Decision Support Systems 116 (2019) 1–12

Contents lists available at ScienceDirect

Decision Support Systems

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

The turf is always greener: Predicting decommitments in college football recruiting using Twitter data

Kristina Gavin Bigsby*, Jeffrey W. Ohlmann, Kang Zhao

Department of Management Sciences, University of Iowa, Iowa City, IA 52242, United States of America

ARTICLE INFO

Keywords:
Social networks
Twitter
Social media
Predictive analytics
Turnover
College football

ABSTRACT

We utilize the wealth of data related to American college football recruitment as a laboratory for studying the impact of social networks on organizational turnover. We combine data about athletes' recruiting activities and college choices with data from their Twitter social networks to predict decommitments over time - specifically, which athletes will decommit from their current college in a given month. Our results demonstrate the value of considering online social networks for decommitment predictions. Models incorporating social media data consistently outperform the baseline model containing only features derived from recruiting and institutional data. In the realm of athletic recruiting, our research can help coaches identify recruits who are more likely to decommit and enable them to proactively adjust recruiting strategies.

1. Introduction

More than a game, college football is often a significant part of the public face of an institution of higher education. In 2017 alone, over 42 million fans attended Division I football games [1]. Successful athletic departments turn multimillion-dollar profits, and on-field performance has been linked to trends in general student body enrollment [42], institutional reputation [43], and donor behaviors [44]. The performance of a team is highly dependent on recruiting the most talented athletes. An analysis of college football data from 2002 to 2012 found that each five-star recruit increased the number of team wins by 0.306 [45], and the worth of a premium college football player has been estimated as high as $2.3 million per season [2]. Thus, it is no surprise that top programs spend upwards of $1 million annually on recruiting, with total football budgets increasing approximately 30% over the past several years [3].

Despite increasing quantities of money, time, and energy devoted to recruitment, media reports indicate that decommitments (instances where an athlete reneges on a non-binding, verbal commitment to a college) are on the rise [4]. Our research is the first to examine decommitments, and can assist coaches by identifying vulnerable commitments and informing recruiting and retention strategies. Specifically, we utilize data from college football recruits' Twitter networks to construct an explanatory model of decommitment decision-making and predict decommitments over time. Using Twitter data is especially appropriate for the college football context, where “nothing has impacted recruiting more in the last 20 years than social media” (as quoted in [5]). The power of online social network data to predict offline outcomes has been explored in the context of box office success [6], event attendance [7], and health outcomes [39], among others. Yet there is very little empirical research focusing on the relationship between athletes' online social media and offline recruiting decisions.

Our work also has more general implications about the relationship between online social networks and offline organizational turnover. Recruitment occurs in many contexts, and we draw comparisons between turnover in college football and human resources (HR). As with college athletics, HR recruiting and retention is a high-stakes issue; good employees add value to an organization through their work product, knowledge, and even their professional and personal networks. However, in a recent survey of 321 U.S. and Canadian companies, 35% reported difficulty retaining top employees (Towers [8]). At the same time, a Deloitte research team found that companies spent an average of approximately $4000 per hire in 2014, an increase of 7% on average over the previous year [9]. High rates of turnover may be interpreted as a signal of poor organizational culture, working conditions, or leadership, negatively affecting the reputation of the organization [10]. Bad reputation has been linked to reduced pride in membership and employee tenure [11], suggesting a feedback loop of adverse outcomes.

Abbreviations: NCAA, National Collegiate Athletic Association; NLI, National Letter of Intent; FBS, Football Bowl Subdivision; CART, Classification and Regression Tree; SVM, Support Vector Machine; ANN, Artificial Neural Network; AUC, Area Under Curve; ROC, Receiver Operating Characteristic

* Corresponding author.

E-mail addresses: kristina-gavin@uiowa.edu (K.G. Bigsby), jeffrey-ohlmann@uiowa.edu (J.W. Ohlmann), kang-zhao@uiowa.edu (K. Zhao).

https://doi.org/10.1016/j.dss.2018.10.003
Received 23 February 2018; Received in revised form 11 September 2018; Accepted 8 October 2018
Available online 09 October 2018

Our work makes a unique contribution to the body of knowledge on organizational turnover by leveraging Twitter data. Social media facilitates the collection of large quantities of network data in situations where observing offline social ties would be expensive or impossible. Our study provides a holistic view of an entire labor market, tracking the recruiting activities and Twitter networks of 2644 athletes in the class of 2016.

In the remainder of this paper we provide a review of the related literature. We then describe our data, features, explanatory model, and predictive performance. Finally, we present an example application to recruiting decision support and discuss the implications of our findings, limitations, and directions for future research.

2. Related work

Drawing on the parallels between HR and college football recruiting, we review selected literature on predictors of employee turnover and previous research on athletic recruiting.

2.1. Individual, organizational, and environmental predictors

An employee's decision to leave an organization is influenced by many factors, and one branch of turnover research focuses on the effects of individual, organizational, and environmental variables. While the evidence supporting the effect of demographic features is mixed, meta-analyses consistently show that age and tenure are negatively correlated to job turnover [12,13]. As our study focuses on high school seniors, we assume that age is constant. We do consider the length of time that an athlete has been committed, as well as individual characteristics that affect his recruitment prospects such as star rating.

Regarding organizational factors, an employee's turnover intention is strongly correlated to the desirability of leaving her current organization [14]. Indeed, job satisfaction is often called “the single most reliable predictor of turnover,” ([15], p. 208), and multiple meta-analyses [12,13] confirm this relationship. Similar to Cotton & Tuttle Cotton & Tuttle [16], we represent satisfaction via a collection of variables. We consider domain-specific cost/benefit factors impacting school desirability such as the college's amenities, geographic distance, academic ranking, and football team performance [17]. We also examine how demonstrated affinity between the athlete and school (via unofficial, official, and coach visits) may signal satisfaction.

Turnover is also linked to environmental conditions, including the job market and perceived ease of leaving the organization [14]. Meta-analyses demonstrate that variables related to specific job alternatives and comparisons to an employee's current position are superior predictors of turnover [12,13]. In our study, we consider features related to the availability of college options (number of scholarship offers) as well as their attractiveness in comparison to the athlete's current school.

2.2. Social network predictors

The second branch of turnover literature considers the effect of social and relational factors. In a five-year survival analysis of 176 healthcare workers, Mossholder et al. Mossholder et al. [18] find that network centrality and interpersonal citizenship behavior are significant predictors of turnover, after controlling for tenure, age, gender, and job satisfaction. Feeley, Hwang, and Barnett Feeley, Hwang, and Barnett [19] discover that subjects who report greater numbers of friendships in the workplace have a lower likelihood of turnover. While Moynihan and Pandey Moynihan and Pandey [15] do not track actual social ties, they survey 326 non-profit employees about perceived coworker support and obligation toward coworkers, finding that both are negatively correlated to short-term turnover intentions.

In the athletic domain, two studies use different social network features to predict commitments. Mirabile and Witte Mirabile and Witte [20] find that familial connections (father, brother, cousin, or uncle who played football at the same college) are significant predictors of school choice. Bigsby, Ohlmann, and Zhao Bigsby, Ohlmann, and Zhao [21] discover that incorporating social media features, including the number and affiliation of Twitter friends and followers and the hashtags posted by athletes, improves the accuracy of school choice predictions. Additionally, qualitative studies of sports recruiting [22] and surveys of high school and college athletes [23] underscore the influence of coaches, parents, and other athletes on school choice. In our study, we capture an athlete's social ties through his Twitter network and track the number of in-links and out-links to coaches, current players, and other recruits affiliated with the athlete's current school.

Our work also provides a unique addition to the literature by investigating the role of inter-organizational networks. We use Twitter data to observe each athlete's in-links and out-links to coaches, players, and recruits from other schools that have offered him a scholarship. With the exception of Moynihan and Pandey Moynihan and Pandey [15], previous research on employee turnover has largely ignored external networks. Their analysis of professional networking activities (e.g., conferences, professional society memberships) fails to find strong support for the hypothesis that inter-organizational networks influence turnover intention via perceived ease of movement.

We also investigate whether decommitments spread through the social network. In a two-year study of referral hiring at a telephone customer service center, Fernandez, Castilla, and Moore Fernandez, Castilla, and Moore [24] find that employees hired via referrals were more likely to leave the organization after their referrers, suggesting a diffusion effect in turnover behavior. Felps et al. Felps et al. [25] propose a theory of turnover contagion, and their results suggest that coworker job embeddedness (a summary measure combining person-job fit and network centrality) is a significant, negative predictor of turnover whose effect was mediated by coworkers' job search activities. Though there are no empirical studies of turnover contagion in college athletic recruiting, anecdotal evidence suggests that athletes may be more likely to decommit after their peers [26], and we expect that the commitment status of an athlete's reciprocated social ties will impact his likelihood of decommitment.

3. Data

We scraped data on 2644 high school football players in the 2016 recruiting class from the 247Sports.com recruiting database [27]. For each individual athlete, we collected personal information (e.g., height, weight, hometown, position) and timelines of recruiting events (e.g., scholarship offers, visits, commitments, decommitments). College coaches may begin directly contacting high school players during their junior year [28], with the recruiting process intensifying at the start of senior year. Different recruiting events tend to occur at specific times over the course of the recruitment, and Fig. 1 displays a timeline of scholarship offers, visits, commitments, and decommitments over the final six months of recruitment.

While the vast majority of offers are extended during junior year (78% of all offers for the class of 2016), there is a slight uptick during the final months of recruitment. At this time, many coaches scramble to fill vacancies in their recruiting classes by offering scholarships to athletes they did not consider before or attempting to “poach” recruits from other schools. Unofficial visits, which are paid for by the athlete and his family, decrease during this period while official visits increase. Recruits in the class of 2016 were not allowed to take official visits financed by the school until the start of their senior year [28]. Coach visits (including evaluations and in-home visits) occur throughout recruitment, but increase during the last six months, when coaches seek to maintain relationships with committed athletes and secure additional recruits. 38% of all coach visits occurred between September 2015 and February 2016. Commitments also increased during this period. 45% of athletes in the class of 2016 committed during the final six months of recruitment. An athlete may make a non-binding verbal

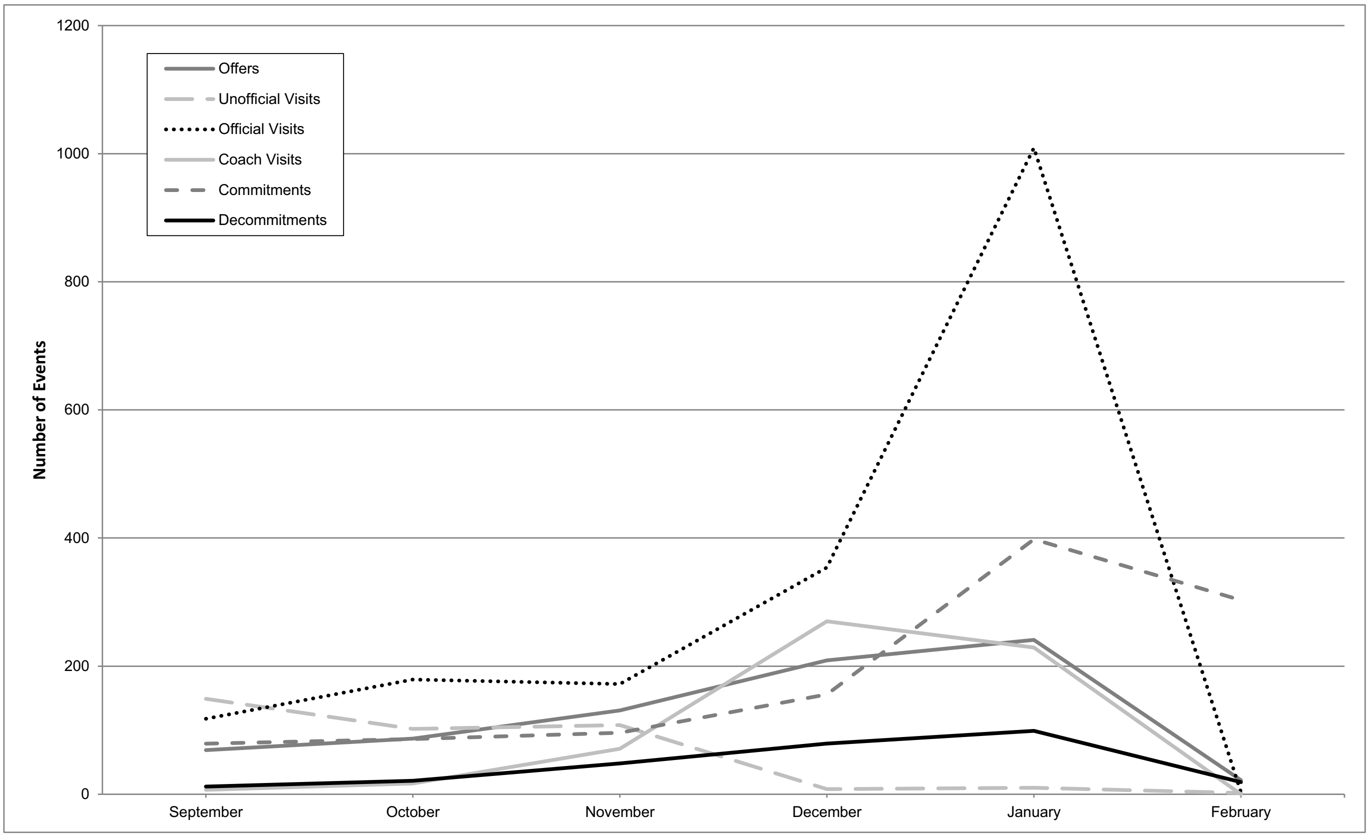

Fig. 1. Timeline of recruiting events (9/1/2015–2/29/2016).

commitment at any time any time, but can then revoke his pledge and commit to another school freely until signing a National Letter of Intent (NLI). Of the 2644 athletes in the class of 2016, 536 unique athletes (20.3%) decommitted at some point during recruitment. 34 athletes decommitted twice and 1 athlete decommitted 3 times, resulting in 572 total decommitment events. This rate is greater than the 12.4% observed previously [29], and supports claims by sports media outlets that decommitments are becoming more frequent [4]. Decommitments increased during the final months of recruitment; 71% of all decommitments occurred between September 2015 and February 2016.

We also obtained basic information about 682 Division I, II, III and junior college teams recruiting these athletes. This included dynamic features which changed from month to month, such as football team ranking [30], NCAA disciplinary record [1], and team performance, as well features that did not change over the final six months of recruitment, such as location and academic ranking [31].

In addition to being a comprehensive source of recruiting data, 247Sports.com provides links to the Twitter profiles of many athletes. We obtained Twitter IDs for 1629 athletes from the site, and performed a manual search for the remaining athletes to conclude that 2329 (88%) possessed public Twitter accounts, while an additional 160 (6%) had protected accounts. For athletes with public accounts, we gathered profile information, friends and followers lists, and tweets using the REST API [32]. Because the API does not return the date when two users became connected, we collected Twitter data monthly between September 2015 and March 2016 and compared the friends and followers lists retrieved over time in order to track changes in the online social network. Twitter IDs for 651 Division I coaches and 2225 current college athletes were also obtained from 247Sports.com in order to identify connections between recruits and college football programs.

The vast majority of previous turnover studies are either a) cross-sectional, collecting data on predictors and turnover intentions at a single point in time, or b) time-lagged, measuring predictors at one time and outcomes at a later time. In contrast, we construct a dataset of "athlete-month" instances. For each athlete with a public Twitter account, we create an instance for each month between October 2015 and February 2016. While our Twitter data spans September 2015 to March 2016, we begin with October so that at least one month of retrospective data is available. We only include instances where the athlete is verbally committed at the beginning of the specified month; assuming that he cannot decommit if he is not currently committed or has signed an NLI. This process yields a dataset of 7128 instances. Each instance has features related to the corresponding athlete's personal characteristics, satisfaction, alternatives, and Twitter network, as recorded up to the end of the specified month *m*, and a binary outcome tracking whether the athlete decommitted during the next month $m + 1$. Over the 7128 athlete-month instances, there are 1785 different athletes, and there are 370 athlete-month instances corresponding to decommitments.

4. Feature engineering

The performance of a predictive model depends on the features it considers. In order to isolate the value added by incorporating information from different dimensions of an athlete's online social network, we present four groups of features: one group of baseline features using recruiting and institutional data, and three groups of features based on Twitter data.

4.1. Baseline features

As with employee turnover, athletic decommitments are likely to be impacted by the individual, organizational, and environmental factors. Using data from each athlete's 247Sports.com recruiting profile, we construct 9 features related to the athlete's personal characteristics, including star rating, length of his current commitment, and whether he has previously decommitted (Table 1). We construct 17 features related

Table 1
Baseline features.

Type	Feature	Description
Personal	Height	Numeric; height in inches
	Weight	Numeric; weight in pounds
	Bmi	Numeric; body mass index
	Position	Categorical; recruit position
	Star	Categorical; recruit star rating (0, 2, 3, 4, 5)
	Region	Categorical; U.S. Census region of recruit hometown
	Hotbed	Binary; hometown located in recruiting "hotbed" state (CA, FL, or TX)
	Days committed	Numeric; days since verbal commitment
	Prior decommitment	Binary; athlete has previously decommitted
Satisfaction	Distance (original)	Numeric; distance between hometown and original school
	In-state (original)	Binary; hometown in same state as original school
	Type (original)	Categorical; institutional type (military, private, public)
	US News (original)	Numeric; U.S. News academic ranking
	Division (original)	Categorical; NCAA division (FBS, FCS, II, III, JUCO)
	Power (original)	Binary; original school member of "Power 5" conference
	AP (original)	Numeric; ranking in Associated Press poll
	Postseason (original)	Binary; played in postseason bowl during 2014 season
	Historic win% (original)	Numeric; winning percentage over past five seasons (2010–2014)
	Current win% (original)	Numeric; winning percentage during 2015 season
	Commits (original)	Numeric; recruits committed to original school
	Coach change (original)	Binary; head coach change during 2015 season
	Sanctions (original)	Binary; team under active NCAA sanction or probation
	First offer (original)	Binary; original school was first to offer recruit
	Unofficial (original)	Numeric; unofficial visits to original school
	Coach visit (original)	Numeric; coach visits from original school
	Official (original)	Binary; recruit has taken official visit to original school
Alternatives	Offers (other)	Numeric; offers from other schools
	Closer (other)	Numeric; offering schools closer than original school
	US News (other)	Numeric; offering schools with higher U.S. News ranking
	FBS (other)	Numeric; offering schools from FBS
	Power (other)	Numeric; offering schools from "Power 5" conference
	AP (other)	Numeric; offering schools with higher AP poll ranking
	Postseason (other)	Numeric; offering schools that played in bowl during 2014 season
	Historic win % (other)	Numeric; offering schools with greater winning percentage over last five seasons (2010–2014)
	Current win % (other)	Numeric; offering schools with greater winning percentage during 2015 season
	Coach change (other)	Numeric; offering schools with head coach change during 2015 season
	Sanctions (other)	Numeric; offering schools with team under active NCAA sanction or probation
	Offers since (other)	Numeric; offers from other schools since commitment
	Unofficial (other)	Numeric; unofficial visits to other schools
	Unofficial since (other)	Numeric; unofficial visits to other schools since commitment
	Coach visit (other)	Numeric; coach visits from other schools
	Coach visit since (other)	Numeric; coach visits from other schools since commitment
	Official (other)	Numeric; official visits to other schools
	Official since (other)	Numeric; official visits to other schools since commitment
	Days to NSD	Numeric; days remaining until National Signing Day

to satisfaction, and expect that features that increase the benefits and decrease the costs of attendance at the original commitment school (termed "original" in Table 1) will decrease the likelihood of decommitment. While environmental predictors in HR studies often represent broad trends, we create features that measure the availability and attractiveness of an athlete's concrete alternatives, i.e., the other colleges that have offered him a scholarship (termed "other" in Table 1). We construct 19 features that measure the relative desirability of these alternatives in comparison to the original commitment school (Table 1). We note that these features are chronologically consistent. For example, to predict if an athlete will decommit in January, we count only official visits to other schools that occurred before January 1.

4.2. Out-links

The second set of features is designed to capture the relationship between Twitter "friends," i.e., other users followed by the athlete, and the likelihood of decommitment. Number of friends varied widely by athlete-as of September 2015, recruits had between 0 and 3841 friends, with an average of 658. In addition, because Twitter users tend to accrue friends over time, a connection made in 2013 might not be indicative of commitment strength in 2015. Thus, we focus on tracking new out-links in the previous month. For example, if predicting whether an athlete will decommit in January, the lists of Twitter friends retrieved January 1 and December 1 are compared to determine the number of friends added or dropped in December. We construct six features tracking the number, affiliation, and type of the athlete's Twitter out-links (Table 2). Ties to coaches, recruits, and current players at the athlete's original commitment school are denoted as "original," and connections to individuals from other schools that have offered a scholarship to the athlete are termed "other."

4.3. In-links

The third group of features focuses on the in-links from other Twitter users, i.e., "followers." As of September 2015, recruits had between 0 and 14,284 followers, with an average of 1185. We track new in-links from coaches, current college football players, and other recruits during the previous month. Followers from the athlete's original commitment school are described as "original," and followers from other schools that have offered a scholarship are termed "other." We construct six features recording the type, number, and affiliation of Twitter followers (Table 2).

Table 2
Online social network features.

Type	Feature	Description
Out-links	Coach friends (original)	Numeric; coaches from original school followed by athlete during previous month
	Recruit friends (original)	Numeric; 2016 recruits from original school followed during previous month
	College friends (original)	Numeric; current players from original school followed during previous month
	Coach friends (other)	Numeric; coaches from other schools followed by athlete during previous month
	Recruit friends (other)	Numeric; 2016 recruits from other schools followed during previous month
	College friends (other)	Numeric; current players from other schools followed during previous month
In-links	Coach followers (original)	Numeric; coaches from original school following athlete during previous month
	Recruit followers (original)	Numeric; 2016 recruits from original school following during previous month
	College followers (original)	Numeric; current players from original school following during previous month
	Coach followers (other)	Numeric; coaches from other schools following athlete during previous month
	Recruit followers (other)	Numeric; 2016 recruits from other schools following during previous month
	College followers (other)	Numeric; current players from other schools following during previous month
Diffusion	Total committed	Numeric; reciprocated connections to committed 2016 recruits who have never decommitted
	Total decommited	Numeric; reciprocated connections to 2016 recruits that have previously decommitted

4.4. *Diffusion*

The fourth group of features examines the actions of athletes' social network neighbors. A large body of research in psychology and sociology has explored how social networks impact individuals' decisions (e.g., [33]), and coworker behavior has previously been linked to personnel turnover [24,25]. We construct two features related to the commitment status of the athlete's reciprocated Twitter connections (Table 2). Consistent with a threshold model of diffusion, we use the total number of committed and decommitted peers prior to the prediction month. For example, to predict if a player will decommit in January, we count reciprocated Twitter connections that have decommitted up to January 1.

5. Explanatory modeling

In order to evaluate the marginal effects of different predictors on the likelihood of decommitment, we construct a series of fitted logistic regression models using the full set of 7148 athlete-month instances. Though logistic regression is fairly robust to class imbalance given sufficient training data, it is sensitive to multicollinearity. We conduct feature selection using a lasso regression with L1 penalty (C = 0.1), parameters selected via grid search (see Appendix A). The lasso regression is solely used for feature selection. After manually removing predictors whose coefficients reduce to 0, we re-fit the logistic regression implemented without regularized maximum likelihood or penalty. This allows consistent comparison of the coefficients and performance across the five models that use different sets of independent variables. Model 0 uses the baseline features, while Models 1, 2, and 3 add the variables tracking the athlete's new out-links, in-links, and network diffusion, respectively. We also create a combined model (Model 4).

5.1. *Factors related to decommitment*

Table 3 reports the coefficients and significance for each feature across the five models. By applying the exponential function to the coefficient, we can quantify how a predictor impacts the odds of decommitment.[1] We also list the pseudo R-squared [34] for each model.

Model 0 contains features related to the athlete's individual characteristics; cost/benefit factors impacting satisfaction with his current school; and the availability and attractiveness of alternatives. After applying lasso regression, the size of the baseline is reduced to 20 features. While most of the individual factors were removed during feature selection, a previous decommitment is associated with a 135% increase in the odds of decommitment. We find that features that decrease costs of attendance at the original school are associated with decreased odds. An athlete's odds of decommitment decrease by 31% when he is committed to a college in his home state. Attendance at an in-state school has been linked not only to lower travel costs, but also an increased sense of satisfaction and fit [35]. Interestingly, the number of committed recruits has a negative coefficient; each additional commit is associated with a 4% decrease in the odds. While one might expect that competition for playing time would encourage an athlete to decommit, a solid recruiting class may function as a signal of the quality and stability of a football program. Features that represent decreased benefits are associated with an increased likelihood of decommitment. Commitment to a team that has experienced a recent head coaching change has the largest impact, increasing the odds by 148%. Recruiting activities are also significant predictors. Each official and unofficial visit to the original commitment school is associated with a 17% and 30% decrease in the odds of decommitment, respectively. The sequence of recruiting events may also signal the athlete's attachment to a school [21]. Indeed, we find that the odds of leaving the original school decrease 69% when it was the first offer.

The regression results also indicate that greater availability and attractiveness of alternatives is associated with increased likelihood of decommitment. Each scholarship offer received from another school after committing is associated with a 17% increase in the odds. Visits taken by the athlete post-commitment are highly significant. The odds of decommitment increase 52% for each unofficial visit to another school after committing and 87% for each official visit. Furthermore, each coach visit from another school after committing is associated with a 37% increase in the odds. In an unexpected result, the odds of decommitment increase 1% for each day closer to National Signing Day. While we expected that limited time to find a new team would discourage athletes from decommitting, this result could be related to coaches' last-minute efforts to “poach” athletes committed to other schools in order to fill out their rosters. 25% of all decommitments occurred during January 2016, the last full month of recruitment. Finally, several features in this model are not significant, although their coefficients show the expected signs.

Model 1 adds features related to the number and affiliation of new out-links in the athlete's Twitter network to Model 0. We find that following users associated with the commitment school is associated with a lower likelihood of decommitment in the next month. The odds decrease 22% for each new friendship with a committed recruit in the class of 2016. Conversely, following users from other schools is associated with an increased likelihood of decommitment. An athlete's odds of decommitting increase 7% for each new friend committed to another school and 17% for each new friend currently attending another school. The features tracking friendships with coaches are not significant, although their coefficients have the expected signs.

Model 2 investigates the effect of in-links, adding features tracking

[1] e^x. For example, the coefficient for prior decommitment is 0.8555 in Model 0 ($e^{0.8555} = 2.3525$).

Table 3
Fitted logistic regressions.

Feature	Model 0		Model 1		Model 2		Model 3		Model 4	
Intercept	−1.5812	***	−1.6692	***	−1.6146	***	−1.5089	***	−1.6395	***
Prior decommitment	0.8555	***	0.8006	**	0.7672	**	0.8627	***	0.7711	**
In-state (original)	−0.3664	**	−0.4014	**	−0.4100	**	−0.3975	**	−0.4117	**
US News (original)	−0.0002		−0.0000		0.0000		−0.0004		0.0000	
AP (original)	0.0034		0.0040		0.0039		0.0028		0.0040	
Current win % (original)	−0.0040		−0.0042		−0.0036		−0.0043		−0.0035	
Commits (original)	−0.0362	**	−0.0255	°	−0.0273	*	−0.0219		−0.0261	°
Coach change (original)	0.9091	***	0.9312	***	0.8349	***	0.7990	***	0.8430	***
Sanctions (original)	0.0985		0.0778		0.0635		−0.0197		0.0571	
First offer (original)	−1.1874	***	−1.1203	***	−1.1361	***	−1.1578	***	−1.1188	***
Unofficial (original)	−0.1889	*	−0.2122	*	−0.2142	*	−0.1876	*	−0.2080	*
Official (original)	−0.3619	*	−0.3053	°	−0.2517		−0.3779	*	−0.2701	
Offers (other)	−0.0117		−0.0136		−0.0070		−0.0145		−0.0051	
Closer (other)	−0.0007		0.0011		0.0006		−0.0044		0.0003	
Higher US News (other)	0.0373		0.0433		0.0464		0.0513		0.0458	
Sanctions (other)	−0.0786		−0.0637		−0.0529		−0.1118		−0.0534	
Offers since (other)	0.1556	**	0.1497	**	0.1294	*	0.1427	**	0.1263	*
Unofficial since (other)	0.4179	***	0.4017	***	0.4040	***	0.4278	***	0.3985	***
Coach visit since (other)	0.3142	*	0.2797	°	0.2022		0.3387	*	0.2113	
Official since (other)	0.7519	***	0.6595	***	0.4936	***	0.7291	***	0.5023	***
Days to NSD	−0.0138	***	−0.0150	***	−0.0167	***	−0.0127	***	−0.0170	***
Coach friends (original)			−0.0431							
Recruit friends (original)			−0.2427	***					−0.1739	*
College friends (original)			−0.1288						−0.0707	
Coach friends (other)			0.0180							
Recruit friends (other)			0.0694	***					0.0068	
College friends (other)			0.1590	*					0.1553	°
Coach followers (original)					−0.0992				−0.0807	
Recruit followers (original)					−0.1970	**			−0.0566	
College followers (original)					−0.0988					
Coach followers (other)					0.3122	***			0.3073	***
Recruit followers (other)					0.0902	***			0.0844	**
College followers (other)					0.2368	°			0.1235	
Total committed							0.0784	***		
Total decommitted							−0.0371	***		
Pseudo R-Squared	0.1326		0.1516		0.1687		0.1400		0.1715	

° $p < 0.1$.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

the number of new Twitter followers associated with the commitment school and other schools during the prior month. An increase in followers from the original school is associated with decreased likelihood of decommitment. Each new recruit following the athlete in the prior month decreases the odds by 18%. Additional followers from other schools are associated with increased odds of decommitment-37% for each coach, 9% for each recruit, and 27% for each current player. While the other features tested in this model are not significant, their coefficients carry the expected signs.

Model 3 investigates diffusion, adding features tracking the behavior of an athlete's social network neighbors to Model 0. Each reciprocated friend who has decommitted increases the athlete's odds of decommitting by 8%. Each reciprocated friend with a current, stable commitment is associated with a 4% decrease in the odds. These results suggest that athletes may be influenced by behavior of their peers and that the decision to decommit can be viral in a social network.

Model 4 adds the online social network features tested in Models 1, 2, and 3 to the baseline. Applying lasso regression (C = 0.1) again to account for potential collinearity issues and spurious effects reduces model to the 29 features shown in Table 3 (20 baseline features and 9 Twitter network features). We then refit the reduced model without regularized maximum likelihood or penalty. The coefficients of the final model suggest that, after considering personal, organizational, and environmental factors, Twitter network features do have a significant impact on likelihood of decommitment. For example, each new friendship with a fellow recruit from the original school is associated with a 16% decrease in the odds. The odds increase 17% for each new friendship with a current football player at another school. The odds of decommitment increase 36% for each coach and 9% for each recruit from another school following the athlete in the prior month. Model 4 excludes both of the proposed diffusion features. This result suggests that the structure of the athlete's Twitter network (the number and affiliation of his friends and followers) may be more useful for predicting decommitments than the behavior of his peers.

6. Predictive modeling

6.1. Classification algorithms

In addition to constructing fitted models to better understand athletes' decommitment decisions and the marginal effects of different features, we seek to predict the occurrence of decommitments over time. Because different classification approaches are suited to different types of problems, model selection may greatly impact prediction quality. We consider five standard algorithms for supervised machine learning from the Python scikit-learn package [36]. We perform feature selection specific to each method, beginning with the same set of 45 baseline features (Table 1). Appropriate parameters were selected via grid search (see Appendix A). We provide a brief description of each method, its parameters, and feature selection process below.[2]

[2] More detailed descriptions of each method can be found in textbook [41] and online resources for machine learning (e.g., [40]).

Logistic regression is generalized linear model that estimates the probability of a binary outcome based on one or more independent predictors, and can be used as a classifier by setting a decision rule. This method offers the advantages of fast training time and interpretability, and is suitable for problems where a linear relationship between the predictors and log odds can be assumed. Because logistic regression is sensitive to multicollinearity, we conduct feature selection using a lasso regression with L1 penalty ($C = 0.1$). After removing the extraneous features, we compare logistic regression without regularization to the other classifiers.

A *decision tree* represents classification problems as a series of binary decision nodes, each testing the value of a given feature (e.g., prior decommitment = 1) and resulting in two branches (True/False). Branches may yield additional decision nodes or class-labeled "leaves." The CART decision tree algorithm works top-down to determine the "best" split based on a specified metric. Using a grid search, we select entropy as the split criterion, and the maximum number of features considered was set to $\sqrt{n\_features}$. Because this method can be vulnerable to overfitting, we set the maximum tree depth to 5 levels. Like logistic regression, a decision tree has the advantages of fast training and interpretability. It can output both class labels and probabilities, and performs well in situations where data is not linearly separable.

A *support vector machine* (SVM) determines the optimal hyperplane to separate instances of one class from other class(es) by as large a margin as possible. SVM may construct linear or non-linear hyperplanes depending on the specified "kernel." This feature enables SVM to succeed in situations where linearity cannot be assumed. Based on a grid search, we use a radial basis function kernel with a penalty weight of 1. SVM classifiers generally take longer to train than a logistic regression or decision tree model. To avoid overfitting and accelerate training speed, we apply univariate feature selection (based on an ANOVA F-test of each feature). This process reduces the size of the baseline model from 45 to 29 features. Unlike the logistic regression and decision tree models, SVM is "black box" process that does not tend to be easily interpretable and does not directly yield probability estimates.

An artificial neural network (*ANN*) consists of neurons (nodes) and synapses (connections). Each layer takes input from its predecessors, transforms the input via a transfer function, and provides output to be used by the next layer. The ANN is trained iteratively by adjusting each neuron's weight via back-propagation and does not require feature selection. Based on a grid search, we select the hyperbolic tangent function as the transfer function between the layers, with the Adam algorithm for weight optimization. This method is well-suited to complex, non-linear classification problems but requires large amounts of training data and tends to have longer training times. Like SVM, it is a "black box" process.

Combining the output of multiple decision trees, a *random forest* is a type of ensemble learning method. Class labels are assigned based upon the "votes" of the individual trees. Using a grid search, we determine an ensemble size of 35 trees. This method attempts to correct for overfitting via randomization; each tree is trained on a random subset of the training data, and the decision nodes are based on a random subset of features. Because of this, it is not necessary to specify the maximum depth of the tree. For each individual tree, splits were based upon entropy, with consideration of all features. Generally, a random forest provides higher accuracy and more generalizability than a single decision tree. Like the logistic regression and decision tree models, random forest is highly interpretable and can output both binary class labels and predicted probabilities (average of probabilities from each tree).

6.2. Performance evaluation

To assess the performance of different classification algorithms and models, we use stratified Monte Carlo cross-validation. For each trial, the full data is randomly split into two equal subsets, each with the same proportion of decommitments. We also ensure that instances corresponding to the same athlete are kept together. Given that the distribution of the outcome is highly unbalanced (only 5% of instances are occurrences where the specified athlete decommitted during the specified month), we use oversampling to decrease bias in the training data and improve classifier performance. We create a balanced training set (50% decommitments) by randomly copying and inserting positive instances. The classifier is then trained on the oversampled data and evaluated on the unaltered testing data. This process is repeated for 100 trials.

As different classification algorithms may have different posterior probabilities, the decision rule may have a significant effect on the proposed evaluation metrics. We explore two methods for calibrating the classification threshold, comparing the default fixed threshold ($p = 0.5$) with a rate-driven threshold, where output is ranked by predicted probability and the proportion of predicted positives is determined by the prevalence of the outcome in the total population [37]. We test two rates based on the proportion of decommitments in the data. Because 5% of all athlete-month instances correspond to situations where the specified recruit decommitted during the specific month, we begin with a 5% classification rate. We compare this to a 20% classification rate, representing the proportion of recruits that decommitted at any time during recruitment. The latter approach is likely to result in higher recall and a lower chance of missing potential decommitments. SVM output is transformed to a predicted probability using Platt scaling [38].

We evaluate the performance of each classifier and model based on standard metrics: precision (ratio of true positives to predicted positive instances), recall (ratio of true positives to actual positive instances), and F1 score (harmonic mean of precision and recall). We also measure area under the receiver operating characteristic curve (AUC). The receiver operator characteristic curve (ROC) plots the false positive rate of a classifier against the true positive rate. Thus, an AUC score of 0.5 corresponds to a random classifier, where false positives and true positives are equally likely, while a score of 1.0 represents a perfect classifier. Because decommitments are fairly rare, balanced metrics like AUC and F1 score are more appropriate for assessing the predictive power than accuracy (sum of true positives and true negatives divided by total instances).

Because we seek to predict whether a specific athlete will predict during a specific month, it is possible that some errors results from situations where a model correctly predicts that an athlete is in danger of decommitting, but during the wrong month. In the context of athletic recruiting, early warnings are likely to be a welcome result, giving coaches more time to salvage a vulnerable commitment or recruit a replacement athlete. Similarly, early warning of an employee's turnover intention can give a manager valuable lead time to compose a counter offer or adjust workflow. Thus, we investigate an adjusted true positive rate, accounting for early warnings, in comparison to the standard definition, referring to instances where the classifier predicts a decommitment and the athlete decommitted during the next month.

6.3. Predictive performance

In this section, we examine the fitted logistic regression models, present the results of the initial test of different classification algorithms, and compare the predictive performance of models incorporating Twitter data with the baseline model.

6.3.1. Comparison of classification algorithms

Table 4 displays the average performance of each classifier over 100 trials using the baseline recruiting features. For each classifier, probabilistic output was transformed into a binary prediction using three different decision rules: a fixed probability threshold of 0.5 (FT), a rate-driven threshold where the top 5% of instances ranked by predicted probability were classified as decommitments (RD5), and a rate-driven

Table 4
Comparison of classifier performance (baseline features).

		AUC		Precision		Recall		F1	
Rule	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
FT	Logistic	0.686	0.016	0.116	0.014	0.646	0.063	0.196	0.018
	Decision tree	0.636	0.032	0.094	0.017	0.615	0.132	0.160	0.022
	SVM	0.631	0.027	0.130	0.020	0.416	0.058	0.197	0.026
	ANN	0.594	0.016	0.189	0.035	0.251	0.048	0.211	0.024
	Random forest	0.697	0.030	0.131	0.027	0.640	0.096	0.215	0.033
RD5	Logistic	0.598	0.031	0.230	0.034	0.243	0.082	0.229	0.040
	Decision tree	0.551	0.030	0.154	0.048	0.147	0.085	0.141	0.048
	SVM	0.571	0.029	0.180	0.033	0.191	0.077	0.179	0.040
	ANN	0.582	0.025	0.202	0.032	0.212	0.071	0.200	0.034
	Random forest	0.602	0.038	0.235	0.034	0.250	0.095	0.234	0.047
RD20	Logistic	0.667	0.031	0.133	0.021	0.544	0.127	0.210	0.024
	Decision tree	0.614	0.043	0.114	0.023	0.425	0.176	0.171	0.032
	SVM	0.640	0.040	0.119	0.018	0.493	0.147	0.188	0.024
	ANN	0.648	0.029	0.124	0.018	0.509	0.128	0.195	0.021
	Random forest	0.688	0.042	0.142	0.022	0.583	0.143	0.224	0.028

threshold of 20% (RD20).

These results indicate that, given the same data and initial set of features, random forest achieves the highest performance. Using the fixed threshold of 0.5, random forest has the highest AUC score, outperforming the other methods by 9.8% on average. This represents a significant difference compared to the other methods ($p = 0.006$ for logistic, $p = 5.41 \times 10^{-31}$ for decision tree, $p = 1.98 \times 10^{-39}$ for SVM, and $p = 2.24 \times 10^{-76}$ ANN). Random forest also has the highest F1 score and the second-highest precision and recall. Logistic regression is the second-best performer, achieving the top recall score and second-highest AUC. For both of the rate-driven thresholds (top 5% and top 20%), random forest has the highest scores across the board, while the logistic classifier has the second-highest.

We speculate that the data may not be well-suited to SVM because of the relatively small number of features. In addition, not all of the proposed features are equally useful (discussed in Section 4.1), which can negatively affect the performance of the decision tree classifier. Ultimately, the size of our dataset may not be large enough to support ANN. While random forest has the best overall performance in this preliminary test, logistic regression has the advantage of being easily interpreted, which is paramount in real-world sports applications. Interpretability is also useful when comparing results with previous studies on HR turnover. Based on this preliminary test, we present results of random forest and logistic classifiers for the remainder of the paper.

6.3.2. Value added by social media data

We compare the predictive performance of the baseline model containing only recruiting and institutional data (Model 0) with models incorporating Twitter data (Models 1–4). Fig. 2 displays the average predictive performance of the logistic and random forest classifiers over 100 random trials, using the default classification threshold of 0.5.

For the random forest classifier, models 1, 2, 3, and 4 dominate the baseline in terms of AUC, precision, and F1 score. For the logistic classifier, the models incorporating Twitter data dominate the baseline across all metrics. These results demonstrate that incorporating Twitter data improves predictive performance. Among the individual sets of online social network features, Model 2-tracking in-links from coaches, recruits, and current college football players-demonstrates the largest performance increase relative to the baseline. For the random forest classifier, Model 2 achieves a 2.2% increase in AUC, 9.2% increase in precision, 1.9% increase in recall, and 7.9% increase in F1 score over the baseline. For the logistic classifier, Model 2 achieves gains of 3.7%, 14.4%, 2.5%, and 12.5%, respectively. Model 1, tracking out-links, also outperforms the baseline for both classification algorithms. Model 3, with features tracking decommitment diffusion, fails to achieve significant improvements over the baseline. These results indicate that features focusing on Twitter network structure add more value to decommitment predictions than those related to diffusion and social influence.

Ultimately, Model 4 is the top performer. For the random forest classifier, Model 4 achieves a 2.2% improvement in AUC, 17.3% improvement in precision, and 13.3% improvement in F1 score over the baseline with only recruiting and institutional data. For the logistic classifier, Model 4 achieves a 4% improvement in AUC, 19.5% improvement in precision, 0.8% improvement in recall, and 16.4% improvement in F1 score. These results suggest that a combination of features measuring different aspects of online social network structure is more useful for predicting decommitments than any individual set of features.

6.3.3. Evaluation of early predictions

Because we use data up to a given month to predict whether an athlete will decommit during the next month, some false positives are the result of predicting a decommitment too early. For example, Model 4 predicted that Kevin Harmon, a wide receiver committed to the University of South Carolina, would decommit in November when in fact he decommitted in December. Thus, simply evaluating the predictions based on decommitments in the next month may actually underestimate the true utility of our models. Table 5 presents the results of evaluating the random forest and logistic classifiers based on two definitions of true positive. First, instances where the classifier predicts a decommitment and the athlete decommitted during the next month ($m + 1$). Second, instances where the classifier predicts a decommitment and the athlete decommitted during any future month ($m + n$).

Treating early predictions as true positives yields notable increases across all metrics. For Model 4 and the random forest classifier, AUC, precision, recall, and F1 score increase 10.8%, 92.6%, 22.2%, and 72.7%, respectively. The logistic classifier achieves a 13.6% increase in AUC, 131% increase in precision, 25.6% increase in recall, and 101.1% increase in F1 score.

7. Example application to recruiting decision support

In addition to demonstrating the utility of online social network features for predicting turnover, we seek to provide practical decision support for recruiters. In Table 6, we produce a sample report for the University of Utah, tracking each committed recruit's predicted probability of decommitment calculated by the logistic classifier as of November 1, 2015. Because of the large effect of head coaching change on the predicted odds of decommitment, we selected a program with a stable coaching staff during the 2015–2016 season.

According to Model 4 (which combines both recruiting and Twitter date), Micah Croom, Jay Griffin, Devontae Henry-Cole, and Lorenzo Neal were predicted to decommit during the month of November. For Croom, his estimated probability of decommitment was increased by the fact that he was out-of-state commit who had not taken any visits to Utah. Similarly, Griffin was an out-of-state commit who had not taken any visits to Utah. He had been followed on Twitter by 5 recruits from other schools during the month of October, and had not followed any additional Utah coaches, current players, or recruits. Armed with the information that both athletes had an estimated 60% chance of decommitting in the next month, it would not be unreasonable for a coach to conclude that they were "lost causes," and turn the attention of his recruiting staff toward securing replacement players. While Croom did not decommit in November as predicted, he did decommit in December, signing with Dartmouth. Griffin decommitted in November, eventually signing with New Mexico.

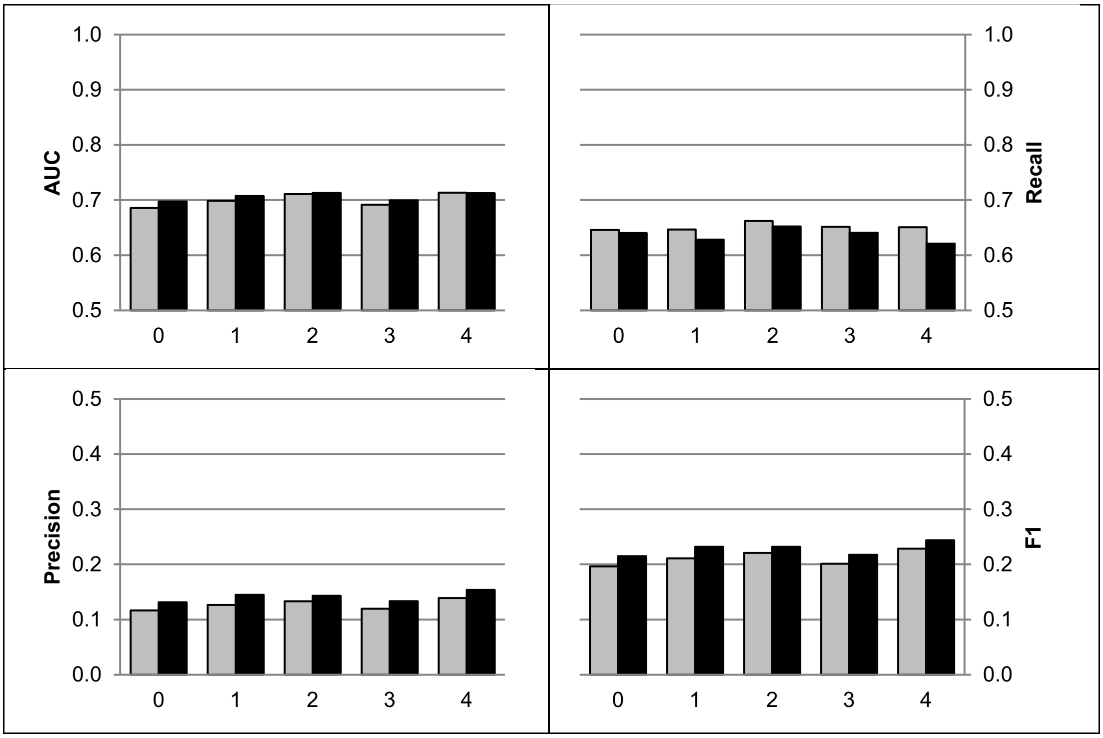

Fig. 2. Comparative model performance (logistic = gray, random forest = black).

The results of the probabilistic classifier can provide coaches with nuanced information to shape recruiting and retention strategies. Henry-Cole and Neal's predicted probabilities of decommitting in November were lower than Croom and Griffin's. In this situation, a coach might decide to hedge his bets, simultaneously taking actions to salvage the relationship and discreetly pursuing other athletes. Neal decommitted in November and signed with Purdue.

A coach could also use a report like this to note recruits who are not expected to decommit in the next month, but whose probability is still relatively high. For example, our model predicts that Demari Simpkins and Ohaji Hawkins had almost a 50% chance of decommitting. Hawkins eventually decommitted in February, signing with Eastern Michigan. Information about which recruits may be on the cusp of decommitting could enable a coach to take action to strengthen the commitment, such as increasing communication and/or visiting the recruit. A risk-averse coach might be concerned by Alema Pilimai, Daevon Vigilant, and Kahi Neves, each of whom was predicted to have a > 20% chance of decommitment. Indeed, all three athletes decommitted later. While this example uses a default 50% fixed classification threshold, coaches could adjust this figure as desired, and we present other possible decision rules in Section 6.2.

Ultimately, the threshold at which a coach determines a commitment to be vulnerable may depend on his individual preferences or on other factors, such as the value of the recruit, playing position, remaining time until National Signing Day, or personal communications. This report is intended to illustrate how our model, incorporating both recruiting and Twitter data to predict decommitments over time, may assist coaches in shaping recruiting strategies during this process. In application, data could be gathered to produce new predictions on a bi-weekly or weekly basis, giving recruiters up-to-date information. Furthermore, our model could also be used by coaches to monitor changes in athletes' predicted probability of decommitment over time, paying special attention to large increases over the course of one or more months.

8. Discussion and conclusions

This study represents a novel addition to literature on organizational turnover, investigating the application of established predictors of personnel turnover to the athletic domain. Specifically, our results support the importance of satisfaction and perceived alternatives. In the fitted logistic regression significant predictors of decommitment include, features representing the costs and benefits of enrollment at the original commitment school in comparison to other schools that have offered scholarships. However, we find mixed evidence for the effectiveness of athletes' personal characteristics in predicting decommitments. Only the feature tracking past decommitments was significant. This may be due to our focus on predicting decommitments over time, where static features like star rating or height may be less useful.

We also explore the value of social network features. Our work takes a unique approach to exploring the intersection between networks and turnover by utilizing Twitter data. Though one's online connections are

Table 5
Performance of early predictions (Model 4).

		AUC		Precision		Recall		F1	
Method	Period	AUC	SD	Mean	SD	Mean	SD	Mean	SD
Logistic	m + 1	0.713	0.016	0.139	0.015	0.651	0.047	0.228	0.019
	m + n	0.810	0.014	0.321	0.036	0.817	0.038	0.460	0.033
Random forest	m + 1	0.712	0.026	0.154	0.026	0.621	0.091	0.244	0.029
	m + n	0.790	0.025	0.297	0.049	0.759	0.095	0.421	0.041

Table 6
Predicted probability of decommitment for University of Utah (November 2015).

Name	Star	Position	Commit date	P(Decommit)	Decommit Month
Micah Croom	3	Safety	12/21/14	0.612	December
Jay Griffin	3	Wide receiver	6/29/15	0.598	November
Devontae Henry-Cole	3	Running back	7/11/15	0.557	–
Lorenzo Neal	3	Defensive tackle	8/10/15	0.553	November
Demari Simpkins	3	Wide receiver	8/2/15	0.497	–
Ohaji Hawkins	3	Safety	8/30/15	0.435	February
Alema Pilimai	3	Athlete	10/1/2015	0.391	February
Daevon Vigilant	3	Running back	6/26/15	0.348	January
Kahi Neves	3	Quarterback	10/6/14	0.271	December
Tucker Scott	3	Offensive tackle	3/30/15	0.166	–
Cole Fotheringham	3	Tight end	7/6/15	0.159	–
RJ Hubert	3	Wide receiver	10/1/2015	0.152	–

an imperfect proxy for offline social ties, we find that-after considering personal, organizational, and environmental factors-online social network features had a significant impact on the odds of decommitment. Furthermore, the addition of Twitter data improved the fit of the logistic regression, increasing the pseudo R-squared by 29%. This result is consistent with previous research on offline, workplace social networks (e.g., [19]). Overall, our results were consistent with the turnover framework, suggesting that the findings of this work may be generalizable to other organizational settings.

After constructing explanatory models to better understand the effect of different features on athletes' decommitment decisions, we predict the occurrence of decommitments over time. Because different classification approaches are suited to different types of problems, we first compare the performance of five different algorithms on the same data, using the same set of baseline features derived from recruiting and institutional data. Experiments with different classification thresholds ($p = 0.5$, top 5%, and top 20%), reveal that random forest achieves the best overall performance, with logistic regression as a close second. In other words, both algorithms are well-suited to the data, with random forest offering slightly more predictive power, while logistic has the advantages of greater simplicity and interpretability.

Our results indicate that Twitter data consistently added value to predictive models. Among the three groups of network features explored, we find those focusing on out-links and in-links (Models 1 and 2) contribute more to predictive performance for both the random forest and logistic classifiers than those based on network diffusion. Thus, the effect of turnover contagion [25] was not supported in our application to athletic decommitments. However, a more detailed analysis may yield different results. Specifically, focusing on the behavior of athletes committed to the same school or differentiating by tie strength may be useful for exploring the question of decommitment diffusion in future work. Model 4, which combines features related to different aspects of the athlete's Twitter network, was the highest performer. For the random forest classifier, Model 4 achieves an AUC score of 0.713, a 2.2% over the baseline with only recruiting and institutional data. Interestingly, as more Twitter data is added, the performance gap between the random forest and logistic classifiers narrows. The logistic classifier also achieves an AUC of 0.713, a 4% improvement over the baseline. These results suggest that a combination of features measuring different aspects of online social network structure is more useful for predicting decommitments than any individual set of features. Our findings are especially significant in light of the prohibitive cost of tracking offline social networks. Network data may be easily retrieved from social media sites like Twitter, offering more opportunities to perform large-scale, holistic analyses.

We note potential limitations of this work. First, the accuracy of the recruiting data in the study is dependent on its source. 247Sports.com incorporates both user-supplied data and expert data, which we believe makes it both a comprehensive and up-to-date source of recruiting information. Second, we use data on decommitments among a single recruiting class during the final months of recruitment. It is possible that predicting earlier decommitments or other recruiting classes may yield different results. Third, the scope of this project is limited to athletes with public Twitter profiles and may not be generalizable to athletes without a social media presence (though that situation has become even rarer over past two years). Fourth, this project utilizes only publicly available data. We contend that this is a strength because our methods can be implemented and replicated by other academic researchers or industry practitioners. However, there are certainly other factors that may influence decommitments that are not represented in our data, such as academic performance or private correspondence between athletes and coaches. Nevertheless, our models achieve reasonable performance without such features. In a realistic application to athletic recruiting, coaches are more likely to possess additional data which can be added to our models and improve decommitment predictions. Finally, as this is not an experimental study, no causality can be inferred from our results. While our findings are largely consistent with existing frameworks of turnover and social network theories, we cannot state that specific variables cause decommitments.

There are several interesting directions for future work. This study focuses specifically on structural features (in-links and out-links), but social media is a rich source of data that can be further explored. Analysis of an athlete's actions online, including replying and retweeting, may provide more nuanced information about the strength of his social network ties, and text data derived from tweets may give insight into the athlete's satisfaction with his current school and decommitment intentions. Further analysis of centrality, connectivity, and community structure in the online network could be fruitful.

Overall, this work represents both a promising first step in predicting decommitments in college football, and using online social network data to explain and predict turnover in other organizational settings. Our findings suggest that recruiters in college football, HR, or other domains, could benefit from considering the information communicated by candidates' online social networks.

Declaration of interests

None.

Acknowledgements

The authors would like to thank the anonymous reviewers whose feedback was instrumental in improving this paper's content and presentation.

Appendix A. Grid search parameters

Table A.7 details which parameters were selected during a grid search, using area under the receiver operating characteristic curve (AUC) as the scoring function.

Table A.7
Range of grid search values for classifier parameters.

Method	Parameter	Range of values
Logistic	Regularization penalty	L1, L2
	Regularization weight	0.001–0.5, in increments of 0.1; 0.5–2.5, in increments of 0.5
Decision Tree	Criterion	Gini, entropy
	Max features	$\sqrt{n\_features}$, $\log_2 n\_features$, none
	Max depth	5–50, in increments of 5; 50–250, in increments of 50
SVM	Kernel	Linear, polynomial, radial basis function, sigmoid
	Error penalty	0.001–0.5, in increments of 0.1; 0.5–2.5, in increments of 0.5
ANN	Transfer function	Logistic, hyperbolic tangent, rectified linear unit, identity
	Weight optimization	Adam, L-BFGS, stochastic gradient descent
Random Forest	Ensemble size	5–50, in increments of 5; 50–250, with increments of 50
	Criterion	Gini, entropy
	Max features	$\sqrt{n\_features}$, $\log_2 n\_features$

References

[1] National Collegiate Athletic Association, NCAA Football Attendance, http://www.ncaa.org/championships/statistics/ncaa-football-attendance, (2018) , Accessed date: 1 March 2018.

[2] P.K. Hunsberger, S.R. Gitter, What is a blue chip recruit worth? Estimating the marginal revenue produce of college football quarterbacks, Journal of Sports Economics 16 (6) (2015) 664–690, https://doi.org/10.1177/1527002515580938.

[3] E. Brady, J. Kelly, S. Berkowitz, Schools in power conferences spending more on recruiting, USA Today Sports, 2015 https://www.usatoday.com/story/sports/ncaaf/recruiting/2015/02/03/college-football-recruiting-signing-day-sec-power-conferences/22813887/ , Accessed date: 1 January 2016.

[4] C. Carson, Introducing the 247 sports decommitment tracker, 247 Sports, 2016 http://247sports.com/Bolt/Introducing-the-247Sports-Decommitment-Tracker-44280624 , Accessed date: 1 June 2016.

[5] J. Crabtree, The 'Social' Science of Recruiting, ESPN http://espn.go.com/college-football/recruiting/story/_/id/14646545/social-media-becomes-powerful-aide-dangerous-connection-recruiting, (2016) , Accessed date: 1 June 2016.

[6] C. Ding, H.K. Cheng, Y. Duan, Y. Jin, The power of the 'like' button: the impact of social media on box office sales, Decision Support Systems 94 (2016) 77–84, https://doi.org/10.1016/j.dss.2016.11.002.

[7] M. Bogaert, M. Ballings, D. Van den Poel, The added value of Facebook friends data in event prediction, Decision Support Systems 82 (2015) 26–34, https://doi.org/10.1016/j.dss.2015.11.003.

[8] Towers Watson, The Targeted Employee Value Proposition: Drive Higher Performance through key Talent and Differentiated Rewards, https://www.towerswatson.com/en/Insights/IC-Types/Survey-Research-Results/2013/12/2013-2014-talent-management-and-rewards-study-north-america, (2014) , Accessed date: 1 March 2018.

[9] J. Krider, K. O'Leonard, R. Erickson, What Works Brief: Talent Acquisition Factbook (Bersin by Deloitte), https://legacy.bersin.com/uploadedfiles/042315-ta-factbook-wwb-final.pdf, (2015) , Accessed date: 1 January 2017.

[10] D.M. Cable, D.B. Turban, The value of organizational reputation in the recruitment context: a brand-equity perspective, Journal of Applied Social Psychology 33 (11) (2003) 2244–2266.

[11] S. Helm, A matter of reputation and pride: associations between perceived external reputation, pride in membership, job satisfaction and turnover intentions, British Journal of Management 24 (2013) 542–556, https://doi.org/10.1111/j.1467-8551.2012.00827.x.

[12] R.W. Griffeth, P.W. Hom, S. Gaertner, A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium, Journal of Management 26 (2000) 463–488, https://doi.org/10.1177/014920630002600305.

[13] A.L. Rubenstein, M.B. Eberly, T.W. Lee, T.R. Mitchell, Surveying the forest: a meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover, Personnel Psychology 71 (1) (2018) 23–65, https://doi.org/10.1111/peps.12226.

[14] J.G. March, H.A. Simon, Organizations, Wiley, New York, NY, 1958.

[15] D.P. Moynihan, S.K. Pandey, The ties that bind: social networks, person-organization value fit, and turnover intention, Journal of Public Administration Research and Theory 18 (2007) 205–227, https://doi.org/10.1093/jopart/mum013.

[16] J.L. Cotton, J.M. Tuttle, Employee turnover: a meta-analysis and review with implications for research, Academy of Management Journal 11 (1) (1986) 55–70.

[17] J.M. Dumond, A.K. Lynch, J. Platania, An economic model of the college football recruiting process, Journal of Sports Economics 9 (1) (2008) 67–87, https://doi.org/10.1177/1527002506298125.

[18] K.W. Mossholder, R.P. Settoon, S.C. Henagan, A relational perspective on turnover: examining structural, attitudinal, and behavioral predictors, Academy of Management Journal 48 (4) (2005) 607–618.

[19] T.H. Feeley, J. Hwang, G.A. Barnett, Predicting employee turnover from friendship networks, Journal of Applied Communication Research 36 (1) (2008) 56–73, https://doi.org/10.1080/00909880701799790.

[20] M.P. Mirabile, M.D. Witte, A discrete-choice model of college football recruit's program selection decision, Journal of Sports Economics 18 (3) (2015) 211–238.

[21] K.G. Bigsby, J.W. Ohlmann, K. Zhao, Online and off the field: predicting school choice in college football recruiting from social media data, Decision Analysis 14 (4) (2017) 261–273, https://doi.org/10.1287/deca.2017.0353.

[22] R. Elliot, J. Maguire, Getting caught in the net: examining the recruitment of Canadian athletes in British professional ice hockey, Journal of Sport & Social Issues 32 (2) (2008) 158–176, https://doi.org/10.1177/0193723507313927.

[23] C. Croft, Factors Influencing Big 12 Conference College Basketball Male Student-athletes' Selection of a University (Doctoral Dissertation), Digital Commons at University of Texas El Paso, 2008 (AAI3313419).

[24] R.M. Fernandez, E.J. Castilla, P. Moore, Social capital at work: networks and employment at a phone center, American Journal of Sociology 105 (5) (2000) 1288–1356.

[25] W. Felps, T.R. Mitchell, D.R. Hekman, T.W. Lee, B.C. Holtom, W.S. Harman, Turnover contagion: how coworkers' job embeddedness and job search behaviors influence quitting, Academy of Management Journal 52 (2009) 545–561, https://doi.org/10.5465/AMJ.2009.41331075.

[26] S. Khan, A&M assistant's tweets add fuel to fire of no. 3 QB's decommitment, ESPN, 2016 http://espn.go.com/college-football/recruiting/story/_/id/15465600/texas-aggies-lose-second-recruit-night-coach-tweets , Accessed date: 1 June 2016.

[27] CBS Sports. (n.d.). 247 Sports. http://247sports.com. Accessed 1 August 2015.

[28] National Collegiate Athletic Association, 2015–2016 NCAA Division I Manual, National Collegiate Athletic Association, Indianapolis, 2016.

[29] A. Staples, The commitment project: studying recruits, college football edition, Sports Illustrated, 2012 http://www.si.com/more-sports/2012/01/20/commitment-project , Accessed date: 1 August 2015.

[30] SB Nation College News, The regular season's last football rankings: playoff, polls, and computers, SB Nation, 2014 http://www.sbnation.com/college-football/2014/12/7/7347525/college-football-rankings-top-25-playoff-ap-coaches , Accessed date: 1 January 2016.

[31] U.S. News and World Report, Best Colleges, U.S. News and World Report, Washington, 2014.

[32] Twitter, REST APIs, https://dev.twitter.com/rest/public, (2016) , Accessed date: 1 August 2015.

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

[34] D. McFadden, Conditional logit analysis of qualitative choice behavior, in: P. Zarembka (Ed.), Frontiers in Econometrics, Academic Press, Berkeley, 1974, pp. 105–142.

[35] J.Q. Barden, D.J. Bluhm, T.R. Mitchell, T.W. Lee, Hometown proximity, coaching change, and the success of college basketball recruits, Journal of Sport Management 27 (2013) 230–246.

[36] F. Pedregosa, et al., Scikit-learn: machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[37] J. Hernandez-Orallo, P. Flach, C. Ferri, A unified view of performance metrics: translating threshold choice into expected classification loss, Journal of Machine Learning Research 13 (2012) 2813–2869.

[38] J. Platt, Probabilistic outputs for support vector machines and comparison to

regularized likelihood methods, in: A.J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, 1999, pp. 61–74.

[39] K. Zhao, X. Wang, S. Cha, A.M. Cohn, G.D. Papandonatos, M.S. Amato, J.S. Pearson, A.L. Graham, A multirelational social network analysis of an online health community for smoking cessation, Journal of Medical Internet Research 18 (2016) e233, https://doi.org/10.2196/jmir.5985.

[40] G. Ericson, W.A. Rohm, A. Jenks, J. Martens, B. Rohrer, How to choose algorithms for Microsoft Azure machine learning, Microsoft Azure, 2017 https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice , Accessed date: 1 August 2018.

[41] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Education, Inc., Boston, 2006.

[42] D.G. Pope, J.C. Pope, The impact of college sports success on the quantity and quality of student applications, Southern Economic Journal 75 (3) (2009) 750–780.

[43] K. Goidel, J. Hamilton, Strengthening higher education through gridiron success? Public perceptions of the impact of national football championships on academic quality, Social Science Quarterly 87 (4) (2009) 851–862, https://doi.org/10.1111/j.1540-6237.2006.00439.x.

[44] R.A. Baade, J.O. Sundberg, Fourth down and gold to go? Assessing the link between athletics and alumni giving, Social Science Quarterly 77 (4) (1996) 789–803.

[45] S.A. Bergman, T.D. Logan, The effect of recruit quality on college football team performance, Journal of Sports Economics (2014) 1–23, https://doi.org/10.1177/1527002514538266.

Kristina Gavin Bigsby is Visiting Assistant Professor in the Department of Management Sciences in the Tippie College of Business at the University of Iowa. She earned her Ph.D. in Information Science from the University of Iowa in August 2018. Her research focuses on the intersection of social network analysis and individual decision-making, especially in organizational contexts, and her previous work on college football recruiting and social media been received national media attention. Email: kristina-gavin@uiowa.edu.

Jeffrey W. Ohlmann is Associate Professor of Management Sciences and Huneke Research Fellow in the Tippie College of Business at the University of Iowa. He earned his Ph.D. in Industrial & Operations Engineering from the University of Michigan. Professor Ohlmann's research interests on the modeling and solution of decision-making problems spans applications in sports analytics, transportation and logistics, and agriculture. Methodologically, Professor Ohlmann is interested in developing heuristic search methods, particularly those for dynamic and stochastic optimization problems. Due to the relevance of his work to industry, he was bestowed the George B. Dantzig Dissertation Award in 2004 and was recognized as a finalist for the 2008 Daniel H. Wagner Prize for Excellence in Operations Research Practice. Email: jeffrey-ohlmann@uiowa.edu.

Kang Zhao is an Assistant Professor at the Tippie College of Business, The University of Iowa. He obtained his Ph.D. from Penn State University. His current research focuses on data science and social computing, especially the analysis, modeling, mining, and simulation of social/business networks and social media. His research has been featured in public media from more than 20 countries. He served as the Chair for the INFORMS Artificial Intelligence Section 2014–2016. Email: kang-zhao@uiowa.edu.