	**Number (videos)**	0.0244	B	**Number (videos)**	0.0230	B	**MIET (movies)**	0.0212	B
3	**MIET (movies)**	0.0153	B	**MIET (movies)**	0.0197	B	**Number (movies)**	0.0117	B
4	**REC(movies)**	0.0016	B	IND (gender == female)	0.0069	P	Category like (company)	0.0109	I
5	**Category music (musician)**	0.0012	I	Mean (photo count)	0.0023	B	**Number (interests)**	0.0101	I
6	IND (gender == female)	0.0009	P	Profile completeness	0.0022	P	Age	0.0051	P
7	Category book (series)	0.0009	I	Category television (TV show)	0.0021	I	Like category (school)	0.0038	I
8	Number (books)	0.0008	I	**Category music (musician)**	0.0018	I	MIET (comments checkins)	0.0030	B
9	Age	0.0007	P	**Number (interests)**	0.0016	I	MIET (status comments)	0.0027	B
10	Category book (movie)	0.0006	I	Like category (health)	0.0015	I	Number (status likes)	0.0026	B
11	Number (favorite teams)	0.0006	I	Category music (club)	0.0012	I	Number (public albums)	0.0023	B
12	SDIET (likes updated)	0.0005	B	SDIET (comments videos)	0.0012	B	Number (friend albums)	0.0023	B
13	SDIET (status comments)	0.0005	B	PERCENT (friends watch movies)	0.0012	N	Number (books)	0.0020	I
14	Number (categories movies)	0.0004	B	Age	0.0011	P	Category book (TV)	0.0019	I
15	MIET (likes)	0.0004	B	Category books (serie)	0.009	I	Category book (book)	0.0019	I

Note: B represents behavioral data, I interests, N network data and P profile data.

To obtain a better understanding of the relationship between whether or not a user will (report to) watch a movie genre and its predictors, we build partial plots for the top 6 predictors (Figure 5 and bold fonts in Table 7). The most important predictors of movie watching behavior are the number of movies a user watched and the number of videos a user has reported to watch (see Figure 5a and 5b). The difference is that videos includes movies, as well as other types of videos such as commercials and home-made movies. For both predictors we observe a positive relationship with the propensity of watching the average and the action movie. For a documentary the relationship is also positive but less steep, especially in the case of the number of movies (Figure 5a). A user who has watched a lot of movies before is likely to be a habitual movie watcher and therefore her frequency of watching movies will be higher. This is in line with the movie recommendations literature where previously watched shows on Facebook are a good base for recommendations (David et al., 2012). Other studies, for example in the direct marketing domain, also confirm that frequency is indeed very valuable when predicting the probability of a user repeating a certain action (Van den Poel, 2003).

Another important variable is the average time between watching consecutive movies (Figure 5c[2]). We notice a negative relationship between the mean inter-event time of watching movies and the probability of watching a specific movie. This implies that the longer the time between watching two movies on average, the lower the chances of watching a movie. This finding can be explained by the lag or the spacing effect (Cepeda et al., 2009). This theory states that an increase in lag is associated with a decline of recall in memory. Also, the MIET can be seen as a measure of intensity where a lower MIET signifies a high level of intensity. In our case this means that users who watch movies with high intensity, have a higher chance of watching a given movie. Closely related to the MIET is the time since a user last watched a movie (Figure 5d). We find that the higher the recency of the last movie someone has watched, the higher the probability of watching a certain movie genre for the average and a documentary. We notice that for an action movie we see that between 800 and 1200 days the probability is higher for a lower and higher recency.

Two other important variables are the number of musicians or bands a user has liked (see Figure 5e) and the number of interests in general (Figure 5f), which are positively related to the response. This is in line with Gupta et al. (2008) who found a correlation between people's taste in music

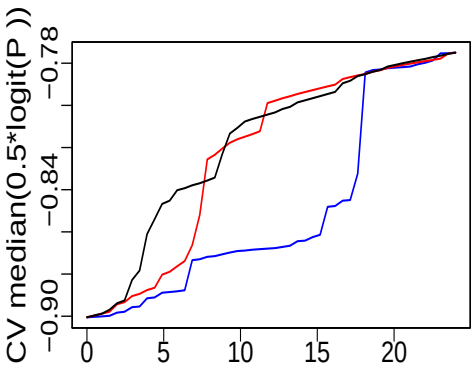
[2]All time variables are expressed in days

and their taste in books. Also, people who are more inclined to indicate the movies they watch, are also more inclined to like more media-related and interest pages, such as music or actors. For the documentary the relationship between the number of music-related likes and movie watching behavior is less positive in the beginning but becomes more positive afterwards (see blue line in Figure 5e).

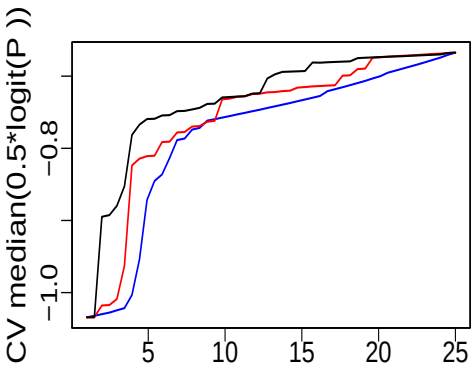
5. Discussion and implications

In this study we have used a data analytical methodology, which identifies users with the highest probability of watching a certain movie genre. To evaluate the viability of predicting self-reported movie watching behavior, our data analytical framework consisted of two stages: a predictive and a descriptive stage. For the predictive stage, we used five classification models (i.e., logistic regression, naive Bayes, random forest, adaboost and rotation forest) to estimate movie watching behavior and evaluate performance. This process was cross-validated for the 10 most popular movies in our database. For the descriptive stage, we applied information-fusion sensitivity analysis to evaluate the most important predictors and their relationship with the response. Next to the cross-validated results, we also reported the results of the sensitivity analysis of individual movie genres to provide insight in the differences between movie genres. In order to cope with the sample selection effects in our data, we resampled the data to be representative of the age and gender characteristics of the whole Facebook population and used randomized oversampling and SMOTE to account for class imbalance.

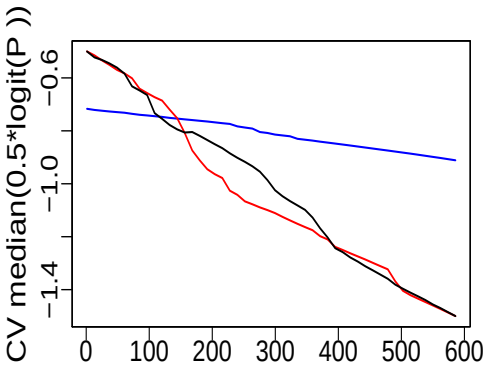
Our findings indicate that we can identify potential movie watchers with high predictive performance. We report a median 10-fold cross-validated AUC from 64.57% to 82.68% across different algorithms. In terms of both accuracy and AUC adaboost was found to be the top performer followed by random forest, rotation forest, logistic regression and naive Bayes. Random forest and rotation forest performed equally well in statistical terms for both ROS and SMOTE. These insights are important for movie producers and advertisers. Before, movie producers had to rely on descriptive studies of market research firms that described their target group in terms of socio-demographics, location and preferences. A problem with this profiling is that it does not take into account the probability that the target group will actually watch a certain type of movie. With our model, movie producers now have the possibility to implement a predictive targeted



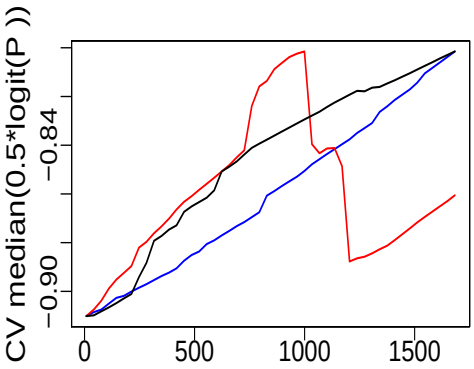
(a) Number (movies)



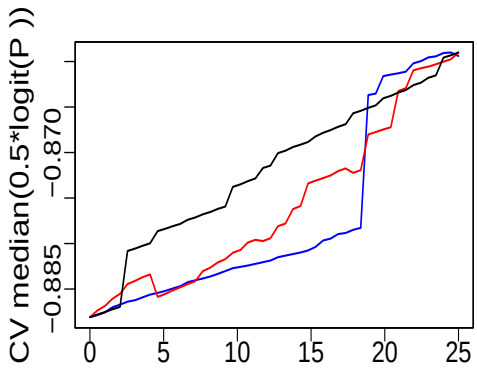
(b) Number (videos)



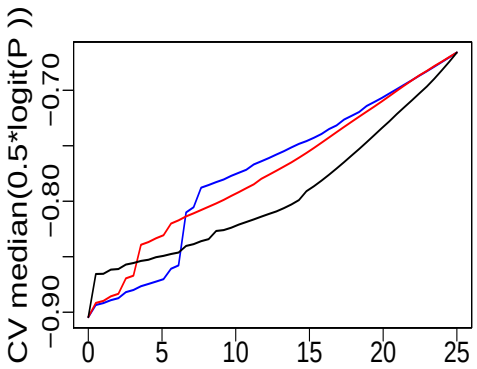
(c) MIET (movies)



(d) REC (movies)



(e) Category music(Musician/bands)



(f) Number (interests)

Figure 5: Partial dependence plots of 6 predictors. The black line represent the average relationship between the predictor and response for the cross-validated models, the red line for an average action movie and the blue line for an average documentary.

marketing approach. Instead of focusing on high-level features, our model identifies users with a high probability of watching a certain type of movie based on the Facebook behavior of similar movie watchers. Hence, producers do not need to rely on expensive market research reports, but can immediately target the users directly. We also build several models across five algorithms to find out which algorithms performs best in predicting movie watching behavior. Moreover, we have an unprecedented number of Facebook variables and give insight into the model based on these variables. This list of top predictors and algorithms assists practitioners in building the best possible predictive model for movie watching behavior. Based on this list, companies can calculate the probability of the user to watch a movie and estimate the extra profits of setting up a targeted marketing approach in contrast to gut feeling (Burez and Van den Poel, 2007). A company could select the top-performing algorithm (adaboost in our case) and conduct a ROC-curve analysis and identify several scenarios for implementing a one-to-one strategy (Bogaert et al., 2017). For example, if a company has a restricted budget it will want to minimize the number of false positives, since this induces a cost for the focal firm.

Our sensitivity analysis revealed that the top predictors were related to behavioral, interest

and profile data. For the overall average movie and the average documentary, interest variables were more important than profile variables, and the opposite was true for the average action movie. Especially variables related to previous movie watching behavior (e.g., the number of movies watched and the time since a given user watched his last movie) were important. The total number of movies and videos in general had a positive influence. However, this relationship was less strong for a documentary. For MIET of watching movies we found that the longer it has been since someone has watched a movie or the longer the time span between watching movies, the lower the chances of watching again. The opposite was found for the recency of watching a movie. Finally, interests in music, books, and in general were found to be important and positively related to the chances of watching a movie. For example, using our results we can say that a user who is a frequent and intense movie watcher, who has declared to have a lot of interests on Facebook, has a higher probability of watching a documentary and indicating it on Facebook. Again, these findings provide important insights for movie producers and advertisers who want to replicate our results. Practitioners want to build predictive models that are both accurate and efficient. Instead of creating all possible variables, our list of top predictors provides guidance as to which predictors to include for several genres. Also, our variable importances and partial dependence plots help practitioners in pinpointing which variables to monitor when targeting potential movie watchers. For example, users who have declared to be interested in various media-related topics (such as music or TV shows), also have a greater probability of watching an action movie.

6. Limitations and future research

A first limitation of this study is that some of our variables are limited in the number of values. Facebook restricts the number of values per variable that can be extracted through an application to the 25 most recent entries of that variable. This restriction mostly impacts frequency variables. In order to deal with this limitation, we calculated the frequency within a specific time period and determined the length of this time window for each variable, since no user in our data set reached the limit of 25 entries. The frequency of status updates, photo uploads and links created was computed for the last 7 days, album uploads and check-ins for the last 4 months and notes and video uploads for the last year.

A second limitation is related to the the lack of more movie-specific variables such as someone's

favorite movies. An interesting avenue could be the application of text mining to obtain meaningful data from various pieces of unstructured textual information. However, preliminary analysis did not achieve extra predictive performance.

A third limitation is that our study does not include rating data of movies. Unfortunately our data set did not include this information. A future research path could be to include these data and see how they influence predictive performance and the variable importances.

A fourth limitation is related to the self-selection bias in our data set. We gathered our data via a customized Facebook application for a European soccer team, which was advertised several times on their Facebook page. To stimulate participation we offered a signed jersey. To avoid privacy issues we made the users aware their data were extracted and included a rules and regulations section containing our contact information. Furthermore, we ensured the participant that all extracted data would be anonymous. Yet, our data suffer from self-selection problems. First, users should be interested in the prize to participate in the contest. Second, users who do not like the European soccer team on Facebook (or more in general soccer) have a lower probability of being in our sample, since it was advertised via the Facebook page of the European soccer team. Users not interested in soccer could still see the app in the News Feed, however sign-up rates in this group would be lower. Third, users may not be willing to share their data with Facebook. As a result, our data set is not fully representative for the whole population of Facebook users. We tried to resolve these selection biases by resampling the data such that the age demographics per gender in our sample have the same distribution as the general Facebook population. However, we acknowledge that resampling only decreases and does not remove all the self-selection bias. However, companies that want to replicate our results will have the same limitation: they will advertise their application through Facebook, have users who are not willing to share their data and will resample the data according to the age and gender demographics. Nevertheless, we are offering a valuable case study that practitioners, and Facebook, can use as a road map to implementing a targeted advertising approach for movie watchers. It is important to note that in this specific case; we do not have data about the users that saw the app and decided not to engage with the app. Since the News Feed Algorithm controls who sees the application in their personal News Feed, it is therefore not possible to alleviate these biases statistically. Therefore, we choose the next best option and resample the data according to the user demographics in order to make our results more generalizable to the

whole population. Any and all studies using Facebook data have this problem and therefore these studies should be considered case studies. It is our hope that our findings stimulate research in this field that can replicate our findings. More researchers need to collect their own data sets and determine if our findings hold, to be able to approach a more generalizable meta conclusion (Hanssens, 2018). In that sense our study is the first piece of the puzzle.

A fifth limitation is related to the choice of our dependent variable. We model whether or not a user has declared to have watched a movie from a certain genre. By doing so, we neglect the people that have watched a certain type of movie but did not share this on Facebook. Again, there are different kinds of biases introduced here. First, there is a social desirability bias. For example users could be more or less inclined to share their movie watching behavior because of their friends (dis-)approving it. Second, there is the issue of availability. Users need to remember whether or not they watched a movie and on top of that share it. Hence, we only model a subset of the target population (i.e., all movie watchers). However, we believe that our choice of dependent variable is the best possible proxy for movie watching behavior. The worst solution would be to target no one or at random. However, these approaches rarely induce an increase in firm performance. The best solution would be to target everyone who actually watched a movie from the target genre. However, this last solution is not possible using social media data and hence researchers are always forced to work with a subsample of the target population. Hence, in terms of prospecting, the best proxy for real movie watching behavior is self-reported movie watching behavior. From a purely predictive perspective it is not a problem if we are only targeting a part of all movie watchers on Facebook (in this case we are targeting the users who watched a certain movie type and at the same time have this listed in their profile).

Although our study has the aforementioned problems, it is the first to come up with a model that predicts movie watching behavior as opposed to traditional movie recommendations and box-office revenue studies. This study answers several important questions managers struggle with today: 'Which consumers/users to target?', 'Where to get prospects and the data to do so?', and 'Is it feasible to execute a one-on-one strategy?'. By solving these questions in the movie industry, movie producers and advertisers now have insights into the potential of a major database (e.g., Facebook contains 25% of the world population), state-of-the-art algorithms and numerous variables to predict movie watching behavior. Therefore, we are confident that this study makes a

significant contribution to existing literature.

Acknowledgements

The authors are thankful to the two anonymous reviewers whose comments have helped significantly improve this paper. The authors are also grateful to the Guest Co-Editor of the Data Mining & Decision Analytics Special Issue, Dr. Asil Oztekin, for the very timely management of this manuscript.

References

Apala, K. R., Jose, M., Motnam, S., Chan, C. C., Liszka, K. J., Gregorio, F. d., Aug. 2013. Prediction of movies box office performance using social media. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 1209–1214.

Arias, M., Arratia, A., Xuriguera, R., Jan. 2014. Forecasting with Twitter Data. ACM Trans. Intell. Syst. Technol. 5 (1), 8:1–8:24.

Asur, S., Huberman, B. A., Aug. 2010. Predicting the Future with Social Media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Vol. 1. pp. 492–499.

Babyak, M. A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine 66 (3), 411–421.

Ballings, M., Van den Poel, D., Dec. 2012. Customer event history for churn prediction: How long is long enough? Expert Systems with Applications 39 (18), 13517–13522.

Ballings, M., Van den Poel, D., Jul. 2015a. CRM in social media: Predicting increases in Facebook usage frequency. European Journal of Operational Research 244 (1), 248–260.

Ballings, M., Van den Poel, D., May 2015b. rotationForest: Fit and Deploy Rotation Forest Models.
URL `https://cran.r-project.org/web/packages/rotationForest/index.html`

Ballings, M., Van den Poel, D., Bogaert, M., Mar. 2016. Social media optimization: Identifying an optimal strategy for increasing network size on Facebook. Omega 59, 15–25.

Basuroy, S., Chatterjee, S., Ravid, S. A., Oct. 2003. How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. Journal of Marketing 67 (4), 103–117.

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

Benedek, G., Lublóy, Á., Vastag, G., Feb. 2014. The Importance of Social Embeddedness: Churn Models at Mobile Providers. Decision Sciences 45 (1), 175–201.

Berk, R. A., 2008. Statistical learning from a regression perspective. Springer.

Biau, G., 2012. Analysis of a random forests model. The Journal of Machine Learning Research 13 (1), 1063–1095.

Bogaert, M., Ballings, M., Hosten, M., Van den Poel, D., Nov. 2017. Identifying Soccer Players on Facebook Through Predictive Analytics. Decision Analysis 14 (4), 274–297.

Bogaert, M., Ballings, M., Van den Poel, D., Feb. 2016a. The added value of Facebook friends data in event attendance prediction. Decision Support Systems 82, 26–34.

Bogaert, M., Ballings, M., Van Den Poel, D., Aug. 2016b. Evaluating the importance of different communication types in romantic tie prediction on social media. Annals of Operations Research Forthcoming, 1–27.

Borsato, F. H., Polato, I., Oct. 2012. May Social Behavior Reveal Preferences on Different Contexts? Recommending Movie Titles Based on Tweets. In: 2012 Brazilian Symposium on Collaborative Systems. pp. 121–126.

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

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

Burez, J., Van den Poel, D., 2007. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Syst. Appl. 32 (2), 277–288.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., Pashler, H., 2009. Optimizing distributed practice: theoretical analysis and practical implications. Experimental Psychology 56 (4), 236–246.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357.

Coussement, K., Van den Poel, D., Jan. 2008. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34 (1), 313–327.

Cox, D. R., 1958. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological) 20 (2), 215–242.

Croux, C., Joossens, K., Lemmens, A., Sep. 2007. Trimmed bagging. Computational Statistics & Data Analysis 52 (1), 362–368.

Culp, M., Johnson, K., Michailidis, a. G., Jun. 2012. ada: an R package for stochastic boosting.

Dag, A., Oztekin, A., Yucel, A., Bulur, S., Megahed, F. M., 2016. Predicting heart transplantation outcomes through data analytics. Decision Support Systems.

David, J., Bajaj, S., Jazra, C., 2012. A Facebook Profile-Based TV Recommender System. vectors 1, u2.

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

Dellarocas, C., Zhang, X. M., Awad, N. F., 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive marketing 21 (4), 23–45.

Demšar, J., Dec. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7, 1–30.

Dietterich, T. G., 1996. Statistical tests for comparing supervised classification learning algorithms. Oregon State University Technical Report 1, 1–24.

Ding, C., Cheng, H. K., Duan, Y., Jin, Y., 2016. The power of the "like" button: The impact of social media on box office. Decision Support Systems.

Dipak Damodar Gaikar, Bijith Marakarkandy, Chandan Dasgupta, Oct. 2015. Using Twitter data to predict the performance of Bollywood movies. Industrial Management & Data Systems 115 (9), 1604–1621.

Du, J., Xu, H., Huang, X., Mar. 2014. Box office prediction based on microblog. Expert Systems with Applications 41 (4, Part 2), 1680–1689.

Duan, W., Gu, B., Whinston, A. B., Nov. 2008. Do online reviews matter? — An empirical investigation of panel data. Decision Support Systems 45 (4), 1007–1016.

El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., Fischer, F., Simon, S., Schreck, T., Keim, D. A., 2013. Visual analytics for the prediction of movie rating and box office performance. IEEE VAST Challenge USB Proceedings.

Eren Demir, Oct. 2014. A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines. Decision Sciences 45 (5), 849–880.

Estabrooks, A., Jo, T., Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. Computational intelligence 20 (1), 18–36.

Facebook, 2016. Audience Targeting Options.
URL https://www.facebook.com/business/help/633474486707199

Facebook, 2018. Facebook-advertising.
URL https://nl-nl.facebook.com/business/products/ads

Forbes, 2014. How Has Movie Marketing And Distribution Evolved Over Time?
URL http://www.forbes.com/sites/quora/2014/02/11/how-has-movie-marketing-and-distribution-evolved-over-time/#45d1e9ac4733

Freund, Y., Schapire, R., Abe, N., 1999. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14 (771-780), 1612.

Freund, Y., Schapire, R. E., others, 1996. Experiments with a new boosting algorithm. In: ICML. Vol. 96. p. 148â€“156.

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

Friedman, J. H., Feb. 2002. Stochastic Gradient Boosting. Comput. Stat. Data Anal. 38 (4), 367–378.

Friedman, J. H., Meulman, J. J., May 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22 (9), 1365–1381.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., Watts, D. J., Oct. 2010. Predicting consumer behavior with Web search. Proceedings of the National Academy of Sciences 107 (41), 17486–17490.

Golbeck, J., Jan. 2006. Generating Predictive Movie Recommendations from Trust in Social Networks. In: Trust Management. Springer Berlin Heidelberg, pp. 93–104.

Guisan, A., Edwards, T. C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological modelling 157 (2), 89–100.

Guo, H., Pathak, P., Cheng, H. K., Feb. 2015. Estimating Social Influences from Social Networking Sites—Articulated Friendships versus Communication Interactions. Decision Sciences 46 (1), 135–163.

Gupta, A., Jain, R., Song, S., 2008. Movie Recommendations Using Social Networks. Stanford University Stanford, CA.

Hand, D. J., Anagnostopoulos, C., Apr. 2013. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognition Letters 34 (5), 492–495.

Hanley, J. A., McNeil, B. J., Apr. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.

Hanssens, D. M., Jan. 2018. The value of empirical generalizations in marketing. Journal of the Academy of Marketing Science 46 (1), 6–8.

He, H., Garcia, E. A., Sep. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21 (9), 1263–1284.

Hennig-Thurau, T., Wiertz, C., Feldhaus, F., Jun. 2014. Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. Journal of the Academy of Marketing Science 43 (3), 375–394.

Hernandez-Orallo, J., Flach, P., Ferri, C., Oct. 2012. A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. Journal of Machine Learning Research 13, 2813–2869.

Hernandez-Orallo, J., Flach, P., Ferri, C., Oct. 2013. ROC Curves in Cost Space. Mach. Learn. 93 (1), 71–91.

Jain, V., 2013. Prediction of movie success using sentiment analysis of tweets. The International Journal of Soft Computing and Software Engineering 3 (3), 308–313.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Vol. 103 of Springer Texts in Statistics. Springer New York, New York, NY.

Janitza, S., Strobl, C., Boulesteix, A.-L., 2013. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics 14, 119.

Kim, T., Hong, J., Kang, P., Apr. 2015. Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting 31 (2), 364–390.

Kuncheva, L. I., Rodriguez, J. J., 2007. An experimental study on rotation forest ensembles. In: Multiple Classifier Systems. Springer, pp. 459–468.

Langley, P., Iba, and, W., Thompson, K., 1992. An Analysis of Bayesian Classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press, San Jose, California, pp. 223–228.

Lee, K., Park, J., Kim, I., Choi, Y., Aug. 2016. Predicting movie success with machine learning techniques: ways to improve accuracy. Information Systems Frontiers, 1–12.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.

Liu, T., Ding, X., Chen, Y., Chen, H., Guo, M., Oct. 2014. Predicting movie Box-office revenues by exploiting large-scale social media content. Multimedia Tools and Applications 75 (3), 1509–1528.

Liu, Y., Jul. 2006. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. Journal of Marketing 70 (3), 74–89.

Liu, Y., Chen, Y., Lusch, R., Chen, H., Zimbra, D., Zeng, S., 2010. User-Generated Content on Social Media: Predicting Market Success with Online Word-of-Mouth. SSRN Scholarly Paper ID 2655800, Social Science Research Network, Rochester, NY.

Liu, Y., Huang, X., An, A., Yu, X., 2007. ARSA: A Sentiment-aware Model for Predicting Sales Performance Using

Blogs. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07. ACM, New York, NY, USA, pp. 607–614.

Meire, M., Ballings, M., Van den Poel, D., Sep. 2016. The added value of auxiliary data in sentiment analysis of Facebook posts. Decision Support Systems 89, 98–112.

Mestyán, M., Yasseri, T., Kertész, J., Aug. 2013. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLOS ONE 8 (8), e71226.

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

Mishne, G., 2006. Predicting movie sales from blogger sentiment. In: In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW). pp. 155–158.

Moon, S., Bergey, P. K., Iacobucci, D., Jan. 2010. Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction. Journal of Marketing 74 (1), 108–121.

Oh, C., Roumani, Y., Nwankpa, J. K., Hu, H.-F., 2016. Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. Information & Management.

Oztekin, A., Sep. 2016. A hybrid data analytic approach to predict college graduation status and its determinative factors. Industrial Management & Data Systems 116 (8), 1678–1699.

Oztekin, A., Delen, D., Turkyilmaz, A., Zaim, S., Dec. 2013. A machine learning-based usability evaluation method for eLearning systems. Decision Support Systems 56, 63–73.

Oztekin, A., Kizilaslan, R., Freund, S., Iseri, A., Sep. 2016. A data analytic approach to forecasting daily stock returns in an emerging market. European Journal of Operational Research 253 (3), 697–710.

Pham, X. H., Jung, J. J., Le Anh Vu, S.-B. P., 2014. Exploiting social contexts for movie recommendation. Malaysian Journal of Computer Science 27 (1), 68–79.

Prinzie, A., Van den Poel, D., 2007. Random multiclass classification: generalizing random forests to random MNL and random NB. In: LECTURE NOTES IN COMPUTER SCIENCE. Vol. 4653. Springer, pp. 349–358.

Reddy, A. S. S., Kasat, P., Jain, A., 2012. Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining. International Journal of Computer Applications 56 (1).

Rish, I., 2001. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. IBM New York, pp. 41–46.

Rodriguez, J., Kuncheva, L., Alonso, C., Oct. 2006. Rotation Forest: A New Classifier Ensemble Method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10), 1619–1630.

Rui, H., Liu, Y., Whinston, A., Nov. 2013. Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems 55 (4), 863–870.

Said, A., De Luca, E. W., Albayrak, S., 2011. Using Social and Pseudo-Social Networks for Improved Recommendation Quality. In: Workshop chairs. p. 45.

Sandri, M., Zuccolotto, P., 2006. Variable selection using random forests. In: Data analysis, classification and the forward search. Springer, pp. 263–270.

Sevim, C., Oztekin, A., Bali, O., Gumus, S., Guresen, E., Sep. 2014. Developing an early warning system to predict currency crises. European Journal of Operational Research 237 (3), 1095–1104.

Shapira, B., Rokach, L., Freilikhman, S., Sep. 2012. Facebook single and cross domain data for recommendation systems. User Modeling and User-Adapted Interaction 23 (2-3), 211–247.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

Van den Poel, D., 2003. Predicting mail-order repeat buying: which variables matter? Tijdschrift voor economie en management 48 (3), 371–404.

Venkatesan, M., Mai, A., 2012. Recommendation of TV shows and Movies based on Facebook data. Citeseer.

Wolpert, D. H., Oct. 1996. The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation 8 (7), 1341–1390.