

MIT Open Access Articles

Behavioral attributes and financial churn prediction

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Kaya, Erdem, et al. "Behavioral Attributes and Financial Churn Prediction." EPJ Data Science, vol. 7, no. 1, Dec. 2018. © 2018 The Authors

As Published: <https://doi.org/10.1140/epjds/s13688-018-0165-5>

Publisher: Springer Berlin Heidelberg

Persistent URL: <http://hdl.handle.net/1721.1/118772>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution





Behavioral attributes and financial churn prediction

Erdem Kaya^{1*} , Xiaowen Dong^{2,3}, Yoshihiko Suhara^{2,4}, Selim Balcisoy¹, Burcin Bozkaya¹ and Alex “Sandy” Pentland²

*Correspondence:

erdemkaya@sabanciuniv.edu

¹Behavioral Analytics and Visualization Lab, Sabanci University, Istanbul, Turkey

Full list of author information is available at the end of the article

Abstract

Customer retention is crucial in a variety of businesses as acquiring new customers is often more costly than keeping the current ones. As a consequence, churn prediction has attracted great attention from both the business and academic worlds. Traditional efforts in the financial domain mainly focus on domain specific variables such as product ownership or service usage aggregation, however, without considering dynamic behavioral patterns of customers' financial transactions. In this paper, we attempt to fill in this gap by investigating the spatio-temporal patterns and entropy of choices underlying the customers' financial decisions, and their relations to customer churning activities. Inspired by previous works in the emerging field of computational social science, we built a prediction model based on spatio-temporal and choice behavioral traits using individual transaction records. Our results show that proposed dynamic behavioral models could predict churn decisions significantly better than traditionally considered factors such as demographic-based features, and that this effect remains consistent across multiple data sets and various churn definitions. We further study the relative importance of the various behavioral features in churn prediction, and how the predictive power varies across different demographic groups. More generally, the proposed features can also be applied to churn prediction in other domains where spatio-temporal behavioral data are available.

Keywords: Churn prediction; Customer behavior; Spatio-temporal patterns; Credit card data

1 Introduction

We live in the era of “Big Data”. The recent availability of large-scale quantitative behavioral data provides opportunity to study human and social behavior at an unprecedented scale, leading to the emerging field of computational social science [1]. The breadcrumbs that we leave behind with everyday activities seem to reveal more about our conscious behaviors and decisions than our socio-demographic characteristics. For example, only little information about when and where we make purchases can predict our financial well-being [2], and the implicit patterns in our communication networks can determine our performance at both individual and group levels [3]. Despite the promising results of these efforts, however, current business analytics solutions and the related body of research still seem to lack consideration of customers' spatio-temporal behavior.

In many businesses, predicting the next set of customer actions is of central importance as this capability enables companies to forestall undesirable decisions of the customers. Among those, *churn prediction* has attracted increasing attention from a variety of domains such as telecommunication and banking industries as well as researchers in academia, as retaining customers is far less costly than acquiring new ones [4]. The cost of making a new customer as profitable as a current customer could be up to 16 times higher than the cost of retaining efforts [5], and decreasing the churn rate by only 5% can increase the profitability by 25–125% [6]. According to a survey carried out with over 24 thousand customers in 33 countries [7], 68% of the churning customers expressed that they would not do business again with the companies that they left. The cost of such provider switches of customers is estimated to be \$1.6 trillion for the United States. However, predicting whether the customer will quit his or her contract seems to be a rather daunting task, due to the unpredictable nature of active decisions of customers, such as quitting contract due to unsatisfactory service [8], and incidental or non-voluntary events, such as change of home/work locations or financial troubles [9]. Complexity becomes even worse for the financial churn prediction due to the relative sparsity of the transactions compared to other domains such as telecommunication. Furthermore, the financial decisions might require longer investigation periods (e.g., loan) leading to the development of heuristics for the churn prevention efforts rather than prediction models based on transactional data.

The problem of churn prediction has been tackled in many different domains such as telecommunication [10–15], banking [16, 17], subscription services [18], game businesses [19], and retailing [20]. In general, most of the efforts in churn studies involve prediction with different definitions of the churn event [14, 21], evaluating new data mining algorithms [15, 22, 23], and introducing ways to deal with large volumes of data [24]. In these studies, churn prediction has usually been considered as a benchmark against which novel analytic approaches are evaluated, and domain-specific historical actions (e.g., aggregation of service usage in unit time) of customers are typically involved as features for prediction. Neslin et al. [25] analyze the effect of the methodological factors on the accuracy of churn prediction models across top-decile lift and Gini criteria. Verbraken et al. [26] further investigated the techniques to employ for churn prediction problem by analyzing effectiveness of Bayesian Network classifiers. They suggest that Bayesian approaches could produce lean models which can aid the analysts in making judgments on the dependencies between explanatory variables. Verbeke et al. [27] argue that, in order to minimize the cost associated with customer churn, the statistical evaluation approaches tend to yield suboptimal prediction models. They introduce a profit-driven data mining approach which determines the optimal model by targeting the customers with whose retention profit is maximized.

As a more related research, Tang et al. [28] conducted a study where they employed demographic features, macroeconomic variables, and financial information such as policy purchase in order to predict customer churn decision. Compared to these domain-related features, we consider the dynamic spatio-temporal “patterns” (diversity, loyalty, regularity) of spending activities, and entropy of fund transfer and purchase transactions. In terms of prediction, Tang et al. converted the original financial features into the derived features by applying the orthogonal polynomial approximation approach. In comparison, we developed novel features based on spatio-temporal and choice behaviors of customers, and built a Random Forest model which is an ensemble method of non-linear tree classifiers.

The model is capable of taking into account various combinations of different features, and the *maximum depth* hyper-parameter corresponds to the degree of polynomial terms. From this vantage point, our prediction methodology is similar to Tang et al.'s in terms of incorporating higher-degree feature combinations.

On the other hand, a number of approaches developed by the spatial data community [29, 30] has also been applied to financial applications. Fernandes et al. [31] develop a credit scoring model that takes into account spatial proximity. Agarwal et al. [32] investigate spatial correlation of defaults in subprime mortgage, and Bourassa et al. [33] compare the methodologies for house price prediction based on spatial features. Along with others, these studies show that spatial properties could be important indicators in the financial domain.

Recent studies suggest that behavioral traits in our everyday activities may better explain the phenomenon under investigation. For example, diversity of phone communication or interaction within social networks has been shown as a strong indicator of economic development of communities [34] and financial status of individuals [35]. Behavioral traits in customers' daily purchases, which are computed based on individual financial transaction data, can predict financial well-being of the customers significantly better than demographic features [2]. These findings highlight the potential of making use of patterns behind social interaction (e.g., phone calls, face-to-face meetings) and decision-making (e.g., expenditures) in financial outcome prediction.

In this article, we focus on spatio-temporal patterns of customer spending behavior as well as the entropy of their choices in fund transfer and purchase activities, and investigate whether features extracted from such financial behaviors could be utilized in the prediction of churn decisions of bank customers. To this end, based on individual credit card transaction records of a large set of customers, we develop spatio-temporal behavioral features, namely diversity, loyalty, and regularity, that were introduced by Singh et al. [2] with minor modifications. We also introduce a set of novel entropy of choice features which reflect the diversity in the financial choices made by the customers, such as the merchants to purchase from, the shopping categories, and the addressee of the fund transfers. It is worth noting that unlike the traditional features, the behavioral features employed in this study mainly capture the dynamic behavioral patterns that the customer follows in both temporal and spatial domains, and the entropy of the choices that constitute these behavioral patterns. We argue that such features are more important in differentiating the customers as opposed to what can be achieved with demographic attributes which are usually constant over long periods of time. In general, we name these implicit mobility and entropy of choice patterns as spatio-temporal and choice (STC) features.

Our findings suggest that the proposed STC features are significantly better than demographic features in bank customer churn prediction. We report that diversity and regularity in customers' spatio-temporal activities and entropy of financial choices are more important than other behavioral traits in financial churn prediction. In particular, the exploration levels of the eventual churners, reflected by the diversity patterns, had a decreasing trend during the observation window, while this was not the case for non-churners. This seems to suggest that deviations from a customer's usual spatio-temporal spending patterns could be indicators of the presence of the financial stress that could be associated with his churn decision, in a way similar to the increasing vulnerability of living organisms as a reaction to persistent stress [36]; however, validity of such relationship needs to be

further investigated in a future research, by seeking a relationship between deteriorating financial wellbeing of a customer and her spatio-temporal behavior change. Furthermore, we conducted the same study for different demographic groups and found out that churn prediction seems to be relatively easier for the group of younger customers, while gender-based difference is not significant. These findings also remain consistent over a number of different data sets that are generated based on different sampling strategies and observation windows applied on a much larger and common customer base.

Our paper contributes to a growing body of research on data-driven behavior understanding, and, in particular, provides novel insights into the understanding of the challenging problem of financial churn prediction. Usage of STC behavioral features can improve the existing churn prediction models employed by both financial industry and academia. Furthermore, the spatio-temporal and choice models utilized in this paper can also be applied to churn prediction problems of other domains such as telecommunication industries, where similar information on spatial and temporal activities as well as choice decisions are readily available. In summary, the main contributions of our paper can be summarized as follows:

- We show that spatio-temporal and choice features are superior to demographic features in financial churn decision prediction.
- We demonstrate the performance comparison of demographic and spatio-temporal and choice features based on stratification of demographic groups, which implies that churning decisions of younger people seem to be more easily predicted.
- We introduce entropy of choice features characterizing the behavioral patterns in selecting products, merchants or transfer addressees. Moreover, we analyze the relative performance of each behavioral feature to investigate feature importance.
- We contribute by introducing novel financial data and churn definitions.

2 Materials and methods

2.1 Data

A major financial institution in an OECD country donated two de-identified samplings of their data that were collected over the period between July 2014 and July 2015. The samples comprise demographic information, credit card transactions, money transfers, and electronic fund transfers (EFT) of over 100 thousand (Sample-A) and 60 thousand (Sample-B) customers. Sample-A and Sample-B contain in total roughly 45 millions and 22 millions of transactions, respectively. Both samples were drawn from a much larger sampling of 450 thousand customers who were located in a major metropolitan city, updated their home and work addresses since January 2012, and made at least one credit card transaction during the sampling period. Sample-A was drawn randomly from this larger set whereas Sample-B was drawn from the same set such that each customer has at least 10 credit card transactions, and in total around 60% of all the credit card transactions were performed with point of sale (POS) machines of the bank donating the data. Bank officials reported that de-identification of Sample-A and B was done independently precluding determination of the number of customers contained by both of the samplings. Customers may prefer to use their credit cards on the POS machines of other banks. In that case, some part of the transaction information such as location cannot be collected by the bank issuing the credit card. The bank also donated monthly segmentation information for each customer. We further elaborate on the segmentation information and the way we utilize it for label generation in the Labeling subsection.

Table 1 Data set characteristics

Data Set	Source	Observation Win.	Labeling Win.	# of TXs	# of Cust.	Label Sets	Churn (%)
A1	Sample A	07/2014–06/2015	07/2015–11/2015	8.5M / 3.3M	55K	SB	1.97
A2	Sample A	07/2014–03/2015	04/2015–06/2015	6.3M / 2.4M	53K	SB, CC, CA	0.99
B1	Sample B	07/2014–06/2015	07/2015–11/2015	4.2M / 2.6M	43K	SB	2.27
B2	Sample B	07/2014–03/2015	04/2015–06/2015	3.1M / 1.9M	42K	SB, CC, CA	1.42

Based on samples A and B, four data sets with different characteristics have been generated. The summary includes the sampling source of the data set, observation window for feature generation, labeling window for churn decision of the customers, count of all transactions and the transactions with POS location information (# of TXs), number of customers (# of Cust.), the label sets generated for the related data set, where SB, CC, and CA stand for segmentation-based, credit card usage-based, and checking account usage-based labeling, and finally percentage of churning customers in the data set (Churn(%)) according to label *inac-full*. The transaction and customer counts represent the state after the data filtering process.

The customer transactions and demographic information were anonymized by the bank officials by masking the unique identifier and names of the customers. For each customer, a pseudo-unique identifier has been generated for cross-referencing of different transaction sets. The credit card transactions of customers with missing home and work location information, and the transactions without merchant or location information were also not included in the calculations. We publish only the demographics information along with the calculated behavioral features rather than the transaction data in order to prevent re-identification of customers.

In addition to customer spending transaction data, the bank donated customer segment information which is generated for each month. Customer segmentation is basically a mapping from the customer set to a segment set which is formed of customer types such as salary customer, loan customer, or credit card user. We further elaborate on the customer segmentation in Labeling subsection. While the samplings (A and B) of the customer transactions were collected over a 12-month period, the customer segment information covers a 23-month period including the sampling period such that the segmentation information of the customers were available for an additional five-month period following the sampling window. This additional information enabled us to generate two variants of each of the samplings by defining 12- and 9-month observation windows for feature extraction, and 5- and 3-month churn decision windows for label (churner or non-churner) extraction. This translates into four distinct data sets, namely data sets A1 and A2 generated from Sample-A, and data sets B1 and B2 derived from Sample-B. Please see Table 1 for the characteristics of the data sets.

2.2 Features

In order to establish the relationship between customers' churn decisions and their behavior patterns, we prepared demographic and behavioral features from the donated data set. For the characterization of the behavioral features, we employ a slight variant of the pattern extraction technique described in [2], and in addition, we introduce a new set of features, namely *entropy of choice*, that explains the expenditure and transfer tendencies of the customers.

2.2.1 Demographic features

The demographic features of the customers included gender, marital status, educational status, job type, income, and age of the customers. Except for income, all the demographic information of the customers were available. The missing part of the income information

(less than 2% of the customers) were filled with the mean income of the rest of the customers.

2.2.2 Spatio-temporal and choice (STC) patterns

The behavioral features comprise implicit spatio-temporal expenditure patterns and financial choice patterns. Spatio-temporal expenditure patterns, namely *diversity*, *loyalty*, and *regularity*, refer to the measures of how diverse or loyal customers are in their spending patterns from time and location perspectives, whereas financial choice patterns indicate how customers distributed their financial activities (i.e. online/offline credit card purchases, fund transfers) with respect to merchants, spending categories, and the addressees of the fund transfers.

For the spatio-temporal expenditure patterns, our study benefited from the formulations introduced in [2], but with minor necessary modifications. First, *Diversity* represents the extent of the customers' tendency to make purchases at different locations or times. A high score of diversity means that customer spreads his or her transactions to a large number of *bins*, which can be considered as the slices of time or space, and will be further discussed below. Mathematically, diversity D_i is the normalized entropy of the transactions of customer i with respect to space and time slots (i.e., bins): $D_i = -\sum_{j=1}^N p_{ij} \log_M p_{ij}$, where p_{ij} is the probability of customer i having transaction in bin j , N is the total number of bins, and M is the number of non-empty bins. Our modification to this calculation was that, as the normalization factor of the entropy, we used the *total* number of bins, rather than the number of *non-empty* bins. The downside of using the number of *non-empty* bins is that it would calculate the same diversity scores, for example, for two customers evenly distributing their transactions into *different* number of bins. With our approach, for the same scenario, the customer evenly spreading her transactions into larger number of bins can get a higher diversity score, as expected.

Second, *Loyalty* is the fraction of the customers' transactions in their k -most frequented bins. If f_i is the total number of the expenditures that happened in the top three bins of the customer, then the loyalty for customer i is calculated as $L_i = f_i / \sum_{j=1}^N p_{ij}$.

Finally, *Regularity* represents the level of the similarity of the customers' diversity and loyalty scores over shorter and longer terms. In principle, it is one-complement of the mean Euclidean distances between shorter and longer term score vectors for diversity and loyalty. Regularity is calculated as $R_i = 1 - \sqrt{((D_i^S - D_i^L)^2 + (L_i^S - L_i^L)^2)/2}$, where D_i^S and D_i^L stand for shorter and longer term diversity, respectively. Likewise, L_i^S and L_i^L stand for shorter and longer term loyalty. Regularity scores closer to 1 represent higher regularity indicating having similar diversity and loyalty scores in shorter and longer term periods. For our study, the duration of the shorter term is selected as the one third of the observation window (Please see Table 1 for the observation windows.).

The *bins* mentioned in the calculation of spatio-temporal features can be considered as the slices of time or space. In this study, the space is organized as a collection of square grids (grid bins), and concentric annular areas centered at home and work addresses of the customers (radial-home and radial-work bins). We selected the edge size of the square bins to be 0.1 degree units, and radii of the annular areas to be 0.5, 1, 2, 3, 4, 5, 10, 15, 30, 50, 100, 150, 300, and 500 kilometers. Temporal hourly and weekly bins are the hour of the day and day of the week of a given transaction, respectively. Hence, there are in total 24 temporal hourly and 7 weekly bins.

Three spatio-temporal behavioral traits with five different spatial and temporal variants translates into a set of 15 behavioral features whose names we abbreviated for better readability as follows. For the three behavioral traits diversity, loyalty, and regularity, we used the prefixes *div-*, *loy-*, and *reg-*, respectively. Similarly, for the five bin variants grid, radial-home, radial-work, hourly, and weekly, we considered the suffixes *-g*, *-rh*, *-rw*, *-ho*, and *-we*, respectively. For example, given the *loyalty* trait and the *grid* variant, we abbreviated the feature *grid-based loyalty* as *loyg*.

In addition to the features merely based on temporal and spatial patterns, we introduce the *entropy of choice* behavioral trait representing the variety of the selections that the customers make in their purchase or fund transfer behavior. In other words, it is the entropy of their choices of *products* and *merchants* when shopping or *peers* when making transfers, and calculated as follows:

$$C_i = - \sum_{j=1}^N p_{ij} \log_M p_{ij},$$

where p_{ij} is the probability of customer i making selection j , N is the number of all possible distinct selections, and M is the number of unique selections customer i made. For example, the set of the selections for a particular customer would be all other banks that she has ever made money transfers to, or all the merchants from which she has ever made purchases. Choice behavioral features are money transfer entropy (*transe*), EFT entropy (*efte*), entropy of credit card transactions with respect to merchants (*ecctmer*) and merchant category code (MCC) (*ecctmcc*), and entropy of offline credit card transactions with respect to merchants (*efctmer*) and MCC (*efctmcc*).

2.3 Labeling

There exists numerous definitions of customer churn in the literature, and most of the definitions represent very specific aspect of the customers' activity such as whether they used their credit card for a specific period of time [37], whether they made a call during an observation window [15], as a fuzzy concept description [14], or some other metrics based on sliding window methodology [12]. However, the decision of whether a customer churned or not is quite subjective and is usually defined based on heuristic rules set by the industry officials. We followed the suggestions of the bank officials, and developed a set of churn decision models based on the segmentation information, credit card expenditure patterns, and checking account activities.

The bank donated monthly segmentation information of each customer for a 23-month period, which also covers the time window for the transaction data provided. Monthly customer segmentation is generated by the bank in order to facilitate productive and convenient management of business processes such as advertising and churn prevention. Being one of the more than a dozen of such segments, the segment *inactive* is applied to the months of a customer for which he/she owns no bank products or utilizes his/her products under some predefined aggregated activity level.

By adopting the bank's approach, we developed a set of churn definitions and corresponding labels (*churner* and *non-churner*) for each customer based on a set of rules pertaining to the order of the *inactive* months of each customer. For example, a customer who was tagged as *inactive* for all of the months in labeling window was considered as churner based on the churn definition that we named as *inac-full*. For the rest of

such segmentation-based (SB) labels, please refer to Appendix 1 in Additional file 1. The segmentation-based labels were generated for all the data sets listed in Table 1.

Unlike for data sets A1 and B1, in addition to segmentation information, we were also provided credit card transaction and checking account balance data for the labeling windows of data sets A2 and B2. This enabled us to develop several additional churn definitions based on credit card and checking account usage patterns forming the label sets *credit card usage-* and *checking account usage-based* labels (CC and CA in Table 1).

The results reported in this study are based on the churn definition *inac-full*. Other segmentation-based definitions along with the credit card- and checking account usage-based churn definitions were also considered in the context of the study. However, we find that the results do not significantly change as can be seen in the results listed in Appendix 1 (see Additional file 1).

2.4 Experimental settings

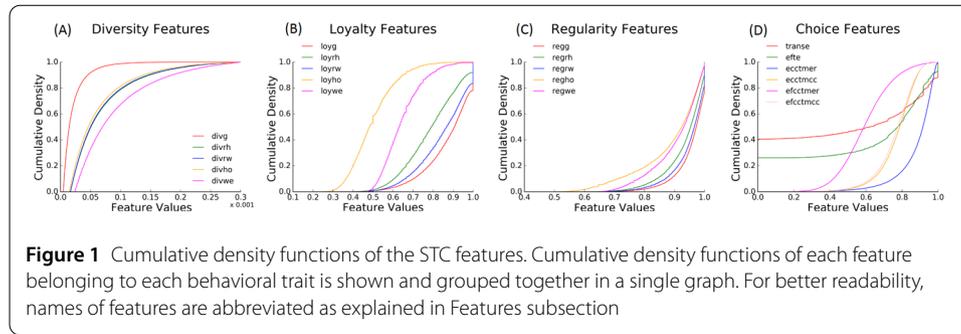
In order to evaluate the performance of STC behavioral features, we performed analyses with four different data sets as discussed in the previous section (Table 1). The data sets B1 and B2 are biased more towards higher credit card usage whereas data sets A2 and B2 are generated for shorter term prediction. The same methodology independently applies to each of these data sets. Prior to the generation of predictive models, the missing values for each feature are assigned the mean values, and all numerical features are standardized by removing the mean and scaling the values to unit variance. Dummy encoding has been applied on the categorical variables so that each of them is represented with as many binary variables as one less of the number of their levels.

We adopted Random Forests [38] as the classification training technique for our study. We trained our classification models with 500 trees and maximum two features per tree. We evaluated our models with stratified 8-fold cross-validation so that, in each iteration, almost the same proportions of churners and non-churners were involved in the evaluation process. In order to estimate the stability of results due to random splitting of samples into training and testing sets, bootstrap simulation approaches could also be adopted as suggested by Tang et al. [28]. It should be noted, however, that cross validation can be considered as a special case of the bootstrapping approach. To our knowledge, these two approaches are chosen based on the trade-off between statistical rigidity and computational cost. In our study, the data sets was so large that we decided to use 8-fold cross validation for evaluation.

In order to mitigate the risk of imbalance between the number of churners and non-churners in our data sets (Table 1), we applied SVM-SMOTE [39] with the ratio of 0.25, meaning that the minority class is oversampled until its cardinality reached to a quarter of that of the majority class. We also conducted our analysis with the regular [40] and borderline SMOTE [41]; however, our reported results did not significantly change. All the pre-processing and classification implementations were done in Python language mainly with Scikit-learn package [42], and for SVM-SMOTE, the open source package Imbalanced-learn by Lemaître et al. [43] was adopted.

3 Results

We analyzed two anonymous credit card transaction samplings collected over a one-year period by a major financial institution. We treated each sampling separately to gener-

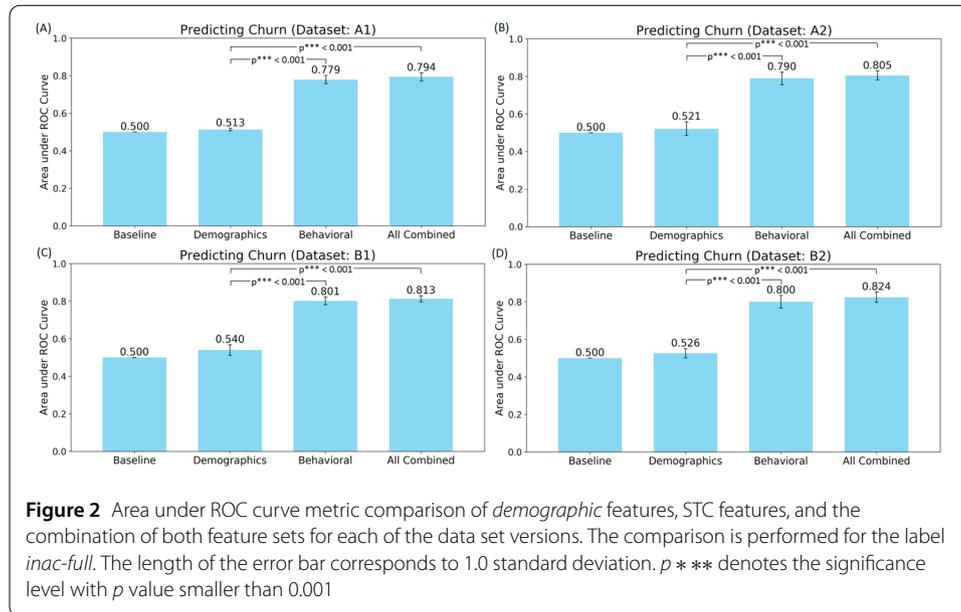


ate four data sets by applying different observation and labeling windows. Around 40–50 thousand customers along with 1.9 to 3.3 million transactions have been considered. The results show that the spatio-temporal patterns and entropy of choice are significantly related with customers' churn decisions. The results remain significant even for 11 different definitions of churn, and various versions of the data sets (e.g., using just weekend transaction data).

Figure 1 illustrates the cumulative density functions (CDF) of the diversity, loyalty, regularity, and choice features of the customers for data set A1. The CDFs of other data sets have similar characteristics as can be seen in Fig. S4 (see in Additional file 4).

The noticeable steepness of diversity CDFs (Fig. 1A) is due to the high normalization denominators that are based on the number of all bins rather than the non-empty ones. It shows that customers were a lot more diverse about where they shop based on grid-based characterization. Figure 1B shows that the customers' three most preferred locations account for a large population (~75–85%) of all their shopping. While less than 65% of the purchases were made on the preferred days, around 50% of the transactions were made during the top three time slots. Figure 1C shows that customers have similar regularity patterns in terms of both location and time. More than around 80% of the customers seem to have high regularity scores meaning that, both temporally and spatially, they presented similar behaviors in shorter (3–4 months) and longer (9–12 months) terms. The CDFs of entropies of choice are illustrated in Fig. 1D. Almost identical curves of *ectmcc* and *efctmcc* imply that the customers distributed their transactions into different shopping categories very similarly regardless of whether they made their purchases online or in person. More than 80% of the customers distributed their purchases to different merchants with high entropy value as high as 0.8 (*ectmer*), whereas they showed relatively more deterministic pattern in their offline purchases (*efctmer*). The money and EFT transactions were not as frequent as credit card transactions in the data sets leading to 0 entropy for around 25% and 40% of customers having very low number of transactions.

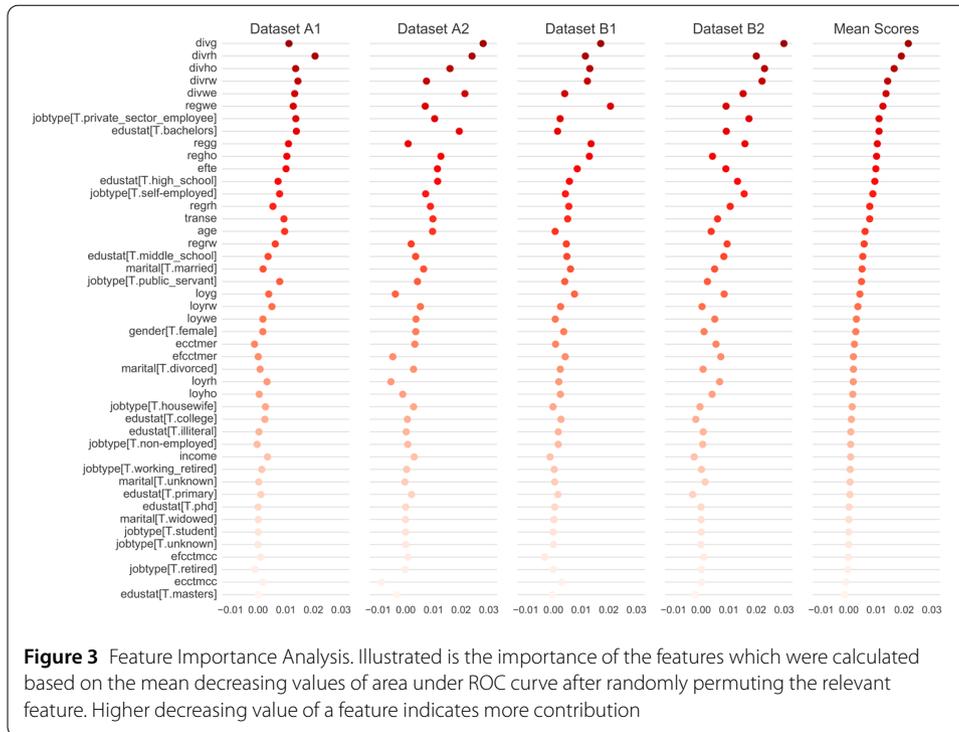
Prediction performances of demographic, STC behavioral features and their combinations are compared in Fig. 2. Due to high level of imbalance of labels (churners vs. non-churners) in our data sets, we adopted the area under ROC curve metric (AUROC score hereinafter) for the evaluation of our models as suggested by [44, 45]. However, it should be noted that usage of AUROC as an evaluation metric could lead to selection of suboptimal prediction model due to the unrealistic assumptions made by AUROC models about the misclassification costs. As suggested by Verbraken et al. [21], profit-driven evaluation approaches such as maximum profit (MP) should be adopted in order to determine the profitable part of the churning customer base.



For all data sets employed in the study, STC behavioral features were significantly better than the demographic features in terms of area under the ROC curve metric. The results were very similar for the other 10 different definitions of the churn. (Please see Fig. S2 in Additional file 2, Fig. S3 in Additional file 3, and Appendix 1 in Additional file 1 for the accuracy scores and significance test results.) The combination of all models were slightly better than STC behavioral features for each of the data set; however, this superiority was not significant. Similarly, demographic model was better than the baseline model, which is merely random guessing, without any significance. For example, for data set A1 (not biased towards credit card users and usage of longer observation window), the behavioral model predicting the churners reached to AUROC score of 77.9% as compared to demographic model which obtained 51.3%, and the baseline model of 50.0%. Hence, the behavioral model performed 55.8% better than the baseline, and 51.9% better than the demographic model for churn prediction in the data set A1. For the data set A2 (not biased towards credit card users and usage of shorter observation window), similar results with slightly higher AUROC scores were observed: STC behavioral features were 58.0% and 51.6% better than baseline and demographic models, respectively.

For data sets B1 and B2 (biased towards credit card users), the AUROC scores were higher around 1–2% compared to those of data sets A1 and A2, as expected due to the possible high resolution of the behavioral models that were generated with higher number of transactions. For all data sets, the AUROC score of the combined model was 1.3–2.4% higher than the behavioral models showing that the demographic features may have a positive effect on the predictive power of the STC behavioral features. The AUROC scores of the models did not significantly change when the same analyses were conducted with only weekend transaction data.

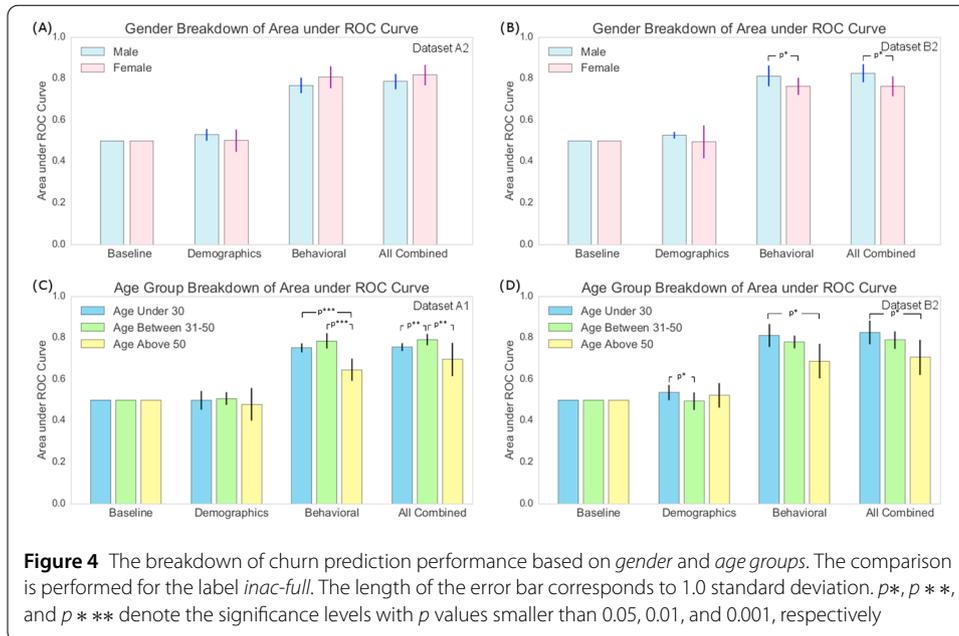
To understand the individual contributions of the features to the prediction power of the combined model (demographic and STC behavioral features), we applied mean decrease accuracy analysis for each of the features in each of the data sets. The importance of the features are calculated as the decrease in the AUROC score of the combined model after having permuted the values of each feature once at a time. In other words, for each



iteration of the cross validation, the combined model has been trained and tested with the current bag, and the AUROC score A_g for the current bag is saved. Then, for each feature f_i of the combined model, the column c_i of the test portion of the current bag is randomly permuted, and the AUROC score A_i for the modified test data set is re-calculated. Importance I_i for feature f_i is calculated as $(A_g - A_i)/A_g$.

It should be noted that STC features as the independent variables of the classification model are not independent from each other; in contrast, they are mutually related. Moreover, the machine learning technique that we employed in our analysis (i.e., random forest) evaluates various combinations of these features. Due to these reasons, we adopted mean-decrease AUROC as a common feature importance calculation technique used for random forest models.

The importance values calculated for each feature in each of the data sets and data set-wise mean of the scores are plotted in Fig. 3. For each feature, we report the mean importance value calculated over 8 random permutation of the feature values. Sorted by the importance value, the plot shows that the importance values vary roughly in the range of -0.01 and 0.03 , and the importance of the features follows similar patterns in each of the data sets meaning that high-scored features in the mean score column tend to be high in other data sets as well and vice versa. Clearly, all of the diversity features are more important than the rest of the features, and except for a few demographic features in between, regularity and loyalty features are following. The entropy of choice features are distributed to the different positions along the ordered list. It is notable that the educational statuses *bachelors*, *high*, and *middle school* were found to be important in this respective order, implying that as the customers' education level increase their churn decision might be more easily predicted. However, this inference is valid only when the customers with educational



statuses *illiteral*, *college*, and *masters*, comprising about 12–13% of all the customers, are not considered.

To identify the effect of gender and age on the predictability of the churners, we prepared separate models for age (under 30, between 30 and 50, above 50) and gender (males and females) groups. In doing so, we divided the data into subgroups each of which have the data for the particular group members (e.g. males), and built the prediction models as described previously. The plot of the evaluation of the age group models based on data sets A1 and B2, and the gender group models based on data sets A2 and B2 are shown in Fig. 4. Please see Fig. S5 in Additional file 5 and Fig. S6 in Additional file 6 for the results generated with other data sets.

As shown in Fig. 4A, for the unbiased data set, male and female customer groups have no significant superiority over each other in terms of predictability. On the other hand, for the data set B2 male customers seem to be more easily predicted with both behavioral ($t(7) = 2.19, p < 0.05$) and combined models ($t(7) = 2.74, p < 0.05$) as shown in Fig. 4B. For the age group analysis, we observe that the elderly customers (above age 50) seem to be significantly more difficult to predict (all p 's < 0.05 , except for data set A2). In particular, this finding is more significant in the analysis with behavioral models based on data set A1, as shown in Fig. 4C.

In order to observe the diversity and loyalty trends of churners and non-churners, we plotted aggregated diversity and loyalty feature values with respect to 3-month periods as shown in Fig. S7 of Additional file 7. The churners seem to have increasing loyalty and decreasing diversity trend towards the time they decide to churn whereas the non-churners have flat trends for the same feature sets. This finding might lead to an intuition such that the descending diversity trend of churners could be explained with the general financial activity decrease of the churners.

In order to differentiate the effects of diversity decrease from the diminishing financial activity, we carried out a study on the comparison of the financial activity trends of churners and non-churners. To this end, we aggregated overall financial activity of each

customer into monthly bins and applied a linear fitting to his/her monthly financial activity levels (in terms of both number of transactions and spending amount). We then considered the slope of the fitted line as a feature for predicting the churning decision of each customer. The slope values for churners and non-churners are shown in Fig. S8 of Additional file 8. Based on the prediction tests that we performed with a single-node Decision Tree classifier trained with financial activity trends of the customers, we found out that the predictive performance of financial activity trends in terms of expenditure amounts was similar to that of demographic features (0.52 AUROC score), and it was statistically significantly worse than the performance of STC behavioral features in general ($t(7) = 28.02$, $p^{***} < 0.001$). The performance based on activity trends in terms of transaction counts was relatively better than the demographic features (0.65 AUROC score); however, it was also significantly worse than the performance of the STC behavioral features ($t(7) = 13.73$, $p^{***} < 0.001$). From the results, we confirm that STC behavioral features capture the signal of customer's churning behavior more precisely than a simple feature of inactivation trend of credit card usage.

4 Discussion

In this article, we show that churn decision can be predicted to a large extent by analyzing dynamic behavioral patterns that the customer follows in both temporal and spatial domains, and the entropy of the choices she makes while performing financial activities. This result not only solidifies previous results in the literature about the relationship between spatio-temporal mobility patterns and individual financial well-being [2], but also serves as a first step towards effective modeling of churning behavior using large-scale financial transaction data.

The diversity and regularity features seem to have a systematic and large effect on the prediction performance. As shown in Fig. S7 of Additional file 7, the diversity scores of churners have a decreasing trend while this is not the case for non-churners. This seems, in one sense, analogous to Selye's characterization of the response to stress of high-level organisms, according to which the body of the organism gets exhausted and vulnerable at the later stages of the persistent stress [36]. Similarly, actual reasons behind the customers' churn decision might affect their *energy* to explore new products or determination for staying as the bank's customer. The decreasing diversity trend towards the churn moment of the customers shown in Fig. S7 of Additional file 7 seems to provide empirical evidence for this statement, although further research needs to be conducted in order to claim the existence of such relationship.

Our results also suggest that younger people seem to be relatively more predictable compared to elderly people as far as the behavioral patterns are concerned (Fig. 4), an observation consistent across several data sets that are considered. This difference in prediction performance for different age groups is consistent with the relatively high ranking of the feature *age* as shown in Fig. 3. Compared to age, however, we find out that there seems no significant difference between the predictive power of behavioral features for male and female customers except for the weak significance signals in the analyses made with data sets B1 and B2 (Fig. S6 in Additional file 6). We also performed the analyses for various data sets generated with different churn definitions (based on credit card and checking account usage), different time frames (with only weekend data and customer-specific observation window), and different parameters for STC behavioral feature extraction (e.g.,

edge size of grid bins and radii of annular areas for radial bins); however, the reported results did not significantly change.

It might be suggested that the usage of online and offline credit card transactions should be evaluated differently in the choice behavioral trait. Nevertheless, as can be seen in the importance ranking of the features, merchant-wise and merchant type-wise entropy (*ecctmer* and *ecctmcc*) and their offline variants (*efctmer* and *efctmcc*) are ranked close to each other (Fig. 3). This implies that the distinction between online and offline transactions for the entropy of choice does not seem to be important in the present data set.

Understanding reasons behind churning activities is extremely important for financial institutions to accurately deliver more engaging and rewarding experience for customers. However, finding causal relations between certain factors and churning decisions could be a very difficult task, as churning might be due to a wide range of personal circumstances such as job loss or unsatisfying customer service. It is therefore worth noting that the results presented in this paper reveal statistical correlations between spatio-temporal and choice patterns and churning activities, and do not support a causal relation between the two. More research needs to be done, potentially by combining data-driven approaches presented in this paper and traditional methodologies based on surveys and questionnaires, to fully understand why and when people decide to churn away from banking products and services. However, our results at least suggest that, even without taking into account the actual financial activities of the customers, such as monthly spendings or savings, greater accuracy in churn prediction may still be achieved merely based on mobility, temporal, and choice entropy patterns that could be extracted from customers' behavior data.

It is interesting to compare the spatio-temporal features adopted in this paper with traditional features used for churn prediction, such as aggregated statistics on customer activities (e.g., number of phone calls and monthly billing/subscription) and marketing related variables (e.g., interactions with operators) [27]. They also bear certain similarity with the recency, frequency and monetary (RFM) features, which are commonly used to evaluate customer values [46], and have also been adopted for predicting customer churn behaviors [47]. In this paper, we do not consider using the RFM features since they are primarily used for assessing customers' loyalty by directly evaluating the transaction-based information. Instead, unlike transaction-based features or aggregated usage statistics, our main goal is to identify domain-independent signals, particularly those based on the spatio-temporal distributions of human activities, which are correlated with customer churn behaviors. However, we would like to emphasize that the proposed features complement, rather than replace, the traditional features, and a combination of the two may lead to even greater prediction accuracy and is therefore worth further studies.

Spatial and temporal signatures of customers' mobility patterns employed in the present paper and [2] depend only on the data that have references to time and space, thus making them independent of the particular application domains. We therefore expect that the generality of these features enables the applicability of our churn prediction methodology in other business domains or industries, such as telecommunication or insurance services. As future work, the proposed STC behavioral features might be combined with other advanced characteristics such as efficiency on social media [48] as a collective churn decision behavior. Furthermore, impact of behavioral properties on early detection of possible churn could be of importance as such functionality would allow for customer retainment

action. On the other hand, for our methodology to work efficiently and effectively, especially from a streaming data point of view, organizations must employ mechanisms and technological frameworks to streamline the applicability of our approach on a daily basis, given the constraints on data availability, data quality and data confidentiality.

Our study has several limitations. The data sets of credit card transaction records used in this study are based on samples of the full customer set of the financial institution. Therefore, sampling bias could exist and potentially influence the results. Another limitation of using credit card transaction data is that, credit card holders may only represent a certain fraction of the population, and customers may choose to pay by cash under certain circumstances. However, our data sets do cover a relatively large period in time, which makes our results robust against external factors that might influence customers' financial activities such as seasonality and economic instability.

There is a growing research community working on utilizing quantitative data for understanding human and social behavior, with several dedicated academic venues [49, 50]. However, many works have focused on problems with behaviors that are correlated with poverty and changes in behavior as a signal for future financial problems. This may be in part because it is still rare to have both financial information and detailed mobility information, and in part because churn is more a commercial problem than a social problem. Consequently, the current study is unique within the computational social science literature. It is also significant because, to our knowledge, it is the first study to find spatio-temporal and choice behavior signals that predict significant changes in a daily habitual behavior. It may be, for instance, that these or other related behavioral signals can also predict changes in other daily habitual behaviors such as shopping, dining, or work-related patterns.

Additional material

Additional file 1: Appendix 1. *Performance Comparison of Models Trained with Various 1 Churn Definitions*. Various churn definition variants are listed and performance summaries of models built with these churn definitions are provided. The significance test results of the pair-wise comparisons of feature sets as well as a detailed list of parameters applied in the prediction models are included. (PDF 113 kB)

Additional file 2: Figure S2. *Comparison of prediction scores of feature sets*. Area under ROC curve metric comparison of *demographic* features, *STC* features, and the combination of both feature sets for each of the data set versions. The length of the error bar corresponds to 1.0 standard deviation. The comparison is performed for the label *inac-12m*. $p * **$ denotes the significance level with p value smaller than 0.001. (TIF 422 kB)

Additional file 3: Figure S3. *Comparison of prediction scores of feature sets*. Area under ROC curve metric comparison of *demographic* features, *STC* features, and the combination of both feature sets for each of the data set versions. The length of the error bar corresponds to 1.0 standard deviation. The comparison is performed for the label *cc-inac13m*. $p * **$ denotes the significance level with p value smaller than 0.001. (TIF 234 kB)

Additional file 4: Figure S4. *Cumulative density functions for data sets A2, B1, and B2*. In general, distributions of the features are similar across all the data sets. (TIF 3.6 MB)

Additional file 5: Figure S5. *Age group breakdown of feature set performance for data sets A2, B1, and B2*. Area under ROC curve metric comparison of *demographic* features, *STC* features, and the combination of both feature sets for the portions of data sets generated based on age groups under 30, between 30–50, and above 50. The length of the error bar corresponds to 1.0 standard deviation. The comparison is performed for the label *inac-full*. $p *$, $p * *$, and $p * **$ denote the significance levels with p values smaller than 0.05, 0.01, and 0.001, respectively. (PNG 841 kB)

Additional file 6: Figure S6. *Gender breakdown of feature set performance for data sets A2, B1, and B2*. Area under ROC curve metric comparison of *demographic* features, *STC* features, and the combination of both feature sets for the portions of data sets generated based on gender. The length of the error bar corresponds to 1.0 standard deviation. The comparison is performed for the label *inac-full*. $p *$ denotes the significance level with p value smaller than 0.05. (PNG 919 kB)

Additional file 7: Figure S7. *Diversity and loyalty trends of churners and non-churners for data set A1*. For the observation window, the diversity and loyalty feature values of churners have decreasing and increasing trend, respectively, whereas this is not the case for non-churners. (PNG 600 kB)

Additional file 8: Figure S8. *Monthly financial activity trend plots of churners and non-churners*. For all the data sets, churners and non-churners seem to have both positive and negative financial activity slopes with an exception for some of the churners with negative slopes. (PNG 1.9 MB)

Acknowledgements

The authors would like to thank Dr. Attila Bayrak of Akbank, Turkey, for the business know-how, ideas and feedback during the study design and implementation.

Funding

Erdem Kaya is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) BIDEB 2214A fellowship program.

Abbreviations

OECD, Organization for Economic Co-operation and Development; EFT, Electronic Fund Transfer; POS, Point of Sale; MCC, Merchant Category Code; SMOTE, Synthetic Minority Over-sampling Technique; CDF, Cumulative Density Function; ROC, Receiver Operating Characteristic; AUROC, Area Under Receiver Operating Characteristic; MP, Maximum Profit; RFM, Recency, Frequency, and Monetary.

Availability of data and materials

All relevant data (feature and label sets) will be included in the paper's additional files upon acceptance.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the experiments: EK XD YS BB AP. Analyzed the data: EK XD YS. Contributed reagents/materials/analysis tools: EK SB BB AP. Wrote the paper: EK XD YS BB. All authors read and approved the final manuscript.

Authors' information

Erdem Kaya, Ph.D. Erdem Kaya has received M.Sc. degree in Computer Science from University of Wisconsin-Madison in 2012, and Ph.D. degree in Computer Science from Sabanci University in 2017. He was honored with IS Academic Achievement Award in 2012 by International Student Services of the UW-Madison. He conducted behavioral analytics research as a visiting student with Human Dynamics group of MIT Media Lab under supervision of Prof. Alex "Sandy" Pentland. He is interested in behavioral analytics, information visualization, visual analytics, human-computer interaction, and collaborative analytic systems. *Xiaowen Dong, Ph.D.* Xiaowen Dong is a Departmental Lecturer in the Department of Engineering Science, University of Oxford and a Research Affiliate at the Media Lab, Massachusetts Institute of Technology. His research focuses on emerging signal processing and machine learning techniques on graphs, and their applications in understanding human behavior, decision making and societal changes. *Yoshihiko Suhara, Ph.D.* Yoshihiko Suhara is a Research Scientist at Megagon Labs and a Visiting Scientist at the MIT Media Lab. His research interests include machine learning, natural language processing, and computational social science. *Selim Balcisoy, Assoc. Prof. Assoc. Prof.* Dr. Selim Balcisoy received his PhD on Computer Science in 2001 from Swiss Federal Institute of Technology, Lausanne (EPFL). Between 2001 and 2004 he was Senior Research Engineer at Nokia Research Center USA, where he conducted research on mobile graphics. Since 2004 Dr. Balcisoy is a Faculty member at Sabanci University and is one of the Directors of Behavioral Analytics and Visualization Lab, a joint Lab with MIT Media Lab. His research interests include Data Science, Visual Analytics, Augmented Reality and Virtual Environments. Dr. Balcisoy (co)authored over 50 publications at refereed international journals and conferences, and has been granted with one U.S. patent. *Burcin Bozkaya, Prof.* Burcin Bozkaya earned his Ph.D. in Management Science at the University of Alberta, Canada and then joined ESRI, Inc., a California based software company specializing in geographic information systems and location analytics. In 2004, Dr. Bozkaya joined Sabanci University School of Management as a full-time faculty member. He has conducted research and published in the field of Operations Research using various analytical techniques on location analysis, location-based services, decision support systems, logistics system design and route optimization. He has also completed numerous industry projects in these fields. Recipient of many international awards, Dr. Bozkaya's current research fields are business and big data analytics, and their applications. He is also a cofounder of VisioThink, Inc., established in 2006. *Alex "Sandy" Pentland, Prof.* Professor Alex "Sandy" Pentland directs the MIT Connection Science and Human Dynamics labs and previously helped create and direct the MIT Media Lab and the Media Lab Asia in India. He is one of the most-cited scientists in the world, and Forbes recently declared him one of the "7 most powerful data scientists in the world" along with Google founders and the Chief Technical Officer of the United States. He has received numerous awards and prizes such as the McKinsey Award from Harvard Business Review, the 40th Anniversary of the Internet from DARPA, and the Brandeis Award for work in privacy.

Author details

¹Behavioral Analytics and Visualization Lab, Sabanci University, Istanbul, Turkey. ²Media Lab, Massachusetts Institute of Technology, Cambridge, USA. ³Department of Engineering Science, University of Oxford, Oxford, United Kingdom. ⁴Megagon Labs, Mountain View, USA.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 16 March 2018 Accepted: 7 October 2018 Published online: 19 October 2018

References

1. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M et al (2009) Computational social science. *Science* 323(5915):721

2. Singh VK, Bozkaya B, Pentland A (2015) Money walks: implicit mobility behavior and financial well-being. *PLoS ONE* 10(8):e0136628
3. Pentland A, Heibeck T (2010) *Honest signals: how they shape our world*. MIT Press, Cambridge
4. Reinchheld FF (1996) The loyalty effect: the hidden force behind growth, profits, and lasting value. *Long Range Plan* 6(29):909
5. The standout customer loyalty stats of 2017. <http://www.socialannex.com/blog/2017/01/10/standout-customer-loyalty-stats-2017/>
6. Customer acquisition vs. retention costs—statistics and trends. <https://www.invespcro.com/blog/customer-acquisition-retention/>
7. Accenture global consumer pulse survey. <https://www.accenture.com/us-en/insight-digital-disconnect-customer-engagement>
8. Nath SV, Behara RS (2003) Customer churn analysis in the wireless industry: a data mining approach. *Proceedings-Annual Meeting of the Decision Sciences Institute*
9. Patil NV, Dixit AM (2014) Survey on profit maximizing metric for measuring classification performance of customer churn prediction models. *Int J* 4(12)
10. Huang B, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. *Expert Syst Appl* 39(1):1414–1425
11. Lu N, Lin H, Lu J, Zhang G (2014) A customer churn prediction model in telecom industry using boosting. *IEEE Trans Ind Inform* 10(2):1659–1665
12. Huang Y, Zhu F, Yuan M, Deng K, Li Y, Ni B, Dai W, Yang Q, Zeng J (2015) Telco churn prediction with big data. In: *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. ACM, New York, pp 607–618
13. Liu Y, Zhuang Y (2015) Research model of churn prediction based on customer segmentation and misclassification cost in the context of big data. *J Comput Commun* 3:87–93
14. Bi W, Cai M, Liu M, Li G (2016) A big data clustering algorithm for mitigating the risk of customer churn. *IEEE Trans Ind Inform* 12(3):1270–1281
15. Wangperawong A, Brun C, Laudy O, Pavasuthipaisit R (2016) Churn analysis using deep convolutional neural networks and autoencoders. *arXiv preprint arXiv:1604.05377*
16. Xie Y, Li X, Ngai EWT, Ying W (2009) Customer churn prediction using improved balanced random forests. *Expert Syst Appl* 36(3):5445–5449
17. Larivière B, Van den Poel D (2005) Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst Appl* 29(2):472–484
18. Coussement K, Van den Poel D (2008) Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. *Expert Syst Appl* 34(1):313–327
19. Castro EG, Tsuzuki MS (2015) Churn prediction in online games using players' login records: a frequency analysis approach. *IEEE Trans Comput Intell AI Games* 7(3):255–265
20. Buckinx W, Van den Poel D (2005) Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *Eur J Oper Res* 164(1):252–268
21. Verbraken T, Verbeke W, Baesens B (2013) A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Trans Knowl Data Eng* 25(5):961–973
22. Au W-H, Chan KC, Yao X (2003) A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Trans Evol Comput* 7(6):532–545
23. Hung S-Y, Yen DC, Wang H-Y (2006) Applying data mining to telecom churn management. *Expert Syst Appl* 31(3):515–524
24. Chen CP, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inf Sci* 275:314–347
25. Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: measuring and understanding the predictive accuracy of customer churn models. *J Mark Res* 43(2):204–211
26. Verbraken T, Verbeke W, Baesens B (2014) Profit optimizing customer churn prediction with Bayesian network classifiers. *Intell Data Anal* 18(1):3–24
27. Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. *Eur J Oper Res* 218(1):211–229
28. Tang L, Thomas L, Fletcher M, Pan J, Marshall A (2014) Assessing the impact of derived behavior information on customer attrition in the financial service industry. *Eur J Oper Res* 236(2):624–633
29. Miller HJ, Han J (2009) *Geographic data mining and knowledge discovery*. CRC Press, Boca Raton
30. Mennis J, Guo D (2009) Spatial data mining and geographic knowledge discovery—an introduction. *Comput Environ Urban Syst* 33(6):403–408
31. Fernandes GB, Artes R (2016) Spatial dependence in credit risk and its improvement in credit scoring. *Eur J Oper Res* 249(2):517–524
32. Agarwal S, Ambrose BW, Chomsisengphet S, Sanders AB (2012) Thy neighbor's mortgage: does living in a subprime neighborhood affect one's probability of default? *Real Estate Econ* 40(1):1–22
33. Bourassa S, Cantoni E, Hoesli M (2010) Predicting house prices with spatial dependence: a comparison of alternative methods. *J Real Estate Res* 32(2):139–159
34. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science* 328(5981):1029–1031
35. Pan W, Aharony N, Pentland A (2011) Fortune monitor or fortune teller: understanding the connection between interaction patterns and financial status. In: *2011 IEEE third international conference on privacy, security, risk and trust (PASSAT) and 2011 IEEE third international conference on social computing (SocialCom)*. IEEE Press, New York, pp 200–207.
36. Selye H (1946) The general adaptation syndrome and the diseases of adaptation. *J Clin Endocrinol Metab* 6(2):117–230
37. Prasad UD, Madhavi S (2012) Prediction of churn behavior of bank customers using data mining tools. *Bus Intell J* 5(1):96–101

38. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
39. Tang Y, Zhang Y-Q, Chawla NV, Krasser S (2009) SVMs modeling for highly imbalanced classification. *IEEE Trans Syst Man Cybern, Part B, Cybern* 39(1):281–288
40. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
41. Han H, Wang W-Y, Mao B-H (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: *International conference on intelligent computing*. Springer, Berlin, pp 878–887
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
43. Lemaitre G, Nogueira F, Aridas CK (2016) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *CoRR abs/1609.06570* <http://arxiv.org/abs/1609.06570>
44. Burez J, Van den Poel D (2009) Handling class imbalance in customer churn prediction. *Expert Syst Appl* 36(3):4626–4636
45. Chawla NV (2005) Data mining for imbalanced datasets: an overview. In: *Data mining and knowledge discovery handbook*. Springer, Berlin, pp 853–867
46. Fader PS, Hardie BGS, Lee KL (2005) RFM and CLV: using iso-value curves for customer base analysis. *J Mark Res* 42(4):415–430
47. Coussement K, Van den Poel D (2009) Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Syst Appl* 36(3):6127–6134
48. Morales AJ, Borondo J, Losada JC, Benito RM (2014) Efficiency of human activity on information spreading on Twitter. *Soc Netw* 39:1–11
49. IC2S2: international conference on computational social science. <https://ic2s2.org/>
50. NetMob: the conference on the scientific analysis of mobile phone datasets. <http://netmob.org/>

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com
