Universiteit Leiden
The Netherlands

Opleiding Informatica

Improving User Engagement Prediction on Unsplash Using Artificial Intelligence

Dylan Dingjan

Supervisors:
Rob Saunders & Derya Soydaner

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

09/07/2026

Abstract

Can user engagement on the photography platform Unsplash be predicted and explained with machine learning? And do features based on artificial intelligence help? Those are the two questions this thesis sets out to answer. Engagement is treated as a regression target that combines downloads, views, and search conversions, and a Random Forest predicts it from feature groups built out of the Unsplash dataset. The approach is two-step. An exhaustive sweep over feature-group combinations comes first, then the hyperparameters of the best configuration are tuned. The resulting model explains about 86% of the variance under cross-validation and reaches a test coefficient of determination of 0.857. Nearly all of that predictive power traces back to curation and textual metadata. Low-level visual features barely register. Two feature groups based on artificial intelligence are tried as well, emotion features inspired by ArtEmis and aesthetic features inspired by EVA, and neither helps. The emotion labels are often wrong when checked by hand, and the aesthetic scores add nothing on a platform whose photographs are already curated for quality. SHAP then explains individual predictions. It also splits the features that drive engagement, from the photographer's point of view, into what they can change after the shot, such as keywords and descriptions, what they decide before it, such as camera and location, and what they cannot touch at all, such as collection membership. Almost all of the photographer's influence turns out to lie after the shot. That split makes per-photograph guidance possible, yet it also shows exactly where causal interpretation starts to break down.

Contents

1 Introduction **1**
- 1.1 Background and Motivation . . . 1
- 1.2 Problem Statement . . . 1
- 1.3 Research Aim . . . 2
- 1.4 Research Questions . . . 2
- 1.5 Contributions . . . 2
- 1.6 Thesis Structure . . . 3

2 Related Work **4**
- 2.1 Image Popularity and Engagement Prediction . . . 4
- 2.2 Visual Features in Engagement Prediction . . . 4
- 2.3 Metadata and Platform Factors . . . 4
- 2.4 Computational Image Aesthetics . . . 5
- 2.5 Emotion Based Image Analysis . . . 5
- 2.6 Explainable Artificial Intelligence . . . 5
- 2.7 Research Gap . . . 5

3 Methodology **7**
- 3.1 Experiment Design . . . 7
- 3.2 Dataset . . . 7
- 3.3 Target Variable . . . 7
- 3.4 Feature Sets . . . 8
  - 3.4.1 Baseline Features . . . 8
  - 3.4.2 ArtEmis Emotion Features . . . 8
  - 3.4.3 EVA Aesthetics Features . . . 9
- 3.5 Model Selection . . . 9
- 3.6 Hyperparameter Tuning . . . 9
- 3.7 Evaluation Metrics . . . 10
- 3.8 Explainability with SHAP . . . 10
- 3.9 Experimental Setup . . . 10

4 Results **11**
- 4.1 Baseline Model Performance . . . 11
- 4.2 Random Forest Tuning Results . . . 11
- 4.3 ArtEmis Feature Results . . . 15
- 4.4 EVA Feature Results . . . 16
- 4.5 Model Comparison . . . 18
- 4.6 Feature Importance Results . . . 19
- 4.7 SHAP Explanation Results . . . 21

5 Discussion **23**
- 5.1 Effects of Emotion Based Features . . . 23
- 5.2 Effects of Aesthetic Features . . . 23
- 5.3 Explainability and Model Based Image Improvement . . . 23

5.4 Comparison with Related Work 24
5.5 Limitations 24
5.6 Future Work 25

6 Conclusion **26**

References **30**

A ArtEmis Emotion Examples **31**

B EVA Aesthetic Examples **33**

1 Introduction

Image sharing platforms produce enormous amounts of visual and engagement data, and that makes them a useful place to study how users interact with online images. This thesis looks at one such platform, Unsplash, which is built around high quality photography. The question it asks is whether artificial intelligence can improve and explain the prediction of user engagement there. The rest of this chapter sets up that question: it gives the motivation, defines the research problem, states the aim and the research questions, lists the main contributions, and describes how the thesis is organised.

1.1 Background and Motivation

Online image sharing platforms have become important places for visual communication, creative discovery, and digital visibility. On a platform such as Unsplash, photographs are not just viewed. They are liked, downloaded, searched for, and reused in all sorts of contexts. Engagement, then, may hinge on several things at once: the subject of the image, its aesthetic quality, its composition, the emotional impression it leaves, its metadata, and how visible the platform chooses to make it. Why does predicting engagement matter? It matters for both research and practice. A photographer who knows which image characteristics go together with higher engagement can make better creative decisions and give their work more visibility. For a platform, engagement prediction feeds into ranking, recommendation, and content understanding. And for a researcher, Unsplash is an interesting case precisely because it concentrates on high quality photographic content, whereas most earlier work on image popularity has dealt with broader social media platforms.

Earlier work has shown that image popularity depends on both visual and contextual factors. Khosla et al. studied image popularity prediction on Flickr and found that image content and social context both contribute to it [KDSH14]. Gelli et al. carried this further by examining the role of sentiment and context features in social image popularity prediction [GUB+15]. Taken together, these studies indicate that engagement with online images cannot be explained by visual content on its own. It is better understood as the joint outcome of several interacting feature groups.

Artificial intelligence is well suited to this problem because machine learning models can bring together very different kinds of information. Here, engagement prediction is set up as a supervised regression task. A Random Forest regression model serves as the main predictor, and several feature groups based on artificial intelligence are added to see whether prediction performance and interpretability can be pushed further.

1.2 Problem Statement

Unsplash holds a great deal of image and engagement data, yet it is still hard to say which characteristics make an image more likely to attract attention. Many things may play a part. Visual properties, aesthetic quality, composition, lighting, metadata, photographer information, platform-specific exposure: the list is long. And these factors are hard to pull apart, because they interact with one another and can push engagement in opposite directions.

There are really two challenges here. The first is predictive: a model has to estimate engagement well enough to account for a meaningful part of the variation between images. The second is interpretive. A model can perform well and still be a black box, so it should also be possible to see which features

drive its predictions. That matters all the more once the model is used for more than prediction, namely to suggest how an image might be improved.
This thesis tackles both challenges by building and testing machine learning models for engagement prediction on Unsplash. The project starts from a Random Forest baseline that uses only features taken from the Unsplash dataset. It then asks whether extra features extracted with artificial intelligence improve performance. Emotion based features inspired by ArtEmis are tested first, then EVA style aesthetic features inspired by the Explainable Visual Aesthetics dataset. Finally, SHAP is applied to the final model to interpret its predictions and to identify the features that push predicted engagement up or down.

1.3 Research Aim

The purpose of this thesis is to find out how far user engagement on Unsplash can be predicted from features taken from the platform's images and metadata, and whether artificial intelligence can improve both the accuracy and the interpretability of such models. In concrete terms, the work builds a Random Forest regression model for an engagement target, works out systematically which feature groups carry the predictive performance, checks whether extra feature groups based on artificial intelligence (emotion features inspired by ArtEmis and aesthetic features inspired by EVA) add anything beyond a metadata and visual baseline, and then uses SHAP to interpret the final model. A second purpose is to judge whether those interpretations can be turned into concrete, model-based suggestions for raising the engagement of a single photograph.

1.4 Research Questions

To pursue this purpose, the thesis is organised around four research questions:

- **RQ1.** To what extent can user engagement on Unsplash be predicted from features derived from the Unsplash dataset using a Random Forest regression model?
- **RQ2.** Which feature groups contribute most to predictive performance, and which contribute little or nothing?
- **RQ3.** Do artificial-intelligence-based feature groups, emotion features inspired by ArtEmis and aesthetic features inspired by EVA, improve predictive performance beyond a metadata and visual baseline?
- **RQ4.** Can an explainability method (SHAP) identify which features drive predicted engagement, and to what extent can these explanations support actionable suggestions for improving the engagement of a photograph?

1.5 Contributions

This thesis makes the following contributions:

- It defines a reproducible engagement target for Unsplash that corrects for how long a photo has been online, and builds a complete Random Forest prediction pipeline around it.

- It carries out an exhaustive feature-group ablation, training every combination of the available feature groups (over two thousand models for the baseline, and over four thousand once an emotion feature group is added), so that the signals that actually drive engagement prediction are found empirically rather than by picking a single feature set by hand.
- It evaluates two feature groups based on artificial intelligence in detail. In particular, it reports a negative result for the ArtEmis emotion features, backed by manual validation on a sample of 250 images, and shows that an intuitively appealing feature group can fail to add value while explaining why.
- It applies SHAP to the final model to interpret its predictions and to test whether those explanations can be turned into practical, image-level advice for photographers, separating what a photographer can change after the shot from what they decide before it and what stays outside their control.

1.6 Thesis Structure

The rest of the thesis is laid out as follows. Section 2 reviews related work on image popularity and engagement prediction, on visual and metadata features, on computational aesthetics, on emotion-based image analysis, and on explainable artificial intelligence, and it marks out the research gap. Section 3 sets out the research design, the dataset, the target variable, the feature sets, model selection, hyperparameter tuning, the evaluation metrics, the SHAP analysis, and the experimental setup. Section 4 reports the experimental results: the baseline performance, the feature-group ablation, the evaluation of the ArtEmis and EVA feature groups, the overall model comparison, and the explainability analysis. Section 5 interprets these findings, ties them back to prior work, and considers limitations and future work. Section 6 closes the thesis and answers the research questions.

2 Related Work

Predicting engagement draws on several strands of research, this section reviews each in turn before drawing out where the gaps lie. It starts with image popularity and engagement prediction, the body of work that frames the task and establishes that attention rests on many interacting signals rather than on any single one. It then looks at the feature groups that feed such models, first the visual content of an image and then the metadata and platform context around it. After that it turns to two areas that supply the features based on artificial intelligence examined here, computational image aesthetics and emotion-based image analysis. Finally it covers explainable artificial intelligence, which makes model predictions legible and ties them to practical advice. These areas are the relevant ones because, taken together, they span both what engagement prediction rests on and how its predictions can be understood and acted upon.

2.1 Image Popularity and Engagement Prediction

How much attention an online image will draw is a question that has been studied for more than a decade. Khosla et al. framed image popularity prediction on Flickr and showed that both image content and social context feed into popularity [KDSH14]. McParlane et al. looked at popularity prediction in a cold-start setting, drawing on image context, visual appearance, and user context [MMJ14]. Gelli et al. then showed that sentiment and contextual features improve social image popularity prediction [GUB+15]. The Social Media Prediction (SMP) Challenge is more recent still: it set up a large-scale benchmark and a standard supervised version of the task, and in doing so settled popularity prediction as a well-defined regression problem over heterogeneous features [WLC+23]. Read together, this body of work supports treating engagement as a predictable quantity, one that rests on several interacting feature groups rather than on visual content alone.

2.2 Visual Features in Engagement Prediction

Much of the engagement-prediction literature centres on visual content. Early work tied image content to popularity by combining low-level descriptors such as colour, texture, and gradients with object and scene recognition [KDSH14], and later studies brought in affective and semantic visual cues [GUB+15]. Compact image representations serve as lightweight visual features as well; in this thesis the BlurHash string is decoded into a small set of colour and brightness descriptors that act as an efficient stand-in for the coarse appearance of the image. As the results show later on, though, these low-level visual features add little once curation and textual signals are on the table, which fits the view that visual content by itself does not settle engagement.

2.3 Metadata and Platform Factors

Apart from the pixels, the metadata and platform context around an image are strong predictors of attention. Tags, titles, and descriptions, the identity and following of the uploader, the time of posting, and group or collection membership have all been reported to be associated with popularity on social platforms [KDSH14, GUB+15, MMJ14]. The SMP Challenge benchmark puts this multimodal setting on a formal footing, combining textual, user, and temporal features with visual ones [WLC+23]. The Unsplash dataset offers a comparable set of contextual signals, among

them keywords, descriptions, photographer information, locations, and collections, and these form the backbone of the feature groups used here.

2.4 Computational Image Aesthetics

Automatic assessment of photographic aesthetics has advanced on the back of large annotated datasets and learned models, a body of work surveyed in depth by Deng et al. [DLT17]. The AVA dataset supplied over 250,000 images with aesthetic scores and photographic-style labels, which opened the door to data-driven aesthetic analysis [MMP12]. NIMA then showed that convolutional networks can predict the whole distribution of human aesthetic ratings instead of a single score [TM18]. The EVA dataset went a step further with attribute-level annotations and per-attribute importance, so that aesthetic judgements become more explainable [KVD20]. More recent work shows that vision-language models such as CLIP pick up aesthetic information directly: Hentschel et al. report that CLIP can score image aesthetics with little task-specific training [HKH22], and Soydaner and Wagemans pair attribute prediction with explainable AI to expose which factors drive aesthetic preference [SW26]. These methods sit behind the EVA-style aesthetic features tested in this thesis.

2.5 Emotion Based Image Analysis

A parallel line of research deals with the emotional or affective content of images. Borth et al. put forward a large-scale visual sentiment ontology together with a set of adjective-noun-pair detectors for recognising affective concepts in images [BJC+13]. ArtEmis carried affective analysis over to visual art, collecting language that describes the emotions artworks evoke along with models that predict these affective responses [AOH+21]. Work of this kind motivates the hypothesis tested here, that the emotional impression of a photograph might help account for engagement on top of its subject and metadata.

2.6 Explainable Artificial Intelligence

As machine-learning models are used more and more to support decisions, methods that explain their predictions have grown in importance. SHAP offers a unified, theoretically grounded framework that gives each feature a signed contribution to an individual prediction [LL17]. TreeSHAP makes the same computation exact and efficient for tree ensembles such as Random Forests, and it also lets local explanations be aggregated into global importance [LEC+20]. Explainable AI has been brought to bear on aesthetics in particular, with attribute-level models used to expose which factors drive aesthetic preference [SW26]. This thesis uses SHAP for two things at once: to interpret the model, and to sort the features it relies on, from the photographer's standpoint, into what they can change after the shot, what they fix before it, and what lies beyond their control, which is what links the explanation to actionable advice.

2.7 Research Gap

Most earlier work on image popularity and engagement targets broad social-media platforms such as Flickr, where content quality ranges widely, and aims mainly at maximising predictive accuracy.

Curated, high-quality photography platforms such as Unsplash have received far less attention, and so has the question of whether emotion and aesthetic features based on artificial intelligence are worth anything in that setting. On top of that, explainability methods are well established yet seldom used to tell apart, from the creator's point of view, what they can change after the shot, what they decide before it, and what stays beyond their reach. This thesis works on these gaps. Keeping the photographer rather than the platform at the centre, it predicts and explains engagement on Unsplash, puts two feature groups based on artificial intelligence through a careful test, and uses SHAP to judge whether the resulting explanations can back practical, photographer-facing suggestions for improving engagement.

3 Methodology

Engagement prediction is framed here as a supervised regression problem. Every photograph on Unsplash is described by a set of features taken from its metadata and visual content, and the task is to predict a continuous engagement score from them. A Random Forest regression model does the predicting. It is chosen for a few practical reasons: it copes well with a mix of numeric, categorical, and text-derived features, it holds up against outliers and non-linear relationships, and it gives a natural starting point for feature-importance and SHAP analysis [Bre01].

3.1 Experiment Design

There are three stages. The first sets up a predictive baseline from features that can be read straight off the Unsplash dataset. Rather than commit to a single feature set, an exhaustive sweep trains a separate model for every possible combination of feature groups, so the contribution of each group can be measured empirically (RQ1, RQ2). The second stage widens the feature space with two groups based on artificial intelligence, emotion probabilities inspired by ArtEmis and aesthetic attributes inspired by EVA, then re-runs the sweep to see whether either group beats the baseline (RQ3). The third stage turns to SHAP. Applied to the best model, it interprets individual predictions and shows whether the explanations can support concrete suggestions for raising engagement (RQ4).

3.2 Dataset

The study draws on the official Unsplash datasets, version 1.3.0 [Unsnd]. Unsplash releases the data in two forms: a *Lite* dataset of about 25,000 photographs meant for fast experimentation, and a *Full* dataset of roughly 1.5 million photographs meant for large-scale analysis. Both come as relational tables that describe photos, photographers, keywords, collections, conversions (search and download events), and per-photo statistics.
Every experiment in this thesis is run on the Lite dataset [Unsnd]. Once the relevant tables are joined and the target variable is built, the Lite dataset gives 25,000 usable photographs. These are split into 20,000 training and 5,000 test examples (an 80/20 split), and model selection is done with 3-fold cross-validation on the training partition. The Full dataset [Unsnd] holds roughly 1.5 million photographs and grows to about six million rows once the relational tables are joined. It is not used for the experiments reported here, because processing and repeatedly training models on data of that size needs far more memory and computing time than this project has on hand. The Lite dataset therefore stands in throughout as a representative sample. A full-scale replication on the Full dataset is left for future research, where the larger sample would let the findings of this thesis be checked at scale and would supply enough data for more complex models.

3.3 Target Variable

The prediction target is an *engagement score*. It brings together several signals of user interaction while correcting for how long a photograph has been online. For each photograph, the number of days since submission, its age, clipped to a minimum of one day, turns raw counts into per-day rates. The score is then defined as

$$\text{engagement} = \log(1 + d_{\text{day}}) + 0.3\ \log(1 + v_{\text{day}}) + 0.7\ \log(1 + c), \tag{1}$$

where d_{day} is downloads per day, v_{day} is views per day, and c is the number of search conversions. The $\log(1 + x)$ transform squashes the heavy-tailed distribution of engagement counts so that a handful of extremely popular photographs do not take over the objective. The weights mirror how strongly each signal expresses user intent: a download is the strongest signal and gets full weight; search conversions get a high weight of 0.7 because they show that an image satisfied an explicit search; and passive views get the lowest weight of 0.3. Working with per-day rates instead of raw totals stops the model from merely learning that older photographs have piled up more interactions. Folding several interaction signals into one weighted engagement score is standard practice in social media management, where engagement is routinely measured as a composite of distinct interaction types rather than any single count, and that makes the chosen target a realistic objective for this domain.

3.4 Feature Sets

Features are grouped so that their individual and combined contributions can be measured. The groups split into a baseline set read directly off the Unsplash dataset, and two sets based on artificial intelligence that are added in the second stage.

3.4.1 Baseline Features

The baseline is made up of eleven feature groups taken from the Unsplash tables: *Photo Dimensions* (width, height, aspect ratio); *EXIF / Camera* (camera make and model and related capture metadata); *Temporal* (submission year, month, and weekday); *Descriptions* (lengths of the human-written and AI-generated descriptions); *Photographer* (photographer-level attributes); *Keywords* (keyword-derived signals); *Colors* (dominant colour information); *Collections* (whether and how often a photo is included in user collections); *Location / Landmark* (presence of location and recognised-landmark information); *BlurHash Visual*, a group of twelve features decoded from the compact BlurHash representation that capture the coarse colour and brightness structure of the image; and *Text / TF-IDF*, a text feature formed by concatenating descriptions, keywords, collection titles, location names, and camera information into one text field and vectorising it with TF-IDF term weighting [SB88]. Inside the model pipeline, numeric features are imputed and passed through as they are, categorical features are one-hot encoded, and the text field goes through a TF-IDF vectoriser.

3.4.2 ArtEmis Emotion Features

The first group based on artificial intelligence draws on ArtEmis, a dataset and model of affective responses to visual art [AOH+21]. For each photograph, an emotion classifier returns probabilities over nine affect categories. From these, two summary features are computed as well, an overall valence and the entropy of the emotion distribution, for eleven emotion features in total. The values are pre-computed and stored in a CSV file that the pipeline joins on photo identifier. The idea behind this group is that the emotional impression of an image, on top of its subject and metadata, helps explain why some photographs draw more engagement than others.

3.4.3 EVA Aesthetics Features

The second group based on artificial intelligence draws on the Explainable Visual Aesthetics (EVA) dataset [KVD20], which supplies human aesthetic ratings together with attribute-level annotations and per-attribute importance weights. Unsplash photographs carry no ground-truth aesthetic labels, so the EVA-style features are predicted rather than annotated. This follows the explainable-aesthetics approach of Soydaner and Wagemans [SW26], who predict an overall aesthetic mean-opinion-score from attribute scores and then use SHAP to expose which attributes drive aesthetic judgement. For each photograph, an overall aesthetic score comes from a linear head trained on human aesthetic-preference data on top of a CLIP ViT-L/14 backbone [RKH+21, shu24]. Four aesthetic attribute scores come from CLIP zero-shot scoring of each image against positive and negative textual prompts for each attribute, a technique that Hentschel et al. showed to work well [HKH22], and four matching attribute-importance weights are derived to echo the importance annotations collected in EVA. The result is a predicted aesthetic score, four attribute scores, and four attribute-importance values per photograph. The reasoning is that aesthetic quality is plausibly a direct driver of engagement on a platform built around high-quality photography. Section 4 reports what this group actually contributes.

3.5 Model Selection

The main model is a Random Forest regressor [Bre01], implemented with scikit-learn [PVG+11]. Random Forests are picked for three reasons: the feature space mixes numeric, categorical, and high-dimensional sparse text features; the engagement target stays non-linear and heavy-tailed even after the log transform; and the ensemble gives stable feature-importance estimates and fits cleanly with SHAP for the explainability stage. The same model family runs throughout, so that differences in performance can be put down to the feature sets and not to changes in the learning algorithm.

3.6 Hyperparameter Tuning

To keep the exhaustive feature-group sweep within reach computationally, the sweep uses a fixed and deliberately modest Random Forest configuration, so that all feature combinations are judged on an equal footing. That configuration uses 40 trees (`n_estimators` $= 40$), no maximum depth (trees grown until leaves are pure or hold too few samples), a minimum of two samples per leaf (`min_samples_leaf` $= 2$), and a fixed random seed for reproducibility. Categorical variables are one-hot encoded, with rare categories grouped together to keep the dimensionality in check.
Once the feature analysis is done, the Random Forest hyperparameters are tuned more carefully on the full feature set with a dedicated grid search. The grid varies the number of trees (`n_estimators` $\in \{200, 400, 800\}$), the maximum tree depth (`max_depth` $\in \{\text{none}, 12, 24, 32\}$), the minimum number of samples per leaf (`min_samples_leaf` $\in \{1, 2, 4\}$), and the fraction of features considered at each split (`max_features` $\in \{\text{sqrt}, 0.5, 1.0\}$). That comes to 108 configurations per feature combination, each scored with 3-fold cross-validation. The grid search covers eight feature combinations: the full feature set, the full photographer-changeable subset, and the pre-shot and post-shot parts of that subset on their own, each with and without the EVA aesthetic features. This way, hyperparameter tuning, the value of the EVA group, and the relative strength of what a photographer can change

before and after the shot can all be weighed at the same time. Section 4 reports the resulting tuning and comparison results.

3.7 Evaluation Metrics

Model performance is measured with four standard regression metrics: the coefficient of determination (R^2), the root mean squared error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). The main model-selection criterion is the mean cross-validated R^2 across the three folds, because it captures the percentage of variance in engagement that the model explains and stays comparable across feature sets. Generalisation is then reported on the held-out test set. Putting the cross-validated and test metrics side by side also brings any over-fitting into view, as a gap between training, cross-validation, and test performance.

3.8 Explainability with SHAP

The final model is interpreted with SHAP (SHapley Additive exPlanations) [LL17]. SHAP gives each feature a signed contribution to each individual prediction, and those contributions add up to the difference between the prediction and the average prediction over the dataset. For the Random Forest model, the exact and efficient TreeSHAP algorithm does the work; it computes these contributions in polynomial time for tree ensembles and lets local explanations be aggregated into global feature-importance estimates [LEC+20]. Adding up these contributions across the test set shows which features most strongly raise or lower predicted engagement, and in which direction. This speaks straight to RQ4: when a feature is both influential and modifiable, its SHAP contribution reads as a model-based suggestion for raising a photograph's engagement. Section 4 reports the results of this analysis.

3.9 Experimental Setup

At the heart of the experimental setup is an exhaustive feature-group sweep. Given k candidate feature groups, the sweep trains a model for each non-empty subset, that is, up to $2^k - 1$ combinations, and records for each one the cross-validated and test metrics, the number of features of each type, and the training time. For the baseline this comes to 2,047 models ($2^{11} - 1$); adding the ArtEmis emotion group brings it to 4,095 models ($2^{12} - 1$). Each model is a scikit-learn pipeline that joins imputation, one-hot encoding, and TF-IDF vectorisation with the Random Forest regressor, so that all preprocessing is fitted inside cross-validation and nothing leaks from the test set. The output of every run is written to a structured JSON report that records the best model by cross-validated R^2 together with the full ranking, and this report is what the analysis in Section 4 builds on.

4 Results

This section reports the results of the feature-group sweeps set out in Section 3. All results come from the Lite dataset of 25,000 photographs (20,000 training, 5,000 test), with model selection by mean 3-fold cross-validated R^2. Unless noted otherwise, "CV R^2" stands for the mean cross-validated R^2 on the training partition and "test R^2" for the score on the held-out test set.

4.1 Baseline Model Performance

The baseline sweep trains 2,047 models, one for every non-empty combination of the eleven baseline feature groups. The best model by cross-validated R^2 brings together eight groups: Photo Dimensions, Temporal, Photographer, Keywords, Colors, Collections, Location / Landmark, and Text / TF-IDF, for 27 numeric, 6 categorical, and 1 text feature in all. Table 1 summarises how it performs.

Table 1: Performance of the best baseline model (best of 2,047 feature-group combinations), Lite dataset.

Split	R^2	RMSE	MAE	MAPE
Train	0.976	0.424	0.310	7.16%
Cross-validation	0.864	1.010	0.774	17.82%
Test	0.744	1.117	0.849	19.83%

The best baseline model explains about 86% of the variance in engagement under cross-validation and about 74% on the held-out test set. The gap between the near-perfect training R^2 (0.976) and the test R^2 (0.744) points to some over-fitting that the modest sweep configuration does not fully rein in, and that is what motivates the final-model tuning below. The top ten feature combinations sit very close together: their cross-validated R^2 values all land between 0.8639 and 0.8643, and every single one includes the Collections and Text / TF-IDF groups. That is an early sign that a small number of groups carry most of the performance.

4.2 Random Forest Tuning Results

The hyperparameter grid search described in Section 3 is run on the full feature set, scoring all 108 configurations with 3-fold cross-validation. The best configuration uses 800 trees, no depth limit, a minimum of one sample per leaf, and half of the features considered at each split (`max_features` $= 0.5$). Tuning improves generalisation a great deal. The best untuned model from the feature sweep reaches a test R^2 of 0.744, whereas the tuned model reaches a test R^2 of **0.857** (cross-validated R^2 of 0.854). The cross-validation and test scores now line up almost exactly, which tells us that the earlier gap between them is largely an artefact of the deliberately small sweep configuration and not genuine over-fitting.

Table 2 summarises the effect of each hyperparameter, averaged over all configurations in which it holds a given value. The most influential setting by far is `max_features`: holding splits to $\sqrt{p}$ features drops the mean cross-validated R^2 to 0.742, whereas considering half or all of the features gives about 0.853. This fits the feature-importance findings in Section 4: the predictive signal sits in a few groups, so sampling only a handful of features per split often misses them. Greater tree

depth helps up to a point (mean R^2 rises from 0.803 at depth 12 to 0.821 when unlimited), and smaller leaves are slightly better. Raising the number of trees from 200 to 800, by contrast, barely moves mean performance, which suggests that 200 trees would already be close to enough and that the gains from 800 trees come at a heavy computational cost.

Table 2: Effect of each Random Forest hyperparameter on the full feature set, as mean cross-validated R^2 over all configurations holding that value fixed (108 configurations, 3-fold CV).

Hyperparameter	Value	Mean CV R^2
`max_features`	sqrt	0.742
	0.5	0.853
	1.0	0.853
`max_depth`	12	0.803
	24	0.819
	32	0.820
	none	0.821
`min_samples_leaf`	1	0.819
	2	0.816
	4	0.812
`n_estimators`	200	0.815
	400	0.816
	800	0.816

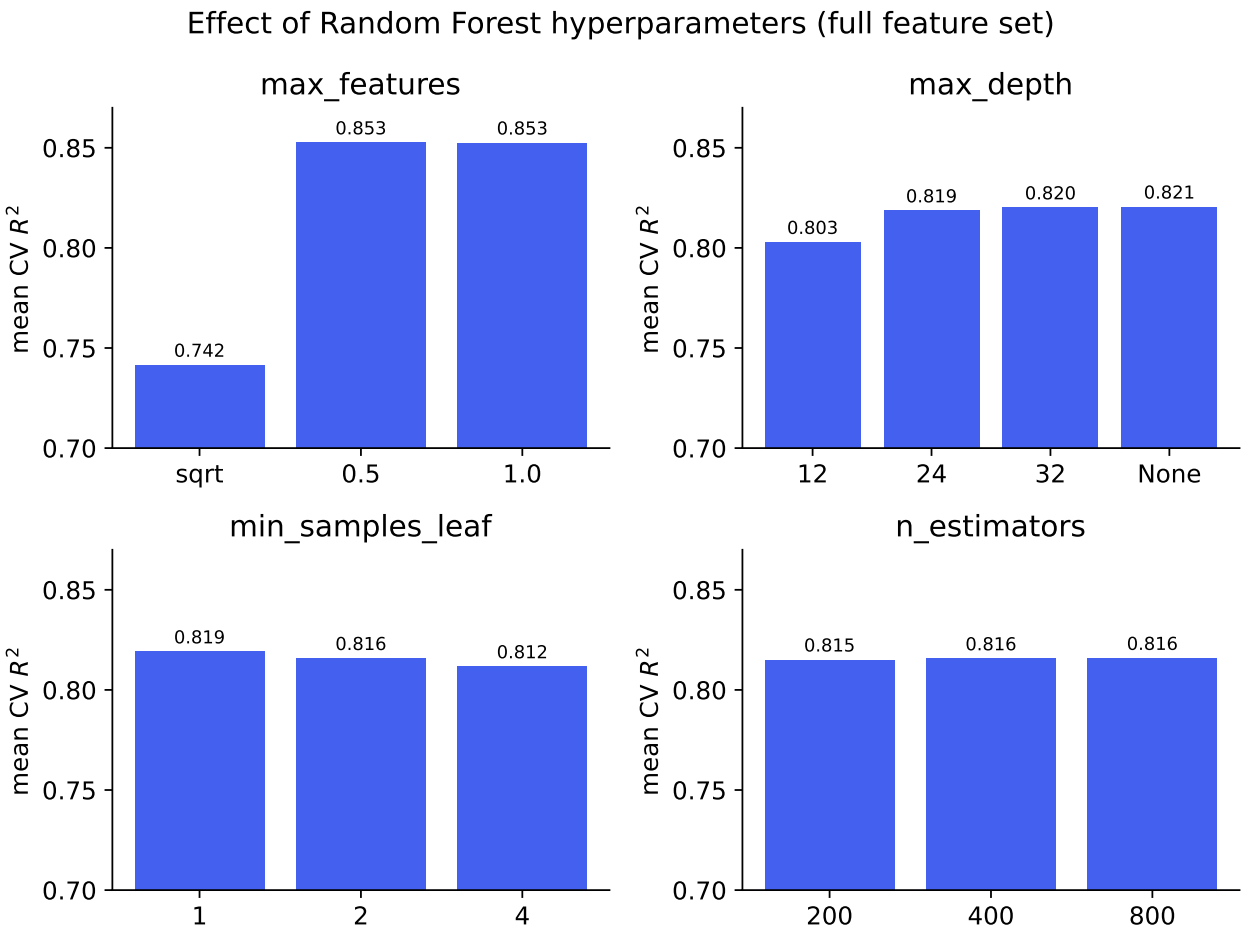

Figure 1: Effect of each Random Forest hyperparameter on the full feature set, as mean cross-validated R^2 over all configurations holding that value fixed. The choice of `max_features` is by far the most consequential.

The same pattern shows up across all eight feature sets that are tuned, namely the full feature set and the photographer-changeable subsets (the full user-changeable set and its pre-shot and post-shot parts), each with and without the EVA features. Figure 2 plots the cross-validated R^2 of every one of the 108 hyperparameter configurations for each of these eight sweeps, coloured by `max_features`, alongside the test R^2 of the best configuration. Three things stand out at once. Within every sweep, the configurations using $\sqrt{p}$ features (red) form a clearly lower cluster, while the 0.5 and 1.0 settings sit together at the top. The full-feature sweeps reach the highest scores, the post-shot and user-changeable sweeps follow close behind, and the pre-shot-only sweeps sit far lower. And adding the EVA features leaves each pair all but unchanged.

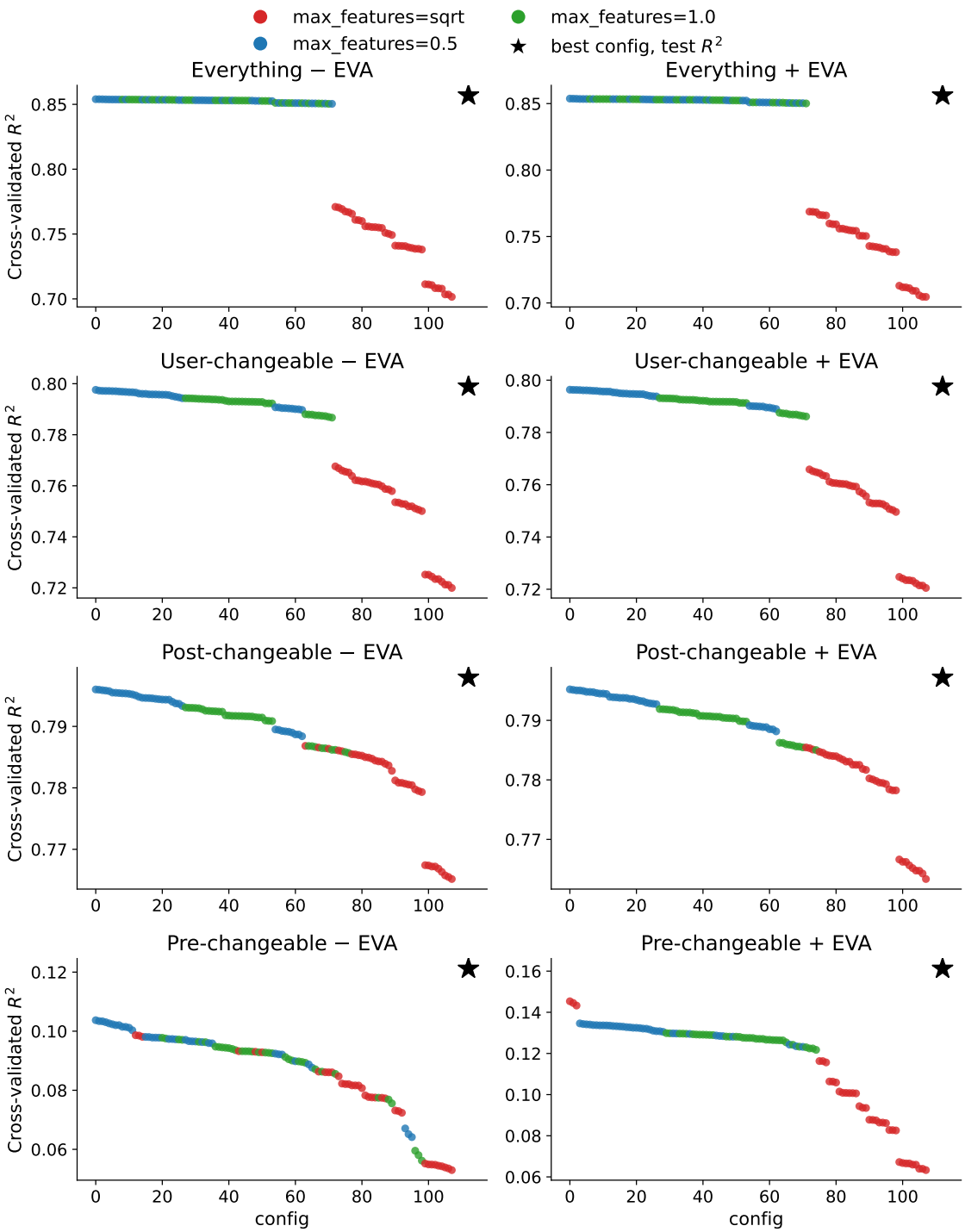

Figure 2: Cross-validated R^2 for all 108 hyperparameter configurations in each of the eight parameter sweeps (one dot per configuration, coloured by `max_features`), with the best model's held-out test R^2 marked by a star. The $\sqrt{p}$ configurations cluster low, the full and post-shot feature sets outperform the pre-shot-only subset by a wide margin, and the EVA features make no difference.

4.3 ArtEmis Feature Results

Adding the ArtEmis emotion group widens the search to 4,095 models ($2^{12}-1$). The best model that *includes* the ArtEmis features reaches a cross-validated R^2 of 0.8637 and a test R^2 of 0.746, while the best model in the same sweep that *excludes* them reaches a cross-validated R^2 of 0.8628. The best baseline model from the earlier sweep reaches 0.8643. Put differently, the best cross-validated performance one can reach with and without the emotion features is the same to three decimal places (a difference of about 0.001, well inside the fold-to-fold variation), so the emotion group does not lift the ceiling of the model.

Averaged over all combinations, the presence of the ArtEmis group goes together with a mean rise in cross-validated R^2 of about 0.011. That small average comes from the weak combinations, though, and it vanishes among the strongest models, where the metadata and text groups already hold the relevant signal. To see why the emotion features add no reliable value, a manual validation is carried out on a sample of 250 photographs, comparing the ArtEmis-derived emotion label against the evident content and mood of each image. Only 156 of the 250 labels (62.4%) are judged correct, so close to four in ten predictions are wrong. Figure 3 shows two representative cases where the predicted emotion clearly mismatches the subject. An error rate this high shows that the emotion signal is too noisy on Unsplash photography to be of use, and Unsplash differs sharply from the artwork ArtEmis was built on. On these grounds the ArtEmis feature group is dropped from the final model.

a grey kitten in a doorway — ArtEmis: *fear*
Photo: Mukesh Jain/Unsplash [Jai16]

a green snake on a branch — ArtEmis: *awe*
Photo: Tyler B/Unsplash [B18]

Figure 3: Two ArtEmis emotion labels that clash with the evident content of the photograph, both graded incorrect in the manual check: a kitten in a doorway labelled *fear*, and a green snake on a branch labelled *awe*. Mismatches of this kind are common and are a direct illustration of why the emotion features carry little usable signal. Further examples are given in Appendix A.

Table 3 reports the model's full output for these two photographs, which shows how the misclassification arises: for the kitten the probability mass falls on *fear* and *sadness*, and for the snake on *awe*, rather than on a neutral or content label. The full output for the appendix example photographs is given in Appendix A (Table 8).

Table 3: Full ArtEmis output for the two in-text example photographs (Figure 3). Values are the model's probabilities over the nine affect categories, plus the summary valence and entropy, the top label, and the manual verdict.

Output	Kitten	Snake
Amusement	0.01	0.03
Awe	0.11	0.52
Contentment	0.05	0.10
Excitement	0.00	0.01
Anger	0.13	0.05
Disgust	0.01	0.02
Fear	0.34	0.15
Sadness	0.30	0.06
Something else	0.05	0.08
Valence	-0.61	0.38
Entropy	1.62	1.58
Top label	fear	awe
Manual verdict	incorrect	incorrect

4.4 EVA Feature Results

The EVA aesthetic features are tested inside the tuned grid search by comparing the full feature set with and without the EVA group, each tuned over the same 108 configurations. Adding the EVA features leaves performance essentially unchanged: the full feature set without EVA reaches a cross-validated R^2 of 0.8540 and a test R^2 of 0.8569, while the same set with EVA reaches a cross-validated R^2 of 0.8537 and a test R^2 of 0.8562. These differences sit well within the fold-to-fold variation, and the best overall model in the whole grid search leaves the EVA group out. The same pattern holds for the photographer-changeable feature subset, where adding EVA leaves the test R^2 all but unchanged at 0.799.

Unlike the ArtEmis labels, the EVA scores themselves are reasonable: Figure 4 shows two photographs the model rates highly that are indeed visually striking, and further high-scoring examples appear in Appendix B. The issue is therefore not that the aesthetic scores are wrong but that, on a platform of uniformly high-quality photography, a well-behaved aesthetic score has little power to tell one strong image from another.

aerial view of a cliff — EVA 6.67/10
Photo: kiwi thompson/Unsplash [kt19]

a wooden house under a starry night — EVA 6.51/10
Photo: Evgeni Tcherkasski/Unsplash [Tch20b]

Figure 4: Two photographs that the EVA model scores near the top of its range (6.7 and 6.5 out of 10). Both are visually striking, so the high scores are well placed; the difficulty is that almost every Unsplash photograph scores highly, which is why the aesthetic features do not help separate more- from less-engaging images. Further examples are given in Appendix B.

Table 4 lists the full EVA output for these two photographs: the predicted aesthetic score, the four attribute scores, and their importance weights. The full output for the appendix example photographs is given in Appendix B (Table 9).

Table 4: Full EVA output for the two in-text example photographs (Figure 4): the predicted aesthetic score, the four attribute scores, the four attribute-importance weights, and their mean and standard deviation.

Output	Cliff	House
Aesthetic score	6.67	6.51
Light & colour	0.92	0.99
Composition & depth	0.27	0.42
Quality	0.60	0.38
Semantics	0.41	0.87
Imp. light & colour	0.63	0.89
Imp. composition & depth	0.03	0.01
Imp. quality	0.33	0.04
Imp. semantics	0.02	0.06
Attribute mean	0.55	0.67
Attribute std	0.24	0.27

So the predicted aesthetic features add no measurable value once metadata and text features are in place, just as the earlier ArtEmis result does not. One plausible reason is that aesthetic quality is already partly captured by signals the model can see more directly, such as whether a photograph has been added to collections, and that on a platform where most photographs share a uniformly high photographic quality, a predicted aesthetic score has little power to tell images apart. On these grounds the EVA group, like ArtEmis, is dropped from the final model.

4.5 Model Comparison

Table 5 compares the tuned models from the grid search, with the untuned feature-sweep results kept in for reference. Three conclusions stand out. First, hyperparameter tuning is the single biggest source of improvement, raising the test R^2 of the full feature set from 0.744 to 0.857. Second, neither feature group based on artificial intelligence adds value: the untuned ArtEmis-augmented model ties the untuned baseline, and the tuned EVA-augmented model is indistinguishable from the same model without EVA. Third, and most relevant to the photographer, the part of the model a photographer can change carries the bulk of the signal only once the shot has been taken: the post-shot-changeable subset (descriptions, keywords, colours, dimensions, and the image-derived text and BlurHash features) reaches a test R^2 of 0.798 on its own, the full photographer-changeable subset adds almost nothing on top (0.799), and the pre-shot-changeable subset (camera settings and the chosen location or landmark) is weak in isolation (0.121).

Table 5: Model comparison on the Lite dataset. The first two rows are the best models from the untuned feature sweeps; the remaining rows are the best tuned models from the grid search. "Full" denotes all baseline feature groups. The photographer-changeable subsets split into a *pre-shot* part (what the photographer decides before or at capture: camera settings and the chosen location/landmark), a *post-shot* part (what they edit afterwards on the platform: descriptions, keywords, colours, dimensions, and the image-derived text and BlurHash features), and their union (*user-changeable*).

Configuration	Tuned	CV R^2	Test R^2	Test RMSE
Baseline (feature sweep)	no	0.864	0.744	1.117
Baseline + ArtEmis (sweep)	no	0.864	0.746	1.112
Full feature set	yes	0.854	**0.857**	1.029
Full feature set + EVA	yes	0.854	0.856	1.032
User-changeable subset	yes	0.798	0.799	1.220
Post-shot-changeable subset	yes	0.796	0.798	1.223
Pre-shot-changeable subset	yes	0.104	0.121	2.550

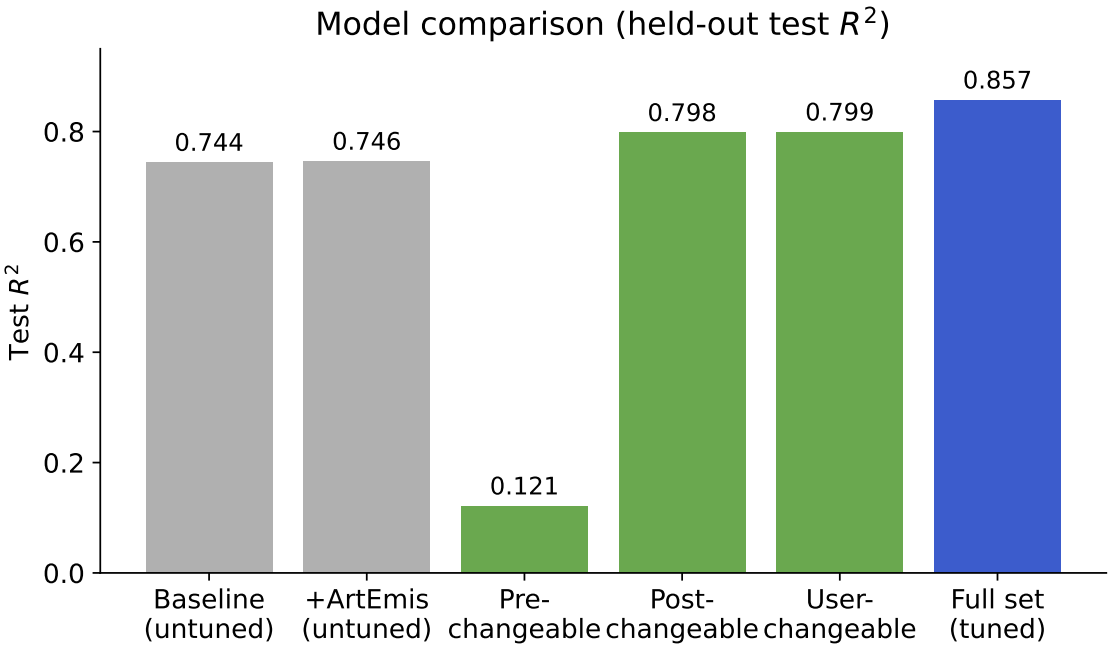

Figure 5: Held-out test R^2 for each configuration. Hyperparameter tuning is the largest single improvement (blue), and among the photographer-changeable subsets (green) the post-shot and full user-changeable sets carry nearly all the signal while the pre-shot-only set is weak. Neither artificial-intelligence feature group adds value.

The contrast between these configurations bears directly on whether the model can guide photographers (RQ4). What a photographer edits after the shot, namely the description, keywords, colours, dimensions, and other image-derived content, already accounts for a large share of the variation in engagement (post-shot-changeable test $R^2 = 0.798$). What they decide before pressing the shutter, namely camera settings and the chosen location or landmark, is by contrast almost inert on its own (pre-shot-changeable test $R^2 = 0.121$), and adding it on top of the post-shot features lifts the user-changeable subset only marginally, to 0.799. Bringing in the features a photographer cannot change, such as whether the photo is placed in collections, the photographer's own profile, and temporal context, lifts the test R^2 to 0.857. The reading is that the photographer's most effective levers are the ones they apply after the shot, that the capture-time choices captured here add little once the post-shot metadata is in place, and that a meaningful part of engagement is still fixed by platform and context factors beyond any one photographer's control, which caps how much model-based suggestions can be expected to help.

4.6 Feature Importance Results

Since every feature-group combination is trained, the marginal contribution of each group can be estimated straight away, by comparing the mean cross-validated R^2 of all models that include the group with the mean of all models that leave it out. This ablation-based measure is reported here rather than the Random Forest's built-in impurity importance, which is known to lean towards high-cardinality and continuous features and can therefore mislead [SBZH07]; the SHAP analysis later on adds a complementary, per-feature view. Table 6 gives this difference for the baseline sweep.

Table 6: Marginal contribution of each baseline feature group, measured as the difference in mean cross-validated R^2 between models that include the group and models that exclude it (baseline sweep, 2,047 models).

Feature group	Mean CV R^2 (incl.)	Mean CV R^2 (excl.)	ΔR^2 [a]
Collections	0.833	0.532	+0.300
Text / TF-IDF	0.816	0.549	+0.266
Keywords	0.727	0.638	+0.088
Temporal	0.692	0.672	+0.020
Location / Landmark	0.687	0.678	+0.009
Photo Dimensions	0.686	0.679	+0.007
Descriptions	0.685	0.680	+0.005
Colors	0.685	0.680	+0.005
EXIF / Camera	0.684	0.680	+0.004
Photographer	0.684	0.680	+0.004
BlurHash Visual	0.684	0.681	+0.003

[a] ΔR^2 is computed from the unrounded values, which may differ slightly from the difference of the two rounded columns shown here.

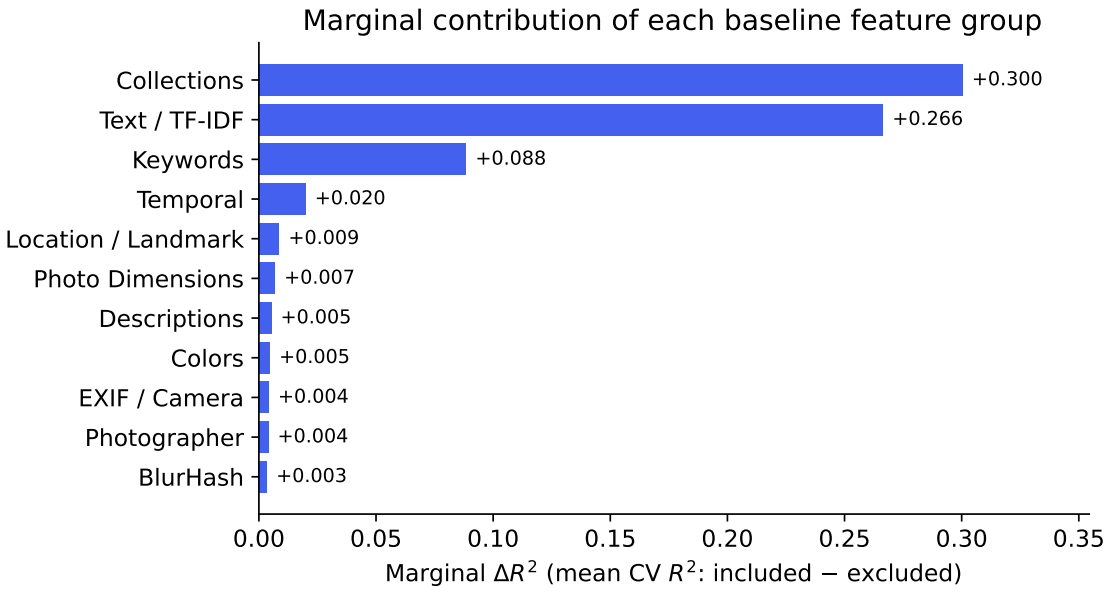

Figure 6: Marginal contribution of each baseline feature group, measured as the difference in mean cross-validated R^2 between models that include and exclude the group. Collections and text dominate.

Two groups clearly lead: *Collections* (+0.300) and *Text / TF-IDF* (+0.266), with *Keywords* (+0.088) some way behind. Each of the other eight groups shifts cross-validated R^2 by at most about 0.02, and the purely visual signals, Colors and the decoded BlurHash features, are among the weakest of all. This says that engagement on Unsplash, as the present target captures it, owes far more to curation and textual metadata (whether a photo is collected, and the descriptive text and keywords

tied to it) than to low-level visual properties. It also helps account for the negative ArtEmis and EVA results: feature groups that describe the image on its own add little once curation and text signals are already there.

4.7 SHAP Explanation Results

To interpret individual predictions, the tuned full-feature model is explained with TreeSHAP [LEC+20]. A dedicated explanation script sets aside a group of photographs that fall in neither the training nor the test split, retrains the best model, and computes SHAP contributions for each held-out photograph. Every feature is then tagged by how much control the photographer has over it: *post-shot changeable* (the description, keywords, colours, dimensions, and the image content summarised by the text and BlurHash features), *pre-shot changeable* (the camera settings and the chosen location or landmark), or *uncontrollable* (collections, the photographer's profile, and timing), so that the explanation can be read as practical guidance. To get a robust picture rather than lean on a single image, the analysis is run as a batch over 50 held-out photographs; the model used for these explanations reaches a test R^2 of 0.860, in line with the tuned result reported above.
Table 7 reports the global importance of each feature group across the 50 photographs, measured as the mean absolute SHAP contribution. At the level of individual explanations, this mirrors the feature-importance analysis of Section 4: collection membership is the most influential single factor by a wide margin (mean absolute contribution 1.35), then the keyword features (0.76) and the free-text features (0.28), while every other group, the colour, BlurHash, and description features included, adds little. The mean signed contributions come out negative for the leading groups because the held-out sample happens to be a little below average in predicted engagement (mean predicted 4.75 against a base value of 5.28); for these under-engaged photographs, sparse collections and keywords drag the prediction down, whereas for above-average photographs the same features push it up.

Table 7: Global SHAP importance by feature group across 50 held-out photographs, measured as the mean absolute SHAP contribution (larger means more influential). "Control" marks how much the photographer can change a group: *post* = after the shot, *pre* = before/at the shot, *none* = outside their control.

Feature group	Control	Mean |SHAP|
Collections	none	1.348
Keywords	post	0.755
Text / TF-IDF	post	0.282
Photo Dimensions	post	0.137
Location / Landmark	pre	0.083
Temporal	none	0.032
BlurHash Visual	post	0.023
Colors	post	0.017
Descriptions	post	0.015
EXIF / Camera	pre	0.010
Photographer	none	0.004

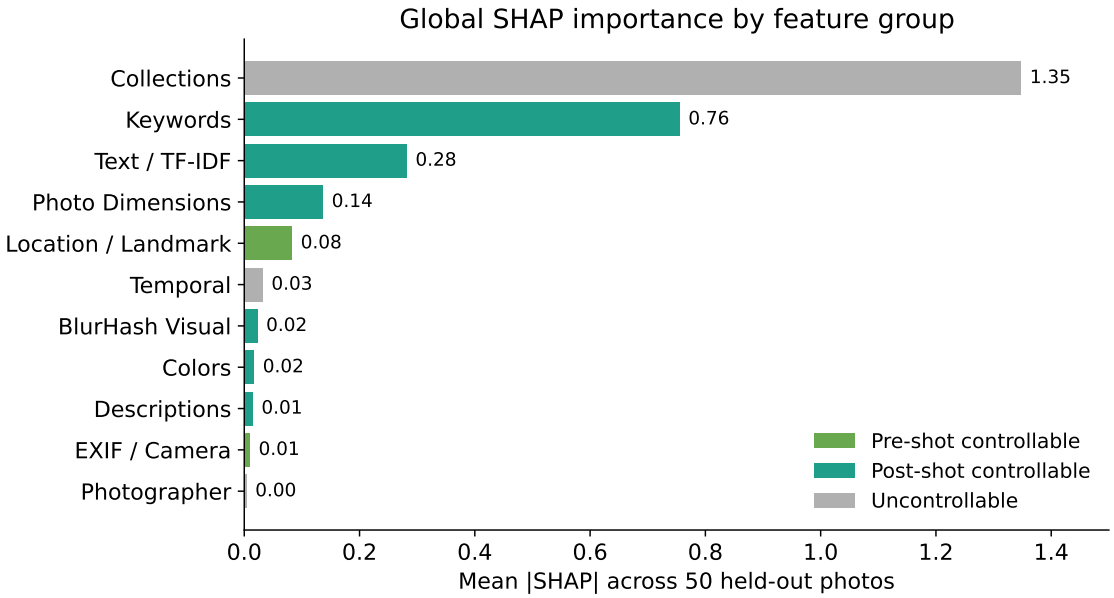

Figure 7: Global SHAP importance by feature group across 50 held-out photographs (mean absolute SHAP contribution), coloured by how much the photographer can control the group: pre-shot, post-shot, or not at all. Collections dominates but is uncontrollable; the post-shot keyword and text features are the leading controllable drivers, while the pre-shot camera and location features barely register.

Splitting these contributions into pre-shot controllable, post-shot controllable, and uncontrollable factors gives a direct, quantitative answer to whether the model can guide photographers. The features a photographer can change make up about 49% of the total explanatory weight across the batch, and almost all of that is post-shot: post-shot-changeable features account for about 45% of the weight against just 3% for pre-shot-changeable ones, with the remaining 51% coming from collections and other context. Per photograph the changeable share is higher still, averaging 63% and running from 31% to 96%, so for a large fraction of photographs most of the explanation falls within the photographer's reach, even though the single strongest lever overall, collection membership, does not. What drives that changeable signal is, again, almost entirely post-shot: the richness of the textual metadata, namely how many descriptive keywords and how much descriptive text go with the photograph. The capture-time choices captured here, the camera and the location, contribute very little. This points to a clear, actionable suggestion: photographs with sparse keywords or descriptions could raise their predicted engagement by enriching this metadata after the shot. The most important uncontrollable driver, the number of collections a photograph appears in, is taken up as a key caveat in Section 5.

5 Discussion

User engagement on Unsplash can be predicted with a Random Forest, and to a substantial degree: the tuned model explains about 86% of the variance under cross-validation and reaches a test R^2 of 0.857. So engagement is far from random. It is strongly tied to observable properties of a photograph and its context. What the feature-group analysis adds is that the predictability sits in only a handful of signals. Two groups lead. One is whether and how often a photograph is placed in collections; the other is the textual metadata that goes with it, namely its description, keywords, and collection titles, captured both directly and through the TF-IDF features. Low-level visual descriptors such as colour statistics and the decoded BlurHash features add very little once those two signals are present. Put plainly, on Unsplash what is written about a photograph and how it is curated seems to matter more than its raw pixel-level appearance, at least as far as the features in the dataset can reach.

5.1 Effects of Emotion Based Features

The ArtEmis emotion features do not improve prediction. The best model with them matches the best model without them. A manual inspection of 250 photographs then finds only 62.4% of the predicted emotion labels to be correct, so nearly four in ten clash with the evident content and mood of the image. Why? The likely cause is a domain mismatch. ArtEmis was built on visual art, where affective interpretation is central and often deliberate. Unsplash is different. It holds stock and lifestyle photography, with a different spread of subjects and intentions. The emotion model therefore gives a noisy signal that carries little information about engagement, and the negative result is best read as evidence that an affective signal carried over from the art domain does not transfer cleanly to a stock-photography platform.

5.2 Effects of Aesthetic Features

The EVA aesthetic features likewise add no measurable value, and the best overall model leaves them out. Two things likely explain this. First, Unsplash is curated for high photographic quality, so the spread of aesthetic quality across photographs is narrow, and a predicted aesthetic score then has little power to tell images apart. Second, aesthetic quality is to some extent already carried by signals the model observes more directly, in particular whether a photograph has been added to collections. One caveat is worth flagging. The EVA features are predicted rather than measured, so part of the null result may come down to the limits of predicting aesthetics with a CLIP-based model rather than to aesthetics being irrelevant as such. Even so, within the scope of this study the aesthetic features give no benefit.

5.3 Explainability and Model Based Image Improvement

The SHAP analysis shows that the model can be explained at the level of individual photographs, and that its explanations can be split into what a photographer changes after the shot, what they decide before it, and what they cannot touch at all. Across a batch of 50 held-out photographs, the changeable share of the explanation averages 63% per photograph and runs from 31% to 96%, and almost all of it is post-shot: the post-shot features (chiefly the richness of the keywords and

description) carry about 45% of the global weight, against only 3% for the pre-shot camera and location features. That is encouraging for the goal of turning predictions into advice, and it sharpens the message for the photographer: the most accessible lever for raising predicted engagement is not the equipment or the place, but how the photograph is described and tagged once it has been uploaded.

This optimism has to be tempered by an important caveat about the strongest non-controllable driver, the number of collections a photograph appears in. Collection membership is plausibly not a cause of engagement but partly a consequence of it: users add photographs they already like to their collections, so this feature may record engagement that has already happened rather than a property that can be changed beforehand. The same worry applies, more weakly, to other popularity-correlated signals. The model is therefore better understood as flagging features associated with engagement than as a causal recipe, and the controllable suggestions it gives (richer keywords and descriptions) should reach users as informed hypotheses rather than guarantees. Two further limits are that SHAP explains the model rather than reality, and that it works at the level of features rather than the image itself, so it can say that more descriptive keywords go together with higher predicted engagement but cannot spell out the exact content of a better photograph.

5.4 Comparison with Related Work

The findings agree with earlier work showing that image engagement rests on both content and context rather than visual content alone [KDSH14, GUB+15, MMJ14], and with the broader view of popularity prediction as a supervised task over heterogeneous features [WLC+23]. The strong role of textual and curation metadata fits the importance of social and contextual features reported on Flickr and other social platforms. The negative results for the aesthetic features seem at first to cut against the computational-aesthetics literature, where learned models predict human aesthetic judgements well [MMP12, TM18, KVD20, HKH22]. The difference is that those studies predict aesthetic ratings, whereas this thesis asks whether predicted aesthetics improve engagement prediction on an already high-quality, curated platform, where they plainly do not. The explicit split of predictions into controllable and non-controllable factors, which builds on explainable-AI methods for aesthetics and tree models [SW26, LEC+20], is a contribution tied to the practical, photographer-facing goal of this work.

5.5 Limitations

A few limitations should be kept in mind. The experiments are run only on the Lite dataset of 25,000 photographs, because the Full dataset calls for more memory and compute than is available; the Lite dataset is a representative sample, but the results have not been checked at the scale of the full platform. The engagement target is a constructed proxy that combines downloads, views, and search conversions with chosen weights and a logarithmic transform, and different weightings could change which features look important. The data form a single cross-sectional snapshot, so the model captures associations at one point in time, not dynamics. As noted above, some of the most predictive features, especially collection membership, may be partly consequences of engagement rather than causes, which limits causal interpretation and the reliability of the suggestions. The feature groups based on artificial intelligence are predicted rather than measured, so their null results depend in part on the quality of those predictions. Finally, Random Forest impurity importances

are known to be biased [SBZH07], which is why ablation and SHAP are used instead, and SHAP itself gives associational, model-level explanations rather than causal ones.

5.6 Future Work

The most immediate next step is to replicate the study on the Full dataset [Unsnd] once enough computing resources are at hand, both to confirm the findings at scale and to supply enough data for more expressive models such as gradient-boosted trees or neural networks. The reverse-causality worry around collection membership calls for a temporal or causal design, for instance predicting future engagement from features measured at upload time, which would make the controllable suggestions more trustworthy. The feature groups based on artificial intelligence could be revisited with extractors better matched to stock photography, and the explanation tool could grow into a deployed, customer-facing service that returns per-photograph, controllability-aware suggestions. Finally, the constructed engagement target could be studied in its own right, comparing alternative definitions and weightings to see how sensitive the conclusions are to the choice of objective.

6 Conclusion

This thesis sets out to find whether user engagement on Unsplash can be predicted and explained with machine learning, and whether features based on artificial intelligence can improve the result. With a Random Forest regression model and an engagement target built from downloads, views, and search conversions, an exhaustive feature-group sweep followed by hyperparameter tuning produces a model that explains about 86% of the variance under cross-validation and reaches a test R^2 of 0.857. The predictive power sits in curation and textual metadata, while low-level visual features, the emotion features inspired by ArtEmis, and the aesthetic features inspired by EVA add little or nothing. SHAP explanations make individual predictions interpretable and let each contributing feature be labelled, from the photographer's standpoint, as changeable after the shot, fixed before it, or outside their control, with almost all of their influence falling after the shot. The research questions set out in the introduction of this thesis have been addressed as follows:

- **RQ1.** User engagement can be predicted to a substantial degree: the tuned Random Forest reaches a test R^2 of 0.857 and so explains a large share of the variation between photographs.
- **RQ2.** Two feature groups carry the predictive performance, collection membership and textual metadata (descriptions, keywords, and the TF-IDF text features), with keywords adding more on top; the rest, all low-level visual descriptors included, matter only marginally.
- **RQ3.** The feature groups based on artificial intelligence do not improve performance. The ArtEmis emotion features add no value and are often inaccurate on manual inspection, and the EVA aesthetic features adds no measurable value, with the best overall model leaving them out.
- **RQ4.** SHAP can pinpoint which features drive a given prediction and in which direction, and splitting these into pre-shot changeable, post-shot changeable, and uncontrollable factors produces actionable, per-photograph suggestions. The photographer's effective levers turn out to be almost entirely post-shot, chiefly enriching keywords and descriptions, while their pre-shot choices of camera and location carry little weight. The strongest uncontrollable driver, collection membership, is plausibly a consequence of engagement, though, so the suggestions are best treated as informed hypotheses rather than guarantees.

To conclude, the practical message is simple. On a platform of uniformly high-quality photography, engagement is shaped more by how a photograph is described and curated than by any sophisticated visual or affective analysis. And what the work delivers is more than a predictive model. It is an explainable one, able to tell apart what a photographer can change after the shot from what they fix before it and what they cannot touch, which is a useful foundation for a photographer-facing tool. Confirming these findings on the full platform and pinning down the causal status of the curation features are the natural next steps.

References

[Aba17] Reign Abarintos. Photograph: a star trail over water. Unsplash, 2017. https://unsplash.com/photos/-cKXtsJWU-I.

[AOH+21] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. https://doi.org/10.1109/CVPR46437.2021.01140.

[Aus16] Kelly Austin. Photograph: a man in the ocean. Unsplash, 2016. https://unsplash.com/photos/NnnsN-cQUSw.

[B18] Tyler B. Photograph: a green snake on a branch. Unsplash, 2018. https://unsplash.com/photos/72vLY-QMd3w.

[Bas14] Pahala Basuki. Photograph: a canoe on calm water. Unsplash, 2014. https://unsplash.com/photos/B2mq60Ksrsg.

[BJC+13] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 223–232, New York, NY, USA, 2013. Association for Computing Machinery. https://doi.org/10.1145/2502081.2502282.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. https://doi.org/10.1023/A:1010933404324.

[D.19] Valdemaras D. Photograph: a bison. Unsplash, 2019. https://unsplash.com/photos/KofzUvmSHks.

[DLT17] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. https://doi.org/10.1109/MSP.2017.2696576.

[dV19] Willian Justen de Vasconcellos. Photograph: a building by trees and water. Unsplash, 2019. https://unsplash.com/photos/y5Tk8f7TBqw.

[Ear19] Joshua Earle. Photograph: a man standing on water. Unsplash, 2019. https://unsplash.com/photos/4UexhWLVcnY.

[GUB+15] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 907–910, New York, NY, USA, 2015. Association for Computing Machinery. https://doi.org/10.1145/2733373.2806361.

[HKH22] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. CLIP knows image aesthetics. *Frontiers in Artificial Intelligence*, 5:976235, 2022. https://doi.org/10.3389/frai.2022.976235.

[Hol19] Andy Holmes. Photograph: mountains under a blue sky. Unsplash, 2019. https://unsplash.com/photos/EgWxUsZByFw.

[Jai16] Mukesh Jain. Photograph: a grey kitten in a doorway. Unsplash, 2016. https://unsplash.com/photos/fi9PSKk7UyE.

[jn18] jcob nasyr. Photograph: a swing by the seashore. Unsplash, 2018. https://unsplash.com/photos/67sVPjK6Q7I.

[jS18] jabez Samuel. Photograph: a galaxy. Unsplash, 2018. https://unsplash.com/photos/Bwkl85szNM.

[KDSH14] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 867–876, New York, NY, USA, 2014. Association for Computing Machinery. https://doi.org/10.1145/2566486.2567996.

[kt19] kiwi thompson. Photograph: aerial view of a cliff. Unsplash, 2019. https://unsplash.com/photos/xxZziWRUVVc.

[KVD20] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. Eva: An explainable visual aesthetics dataset. In *Proceedings of the Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends*, pages 5–13, New York, NY, USA, 2020. Association for Computing Machinery. https://doi.org/10.1145/3423268.3423590.

[LEC+20] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. https://doi.org/10.1038/s42256-019-0138-9.

[Leo17] Leo_Visions. Photograph: trees and water near a snowy mountain. Unsplash, 2017. https://unsplash.com/photos/Dpo2Z7WgxIk.

[LL17] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.

[Lop19] Helena Lopes. Photograph: a rabbit on a leash. Unsplash, 2019. https://unsplash.com/photos/YUwcEo-zrUA.

[Mat21] Ric Matkowski. Photograph: a desert landscape with a mountain. Unsplash, 2021. https://unsplash.com/photos/OKv34B79Ng0.

[MMJ14] Philip J. McParlane, Yashar Moshfeghi, and Joemon M. Jose. "nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 385–391, New York, NY, USA, 2014. Association for Computing Machinery. https://doi.org/10.1145/2578726.2578776.

[MMP12] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. https://doi.org/10.1109/CVPR.2012.6247954.

[Nel19] Eugene Nelmin. Photograph: a wooden house facing a mountain. Unsplash, 2019. https://unsplash.com/photos/MTcz2q0HlNY.

[Nor15] Natasha Norton. Photograph: a house in the snow. Unsplash, 2015. https://unsplash.com/photos/z7tQUhBVOrY.

[Par18] Ian Parker. Photograph: a flock of penguins. Unsplash, 2018. https://unsplash.com/photos/JRhVvF5VHG4.

[Piw23] Marek Piwnicki. Photograph: a snowy field with a mountain. Unsplash, 2023. https://unsplash.com/photos/UEbFDRf9R-c.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://jmlr.org/papers/v12/pedregosa11a.html.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[SB88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. https://doi.org/10.1016/0306-4573(88)90021-0.

[SBZH07] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007. https://doi.org/10.1186/1471-2105-8-25.

[shu24] shunk031. aesthetics-predictor-v2-sac-logos-ava1-l14-linearmse. Hugging Face model repository, 2024.

[SPE19] Nathalie SPEHNER. Photograph: a black dog. Unsplash, 2019. https://unsplash.com/photos/WNCdElkH9pw.

[SW26] Derya Soydaner and Johan Wagemans. Unveiling the factors of aesthetic preferences with explainable ai. *British Journal of Psychology*, 117(2):444–478, 2026. https://doi.org/10.1111/bjop.12707.

[Tch20a] Evgeni Tcherkasski. Photograph: a boat on a lake at night. Unsplash, 2020. https://unsplash.com/photos/v41Z5HCahzo.

[Tch20b] Evgeni Tcherkasski. Photograph: a wooden house under a starry night. Unsplash, 2020. https://unsplash.com/photos/IUPiWRNKNm8.

[Tch21] Evgeni Tcherkasski. Photograph: a rocky mountain under a starry night. Unsplash, 2021. https://unsplash.com/photos/c_J3luSi8lc.

[TM18] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. https://doi.org/10.1109/TIP.2018.2831899.

[Unsnd] Unsplash. Unsplash Lite Dataset 1.3.0 and Unsplash Full Dataset 1.3.0. Dataset, n.d. Version 1.3.0. Retrieved from Unsplash Data.

[Vit18] Ryan Vitter. Photograph: Antelope cave, arizona. Unsplash, 2018. https://unsplash.com/photos/kFMQUwT827k.

[WLC+23] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. SMP challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, New York, NY, USA, 2023. Association for Computing Machinery. https://doi.org/10.1145/3581783.3613853.

A ArtEmis Emotion Examples

This appendix gives further examples of the ArtEmis emotion labels discussed in Section 4, drawn from the manual-validation sample. The content of each photograph is summarised by its AI-generated description, and the predicted ArtEmis emotion is shown below it. The first group collects faulty classifications, where the predicted emotion clashes with the evident subject; the second collects correct ones, where it is a reasonable fit.

a canoe on calm water — ArtEmis: *fear*
Photo: Pahala Basuki/Unsplash [Bas14]

a bison — ArtEmis: *anger*
Photo: Valdemaras D./Unsplash [D.19]

a rabbit on a leash — ArtEmis: *fear*
Photo: Helena Lopes/Unsplash [Lop19]

a man in the ocean — ArtEmis: *fear*
Photo: Kelly Austin/Unsplash [Aus16]

Figure 8: Four ArtEmis classifications graded incorrect in the manual inspection: the predicted emotion does not match the evident content of the photograph.

Antelope Cave, Arizona — ArtEmis: *awe*
Photo: Ryan Vitter/Unsplash [Vit18]

a star trail over water — ArtEmis: *awe*
Photo: Reign Abarintos/Unsplash [Aba17]

a galaxy — ArtEmis: *awe*
Photo: jabez Samuel/Unsplash [jS18]

a house in the snow — ArtEmis: *contentment*
Photo: Natasha Norton/Unsplash [Nor15]

a black dog — ArtEmis: *contentment*
Photo: Nathalie SPEHNER/Unsplash [SPE19]

a swing by the seashore — ArtEmis: *contentment*
Photo: jcob nasyr/Unsplash [jn18]

Figure 9: Six correct ArtEmis classifications: the predicted emotion is a reasonable fit for the subject (grand natural scenes labelled *awe*, calm or homely scenes labelled *contentment*).

Table 8 lists the complete ArtEmis output for the ten photographs shown above: the probability assigned to each of the nine affect categories, the summary valence and entropy, the predicted top label, and the manual verdict.

Table 8: Full ArtEmis output for the ten photographs shown in this appendix: probabilities over the nine affect categories (Amu amusement, Awe, Con contentment, Exc excitement, Ang anger, Dis disgust, Fea fear, Sad sadness, Els something-else), valence (Val), entropy (Ent), the top label (Top) and the manual verdict (OK: yes = correct, no = incorrect).

Photograph	Amu	Awe	Con	Exc	Ang	Dis	Fea	Sad	Els	Val	Ent	Top	OK
a canoe on calm water	0.03	0.30	0.05	0.02	0.05	0.03	0.31	0.11	0.09	-0.11	1.79	Fea	no
a bison	0.01	0.21	0.12	0.05	0.22	0.03	0.12	0.06	0.17	-0.05	1.96	Ang	no
a rabbit on a leash	0.07	0.08	0.11	0.07	0.17	0.02	0.30	0.07	0.10	-0.23	1.98	Fea	no
a man in the ocean	0.04	0.12	0.12	0.01	0.05	0.01	0.30	0.17	0.17	-0.24	1.87	Fea	no
Antelope Cave, Arizona	0.00	0.90	0.02	0.01	0.01	0.00	0.01	0.01	0.03	0.90	0.50	Awe	yes
a star trail over water	0.01	0.82	0.02	0.01	0.02	0.00	0.05	0.01	0.05	0.79	0.79	Awe	yes
a galaxy	0.01	0.68	0.06	0.01	0.03	0.01	0.06	0.07	0.07	0.59	1.22	Awe	yes
a house in the snow	0.02	0.06	0.68	0.11	0.01	0.01	0.01	0.01	0.08	0.83	1.15	Con	yes
a black dog	0.01	0.08	0.75	0.02	0.02	0.01	0.02	0.07	0.01	0.74	0.99	Con	yes
a swing by the seashore	0.02	0.04	0.69	0.03	0.01	0.01	0.01	0.02	0.17	0.73	1.06	Con	yes

B EVA Aesthetic Examples

This appendix shows ten Unsplash photographs with high predicted EVA aesthetic scores (out of 10). The scores are well placed in that all ten are visually strong; the point made in Section 4 is that almost every Unsplash photograph scores in this upper range, which is why the aesthetic features cannot separate more- from less-engaging images.

a flock of penguins — EVA 6.70/10
Photo: Ian Parker/Unsplash [Par18]

a desert landscape with a mountain — EVA 6.61/10
Photo: Ric Matkowski/Unsplash [Mat21]

a man standing on water — EVA 6.59/10
Photo: Joshua Earle/Unsplash [Ear19]

a wooden house facing a mountain — EVA 6.52/10
Photo: Eugene Nelmin/Unsplash [Nel19]

a building by trees and water — EVA 6.50/10
Photo: Willian Justen de Vasconcellos/Unsplash [dV19]

a boat on a lake at night — EVA 6.50/10
Photo: Evgeni Tcherkasski/Unsplash [Tch20a]

a rocky mountain under a starry night — EVA 6.49/10
Photo: Evgeni Tcherkasski/Unsplash [Tch21]

mountains under a blue sky — EVA 6.45/10
Photo: Andy Holmes/Unsplash [Hol19]

trees and water near a snowy mountain — EVA 6.44/10
Photo: Leo_Visions/Unsplash [Leo17]

a snowy field with a mountain — EVA 6.43/10
Photo: Marek Piwnicki/Unsplash [Piw23]

Figure 10: Ten Unsplash photographs with high predicted EVA aesthetic scores. All are visually strong, illustrating that the scores behave sensibly while clustering in a narrow upper band across the platform.

Table 9 lists the complete EVA output for the ten photographs shown above: the aesthetic score, the four attribute scores, the four attribute-importance weights, and the attribute mean and standard deviation.

Table 9: Full EVA output for the ten photographs shown in this appendix: aesthetic score (Aes), the four attribute scores (L&C light and colour, C&D composition and depth, Qua quality, Sem semantics), the four attribute-importance weights (iL&C, iC&D, iQua, iSem), and the attribute mean and standard deviation.

Photograph	Aes	L&C	C&D	Qua	Sem	iL&C	iC&D	iQua	iSem	Mean	Std
a flock of penguins	6.70	1.00	0.22	0.43	0.93	0.96	0.03	0.00	0.01	0.64	0.33
a desert landscape with a mountain	6.61	0.91	0.37	0.18	0.16	0.95	0.02	0.02	0.01	0.40	0.30
a man standing on water	6.59	1.00	0.55	0.84	0.75	0.94	0.03	0.02	0.01	0.78	0.16
a wooden house facing a mountain	6.52	0.90	0.58	0.18	0.55	0.80	0.09	0.07	0.04	0.55	0.26
a building by trees and water	6.50	1.00	0.43	0.34	0.37	0.95	0.01	0.03	0.00	0.53	0.27
a boat on a lake at night	6.50	0.99	0.55	0.75	0.83	0.68	0.02	0.19	0.11	0.78	0.16
a rocky mountain under a starry night	6.49	0.94	0.40	0.71	0.87	0.82	0.00	0.15	0.03	0.73	0.21
mountains under a blue sky	6.45	0.94	0.21	0.36	0.19	0.25	0.09	0.58	0.09	0.43	0.30
trees and water near a snowy mountain	6.44	1.00	0.72	0.29	0.60	0.95	0.01	0.02	0.02	0.65	0.25
a snowy field with a mountain	6.43	1.00	0.21	0.30	0.68	0.96	0.00	0.02	0.02	0.55	0.32