Have It Both Ways—From A/B Testing to A&B Testing with Exceptional Model Mining

Wouter Duivesteijn[1(✉)], Tara Farzami[2], Thijs Putman[2], Evertjan Peer[1], Hilde J. P. Weerts[1], Jasper N. Adegeest[1], Gerson Foks[1], and Mykola Pechenizkiy[1]

[1] Technische Universiteit Eindhoven, Eindhoven, the Netherlands
{w.duivesteijn,m.pechenizkiy}@tue.nl,
{e.peer,h.j.p.weerts,j.n.adegeest,g.foks}@student.tue.nl
[2] StudyPortals B.V., Eindhoven, the Netherlands
{tara,thijs}@studyportals.com

Abstract. In traditional A/B testing, we have two variants of the same product, a pool of test subjects, and a measure of success. In a randomized experiment, each test subject is presented with one of the two variants, and the measure of success is aggregated per variant. The variant of the product associated with the most success is retained, while the other variant is discarded. This, however, presumes that the company producing the products only has enough capacity to maintain one of the two product variants. If more capacity is available, then advanced data science techniques can extract more profit for the company from the A/B testing results. Exceptional Model Mining is one such advanced data science technique, which specializes in identifying subgroups that behave differently from the overall population. Using the association model class for EMM, we can find subpopulations that prefer variant A where the general population prefers variant B, and vice versa. This data science technique is applied on data from StudyPortals, a global study choice platform that ran an A/B test on the design of aspects of their website.

Keywords: A/B testing · Exceptional Model Mining · Association
Online controlled experiments · E-commerce · Website optimization

1 Introduction

A/B testing [20] is a form of statistical hypothesis testing involving two versions of a product, A and B. Typically, A is the control version of a product and B represents a new variation version, considered to replace A if it proves to be more successful. An A/B test requires two further elements: a pool of test subjects, and a measure of success. Each test subject in the pool is presented with a randomized choice between A and B. The degree to which this product version is successful with this test subject is measured. Having collected results over the full pool of test subjects, the success degree is aggregated per version. Subsequently, a decision is made whether the new variation version B is a

© Springer International Publishing AG 2017
Y. Altun et al. (Eds.): ECML PKDD 2017, Part III, LNAI 10536, pp. 114–126, 2017.
https://doi.org/10.1007/978-3-319-71273-4_10

(substantial) improvement over the control version A. For making this decision, a vast statistical toolbox is available [6,7].

Since the rise of the internet, A/B tests have become ubiquitous. It is a simple, cheap, and reliable manner to assess the efficacy of the redesign of a web page. Running two versions of a web page side by side is not too intrusive to your online business, and standard web analytics suites will tell you all you need to know on which of the versions deliver the desired results. In fact, through proper web analytics tools, we can obtain substantially more information on the factors that influence the success of versions A and B.

Having performed an A/B test, the standard operating procedure is the following. An assessment is made whether the new variation version B performs (substantially) better than the current control version A. From that assessment, a hard, binary decision is made: either version A or version B is the winner. The loser is discarded, and the winner becomes the standard version of the web page that is rolled out and presented to all visitors from this moment onwards. There is beauty in the simplicity, and this 'exclusive or' procedure inspires the slash in the name of the A/B test.

For large companies, making such a coarse decision leaves potential unused. If you own a high-traffic website, then even a small increase in click-through rate gets multiplied by a large volume of visitors, which results in a vast increase in income. It makes sense to use the traditional conclusion of an A/B test to determine the default page that should be displayed to a visitor of which we know nothing. But it is not uncommon to have some meta-information on the visitors to your website: which language setting does their browser have, which OS do they use, in which country are they located, etcetera. If we can identify subpopulations of the dataset at hand, defined in terms of such metadata, for which the A/B test reaches the opposite conclusion from the general population, then we can generate more revenue with a more sophisticated strategy: we maintain both versions of the web page, and present a visitor with either A or B depending on whether they belong to specific subgroups. Rather than choosing either A or B, we can instead choose to have it both ways: this paper turns the A/B test into an A&B test.

2 Related Work

First, we provide a brief summary on the current state of the art in mining of A/B testing results. Thus we explain how our problem formulation is different from existing body of work. Then we overview relevant research in the areas of local pattern mining and exceptional model mining that motivate our approach for the chosen problem formulation.

2.1 Utility of A/B Testing

In a marketing context, A/B testing has been studied extensively [20]. Analysis of the results from an A/B test has made it to the Encyclopedia of Machine

Learning and Data Mining [6], and an extensive survey on experiment design choices and results analysis is available [7]. This last paper encompasses a discussion of accompanying A/B tests with A/A tests to establish a proper baseline, extending the test to the multivariate case (more than two product versions), result confidence intervals, randomization methods to divide the test subjects fairly over the versions, sample size effects, overlapping experiments, and the effect of bots on the process. Regardless of the setting of all of these facets, the goal of A/B testing always remains to make a crisp decision at the end, selecting either A or B and discarding the alternative(-s).

If the main business goal is to increase the average performance with respect to e.g. a click through rate (CTR) rather than really find our whether A or B is statistically significantly better, then the Contextual Multi-Armed Bandits (cMAB) is the commonly considered alternative optimization approach to A/B testing. cMABs help to address an exploration-exploitation trade-off: using, i.e. exploring effectiveness, of A and B provides feedback about its effectiveness (exploration), but collecting that feedback on both A and B is an opportunity cost of exploitation, i.e. using one of the variants we already know is effective. To balance exploration with exploitation lots of policy learning bandit algorithms were considered, particularly in web analytics, e.g. [22,23].

In data mining for user modeling and convergence prediction two related problem formulations have been studied – predictive user modeling with actionable attributes [26] and uplift prediction [18]. While in traditional predictive modeling, the goal is to learn a model for predicting accurately the class label for unseen instances, in targeting applications, a decision maker is interested not only to generate accurate predictions, but to maximize the probability of the desired outcome, e.g. user clicking. Assuming that possibly neither of marketing actions A and B is always best, the problem can be formulated as learning to choose the best marketing action at instance level (rather than globally).

The paper that you are currently reading does not have a mission to promote either A/B testing or cMABs or uplift prediction; we merely observe that A/B tests are performed anyway, and strive to help companies performing such tests to learn more actionable insight from their data that would allow to domain experts to decide whether to stay with A, or switch to B or use both A and B, each for a particular context or customer segment.

2.2 Local Pattern Mining

The subfield of Data Mining with which this paper is concerned is Local Pattern Mining [4,17]: describing only part of the dataset at hand, while disregarding the coherence of the reminder. The Local Pattern Mining subtask that is particularly relevant here, is Theory Mining [15], where subsets of the dataset are sought that are *interesting* in some sense. Typically, not just any subset is sought. Instead, the focus is on subsets that are easy to interpret. A canonical choice to enforce that is to restrict the search to subsets that can be described as a conjunction of a few conditions on single attributes of the dataset. Hence, if the dataset concerns people, we would find subsets of the form

"Age $\geq 30 \wedge$ Smokes = yes $\Rightarrow$ (interesting)". Such subsets are referred to as *subgroups*. Limiting the search to subgroups ensures that the results can be interpreted in terms of the domain of the dataset at hand; the resulting subgroups represent pieces of information on which a domain expert can act.

Many choices can be made to define 'interesting'. One such choice is to make this a supervised concept: we set apart one attribute of the dataset as the *target*, and seek subsets that feature an unusual distribution of that target. This is known as Subgroup Discovery (SD) [9,11,25]. In the running example of a dataset concerning people, if the target would be whether the person develops lung cancer or not, SD would find results such as "Smokes = yes $\Rightarrow$ Lung cancer = yes". This of course does not mean that all smokers fall in the 'yes' category; it merely implies a skew in the target distribution.

2.3 Exceptional Model Mining

Exceptional Model Mining (EMM) can be seen as a generalized form of SD. Instead of singling out one attribute of the data as the target, in EMM one typically selects several target attributes. The exceptionality of a subgroup is no longer evaluated in terms of an unusual distribution of the single target, but instead in terms of an unusual interaction between the multiple targets. This interaction is captured by some kind of modeling, which inspired the name of EMM. Exceptional Model Mining was first introduced in 2008 [13]. An extensive overview of the *model classes* (types of interaction) that have been investigated can be found in [3]; as examples, one can think of an unusual correlation between two targets [13], an unusual slope of a regression vector on any number of targets [2], or unusual preference relations [19].

Algorithms for EMM include a form of beam search [3] that works for all model classes, a fast sampling-based algorithm for a few dedicated model classes [16], an FP-Growth-inspired tree-based exhaustive algorithm that works for almost all model classes [14], a tree-constrained gradient ascent algorithm for linear models using sofy subgroup membership [10], and a compression-based method that improves the resulting models at the cost of interpretability [12].

3 The StudyPortals A/B Test Setting

Since the Bologna process contributed to harmonizing higher-education qualifications throughout Europe, locating (part of) one's study programme in another country than one's own has become streamlined. This offers opportunities for students to acquire international experience while still studying, which is something from which both the students and the higher education institutions can benefit. The harmonization of how higher education is structured enables a fair comparison of programmes across country boundaries.

Such a comparison being possible does not necessarily imply that it is also easy. In 2007, three (former) students identified that there was a hole in the information market, and they filled that hole with a hobby project that eventually resulted in StudyPortals [21].

3.1 StudyPortals

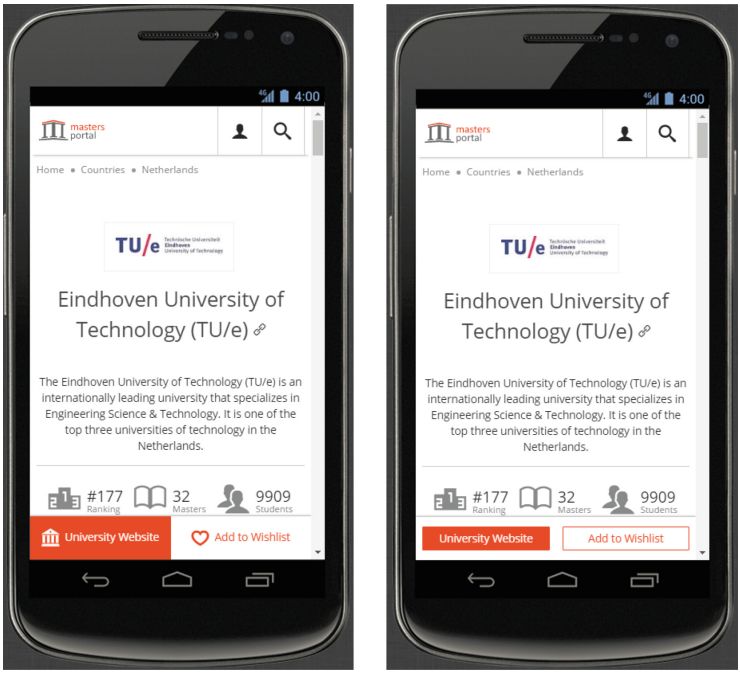
In 2007, two alumni from the Technische Universiteit Eindhoven and one from the Kungliga Tekniska Högskolan created MastersPortal: a central database for European Master's programmes. The goal was to become the primary destination for students wanting to study in Europe. In April 2008, the website presented 2 700 studies at 200 universities from 30 countries, and attracted 80 000 visits per month. Since then, the scope of the website has expanded. The subject ranges beyond Master's programmes, also encompassing Bachelor's and PhD programmes, short courses, scholarships, distance learning, language learning, and preparation courses. The website is no longer restricted to Europe, but expanded globally. In September 2016, MastersPortal presented 56 000 studies at 2 000 universities from 100 countries, and attracted 1.4 million unique sessions per month. The overarching company StudyPortals logged 14.5 million unique visitors in the first nine months of 2016, with 7 page views per second during the busiest hour of the year. This growth allows the company to employ 150 team members in five offices on three continents.

StudyPortals generates revenue from the visitors to their websites through the universities, who pay for activity on the pages presenting their programmes. A study programme's web page generates revenue in three streams: (1) Cost Per Mille (thousand page views); (2) Cost Per Lead; (3) Cost Per Click. The first revenue stream depends on the attractiveness of links towards the programme's web page. The second revenue stream depends on whether the person viewing the programme's web page fill their personal information in the university lead form. The design of a programme's web page has a low impact on these two revenue streams. The third revenue stream is the one that StudyPortals can influence directly through appropriate web page design.

3.2 The Third Revenue Stream and the A/B Test

Figure 1 displays the mobile version of a university's web page on the MastersPortal website. The orange button at the bottom left of the page links through to the website of the university itself. When a user clicks on that button, StudyPortals receives revenue in the Cost Per Click revenue stream. With the volume of web traffic StudyPortals experiences, a small increase in the click-through rate represents a substantial increase in income.

The advance of smartphones and tablets has vastly increased the importance of the mobile version of websites. These versions come with their own UI requirements and quirks. Figure 1a displays the page design that was in use in September 2016; having an orange rectangle that is clickable is one of those UI design elements that is typical of mobile websites as opposed to desktop versions. However, the website visitors, being human beings, are creatures of habit. They might prefer clickable elements of websites to resemble traditional buttons, as they remember from their desktop dwelling times. To test this hypothesis, StudyPortals designed an alternative version of their mobile website (cf. Fig. 1b). These variants become the subject of our A/B test: the rectangular version is the control version A, and the more buttony version is the variation B.



(a) Control version (b) Buttony variation

Fig. 1. The A and B variants of the A/B test at hand: two versions of buttons on university profile pages of the mobile version of the MastersPortal website.

3.3 The Data at Hand

StudyPortals collected raw data on the A/B test results for a period of time. From this raw, anonymized data, a traditional flat-table dataset was generated through data cleaning and feature engineering. The full process is beyond the scope of this paper; it involved removing redundant information, removing the users that have seen both versions of the web page (as is customary in A/B testing), aggregating location information (available on city level) to country level, merging various versions of the distinct OSs (e.g., eight distinct versions of iOS were observed; these sub-OSs were flattened into one OS), etcetera. In the end, the columns in the dataset include device characteristics, location information, language data, and scrolling characteristics. The dataset spans 3 065 records.

Finally, we are particularly interested in two columns: the one holds the information with which version of the web page (A/B) the visitor was presented, and the other holds whether the visitor merely viewed or also clicked. The goal of traditional A/B testing is to find out whether version A or B leads to more clicks; the main contribution of this paper is to identify subpopulations where these two columns display an unusual interaction: can we find subgroups where the click rate interacts exceptionally with the web page version?

4 Data Science to Be Applied

Finding subsets of the dataset at hand where several columns of special interest interact in an unusual manner is the core task of Exceptional Model Mining (EMM). This interaction can be gauged in many ways. This section discusses the EMM framework and its specific instantiation for the problem at hand.

4.1 The Exceptional Model Mining Framework

EMM [3,13] assumes a flat-table dataset Ω, which is a bag of N records of the form $r = \{a_1, \ldots, a_k, t_1, \ldots, t_m\}$. We call the attributes $a_1, \ldots, a_k$ the *descriptors* of the dataset. These are the attributes in terms of which subgroups will be *defined*; the ones on the left-hand side of the $\Rightarrow$ sign in the examples of Sect. 2.2. The other attributes, $t_1, \ldots, t_m$, are the *targets* of the dataset. These are the attributes in terms of which subgroups will be *evaluated*; the most exceptional target interaction indicates the most interesting subgroup.

Subgroups are defined in terms of conditions on descriptors. These induce a subset of the dataset: all records satisfying the conditions. For notational purposes, we identify a subgroup with that subset, so that we write $S \subseteq \Omega$, and denote $|S|$ for the number of records in a subgroup. We also denote S^C for the complement of subgroup S in dataset Ω, i.e.: $S^C = \Omega \backslash S$.

To instantiate the EMM framework, we need to define two things: a model class, and a quality measure for that model class. The model class specifies what type of interaction we are interested in. This can sometimes be fixed by a single word, such as 'correlation'; it can also be a more convoluted concept. The choice of model class may put restrictions on the number and type of target columns that are allowed: if one chooses the regression model class [2], one can accommodate as many targets as one wishes, but if one chooses the correlation model class [13], this fixes the number of targets $m = 2$ and demands both those targets to be numeric. Once a model class has been fixed, we need to define a quality measure (QM), which quantifies exactly what in the selected type of interaction we find interesting. For instance, in the correlation model class, maximizing ρ as QM would find those subgroups featuring perfect positive target correlation, minimizing $|\rho$ would find those subgroups featuring uncorrelated targets, and maximizing $|\rho_S - \rho_{S^C}|$) would find those subgroups S for which the target correlation deviates from the target correlation on the subgroup complement S^C.

4.2 Instantiating the Framework: The Association Model Class

As alluded to in Sect. 3.3, the StudyPortals dataset comes naturally equipped with $m = 2$ nominal targets: t_1 is the binary column representing whether the page visitor merely viewed or also clicked, and t_2 is the binary column representing whether the visitor was presented with web page version A or B. Therefore, the natural choice of EMM instance would be the association model class [3, Sect. 5.2]. Essentially, this is the nominal-target equivalent of the correlation model class [13, Sect. 3.1]: we strive to find subgroups for which the association between view/click and A/B is exceptional.

Table 1. Target cross table

	View	Click
A	n_1	n_2
B	n_3	n_4

4.3 Instantiating the Framework: Yule's Quality Measure

Having fixed the model class, we need to define an appropriate quality measure. As has been observed repeatedly [3,13,19], one can easily achieve huge deviations in target behavior for very small subgroups. To ensure the discovery of subgroups that represent substantial effects within the datasets, a common approach is to craft a quality measure by multiplying two components: one reflecting target deviation, and one reflecting subgroup size.

The Target Deviation Component. For the quality measure component representing the target deviation, we build on the cells of the target contingency table, depicted in Table 1. Given a subgroup $S \subseteq \Omega$, we can assign each record in S to the appropriate cell of this contingency table, which leads to count values for each of the n_i such that $n_1 + n_2 + n_3 + n_4 = |S|$. From such an instantiated contingency table, we can compute Yule's Q [1], which is a special case of Goodman and Kruskal's Gamma for 2×2 tables. Yule's Q is defined as $Q = {}^{(n_1 \cdot n_4 - n_2 \cdot n_3)}/_{(n_1 \cdot n_4 + n_2 \cdot n_3)}$. A positive value for Q implies a positive association between the two targets, i.e. high values on the diagonal of the contingency table and low values on the antidiagonal. Hence, a positive value for Q indicates that people presented with web page variant B click the button more often than people presented with web page variant A. We denote by Q_S the value for Q instantiated by the subgroup S.

Analogous to the component developed for Pearson's ρ in the correlation model class [13, Sect. 3.1], we contrast Yule's Q instantiated by a subgroup with Yule's Q instantiated by that subgroup's complement: $\varphi_{\mathrm{Q}}(S) = |Q_S - Q_{S^C}|$. Hence, this component detects schisms in target interaction: subgroups whose view/click-A/B association is markedly different from the rest of the dataset.

The Subgroup Size Component. To represent subgroup size, we take the entropy function φ_{ef} as described in [13, Sect. 3.1] (denoted $H(p)$ there). The components rewards 50/50 splits between subgroup and complement, while punishing subgroups that either are tiny or cover the vast majority of the dataset.

Combining the Components: Yule's Quality Measure. Combining the components into an association model class quality measure is straightforward:

$$\varphi_{\mathrm{Yule}}(S) = \varphi_{\mathrm{Q}}(S) \cdot \varphi_{\mathrm{ef}}(S)$$

Multiplication is chosen to ensure subgroups score well on both components.

5 Experiments

On the entire dataset, Yule's Q has a value of $\varphi_{\mathrm{Q}}(\Omega) = -0.031$. Hence, the results of the traditional A/B test would be a resounding victory for variant A: the less buttony control version of Figure 1a generates more clicks than the more buttony variation of Fig. 1b. Whether the difference is significant is another question, but the new variation is clearly not significantly better than the already-in-place control version. In traditional A/B testing, that would be the end of the analysis: the new variant B does not outperform the current variant A, so we keep variant A and discard variant B. The main contribution of this paper is that with EMM, we can draw more sophisticated conclusions.

5.1 Experimental Setup

For empirical evaluation, we select the beam search algorithm for EMM whose pseudocode is given in [3, Algorithm 1], parametrized with $w = 10$ and $d = 2$. We have also trialed more generous values for the beam width w, which did not affect the results much. The search depth d is deliberately kept modest: this parameter controls the number of conjuncts allowed in a subgroup description, hence modest settings guarantee good subgroup interpretability.

The beam search algorithm, the association model class, and Yule's quality measure have been implemented in Python as part of a Bachelor's project in a course on Web Analytics. The code will be made available upon request. In the following section, we report the top-five subgroups found with the thusly parametrized and implemented EMM algorithm.

5.2 Found Subgroups

The top-five subgroups found are presented in Table 2, in order of descending quality. Subgroup definitions are provided along with the values for the compound quality measure φ_{Yule}, the value of the Yule's Q component on both the

Table 2. Top-five subgroups found with the association model class for Exceptional Model Mining. The subgroup definitions are listed along with their values for Yule's quality measure, the within-subgroup value for Yule's Q, the outside-subgroup value for Yule's Q, and the subgroup size.

Subgroup definition	$\varphi_{\mathrm{Yule}}(S)$	$\varphi_{\mathrm{Q}}(S)$	$\varphi_{\mathrm{Q}}(S^C)$	$\lvert S\rvert$
Browser_lang = EN-GB	0.1540	0.1287	−0.1172	979
Browser_lang = EN-GB ∧ Viewheight = small	0.1300	0.2852	−0.0722	363
Browser_lang = TR	0.0859	−1.0000	−0.0164	53
Browser_lang = EN-GB ∧ OS_name = iOS	0.0797	0.2661	−0.0599	204
Country = NG	0.0783	0.2000	−0.0554	281

subgroup and its complement, and the subgroup size. Recall that the total number of records in the dataset is 3 065, and the value for Yule's Q on the whole dataset is $\varphi_{\mathrm{Q}}(\Omega) = -0.031$.

The best subgroup found, S_1, is defined by people having British English set as their browser language. More extreme values for Yule's Q itself can be found elsewhere in the table; S_1 has other distinctive qualities. What sets it apart, is that there is a clear dichotomy in Q-values between subgroup and complement: the Q-value on S_1 is substantially (though not spectacularly) elevated from the behavior on the whole dataset, and *at the same time*, the Q-value on the S_1^C is substantially *depressed* from the behavior on the whole dataset. This means that people using British English as their browser language generate markedly more revenue when presented with version B of the web page, whereas people using any other browser language generate markedly more revenue when presented with version A of the web page. Moreover, S_1 has a substantial size. These two factors make S_1 the subgroup for which business action is most apposite: we have clearly distinctive behavior between two sizeable groups of website visitors, and presenting each group with the version of the web page appropriate for that group stands to substantially increase overall revenue.

The second- (S_2) and fourth-ranked (S_4) subgroups are specializations of S_1. S_2 specifies visitors that view the website using a relatively small mobile browser screen; they strongly prefer version B. Small screens can be found in relatively old smartphones, so this population contains people that are relatively slow in adopting new technology. It stands to reason that this population would also prefer a more traditionally-shaped button. S_4 specifies visitors that run the iOS operating system. They too strongly prefer version B, which is remarkable, since the buttons of version B do not conform to Apple's design standards. Perhaps the unusual button design draws more attention.

The third-ranked subgroup are those people that have set their browser language to Turkish. This subgroup may be too small to deliver actionable results, covering less than 2% of the dataset. However, the Q-value measured on this subgroup is strong: this subgroup displays a crystal clear preference for version A. This is a marked departure from the previously presented subgroups.

The final subgroup presented in Table 2, ranked fifth, concerns people from Nigeria. Yule's Q indicates that these people prefer version B. Given that the official language of Nigeria is English, the version preference is unsurprising: this subgroup overlaps substantially with S_1.

6 Conclusions

Having performed an A/B test—where a pool of test subjects are randomly presented with either version A or version B of the same product, a measure of success is aggregated by version, and the experimenter is presented with the results—the typical subsequent action is to make a crisp decision to either maintain the control version A, or replace it with the new variation version B, while the losing alternative is discarded. In this paper, we argue that that action can be overly coarse. Instead, we present an alternative approach: A&B testing.

The procedure of the A&B test is the exact same as that of a traditional A/B test, but the subsequent action is much more sophisticated. We analyze the results of the traditional A/B test with Exceptional Model Mining, to find coherent subgroups of the overall population that display an unusual response to the A/B test: the resulting subgroups feature an unusual association between the A/B decision and the measure of success at hand. Hence, while the general population might generate more revenue when presented with the one version, the resulting subgroups might generate more revenue when presented with the other version. If the company performing the A/B test can afford the upkeep of both versions, then knowledge of these subgroups can be invaluable.

As proof of concept, we roll out the A&B test on data generated by StudyPortals, an online information platform for higher education. From the results of the A/B test (cf. Fig. 1), we derive several subgroups displaying unusual behavior (cf. Table 2). The largest schism lies between people using British English as browser language ($\sim 1/3$ of the population, preferring version B), and people using any other browser language ($\sim 2/3$ of the population, preferring version A). In other words, the results suggest that British prefer buttony buttons.

A natural next step would be to verify empirically whether identified subgroups lead to effective personalization serving either A or B version to corresponding web portal visitors. Since it is common for StudyPortals and other companies to run a number of A/B testing experiments, and there is a motivation to provide personalized content and personalized layout, it is interesting to develop a framework for automation of website personalization based on findings of EMM. It would also make sense to extend this paper by refining the employed quality measure, incorporating the economics of the underlying decision problem directly [8].

While the main application within this paper lies in the context of web analytics, it is important to notice that the methodology of A&B testing is applicable on any controlled experiment. Hence, A&B testing is relevant in diverse fields such as medical research [5], education [24], etcetera. In future work, we plan to roll out A&B testing in clinical trials near you.

References

1. Adeyemi, O.: Measures of association for research in educational planning and administration. Res. J. Math. Stat. **3**(3), 82–90 (2010)
2. Duivesteijn, W., Feelders, A., Knobbe, A.J.: Different slopes for different folks – mining for exceptional regression models with cook's distance. In: Proceedings of KDD, pp. 868–876 (2012)
3. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional model mining – supervised descriptive local pattern mining with complex target concepts. Data Min. Knowl. Disc. **30**(1), 47–98 (2016)
4. Hand, D.J.: Pattern detection and discovery. In: Hand, D.J., Adams, N.M., Bolton, R.J. (eds.) Pattern Detection and Discovery. LNCS (LNAI), vol. 2447, pp. 1–12. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45728-3_1

5. Jakowski, M., Jaroszewicz, S.: Uplift modeling for clinical trial data. In: Proceedings of ICML 2012 Workshop on Machine Learning for Clinical Data Analysis (2012)
6. Kohavi, R., Longbotham, R.: Online controlled experiments and A/B tests. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning and Data Mining, pp. 1–8. Springer, New York (2016). https://doi.org/10.1007/978-1-4899-7502-7_891-1
7. Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M.: Controlled experiments on the web: survey and practical guide. Data Min. Knowl. Discov. **18**(1), 140–181 (2009)
8. Kleinberg, J., Papadimitrou, C., Raghavan, P.: A microeconomic view of data mining. Data Min. Knowl. Disc. **2**(4), 311–324 (1998)
9. Klösgen, W.: Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271 (1996)
10. Krak, T.E., Feelders, A.: Exceptional model mining with tree-constrained gradient ascent. In: Proceedings of SDM, pp. 487–495 (2015)
11. Lavrač, N., Kavšek, B., Flach, P.A., Todorovski, L.: Subgroup discovery with CN2-SD. J. Mach. Learn. Res. **5**, 153–188 (2004)
12. van Leeuwen, M.: Maximal exceptions with minimal descriptions. Data Min. Knowl. Discov. **21**(2), 259–276 (2010)
13. Leman, D., Feelders, A., Knobbe, A.: Exceptional model mining. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5212, pp. 1–16. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87481-2_1
14. Lemmerich, F., Becker, M., Atzmueller, M.: Generic pattern trees for exhaustive exceptional model mining. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 277–292. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_18
15. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Min. Knowl. Discov. **1**(3), 241–258 (1997)
16. Moens, S., Boley, M.: Instant exceptional model mining using weighted controlled pattern sampling. In: Blockeel, H., van Leeuwen, M., Vinciotti, V. (eds.) IDA 2014. LNCS, vol. 8819, pp. 203–214. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12571-8_18
17. Morik, K., Boulicaut, J.-F., Siebes, A. (eds.): Local Pattern Detection. Springer, Heidelberg (2005). https://doi.org/10.1007/b137601
18. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowl. Inf. Syst. **32**(2), 303–327 (2012)
19. Rebelo de Sá, C., Duivesteijn, W., Soares, C., Knobbe, A.: Exceptional preferences mining. In: Calders, T., Ceci, M., Malerba, D. (eds.) DS 2016. LNCS (LNAI), vol. 9956, pp. 3–18. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46307-0_1
20. Siroker, D., Koomen, P.: A/B Testing: The Most Powerful Way to Turn Clicks Into Customers. Wiley, Hoboken (2013)
21. StudyPortals. www.studyportals.com
22. Tang, L., Jiang, Y., Li, L., Li, T.: Ensemble contextual bandits for personalized recommendation. In: Proceedings of RecSys, pp. 73–80 (2014)
23. Tang, L., Rosales, R., Singh, A.P., Agarwal, D.: Automatic ad format selection via contextual bandits. In: Proceedings of CIKM, pp. 1587–1594 (2013)

24. Williams, J.J., Li, N., Kim, J., Whitehill, J., Maldonado, S., Pechenizkiy, M., Chu, L., Heffernan, N.: MOOClets: A Framework for Improving Online Education through Experimental Comparison and Personalization of Modules. Working Paper No. 2523265 (2014). http://tiny.cc/moocletpdf
25. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Komorowski, J., Zytkow, J. (eds.) PKDD 1997. LNCS, vol. 1263, pp. 78–87. Springer, Heidelberg (1997). https://doi.org/10.1007/3-540-63223-9_108
26. Žliobaitė, I., Pechenizkiy, M.: Learning with actionable attributes: attention - boundary cases! In: Proceedings of ICDM Workshops, pp. 1021–1028 (2010)