Does UI Labeling Data Quality Matter for Predicting Website Aesthetics

Elnur Abbasov, Maxim Bakaev [0000-0002-1889-0692]

Novosibirsk State Technical University, Pr. K. Marksa 20, Novosibirsk 630073, Russia
bakaev@corp.nstu.ru

Abstract. The adoption of today's data-intensive digital services relies on the overall user experience (UX), which is shaped not just by "hard" functionality, but also by "soft" subjective satisfaction. In the latter, aesthetic impression plays an important role (particularly since visually pleasing products are known to be perceived as more usable) and became a popular prediction objective for Machine Learning (ML) based user behavior models. Since datasets in the field of Human-Computer Interaction are generally too scarce for application of deep learning methods that could operate on raw website screenshots, they often undergo preliminary labeling. Although the common notion is that the quality of the labeling is important for the end quality of the predictive models, there were few attempts to quantify the effect. In a previous study, we unexpectedly found significant **negative** correlations between the input data quality and the models' quality for Aesthetics and Orderliness subjective impressions. Our current paper is dedicated to validating the findings with another 557 website screenshots, 31 human participants labeling them, and 22 participants verifying the quality of their work. The non-parametrical models (Nadaraya-Watson kernel regression) with feature selection demonstrated somehow better performance, and the combined dataset better aligned with the expected effect of the labeling quality. Although our overall results are inconclusive, they might be of interest to ML practitioners and web designers who seek to automate the prediction of UX dimensions.

Keywords: Data Quality; Machine Learning; Image Recognition, User Experience

1 Introduction

Digital transformation is a constantly evolving field that influences all parts of our live. With the rise of technology, it has become crucial for businesses and organizations to adapt to this new digital landscape. However, simply adopting new technologies is not enough to drive successful transformation: "The development of evidence-based methods of management decisions and the transition to data-driven management, the evaluation of the effectiveness of state programs and public policy measures require high-quality data, as well as the ability to link data from different sources" [1].

Indeed, it is important to use quality data in order to create and develop digital products, and e-government is no exception [2]. In particular, user behavior-related data is essential for designing appealing websites and mobile applications. Of the

subjective perception dimensions, aesthetic impression has been in focus lately [3]. Web aesthetics modeling involves designing websites that are visually appealing to users while maintaining functionality and usability. It is based on the premise that visual appearance of a website can significantly affect user engagement and subjective satisfaction level after the interaction. Correspondingly, user behavior models (UBMs) that specify the influencing factors or just predict aesthetic impression based on a graphical user interface (UI) image see intensive development [4]. However, collecting enough data of appropriate quality remains a challenge in the field [5], while their positive effect on the end quality of the models remains unmeasured.

In our previous research [6], we explored the relation between the quality of the input data produced by 11 UI labelers and the quality of the ensuing UBMs constructed for the assessed Complexity, Aesthetics and Orderliness subjective scales. Rather unexpectedly, we found statistically significant **negative** correlations for Aesthetics and Orderliness, which suggested that the technically neglectable labelers supplied the data that were more beneficial with respect to predicting users' subjective perception of web UIs.

In the current study, our goal is to validate the previous results with a different dataset consisting of 557 webpage screenshots. The UI elements in them were labeled by 31 participants, and then the verification by another 22 participants was used to evaluate the quality of the input data. In the current study, we added non-parametrical Nadaraya-Watson kernel regression in addition to the linear regression used in [6], but the results are still inconclusive, as we could not find a significant correlation with the end quality for either group of the models.

The remaining part of the paper is organized as follows. In Section 2, we describe the experimental study, which involved the three major components: subjective assessment, labeling and verification. In Section 3, we analyze the experimental data and construct 58 parametrical and non-parametrical models. In the final section, we summarize and discuss our findings and specify limitations of our study and plans for further research.

2 The Experiment Description

2.1 Material

The material in our experiment was screenshots of website homepages belonging to 7 domains: culture, food, games, government, health, news, universities. The scope was the entire world, but we only used their English versions of the websites. We used a dedicated Python script to automatically collect 10639 screenshots in PNG format [7], of which 557 were manually selected for the experiment (we hereafter refer to them as web UIs). In order to enhance the variety of UI components, complete web pages were captured in the screenshots rather than solely focusing on the above-the-fold section or using a predetermined size.

2.2 Procedure

The UI Aesthetics Assessment. The subjective evaluations of the participants for each UI were obtained through a specialized online survey (see [7] for details). To assess the subjects' visual aesthetics impressions of a website, Likert ratings were used (1 – lowest, 7 – highest). The participants were asked to provide their honest subjective evaluations, as there were no correct or incorrect answers. Screenshots were assigned to each participant in a randomized sequential order, with the default number of UIs per participants being 50 or 100.

The UI Labeling. The labelers used `Crowd HTML Elements` library, provided by Amazon Mechanical Turk (AMT). The `crowd-bounding-box` widget on MTurk displays the screenshot, providing zoom and pan functionality, along with keyboard shortcuts for creating bounding boxes of various types to quickly label numerous objects. Crowd workers could preview and skip HITs as needed.

The written instruction of the labeling process was given to the participants, who were then asked to label as many UI elements as possible for each UI. The screenshots were distributed among them fairly equally, however, there was no random assignment. For each UI element they would mark out with a rectangular, the labelers were asked to choose one of the 10 pre-defined classes: interactive (button, check, input, link, dropdown, navigation), non-interactive (image, background image) or container objects (table, panel). It is important to note that in our first experiment there were 20 pre-defined classes [6].

The Labeling Verification. The verifiers were asked to choose for each UI element whether the labeling was *correct* or *incorrect* and then subjectively assess the completeness of the labeling for each UI (i.e. whether the labeler had marked out all the visible UI elements). All of the necessary instructions for making the correct/incorrect decisions were provided to the verifiers, and they were briefed on how the labeling process was supposed to be performed.

Fig. 1 demonstrates an example of a "good" website labeling from [6] (a fragment of the website screenshot loaded into LabelImg tool): it was verified as having SC = 100% and Precision = 100%. Fig. 2, on the contrary, presents an example of a clearly "neglectable" labeling: with SC = 20% (a major image in the center is ignored, while the *text* class elements should have been labeled individually) and Precision = 33% (note the inaccurate borders around the UI elements).

2.3 Subjects

The aforementioned activities were carried out by three groups of human participants, who were mostly Bachelor's students of Novosibirsk State Technical University (NSTU):

1. *The UI assessment* was done by 137 participants (67 females, 70 males), whose age ranged from 17 to 46 (mean 21.18, SD = 2.68).
2. *The UI labeling* was performed by 31 participants (14 male, 17 female), with the age ranging from 20 to 22 (mean 21.03, SD = 0.5).

3. *The verification* of the labelers' output was performed by another 22 participants (20 male, 2 female), whose age ranged from 20 to 22.

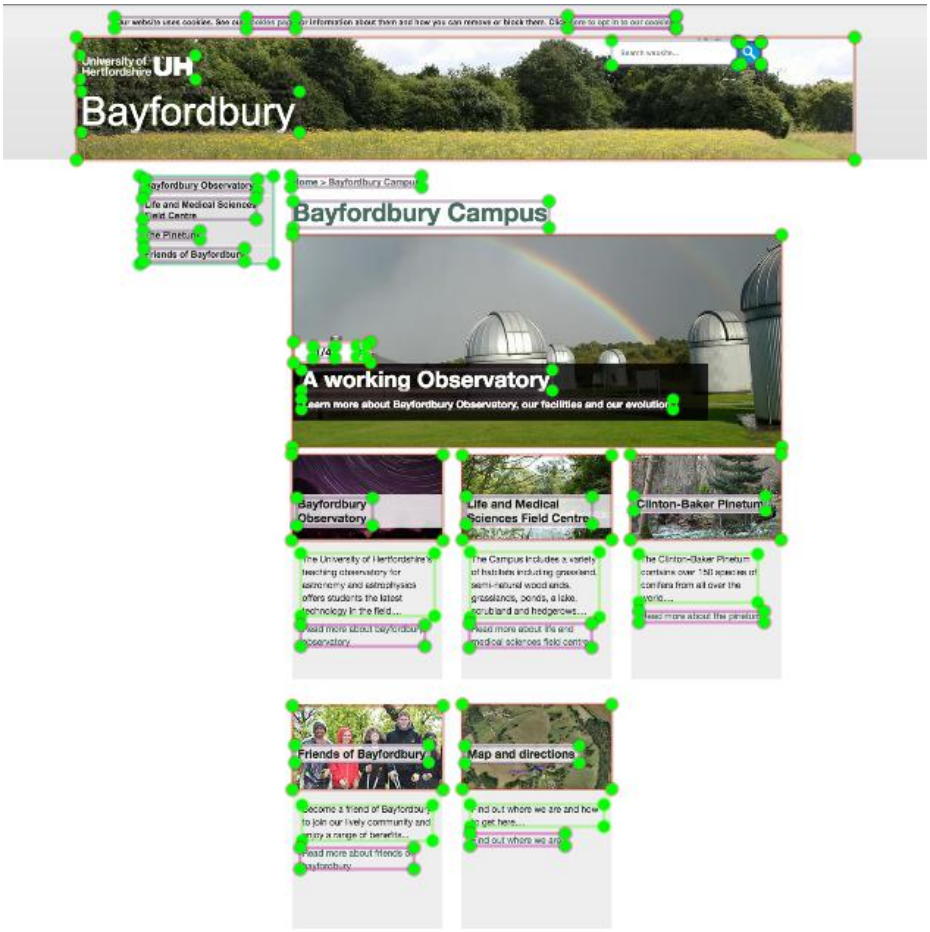

Fig. 1. An example of a "good" labeling: both Subjective Completeness and Precision are high.

2.4 Design and Modelling

In our previous research work [6], we only used linear regression (LR). So, the somehow controversial results we have got might be due to the common parametric models' problems, such as multicollinearity, heteroscedasticity and autocorrelation. The initial data included 24 factors extracted from web UIs, of which 8 factors were manually selected: number of UI elements, number of images, share of the text elements' area, share of whitespace, etc. However, the average number of UIs processed by each labeler in the current study was only 23.7 (SD = 8.71), unlike 44.3 (SD = 3.41) in the previous one. On the other hand, now we had 31 labelers compared to the

previous 11. Thus, in the current study we decided to conduct feature selection and use Nadaraya-Watson kernel regression (KR) to minimize the effect of the violation of the Gauss-Markov assumptions [8]. Hence, we infer that the number of factors for the KR must not exceed 3, which suggests the need for feature selection. For each labeler we would build a user behavior model (LR or KR) with the mean aesthetics rating as the output variable and the factors as the predictors.

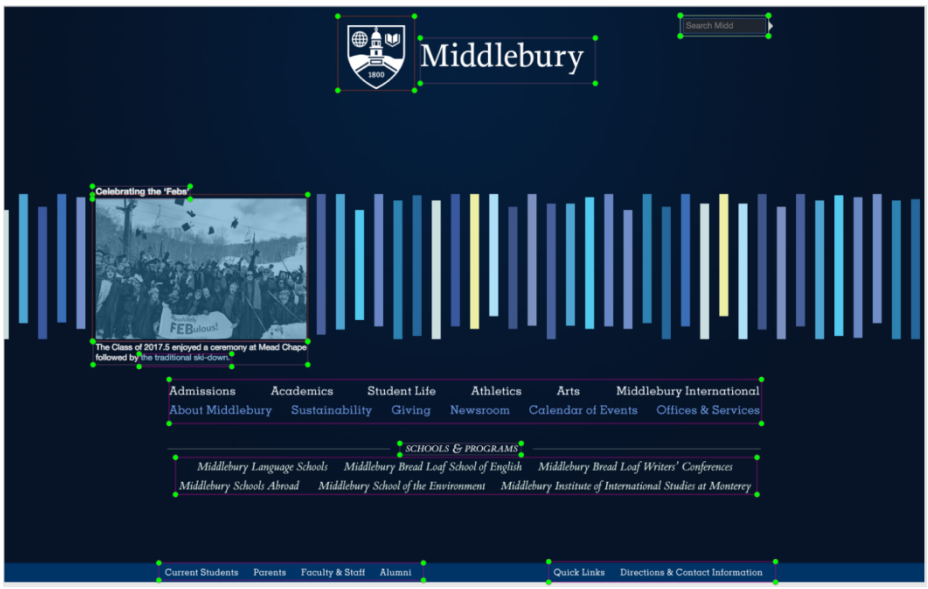

Fig. 2. An example of a "neglectable" labeling: both Subjective Completeness and Precision are low.

In the labeling verification, we had two variables: subjective completeness (SC) and Precision, averaged for each of the 31 labelers (see [6] for detail).

To assess the models' quality, we employed two criteria:

1. Coefficient of determination: $\mathcal{R}^2$
2. Coefficient of determination using PRESS statistics: $\mathcal{R}^2_{PRESS}$

In order to refute the results that we got in our first experiment, we decided to test one more time the following hypothesis: better quality of the labeling data, indicated by higher SC and Precision values, is expected to translate into improved model quality.

3 Results

3.1 Descriptive Statistics and Outliers

In total, we collected 3205 aesthetics assessments and averaged them per each website. So, we had one average assessment value per each of the 557 UIs.

Further, the 31 labelers marked out 27564 elements in 557 UIs, and the quality of their work was evaluated by the 22 verifiers. Two of the labelers, with SCs of 2.3% and 33.84%, were considered outliers, as the mean SC for the other labelers was 68.9% (SD = 11.97%). That is why they were removed from the further analysis, so we remained with 29 labelers (A1...A29 in Table 1) who altogether processed 544 unique UIs.

Table 1. The descriptive statistics per the labelers (M±SD for the respective UIs is shown).

	UI Labeling		**UI Assessment**
Labeler's ID	**# of UIs**	**# of Elements**	**Aesthetics**
A1	18	588	4.03±0.84
A2	20	306	4.31±0.83
A3	17	891	4.28±0.94
A4	45	1658	4.11±1.00
A5	17	1206	3.91±1.08
A6	17	1313	4.01±0.96
A7	46	1334	4.06±0.94
A8	27	623	3.94±0.80
A9	19	587	3.87±0.93
A10	16	631	4.18±1.08
A11	22	1400	4.07±0.86
A12	14	619	3.97±1.02
A13	28	1888	4.00±1.01
A14	20	1106	3.82±0.91
A15	23	810	4.23±0.77
A16	27	2105	3.56±0.86
A17	26	1051	4.25±0.86
A18	33	1120	3.82±0.82
A19	43	745	4.00±1.08
A20	15	367	4.04±1.10
A21	15	670	3.94±0.89
A22	23	1132	4.19±0.97
A23	25	498	4.15±0.74
A24	26	1162	3.87±0.91
A25	20	630	4.10±1.30
A26	23	297	4.26±1.17
A27	14	780	4.01±1.21
A28	21	1037	3.88±0.84
A29	28	396	4.30±0.90
	544	**26950**	**4.04±0.95**

Among the elements in the remaining UIs, 23484 were verified as correct and 3466 as incorrect, and the mean Precision per labelers was 84% (SD = 10.69%), which indicates a reasonably good work quality and is comparable to the precision value of 88.7% obtained in [6]. Pearson correlation between Precision and SC per labelers was significant ($r_{29} = 0.506$, $p = 0.05$), which suggests that these two aspects of UI labeling quality are related, but still distinct.

3.2 The Effect of the Input Data Quality in the Models

In order to select the factors, we used LASSO and partial correlations (PC). In LASSO regularization, the later the parameter's estimation goes to zero, the stronger is the factor's influence. After using five-fold cross-validation method, we found the optimal regularization parameter $\lambda = 0.0624$ where the minimum of mean squared error (MSE) is reached. So, the straightforward LASSO application suggested that we use four factors: SInE, SImE, TE and BE (see in Table 2). To further decrease the number of factors, we increased the regularization parameter to 0.0625, which resulted in decrease of MSE by 0.02% and exclusion of the BE factor.

Despite the fact that the correlations for BE, LE, NE, IE and PE were higher than for the SInE, SIme (see in Table 2), they were constant for most of the labelers, which makes it impossible to build most of the non-parametric models and badly affects parametric ones. That is why the same three factors that were obtained by applying the LASSO method were selected to predict the aesthetic impressions: number of text elements (TE), share of input element' areas in the screenshot (SInE) and share of the image elements' area in the screenshot (SImE).

Table 2. Partial correlations between aesthetics and the independent factors.

Variable name	Variable indicator	PC value
Number of text elements	TE	0.099
Number of button elements	BE	0.096
Number of link elements	LE	0.096
Number of navigation elements	NE	0.091
Number of input elements	IE	0.083
Number of panel elements	PE	0.080
Share of the input elements' area in the screenshot	SInE	0.070
Share of the image elements' area in the screenshot	SImE	0.048

To construct the user behavior models, we used simple LR and KR. So, we built 58 models, each having the same 3 factors calculated from each labeler's output and 2 aggregated models without splitting into labelers. The models' quality metrics and the mean labelers' quality parameters obtained from the UI verifications are shown in Table 3.

The models' average $\mathcal{R}^2_{kernel} = 0.244$ was considerably higher than the average $\mathcal{R}^2_{linear} = 0.172$. To compare $\mathcal{R}^2_{PRESS_kernel}$s and $\mathcal{R}^2_{PRESS_linear}$s, we used t-test for paired samples, which found highly significant difference ($p < 0.001$).

Table 3. The labelers' work quality and the models' end quality.

Labeler's ID	$\mathcal{R}^2_{PRESS\_linear}$	$\mathcal{R}^2_{linear}$	$\mathcal{R}^2_{PRESS\_kernel}$	$\mathcal{R}^2_{kernel}$	SC	Precision
A1	-0.53	0.08	0.00	0.10	86.0%	100.0%
A2	-0.51	0.08	0.00	0.12	67.5%	86.1%
A3	-0.12	0.30	0.16	0.21	89.8%	91.4%
A4	-2.44	0.01	0.49	0.56	70.0%	85.7%
A5	-0.74	0.09	0.00	0.05	88.9%	95.6%
A6	-0.60	0.16	0.16	0.29	61.0%	86.5%
A7	-0.36	0.04	0.41	0.51	55.3%	79.4%
A8	-0.32	0.09	0.38	0.43	58.8%	88.9%
A9	-0.40	0.03	0.00	0.08	70.0%	81.2%
A10	-1.51	0.28	0.17	0.31	85.6%	89.0%
A11	-0.40	0.08	0.13	0.15	80.9%	89.9%
A12	-3.10	0.07	0.34	0.45	54.9%	82.9%
A13	-0.21	0.20	0.07	0.14	66.6%	95.4%
A14	-0.53	0.22	0.19	0.34	77.4%	95.4%
A15	0.17	0.35	0.25	0.28	77.0%	87.3%
A16	-0.14	0.13	0.05	0.13	72.3%	84.6%
A17	-0.23	0.14	0.05	0.19	71.7%	79.4%
A18	0.04	0.25	0.15	0.23	57.7%	76.6%
A19	-0.26	0.12	0.00	0.03	52.4%	85.1%
A20	-0.34	0.25	0.11	0.34	53.9%	79.9%
A21	-1.21	0.52	0.66	0.68	55.3%	94.0%
A22	-0.21	0.23	0.00	0.07	71.0%	91.9%
A23	-0.36	0.11	0.07	0.20	53.4%	64.2%
A24	-0.23	0.06	0.17	0.28	84.4%	94.6%
A25	0.04	0.29	0.00	0.20	70.1%	78.3%
A26	0.08	0.33	0.00	0.03	70.1%	85.4%
A27	-1.64	0.14	0.12	0.17	71.1%	58.1%
A28	-0.39	0.23	0.20	0.23	75.8%	74.4%
A29	-0.26	0.12	0.17	0.29	48.9%	55.8%
Avg.	**-**	**0.17**	**-**	**0.24**	**68.9%**	**84.0%**
Agg.	**0.02**	**0.03**	**0.04**	**0.05**	**-**	**-**

We acknowledge the potential imprecision in our quality evaluations and opted to consider all the metrics as ordinal variables. This approach is highly useful because those requesting tasks typically only accept completed work from the most skilled labelers and discard the output of the neglectable ones. So, we used both Kendall and Pearson correlations to find the connection between the input data quality per labelers and the models' quality criteria (see in Table 4).

Although no significant correlations between the metrics of data labeling and the models' quality criteria were found, we can note that most correlations for KR models, which are more accurate, are negative (see in Table 4).

3.3 Consideration of Our Previous Study Data

To extend the volume of the dataset, we combined the data from the current study and the previous one [6], as demonstrated in Table 5. Moreover, we considered all the

three scales: Aesthetics (A), Complexity (C) and Orderliness (O), with the corresponding R^2 criteria values for the respective LR models.

Table 4. Correlations between the quality of the labeling and the models.

Model	Criterion	Pearson (correlation/p-value)		Kendall (correlation/p-value)	
		SC	Precision	SC	Precision
Linear regression (LR)	$\mathcal{R}^2_{PRESS\_linear}$	0.08/0.69	0.04/0.83	-0.002/1.00	-0.08/0.56
	$\mathcal{R}^2_{linear}$	0.02/0.92	0.13/0.49	0.06/0.65	0.02/0.87
Kernel regression (KR)	$\mathcal{R}^2_{PRESS\_kernel}$	-0.29/0.13	0.06/0.77	-0.10/0.48	-0.02/0.90
	$\mathcal{R}^2_{kernel}$	-0.35/0.06	-0.002/0.99	-0.17/0.22	-0.08/0.56

Table 5. The labelers' and the LR models' quality based on the combined data.

Labelers' ID	$\mathcal{R}^2_A$	$\mathcal{R}^2_C$	$\mathcal{R}^2_O$	SC	Precision
A1	0.08	0.14	0.01	86.0%	100.0%
A2	0.08	0.09	0.08	67.5%	86.1%
A3	0.30	0.07	0.33	89.8%	91.4%
A4	0.01	0.13	0.01	70.0%	85.7%
A5	0.09	0.58	0.08	88.9%	95.6%
A6	0.16	0.20	0.10	61.0%	86.5%
A7	0.04	0.14	0.01	55.3%	79.4%
A8	0.09	0.16	0.07	58.8%	88.9%
A9	0.03	0.27	0.08	70.0%	81.2%
A10	0.28	0.04	0.24	85.6%	89.0%
A11	0.08	0.09	0.02	80.9%	89.9%
A12	0.07	0.16	0.00	54.9%	82.9%
A13	0.20	0.22	0.23	66.6%	95.4%
A14	0.22	0.17	0.17	77.4%	95.4%
A15	0.35	0.23	0.44	77.0%	87.3%
A16	0.13	0.07	0.14	72.3%	84.6%
A17	0.14	0.34	0.09	71.7%	79.4%
A18	0.25	0.13	0.13	57.7%	76.6%
A19	0.12	0.05	0.06	52.4%	85.1%
A20	0.25	0.10	0.14	53.9%	79.9%
A21	0.52	0.30	0.15	55.3%	94.0%
A22	0.23	0.18	0.11	71.0%	91.9%
A23	0.11	0.08	0.16	53.4%	64.2%
A24	0.06	0.05	0.03	84.4%	94.6%
A25	0.29	0.25	0.08	70.1%	78.3%
A26	0.33	0.08	0.11	70.1%	85.4%
A27	0.14	0.47	0.27	71.1%	58.1%
A28	0.23	0.55	0.45	75.8%	74.4%
A29	0.12	0.02	0.16	48.9%	55.8%
AA	0.15	0.11	0.11	73.0%	89.0%
GD	0.35	0.26	0.22	84.3%	89.9%
KK	0.25	0.26	0.15	82.5%	95.5%

MA	0.49	0.36	0.30	75.1%	72.0%
NE	0.49	0.32	0.42	78.3%	85.1%
PV	0.29	0.36	0.20	81.7%	91.6%
PE	0.57	0.17	0.61	72.0%	77.9%
SV	0.18	0.28	0.21	80.4%	97.4%
ShM	0.32	0.34	0.22	77.5%	89.5%
SoM	0.31	0.30	0.20	56.0%	95.9%
VY	0.11	0.20	0.17	95.5%	92.8%

Again, no significant correlations could be found for any of the scales (see in Table 6). However, we can see that most of the correlation coefficients for the combined dataset are now positive, which better aligns with the theoretically expected results.

Table 6. Correlations between the quality of the labeling and the LR models (combined with [6]).

Criterion	Pearson (correlation/p value)		Kendall (correlation/p value)	
	SC	Precision	SC	Precision
$\mathcal{R}_A^2$	0.09/0.57	0.03/0.89	0.10/0.37	0.01/0.95
$\mathcal{R}_C^2$	0.27/0.10	-0.01/0.95	0.17/0.12	0.08/0.48
$\mathcal{R}_O^2$	0.24/0.14	-0.19/0.24	0.19/0.09	-0.04/0.71

4 Discussion and Conclusions

In our previous study [6], we found a significant negative correlation between labeling precision and the quality of the predictive models for the Aesthetics dimension of UX. However, these results turned out to be quite surprising for us, even though we kept in mind that the object of our study is rather a philosophical concept. Common sense told us that the better everything is labeled, the easier it will be to predict aesthetics. To validate the results of our original study, we now considered another 557 web UIs.

In [6], we built simple linear regression (LR) models on a fairly small sample size, which could have been affected by heteroscedasticity, autocorrelation, and multicollinearity. In the current work, we selected the most significant factors and used the Nadaraya-Watson kernel regression (KR), because in this case the main problems of structural regression models do not directly affect the estimation results. The employment of the kernel regression indeed allowed us to increase the R^2s of the models (0.244 for KR vs. 0.172 for LR). Still, the correlations between the R^2s and the labeling quality were mostly negative (see in Table 4), although not significant. Extending the dataset with the data from [6] lead to mostly positive correlations, which are easier to interpret. Overall, the previous results obtained in [6] still hold, since in the current study we did not achieve the appropriate level of statistical significance. Still, we see the main contributions of the current study and their importance as follows:

1) We demonstrated that kernel regression (KR) might be applicable to predict aesthetic impressions even with a small number of factors. Although a considerable share of existing publications focus on increasing the number of the input variables (see e.g. in [4]), the problem of making the most with the limited data, which do not come for free in the field of Human-Computer Interaction [9], remains rather urgent [10].
2) We highlighted the sophistication of the concept of training data quality in ML and certain counter-intuitiveness with its practical application. While a lot of research publications consider the benefits of more accurate labelling data to be obvious (see reviews in [11] or [12]), relatively few put forward the importance of its effect's quantification [13]. We demonstrated that the labeling quality might have no effect on the resulting models' quality or even a negative correlation. This can be explained by the high subjectivity of the prediction object: the aesthetic impression, for which "less could be more", as we reasoned in [6].

The main limitation of our study is arguably the relatively small number of labeled UIs per subject and the associated low R^2s in LR and KR models. Our further research plans include obtaining at least 20 observations per labeler and carrying out a more sophisticated selection of the factors that also considers interfactorial interactions. In this case models will more accurately describe the data and the correlation can become significant for quality. The sample of the verifiers in our study – over 90% of them were men – might also be problematic with respect to external validity. This is particularly remarkable since the output variables included such gender-dependent subjective dimension as aesthetic impression.

Although we do not propose a significant and final model for the relation between labeling quality and the final quality of predictive user behavior models, we believe that our results might be of interest to UI designers and ML practitioners who collect training data. The main take-away is that the costs of obtaining more quality data should be weighed against its actual effect on the end quality of the models. So, it generally makes sense to collect the data in iterations (i.e., several batches), carefully measuring the models' quality dynamics.

Acknowledgment. We would like to thank V. Shchekoldin (PhD, Assoc. Prof. of the NSTU's department of Marketing and Service) for consulting us on the modeling issues.

References

1. Orlova, A.: Data quality became a topic of discussion at the Gaidar Forum (in Russian). Training Center for Leaders and Digital Transformation Teams of the Russian Presidential Academy of National Economy and Public Administration (RANEPA), (2022). https://cdto.ranepa.ru/sum-of-tech/materials/37, last accessed 2023/08/28.
2. Chang, C., Almaghalsah, H.: Usability evaluation of e-government websites: A case study from Taiwan. International Journal of Data and Network Science, 4(2), 127-138 (2020).

3. Miniukovich, A., Marchese, M.: Relationship between visual complexity and aesthetics of webpages. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020, pp. 1–13 (2020).
4. Wan, H., Ji, W., Wu, G., Jia, X., Zhan, X., Yuan, M., Wang, R.: A novel webpage layout aesthetic evaluation model for quantifying webpage layout design. Information Sciences, 576, pp. 589-608 (2021).
5. Lima, A.L.D.S., Gresse von Wangenheim, C.: Assessing the visual esthetics of user interfaces: A ten-year systematic mapping. International Journal of Human–Computer Interaction, 38(2), 144-164 (2022).
6. Bakaev, M., Khvorostov, V.: Quality of Labeled Data in Machine Learning: Common Sense and the Controversial Effect for User Behavior Models. Engineering Proceedings, 33(1):3. (2023). https://doi.org/10.3390/engproc2023033003
7. Boychuk, E., Bakaev, M.: Entropy and compression-based analysis of web user interfaces, in: International Conference on Web Engineering, Springer, pp. 253–261 (2019).
8. Ali, T.H.: Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study). Communications in Statistics-Simulation and Computation, 51(2), pp. 391-403 (2022).
9. Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M.: Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Computing Surveys (CSUR), 51(1), pp. 1-40 (2018).
10. Bakaev, M., Speicher, M., Heil, S., Gaedke, M.: I Don't Have That Much Data! Reusing user behavior models for websites from different domains. In International Conference on Web Engineering, pp. 146-162 (2020).
11. Geiger, R.S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., Tang, R.: "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? Quant. Sci. Stud., 2, pp. 795–827 (2021).
12. Whang, S.E., Roh, Y., Song, H., Lee, J.G.: Data collection and quality challenges in deep learning: A data-centric AI perspective. The VLDB Journal, 32(4), pp. 791-813 (2023).
13. Mitchell, M. et al.: Measuring data. arXiv preprint arXiv:2212.05129 (2022).