ERASMUS UNIVERSITEIT ROTTERDAM

What is the beauty in Instagram?

Measuring consumer engagement by examining visual and textual elements of beauty influencers on Instagram

Author: Danique Osinga
Student number: 449195

A thesis presented for the degree of
MSc Data Science and Marketing Analytics

Erasmus School of Economics
Erasmus University Rotterdam

Supervisor: Dr. A. Tetereva
Second assessor: H. Deng

September 9, 2021

Abstract

In the modern age advertisements commonly take form in an online environment and social media is dominated by influencers. This research has a focus on the cosmetic industry as it is well known for online presence and use of influencers in marketing strategies. By means of a quantitative study, I analyse Instagram posts of the top ten beauty influencers on Instagram. The study is based on an ordinal logistic regression model to identify what post attributes enhance consumer engagement. The effect of enhancing the consumer engagement on the customer life cycle is explained. This research focuses on five dimensions of visual attributes, namely: colour, objects, person, content, and text. This study bridges the gap between marketing and data analytics literature and aims to develop a helpful tool in generating social media marketing strategies for both influencers and marketing managers. The most essential recommendations from this study are that influencers should focus on entertaining and educational content. Full body visuals enhance engagement, as do brighter hair colours.

Contents

1 Introduction

1.1 Introduction on the topic

People spend on average 145 minutes of their daily life on social media platforms (Statista, 2021a). One industry that seems to adapt to the new market is the beauty industry. Prior to social media, the beauty industry mostly relied on magazine advertisements to target consumers (Osman, 2018). Nowadays beauty brands use influencers as a marketing mean to target potential consumers. Almost half of consumers discover new brands or products via social media ads (R., 2019). A niche and competitive industry became even more competitive by going online. Long-time leaders such as Lancôme and Chanel Beauty are joined by fast growing new brands, including Glossier and Kylie Cosmetics. The growing cosmetic market is expected to reach a global market size of more than 400 billion US dollars by 2024 (Goldstein, 2020).

A popular platform among beauty marketers is Instagram. The app is known for sharing images and videos. As the platform is free of charge (not taking the advertisements into account), it is a popular platform for brands to engage with customers. Young consumers watch influencers on Instagram to see what they are wearing or using (Knowledge, 2019). With more than a billion active users (Statista, 2021b) it is among the most popular social media platforms used by the cosmetics industry. Four out of every five companies use Instagram influencers for their advertisements (Cvetkovska, 2021).

Using influencers is a form of word-of-mouth advertisement and found to be highly effective and of increasing influence due to digitalisation (H. Liu et al., 2021). On Instagram users can engage by following accounts, 'liking' pictures and video's and commenting on posts by other users. This study investigates what post features result in the highest consumer engagement level. This research uses the 100 most recent posts of the 10 most followed beauty influencers worldwide.

The effect that influencers can have on revenue is found to be highly effective (V. Kumar & Mirchandani, 2013). The general favourable effect of social media engagement on brand performance has been confirmed over and over again as well. Online consumer engagement increases brand awareness, purchase intention, word of mouth marketing along with willingness to pay (Jahn & Kunz, 2012; Hutter et al., 2013).

1.2 Problem definition

The transformation of the beauty market due to digitalisation asks for a new view on marketing strategies. Turning towards social media makes room for the implementation of data in marketing strategies. As online advertisements are expected to take on half of the total advertisement budget of beauty brands by 2021, researching content attributes is of high importance (Lashbrook, 2019). Influencers want to create content that generates the most engagement in order to make deals with beauty brands. As the beauty industry targets a young consumer group (S. Kumar et al., 2006), and over two thirds of the total Instagram users is under 35 it is useful to investigate in Instagram. Brands can use Instagram to enhance consumer engagement and possibly market share growth and increase revenue (Statista, 2021c). This is where machine learning comes in. By analysing non-textual data (e.g., images), consumers engagement can be monitored and advise could be given in what attributes of visuals enhance consumer engagement. Therefore, the following research question is proposed:

"Can data analytics be used to predict Consumer engagement and generate marketing strategies for beauty brands on Instagram?"

The dimensions of the post content that will be looked at in answering the research question can be categorised in dimensions (Appendix: Table 8.1), namely *Colour, Objects, Person, Content*, and *Text*.

This research will help marketers and influencers to understand what features drive interaction and make informed decisions on what type of content to create. This way engagement, brand awareness, brand imaging and possibly sales can be enhanced and it is therefore of great marketing relevance.

1.3 Academic relevance

The importance of the research community putting more focus on Instagram has been mentioned at the Conference on Weblogs and Social Media (Hu et al., 2014). Literature does exist on consumer interaction of Instagram posts of beauty brands (Evensson & Jansson, 2020). These studies only focus on the impact of influencers and not on what type of content influencers should create. Barger et al. (2016) suggested to investigate what content factors lead to engagement on social media. There are marketing papers on colour in advertisements, but none seem to examine the effect of colour (no filters) in a social media setting. No literature has examined the effects of objects in images on engagement on a social platform for a specific type of influencer category. While Instagram captions have been examined (Baker & Walsh, 2018; among others), no analysis has been done on the effects of hashtags and tagging accounts on engagement. At the time no study analysed what effect attributes of an Instagram post have on consumer engagement in the beauty industry (Highfield & Leaver, 2016).

2 Theoretical framework

The literature review will focus on general digital marketing strategies (for beauty companies) before diving into the five dimensions of content characteristics and data analytics.

2.1 Literature on digital marketing strategies

The level of consumer engagement in social media is a useful and actionable key performance indicator (KPI), that can help brands to select influencers and promote brand awareness (Cox, 2021). The KPI can also be used to retool posts for optimal engagement. The news feed Instagram users see is based upon algorithms that show posts contingent on interests and engagement. A higher level of engagement therefore results in more showings on the news feeds of users (Rodriguez, 2021). Therefor more engagement leads to a bigger reach, which in terms can lead to more engagement. This vicious circle continues. Reaching more possible consumers is a goal for beauty brands, as this can generate an increase in sales.

KPI's based on Instagram as a social media platform can be used to measure and evaluate the performance of a cosmetic companies' campaign using influencers. The customer journey can be broken down into five parts: awareness, consideration, decision, adaption, and advocacy (Gregory, 2020). Each stop of the customer journey has a different strategy and social KPI (Gregory, 2020). Influencers can improve their engagement by customizing content with features that appeal to users the most. This will result in a broader reach and more impressions as their post will show up more on Instagram's news feed. These actions could result in more clicks that could result in more conversions, probably the most valuable KPI of cosmetic companies. This research could help beauty influencers and (luxury) cosmetic brands to determine the success of the campaign via digital marketing KPI's.

A study that investigated the business strategies of four top players in the cosmetic industry found that every company has a unique marketing strategy (S. Kumar et al., 2006). The companies differ not only in products, but also in advertisement, client base and marketing strategy. One common change in advertisement strategy is to target a broader group of consumers regardless of sex or ethnicity. This marketing strategy could well be seen in the variety of Instagram posts of beauty influencers and influencers themselves.

Besides uniqueness, marketing strategies often focus on creating brand loyalty. Loyal customers have an extended life cycle that results in growing market share and increased revenue. Chan and Mansori (2016) investigated what influences brand loyalty in the cosmetic market of Malaysia. They measured brand loyalty through consumer satisfaction. Chan and Mansori discovered that perceived quality and promotion positively influence consumer satisfaction. Influencers use Instagram to promote a beauty product. The same result is found by Forbes (2016), whom analysed the use of influencers by beauty brands on YouTube. Forbes found that using influencers creates a connection between the consumers and the influencer. This connection results in reliability because of the informal tone and more personal touch of the influencer. Bazi et al. (2020) found that using celebrities creates credibility. Using influencers can be a powerful marketing tool for beauty brands and analysing their content is therefore of high academic relevance.

Chen (2017) researched consumers perception in social media marketing by interviewing consumers aged between 18 and 23 years old. Chen found that celebrity endorsement could be of influence on brand perception and building brand awareness. This does not necessarily lead to a higher level of purchase intention. Chen reported that subtle marketing is preferred over obvious advertisements. He also states that consumers want to see natural settings such as backstage pictures. This indicates why beauty brands often work together with influencers and why the subtle advertisements of influencers work well with the audience. The more creative and natural settings of their posts can appeal more to consumers than the obvious advertisements of the beauty brands itself.

Another research besides that of Chen that focused on exclusivity is by Ashley and Tuten (2014). They investigated creative social media strategies and its influence on consumer engagement of the most valuable brands globally. They found that exclusivity and making use of images of users appeal to consumers and have a positive effect on consumer engagement.

Most influencers are no models, but look more such as ordinary people and this could appeal to consumers. Influencers have an exclusive standing and consumers want to use the products influencers advertise, because they want to feel the same level of exclusivity.

Besides exclusivity, colourfulness is another aspect that influences engagement. Y. Li and Xie (2019) investigated the level of colourfulness on engagement in a social media setting and their result suggests that pictures with a higher level of colourfulness enhance a viewer's attention and therefore lead to a higher level of engagement.

Besides Y. Li and Xie the effects of colours in advertisements were studied by Bellizzi and Hite (1992)[p. 361]. Their research is based upon experiments, where subjects were exposed to either a red or a blue retail environment. They found that purchases were positively affected by a blue display and negatively affected by a red display. The colour blue was identified as calm, cool, and positive, while the colour red related to a tense negative feeling. This contradicts literature by Zhang, Lee, Singh, and Srinivasan (2017) as their research suggests a positive effect of using warmer colours, such as red, as these elicit higher levels of excitement.

One more study that focused on the impact of colours, specifically in a social media setting was by Zailskaite-Jakste et al. (2017)[p 1377] by analysing images of 35 of the most popular brand Facebook groups. The colours black, brown, and grey were found to be used most often and also used in the more popular images. Zailskaite-Jakste et al. state that this finding is in line with the claim that the colours black and blue are among the preferred colours for Generation Y. This generation consists of people aging from 18 to 24 years old and are considered to be the most active group on social media.

Apart from studies on the influence of colours, Hu et al. (2014) focused on the objects seen in Instagram posts. Hu et al. randomly sampled public Instagram users from Instagram's public timeline, consisting of mostly celebrities. Their research found that nearly half of the pictures can be classified as selfies or pictures with friends. Another popular photo category is food, while pictures with pets or fashion are the least common categories. As Hu et al. only looked at the proportion of categories but did not do advanced analysis nor did they look at videos, this allows for further research as is done in this study.

Background setting can be seen as an element of object categories as studies by Hu et al.. Jaakonmäki et al. (2017) analysed to what extent certain content features from Instagram posts influence user engagement. They used quantitative data from Instagram and LASSO regression models together with a Clarifai image recognition API to classify objects in images. The conference paper by Jaakonmäki et al. (2017) found a positive relationship between nature scenes and social media engagement.

Besides objects, people can appear in visuals as well. Y. Li and Xie (2019) found that human presence in images increases engagement. This is in line with a study by Jaakonmäki et al. (2017). The paper by Y. Li and Xie also found that all facial expressions have a positive effect on engagement except for the facial expression “happy”. This is due to the fact that these images are often selfies and lack of relevance for consumers.

Other than objects, visual quality is another aspect of Instagram posts. Y. Li and Xie (2019) found that picture quality has a positive effect on engagement, professional shot pictures lead to higher engagement. Other studies that found a positive effect of perceived quality on consumer satisfaction are from Chan and Mansori (2016), Zhang et al. (2017) and Bazi et al. (2020). Previous research (Liu-Thompkins and Rogerson (2012); Ashley and Tuten (2014)) contradicts this finding and concludes that the quality of the visual is irrelevant.

Ashley and Tuten (2014) focused on the different types of content and concluded that incentives for participation result in higher engagement scores. Interestingly emotional appeals do not lead to more engagement. They also found that although interactive post led to higher engagement, most posts can still be qualified as functional. Previous research by Liu-Thompkins and Rogerson (2012) contradicts these findings and discovered that emotional content are shared more often in a social media setting.

Content types were examined along with Ashley and Tuten by X. Liu et al.. X. Liu et al. (2021) examined the dimensions of social media marketing and their effect on consumer engagement among luxury brands. The research focused on the effect of entertainment, interaction, trendiness, and customization on customer engagement. They used panel data of luxury brands on Twitter and found a positive relationship between entertainment and engagement. Influencers often do not only post advertisements but give the consumer a view of their lifestyle in an entertaining way. Therefore, their post can often be classified as entertaining. Another finding by X. Liu et al. (2021) is that interaction to have a positive effect on consumer engagement. The same effect was found by Ashley and Tuten (2014), who concluded that interactive posts generally increase consumer satisfaction.

Moreover, Bazi et al. (2020) found that entertainment also has a positive effect on interaction. Liu-Thompkins and Rogerson (2012) also found a positive effect of entertaining content and social media interaction.

Another aspect of entertainment is emoji use. According to research by Yakin and Eru (2017) emoji are effective in social media advertisement campaigns. Emoji use tends to create a more attractive and creative campaign. Yakin and Eru drew these conclusions based upon a questionnaire about advertisement campaigns distributed among 400 students.

2.2 Literature on data analytics

Besides a substantial portion of marketing literature, literature on data analytics and machine learning (ML) is analysed as well. This is of importance as the textual and non-textual elements of Instagram posts will be analysed using ML algorithms.

Ma and Sun (2020) highlighted the importance of combining marketing insights and ML algorithms to get more insights out of data. They call attention to the success of convolutional neural networks (CNN) on image processing. Ma and Sun address that ML algorithms can be used to extract the factors of text and images that enhance consumer engagement. One major disadvantage of deep learning (DL) algorithms is that they are often hard to interpret.

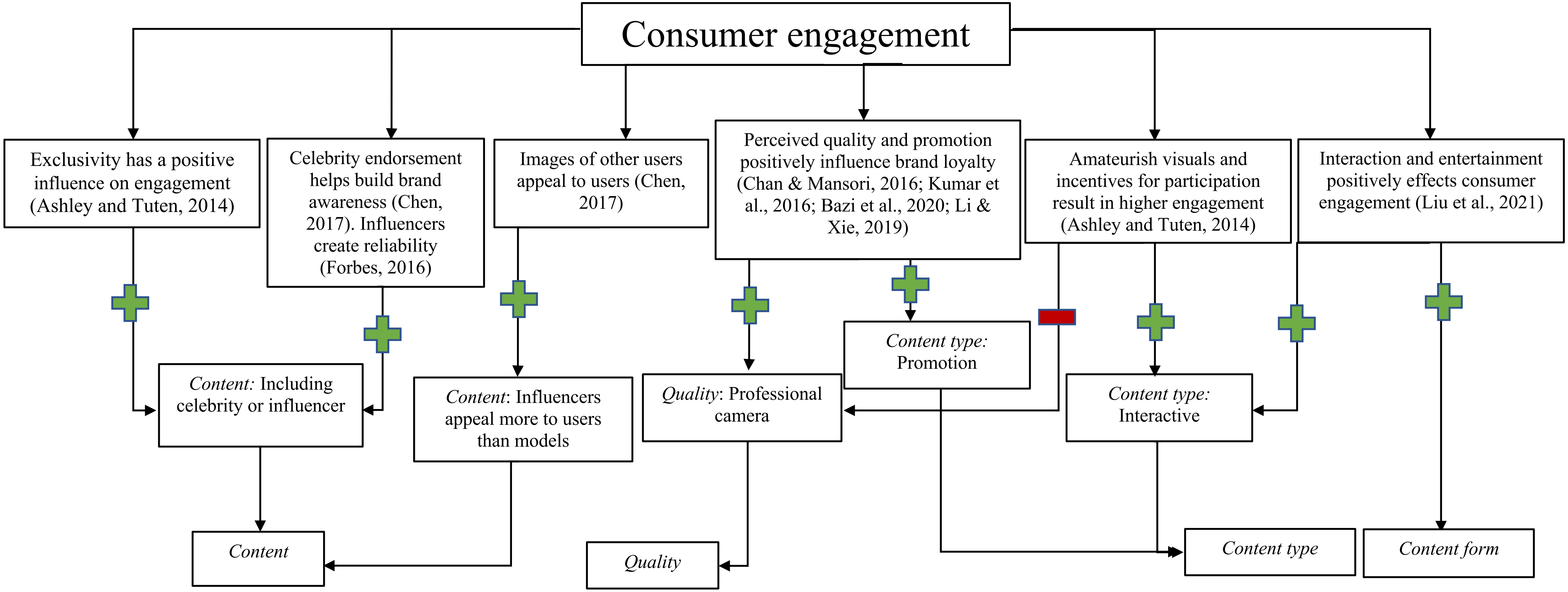
A study by Zhang et al. (2017) investigated images of Airbnb properties and the effect of image quality on property demand. Their research is based upon CNN models for object recognition and image classification. The CNN model is used to classify image quality and resulted in a prediction accuracy of over 90%. With the use of a difference-in-difference (DID), the effect of image attributes were estimated. This is done by comparing changes in the property demand between a treatment (verified pictures) and a control group (non-verified pictures). The DID analysis uses a weighted least squares (WLS) regression to estimate the effects of image factors on demand. WLS is an ordinary least squares (OLS) model with added sampling weights to balance data and account for the probability of a unit receiving the treatment. Their research suggests that using warmer colours (red and yellow) and image brightness positively effect property demand. These factors elicit higher levels of excitement. A photo taken by a professional photographer also has a positive effect on property demand.

Along with Zhang et al. ML insights in a social media setting were used in research by Arora et al. (2019). They investigated the impact of influencers on social media engagement analysing Facebook, Twitter, and Instagram (2019, p. 90). Arora et al. wanted to find out what influence different celebrities have on social media channels. To investigate this, they used ols regression, K-nearest neighbours (KNN), support vector regression and lasso regression. Influence is measured by the reactions of consumers, such as likes and shares, on the posts. Regression models are excellent to solve this problem, as the model estimates the average effect of the features on the continuous output variable. Besides this, the models are straightforward to interpret.

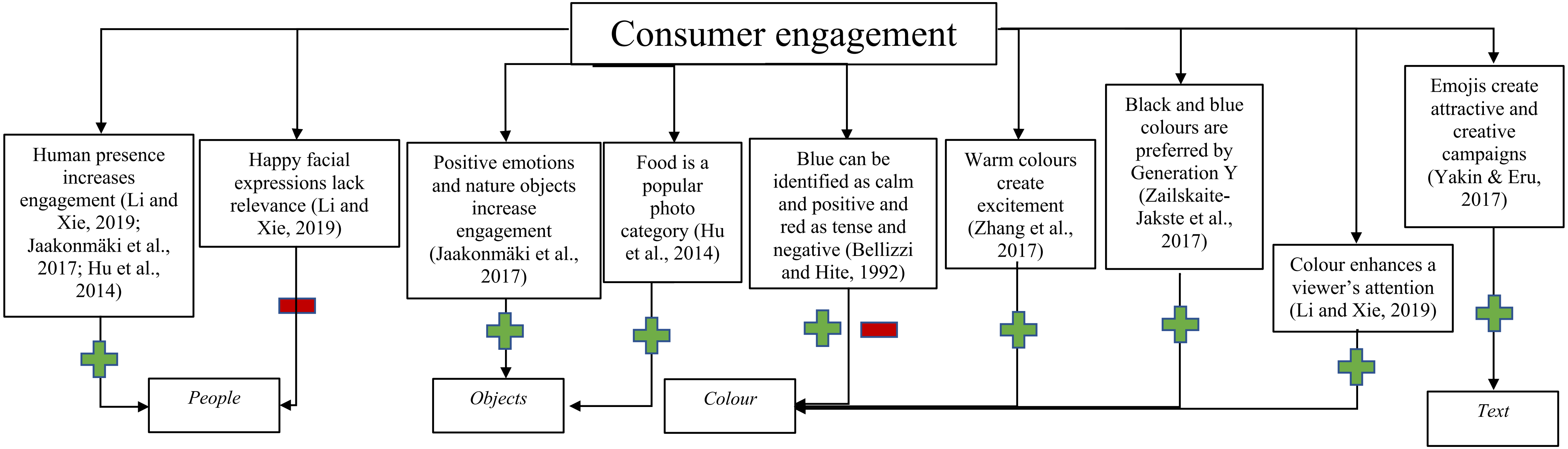
As several variables need to be generated for this research, a few cloud service platforms for feature engineering are compared. Singh et al. (2019) used the metrics: accuracy, speed, and cost to analyse different cloud providers: Google Cloud Vision API, Microsoft Computer Vision API and Amazon Image Recognition. All providers are powered by neural networks (NN) and can classify images into categories using object, colour, and face detection. NN score high on accuracy and perform well on multiple tasks Simonyan and Zimmerman (2015). Therefore, they can be used for a wide range of problems and will be used to identify colours, objects, and people in images in this study. Microsoft scores overall best on speed, followed by Google. Google scored best based on accuracy followed by Amazon. Amazon is cheapest, but all cloud providers offer a free trial period that will be enough for this study. Based on research done by Singh et al., I will continue feature engineering using the Google Cloud Vision API.

3 Conceptual framework

The conceptual framework on how using influencers can affect customer engagement is build based on the discussed literature. This framework can be found in Figure 3.1. The expected relationships between the post characteristics and customer engagement are based on the discussed literature and can be found in Figure 3.2.


Consumer engagement
Exclusivity has a positive influence on engagement (Ashley and Tuten, 2014)
Celebrity endorsement helps build brand awareness (Chen, 2017). Influencers create reliability (Forbes, 2016)
Images of other users appeal to users (Chen, 2017)
Perceived quality and promotion positively influence brand loyalty (Chan & Mansori, 2016; Kumar et al., 2016; Bazi et al., 2020; Li & Xie, 2019)
Amateurish visuals and incentives for participation result in higher engagement (Ashley and Tuten, 2014)
Interaction and entertainment positively effects consumer engagement (Liu et al., 2021)
Content: Including celebrity or influencer
Content: Influencers appeal more to users than models
Quality: Professional camera
Content type: Promotion
Content type: Interactive
Content
Quality
Content type
Content form

3.1: Conceptual framework: How using influencers affects Consumer engagement.


Consumer engagement
Human presence increases engagement (Li and Xie, 2019; Jaakonmäki et al., 2017; Hu et al., 2014)
Happy facial expressions lack relevance (Li and Xie, 2019)
Positive emotions and nature objects increase engagement (Jaakonmäki et al., 2017)
Food is a popular photo category (Hu et al., 2014)
Blue can be identified as calm and positive and red as tense and negative (Bellizzi and Hite, 1992)
Warm colours create excitement (Zhang et al., 2017)
Black and blue colours are preferred by Generation Y (Zailskaite-Jakste et al., 2017)
Colour enhances a viewer's attention (Li and Xie, 2019)
Emojis create attractive and creative campaigns (Yakin & Eru, 2017)
People
Objects
Colour
Text

3.2: Conceptual framework: What post characteristics affect Consumer engagement.

4 Data

4.1 Data scraping

Data from Instagram is scraped to investigate what characteristics of an Instagram post lead to the highest *Consumer engagement rate* (*CER*). As the set contains data of ten influencers in a tie frame of 100 posts, measured on March 26, 2021, the data can be defined as panel-data.

4.2 Top 10 beauty influencers on Instagram

The Instagram posts of the following beauty influencers will be analysed, with their username and number of followers in millions between parentheses: James Charles (@jamescharles, 27.2), Becky G (@iambeckyg, 25.7), Chiara Ferragni (@chiaraferragni, 22.8), Bretman Rock (@bretmanrock, 15.5), Nikkie Tutorials (@nikkietutorials, 14.5), Jeffree Star (@jeffreestar, 13.7), Nikita Dragun (@nikitadragun, 9.1), Makeup By Mario (@makeupbymario, 8.3), Naomi Giannopoulos (@vegas_nay, 6.4), Jaclyn Hill (@jaclynhill, 6.3). These influencers have the highest number of Instagram followers in the beauty industry worldwide measured on March 26, 2021.

4.3 Variable clarification

The post attributes are divided into five dimensions: *Colour, Objects, Person, Content,* and *Text* (Appendix: Table 8.1). The information captured in metadata is scraped from Instagram using JavaScript, Google Vision API and partly coded or collected by hand.

The dependent variable *Consumer engagement* will be measured by interaction with the post on Instagram. Engagement will be measured by the number of likes, comments, shares, and followers. This engagement rate will be measured separately for every post a company made and can be formulated in the following way:

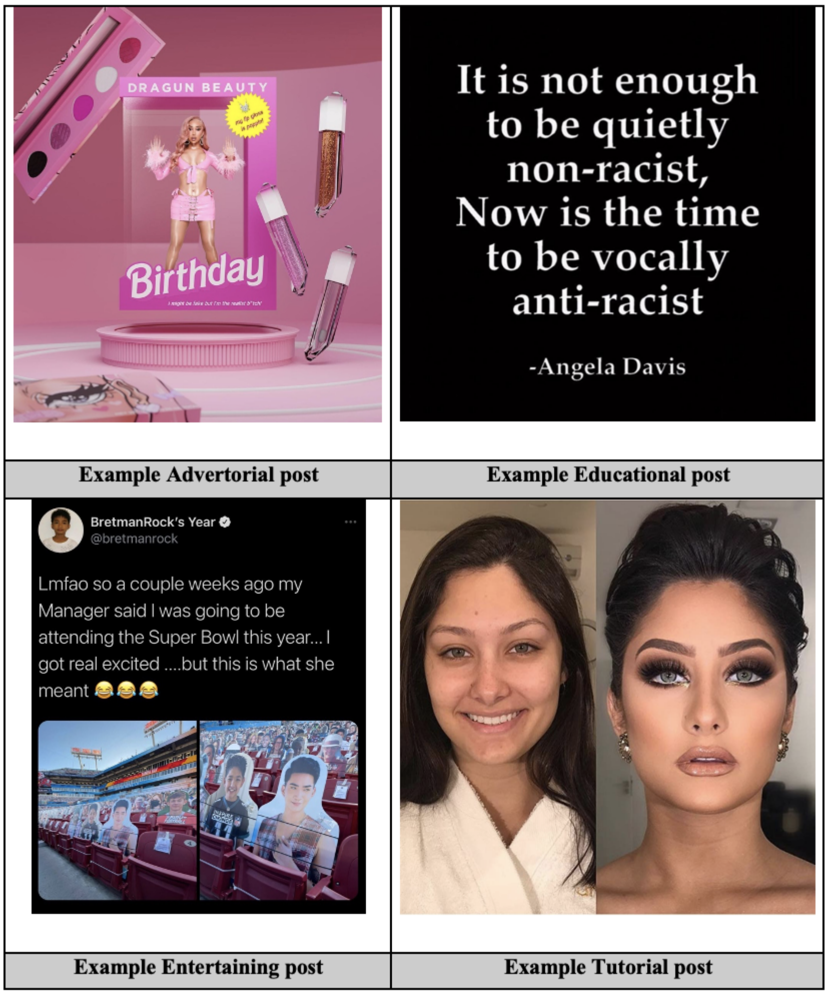
$$\textit{Consumer engagement rate} = \frac{\textit{(Number of comments + Number of likes)}}{\textit{Number of followers}} \quad * \quad \textit{100\%} \tag{1}$$

The first parameter, *Number of comments* refers to the number of individual comments on the Instagram post left by users. Users can leave multiple comments and therefore this parameter can include multiple comments by the same account. *The Number of likes* indicates the number of unique users that liked the Instagram post. Users can leave a like on a post to easily interact with the post. This way users can show they enjoy seeing the post, a user can only like a post once. The last parameter is the *Number of followers*. This parameter is unique for every brand and kept constant during this research. The *Number of followers* is measured on the 26th of March 2021 and rounded to a hundred thousand. Users can follow a profile to see all new posts of the profile on their feed, and that way keep updated on new content shared by that profile.

The independent variables are divided into five dimensions. The first dimension is *Colour*, the variable that belongs to this dimension indicates the most dominant colour used in the image. This is determined using a ML algorithm developed by Google Cloud Vision API. The second dimension is *Objects* and contains the variables *Beauty product visible?, Type of beauty product?* and *Main object*. The variable *Beauty product visible?* is a dummy variable specifying if a beauty product is visible in the photo or video. The most visible make-up product used on a person or seen in the visual is indicated by the *Type of beauty product*. The *Main object* represents the most dominant object included in the visual and is determined via Google Cloud Vision API and Python.

The third dimension *Person* includes four variables. *Person* is a categorical variable that indicates whether there are no people in the visual (value = 0), if one person is visible (value = 1) or if multiple people appear (value = 2). The second variable *Facial expression* is a categorical variable to categorise the emotion captured in the face of a person in the image. It has the following categories: anger, joy, and surprise. This variable is also created using Google Cloud Vision API and coding in Python. The variable *Full body visible?* Is a dummy indicating if the full body is visible in the visual. The final variable in this dimension is *Hair colour*, a categorical variable specifying the hair colour(s) that can be seen in the visual.

The fourth dimension is *Content* and contains the variables *Background setting, Professional visual?, Black and white visual?, Content type,* and *Video*. The categorical variable *Background setting* describes the background setting of the visual, for example if the photo or video is taken inside or in nature. The professionality is defined by the dummy *Professional visual?*, taking on a value of Yes if the visual quality is high and No if this is not the case. As pictures taken with a camera phone can have high quality as well, the pictures do not necessarily need to be taken with a professional camera. Grayscale visuals are accounted for by the variable *Black and white* *visual?*. The variable *Video* is a dummy that takes on a value of Yes for videos and a value of No for pictures. The final variable *Content type* is a categorical variable indicating if the visual can be considered to be advertorial, educational, entertaining or a (make-up) tutorial. These categories are partly based upon the Content Marketing Matrix (Chaffey, 2021) and previous literature (Chapter 2). Content classified as advertorial contains either product features, commercials, promotions or introduces a new product or tv-show. Educational content tries to teach the user about a certain social-cultural problem, such as the Black Lives Matter campaign. Entertaining content consists of (Tiktok) videos, games, viral and funny messages, or visuals that cannot be qualified under one of the other categories. Besides these categories, beauty influencers post tutorials where they show how to apply beauty products and how to recreate certain looks. Examples of the different types of content can be seen in Figure 4.1.



It is not enough to be quietly non-racist, Now is the time to be vocally anti-racist

-Angela Davis

Example Advertorial post

Example Educational post

BretmanRock's Year
@bretmanrock

Lmfao so a couple weeks ago my Manager said I was going to be attending the Super Bowl this year... I got real excited ....but this is what she meant 😂😂😂

Example Entertaining post

Example Tutorial post

Figure 4.1: Examples of *Content types*

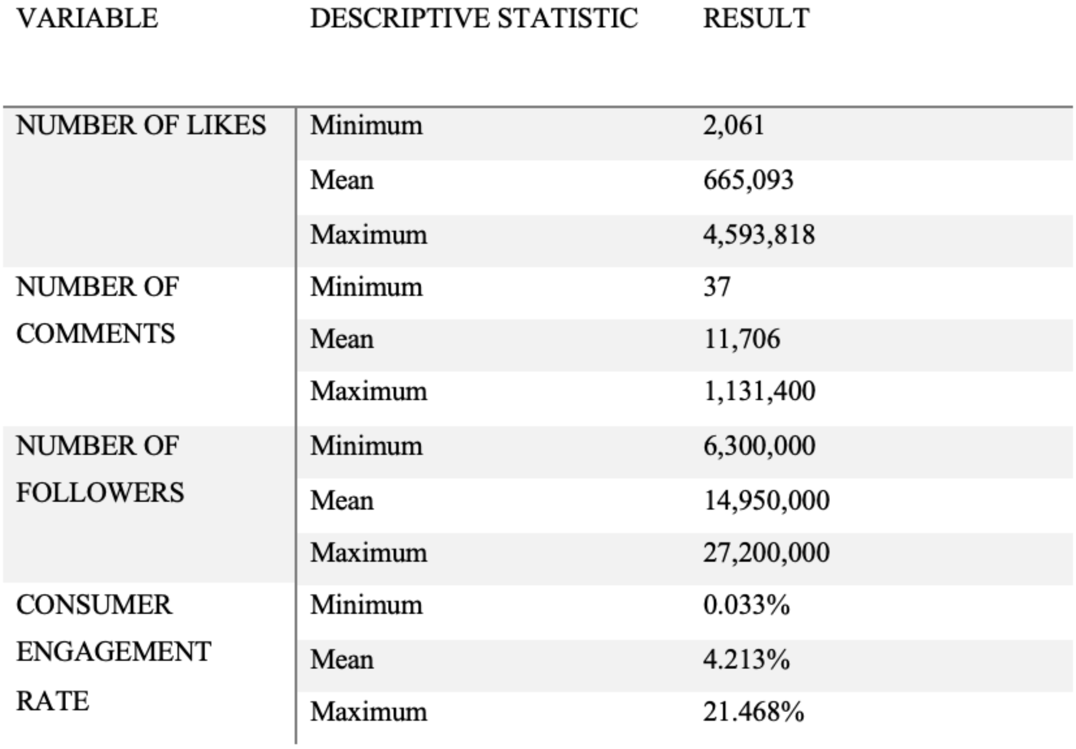
The final dimension is *Text*. The first variable *Caption* indicates whether a visual has a caption or not. If a caption contains emotions the variable *Emoji* gets a value of Yes. The variable *Tagged accounts* is a dummy indicating if other users are tagged in the photo or caption. The final variable *Hashtags* equals Yes if the influencer uses hashtags in the caption. A visual representation of the dimensions and the independent variables can be found in the summary table in the Appendix (Table 8.1).

The names of the beauty influencers are not taken into the ML model as a (dummy) variable. The *Consumer engagement* is relative to the *Number of followers*, thus the model already controls for fixed effects of certain beauty influencer having more engagement as. Therefore, there is no need to add another variable to control for this.

4.4 Descriptive statistics

The data set includes 1,000 posts of the top ten beauty influencers on Instagram, of which 761 of these posts include an image and 239 posts include a video. The descriptive statistics of the full data set in the interest of general data exploration can be found in Table 4.1.

Table 4.1: Descriptive statistics



VARIABLE	DESCRIPTIVE STATISTIC	RESULT
NUMBER OF LIKES	Minimum	2,061
	Mean	665,093
	Maximum	4,593,818
NUMBER OF COMMENTS	Minimum	37
	Mean	11,706
	Maximum	1,131,400
NUMBER OF FOLLOWERS	Minimum	6,300,000
	Mean	14,950,000
	Maximum	27,200,000
CONSUMER ENGAGEMENT RATE	Minimum	0.033%
	Mean	4.213%
	Maximum	21.468%

As can be seen in Table 4.1, the range of the (parameters of) engagement is quite big. On average a little bit over 650 thousand users like a post. The range of likes includes a minimum of 2,061 likes and a maximum of over 4.5 million likes. The spread of the number of comments a post in the data set has ranges from 37 to over 1 million comments, with an average of 11,706. The number of followers also entails a large range. On average the beauty influencers have almost 15 million followers. The beauty influencer with the least number of followers has 6.3 million followers (Jaclyn Hill) and James Charles tops the list with 27.2 million followers.

The average engagement rate for Instagram posts of celebrities found in a study by Naumanen and Pelkonen (2017) is 7.4%. They defined the average engagement rate as the sum of the engagement rates of every picture posted by a celebrity divided by the total number of photos in the data set. As illustrated in Figure 4.2 and Table 4.1, there can be found a tremendous difference in the *CER* as formulated by Formula 1 among beauty influencers. The highest average *CER* is achieved by Nikita Dragun, averaging on almost 11%, scoring better than the celebrities studied by Naumanen and Pelkonen (2017). The lowest average engagement rate belongs to Makeup by Mario, ending the list with an average of 1.78%. Over the whole data set the average *CER* is found to be 4.21%. The result found is significantly lower than that of Naumanen and Pelkonen (2017).

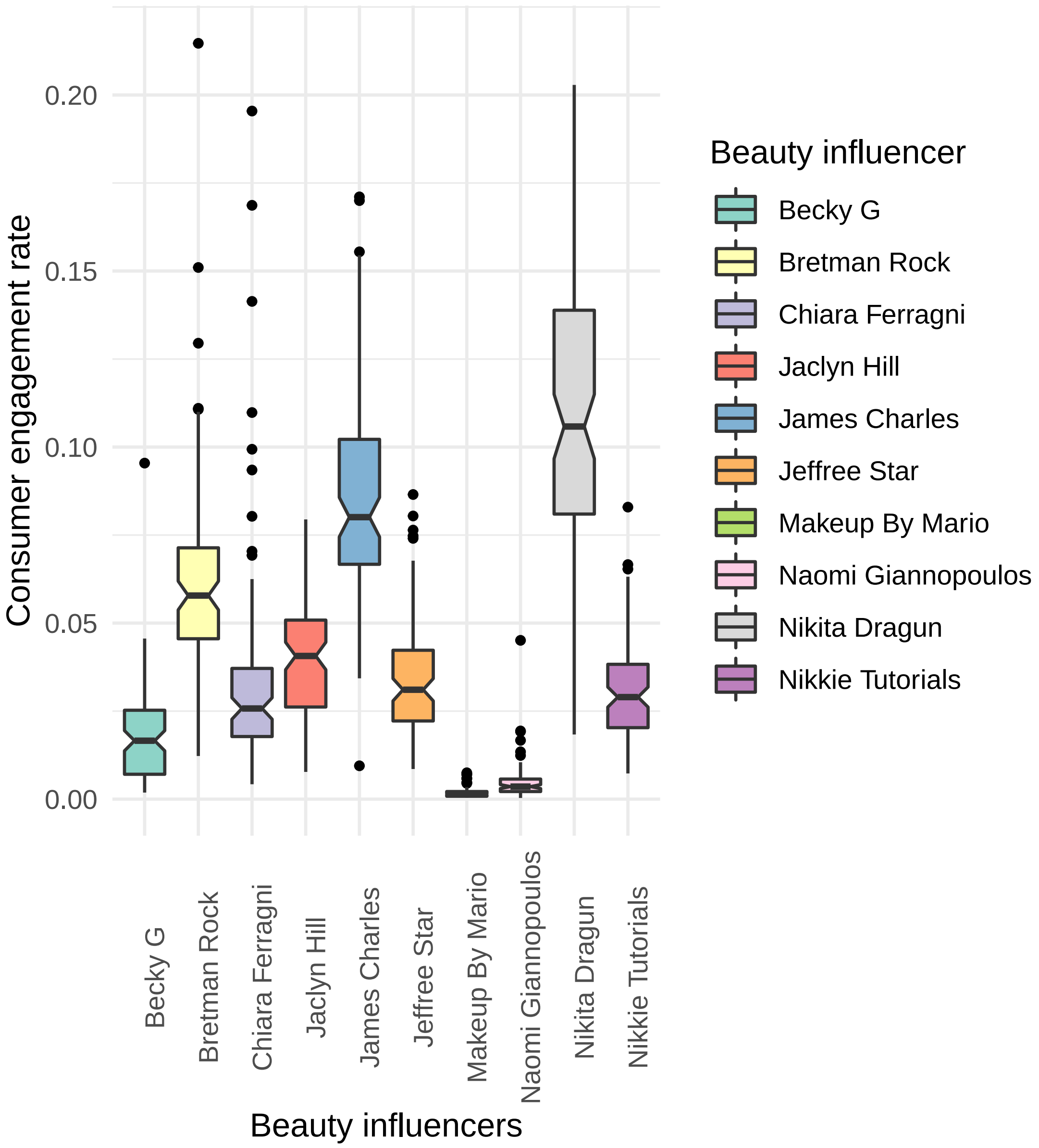


Figure 4.2: *Consumer engagement rate* per beauty influencer

The type of content of the Instagram posts analysed in this research are classified as either advertorial, educational, entertaining or a tutorial. More than 60 percent of the posts are found to be entertaining, as can be seen in Figure 4.3. This category includes all posts that are found to be funny or cannot be classified under one of the other categories. The second biggest category is advertorial.

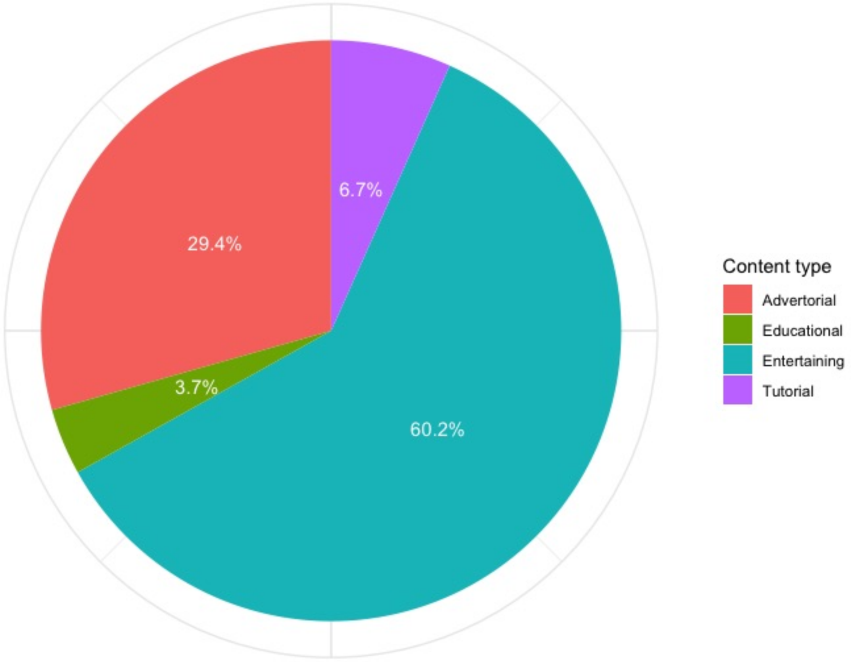


Figure 4.3: Distribution of *Content types* of Instagram posts

4.5 Data preparation

Due to the fact that the machine learning model (ML) implemented by Google Cloud Vision API can only work on images, some variables of the observations containing videos were coded by hand.

4.5.1 Creating the variables: *Dominant colour, Main object, Person* and *Facial expression*

The variables *Dominant colour, Main object, Person* and *Facial expression* are determined using Google Cloud Vision API (Rash, 2019). The API was trained on large datasets and therefore performs well in the classification phase. The Google Cloud Vision API is based on a deep neural network to analyse images and assign labels within a confidence level [0:1] to classify properties of an image, for example the most dominant colour used or the content of the image. The network does this by sending the base64 encoded string of the content of the image file through the network (Google Cloud, 2021). To identify the dominant colour the Vision API analysing all Red-Green-Blue (RGB) colour values and detecting the RGB value that is most dominant (Google Cloud, 2021). For this study the variable category with the highest confidence score will be selected for every image. Similar objects of the *Main object* variable were converted to standardized categories, such as all flowers and plants are converted to the category plant and all clothing items were converted to the category fashion. Google Cloud Vision API was used together with Python and RStudio. In order to quantify visual content features, this research used word tagging. Word tagging is a textual description of a content feature existent in a visual. The videos were manually watched to fill in the values for the variables for the visual as the API does not work on videos. The API labels *Facial expressions* relating to Anger, Joy, or Surprise. If no person is included in the image, this variable received a value of None.

4.5.2 General patterns

Videos are found to be more interactive (de Vries, Gensler, & Leeflang, 2012). As interactive posts lead to more *Consumer engagement* as proposed by X. Liu et al. (2021), it is of interest to find out if this is true for this research. A link to an example of an interactive video calling for users to participate in a dance challenge can be found under References (@iambeckyg, 2021). On average, about 23.9% of the posts include a *Video.* The relative use of videos differs tremendously among influencers. Nikkie Tutorials posting relatively the most recordings, 50% of her posts and Chiara Ferragni posting the least videos, only 6% of her posts. The result of this research contradicts previous literature by X. Liu et al. (2021) that interactive posts lead to a higher level of customer engagement. As shown in Table 4.2, the average *CER* is greater for posts including an image instead of a video. This can be concluded as the one-way ANOVA test indicated that the mean of the groups differ. This result suggests that for marketing strategies, images are preferred by followers and lead to a higher level of *Consumer engagement.*

Table 4.2: Average Consumer engagement rate videos vs. images

Includes *Video*	Average *Consumer engagement rate*
Yes	0.0357
No	0.0442

Colour use is another interesting characteristic to look at and find out if a general pattern can be found among. As reflected in Figure 4.4, some colours are used more often than other by beauty influencers in their Instagram images. The *Dominant colours* used are black, blue, orange, and pink. Black is the *Most dominant colour* in almost 200 visuals in total. As illustrated in Figure 4.4 most influencers use a variety of colours in their posts and none seem to opt for a certain colour palette. The effect of colour use in the prediction of *Consumer engagement* will be further analysed in Paragraph 6.4.1.

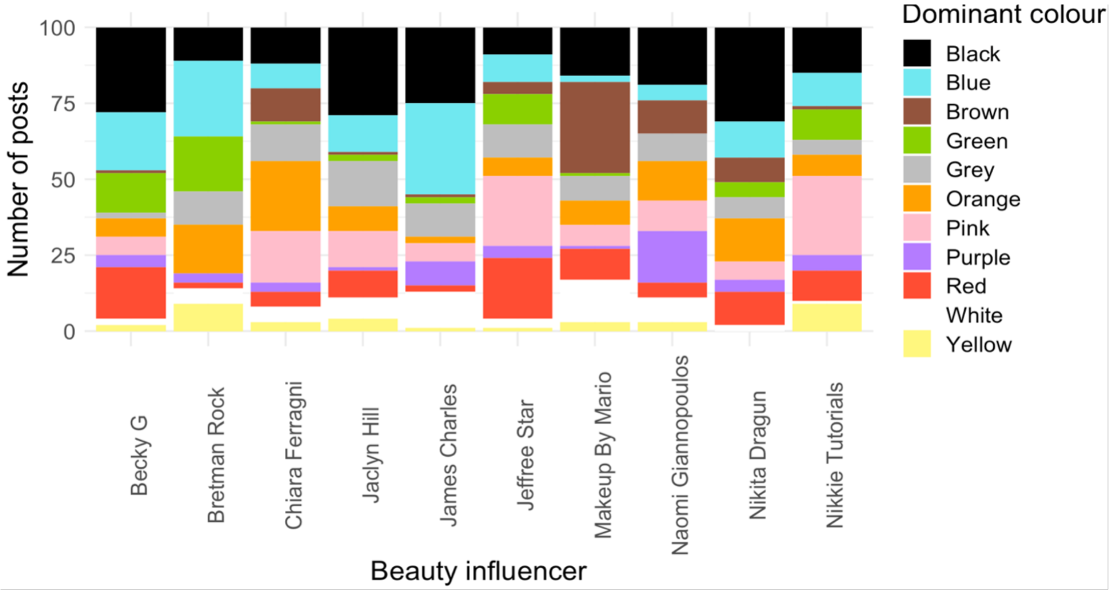


Figure 4.4: Most dominant image colour per beauty influencer

Content type is a third objective in which their might be a common pattern. The type of content posted differs tremendously per beauty influencer (Figure 4.5.2). No general trend can be found except that most influencers post predominantly entertaining or advertorial content. For instance, James Charles only posts content that can be found to be entertaining. Unlike Nikkie Tutorials, who posts mostly advertorial content or tutorials. I expected to find more interactive content such as tutorials on account of research done by X. Liu et al. (2021). But as the same research concluded that entertaining content also leads to a higher consumer engagement, it makes sense this category is the most popular. I investigated if these two categories of *Content type* influence the *CER* in a positive way.

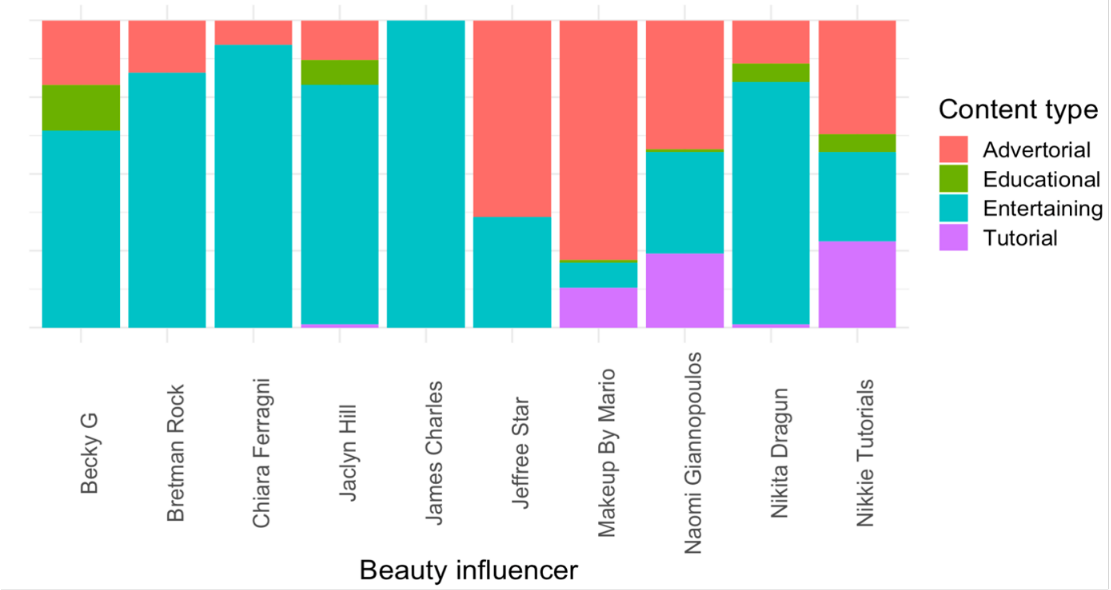


Figure 4.5: Distribution of Content type per beauty influencer

Influencers try to stand out and are typically known for something that characterizes them. As this study focuses on beauty influencers, I investigated if these influencers try to stand out by using certain beauty products that characterise them. Although all influencers use multiple types of beauty products, some do favour certain products. For example, Nikkie Tutorials has a thing for heavy eye make-up (eye-shadow), blush is a product that characterises Chiara Ferragni. The most used beauty products are eye-shadow, and lipstick. These products , together with wearing no make-up. These are used in the far majority of the posts (71.75%). Therefore, we can conclude that a general pattern can be found among the *Type of beauty products* used. Taking into account that some influencers do tend to distinguish themselves and their looks by opting for a certain beauty product.

Moving on to the textual elements, there can be found a general pattern among *Caption* use. All influencers use *Captions* for almost every post. Another trend can be found in *Emoji* use. Nikkie Tutorials uses *Emoji* the most, a total of 99 times out of a hundred. While Nikita Dragun uses *Emoji* the least (2/3$^{\text{rd}}$ of posts). Thus, in general beauty influencers use *Emoji* in their captions. The use of *Hashtags* is a lot more versatile. James Charles uses no *Hashtags* at all, while on the other hand Naomi Giannopoulos uses *Hashtags* in 71 out of a hundred captions. No general trend can be found in the number of times influencers *Tag* other accounts in their posts.

It can be concluded that influencers generally have different marketing strategies and mostly no patterns can be found. But the data does give some information away. It can be inferred that black is the most popular *Dominant colour* in the visuals. One may also conclude that most influencers post either entertaining or advertorial *Content.* It can be said that wearing eye-shadow, lipstick or no make-up is most popular among beauty influencers. Influencers do tend to distinguish themselves by focusing on one *Type of beauty product. Captions* and *Emoji* use are popular among beauty influencers, while no pattern can be found in use of *Hashtags.*

As influencers work together with different brands, it is of interest to find out if tagging these brands or using a hashtag with the cosmetic brands name is of influence on the *CER.* To analyse this, a data set is created with all beauty influencers and the cosmetic brands they name. The *Consumer engagement rate* is calculated as in Formula 1, with the difference that the *Number of followers* of the cosmetic brand is taken into account instead of the *Number of followers* of the influencer. Cosmetic brands that are named often are wide known beauty companies and brands of the influencers themselves, such as Jeffree Star Cosmetics. The fixed effects a certain influencer has is controlled for by the model. The model does not control for the fixed effects a cosmetic brand might have. Therefore it is of interest to see if the *CER* differs among these brands. It could be that engagement is higher, because a cosmetic brand already has a higher level of visibility than other brands. Intriguingly, the *CER* does not differ significantly among naming different cosmetic brands in Instagram posts. It can be concluded that the relative amount of interaction is different based on the beauty influencer itself and not the cosmetic brand they advertise for. The conclusions in this paper apply to all (luxury) cosmetic brands and beauty influencers.

5 Methodology

5.1 Method selection

Based upon previous research by Singh et al. (2019) as discussed in Chapter 2, Convolutional Neural Networks were used together with Google Cloud Vision API to create the variables for *Dominant colour, Objects, Person,* and *Facial expression.* Subsequently, a model needs to be built in order to predict the *Consumer engagement rate.* The models that were evaluated are a multivariate linear regression, an ordinal logistic regression and a decision tree. These models were selected based upon research discussed in Chapter 2, appropriateness for the data, prediction accuracy and ease of interpretation.

5.2 Convolutional Neural Networks used by Google Cloud Vision API

To create the variables for objects, person, and facial expression Google Cloud Vision API was used. The machine learning models imposed by Google are trained on large image data sets and therefore highly accurate. The models used in this research can classify pictures into categories to discover objects, people and facial expressions based on a confidence value. The model used for image classification is called a Convolutional Neural Network (CNN). These models can be used for either classification, clustering, or prediction (Smith, 2015). The model has an input layer, several layers that transform the input and an output layer that gives the confidence scores of each classification label (Y. Li & Xie, 2019). All layers are made up of a set of neurons, connected to the neurons in the preceding layer, the structure of a CNN model can be seen in Figure 5.1.

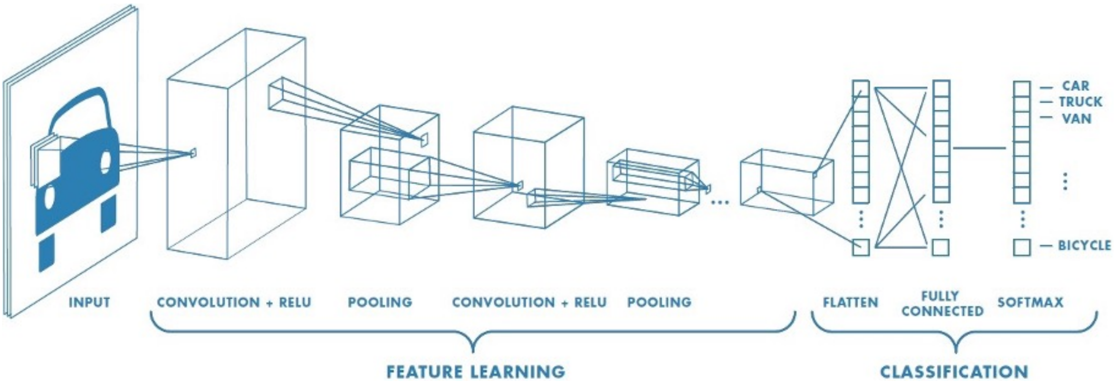


Figure 5.1: Structure of a CNN model (Asiri, 2021)

A common drawback of CNN models is that the models are very time consuming (Bhuiya, 2020). This is resolved by Google Vision API, since their CNN models are trained on huge image data sets. The second drawback of CNN models is the need for large data sets to train the models. This limitation is also eliminated by using the pre-trained models provided by Google. A final limitation of CNN models, is the interpretability. Black box models have low interpretability, but this apposes no real limitation for this research as it only uses the classification labels.

In this research CNN models will be used to label objects in images, identify the most dominant colour in images, identify whether a person is included in the image and what facial expression(s) are present.

5.3 Regression Models

To analyse what content factors of an image drive *Consumer engagement* on Instagram, a model that predicts engagement needs to be created. Popular models to predict the level of engagement are regression models (Sadeque & Bethard, 2019).

5.3.1 Multivariate Linear Regression Model

Since the data set used in this research has multiple predictor variables a multivariate regression model can be used to estimate the *CER*. The most common method used to estimate the β_{n} coefficients of such a model is ordinary least squares (OLS). This method tries to fit the best line for each predictor variable and therefore minimizes the sum of squared residuals. The regression model assumes a normal distribution of the data and a linear fit through the data points (Foley, 2020a).

In all likelihood the most substantial drawback of using OLS is the linearity assumption (Hutcheson & Sofroniou, 1999). This assumption is met since the data is recoded into dummies and that way have a linear relationship by definition. Another assumption is that of no multicollinearity among the variables. Due to recoding into dummies, correlation might exist. A third assumption is homoscedasticity. These assumptions have been tested and met.

5.3.2 Ordinal Logistic Regression Model

As proposed by Sadeque and Bethard (2019) logistic regression is a simple way to make predictions and understand what attributes are important in the prediction task. Logistic regression models have the advantage that the linearity assumption is omitted and therefore have a better fit than OLS models. To use logistic regression, the dependent variable *Consumer engagement rate* needs to be transformed into an ordinal variable with multiple ranges indicating the engagement levels. The outcome groups are ordered since a high level of engagement is considered better than a lower level. The OLR model can be expressed in the following way (Foley, 2020b):

$$\mathit{Logit[}P(y_{\mathrm{i}} \leq j)] = \mathit{log}\left[\frac{P((y_i) \leq j)}{P((i' > j'))}\right] = \alpha_j - \beta X, \quad j\varepsilon[1, J-1] \qquad (2)$$

Where $j\varepsilon[1, J-1]$ in Formula 2 indicates the levels of *Consumer engagement* as an ordinal outcome. β indicates a set of coefficients for the predictor variables X. The intercept of each outcome level is specified by α_j. Formula 2 indicates the log-odds of falling into categories $y_i > j$ minus the log-odds of falling into categories $y_i \leq j$. The log-odds can be transformed into odds and (cumulative) probabilities for ease of interpretation.

This ordinal logistic regression (OLR) model predicts the probability of an Instagram post taking on a certain level of *Consumer engagement* based on its features. Since all independent variables are categorical, the coefficients can be interpreted in the following way: The probability of falling into category j when using variable category x instead of the base category changes by a percentage. This way advise can be given on what features to use and how this exactly translates to probabilities of higher engagement levels.

A limitation of OLR models is that the data needs to be transformed to ordinal categories. This will cause some loss of information as no exact engagement rate can be predicted. This imposes no real limitation as engagement levels are sufficient in giving advice to influencers and brands. One assumption of OLR is the proportional odds assumption, therefore only one set of coefficients is needed (Foley, 2020b). Another disadvantage of OLR is the interpretability of the coefficients, since these need to be transformed to probabilities. Transformation is easy and therefore this imposes no real limitation.

5.4 Decision trees

Another popular method for supervised learning problems is decision trees (Sadeque & Bethard, 2019). Decision trees can handle numerical and categorical data and are easy to interpret. A downside of decision trees is that they are prone to overfitting the data. They often outperform OLS as they do not assume linearity. No sign of overfitting was found in this study.

The algorithm forms trees by recursive partitioning of the data into homogeneous subgroups. These subgroups have a similar value of *Consumer engagement rate*. The algorithm fits the mean value of the within group dependent variable, by recursively splitting the data based on a binary question about each feature. These splits are called nodes. The splits are based upon an impurity measure and for regression the weighted mean squared error (MSE) is used:

$$MSE(t) = \frac{1}{N_t} \sum_{i \varepsilon D_t} (y_i - \widehat{y_t})^2 \tag{3}$$

Where, N_t indicates the number of training samples at node t, D_t can be seen as the training subset at node t. y_i and $\widehat{y_t}$ indicate the true and the predicted response value.

The first split at the root node is based upon the feature that results in the largest information gain. The objective function used to maximise the information gain at each split can be expressed in the following way:

$$IG(D_r, f) = I(D_r) - \left(\frac{N_{left}}{N_r} (D_{left}) + \frac{N_{right}}{N_r} I(D_{right}) \right) \tag{4}$$

f indicates the feature the split is based on. D_p is the data set of the root node and D_{left} and D_{right} are the data sets of the internal nodes. I refers to the impurity measure based upon Formula 3. N_r is the number of samples taken at the root node and N_{left} and N_{right} indicate the number of samples taken at the splits (L. Li, 2019).

For classification the Gini index is used as an impurity measure. Gini measures the chances of misclassifying an observation based on the assumption that the observation is randomly assigned to a class according to the class distribution. The Gini index is defined as follows:

$$Gini = 1 - \sum_{i=1}^{n} (p(c_i|t))^2 \tag{5}$$

In every node t, the Gini index assigns an observation to a class c_i, so that the Gini index can be expressed as 1 minus the sum of probabilities of an observation being classified to a certain label. As we want splits to be based on classifying all elements to a specified class, we want to minimize the Gini index at each split. The value of the Gini index can range between [0:1] (Raileanu & Stoffel, 2004).

Splitting is done until the stopping criteria is reached. A frequently used criteria is the maximum depth of a tree (Mantovani et al., 2019). The model predicts the response variable based upon the mean dependent variable of observations that belong to that subgroup, the final node that holds the prediction is called a leaf node. The leaf node holds the predicted engagement level based upon certain Instagram post attributes.

5.5 Method application

5.5.1 Dividing data into a test and train set

Before splitting the data into a test and train set, all categorical variables are set to factors so that the algorithms will treat them correctly. As the data used in this study consists of only 1,000 observations and 17 explanatory variables, splitting the data will drastically reduce the availability of data in the training process. A solution for this is k-fold cross-validation. In this procedure the data is split into a train (70%) and a test set (30%) and the model is trained and validated repeatedly on k number of samples of the training data. By introducing cross-validation, hyperparameter tuning can improve the predictive performance without loss of training data. As proposed by Kuhn and Johnson (2013), the number of folds depends on the size of the data set. Considering the size, the number of folds used in the training of the algorithms is equated to ten.

5.5.2 Reshaping the data for multivariate linear regression

Before running the multivariate model, the number of category levels of the variable *Main object* will be reduced by combining similar levels. This way no unnecessary levels will remain, and the sample sizes of each category are larger which will increase the performance. For example, all limbs are recoded into *body parts.*

5.5.3 Reshaping the data for ordinal logistic regression

As stated in Paragraph 5.5.2 the dependent variable needs to be transformed into ordinal categories. *Consumer engagement rates* with a value of less than 1.5% are recoded into the level *Low*, rates with a value between 1.5% and 3.5% are considered to be *Medium-low*, if the engagement rate lays within 3.5% and 6.5% it is considered as *Medium-high.* Rates above 6.5% are revalued as *High.* These intervals were chosen to create balanced classes with roughly equal sample sizes in each interval category.

6 Results

The results of the models discussed in Chapter 5 will be debated here. To select the final model hyperparameter tuning, assumptions and ease of interpretation will be taken into account. Besides these features, the Root Mean Squared Error (RMSE) will be used for the models with a continuous response variable. Balanced accuracy will be used as an accuracy measure for the models with an ordinal response variable. The RMSE is calculated using the following formula (Hyndman & Koehler, 2006):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}} \tag{6}$$

Here n is equal to the number of observations in the data set. $\widehat{y_i}$ are the predicted values of the *CER* and y_i the actual values of the *Consumer engagement rate*.

Based upon research by García et al. (2009), it is best to look at the balanced accuracy when the data might contain imbalanced classes. This accuracy measure takes all four classes into account and does not mislead when imbalance occurs. The balanced accuracy is defined as the mean accuracy derived by each class (Brodersen et al., 2010).

6.1 Results of the multivariate linear regression model

The multivariate regression model has a R-squared value of 0.471 and an adjusted R-squared of 0.4158. This value is considered to be moderate in marketing research according to Sarstedt and Mooi (2019). Others already consider a value of 0.20 as high in consumer behaviour studies, especially in a social media setting (Vock et al., 2013). As no definitive criteria can be found in literature, predictive performance of the model is evaluated. The average *CER* of the test set equals 4.405%, while our model predicted an average of 4.367%. The RMSE equals 3.291%. Overall, we can conclude that the multivariate model predicts quite well on the test set.

6.2 Results of the ordinal logistic regression model

After reforming the *Consumer engagement rate* into an ordinal variable, the OLR model is built on the train set. The final OLR model is based upon all explanatory variables. The accuracy on the test set equals 50% and the balanced accuracy is equal to 66.527%. The balanced accuracy can be considered quite good in marketing research. The confusion matrix of the predictions on the test set can be seen in Table 6.1. The model succeeds in predicting all classes.

Table 6.1: Confusion matrix predictions on the test set ordinal logistic regression model

	Low	Medium-low	Medium-high	High
Low	59	18	10	3
Medium-low	14	29	16	4
Medium-high	12	21	26	17
High	5	8	22	36
Accuracy	75.40%	61.49%	56.51%	72.71%

6.3 Results of the decision tree model

The decision tree models are built in RStudio using the rpart package (2019). In order to optimize the model, a grid search is done to tune the tree using the default of 10-fold cross validation. First, the model is built using the continuous response variable *Consumer engagement rate*. This model with the lowest RMSE is a decision tree with a cost complexity (α) of 0.01, a depth of 13 nodes and a minimum of 13 observations per split. The final model has a RMSE of 3.309%. Based on the RMSE it can be concluded that the decision tree model predicts almost as well as the multivariate linear regression model.

Secondly, the model is trained on the ordinal response variable *CER* as proposed under Paragraph 5.5.3 and splits are made based upon the Gini index. This resulted in a decision tree model with the following hyperparameters: a cost complexity (α) of 0.01, a depth of nine nodes and minimum 12 observations per split. The confusion matrix of this optimal decision tree with an ordinal response variable can be seen in Table 6.2. The structure of this tree can be seen in Figure 6.1 and 6.2, divided over two pages for the left and right side of the tree. The leaf nodes hold the predicted engagement level. The lower engagement levels are represented by red colours. The higher engagement levels are represent by green colours. The accuracy of the model on the test set equals 46.33% and the balanced accuracy is equal to 63.903%. The resulting decision tree model predicts less well than the ordinal logistic regression model.

Table 6.2: Confusion matrix class predictions on the test set decision tree model

	Low	Medium-low	Medium-high	High
Low	51	14	7	3
Medium-low	16	36	20	8
Medium-high	9	13	21	18
High	14	13	26	31
Accuracy	72.62%	63.86%	55.34%	64.79%

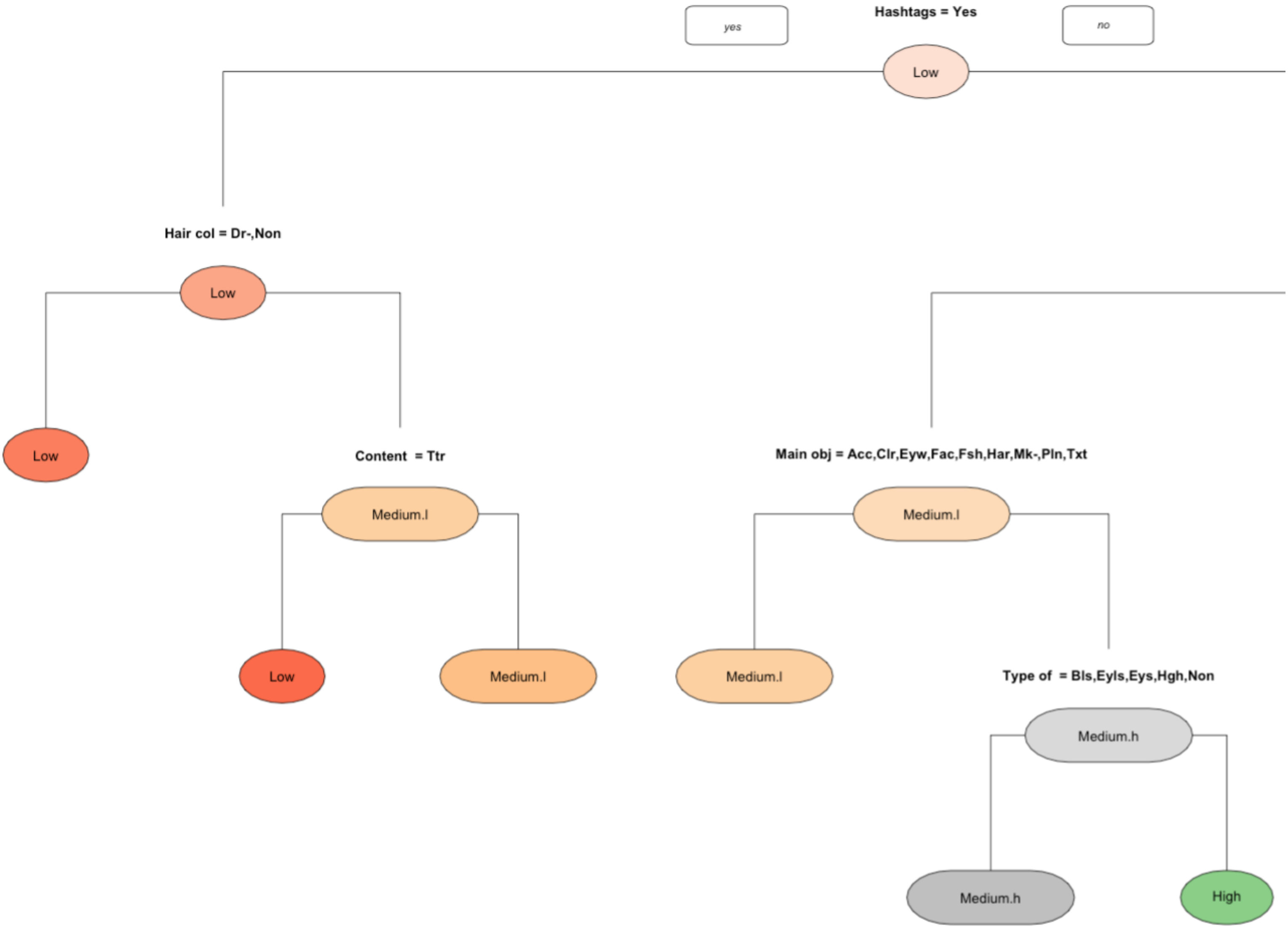


Figure 6.1: Structure of the decision tree classification model left side

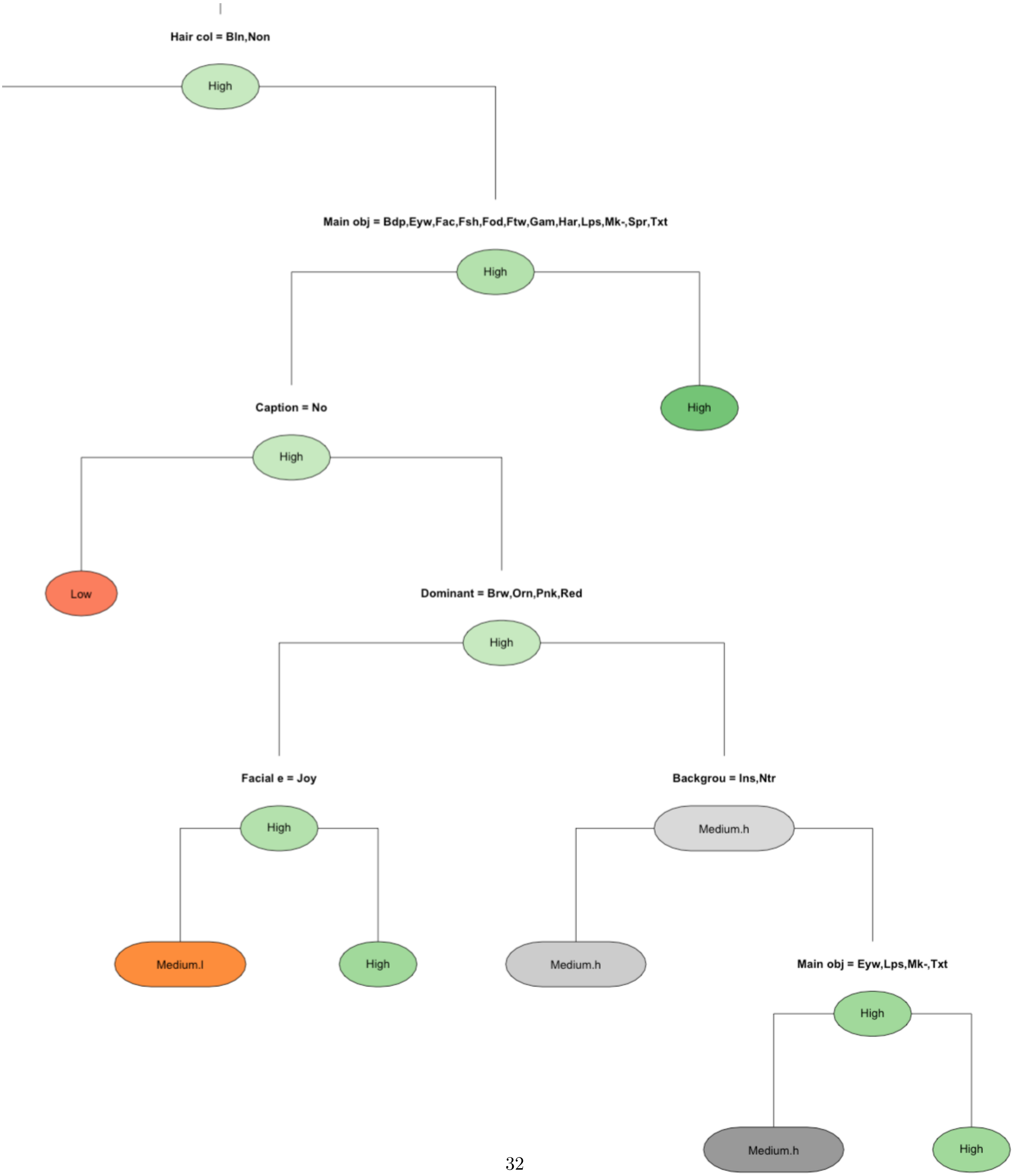


Figure 6.2: Structure of the decision tree classification model right side

The importance of the variables used in the model can be accessed using the goodness of split measure. This equals the second part of Equation 5. As is apparent in Table 6.3, the use of *Hashtags* in post captions is considered to be the most important variable in the decision tree model, followed by the *Hair colour.* The high influence of *Hashtags* could be due to the increased visibility.

Table 6.3: Top five important decision tree classification

	Sum of goodness of split measure
Hashtags	48.437
Hair colour	35.952
Main object	34.147
Content type	21.130
Background setting	17.870

Based upon the balanced accuracy the OLR model can be considered to be the best prediction model. As this research does not only focus on prediction itself but on finding out what influence certain features have on the prediction, the interpretation of the coefficients is most important. For these reasons the OLR model is the optimal model in answering the research question, but the decision tree classification model can be of major importance to influencers in deciding what content to post. The model can be a guideline in choosing what to post in order to achieve the highest level of *Consumer engagement.* For example, choosing the *Dominant colour* based upon a chosen *Main object.* The significant coefficients of the final model can be found in Table 6.4. For all coefficients, see Table 8.2 in the Appendix.

Table 6.4: Significant coefficients of the full ordinal logistic regression model

	Value	Std. Error	*t value*	*p value*
‘Hair colour‘Colourful	0.914	0.322	2.836	0.005 ∗ ∗
‘Hair colour‘None	−1.645	0.452	−3.641	0.0003 ∗ ∗∗
‘Full body visible?‘Yes	0.429	0.218	1.969	0.049∗
‘Type of beauty product‘Eyebrows	1.714	0.763	2.247	0.025∗
‘Type of beauty product‘Eyelashes	1.294	0.489	2.645	0.008 ∗ ∗
‘Type of beauty product‘Eyeliner	1.121	0.565	1.983	0.047∗
‘Type of beauty product‘Eye-shadow	0.990	0.444	2.227	0.026∗
‘Type of beauty product‘Lipstick	1.346	0.441	3.054	0.002 ∗ ∗
‘Type of beauty product‘Nails	1.530	0.711	2.153	0.031∗
‘Type of beauty product‘None	1.203	0.444	2.708	0.007 ∗ ∗
‘Professional visual?‘Yes	0.460	0.213	2.157	0.031∗
‘Content type‘Educational	2.027	0.700	2.893	0.004 ∗ ∗
CaptionYes	3.092	0.858	3.604	0.0003 ∗ ∗∗
HashtagsYes	−2.397	0.229	−10.468	0 ∗ ∗∗
Medium-low|Medium-high	4.472	1.862	2.402	0.016 ∗ ∗
Medium-high|High	6.059	1.870	3.241	0.001 ∗ ∗

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

6.4 Interpretation of the models

6.4.1 Colours

To give meaningful advice to influencers, it is important to analyse what attributes lead to more *Consumer engagement* to improve this KPI. Based upon the insignificant coefficients of the OLR model in Table 8.2 (Appendix), we cannot draw meaningful conclusions about the influence of *Dominant colour* on *CER*.

The decision tree model in Figure 6.1 and 6.2 tell us what colours lead to a higher level of *Consumer engagement* under certain conditions. Based upon Figure 6.2, we can conclude that if the colour brown, orange, pink, or red is the most dominant, the split is made to a higher *Consumer engagement level.* This conclusion can only be drawn based upon the former splits. Contingent on simple analysis, it can be found that the colours black and blue have, in general, have a higher average *CER* than some other colours. This can not be said about the colour brown. This result is in line with previous research by Zailskaite-Jakste et al. (2017), who concluded that black and blue colours are preferred by Generation Y. Another research that found a positive influence of the colour blue is by Bellizzi and Hite (1992). Based upon the results of this study, no clear conclusion can be drawn about the influence of colour use in visuals of beauty influencers on Instagram.

6.4.2 Objects

Based upon the OLR model (Appendix: Table 8.2), we can only draw meaningful conclusions and make recommendations based upon the coefficients of the variables that are found to be significant. None of the categories of *Main object* are found to have a significant influence on the *Consumer engagement rate.*

The decision tree model in Figure 6.2 can tell us what the influence in the prediction of a classification level is in a specific setting. For instance, body parts, eye wear, face, fashion, footwear, games, hair, lips, make-up, sports, or texts will probably lead to the highest level of *Consumer engagement.* This conclusion can only be drawn if the *Hair colour* is either blonde or none. As this study focuses on beauty influencers, it is of no surprise that visuals focusing on eye wear, faces, hair, lips, or make-up tend to score better on *Consumer engagement.* A lot of these categories focus on human presence, the positive influence of human presence were already been established in literature by Y. Li and Xie (2019) and Jaakonmäki et al. (2017). The decision tree model (Figure 6.2) tells us that including fashion or footwear as a dominant object increases the probability of falling into a high engagement level. As was expected, these findings contradict previous research by Hu et al. (2014) that fashion is one of the least popular content categories.

The second variable of the dimension *Object* is *Beauty product visible?.* Although including beauty products in visuals has a positive effect on the *CER* (Appendix: Table 8.2), no meaningful conclusions can be drawn, due to insignificance in the OLR model.

The final *Object* variable that this research investigates is the *Type of beauty product.* As most beauty influencers tend to wear make-up, it is of interest to find out what beauty products lead to more engagement on social media. The coefficients of Table 8.2 (Appendix) can be read as reference to the base level (blush). Make-up that focuses on the eyes such as wearing fake eyelashes, have a significant positive influence on the *CER*. The same goes for wearing lipstick, nail polish, or fake nails. For example, the odds of falling into a higher *Consumer engagement rate level* when wearing fake eyelashes is estimated to be 2.647 times ($exp(1.294) - 1$) higher than if someone is wearing blush. This equals an increase in the probability of belonging to a higher *Consumer engagement rate level* of over 78% ($\frac{exp(1.294)}{(1+exp(1.294))} * 100\%$). Interestingly, opting for an all-natural look by wearing no make-up also positively affects the *CER*. The increase in the probability of falling into a higher *Consumer engagement rate level* compared to the reference level of the significant *Type of beauty product* categories can be seen in Table 6.5. These conversions were made by transforming the log-odds to probabilities. Using make-up to enhance the appearance of your eyebrows has the biggest positive influence on the *Consumer engagement* in reference to wearing blush.

Table 6.5: Significant effects of different beauty products on Consumer engagement

Type of beauty product	Coefficient	Increase in probability
Eyebrows	1.714	84.735%
Nails	1.530	82.201%
Lipstick	1.346	79.347%
Eyelashes	1.294	78.482%
None	1.203	76.906%
Eyeliner	1.121	75.417%
Eye-shadow	0.990	72.392%

We can conclude that although no significant influence is found between the categories of the variable *Main object* and the *CER*, some conclusions can be drawn based upon the tree as illustrated in Figure 6.2. Under the condition of the *Hair colour* being blonde or bald, including body parts, eye-wear, face, fashion, footwear, games, hair, lips, make-up, sports, or texts, lead to the highest level of *Consumer engagement*. Focalising on the make-up used by influencers, the following *Types of beauty products* have a significant positive effect on the *CER*: eyebrows, eyelashes, eyeliner, eye-shadow, lipstick, and nails. Also wearing no make-up increases the chances of being classified as one of the higher *Consumer engagement rate levels*.

6.4.3 Persons

The influence of people in Instagram posts is analysed in four different ways: by analysing the presence of people, their facial expression, if their full body is visible and by analysing the influence of the colour of someone's hair.

The OLR model found no significant influence of including one or multiple people in a photo on Instagram on the *CER*. Simple analysis gives that the mean of the *CER* is significantly higher when one or multiple people are present in the visual, compared to none. No conclusion can be drawn about the difference in effect of including one or more people. It can only be concluded that human presence has a positive effect.

Based upon simple analysis and the insignificant coefficients of the OLR model (Appendix: Table 8.2), nothing meaningful can be said about the influence of different *Facial expressions* on the *CER*. It can be concluded that visuals with a joyful *Facial expression* tend to be classified as medium-low *Consumer engagement level* (Figure 6.2). This presumption can only be made upon former splits. This negative influence is in line with research by Y. Li and Xie (2019), as they concluded that happy facial expressions lack relevance and therefore do not necessarily increase engagement.

Another aspect of *Person* in Instagram posts is whether their full body is visible. The coefficient for the variable *Full body visible?* taking on a value of Yes, is 0.429 (Table 6.4). The probability of falling into a higher category of the *CER* increases by 60.563% ($\frac{exp(0.429)}{(1+exp(0.429))} * 100\%$) if the beauty influencer shows his/her full body. This is in line with literature by Y. Li and Xie (2019) and Jaakonmäki et al. (2017), indicating that human presence in visuals has a positive influence on engagement.

The final variable of the dimension *Person* is *Hair colour.* The reference category for this variable is blonde. Having colourful hair as opposed to blonde hair has a highly significant positive influence on the *CER*. Hair colours that are considered colourful can range from red to green or from blue to orange. The probability of falling into a higher *Consumer engagement level* increases by 71.382% ($\frac{exp(0.914)}{(1+exp(0.914))} * 100\%$) if the person in the visual has colourful hair instead of blonde (Table 6.4). As some beauty influencers are bald in some of the posts, this is taken into the model as *Hair colour* taking on a value of none. The coefficient for the variable *Hair colour* being none is negative (Table 6.4). To be precise, the odds of falling into a higher category of *Consumer engagement* is estimated to be 0.806 times $((exp(-1.645))-1)$ lower when an influencer decides to go bald as opposed to blonde. No meaningful conclusions can be drawn about the effects of dark or mixed hair (different hair colours visible) as these coefficients appear to be insignificant.

Based upon the analysis the following conclusions can be made on the influence of including a person in the Instagram visual and their characteristics:

- Including at least one *Person* in the visual increases the *CER*.
- Visuals that include a joyful expression tend to fall into the medium-low *Consumer engagement level*. This conclusion can be drawn based upon (former splits of) the decision tree classification model (Figure 6.2).
- Showing of your full body in an Instagram post positively effects the chances of falling into a higher *Consumer engagement level*.
- In reference to having blonde hair, being bold has a negative influence on the odds of falling into a higher *CER level*, while having colourful hair has a positive influence.

6.4.4 Content

This dimension consists of the following variables: *Background setting, Professional visual?, Black and white visual?, Content type* and *Video*.

As the variable *Background setting* has no significant coefficients in the OLR model, simple analysis is used to find out if there is a difference in mean of the numerical *CER* between the different *Background settings*. This analysis states that the mean *CER* of the categories nature and outside are significantly higher than of the categories inside, studio and text. The average *CER* of visuals taken inside is also higher than those visuals taken in a studio. The positive relationship found between nature and *Consumer engagement* was expected based upon literature by Jaakonmäki et al. (2017).

As many researchers investigated the influence of the professionality of visuals on consumer satisfaction, the positive effect found was expected. The probability of belonging to one of the higher category levels of *Consumer engagement* increases by 61.301% ($\frac{exp(0.460)}{(1+exp(0.460))} * 100\%$) if the visual is considered to be *Professional* as opposed to amateurish (Table 6.4).

The expectation of *Black and white visuals* having a positive influence on *Consumer engagement* is derived from Zailskaite-Jakste et al. (2017). As the effect is insignificant (Appendix: Table 8.2), a one-way ANOVA test is used to see if the average *CER* differs for *Black and white visuals* as opposed to colour visuals. As (α) is even lower than 0.001, we can conclude there is a difference in the mean among these groups. The mean *CER* for *Black and white visuals* equals 7.742% while the average for colourful visuals is 4.159%. On average *Black and white visuals* tend to have a higher *CER* than colourful visuals. This result is in line with the 2017 study by Zailskaite-Jakste et al. and my expectations.

The third variable in the *Content* dimension is *Content type*. Based upon Table 6.4 and simple analysis, the categories educational and entertaining positively affect the *CER*, while posts classified as tutorials generally have lower interaction. This lower interaction was not expected based upon literature by Ashley and Tuten (2014). Posting educational content as opposed to advertorial content increases the probability of being classified as one of the higher *Consumer engagement* categories by 88.360% ($\frac{exp(2.027)}{(1+exp(2.027))} * 100\%$). The positive effect of entertaining content on the *Consumer engagement* was expected based upon literature by X. Liu et al. (2021) and Liu-Thompkins and Rogerson (2012). According to X. Liu et al., among others (Figure 3.1), a positive relation was anticipated between advertorial content or tutorials and *Consumer engagement*. Although the first effect can be found in simple analysis, the positive coefficient in Table 8.2 (Appendix) is insignificant. Regarding the influence of interactive visuals, tutorials tend to negatively influence the *CER* instead of the anticipated positive influence.

The analysis that *Videos* are found to have a lower average level of interaction than photographs has already been done in Paragraph 4.5.2.

Interpretation of the *Content* dimension can be summarised by the following conclusions:

˘ Influencers should shoot their visuals outside or in nature as a *Background setting* in order to achieve a higher *CER*.

˘ *Professional visuals* lead to more interaction than amateurish visuals.

˘ *Black and white visuals* have a higher mean *CER* than colourful visuals.

˘ The mean *CER* of entertaining content is higher than of advertorial or tutorial content. Tutorials tend to have the lowest mean *CER*. Educational content increases the probability of falling into one of the higher *Consumer engagement levels*.

˘ *Videos* tend to score lower on *CER* than photographs.

6.4.5 Textual elements

The final dimension is *Text* and includes variables in context of the Instagram post caption. Users can add captions to elaborate on what is in the post (via hashtags), add emoji or tag users. As this field was not (extensively) researched before, this analysis is of high academic relevance. As seen in Table 6.4, the effect of *Captions* is positive. *Captions* increase the probability of being classified to a higher *CER level* by 95.656% ($\frac{exp(3.092)}{(1+exp(3.092))} * 100\%$). Including a *Caption* has the largest positive impact of all variables. This information is meaningful for beauty influencers and companies that want to maximise their KPI's on social media interaction and *Consumer engagement*.

No significant conclusions can be drawn about the influence of *Emoji* based upon the insignificant coefficient in the OLR model (Appendix: Table 8.2). The results of the F-test are insignificant as well, indicating that there is no difference in mean among the posts that do and do not contain *Emoji*. As stated in the Conceptual framework (Chapter 3), a positive effect of emoji use on engagement was expected.

Another textual element is *Tagged accounts*. Tagging other users creates awareness by increasing the reach and impressions (KPI's). As the tagged accounts are mostly cosmetic companies, these posts are generally classified as advertorial and can be considered impersonal. Although the coefficient of the OLR model is insignificant (Appendix: Table 8.2), the one-way ANOVA test indicates that there is a significant difference in the mean *CER*. The average *CER* for posts with no tags is 5.392%, while the mean *CER* for posts with tags is 3.440%.

At last, the influence of *Hashtags* is analysed. Based upon the results in Table 6.4 it can be assumed that using *Hashtags* negatively affects the *Consumer engagement*. Adding *Hash-*

tags to a post lowers the odds of falling into a higher *Consumer engagement level* by 0.909 $((exp(-2.397))-1)$ times. Using *Hashtags* therefore has the largest negative effect on the *Consumer engagement.*

Based upon research by X. Liu et al. (2021), a positive effect was expected of (*Tagged accounts*) and using *Hashtags*, as these aspects make the Instagram post more interactive. Based upon literature in Chapter 2 interactive posts tend to capture the attention of users and therefore lead to more interaction and more *Consumer engagement.* Both these expectations are inaccurate, and the results of this investigation contradict previous literature by X. Liu et al. (2021).

This subsection can be summarised by the following bullet points:

˘ Beauty influencers should certainly add a *Caption* to their post to enhance interaction and therefore maximise the *CER* and KPI's.

˘ Including *Emoji* has no influence on the *Consumer engagement rate.*

˘ The average *CER* is lower for posts that have *Tagged accounts.*

˘ Using *Hashtags* has a significant negative effect on *Consumer engagement.*

7 Conclusion and Discussion

7.1 Research question and main results

The purpose of this thesis is answering the following research question:

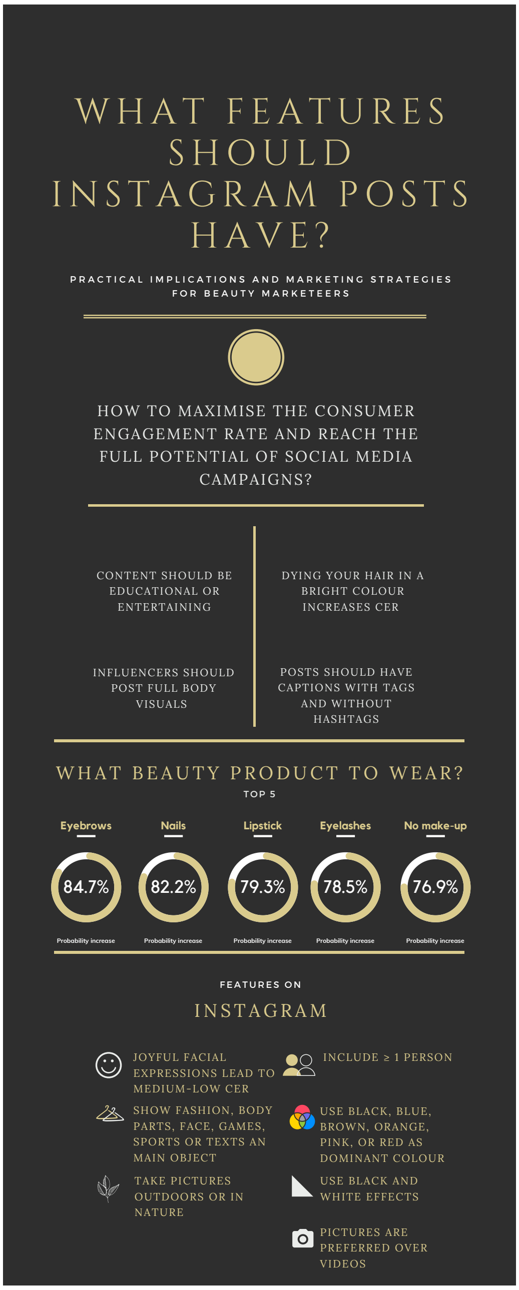
"Can data analytics be used to predict Consumer engagement and generate marketing strategies for beauty brands on Instagram?"

Data analytics can be used to predict *Consumer engagement* and machine learning methods can help generate marketing strategies. I will elaborate more on practical implications in Paragraph 7.1.1. This research is based upon an OLR model that predicts the *Consumer engagement level* of each Instagram post. The *Consumer engagement* can be classified as either low, medium-low, medium-high, or high. The model has a balanced accuracy of 66.527%. The most important post attributes in the class prediction of the decision tree model are *Hashtags*, *Hair colour* and *Main object.* The use of *Hashtags* has the largest negative impact on belonging to one of the higher *CER classes*, while using a *Caption* has the biggest positive impact.

7.1.1 Practical implications and marketing strategies for beauty marketers

The proposed marketing strategy for beauty influencers and cosmetic companies focuses on maximizing the probability of falling into a high *Consumer engagement class.* By striving to enhance *Consumer engagement* on Instagram, both beauty influencers and cosmetic brands can reach the full potential of their social media marketing strategy. Companies will have more customers successfully going through the customer journey and influencers will gain more online interaction and more collaborations with companies, as these prove to be successful.

The infographic on the following page (Figure 7.1) shows a short overview of the most important findings of this study in answering the research question. It also gives practical insights of what content features beauty influencers should include in their Instagram posts in order to have the highest probability of falling in a higher *Consumer engagement level.* On top are the eminently principal results that can be said with greatest certainty. While at the bottom are recommendations that can be made with less certainty or are subject to a limited number of assumptions.



WHAT FEATURES SHOULD INSTAGRAM POSTS HAVE?

PRACTICAL IMPLICATIONS AND MARKETING STRATEGIES FOR BEAUTY MARKETEERS

HOW TO MAXIMISE THE CONSUMER ENGAGEMENT RATE AND REACH THE FULL POTENTIAL OF SOCIAL MEDIA CAMPAIGNS?

CONTENT SHOULD BE EDUCATIONAL OR ENTERTAINING

DYING YOUR HAIR IN A BRIGHT COLOUR INCREASES CER

INFLUENCERS SHOULD POST FULL BODY VISUALS

POSTS SHOULD HAVE CAPTIONS WITH TAGS AND WITHOUT HASHTAGS

WHAT BEAUTY PRODUCT TO WEAR?

TOP 5

Eyebrows	Nails	Lipstick	Eyelashes	No make-up
84.7%	82.2%	79.3%	78.5%	76.9%
Probability increase	Probability increase	Probability increase	Probability increase	Probability increase

FEATURES ON

INSTAGRAM

☺ JOYFUL FACIAL EXPRESSIONS LEAD TO MEDIUM-LOW CER

INCLUDE ≥ 1 PERSON

SHOW FASHION, BODY PARTS, FACE, GAMES, SPORTS OR TEXTS AN MAIN OBJECT

USE BLACK, BLUE, BROWN, ORANGE, PINK, OR RED AS DOMINANT COLOUR

TAKE PICTURES OUTDOORS OR IN NATURE

USE BLACK AND WHITE EFFECTS

PICTURES ARE PREFERRED OVER VIDEOS

Figure 7.1: Infographic practical implications and marketing strategies beauty marketers

7.2 Limitations of the research

The research has several limitations. The primary limitation is the relatively small size of the data set. This is not favourable for machine learning techniques. Due to the fact that Instagram changed its API, it is harder to scrape posts and meta data from Instagram. As some variables can be qualified as self-reported data, this leads to possible misclassification.

7.3 Suggestions for future research

As stated in Paragraph 7.2, accuracy could be improved by utilising a more extensive data set. This data set could be obtained by looking at more influencers or adding variables, for instance using features obtained from text mining analysis on the Instagram post captions. Another suggestion would be to observe longitudinal effects, extending the time frame of the data collection enlarges the data set and could yield insights in trends over time.

Another suggestion is to investigate if it is possible to match influencers to cosmetic brands by performing canonical correlation analysis on a data set of posts by beauty influencers and a selection of cosmetic brands. If a certain influencer is correlated with a brand, their visuals have the same characteristics, and they could be a good match for social media marketing programs. As this would be time consuming and extend the research a lot, I leave this open for future research.

References

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index- insights from facebook, twitter and instagram. , *49*, 86–101. doi: 10.1016/j.jretconser.2019.03.012

Ashley, C., & Tuten, T. (2014). Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. , *32*(1), 15–27. doi: 10.1002/mar.20761

Asiri, S. (2021). *Building a convolutional neural network for image classification with tensorflow.* Retrieved from https://medium.com/@sidathasiri/building-a-convolutional-neural-network-for-image-classification-with-tensorflow-f1f2f56bd83b

Baker, S. A., & Walsh, M. J. (2018). 'good morning fitfam': Top posts, hashtags and gender display on instagram. , *20*(12), 4553–4570. doi: 10.1177/1461444818777514

Barger, V., Peltier, J. W., & Schultz, D. E. (2016). Social media and consumer engagement: a review and research agenda. , *10*(4), 268–287. doi: 10.1108/jrim-06-2016-0065

Bazi, S., Filieri, R., & Gorton, M. (2020). Customers' motivation to engage with luxury brands on social media. , *112*, 223–235. doi: 10.1016/j.jbusres.2020.02.032

Bellizzi, J. A., & Hite, R. E. (1992). Environmental color, consumer feelings, and purchase likelihood. , *9*(5), 347–363. doi: 10.1002/mar.4220090502

Bhuiya, S. (2020). *Disadvantages of cnn models.* Retrieved from https://iq.opengenus.org/disadvantages-of-cnn/

Brodersen, K. A., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In (pp. 3121–3124). doi: 10.1109/icpr.2010.764

Chaffey, D. (2021). *The content marketing matrix.* Retrieved from https://www.smartinsights.com/content-management/content-marketing-strategy/the-content-marketing-matrix-new-infographic/

Chan, Y. Y., & Mansori, S. (2016). Factor that influences consumers' brand loyalty towards cosmetic products. , *1*(1), 12–29. Retrieved from https://www.researchgate.net/profile/Shaheen-Mansori/publication/337907713_Factor_that_influences_consumers'_brand_loyalty_towards_cosmetic_products/links/5df22fe74585159aa476da58/Factor-that-influences-consumers-brand-loyalty-towards-cosmetic-products.pdf

Chen, H. (2017). College-aged young consumers' perceptions of social media marketing: The story of instagram. , *39*(1), 22–36. doi: 10.1080/10641734.2017.1372321

Cox, A. (2021, March 16). *6 social media kpis for social media marketing mastery (infographic + slideshare).* Retrieved from `https://www.brafton.com/blog/social-media/6-social-media-kpis-for-social-media-marketing-mastery/`

Cvetkovska, L. (2021, October 2). *45 absolutely astonishing beauty industry statistics for 2021.* Retrieved from `https://loudcloudhealth.com/resources/beauty-industry-statistics/`

de Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. , *26*(2), 83–91. doi: 10.1016/j.intmar.2012.01.003

Evensson, F., & Jansson, E. (2020, June). *"this is my favorite beauty product... do you like it too?"* (No. Thesis). Retrieved from `https://www.diva-portal.org/smash/get/diva2:1437981/FULLTEXT01.pdf`

Foley, M. (2020a). *4.1 linear regression model | my data science notes.* Retrieved from `https://bookdown.org/mpfoley1973/data-sci/ordinal-logistic-regression.html`

Foley, M. (2020b). *5.3 ordinal logistic regression | my data science notes.* Retrieved from `https://bookdown.org/mpfoley1973/data-sci/ordinal-logistic-regression.html`

Forbes, K. (2016). Examining the beauty industry's use of social influencers. , *7*(2), 78–87. Retrieved from `https://www.elon.edu/u/academics/communications/journal/wp-content/uploads/sites/153/2017/06/08_Kristen_Forbes.pdf`

García, V., Mollineda, R. A., & Sánchez, J. S. (2009). Index of balanced accuracy: A performance measure for skewed class distributions. , 441–448. doi: 10.1007/978-3-642-02172-5_57

Goldstein. (2020). *Global cosmetics industry report: By country, by product type (fragrances, color cosmetics, hygiene, hair skin care products), by pricing (low and medium, premium) by distribution chanel (online, offline), by geography with covid-19 impact | forecast period 2017-2030.* Retrieved from `https://www.goldsteinresearch.com/report/cosmetics-industry-beauty-market-size-share-trends-demand`

Google Cloud. (2021). *Detect labels | cloud vision api | .* Retrieved from `https://cloud.google.com/vision/docs/labels`

Gregory, S. (2020, November 4). *19 digital marketing metrics for measuring success in 2020.* Retrieved from `https://freshsparks.com/digital-marketing-success/`

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: studying visual social media, from selfies and gifs to memes and emoji. , *2*(1), 47–62. doi: 10.1080/

22041451.2016.1155332

Hu, Y., Manikonda, L., & Manikonda, S. (Eds.). (2014). *What we instagram: A first analysis of instagram photo content and user types* (Vol. 8) (No. 1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14578

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist.* SAGE Publications. doi: 10.4135/9780857028075.d49

Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013). The impact of user interactions in social media on brand awareness and purchase intention: the case of mini on facebook. , *22*(5/6), 342–351. doi: 10.1108/jpbm-05-2013-0299

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. , *22*(4), 679–688. doi: 10.1016/j.ijforecast.2006.03.001

@iambeckyg. (2021, March 11). *Lasnenas.* Retrieved from https://www.instagram.com/p/CMQkcqIJx1_/

Jaakonmäki, R., Müller, O., & vom Brocke, J. (2017). The impact of content, context, and creator on user engagement in social media marketing. Hawaii International Conference on System Sciences. doi: 10.24251/hicss.2017.136

Jahn, B., & Kunz, W. (2012). How to transform consumers into fans of your brand. , *23*(3), 344–361. doi: 10.1108/09564231211248444

Knowledge, H. W. (2019, December 4). *How influencers are making over beauty marketing.* Retrieved from https://www.forbes.com/sites/hbsworkingknowledge/2019/12/13/how-influencers-are-making-over-beauty-marketing/?sh=5367c4301203

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer Publishing.

Kumar, S., Massie, C., & Dumonceaux, M. D. (2006). Comparative innovative business strategies of major players in cosmetic industry. , *106*(3), 285–306. doi: 10.1108/02635570610653461

Kumar, V., & Mirchandani, R. (2013). Increasing the roi of social media marketing. , *41*(3), 17–23. doi: 10.1109/emr.2013.6596535

Lashbrook, J. (2019, November 22). *Beauty industry increasingly turning to digital advertising, spurning magazines and tv.* Retrieved from https://www.marketingcharts.com/industries/cpg-and-fmcg-111101

Li, L. (2019). *Classification and regression analysis with decision trees.* Retrieved from https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

Li, Y., & Xie, Y. (2019). Is a picture worth a thousand words? an empirical study of image content and social media engagement. , *57*(1), 1–19. doi: 10.1177/0022243719881113

Liu, H., Jayawardhena, C., Shukla, P., Osburg, V., & Yoganathan, V. (2021). Old wine in new bottles? revisiting electronic word-of-mouth (ewom).. Retrieved from https://www.journals.elsevier.com/journal-of-business-research/call-for-papers/old-wine-in-new-bottles-revisiting-electronic-word-of-mouth

Liu, X., Shin, H., & Burns, A. C. (2021). Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. , *125*, 815–826. doi: 10.1016/j.jbusres.2019.04.042

Liu-Thompkins, Y., & Rogerson, M. (2012). Rising to stardom: An empirical investigation of the diffusion of user-generated content. , *26*(2), 71–82. doi: 10.1016/j.intmar.2011.11.003

Ma, L., & Sun, B. (2020). Machine learning and ai in marketing – connecting computing power to human insights. , *37*(3), 481–504. doi: 10.1016/j.ijresmar.2020.04.005

Mantovani, R. G., Horváth, T., Cerri, R., Barbon Junior, S., Vanschoren, J., & de Carvalho, A. C. P. d. L. F. (2019). *An empirical study on hyperparameter tuning of decision trees.*

Naumanen, E., & Pelkonen, M. (2017). *Celebrities of instagram - what type of content influences followers' purchase intentions and engagement rate?* Retrieved from https://aaltodoc.aalto.fi/bitstream/handle/123456789/27277/master_Naumanen_Emma_2017.pdf?sequence=1&isAllowed=y

Osman, M. (2018, July 2). *How has social media changed the beauty industry?* Retrieved from https://mayranosman.atavist.com/socialmediaandmakeupbrands#chapter-4131919

R. (2019, October 25). *How social media changes the beauty industry landscape.* Retrieved from https://reputationdefender.medium.com/how-social-media-changes-the-beauty-industry-landscape-a3f9b7bbecc1

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. , *41*(1), 77–93. doi: 10.1023/b:amai.0000018580.96245.c6

Rash, W. (2019, September 30). *Vision ai | derive image insights via ml | cloud vision api.* Retrieved from https://cloud.google.com/vision/

Rodriguez, D. (2021, February 20). *How to beat the 2021 instagram algorithm.* Retrieved from https://dani-the-explorer.com/2019-instagram-algorithm/

Sadeque, F., & Bethard, S. (2019). *Predicting engagement in online socialnetworks: Challenges and opportunities.* Retrieved from https://www.researchgate.net/publication/334457576_Predicting_engagement_in_online_social_networks_Challenges_and_opportunities

Sarstedt, M., & Mooi, E. (2019). *A concise guide to market research: The process, data, and methods using ibm spss statistics (springer texts in business and economics)* (Softcover reprint of the original 3rd ed. 2019 ed.). Springer. doi: 10.1007/978-3-642-12541-6

Simonyan, K., & Zimmerman, A. (2015). *Very deep convolutional networks for large-scale image recognition.* ICLR. Retrieved from https://arxiv.org/pdf/1409.1556.pdf

Singh, J., Wheeler, J., Fong, N., & Chaudhary, S. (2019, September). *A comparision of public cloud computer vision services.* Retrieved from osf.io/9t5qf

Smith, K. A. (2015). Neural networks for prediction and classification. , 865–869. doi: 10.4018/978-1-59140-557-3.ch164

Statista. (2021a, February 8). *Daily social media usage worldwide 2012-2020.* Retrieved from https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista. (2021b, September 2). *Global social networks ranked by number of users 2021.* Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Statista. (2021c, February 10). *Instagram: distribution of global audiences 2021, by age group.* Retrieved from https://www.statista.com/statistics/325587/instagram-global-age-group/

Therneau, T., & Atkinson, B. (2019). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rpart (R package version 4.1-15)

Vock, M., Dolen, W. V., & Ruyter, K. D. (2013). Understanding willingness to pay for social network sites. , *16*(3), 311–325. doi: 10.1177/1094670512472729

Yakin, V., & Eru, O. (2017). An application to determine the efficacy of emoji use on social marketing ads. , *3*(1), 230–230. doi: 10.24289/ijsser.270652

Zailskaite-Jakste, L., Ostreika, A., Jakstas, A., Staneviciene, E., & Damasevicius, R. (2017). Brand communication in social media: The use of image colours in popular posts. In (pp. 1373–1378). IEEE. doi: 10.23919/mipro.2017.7973636

Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2017). How much is an image worth? airbnb property demand estimation leveraging large scale image analytics. , 1–32. doi: 10.2139/ssrn.2976021

8 Appendix

Table 8.1: Independent variables summary table

Dimension	Variable name	Variable type	Values
Colour	*Dominant colour*	Categorical variable	Black, Blue, Brown, Green, Grey, None, Orange, Pink, Purple, Red, White, and Yellow
Objects	*Beauty product visible*	Dummy variable	Yes or No
	Type of beauty product?	Categorical variable	Blush, Eyebrows, Eyelashes, Eyeliner, Eyeshadow, Foundation, Highlighter, Lipstick, Nails, and None
	Main object	Categorical variable	Accessory, Art, Body parts, Colour, Dog, Eye-wear, Face, Fashion, Food, Footwear, Furniture/Building, Hair, Lingerie, Make-up, Nails, Nature, None, Plant, Text, and Vehicle

Person	*Person*	Categorical variable	0 = No, 1 = 1 person included, 2 = multiple people included
	Facial expression	Categorical variable	Anger, Joy, None, and Surprise
	Full body visible?	Dummy variable	Yes or No
	Hair colour	Categorical variable	Blonde, Colourful, Dark-haired, Mixed, and None
Content	*Background setting*	Categorical variable	Inside, Nature, Outside, Studio, and Text
	Professional visual?	Dummy variable	Yes or No
	Black and white visual?	Dummy variable	Yes or No
	Content type	Categorical variable	Advertorial, Educational, Entertaining, Tutorial
	Video	Dummy variable	Yes or No
Text	*Caption*	Dummy variable	Yes or No
	Emoji	Dummy variable	Yes or No
	Tagged accounts	Dummy variable	Yes or No
	Hashtags	Dummy variable	Yes or No

Table 8.2: Coefficients of the full ordinal logistic regression model

	Value	Std. Error	*t value*	*p value*
‘Dominant colour‘Blue	0.222	0.314	0.706	0.480
‘Dominant colour‘Brown	−0.394	0.375	-1.049	0.294
‘Dominant colour‘Green	-0.226	0.367	-0.616	0.538
‘Dominant colour‘Grey	0.228	0.318	0.718	0.473
‘Dominant colour‘Orange	0.058	0.333	0.173	0.863
‘Dominant colour‘Pink	0.094	0.314	0.298	0.766
‘Dominant colour‘Purple	-0.092	0.379	-0.244	0.807
‘Dominant colour‘Red	-0.156	0.358	-0.436	0.663
‘Dominant colour‘White	-0.159	0.393	-0.404	0.686
‘Dominant colour‘Yellow	0.188	0.524	0.358	0.720
‘Main object‘Art	2.229	1.488	1.498	0.134
‘Main object‘Body parts	0.280	0.858	0.326	0.744
‘Main object‘Colour	-0.200	0.980	-0.204	0.838
‘Main object‘Dog	1.469	0.962	1.527	0.127
‘Main object‘Eyewear	-0.641	0.905	-0.708	0.479
‘Main object‘Face	-0.448	0.835	-0.537	0.592
‘Main object‘Fashion	-0.084	0.843	-0.100	0.921
‘Main object‘Food	1.109	1.307	0.848	0.396
‘Main object‘Footwear	0.210	0.883	0.238	0.812
‘Main object‘Furniture/Building	0.884	0.884	1.000	0.317
‘Main object‘Game	-0.260	1.484	-0.175	0.861
‘Main object‘Hair	-0.096	0.843	-0.114	0.909
‘Main object‘Lingerie	2.127	1.406	1.512	0.130
‘Main object‘Lips	-0.139	0.898	-0.155	0.877
‘Main object‘Make-up	-0.445	0.885	-0.503	0.615
‘Main object‘Nails	−0.483	1.418	-0.341	0.733

	Value	Std. Error	t value	p value
‘Main object‘Nature	0.632	0.882	0.717	0.473
‘Main object‘Plant	0.930	0.921	1.010	0.313
‘Main object‘Sports	0.565	1.213	0.466	0.641
‘Main object‘Text	0.075	1.080	0.070	0.944
‘Main object‘Vehicle	0.273	0.919	0.297	0.766
‘Facial expression‘Joy	0.050	1.182	0.043	0.966
‘Facial expression‘None	0.640	1.171	0.547	0.585
‘Facial expression‘Sad	−0.704	1.555	-0.453	0.651
‘Facial expression‘Suprise	1.145	1.257	0.911	0.362
VideoYes	-0.148	0.269	-0.551	0.582
‘Hair colour‘Colourful	0.914	0.322	2.836	0.005 ∗ ∗
‘Hair colour‘Dark-haired	0.085	0.235	0.363	0.717
‘Hair colour‘Mixed	0.204	0.468	0.437	0.662
‘Hair colour‘None	−1.645	0.452	−3.641	0.0003 ∗ ∗∗
Person1	−0.301	0.470	-0.641	0.521
Person2	−0.520	0.512	−1.014	0.311
‘Background setting‘Nature	0.149	0.341	0.438	0.661
‘Background setting‘Outside	0.250	0.247	1.012	0.311
‘Background setting‘Studio	0.379	0.264	1.433	0.152
‘Background setting‘Text	−0.943	0.983	−0.960	0.337
‘Full body visible?‘Yes	0.429	0.218	1.969	0.049∗
‘Beauty product visible?‘Yes	0.257	0.351	0.734	0.463
‘Type of beauty product‘Eyebrows	1.714	0.763	2.247	0.025∗
‘Type of beauty product‘Eyelashes	1.294	0.489	2.645	0.008 ∗ ∗
‘Type of beauty product‘Eyeliner	1.121	0.565	1.983	0.047∗
‘Type of beauty product‘Eye-shadow	0.990	0.444	2.227	0.026∗
‘Type of beauty product‘Foundation	0.478	1.221	0.391	0.695
‘Type of beauty product‘Highlighter	0.870	0.476	1.829	0.067

	Value	Std. Error	t value	p value
‘Type of beauty product‘Lipstick	1.346	0.441	3.054	0.002 * *
‘Type of beauty product‘Nails	1.530	0.711	2.153	0.031*
‘Type of beauty product‘None	1.203	0.444	2.708	0.007 * *
‘Professional visual?‘Yes	0.460	0.213	2.157	0.031*
‘Black and white visual?‘Yes	0.663	0.869	0.763	0.446
‘Content type‘Educational	2.027	0.700	2.893	0.004 * *
‘Content type‘Entertaining	0.398	0.238	1.674	0.094
‘Content type‘Tutorial	-1.064	0.489	-2.175	0.030*
CaptionYes	3.092	0.858	3.604	0.0003 * **
‘Emoji’s‘Yes	−0.017	0.266	−0.064	0.949
‘Tagged accounts‘Yes	-0.150	0.191	-0.782	0.434
HashtagsYes	−2.397	0.229	−10.468	0 * **
Low|Medium-low	2.755	1.857	1.483	0.138
Medium-low|Medium-high	4.472	1.862	2.402	0.016 * *
Medium-high|High	6.059	1.870	3.241	0.001 * *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1