Visual Analysis and Personalisation in Advertisement

Daniel Christopher Jakobsen

Supervisor: Assoc. Prof. Dr. Mehdi Elahi
Co-supervisor: Chief Data Scientist Dr. Igor Pipkin

Master's Thesis
Department of Information Science and Media Studies
University of Bergen

December 1, 2022

Acknowledgements

First, I want to thank my supervisors, Prof. Dr. Mehdi Elahi and Dr. Igor Pipkin, for their amazing support and motivation throughout this thesis. Their expertise has been invaluable and when the work, and life in general, have been challenging they have shared their own experiences to light up the mood and encourage me to continue. I could not have asked for any better supervisors. I must also thank Amedia for this incredible opportunity to work with real data. I also want to thank my fellow students, David and Evy-ann for their time and help during the thesis. I want to thank my family for their support and motivation, be it helping with the thesis or just bringing food. Lastly, I want to thank my girlfriend Emma, for her endless love and support, motivation and encouragement and of course all the humour and the ability to calm me down when needed; I could not have done this without you.

This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through The Centres for Research-based Innovation scheme, project number 309339.

Daniel Christopher Jakobsen
Bergen, Date

Abstract

Advertisement is one of the primary sources of revenue for companies. Advertisement can result in increased sales for the companies by promoting their products or by hosting the advertisement of the other companies. While there have been many studies investigating the effectiveness of the advertisement in various domains, there are still challenges to be addressed. Notable challenges are overabundance, privacy, spamming, or irrelevant placement of advertisement. The results of such challenges can be poor quality of advertisement irrelevant to the audience interest or irrelevant in the context of the content in which it has been shown. A colourful and shiny image in an advertisement, promoting a luxury product that is placed together with a news article discussing the growing poverty in certain regions of the world can be an example. This can lead to dissatisfaction of the audience when experiencing a particular irrelevant advertisement.

Personalisation techniques can be useful in addressing such challenges. For example it can be used to match the advertisement shown in news outlets to the article. Such techniques can analyse the content of the news articles and find right advertisement that is better suited to the article. However, this requires rich data of displayed advertisement which are not always available. That can be one of the reasons why viewing an irrelevant advertisement is still part of our everyday experience.

In this thesis, I have explored the potential of using automatic visual analysis to obtain better representation of the advertisement and to improve the relevance of advertisement to the audience. For such analysis, a number of visual features has been extracted from the images of the advertisements and analysed to better understand their correlations with audience behaviour online when interacting with advertisement campaigns. I have received the data set from Amedia, one of the largest media companies in Norway. The data set includes campaigns and audience behaviour, and the images associated with the campaigns. Four analyses have been conducted in the thesis including, exploratory analysis and correlation analysis. Machine learning models have been built (using *CatBoost*) based on the audience behaviours and visual features extracted from advertisement. In addition a method called *SHAP (SHapley Additive exPlanations)* has been used to find explanation of model prediction. The contribution of this thesis includes preprocessing and exploratory analyses of the industry data, a novel data set containing the visual features extracted from the images of the advertisement, a Python prototype, and the results of analyses based on this prototype.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Advertisement is one of the main approaches for companies to earn revenue. While advertisement companies can follow the a general approach of one-size-fits-all, personalisation of advertisement can beneficial and potentially improve the satisfaction of consumers and increase the revenue of industry [30].

Personalisation (or contextualisation) in advertisement is not new and has come a long way already. However, there are still challenges that needs to be addressed, notably, overabundance, privacy and irrelevance of advertisement on online news outlets. This can get further challenging with policies of the big platforms. As an example, Apple has been removing ways that companies could personalise advertising on iPhone — such as introducing the option to turn off cookies, cross-site tracking and e-mail protection from third party apps [1]

Obtaining more information from images of the advertisement can help to better match the advertisement with the news content. Visual analysis can be a solution in that direction. In such analyses, image of advertisement are analysed and a set of visual features are extracted. Example of such features are colourfulness, sharpness and saturation and prior works have reported their effectiveness on other domains to deal with other challenges (e.g., to tackle cold start in movie personalisation and recommendation [15]).

This thesis addresses the challenge of personalisation in advertisement by proposing a method which can help news sharing platforms and industry to improve their advertisement process. While the thesis addresses the needs of such industry players, it also addresses the needs of the audience when reading news online. This can hence provide a more healthy and fair environment and improve the satisfaction of the audience and industry.

The method proposed in this thesis is a combination of extracting visual features from the images of advertisement and using it with audience behavioural data to improve personalisation in advertisement. Several experiments were conducted on a data set provided by Amedia, one of the largest media companies in Norway [1]. These experiments included a preliminary analysis of the data, a correlation analysis of the visual features data combined with the click through rate (CTR) of the campaign associated

[1] https://www.amedia.no/

with image from which the visual features were extracted, a machine learning analysis of both of these data sets using the machine learning model called CatBoost and an analysis aimed at explaining how the machine learning model arrived at the predictions it did with the use of SHAP, a game theory method of explaining how much impact an "player" has on the "game". The prototype created in this thesis is the machine learning model trained on the visual features data set combined with audience behaviour, CTR.

1.2 Research Questions

The research questions this thesis attempt to answer is the following:

RQ 1: What is the level of correlation between visual features extracted from images of advertisement and metrics indicating the behaviour of the media audience?

RQ 2: How do different media audiences differ in terms of their reaction to the advertisement exposed to them online?

- **RQ 2.1:** What are the differences of visual features importance when predicting metrics such as Click Through Rate (CTR)?
- **RQ 2.2:** What can be learned from models that can provide explanation for the prediction of metrics such as Click Through Rate (CTR)?

The first research question is addressed by the correlation analysis looking at the different correlations between visual features and CTR, audience behaviour in the form of clicking as a reaction to the advertisement. This continues to the second research question of how media audiences differ in terms of their reaction. The second research question and its subsections are addressed by further exploratory analysis, feature analysis and SHAP analysis.

1.3 Contributions

The main contributions of this thesis are:

- A novel analysis method of the combination of visual features with audience behavioural data.
- A prototype[2] for the feature extraction using the method proposed.
- A novel data set of visual features extracted from multiple different advertisement campaign together with the click through rate for each of them.
- An extensive analysis of the method using machine learning.
- An extensive analysis of the machine learning model in how it made predicted with the use of SHAP.

[2] https://github.com/Dennydc007/Visual_feature_extraction_prototype

1.4 Thesis Outline

- *Chapter 2 Background:* Chapter 2 is an overview of the background related to this thesis. Section 2.1 describes previous works using visual features to improve recommender systems. Section 2.2 reviews the importance of advertising, how it is changing and what challenges it brings. Section 2.3 details how personalisation systems work and what methods are used.
- *Chapter 3 Methodology and Methods:* Chapter 3 describes what methodology and methods have been used in this thesis and details the data set. Section 3.1 and 3.2 describe the methodologies, design science and experiment. Section 3.3 is an overview of the data set provided by Amedia. Section 3.4 describes the machine learning model used (CatBoost) and the tools used to process the data and analyse the machine learning model (SHAP).
- *Chapter 4 Experiments and Results:* Chapter 4 describes and presents the results of the different experiments performed in this thesis. Section 4.1 presents the preliminary analysis and initial findings of the data. Section 4.2 presents the correlation analysis of the visual features data and the results of the analysis. Section 4.3 is the analysis of the audience and campaign data, and section 4.4 is the analysis of the visual features data.
- *Chapter 5 Discussion and Conclusion:* Chapter 5 presents the discussion and conclusion of this thesis. Section 5.1 is the discussion of the results of the experiments and section 5.2 presents the limitations of this thesis. Section 5.3 presents the contribution of the thesis and section 5.4 presents what future work can be done.

Chapter 2

Background

In this chapter previous works related to this thesis will be presented. Section 2.1 describes a background regarding how visual features can be used within personalisation and some examples of extraction methods possible. Section 2.2 presents the importance of advertisement and how personalisation is a powerful tool for advertising. This section also introduces contextual advertising which might become an important approach for personalisation if more companies follow Apples lead in privacy protection. Lastly, section 2.3 presents personalisation through recommendation systems.

2.1 Visual Features

The ability to automatically extract and process visual features is made possible by computer vision technology and machine learning. By analysing media content such as images, videos or movies we can extract visual features such as colour, brightness, sharpness, contrast and more [8, 6, 12, 15, 32, 17].

Another method of visual feature extraction possible is to train a machine learning algorithm to recognise different visual elements, for example in [17] the machine learning model was trained to recognise different types of scenes in movies such as landscapes, face focused conversation or title screens and end screens.

Visual tagging proposed by [15] is another method of visual feature extraction which is based on user tagging, used for categorisation. Tagging means to give an item a keyword which describes the item in one way, with multiple keywords an item can be described greatly to the benefit of information systems. Visual tagging is done automatically with machine learning models, doing the work users could have done. These visual tags can then be used for personalised recommendation [15]. Visual features have a huge potential when it comes to personalisation and recommendation. Multiple studies have shown the use of visual features extracted from movies to improve movie or video recommendation systems [17, 6, 32, 8, 20, 15].

Multiple studies have done visual feature extraction from key frames of movies or movie trailers as illustrated in figure 2.1. They have analysed each scene of the movie and taken the most representative frame of the scene to extract the visual features from. This is very similar to just extracting visual features from a still image, difference being an extra step of scanning through a movie or movie trailer[3, 17]. Similar methods used in [3, 17] can be used for different purposes such as extracting visual features from

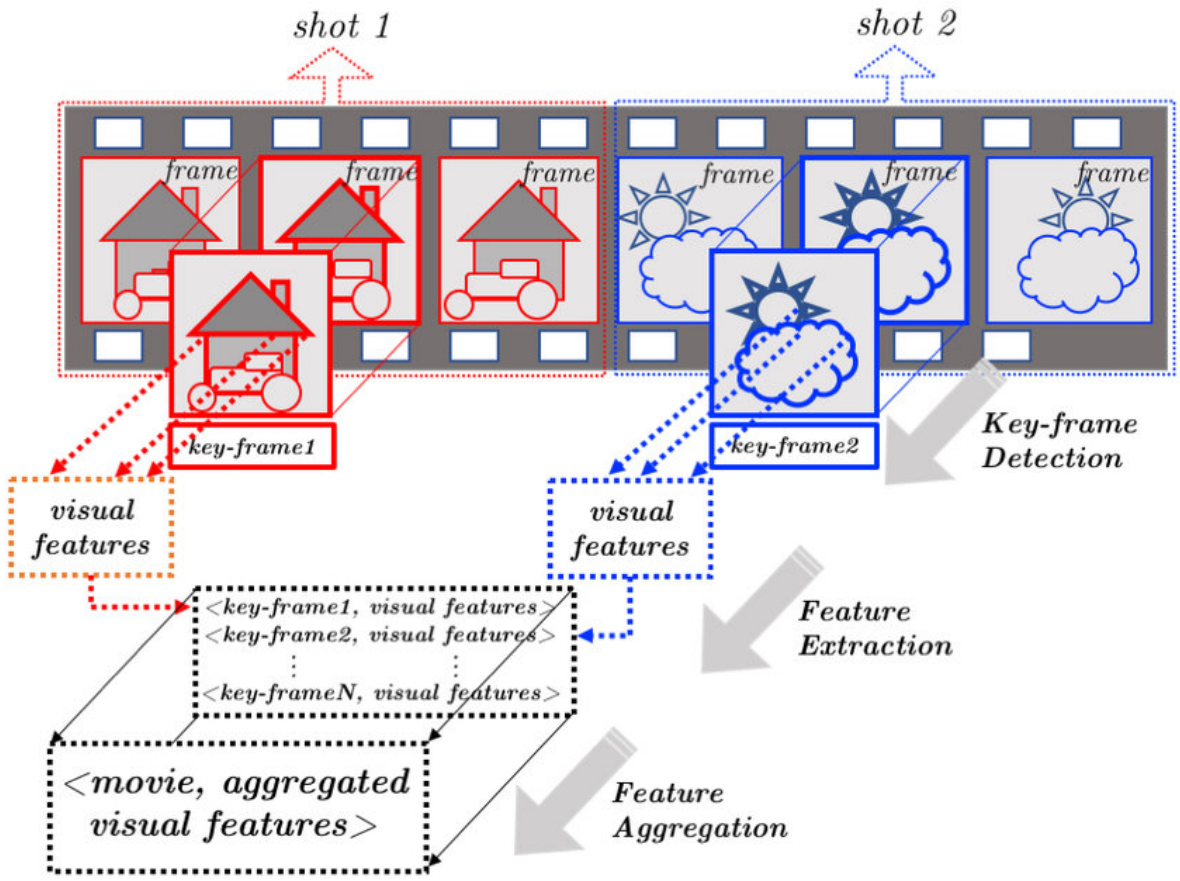

Figure 2.1: The method of feature extraction and aggregation from movie trailers as presented by Moghaddam [3]

advertisement images, skipping the step of scanning for key frames.

Colour has considerable impact on us as humans and can truly affect how we feel about a product in a setting. There has been several research studies focused on various ways of characterising colours [38]. Not only is colour often associated with a brand name through how companies make themselves recognisable with colour, but also how they use colour in advertisement to make their audience feel something and to influence purchasing behaviour [11, 37]. In figure 2.2 the different colours and the associated feeling with those colours are illustrated.

2.2 Advertisement

Advertisement is an essential part of any business providing a product or a service. Without any promotion of one's business, product and/or service it is unlikely that it will be successful. Before Internet and TV, advertisement was made to reach as many as possible, meaning that they were made with a general, one-size-fits-all approach [30].

Personalisation has increased the effectiveness of advertising with the progression of technology. With personalisation advertisement can be better suited the interest of the consumer, which increases the satisfaction of the consumer and increases revenue of the business [30]. Personalisation seems to be desired over general and possibly

Figure 2.2: Chart illustrating colours and their corresponding feeling associated [37]

irrelevant advertisement according to a study made by Pauzer. In this study 71% of the respondents preferred personalised advertisement. The advantages presented by Pauzers study showed that 46% of the respondents reported reduced irrelevant advertisement, 25% reported a way to discover new products and 19% reported that it made online searching and shopping faster and easier [29].

With personalisation comes contextual advertisement, where the aim is to better understand the context of the content a consumer is consuming and to give the consumer a relevant advertisement to that content. In contextual advertising one assumes that the audience interest in content is similar to their purchase interest, so analysing the context of the content to better understand what advertisement would be most effective [40].

Contextual advertising might become more important now that the giant electronic company Apple has shifted to a focus on protecting the privacy of their customer. By giving the customers the option to turn of tracing technology such as cookies, cross-site tracing, and protecting e-mail from third-party apps [1, 33]. This means that getting an overview of what advertisement works best with the context of the content could be a solution, there is still the possibility to get audience data through them signing up for accounts which can store the information they are willing to share [9, 40].

One problem that can occur as a limitation of technology is persons who have more challenges than just learning new technology and this is true for touch technology [2, 28]. In this thesis it is shown that there is a possibility of unintentional clicking by older age groups and it is important to note that there have been other studies which shows older people struggling with the use of touch technology.

2.3 Personalisation

Personalisation is widely used to help us process the huge amount of information out there [16, 15, 17, 8, 12, 20, 6, 3]. Since The research on personalisation is closely connected with the research on recommendation. Hence, I briefly review the literature

on Recommender systems.

Recommender systems can be classified into at least three distinct methods, collaborative filtering models, content-based filtering models and hybrid models [31, 5]. Collaborative filtering models learn what users like through their interaction with the system. Then it clusters together users based on what they like e.g., users that like the same items. The model makes recommendations based on how similar a user is to these clusters of users. Typically, you can observe this on online shopping sites when you get the recommendations such as "other users also bought these items" when you are either buying or browsing. The advantage of collaborative filtering is that it does not need any metadata concerning the items, only information regarding users. The necessary information about users is a disadvantage especially if the system is completely new and does not have any information about any users. Not having any information means that it cannot make any personalised recommendations to users. If the system already has the necessary information about its user base and a new user is introduced to the system, we still will encounter the user cold start problem since there is no information about the user until they have interacted with the system a bit. Before that the system can still recommend popular items [16, 17, 25, 8].

Content-based filtering models exploit the metadata about items to recommend items to users. For example, a movie recommender system would exploit the metadata about movies such as genre, actors in the movie or the director of the movie to recommend to users. Content-based filtering models still need information about its users but do not need it to compare the users such as collaborative filtering models do. Instead, it is important that there is metadata about the items in the system or else it would fail to recommend the item to any user [14]. Content-based filtering compares the similarity of items to items a user has shown interest in by interacting with them either directly such as rating a movie or indirectly by watching it. While a content-based model still needs information about the user, about what that user is interested in, it does not need information of a user base such as the collaborative filtering model does [16]. Hybrid models combine both collaborative filtering and content-based filtering to exploit the advantages of both while levering the disadvantages [5]

Chapter 3

Research Methodology

In this section, the methodologies, and methods used in this thesis will be introduced. In general, methodologies used in this thesis are consisted of design-science and experiment methods which will be discussed. Section 3.3 presents the data set provided by Amedia and section 3.4 details the tools used to conduct the analysis.

3.1 Design Science

The goal of a design-science approach to a problem is to produce an artefact (human made objects) which is useful and can solve the problem for which it is designed to solve. While there is much debate about the differences between classical science and design-science because their similarities, the difference being classical science has focus on phenomena from the real world and design-science focus is on the artefact produced from the research [21]. Four different forms have been proposed in which a product of design-science can be by Henver et al. [21], but for this thesis the form of instantiation is the most appropriate. All four forms are described by Hevner et al. [21]. The descriptions have been taken from Johanson and Williamson (2018)[21]:

> **Construct:** "A construct is a conceptual object that researchers create as a means of describing and representing some type of phenomena in the world, such as classes of things (e.g., businesses), subclasses of things (e.g., small businesses), components of things (e.g., employees of a business), properties of things (e.g., the level of profitability of businesses), states of things (e.g., a business is either liquid or bankrupt), events that occur to things (e.g., a business makes a sale), and processes that things undergo (e.g., a business receives, fills, and despatches an order)"[21].

> **Models:** "A model is a conceptual object that comprises constructs and associations among these constructs as a way to describe and represent some subset of real-world phenomena. Depending on the nature of the constructs linked by an association, the association can have different types of meanings." [21].

> **Methods:** "A method is a set of actions (the actions are often ordered) that is used to achieve some outcome (a product or service)"[21].

> **Instantiation:** "An instantiation is a hardware/software system that researchers produce using some method to implement a construct or model"[21].

Hevner et al. (2004)[21] also proposed seven guidelines for design-science which is supposed to assist in understanding the requirements for effective design-science research.

The first guideline; "creation of an innovative and purposeful artefact", meaning that the result of design-science research must be a practical, workable artefact in one of the four forms described above [21]. The prototype and method made during this thesis combined with machine learning analysing audience data and advertisement data and for a deeper understanding of audience behaviour goes under Henver's [21] guideline one.

The second guideline; "ensure that the artefact is purposeful", is concerned with the relevance and ability to resolve a problem to some stakeholder community. The stakeholders for this thesis would be the advertising company and the companies advertising through them and the audience who gets exposed to the advertisement. The analysis method proposed in this thesis aims to improve advertisement and recommendation algorithms fairly for all stakeholders, both industry and audience.

The third guideline; "rigorously evaluate the artefact", evaluating the artefacts effectiveness and efficiency with rigorous methods to mitigate threats to validity. The method proposed in this thesis is not a fully developed method and therefore has not been rigorously evaluated, the machine learning models have been evaluated in the sense of validating the predictions performed by it.

The fourth guideline; "produce an artefact that makes a research contribution", is to ensure how important it is that the artefact contributes to knowledge and must be novel for this. To my knowledge there are no research papers which have explored a method specifically like the one presented in this thesis.

The fifth guideline; "follow rigorous construction methods", replication is possible for other researchers, it is important that the construction methods are rigorous meaning that they are sufficiently well specified and formalised. Because of the nature of machine learning, it will not be possible to replicate an exact model of the one used in this thesis, but how to go forward with the method has been documented in this thesis.

The sixth guideline; "show the artefact is the outcome of a search process", is the description of the search process from the initial state to the end state of the artefact design. This search process concerns the actions, resources, natural or social laws to get to the end state. The method proposed in this thesis is a result of a search process and this is explained in the different experiments in section 4.

The seventh guideline; "clearly communicate the research process and outcome", so any stakeholder understands how the artefact is to be constructed and used effectively and the resources needed in construction and use of the artefact [21]. The communication of this method and machine learning model have been to the best of abilities but is of course a matter of the readers judgement if it has been clearly and effectively communicated.

Evaluation in design-science does not mean only evaluating the artefact produced by the research, but also the research process, problem specification contribution to knowledge and the solution that was derived. Evaluating all of these aspects of design

science ensures the quality of the research. To evaluate the artefact, or in the case of this thesis the recommender system, an experiment will be conducted to test the artefact's usefulness in a situation more suited for the artefact.

As noted, the method proposed in this thesis is highly exploratory within a new way of analysing available data and to test for correlations between the data and audience behaviour. The thesis fits well in for the design science paradigm because it is an artefact (method and machine learning model) produced to analyse data in a new way

3.2 Experiment

When using an experiment as evaluation of an *artefact*, produced by the design-science research, it will be focused on field experiment as being an efficient way to understand if the artefact does solve the proposed problem it is designed to solve. While that problem is not yet proposed for this thesis, it is still the end goal of the thesis. Conducting a field experiment compared to a laboratory experiment gives us higher external validity and reflects a real-world setting to a higher degree. However, it has the disadvantage of variables being harder to control thus more difficult to verify the effects being caused by the experiment or other variables Johanson and Williamson (2018). The pre-test/post-test non-equivalent control group design would be the ideal field experiment and is what will be aimed to be done for this thesis. Having a control group as a baseline to compare the effect of the artefact will help the internal validity of the experiment and better validate the artefact itself.

3.3 Data set

In this section, the details about the data set have been provided. The data set used in this thesis was provided by Amedia, one of the largest media companies in Norway[1]. Table 3.1 provides a summary of the data sets. As can be seen, the data included details about the different campaigns and their audience behaviour, as well as the advertisement images. The data set included nineteen features and 685553 observations. The details about the campaigns include category, industry, target groups and more, and audience behaviour includes how many who saw the advertisement and how many of them clicked on them, which is detailed later in this section. It is also assumed that most of the data collected is mainly from mobile phone use but does not exclude other sources such as personal computers or tablets. Even though there are campaigns with young target age groups, 18-25, Amedia has mainly an older audience which can skew the results. Most of the advertisements Amedia's has are also not personalised, and since they have an older audience, most of the advertisements are targeted towards an older audience, meaning that they can be irrelevant to younger age groups. Amedia also has a balanced audience in terms of gender which is reflected in the analysis 4.3. This data set had to be cleaned before extensive analysis could be conducted; the preliminary analysis in section 4.1 details how this was done.

The data set included several numerical and categorical features describing the user behaviours and campaign data. A snapshot of the data set has been illustrated in table

[1] https://www.amedia.no/

Table 3.1: Data set description.

	Datatype	Number of Features			Number of Observations	Comments
		Numerical	*Categorical*	*Others*		
Provided data set	*Audience*	7	2	0	685553	Provided by Amedia [2]
	Ads Images	3	5	2	685553	
Extracted data set	Visual Features	10	0	0	132	Extracted by OpenCV [3]

3.2.

Table 3.2: Sample of Amedia data set

page_type	industry	format	hb_size	gender	age_group	cat20_maxlabel	word_count	n_obs	n_impressions _measurable	n_impressions _viewable	n_click	ctr	n_obs_total	ctr_total
contentpage	Øvrige	midtbanner	NaN	F	60-64	Kultur og underholdning	514.000000	33	33	24	1	0.030303	135631	0.000833
contentpage	Øvrige	midtbanner	320x400	F	60-64	Økonomi og næringsliv	464.555556	9	9	8	1	0.111111	98743	0.005631
homepage	Øvrige	midtbanner	NaN	M	65-69	NaN	NaN	2074	2072	1732	6	0.002893	111407	0.002208
contentpage	Øvrige	midtbanner	300x250	F	45-49	Ulykker og naturkatastrofer	71.000000	69	69	59	1	0.014493	26772	0.001046
contentpage	Dagligvare/Mat	midtbanner	300x250	M	50-54	Samferdsel	448.862745	138	138	128	1	0.007246	134405	0.002552
homepage	Øvrige	midtbanner	300x300	F	45-49	NaN	NaN	118	118	38	2	0.016949	422563	0.007275
contentpage	Øvrige	midtbanner	320x400	F	65-69	Samferdsel	434.500000	8	8	8	1	0.125000	50445	0.000773
homepage	Øvrige	midtbanner	320x400	M	65-69	NaN	NaN	66	66	62	4	0.060606	141270	0.010561
homepage	Øvrige	midtbanner	NaN	F	45-49	NaN	NaN	358	358	302	1	0.002793	53314	0.001350
contentpage	Øvrige	midtbanner	300x250	M	65-69	Politikk	821.500000	2	2	2	1	0.500000	20927	0.001099
homepage	Øvrige	midtbanner	NaN	F	55-59	NaN	NaN	889	889	740	6	0.006749	147718	0.003777
homepage	Øvrige	midtbanner	980x600	M	55-59	NaN	NaN	39	39	28	1	0.025641	29734	0.012309
homepage	Øvrige	midtbanner	320x400	M	55-59	NaN	NaN	37	37	33	1	0.027027	29734	0.012309
contentpage	Øvrige	midtbanner	300x250	M	50-54	Kriminalitet og rettsvesen	310.800000	41	41	38	1	0.024390	31370	0.002646
contentpage	Øvrige	midtbanner	300x250	M	40-44	Bolig og eiendom	841.250000	4	4	3	1	0.250000	32520	0.003137
homepage	Øvrige	midtbanner	320x250	F	60-64	NaN	NaN	131	131	116	2	0.015267	126498	0.002814
contentpage	Øvrige	midtbanner	320x480	M	70-74	Kriminalitet og rettsvesen	276.062500	36	36	32	1	0.027778	32520	0.003137
homepage	Øvrige	midtbanner	NaN	F	75+	NaN	NaN	1226	1226	1056	15	0.012235	422563	0.007275
homepage	Øvrige	midtbanner	300x300	F	65-69	NaN	NaN	83	83	63	1	0.012048	422563	0.007275
homepage	Øvrige	midtbanner	300x250	M	35-39	NaN	NaN	151	151	128	1	0.006623	27669	0.000867
homepage	Elektrisk/Brune- og hvitevarer/Lyd/Bilde	midtbanner	NaN	M	50-54	NaN	NaN	430	430	338	2	0.004651	52333	0.001051
contentpage	Øvrige	midtbanner	320x250	M	65-69	Kriminalitet og rettsvesen	1591.820000	103	103	95	2	0.019417	40132	0.001370
contentpage	Øvrige	toppbanner	980x300	F	70-74	Ulykker og naturkatastrofer	339.529412	18	18	16	1	0.055556	126498	0.002814
contentpage	Øvrige	midtbanner	320x250	M	65-69	Medisin og helse	906.714286	7	7	7	1	0.142857	98743	0.005631
homepage	Øvrige	toppbanner	980x300	M	70-74	NaN	NaN	43	43	35	1	0.023256	122238	0.001653

As it can be seen, the data set has the following features:

- *page_type:* This is whether the advertisement is on the homepage or a contentpage. Homepage is the main page all audience will first see when visiting a newspaper cite and is where they can scroll through to get to contentpages.
- *annonsornavn:* This is the name of the company running a campaign with Amedia. This feature has been cut out from examples.
- *industry:* This is what type of industry the advertisement belongs in, which is related to the company running the campaign.
- *format:* This is where the advertisement is placed on the page. For example, is toppbanner in the top section of the page, which is often one of the first things the audience can see. The five most frequent formats are midtbanner, toppbanner, netboard, skyscraper and artikkelboard.
- *hb_size:* This is the size of the advertisement image.
- *gender:* This is the target gender of the advertisement, which can be either male or female.
- *age_group:* This is the target age group of the advertisement in 5-year increments from 18-25 to 75+.
- *cat20_maxlabel:* This is what category the advertisement is and determines what different sections of content the advertisement will be shown in.

- *word_count:* This is the number of words in an article.
- *n_obs:* This is the number of observations the audience have had of the advertisement.
- *n_impressions_measurable:* This is the number of measurable impressions as defined by Google [18], meaning the system detected that the advertisement was on the screen of the audience.
- *n_impressions_viewable:* This is the number of the audience where 50% of the advertisement was visible for at least two seconds, as defined by Google [18].
- *n_click:* This is the number of the audience who clicked on the advertisement.
- *ctr:* CTR represents Click Through Rate, which is the percentage of the audience who clicked on the advertisement.
- *n_obs_total:* This is the total number of observations audience members have of the advertisement for the whole campaign.
- *ctr_total:* This is the total click through rate for the whole campaign.

Table 3.3 shows the top twenty-five rows of who got the most clicks, where most of the features has similar values such as *page_type*, *industry*, *format* and *hb_size*. Also, all of these rows are for the same campaign just for the different target groups such as *gender* and *age_group*.

Table 3.3: Top 25 rows of highest clicks

	page_type	industry	format	hb_size	gender	age_group	cat20_maxlabel	word_count	n_obs	n_impressions_measurable	n_impressions_viewable	n_click	ctr	n_obs_total	ctr_total
7137	homepage	Øvrige	midtbanner	300x250	M	65-69	NaN	NaN	11644	11641	9900	176	0.015115	678919	0.009475
7098	homepage	Øvrige	midtbanner	300x250	M	75+	NaN	NaN	7070	7069	5971	169	0.023904	678919	0.009475
7127	homepage	Øvrige	midtbanner	300x250	F	60-64	NaN	NaN	8693	8693	7308	148	0.017025	678919	0.009475
7133	homepage	Øvrige	midtbanner	300x250	M	70-74	NaN	NaN	9490	9488	8039	147	0.015490	678919	0.009475
7162	homepage	Øvrige	midtbanner	300x250	M	60-64	NaN	NaN	11240	11239	9488	146	0.012989	678919	0.009475
7106	homepage	Øvrige	midtbanner	NaN	M	75+	NaN	NaN	6894	6888	5725	142	0.020598	678919	0.009475
7200	homepage	Øvrige	midtbanner	300x250	M	50-54	NaN	NaN	13806	13803	11751	141	0.010213	678919	0.009475
7186	homepage	Øvrige	midtbanner	300x250	F	50-54	NaN	NaN	12016	12015	10185	131	0.010902	678919	0.009475
7153	homepage	Øvrige	midtbanner	NaN	M	65-69	NaN	NaN	9260	9255	7738	130	0.014039	678919	0.009475
7126	homepage	Øvrige	midtbanner	300x250	F	65-69	NaN	NaN	7573	7573	6355	129	0.017034	678919	0.009475
7154	homepage	Øvrige	midtbanner	NaN	M	70-74	NaN	NaN	8692	8685	7111	122	0.014036	678919	0.009475
7167	homepage	Øvrige	midtbanner	300x250	F	55-59	NaN	NaN	9768	9767	8297	122	0.012490	678919	0.009475
7197	homepage	Øvrige	midtbanner	300x250	M	55-59	NaN	NaN	11529	11527	9722	121	0.010495	678919	0.009475
7101	homepage	Øvrige	midtbanner	300x250	F	70-74	NaN	NaN	5013	5009	4199	116	0.023140	678919	0.009475
7138	homepage	Øvrige	midtbanner	NaN	M	60-64	NaN	NaN	7752	7746	6420	116	0.014964	678919	0.009475
7142	homepage	Øvrige	midtbanner	320x250	M	65-69	NaN	NaN	6940	6937	5828	102	0.014697	678919	0.009475
7125	homepage	Øvrige	midtbanner	NaN	F	65-69	NaN	NaN	5732	5730	4741	98	0.017097	678919	0.009475
7100	homepage	Øvrige	midtbanner	NaN	F	75+	NaN	NaN	4021	4019	3313	95	0.023626	678919	0.009475
7109	homepage	Øvrige	midtbanner	300x250	F	75+	NaN	NaN	4676	4676	3927	94	0.020103	678919	0.009475
7129	homepage	Øvrige	midtbanner	320x250	M	70-74	NaN	NaN	5632	5630	4645	93	0.016513	678919	0.009475
7222	homepage	Øvrige	midtbanner	300x250	F	45-49	NaN	NaN	10415	10413	8841	88	0.008449	678919	0.009475
7110	homepage	Øvrige	midtbanner	NaN	F	70-74	NaN	NaN	4385	4384	3622	88	0.020068	678919	0.009475
7113	homepage	Øvrige	midtbanner	320x250	F	65-69	NaN	NaN	4397	4395	3557	86	0.019559	678919	0.009475
7105	homepage	Øvrige	midtbanner	320x250	M	75+	NaN	NaN	4151	4150	3440	86	0.020718	678919	0.009475
7152	homepage	Øvrige	midtbanner	NaN	F	55-59	NaN	NaN	6036	6033	5005	85	0.014082	678919	0.009475

It is worth while to explain how Google's Active View technology [18] and how they define the Active View measurement terminology, which has been explained as so by Google:

- *Eligible impressions:* This is all impressions which successfully communicated with Google's servers associated with an Active View-enabled ad tag [18].
- *Measurable impressions:* This is all impressions which could actually be measured by the Active View technology [18]

- *Viewable impressions:* This is all the impressions, which based on MRC standards(ref) were considered viewable [18]. MRC standards shortly explained is that 50% of the advertisement image could be viewed on the screen of the audience members device [19].

As seen in figure 3.1 is an example of how Active View functions. Where the total amount of impressions is 1,000,000 of them 900,000 is eligible impressions. Of the eligible impressions, 800,000 of them are measurable impressions, of which 350,000 are Viewable impressions. Almost 44% of the measurable impressions are viewable, meaning this is how many of the audience who had the opportunity to see the advertisement on their device [18].

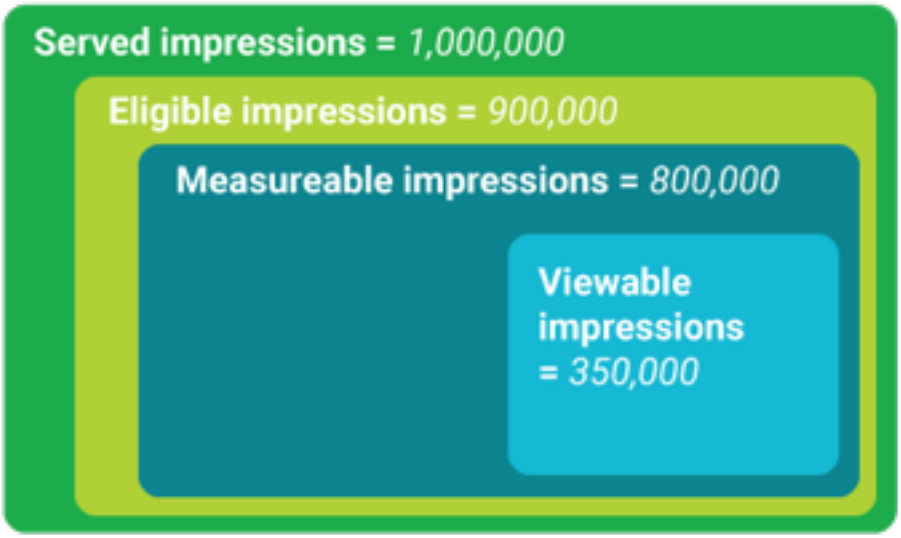

Figure 3.1: Illustration of Active View measurement by Google [18]

3.3.1 Visual Feature Extraction.

In this thesis a prototype has been created to extract a set of visual features from the images of advertisements [4]. To preform effective capturing the visual OpenCV, Scikit-Image and Scikit-Learn have been used to extract visual features from the images. The features extracted is detailed as follows:

- *Dominant colour:* This feature has been extracted through a method using a *K-Means* algorithm, provided by Scikit-Learn Python library[7]. The dominant colour method will return the most centre colour of the biggest cluster of pixels colour, by converting an image to a list of pixels and then labelling them for the *K-Means* algorithm to analyse and return the most popular label or colour [36] [5]. The colour space used for this feature has been RGB as this is the standard colour space used with monitors and the internet [23].

[4] Link to the code of the prototype: https://github.com/Dennydc007/Visual_feature_extraction_prototype

[5] https://github.com/AdamSpannbauer/iphone_app_icon

- *Red:* This feature is the red RGB colour value form the dominant colour feature
- *Green:* This feature is the green RGB colour value form the dominant colour feature
- *Blue:* This feature is the blue RGB colour value form the dominant colour feature
- *Entropy:* Entropy can indicate the amount of complexity in an image, the amount of information that is needed to be compressed[15]. The skimage entropy ranker algorithm[39] has been used to extract entropy.
- *Saturation:* Saturation can be described as the amount of grey in the colour, where a value of 0 would be grey and one would be the primary colour [4], or how colourful it is [15]. Elahi et al[15] describes how they calculated the saturation of a frame with estimation in the HSV colour space can be calculated via the RGB approximation of

$$image\_saturation = \frac{1}{N}\sum_{x,y} S_{xy}, \text{ with} \tag{3.1}$$
$$S_{xy} = max(R_{xy}, G_{xy}, B_{xy}) - min(R_{xy}, G_{xy}, B_{xy})$$

  where N is the number of pixels in an image and R_{xy}, G_{xy} and B_{xy} are the coordinates of the colour of the pixel in sRGB space [15].
- *Brightness:* Brightness: Brightness or value, V, describes the luminosity or intensity of a colour in conjunction with saturation [4] [15].
- *Sharpness:* Sharpness measures the level of details and clarity in the elements of an image. This feature is related to the brightness contrast of edges in an image [15] [35].
- *Contrast:* Contrast measures the relative difference in the colour of local features or brightness in an image. Assessment of the difference in appearance of two or more parts of a field seen simultaneously or successively is a typical definition of contrast. The root mean square contrast (RMS-contrast) is often used to compare images [15].
- *colourfulness:* colourfulness which is the measure of the individual colour distance between pixels in an image which is done by taking the mean and standard deviation, $rg = RG$ and $yb = 1/2(R+G) - B$, in the RGB colour space [15].
- *Edge_ratio:* This feature was extracted in a method using the Canny edge detection algorithm [24] and then the calculation of how much of the image is edges, x, compared to non-edges, y, is used as the visual feature in the primary extraction:

$$x/x+y$$

All of the values from the data set created by the visual feature extraction prototype can be found in table 3.4. These are values after normalisation has been applied to them for the lower deviation between values.

Table 3.4: Visual feature extraction table

	Red	Green	Blue	Entropy	Brightness	Saturation	Sharpness	Contrast	colourfulness	Edge ratio
1	0.203252	-0.321739	0.504645	-0.336713	0.965279	1.716623	3.721408	0.257873	1.987981	3.790964
2	-0.840961	2.286289	-0.381056	0.042266	0.988469	0.607161	0.005689	0.783375	1.392529	0.600926
3	0.203252	-0.321739	0.503493	-0.336713	0.964912	1.716083	3.721108	0.257664	1.987943	3.789015
4	0.064024	2.072516	-0.499888	0.093609	1.568357	1.360137	-0.605598	-0.333598	1.900355	1.141764
5	0.050101	2.115271	-0.509340	-0.008123	1.561153	0.931439	-1.055681	-0.133423	1.911671	-0.337549
6	-0.840961	2.286289	-0.381070	0.042262	0.988620	0.602516	0.005643	0.782806	1.392333	0.600772
7	0.064024	2.129522	-0.555213	-0.303422	1.926348	1.477644	-0.592149	0.109751	1.811225	1.233263
8	0.050101	2.115271	-0.509384	-0.008137	1.559901	0.931795	-1.055683	-0.133200	1.911861	-0.336999
9	0.203252	-0.321739	0.501406	-0.336688	0.965003	1.715811	3.721042	0.257706	1.988001	3.791180
10	0.077947	2.072516	-0.500022	0.093607	1.568643	1.359701	-0.605509	-0.333380	1.900432	1.145502
11	0.077947	2.414552	-0.507211	-0.350323	1.967927	1.068268	-1.051174	0.499836	2.012117	-0.205441
12	0.064024	2.129522	-0.035333	0.405256	0.967579	0.569732	0.399310	0.501012	1.796830	0.801925
13	0.050101	2.115271	-0.768681	0.557743	2.629426	-1.348845	-1.109901	-4.529027	0.485987	-1.802449
14	0.593092	-0.578266	1.183381	0.660723	0.435138	-0.094945	-0.738988	1.098516	2.150056	-0.231664
15	0.356404	-0.549763	2.265836	0.910255	0.116894	-0.211521	1.589620	0.940561	1.964644	0.553973
16	0.579169	-0.592518	1.320524	0.525021	0.466421	-0.203789	0.270247	1.186397	2.086773	-0.115748
17	0.064024	2.101019	-0.749432	0.553348	2.214182	-0.661023	-0.955138	-2.166050	1.566617	-0.261909
18	0.050101	2.101019	-0.747840	0.547211	2.175861	-0.622819	-0.929590	-2.028408	1.643460	-0.163868
19	0.676629	-0.649524	-0.788244	1.445601	-1.256691	-1.938154	0.005463	-1.996832	-1.321091	-0.921242
20	0.676629	-0.649524	-0.788244	1.445601	-1.256691	-1.938154	0.005463	-1.996832	-1.321091	-0.921242
21	0.648783	-0.649524	-0.781694	0.691486	-1.058003	-0.967735	0.470095	0.363663	-1.195999	-0.093504
22	0.648783	-0.649524	-0.784327	0.717162	-1.078151	-1.088288	0.393426	0.516924	-1.150049	-0.359609
23	0.453864	-0.392996	0.007222	0.471772	0.228501	0.316931	-0.896100	-0.320555	0.718134	-0.331490
24	0.453864	-0.392996	-0.025334	0.427779	0.213948	0.327405	-0.892682	-0.310783	0.644149	-0.083328
25	0.398172	-0.364493	0.025086	0.237779	0.458414	0.538697	-0.874543	-0.159394	0.850416	0.033659

3.4 Models and Tools

In this thesis, a range of tools and models have been used. First of all, I have implemented this thesis project using Python language. For the feature extraction OpenCV have been used in combination with Scikit-Learn and Scikit-Image. For data analysis a machine learning algorithm called CatBoost have been used, which use gradient boosting on decision trees. For further analysis of the data SHAP was used to explain what the CatBoost model had learned from analysing the data.

3.4.1 OpenCV

OpenCV is an open-source computer vision library with interfaces for C++, Python, Java, and MATLAB and supports Windows, Linux, Android and Mac OS[27]. The aim of OpenCV is to provide a common infrastructure for computer vision applications, which can accelerate the use of machine perception for open use. OpenCV has over 2500 optimized algorithms, which can be used for many different computer vision tasks such as recognition, identification, tracking and so on [27]. In this thesis OpenCV has been used for tools such as colour space conversions from RGB to HSV or to grey-scale, or detection of edges or sharpness as previously described. OpenCV together with KMeans algorithm from Scikit-Learn[7] is what made up the dominant colour feature extraction. From scikit-Image library, which is an open-source image processing library for Python similar to OpenCV but with more focus on research and education [39], the entropy ranker was utilised, as previously mentioned.

3.4.2 Decision Trees

Decision Trees can be seen as tools that can support the decision making processes. This is done by breaking up the process into possible options and the outcomes of these options. The different options are called branches and their final outcomes are called leaf nodes [22].

Decision trees as a machine learning model works on the same principals. The main difference is that a machine learning model uses features from a data set in either a classification problem or regression problem. The decision tree splits from the root node out in branches based on how it can most efficiently split the data set based on conditions of the features. Then entropy and information gain are calculated based on the feature. The decision tree is then split based on the feature with the highest information gain [10, 34]. One of the main problems with decision trees is that they tend to be over-fitted to the training data, to which there are several possibilities to combat such as CatBoost does [13].

CatBoost

CatBoost is an open-source machine learning model which uses gradient boosting on decision trees to manage categorical features easier and faster than existing publicly available implementations of gradient boosting [13]. CatBoost can automatically process categorical features, uses different methods in dealing with them and they are considered in each iteration of the model's analysis. To overcome the bias overfitting problem other gradient boosting algorithms has, CatBoost include new trees in each iteration which has not seen any data from before, combine this with the use of generating random permutations of the training data set decreases overfitting [13].

3.4.3 Shapley Values

Shapley values is a method from coalitional game theory, which assumes that each feature value is a "player" in a game where the prediction is the pay-out [26]. Molnar [26] gives an example on how Shapley values work with a scenario of a machine learning model trained on predicting apartment prices. In the scenario, the contribution of a feature called cat-banned can be found through the coalitions of all of the other features without the cat-banned feature and we compute the predicted apartment price of each of these coalitions and take the difference to get the marginal contribution of cat-banned. The math for shapley values is presented by Molnar [26] as such:

The Shapley Value

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\}) - val(S)) \tag{3.2}$$

" where S is a subset of the features used in the model, x is the vector of feature values of the instance to be explained and p the number of features. $val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S:"[26]

$$val_x(S) = \int \hat{f}(x_1, ..., x_p) d\mathbb{P}_{x \notin S} - Ex(\hat{f}(X)) \tag{3.3}$$

"**Efficiency** The feature contributions must add up to the difference of prediction for x and the average."[26]

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - Ex(\hat{f}(X)) \tag{3.4}$$

"**Symmetry** The contributions of two feature values j and k should be the same if they contribute equally to all possible coalitions."[26] If

$$val(S \cup \{j\}) = val(S \cup \{k\}) \tag{3.5}$$

for all

$$S \subseteq \{1, ..., p\} \backslash \{j, k\} \tag{3.6}$$

then

$$\phi_j = \phi_k \tag{3.7}$$

"**Dummy** A feature j that does not change the predicted value regardless of which coalition of feature values it is added to should have a Shapley value of 0."[26] If

$$val(S \cup \{j\}) = val(S) \tag{3.8}$$

for all

$$S \subseteq \{1, ..., p\} \tag{3.9}$$

then

$$\phi_j = 0 \tag{3.10}$$

"**Additivity** For a game with combined pay-outs $val + val^+$ the respective Shapley values are as follows:"[26]

$$\phi_j + \phi_j^+ \tag{3.11}$$

Chapter 4

Experiments and Results

In this chapter, the results of the analyses performed within this thesis have been described. First, a preliminary analysis of the data has been provided in section 4.1. This includes the data exploration (with examples) and subsequent data cleaning of the Amedia data set. Then in section 4.2, correlation analysis has been explained followed by section 4.3 where the audience and campaigns data analysis is presented. Then in section 4.4 visual features analysis is presented with description of observations.

4.1 Preprocessing and Preliminary Analysis of Data

In the beginning, preprocessing of the behavioural data as well as visual features has been performed. This has been followed by a set of experiments, conducted to understand the data set better, to explore some of its characteristics, and to ensure that the feature extraction process has been conducted correctly. As an example, figure 4.1 and 4.2 shows the visual feature extraction of dominant colour feature and edge feature from two sample images. The visual features have been extracted using OpenCV, an image processing library for Python and C, and scikit-image for different algorithms. The exploration of the data set was done using libraries such as PANDAS. While exploring, the data was filtered and looked closely at, such as the different features, to better understand the data set.

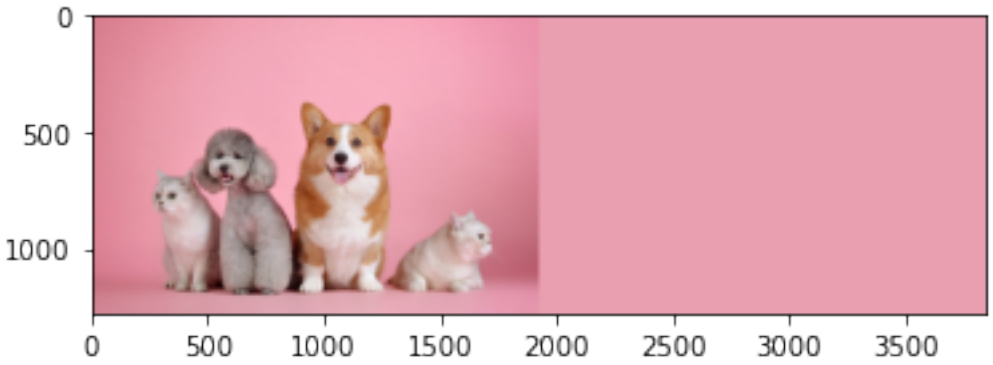

Figure 4.1: Dominant colour detection example

The dominant colour feature was extracted as defined in 3.3.1 hence, the method operates different from other potential methods such as taking the average of the colour

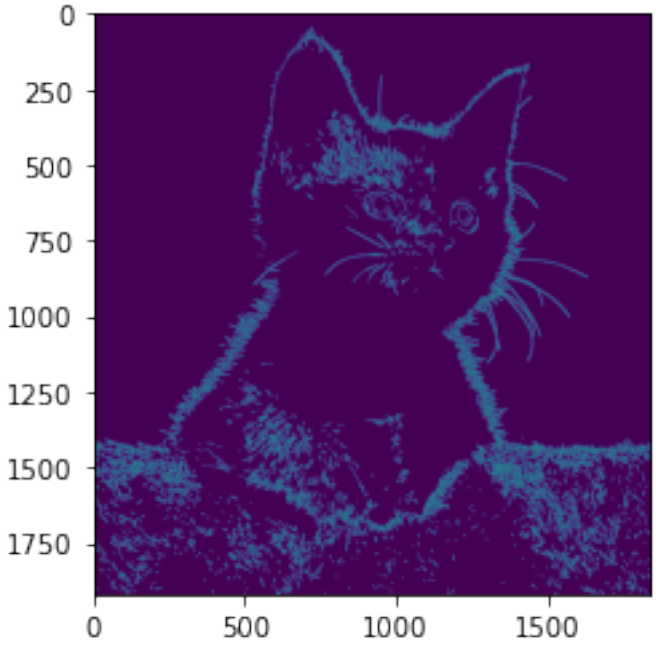

Figure 4.2: Edge detection example

in an image which will often give a different colour than expected. After the dominant colour feature has been extracted from the image, the feature is split into each of the RGB colour value, *red*, *green*, and *blue* as their own features. Companies may use a range of colours for the image they use for advertisement campaigns so that it can get easier to recognize them and their intention, and this is often a single colour which is one of the reasons it has been chosen as a feature to be extracted. As mentioned in section 2.1 colour can be used to influence our decisions, which means that the choice of colour made for advertisement is important [11, 37]. The *edge ratio* feature for which the edge detection method has been used is another visual feature that has been extracted, as shown in the example in figure 4.2.

After obtaining an initial understanding, a number of steps were taken to prepare the data set for further analysis. For example, all rows without any clicks were filtered out to get an understanding of how many of the advertisements had gotten clicks. One of the main reasons to remove all of the rows without any clicks is that since there where so many of them they could have affected the machine learning model substantially. We are more interested in what made audience click rather then what they did not click on. As noted in section 3.3, the data set included nineteen features and 685553 observations. Not all of the features in the data set would be useful for further analysis such as the IDs of advertisement campaign and their corresponding image ID. These IDs had heavy influence during early analysis, while not contributing to the prediction other than showing what campaign was doing well in terms of number of clicks which was not interesting for this thesis. Other features where missing from some of the observations which had to be addressed by giving them an -999 value so that they would have minimal influence on the analysis. Another minor issue was adding CTR to the right observations in the visual features data set.

After that, a close look at the provided data has revealed interesting findings. For example, it was figured out that many advertisements have not received any click! Ac-

cording to the data, the number of observation for such advertisements is 4989 and the total amount of clicks these got was 16213 which ranges from 1 to 150. One of the reasons that so many of the advertisements did not get any click could be because of the brief time period the data was collected.

Checking the advertisements that received the largest number of clicks, it is observed that almost all the advertisements are from the same company and most of them are on the homepage. The age groups in these advertisements were older than 40, here we must consider the possibility of people, particularly older people, unintentionally clicking the advertisements while scrolling through the homepage. The differences between genders, male and female, were also closely looked at. However, no meaningful difference between them could be found in the data. The lack of differences in the interest of different genders reveals that, at least in this particular data, genders have nearly equal interests. After conducting further careful comparison, the findings showed slight differences between the age groups and no particular differences in genders.

For some features, a further analysis has been necessary to obtain better understanding of them. An example is the feature *n_impressions_measurable* that refers to a subset of all eligible impressions that can actually be measured by the system which was described in section 3.3. This may indicate that the system has actually detected that the advertisement was on the screen of the audience. After looking closer at the data, around 684573 observations collected form campaign got measurable impressions. I have further looked at the number of clicks to viewable impressions, indicating the number of the audience where 50% of the advertisement viewed for at least two seconds. When looking at viewable data with 623885 observations, there are still many of the advertisements which got viewable impressions. However, this only indicates if the advertisement received any viewable impressions from the audience and does not say much about how many.

4.2 Experiment A: Correlation Analysis

In this section, the results for correlations analysis have been provided. Initial experiments have focused on exploratory data analyses. First, investigation of the potential correlation between visual features extracted from advertisement images and the clicking behaviour of the audience. Hence, a scatter plot of the visual features was created in order to understand the correlation between visual features and a target feature, i.e., Click Through Rate (CTR). The plot can be seen in figure 4.3.

As a starting point, I checked the potential correlations of the features with each other. From the scatter plot (figure 4.3) there are different groups of features which have some interesting correlations with each other. First group of features are *brightness*, *colourfulness*, *green* and *blue*. Checking *brightness* and *colourfulness*, seems they are positively correlated. This is an expected results and indicate that advertisement images with higher *colourfulness* are also brighter. Comparing *green* with *brighness*, and *green* with *colourfulness*, a slight positive correlation can be observed. This is while *blue* reflects a slightly lower but negative correlation with all three of them (*green*, *brighness*, and *colourfulness*). This means that images with greener dominant colour are brighter and more colourful while the opposite is true for dominant colours which is bluer.

The features in this group (overall) do not indicate a considerable and meaningful

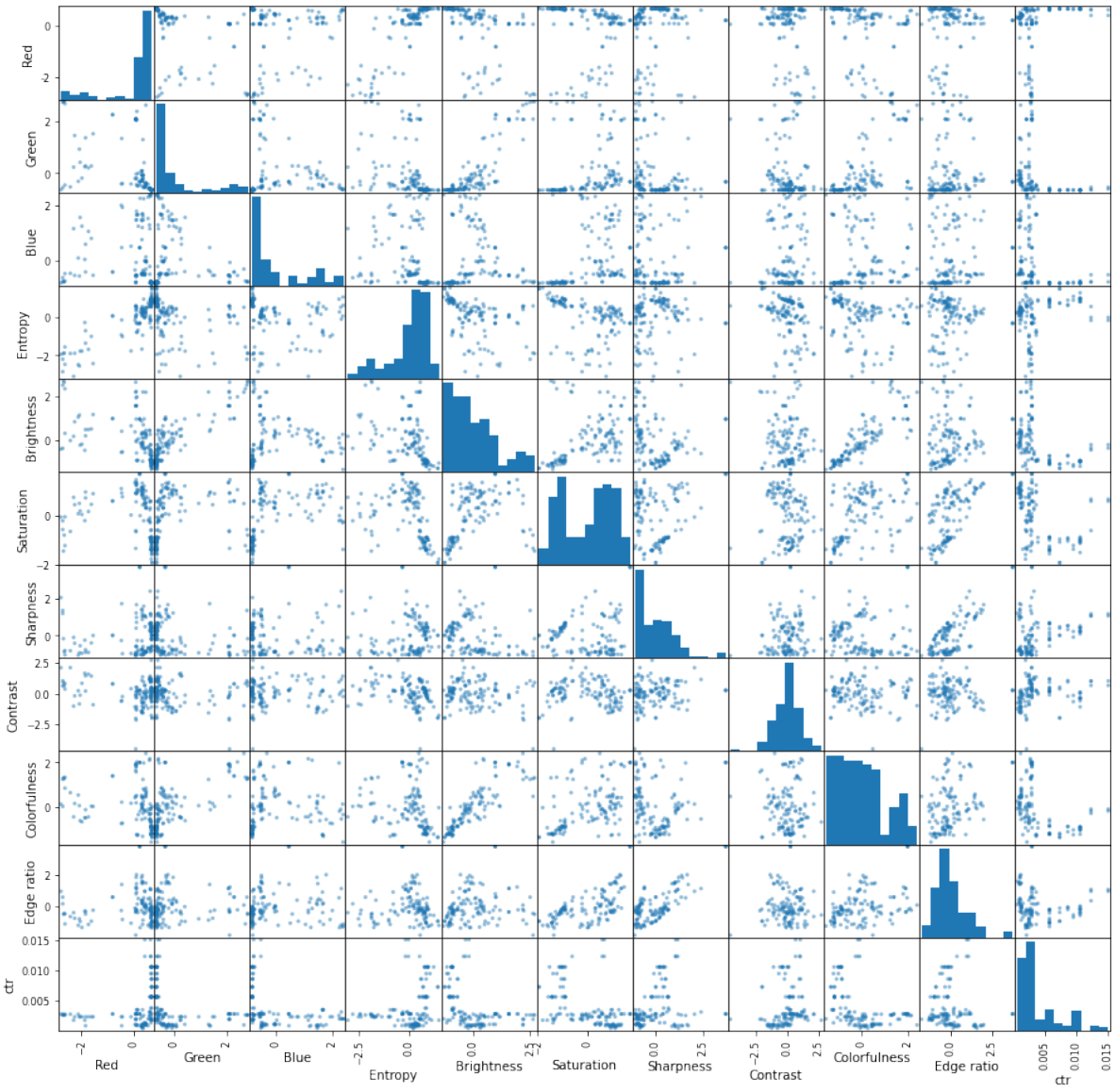

Figure 4.3: Scatter plot of visual features analysis

correlation with CTR. However, there might still be indirect effect of the colours with CTR but this seems to be hard to observe with this sample of data. A larger sample may reveal such correlation.

Another group of features is *red*, *saturation* and *edge ratio* which show slight positive correlation with each other. This means that images with a redder dominant colour are also more saturated and have a higher *edge ratio*. Similar to *green* and *blue*, *red* colour does not show any major and meaningful correlation with CTR. However, there can again be indirect relations by other features such as *saturation* and *edge ratio* which show some positive correlation with CTR.

Interestingly, *contrast* does not show much of correlation with other features. This results of the analysis may indicate that the *contrast* for different images of advertisement in this data set varies without necessarily revealing meaningful information about how other features (including CTR) vary.

4.3 Experiment B: Feature analysis - Audience and Campaign Data

The previous analysis has focused on visual features. In the next analyses, an exploratory analysis of the audience and campaign data has been performed using the *CatBoost* machine learning model, introduced earlier in section 3.4.

After preprocessing the data, a CatBoost model has been built to perform a further in-depth analysis of the data set. The model is called *AD model*. In order to find the optimal model setting, different values have been experimented with different parameters such as learning rate, number of iterations, etc. However, after a few checks of different CatBoost models, the default settings has been found to be sufficient fine for the purpose of this thesis. Moreover, CatBoost can be used for different goals, including making predictions for a target features, measuring the learning performance of the model, or even computing the feature importance scores. In addition to former tasks, the latter task has been particularly interesting for me to investigate.

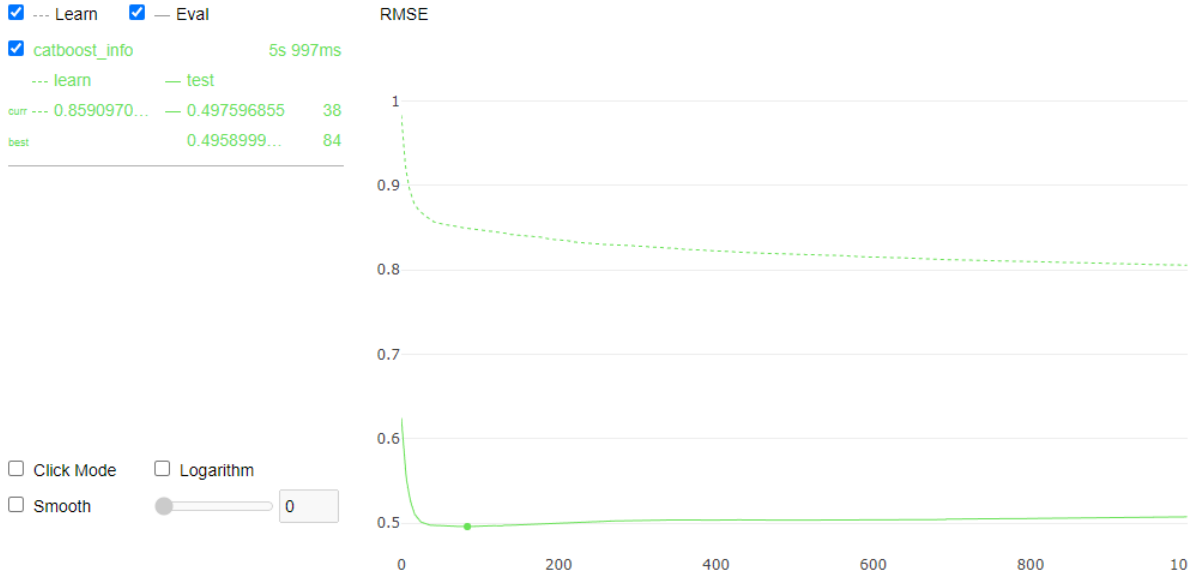

Figure 4.4: AD model

Figure 4.4 represents the values for the prediction error, in terms of Root Mean Square Error (RMSE), computed by the AD model. Looking at the evaluation curves, it can be seen the curves gets flattens out around iteration one hundred, which means that the model could not improve the predictions after this iteration. A reduction in computing power and time could have taken place by reducing the number of iterations the model would go through. According to the values, the model could perform the prediction around 50% accuracy.

Again, an analysis focused on computing the feature importance scores can be interesting as it would reveal information about the characteristics of the audience. The task here is click prediction and the target feature is the number of clicks in the data set, which is an important factor for the *industry* players when evaluating the advertising campaigns. Table 4.1 presents the results of this analysis. As it can be seen, the most important feature with a score of 29.543 in predicting number of clicks is the feature *n_impressions_viewable*, indicative of the audience who has been able to see 50%

Table 4.1: AD model feature importance score

Feature	Feature Importance Score
n_impressions_viewable	29.543
age_group	17.176
hb_size	14.832
n_obs	9.918
n_impressions_measurable	9.684
gender	5.397
industry	4.740
page_type	3.851
format	2.287
cat20_maxlabel	1.168
word_count	1.126
n_content ids	0.272

of the advertisement for at least 2 seconds before they clicked it. The least important feature has been *word_count* and *n_content_ids*.

In addition to the previous analysis, I have also performed SHAP analysis that can provide explanation for the CatBoost models. The results are plotted in figure 4.5.

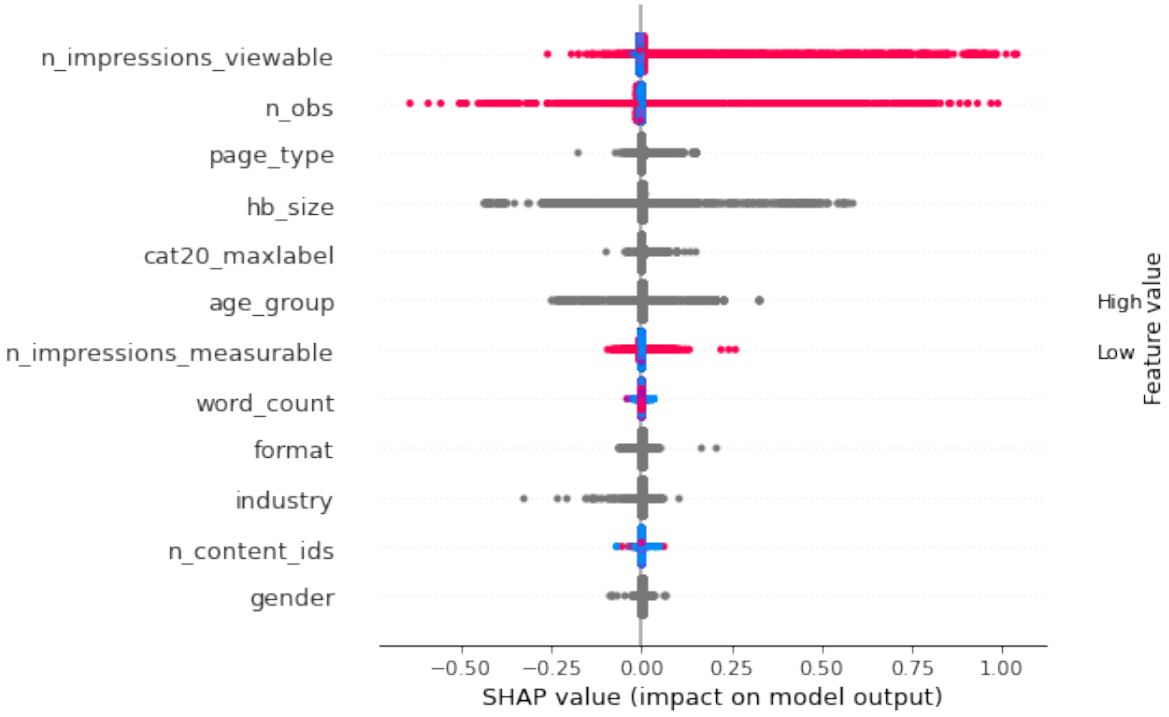

Figure 4.5: SHAP features corresponding

Starting from the top with is *n_impressions_viewable* which is the most impactful feature according to the SHAP analysis. Below *n_impression_viewable* are the rest of the features in descending ranked order according to their impact on the model. In addition to the ranking of each feature, each individual data point, represented as a coloured dot, for each feature is shown with how they impact the prediction of the

model and if they have a positive- (red dots), negative- (blue dots) or undefined-value (grey dots). The advantage of knowing the values of each of the data points is that it is possible to see what impact a feature has when it is positive or negative which can give a better indication of what increases or decreases click through rate. There are some features which have grey dots because these are categorical features and not numerical features. Unfortunately, the summary plot of the AD model in figure 4.5 is not very informative as the one for the VF model shown in figure 4.8, in section 4.4.

Looking at the results (i.e., red and blue dots on top of the the figure) for *n_impression_viewable* feature, almost all of them are red, perhaps since this feature is either positive or zero in the data set. Indeed, the blue dots are most likely seen for features with data points where the values are zero and for some reason SHAP has interpreted them as negative values. The distribution of the dots for *n_impression* viewable shows that most of them has a positive impact with some having negative and some having no impact.

A force plot which shows an example of a prediction of the model, can give us a better picture of how the different features affected the models predictions. Figure 4.6 shows an example of a prediction of the AD model, the red colour shows features which have pushed the model in a positive direction while the blue colour pushes the model in a negative direction. In figure 4.6 we can see that when *age_group* has the value 65-69, it pushes the model to have a higher prediction value. The same is for *hb_size* when it has the value 320x250, but *n_impression_viewable* pushes the model in a negative direction when it has a value of 1 and the rest of the features also pushes the model in a negative direction.

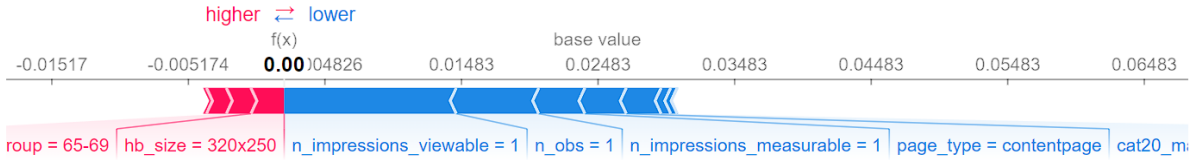

Figure 4.6: Force plot, detailing how the AD model did a single prediction

After getting an initial explanation we can check the second most important indicator for click prediction task. This can be seen in the results introduced before in figure 4.1. The second feature in terms of importance scores is *age_group* with a score of 17.176 which is also supported by earlier analysis and the speculation of older age groups unintentionally clicking ads described in section 4.1. But in the SHAP summary plot (figure 4.5) its ranked number six is surprising considering earlier analysis showing that older age groups are more likely to click on advertisements. Why *age_group* is having less of an impact according to SHAP will need further research. Looking at the distribution of *age_group* it falls in between *cat20_maxlabel*, having more impact in both directions than *cat20_maxlabel*, and *hb_size*, having less impact in both directions than *hb_size*.

hb_size comes in as third most important feature with 14.832 which like *age_group* supports the speculations of older age groups unintentionally clicking ads. The reason for why *hb_size* is so high can be because a lot of the high click rows had the same or similar *hb_size* which can be because of the specific *format* midbanner. Sim-

ilarly, *hb_size* is the fourth most impactful feature in the SHAP summary plot (figure 4.5), and earlier analysis showed that there could be a correlation between *hb_size* and *page_type*. This correlation is between homepage and the fact that a lot of the high click data had the same or very similar *hb_size* ranging from 300x250 to 320x250 and can future support the theory of unintentional clicking while scrolling. The distribution of *page_type* is similar to *n_obs* but less impactful in both positive and negative.

Ranked as fourth is *n_obs* with a score of 9.918, but is the second highest ranked feature in the SHAP summary plot. *n_obs* is a very similar feature to *n_impressions_viewable*, the difference being *n_obs* includes all possible impressions the audience could have had of the advertisement. Further down is also *n_impressions_measurable* which is also very similar to both *n_impressions_viewable* and *n_obs* but is in between them being all the impressions which could be measured by the system, meaning the advertisement was registered on the screen of the audience. Interestingly the model favoured *n_impressions_viewable* and *n_obs* and not *n_impressions_measurable*, even though all three are very similar features, which could just be because of how decision tree machine learning models work even if CatBoost have means of combating bias under the learning process. It is possible since all three are so similar that *n_impressions_measurable* did not have any other contribution after *n_impressions_viewable* and *n_obs* had contributed to the output of the CatBoost model which caused it to have less impact than them. Further analysis can shed light on which features are important. Continuing with the distribution of *n_obs*, it is similar to *n_impressions_viewable* but moved further towards a negative impact.

Fifth in the improtance score is *n_impressions_measurable* with and 9,684 but is the seventh ranked feature in the SHAP summary plot. *n_impressions_measurable* have already been mentioned earlier with *n_impressions_viewable* and *n_obs*. Already noted is the interesting favour the model had with the other two features and not with *n_impressions_measurable*, and the impact it has according to the SHAP figure (4.5) is as low as *page_type* and *cat20_maxlabel*. Since *n_impressions_measurable* is a numerical feature, SHAP shows if the dots have a positive or negative value, but *n_impressions_measurable* in likeness as *n_impressions_viewable* and *n_obs* only have positive values so all the blue dots likely have values of zero.

Then comes *gender* with 5.397 at sixth in the feature importance score, the very last ranked feature is *gender* according to the SHAP summary plot. This is highly surprising, especially when it was in the middle of the importance score. In the preliminary analysis sec 4.1, it was discovered that there was little difference in the interest of advertisement of the two different genders, this could of course be a limitation of the data in either too low of diversity of advertisement or too short of data collection period. The distribution of the dots is low in both directions.

After gender, comes *industry* with a feature importance score of 4.740. These being so low is surprising as these was expected to have some indication of audience behaviour. In the SHAP summary plot it is ranked tenth. *industry* was initially thought to give more information about audience behaviour and interests. This low importance of *industry* can also be a limitation of the data as with all the results. The distribution of *industry* shows the first feature where more dots have a negative impact on the model than positive.

Under *industry* is *page_type* with a score of 3.851, which is very low considering that it is ranked as the third most impactful feature in the SHAP summary plot. Since

page_type is a categorical feature all of the dots are grey which limits the usefulness of the SHAP analysis. However, *page_type* still got a high ranking, and we can see how it impacted the model. As described in section 3.3, *page_type* has two main categories which are homepage and contentpage. Observations in earlier analysis of *page_type* revealed that homepage was often in the most clicked for the *page_type* feature, this can go back to the speculation of older age groups unintentionally clicking advertisements when scrolling. Looking at the distribution of *page_type*, one can observe that it has a smaller negative and positive impact on the model compared to *n_impressions_viewable* and *n_obs*. There is a possibility that all of the negative dots for *page_type* is content page while all of the positive dots are homepage, but most likely it is a mix, this could be revealed with further analysis.

Ranked ninth by both feature importance and SHAP summary plot is *format* which is where the advertisement is placed. Initial thought hoped that this feature would give more information about audience behaviour which through all analysis has proven to not be the case. The distribution of the dots shows low impact in either direction with two outliers having higher positive impact.

The tenth ranked feature by the feature importance score is *cat20_maxlable* which describes what the content of the page is about. But it is ranked fifth by the SHAP summary plot which is interesting to see, since it is higher up than the *industry* feature because it can mean that where the advertisement is displayed is more important than what the advertisement is about. The distribution of the dots of *cat20_maxlabel* is similar to *page_type* which is not high impact in either the negative or positive direction contrasting to *n_impression_viewable* and *n_obs*.

Second last in the feature importance score and the eight ranked feature in the SHAP summary plot is *word_count* which, as the name implies, is the number of words in the article. It is interesting that *word_count* has been ranked so highly by the SHAP summary plot compared to feature importance score where it was ranked very low. However, it is a numerical feature it might have had more impact than other features. However, looking at the distribution of the dots for *word_count* shows the lowest impact in either negative or positive direction on the SHAP figure (4.5) of all the features.

The last feature in the feature importance score and the second last ranked feature in the SHAP summary plot is n content id which is only the id of the advertising campaign. The n content id feature is not necessarily useful and could have been dropped and by the ranking it got in the SHAP analysis and how low of an important score it got in sec 4.3, it could have been dropped. The distribution of the dots is similar to format, low impact in either direction.

4.4 Experiment C: Visual Feature Analysis

As with audience and campaign data in section 4.3, a CatBoost machine learning model was created and trained on the visual features data set created by the prototype mentioned in section 3.3. This model will be called VF model and as with experiment B, a feature importance analysis and SHAP explanation analysis have been conducted.

Figure 4.7 represents the values for the prediction error (in terms of RMSE-Root Mean Square Error) computed by VF model. According to the values, the model could perform the prediction with a good accuracy. Looking at the evaluation curves, it can

be seen the curves gets flattens out around iteration hundred and fifty.

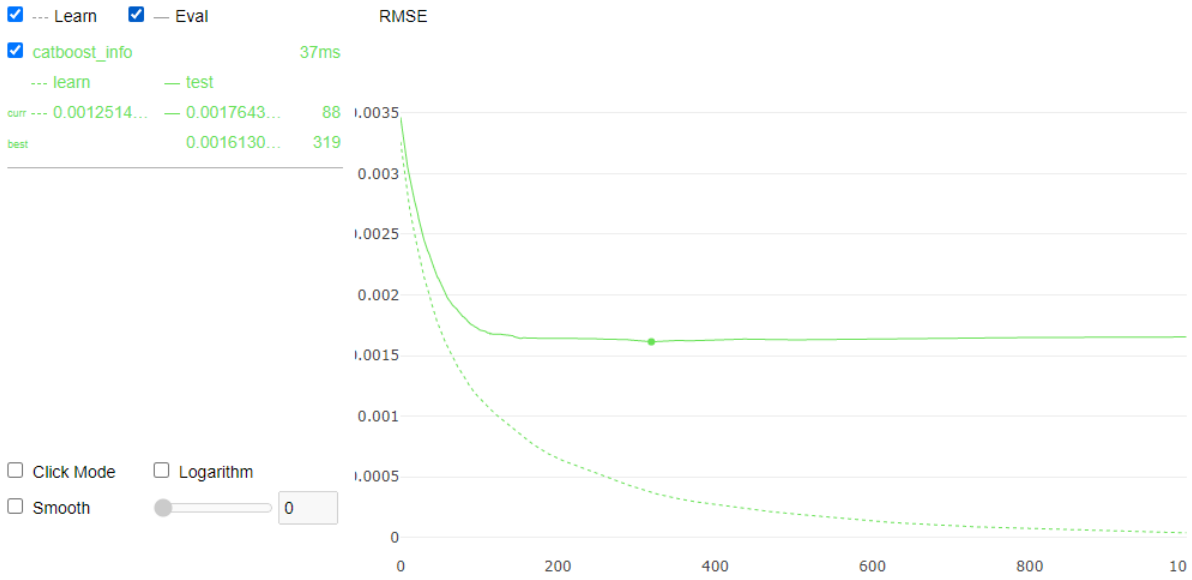

Figure 4.7: VF model of the visual features data

Table 4.2: VF model feature importance score

Feature	Feature importance score
Blue	17.511
Sharpness	13.242
Red	12.607
Colourfulness	11.987
Contrast	11.889
Edge ratio	8.436
Brightness	8.065
Green	6.999
Saturation	5.093
Entropy	4.168

The results of the feature importance score have been shown in Table 4.2. The table shows the ranking computed by the VF model for the different features for the task of predicting CTR. In figure 4.8 the SHAP summary of the VF model is shown.

According to the results of both the feature importance score and SHAP summary, *blue* is significantly more important than any other feature with a score of 17.511 in the feature importance score. This can be explained by images with more and less blue in the dominant colour being better for prediction for the model. This could be because of the differences in CTR for the different campaigns. If the advertisement of companies or *industry* with dominantly blue colour seems has higher CTR, then higher values of *blue* will mean higher CTR and vice versa. What is interesting to see in the SHAP summary here is the distribution of the dots, where none have zero impact, but all the positive impact dots are negative values while most of the negative are positive values. This result tells us that low values of *blue* in the dominant colour contributes to a higher

predicted ctr by the model which means according to this data and this model less *blue* gives more clicks on advertisement.

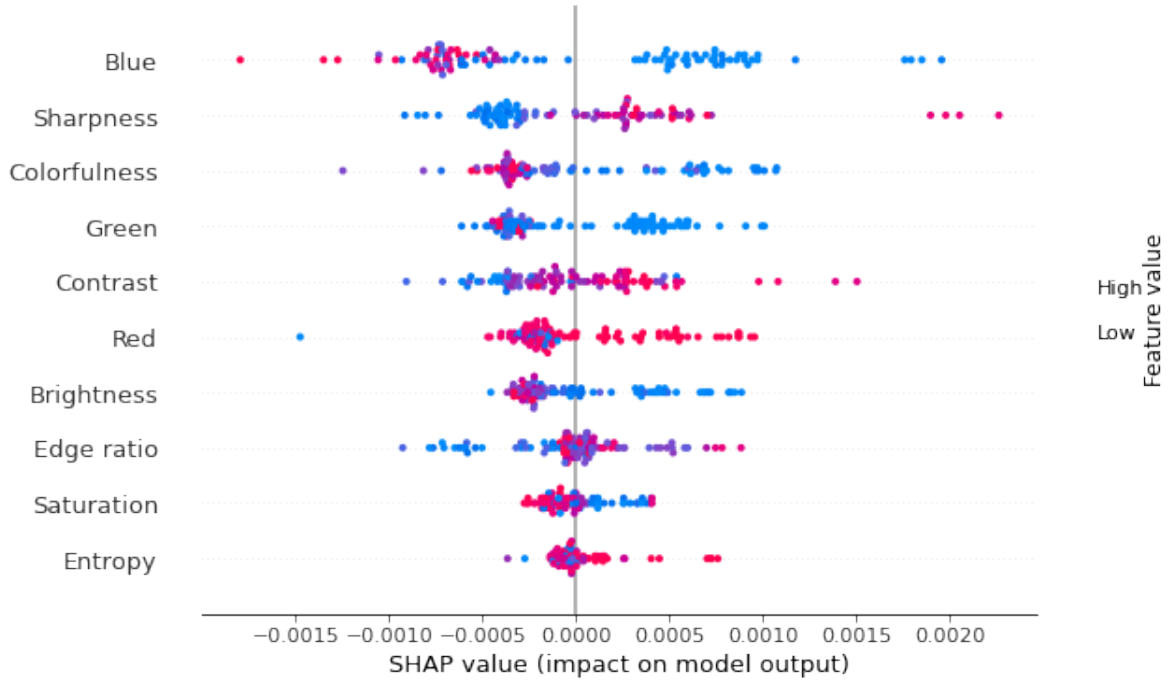

Figure 4.8: SHAP visual features corresponding

Next comes *sharpness* as ranked second and have a feature importance score of 13.242. It is important to note again that due to the collection method of the data some of the images of the advertisements were blurrier than they should have been. The ranking of *sharpness* in the SHAP analysis matches the feature importance score from sec 4.3. Looking at the distribution of *sharpness* positive values has higher positive impact while negative values have negative impact which means that having sharp images in advertisement is more effective. *red* is ranked third with a score of 12.607 in the feature importance score but is the sixth ranked feature in the SHAP summary. What is interesting is that most of the values of *red* are positive which means that from the data set, advertisements have redder dominant colour, which could be a limitation of the data set. From the distribution of the dots the most notable is that negative values have negative impact while positive values are more spread out.

colourfulness have a feature importance score of 11.987 which is the third ranked feature in the SHAP summary. The distribution of *colourfulness* is interesting since it shows that higher *colourfulness* is less effective, and that lower *colourfulness* is more effective which means less variation in colours in the advertisement can increase higher click through rate. The next feature ranked fifth is *contrast* which has the same ranking in both the feature importance score, scoring 11.889, and the SHAP summary. When looking at the distribution of the dots starting from the left, the negative side, to the right, the positive side, we can see a gradual shift from negative values of *contrast* to positive values which means that advertisements with good *contrast* has higher click through rates than those with lower *contrast* according to this data set and this model. Next comes *edge ratio* in the feature importance score with a score of 8.436 but is ranked eighth in the SHAP summary. Most of the dots for *edge ratio* are centred in the

middle meaning that they do not have much or any impact on the model, but some of the negative values have negative impact.

Ranked seventh by both feature importance score and SHAP summary is *brighness*, having a feature importance score of 8.065. From the distribution of the dots, we can see that the positive impact *brightness* has comes from negative values while all the positive values have negative impact on the model. This could mean that not having too bright images in the advertisement is better for click through rate. *green* is ranked eighth with a score of 6.999 by the feature importance score but in the SHAP summary it is ranked fourth which is contrasting. The distribution of the dots shows two clusters where one has a positive impact on the model and the other has a negative impact. The positive cluster has all negative values while the negative cluster has a mix of negative and positive values.

The second last ranked feature is *saturation* for both feature importance score, scoring 5.093, and SHAP summary. From the distribution we can see that positive values of *saturation* have negative impact and negative values have positive impact which means that having not too high *saturation* in the image of the advertisement is better for click through rate. The last ranked feature is *entropy* in both the feature importance score, with a score of 4.168, and SHAP summary. As almost all the dots are positive values, it is difficult to see any usefulness in the analysis.

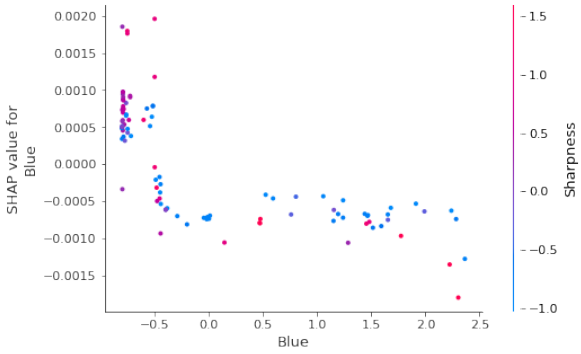

Figure 4.9: SHAP dependency plot of the feature Blue

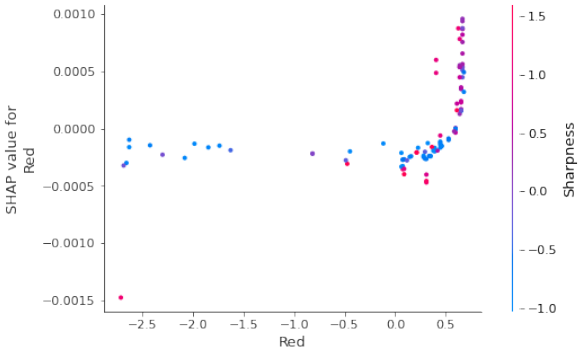

Figure 4.10: SHAP dependency plot of the feature Red

SHAP offers also deeper analysis which looks at each feature in detail in dependency plots. Two examples can be seen in figure 4.9 and 4.10, which are the dependency plots of the features *blue* and *red*. As can be seen in the figures, both features have highest interaction with the feature *sharpness*. In the dependency plots, we can see the SHAP value of the feature beeing looked at on the y axis. The actual value of the feature is on the x axis. The colour of the dots (red to blue) represent the value of the feature which it interacts with. Looking at figure 4.9, we can see that most of the higher SHAP values are negative values of *blue*, which the SHAP summary have already shown. Interestingly we can see that here are more higher values of *sharpness* when the SHAP value of *blue* is high. This means that images with lower values of *blue* are sharper. When we then look at figure 4.10, we can see the opposite happening with *red*, when an image has higher value of *red* it means that it is sharper.

Chapter 5

Discussion and Conclusion

In this chapter, a discussion on the overall results of the analyses, section 5.1, and the limitations of them is provided in section 5.2. Moreover, in section 5.3 is the conclusion of the thesis, and in section 5.4 future work are presented.

5.1 Discussion

This section summarises the contributions of the thesis and provides a discussion for the findings discussed in chapter 4, and suggests potential directions for future work. The main contributions of these thesis is summarised as follows:

- *A prototype for the feature extraction has been created using the method proposed:* In section 4.1 there are examples of how the prototype works and the code for the prototype can be found here [1].
- *A novel data set of visual features extracted from multiple different advertisement campaign together with the click through rate for each of them have been created:* Section 3.3 also shows the first twenty-five rows of the data set which was created with the prototype from the data set provided by Amedia.
- *A novel analysis method of the combination of visual features with audience behaviour data:* Section 3.3 describes the different visual features which have been extracted from the images of the advertisements and details how they were extracted.
- *An extensive analysis of the method using machine learning:* Section 4.2 details the correlation analysis and the results.
- *An extensive analysis of the machine learning model in how it predicted with the use of SHAP:* Section 4.4 details the SHAP analysis and the results.

Throughout chapter 4, several findings have been presented from the results of the experiments. As expected, *n_impressions_viewable* is the best indicator for clicks. The audience must see the advertisement in order to click it. Even though there is a high correlation with *n_impressions_viewable*, *n_obs* and *n_impressions_measurable*, but

[1] https://github.com/Dennydc007/Visual_feature_extraction_prototype

n_impressions_measurable being so much lower than the other two shows the value of the advertisement is on screen, that the audience views the advertisement. Unfortunately, because of the sparsity and division of the data set, calculating the frequency (*n_impressions_viewable* divided by *n_obs*) is not accurate, which is an important metric in adverting.

In section 3.3, it is described who the audience of Amedia is, which is an older audience. Having an older audience has skewed the analysis results as we can see that a majority of older age groups have clicked. Interestingly there is a correlation between *age_group*, *page_type*, *hb_size* and *format*, where some of the values of these features got the highest number of clicks. The values of these features have been older age groups, forty and up, homepage of *page_type*, *hb_sizes* of 300x250 and 320x250 and the *format* midtbanner, which all are for few advertisement campaigns, table 3.3 in section 3.3 shows the top twenty-five of these where all rows where from the same campaign.

At first glance, these findings tell us that these values for these features perform well, but this might not be the case. It can be argued that these findings might be because of unintentional clicking on the advertisements while scrolling [2, 28] and looking closer at the data, there is no other combination of feature values which has gotten as many clicks as these have.

Another reason for the consideration of unintentional clicking is the fact that it is the midtbanner which gets clicked; this is an advertisement placement which is placed in the "middle" and does not appear before the audience member has scrolled down the page a bit. It is then possible that when the advertisement appeared on their screen while they scrolled, they might have intended to continue scrolling but clicked the advertisement instead, which could be argued to be a limitation of the touch screen technology or audience error [2, 28].

Both the preliminary analysis, section 4.1, and feature analysis, section 4.3, show the importance of the features *page_type*, *hb_size* and *age_group* to the model, which is most likely because of these advertisement campaigns but might be a limitation of the data and therefore is a limitation of the model. In the preliminary analysis, there was found to be no difference in the gender's interests. However, with the findings of the possibility of unintentional clicking in mind, the data set may be limited because of that, and there could be a difference between the genders which more data or another data set could reveal.

Assuming the data set reflects reality, we can see that age is a better indicator of audience behaviour than gender, with a distinct difference in age groups and no difference between genders. As noted in section 3.3, Amedia has more advertisements targeting an older audience, and with little personalisation, it can be irrelevant to younger audiences, which can mean that younger audiences do not click advertisements as much as older audiences do.

The fact that homepage where more frequent than contentpage is a bit unfortunate but expected because audience will frequent the homepage more than they will be on contentpages, reading articles. Collecting new data which only include contentpages would give us more information about the behaviour of audience when it comes to advertising on contentpages. With a data set based only on contentpage we would be able to better analyse how context affects click behaviour. *format* and *hb_size* are very similar to each other, but *hb_size* is more important than *format* for the model (section 4.3).

hb_size is a more granular feature giving the model more information about the data set.

In section 4.2 a correlation analysis of the visual features data was presented, and some interesting correlations were observed. Elahi et al and others [15, 8, 6, 12, 32, 17] have shown that with the use of visual features they found that they could more accurately recommend movies which fit with the audience taste in movies, and this thesis has attempted to apply the same method for advertisement.

Unfortunately, an audience test was not possible and would have been the next step for this thesis. However, by analysing the visual features data and combine it with the click through rate data for each of the advertisement campaigns can represent the possibility of using visual features to improve advertising and is close to how a proper audience test would have been.

When looking at the correlation analysis, some interesting correlations can be observed such as a correlation between the features *green*, *blue*, *brightness* and *colourfulness*. *green*, *brightness*, and *colourfulness* has a positive correlation while *blue* has a negative correlation with all three which does mean that images with a higher value of *green* in the dominant colour is both more colourful and brighter than images with a higher value of *blue* in the dominant colour. This correlation is just an observation of the images in the data set but in combination with CTR we can see that *brightness* and *colourfulness* have a slight correlation with CTR meaning that greener, brighter, and more colourful images get more clicks than bluer, dimmer and less colourful images. But after looking at the SHAP analysis in relation to these correlations it seems that images with less *blue*, less *green*, less *brightness* and less *colourfulness* do better as can be seen by the blue dots having positive impact on the machine learning model.

Further, the correlation analysis also showed a positive correlation between *red*, *saturation*, *sharpness* and *edge ratio* which means that images with higher values of *red* in their dominant colour is more saturated, sharper and has a higher *edge ratio* than other images. *red* and *sharpness* do not have a direct correlation with each other, but it can have an indirect correlation through the features *saturation* and *edge ratio*. *red* has the same problem as *green* and *blue* when it comes to CTR, but *saturation*, *sharpness* and *edge ratio* does have positive correlation with CTR which in turn can indirectly be affected by the correlation with *red*. By looking at the SHAP analysis in section 4.4 we can see that both *red*, *sharpness* and to some extent *edge ratio* all have positive values which has positive impact on the model.

An interesting note from the SHAP analysis (section 4.4) is that images with lower values of *green* or *blue* or higher values of *red* in their dominant colour has a higher CTR according to the model which could mean that warmer, redder colours, images get higher CTR than colder images, greener or bluer colours. This could also just be a result of the data set and that other data could have different results. Another interesting part of the correlation analysis is that contrast did not have any meaningful correlation with any other feature, but it does have a majority of the positive values with positive impact on the model according to the SHAP analysis.

5.2 Limitations

As already discussed in this section (i.e., section 5.1) there have been several limitations of this work in terms of the data set and method. The data set has had some limitations in terms of how it was collected, images being blurry, missing parts or corrupt and the short time period the data of audience behaviour was collected. The short time period of data collection may partially represent the audience's behaviour.

Another limitation of the data set is one of the findings the preliminary analysis revealed which was the possibility of unintentional clicks of particularly older age groups, a longer time period could give a better view of this possibility. The data could also have been analysed with other analysis methods which could have revealed other information about it.

As mentioned, using the visual features extraction method to improve advertisement personalistion could have been evaluated further in a user study to more accurately understand its effectiveness usefulness. The task could have been formulated in the context of recommendation. There was also an aim in this thesis to see if some contextual information could be used to help improve personalisation and recommendation. However, as the data revealed, there could more predictive power from features such as category or industry, but not observed in the data. This could also be a limitation of the thesis and the analysed (majority of clicks happened on homepage), or the method used in this thesis. Further research on the usage of contextual advertisement in combination with the use of visual features is still needed to get more promising results in improving advertising.

5.3 Conclusion

In conclusion, this thesis has addressed the challenge of personalization in advertisement domain by proposing a technique to visually analyse images of advertisements and obtain more information from them. A industry data set has been provided by Amedia, one of the largest media companies in Norway [2]. The data set has been thoroughly analysed (visually) and Machine Learning models have been trained on a set of extracted visual features. The features have been useful to obtain an understanding of the images of advertisement and to predict certain target features (e.g. CTR). Such feature are widely used by industry as indicators of the behaviours of audience online.

Results have shown that visual features can relatively be useful and predictive of audience behaviour. This shows that the use of visual features can possibly improve personalization of advertisement. There were also many different findings of the data set and the data set created by the visual features extraction prototype.

5.4 Future Work

Several extensions can be considered for further work. First of all, a larger data set collected over a longer period could give more results of audience behaviour. Using other

[2] https://www.amedia.no/

methods of analysis could also reveal more about the existing data set. It is also possible to transfer this method to other fields as this has already been transfer from the domain of movie recommendation [15]. There is also the point of contextual advertisement where personalization is done by assuming the audience would be interested in products which is relevant to the content they are viewing. Creating a new data set which only includes content page as suggested in the discussion could give us more information about context.

Chapter 6

Appendix A

Visual features extraction code

```
import cv2
import numpy as np
import scipy
from scipy import spatial
import imageio
import pickle
import random
import os
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from collections import Counter
import pandas as pd
from skimage import data
from skimage.util import img_as_ubyte
from skimage.filters.rank import entropy
from skimage.morphology import disk

# Calculates brightness by splitting HSV color space into
# hue, saturation, and value. The value is synonymous with
brightness.
def get_brightness(image):
    hsv = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
    #cv2.imshow('Image', hsv)
    _, _, v = cv2.split(hsv)
    sum = np.sum(v, dtype=np.float32)
    num_of_pixels = v.shape[0] * v.shape[1]
    return (sum * 100.0) / (num_of_pixels * 255.0)

# Calculates saturation by splitting HSV color space into
# hue, saturation, and value. Saturation is extracted and represents
# saturation
def get_saturation(image):
    hsv = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
```

```
    #cv2.imshow('Image', hsv)
    _, s, _ = cv2.split(hsv)
    sum = np.sum(s, dtype = np.float32)
    num_of_pixels = s.shape[0] * s.shape[1]
    return (sum * 100.0) / (num_of_pixels * 255.0)

# Calculates entropy
def get_entropy(image):
    gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    entropy_img = entropy(gray,disk(5))
    all_sum = np.sum(entropy_img, dtype = np.float32)
    num_of_pixels = entropy_img.shape[0] * entropy_img.shape[1]
    return all_sum / num_of_pixels

# Calculates image sharpness by the variance of the Laplacian
def get_sharpness(image):
    img2gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    return cv2.Laplacian(img2gray, cv2.CV_64F).var()

# Return contrast (RMS contrast)
def get_contrast(image):
    img_gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    return img_gray.std()

def get_colorfulness(image):
# split the image into its respective RGB components
(B, G, R) = cv2.split(image.astype("float"))
# compute rg = R - G
rg = np.absolute(R - G)
# compute yb = 0.5 * (R + G) - B
yb = np.absolute(0.5 * (R + G) - B)
# compute the mean and standard deviation of both `rg` and `yb`
(rbMean, rbStd) = (np.mean(rg), np.std(rg))
(ybMean, ybStd) = (np.mean(yb), np.std(yb))
# combine the mean and standard deviations
stdRoot = np.sqrt((rbStd ** 2) + (ybStd ** 2))
meanRoot = np.sqrt((rbMean ** 2) + (ybMean ** 2))
# derive the "saturation" metric and return it
return stdRoot + (0.3 * meanRoot)

# Gotten from (ref)
def get_dominant_color(image_path, k=4, image_processing_size = None):
    image = cv2.imread(image_path)
    """
    takes an image as input
    returns the dominant color of the image as a list
```

```
    dominant color is found by running k means on the
    pixels & returning the centroid of the largest cluster

    processing time is sped up by working with a smaller image;
    this resizing can be done with the image_processing_size param
    which takes a tuple of image dims as input

    >>> get_dominant_color(my_image, k=4, image_processing_size = (25, 25))
    [56.2423442, 34.0834233, 70.1234123]
    """
    #resize image if new dims provided
    if image_processing_size is not None:
        image = cv2.resize(image, image_processing_size,
                           interpolation = cv2.INTER_AREA)

    hsv_image = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)

    #reshape the image to be a list of pixels
    hsv_image = hsv_image.reshape((hsv_image.shape[0] *
    hsv_image.shape[1], 3))

    #cluster and assign labels to the pixels
    clt = KMeans(n_clusters = k)
    labels = clt.fit_predict(hsv_image)

    #count labels to find most popular
    label_counts = Counter(labels)

    #subset out most popular centroid
    dominant_color = clt.cluster_centers_[label_counts.most_common(1)[0][0]]

    return list(dominant_color)

def get_edges(image, threshold_lower=100, threshold_upper=200):

    # converts color to gray scale
    img_gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)

    # noise reduction by a Gussian blur
    img_blur = cv2.GaussianBlur(img_gray, (3,3), 0)

    # mounts Canny edge detection algorithm. Threshold can be tuned to
    include more or less edges depending on the image.
    edges = cv2.Canny(image=img_blur,
    threshold1=threshold_lower, threshold2=threshold_upper)

    return edge_ratio(edges)
```

```
def edge_ratio(edge_array):
    # counts the pixels which has either a value of 255  in edges or 0 in
    empty
    edges = 0
    empty = 0

    for row in edge_array:

        for pixel in row:

            if pixel == 255:
                edges += 1
            else:
                empty += 1

    return edges / (edges + empty)

def batch_extractor(images_path, pickled_db_path="dominant_color.pck"):
    files = [os.path.join(images_path, p) for p in
    sorted(os.listdir(images_path))]

    result = {}
    for f in files:
        print('Extracting features from image %s' % f)
        name = f.split('/')[-1].lower()
        image = cv2.imread(f)
        name = name.replace('.png', '')

        # Dominant color
        result_dom_color = get_dominant_color(f)
        color_change = result_dom_color
        color_change[0], color_change[-1] = color_change[-1], color_change[0]
        for i in range(0, len(color_change) - 1):
            color_change[i] = round(color_change[i])

        brightness = get_brightness(image)

        saturation = get_saturation(image)

        entropy = get_entropy(image)

        sharpness = get_sharpness(image)

        contrast = get_contrast(image)

        colorfulness = get_colorfulness(image)
```

```
        edge = get_edges(image)

        result[name] = [color_change[0], color_change[1], color_change[2],
        brightness, saturation, entropy, sharpness, contrast, colorfulness,
        edge]

    # saving all our feature vectors in pickled file
    with open(pickled_db_path, 'wb') as fp:
        pickle.dump(result, fp)

    return result

# rename column names
df = pd.DataFrame.from_dict(data=colors, orient='index', columns=['Red',
'Green', 'Blue', 'Entropy', 'Brightness', 'Saturation', 'Sharpness',
'Contrast', 'Colorfulness', 'Edge ratio'])

# normalization of data-frame
normalized_df=(df-df.mean())/df.std()
```

Bibliography

[1] Apple. Clear the history and cookies from safari on your iphone, ipad, or ipod touch. https://support.apple.com/en-us/HT201265, 2021. Accessed: 31.05.2022. 1.1, 2.2

[2] S. Baker, J. Waycott, S. Pedell, T. Hoang, and E. Ozanne. Older people and social participation: From touch-screens to virtual realities. In *Proceedings of the International Symposium on Interactive Technology and Ageing Populations*, ITAP '16, page 3443, New York, NY, USA, 2016. Association for Computing Machinery. 2.2, 5.1

[3] F. Bakhshandegan Moghaddam, M. Elahi, R. Hosseini, C. Trattner, and M. Tkali. Predicting movie popularity and ratings with visual features. page 6, 06 2019. (document), 2.1, 2.3

[4] J. H. Bear. The hvs color model in graphic design. https://www.lifewire.com/what-is-hsv-in-design-1078068, 2020. Accessed: 22.10.2021. 3.3.1, 3.3.1

[5] A. Beheshti, S. Ghodratnama, M. Elahi, and H. Farhood. *Social Data Analytics*. CRC Press, 2022. 2.3

[6] K. Bougiatiotis and T. Giannakopoulos. Enhanced movie content similarity based on textual, auditory and visual information. *Expert Systems with Applications*, 96:86–102, 2018. 2.1, 2.3, 5.1

[7] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013. 3.3.1, 3.4.1

[8] P. Cermonesi, Y. Deldjoo, M. Elahi, F. Garzotto, P. Piazzolla, and M. Quadrana. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, 5:99–113, 2016. 2.1, 2.3, 5.1

[9] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 417426, New York, NY, USA, 2008. Association for Computing Machinery. 2.2

[10] N. S. Chauhan. Decision tree algorithm, explained. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html, 2022. Accessed: 30.12.2022. 3.4.2

[11] K. Cherry. Color psychology: Does it affect how you feel? how colors impact moods, feelings, and behaviors. https://www.verywellmind.com/color-psychology-2795824#toc-what-is-color-psychology, 2020. Accessed: 29.05.2022. 2.1, 4.1

[12] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi. Recommender systems leveraging multimedia content. *ACM Comput. Surv.*, 53(5), sep 2020. 2.1, 2.3, 5.1

[13] A. V. Dorogush, V. Ershov, and A. Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, page 7, 2018. 3.4.2, 3.4.2

[14] M. Elahi, M. Braunhofer, T. Gurbanov, and F. Ricci. User preference elicitation, rating sparsity and cold start., 2018. 2.3

[15] M. Elahi, R. Hosseini, M. H. Rimaz, F. B. Moghaddam, and C. Trattner. *Visually-Aware Video Recommendation in the Cold Start*, page 225229. Association for Computing Machinery, New York, NY, USA, 2020. 1.1, 2.1, 2.3, 3.3.1, 3.3.1, 5.1, 5.4

[16] A. Felfernig, G. Friedrich, D. Jannach, and M. Zanker. *Recommender system an introduction*. Cambridge University Press, 32 Avenue of the Americas, New York, 2011. 2.3

[17] R. J. R. Filho, J. Wehrmann, and R. C. Barros. Leveraging deep visual features for content-based movie recommender systems. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 604–611, May 2017. 2.1, 2.1, 2.3, 5.1

[18] Google. How active view metrics are calculated. https://support.google.com/admanager/answer/6233478?hl=en, 2022. Accessed: 18.05.2022. (document), 3.3, 3.3, 3.1

[19] D. Gunzerath and IAB's Emerging Innovations Task Force. Mrc viewable ad impression measurement guidelines. Technical report, Media Rating Council, 420 Lexington Avenue, Suite 343, New York, NY 10170, June 2014. 3.3

[20] N. Hazrati and M. Elahi. Addressing the new item problem in video recommender systems by incorporation of visual features with restricted boltzmann machines. *Expert Systems*, 38(3):e12645, 2021. 2.1, 2.3

[21] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004. 3.1

[22] Indeed Editorial Team. How a decision tree works (definition, guide and tips). https://au.indeed.com/career-advice/career-development/how-decision-tree-works, 2022. Accessed: 30.12.2022. 3.4.2

[23] International Electrotechnical Commission. Iec 61966-2-1:1999 multimedia systems and equipment - colour measurement and management - part 2-1: Colour management - default rgb colour space - srgb. https://webstore.iec.ch/publication/6169#additionalinfo, October 1999. Accessed: 30.05.2022. 3.3.1

[24] S. Mallick and LearnOpenCV. Edge detection using opencv. https://learnopencv.com/edge-detection-using-opencv/, 2022. Accessed: 30.08.2021. 3.3.1

[25] F. B. Moghaddam and M. Elahi. Cold start solutions for recommendation systems. In *Big Data Recommender Systems: Recent Trends and Advances*. IET, 2019. 2.3

[26] C. Molnar. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. WestendstraSSe 78, 80339 München, Germany, 2 edition, 2022. 3.4.3, 3.4.3, 3.4.3, 3.4.3, 3.4.3, 3.4.3

[27] OpenCV. About open cv. https://opencv.org/about/, 2022. Accessed: 30.05.2022. 3.4.1

[28] T. Page. Touchscreen mobile devices and older adults: a usability study. *International Journal of Human Factors and Ergonomics*, 3(1):65–85, 2014. 2.2, 5.1

[29] H. Pauzer. 71% of consumers prefer personalized ads. https://www.adlucent.com/resources/blog/71-of-consumers-prefer-personalized-ads/, 2022. Accessed: 29.05.2022. 2.2

[30] T. Quick. Advertising personalization and landing pages: The next wave in digital marketing. https://instapage.com/blog/what-is-personalized-advertising, 2022. Accessed: 29.05.2022. 1.1, 2.2

[31] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997. 2.3

[32] M. H. Rimaz, M. Elahi, F. Bakhshandegan Moghadam, C. Trattner, R. Hosseini, and M. Tkalčič. Exploring the power of visual features for the recommendation of movies. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '19, page 303308, New York, NY, USA, 2019. Association for Computing Machinery. 2.1, 5.1

[33] J. Rossignol. Apple highlights iphone's latest privacy features in new 'data auction' ad. https://www.macrumors.com/2022/05/18/apple-data-auction-iphone-privacy-ad/, 2022. Accessed: 29.05.2022. 2.2

[34] A. Roy. A dive into decision trees. https://towardsdatascience.com/a-dive-into-decision-trees-a128923c9298, 2020. Accessed: 30.12.2022. 3.4.2

[35] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 771780, New York, NY, USA, 2009. Association for Computing Machinery. 3.3.1

[36] A. Spannbauer. Finding and using images' dominant color using python opencv. https://adamspannbauer.github.io/2018/03/02/app-icon-dominant-colors/, 2018. Accessed: 30.08.2021. 3.3.1

[37] TastyAd. Color psychology in billboard advertising. https://www.tastyad.com/color-psychology-in-billboard-advertising/, 2019. Accessed: 29.05.2022. (document), 2.1, 2.2, 4.1

[38] M. Tkalcic and J. F. Tasic. *Colour spaces: perceptual, historical and applicational background*, volume 1. IEEE, 2003. 2.1

[39] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014. 3.3.1, 3.4.1

[40] K. Zhang and Z. Katona. Contextual advertising. *Marketing Science*, 31:873–1025, 2012. 2.2