Happiness, an inside job?

Turnover prediction using employee likeability, engagement and relative happiness

Jose Berengueres
College of IT
UAEU
Al Ain, UAE
jose@uaeu.ac.ae

Guillem Duran
R&D Department
Source{d}
Madrid, Spain
guillem@sourced.tech

Dani Castro
IT Department
Happyforce S.L
Barcelona, Spain
dani@myhappyforce.com

In this paper, we describe how to rank employees for risk of turnover by using data obtained from a happiness self-reporting app. Two data sources are used: daily happiness and social interactions. The data spans 2.5 years and 4,356 employees of 34 companies based in Barcelona. For each employee, we build features at three levels: individual, company level and social interaction graph level. We develop various turnover risk models and we compare how different features affect performance prediction. The results show that the top three features that explain turnover risk are: ratio of likes received (likeability), posting frequency (engagement), and relative happiness (employee happiness normalized by company mean). Surprisingly, a priori expected explanatory features such as mean happiness level and the ratio of likes (positivity), were not significant. Precision@50 = 80% out of a test set with 116 churns, sample size N=2k.

Keywords—Learning-to-Ranking; Information Retrieval; Happiness; Privacy.

I. INTRODUCTION

Here we describe how we used: (i) employee self-reported daily happiness at work and (ii) social interaction data from an **anonymous** forum to estimate which employees will churn in the next weeks. Data was collected via an app which is deployed by companies who want to monitor employee happiness and obtain feedback with the purpose to improve the workplace. To predict if a given employee would churn, first we generated various types of features to create a training set. For each employee, a vector of features was generated. According to how they were generated the features can be divided into five groups:

(i) **Employee** features (**E**): **individual** features such as average happiness. These features are unique to each employee.

(ii) **Company-wide** features (**C**): features such as the average happiness of the company to which the employee belongs. These features are **common** to employees that belong to the same company. (Also referred as entity-faceted features elsewhere.)

(iii) **Social graph** features (**NMF**): A magnitude that indicates the employee affinity to a social graph cluster.

(iv) **EC** features: individual features (**E**) normalized by the corresponding company mean (**C**). These features are **unique** to each employee.

(v) Company ID (as dummy variable).

Fig.1 shows a visual description of the features and timeline.

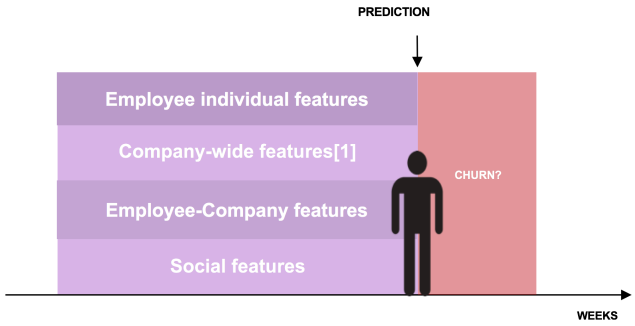


Fig. 1 Individual features as well as so-called entity-faceted features are used to predict if an employee churns within 12 weeks from a given date.

A. Motivation

Our main goal here is **not** to optimize an employee turnover model but to clarify which features are related to turnover and turnover risk so that company managers can get a sense of what percentage of their workforce is likely to churn. In other words, to understand the risk factors of turnover. In addition, we are interested in identifying which feature group is the appropriate unit of analysis: the individual, the company, or the social graph.

The cost of employee turnover has been pointed out extensively in the literature. A high turnover rate not only increases human resource costs, which can reach up to 150% of the annual salary per replaced employee, but it also has social costs, as it is correlated with lower wages, lower productivity per employee, and not surprisingly, a less loyal workforce [1]. For reference, in 2006, turnover at *Walmart's Sam's Club* was 44% with an average hourly pay of $10.11, while at *Costco* it was a much lower 17% with a higher $17.0 hourly wage [2]. In addition, a more recent study correlated companies with low

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ASONAM '17, July 31-August 03, 2017, Sydney, Australia
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4993-2/17/07…$15.00
http://dx.doi.org/10.1145/3110025.3110132

turnover with a series of socially positive characteristics dubbed ***high-involvement*** work practices [3].

B. Turnover prediction

On the other hand, research on employee turnover (churn) is not a prolific topic in the engineering community. In IEEE publications, one can find just over 278 publications with titles containing the keyword *churn*, and the bulk of those focus on ***customer*** churn, and specifically churn in the telecommunications industry, while on the topic of *employee* churn there is just one title indexed [4]. Here, we address employee turnover in a variety of companies in fields such as: education, retail and IT services. The goal is to clarify the characteristics of employees that will churn (or that are at risk of churning), to help companies understand the causes so they can reduce the turnover rate.

The paper is organized as follows. In the following section we review previous literature, particularly focusing on entity-faceted features. In section 3, we explore the data set from various points of view (company level, social interaction level). In section 4, we explain feature engineering. In section 5, we compare the predictive power of various feature groups. In section 6, we interpret the results.

II. Related Work

A. Customer turnover prediction

In the past, most turnover and churn research has been directed towards customer churn in telecommunications. A wide range of modeling techniques such as decision trees, evolutionary algorithms [5], support vector machine [6] and logistic regression among many others have been successfully applied to predict churn. Similarly, various kinds of input data have been tried. Besides the traditional customer data available to carriers which includes phone, data usage, and CRM data, previous works have also utilized data from social networks [7,8], specialized feature engineering [9], and phone call data [10], all of which have been successfully used to improve churn prediction. However, the field has advanced at a quick pace. An instructive overview of the state-of-the art in churn prediction modeling (albeit based on financial services) is available at the kernels section of Kaggle's American Express competition that ended on Jun 5th 2014 [11]. In addition, it is worth mentioning that a useful predictor is not just the one that discriminates churn correctly, but one that can estimate as well the likelihood of churn. This is economically important so companies can act before a value customer churns [5].

B. Entity-faceted features

If we consider our churn prediction problem as an information retrieval problem, we can apply and take advantage of the feature strategy used at LinkedIn [12]. The authors describe how to improve the matching of jobs posted on LinkedIn to job searches. Performance was measured via an A/B test and click through rates on the returned result pages. The authors took cues from the problem of ranking results of a web query. In a web query a list of possible results (pages) is returned and it is important to return the relevant results at the top. Hence the use of metrics such as Precision@, and NGCC. The authors used previous techniques and enhanced them successfully (by about 20% to 8%) by adding so-called **entity-dependent** features. To match a query with relevant results, the authors not only used features based on the query's keywords, and the user profile, but also added features **dependent** on the company that the user targeted (entity-faceted features). In our case, we can consider that we just have the same query all the time, *Who will most likely churn in the next three months?*, and that we need to rank the results (a list of employees) by risk of churn from high to low. As we will see later, adding company (entity) specific features, in addition to employee only dependent features improved baseline metrics by a similar amount as reported in [12].

III. Data

The data and code (R and python notebooks) is public at https://www.kaggle.com/harriken/employeeturnover. Data used here was collected from 2014-05-10 to 2017-03-08. The data was generated by employees who work(ed) at 34 different companies. The companies belong to one of the following sectors: e-payment startup, IT consulting services, retail, manufacturing, services, tourism or education. About half of the companies are multinational and the other half are Barcelona-based companies. Employee data was collected in the framework of corporate feedback programs as part of their companies' improvement initiatives (kaizen). The bulk of the employees used the app in Spain (Barcelona area). More than 90% are Spanish nationals. The comments are written in various languages: 97% in Spanish, 2% in English, 1% Catalan. 1.6% of the comment file (measured in bytes) corresponded to emoji symbols. Churn was detected when the employee was removed from the ERP system which is linked to the backend of the app provider (Happyforce S.L). See section 3.3 for more details on churn detection.

A. Overview

The data consisted of four tables: votes, comments, interactions and churn. A *vote* was obtained when an employee opened the app and answered the question: *How happy are you at work today?* To vote the employee indicates their feeling by touching one of four icons that appeared on the screen. See 1st screen, Table 1. The User Interface of the English version is shown in the table. The default UI was in Spanish. After the employee indicates their happiness level, a second screen appears where they can input a text explanation (usually a complaint, suggestion or comment), this is the comments table. Out of 4,356 employees, 2,638 employees commented at least once. Finally, in a third screen the employee can see their peers' comments and like or dislike them. This data is stored in the interactions table. 3,516 employees liked or disliked at least one of their peers' comments. In the following, we describe and visualize key metrics of each of the tables.

B. Self-reported happiness

Table 2 shows basic statistics about happiness votes aggregated by company. A four-grade scale is used to convert the happiness categories to numeric form (see table 1). The mean of votes is **2.85** (mean employee happiness). The mean happiness of companies is **3.08** (N=34). Fig. 2 contains a histogram of votes recorded. The dips in August and January correspond to the Summer and Christmas vacation periods in Spain. Weekly dips are weekends. Fig. 3 is a histogram of votes by company. Company names are anonymized with a one or

two-letter code. The most common answer to the question "How happy are you at work today?" is *good*. Fig. 4 is a histogram of app usage derived from the votes' timestamps. 0.5 means that the employee used the app once every two days, (days include working days, weekends and holidays). About 10% of the employees used the app every day to report their happiness level.

Description of tables:

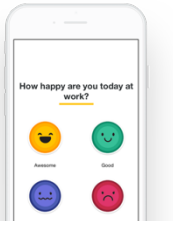
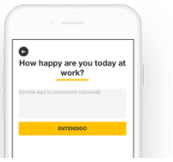
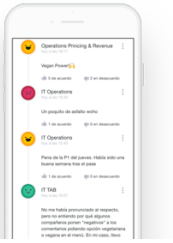
Table **(Rows)**	**Feed-back**	**UI flow**
Happiness votes (221k)	**How happy are you at work today?** - 4: Great - 3: Good - 2: So-so - 1: Pretty Bad	1st screen 
Comments (29.5k)	**Comment box** (optional)	2nd screen 
Likes (284k) Dislikes (52k)	**Anonymous forum** Users can: - **view** comments - **like** a comment - **dislike** a comment	3rd screen 

TABLE I. STATS ON HAPPINESS BY COMPANY

Item (by company N=34)	**Max**	**Min**	**Mean**	**Median**
N of employees participating	569	9	122	74
Count of votes by company (all period)	49k	38	6k	2k
Mean happiness (by company)	**3.62**	**2.48**	**2.99**	**2.99**
Average count of votes per employee (by company)	125	2.33	33	22
Average activity period in weeks of an employee	106	0.14	14.7	8.42

Nvotes=221k votes

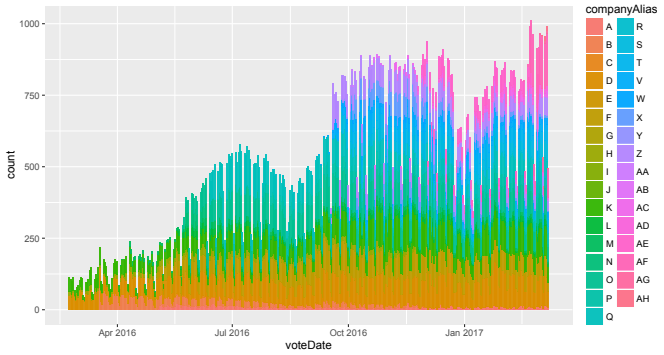


Fig. 2 Count of votes caste daily. Small dips are weekends.

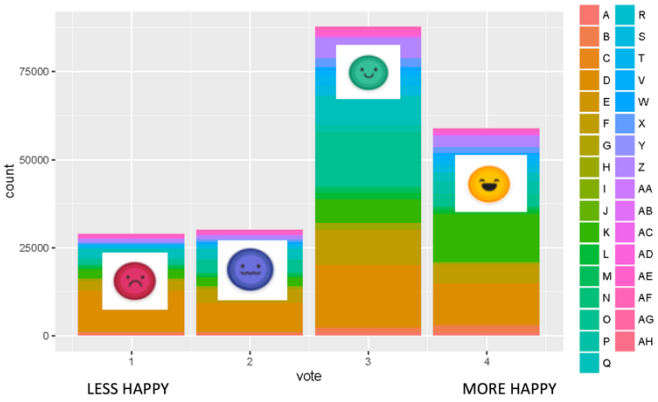


Fig. 3 Count of answers to the question: 'How happy are you at work today?'

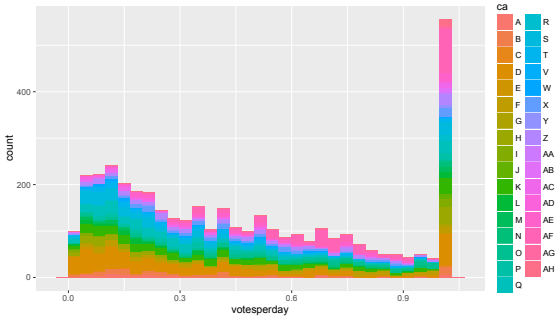


Fig. 4 Usage of the app by vote activity. About 10% of employees use the app every day.

Fig. 5 shows the usage by weekday. By design, employees can only vote once a day. Fig. 6 is a radar chart of employee self-reported (numeric) happiness averaged by weekday. Significance: A Kolmogorov test was used to compare each employee's mean vote on Saturdays vs. Tuesdays. (N=2,638 employees). The hypothesis that CDF of Tuesdays' (numeric) happiness mean lies below that of Saturdays' shows a p-value = 2.2^-16.

C. *Comments*

Table 3 visualizes basic statistics of the comments table by company. In total, 29k comments were recorded with a median length of 55 characters per comment. Emoticons were not considered when calculating length. Fig. 7 shows the correlation between the comment publication day by weekday and how

many likes the comment receives. Comments posted on Mondays are liked less frequently. (A comment can be liked for a few days after being posted). Kolmogorov test comparing Monday and Saturday yields p-value = 3.5^-05.

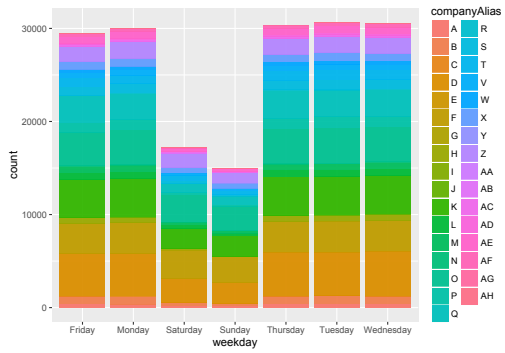


Fig. 5 Vote count by weekday. On Weekends people use the app less.

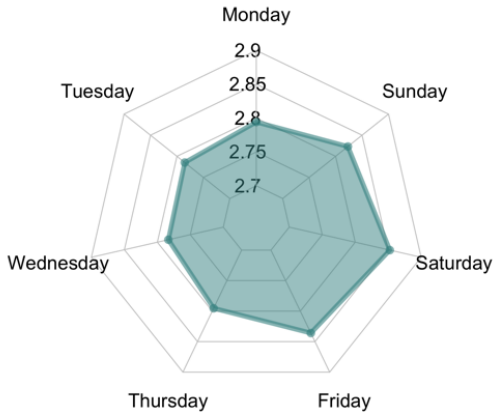


Fig. 6 Radar chart of mean happiness by weekday. Saturdays are the happiest days and Tuesdays are the least "happy" days. Kolmogorov p<10^-15

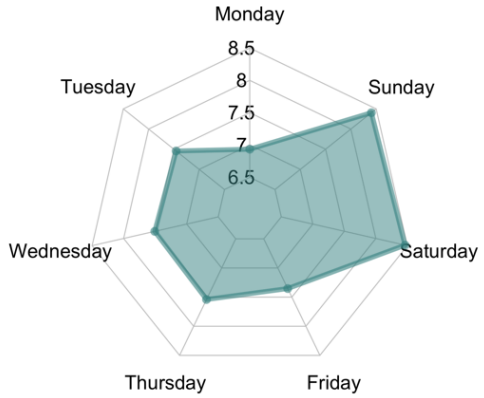


Fig. 7 Number of likes received per comment posted: A comment posted on a Monday receives on average 1.5 less likes than a comment written on a Saturday, Kolmogorov p-value = 2.1e-07

TABLE II. OVERVIEW OF COMMENTS BY COMPANY

Item (by company)	Max	Min	Mean	Median
Length of comment (characters, excluding emoticons, by company)	29k[a]	1	103.8	58
Likes received per comment	85	-	6.9	4
Dislikes received per comment	53	-	1.26	0
Number of employees that wrote at least one comment by company	394	5	87.7	55
Average number of comments per employee **who commented**, aggregated by company	19.32	1	10	10
% employees who voted but never commented aggregated by company (**writer dis-engagement**)	87.2	38.2	69.7	72.2

[a] A copy pasted comment (outlier)

D. *Anonymous forum interactions*

Table 4 shows some basic statistics about the likes and dislikes on comments appearing in the company forum.

TABLE III. OVERVIEW OF INTERACTIONS

Item (by company)	Max	Min	Mean	Median
Length of comment by company	29k*	1	103.8	58
Likes received per comment	85	-	6.9	4
Dislikes received per comment	53	-	1.26	1

In order to study the social interactions among users in the forum we build a directed network representation. We consider each employee as a node in a network. Given two employees A and B, if the employee A has liked or disliked a comment made by B, an edge linking A to B will be added to the network. The adjacency matrix from the resulting graph is used to perform (a two component) non negative matrix factorization (NMF) clustering [15-17]. We use two values as edge weight (2 matrix): (i) the total number of interactions, (ii) the number of likes divided by the total number of likes+dislikes (**agreement ratio**). In addition, we can consider the directed and undirected versions of the adjacency matrices. Thus, we have (2x2x2 combinations) eight different graph-based features whose numeric value is the NFM component of one of the clusters.

Fig. 8 shows interactions for the undirected version of the graph and feature (i). The value of the clustering is not shown. Instead nodes are colored according to churn for visualization purposes. Each company appears as a separate group.

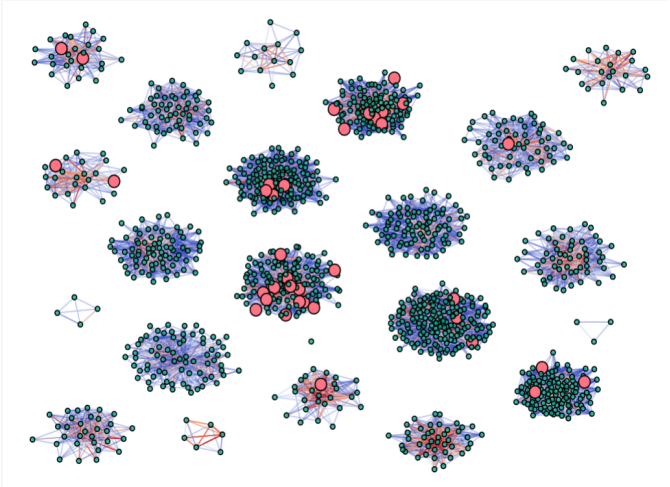


Fig. 8 Visualization of the number of interactions between employees in the anonymous forum. Each edge is colored according to its percentage of likes to dislikes. Blue means the majority of interactions are likes. Red means the majority of interactions are dislikes. Red dots are employees who churned, green dots are emloyees who did not churn. Each group is a company

E. *Turnover*

The process to detect that an existing employee has churned is done via an integration with the company ERP that updates the app backend automatically via an API. Figs. 9, 10 show churn timelines. The churn table contains 3,881 employees of whom 238 or 6% churned during the observation period.

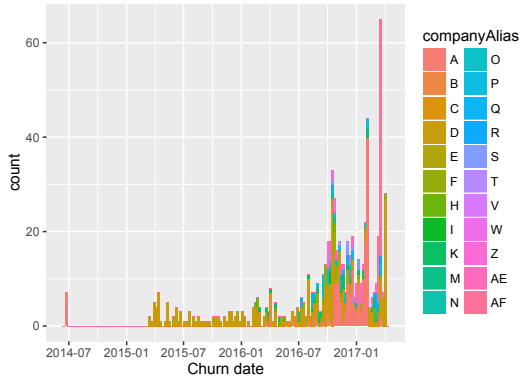


Fig. 9 Distribution of employees that churned by date

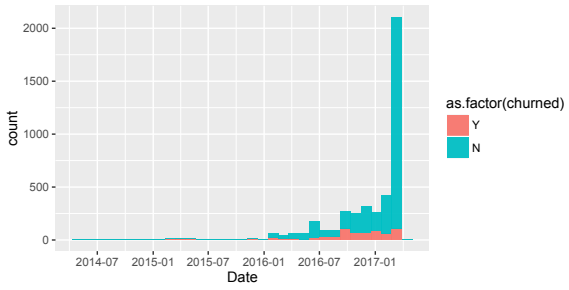


Fig. 10 Proportion of churn vs. still active users

IV. TRAINING

To model the data, we used GBM and a procedure similar to [13]. As we mentioned in the introduction, it is not our goal at this stage to optimize prediction performance. Other methods such as SVM, xgboost and blending methods [13] could be tested to incrementally improve performance. The focus here, is to find features correlated with turnover to help the company managers understand the profile of employees who churn.

We set a prediction date on 2016-12-26 (to maximize the history of usable records), 12 weeks before the data collection period ended. The objective is to rank employees by risk of churn during the next 12 weeks starting from the prediction day using past data only. The length of the period (12 weeks) is chosen arbitrarily. However, as reference we know that accuracy of similar models decreases when the period lengthens and that it increases when the period shortens [13].

For each employee, we generate three kinds of features:

A. *List of features*

- **Employee features (N=13)**
  - **happiness:** mean, standard deviation (sd),
  - **length** of employee comments: sum, mean, sd.
  - count of **comments** posted per day of observation, count of characters written per day of observation
  - likes **given** in the forum: sum, daily mean, sd.
  - count of likes + dislikes **received** by the employee's posted comments in the forum, ratio of likes to likes + dislikes.
- **Company-wide features (N=18 +34).** Same as previous but taking the mean of the company. Additional features: number of employees of company, days since last interaction, average daily usage of app by employees, average daily posts per employee, sum of total accumulated happiness votes. In addition, the company ID as a dummy variable was added (34).
- **Employee-Company features (N=4)** Employee features normalized by the corresponding company feature Example: employee mean happiness/ company mean. mean, standard deviation, comment average length, and comment posting frequency, among others, were tested.
- **Social graph features (N=8)** In addition, eight features derived from clustering the forum social graph were tested.

The target variable churn is set to 1 if the employee was active beyond the prediction date and churned in the next 12 weeks from the prediction date. It is set to 0 otherwise. The reason is to ensure that no churns that might have happened before the prediction date are considered in the test set. This situation would arise if there was a churn but the notification to the back end was delayed.

B. *Training*

Therefore, after feature building, we have a matrix of dimension 3.8k employees x 77. We train a GBM model [14]. We split the matrix in two: a train set with the odd rows, and a test set with the even rows. GBM parameters used: n.trees= 1000, shrinkage= 0.05, interaction.depth= 3, bag.fraction = 0.5, train.fraction = 0.5, cv.folds=5.

V. EVALUATION

The test set was used to evaluate the models and compare features. It contains 129 churns out of 1,944 employees. Fig. 11 shows the output of a GBM model using all the features available. The X-axis ranks the employees by risk of churn. Red color indicates that the employee churned in the 12 weeks following the prediction date; green means it did not churn. For the period considered, Precision@50 is about 80%. Table 5 shows the most influential feature in the average GBM model. Fig. 12 shows a scatter plot of F1 vs. F3 (using the test set). We can see clusters and why this feature is useful to predict churn for a decision-tree based algorithm such as GBM.

A. Likeability

We define likeability as likes/(likes+dislikes). See table 5, F1. Fig. 13., which shows the relation between likeability and churn. 20% of the employees with likeability <0.25 churn (vs. 6% in general). If a linear fit is used, every percentage point increase in this ratio will decrease the churn risk by 0.17 points. %. R-squared = 0.08, (based on the test + the train set, N=3.8k).

Fig. 14 shows the same as a smooth density plot. Here note that employees can be separated in two clusters (bulges). One cluster can be labeled as ***liked*** employees (right side of chart). Their comments receive on average a ratio of 0.92 likes per interaction. On the left side of the chart, we label the cluster of employees as ***unliked***, their comments receive on average a ratio of 0.02 likes per interaction. Employees who belong to the unliked cluster are **three** times more likely to churn than likeable ones. In addition, employees are polarized between these two clusters. There are very few employees whose comments receive a ratio between 0.1 and 0.5. For reference, for every unlike (dislike) in the forum there are 5 likes and three in four employees belong to the likeable cluster. One in four belong to the unliked cluster.

B. Comparison of feature groups

Table 6 summarizes the effect on prediction performance when a feature group is removed from the model. To compare we use Precision@50 and Precision@100. The total count of churns in the test set was 129. The best model combines all features except the company as dummy ID.

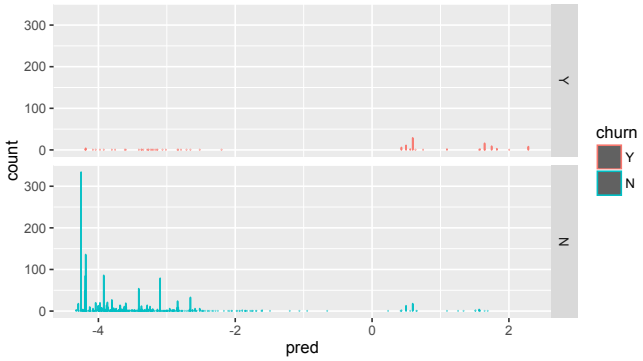


Fig. 11 X-Axis (pred) is risk of churn assigned by the model to each employee. Greater means more risk. Y-axis is count of employees in each risk bin. P@50 measured for pred = -1.0 where employess that have a pred>1 are predicted to churn by the model.

TABLE IV. RANK OF TOP FEATURES FOR A CHURN PREDICTION 3 MONTHS IN ADVANCE

F# **Name of Feature**: Description of feature	**Group**	**Influence** [a]
F1 **Likeability**: Count of likes received / (count of likes + count of dislikes) on the all the comments written by the employee in the anonymous forum.	E[b]	33
F2[c] **Posting frequency**: Average comments per day per employee. This is a company-wide feature.	C	9.6
F3 **Relative Happiness**: Employee mean happiness vote/company average happiness	EC	4.2
F4 **Relative Variability of Happiness**: Standard deviation of employee happiness / standard deviation of the company to which the employee belongs.	EC	2.4
F5 NMF Clustering of employees according to their **interactions** of the undirected graph of likes and dislike in the forum.	NFM	2.2
F6 Mean **Happiness** of the employee.	E	1.8

[a] Average of 5 runs.

[b] Can be considered an EC feature or a social graph feature.

[c] Company size feature was excluded from the table.

[d]

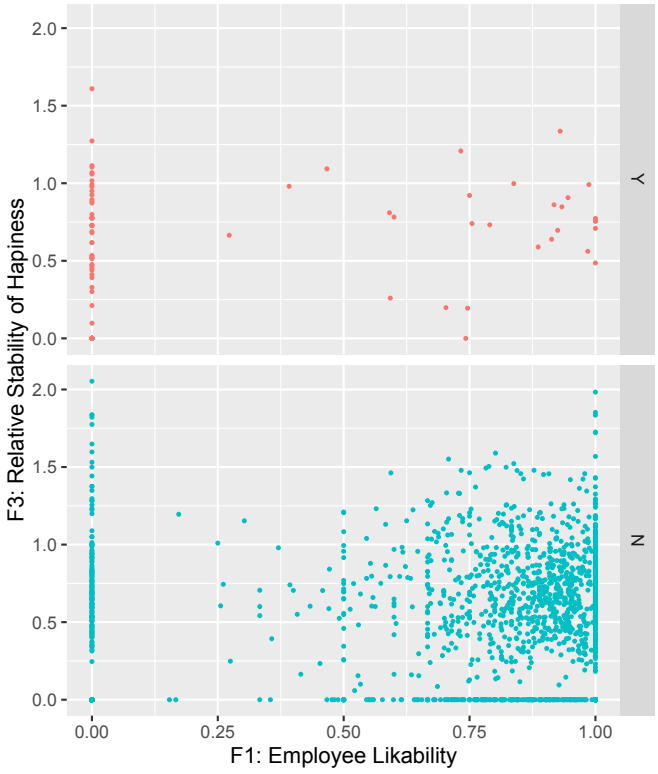


Fig. 12 Scatter plot of the two top features. Each red dot represents a churned employee. Very stable employees seldom churn, employees that receive more likes than dislikes churn less.

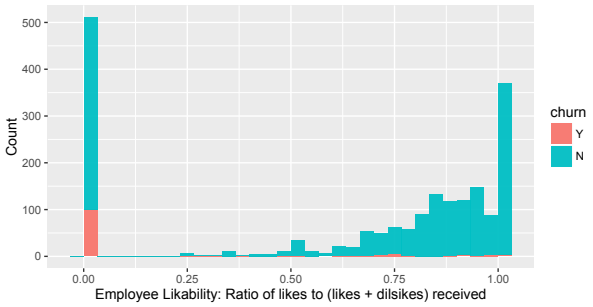


Fig. 13 Count of churn vs. likeability

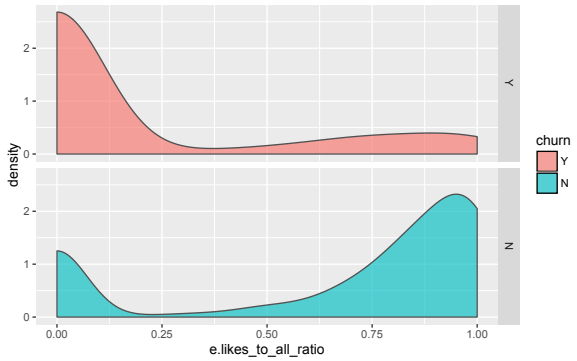


Fig. 14 Density plot of churn density vs likeability. There are two kinds of employees: The bulge on the left represents employees whose comments seldom receive a like. The bulge to the right represents employees whose comments receive on average 0.92 likes per interaction.

TABLE V. COMPARISON OF FEATURES

Features used	P@100	P@50	Change @50
E	41%	47%	-29%
E+C	53%	78%	-6%
EC+E+C	65%	80%	-4%
EC+E+C+NFM	**66%**	**84%**	-
EC+E+C+NFM+Dummy Id	70%	83%	-1%
EC+E+C +Dummy Id	48%	82%	-2%

* average 5 models.

VI. CONCLUSION

We have considered the turnover prediction problem from a learn-to-ranking point of view. We have shown that applying entity-faceted features improves the model performance by 10% to 20%, a similar improvement as reported by [12]. In addition, we have clarified how features such as comment-likes interactions are related to employee churn risk. The top four explanatory features correlated with churn likelihood are as follows:

A. *Top features*

1. **Likeability**. The ratio of likes/(likes+dislikes) received by comments posted by an employee in the forum. This is a measure of how liked are an employee's comments by their peers. Employees with a low ratio churn three times more often (F1).
2. How **often** employees post in the company forum (company average). This is a measure of how engaged the company is as a whole (F2).
3. The mean **happiness** of the employee relative to their company's mean happiness. The least happy employees in a group churn more (F3).
4. Standard **deviation** of the happiness normalized by company standard deviation of happiness: relatively more stable employees churn less (F4).

Surprisingly, a priori expected relevant features were not confirmed to be relevant. For example, we expected the **mean happiness** of the employee to be relevant based on the presumption that unhappy employees would churn more. We also expected the **positivity** of the employee (measured as ratio of likes to dislikes given to others) to be relevant based on the presumption that positive people are happier and would churn less. Finally, we expected employee individual **engagement** (measured as posting frequency) to be relevant based on the presumption that engaged employees would churn less. Neither of these features was significantly relevant to our model. An interesting result on the likeabiliy feature is that employees can be cleanly clustered in two groups according to the likes/(likes+dislike) ratio. One group churns about three times faster than the other.

B. *Entity-faceted features*

The results here confirm the effectiveness using entity-faceted features. Another advantage is that entity-faceted features make it easier for humans to interpret the results, whereas more complex approaches, such as dimensionaility reduction by performing k-means clustering at various k number that we previously used in airline prediction, yield less interpretable results.

C. *Happiness, an inside job?*

Of the top six features, only the ***mean happiness*** (ranked 6th) is a purely individual feature of the employee. The other five features are relative measures, graph based or entity-faceted features. Similarly, even though we labeled the most influential feature (***likeability***) as an individual feature (group E), it is not a feature 100% intrinsic to the employee because the likes and dislikes received are given by **peers**. In other words, the majority of the features that predict employee turnover depend on factors that are outside the control of the employee. If we hypothesize that a happy workforce is correlated with less turnover, then so should the features, which leads to the conclusion that when it comes to work, happiness is not much of *an inside job* as it is claimed by some self-help literature. It is rather significantly dependent on (social) factors outside the control of the employee.

D. *Clinical interpretation*

Finally, we have shown how to use machine learning tools to **discover** risk factors of the employees that turn over. Moreover, we can translate the Precision@50 achived by the modelto clinical test metrics. In such case we have: **Sensitivity 34%** and **Specificity 99.4%.** The predictive value of the

positive test is **80%**. The predictive value of the negative test is **96%**.

REFERENCES

[1] Bliss, W.G., 2004. Cost of employee turnover. The Advisor.

[2] Cascio, W.F., 2006. The high cost of low wages. Harvard Business Review, 84(12), p.23.

[3] Guthrie, J.P., 2001. High-involvement work practices, turnover, and productivity: Evidence from New Zealand. Academy of management Journal, 44(1), pp.180-190.

[4] Lucky, R. W. 1998. Job churn [Reflections]. In *IEEE Spectrum*, vol. 35, no. 11, pp. 17-17, Nov. 1998. DOI=: 10.1109/MSPEC.1998.730513

[5] Wai-Ho Au, K. C. C. Chan and Xin Yao, "A novel evolutionary data mining algorithm with applications to churn prediction," in IEEE Transactions on Evolutionary Computation, vol. 7, no. 6, pp. 532-545, Dec. 2003. doi: 10.1109/TEVC.2003.819264

[6] Farquad, M.A.H., Ravi, V. and Raju, S.B., 2014. Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing, 19, pp.31-40.

[7] Verbeke, W., Martens, D. and Baesens, B., 2014. Social network analysis for customer churn prediction. Applied Soft Computing, 14, pp.431-446.

[8] Ngonmang, B., Viennet, E. and Tchuente, M., 2012, August. Churn prediction in a real online social network using local community analysis. In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on (pp. 282-288). IEEE.

[9] Coussement, K., Lessmann, S. and Verstraeten, G., 2016. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems.

[10] Wei, C.P. and Chiu, I.T., 2002. Turning telecommunications call details to churn prediction: a data mining approach. Expert systems with applications, 23(2), pp.103-112.

[11] https://www.kaggle.com/c/risky-business

[12] Li, J., Arya, D., Ha-Thuc, V. and Sinha, S., 2016, August. How to Get Them a Dream Job?: Entity-Aware Features for Personalized Job Search Ranking. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 501-510). ACM.

[13] Berengueres, J. and Efimov, D., 2014. Airline new customer tier level forecasting for real-time resource allocation of a miles program. Journal Of Big Data, 1(1), p.3.

[14] Generalized Boosted Regression Models. R package version 2.0–8 Greg Ridgeway with contributions from others 2013. http://CRAN.R-project.org/package=gbm

[15] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

[16] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. Neural Computation, 19(2007), 2756-2779. http://www.csie.ntu.edu.tw/~cjlin/nmf/

[17] Bokeh Development Team (2014). Bokeh: Python library for interactive visualization. URL http://www.bokeh.pydata.org.