Future User Engagement Prediction and Its Application to Improve the Sensitivity of Online Experiments

Alexey Drutsa
Yandex
Moscow, Russia
adrutsa@yandex.ru

Gleb Gusev
Yandex
Moscow, Russia
gleb57@yandex-team.ru

Pavel Serdyukov
Yandex
Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

Modern Internet companies improve their services by means of data-driven decisions that are based on online controlled experiments (also known as A/B tests). To run *more* online controlled experiments and to get statistically significant results *faster* are the emerging needs for these companies. The main way to achieve these goals is to improve the sensitivity of A/B experiments. We propose a novel approach to improve the sensitivity of user engagement metrics (that are widely used in A/B tests) by utilizing prediction of the future behavior of an individual user. This problem of prediction of the exact value of a user engagement metric is also novel and is studied in our work. We demonstrate the effectiveness of our sensitivity improvement approach on several real online experiments run at Yandex. Especially, we show how it can be used to detect the treatment effect of an A/B test faster with the same level of statistical significance.

Categories and Subject Descriptions: H.1.2 [User/Machine Systems]: Human information processing; H.5.2 [User interface]: Evaluation/methodology

General Terms: Measurement, Experimentation

Keywords: User engagement; engagement prediction; online controlled experiment; sensitivity; quality metrics

1. INTRODUCTION

Online controlled experiments, or A/B testing, have become the state-of-the-art technique to improve web services based on data-driven decisions [16]. It is utilized by many web companies, e.g., web search engines such as Bing [6, 16] and Google [28], social networks like Facebook [1], etc. The largest ones have designed special experimental platforms that allow them to run A/B tests at large scale (e.g., hundreds of concurrent experiments per day) [28, 15].

An A/B test compares two variants of a service[1] at a time by exposing them to two user groups and by measuring the difference between them in terms of a key metric (e.g., the revenue, the number of visits, etc.), also known as an overall evaluation criterion [18]. The ability of the metric to detect the statistically significant difference when the treatment effect exists is referred to as the *sensitivity* of the test. Improvement of the sensitivity is quite challenging, because the better it is, the smaller changes of the service could be detected by an experiment.

The straightforward way to improve the sensitivity, or the *power*, of a controlled experiment is to increase the amount of the observed statistical data that could be done either by increasing the population of users, participating in the experiment, or by conducting the experiment for a sufficient period of time. Both approaches have disadvantages. First, the amount of data is upper-bounded by the available web service traffic, and the growth of traffic per experiment reduces the number of experiments conducted per year [28, 7]. Second, increasing the size of a treatment group in the treatment group is always not desirable, as any treatment may have a negative effect[2]. Third, the sooner an experiment finishes, the sooner the treatment variant of the system is shipped or is returned for rework. So, the described ways of improving the sensitivity of an A/B test are not preferred, because they are contrary to the main purposes of an experimentation platform: to run many experiments and to get their results fast [28].

The power of a controlled experiment depends on the variance of the key metric used in the test [18], hence, a variance reduction could serve as another way to improve the sensitivity. The basic examples of such approaches are discussed in [18]: the use of a key metric with lower variance and the elimination of users who were not affected by the service change in the treatment group[3]. The applicability of these approaches is limited and strongly depends on the case of a conducted experiment. Namely, an experimental platform usually has a standardized success criteria, which utilizes a small set of key metrics to make the final decision on the treatment variant of the service [15]. These metrics are usually selected with respect to some business-related criteria of a considered service and are aligned with its long-term goals (like the number of sessions per user for a search engine [14]). Hence, finding an alternative for them is non-trivial

[1] e.g., a current version of the service (the control variant) and a new one (the treatment variant)

[2] Actually, experiments with negative result form a noticeable fraction among all experiments [13].

[3] For instance, if an update of the ranking algorithm of a search engine is made only for the small amount of queries, then only the users submitted these queries should be kept in the treatment group.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741116.

and challenging [4], that is why a modification of an existing standardized metric is preferred.

The variance reduction techniques (the stratification and control variates), used in Monte Carlo simulations, were also applied [7] in the A/B testing. Though, they do not have the above-mentioned drawbacks and lead to a noticeable variance reduction, these techniques are based on the utilization of pre-experiment data and, hence, are limited in their applicability as well.

In our paper, we propose a novel method to improve sensitivity of online controlled experiments that exploits the prediction of the future behavior of an *individual* user based solely on the user's behavior observed during the experiment. The intuition of this method is the following. We know (as mentioned above) that the longer an A/B experiment is conducted, the higher its sensitivity. Therefore, we can try to peek into the future behavior of each user participated in the experiment, and, based on the predicted user behaviors, calculate the evaluation metric for the experiment as if it was conducted beyond its actual time period. Hence, we expect improvement of the sensitivity as if the experiment was extended, while the actual experiment duration was not increased.

While our approach could be applied to any online metric of system performance (which is calculated over user data) we investigate it for the case of user engagement metrics. User engagement reflects how often the user solves her needs (e.g., to search something) by means of the considered service (e.g., a search engine). On the one hand, these metrics are measurable in the short-term experiment period, and, on the other hand, they are predictive of the long-term success of the company [14, 15, 16, 25]. That is why engagement metrics are often considered to be most appropriate for online evaluation. In this study, we pay special attention to the metrics that reflect the loyalty aspect of engagement: the state-of-the-art *number of sessions per user* metric [14, 27] (which is accepted as the "North-star" for A/B testing evaluation in major search engine companies like Bing [15, 16]) and the recently proposed *absence time* metric [9]. Our approach raises the problem of prediction of the future individual user engagement. To the best of our knowledge, no existing study on user engagement investigated such a prediction problem[4].

Finally, our approach could be used simultaneously with other sensitivity improvement methods (i.e., with stratification, control variates [7], transformation of the metric, filtration of the treatment group [18], etc.) and, hence, is not an a substitute, but rather a complement to them.

To sum up, our study considers the problem, which coincides with the *emerging Internet companies' needs*: to run *more* online controlled experiments and to get their results on a significant level *faster*. Specifically, the major contributions of our work include:

- The solution to the problem of predicting the exact future values of engagement metrics of individual user behavior.
- Utilization of this predictor for the sensitivity improvement of online controlled experiments.
- Validation of our sensitivity improvement approach on real online experiments run at Yandex, first, demonstrating the growth of the number of statistically significant A/B experiments and, second, detecting the treatment effect of an A/B test faster (up to 50% of saved time) with the same level of statistical significance.

The rest of the paper is organized as follows. In Section 2, the related work on A/B testing and user engagement is discussed. In Section 3, our sensitivity improvement approach is presented. In Section 4, the six studied user engagement measures are introduced and a brief analysis of them is presented. In Section 5, our user engagement prediction problem is stated and its solution is evaluated by a large series of experiments. In Section 6, we show the results of applying our sensitivity improvement approach to several A/B experiments and discuss its possible extensions. In Section 7, the study's conclusions and our plans for the future work are presented.

2. RELATED WORK

We compare our research with other studies in two aspects. The first one relates to online controlled experiments and, specifically, to the sensitivity improvement. The second one concerns user engagement with web services.

Online controlled experiment studies. The theoretical details of the online controlled experiment methodology were widely described in the existing works [22, 17, 18], and we conduct our experiments in accordance with them. A rich practical experience of using this methodology for different evaluation cases in many web companies was studied recently [13, 19, 28, 15]. These studies concern, inter alia, experiments with different components of a web service (the user interface [13, 8], ranking algorithms [27, 8], etc.), large-scale experimental infrastructure [28, 15], different aspects of user behavior and interaction with a web service (clicks [14, 16], speed [19, 16], abandonment [16], periodicity [8], etc.), and so on. Some of existing works focused on the study of the trustworthiness of the results of an A/B test. Various pitfalls and puzzling results of online controlled experimentation were shared in [4, 14] and several "rules of thumb" were discussed in [16]. In our work, we were aware of all these experiences and pitfalls during our A/B experimentation.

Basic sensitivity improvement techniques for an online controlled experiment were concerned in [18]: increasing of the experiment duration or of the user population, participated in the experiment [5]; the use of another evaluation metric with lower variance [8]; and the elimination of users who were not affected by the service change in the treatment group [27]. More sophisticated variation reduction techniques were proposed in [7]: the stratification and the usage of control variates based on pre-experiment data. The authors of [16] noted that the reduction of skewness of the evaluation metric used in an A/B experiment could also improve the sensitivity (e.g., transformation of the metric or capping its values [16]). The sensitivity improvement for two-stage online controlled experiments was proposed in [6]. All these methods have their own disadvantages and limitations in their applicability (as discussed in Section 1). We propose an alternative approach to improve the sensitivity, which is based on predicting future user behavior and does not possess the limitations of the ones described above: no

[4] The only study [27] was devoted to the binary prediction of user engagement increase/decrease in the future week. In our study, we focus on predicting the exact values of the user engagement metrics.

use of pre- or post-experiment data; no dependance on the particular case of the service change (or comparison) under evaluation; and applicability to a wide range of metrics that could be predicted for an individual user.

User engagement with web services. Existing studies of user engagement with a web service and, particularly, a search engine are three-fold. First, several studies focused on analysis of user behavior. Some of them discovered the relationship between search success and search engine reuse [30, 11]. Some others concerned behavioral patterns of users (e.g., simultaneous usage of several search engines [30] or periodicity of user engagement with a search engine usage [8]) and models of web sites with respect to user behavior (e.g., w.r.t. multitasking user behavior [20] or w.r.t. popularity, activity, and loyalty among users [21]). In our work, we present a brief analysis of the correlations between several user engagement metrics including the state-of-the-art "number of sessions" and the recently proposed "absence time".

Second, some studies focused on the prediction of future changes in user engagement. Prediction of how a user switches (no switch, persistent switch, or oscillating behavior) between search engines during 26 weeks was studied in [30]. The authors of [27] predict user engagement increase/decrease in the future week (they did not attempt to put this into any practical use, e.g., for sensitivity improvement). Among the features of the classifier they built, there were both engagement measures (the numbers of sessions, queries, clicks) and some non-engagement ones (query types, user satisfaction, etc.) observed during either one or three preceding weeks. The output of such a classifier could not be straightforwardly integrated in the A/B testing methodology to improve the sensitivity of evaluation metric due to the lack of the interpretable quantity of that output (see Sec. 3 for details). The prediction of the exact values of a user engagement metric provides an approximation of the original evaluation metric, which thus can serve as an interpretable metric itself, but it was not investigated in existing studies to the best of our knowledge. In our work, we introduce such prediction methodology for different user engagement metrics including the commonly used numbers of sessions, queries, clicks, and the recently proposed absence time.

Third, there are papers devoted to user engagement as an evaluation metric in online controlled experiments. In [14, 15, 16] the number of sessions per user was stated to be one of the key metrics used in Bing's experimentation platform. The authors of [27] evaluated different changes in search relevance of a popular search engine by means of A/B testing with respect to this engagement metric and several non-engagement measures reflecting query types and user satisfaction. The absence time (the time between two user visits) on a par with other engagement metrics was applied to compare different ranking algorithms used at Yahoo! Answers [9] and to evaluate a web search engine changes in ranking algorithm and user interface [3]. The novel periodicity engagement metrics of user behavior (resulted from the Discrete Fourier transform of 4 state-of-the-art engagement measures) were applied in [8] to evaluate, by means of A/B experiments, different changes of the search engine ranking algorithm, changes of the user interface, and changes of the engine's efficiency. In our work, we study the sensitivity improvement of controlled experiments with several user engagement metrics that are widely used in the described literature (see details in Sec. 4).

3. FRAMEWORK

Background. In A/B testing, users, participated in an experiment, are randomly exposed to one of two variants of the service, the control (A) and the treatment (B) variants (e.g., the current production version of the service and its update), in order to compare them [18, 14, 16]. The comparison is based on *the evaluation metric* (also known as the online service quality metric, the overall evaluation criterion, etc.). In this paper, we consider "per user metrics" [4], which are calculated for each individual user and, then, are averaged over all users in each experiment variant[5]. Namely, let us consider a measure $\mathtt{M}(u, \mathbb{T})$ of user interaction with the service (e.g., the number of clicks) during the experiment period $\mathbb{T}$. For each user group $\mathcal{U}_A$ and $\mathcal{U}_B$, we calculate the average values[6] $\mathtt{M}^X(\mathbb{T}) = \mathrm{avg}_{u \in \mathcal{U}_X} \mathtt{M}(u, \mathbb{T}), X = A, B$. After that, their absolute and relative differences are calculated[7]: $\Delta_{\mathbb{T}} = \mathtt{M}^B(\mathbb{T}) - \mathtt{M}^A(\mathbb{T})$ and $\mathrm{Diff}_{\mathbb{T}} = \chi \cdot \Delta_{\mathbb{T}} / \mathtt{M}^A(\mathbb{T})$.

Nonetheless, the quantities $\Delta_{\mathbb{T}}$ and $\mathrm{Diff}_{\mathbb{T}}$ could not serve themselves as indicators of positive or negative consequences of the evaluated treatment variant of the service. The difference between the evaluation metrics over groups should be controlled by a statistical significance test. In our study, we utilize widely applicable *two-sample t-test* (as in [7, 27, 6]) to decide weather the metric of the treatment user group B is significantly larger or smaller than that of the control one. This test is based on *t-statistic*:

$$\Delta_{\mathbb{T}} / \sqrt{\sigma_A^2(\mathtt{M}(\mathbb{T})) \cdot |\mathcal{U}_A|^{-1} + \sigma_B^2(\mathtt{M}(\mathbb{T})) \cdot |\mathcal{U}_B|^{-1}}, \tag{1}$$

where $\sigma_X^2(\mathtt{M}(\mathbb{T}))$ is the standard deviation of the measure $\mathtt{M}(\cdot, \mathbb{T})$ over users $\mathcal{U}_X, X = A, B$. The larger the absolute value of the t-statistic, the lower the probability (also known as *p-value*) of *the null hypothesis*, which assumes that the observed difference is caused by random fluctuations, and the variants are not actually different. If the p-value is lower than the threshold $p_{\mathrm{val}} \leq 0.05$ (commonly used [18, 14, 7, 27, 16]), then the test rejects the null hypothesis, and the difference $\Delta_{\mathbb{T}}$ is assumed to be statistically significant. The additional details of the A/B testing framework could be found in the survey and practical guide [18].

The future user behavior prediction approach. As it was shown above, any evaluation metric is calculated based on some time period $\mathbb{T} = [0, T)$, i.e., on the experiment period. One way to improve the sensitivity of an A/B test is to conduct it longer [18], i.e., to increase the length of the period $\mathbb{T}$. *The main idea of this paper is to make it virtually by peeking into the future user behavior.*

Suppose that, for a considered measure $\mathtt{M}$ and for each user $u \in \mathcal{U}$, we are able to forecast the value of the measure $\mathtt{M}$, calculated over some forecast T_f-day period based on some user behavior data $\mathbf{F}_{T_p}(u)$ observed during the first T_p days of the target period. We denote this predicted value of the true measure by $\mathrm{P}_{\mathtt{M}, T_f}(\mathbf{F}_{T_p}(u))$. If the length T of an A/B experiment equals to T_p, then we are able to predict the user

[5] This type of metrics (e.g., the number of sessions per user) is very popular in the online controlled experimentation [14]. However, there are also frequently used non-per user metrics [4] like the annual revenue [16].

[6] In order to distinguish the method of calculating the value $\mathtt{M}(u, \mathbb{T})$ for a user, from the value $\mathtt{M}^X(\mathbb{T})$, calculated for a user group X, the first one is referred to as *a measure* and the second one is referred to as *a metric* in our work.

[7] The factor χ is randomly chosen once in our study in order to hide real values for confidentiality reasons.

measure $\mathtt{M}(u, \cdot)$ over *the post-experiment period* $\mathbb{T}_f = [0, T_f)$ and calculate the evaluation metric $\mathtt{M}^X(\cdot), X = A, B$, (based on the predicted values) as if the A/B test was conducted beyond its actual time period. We will use the notations $\widetilde{\mathtt{M}}(u, \mathbb{T}_f) = \mathrm{P}_{\mathtt{M}, T_f}(\mathbf{F}_T(u)) \approx \mathtt{M}(u, \mathbb{T}_f)$ and $\widetilde{\mathtt{M}}^X(\mathbb{T}_f) \approx \mathtt{M}^X(\mathbb{T}_f)$ for the predicted user measure and its average values over the user groups $X = A, B$ respectively. Thereby, we forecast the metric $\mathtt{M}(\cdot, \mathbb{T}_f)$, which would be calculated, if the A/B test period has been $[0, T_f)$ instead of $[0, T_p)$. We refer to this method of improving an evaluation metric as *the future user behavior prediction approach* (***FUBPA***).

First, our approach differs from the naive one, where prediction of the metric value $\mathtt{M}^X(\mathbb{T}_f)$ for each group $X = A, B$, (or of their difference $\Delta_{\mathbb{T}}$) is based on their trends, observed during the A/B experiment (i.e., on the metric values $\mathtt{M}^X([0, t))$ for some $t \leq T$) [14]. This method does not provide the significance level of the forecast difference due to the absence of data per experimental unit (i.e., per user in our case), and, thus, could not be used to improve sensitivity of A/B tests. On the contrary, our approach provides data for each user and, therefore, allows to calculate t-statistic.

Second, our approach differs from the utilization of a binary classifier of increase/decrease of a considered user measure in a post-experiment period (as it is proposed in [27] for the number of sessions). A fundamental shortcoming of such a classifier is that it could not be straightforwardly integrated in the A/B testing methodology. In fact, its outputs should be transformed to an evaluation metric first. One clear solution is to take the fraction of users whose engagement is predicted to grow. Another solution could be the averaged probability of growth estimated by the classifier. In any case, there is no reason why the obtained metric will correlate with the original user engagement metric accepted in the company. Hence, its difference $\mathrm{Diff}_{\mathtt{T}}$ could not provide an insight on the difference of the original metric, which may be fatal both for understanding the A/B experiment's result and for the taken decision. For instance, a drop in the fraction of users with increased number of clicks on ads could be observed simultaneously with a growth in the number of clicks per user. The latter evaluation metric has the direct connection with the annual revenue of a considered company. On the contrary, in our approach, the approximation $\widetilde{\mathtt{M}}^X(\mathbb{T}_f)$ is trained to be as close as possible to the baseline metric $\mathtt{M}$.

Finally, the approximation is a surrogate of different simple measures $\mathbf{F}_{T_p}(u)$ (which are used as features in the predictor). Therefore, *the FUBPA could be also treated as a method to combine a series of evaluation metrics ([4]), targeting the combined metric to be similar to a desirable evaluation metric in future*. The rest of this section is devoted to the theoretical analysis of the FUBPA approach.

Relation between prediction quality and sensitivity. Denote $m(u) = \mathtt{M}(u, \mathbb{T}_f)$ the metric over users $u \in \mathcal{U}$. Denote $\widetilde{m}(u) = \widetilde{\mathtt{M}}(u, \mathbb{T}_f)$ its predictor. W.l.o.g., we assume that the mean value $\mathbb{E}(m)$ equals 0. We also assume that our prediction is unbiased, i.e., its mean value $\mathbb{E}(\widetilde{m})$ is also equal to zero. In this case, the standard deviation σ_e of the prediction error $e := m - \widetilde{m}$ is, by definition, the Root Mean Square Error (RMSE) of the predictor $\widetilde{m}$. We have

$$\sigma_m^2 = \sigma_e^2 + \sigma_{\widetilde{m}}^2 + 2\mathrm{Cov}(e, \widetilde{m}), \tag{2}$$

where σ_ξ^2 denotes the standard deviation of a variable ξ.

Note that, for any predictor m', the optimal predictor in the class $\{cm' \mid c \in \mathbb{R}\}$ is $\widehat{m'} := (\mathrm{Cov}(m, m')/\sigma_{m'}^2)m'$, since it provides the minimum of the mean squared error $\sigma_{m-cm'}^2 = c^2\sigma_{m'}^2 - 2c\mathrm{Cov}(m, m') + \sigma_m^2$. Therefore, if the class of prediction models is closed under scalar multiplication[8], for the optimal predictor $\widetilde{m}$, we have $\mathrm{Cov}(m, \widetilde{m}) = \sigma_{\widetilde{m}}^2$, and thus

$$\mathrm{Cov}(e, \widetilde{m}) = \mathrm{Cov}(m - \widetilde{m}, \widetilde{m}) = \mathrm{Cov}(m, \widetilde{m}) - \sigma_{\widetilde{m}}^2 = 0.$$

Therefore, the identity (2) implies that, in the case of an optimal predictor, the RMSE and the standard deviations of the metric σ_m and its predictor $\sigma_{\widetilde{m}}$ are connected by the following identity

$$\sigma_{\widetilde{m}}^2 = \sigma_m^2 - \mathrm{RMSE}^2(m, \widetilde{m}). \tag{3}$$

This implies the following clear finding. The better the prediction quality in terms of RMSE, the greater the variance of the obtained surrogate measure $\widetilde{m}$. On the one hand, increasing of the prediction quality provides a better approximation of the original measure $m(u)$ and may lead to a better online metric $\mathtt{M}(\mathbb{T}_f)$. On the other hand, better prediction implies greater variance, which may affect the sensitivity of the metric (see t-statistics in Eq. (1)). Hence, we conclude that the prediction quality does not directly translate into the growth of the metric sensitivity. For example, we may expect that a 10% growth of prediction quality over a baseline may lead to a significant sensitivity improvement, while further growth of quality does not improve the obtained metric.

4. USER ENGAGEMENT

Engagement measures. We use the logs of Yandex[9], one of the most popular global search engines, in order to study user engagement. For each user[10], we study 6 popular engagement measures, which represent both loyalty and activity aspects of user engagement. For a considered period of time (a day, a week, a month, etc.), we study the following engagement measures calculated over this time period:

- the number of sessions (`S`);
- the number of queries (`Q`);
- the number of clicks (`C`);
- the presence time (`PT`);
- the number of clicks per query (`CpQ` or `CTR`);
- the absence time per session (`ATpS`).

Following common practice [12, 14, 9, 27, 3, 8], we define a session as a sequence of user actions whose dwell times are less than 30 minutes. The presence time `PT` is measured as the sum of all session lengths (in seconds) observed during the considered time period, while the total absence time is measured as the length of the considered time period minus the presence time. Note that the measures `S`, `Q`, `C`, and `PT` are additive with respect to the time period. The measure `ATpS` is calculated as the total absence time divided by the

[8] This condition is satisfied by a wide range of prediction models including linear model and the state-of-the-art Friedman's gradient boosting decision tree model [10], which are applied in Section 5.

[9] `http://www.yandex.com`

[10] We use cookie IDs [18] to identify users as done in other studies on user engagement [9, 20, 27, 8].

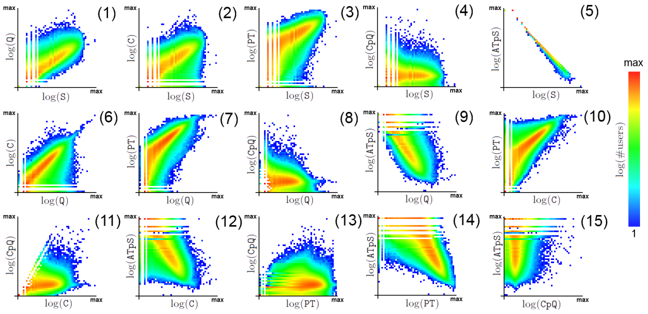

Figure 1: The joint distributions of users $\mathcal{U}$ w.r.t. each pair of the studied measures $\mathfrak{M}$ calculated over a 2-week period.

number of user sessions S[11]. The number of clicks per query CpQ could be regarded as *the CTR of the search engine result pages* as well. The measures S and ATpS represent the *user loyalty* [25, 14, 9, 27, 8], whereas the measures Q, C, PT, and CpQ represent the *user activity* [18, 25, 9, 7, 8] aspects of user engagement [25]. The set of all measures is denoted by $\mathfrak{M} = \{\mathtt{S}, \mathtt{Q}, \mathtt{C}, \mathtt{PT}, \mathtt{CpQ}, \mathtt{ATpS}\}$. Before proceeding to the main problems studied in this paper, we present a brief analysis of these measures. We investigate the relationships between them and their persistence across time in order to have a better interpretation of the prediction quality obtained in the next section.

Correlation between measures. We calculate the correlations between the studied measures in the following way. Let one consider two measures $\mathtt{M}_1, \mathtt{M}_2 \in \mathfrak{M}$, a certain set of users $\mathcal{U}$, and a certain time period $\mathbb{T}$. Then, these measures $\mathtt{M}_i(u, \mathbb{T}), i = 1, 2$, are calculated for each user $u \in \mathcal{U}$ *over the whole* time period $\mathbb{T}$ (e.g., the total number of sessions during the entire period). We consider a user as a random event and calculate the Pearson's correlation coefficient $\mathrm{Corr}_{\mathcal{U}}$ over users $\mathcal{U}$.

In the analysis of this section, we use a *2-week period*[12] from March, 2013 as the time period $\mathbb{T}$, and we consider all active users of the search engine during the period $\mathbb{T}$ as the set $\mathcal{U}$ (its size $|\mathcal{U}| \gg 10^7$). Table 1 reports the values of the correlations $\mathrm{Corr}_{\mathcal{U}}$ between all measures from $\mathfrak{M}$. We highlighted in **boldface** the ones with the highest absolute value in each column, while the smallest ones are underlined. First, one sees that all additive measures S, Q, C, and PT are noticeably well correlated ($\mathrm{Corr}_{\mathcal{U}} \geq 0.81$). Second, the measure CpQ has low correlation with other measures ($|\mathrm{Corr}_{\mathcal{U}}| \leq 0.083$) except with the number of clicks C ($\mathrm{Corr}_{\mathcal{U}} = 0.175$). Third, the absence time ATpS is mostly negatively correlated with the other measures. This observation coincides with the intuition that the more frequently a user utilizes the service (higher the number of sessions, queries, etc.), the lower the absence time.

Next, we plot the joint distributions[13] of users $\mathcal{U}$ w.r.t. each pair of the studied measures $\mathfrak{M}$ calculated over the 2-week period $\mathbb{T}$ (i.e., 15 heat maps in total) in Fugure 1. First, the previously observed close relationships between all additive measures S, Q, C, and PT are also seen in their joint distributions: the user population is concentrated near the main diagonal in the heat maps in Fig. 1 (1, 2, 3, 6, 7, and 10). The most consistent pattern is observed for the distribution of the number of queries Q and the number of clicks C (Fig. 1 (6)), which reveals a clear linear dependence between their logarithms. The maps Fig. 1 (4, 8, 11, 13, and 15) explain the very small correlation of CpQ with other measures reported in Table 1. Second, the negative correlation of the absence time ATpS with the other measures is also seen in Fig. 1 (5, 9, 12, 14, and 15), where the user population is concentrated along the secondary diagonal of the heat maps.

Table 1: Correlations $\mathrm{Corr}_{\mathcal{U}}$ between all engagement measures $\mathfrak{M}$ calculated over a 2-week period.

$\mathrm{Corr}_{\mathcal{U}}$	S	Q	C	PT	CpQ	ATpS
S	—	0.843	0.810	0.831	0.028	**−0.595**
Q	**0.843**	—	0.888	**0.910**	−0.001	−0.483
C	0.810	0.888	—	0.904	**0.175**	−0.478
PT	0.831	**0.910**	**0.904**	—	0.081	−0.461
CpQ	0.028	−0.001	0.175	0.081	—	−0.083
ATpS	−0.595	−0.483	−0.478	−0.461	−0.083	—

Long-term measure persistence. We also identify the relationship between the values of the same measure calculated over two different time periods. Namely, the Pearson's correlation coefficient over users $\mathcal{U}$ is calculated for the values $\mathtt{M}(\cdot, \mathbb{T}_1)$ and $\mathtt{M}(\cdot, \mathbb{T}_2)$ of each measure $\mathtt{M} \in \mathfrak{M}$, where $\mathbb{T}_1$ and $\mathbb{T}_2$ are two *consecutive 2-week periods*[14]. The correlation are presented at the bottom of Fig. 2. The highest value is highlighted in **boldface**, and the lowest one is underlined. The joint distributions of users $\mathcal{U}$ w.r.t. each engagement measure from $\mathfrak{M}$ calculated over the 2-week periods are also captured in Fig. 2. Almost all measures (except for CpQ) have a high persistence over time ($\mathrm{Corr}_{\mathcal{U}} \geq 0.61$). Hence, *we expect that the values of these measures over the future may be better predicted by their values over the observed period than the ones of* CpQ.

In this section, we presented the main user engagement measures under our study and briefly analyzed them in different ways. These observations will help us clearly understand the findings in the next section devoted to the prediction of user engagement.

5. USER ENGAGEMENT PREDICTION

In our work, we study the user engagement prediction problem in the following setting. We suppose that one has user data observed during a period of time $\mathbb{T}_p$ (the *past*, or the *observed time period*), and, based on these data, one needs to predict *the exact value* of a target engagement measure calculated over an increased period of time $\mathbb{T}_f \supseteq \mathbb{T}_p$ (the *forecast time period*) for each *individual user*. We consider the 6 engagement measures $\mathfrak{M}$ presented in the previous section as our targets.

The models. We utilized two models to predict the exact values of the targets. The first one is a state-of-the-art Friedman's gradient boosting decision tree model [10]. We used a proprietary implementation of the machine learning algo-

[11] In our study, we also considered another definition of this measure: the sum of times between consecutive user sessions divided by the number of absences (i.e., by $\mathtt{S} - 1$), and, if only one session is observed, then it is equal to 0. However, such variant of measure demonstrated lower persistence over time, noticeably lower predictability, and the OEC based on it failed unacceptable number of A/A tests.

[12] It is a popular length of A/B experiments [14, 27, 16, 8].

[13] We hide all absolute values for confidentiality reasons.

[14] In our study, we consider all users $\mathcal{U}$ that used the search engine during the period $\mathbb{T}_1$ (and may use it during the forecast period), while users who started using it in the period $\mathbb{T}_2$ are not included in the set $\mathcal{U}$.

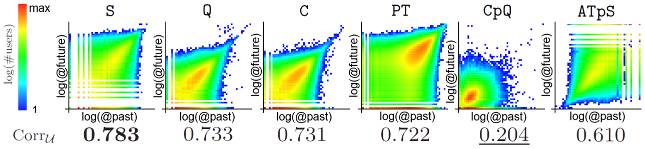

Figure 2: Correlations and the joint distributions of users $\mathcal{U}$ w.r.t. to each engagement measure from the set $\mathfrak{M}$ between two consecutive 2-week periods.

rithm with 1000 iterations and 1000 trees, which appeared to be the best settings during validation. The second model is a linear regression model with L_2 regularization. The learning rate of the decision tree model and the regularization parameter of the linear model were adjusted by means of cross-validation technique during optimization. The optimal settings were found with respect to *the mean squared error* used as the loss function.

The data. In order to conduct our experiments, we used the logs of user search activity on the popular search engine Yandex. We collected user data from two non-intersecting 34-day periods of 2013 year, the earlier period was used to form the training data set and the later one was used to collect the test data set. Since the day of the week may have a considerable impact on the user behavior, we would not like the data set to be biased to some particular day of the week the observed period $\mathbb{T}_p$ begins. Therefore, for each of the 34-day periods, each user is represented by 7 feature vectors in the data set (treated as 7 different examples) that are calculated based on the raw user behavior data truncated at $0, 1, \dots, 6$ first days, so that the observed period begins on the 1st, 2nd, …, 7th day of the 34-day period. Further, we randomly sampled these data sets 20 times, obtaining smaller training and test data sets of equal size (of $> 10^5$ users) in order to, first, reduce the data size, which should comply with computational constrains, and, second, apply *the paired two-sample t-test* to measure the significance level of the obtained results.

The performance measure. Let us consider the naive average prediction model ($\text{avg}_{\mathcal{U}}$), which, for each user from the test data set, predicts the target value as *the average* of the values of that target *over all users* in the training data set. One could treat it as a model, which utilizes *zero features*. Denote the RMSE value calculated over the test set for a given observed and forecast periods ($\mathbb{T}_p$ and $\mathbb{T}_f$), for a given target (**tg**), for a given prediction model (m), and for a given feature set (Φ), by $\text{RMSE}(\mathbb{T}_p, \mathbb{T}_f, \mathbf{tg}, m, \Phi)$. In the results reported below, we use the *normalized RMSE* (nRMSE) defined as

$$\text{nRMSE}(\mathbb{T}_p, \mathbb{T}_f, \mathbf{tg}, m, \Phi) = \frac{\text{RMSE}(\mathbb{T}_p, \mathbb{T}_f, \mathbf{tg}, m, \Phi)}{\text{RMSE}(\mathbb{T}_p, \mathbb{T}_f, \mathbf{tg}, \text{avg}_{\mathcal{U}}, \varnothing)}.$$

This performance measure allow us, first, to hide the RMSE values due to confidential reasons, second, to compare the prediction quality between different targets and forecast periods, and, finally, to improve our understanding of weather a studied prediction model is better or not than a naive baseline model (the average prediction) by comparing the nRMSE w.r.t. 1.

The observed and target periods. In our experimentation, we compare different feature sets and prediction models for 14-day observed periods $\mathbb{T}_p$ and for 28-day forecast period $\mathbb{T}_f$. We remind that these are popular lengths of A/B experiments [14, 27, 16, 8]. However, the dependence of the prediction quality on the period lengths is analyzed at the end of this section.

Table 2: The top-20 features w.r.t. the improvement over the baseline feature set in terms of RMSE for the measure S with $|\mathbb{T}_p| = 14$ and $|\mathbb{T}_f| = 28$.

Decision Trees			Linear Regression		
M	Feature	Impr.	M	Feature	Impr.
S	lg**avg**	−0.43%	S	$\mathbf{DFT_A}$[1]	−0.28%
ATpS	lg$\mathbf{TS_c}$[14]	−0.40%	S	lg$\mathbf{DFT_A}$[1]	−0.25%
ATpS	$\mathbf{Ent_{Perm}}$[5]	−0.20%	C	$\mathbf{GrPos_1}$	−0.24%
S	$\mathbf{Ent_{Ap}}$[2]	−0.20%	PT	$\mathbf{GrPos_1}$	−0.23%
ATpS	lg$\mathbf{TS_c}$[13]	−0.19%	PT	$\mathbf{GrPos_2}$	−0.23%
S	$\mathbf{sum_2}$	−0.18%	S	lg**max**	−0.22%
ATpS	lg$\mathbf{Ent_{Perm}}$[5]	−0.18%	S	**std**	−0.21%
Q	$\mathbf{Ent_{Smpl}}$[6]	−0.18%	Q	$\mathbf{GrPos_1}$	−0.21%
ATpS	$\mathbf{Ent_{Perm}}$[6]	−0.17%	S	$\mathbf{GrPos_1}$	−0.21%
ATpS	lg**avg**	−0.17%	Q	lg$\mathbf{GrPos_1}$	−0.21%
S	$\mathbf{DFT_A}$[1]	−0.17%	S	lg$\mathbf{GrPos_1}$	−0.21%
Q	$\mathbf{Ent_{Ap}}$[2]	−0.16%	PT	$\mathbf{Ent_{Sh}}$	−0.20%
ATpS	$\mathbf{Ent_{Perm}}$[8]	−0.16%	ATpS	$\mathbf{Ent_{Sh}}$	−0.20%
ATpS	$\mathbf{Ent_{Perm}}$[8]	−0.15%	S	lg**std**	−0.20%
S	lg$\mathbf{Ent_{Ap}}$[2]	−0.14%	ATpS	lg$\mathbf{Ent_{Sh}}$	−0.19%
S	lg$\mathbf{sum_2}$	−0.14%	S	$\mathbf{Ent_{Ap}}$[2]	−0.18%
ATpS	$\mathbf{Ent_{Sort}}$[8]	−0.13%	ATpS	$\mathbf{DFT_A}$[1]	−0.18%
Q	$\mathbf{Ent_{Ap}}$[6]	−0.13%	S	lg**max**	−0.18%
S	$\mathbf{GrPos_4}$	−0.13%	PT	lg$\mathbf{Ent_{Sh}}$	−0.17%
ATpS	$\mathbf{Ent_{Perm}}$[7]	−0.13%	ATpS	$\mathbf{Ent_{Perm}}$[2]	−0.17%

Features. Now we discuss the features we utilize in the user engagement prediction. Each engagement measure $\mathtt{M} \in \mathfrak{M}$ can be translated into several scalar features in different ways. Specifically, we consider the following *calculation and transformation methods*.

The total feature. First, we calculate the measure M over the observed time period $\mathbb{T}_p$ (e.g., for $\mathtt{M} = \mathtt{S}$, it is the total number of sessions over the time period $\mathbb{T}_p$). This is the same as the predicted target, but over the observed time period $\mathbb{T}_p$ instead of the forecast one $\mathbb{T}_f$. We denote this feature by **Total**.

The time series. Second, we calculate the measure M over each day of the time period $\mathbb{T}_p$ and obtain a daily time series of length $|\mathbb{T}_p|$. We denote such feature vector by $\mathbf{TS_d}$. Then, for each day $t \in \mathbb{T}_p$, we calculate the measure M over the time period that starts on the first day of $\mathbb{T}_p$ and finishes on the day t. Thus, we obtain a cumulative time series of length $|\mathbb{T}_p|$. This feature vector is denoted by $\mathbf{TS_c}$.

The statistics features. Next, we consider the daily time series $\mathbf{TS_d} = \{x_t\}_{t=1}^{|\mathbb{T}_p|}$ and get basic statistics over it: the minimal (**min**), maximal (**max**), average (**avg**) values, the standard deviation (**std**), the median (**med**), the sum (**sum**), the sum of squares ($\mathbf{sum_2}$), and the sum of cubes ($\mathbf{sum_3}$) over days. Additionally, we calculate the variation of the time series (i.e., $\mathbf{var} := \sum_{t=1}^{|\mathbb{T}_p|-1} |x_{t+1} - x_t|$) and the number ($\mathbf{GrPos}_\tau$) of the values of the series not less than the threshold τ, which is chosen from among $1, 2, 4, 10, 20, 40,$ and 100. Thus, we consider here 16 scalar features in total.

The derivative features. We calculate the finite differences of the first order $\mathbf{FD_1} := \{x_{t+1} - x_t\}_{t=1}^{|\mathbb{T}_p|-1}$ and of the second order $\mathbf{FD_2} := \{x_{t+2} - 2x_{t+1} + x_t\}_{t=1}^{|\mathbb{T}_p|-2}$ over the daily time series $\mathbf{TS_d}$. These $2|\mathbb{T}_p| - 3$ scalar features are analogs of the first and second derivatives in the discrete case. Similar first order finite differences over weekly time series were used in the classifier [27] and had one of the highest weights in the prediction settings. We also sort in descending order the differences $\mathbf{FD_1}$ ($\mathbf{FD_2}$) and replace each difference by its index in the original series, obtaining the integer sequence

$\mathbf{FD_1^{rank}}$ ($\mathbf{FD_2^{rank}}$), which describes the day of the highest difference, the day of the lowest difference, etc.

The periodicity features. We apply the *discrete Fourier transform (DFT)* [29] to the daily time series $\mathbf{TS_d}$: $X_k = \sum_{t=1}^{|\mathbb{T}_p|} x_t e^{-\mathrm{i}\omega_k(t-1)}$, $\omega_k = \frac{2\pi k}{|\mathbb{T}_p|}$, $k \in \mathbb{Z}_{|\mathbb{T}_p|}$. Thus, we obtain $|\mathbb{T}_p|$ complex numbers ($\mathbf{DFT}$), then we get their real parts (amplitudes $\mathbf{DFT_A}$) and their imaginary parts (phases $\mathbf{DFT_{Ph}}$). These $2|\mathbb{T}_p|$ scalar features encode periodicity in user engagement and were studied in the paper [8], where their long-term persistence is found to be better than the one of the daily time series. We also sort in descending order the amplitudes $\mathbf{DFT_A}$ and replace each amplitude by its index in the original series, obtaining the integer sequences $\mathbf{DFT_A^{rank}}$. This sequence describes which frequencies $\omega_k, k \in \mathbb{Z}_{|\mathbb{T}_p|}$, dominate among others, which one has the lowest amplitude, etc.

The entropy features. In order to get the average amount of information contained in daily user engagement series, we calculate entropy for our series $\mathbf{TS_d}$ in the following ways: (a) the Shannon entropy $\mathbf{Ent_{Sh}}$ as in [26]; (b) the Permutation entropies $\mathbf{Ent_{Perm}}$ as in [2], and the Sorting entropies $\mathbf{Ent_{Sort}}$ as in [2] of orders $n = 2, .., |\mathbb{T}_p| - 1$; (c) the Approximate entropies $\mathbf{Ent_{Ap}}$ as in [23, 24], and the Sample entropies $\mathbf{Ent_{Smpl}}$ as in [24] for $m = 2, .., |\mathbb{T}_p| - 2$.

Thus, we consider $14|\mathbb{T}_p| + 2$ scalar features for each of 6 engagement measures $\mathfrak{M}$ (i.e., 1188 in total). All features described above are considered both in normal and in logarithmic (where applicable) scales in order to better catch the differences between values of different magnitudes. Besides, we consider the day-of-the-week of the first day of the observed period $\mathbb{T}_p$ as an additional categorical feature **DoW** in order to better catch the influence of the position of the period w.r.t. the week cycle.

Feature selection. Note that statistics, derivative, periodicity, and entropy features are derivable from the source time series. Therefore, we expect that not all these features could provide a significant profit in prediction quality w.r.t. the source time series. Besides, utilization of the large number of features may lead to overfitting of the model[15] and unreasonable consumption of computational resources. Hence, we conduct feature selection by studying the profit in prediction quality of each scalar feature w.r.t. a baseline feature set. We consider the set $\{\mathbf{Total}, \mathbf{TS_d}\}$ calculated for all user engagement measures $\mathfrak{M}$ as the baseline feature set (it consists of $6|\mathbb{T}_p| + 6 = 90$ scalar features) in our feature selection procedure. Then, we train our models under the setup described at the beginning of this section for the baseline set and for this set extended by one of the other scalar features. We present the top-20 features w.r.t. the improvement of the baseline prediction quality in terms of RMSE for the number of sessions measure `S` in Table 2[16].

We see that both the periodicity and entropy features improve prediction quality, contrariwise, none of the derivative features shows any significant improvement with p-value $\leq$ 0.05. We selected the features that showed noticeable and significant improvement in several prediction tasks (i.e., targets and models) or for several user engagement measures. These features are: lg$\mathbf{TS_d}[13-14]$, $\mathbf{TS_c}[12-13]$, lg$\mathbf{TS_c}[12-14]$, $\mathbf{GrPos}_\tau$ for $\tau = 1, 2, 4, 10, 20, 40$, and 100, **min**, **max**, lg**max**, **avg**, lg**avg**, **std**, **med**, $\mathbf{sum_2}$, lg$\mathbf{sum_2}$, $\mathbf{sum_3}$, **var**, $Re\mathbf{DFT}[1]$, $Im\mathbf{DFT}[1]$, $\mathbf{DFT_A}[1]$, $\mathbf{DFT_A^{rank}}[1]$, $\mathbf{Ent_{Sh}}$, $\mathbf{Ent_{Perm}}[3-6]$, $\mathbf{Ent_{Sort}}[2, 3, 4-8]$, $\mathbf{Ent_{Ap}}[2-9]$, $\mathbf{Ent_{Smpl}}[3-6]$, and **DoW**. All these features together with the baseline ones (**Total** and $\mathbf{TS_d}$) calculated over all user engagement measures are denoted by **Best** (307 scalar features in total).

Table 3: Comparison of feature sets in terms of nRMSE (relative improvement w.r.t. the first column) of prediction of each of the measures $\mathfrak{M}$ for $|\mathbb{T}_p| = 14$ and $|\mathbb{T}_f| = 28$.

Measures:		Target measure		All (i.e., $\mathfrak{M}$)		
Features:		**Total**	**Total**, $\mathbf{TS_d}$	**Total**	**Total**, $\mathbf{TS_d}$	**Best**
Target & Model		1 feat.	15 feat.	6 feat.	90 feat.	307 feat.
`S`	LR	0.334	−4.92%	−0.12%	−5.11%	−5.58%
	DT	0.34	−6.38%	−0.32%	−6.48%	−7.59%
`Q`	LR	0.37	−2.65%	−0.64%	−3.59%	−5.04%
	DT	0.411	−4.49%	−1.17%	−4.28%	−5.62%
`C`	LR	0.375	−2.52%	−1.12%	−4.21%	−5.4%
	DT	0.394	−5.31%	−1.9%	−5.23%	−6.49%
`PT`	LR	0.372	−2.49%	−1.35%	−4.47%	−4.94%
	DT	0.388	−4.82%	−1.86%	−5.41%	−6.57%
`CpQ`	LR	0.423	0%	−0.12%	−0.08%	−0.46%
	DT	0.44	−0.03%	−1.18%	−1.02%	−2.5%
`ATpS`	LR	0.458	−0.3%	−0.2%	−0.64%	−1.78%
	DT	0.456	−1.9%	−0.67%	−2.63%	−2.96%

Comparison of feature sets and prediction models. We evaluate the feature set **Best** for both the linear regression model and the decision tree model with respect to the following 4 baselines. The first baseline set consists of one scalar feature **Total** calculated for the target measure we predict. The second baseline includes 6 scalar features **Total** calculated for each user engagement measure from $\mathfrak{M}$. The last two baseline sets contain both **Total** and the daily time series $\mathbf{TS_d}$ calculated either for the target measure (15 scalar features in total), or for each measure from $\mathfrak{M}$ (90 scalar features). The nRMSE values for these 5 feature sets, for 2 models, and for 6 different targets are presented in Table 3, all differences are significant with p-value $\leq$ 0.05.

First, we see that our feature set **Best** outperforms all baselines. Second, we see that the feature set **Best** provides a higher improvement over the fourth baseline set than the addition of each scalar feature individually, what we did in our feature selection procedure (compare the top of Table 2 and two first rows in Table 3). Therefore, we conclude that the new features carry different information for prediction and are not interchangeable.

Third, the decision trees and the linear models show similar results. On the one hand, the linear regression model sometimes outperforms the former one for several targets, especially for a low number of features. On the other hand, we see that the decision tree model demonstrates a higher rate of improvement with the growth of the size of the feature set than the linear one. In the next experiments we mostly report the results only for the decision trees model, because it demonstrates the best results for loyalty measures (`S` and `ATpS`) and it is the state-of-the-art model.

Fourth, we see that the number of sessions `S` is the most predictable target measure in terms of nRMSE for any used feature set (e.g., the use of only one feature **Total** reduces

[15] In our work, we tried to train models based on all features. However, these models were outperformed by the models based on the time series features solely.

[16] We removed the identical features, such as **avg**, $Re\mathbf{DFT}[0]$, $\mathbf{DFT_A}[0]$, etc. from the lists.

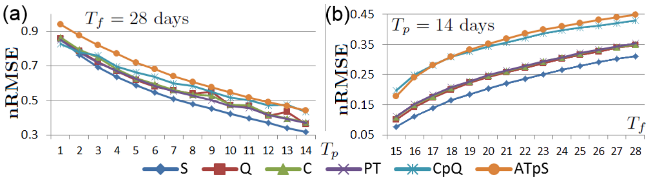

Figure 3: The dependence of prediction performance on (a) different $T_p = 1, .., 14$ for $T_f = 28$ and (b) different $T_f = 15, .., 28$ for $T_p = 14$ in terms of nRMSE.

the RMSE of the naive average baseline by 66.6%). The other additive target measures Q, C, and PT have almost the same nRMSE values and they are a little bit worse than the ones of S. The ratio target measures CpQ and ATpS have the worst prediction quality w.r.t. the naive average baseline. Moreover, the prediction quality for these two targets is very difficult to improve in comparison with the others. These observations coincide with our findings on the measures' persistence across time (compare the correlations $\mathrm{Corr}_{\mathcal{U},\mathbb{T}_{p/f}}$ in Fig. 2 and the first column of Table 3).

Fifth, the most noticeable quality improvement is observed, when we add the daily time series $\mathbf{TS_d}$ into the feature set (see col. 2 and 4 in the table) for all additive engagement measures (S, Q, C, and PT). This effect is expected, because all our novel features are derived from these time series (even the cumulative time series due to additivity of these measures). Note also that the quality for CpQ is improved noticeably when the feature set is extended by the other measures or the novel features, while the quality for ATpS also increases with the addition of the daily time series $\mathbf{TS_d}$ in the feature set.

At last, in order to understand the dependence of the prediction quality on the UE measures $\mathfrak{M}$ used in the set of features, we conduct the experiments to evaluate the drop in the performance caused by removal of each measure $\mathtt{M} \in \mathfrak{M}$. For all targets, we observe the highest performance drop, when we remove the features calculated for the target measure (e.g., ablation of all S-features decreases the nRMSE for prediction of the number of session S by 0.94%). In the other cases, the ablation of one of the measures connected with the activity aspect of user engagement (Q, C, PT, and CpQ) does not have a noticeable effect on the quality, while the user loyalty measures S and ATpS have a noticeable and significant influence on the prediction of any studied target (e.g., the ablation of all S-features decreases the nRMSE for prediction of the presence time PT by 0.24%). Finally, Fig. 3 demonstrates the nRMSE values for our best predictors learned for different combinations of the length T_p of the observed time period and the length T_f of the forecast one. We see that the closer these lengths are, the better the prediction quality.

To sum up, in this section, we studied the problem of prediction of the exact value of one of the six user engagement measures. We investigated the dependance of the prediction quality on the utilization of different user engagement measures, on the calculation/transformation techniques of translation them into features, and on the sizes of the observed and the forecast time periods. We conclude that *the use of almost all studied features could significantly improve the quality of user engagement prediction, and the dependance of this quality on the parameters of the prediction task agrees with the user engagement analysis provided in Section 4.*

6. A/B EXPERIMENTS

Experimental setup. In our paper, we consider 32 A/B experiments conducted on *real users* of the search engine (Yandex) in order to validate our approach of improving the sensitivity of key metrics (FUBPA, see Section 3 for details). Each experiment ran for at least two weeks. The user samples used in the A/B tests are all uniformly randomly selected, and the control and the treatment groups are of approximately the same size (at least, hundreds of thousands of users). Each experiment evaluates a change in one of the main components of the search engine: a change in the ranking algorithm, the engine response time, or a change in the user interface. During our A/B experimentation, 23 control experiments (so-called A/A tests) were conducted in order to check the correctness of the experimentation [18, 4]. We considered a commonly used threshold $p_{\text{val}} = 0.05$ for the p-value of the statistical significance test (i.e., of two-sample t-test [7, 27, 6], see Sec. 3 for details).

In this section, the per-user metrics based on six user engagement measures from $\mathfrak{M}$ (introduced and analyzed in Sec. 4) are considered as our baseline evaluation criteria. Then, we study the modified variants of these metrics obtained by means of *the future user behavior prediction approach* (FUBPA) described in Sec. 3. Namely, for each user participated in an A/B test, we predict her future engagement based only on the data observed during the A/B test's period by means of the best decision tree model from Sec. 5, that have been trained in advance[17]. We consider forecast time periods up to the 28-th day since the experiment start, and, hence, we get several novel modifications for each metric $\mathtt{M} \in \mathfrak{M}$. In this section, we refer to them as $\mathtt{FUBPA}_{X \to Y}$, where X is the considered duration of the experiment in days and Y is the length of the forecast period in days. For instance, $\mathtt{FUBPA}_{7 \to 21}$ is the evaluation metric, which predicts (based on the first 7 days of the experiment) the value of the considered metric M over 21 days since the experiment start, as if it was conducted for 21 days. Additionally, we predicted the metric value at the post-experiment periods. Such metric modifications, which are referred to as $\mathtt{FUBPA}^{post}_{X \to Y}$, predict the value of M over $(Y - X)$ days after the end of X-day experiment period.

A/A tests. First of all, we check our metrics against 23 A/A tests. An A/A test should be failed about 5% of the time for the p-value threshold 0.05 [18, 4], which is used in our work. Each of the metrics C, PT, CpQ, each of all their FUBPA modifications, and each of the post-experiment modifications (i.e., $\mathtt{FUBPA}^{post}_{X \to Y}$) of ATpS failed one A/A test. All other metrics and all other modifications did not fail any A/A test. Hence, all our metrics and all our modifications have an acceptable rate of A/A fails.

Example. We start our study from consideration of an example A/B experiment. It evaluates a treatment, which is an improvement of the ranking algorithm of the search engine. We look at the absence time metric ATpS. The results for this experiment are presented in Fig. 4. We see that the metric does not detect the treatment effect during the first 7 days of the experiment. However, its modification $\mathtt{FUBPA}_{4 \to 17}$ shows a statistically significant difference -0.4%

[17] They have been trained on the data obtained during the time period of February and March, 2013, before all our A/B experiments that were conducted from April, 2013 to September, 2014.

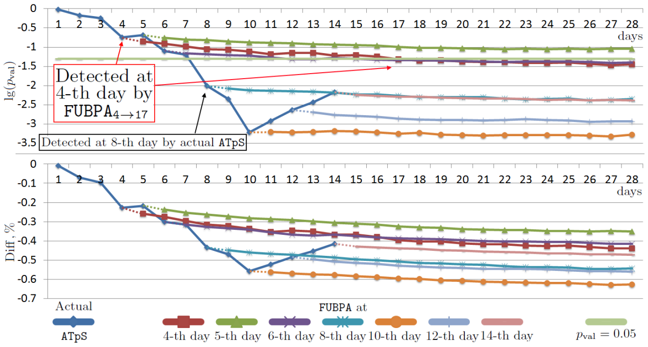

Figure 4: The Diff **and** p_{val} **of** ATpS **observed during an example A/B test and of the estimations of** ATpS **by the** $\mathtt{FUBPA}_{X\to Y}$ **with different values of** X **and** Y**.**

($p_{\text{val}} \leq 0.048$). Thus, *the treatment effect is correctly*[18] *detected at the 4-th day since the experiment start by means of FUBPA, while the baseline metric* ATpS *does not detect the effect till the 8-th day* (i.e., one saves 50% of the time for this example). Note that the signs of the actual and the predicted differences Diff are the same. Moreover, it is observed for all FUBPA modifications. The magnitudes of the predicted Diff for $\mathtt{FUBPA}_{X\to Y}$ are also of the same order. However, their exact values are sometimes far from the actual ones and depend noticeably on the difference observed at the X-day for the actual metric. This finding correlates with the fact that the information from the last days of the observation period has a large impact on the prediction quality (see Sec. 5).

Overall detection of the treatment effect. Next, we discuss the results of application of our approach to our six user engagement metrics $\mathfrak{M}$ in order to improve sensitivity of 32 studied A/B experiments with 14-day duration. Table 4 summarizes the number of A/B experiments, whose treatment effect is detected (i.e., $p_{\text{val}} \leq 0.05$) by each of the actual key metrics and each of its FUBPA modifications[19]. The best result in each column is underlined. Additionally, the number of A/B experiments, whose treatment effect is detected by a considered modification and is not detected by the corresponding baseline metric, is indicated in brackets. First, we see that the metrics based on the number of sessions measure S (i.e., the actual per-user metric and its FUBPA modifications) detect the treatment effects in a fewer number of A/B tests than the ones based on the absence time measure ATpS, which is assumed to be a novel alternative to S [9]. Thus, *the absence time appears to be more sensitive than the state-of-the-art number of sessions per user*. Second, the click-based metrics (i.e., the number of clicks per user C and the number of clicks per query per user CpQ) are more sensitive than the others. This observation is expected and exactly correlates with the "rule of thumb" #5 in [16], which states that clicks are easy to shift, while the number of sessions is hard to change.

Third, one could note that, for instance, the actual metric C detects the treatment effects in 8 A/B tests, while its modification $\mathtt{FUBPA}^{post}_{14\to21}$ detects the ones in 8 A/B tests also, but one of them is new (i.e., it is not detected by C). It means that a FUBPA modification does not always detect the treatment effect in a test, where the baseline metric detects it. On the one hand, if the change of the service is already detected in such experiments, then we do not need to utilize sensitivity improvement approaches. On the other hand, one could use the FUBPA technique to ensure that the treatment effect will not disappear in the future (see further in this section). Nonetheless, in all those treatment effects that were detected both by a baseline metric and its FUBPA modifications, the sign and the magnitude of the relative difference value Diff are *the same*. This means that our modifications at least do not harm the decision taken after a successful experiment: to accept the evaluated change of the web service, or to reject it. In total, the baseline metrics detected the treatment effects in 17 A/B experiments (i.e., $p_{\text{val}} \leq 0.05$ for at least one of the metrics), while, *after applying the FUBPA, we additionally detect the treatment effect in* 3 *tests.* Thereby, the FUBPA metrics increase this number from 53.125% to 62.5%, w.r.t. the number of all A/B tests. Thus, we conclude that *our FUBPA technique improves the sensitivity of the studied user engagement metrics and helps us to detect the treatment effect in more online controlled experiments.*

Table 4: The number of A/B tests whose treatment effect is detected by each UE metric and its FUBPA modifications.

Metric	S	Q	C	PT	CpQ	ATpS
Actual@14 day	2	$\underline{1}$	8	$\underline{4}$	$\underline{15}$	$\underline{4}$
$\mathtt{FUBPA}_{14\to15}$	$\underline{3}$ (+1)	$\underline{1}$	$\underline{9}$ (+1)	$\underline{4}$	14	$\underline{4}$
$\mathtt{FUBPA}_{14\to18}$	$\underline{3}$ (+1)	0	$\underline{9}$ (+1)	$\underline{4}$	14	$\underline{4}$
$\mathtt{FUBPA}_{14\to21}$	$\underline{3}$ (+1)	0	$\underline{9}$ (+1)	$\underline{4}$	$\underline{15}$ (+1)	$\underline{4}$
$\mathtt{FUBPA}_{14\to25}$	$\underline{3}$ (+1)	0	$\underline{9}$ (+1)	$\underline{4}$	$\underline{15}$	$\underline{4}$
$\mathtt{FUBPA}_{14\to28}$	$\underline{3}$ (+1)	0	$\underline{9}$ (+1)	3	14	$\underline{4}$
$\mathtt{FUBPA}^{post}_{14\to21}$	2	0	8 (+1)	2	8	1
$\mathtt{FUBPA}^{post}_{14\to28}$	2	0	6	3	8	3

Discover of future metric sensitivity. Let us see how the FUBPA detects the actual future sensitivity of a baseline metric. We consider the metric CpQ, which detected the largest number of treatment effects. For these purposes, we consider our 32 A/B experiments as if they ran only for the first 7 days of their duration, and, based on these observations, we apply the FUBPA to our baseline metric. Further, we *consider only those A/B tests* that *have no detected* treatment effect w.r.t. the baseline metric over 7 days. For each FUBPA modification and for each A/B test, we look at the FUBPA modification and check if it expects appearing of the treatment effect in the future measurement of the baseline metric (i.e., in the future 7 days, 14 days, etc., depending on the type of the FUBPA). After that, we check it against the actual effect in the baseline metric measured for the considered A/B test over the 14 days of its duration. Finally, Table 5 summarizes this information over considered A/B tests and the baseline metric CpQ obtaining the statistics for each type of the FUBPA modification. Thus, we are interested in how good our approach is at prediction of the detection of the treatment effect in the future (i.e., at the 14-th day of an A/B test) in the case of the non-significant difference observed at the current moment (i.e., at the 7-th day of an A/B test). From Table 5, one sees that almost all FUBPA modifications have very good positive and negative predictive values (i.e., Col. 1 and Col. 4). So, one could conclude that *our FUBPA technique can be used in practice as an additional flag* to make the decision (e.g., at a half of time period of an A/B test which had not detected a

[18] The absence time should decrease for an improvement [9].

[19] We present several representative forecast period lengths Y, because the results for all the other are similar to them.

Table 5: The prediction of appearing of the treatment effect of 17 A/B tests w.r.t. CpQ.

Actual 14-day effect:	appears (1)		none (16)	
FUBPA expects:	appears	none	appears	none
$\mathtt{FUBPA}_{7\to 8}$	0 (0%)	1	1	15 (94%)
$\mathtt{FUBPA}_{7\to 11}$	0 (0%)	1	1	15 (94%)
$\mathtt{FUBPA}_{7\to 14}$	0	1	0	16 (94%)
$\mathtt{FUBPA}_{7\to 21}$	0	1	0	16 (94%)
$\mathtt{FUBPA}_{7\to 28}$	0 (0%)	1	1	15 (94%)
$\mathtt{FUBPA}^{post}_{7\to 14}$	0 (0%)	1	1	15 (94%)
$\mathtt{FUBPA}^{post}_{7\to 21}$	1 (50%)	0	1	15 (100%)
$\mathtt{FUBPA}^{post}_{7\to 28}$	1 (50%)	0	1	15 (100%)

Table 6: The persistence of the treatment effect of 15 A/B tests w.r.t. CpQ.

Actual 14-day effect:	remains (14)		disappears (1)	
FUBPA expects:	remains	disappears	remains	disappears
$\mathtt{FUBPA}_{7\to 8}$	13 (93%)	1	1	0 (0%)
$\mathtt{FUBPA}_{7\to 11}$	13 (93%)	1	1	0 (0%)
$\mathtt{FUBPA}_{7\to 14}$	13 (93%)	1	1	0 (0%)
$\mathtt{FUBPA}_{7\to 21}$	14 (93%)	0	1	0
$\mathtt{FUBPA}_{7\to 28}$	14 (93%)	0	1	0
$\mathtt{FUBPA}^{post}_{7\to 14}$	8 (**100%**)	6	0	1 (14%)
$\mathtt{FUBPA}^{post}_{7\to 21}$	9 (**100%**)	5	0	1 (**17%**)
$\mathtt{FUBPA}^{post}_{7\to 28}$	9 (**100%**)	5	0	1 (**17%**)

treatment effect at this time): to stop the experiment (and, hence, *save some experimentation platform's resources for other A/B tests*, e.g., a fraction of an experimentation traffic) or to continue it.

Control of the current effect persistence. On the contrary, we can consider the opposite situation. Suppose that a key metric detects the treatment effect at the 7-th day of an A/B test, then we are interested to know if this effect remains at the 14-th day or disappears. This case is very important in practice, because the statistically significance difference in the first days of an A/B experiment may be caused by the *Primacy* and *Novelty* effects [18, 14, 16], and, as a result, the true treatment effect might be delayed or be absent at all. In order to check the persistence of the treatment effect observed at the 7-th day to the 14-th day, we apply our FUBPA technique in the same way as in the previous paragraph, but we *consider only those A/B tests* that *have detected* treatment effect w.r.t. to the baseline metric over 7 days. We present these results for predicting the current treatment effect persistence in Table 6 for the metric CpQ. First, one sees that all types of the FUBPA have a very good precision (i.e., $\geq$ 93%, see Col. 1), however the negative predictive values (i.e., Col. 4) are very low, i.e. the biggest one is 17%. The recall values (Col. 1 divided by the sum of Col. 1 and Col. 2) for all types of the FUBPA are higher than 57%, and the best recall value 93% is observed for $\mathtt{FUBPA}_{7\to Y}$, $Y = 8, 11$, and 14. So, one could conclude that *our FUBPA technique can be used in practice as an additional flag* to make the decision after obtaining a treatment effect earlier than the predefined experiment duration: to continue the experiment (e.g., *to ensure the result against the Primacy and Novelty effects*), or to stop the experiment (and, hence, *save some experimentation platform's resources for other A/B tests*, e.g., a fraction of an experimentation traffic).

Possible extensions and combinations with other techniques. Our future user behavior prediction approach could be applied in many cases and could be combined with other sensitivity improvement techniques. For example, we could combine the FUBPA with the stratification technique, which is proposed in [7]. On the one hand, we could directly apply stratification of users (e.g., w.r.t. the used browser) to the FUBPA modification of the key metric. On the other hand, we could stratify users in the training set of the user behavior predictor and obtain an individual predictor for each stratum (e.g., own predictor for each user preference to a particular browser). After that, during an A/B test, we would apply the proper predictor for each stratum.

Besides, a FUBPA modification of a metric could be regarded as a self-sufficient evaluation metric, because it is the average value of a user's feature, which is calculated based on the data observed in the experiment period solely (as it is shown in Sec. 3). The only difference between this feature and a simple measure, like the number of sessions, consists in that the first one is calculated in a sophisticated way by means of the predictor. Therefore, we can apply any previously proposed approaches of sensitivity improvement, treating a FUBPA modification as a key metric. For instance, one can (a) remove those users from the treatment group, who were not affected by the evaluated change [18, 5]; (b) transform or cap this metric [16]; and (c) consider this metric in two-stage controlled experiments, applying proper variation reduction techniques, that are proposed in [6]. Finally, note, that some techniques (e.g., the point (b)) could be applied during the prediction training stage (i.e., train a future predictor of an improved metric).

7. CONCLUSIONS AND FUTURE WORK

In our work, we consider the problem of prediction of user engagement in terms of exact values of several state-of-the-art UE metrics. To the best of our knowledge, this problem is assumed to be novel and never investigated in existing studies. We performed a deep study of the influence on the prediction quality of utilized engagement measures and their transformations. Then, we applied the obtained predictor for the purpose of improving the sensitivity of online controlled experiments. We found that this approach increases the number of online controlled experiments with detected a treatment effect. The evaluations on real online experiments run at Yandex show that the approach can be used to detect the treatment effect of an A/B test faster (up to 50% of saved time) with the same level of statistical significance. Our technique can also be used in practice as an additional flag to make the decision during an A/B test to continue or to stop it based on the probabilities of obtaining the significant treatment effect. Hence, the results of our study coincide with the emerging needs of modern web companies to run more and fast controlled experiments on a limited number of their users.

Future work. We believe that our sensitivity improvement approach will be of interest to researchers and practitioner in online controlled experiments. As future work we can, first, extend the set of user engagement measures and other user behavior data for further improvement of the user engagement prediction quality and, hence, the controlled experiment sensitivity. Second, we can study more sophisticated prediction models to be applied in our sensitivity improvement approach. Finally, it would be interesting to combine different sensitivity improvement techniques with our approach.

8. REFERENCES

[1] E. Bakshy and D. Eckles. Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods. In *KDD'2013*, pages 1303–1311, 2013.

[2] C. Bandt and B. Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.

[3] S. Chakraborty, F. Radlinski, M. Shokouhi, and P. Baecke. On correlation of absence time and search effectiveness. In *SIGIR'2014*, pages 1163–1166, 2014.

[4] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In *KDD'2009*, pages 1105–1114, 2009.

[5] A. Deng and V. Hu. Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *WSDM'2015*, pages 349–358, 2015.

[6] A. Deng, T. Li, and Y. Guo. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *WWW'2014*, pages 609–618, 2014.

[7] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM'2013*, pages 123–132, 2013.

[8] A. Drutsa, G. Gusev, and P. Serdyukov. Engagement periodicity in search engine usage: analysis and its application to search quality evaluation. In *WSDM'2015*, pages 27–36, 2015.

[9] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *WSDM'2013*, pages 173–182, 2013.

[10] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 2001.

[11] V. Hu, M. Stone, J. Pedersen, and R. W. White. Effects of search success on search engine re-use. In *CIKM'2011*, pages 1841–1846, 2011.

[12] B. J. Jansen, A. Spink, and V. Kathuria. How to define searching sessions on web search engines. In *Advances in Web Mining and Web Usage Analysis*, pages 92–109. Springer, 2007.

[13] R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, and T. Melamed. Online experimentation at microsoft. *Data Mining Case Studies*, page 11, 2009.

[14] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD'2012*, pages 786–794, 2012.

[15] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *KDD'2013*, pages 1168–1176, 2013.

[16] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *KDD'2014*, 2014.

[17] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD'2007*, pages 959–967, 2007.

[18] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[19] R. Kohavi, D. Messner, S. Eliot, J. L. Ferres, R. Henne, V. Kannappan, and J. Wang. Tracking users' clicks and submits: Tradeoffs between user experience and data loss, 2010.

[20] J. Lehmann, M. Lalmas, G. Dupret, and R. Baeza-Yates. Online multitasking and user engagement. In *CIKM'2013*, pages 519–528, 2013.

[21] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret. Models of user engagement. In *User Modeling, Adaptation, and Personalization*, pages 164–175. Springer, 2012.

[22] E. T. Peterson. *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram, 2004.

[23] S. M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[24] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[25] K. Rodden, H. Hutchinson, and X. Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *CHI'2010*, pages 2395–2398, 2010.

[26] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[27] Y. Song, X. Shi, and X. Fu. Evaluating and predicting user engagement change with degraded search relevance. In *WWW'2013*, pages 1213–1224, 2013.

[28] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD'2010*, pages 17–26, 2010.

[29] W. W.-S. Wei. *Time series analysis*. Addison-Wesley Redwood City, California, 1994.

[30] R. W. White, A. Kapoor, and S. T. Dumais. Modeling long-term search engine usage. In *User Modeling, Adaptation, and Personalization*, pages 28–39. Springer, 2010.