Data + Intuition: A Hybrid Approach to Developing Product North Star Metrics

Albert C. Chen
LinkedIn
1000 West Maude Avenue
Sunnyvale, CA, USA
abchen@linkedin.com

Xin Fu
LinkedIn
1000 West Maude Avenue
Sunnyvale, CA, USA
xfu@linkedin.com

ABSTRACT

"You make what you measure" is a familiar mantra at data-driven companies. Accordingly, companies must be careful to choose North Star metrics that create a better product. Metrics fall into two general categories: direct count metrics such as total revenue and monthly active users, and nuanced quality metrics regarding value or other aspects of the user experience. Count metrics, when used exclusively as the North Star, might inform product decisions that harm user experience. Therefore, quality metrics play an important role in product development. We present a five-step framework for developing quality metrics using a combination of machine learning and product intuition. Machine learning ensures that the metric accurately captures user experience. Product intuition makes the metric interpretable and actionable. Through a case study of the Endorsements product at LinkedIn, we illustrate the danger of optimizing exclusively for count metrics, and showcase the successful application of our framework toward developing a quality metric. We show how the new quality metric has driven significant improvements toward creating a valuable, user-first product.

Keywords

metric; endorsement; reputation system; survey

1. INTRODUCTION

It is a common practice for Web companies to evaluate their products and businesses using metrics. While this practice has brought rigor to product development, wrong metrics have the potential to mislead the business. For example, at Bing, the Microsoft-owned search engine, two key metrics, queries per user and revenue per user, actually increased when a bug degraded search relevance [12]. The team reconsidered their choice of metrics and switched to sessions per user instead.

There is a clear need for the right North Star metric to guide product development. In this paper, we present a framework to develop such metrics, and showcase its successful application for the professional networking site LinkedIn. The metric development framework generalizes quite well to multiple products at LinkedIn, so we are confident that teams in other organizations will be able to reuse or adapt it successfully to create quality metrics regarding value or other aspects of the user experience.

Our work contributes to the data science community in multiple ways:

- While most work applies machine learning toward recommender systems or prediction problems, our work represents one of the first that leverages machine learning in metric development for products.
- In addition to the usual emphasis on prediction accuracy, our approach puts a heavy emphasis on making the metric actionable and intuitive, an important concern for a metric to be adopted in product development. We show how product intuition can be used jointly with machine learning for this purpose.
- Internet companies typically collect training data from user behavior web logs for machine learning, but these do not always provide a clear signal of user experience. We demonstrate that an in-product survey can be a scalable way to collect high-quality labeled data.

The rest of this paper is organized as follows. Section 2 introduces LinkedIn Endorsements and reviews existing literature around metric development for Internet products. Section 3 proposes the metric definition framework and its application to LinkedIn Endorsements. The resulting metric and its performance are presented in Section 4. We then showcase the product impact of switching to this new North Star metric in Section 5. Finally, we end with concluding remarks in Section 6.

2. BACKGROUND AND RELATED WORK

In this section, we introduce the LinkedIn Endorsements product and review related metrics literature. We then categorize metrics into two broad categories, those measuring raw counts and those measuring quality of experience, and explain the advantages and disadvantages of each.

2.1 LinkedIn Endorsements

LinkedIn is a professional networking site with over 450 million users [16]. Users connect with other professionals,

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054199

share content, and participate in discussions. They can apply to jobs and create profiles to represent themselves to recruiters and other people.

The Endorsements product touches upon both the social and hiring aspects of LinkedIn. The vision of Endorsements is to build the largest and most trusted professional peer validation system. Endorsements allow users to vouch for the expertise of other users. After LinkedIn users add skills (e.g. 'Data Mining' or 'Java') to their profile, their connections can endorse them for those skills. These endorsements are displayed alongside the skill and serve as social validation of the user's expertise. Endorsements are used as one input to rank search results for skill experts, so that recruiters can find candidates with the right skills [8].

The Endorsements product was first introduced to LinkedIn in 2012, and has been heavily used by LinkedIn users since inception. Today there are more than 10 billion total endorsements, with an average of 43 million new endorsements given each week.

This work was motivated by the realization that not all endorsements are equally valuable. For example, an endorsement from an adviser or colleague is likely more significant than an endorsement from a social acquaintance. We wanted to identify the ones which best serve the product vision.

2.2 Literature on what constitutes a good endorsement

While the Endorsements product is unique to LinkedIn, there is literature around related products such as reviews and (up/down) votes. These products have important differences (e.g. endorsements are always positive), but there are enough similarities to learn from existing work. Reviews, like endorsements, are public assessments of an entity. They usually include both a quantitative rating and qualitative text, and are widely used in online marketplaces. Many review sites allow customers to mark reviews as helpful or not helpful. This allows them to determine the most helpful reviews and prioritize which ones to show. One study found that review rating extremity, review depth, and product type affect perceived review helpfulness on Amazon [17]. Another study examined how the reviewer's number of reviews and number of friends affect the perceived credibility of a Yelp review [15]. As of 2013, Yelp classified about 75% of its reviews as 'Recommended Reviews' based on "quality, reliability, and user activity" [21].

Yet endorsements are not entirely like reviews; they are much more lightweight and other users cannot comment on whether or not they are helpful. In that sense, endorsements are more similar to the upvotes on question and answer sites like Quora and StackOverflow, which allow a users to rate the helpfulness of questions and answers. Quora's ranking of answers to display depends upon "upvotes and downvotes on the answer, in part based on how trustworthy voters are in a given topic" [18]. The common theme across reviews and upvotes is that the credibility of the reviewer or voter determines the value of their opinion.

2.3 Literature on using metrics to guide product decisions

Metrics serve multiple goals in a product organization. First, they represent the North Star, i.e. the success of the business, such that optimizing for those metrics leads product teams toward the business objectives. They rally a team around a clear target that they can hold themselves accountable to. Secondly, many of these metrics are reported to internal and external stakeholders as a measure of product performance. Lastly, these metrics are often monitored to identify issues in the product, such as server outages (short term metric move) or secular changes in user needs (long term metric move), and to evaluate product changes through controlled experiments [12].

In general, we can divide metrics into two broad classes, depending on what they intend to measure and the complexity of how they are defined.

2.3.1 Measuring volume - count metrics

Many articles on professional sites such as LinkedIn and Medium discuss how companies define their North Star metrics to represent the overall success of their products (e.g. [1]). Most start with some kind of user engagement metrics. These include the number of active users, page views, and time spent on a web site [13]. Metrics based on active users, such as daily active users (DAU) and monthly active users (MAU), are commonly used and reflect the most basic way of measuring user engagement. For instance, Facebook monitors MAU internally and reports it externally [4]. Engagement metrics are easy to track and give a good baseline of overall performance, but a common problem is that raw activity doesn't necessarily translate to the quality of the user experience or the business objectives. For this reason, some have declared these to be "vanity metrics" [20]. In response, some companies measure more valuable indicators of engagement, such as "messages sent" for Whatsapp and "nights booked" for AirBnB [4]. Others propose to measure "stickiness," which is the ratio of DAU to MAU [3].

2.3.2 Measuring experience - quality metrics

When the construct to be measured is more complex or subjective, simple aggregation (e.g. count, ratio) of user actions may be insufficient. For example, the Web Search community often wants to measure search success or searcher satisfaction. Neither of these can be reliably defined by counting intuitive actions like search queries or result clicks; for example, in the case of "good abandonment," the searcher finds the information on the search result page and has no need to reformulate the query or click a result [12, 14]. To better predict search success, researchers resort to the combination of multiple signals, such as result clicks, dwell time, and query reformulation. They leverage predictive modeling techniques to characterize the relationship between these signals and the overall success criterion. For example, Hassan et al. predicted searcher satisfaction from clicks, query reformulation, and inter-query time [9].

An important difference between our work and most of the above literature is that the ultimate goal of our work is to create a metric that can be used to drive product development, while the focus of most of the cited work here (e.g. [9]) is on prediction accuracy. To drive product development, the metric must be intuitive and actionable [10]. Product teams need to understand how improvement of each product feature relates to the overall metric movement. In practice, given the black-box structure of most machine learning models (e.g. logistic regression predicting the log-odds, instead of the raw outcome variable), they should not be used directly as metrics. So extra work is needed to transform measurable signals and their relative importance learned from the

model into metrics that are intuitive and actionable. We found only one work that took a modeling-based approach to product metric development [5]. The authors argued that the metrics should be "debuggable" and "decomposable." To do so, they suggested that "we can first have some machine-learned models, and then we simplify and hand-tune the machine-learned models into clear, human-readable models for online metrics." However they did not provide details on how they hand-tuned the model or its performance. Our work presents a framework for how this can be done.

Our work expands upon the modeling-based work reviewed in this section by presenting a concrete framework for developing accurate, intuitive, and actionable metrics that capture user experience. The framework shows how data insights and product intuition play complementary roles in this process. It has been applied to multiple products at LinkedIn to develop metrics that aim to measure quality; we illustrate the approach in detail as applied to LinkedIn Endorsements.

3. METRIC DEFINITION FRAMEWORK

Here we outline a modeling-based approach to developing product North Star metrics. There are five steps in the framework:

1. Collect labeled data on the True North measure for success (e.g. quality of user experience)
2. Identify signals that could be predictive of that True North success measure
3. Apply machine learning to select the most predictive features
4. Develop 2-3 accurate and intuitive metrics based on the top features
5. Select a winning metric definition using product intuition

In the rest of this section, we will describe each step in detail and how we applied the framework to the case of LinkedIn Endorsements. Since the goal of LinkedIn Endorsements is to validate user's skills, the North Star metric should measure the endorsements that serve this purpose. We call these the Quality Endorsements. We present the results in Section 4.

3.1 Collect labeled data

The first step is to collect labeled data that measures the True North success of the product (e.g. quality of user experience). In our example, these are the Quality Endorsements.

In the best case, behavioral web log data can be used. This provides the largest amount of data and the data are readily available for analysis. For example, suppose the metric aims to determine whether an action is valuable. Then we could label the data based on whether the action results in a certain desired outcome. The outcome may be realized over time or depend upon the reaction of others.

In many cases, clickstream data are unable to measure quality, so a survey may be appropriate. Although smaller in scale, surveys can still reach a sizable number of users. Furthermore, the responses provide a more direct signal than inferences made from user behavior logs. Related to this approach is crowdsourcing, in which a third party labels data by applying human understanding [6].

A final category, user research, is typically more qualitative and covers a small group of users. In this approach, users are interviewed individually or in focus groups to understand their thoughts about the product, and to observe their interaction with it. This is helpful for building product intuition. However, metrics based on user research alone are difficult to justify because they rely on the experiences of only a few users.

In our case, we want to know which endorsements serve the purpose of validating a LinkedIn user's skills. We believe it is difficult to infer quality from clickstream logs. We could try to measure it based on whether an endorsement leads to more interest from recruiters. However, from discussions with our recruiter colleagues at LinkedIn, the presence of an endorsement is a small part of the decision, whereas past experience bears more weight. Another approach could be to check how the endorsement recipient reacts to the endorsement; in particular, the recipient can decide not to show the endorsement on his profile. However, very few endorsements are hidden, even though users complain about their endorsements not being valuable. So neither recruiter interest nor hidden endorsements is a good signal to generate training data.

We therefore took the survey approach, asking endorsement recipients one of two questions when they are notified about receiving the endorsement. We showed this survey to users of LinkedIn's iOS mobile app in the Notifications tab (Figure 1).

- Q1: How familiar is [endorser] with your skill in [skill name]?

  Not familiar, Slightly familiar, Somewhat familiar, Moderately familiar, Extremely familiar

- Q2: Does [endorser]'s endorsement improve your reputation for [skill name]?

  Strongly disagree, Disagree, Neutral, Agree, Strongly agree

We designed the questions to capture different aspects of what makes an endorsement valuable. The first question distinguishes valid endorsements from social ones. The endorser should be able to assess the recipient's skill level well enough to give a lightweight recommendation. The second question asks which endorsements satisfy the recipient's goal to improve their reputation. We took care to be specific in our wording so that the questions are clear, explicitly stating both the endorser's name and the skill name. We used the familiar five-point rating scale with standard Likert scale choices (visible upon tapping on stars).

We decided to survey the recipients because they have the best context on who the endorser is and why they gave the endorsement. The questions are designed so that a five-star endorsement is one that the profile viewer (e.g. recruiter or other user) would trust and find valuable. Thus, we capture the perspectives of recipient and viewer, the ones directly served by the goal of endorsements to provide meaningful skill validation.

Users were randomly selected for the survey and shown one of the questions on 10% of their endorsement notifica-

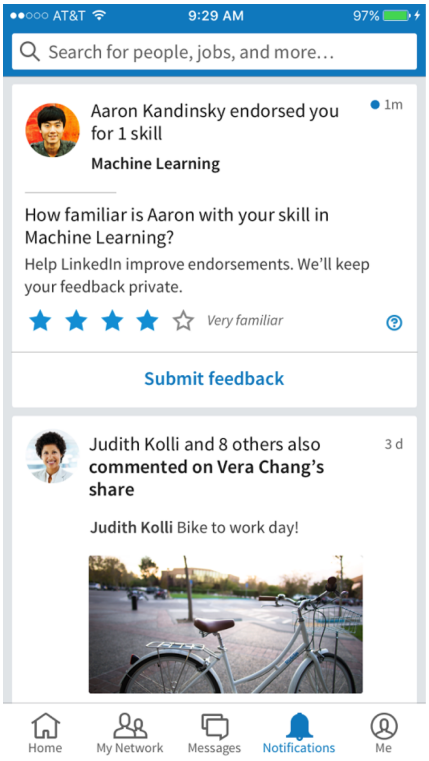

Figure 1: In-App Survey

Table 1: Volume of survey responses. For the training set, we used responses from the first 12 days. The validation set consists of responses from the last 6 days.

	Q1	Q2
Training Set	11,138	9,359
Validation Set	5,510	4,556

tions. We limited the survey frequency so as not to overwhelm our users. For the purposes of metric development, we collected survey data over 18 days, for a total of 30,563 responses from 25,422 users. The responses collected are shown in Table 1.

3.2 Identify broad set of relevant signals

With data in hand, the next step is to generalize the metric. The goal is to identify all the quality endorsements without surveying all recipients. To do this, we need signals indicative of quality that are known at the time an endorsement is given. In this step, we identify as many relevant signals as possible. Not all will be used in the final metric definition.

By discussing with the product experts, we identified a total of 84 signals, broadly falling into three categories: properties about the endorser, about the recipient, and about their relationship (Table 2). For example, quality may depend on how the endorser and recipient know each other and the endorser's years of experience.

Table 2: Candidate signals for defining the metric.

Category	Example Signals	# of signals
Endorser	Endorsements given, # of skills, seniority, years in career	39
Recipient	Endorsements given, # of skills, seniority, years in career	21
Overlap	Same company, education, region	23
Endorsement	Whether recipient has skill on profile	1

3.3 Apply machine learning to identify top signals

Not all of the signals should be part of the definition. Some may not be useful, and a metric with too many signals becomes hard to grasp. Taking a data-driven approach, machine learning can be applied to rigorously narrow down the set of signals to the ones most predictive of the labeled data.

To prepare the survey responses for modeling, 1-3 star endorsements were considered 'not quality' and 5-star were considered 'quality.' We discarded 4-star endorsements as a neutral buffer between clearly positive and negative signal, akin to the neutral 'passives' group in the calculation of Net Promoter Score [11]. We used the first 12 days of responses as the training set, leaving the last six for validation.

We built two different models on the training data to make feature selection more robust: boosted trees (nonparametric) and L1-regularized logistic regression (parametric) [19, 7]. Because the goal is to use the features to construct a simple definition, it is best to use basic models to assess main effects and low order interactions. Boosted trees allowed us to easily handle categorical variables and capture interactions. For this model, we used a low interaction depth and enforced numeric variables to be monotonically increasing with the log-odds. Feature importance was assessed by contribution to reducing the loss function. For logistic regression, L1-regularization gave a natural approach for feature section. We left out interaction terms to capture only the main effects. We standardized numeric, non-binary features, and assessed feature importance by the coefficient magnitudes. The top features of each model were selected, looking for a natural cutoff in feature importance. As a result, this narrowed our original set of 84 features down to 12 (detailed results in Section 4).

This step illustrates the advantages of combining product intuition with data. It is easy for the product experts to come up with a list of relevant signals, but difficult for them to know which are most important. People may have different thoughts as to which signals should be used to define the final metric, so machine learning eliminates the need for speculation.

3.4 Propose candidate definitions using top signals

After finding the top signals, the next step is to craft candidate metrics. The goal is to find 2-3 intuitive metrics that

match the sensitivity and accuracy of the machine-learned model. This is an iterative process in which both product intuition and data come into play. Product intuition informs how the top signals could be combined into sensible definitions. The precision and recall of candidate definitions are then checked against the data to evaluate their success. By exploring various candidate definitions and checking their performance against the data, we identify a few potential definitions for the metric that are intuitive and accurate.

For Endorsements, product intuition was guided by user research. From interviewing users, we learned that the endorser's reputation and relationship to the recipient affect how the endorsement is perceived. Machine learning confirmed these perspectives, and identified the most useful signals about the endorser's expertise and the relationship between endorser and recipient. With these categories of 'knowing the skill' and 'knowing the person' in mind, we created over 30 sensible definitions with varying compositions and conditions on the top signals. For example, a quality endorsement could be one where the boolean signals 'A' and 'B' are both true, and numeric signal 'C' is above a certain cutoff. After crafting these candidates, we evaluated their precision and recall against the training data. We then evaluated the best ones against the validation data to guard against overfitting and measure final performance.

Evaluating candidate definitions in this way helped narrow down to a few of the best ones. The result was a set of three candidate definitions of Quality Endorsements, all of which had reasonable performance from a model evaluation perspective. Section 4 describes the candidate definitions and evaluation process in detail.

3.5 Pick winning definition using human judgment

The final step of the framework is to apply human judgment to pick the winning definition. There will be some tradeoffs among the candidate definitions, but by design, they are all sufficiently accurate, intuitive, and actionable. So this step comes down to human judgment. For Endorsements, we presented the product team with all three variants of the Quality Endorsements metric along with their predictive performance against the validation data. By letting the team choose the final metric, we secured buy-in from the stakeholders who will be using and relying on the metric.

4. RESULTS AND DISCUSSION

In this section, we first present the machine-learned models that we trained in Step 3 of the framework and evaluate their predictive performance on the validation data. Then we describe the three candidate definitions developed based on the models, and compare them against two baseline definitions.

4.1 Model performance

We trained a nonparametric and a parametric model for each question: boosted trees and L1-regularized logistic regression. The validation set AUC was 0.68-0.71 for Q1 ("familiarity") and 0.56-0.58 for Q2 ("reputation"). Understandably, perceived impact of endorsements on reputation is harder to predict with our features. Whether a user thinks an endorsement improves their reputation depends not only on the endorser and skill, but also on the user's overall perception of the Endorsements product. Nevertheless, we did identify a few useful signals. From Q1, top signals included the endorser's skill expertise score and endorser-recipient overlap at a company, education, or industry. For Q2, endorsements from large company managers, senior leaders, and highly endorsed users tended to receive 5-star ratings. In total, this narrowed our original set of 84 features down to 12. Due to business sensitivity, we do not report the full list of features.

4.2 Comparison of definitions

In this section, we compare the three candidate definitions from our hybrid framework to baseline definitions that rely solely on product intuition or solely on machine learning.

Based on the model output, we constructed three candidate definitions of Quality Endorsements, each of which included the top signals identified from the models. We compared them with two baseline definitions. The first baseline definition is effectively what the team used before this work took place: treating every endorsement as quality (B1). The second baseline definition is based solely on product intuition (B2): we discussed with the product experts before surveying users and hypothesized that a quality endorsement is one given by a coworker, classmate, or senior-level user, who is a top expert in the skill area. M1, M2, and M3 are the final three candidates crafted through the metric development framework. They each include the components of knowing the person and knowing the skill. But they vary in strictness of what it means to satisfy each condition. To illustrate how we adjusted strictness, suppose we define M1 as the endorsements given by coworkers that are in the top 20% at the skill, according to some measure of their skill level. We can make the 'knowing person' requirement tighter by also requiring the endorser and recipient to have the same title within the company. We can make the 'knowing skill' requirement tighter by taking only the top 10% of skill experts.

- B1: all endorsements are quality
- B2: product intuition alone
- M1: top signals for knowing skill and knowing person
- M2: M1 with tighter requirement for knowing person
- M3: M1 with tighter requirement for knowing skill

We compared the precision and recall of each of these definitions, evaluated against the validation data from the survey. Table 3 summarizes the results. Out of business sensitivity, we present the relative precision and recall of each definition, rather than absolute values. For example, on question 1, definition M1 attained precision that was 4.8 percentage points (pp) higher than B1 and recall 13pp higher than B2.

The survey responses helped inform definition construction toward one with high precision and recall. M1 far outperforms B2, with similar precision but over 11pp greater recall. By using a tighter requirement for knowing the person, M2 attains the highest precision on Q1. M3 has higher precision and recall than the baseline B2 for both questions. These definitions from our hybrid approach outperform a definition based on product intuition alone. The results show that intuition-based definition B2 was far too narrow in what it considered to be quality.

Table 3: Comparison between baseline and model-informed definitions. Precision gain shown as percentage point increase relative to B1. Recall shown as percentage point increase relative to B2.

	Precision (pp)		Recall (pp)	
Definition	Q1	Q2	Q1	Q2
B1 (all quality)	0.0	0.0	-	-
B2 (intuition)	4.1	4.3	0.0	0.0
M1 ('person' and 'skill')	4.8	3.7	13.0	11.8
M2 (stricter 'person')	6.3	3.5	9.3	7.7
M3 (stricter 'skill')	5.2	4.7	4.2	3.9

Table 4: Precision of final definition and machine-learned models (percentage point increase relative to B1). Final definition achieves similar level of precision to machine-learned models. Models evaluated at the same level of recall as the final definition.

Question	Boosted Trees	Logistic Regression	Final Choice
Q1	5.6	6.6	4.8
Q2	3.0	3.4	3.7

After seeing these results, the product team chose M1 as the final definition over M2 and M3 because it achieves the highest recall while maintaining high precision. Comparatively, the slight increase in Q1 precision by M2 and Q2 precision by M3 is not worth the sharp loss in recall. From a product standpoint, the goal is to increase the volume of quality endorsements in the system so that members receive useful validation of their skills. By selecting a definition with high recall, we optimize for a wide net of what users consider to be quality. An overly narrow definition like M3 would encourage the team to pursue the very best endorsements, while missing many others that users consider valuable. We realized that our goal to improve Endorsements is best served by having more good endorsements in the system.

Besides improving upon product intuition, the Quality Endorsement definition from our hybrid framework achieves similar levels of precision to the machine-learned models while being easily interpretable and intuitive (Table 4). The definition has slightly lower precision for Q1 and slightly higher precision for Q2 when compared with the models at the same level of recall. Furthermore, the definition uses only five features, far fewer than the black-box models, and it follows the intuitive framework of knowing the person and the skill (Table 5).

4.3 Discussion

Our metric improves upon intuition alone (Definition B2) or machine learning alone. If we had simply taken the intuition-based definition, we would have been too strict on what constitutes quality, with much lower recall. While B2 applied the same framework of knowing the skill and knowing the person, it sets the bar too high. Furthermore, the data pointed out how B2 could be improved; we had supposed that all endorsements from high seniority users are equally valuable. However, we learned that endorsements from senior users in large companies are valued more than those from senior users in small companies. More fundamentally, even if an intuition-based metric happens to be accurate, it is difficult to trust without data to back it up.

Although it was possible to use the machine-learned model directly for an accurate and sensitive metric, this type of metric is a black-box and is difficult to use to suggest actionable product decisions. For example, with logistic regression, a linear combination of many features is used to model the log-odds. There are many components to the model and no direct interpretation of how the decision boundary is defined. With an ensemble of trees, each classification is based on the output of many hundreds of decision trees. Explaining what makes one endorsement quality and another not becomes hard. This is why we use machine learning for feature selection but not for the final metric. Our final metric can be easily visualized as a single decision tree, and the criteria for quality are clear. This way, the metric is not only clear but also actionable, leading to direct product interpretation about which endorsements matter most. By combining machine learning with product intuition, the result is an accurate and sensitive metric that is easy to communicate and use.

5. PRODUCT IMPACT

In a data-driven organization, it is important to have the right metrics to create the right product. In this section, we show how optimizing for total endorsements influenced the existing Endorsements product, and how the new Quality Endorsement metric is driving changes in the right direction.

5.1 Past Metrics Drove the Wrong Goal

When Endorsements was introduced in 2012, the North Star metric was total endorsements given. Secondary metrics included unique endorsers and unique recipients. Each of these is a simple count metric. For a new feature, it is natural to measure success by user engagement. Greater usage implies that users like the feature; in our case, more endorsements means that more social validation is available in the system. Furthermore, the total endorsements metric is easy to compute directly from user action logs. The simplicity and intuitiveness made it the right metric at product launch.

Endorsements gained traction with our users very quickly. To help increase usage, LinkedIn suggests endorsements for a user to give to his connections. For example, when a user views a connection's profile, he might be shown a list of five skills to endorse (Figure 2). He is not required to have those five skills on his own profile. In our user research studies, users expressed that they would not have thought to endorse for some skills, but do so anyway after being prompted. These types of promos were poorly targeted and made it too easy to endorse in bulk, thus devaluing the meaning of an endorsement. While the total endorsements metric certainly increased as a result of these promos, the value of the Endorsements product did not necessarily improve.

The total endorsements metric also affected product design. When displaying skills on someone's profile, we showed the total endorsements received, the skill name, and a list of profile pictures of all the endorsers (Figure 3). This format emphasized the total number rather than the context of each endorsement. In user research, we found that people

Table 5: Comparison of definition complexity. Showing number of features and how they are combined to make a prediction. Final definition is more intuitive than the black-box models, with fewer features and easy visualization as a single decision tree.

Complexity	Boosted Trees	Logistic Regression	Final Choice
Description	Ensemble of 900+ trees	Linear combination for the log-odds	Single decision tree
Q1 # features	81	33	5
Q2 # features	79	29	5

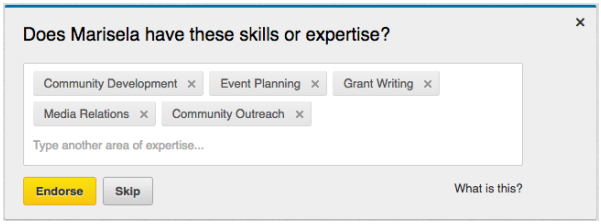

Figure 2: Existing suggested endorsement prompt encouraged high volume.

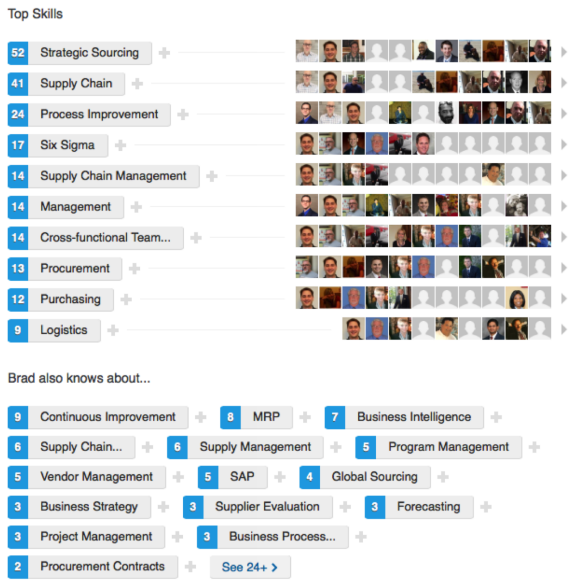

Figure 3: Existing endorsements display on the profile was difficult to parse and emphasized endorsement volume rather than meaningful endorsements.

viewing the skill section wanted to see who the endorsers are, which is time-consuming to do in the existing design.

While endorsements are meant to be lightweight recommendations, interviews with users revealed that they sometimes endorse for other reasons: to return an endorsement, to be endorsed by others, to keep in touch with a connection, or to help out a friend. Often they are not qualified to validate the user's expertise. In contrast, the recipient is concerned about building a strong reputation, and the profile viewer (e.g. hiring manager) is interested in making an evaluation.

Although total endorsements was the right metric at product launch, it became a misleading metric over time. Focus on a misleading metric blinded us to a user experience that drifted away from the main purpose to validate user's skills and provide the viewer a way to assess expertise. Users expressed skepticism to trust endorsements as a measure of expertise, because it was hard to tell the signal from the noise.

Besides total endorsements, we also monitored click through rate (CTR) on suggested endorsements. It would seem that optimizing for high CTR would serve the product vision, because we suggest the endorsements users would like to give. However, this is misleading because, as we learned, users endorse for many reasons. CTR focuses on the endorser without capturing the value to the recipient or the profile viewer.

5.2 New Product Direction

Feedback from our users motivated the need for a new metric that aligns with our vision to build the largest and most trusted professional peer validation system. The Quality Endorsement metric has shaped how our team approaches the product.

For example, the suggested endorsement models have been rebuilt to optimize for Quality Endorsements rather than just the CTR. In addition to updating our models on the backend, we changed how we present the suggestions to users. We explain the reason for the suggestion in the context of the endorser's skill and relationship with the recipient (Figure 4). Our A/B tests indicate that these changes increased Quality Endorsements given by over 50%.

In addition, we have changed the presentation of endorsements on user profiles to focus less on the total number given, and instead showcase the Quality Endorsements (Figure 5). This comes in the form of highlights that show key insights into the endorsers who have vouched for the user's skill. For example, we showcase endorsements from top skill experts, from senior leaders, and from people the profile viewer is connected with. These highlights help profile viewers to draw quick insights from the skills section, giving them a clearer signal for evaluating expertise.

The new Endorsements experience is now available on both mobile and desktop. As a result of these improvements, survey responses from members have improved noticeably over the last eight months (Figure 6). In absolute terms, the percentage of 5-star responses have increased by nearly 5 percentage points, while 1-star responses decreased by around 1pp. Based on this member feedback and plenty of positive press (e.g. [2]), we are confident that our product changes have made endorsements increasingly valuable, and

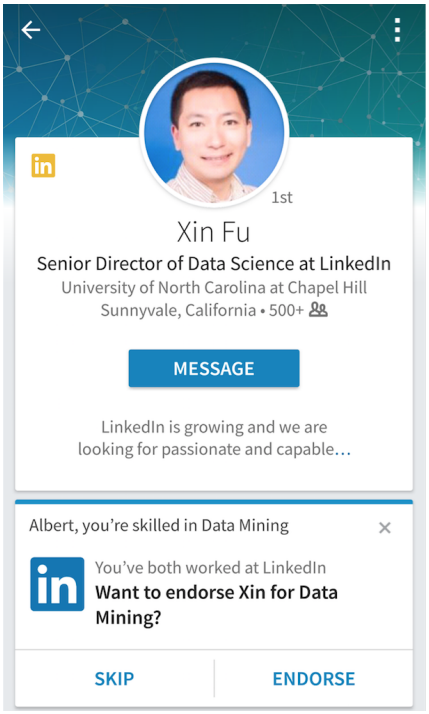

Figure 4: New suggested endorsement prompt presents the suggestion in context of skill expertise and the relationship between users.

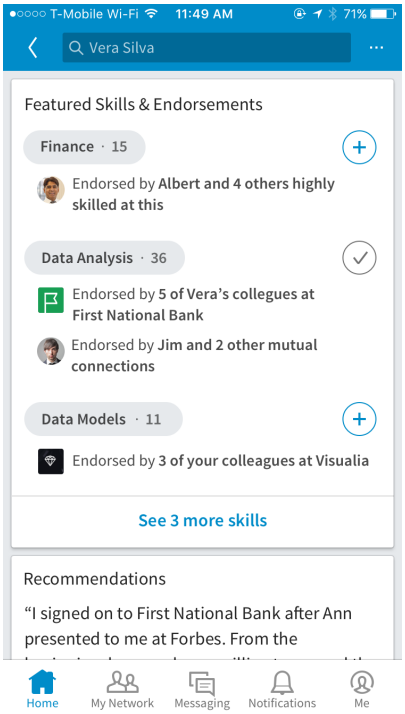

Figure 5: New endorsement display on the profile highlights the meaningful endorsements. Each highlight can be clicked for details about the endorsers.

that we have a good North Star metric to guide continued improvements.

6. CONCLUSIONS

Successful data-driven product development depends upon measuring and improving the right metrics. Simple count metrics are not always best, nor should a metric be defined using product intuition alone. In this paper, we presented a framework for developing North Star metrics that align with the product vision, through using a combination of machine learning and product intuition. We applied this framework to the specific example of LinkedIn Endorsements and showed how two metrics, total endorsements and Quality Endorsements, led the product in different directions. To create the metric, we leveraged an in-product survey for training data after realizing that behavioral web logs do not capture the value of an endorsement. The new metric is similar in precision to the black-box machine-learned models, but additionally places a heavy emphasis on interpretability and actionability.

As we demonstrate, the two perspectives of data and intuition complement each other well. In our framework, product intuition plays the role of guiding the metric development process, and data acts as the measuring stick to inform the product team's choices. Specifically, product vision sets the scope of what to measure and identifies candidate signals. Then, machine learning selects the best signals to include in the metric definition. The final steps in the process are an exercise of combining product intuition and data, through developing sensible definitions and choosing one with high accuracy. Just as in the entire product development lifecycle, the partnership of product and data teams is key to creating North Star metrics.

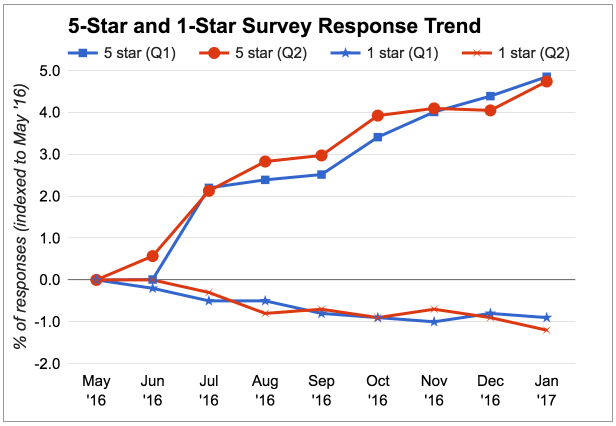

Figure 6: Survey responses over time (% of total responses, reported as percentage point difference relative to May 2016). The first product changes were introduced starting June 2016. Since then, 5-star responses increased by nearly 5pp, while 1-star responses by nearly 1pp.

We hope other organizations will take the lessons learned from LinkedIn Endorsements and seek a more comprehensive and nuanced perspective on quality beyond simple engagement metrics, and utilize our metric development framework to create meaningful measures of success.

7. ACKNOWLEDGMENTS

The authors would like to thank the Endorsements team for recognizing the need for a new metric, supporting the idea of an in-product survey, and executing to improve the product. Hari Srinivasan and Yolanda Yeh drove the product vision from start to finish. Jaewon Yang and How Jing improved the suggested endorsement relevance models to optimize for quality. Maria Iu created the new design that highlights quality endorsements. Nicole Lee led the user research that informed product intuition. Victor Kabdebon, Joey Bai, Jie Zhang, Rick Ramirez, and Kyle Brickman led the engineering efforts, including the survey implementation. We would also like to thank the Analytics team at LinkedIn for their helpful feedback on the metric definition framework, and for applying the framework to other products at LinkedIn.

8. REFERENCES

[1] Aim4Global. 9 types of metrics you need to track for your product's success. http://goo.gl/1ykRhM, March 2016.

[2] Bain and Company. Measuring your net promoter score. http://netpromotersystem.com/about/measuring-your-net-promoter-score.aspx.

[3] P. Boyce. 5 awesome user engagement metrics for growth. http://blog.popcornmetrics.com/5-user-engagement-metrics-for-growth/, June 2015.

[4] Z. Bulygo. Facebook's vp of growth gives you tips on growing your product. https://blog.kissmetrics.com/alex-schultz-growth/, February 2015.

[5] A. Deng and X. Shi. Data-driven metric development for online controlled experiments: Seven lessons learned. KDD 2016.

[6] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, Apr. 2011.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[8] V. Ha-Thuc, G. Venkataraman, M. Rodriguez, S. Sinha, S. Sundaram, and L. Guo. Personalized expertise search at linkedin. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 1238–1247. IEEE, 2015.

[9] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2019–2028. ACM, 2013.

[10] P. Hill. Translating business goals to specific objectives and kpis. http://www.smartinsights.com/goal-setting-evaluation/goals-kpis/translating-business-goals-kpis-aa-03/, October 2014.

[11] A. Hutchinson. Linkedin unveils improved endorsements system, continues to refine data accuracy. http://www.socialmediatoday.com/social-networks/linkedin-unveils-improved-endorsements-system-continues-refine-data-accuracy.

[12] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*, pages 786–794. ACM, 2012.

[13] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret. Models of user engagement. http://ir.dcs.gla.ac.uk/ mounia/Papers/umap_CRC.pdf, April 2012.

[14] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–50. ACM, 2009.

[15] Y.-s. Lim and B. Van Der Heide. Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on yelp. *Journal of Computer-Mediated Communication*, 20(1):67–82, 2015.

[16] LinkedIn. About linkedin. https://press.linkedin.com/about-linkedin, July 2016.

[17] S. M. Mudambi and D. Schuff. What makes a helpful review? a study of customer reviews on amazon. com. *MIS quarterly*, 34(1):185–200, 2010.

[18] Quora. How does the ranking of answers on quora work? https://www.quora.com/How-does-the-ranking-of-answers-on-Quora-work/, June 2015.

[19] G. Ridgeway. Generalized boosted models: A guide to the gbm package, 2005.

[20] E. Schonfeld. Don't be fooled by vanity metrics. https://techcrunch.com/2011/07/30/vanity-metrics/, July 2011.

[21] Yelp. Why does yelp recommend reviews? https://www.youtube.com/watch?v=PniMEnM89iY, November 2013.